ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΚΑΙ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ
ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΟΛΟΓΙΑΣ ΓΙΑ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΦΑΣΕΩΝ ΕΠΕΝΔΥΣΕΩΝ ΣΤΗΡΙΖΟΜΕΝΟΙ ΣΕ ΤΕΧΝΙΚΕΣ SENTIMENT ANALYSIS ΚΑΙ ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΥΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΝΙΚΟΛΑΟΣ ΕΜΜ. ΣΥΜΒΟΥΛΑΚΗΣ

Επιβλέπων: Βασίλειος Ασημακόπουλος
Καθηγητής, Ε.Μ.Π.

Υπεύθυνος: Ευάγγελος Σπηλιώτης
Υποψήφιος Διδάκτωρ, Ε.Μ.Π.

Αθήνα, Ιούλιος 2015

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΚΑΙ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

ΑΝΑΠΤΥΞΗ ΜΕΘΟΔΟΛΟΓΙΑΣ ΓΙΑ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΦΑΣΕΩΝ ΕΠΕΝΔΥΣΕΩΝ ΣΤΗΡΙΖΟΜΕΝΟΙ ΣΕ ΤΕΧΝΙΚΕΣ SENTIMENT ANALYSIS ΚΑΙ ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΥΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΝΙΚΟΛΑΟΣ ΕΜΜ. ΣΥΜΒΟΥΛΑΚΗΣ

Επιβλέπων: Βασίλειος Ασημακόπουλος
Καθηγητής, Ε.Μ.Π.

Υπεύθυνος: Ευάγγελος Σπηλιώτης
Υποψήφιος Διδάκτωρ, Ε.Μ.Π.

Εγκρίθηκε από την τριμελή επιτροπή την α Ιουλίου 2015

…….....	…….....	…….....
Βασίλειος Ασημακόπουλος	Ιωάννης Ψαρράς	Δημήτριος Ασκούνης
Καθηγητής, Ε.Μ.Π	Καθηγητής, Ε.Μ.Π	Επίκουρος Καθηγητής, Ε.Μ.Π

Αθήνα, Ιούλιος 2015

…….……
ΝΙΚΟΛΑΟΣ ΕΜΜ. ΣΥΜΒΟΥΛΑΚΗΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία πραγματεύεται τη μέτρηση της επίδρασης των δημοσιευμένων οικονομικών νέων στις τιμές των οικονομικών προϊόντων καθώς και τις αποφάσεις κριτικής πρόβλεψης που πρέπει να πάρει ένας επενδυτής γύρω από αυτά. Παρουσιάζεται η σχέση του κλίματος που διαμορφώνουν τόσο η οικονομική ειδησεογραφία όσο και οι απόψεις που αναμεταδίδονται στον πλανήτη γύρω από την αγορά με τις αποδόσεις των χρεογράφων.

Σκοπός της αποτελεί η ανάδειξη του προβλήματος αδυναμίας του ανθρώπου να λάβει και να επεξεργαστεί χωρίς τη βοήθεια υπολογιστικών συστημάτων όλη την απαιτούμενη ειδησεογραφία για τις επενδύσεις του κυρίως λόγω του περιορισμού του χρόνου του και κατόπιν τη πρόταση επίλυσής του.

Η λύση που προτείνεται εδράζεται στην ανάλυση της φυσικής γλώσσας που περιέχεται στα οικονομικά δημοσιεύματα μέσω προγραμμάτων επεξεργασίας του financial sentiment. Η συγκέντρωση όλων των αξιόπιστων ειδήσεων από σοβαρές οικονομικές πηγές, η κατηγοριοποίησή τους καθώς και το φιλτράρισμά τους παίζουν σοβαρό ρόλο. Ο συνδυασμός των προγραμμάτων ανάλυσης του sentiment των κειμένων με λεξικά ειδικευμένα τόσο στον τομέα των χρηματοικονομικών όσο και του οικονομικού συναισθήματος κρίνεται αναγκαία.

Εξετάζεται επίσης πειραματικά η επίπτωση της ειδησεογραφίας, επεξεργασμένης ως προς το sentiment που εκπέμπουν, στις τιμές μετοχών και βάσει αυτών λαμβάνονται αποφάσεις για ένα απλοποιημένο υποθετικό χαρτοφυλάκιο. Πραγματοποιείται σύγκριση με τις αποφάσεις που θα λαμβάνονταν αν γινόταν χρήση μόνο μαθηματικών μεθόδων πρόβλεψης (SES, Holt, ARIMA) καθώς και ο συνδυασμός των δύο.

ABSTRACT

This thesis examines the measurement of the impact that published economic news have on the prices of financial products as well as the decisions through judgmental forecasting taken by investors when driven by these news. It studies the relationship of the yields of securities with the climate shaped by both the economic news and the opinions formed about the market around the world.

The purpose is to highlight the problem of human inability to receive and process all the required news for investment, without the help of computer systems, primarily due to the limitation of time.

The solution proposed is based on the Natural Language Processing science by analyzing the contained linguistic fesatures in financial reports via processing programs of financial sentiment. The concentration of all credible news from reliable financial sources, their classification and their filtering plays a serious role. The combination of programs of sentiment analysis with specialized dictionaries of both financial and economic terminology is necessary.

We also experimentally examine the impact of the news, processed regarding the sentiment they instigate, on the prices of shares. A simplified hypothetical portfolio is used to examine the success of hypothetical decisions taken based on these prices. A comparison is then made between the aforementioned decisions based on sentiment analysis and judgmental forecasting and decisions that an investor could take if he was using only simple mathematical forecasting methods (SES, Holt, ARIMA). Finally, we examine the effect of the combination of the two methods.

Πρόλογος

Η διπλωματική αυτή εργασία εκπονήθηκε στα πλαίσια των ερευνητικών δραστηριοτήτων της Μονάδας Προβλέψεων και Στρατηγικής κατά το ακαδημαϊκό έτος 2014-2015. Η μονάδα υπάγεται στον Τομέα Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ, του Εθνικού Μετσόβιου Πολυτεχνείου.

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Βασίλη Ασημακόπουλο για τη στήριξη που μου παρείχε στην ολοκλήρωση της διπλωματικής μου εργασία. Ακόμα, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Ιωάννη Ψαρρά και τον Επίκουρο Καθηγητή κ. Δημήτριο Ασκούνη για την τιμή που μας έκαναν να συμμετάσχουν στην επιτροπή εξέτασης της εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τον Υποψήφιο Διδάκτορα κ. Ευάγγελο Σπηλιώτη για την επαφή στην οποία με έφερε με τον κλάδο των προβλέψεων και για το ιδιαίτερο ενδιαφέρον που έδειξε στην εξέλιξη, την καθοδήγηση και ολοκλήρωση της διπλωματικής μου εργασίας, όπως και τους υπόλοιπους συναδέλφους που δραστηριοποιούνται στη μονάδα.

Νικόλαος Εμμ. Συμβουλάκης
Αθήνα, Ιούλιος 2015

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ ΕΙΚΟΝΩΝ, ΠΙΝΑΚΩΝ ΚΑΙ ΓΡΑΦΗΜΑΤΩΝ

Εικόνες

Πίνακες

Γραφήματα

ΚΕΦΑΛΑΙΟ 1: ΕΥΡΕΙΑ ΠΕΡΙΛΗΨΗ

1.1 Εισαγωγή

Η χρηματοοικονομική αγορά διαθέτει ένα εξαιρετικά μεγάλο σύνολο από πληροφορίες και εργαλεία για να ανταποκριθεί στις προσδοκίες και να εκπληρώσει τους στόχους των επενδυτών και παικτών της.

Η σύγχρονη παγκοσμιοποιημένη τραπεζοοικονομική αγορά εισήλθε στον νέο αιώνα έχοντας τα εφόδια της επιστήμης των μαθηματικών, της επιστήμης των υπολογιστών και του προγραμματισμού στο πλευρό της. Οι επενδυτικές εταιρίες και οι τραπεζικοί όμιλοι επενδύουν όλο και περισσότερα χρήματα στην έρευνα ώστε οι αναλύσεις των κινήσεων των χρηματοδοτικών προϊόντων να γίνεται με όλο και περισσότερη ακρίβεια σε συνδυασμό με όλο και μεγαλύτερη ταχύτητα.

Στον τομέα αυτό έχουν επιστρατευθεί και οι επιστήμες που έχουν ως αντικείμενο την ανθρώπινη συμπεριφορά. Άλλωστε είναι ευρέως γνωστό ότι «η αγορά είναι ψυχολογία».

Το σύγχρονο παγκόσμιο σύστημα μέσων μαζικής ενημέρωσης δε θα μπορούσε να λείπει από την εξέλιξη των σύγχρονων χρηματοοικονομικών. Οι οικονομικές ειδήσεις δημοσιεύονται με ασύλληπτη ταχύτητα και σε απίστευτο όγκο σε όλα τα μήκη και πλάτη του πλανήτη και διαδίδονται σε κλάσματα δευτερολέπτου. Σε φυσική συνέχεια οι πληροφορίες αυτές διαμορφώνουν το κλίμα της αγοράς, σφυρηλατούν απόψεις και αυτές με τη σειρά τους και με την ίδια ταχύτητα αναμεταδίδονται σε όλον τον κόσμο.

Συνεπώς ένας παίκτης της επενδυτικής αγοράς χρειάζεται όλα τα οικονομικά δεδομένα ώστε να διαμορφώσει άποψη και να λάβει τις σωστές, δηλαδή τις επικερδείς, επενδυτικές αποφάσεις. Η τεχνολογία του έχει λύσει τα χέρια ως προς την ανάλυση των μαθηματικών και ποσοτικών δεδομένων. Η τεχνολογία του έχει επίσης λύσει τα χέρια ως προς την παροχή όλου του οικονομικού δημοσιογραφικού φάσματος και διαθέτει ανά δευτερόλεπτο όλη την πληροφορία ενώπιον του.

Το μόνο που δε διαθέτει είναι ο χρόνος. Για μια απλή τοποθέτηση κεφαλαίου σε μία απλώς μετοχή οι σύγχρονοι επενδυτές αναλύουν χιλιάδες συνιστώσες όπως λόγου χάρη η τάση των τιμών, το αν ιστορικά βρίσκεται σε χαμηλά ή υψηλά επίπεδα, το μέγεθος του τζίρου που διαμορφώνεται στην διαπραγμάτευση της μετοχής, τα επίπεδα των

κερδών ή των πωλήσεων της εταιρίας και άλλα πολλά. Τα δεδομένα αυτά έχουν όμως το χαρακτηριστικό ότι μπορούν σε απειροελάχιστο χρόνο να συγκεντρωθούν αυτόματα σε έναν πίνακα. Το κλίμα της αγοράς, τα συναισθήματα που περιβάλλουν τους υπόλοιπους επενδυτές για τη συγκεκριμένη μετοχή, την υποκειμενική άποψη του κάθε συμμετέχοντος στο χρηματοοικονομικό γίγνεσθαι δεν δύναται ακόμα να τοποθετηθεί σε έναν πίνακα αυτομάτως. Για αυτό το λόγο και ο εν λόγω επενδυτής μπαίνει σε μία διαδικασία να ερευνήσει την ειδησεογραφία που αφορά στην επένδυση που επιθυμεί να προβεί. Το ερώτημα που γεννάται εύλογα είναι ποιες ειδήσεις και απόψεις πρέπει να διαβάσει. Πόσες δημοσιεύσεις θα προλάβει να λάβει υπόψιν; Πόσο χρόνο τελικά μπορεί να διαθέσει;

Η παρούσα διπλωματική εργασία εντάσσεται σε ένα γενικότερο και μεγαλύτερο εγχείρημα που επιχειρεί να ποσοτικοποιήσει με έναν τρόπο τα υποκειμενικά συναισθήματα όπως αυτά αποτυπώνονται στα δημοσιευμένα νέα του οικονομικού τύπου. Ο σκοπός του εγχειρήματος είναι η δημιουργία ενός ολοκληρωμένου συστήματος το οποίο να αντλεί όλες τις πληροφορίες, ειδήσεις και απόψεις για κάποιο συγκεκριμένο χρεόγραφο ή οικονομικό προϊόν, κατόπιν να τις διατρέχει λέξη προς λέξη αναλύοντας όλες τις συνιστώσες του εγγράφου-είδησης και να εξάγει το συνολικό sentiment σε μία ποσοτικοποιημένη μορφή.

1.2 Συλλογή πληροφοριών

Η πρώτη ερώτηση που διατυπώνεται άμεσα αφορά τις πηγές της πληροφορίας. Όλοι σχεδόν οι εμπλεκόμενοι στην χρηματιστηριακή αγορά χρησιμοποιούν την πλατφόρμα του κορυφαίου ειδησεογραφικού ομίλου Bloomberg. Η πλατφόρμα αυτή έχει σχεδιαστεί ειδικά για τους επενδυτές των χρηματοπιστωτικών ιδρυμάτων και παρέχει μια τεράστια βάση δεδομένων τόσο σε επίπεδο αριθμών όσο και σε επίπεδο ειδησεογραφίας. Έχει τη δυνατότητα να παρουσιάσει στον χρήστη συγκεντρωμένα όλα τα νέα που αφορούν συγκεκριμένους επενδυτικούς στόχους, είτε πρόκειται για μετοχές, ομόλογα, παράγωγα, commodities κτλ. Συγχρόνως υπάρχουν στο Διαδίκτυο τεράστιοι συλλέκτες οικονομικών πληροφοριών (Major Financial Mass Media Web pages) οι οποίοι προσφέρουν έναν πολύ μεγάλο όγκο δεδομένων και ειδήσεων. Η απόκτηση έγκαιρων οικονομικών εγγράφων από αξιόπιστες πηγές στο Διαδίκτυο είναι ένα

κρίσιμο βήμα και είναι μια ποικιλία οικονομικών χώρων συνάθροισης ειδήσεων που παρέχουν αυτή την υπηρεσία. Φυσικά δεν πρέπει να παραλειφθούν τα μέσα κοινωνικής δικτύωσης όπως το Facebook και το Twitter τα οποία αποτελούν γιγάντιους φορείς γνώμης και για τα οικονομικά θέματα, ενώ φιλοξενούν ολόκληρες κοινότητες επενδυτών και συμμετεχόντων στις αγορές που ανταλλάζουν απόψεις και φέρουν γνώμη για όποια οικονομική κίνηση γίνεται ή πρόκειται να συμβεί. Το Διαδίκτυο όμως φιλοξενεί από μόνο του ειδικευμένες ιστοσελίδες επενδυτικής γνώμης είτε μεμονωμένων bloggers είτε ολόκληρα forums που ασχολούνται με τις κινήσεις των χρεογράφων, αξιολογούν επενδύσεις, διατυπώνουν γνώμη και προσφέρουν προβλέψεις για το μέλλον της οικονομίας. Φυσικά υπάρχουν στο Διαδίκτυο ιστοσελίδες καταξιωμένων επενδυτών των οποίων η γνώμη και η πρόβλεψη έχει βαρύνουσα σημασία στην αγορά. Αυτές επίσης εντάσσονται στις πηγές της πληροφορίας. Στην τεράστια κατηγορία των πηγών πληροφορίας εντάσσονται και τα ειδικά δελτία οικονομικών εξελίξεων. Αυτά απηχούν τις γνώμες των αναλυτών κορυφαίων επενδυτικών οίκων για τα χρηματοπιστωτικά προϊόντα της αγοράς και τις οικονομικές εξελίξεις εν γένει. Η κάθε κατηγορία-φορέας γνώμης έχει τα δικά της μοναδικά χαρακτηριστικά και η επεξεργασία της πρέπει να πραγματοποιηθεί με διαφορετικό τρόπο.

1.3 Διάκριση και επιλογή πληροφοριών

Η βάση των προς ανάλυση πληροφοριών όπως αντιλαμβάνεται κανείς είναι χαώδης. Επιβάλλεται λοιπόν η δημιουργία ενός υποσυστήματος το οποίο να έχει τη δυνατότητα να κατηγοριοποιεί, να φιλτράρει, να εγκρίνει και να απορρίπτει τα προς ανάλυση νέα. Οι τεχνικές που μπορούν να χρησιμοποιηθούν βασίζονται κυρίως στη λεξιλογική ανάλυση των κειμένων. Για παράδειγμα αν ένας επενδυτής χρειάζεται πληροφορίες για μια συγκεκριμένη μετοχή τότε οι ειδήσεις που εγκρίνονται σε πρώτη φάση για να οδεύσουν προς επεξεργασία είναι αυτές που είτε περιέχουν το όνομα της εταιρίας είτε αναφέρουν το ειδικό ticker που χρησιμοποιείται από τα χρηματιστήρια. Στη συνέχεια το φιλτράρισμα αλλάζει επίπεδο και πραγματοποιείται εντός του ιδίου του εγγράφου-είδησης. Εκεί εφαρμόζονται τεχνικές δομικής ανάλυσης του κειμένου. Υπάρχουν περιπτώσεις που η δημοσίευση απορρίπτεται διότι η αναφορά της εταιρίας γίνεται ελάχιστες φορές σε σχέση με το μέγεθος του κειμένου πράγμα που υποδεικνύει ότι στη

συγκεκριμένη είδηση απλά έγινε μία αναφορά και ουσιαστικά δεν παρέχεται πληροφορία άξια ανάλυσης. Επίσης οι αναλύσεις δύνανται να γίνονται σε επίπεδο της παραγράφου στην οποία έγινε η αναφορά της εταιρίας, δηλαδή η χρήσιμη και προς ανάλυση πληροφορία να εμπεριέχεται μόνο σε μία παράγραφο ενός δημοσιεύματος και το υπόλοιπο να απορρίπτεται. Επιπροσθέτως, κατά την κατηγοριοποίηση των δημοσιευμένων νέων υπεισέρχεται και ανάλυση σε επίπεδο λεξιλογίου. Η έννοια του θορύβου είναι πολύ σημαντική. Η χρήση λέξεων οι οποίες δεν έχουν αναλυτική αξία σε ένα έγγραφο, δηλαδή λέξεων οι οποίες χαρακτηρίζονται ουδέτερες καθιστούν το έγγραφο αυτό μη άξιο ανάλυσης και το οδηγούν μοιραία στην απόρριψη.

1.4 Τεχνικές εξαγωγής financial sentiment

Στη συγκεκριμένη όμως διπλωματική εργασία η βασική προς ανάλυση έννοια είναι το financial sentiment. Το κυρίως κείμενο μίας είδησης ή ενός δημοσιευμένου άρθρου προκαλεί στον αναγνώστη θετικά, αρνητικά ή ουδέτερα συναισθήματα. Τεχνικές εξαγωγής αυτού του συναισθήματος έχουν ήδη αρχίσει να εφαρμόζονται σε πολλές περιπτώσεις όπως η αξιολόγηση παραγόμενων προϊόντων, βαθμολόγηση υπηρεσιών εξυπηρέτησης ή ακόμα και αξιολόγηση ταινιών. Οι μέθοδοι ανάλυσης του συναισθήματος βρίσκονται κάτω από την ομπρέλα της Επεξεργασίας της Φυσικής Γλώσσας ή αλλιώς NLP (Natural Language Processing). Η διαδικασία αυτή οφείλει να αντιληφθεί μέσω ενός συνδυασμού συντακτικής και λεξιλογικής ανάλυσης το είδος του εγγενούς συναισθήματος που προκαλεί το κάθε κείμενο. Στη συνέχεια δε, η ανάλυση αυτή οφείλει να προσαρμοστεί σε πιο εξειδικευμένα κείμενα και συγκεκριμένα σε κείμενα που αφορούν τη χρηματοδοτική και οικονομική αγορά.

Οι μελέτες έχουν καταδείξει ότι αυτό μπορεί να επιτευχθεί με τη χρήση τριών ειδικών λεξικών συναισθήματος. Το πρώτο είναι ένα λεξικό που περιέχει λέξεις συναισθηματικές αξίες (sentimental value), το δεύτερο λέξεις με πολικότητα (polarity) συναισθήματος και το τρίτο λέξεις συναισθήματος μόνο. Το πρώτο λεξικό είναι αυτό το οποίο παρέχει τις περισσότερες πληροφορίες καθότι συνδυάζει τις δυνάμεις των άλλων δυο. Συγκεκριμένα ο εντοπισμός των λέξεων συναισθήματος από το τρίτο λεξικό είναι το πρώτο βήμα. Κατόπιν ο υπολογισμός της πολικότητας και της εγγύτητας προς μία συναισθηματική λέξη βάσης είναι το ουσιαστικά σημαντικό

στοιχείο. Επι παραδείγματι η λέξη good αποτελεί μία λέξη βάση με σαφή θετική πολικότητα. Η λέξη better ενέχει επίσης θετική πολικότητα αλλά βρίσκεται ένα βήμα παραπάνω από τη λέξη βάσης. Στην περίπτωση της λέξης magnificent, της οποίας η πολικότητα είναι επίσης θετική, η απόσταση της από τη λέξη good είναι ακόμα μεγαλύτερη. Αυτήν ακριβώς την απόσταση καλούμαστε να ποσοτικοποιήσουμε, διαχωρίζοντας με αυτόν τον τρόπο τη συναισθηματική βαρύτητα των λέξεων. Οι έρευνες επίσης έχουν δείξει ότι η συναισθηματική αυτή συσχέτιση δεν περιορίζεται μόνο στα επίθετα αλλά και σε άλλα μέρη του λόγου όπως τα επιρρήματα αλλά και τα ουσιαστικά.

Πέραν όμως της λεξιλογικής ανάλυσης οφείλουμε να λάβουμε υπόψιν και τη συντακτική ανάλυση ενός κειμένου. Οι αρνήσεις, τα αρνητικά επιρρήματα ακόμα και οι σύνδεσμοι έχουν τη δυνατότητα να αλλάξουν την έννοια του νοήματος αντιστρέφοντας την πολικότητα ή ακόμα και αυξάνοντας την αξία του συναισθήματος. Χαρακτηριστικό παράδειγμα είναι η έκφραση not good ή very good.

Στην παρούσα διπλωματική εργασία προτείνεται η χρήση τριών ειδικευμένων λεξικών για την εξαγωγή και ποσοτικοποίηση του financial sentiment.

Συγκεκριμένα το πρώτο είναι το Loughran and McDonald Financial Sentiment Dictionary. Η εργασία των Loughran και McDonald δημιούργησε ένα λεξικό ειδικευμένο σε οικονομικούς και χρηματοοικονομικούς όρους και εισήγαγε σε αυτούς βάρη και πολικότητα. Ειδικά για την ανάλυση του financial sentiment στα προς επεξεργασία δημοσιευμένα νέα το συγκεκριμένο λεξικό προκρίνεται ως βάση.

Επειδή πολλές φορές η πολιτική ειδησεογραφία επηρεάζει τις οικονομικές εξελίξεις η αγνόηση του πολιτικού περιβάλλοντος ίσως να ερμηνεύσει λανθασμένα τα δημοσιευμένα οικονομικά νέα. Για το λόγο αυτό σε αυτήν την εργασία προτείνεται και το Lexicoder Sentiment Dictionary το οποίο δημιουργήθηκε και χρησιμοποιήθηκε στην αποτελεσματική ανάλυση του πολιτικού sentiment με υψηλή προβλεπτική αξία σε εκλογικές αναμετρήσεις.

Τέλος προτείνεται η χρήση του WordStat Sentiment Dictionary το οποίο δεν περιέχει μόνο τις λέξεις που έχουν συναισθηματική φόρτιση αλλά διέπεται και από κανόνες που καλύπτουν τυχούσες αντιστροφές πολικότητας συναισθήματος. Επίσης επεκτείνει τις

λίστες με τις συναισθηματικά φορτισμένες λέξεις αναγνωρίζοντας συνώνυμα και αντώνυμα που προσθέτουν στην αξιοπιστία των αποτελεσμάτων.

Αξίζει να αναφερθεί ότι η χρήση του προγράμματος Linguistic Inquiry and Word Count στα παραπάνω λεξικά κρίνεται αρκετά βοηθητική στην εξαγωγή του πραγματικού συναισθήματος του συγγραφέα καθώς το συγκεκριμένο πρόγραμμα δημιουργήθηκε για να αποδώσει στα κείμενα την ψυχική διάθεσή του.

Όπως έχει προαναφερθεί η ψυχολογία είναι σημαντική συνιστώσα της χρηματοοικονομικής αγοράς και το κλίμα που δημιουργείται γύρω από τα επενδυτικά προϊόντα συνδιαμορφώνει την τιμή τους. Όμως το κλίμα ως απόρροια των ειδήσεων παράγεται από την αναγνωσιμότητά τους. Για παράδειγμα αν μία είδηση αποδεικνύεται από την ανάλυση αρνητική και προβλέπει πτώση της τιμής ενός χρεογράφου, αλλά δεν έχει διαβαστεί από τους επενδυτές, τότε εκ των πραγμάτων δεν συνεισφέρει στη γενικότερη αξιολόγηση. Συνεπώς κρίνεται αναγκαία η ύπαρξη βαρών στα αποτελέσματα που προκύπτουν βάσει της επισκεψιμότητας και αναγνωσιμότητας των εκάστοτε ειδήσεων. Ελαχιστοποιείται κατά αυτόν τον τρόπο το σφάλμα που μπορεί να προκληθεί ακόμα και αν η είδηση αποκωδικοποιηθεί σωστά.

Η βασική μελλοντική πρόκληση είναι ο συγκερασμός των παραπάνω λεξικών και προγραμμάτων ώστε μετά την εκτενή και σε βάθος ανάλυση των δημοσιευμένων ειδήσεων να εξάγεται ποσοτικοποιημένο το συναίσθημα που εκπέμπουν. Με άλλα λόγια, ο στόχος είναι η δημιουργία μίας κλίμακας η οποία θα καθορίζει αν τελικά η συνολική ειδησιογραφία που αφορά φερ΄ ειπείν έναν μετοχικό τίτλο προκρίνει την επένδυση ή την απορρίπτει, και κυρίως, σε ποιόν βαθμό.

Κατ' αυτόν τον τρόπο, θα προστεθεί και ένα νέο όπλο στη φαρέτρα των επενδυτών που καλύπτει την δεδομένη ανάγκη που έχουν στην κριτική πρόβλεψη.

1.5 Τεχνικές προβλέψεων

Στο πλαίσιο της εργασίας αυτής παρουσιάζεται σύντομα η επιστήμη των προβλέψεων. Ειδικότερα παρουσιάζονται συνοπτικά και οι μαθηματικές μέθοδοι που χρησιμοποιήθηκαν στην παρούσα εργασία και το θεωρητικό τους υπόβαθρο. Γίνεται αναφορά στις χρονοσειρές και τον τρόπο με τον οποίο θέτουν τη βάση για την εξαγωγή των προβλέψεων, αναλύοντας τα χαρακτηριστικά τους, καθώς και το πώς

αντιμετωπίζονται. Στη συνέχεια γίνεται λόγος τόσο για τις ποσοτικές όσο και τις ποιοτικές μεθόδους προβλέψεων αφού κατά τη διάρκεια του πειράματος που πραγματοποιήθηκε στα πλαίσια της διατριβής, έγινε χρήση και των δύο. Τέλος γίνεται ειδική αναφορά στα σφάλματα που προκύπτουν από τις διαδικασίες καθώς και τα κριτήρια με τα οποία επιλέγεται η κάθε μέθοδος πρόβλεψης.

1.6 Πείραμα

Στο τέλος της παρούσας διπλωματικής εργασίας εξετάστηκε η επιρροή που θα μπορούσαν δυνητικά να έχουν δημοσιευμένα άρθρα στην αυξομείωση των τιμών των μετοχών αλλά και στη λήψη αποφάσεων ενός επενδυτή. Για τον σκοπό αυτό, συγκεντρώθηκε ειδησιογραφία για τρεις μετοχές του αμερικάνικου χρηματιστηρίου με μεγάλη κεφαλαιοποίηση καθώς και οι τιμές κλεισίματός τους για μια περίοδο τεσσάρων μηνών. Με την χρήση ενός μη εκπαιδευμένου προγράμματος αξιολόγησης sentiment εξήχθη ποσοτικοποιημένο το συναίσθημα για κάθε δημοσιοποιημένο νέο και άρα και η απόφαση για το εάν ο επενδυτής οφείλει να επενδύσει, δηλαδή προσδοκά άνοδο στην τιμή της μετοχής, ή όχι, δηλαδή προσδοκά πτώση. Συγχρόνως εφαρμόστηκαν, ξεχωριστά, στα αριθμητικά δεδομένα απλές ποσοτικές μέθοδοι προβλέψεων. Οι απλές μέθοδοι προβλέψεων εξήγαγαν επίσης συμπεράσματα για την κίνηση της μετοχής και την απόφαση του επενδυτή.

Κατόπιν, χρησιμοποιήθηκε και η υβριδική μέθοδος, η οποία αποτελεί συνδυασμό των δύο παραπάνω μεθόδων εξαγωγής πρόβλεψης. Τα αποτελέσματα έδειξαν ότι το sentiment analysis υπερτερεί, σε αυτό το πείραμα, τόσο της απλής μαθηματικής μεθόδου όσο και της υβριδικής.

Τόσο το παραπάνω πείραμα όσο και η παρούσα διπλωματική εργασία κατέδειξαν την ανάγκη δημιουργίας ενός ολοκληρωμένου συστήματος αυτόματης ανάλυσης όλης της ειδησεογραφίας ως σημαντικό προσθετικό παράγοντα για την ακριβέστερη λήψη επενδυτικών αποφάσεων.

ΚΕΦΑΛΑΙΟ 2: ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

2.1 Βασικές Αρχές

Η ανάλυση sentiment είναι η δραστηριότητα της ανάδειξης των θετικών και αρνητικών γνωμών, συναισθημάτων και αξιολογήσεων. Οι περισσότερες εργασίες για την ανάλυση συναισθήματος έχουν γίνει σε επίπεδο εγγράφου, όπως για παράδειγμα η διάκριση μεταξύ των θετικών και των αρνητικών. Η ανάλυση του συναισθήματος μέσα στην οικονομική έρευνα είναι ένας σχετικά νέος και συναρπαστικός τομέας. Μερικές από τις πιο αξιοσέβαστες και αξιόπιστες δημοσιευμένες ειδήσεις στον κόσμο είναι αφιερωμένες σε οικονομικές και επιχειρηματικές ειδήσεις, οι οποίες διαδραματίζουν καίριο ρόλο στην παροχή οικονομικών ερεθισμάτων στους συμμετέχοντες με τις πληροφορίες που παρέχουν και βοηθώντας τους στη διαμόρφωση των απόψεών και των αποφάσεών τους.

Το αίνιγμα της εξήγησης της υπερβολικής αστάθειας των τιμών των μετοχών (volatility) που δεν μπορεί να ερμηνευθεί από τις θεμελιώδεις ή οικονομικές πληροφορίες είναι ένα ενδιαφέρον παζλ που στερείται μιας οριστικής απάντησης λόγω των δυσκολιών της ποσοτικοποίησης ή μέτρησης των ποιοτικών δεδομένων των μέσων ενημέρωσης (Cutler et al, 1989). Ωστόσο, τα τελευταία χρόνια οι ερευνητές έχουν αρχίσει να μετρούν το συναίσθημα που περιέχεται σε άρθρα στα μέσα ενημέρωσης χρησιμοποιώντας την ανάλυση των κειμένων σε μια προσπάθεια να συλλάβουν πληροφορίες που είναι δύσκολο να ποσοτικοποιηθούν και να προσδιορίσουν την επίπτωσή τους στις τιμές των μετοχών (Tetlock, 2007; Tetlock et al. 2008; Loughran and McDonald, 2010).

Χρησιμοποιώντας το σημασιολογικό περιεχόμενο των άρθρων των ειδήσεων των μέσων μαζικής ενημέρωσης, όπως οι θετικές και αρνητικές λέξεις που περιέχονται σε αυτά, μπορεί να παράσχει πολύτιμες πληροφορίες που τα ποσοτικά στοιχεία σχετικά με τα βασικά οικονομικά μεγέθη δεν μπορούν. Οι αποτιμήσεις των μετοχών θα πρέπει να είναι ίσες με την έκπτωση των αναμενόμενων ταμειακών ροών των επιχειρήσεων, σύμφωνα με το σύνολο των πληροφοριών του επενδυτή (Tetlock et al. 2008). Σε ένα σύνολο τέτοιων πληροφοριών, ωστόσο, περιέχονται επίσης ποιοτικές περιγραφές των προσδοκιών της μελλοντικής απόδοσης μιας επιχείρησης, όπως η ποιότητα της διοίκησης, ίσως μια αλλαγή στα κατώτερα επίπεδα της διοίκησης, μία ενδεχόμενη

συγχώνευση, αγωγές ή νομική δράση που λήφθηκε κατά της επιχείρησης, νέες σειρές προϊόντων ή διαφημιστικών εκστρατειών. Με τη χρήση ενός ποσοτικού μέτρου της σημασιολογίας στη γλώσσα που χρησιμοποιείται στα άρθρα των ειδήσεων, είναι δυνατόν να μετρηθούν οι επιπτώσεις των εν λόγω γεγονότων-ειδήσεων στις αποδόσεις των μετοχών. Επίσης, έχοντας ένα μεγάλο σύνολο δεδομένων που περιέχει πολλά γεγονότα ειδήσεων, επιτρέπει στους ερευνητές να μετρηθεί η αντίδραση της χρηματιστηριακής αγοράς αναλύοντας και τη σοβαρότητα της γλώσσας που χρησιμοποιείται μέσα στα άρθρα των ειδήσεων, ανεξάρτητα από το είδος της εκδήλωσης των ειδήσεων.

2.2 Ο ρόλος της τεχνολογίας

Στην προσπάθεια να προβλεφθούν οι αποδόσεις των μετοχών με βάση το περιεχόμενο των ειδήσεων ή των γνωμών, οι συμμετέχοντες στη χρηματοπιστωτική αγορά στρέφουν όλο και περισσότερο την προσοχή τους στη χρήση καινοτομιών με στόχο την αλγοριθμική ανάλυση και ερμηνεία των αναρτήσεων στα μέσα μαζικής ενημέρωσης.Οι New York Times και το περιοδικό Wired δημοσίευσαν άρθρα που τεκμηριώνουν την άνοδο των υπολογιστικών προγραμμάτων που διαβάζουν και επεξεργάζονται με μεγάλη ταχύτητα ειδήσεις, ειδικά προσαρμοσμένες από την υπηρεσία Dow Jones Lexicon (Bowley, 2010, Salomon και Stokes, 2010). Οι υπολογιστές σήμερα είναι σε θέση να διαβάσουν τις ειδήσεων, τις ιστοσελίδες, τις δημοσιεύσεις στα blog και ακόμη και τα μηνύματα στο twitter με άμεση απόδοση συνολικών πληροφοριών και εμπορικών σημάτων για συναλλαγές. Έχουν ακόμη εξελιχθεί προγράμματα για να μετρήσουν το συναίσθημα, την κατανόηση των αισθημάτων αυτών ακόμα και σε αδόμητα δεδομένα, όπως τα social media buzz.

2.3 Διαφορετικές χώρες – διαφορετικές ειδησεογραφικές αντιλήψεις

Σχεδόν όλες οι ακαδημαϊκές μελέτες που αφορούν το συναίσθημα των μέσων στον τομέα της χρηματοδότησης επικεντρώθηκαν στην αγορά των ΗΠΑ και τις δημοσιεύσεις των αμερικανικών μέσων ενημέρωσης. Ωστόσο, οι διαφορές στις δημοσιογραφικές αντιλήψεις και πρακτικές έχουν αρχίσει να λαμβάνονται σε μεγάλο βαθμό υπόψη, ειδικά σε χώρες που μοιράζονται παρόμοιες δημοσιογραφικές δεοντολογίες και πρακτικές (βλ.Weaver, 1998). Ο Shaw (1999) αναφέρει μερικές από τις διαφορές μεταξύ της κάλυψης των μέσων ενημέρωσης στις ΗΠΑ και το Ηνωμένο

Βασίλειο, τονίζοντας, συγκεκριμένα, ότι τα αμερικανικά μέσα ενημέρωσης έχουν πολύ μεγαλύτερη συμμόρφωση, ενώ τα μέσα μαζικής ενημέρωσης στο Ηνωμένο Βασίλειο έχουν πολύ μεγαλύτερη διασπορά στη γνώμη και ανεξαρτησία.. Δεδομένων των διαφορών στα χαρακτηριστικά μεταξύ των δύο αγορών, των ΗΠΑ και του Ηνωμένου Βασιλείου, δεν μπορούμε υποθέσουμε ότι τα αποτελέσματα του συναισθήματος μέσα στις αποδόσεις των μετοχών είναι ομοιόμορφα σε όλες τις χώρες. Συνεπώς, είναι σημαντικό να διερευνηθεί η επίδραση του συναισθήματος των μέσων ενημέρωσης γενικά, δεδομένης της παγκόσμιας παρουσίας του στις χρηματοοικονομικές αγορές και η διεθνής προσέγγιση των δημοσιευμένων νέων.

Στην κουλτούρα των ΗΠΑ, ως προς την ειδησεογραφία, η έρευνα δείχνει ότι υπάρχει μεγαλύτερη διστακτικότητα στο να αμφισβητήσει κανείς τις πηγές της εξουσίας των επιχειρήσεων και της κυβέρνησης και μια μεγαλύτερη διείσδυση των ταμπλόιτ του Τύπου στην κοινωνία γενικά, καθώς τα μέσα στοχεύουν στην επίτευξη ενός μαζικού ακροατηρίου και άρα στην διασφάλιση μεγαλύτερου μεριδίου της αγοράς των ΜΜΕ και όχι να παρέχουν μια μειοψηφούσα/σοβαρή φωνή (Deuze, 2002). Τα βρετανικά μέσα ενημέρωσης, ωστόσο, δεν φοβούνται να εκφράσουν ισχυρά τη γνώμη τους σε επίμαχα θέματα και γενικά παρέχουν μια μεγαλύτερη και ανεξάρτητη φωνή. Λαμβάνοντας υπόψη τις διαφορές στα χαρακτηριστικά μεταξύ των δύο αγορών, των ΗΠΑ και του Ηνωμένου Βασιλείου, δεν μπορούμε να υποθέσουμε ότι οι συνέπειες των απόψεων των μέσων ενημέρωσης σχετικά με τις αποδόσεις των επιχειρήσεων και των μετοχών είναι ίδιες. Επομένως, είναι σημαντικό να διερευνηθεί η επίδραση των μέσων στην αγορά σε παγκόσμιο επίπεδο.

2.4 Σχέση μεταξύ είδησης και απόδοσης οικονομικών προϊόντων

Σκοπός της παρούσας εργασίας είναι να αναπτυχθεί η μεθοδολογία για την ποσοτικοποίηση της συσχέτισης ανάμεσα στο χρηματοοικονομικό συναίσθημα και τα δημοσιευμένα νέα και των κριτικών προβλέψεων για αποφάσεις που αφορούν οικονομικά δεδομένα και προϊόντα.

Χρησιμοποιούνται τόσο θετικά όσο και αρνητικά μέτρα του συναισθήματος των μέσων ενημέρωσης, καθώς προηγούμενες μελέτες, όπως Tetlock et al. (2008) χρησιμοποίησαν μόνο αρνητικά μέτρα, δηλαδή μέσα απαισιοδοξίας για τη μελέτη των αποδόσεων των μετοχών. Με τη χρήση τόσο θετικών όσο και αρνητικών μέτρων του συναισθήματος

των μέσων ενημέρωσης, η συνολική κατανομή των ειδήσεων μπορεί στη συνέχεια να χρησιμοποιηθεί για διαμορφωθεί μία εικόνα για τη συχνότητα και τις πιθανές προκαταλήψεις που συνδέονται με άρθρα ειδήσεων. Με τον τρόπο αυτό είναι δυνατόν να προσδιοριστεί κατά μέσο όρο, πόσο το συναίσθημα ενσωματώνεται στις αποδόσεις των μετοχών.

Έχει υπάρξει εκτενής έρευνα σχετικά με τον αντίκτυπο των πληροφοριών ή «ειδήσεων» για τις χρηματοδοτικές αγορές. Η έννοια των νέων, παραδοσιακά, ερμηνεύονται ως αιτιώδης μεταβλητή σε χρηματοδοτικά μοντέλα της αγοράς και ένα μεγάλο μέρος των προηγούμενων μελετών δείχνουν ποσοτικές πτυχές των ειδήσεων ως υποκατάστατο για την ίδια η είδηση. Παραδείγματα τέτοιων μελετών περιλαμβάνουν το χρονοδιάγραμμα της άφιξης των ειδήσεων, ο όγκος των ειδήσεων και ο τύπος των ειδήσεων (περιοδικά αναμενόμενες ανακοινώσεις και γενικά διαθέσιμες στο κοινό ειδήσεις). Μεταξύ των συνεισφορών σε αυτή την βιβλιογραφία είναι αυτή των Mitchell και Mulherin (1994), Coval και Shumway (2001) και Antweiler και Frank (2004), που έχουν χρησιμοποιήσει την ενημέρωση του κοινού, την μετάδοση των νέων και τις ειδήσεις στο διαδίκτυο.

Νέες πληροφορίες εισάγονται στην αγορά όλο το χρόνο. Ενώ μία ποικιλία πηγών πληροφοριών μπορούν να κινήσουν όλες ή και μια τιμή της μετοχής, π.χ. φήμες, υποκλοπές και σκάνδαλα, ενώ οι οικονομικές ειδήσεις ήδη θεωρούνται πιο σταθερές και πιο αξιόπιστες πηγές. Η σταθερότητα αυτή ώθησε κάποιους να δηλώνουν ότι οι ειδήσεις είναι μια άλλη μορφή εμπορεύματος (Mowshowitz 1992), η οποία μπορεί να έχει διαφορετικές τιμές (Raban & Rafaeli 2006). Ωστόσο, η ακριβής σχέση μεταξύ των οικονομικών άρθρων των ειδήσεων και της κίνησης των τιμών των μετοχών είναι πολύπλοκη. Ακόμα και όταν οι πληροφορίες που περιέχονται σε οικονομικά άρθρα ειδήσεων μπορεί να έχουν ένα ορατό αντίκτυπο στην τιμή ενός χρεογράφου (Gidofalvi 2001) ενώ απότομες διακυμάνσεις των τιμών μπορεί ακόμα να προκύψουν από άλλες πηγές, όπως είναι οι μεγάλες απρόσμενες συναλλαγές (Camerer & Weigelt 1991).

Ο Engle (2003) έδωσε μια μαθηματική περιγραφή της ασύμμετρης και συναισθηματικής επίπτωσης των ειδήσεων σχετικά με τις τιμές: οι θετικές ειδήσεις συνήθως σχετίζονται με μεγάλες αλλαγές στις τιμές, αλλά μόνο για ένα μικρό χρονικό διάστημα. Αντίθετα η επίδραση των αρνητικών ειδήσεων σχετικά με τις τιμές και τον όγκο των συναλλαγών είναι μεγαλύτερης διάρκειας. Ο αναδυόμενος τομέας της

κοινωνιολογίας των Οικονομικών εξετάζει τις δημοσιονομικές αγορές, όπως τις κοινωνικές δομές και τον τρόπο επικοινωνίας, όπως τα e-mails και τα δελτία ειδήσεων, που μπορεί να είναι φορτωμένα με συναίσθημα και που θα μπορούσαν να στρεβλώσουν τις απόψεις των επενδυτών.

Φαίνεται ότι οι ειδήσεις επηρεάζουν τις αγορές με προφανείς τρόπους, επηρεάζοντας τον όγκο των συναλλαγών, την απόδοση των μετοχών, την αστάθεια των τιμών και ακόμη και τα μελλοντικά κέρδη μιας επιχείρησης. Στον τομέα των επιπτώσεων της ανάλυσης των ειδήσεων στη χρηματοδότηση, κατά τα τελευταία έτη, το επίκεντρο έχει επεκταθεί από το ενημερωτικό στο συναισθηματικό περιεχόμενο των κειμένων σε μία προσπάθεια για να εξηγήσει η σχέση μεταξύ του κειμένου και των αγορών.

2.5 Το πλεονέκτημα που παρέχουν οι υπολογιστές και το Διαδίκτυο

Όλα τα κείμενα, είτε πρόκειται για ειδήσεις, blogs, λογιστικές καταστάσεις ή ποίηση, επικαλούνται το συναίσθημα, παρέχοντας μια προοπτική του πραγματικού περιεχομένου του κειμένου. Με την αύξηση της ισχύος των υπολογιστών καθώς και των λεκτικών και δομικών πόρων φαίνεται να είναι υπολογιστικά εφικτό να ανιχνεύονται ορισμένες από τις συναισθηματικές πτυχές του περιεχομένου του κειμένου αυτόματα (Ahmad et al., 2006). Μια συστηματική ανάλυση των επιπτώσεων της προκατάληψης των ειδήσεων ή της πολικότητας των μεταβλητών της αγοράς απαιτεί μια αριθμητική τιμή για την ένταση του συναισθήματος, καθώς και μια δυαδική ετικέτα για το συναίσθημα της πολικότητας, τον προσδιορισμό των τάσεων στο δείκτη του συναισθήματος καθώς και τα σημεία καμπής.

Η πρώτη πρόκληση ενός χρηματοπιστωτικού συστήματος πρόβλεψης είναι η διαχείριση των μεγάλων ποσοτήτων πληροφοριών κειμένου που υπάρχουν για τις κινητές αξίες. Αυτό το υλικό μπορεί να περιλαμβάνει απαιτούμενες εκθέσεις, όπως περιοδικές εκθέσεις, δελτία τύπου και οικονομικά άρθρα ειδήσεων που αναφέρουν και απρόβλεπτα γεγονότα. Αυτά τα έγγραφα κειμένου μπορούν στη συνέχεια να αναλυθούν χρησιμοποιώντας τεχνικές διαδικασίες ανάλυσης φυσικής γλώσσας, ή αλλιώς Natural Language (NLP) για τον προσδιορισμό των ειδικών όρων του άρθρου ή φράσεων που το πιο πιθανό είναι να προκαλέσουν δραματικές αλλαγές στην τιμή της μετοχής, όπως το "εργοστάσιο εξερράγη" θα έδειχνε πιθανώς μια βουτιά των τιμών στο εγγές μέλλον. Με την αυτοματοποίηση της διαδικασίας αυτής, οι μηχανές μπορούν να

επωφεληθούν από τις ευκαιρίες κερδοσκοπίας ταχύτερα από τον άνθρωπο προβλέποντας επανειλημμένα τις διακυμάνσεις των τιμών και εκτελώντας άμεσες συναλλαγές.

Η απόκτηση έγκαιρων οικονομικών εγγράφων από αξιόπιστες πηγές στο διαδίκτυο είναι ένα κρίσιμο βήμα και υπάρχει μια ποικιλία οικονομικών χώρων συνάθροισης ειδήσεων που παρέχουν αυτή την υπηρεσία. Ένα από αυτά τα sites είναι το Comtex που προσφέρει σε πραγματικό χρόνο οικονομικές ειδήσεις σε μορφή συνδρομής. Μια άλλη πηγή είναι το PRNewswire, το οποίο προσφέρει δωρεάν σε πραγματικό χρόνο και συνδρομητικές υπηρεσίες. Το Yahoo! Finance είναι μια τρίτη τέτοια πηγή και είναι μια συλλογή από 45 διαφορετικές πηγές ειδήσεων όπως το Associated Press, Financial Times και PRNewswire μεταξύ άλλων. Αυτή η πηγή παρέχει μια ποικιλία από προοπτικές και έγκαιρες ειδήσεις σχετικά με τις χρηματοπιστωτικές αγορές.

2.6 Διατύπωση προβλήματος

Οι επενδυτές της σύγχρονης χρηματοοικονομικής αγοράς βρίσκονται σε συνεχή πίεση, στην αναζήτηση αποδόσεων και την μεγιστοποίηση των κερδών των χαρτοφυλακίων τους. Είτε πρόκειται για δημοσίους οργανισμούς, είτε για hedge funds, είτε για τραπεζικούς κολοσσούς, είτε για μεμονωμένους παίκτες της επενδυτικής αγοράς, όλοι διαθέτουν περιορισμένο χρόνο στην απόκριση τους για γρήγορες και προσοδοφόρες κινήσεις.

Η λήψη επενδυτικών αποφάσεων από τους traders βασίζονται σε τρία βασικά χαρακτηριστικά της ανάλυσης των πληροφοριών που λαμβάνουν:

Α. Στο πρώτο μέρος της απόφασης περιλαμβάνεται η ανάλυση των λεγομένων **Fundamental** Financial Terms, δηλαδή των βασικών οικονομικών όρων των υπό επένδυση χρεογράφων. Σε αυτά, μεταξύ άλλων, περιλαμβάνονται ο ισολογισμός και οι ταμιακές ροές μιας εταιρίας, οι δείκτες απόδοσης/εκτίμησης μετοχών, ο δείκτης χρεοκοπίας, η καθαρή παρούσα αξία ενός ομολόγου, ακόμα και το επίπεδο του ΑΕΠ ενός κράτους.

Β. Παράλληλα χρησιμοποιείται η τεχνική ανάλυση (**technical analysis**) η οποία ουσιαστικά αποτελεί μία μέθοδος επεξεργασίας στατιστικών πληροφοριών που γεννά

η αγορά και φανερώνει τάσεις του οικονομικού περιβάλλοντος, αλλά και το προς τα που κινούνται τα χρηματοοικονομικά προϊόντα μέσα σε αυτό.

Γ. Κριτική πρόβλεψη, ανάλυση ειδήσεων και η μέτρηση του financial sentiment της αγοράς. Ως γνωστό το πώς κινείται η χρηματοοικονομική αγορά και γενικότερα η οικονομία είναι άρρηκτα συνδεδεμένο με την ψυχολογία των παικτών που συμμετέχουν. Αυτή σε πολύ μεγάλο βαθμό εκφράζεται από τις δημοσιευμένες απόψεις των αναλυτών στον έντυπο και ηλεκτρονικό τύπο.

Η παρούσα διπλωματική εργασία εστιάζεται στο πρόβλημα της έλλειψης ικανού χρόνου για την ολοκληρωμένη ενημέρωση του εκάστοτε επενδυτή-παίκτη της χρηματοοικονομικής αγοράς από τα δημοσιευμένα νέα - κυρίως στον ηλεκτρονικό τύπο - και ερευνά τις διαδικασίες και τους τρόπους αυτοματοποιημένης ποσοτικοποίησης της οικονομικής αίσθησης (financial sentiment)όλης της αγοράς για τη λήψη επενδυτικών αποφάσεων κριτικής πρόβλεψης για οικονομικά προϊόντα.

2.7 Ορισμοί

Έννοιες που θα απαντηθούν στην παρούσα διατριβή και η σημασία τους:

1. **<u>Forecasting</u>:** Η πρόβλεψη είναι μία από τις σημαντικότερες λειτουργίες μέσα σε μια επιχείρηση και εν γένει σε έναν οργανισμό, για τη λήψη κάθε κρίσιμης απόφασης. Η πρόβλεψη μπορεί να είναι βραχυπρόθεσμη, μεσοπρόθεσμη ή μακροπρόθεσμη ανάλογα με τον χρονικό ορίζοντα στον οποίο αναφέρεται.
   Αρχές της Πρόβλεψης
   i. Καμία πρόβλεψη δεν είναι τέλεια. Περιλαμβάνει το στοιχείο της αβεβαιότητας, η πρόβλεψη θα περιέχει κάποιο σφάλμα (δηλ. τη διαφορά μεταξύ της πρόβλεψης και της πραγματικότητας). Με βάση αυτό, στόχος της διαδικασίας πρόβλεψης είναι η ελαχιστοποίηση του σφάλματος για την όσο το δυνατόν ακριβέστερη προσέγγιση της πραγματικότητας.
   ii. Μια πρόβλεψη είναι περισσότερο ακριβής για ομάδες στοιχείων παρά για μεμονωμένα δεδομένα: π.χ. η πρόβλεψη της συνολικής ζήτησης για βιομηχανικά ορυκτά (καολίνης, μπεντονίτης, περλίτης κτλ.) για το επόμενο έτος θα είναι ακριβέστερη από την ζήτηση για ένα συγκεκριμένο ορυκτό (π.χ. του περλίτη) και η τελευταία θα είναι με τη σειρά της ακριβέστερη από την πρόβλεψη της ζήτησης για ένα ορυκτό με ορισμένη ποιότητα (π.χ. περλίτης

συγκεκριμένης κοκκομετρίας). Αυτό συμβαίνει γιατί οι μέγιστες και ελάχιστες τιμές των διαφόρων στοιχείων (π.χ. ορυκτών) αλληλοεξουδετερώνονται με αποτέλεσμα η ομάδα το στοιχείων να έχει σταθερή συμπεριφορά ακόμα και αν τα μεμονωμένα στοιχεία συμπεριφέρονται με ασταθή τρόπο.

iii. Η πρόβλεψη είναι περισσότερο ακριβής όταν είναι βραχυπρόθεσμη παρά όταν είναι μακροπρόθεσμη: όσο κοντινότερος είναι ο χρονικός ορίζοντας της πρόγνωσης τόσο μικρότερος είναι ο βαθμός αβεβαιότητας και άρα τόσο μικρότερο το σφάλμα που θα περιέχει. Ένα κλασσικό παράδειγμα αφορά στην πρόβλεψη του καιρού: ένα μετεωρολογικό δελτίο για τις επόμενες δύο ή τρεις μέρες είναι πάρα πολύ πιθανό να είναι βγει αληθινό. Αντίθετα, η πρόγνωση για τον καιρό του επόμενου μήνα έχει μεγάλες πιθανότητες να αποδειχτεί λανθασμένη.

2. **Judgmental Forecasting (Κριτική πρόβλεψη):** Η πρόβλεψη που γίνεται βάση υποκειμενικών πληροφοριών ή συναισθημάτων. Συχνά τέτοιου είδους προβλέψεις γίνονται λόγω έλλειψης πληροφοριών ή δεδομένων τα οποία σε κλασσικές εφαρμογές προβλέψεων τα αποτελέσματά τους αποδεικνύονται αναξιόπιστα. Σε πολλές περιπτώσεις η κριτική πρόβλεψη συνεπικουρεί και προσαρμόζει τη μαθηματική μοντελοποιημένη πρόβλεψη ώστε να ελαχιστοποιείται το σφάλμα της.
3. **Financial Sentiment:** Ως financial sentiment ορίζεται η υποκειμενική αίσθηση που απορρέει από τα δημοσιευμένα νέα που αφορούν οικονομικά δεδομένα, προϊόντα και την χρηματοοικονομική αγορά.
4. **Polarity:** Ως polarity ορίζεται η συναισθηματική κατεύθυνση μιας λέξης, μίας πρότασης ή ακόμα και ενός ολόκληρου κειμένου. Διακρίνεται σε θετικό και αρνητικό, ενώ πολλές φορές συναντάται ως ουδέτερο.
5. **Επεξεργασία Φυσικής Γλώσσας:** (Natural Language Processing) Είναι οι μέθοδοι, οι τεχνικές, τα εργαλεία και οι εφαρμογές για τη μοντελοποίηση της χρήσης της ανθρώπινης γλώσσας. Είναι παραδοσιακός όρος και προέρχεται από τον ομώνυμο κλάδο της Τεχνητής Νόησης (Artificial Intelligence). Αποτελεί τον κλάδο της επιστήμης ο οποίος έχει ως σκοπό τη καλυτέρευση της επικοινωνίας ανθρώπου μηχανής.

ΚΕΦΑΛΑΙΟ 3: ΣΥΛΛΟΓΗ ΠΛΗΡΟΦΟΡΙΩΝ

3.1 Συλλογή πληροφορίας και συμπεριφορική χρηματοδότηση

Η πρόσφατη έρευνα στην συμπεριφορική χρηματοδότηση υποδηλώνει ότι πράγματι, μακριά από το να είναι ορθολογική, οι επενδυτές και οι καταναλωτές κάνουν επενδύσεις και λαμβάνουν αποφάσεις κατανάλωσης οδηγούμενοι εν μέρει από το συναίσθημα και τη συγκίνηση, επηρεάζονται από γνωστικές προκαταλήψεις, και συνήθως στηρίζονται σε ελλιπείς, ανακριβείς και «θορυβώδεις» πληροφορίες για την απόφασή τους και στις διαδικασίες λήψης αποφάσεων. Τελικά, η έρευνα κάνει ένα βήμα πίσω και μας θυμίζει ότι «η ψυχολογία δεν ήταν ποτέ εκτός της χρηματοδότησης» και ότι όλη η συμπεριφορά των ανθρώπων που συμμετέχουν στις αγορές και την οικονομία στηρίζονται στην ψυχολογία (Statman, 1999). Ο Statman (1999), κατά την επανεξέταση των συμπεριφορικών οικονομικών (behavioral finance), δημιουργεί μια λίστα με μερικά από τα εργαλεία της συμπεριφορικής χρηματοδότησης για τη μοντελοποίηση της ανθρώπινης συμπεριφοράς, συμπεριλαμβανομένης της «ευαισθησίας σε πλαίσια και άλλα γνωστικά λάθη, διαφορετικές στάσεις απέναντι στον κίνδυνο, αποστροφή για λύπη, ατελή αυτο-έλεγχο και προτιμήσεις ως προς τη χρήση και την αξία των εκφραστικών χαρακτηριστικών».

Οι Baker και Wurgler (2007) καθορίζουν σε γενικές γραμμές την ψυχολογία των επενδυτών ως «μια προσδοκία/πεποίθηση για τις μελλοντικές ταμειακές ροές και επενδυτικούς κινδύνους που δεν δικαιολογείται από τα πραγματικά γεγονότα». Παρουσιάζοντας τη σύγχρονη κριτική τους για το θέμα, οι Baker και Wurgler (2007) υποστηρίζουν ότι «το ζήτημα δεν είναι πλέον, όπως ήταν πριν από μερικές δεκαετίες, αν η ψυχολογία των επενδυτών επηρεάζει τις τιμές των μετοχών, αλλά μάλλον πώς μπορεί να μετρηθεί το επενδυτικό κλίμα και να ποσοτικοποιηθούν οι επιπτώσεις του» - αν και σίγουρα παραδέχονται ότι η μέτρηση αυτή δεν είναι μια απλή υπόθεση. Μια λύση, που οι Baker και Wurgler (2007) προτείνουν, είναι να χρησιμοποιηθεί ένας συνδυασμός από διαφορετικές πληροφορίες για την ψυχολογία που παραμένουν χρήσιμες για τουλάχιστον κάποιο χρονικό διάστημα. Εισάγουν και συνοψίζουν συγκεκριμένες πληροφορίες που πρέπει να συλλέγονται όπως έρευνες του επενδυτή (συμπεριλαμβανομένων των UBS / Gallup δημοσκοπήσεων), μικροοικονομικού επιπέδου δεδομένων λιανικής εμπορικής (λιανικών πωλήσεων), αμοιβαίες (mutual)

ροές κεφαλαίων, όγκος των συναλλαγών, τεκμαρτή μεταβλητότητα λόγω γνώμης (συμπεριλαμβανομένου του δείκτη VIX) και συνολικός όγκος συναλλαγών.

Οικονομικές ιστοσελίδες, περιοδικά, εφημεριδες επενδύσεων και δημοσιεύσεις κρατούν ενήμερο έναν επενδυτή και τον βοηθούν να πληροφορηθεί σχετικά με μία επένδυση, για το τι συμβαίνει στην οικονομία, τα νέα που επηρεάζουν τα χρήματά του, το που πρέπει να επενδύσει τα χρήματά του και που μπορεί να τοποθετήσει τα χρήματά του έτσι ώστε να έχει την υψηλότερη δυνατή απόδοση. Οι πηγές αυτές περιέχουν πολύτιμες πληροφορίες σχετικά με τις επιχειρήσεις εν γένει, καθώς και τις τρέχουσες οικονομικές και χρηματοπιστωτικές εξελίξεις, τις ειδήσεις της χρηματιστηριακής αγοράς και τις σχετικές ειδήσεις, όλες αυτές οι πληροφορίες που επηρεάζουν το επενδυτικό κοινό και το πιο σημαντικό, που επηρεάζουν τις επενδύσεις και τις επενδυτικές αποφάσεις εν γένει.

Υπάρχει μια πληθώρα των χρηματοπιστωτικών και επενδυτικών διαθέσιμων πληροφοριών για τους μεμονωμένους επενδυτές και ένα από τα καθήκοντα κάποιου, αν θέλει να είναι ενημερωμένος, είναι να αποστάξει αυτόν τον όγκο των πληροφοριών, προκειμένου να βρει τις πηγές που μπορεί και καταλαβαίνει ότι θα του είναι χρήσιμες, που παρέχουν πληροφορίες που είναι σαφείς, αξιόπιστες, και, όσο το δυνατόν, ανεξάρτητες από προκαταλήψεις και επιρροές.

Ο Niederhoffer (1971), ακαδημαϊκός και διαχειριστής αμοιβαίου κεφαλαίου, ανέλυσε 20 χρόνια πρωτοσέλιδα των New York Times που κατατάσσονται σε 19 σημασιολογικές κατηγορίες και πάνω σε μία «καλή-κακή» κλίμακα διαβάθμισης για να εκτιμηθεί πώς οι αγορές αντέδρασαν σε καλές και κακές ειδήσεις: ο ίδιος διαπίστωσε ότι οι αγορές αντιδρούν στις ειδήσεις με μια τάση να υπερ-αντιδρούν σε κακές ειδήσεις. Κάπως προφητικά, προτείνει ότι οι ειδήσεις πρέπει να αναλυθούν από τους υπολογιστές για την εισαγωγή περισσότερης αντικειμενικότητας στην ανάλυση. Οι Engle και Ng (1993) προτείνουν την «καμπύλη επίπτωσης ειδήσεων» (News Impact Curve) ως μοντέλο για το πώς οι ειδήσεις επιδρούν στη μεταβλητότητα στην αγορά με τα άσχημα νέα να προκαλούν περισσότερη αστάθεια. Χρησιμοποίησαν τη μεταβλητή της αγοράς, δηλαδή τις αποδόσεις των μετοχών, ως υποκατάστατο για τις ειδήσεις, από μια απρόσμενη πτώση στις αποδόσεις για τις κακές ειδήσεις και την απροσδόκητη άνοδο στις καλές ειδήσεις. Πράγματι, οι πολύ πρώιμες μελέτες χρησιμοποίησαν τέτοιες μεταβλητές της αγοράς ή εύκολα ποσοτικά στοιχεία των ειδήσεων ως υποκατάστατο

της είδησης από μόνη της: όπως π.χ. άφιξη ειδήσεων, το είδος, την προέλευση και τον όγκο (Cutler et al, 1989, Mitchell και Mulherin, 1994). Πιο πρόσφατες μελέτες έχουν προχωρήσει, με τη βοήθεια υπολογιστή, σε ένα πνεύμα αντικειμενικότητας, η οποία συνεπάγεται τον καθορισμό των γλωσσικών χαρακτηριστικών που πρέπει να χρησιμοποιηθούν για την αυτόματη κατηγοριοποίηση του κειμένου σε θετικές ή αρνητικές ειδήσεις. Οι Davis et al (2006) ερεύνησαν τις επιπτώσεις από την αισιόδοξη ή την απαισιόδοξη χρήση της γλώσσας που χρησιμοποιείται στα οικονομικά δελτία τύπου σχετικά με τις μελλοντικές αποδόσεις των επιχειρήσεων. Συμπέραναν ότι α) οι αναγνώστες διαμορφώνουν τις προσδοκίες τους σχετικά με τη συνήθη προκατάληψη των συγγραφέων και β) αντιδρούν έντονα με τις εκθέσεις που παραβιάζουν αυτές τις προσδοκίες, γεγονός που υποδηλώνει έντονα ότι οι αναγνώστες, και κατ' επέκταση οι αγορές, διαμορφώνουν προσδοκίες σχετικά και αντιδρούν όχι μόνο με το περιεχόμενο αλλά και συναισθηματικές πτυχές ενός κειμένου.

Ο Tetlock (2007) διερευνά επίσης πώς ένας παράγοντας απαισιοδοξίας, παράγεται αυτόματα από το κείμενο ειδήσεων μέσω του όρου της ταξινόμησης και ανάλυσης κυρίων συνιστωσών, μπορεί να προβλέψει τη δραστηριότητα της αγοράς, ιδίως τις αποδόσεις των μετοχών. Βρίσκει ότι η υψηλή αρνητικότητα στις ειδήσεις προβλέπει χαμηλότερες αποδόσεις έως και 4 εβδομάδες περίπου. Οι μελέτες καθορίζουν τις σχέσεις μεταξύ της συναισθηματικής προκατάληψης σε μορφή κειμένου και τη δραστηριότητα της αγοράς που οι παίκτες της αγοράς και οι ρυθμιστικές αρχές μπορεί να έχουν να αντιμετωπίσουν

3.2 Πηγές πληροφόρησης, ειδησεογραφίας

3.2.1 Bloomberg Platform / Bloomberg Terminal

Η Bloomberg L.P. αποτελεί μία από τις μεγαλύτερες επιχειρήσεις παγκοσμίως που συνδυάζει την παγκόσμια οικονομική ειδησεογραφία με τις χρηματοπιστωτικές υπηρεσίες και την τεχνική υποστήριξη των δύο.

Σχεδόν όλοι οι traders παγκοσμίως χρησιμοποιούν το Bloomberg Terminal ώστε να έχουν πρόσβαση σε πραγματικό χρόνο σε όλες στις αγορές όλου του πλανήτη.

Η πλατφόρμα του Bloomberg είναι μια βάση δεδομένων πλούσια σε πληροφορίες, που απορρέει από τη μεγαλύτερη βάση δεδομένων που είναι διαθέσιμη στην αγορά. Οι πληροφορίες για περισσότερο από 6 + εκατομμύρια μέσα, σε συνδυασμό με εξελιγμένη αναλυτική λειτουργικότητα και τη δυνατότητα να υπολογίσει προσαρμοσμένες αποδόσεις με βάση τον πελάτη - συγκεκριμένες εισροές (παρακάμψεις) κάνουν τα δεδομένα ένα ισχυρό προϊόν.

Πάνω από 6 εκατομμύρια μέσα σε όλες τις κατηγορίες περιουσιακών στοιχείων, όπως οι εξής:

- σταθερού Εισοδήματος
- κεφαλαίων
- CDS
- Υποθήκες
- Νομίσματα
- Δείκτες
- Μετοχές
- Παράγωγα

Πάνω από 6 εκατομμύρια μέσα σε όλες τις κατηγορίες δεδομένων:

- Εταιρικές πράξεις
- Περιγραφικά δεδομένα
- Τιμές από ανταλλαγές και πολλαπλούς συνεισφέροντες
- Προερχόμενα και υπολογιζόμενα δεδομένα
- Δεδομένα πελατών και αναλυτική λειτουργικότητα
- Ιστορική χρονοσειρά

- Βαθμολογίες
- Βασικές αρχές

Οι πληροφορίες που προσφέρει το Bloomberg platform συνολικά καλύπτουν:

- 250 + παγκόσμιες αγορές σε 126 χώρες
- Πληροφορίες για το επίπεδο της ασφάλειας και των εκδοτών πληροφοριών
- Ένας αριθμός σταθερών - πηγών εισοδήματος των τιμών που είναι απαράμιλλος στην αγορά
- Credit Risk Module

Απευθύνεται σε πελάτες που ψάχνουν για την πιο ολοκληρωμένη, έγκαιρη και ακριβή υπηρεσία δεδομένων των επιχειρήσεων για να τροφοδοτήσουν κρίσιμες εφαρμογές γραφείου και / ή βάσεις δεδομένων.

Η Bloomberg L.P. αποτελεί έναν ειδησεογραφικό κολοσσό, που απασχολεί χιλιάδες δημοσιογράφους σε 192 σημεία όλου του κόσμου που μεταδίδουν οικονομικά νέα τη στιγμή που συμβαίνουν. Επίσης η πλατφόρμα συνδέει επενδυτές από όλα τα μήκη και πλάτη του πλανήτη οι οποίοι διατυπώνουν γνώμες και μεταδίδουν ειδήσεις. Ο συνδυασμός των παραπάνω, με την σύγχρονη ενσωμάτωση των ειδήσεων που μεταδίδονται από τα υπόλοιπα παγκόσμια και τοπικά μέσα ενημέρωσης στην ίδια πλατφόρμα, τοποθετεί το Bloomberg Terminal στην κορυφαία θέση, ίσως και πάνω από το ίδιο το αχανές Διαδίκτυο, για την άντληση δημοσιευμένων νέων και κριτικών προβλέψεων σε όλο τον κόσμο.

Το Bloomberg Terminal έχει τη δυνατότητα να φιλτράρει από την αρχιτεκτονική του τα άρθρα και τα νέα για να παρουσιάσει σε οποιονδήποτε παίκτη της αγοράς αυτό που επιθυμεί να διαβάσει ή να συμβουλευτεί.

Συνεπώς η πλατφόρμα του Bloomberg αποτελεί μία από τις πιο αξιόπιστες λύσεις και πηγές για την άντληση πληροφοριών σε φυσική γλώσσα αλλά και σε πλήθος ώστε να μπορεί να εξαχθεί το financial sentiment για οποιοδήποτε οικονομικό προϊόν αλλά και για τη γενικότερη πορεία των τιμών και των δεικτών των αγορών.

Εικόνα 3.1: Πλατφόρμα Bloomberg, παρουσίαση ειδήσεων όπως ορίζει ο χρήστης

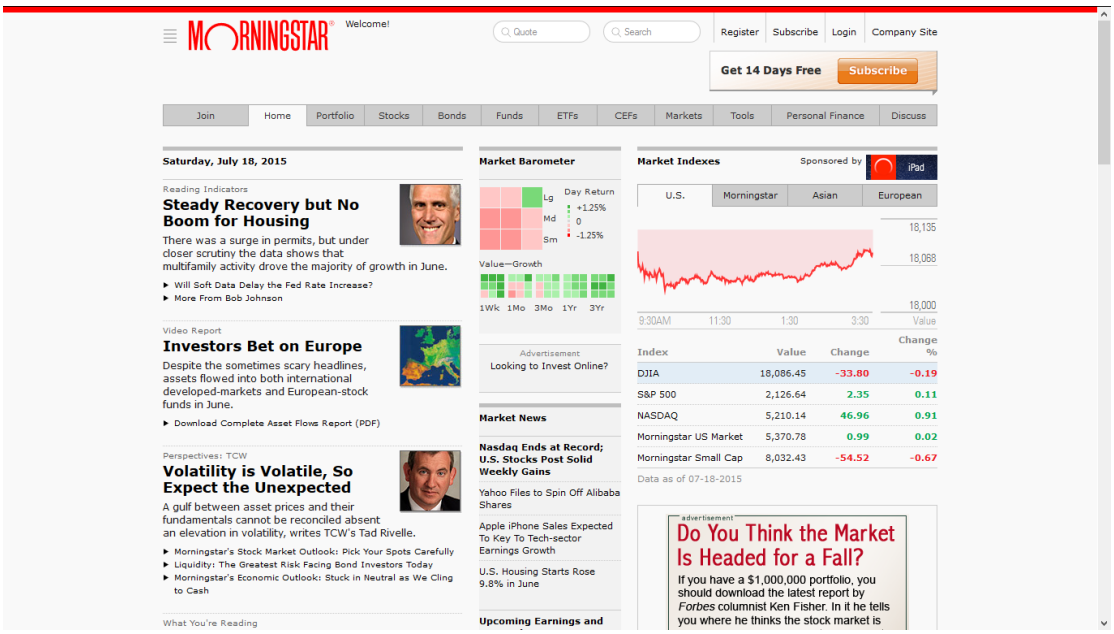
Εικόνα 3.2: Πλατφόρμα Bloomberg, παρουσίαση ειδήσεων που μεταδίδονται παγκοσμίως

3.2.2 Major Financial Mass Media Web pages (συλλέκτες πληροφοριών)

Οι Major Financial Mass Media Web pages είναι ιστοσελίδες οι οποίες είναι ειδικευμένες στην παροχή ολόκληρου του φάσματος στην οικονομική πληροφορία. Ο χρήστης επισκεπτόμενος τις συγκεκριμένες διαδικτυακές σελίδες μπορεί να διαβάσει όλες τις οικονομικές ειδήσεις που διαχέονται στον πλανήτη σε πραγματικό χρόνο. Έχει τη δυνατότητα να αναγνώσει τις πιο πολυδιαβασμένες ειδήσεις αλλά και να

συγκεκριμενοποιήσει το newsfeed του στα προσωπικά του θέλω, είτε πρόκειται για ολόκληρους χρηματοοικονομικούς κλάδους όπως για παράδειγμα ο ασφαλιστικός κλάδος, είτε πρόκειται για μεμονωμένα χρεόγραφα για τα οποία ενδιαφέρεται, δηλαδή για έναν μετοχικό τίτλο, ένα ομόλογο ή ένα παράγωγο. Συγχρόνως έχει πρόσβαση σε ιστορικά δεδομένα όπως τιμές κλεισίματος παρελθουσών ετών, σε μια τεράστια ποικιλία από γραφήματα, αλλά και σε πραγματικό χρόνο τις κινήσεις όλων των οικονομικών προϊόντων. Συν τοις άλλοις, παρέχεται η δυνατότητα στο χρήστη να δημιουργήσει το δικό του «διαδικτυακό» χαρτοφυλάκιο στο οποίο να τοποθετήσει τους τίτλους της επιλογής του και κατόπιν να ενημερώνεται για τις επιλογές του. Το κυριότερο από όλα όμως είναι το γεγονός ότι όλο αυτό το πλήθος των πληροφοριών παρέχεται δωρεάν.

Φυσικά, για το στόχο της εργασίας αυτής η εστίαση βρίσκεται στον τεράστιο πλούτο των ειδήσεων και απόψεων που παρέχονται μέσω των σελίδων αυτών στο κοινό.



Εικόνα 3.3: Η δομή της Major Financial Mass Media Web Page της Morningstar.com

Μερικά χαρακτηριστικά παραδείγματα τα οποία και είναι κορυφαία στην επισκεψιμότητα σύμφωνα με το Alexa.com είναι:

Το MSN Money (moneycentral.msn.com) – προσφέρει οικονομικές ειδήσεις, ένα εκπαιδευτικό / επενδυτικό κέντρο, έρευνα και αξιολόγηση μετοχών, ομολόγων και

αμοιβαίων κεφαλαίων, ώστε να βοηθήσει με τις προσωπικές επενδύσεις και περισσότερα.

Το CNNMoney.com (money.cnn.com) - όπως και το MSN Money, είναι μια ολοκληρωμένη οικονομική μονάδα, συμπεριλαμβανομένων και των τρέχουσων ειδήσεων που αφορούν την οικονομία και την επενδυτική κοινότητα. Η ιστοσελίδα παρέχει επίσης έναν πολύ χρήσιμο, βήμα-βήμα, προσωπικό οδηγό χρήματος.

Το Morningstar (www.morningstar.com) – μια ολοκληρωμένη και αξιόπιστη ιστοσελίδα για τα αποθέματα, τα ομόλογα, το ETF (Exchange Traded Funds) και τα αμοιβαία κεφάλαια, αξιοποιώντας το σύστημα αξιολόγησης Morningstar για τον έλεγχο των επενδύσεων.

3.2.3 Social Networks

Είναι μια σειρά από τρόπους που οι επιχειρήσεις επί του παρόντος εφαρμόζουν την ανάλυση συναισθήματος μεταξύ όλων των κοινωνικών μέσων μαζικής ενημέρωσης. Κάποιοι επιλέγουν για αυτοματοποιημένες λύσεις που εκτελούνται από το κοινωνικό λογισμικό παρακολούθησης των μέσων ενημέρωσης, όπως Radian6, Alterian, Spiral16 και άλλα. Αλλά ακόμη και στο ολοένα ψηφιακό, ηλεκτρονικό κόσμο μας, τα ανθρώπινα όντα εξακολουθούν να διαδραματίζουν σημαντικό ρόλο όταν πρόκειται για την ερμηνεία των εκατομμυρίων κομματιών υποκειμενικού περιεχομένου που οι χρήστες αποστέλλουν κάθε μέρα. Άλλες εταιρείες έχουν στραφεί προς το crowdsourcing, προκειμένου να ολοκληρώσουν τα έργα ανάλυσης συναισθήματος. Αλλά τόσο τα ανοικτά στα πλήθη όσο και το κοινωνικό λογισμικό παρακολούθησης παράγουν διαφορετικό βαθμό ακρίβειας. Η πρόκληση είναι το πώς να συλλάβει κανείς, να αναλύσει και να ερμηνεύσει κανείς ακριβέστερα όλο το κοινωνικό περιεχόμενο των μέσων ενημέρωσης που σχετίζονται με την αγορά και τα χρηματιστηριακά προιόντα.

Τα social media αποτελούν στη σημερινή κοινωνία της πληροφορίας άκρως πολυσύχναστες διαδικτυακές πλατφόρμες και τεράστιες πηγές συναισθήματος. Συγκεκριμένα στους τόπους του Facebook και του Twitter διαθέτουν εκατομμύρια χρήστες, λογαριασμούς τους οποίους χρησιμοποιούν για να εκφράσουν τα

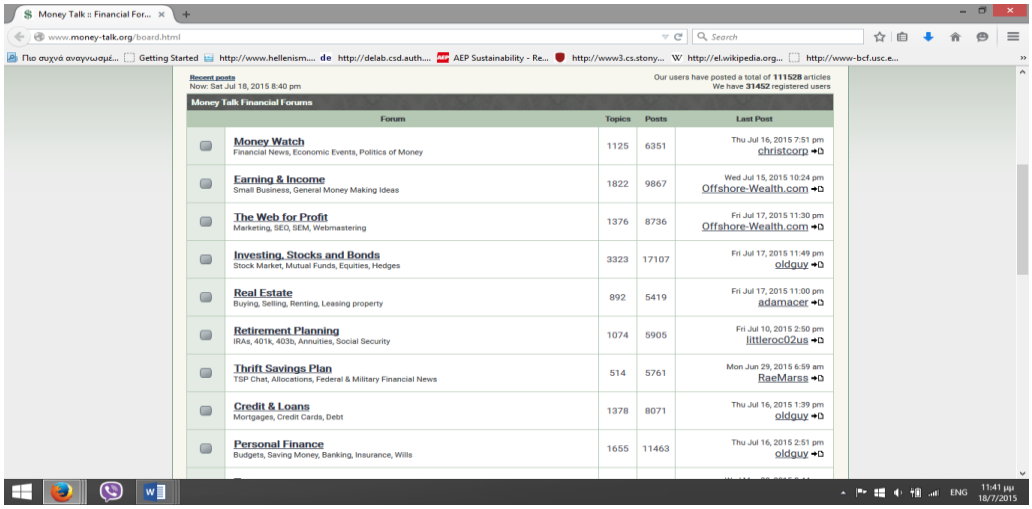
συναισθήματά τους και την αίσθησή τους για τα τεκταινόμενα στις χρηματοπιστωτικές αγορές.

Το Facebook πέραν των προσωπικών λογαριασμών, από τους οποίους μπορεί να μετρηθεί εύκολα η αναγνωσιμότητα και η αναγνωρισιμότητά τους, άρα και το impact της γνώμης του κάθε ενός, υπάρχουν κοινότητες ολόκληρες επενδυτών και φορείς γνώμης για την εξαγωγή και ποσοτικοποίηση του financial sentiment. Οι αναρτήσεις και τα σχόλια που γράφονται στην πλατφόρμα αυτή μπορούν εύκολα να κατηγοριοποιηθούν και να αναλυθούν και να εξαχθούν με αυτόν τον τρόπο συμπεράσματα της αυθόρμητης κρίσης της κοινής γνώμης για την αγορά.

Η πλατφόρμα του Twitter έχει δύο επιπλέον ιδιαιτερότητες. Ενώ αποτελεί και αυτός ο διαδικτυακός τόπος μια τεράστια βάση δεδομένων συναισθήματος και φορέας γνωμών για τα οικονομικά δεδομένα και προϊόντα, ο περιορισμός των χαρακτήρων, άρα και των λέξεων που θα χρησιμοποιηθούν, είναι πεπερασμένος και μικρός. Το γεγονός αυτό ουσιαστικά συμπυκνώνει τη γνώμη που θα εκφραστεί σε ένα μεγάλο βαθμό. Η άλλη ιδιαιτερότητα έγκειται στο γεγονός ότι υπάρχει η δυνατότητα της χρήσης των hashtags, που αναλαμβάνουν το ρόλο της σηματοδότησης του κάθε tweet, ανάλογα με το περιεχόμενο και το σκοπό της κάθε ανάρτησης. Μπορεί για παράδειγμα να είναι το όνομα μιας μετοχής ή και κάποιου χρηματιστηριακού δείκτη. Συγκεκριμένα το twitter χρησιμοποιεί ειδικό hashtag για τις μετοχές αντί της δίεσης (#) το σύμβολο του δολαρίου ($). Η ανάλυση συναισθήματος στα δεδομένα του Twitter και άλλων παρόμοιων microblogs αντιμετωπίζει πολλές νέες προκλήσεις λόγω και της ακανόνιστη δομή του εν λόγω περιεχομένου.

Σημαντική πηγή άντλησης απόψεων για το οικονομικό και χρηματοοικονομικό γίγνεσθαι στον πλανήτη αποτελούν τα εξειδικευμένα φόρα συζήτησης που υπάρχουν στο Διαδίκτυο. Χαρακτηριστικό παράδειγμα αποτελεί το «World Banking Forum» που ανήκει στον διαδικτυακό τόπο του linkedin, που συνδέεται και με την προηγούμενη παράγραφο, και αριθμεί 3750 μέλη της χρηματοδοτικής αγοράς στο οποίο γίνεται καθημερινή ανταλλαγή απόψεων για την τραπεζική, τα αμοιβαία κεφάλαια και τις επενδύσεις. Άλλα παραδείγματα είναι το «Wilmott forums», το http://www.thefinanceforums.com/ και το http://www.money-talk.org/board.html. Η δομή των συγκεκριμένων ιστοτόπων διευκολύνει σε τεράστιο βαθμό το φιλτράρισμα των επιθυμητών πληροφοριών και την επιλογή προς ανάλυση των απόψεων που

εκφράζονται καθώς τα ίδια τα φόρα διαστρωματώνονται σε κατηγορίες και υποκατηγορίες. Φυσικά, σχεδόν όλα τα φόρα γενικού ενδιαφέροντος περιέχουν υποκατηγορίες συζητήσεων για τα χρηματοοικονομικά και μπορούν να αποτελέσουν πηγή γνώμης με εύκολο τον εντοπισμό τους αλλά φυσικά είναι λιγότερο αξιόπιστη.



Εικόνα 3.4: Η δομή του οικονομικού forum money-talk.org

3.2.4 Blogs και Εξειδικευμένα Blogs

Ένας τρόπος διαχείρισης της διαδικτυακής φήμης ενός ατόμου, μιας εταιρίας ή ενός προϊόντος είναι με αναζήτηση σε δημοφιλή blogs, ώστε να εντοπιστούν άρθρα, που μιλούν θετικά ή αρνητικά για το αντίστοιχο θέμα. Ο τομέας, που ασχολείται με την εύρεση κειμένων άποψης από blogs, ονομάζεται εξόρυξη σε blogs και είναι μια υποκατηγορία της εξόρυξης άποψης.

Σκοπός της εξόρυξης από blogs είναι ο εντοπισμός αναρτήσεων σε blogs, που εκφράζουν άποψη για ένα θέμα-στόχο, που έχει δοθεί. Αυτός ο στόχος μπορεί να είναι παραδοσιακά ένα όνομα, είτε αυτό το όνομα αφορά ένα άτομο, μια τοποθεσία ή έναν οργανισμό. Ειδάλλως μπορεί να είναι μια έννοια, ένα όνομα προϊόντος ή ένα γεγονός.

Για να εφαρμοστεί η εξόρυξη απαιτείται μια συλλογή από blogs, στα οποία θα γίνει η προσπάθεια εντοπισμού σχετικών αναρτήσεων. Η αναζήτηση μπορεί να γίνει είτε στις σελίδες των blogs είτε μέσω των RSS feeds τους. Οι πρώτες αναρτήσεις, που θα συγκεντρωθούν σε ένα ανάλογο σύστημα, μπορούν κατόπιν να χρησιμοποιηθούν ως δεδομένα εκπαίδευσης σε κάποια τεχνική ανάλυσης συναισθήματος.

Ένα μεγάλο ζήτημα είναι η εύρεση του κυρίως κειμένου μιας ανάρτησης των blogs.

Υπάρχουν περιπτώσεις, όπου δεν αποτελεί πρόβλημα, όπως όταν η αναζήτηση γίνεται μέσω των RSS feeds. Παρόλο που το στοιχείο description (περιγραφή), που περιέχουν τα feeds, είναι προορισμένο για να περιλαμβάνει μια σύνοψη ή ένα απόσπασμα του άρθρου, έχει καταλήξει να περιέχει όλο το κείμενο, ενισχύοντας με αυτόν τον τρόπο την εξόρυξη. Όμως, στις περιπτώσεις, όπου τα RSS feeds δεν υποστηρίζουν την ενσωμάτωση του κυρίως κειμένου στο feed ή/και η αναζήτηση γίνεται στις σελίδες των blog, τότε χρειάζονται νέες στρατηγικές.

Δυστυχώς οι σελίδες των blogs είναι ανάκατες με κάθε λογής έξτρα πληροφορίας, πέρα από το άρθρο και τα σχόλια των αναγνωστών. Περιέχουν συχνά έναν κατάλογο σχολιασμού των προηγούμενων αναρτήσεων, λίστες με συναφείς σελίδες, μπάρες πλοήγησης, πλάγιες μπάρες, διαφημίσεις κλπ.. Αν κατασκευαστεί ο κατάλογος της σελίδας σα μία κανονική HTML σελίδα, τότε όλο το κείμενο από τα κομμάτια, που αναφέρθηκαν, θα καταλήξει μέσα στον κατάλογο, πράγμα, που θα οδηγήσει σε αποτελέσματα με μικρή σχετικότητα.

Για να εντοπιστεί το κατάλληλο περιεχόμενο της ανάρτησης μέσα σε ένα blog, προτείνονται τρεις στρατηγικές. Η πρώτη στρατηγική είναι η χρήση του στοιχείου content (περιεχόμενο) από το feed, όταν είναι διαθέσιμο. Για την πραγματοποίηση αυτής της στρατηγικής απαιτείται η κατασκευή ενός καταλόγου για τα RSS feeds. Όταν κατατάσσεται στον πίνακα ένας σύνδεσμος του blog, γίνεται έλεγχος αν το feed, από το οποίο προήλθε, περιέχει στοιχείο content, ώστε να χρησιμοποιηθεί αυτό ως κυρίως κείμενο.

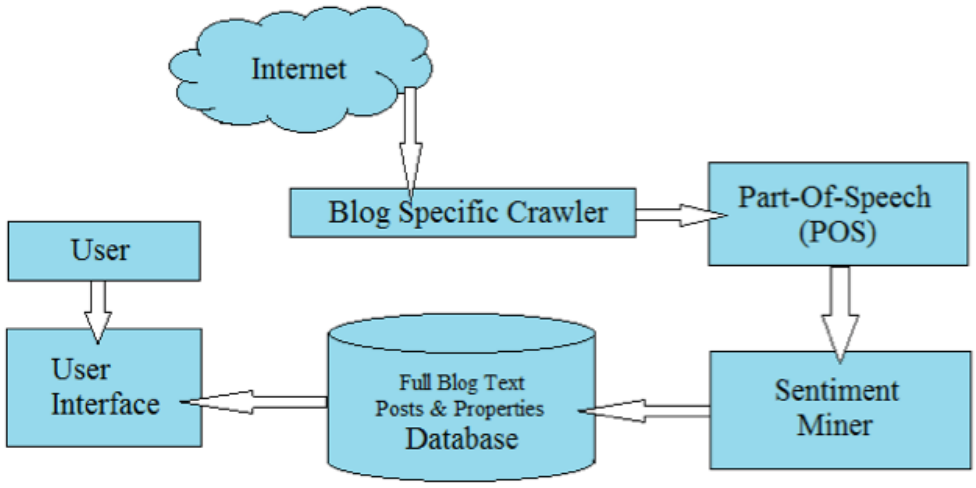
Η δεύτερη στρατηγική προτείνει την ειδική αντιμετώπιση blogs, που παράχθηκαν με τη χρήση προγραμμάτων, που ακολουθούν σαφώς ορισμένες σημάνσεις, επιτρέποντας έτσι στο περιεχόμενο της ανάρτησης να εντοπιστεί.

Η τρίτη στρατηγική χρησιμοποιείται για τον χειρισμό των υπόλοιπων περιπτώσεων, αποκλείοντας στοιχεία, που δε θεωρούνται μέρος της ανάρτησης. Αφαιρούνται στοιχεία div, των οποίων η κλάση περιέχει τη λέξη link ή αρχίσει με τη λέξη side ή τελειώνει με τη λέξη bar. Ακόμη, αποκλείονται στοιχεία div με id, που περιέχουν τη λέξη header ή nav. Ευτυχώς, αρκετά προγράμματα παραγωγής blog χρησιμοποιούν τέτοιου είδους γλώσσα σήμανσης, άρα με μία λίστα 50 περίπου στοιχείων προς αποκλεισμό, αποφεύγουν το μεγαλύτερο κομμάτι των άσχετων τμημάτων.

Είναι δυνατόν να προσφέρεται μία λίστα από blogs, από τα οποία θα μπορεί να διαλέγει ο χρήστης να ενημερώνεται για το θέμα, που έχει ήδη επιλέξει. Ειδάλλως, μπορεί το σύστημα να του προτείνει να διαβάσει κάποια blogs, τα οποία έτυχε να αναφέρουν κάτι σχετικό με το θέμα, που παρακολουθεί.

Οι επιλογές τόσο του θέματος, όσο και των blogs προς παρακολούθηση, μπορούν να παρέχονται με μία καινοτόμα διεπαφή. Επίσης, η διεπαφή θα προσφέρει τα αποτελέσματα με γραφική απεικόνιση ή με διαχωρισμό ομάδων ανάλογα με το συναίσθημα, που εκφράζουν οι αναρτήσεις.

Προτεινόμενο σύστημα για εξόρυξη από blog



Εικόνα 3.5: Προτεινόμενο σύστημα για εξόρυξη ειδήσεων από blog

Για την εξόρυξη γνώμης ως προς τα οικονομικά μεγέθη και προϊόντα, μπορούν επίσης να χρησιμοποιηθούν blogs διακεκριμένων παικτών, σημαινόντων επενδυτών, έγκριτων δημοσιογράφων και καθηγητών του χώρου της οικονομίας και των χρηματοποικονομικών. Η γνώμη αυτών των παικτών έχει βαρύνουσα σημασία και έχει τη δυνατότητα επιρροής στην αγορά.

3.2.5 Δελτία οικονομικών εξελίξεων

Όλοι σχεδόν οι χρηματοπιστωτικοί οίκοι στην κάθε χώρα του πλανήτη διαθέτουν στο οργανόγραμμά τους, τμήματα οικονομικών μελετών. Οι διευθύνσεις αυτές εκδίδουν

ανά τακτά χρονικά διαστήματα ενημερωτικά δελτία για τα χρηματοπιστωτικά προϊόντα της αγοράς και τις οικονομικές εξελίξεις εν γένει. Τα δελτία αυτά, παρατηρείται ανάλογα και με τον εκδότη, μπορούν να είναι ημερήσια, εβδομαδιαία, μηνιαία, τριμηνιαία, εξαμηνιαία ή και ετήσια. Οι εκθέσεις αυτές αντικατοπτρίζουν τη γνώμη κορυφαίων τραπεζικών κολοσσών όπως για παράδειγμα η Goldman Sachs με το εβδομαδιαίο "Weekly Monitor" ή η τεράστια επενδυτική τράπεζα VANGUARD η οποία διαχειρίζεται κεφάλαια ύψους 3 τρισεκατομμυρίων δολαρίων με το ετήσιο **"**Vanguard's economic and investment outlook**"**. Προσφέρουν μία πυκνογραμμένη άποψη για την αγορά και την κατεύθυνσή της, επισημαίνουν τους κινδύνους και τις ευκαιρίες για επενδύσεις σε μεμονωμένα χρεόγραφα ή και ολόκληρους κλάδους. Χαρακτηριστικό των εγγράφων αυτών είναι η εξαιρετικά συμπυκνωμένη πληροφορία και άποψη που εμπεριέχουν καθώς και το γεγονός ότι το περιεχόμενό τους είναι αποκλειστικά περί των χρηματοοικονομικών και χρηματοπιστωτικών αγορών.

ΚΕΦΑΛΑΙΟ 4: ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΠΛΗΡΟΦΟΡΙΩΝ

4.1 Μέθοδοι φιλτραρίσματος ειδησεογραφίας

Κλασικές προσεγγίσεις στην ανάκτηση κειμένου και κατηγοριοποίηση έχει μέχρι στιγμής επικεντρωθεί στην εξόρυξη και ανάλυση πραγματικών στοιχείων, όπως εκδηλώσεις, φορείς και περιουσίες. Βασικά χρησιμοποιούν φυσικές μεθόδους επεξεργασίας του λόγου και των τεχνικών προκειμένου να εξαχθούν αντικειμενικά χαρακτηριστικά που βοηθούν στην ταξινόμηση και κατηγοριοποίηση των κειμένων και των εκφράσεων (Pang et al, 2002) με ιδιαίτερη έμφαση στις γλωσσικές λειτουργίες προκειμένου να αυξηθεί η απόδοση. Καθώς τα γλωσσικά χαρακτηριστικά (Gamon, 2004, Matsumoto et al, 2005) παρουσιάζουν χαρακτηριστικά κίνητρα, τα περισσότερα από αυτά βασίζονται στις πληροφορίες σχετικά με την ενημέρωση της πορείας εξάρτησης. Περαιτέρω γλωσσικά χαρακτηριστικά, όπως μέρη του λόγου, μορφές άρνησης, ρήματα και σημασιολογικές πληροφορίες (από το WordNet για παράδειγμα) έχουν πρόσφατα επίσης διερευνηθεί (Wiegand και Klakow, 2009).

Μία προσέγγιση είναι με τη χειροκίνητη κατασκευή ενός λεξικού με επίθετα γνωστού προσανατολισμού. Με αυτόν τον τρόπο, αν μία πρόταση περιέχει επίθετα συγκεκριμένης πολικότητας, τότε η πρόταση έχει αυτόν τον προσανατολισμό. Αν υπάρχει μια αρνητική λέξη (πχ. «δεν»), τότε αντιστρέφεται η πολικότητα.

Μία άλλη προσέγγιση είναι με την εκπαίδευση ενός κατηγοριοποιητή. Για παράδειγμα, για την κατηγοριοποίηση συναισθήματος για κριτικές ταινιών θα χρησιμοποιηθούν δεδομένα κριτικής από κάποια σχετική ιστοσελίδα, όπως το IMDB (Internet Movie Database). Κριτικές με χαμηλές βαθμολογίες χρησιμοποιούνται για την εκπαίδευση αρνητικού κατηγοριοποιητή και κριτικές με υψηλές βαθμολογία για την εκπαίδευση θετικού. Στη συνέχεια, αυτοί οι κατηγοριοποιητές χρησιμοποιούνται για να καθορίσουν συναισθήματα.

Μία ενίσχυση αυτών των προσεγγίσεων είναι με την ενσωμάτωση ετικετών Part-Of-Speech στις λέξεις των κείμενων, δηλαδή κάθε λέξη του κειμένου παίρνει σαν ετικέτα το μέρος του λόγου, που είναι. Παραδείγματος χάριν, το επίθετο "άριστος" θα πάρει την ετικέτα _JJ για επίθετα και θα γίνει "άριστος". Με αυτόν τον τρόπο διευκολύνεται η εύρεση των κατάλληλων μερών του λόγου, σύμφωνα με τα οποία γίνεται η ανάλυση

συναισθήματος

Ένα μεγάλο μέρος της προηγούμενης έρευνας επικεντρώνεται στον προσδιορισμό των χαρακτηριστικών των μεταφερομένων απόψεων βάσει δεδομένων με μορφή επεξεργασίας κειμένου που κυμαίνεται από λόγια, εκφράσεις και έγγραφα. Σε επίπεδο λέξεων, που θεωρείται ως το πιο εξελιγμένο επίπεδο, θεωρείται ότι υπάρχει μια μονόπλευρη άποψη για ολόκληρη την πρόταση ή την έκφραση, τυπικά η περίπτωση της απάντησης σε ερωτήσεις γνώμης ή τις ανθρώπινες συζητήσεις. Υπάρχουν κυρίως δύο τύποι ερευνητικών προσεγγίσεων που στοχεύουν στην επίλυση αυτού του προβλήματος: οι στατιστικές και σημασιολογικές προσεγγίσεις. Οι στατιστικές προσεγγίσεις κάνουν χρήση τεχνικών εκμάθηση για την κατηγοριοποίηση της σημασιολογικής πόλωση γνώμης σε θετικές και αρνητικές τάξεις και στην προσέγγιση της αξίας της έντασής τους. Οι τεχνικές αυτές διαφέρουν τις μεθόδους πιθανοτήτων (όπως Naive Bayes, Maximum Entropy), τη γραμμική διάκριση (όπως Support Vector machine) και μη παραμετρικές ταξινομητές (όπως το K-Nearest Neighborhood), καθώς και μέθοδοι σκορ ομοιότητα (όπως ταίριασμα φράσεων, απόσταση διανύσματος, συχνότητα και στατιστικές μετρήσεις βάρους). Για το σκοπό αυτό, έχει κατασκευαστεί ένας μεγάλος αριθμός μέτρων για την κατηγοριοποίηση των νέων όπως, για παράδειγμα, MPQL (Wilson et al., 2005), Movie Review Data (Pang et al., 2002), SentiWordNet (Esuli and Sebastiani, 2006), WordNet-Affect (Strapparava and Valitutti, 2004)), Product Review (Yi et al., 2003), Book Review (Gamon and Aue, 2005), Whissell's Dictionary of Affect Language (Whissell, 1989), Linguistic Inquiry and Word Count Dictionary (LIWC2001) (Pennebaker et al., 2001), εκτός από το τεράστιο ποσό των διαθέσιμων δεδομένων προσανατολισμού συναισθημάτων που βρίσκονται σε φόρουμ, blogs, chat rooms, κριτικές, συζητήσεις και e-γνώμεων σε ιστοσελίδες.

Αυτόματες μέθοδοι σχολιασμού των συναισθημάτων σε επίπεδο λέξης μπορεί να ομαδοποιηθούν σε δύο μεγάλες κατηγορίες: (1) προσεγγίσεις «σώματος» και (2) προσεγγίσεις που στηρίζονται σε λεξικά.

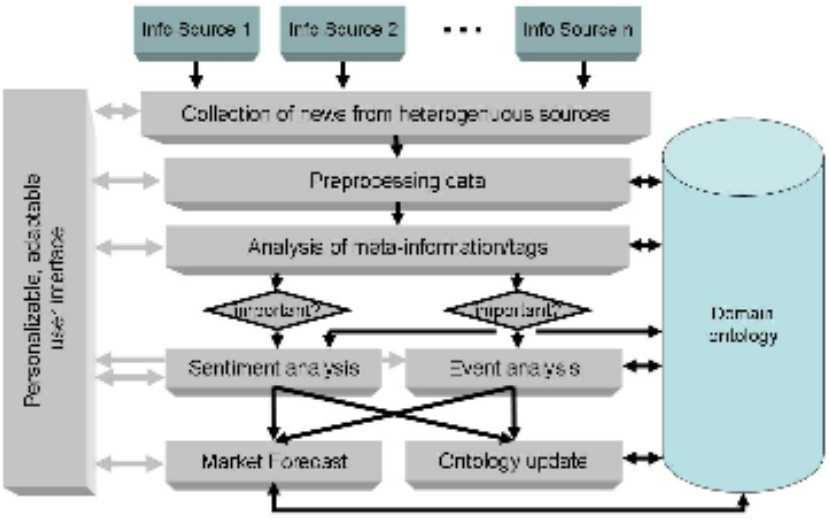
Οι traders των χρηματοπιστωτικών αγορών έρχονται αντιμέτωποι με το πρόβλημα ότι πάρα πολλές πληροφορίες είναι διαθέσιμες από διάφορες, ετερογενείς πηγές, όπως ειδησεογραφικά πρακτορεία, φόρουμ, blogs και συνεργατικά εργαλεία. Για να ληφθούν ακριβείς αποφάσεις για τις συναλλαγές, οι traders πρέπει να φιλτράρουν τις

σχετικές πληροφορίες αποτελεσματικά, έτσι ώστε να είναι σε θέση να αντιδρούν σε νέες πληροφορίες σε εύθετο χρόνο.

Μια ποικιλία συστημάτων για την πρόβλεψη των περιουσιακών στοιχείων εξέλιξης των τιμών με βάση πρόσφατα νέα που δημοσιεύονται έχουν αναπτυχθεί (βλ. [1] για μια γενική επισκόπηση). Αυτά τα συστήματα στηρίζονται στην ταξινόμηση κειμένου, όπου προκύπτουν οι κατηγορίες-στόχοι από οικονομικά στοιχεία. Ωστόσο, έχουν δύο σημαντικές αδυναμίες: (i) τις προσδοκίες, οι οποίες σε μεγάλο βαθμό επηρεάζουν την εξέλιξη των τιμών των περιουσιακών στοιχείων, και (ii) ποσοτικά στοιχεία (όπως η αξία των καταβληθέντων μερισμάτων ή το ποσό του ετήσιου κέρδους), η οποία επιτρέπει την ποσοτικοποίηση της αναμενόμενης μεταβολής των τιμών, δεν θεωρούνται σε αυτά τα συστήματα.

Η online σύνδεση μεθόδων ανίχνευσης, μόνο προσπαθούν να εντοπίσουν νέες εκδηλώσεις κυρίως με την ομαδοποίηση τεχνικών, χωρίς να προσπαθούν να τα επισημοποιήσουν σημασιολογικά.

Ενώ οι Das και Chen χρησιμοποιούν γλωσσικά στοιχεία για την ταξινόμηση των μηνυμάτων σε αρνητικά και θετικά και στη συνέχεια εξετάζουν τη συσχέτιση με τις μεταβολές των τιμών των μετοχών, οι Koppel et al. χρησιμοποιούν τις μεταβολές των τιμών με τη χρήση υλικού για τον προσδιορισμό των θετικών και αρνητικών ειδήσεων από τις οποίες στη συνέχεια περιγράφουν τα χαρακτηριστικά που μπορούν να εξαχθούν. Ένα εργαλείο ανάλυσης της ειδησεογραφίας είναι το ακόλουθο:



Εικόνα 4.1: Εργαλείο ανάλυσης της ειδησεογραφίας σε global επίπεδο

Το πρώτο συστατικό του είναι υπεύθυνο για τη συλλογή των ειδήσεων από το διάφορες, ετερογενείς πηγές. Το στοιχείο αυτό παρακολουθεί έναν τεράστιο αριθμό των σχετικών πηγών για νέες πληροφορίες και καθιστά διαθέσιμη την προ-επεξεργασία των δεδομένων. Το τελευταίο προβάλει τα διαθέσιμα μεταδεδομένα σε μια ενιαία εκπροσώπηση τέτοια ώστε όλα τα δεδομένα είναι επεξεργάσιμα με τον ίδιο τρόπο στα επόμενα στάδια. Καθορίζεται μια οντολογία για κάθε πηγή ειδήσεων μεταδεδομένων και μπορεί να χαρτογραφήσει αυτές τις οντολογίες. Εάν τα μεταδεδομένα είναι διαθέσιμα, εξάγονται ορισμένοι σχολιασμοί του περιεχομένου από τις ειδήσεις. Πάντως, μόνο πολύ αποτελεσματικές μέθοδοι μπορούν να εφαρμοστούν εδώ ως εξαγωγή πληροφοριών.

Η ανάλυση της συνιστώσας μεταδεδομένων θα εξετάσει τα μεταδεδομένα για να φιλτράρει τις σχετικές ειδήσεις. Αποφασίζεται αν μια είδηση περιέχει προσδοκίες – βουλεύσεις που αφορούν μελλοντικά γεγονότα και έτσι υποβάλλεται σε επεξεργασία από την ανάλυση συναισθήματος συνιστώσα και εάν αυτή περιέχει πληροφορίες σχετικά με ένα πραγματικό γεγονός και θα υποβληθεί σε επεξεργασία από τη συνιστώσα της ανάλυσης γεγονότων. Είναι πιθανό μια είδηση να επεξεργάζεται από τα δύο συστατικά ή ότι μια είδηση δεν είναι σημαντική και ως εκ τούτου δεν θα πρέπει να υποστεί περαιτέρω επεξεργασία. Αυτές οι περιγραφές χρησιμοποιούνται από τις προβλέψεις της αγοράς ως συστατικό για την πρόβλεψη του αντίκτυπου της είδησης στην αγορά με την ποσοτικοποίηση της διαφοράς των πληροφοριών που δημοσιεύονται από τις προσδοκίες και την τρέχουσα κατάστασή της.

Η οντολογία domain είναι η ραχοκοκαλιά των τριών εξαρτημάτων που περιγράφηκαν προηγουμένως. Περιγράφει την τρέχουσα κατάσταση της αγοράς και τις προσδοκίες που αφορούν μελλοντικά γεγονότα. Αυτή τη στιγμή, προσπαθούμε να προσδιορίσουμε τα πιο έξυπνη χαρακτηριστικά ειδήσεις. Μπορεί να αναπτυχθεί ένα γραμμικό μοντέλο παλινδρόμησης που να προβλέπει τις απαντήσεις της αγοράς με βάση τα χαρακτηριστικά του κειμένου.

Αυτή είναι η τεχνική της επιλογής, καθώς η προβλεπόμενη επίπτωση μπορεί να ποσοτικοποιηθεί και καθώς η επιρροή του κάθε χαρακτηριστικού γνωρίσματος για το αποτέλεσμα μπορεί εύκολα να βρεθεί. Το μοντέλο αυτό βοηθά στον εντοπισμό των πληροφοριών, ενώ παρέχει επίσης κάποια δυνατότητα πρόβλεψης που μπορεί να χρησιμεύσει ως βασική γραμμή για την αξιολόγηση των πιο περίτεχνων μεθόδων. Το

εργαλείο μπορεί να είναι προσωποποιημένο και προσαρμόσιμο, με την έννοια ότι οι χρήστες μπορούν να καθορίσουν τις προτιμήσεις τους, π.χ. εταιρείες που ενδιαφέρονται ιδιαίτερα.

Ο εντοπισμός του συναισθήματος σε ένα άρθρο μπορεί γενικά να είναι μια δύσκολη και χρονοβόρα διαδικασία. Ευτυχώς στη χρηματοδότηση, το συναίσθημα ενός προϊόντος δεν μπορεί αξιόπιστα να συνδεθεί με την εξέλιξη των σχετικών τιμών της αγοράς μετοχών κατά το χρονικό διάστημα γύρω από τη δημοσίευση. Στον προτεινόμενο αλγόριθμο, ένα άρθρο ταξινομείται ως θετικό όταν αυτό συμβαίνει κατά τη διάρκεια μιας χρονικής περιόδου που συνδέεται με μια ευνοϊκή για αγορά απάντηση. Η ανταπόκριση της αγοράς είναι ποσοτική και μπορεί εύκολα να προσδιοριστεί, η διαδικασία επιτρέπει να εφαρμόζεται αυτόματα σε ένα μεγάλο σώμα των δεδομένων και να κάνει προβλέψεις για το συναίσθημα των προηγουμένως αναγνωσμένων άρθρων των επιχειρήσεων.

Οι πηγές δεδομένων φαίνεται στα τετράγωνα και οι ανάγκες αντιμετωπίζονται με τη μετάβαση από τη μια πηγή στην επόμενη που δείχνονται δίπλα από τα βέλη. Αυτό που δείχνεται σε παρένθεση κάτω από τις πηγές είναι τα κριτήρια που χρησιμοποιούνται για τη συλλογή αντικειμένων. Οι πηγές RSS και το CNN money λαμβάνονται υπόψη για γενικά άρθρα σχετικά με την οικονομία, ενώ το Yahoo! news έψαξε για συγκεκριμένη εταιρεία (ες). Συνολικά, το CNN Money βρέθηκε να έχει ένα μεγάλο όγκο των σχετικών οικονομικών άρθρων που αφορούν την οικονομία, γι' αυτό επιλέχθηκε ως η προτιμώμενη και μοναδική πηγή δεδομένων.

Η συγκομιδή των δεδομένων επιτυγχάνεται με την αναζήτηση στο cnnmoney.com για τα άρθρα των επιχειρήσεων τα οποία, στη συνέχεια, φιλτράρονται κατά ημερομηνία. Μια διεπαφή web χρησιμοποιείται για την εξαγωγή του τίτλου, την περίληψη, ή / και το πλήρες κείμενο κάθε άρθρου και ο απορρέων αλγόριθμος του Porter χρησιμοποιείται μαζί με μια συγκεκριμένη λίστα λεξιλογίου για να δημιουργηθεί ένα ιστόγραμμα μαρκών. Τα σημαντικότερα θετικά λόγια ήταν αυτά που είχαν τη μεγαλύτερη αναλογία μεταξύ $j \mid y = 1$ και $j \mid y = 0$ κατά την κατάρτιση Naïve Bayes και το αντίστροφο ίσχυε για μια πιο σχετική αρνητική λέξη.

Τόσο ο Naïve Bayes και οι Vector Machine Support (SVM) αλγόριθμοι μπορούν να χρησιμοποιηθούν για να εξαχθεί το μοντέλο πρόβλεψης από τα δεδομένα που

συλλέγονται. Κάθε άρθρο σημάνθηκε σύμφωνα με την ημερομηνία της δημοσίευσης χρησιμοποιώντας το ακόλουθο μέτρο που βασίζεται στο δείκτη Dow Jones Industrial Average (DJI): Άνοιγμα αξίας την επόμενη μέρα μείον το κλείσιμο αξίας της προηγούμενης ημέρας. Τα άρθρα που δεν δείχνουν αύξηση ή μείωση μεγαλύτερη από τα κατώτατα όρια που απορρίπτονται και τα όρια προσαρμόζονται μέχρι τον αριθμό των θετικών και αρνητικών παραδειγμάτων που αποκτήθηκαν.

Οι Naïve Bayes δοκιμές δίνουν ένα μέτρο μεταξύ μηδέν και ένα όταν ένα απλό άρθρο κατατάσσεται ως θετικό. Ένας τρόπος για να ενεργήσει σύμφωνα με αυτά τα αποτελέσματα θα ήταν να τοποθετηθεί ένα όριο σε 0,5 ταξινόμησης του άρθρου ως θετικό εάν το αποτέλεσμα είναι μεγαλύτερα από 0,5 και αρνητικό αν είναι διαφορετικά. Μια άλλη επιλογή θα ήταν να επισημανθεί ένα άρθρο ως θετικό, αρνητικό ή αβέβαια περίπτωση κατά την οποία το μέτρο είναι κοντά στο 0,5. Χρησιμοποιώντας το τελευταίο μέτρο, άρθρα των οποίων ο αλγόριθμος είναι αβέβαιος αφαιρούνται από το σύνολο δεδομένων δοκιμής. Τα υπόλοιπα άρθρα επισημαίνονται ως θετική ή αρνητική αυτοπεποίθηση. Λαμβάνοντας υπόψη ότι ένας επενδυτής ενδιαφέρεται για τη λήψη αποφάσεων χρησιμοποιώντας όλα τα άρθρα που δημοσιεύονται σε μια δεδομένη χρονική περίοδο, ο αριθμητικός μέσος όρος όλων των άρθρων σε μια δεδομένη ημέρα των αποδόσεων για μια μέρα. Αυτές οι προβλέψεις μπορούν στη συνέχεια να επικυρωθούν από τα δεδομένα της αγοράς μετοχών.

Η εξέταση των παραμέτρων του προβλήματος, μπορεί να αποτελέσει ένδειξη αυτή του μηχανισμού πρόβλημα ταξινόμησης της μάθησης. Δοκιμές δείχνουν ότι μία συμβολική ανάλυση άρθρων με πλήρες κείμενο δίνει μια υψηλή αυτοπεποίθηση αλλά και χαμηλή ακρίβεια. Μια κίνηση σε περιλήψεις άρθρων μειώνει δραστικά την εμπιστοσύνη. Μια εξήγηση είναι ότι τα άρθρα πλήρους κειμένου παρέχουν πολλές λέξεις που δεν είναι σχετικές με την ανάλυση του συναισθήματος. Προχωρώντας στην εξέταση ζεύγη token και περιλήψεις άρθρων οδηγεί σε μεγάλη αύξηση των επιδόσεων. Ένα ζεύγος λέξεων προσφέρει εγγενώς περισσότερες πληροφορίες επειδή δεσμεύει και τις δύο λέξεις, καθώς και τη σχέση τους. Σαφώς, τα ζεύγη λέξεων δίνουν πιο σχετικές πληροφορίες από τα απλά λόγια για το οικονομικό κλίμα του άρθρου.

Η ανάλυση της επιλογής λεξιλογίου δείχνει ότι οι ενιαίες φράσεις είναι ασαφής, επειδή λαμβάνονται έξω από το πλαίσιο, ενώ τα ζεύγη λέξεων διατηρούν μια πιο ολοκληρωμένη άποψη με περισσότερο νόημα. Περαιτέρω η μείωση του λεξιλογίου με

ανθρώπινη συμβολή μειώνει την εμπιστοσύνη, αλλά και αυξάνει την ακρίβεια. Η διαίσθηση είναι ότι υπάρχει μια μικρότερη σειρά των λέξεων που αποδίδουν νόημα, αλλά αυτές οι λέξεις είναι πιο σχετικές. Λαμβάνοντας ένα υποσύνολο των όρων αναζήτησης. Ο περιορισμός της προσοχής σε άρθρα που δημοσιεύτηκαν την Τρίτη, Τετάρτη και την Πέμπτη έχει μέτρια αύξηση της ακρίβειας. Δεδομένου ότι η DJI περιέχει μόνο πληροφορίες για τις καθημερινές (η διαπραγμάτευση έχει διακοπεί τα Σαββατοκύριακα), η μέτρηση για Παρασκευή, Σάββατο, Κυριακή και Δευτέρα, περιέχει πληροφορίες για τις ημέρες που δεν είναι γειτονικές. Για παράδειγμα, η είδηση που δημοσιεύθηκε την Παρασκευή έχει επισημανθεί σύμφωνα με την τιμή της μετοχής κατά το άνοιγμα της Δευτέρας μείον την τιμή κλεισίματος της Πέμπτης. Τα νέα άρθρα χάνουν τη σχετικότητα μετά από ένα ορισμένο χρονικό διάστημα, οπότε αντιπροσωπεύουν μόνο Τρίτη, Τετάρτη και Πέμπτη οι αυξήσεις ακρίβειας. Τέλος, η ανάλυση που βασίζεται στην παρουσία λέξεων στο άρθρο δείχνει σχεδόν καμία μεταβολή από την ανάλυση με τη συχνότητα της λέξης. Χρησιμοποιώντας περιλήψεις άρθρων, αρκετές σχετικές πληροφορίες κωδικοποιούνται σε ένα σύντομο χρονικό του κειμένου που επαναλαμβάνονται τα λόγια που είναι λιγότερο συχνά και ως εκ τούτου λιγότερο σημαντικά να εξεταστούν.

Ο αριθμός των προτεινόμενων μεθόδων στην πρόβλεψη των οικονομικών χρονοσειρών είναι εξαιρετικά μεγάλος. Αυτές οι μέθοδοι βασίζονται σε μεγάλο βαθμό σε δομημένα και αριθμητικά δεδομένα. Στον τομέα των συναλλαγών, τα περισσότερα εργαλεία ανάλυσης της χρηματιστηριακής αγοράς εξακολουθούν να επικεντρώνονται στη στατιστική ανάλυση των προηγούμενων τιμών.

Όμως, ένας από τους τομείς πρόβλεψης των τιμών της αγοράς προέρχεται από τα δεδομένα κειμένου, με βάση την παραδοχή ότι η πορεία της τιμής της μετοχής μπορεί να προβλεφθεί πολύ καλά με την εξέταση των άρθρων ειδήσεων που εμφανίζονται. Στη χρηματιστηριακή αγορά, οι τιμές των μετοχών μπορεί να επηρεάζονται από πολλούς παράγοντες, που κυμαίνονται από τα δελτία ειδήσεων των επιχειρήσεων και της τοπικής πολιτικής για την είδηση της οικονομίας μιας υπερδύναμης.

Η Ανακάλυψη της Γνώσης σε Βάσεις Δεδομένων (KDD), επίσης γνωστή ως εξόρυξη δεδομένων, εστιάζει στην ηλεκτρονική εξερεύνηση των μεγάλων ποσοτήτων δεδομένων και για την ανακάλυψη των ενδιαφερόντων προτύπων στο εσωτερικό τους. Μέχρι πρόσφατα, οι επιστήμονες των υπολογιστών και οι ειδικοί των συστημάτων

πληροφοριών έχουν επικεντρωθεί στην ανακάλυψη της γνώσης από δομημένες, αριθμητικές βάσεις δεδομένων. Ωστόσο, πολλές πληροφορίες είναι σήμερα διαθέσιμες σε μορφή κειμένου, συμπεριλαμβανομένων των εγγράφων, ειδήσεων, εγχειριδίων, e-mail, κλπ. Η αύξηση του αριθμού των γραπτών δεδομένων έχει οδηγήσει στην ανακάλυψη της γνώσης σε αδόμητα (βάσεων δεδομένων κειμένων) στοιχεία γνωστά ως εξόρυξη κειμένου ή κείμενο εξόρυξης δεδομένων. Η εξόρυξη κειμένου (text miningh) είναι μια αναδυόμενη τεχνολογία για την ανάλυση μεγάλων συλλογών αδόμητων εγγράφων για τους σκοπούς της εξόρυξης ενδιαφερόντων και μη - ασήμαντων μοτίβων ή γνώσεων. Η εξόρυξη κειμένου έχει ως στόχο να αναζητήσει πρότυπα στη φυσική γλώσσα κειμένου και να εξαγάγει αντίστοιχες πληροφορίες.

Μία από τις εφαρμογές της εξόρυξης κειμένου είναι η ανακάλυψη και αξιοποίηση της σχέσης μεταξύ του κειμένου του εγγράφου και εξωτερικών πηγών πληροφόρησης, όπως συγκεκριμένα αποσπάσματα του χρηματιστηρίου. Προβλέποντας τις κινήσεις των τιμών των μετοχών με βάση το περιεχόμενο των άρθρων ειδήσεων είναι μία από τις εφαρμογές των τεχνικών εξόρυξης κειμένου. Πληροφορίες σχετικά με την έκθεση ή έκτακτες ειδήσεις της εταιρείας μπορούν δραματικά να επηρεάσουν την τιμή της μετοχής. Έχουν υπάρξει πολλοί ερευνητές με σκοπό να διερευνήσουν την επιρροή των άρθρων των ειδήσεων σχετικά με την αγορά των μετοχών και την αντίδραση της χρηματιστηριακής αγοράς.

Οι ερευνητές έχουν δείξει ότι υπάρχει μια ισχυρή σχέση μεταξύ του χρόνου που έχουν κυκλοφορήσει τις ειδήσεις και της ώρας όταν κυμαίνονται οι τιμές των μετοχών. Αυτό έκανε τους ερευνητές να εισέλθουν σε ένα νέο τομέα της έρευνας, την πρόβλεψη της κίνησης της τάσης που βασίζεται στο περιεχόμενο των ειδήσεων. Ενώ υπάρχουν πολλές υποσχόμενες μέθοδοι πρόβλεψης για να προβλέπουν τις κινήσεις της χρηματιστηριακής αγοράς με βάση αριθμητικά δεδομένα χρονολογικών σειρών, ο αριθμός των μεθόδων πρόβλεψης σχετικά με την εφαρμογή τεχνικών εξόρυξης κειμένου χρησιμοποιώντας ειδήσεων άρθρων είναι λίγες. Αυτό συμβαίνει επειδή η εξόρυξη κειμένου φαίνεται να είναι πιο περίπλοκη από ό,τι η εξόρυξη δεδομένων, δεδομένου ότι περιλαμβάνει την ασχολία με δεδομένα κειμένου που είναι εγγενώς αδόμητα και ασαφή.

4.2 Κατηγορίες ειδησεογραφίας

Εφαρμογές της κατηγοριοποίησης της ειδησεογραφίας είναι:

- **Εφαρμογές για την κριτική που σχετίζονται με ιστοσελίδες.** Η σύνοψη των κριτικών των χρηστών είναι ένα σημαντικό πρόβλημα. Στην ανάλυση του blog, που χρησιμοποιείται για να εκτελέσει την υποκειμενικότητα και την πολικότητα της ταξινόμησης σε θέσεις blog, πρέπει να ανακαλύπτονται παρατυπίες σε χρονικά πρότυπα διάθεσης (φόβου, του ενθουσιασμού, κλπ) που εμφανίζονται σε ένα μεγάλο τμήμα των blogs, να χρησιμοποιηθεί ο σύνδεσμος πληροφοριών πολικότητας με το μοντέλο εμπιστοσύνης και να υπάρξει επιρροή στη σφαίρα του blog, να αναλυθούν τα συναισθήματα του blog και να συσχετιστούν με τις πωλήσεις.
- **Εφαρμογές ως υποσυνιστώσα τεχνολογία.** Οι αναλύσεις συναισθημάτων και τα συστήματα εξαγωγής της γνώμης έχουν ένα σημαντικό δυναμικό ρόλο ως βάση νέων τεχνολογιών για άλλα συστήματα. Τα online συστήματα που εμφανίζουν τις διαφημίσεις, όπως πλαϊνές μπάρες, είναι χρήσιμο να ανιχνεύουν τις ιστοσελίδες που περιέχουν ευαίσθητα περιεχόμενα ακατάλληλα για τις διαφημίσεις τοποθέτησης. Για πιο εξελιγμένα συστήματα, θα μπορούσαν να είναι χρήσιμο για να φέρουν διαφημίσεις προϊόντων, όταν τα σχετικά θετικά συναισθήματα ανιχνεύονται και ίσως περισσότερο σημαντικότερο, να ανακατευτούν οι διαφημίσεις, όταν σχετικές αρνητικές δηλώσεις ανακαλύπτονται.

Μόλις τα οικονομικά άρθρα των ειδήσεων έχουν συγκεντρωθεί, θα πρέπει να αντιπροσωπεύονται τα σημαντικά χαρακτηριστικά τους σε φιλική μορφή. Μια τεχνική είναι μια προσέγγιση της «τσάντας λέξεων» η οποία έχει χρησιμοποιηθεί ευρέως σε μορφή κειμένου της οικονομικής έρευνας (Gidofalvi 2001). Αυτή η διαδικασία περιλαμβάνει την αφαίρεση του νοήματος λέξεων, όπως οι σύνδεσμοι το κείμενο, χρησιμοποιώντας ό, τι απομένει ως εκπροσώπηση κειμένου. Ενώ η μέθοδος της «τσάντας των λέξεων» είναι δημοφιλής, πάσχει από θέματα θορύβου που συνδέονται με σπάνια χρησιμοποιούμενους όρους και τα προβλήματα της κλιμάκωσης, όπου τεράστια υπολογιστική ισχύς απαιτείται για μεγάλα σύνολα δεδομένων. Ένα βελτιωμένο σύστημα αναπαράστασης είναι οι ονοματικές φράσεις. Αυτή η παράσταση διατηρεί μόνο τα ουσιαστικά και ονοματικές φράσεις από ένα έγγραφο και μπορεί επαρκώς να αντιπροσωπεύει τις σημαντικές έννοιες του άρθρου (Tolle & Chen, 2000). Ως αποτέλεσμα, αυτή η τεχνική χρησιμοποιεί λιγότερους όρους και μπορεί να χειριστεί την κλιμάκωση του άρθρου καλύτερα από ό, τι η «τσάντα των λέξεων». Μια τρίτη

τεχνική αναπαράστασης είναι οι «επώνυμες οντότητες», η οποία αποτελεί επέκταση των ονοματικών φράσεων. Λειτουργεί με την επιλογή των κύριων ονομάτων ενός άρθρου που εμπίπτουν σαφώς καθορισμένες κατηγορίες. Αυτή η διαδικασία χρησιμοποιεί μια σημασιολογική λεξιλογική ιεραρχία (Sekine & Nobata 2004), καθώς και μια συντακτική / σημασιολογική διαδικασία tagging για να ορίσει τους υποψήφιους όρους σε κατηγορίες.

Οι «επώνυμες οντότητες» επιτρέπουν την καλύτερη γενίκευση των ήδη αθέατων όρων και δεν έχουν τα προβλήματα επεκτασιμότητας που συνδέονται με την προσέγγιση της σημασιολογίας μόνο. Μια τέταρτη παραστατική τεχνική είναι τα «σωστά ουσιαστικά». Αυτή η μέθοδος λειτουργεί ως ενδιάμεσος μεταξύ ονοματικών φράσεων και επώνυμων οντοτήτων, όπου υπάρχει ως ένα υποσύνολο των «Φράσεων Ουσιαστικών» επιλέγοντας συγκεκριμένα ουσιαστικά, αλλά και ως ένα υπερσύνολο των επώνυμων οντοτήτων χωρίς τον περιορισμό των προκαθορισμένων κατηγοριών. Η παράσταση καταργεί την ασάφεια που σχετίζεται με τα κύρια ονόματα που θα μπορούσαν να εκπροσωπούνται από περισσότερα από ένα όνομα κατηγορίας οντότητας ή δεν εμπίπτουν σε μία από τις επτά κατηγορίες που ορίζονται. Σε μια μελέτη σύγκρισης αυτών των τεσσάρων τεχνικών αναπαράστασης, βρέθηκε ότι η τεχνική του «σωστού ουσιαστικού» ήταν πιο αποτελεσματική στην αντιπροσώπευση των κειμένων των άρθρων των οικονομικών νέων (Schumaker &Chen, 2006).

4.3 Μέθοδοι κατηγοριοποίησης

Υπάρχουν διαφορετικές προσεγγίσεις για να χαρακτηριστεί το επίπεδο συναισθήματος σε ένα έγγραφο που περιλαμβάνει τις μεθόδους μηχανικής μάθησης. Αυτές οι μέθοδοι μηχανικής μάθησης περιλαμβάνουν Naïve Bayes, την ταξινόμησης Μέγιστης Εντροπίας και τις υποστηρικτικές μηχανές (SVM). Διάφορες μελέτες για την ταξινόμηση συναίσθημα του συναισθήματος έχουν πραγματοποιηθεί και έχουν εξελιχθεί σε διαφορετικά επίπεδα λέξεων), επίπεδο πρότασης και το επίπεδο εγγράφων.

Ο χαρακτηρισμός ενός εγγράφου (π.χ. επανεξέταση, blogs) γίνεται με βάση τη συνολική άποψη που εκφράζεται από τη γνώμη του κατόχου. Αυτό υποθέτει ότι κάθε έγγραφο επικεντρώνεται σε ένα μόνο αντικείμενο και περιέχει απόψεις από ένα μόνο κάτοχο γνώμης. Το κύριο καθήκον στην ταξινόμηση του επιπέδου συναισθήματος του εγγράφου είναι να καθοριστεί ο γενικός προσανατολισμός του συναισθήματος του

εγγράφου που εξαρτάται από τις κατηγορίες που μπορεί να είναι θετικές ή αρνητικές και ουδέτερες. Η κατηγοριοποίηση των φράσεων εξετάζει κάθε πρόταση ως ξεχωριστή μονάδα και υποθέτει ότι η φράση πρέπει να περιέχει μόνο μία άποψη. Η ανάλυση σε επίπεδο συναίσθημα έχει δύο καθήκοντα: την κατάταξη της υποκειμενικότητας και της ταξινόμησης του συναισθήματος.

Ο στόχος της ταξινόμησης σε επίπεδο χαρακτηριστικών είναι η δημιουργία μιας περίληψης των απόψεων που βασίζονται στα χαρακτηριστικά των πολλών σχολίων. Έχει κυρίως τρία καθήκοντα. Το πρώτο καθήκον είναι να εντοπίσει και να εξαγάγει αντικειμενικά χαρακτηριστικά που έχουν σχολιαστεί από των κάτοχο της γνωμοδότησης (π.χ. «εικόνα»). Το δεύτερο καθήκον είναι να καθοριστεί η πολικότητα των απόψεων σχετικά με τις κατηγορίες των χαρακτηριστικών: θετικό, αρνητικό και ουδέτερο και το τρίτο έχει σχέση με τα συνώνυμα της λειτουργίας της ομάδας.

Το τυπικό μοντέλο της ανάλυσης συναισθήματος φαίνεται στο ακόλουθο σχήμα. Το στάδιο προετοιμασίας των δεδομένων εκτελεί τα απαραίτητα στοιχεία προ επεξεργασίας και καθαρισμού για το σύνολο δεδομένων για την επόμενη ανάλυσης. Μερικά χρησιμοποιούμενα συνήθως βήματα προεπεξεργασίας περιλαμβάνουν την αφαίρεση των περιεχόμενων μη κειμένου και τις ετικέτες σήμανσης (για HTML σελίδες), καθώς και την κατάργηση των πληροφοριών σχετικά με τα σχόλια που δεν απαιτούνται για την ανάλυση συναισθήματος, όπως ημερομηνίες εξέτασης και ονόματα σχολιαστών. Το στάδιο της επανεξέτασης της αναλύει τα γλωσσικά χαρακτηριστικά των κριτικών έτσι ώστε ενδιαφέρουσες πληροφορίες, συμπεριλαμβανομένων των γνωμών ή / και των χαρακτηριστικών του προϊόντος, μπορούν να προσδιοριστούν. Δύο κοινά αποδεκτές ενέργειες για την ανάλυση της κριτικής είναι η θετική και αρνητική κατηγοριοποίηση. Μετά από αυτή τη φάση, η κατάταξη της γνώμης γίνεται για να ληφθούν τα αποτελέσματα.

Με την ταξινόμηση συναισθήματος υπάρχουν σχετικά λίγες τάξεις (π.χ., "θετική" ή "3 αστέρων"), που γενικεύουν πολλούς τομείς και χρήστες. Επιπλέον, ενώ οι διαφορετικές τάξεις για το θέμα με βάση την κατηγοριοποίηση μπορεί να είναι εντελώς άσχετες, οι ετικέτες συναισθήματος που ευρέως θεωρούνται συνηθισμένες αντιπροσωπεύουν αντίθετες (αν η εργασία είναι η δυαδική ταξινόμησης) ή τακτικές / αριθμητικές κατηγορίες (αν η κατάταξη είναι σύμφωνα με μια κλίμακα πολλών σημείων). Στην πραγματικότητα, η τύπου παλινδρόμησης φύση της δύναμης του

αισθήματος , ο βαθμός θετικότητας και ούτω καθεξής φαίνεται μάλλον μοναδικός στην κατηγοριοποίηση του συναισθήματος.

Υπάρχουν επίσης πολλά χαρακτηριστικά απαντήσεων σε ερωτήματα προσανατολισμένων στη γνώμη που διαφέρουν από εκείνα των ερωτημάτων που στηρίζονται σε γεγονότα. Ως αποτέλεσμα, η διάχυση των πληροφοριών που είναι προσανατολισμένες στη γνώμη, ως ένας τρόπος για να προσεγγίσει την απάντηση σε ερωτήσεις προσανατολισμένες στη γνώμη , φυσικά, διαφέρει από τη διάχυση των παραδοσιακών πληροφοριών. Οι γνωμοδοτήσεις που είναι προσανατολισμένες στη διάχυση πληροφοριών, επίσης, συχνά γενικεύονται και σε διαφορετικούς τομείς, δεδομένου ότι το ενδιαφέρον είναι περίπου το ίδιο σε σύνολο πεδίων για κάθε έκφραση γνώμης (π.χ., κάτοχος, τύπο, δύναμη), ανεξάρτητα από το θέμα. Σε αντίθεση, τα παραδοσιακά πρότυπα διάχυσης πληροφοριών μπορεί να διαφέρουν σημαντικά από το ένα πεδίο στο άλλο, το τυπικό πρότυπο για την καταγραφή πληροφοριών που σχετίζονται με μια φυσική καταστροφή είναι πολύ διαφορετικό από ένα τυπικό πρότυπο για την αποθήκευση των βιβλιογραφικών πληροφοριών. Οι διακρίσεις αυτές ενδέχεται να κάνουν τα προβλήματα να εμφανίζονται απατηλά απλά από ότι στην πραγματικότητα με βάση την ανάλυση, αλλά αυτό απέχει πολύ από την αλήθεια. Λίγα παραδείγματα έχουν ληφθεί ως δειγματοληψία για να δείξουν τι κάνει αυτά τα προβλήματα είναι δύσκολα σε σύγκριση με την παραδοσιακή ανάλυση κειμένου.

Η πρόκληση σε αυτό το θέμα είναι να καθοριστεί αν ένα έγγραφο ή τμήμα (π.χ. σκέψη ή κατάσταση) είναι υποκειμενικό. Μία μόνο λέξη μπορεί να χρησιμοποιηθεί για να μεταφέρει τρεις διαφορετικές απόψεις, θετική, ουδέτερη και αρνητική αντίστοιχα. Για να καταλήξει σε λογικά συμπεράσματα, πρέπει κανείς να καταλάβει το πλαίσιο της ανάλυσης συναισθήματος.

Όπως αναλύεται στο κεφάλαιο της εισαγωγής, η ανάλυση του συναισθήματος διατυπώνεται ως ένα πρόβλημα. Ωστόσο, η κατηγοριοποίηση μπορεί να προσεγγιστεί διαφορετικές προοπτικές. Ανάλογα με την εργασία και την προοπτική του ατόμου που κάνει την ανάλυση συναισθήματος, η προσέγγιση μπορεί να είναι με γνώμονα το λόγο, με γνώμονα τη σχέση ή τη γλώσσα. Μερικές από τις προοπτικές που μπορεί να χρησιμοποιηθούν στην κατηγοριοποίηση της ανάλυσης συναισθήματος συζητούνται στις επόμενες υποενότητες.

1.Προσέγγιση βασισμένη στη γνώση

Σε αυτή την προσέγγιση, το συναίσθημα θεωρείται ως η λειτουργία ορισμένων λέξεων-κλειδιών. Το κύριο έργο είναι η κατασκευή λεξικών συναισθήματος διακρίσεων που υποδεικνύουν μια συγκεκριμένη κατηγορία, όπως η θετική ή αρνητική κατηγορία. Η πολικότητα των λέξεων στο λεξικό καθορίζεται πριν από την εργασία της ανάλυσης του συναισθήματος. Υπάρχουν παραλλαγές για ο πώς το λεξικό δημιουργείται. Για παράδειγμα, λεξικά μπορούν να δημιουργηθούν ξεκινώντας με μερικές λέξεις-«σπόρους» και στη συνέχεια, χρησιμοποιώντας κάποια γλωσσικά χαρακτηριστικά να προστεθούν περισσότερες λέξεις σε αυτά ή την έναρξη με κάποιες λέξεις- σπόρους και πάνω σε αυτά τα λόγια σπόρων προστίθενται άλλα λόγια, ανάλογα με τη συχνότητα σε ένα κείμενο (βλ. Turney, 2002). Για ορισμένους τομείς των καθηκόντων, υπάρχουν διαθέσιμες στο κοινό λεξικά με διακρίσεις λέξεων για χρήση στην ανάλυση συναισθήματος. http://twitrratr.com/ και http://www.cs.pitt.edu/mpqa/ είναι δύο παραδείγματα. Το http://twitrratr.com/ παρέχει λεξικά συναισθήματος για ανάλυση συναισθήματος στο Twitter.

2.Προσέγγιση βάσει σχέσης

Εδώ η κατηγοριοποίηση μπορεί να προσεγγιστεί από τις διαφορετικές σχέσεις που μπορεί να υπάρχουν σε ή μεταξύ των χαρακτηριστικών και των συστατικών. Οι σχέσεις αυτές περιλαμβάνουν σχέσεις μεταξύ των συμμετεχόντων λόγου, σχέσεις μεταξύ των χαρακτηριστικών του προϊόντος χαρακτηριστικά. Για παράδειγμα, αν κάποιος θέλει να ξέρει το συναίσθημα των πελατών για ένα εμπορικό σήμα του προϊόντος, μπορεί κανείς να το υπολογίσει σε συνάρτηση με τα συναισθήματα για τα διαφορετικά χαρακτηριστικά ή συστατικά του.

3.Μοντέλα γλώσσας

Σε αυτή την προσέγγιση, η κατηγοριοποίηση γίνεται με την οικοδόμηση n-grams μοντέλων γλώσσας. Η παρουσία ή η συχνότητα των n-grams θα μπορούσε να χρησιμοποιηθεί. Στην παραδοσιακή ανάκτηση πληροφοριών και την κατηγοριοποίηση με βάση το θέμα, η συχνότητα των n-grams φαίνεται να καλύτερα αποτελέσματα. Ωστόσο, οι Pang et al (2002), στην κατηγοριοποίηση του συναισθήματος διαπίστωσαν η παρουσία όρων δίνει καλύτερα αποτελέσματα από ό, τι η συχνότητα όρου. Αναφέρουν ότι η παρουσία uni-gram είναι πιο κατάλληλη για την ψυχολογία της

ανάλυσης. Όμως, λίγο αργότερο από ό, τι οι Pang et al. (2002), οι Dave et al. (2003) διαπίστωσε ότι τα bi-grams και tri-grams λειτούργησαν καλύτερα από ό,τι uni-grams στην ανάλυση του συναισθήματος για τις κριτικές για το προϊόν.

4. Δομές λόγου και σημασιολογία

Σε αυτήν την προσέγγιση, η σχέση λόγου μεταξύ των συνιστωσών κειμένου χρησιμοποιείται για την καθοδήγηση της κατηγοριοποίησης. Για παράδειγμα, σε σχόλια, το συνολικό συναίσθημα εκφράζεται συνήθως στο τέλος του κειμένου (Pang et al, 2002). Ως αποτέλεσμα, η προσέγγιση της ανάλυσης συναισθήματος, σε αυτή την περίπτωση, μπορεί να οδηγείται από το λόγο στον οποίο το συναίσθημα ως σύνολο της αναθεώρησης λαμβάνεται ως συνάρτηση του συναισθήματος διαφορετικών στοιχείων λόγου για την εξέταση και τις σχέσεις του λόγου που υπάρχουν μεταξύ τους. Σε μια τέτοια προσέγγιση, το συναίσθημα μιας παραγράφου που βρίσκεται στο τέλος της επανεξέτασης θα μπορούσε να δοθεί περισσότερο βάρος στον καθορισμό του κλίματος του σύνολο της αναθεώρησης. Η σημασιολογία μπορεί να χρησιμοποιηθεί σε αναγνωριστικός ρόλος προσδιορισμού των πρακτόρων, όπου είναι ανάγκη να το πράξουν.

ΚΕΦΑΛΑΙΟ 5: ΛΟΓΙΣΜΙΚΑ ΣΥΣΤΗΜΑΤΑ SENTIMENT ANALYSIS

5.1 Επεξεργασία φυσικής γλώσσας

Το συναίσθημα δεν αναλύεται μέσω τεχνητής αναφοράς, όπως μερικοί άνθρωποι μπορεί να μπουν στον πειρασμό να σκεφτούν. Αυτό αναλύεται μέσω μιας συστηματικής διαδικασίας που περιλαμβάνει τη χρήση ενός λεξικού συναισθημάτων. Το λεξικό αυτό εκχωρεί ένα βαθμό θετικότητας ή αρνητικότητας σε μία λέξη από μόνο του το οποίο στη συνέχεια χρησιμοποιείται για να δώσει νόημα στο σύνολο του κειμένου. Αυτός είναι ένας τρόπος ανάλυσης του συναισθήματος (sentiment). Στη συνέχεια, λαμβάνοντας υπόψη ένα είδος εγγενούς θετικότητας ή αρνητικότητας κάθε λέξης που θα μπορούσε να χρησιμοποιηθεί από κάποιον για να μιλήσει για μια επιχείρηση ή ένα προϊόν. Για παράδειγμα, η λέξη "good" θα πρέπει να θεωρείται μια θετική λέξη, καθώς και το "like" και το "love". Στο αντίθετο άκρο του φάσματος μπορούμε να δούμε λέξεις όπως "hate", "dislike", κλπ. Υπάρχουν δύο προβλήματα με αυτή τη μεθοδολογία, ωστόσο. Το πρώτο πρόβλημα είναι ότι αυτή η εκχώρηση των θετικών και αρνητικών συναισθημάτων αξιολογεί μια λέξη χωρίς το πλαίσιο στο οποίο είναι γραμμένη. Με άλλα λόγια χωρίς να λαμβάνει υπόψη τα συμφραζόμενα. Το λεξικό είναι εξαιρετικά περιορισμένο στον αριθμό των λέξεων που θα αποδίδουν πάντα ένα θετικό ή αρνητικό συναίσθημα σε μια έκφραση. Το δεύτερο πρόβλημα είναι ότι οι ερευνητές μπορεί να αναθέτουν διαφορετικούς βαθμούς θετικότητας ή αρνητικότητας σε μία λέξη. Ιδιαίτερα στην περίπτωση των διφορούμενων εκφράσεων, ένας ερευνητής μπορεί να είναι περισσότερο διατεθειμένος να σημειώσει μια λέξη ως περισσότερο ή λιγότερο θετική.

Η κατηγοριοποίηση ενός κειμένου δεν φαίνεται πάντα στα διάφορα χαρακτηριστικά που αναφέρονται μέσα σε ένα άρθρο. Η ανάλυση συναισθήματος παραδοσιακά πραγματοποιείται με τη χρήση της τεχνολογίας που αξιολογεί ένα άρθρο σε global επίπεδο. Μέσα σε ένα κείμενο, ωστόσο, το θέμα δεν μπορεί να συνδέεται με τις περιγραφές. Για παράδειγμα, αν ληφθεί υπόψη η πρόταση : «Αυτή η ταινία θα πρέπει να είναι λαμπρή, φαίνεται να είναι μία μεγάλη παραγωγή, οι ηθοποιοί είναι πρώτης τάξεως, τα βοηθητικά cast είναι καλά, καθώς και ο Σταλόνε προσπαθεί να έχει μια καλή απόδοση, ωστόσο, δεν μπορεί να κρατηθεί ψηλά.» Η πρόταση θα πρέπει είναι θετική,

δεδομένου του αριθμού των θετικών λέξεων που διαθέτει. Μόνο στο τέλος μπορεί κανείς να προσδιορίσει την τελεσιδικία της αποφάσεως που είναι συνολικά αρνητική. Τα λεξικά που χρησιμοποιούνται αναπτύσσονται μέσω της ανάλυσης διαφόρων παραγόντων, συμπεριλαμβανομένων της πόλωσης του συναισθήματος και των βαθμών θετικότητας (Όπως το «μου αρέσει» έναντι του «δεν μου αρέσει»), προσδιορίζουν ποια μέρη ενός εγγράφου περιέχουν υποκειμενικό περιεχόμενο (ανίχνευση της υποκειμενικότητας και της αναγνώρισης της γνώμης), προσδιορίζουν τα μέρη ενός εγγράφου που υπόκεινται στο ίδιο αντικείμενο ανάλυσης (από κοινού το θέμα της ανάλυσης συναισθήματος), καθώς και τον καθορισμό του «πολιτικού» προσανατολισμού του κειμένου (απόψεις και προοπτικές). Άλλες μη τεκμηριωμένες πληροφορίες στο κείμενο μπορούν επίσης να ληφθούν υπόψη. Για παράδειγμα, υπάρχουν έξι "καθολικά" αισθήματα: θυμός, αηδία, φόβος, χαρά, λύπη και έκπληξη που μπορούν να αναλυθούν, καθώς και η παρουσία ενός όρου, η συχνότητα ενός όρου και η σύνταξη.

Υπάρχουν μια σειρά από διαφορετικές τεχνικές ανάλυσης γλώσσας που εμπίπτουν κάτω από την ομπρέλα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing-NLP), εκ των οποίων μόνο ένα πολύ περιορισμένο υποσύνολο προκύπτει τακτικά στη βιβλιογραφία της εξόρυξης του συναισθήματος. Μέρος της δημιουργίας ετικετών λόγου είναι το πιο συχνά εμφανιζόμενο, αν και υπάρχουν και δημοσιεύσεις λεπτομερών ταξινομητών που χρησιμοποιούν την ανάλυση των συνδιαλέξεων ακόμη και με τη χρήση ενός πλήρους συντακτικού δένδρου.

Μέρος της ετικέτας του λόγου (POS) είναι η διαδικασία της επισήμανσης των λέξεων των περιστατικών, για παράδειγμα, αν μια λέξη εμφανίζεται ως επίθετο, ουσιαστικό ή ρήμα. Η αποτελεσματική απόδοση ετικετών απαιτεί γνώση όχι μόνο της λέξης, αλλά και του πλαισίου της, όπως η θέση στην πρόταση και γύρω τις λέξεις. Τα κρυμμένα μοντέλα Markov είναι μια κοινή τεχνική που χρησιμοποιείται σε POS tagging, αν και το Stanford POS Tagger χρησιμοποιεί μια μέγιστη τεχνική εντροπίας που είναι να επιλεξούμε το πιο ανεπηρέαστο αποτέλεσμα, το οποίο όμως είναι συμβατό με τους περιορισμούς του προβλήματος μας. Έτσι η αρχή της μεγίστης εντροπίας δίνει την πιο ανεπηρέαστη περιγραφή η οποία όμως συμφωνεί με τη διαθέσιμη σχετική πληροφορία. Για να αποφευχθεί συνεχώς αναφορά σε ένα θέμα με βάση το όνομα, φυσικές γλώσσες

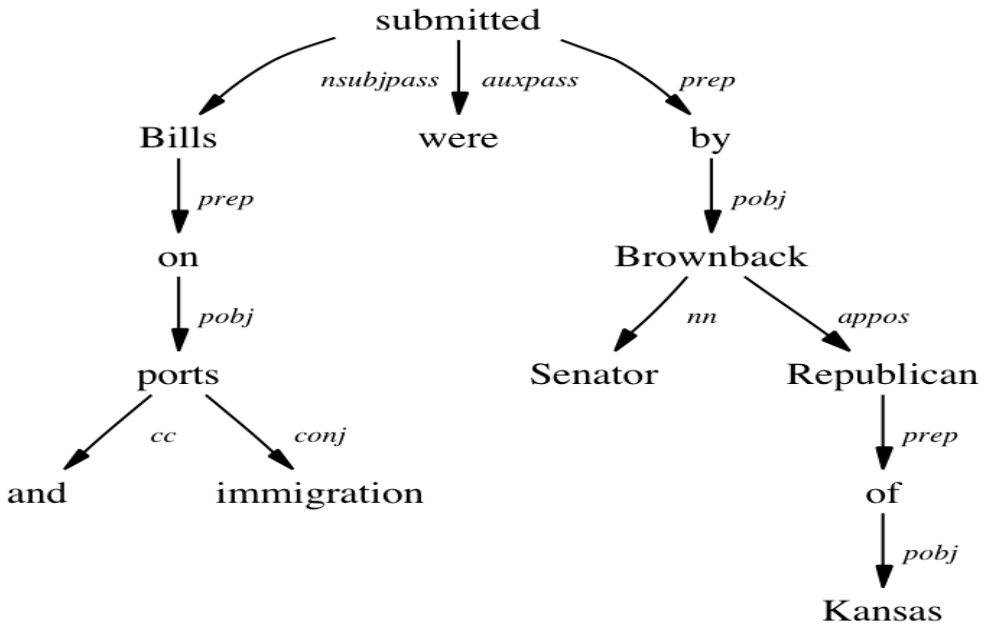
που συνήθως περιέχουν εναλλακτικές λέξεις μπορούν να χρησιμοποιηθούν όταν αναφέρονται σε ένα αντικείμενο που αναφέρθηκε προηγουμένως.

Για παράδειγμα:

«Ο Γιάννης έχει αποδειχθεί τι μεγάλος ηθοποιός που είναι, για άλλη μια φορά παίζει το ρόλο του τέλεια».

Εδώ, οι λέξεις "αυτός" και δύο "του" αφορούν το Γιάννη. Αυτοματοποίηση της διαδικασίας σύνδεσης τέτοιων αναφορών ονομάζεται ανάλυση coreference και προηγούμενες μελέτες έχουν δείξει ότι η αποτελεσματική εφαρμογή της τεχνική στη ανάλυση συναισθήματος μπορεί να βελτιώσει την ακρίβεια της ταξινόμησης περισσότερο από το 10%.

Η δημιουργία δέντρων ανάλυσης για τις προτάσεις φυσικής γλώσσας είναι μια άλλη κεντρική περιοχή της μελέτης NLP. Λόγω της ασάφειας της φυσικής γλώσσας, υπάρχουν συχνά πολλαπλά έγκυρα δέντρα για μεταγλώττιση σε μια δεδομένη περίοδος, οπότε απαιτούν πιθανοτικές τεχνικές που πρέπει να χρησιμοποιηθούν. Η ανάλυση σχετίζεται με το POS tagging, όπως ο καθορισμός της δομής μιας πρότασης απαιτεί τη γνώση των οποίων οι λέξεις με νόημα χρησιμοποιούνται, σε μια απλοποιημένη μορφή της ανάλυσης που δεν αναλύει προτάσεις σε βάθος, μπορεί να χρησιμοποιηθεί για ορισμένες εφαρμογές.



Εικόνα 5.1: Παράδειγμα Part Of Speech-tagging (POS-tagging)

5.2 Εργαλεία Sentiment Analysis

5.2.1 Ειδικά λεξικά

Υπάρχουν τρία είδη λεξικών συναισθήματος: λέξεις με συναισθηματικές αξίες (sentimental value), με πολικότητα συναισθήματος (polarity) και μια συλλογή από λέξεις συναισθήματος μόνο. Οι αξίες συναισθημάτων μας δίνουν τις περισσότερες πληροφορίες μεταξύ των τριών, έτσι ώστε να επικεντρωθούν σχετικά με την κατάρτιση ενός λεξικού με την αξία του συναισθήματος. Για να κατασκευαστεί ένα λεξικό συναισθημάτων με ένα μεγάλο μέγεθος λεξιλογίου, οι δημιουργοί υπολογίζουν συνήθως το συναίσθημα σε αξίες / πολικότητα νέων λέξεων από αυτές στα υφιστάμενα λεξικά αυτόματα τα τελευταία χρόνια.

Τα λεξικά συναισθήματος αποτελούν πολύτιμους πόρους για την ανάλυση συναισθήματος. Μπορούν να χρησιμοποιηθούν για τον εντοπισμό λέξεων συναισθήματος και εκφράσεων και μπορούν επίσης να χρησιμοποιηθούν για να παράγουν κατατοπιστικά χαρακτηριστικά για την κατηγοριοποίηση των συναισθημάτων των δημοσιευμένων νέων. Πολλά λεξικά συναισθήματος έχουν καταρτιστεί για τα αγγλικά (Hu και Liu, 2004, Wilson et al, 2005). Αυτά χρησιμοποιούνται ευρέως στην έρευνα για την ανάλυση συναισθήματος. Αντίθετα, λόγω του υψηλού κόστους της κατασκευής ενός χειρόγραφου λεξικού, τα λεξικά συναισθήματος σε άλλες γλώσσες είναι πολύ λίγα ή δεν είναι ακόμα διαθέσιμα. Η έλλειψη λεξικών συναισθήματος περιορίζει την εξαγωγή του συναισθήματος στην χρηματοοικονομική ειδησεογραφία και δημοσιογραφία που είναι γραμμένα σε άλλες γλώσσες. Εκτιμάται δε ότι το 2012, μόνο το 26,8% των χρηστών του Internet μιλούν αγγλικά.

Πολλή από την αρχική έρευνα στο συναίσθημα επικεντρώθηκε σε επίθετα ή φράσεις επιθέτικών προσδιορισμών, όπως η πρωταρχική πηγή του υποκειμενικού περιεχομένου σε ένα έγγραφο (Hatzivassiloglou και McKeown 1997, Hu και Liu 2004), αν και υπάρχουν κάποιες εξαίρεσεις, ιδιαίτερα πιο πρόσφατες, οι οποίες έχουν συμπεριλάβει επίσης τη χρήση των επιρρημάτων (Benamara et al. 2007), επιθέτων και ρημάτων (Kim και Hovy 2004), φράσεις επιθέτων, φράσεις δύο-λέξεων (Turney και Littman 2003), επίθετα, ρήματα, επιρρήματα και την αποκλειστική χρήση των ρημάτων, χρήση των μη συναισθηματικών επιθέτων και επιρρημάτων, ή λογικές λέξεις και φράσεις που επιλέγονται από ανθρώπινους σχολιαστές. Σε γενικές γραμμές, ο υπολογισμός του

σημασιολογικού προσανατολισμού ενός ολόκληρου εγγράφου είναι το συνδυασμένο αποτέλεσμα των επιθέτων ή ίδιας αξίας λέξεων που βρίσκονται μέσα, με βάση ένα πιθανό λεξικό της κατάταξης λέξεων το οποίο να αποδίδει ένα είδος αυτοτελούς-αυτομάτου σκορ στην κάθε λέξη που ερευνάται. Το λεξικό μπορεί να δημιουργηθεί με διάφορους τρόπους: χειροκίνητα, χρησιμοποιώντας τα υπάρχοντα λεξικά ή ημι-αυτόματα, κάνοντας χρήση των πόρων, όπως το WordNet (Hu και Liu 2004).

Το λεξικό μπορεί επίσης να παραχθεί αυτόματα μέσω της σύνδεσης, όπου η βαθμολογία για κάθε νέο επίθετο υπολογίζεται χρησιμοποιώντας την εγγύτητας του επιθέτου σε σχέση με μία ή περισσότερες λέξεις σπόρους. Οι λέξεις-σπόροι είναι ένα μικρό σύνολο λέξεων με ισχυρή ή απολύτως ευδιάκριτη αρνητική ή θετική συσχέτιση, όπως good ή bad. Καταρχήν, ένα θετικό επίθετο θα πρέπει να εμφανίζεται πιο συχνά παράλληλα με τα θετικά λόγια των σπόρων και έτσι θα αποκτήσει ένα θετικό αποτέλεσμα, ενώ τα αρνητικά επίθετα συμβαίνουν πιο συχνά στην περιοχή των αρνητικών λέξεων- σπόρων, έτσι λαμβάνουν αρνητική βαθμολογία. Η μέθοδος του Turney (Turney και Littman 2003) παρουσίασε έναν απλό και εύκολο τρόπο υπολογισμού του σημασιολογικού προσανατολισμού μέσω της σημασιολογικής συσχέτισης που δεν περιορίζεται μόνο στα επίθετα. Η σύνδεση λοιπόν συνήθως υπολογίζεται σύμφωνα με τη μέθοδο του Turney για τον υπολογισμό του σκορ (Turney και Littman 2003).

Προηγούμενες εκδόσεις του σημασιολογικού προσανατολισμού ενός ολόκληρου εγγράφου (Taboada, Anthony, και Voll 2006) βασίστηκαν σε ένα λεξικό επιθέτων για να προβλέψουν τα συνολικά επίθετα ενός εγγράφου, χρησιμοποιώντας ένα απλό άθροισμα και τη μέθοδο του μέσου όρου: Οι επιμέρους βαθμολογίες για κάθε επίθετο σε ένα έγγραφο, προστίθενται και στη συνέχεια διαιρούνται με το συνολικό αριθμό των επιθέτων στο εν λόγω έγγραφο.

Όπως θα περιγράψουμε στη συνέχεια, η τρέχουσα έκδοση του σημασιολογικού προσανατολισμού παίρνει άλλα μέρη του λόγου υπόψη και κάνει χρήση των πιο εξελιγμένων μεθόδων που καθορίζουν την πραγματική συμβολή της κάθε λέξης. Είναι σημαντικό να σημειωθεί ότι το πώς ένα λεξικό δημιουργείται επηρεάζει τη συνολική ακρίβεια των μετέπειτα αποτελεσμάτων.

Οι Taboada, Anthony και Voll (2006) αναφέρουν σχετικά με τα πειράματα χρησιμοποιώντας διαφορετικές μηχανές αναζήτησης και φορείς που προσπαθούν να δημιουργήσουν λεξικά ημι-αυτόματα. Βρήκαν ότι, αν και μπορούν να χρησιμοποιηθούν, λεξικά που δημιουργήθηκαν με τη μηχανή αναζήτησης του Google ήταν ασταθή. Μια εναλλακτική θα ήταν να χρησιμοποιηθεί ένα αρκετά μεγάλο στατικό σώμα με τη μέτρηση της σχεσιακής ομοιότητας σε ζεύγη λέξεων. Αυτόματα ή ημιαυτόματα δημιουργήθηκαν λεξικά που έχουν κάποια πλεονεκτήματα.

Όλα τα εργαλεία ανάλυσης συναισθήματος έχουν επικαλεστεί, σε διαφορετικούς βαθμούς, τις λίστες των λέξεων και φράσεων με τις θετικές και αρνητικές χροιές ή εμπειρικά σχετίζονται με θετικά ή αρνητικά σχόλια. Οι εν λόγω κατάλογοι δεν μπορούν να χρησιμοποιηθούν όπως είναι, αλλά θα πρέπει να προσαρμοστούν ανάλογα με το αντικείμενο για το οποίο θα χρησιμοποιηθούν, ώστε να παρέχουν αξιόπιστα αποτελέσματα. Μια μεγάλη προσπάθεια είναι απαραίτητη για την ανάπτυξη ενός λεξικού συναισθημάτων συγκεκριμένων όρων και να προσδιοριστεί το σωστό λεξιλόγιο που σχετίζεται με την έκφραση των θετικών και αρνητικών συναισθημάτων. Πολλοί άνθρωποι δεν είναι απαραίτητα πρόθυμοι να περνούν το χρόνο τους στην εκτέλεση τέτοιων προσαρμογών και καθηκόντων επικύρωσης. Θέλουν κάτι που πιστεύουν ότι θα λειτουργήσει αμέσως και θα είναι έτοιμοι να πληρώσουν πολλά για ένα τέτοιο εργαλείο.

Υπάρχει κίνδυνος ορισμένοι άνθρωποι να μπορούν να χρησιμοποιήσουν το λεξικό συναισθήματος, όπως είναι, χωρίς να προσπαθήσουν να το διαμορφώσουν ή να το προσαρμόσουν στο δικό τους τύπο των δεδομένων. Όσοι έχουν επίγνωση των ορίων αυτών των πινάκων μπορεί να εξακολουθούν να μην έχουν ιδέα για το πώς μια τέτοια προσαρμογή θα μπορούσε να επιτευχθεί και χρειάζονται κάποια καθοδήγηση. Ωστόσο, παρά την πιθανή κακή χρήση των λιστών λέξεων της ανάλυσης συναισθήματος, υπάρχει το WordStat λεξικό στη διάθεση του κοινού. Ένας από τους λόγους που έκανε την εταιρία που το εκδίδει να προσαρμοστεί ήταν η δημοσίευση δύο άρθρων.

Το πρώτο από αυτά, γραμμένο από τους Loughran και McDonald (2011), υπογραμμίζει τον κίνδυνο να μην εξάγονται αξιόπιστα αποτελέσματα όταν γίνεται χρήση μη ειδικευμένων λιστών λέξεων στις περιπτώσεις ανάλυσης χρηματοοικονομικών κειμένων, άρα και ειδήσεων. Οι ερευνητές ανέπτυξαν δικά τους λεξικά, ειδικά ανά τομέα για να περιγράψουν το συναίσθημα, σε κάποια λεπτομέρεια, διαδικασία με την

οποία επιλέγονται οι λέξεις οι βαρύτητες τους και να επικυρώνονται τα αποτελέσματά τους. Το δεύτερο άρθρο, που δημοσιεύθηκε από τον Young και Soroka (2011), παρουσιάζει τη διαδικασία κατασκευής και την επικύρωση ενός λεξικού συναισθήματος, αλλά αυτή τη φορά προσαρμοσμένο για την ανάλυση της πολιτικής ειδήσεογραφίας. Και τα δύο αυτά έγγραφα αποτελούν αξιέπαινες προσπάθειες και αξίζουν να διαβαστεί από τον καθένα που θα ήθελε να μάθει πώς να δημιουργήσει ένα συγκεκριμένο πλαίσιο λεξικό ανάλυσης συναισθήματος και τρόπους για να προσαρμόσει λεξικά ώστε να εξυπηρετούν αποδοτικά τους στόχους και τις θεματικές που επιθυμεί να αναλύσει.

5.2.1.1 Loughran and McDonald Financial Sentiment Dictionary

Η εργασία των Loughran και McDonald (2011) παρέχει μια σαφή απόδειξη ότι η εφαρμογή μιας γενικής λίστας λέξεων συναισθήματος για την αξιολόγηση κειμένων που αφορούν την λογιστικά, οικονομικά και κατά επέκταση χρηματοοικονομικά θάματα μπορεί να οδηγήσει σε υψηλό ποσοστό εσφαλμένης ταξινόμησης. Επεξεργάστηκαν ετήσιες αναφορές (10-k forms) της αμερικάνικης επιτροπής χρεογράφων (U.S. Securities and Exchange Commission (SEC)) από το 1994 μέχρι το 2008 και διαπίστωσαν ότι περίπου τα τρία τέταρτα των αρνητικών λέξεων από το Harvard IV λεξικό TagNeg δεν έχουν συνήθως αρνητικό περιεχόμενο μέσα σε ένα οικονομικό κείμενο. Για παράδειγμα, οι λέξεις όπως "mine", "cancer", "vice" ή "capital" χρησιμοποιούνται συχνά για να αναφερθούν σε ένα συγκεκριμένο τμήμα ή σε ένα μέρος του οργανογράμματος των εταιριών. Αυτές οι λέξεις δεν είναι προγνωστικά του τόνου των εγγράφων ή των οικονομικών ειδήσεων και απλά προσθέτουν θόρυβο, αρνητικό κυρίως, στη μέτρηση του συναισθήματος και μετριάζεται η προγνωστική αξία τους.

Οι συγγραφείς έχουν δημιουργήσει πρόσθετες προσαρμοσμένες λίστες των αρνητικών και των θετικών λέξεων ειδικά για τη λογιστική, την χρηματοοικονομική και το δημοσιονομικό τομέα. Λαμβάνοντας υπόψη τα αποτελέσματά συνιστούν τη χρήση του όρων στάθμισης κατά την ανάλυση των λέξεων. Ακόμη και αν η φαινόμενη ισχύς (με τον όρο στάθμισης) των δύο αρνητικών λιστών λέξεων είναι παρόμοια, προτείνουν τη χρήση του καταλόγου τους για να αποφεύγεται το misinterpretation στις λέξεις στη

λίστα H4N που θα μπορούσε να είναι επιβλαβής στην ερμηνεία. Ένα άλλο πλεονέκτημα του λεξικού είναι ότι δείχνει πως η ποσοτική ανάλυση περιεχομένου μπορεί να κινηθεί πέρα από το χαρακτηριστικό των απλών διχοτομικών διαφοροποιήσεων της ανάλυσης συναισθήματος και μπορεί επίσης να χρησιμοποιηθεί για τη μέτρηση πρόσθετων διαστάσεων ενδιαφέροντος. Δύο αξιοσημείωτες προσθήκες είναι η λίστα λέξεων αβεβαιότητας που επιχειρεί να μετρήσει την γενική έννοια της ανακρίβειας (χωρίς ρητή αναφορά σε κινδύνους) και η λίστα φιλόδικων λέξεων που μπορούν να χρησιμοποιηθούν για τον εντοπισμό δυνητικών νομικά προβληματικών καταστάσεων. Περιέλαβαν, επίσης, αδύναμες Modal και Ισχυρές Modal λίστες λέξεων.

5.2.1.2 Lexicoder Sentiment Dictionary (LSD)

Πολλές φορές η πολιτική ειδησεογραφία επηρεάζει τις οικονομικές εξελίξεις. Συνεπώς η αγνόηση του πολιτικού περιβάλλοντος ίσως να ερμηνεύσει λανθασμένα τα δημοσιευμένα οικονομικά νέα. Οι Young και Soroka (2011) εξηγούν τη διαδικασία κατασκευής ενός λεξικού ανάλυσης συναισθήματος βασισμένη στην πολιτική επικοινωνία. Στόχος τους ήταν να διευρύνουν το σκορ της κάλυψης των υφιστάμενων λεξικών συναισθήματος, χωρίς να διακυβεύεται η ακρίβεια τους. Όπως και στο προηγούμενο άρθρο, χρησιμοποιήθηκε ως βάση το Harvard IV λεξικό (Stone et al., 1966), στο οποίο στις αρχικές θετικές και αρνητικές προσέθεσαν και άλλες λέξεις από το θησαυρό Roget, καθώς και από το Λεξικό Imagery του Colin Martindale (δύο λεξικά που είναι διαθέσιμα σε μορφή WordStat). Αφαίρεσαν τις ουδέτερες και διφορούμενες λέξεις και στη συνέχεια εξήγαγαν τις πιο συχνές, με αποτέλεσμα να προκείψει μια λίστα 2.858 αρνητικών και 1.709 θετικών καταχωρήσεων. Μερικά αξιοσημείωτα χαρακτηριστικά του λεξικού τους είναι η εφαρμογή των βασικών λέξεων αποσαφήνισης της αίσθησης με τη χρήση των φράσεων, την αποκοπή και την προεπεξεργασία, καθώς και την προσπάθεια για την αντιμετώπιση των αρνήσεων. Για να εκτιμηθεί η ακρίβεια του LSD, το λεξικό ελέγχθηκε σε ένα σύνολο από 900 κωδικοποιημένες ειδήσεις. Τα αποτελέσματα δείχνουν ότι το λεξικό που δημιούργησαν αποδίδει το νόημα των ειδήσεων με πολύ πιο αξιόπιστο τρόπο από όλα τα άλλα εννέα content analytic λεξικά που ήταν διαθέσιμα και ότι το LSD είναι το πιο υποσχόμενο ως προς την πολιτική επικοινωνία και την ερμηνεία του τόνου-διάθεσης των δημοσιευμένων άρθρων. Οι συγγραφείς επίσης έδειξαν την προγνωστική ισχύ του

λεξικού με την επίδειξη υψηλών συσχετίσεων μεταξύ των αποτελεσμάτων της ανάλυσης των δημοσιευμένων άρθρων και των δημοσκοπήσεων με τα αποτελέσματα κατά την καναδική ομοσπονδιακή προεκλογική εκστρατεία το 2006.

5.2.1.3 WordStat Sentiment Dictionary

Το λεξικό Sentiment WordStat αποτελεί ίσως την πιο διάσημη πηγή πληροφοριών για τις λέξεις στον τομέα του sentiment analysis. Έχει δημιουργηθεί στη βάση τριών τεράστιων και αναγνωρισμένων λεξικών. Συγκεκριμένα από το Harvard IV, το λεξικό Regressive Imagery (Martindale, 2003) και με τη γλωσσικό Λεξικό Word Count (Pennebaker, 2007). Στη συνέχεια αναπτύχθηκε τεχνικά έτσι ώστε να επεκτείνει τις λίστες και τον αριθμό των λέξεων αναγνωρίζοντας πιθανά συνώνυμα και σχετικές εννοιολογικά έννοιες με άλλες λέξεις αλλά έλαβε υπόψη και προσέθεσε τον παράγοντα των γραμματικών κλίσεων των λέξεων. Κατέληξαν οι δημιουργοί του να έχουν ένα σύνολο 14011 λεκτικών μοτίβων εκ των οποίων οι 9164 έχουν αρνητική και οι 4847 έχουν θετική χροιά. Η εξαγωγή του sentiment από αυτό το εγχείρημα προφανώς δε βασίζεται μόνο στις λίστες των λέξεων αλλά από δύο ευρύτατα σύνολα κανόνων που αφορούν και επιδιώκουν να λάβουν υπόψη και τα συμφραζόμενα όπως τις αρνήσεις που επιφέρουν αλλαγή στην πολικότητα όπως τη διαφορά μεταξύ «good» και «not good». Επί παραδείγματι η αρνητική πολικότητα υπολογίζεται με τη χρήση των εξής δύο κανόνων:

1. Πριν τις αρνητικά φορτισμένες λέξεις δεν πρέπει να προκύπτει άρνηση (no, not, never, κτλ) σε απόσταση τριών λέξεων από αυτήν.
2. Πριν από θετικά φορτισμένες λέξεις προκύπτει άρνηση απόσταση τριών λέξεων από αυτήν στην ίδια πρόταση. Λ.χ. not good, not at all good και never going positive.

Το θετικό συναίσθημα μετριέται με την εφαρμογή παρόμοιων κανόνων.

5.2.2 Linguistic Inquiry and Word Count (LIWC)

Οι ερευνητές την τελευταία τριακονταετία έχουν δείξει με στοιχεία τη σχέση μεταξύ ψυχικής υγείας των ανθρώπων και των λέξεων που χρησιμοποιούν είτε στον γραπτό είτε στον προφορικό λόγο (Rosenberg & Tucker, 1978, Stiles, 1992).

Προκειμένου να παρασχεθεί μια αποδοτική και αποτελεσματική μέθοδος για τη μελέτη των διαφόρων συναισθηματικών, και όχι μόνο, συστατικών που υπάρχουν σε προφορικά και γραπτά δείγματα ατόμων, αναπτύχθηκε μια εφαρμογή ανάλυσης κειμένου που ονομάζεται γλωσσική έρευνα και καταμέτρηση λέξεων, ή Linguistic Inquiry and Word Count (LIWC). Η πρώτη εφαρμογή LIWC αναπτύχθηκε ως μέρος μιας διερευνητικής μελέτης της γλώσσας και της εξωτερίκευσης των σκέψεων (Francis, 1993). Όπως περιγράφεται παρακάτω, η δεύτερη έκδοση, το LIWC2007, είναι μια ενημερωμένη αναθεώρηση της αρχικής εφαρμογής. Οι LIWC2007 εφαρμογές έχουν σχεδιαστεί για να αναλύσουν γραπτό κείμενο, επεξεργαζόμενες την κάθε λέξη του κειμένου. Κατόπιν υπολογίζουν το ποσοστό των λέξεων που ταιριάζουν με κάθε μία από τις 82 γλωσσικές «διαστάσεις» οι οποίες περιλαμβάνουν μεταξύ άλλων, πεδία για τις προσωπικές εκφράσεις (I, me, my), άρθρα, μεγάλες λέξεις (άνω των 6 γραμμάτων), θετικά και αρνητικά συναισθήματα και εξάγουν ένα αποτέλεσμα εύκολα επεξεργάσιμο από τον χρήστη.

Η εφαρμογή LIWC2007 περιέχει μέσα της ένα προεπιλεγμένο σύνολο κατηγοριών λέξεων και προεπιλεγμένο λεξικό που καθορίζει ποιες λέξεις θα πρέπει να υπολογίζονται στα κείμενα.

Η LIWC2007 έχει σχεδιαστεί να δέχεται γραπτά ή προφορικά κείμενα που έχουν αποθηκευτεί ως κείμενο ή ASCII αρχείο χρησιμοποιώντας οποιοδήποτε από τα δημοφιλή πακέτα λογισμικού επεξεργασίας κειμένου (π.χ. WordPerfect ή Word) και αναλύει κάθε λέξη διαδοχικά. Ο χρόνος επεξεργασίας για μια σελίδα ενός απλού κειμένου είναι συνήθως ένα κλάσμα του δευτερολέπτου. Το LIWC2007 διαβάζει κάθε καθορισμένο αρχείο κειμένου, μια λέξη λέξη. Το κάθε λοιπόν στοιχείο διαβάζεται και αντιστοιχίζεται στις κατάλληλες κατηγορίες λέξεων με αποτέλεσμα να παράγεται ένα σκορ για την κάθε κατηγορία.

Το LIWC αναπτύχθηκε κυρίως για γενικότερη μελέτη των συναισθημάτων και της ψυχικής διάθεσης των συγγραφέων των κειμένων. Το σημαντικό όμως είναι ότι έχει τη δυνατότητα εύκολης προσαρμογής για την εφαρμογή σε δημοσιευμένα οικονομικά νέα καθώς ήδη διαθέτει πλουσιοπάροχο λεξιλόγιο, τη δυνατότητα να επεξεργάζεται λέξεις με ίδιες ρίζες, αλλά και να ξεχωρίζει την επίσημη/επαγγελματική από την καθημερινή/ανεπίσημη γραφή.

ΚΕΦΑΛΑΙΟ 6: ΤΕΧΝΙΚΕΣ ΠΡΟΒΛΕΨΕΩΝ

6.1 Γενικά για τις προβλέψεις

Εδώ και εκατοντάδες χρόνια ο άνθρωπος είχε την περιέργεια να γνωρίζει τι του επιφυλάσσει το μέλλον. Το ενδιαφέρον του ανθρώπου για το μέλλον και την πρόβλεψη αυτού, πηγάζει κυρίως από την αβεβαιότητα για το τι φέρνει το αύριο. Ανεξαρτήτως της μεθόδου ή του μηχανισμού πρόβλεψης, βασικό μέλημα του ανθρώπου είναι να μειωθεί όσο γίνεται η απόκλιση της πρόβλεψης από την επικείμενη αλήθεια. Από τους ανώτερους διοικητές των επιχειρήσεων και τους επενδυτές έως τους απλούς καθημερινούς πολίτες, όλοι βρίσκονται αντιμέτωποι με την αβεβαιότητα. Η αβεβαιότητα αυτή δεν προκύπτει μονάχα από την αστοχία της πρόβλεψης λόγω μιας μικρής διαφοροποίησης των γεγονότων από το αναμενόμενο, αλλά και από παντελώς απρόβλεπτα γεγονότα που είναι σχεδόν αδύνατον να προβλεφθούν με βάση τα υπάρχοντα δεδομένα.

Η αδυναμία του ανθρώπου να βασιστεί εξ ολοκλήρου στην διαίσθησή του για να διαχειρισθεί αυτήν την αβεβαιότητα οδήγησε στην επιστημονική ανάπτυξη του κλάδου των προβλέψεων που θα μπορούσε να χαρακτηριστεί ως ένας συνδυασμός στατιστικής ανάλυσης δεδομένων και επιχειρησιακής έρευνας. Από το 1980 και μετά η επιστήμη των προβλέψεων αναπτύχθηκε ραγδαία βρίσκοντας εφαρμογές τόσο σε ακαδημαϊκό επίπεδο, όσο επιχειρησιακό.

Είναι αναγκαίο να γνωρίζουμε την πιθανή κατάληξη-αποτέλεσμα κάθε κινήσής μας, ώστε να μπορούμε να διαχειριστούμε τυχούσες αποτυχίες και να εκμεταλλευτούμε στο έπακρο κάθε πιθανή επιτυχία. Η ανάγκη αυτή οδήγησε τους επιστήμονες, να δημιουργήσουν κάποια μαθηματικά μοντέλα τα οποία ερμηνεύουν το «μέλλον». Η εργασία αυτή εξετάζει κατά πόσο μπορούν οι τεχνικές προβλέψεων να συμβουλεύσουν αξιόπιστα έναν επενδυτή ώστε να λάβει τις πιο επικερδής αποφάσεις.

Τα πεδία εφαρμογής των προβλέψεων είναι: Οικονομία και Χρηματοοικονομικά, Περιβάλλον και Κλίμα, Κοινωνικό περιβάλλον, Τουρισμός, Μεταφορές και Μετακινήσεις, Ακίνητα και κτηματικές περιουσίες αλλά και η Πολιτική.

Λόγω του ευρύτατου φάσματος εφαρμογών, οι τεχνικές προβλέψεων έχουν κατηγοριοποιηθεί σε ποσοτικές και ποιοτικές. Ποσοτικές είναι οι μέθοδοι που δίνουν αποτελέσματα με βάση τις τιμές του παρελθόντος (χρονοσειρές). Αντίθετα, ως

ποιοτικές μεθόδους ορίζουμε αυτές που εφαρμόζονται σε περιπτώσεις ανεπαρκών στοιχείων και βασίζονται κυρίως στην εμπειρία, τα συναισθήματα, τα ερεθίσματα και στις γνώσεις των ανθρώπων που πραγματοποιούν την πρόβλεψη. Οι ποιοτικές μέθοδοι συνήθως χρησιμοποιούνται σε συνδυασμό με ορισμένες ποσοτικές.

Πολλές διαφορετικές μέθοδοι προβλέψεων έχουν προταθεί και προταχθεί από τους επιστήμονες και κυρίως τους ακαδημαϊκούς εκ των οποίων μερικές βασίζονται μόνο σε θεωρητικό υπόβαθρο, ενώ άλλες απαιτούν και την συμβολή της τεχνολογίας και μάλιστα με μεγάλη υπολογιστική ισχύ.

6.2 Χρονοσειρές

Το βασικό στοιχείο στον τομέα των προβλέψεων αποτελεί η χρονοσειρά. Ως χρονοσειρά ορίζεται ένα σύνολο από διαδοχικές παρατηρήσεις κάποιου φυσικού ή άλλου μεγέθους. Οι χρονοσειρές παρουσιάζουν την εξέλιξη ενός μεγέθους σε ένα εύρος χρόνου. Οι διαδοχικές αυτές παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους.

Οι χρονοσειρές χωρίζονται σε ντετερμινιστικές και σε στοχαστικές ανάλογα με τον τρόπο προσδιορισμού των μελλοντικών δεδομένων. Στις ντετερμινιστικές χρονοσειρές οι παρατηρήσεις είναι εξαρτημένες μεταξύ τους και έτσι έχουμε τη δυνατότητα, όταν γνωρίζουμε την σχέση της εξάρτησής τους, να υπολογίσουμε με ακρίβεια τις μελλοντικές παρατηρήσεις. Το γεγονός αυτό, ωστόσο, δεν συμβαίνει με τις πραγματικές χρονοσειρές, καθώς το μέλλον δεν καθορίζεται πλήρως από το παρελθόν, αλλά μόνο μερικώς. Σε πραγματικό περιβάλλον τα περισσότερα μεγέθη επηρεάζονται και από τον λεγόμενο "τυχαίο παράγοντα" που αντιπροσωπεύει μια στατιστική μεταβλητή. Τα μοντέλα που περιέχουν την τυχαιότητα ονομάζονται στοχαστικά.

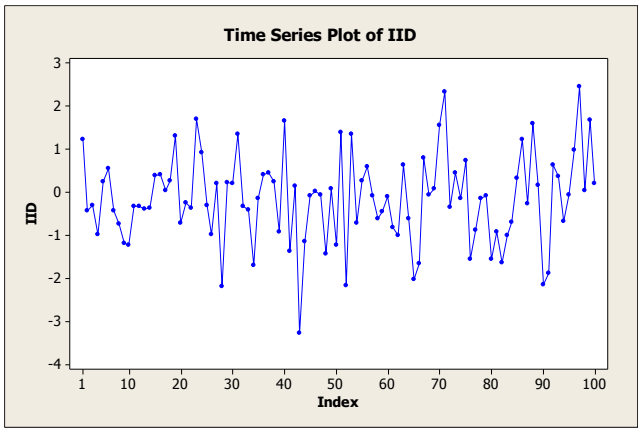
Η εφαρμογή κανόνων πρόβλεψης έχει νόημα για τις στοχαστικές χρονοσειρές, καθώς η πορεία των ντετερμινιστηκών μεγεθών είναι προδιαγεγραμμένη.

6.2.1 Ποιοτικά χαρακτηριστικά χρονοσειρών

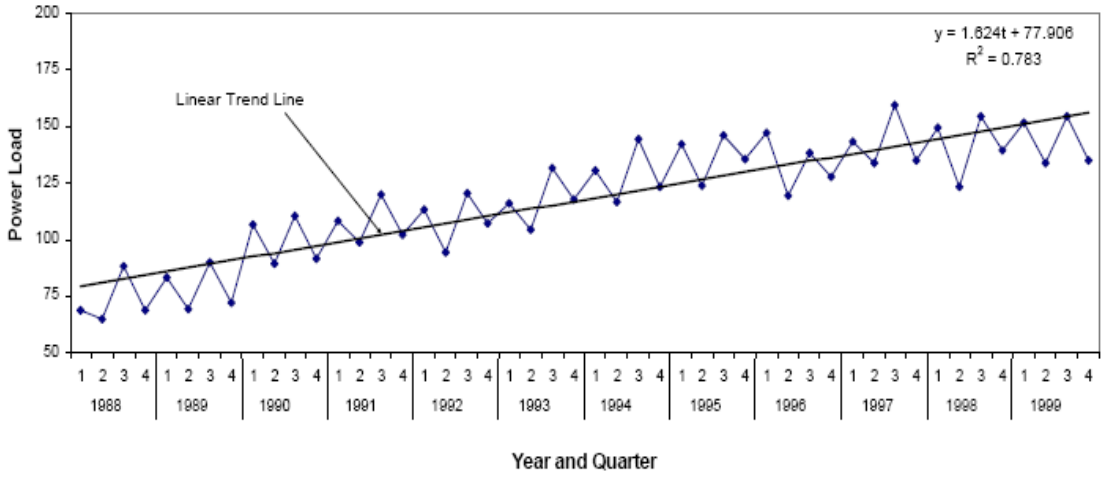
Το πρώτο βήμα για τη συστημική μελέτη μιας χρονοσειράς είναι η επισκόπηση του γραφήματός της στο πεδίο του χρόνου. Οι κλασσικές μέθοδοι ανάλυσης χρονοσειρών αποσυνθέτουν την χρονοσειρά σε τέσσερα βασικά στοιχεία: την τάση, την κυκλικότητα, την εποχικότητα και τις τυχαίες ή μη κανονικές διακυμάνσεις.

i. Τάση

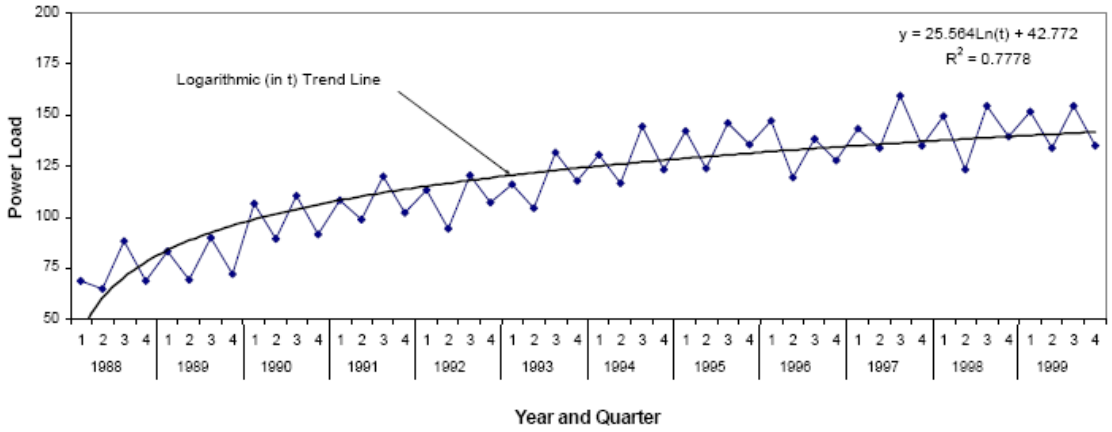
Η τάση (T) ή αλλίως trend υποδεικνύει την μακροπρόθεσμη κατεύθυνση μεταβολής του μέσου επιπέδου τιμών της χρονοσειράς. Μια χρονοσειρά μπορεί να χαρακτηρίζεται από αύξουσα η φθίνουσα τάση. Υπάρχουν περιπτώσεις που το επίπεδο των τιμών δεν μεταβάλλεται σημαντικά στην οποία περίπτωση θεωρούμε πως η χρονοσειρά δεν εμφανίζει τάση και ονομάζεται στάσιμη όπως στο Γράφημα 6.1. Η τάση μπορεί να είναι είτε γραμμική είτε μη γραμμική. Οι πιο συχνές περιπτώσεις χρονοσειρών που εμφανίζουν τάση είναι οι εξής: α) Γραμμική (Γράφημα 6.2), β) Λογαριθμική (Γράφημα 5.3), γ) Εκθετική και δ) Πολυωνυμική. Για την αναγνώριση μιας σαφούς τάσης και τον ακριβή καθορισμό της απαιτείται μεγάλος αριθμός παρατηρήσεων. Είναι πιθανό, ειδικά σε χρονοσειρές που εμφανίζουν κυκλικές μεταβολές, η εξέταση ενός μικρού τμήματος δεδομένων να οδηγήσει σε λανθασμένα συμπεράσματα. Θα πρέπει να υπάρχουν επαρκή δεδομένα ανάλογα με τα εξεταζόμενα μεγέθη και τα χαρακτηριστικά τους και η τάση να αναζητείται σε ανάλογα διαστήματα δεδομένων.



Γράφημα 6.1: Παράδειγμα Στάσιμης Χρονοσειράς



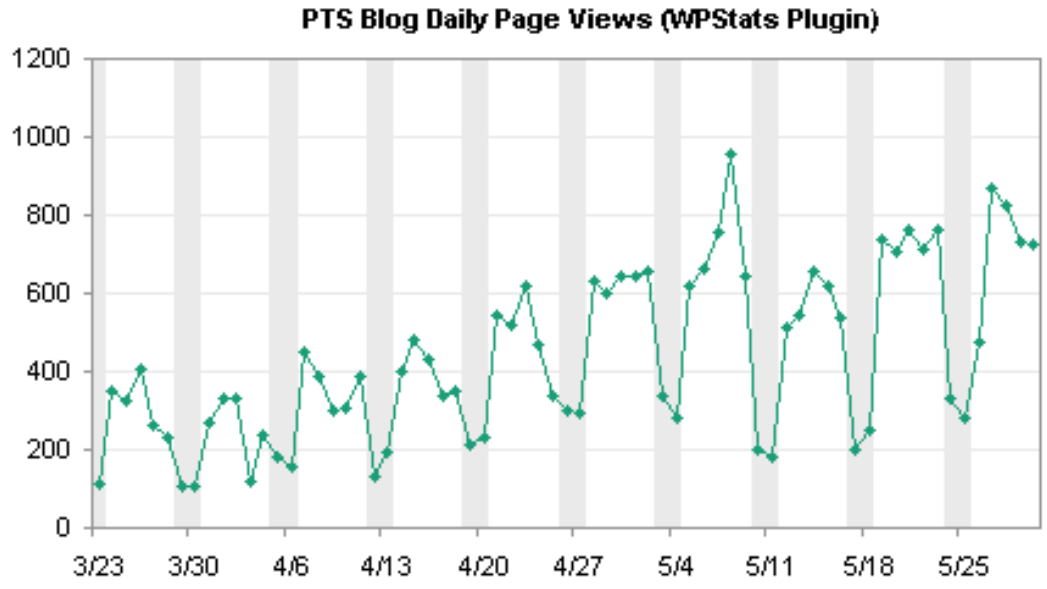
Γράφημα 6.2: Παράδειγμα Χρονοσειράς με γραμμική αύξουσα τάση



Γράφημα 6.3: Παράδειγμα Χρονοσειράς με λογαριθμική αύξουσα τάση

ii. Κυκλικότητα

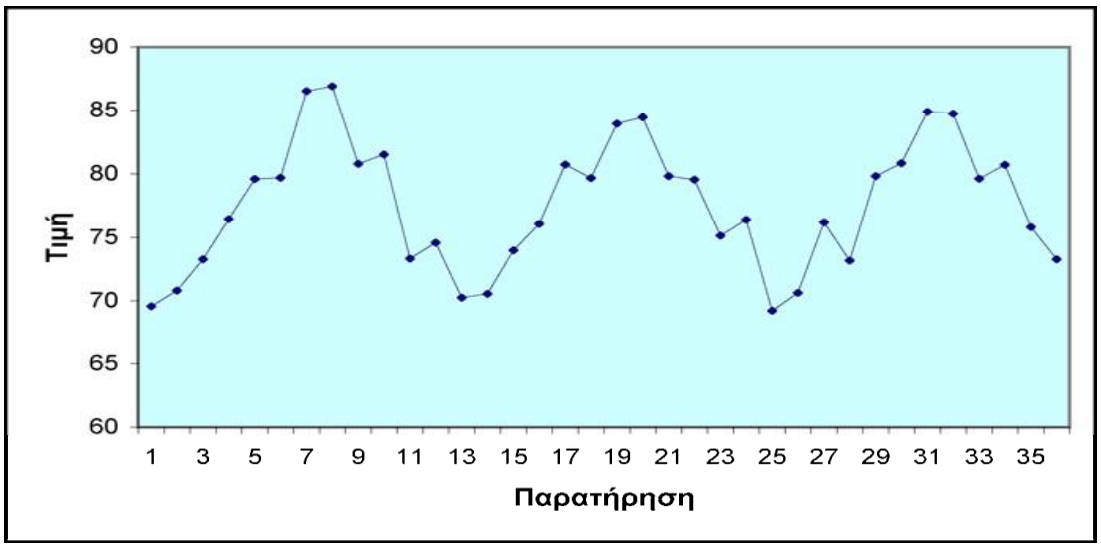
Η κυκλικότητα (C) (ή κυκλικές μεταβολές όπως αναφέρθηκαν παραπάνω) αποτελεί μια περιοδική μεταβολή της χρονοσειράς γενικά μεγάλες χρονικές περιόδους. Συχνά απαντάται με τον όρο κυκλική κύμανση (Cyclical Fluctuation) Τα αίτια αυτών των μεταβολών είναι εξωγενείς παράγοντες ενώ απαιτείται πολύ μεγάλος όγκος δεδομένων για να διαπιστωθεί η κυκλική συμπεριφορά ενός μεγέθους. Κυκλική συμπεριφορά παρουσιάζουν οι περισσότεροι οικονομικοί δείκτες όπως το ΑΕΠ μιας χώρας, καθώς και οικονομικά μεγέθη όπως η τιμή του πετρελαίου και του χρυσού. Η επισκεψιμότητα κάποιων ιστοσελίδων, ειδικά εγχώριων, επίσης παρουσιάζει τάσεις κυκλικότητας με υψηλά νούμερα τη μέρα και χαμηλά τη νύχτα όπως στο Γράφημα 6.4.



Γράφημα 6.4: Παράδειγμα Χρονοσειράς με κυκλικότητα

iii. Εποχικότητα

Η εποχικότητα (S) ή εποχιακή εξάρτηση ορίζεται ως μια περιοδική διακύμανση στη χρονοσειρά με σταθερή περίοδο που είναι κατά κανόνα μικρότερη του έτους. Τα αίτια της εποχικότητας είναι εύκολα αντιληπτά καθώς επηρεάζουν κατά τον ίδιο τρόπο τη χρονοσειρά σε κάθε χρονική περίοδο. Η ευκολία στην αναγνώριση και ο σταθερός τρόπος που επηρεάζεται η πορεία της χρονοσειράς καθιστούν εύκολη την απομόνωση της εποχικότητας για την παραγωγή προβλέψεων.



Γράφημα 6.5: Παράδειγμα Χρονοσειράς με σταθερή εποχικότητα

iv. Ασυνέχεια

Ασυνεχείς ονομάζονται οι απομονωμένες παρατηρήσεις που εμφανίζονται στη γραφική παράσταση μιας χρονοσειράς ως απότομες αλλαγές στο πρότυπο συμπεριφοράς της και δε θα μπορούσαν να έχουν προβλεφθεί από το ιστορικό υπόβαθρό της. Οι μεταβολές αυτές διακρίνονται σε δύο κατηγορίες, α. τις ασυνήθιστες τιμές (outliers) και β. τις αλλαγές επιπέδου, ανάλογα με την χρονική τους διάρκεια. Η πρώτη είναι αυτή των ασυνήθιστων τιμών, της οποίας το χαρακτηριστικό είναι η μικρή διάρκεια. Η δεύτερη είναι αυτή των αλλαγών επιπέδου, η οποία σε μεγάλο χρονικό διάστημα και σε μεγάλο βαθμό είναι η υπαίτιος για την αλλαγή επιπέδου της χρονοσειράς.

v. Μη κανονικές Διακυμάνσεις ή Τυχαιότητα

Οι μη κανονικές διακυμάνσεις είναι οι απρόβλεπτοι παράγοντες κάθε χρονοσειράς αποτελώντας το κύριο στοιχείο σφάλματος και είναι η εναπομένουσα συνιστώσα μετά

την διαδικασία της αφαίρεσης των συνιστωσών της τάσης, της κυκλικότητας και της εποχικότητας. Οι διακυμάνσεις αυτές μπορεί να αντιπροσωπεύουν μια εντελώς τυχαία μεταβλητή που εκφράζει τον τυχαίο παράγοντα μιας στοχαστικής διαδικασίας ή ακόμα κάποια ασυνέχεια που συνδέεται με κάποιο γεγονός.

6.2.2 Διαχείριση κενών και μηδενικών τιμών

Κατά τη συλλογή και διαχείριση των δεδομένων που αποτελούν τις χρονοσειρές υπάρχει το ενδεχόμενο ελλειπουσών ή μηδενικών τιμών οι οποίες δημιουργούν προβλήματα στην εφαρμογή των περισσότερων στατιστικών μεθόδων πρόβλεψης. Οι κενές τιμές αφορούν περιπτώσεις όπου η τιμή κάποιων περιόδων δεν είχε καταγραφεί και αποθηκευτεί στη βάση των δεδομένων είτε λόγω του πληροφοριακού συστήματος είτε από λάθος χειρισμό του υπεύθυνου χρήστη.
Σε περίπτωση ελλείπουσας τιμής ακολουθείται μία από τις παρακάτω διαδικασίες εκτίμησης ανάλογα με την περίπτωση:

- Εύρεση της κενής τιμής από άλλες πηγές ή απευθείας ορισμός αυτής, αν υπάρχει ασφαλής κριτική εκτίμηση
- Η κενή τιμή ορίζεται ως το ημιάθροισμα της προηγούμενης και της επόμενης παρατήρησης, όταν η χρονοσειρά χαρακτηρίζεται από στασιμότητα και δεν παρατηρείται εποχιακή συμπεριφορά.
- Αν η χρονοσειρά παρουσιάζει σαφή εποχιακή συμπεριφορά η κενή τιμή ορίζεται ως ο μέσος όρος των τιμών των αντίστοιχων περιόδων. Για παράδειγμα, αν τα δεδομένα αποτελούνται από μηνιαίες παρατηρήσεις και είναι κενή η τιμή σε κάποιο μήνα ενός έτους, η κενή τιμή ορίζεται ως ο μέσος όρος των ίδιων μηνών.

6.3 Κατηγορίες Μεθόδων Πρόβλεψης

Οι μέθοδοι πρόβλεψης χωρίζονται σε τρεις μεγάλες κατηγορίες: τις ποσοτικές, τις κριτικές και τις τεχνολογικές μεθόδους. Στην εργασία αυτή χρησιμοποιούνται και ποσοτικές και κριτικές μέθοδοι πρόβλεψης.

6.3.1 Ποσοτικές Μέθοδοι Πρόβλεψης

Οι ποσοτικές μέθοδοι προβλέψεων εφαρμόζονται όταν η διαθέσιμη πληροφορία διαμορφώνεται με την μορφή αριθμητικών δεδομένων και με την υπόθεση ότι το πρότυπο συμπεριφοράς των ιστορικών αυτών δεδομένων διατηρείται σταθερό στο μέλλον, κάτι το οποίο αποτελεί παράλληλα και το βασικό μειονέκτημα των μεθόδων αυτών. Η αλλαγή, δηλαδή, του προτύπου λόγω ενός γεγονότος αποτελεί σημαντικό παράγοντα αστοχίας καθώς το μοντέλο αυτό δεν μπορεί να συσχετίσει το προς πρόβλεψη μέγεθος με τους παράγοντες που το επηρεάζουν. Ωστόσο είναι ευρέως διαδεδομένα λόγω της ευκολίας χρήσης τους, του χαμηλού κόστους και της αδυναμίας συσχετισμού ενός μεγέθους με τους παράγοντες που το επηρεάζουν.

Οι ποσοτικές μέθοδοι ανάλογα με το μοντέλο που χρησιμοποιείται ταξινομούνται σε δύο μεγάλα μοντέλα, στα μοντέλα χρονοσειρών και τα αιτιοκρατικά μοντέλα.

6.3.1.1 Μοντέλο χρονοσειρών

Αποτελεί το πιο κλασσικό είδος ποσοτικού μοντέλου πρόβλεψης. Η εφαρμογή του μοντέλου αυτού είναι εφικτή όταν υπάρχουν στοιχεία για την τιμή του υπό πρόβλεψη μεγέθους (ιστορικά δεδομένα) σε προηγούμενες και σταθερές χρονικές περιόδους. Βασική του υπόθεση είναι ότι η μεταβολή του προς πρόβλεψη μεγέθους ακολουθεί ένα λανθάνον πρότυπο συμπεριφοράς που επαναλαμβάνεται στο μέλλον. Στα μοντέλα χρονοσειρών περιλαμβάνονται οι μέθοδοι αποσύνθεσης, οι μέθοδοι εξομάλυνσης καθώς και οι αυτόπαλινδρομικές μέθοδοι κινητού μέσου όρου.

- **Μέθοδοι αποσύνθεσης**

Οι μέθοδοι αποσύνθεσης αναγνωρίζουν τις τέσσερεις ξεχωριστές συνιστώσες που χαρακτηρίζουν τις χρονοσειρές και τις απομονώνουν. Όπως έχει προαναφερθεί πρόκειται για την τάση (T), τον κύκλο (C), την εποχικότητα (S) και την τυχαιότητα (R). Σκοπός των μεθόδων αποσύνθεσης είναι η απομόνωση των συνιστωσών αυτών με τη μέγιστη δυνατή ακρίβεια. Το ποσοστό του λάθους κατά αυτόν τον τρόπο οφείλεται στην τυχαιότητα. Η πιο διαδεδομένη λόγω και της ευκολίας υλοποίησής της είναι η Κλασσική Μέθοδος Αποσύνθεσης.

- **Μέθοδοι εξομάλυνσης**

Οι μέθοδοι εξομάλυνσης εφαρμόζονται εύκολα και παρέχουν ικανοποιητικές προβλέψεις στο βραχυπρόθεσμο ορίζοντα. Στις μεθόδους εξομάλυνσης γίνεται προσπάθεια διάκρισης του βασικού προτύπου από τις τυχαίες διακυμάνσεις εξομαλύνοντας τα δεδομένα. Έτσι, ελαχιστοποιείται η τυχαιότητα που υπάρχει στην χρονοσειρά με την πρόβλεψη πλέον να βασίζεται σε ένα εξομαλυμένο πρότυπο συμπεριφοράς. Υπάρχουν δύο κατηγορίες μεθόδων εξομάλυνσης: οι κινητού μέσου όρου, όπου οι παρελθούσες τιμές συμμετέχουν με τα ίδια βάρη στην παραγωγή προβλέψεων και οι εκθετικής εξομάλυνσης, όπου οι συντελεστές βαρύτητας φθίνουν εκθετικά για τα πιο μακρινά δεδομένα.

- **Αυτοπαλινδρομικές μέθοδοι κινητού μέσου όρου (ARIMA)**

Οι αυτοπαλινδρομικές μέθοδοι κινητού μέσου όρου είναι στοχαστικά μαθηματικά μοντέλα με τα οποία περιγράφουμε την διαχρονική εξέλιξη κάποιου φυσικού μεγέθους. Τα στοχαστικά μοντέλα περιέχουν το τυχαίο παράγοντα, τις τιμές του μεγέθους για τις προηγούμενες χρονικές στιγμές όπως και άλλους στοχαστικούς παράγοντες συνήθως. Το μοντέλο που προκύπτει τελικά είναι ένας γραμμικός συνδυασμός των παραπάνω ποσοτήτων. Τα αυτοπαλινδρομικά μοντέλα βασίζονται στην παραδοχή της αλληλεξάρτησης μεταξύ των τιμών που λαμβάνει η χρονοσειρά τις διάφορες χρονικές στιγμές.

6.3.1.2 Αιτιοκρατικό μοντέλο

Στις επεξηγηματικές μεθόδους, αντί της προσαρμογής κάποιου μοντέλου στην χρονοσειρά (όπως το μοντέλο εκθετικής εξομάλυνσης), αναγνωρίζονται ορισμένες μεταβλητές οι οποίες σχετίζονται με τη σειρά δεδομένων και αναπτύσσεται κάποιο μοντέλο προκειμένου να εκφράσει τη σχέση αυτή. Η πρόβλεψη εκφράζεται ως συνάρτηση κάποιου συγκεκριμένου αριθμού παραγόντων που επηρεάζουν την τελική τιμή της. Δεν είναι απαραίτητο να υπάρχει χρονική εξάρτηση.

Έτσι, αναπτύσσεται ένα μοντέλο το οποίο καθιστά ευκολότερη την κατανόηση των συνθηκών και δίνει τη δυνατότητα πρόβλεψης μελλοντικής τιμής κάποιου μεγέθους μέσω διαφόρων συνδυασμών τιμών των ανεξάρτητων μεταβλητών. Μειονέκτημα

αυτών των μεθόδων είναι το μεγάλο πλήθος δεδομένων καθώς υπάρχει απαίτηση δεδομένων σχετικών με τις ανεξάρτητες μεταβλητές. Ακόμα, πολλές φορές, η πρόβλεψη με βάση αιτιοκρατικά μοντέλα προϋποθέτει πρόβλεψη και των ανεξάρτητων μεταβλητών, κάτι που συνεπάγεται και αυξημένο κόστος εφαρμογής. Στα αιτιοκρατικά μοντέλα ανήκουν οι μέθοδοι παλινδρόμησης και οι οικονομετρικές μέθοδοι.

- **Μέθοδοι παλινδρόμησης**

Υποθέτουμε την ύπαρξη γραμμικής σχέσης ανάμεσα στη μεταβλητή της οποίας την τιμή θέλουμε να προβλέψουμε (εξαρτημένη μεταβλητή) και έναν αριθμό ανεξάρτητων μεταβλητών. Στην περίπτωση της μίας ανεξάρτητης μεταβλητής η μέθοδος ονομάζεται "Απλή Παλινδρόμηση" ενώ στην περίπτωση περισσοτέρων, "Πολλαπλή Παλινδρόμηση".

- **Οικονομετρικές μέθοδοι**

Εάν οι ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους τότε προκύπτει ένα σύστημα ταυτόχρονων εξισώσεων. Αυτό το σύστημα εξισώσεων αποτελεί ένα οικονομετρικό μοντέλο και συναντάται συχνά σε περιπτώσεις οικονομικών ή επιχειρησιακών σχέσεων.

6.3.2 Κριτικές Μέθοδοι Πρόβλεψης

Οι κριτικές μέθοδοι πρόβλεψης (judgmental forecasting methods) βασίζονται στην εμπειρία, τη διαίσθηση, την ψυχολογία και τις γνώσεις των ατόμων που την εκτελούν. Οι μέθοδοι αυτές είναι ευρέως διαδεδομένες σε οργανισμούς και επιχειρήσεις και δεν απαιτούν μεγάλο όγκο δεδομένων. Η πρόβλεψη μπορεί να βασίζεται είτε στις γνώσεις και την κρίση ενός ατόμου (ατομικές μέθοδοι) είτε στο συνδυασμό απόψεων των μελών κάποιας επιτροπής (μέθοδοι επιτροπής). Ένα βασικό πλεονέκτημα των κριτικών μεθόδων πρόβλεψης είναι πως ειδικά γεγονότα και ενέργειες μπορούν να ληφθούν υπόψη σε αντίθεση με τις μεθόδους χρονοσειρών. Το μεγάλο, όμως, πρόβλημα αυτής της μεθόδου είναι η τυχούσα προκατάληψη που μπορεί να υπάρχει κατά την παραγωγή μιας πρόβλεψης.

6.3.3 Τεχνολογικές Μέθοδοι Πρόβλεψης

Οι τεχνολογικές μέθοδοι πρόβλεψης χρησιμοποιούνται για μακροπρόθεσμες προβλέψεις τεχνολογικού, οικονομικού, κοινωνικού και πολιτικού περιεχομένου. Διακρίνονται σε δύο κατηγορίες: στις διερευνητικές (exploratory) και στις κανονιστικές (normative). Οι διερευνητικές μέθοδοι έχουν ως σημείο εκκίνησης το παρελθόν και το παρόν και προχωρούν στο μέλλον διερευνώντας όλες τις πιθανές περιπτώσεις. Από την άλλη, οι κανονιστικές έχουν προκαθορισμένους στόχους και απλά μελετούν τη δυνατότητα πραγματοποίησης με τους υπάρχοντες περιορισμούς και τους διαθέσιμους πόρους.

6.4 Κυριότερες Μέθοδοι Πρόβλεψης

6.4.1 Απλοϊκή Μέθοδος (Naive)

Είναι η πιο απλή στατιστική μέθοδος πρόβλεψης. Ως πρόβλεψη θεωρείται η τελευταία διαθέσιμη παρατήρηση. Δηλαδή:

$$F_t = Y_{t-1}$$

Η τεχνική αυτή ενδείκνυται για περιπτώσεις που τα δεδομένα δεν παρουσιάζουν τάση και για μικρούς ορίζοντες πρόβλεψης. Χρησιμοποιείται κυρίως ως μέτρο σύγκρισης για την ακρίβεια άλλων μεθόδων (benchmark) και για την τεχνική back casting. Η τεχνική back casting χρησιμοποιείται σε περιπτώσεις που εμφανίζονται κενές τιμές σε μια χρονοσειρά και για την συμπλήρωση των κενών αυτών.

6.4.2 Μέθοδοι εκθετικής εξομάλυνσης

Η εκθετική εξομάλυνση είναι μια μέθοδος πρόβλεψης η οποία προεκτείνει στοιχεία του προτύπου των ιστορικών δεδομένων, όπως τάσεις και εποχιακούς κύκλους, στο μέλλον. Οι προβλέψεις υπολογίζονται μετά από εξομάλυνση των δεδομένων, προκειμένου να απομονωθούν τα πραγματικά πρότυπα από τις τυχαίες διακυμάνσεις. Η δημοτικότητα των μεθόδων αυτών οφείλεται στην απλότητα των μοντέλων που υιοθετούν, τις περιορισμένες απαιτήσεις τους σε αποθήκευση δεδομένων και τον μειωμένο υπολογιστικό φόρτο. Εμπειρικές μελέτες αποδεικνύουν ότι οι μέθοδοι εκθετικής εξομάλυνσης παρουσιάζουν ικανοποιητικά ποσοστά ακρίβειας σε σχέση με πιο πολύπλοκες μεθόδους πρόβλεψης. Το γεγονός αυτό οφείλεται στο ότι οι μέθοδοι εκθετικής εξομάλυνσης δεν επηρεάζονται από τις ιδιομορφίες των προτύπων των

δεδομένων ή από περιστασιακά εμφανιζόμενες ακραίες τιμές, οι οποίες παρατηρούνται σε επιχειρησιακά δεδομένα.

6.4.2.1 Απλή Εκθετική Εξομάλυνση (Simple Exponential Smoothing)

Το μοντέλο αυτό ονομάζεται και μοντέλο απλής εκθετικής εξομάλυνσης ή SES. Οι ακόλουθες εξισώσεις περιγράφουν το μοντέλο σταθερού επιπέδου.

$$e_t = Y_t - \mathrm{F_t}$$

$$S_t = S_{t-1} + \mathrm{a} \cdot e_t$$

$$\mathrm{F_{t+1}} = S_t$$

Στις παραπάνω εξισώσεις, το e δηλώνει το σφάλμα πρόβλεψης, το S το επίπεδο, F την πρόβλεψη και α μια σταθερά εξομάλυνσης που λαμβάνει οποιαδήποτε τιμή στο διάστημα [0,1]. Το μοντέλο αυτό είναι κατάλληλο για δεδομένα που δεν παρουσιάζουν έντονο το στοιχείο της τάσης.

Το αρχικό επίπεδο ορίζεται με διάφορούς τρόπους και συνήθως χρησιμοποιείται ο μέσος όρος όλων ή κάποιων αρχικών παρατηρήσεων, η πρώτη παρατήρηση ή το σταθερό επίπεδο από το μοντέλο της απλής γραμμικής παλινδρόμησης. Ο σημαντικότερος όμως παράγοντας είναι ο καθορισμός του συντελεστή εξομάλυνσης α. Ο καθορισμός του α εξαρτάται από τον θόρυβο που έχουν τα δεδομένα και από (όσο περισσότερος τόσο μικρότερη τιμή παίρνει ο α) και από την σταθερότητα του μέσου όρου της χρονοσειράς (μεγάλες μεταβολές αντιμετωπίζονται με μεγαλύτερο α). Υπάρχουν διάφοροι αλγόριθμοι εύρεσης του κατάλληλου α, συνήθως με την εύρεση εκείνου που ελαχιστοποιεί κάποιο δείκτη σφάλματος. Για τις ακραίες τιμές α=1, η πρόβλεψη γίνεται ίδια με την naive ενώ για α=0, η πρόβλεψη παραμένει ίδια και ίση με το αρχικό επίπεδο. Είναι εμφανές πως σε περιπτώσεις που απαιτούνται προβλέψεις ορίζοντα μεγαλύτερου από μια χρονική περίοδο, όλες οι προβλέψεις είναι ίδιες με την τελευταία.

6.4.2.2 Μοντέλο Γραμμικής Τάσης (Holt Exponential Smoothing)

Το μοντέλο γραμμικής τάσης αποτελεί μια εξέλιξη του μοντέλου σταθερού επιπέδου και δίνει τη δυνατότητα διαχείρισης δεδομένων που παρουσιάζουν το στοιχείο της τάσης. Το μοντέλο περιγράφεται από τις παρακάτω εξισώσεις:

$$e_t = Y_t - \mathrm{F_t}$$

$$S_t = S_{t-1} + T_{t-1} + \mathrm{a} \cdot e_t$$

$$T_t = T_{t-1} + \mathrm{b} \cdot e_t$$

$$\mathrm{F_{t+m}} = S_t + m \cdot T_t$$

Η νέα παράμετρος β που εμπεριέχεται στις εξισώσεις ονομάζεται συντελεστής εξομάλυνσης τάσης και λαμβάνει τιμές στο διάστημα [0,1]. Σε αυτό το μοντέλο χρειάζεται αρχικοποίηση τόσο του επιπέδου όσο και της τάσης. Το αρχικό επίπεδο ορίζεται όπως στην απλή εκθετική εξομάλυνση, ενώ η αρχική τάση ως η διαφορά της ν-στής και της πρώτης παρατήρησης διαιρεμένης με ν-1, ή ως η σταθερά κλίσης από το μοντέλο της απλής γραμμικής παλινδρόμησης. Το αρχικό επίπεδο και η αρχική τάση πρέπει να καθορίζονται με προσοχή καθώς επηρεάζουν αρκετά την τελική πρόβλεψη. Η μεγάλη διαφορά του μοντέλου αυτού από τη μέθοδο SES είναι η παραγωγή προβλέψεων με χρονικό ορίζοντα μεγαλύτερο της μονάδας. Λόγω της θεώρησης πως τα δεδομένα έχουν μια σταθερά ανοδική τάση, οι προβλέψεις για ορίζοντα μεγαλύτερο της μονάδας προκύπτουν με τη χρήση των τελευταίων διαθέσιμων τιμών για το επίπεδο και την τάση και αύξηση του δείκτη m.

6.4.2.3 Μοντέλο Μη Γραμμικής Τάσης (Damped Exponential Smoothing)

Το μοντέλο φθίνουσας γραμμικής τάσης είναι μία υποπερίπτωση του μοντέλου μη γραμμικής τάσης. Το μοντέλο μη γραμμικής τάσης έχει τη δυνατότητα μεταβολής της μορφής της χρονοσειράς και της προσαρμογής της σε μη γραμμικές τάσεις. Η προσαρμογή αυτή γίνεται μέσω μιας μεταβλητής που ονομάζεται παράμετρος διόρθωσης της τάσης φ. Το μοντέλο μη γραμμικής τάσης περιγράφεται μαθηματικά από τις παρακάτω εξισώσεις:

$$\begin{aligned}
\mathrm{e_t} &= \mathrm{Y_t} - \mathrm{F_t} \\
\mathrm{S_t} &= \mathrm{S_{t-1}} + \mathrm{T_{t-1}} + \alpha \cdot \mathrm{e_t} \\
\mathrm{T_t} &= \mathrm{T_{t-1}} + \beta \cdot \mathrm{e_t} \\
\mathrm{F_{t+1}} &= \mathrm{S_t} + \sum_{\mathrm{i=1}}^{\mathrm{m}} \varphi^{\mathrm{i}} \cdot \mathrm{T_t}
\end{aligned}$$

Όπου, t η χρονική περίοδος, Yt η πραγματική τιμή των δεδομένων, Ft η πρόβλεψη τη χρονική στιγμή t, et το σφάλμα (απόκλιση πραγματικής τιμής από πρόβλεψη), St το επίπεδο της χρονοσειράς, Tt η τάση της χρονοσειράς, α ο συντελεστής εξομάλυνσης επιπέδου, λαμβάνει τιμές στο διάστημα [0,1], β ο συντελεστής εξομάλυνσης της τάσης, λαμβάνει τιμές στο διάστημα [0,1], φ ο συντελεστής διόρθωσης της τάσης, λαμβάνει τιμές στο διάστημα (0,1) και m χρονικός ορίζοντας της πρόβλεψης.

Εύκολα γίνεται αντιληπτό, ότι οι εξισώσεις είναι πανομοιότυπες με αυτές του γραμμικού μοντέλου πλην της τελευταίας, όπου αντί να υπολογίζεται μια γραμμική αύξηση μέσω του συντελεστή m, πραγματοποιείται ένας μη γραμμικός υπολογισμός αυτής, γεγονός που οφείλεται στην παράμετρο εξομάλυνσης φ. Η παράμετρος φ, σε αντίθεση με τις παραμέτρους α και β, δύναται να λάβει τιμές μεγαλύτερες του μηδενός, χωρίς κάποιο άνω όριο αλλά είναι πολύ σημαντική η επιβολή άνω και κάτω ορίων ανάλογα με την εκάστοτε περίπτωση.

Όπως αναφέρεται και παραπάνω για $0<\varphi<1$ προκύπτει το μοντέλο της φθίνουσας τάσης (Damped Exponential Smoothing). Ανάλογα την τιμή που παίρνει η παράμετρος φ, το μοντέλο της μη γραμμικής τάσης μπορεί να πάρει περεταίρω τις μορφές:

- Για $\varphi=0$ προκύπτει το μοντέλο της απλής εκθετικής εξομάλυνσης (Simple Exponential Smoothing), αφού η τάση δεν συμμετέχει στην παραγωγή προβλέψεων.
- Για $\varphi=1$ προκύπτει το μοντέλο της γραμμικής τάσης (Holt Exponential Smoothing), καθώς στην εξίσωση υπολογισμού της πρόβλεψης, τη θέση του αθροίσματος παίρνει το γινόμενο της μεταβλητής χρονικού ορίζοντα m κα της προηγούμενης τάσης T_t.
- Για $\varphi>1$ προκύπτει το μοντέλο της εκθετικής τάσης, το οποίο χαρακτηρίζεται από μεγάλη προκατάληψη.

Σχετικά με την επιλογή του αρχικού επιπέδου (S_0), της αρχικής τάσης (T_0) και την βελτιστοποίηση των παραμέτρων εξομάλυνσης, ισχύουν τα ίδια που αναφέρθηκαν παραπάνω για την περίπτωση του μοντέλου γραμμικής τάσης. Συγκεκριμένα για την μη γραμμική τάση προτείνεται ωστόσο η εφαρμογή της γραμμικής παλινδρόμησης με ανεξάρτητη μεταβλητή το χρόνο t για τον προσδιορισμό των S_0 και T_0. Για την εύρεση των βέλτιστων συνδυασμών των παραμέτρων α, β, φ εφαρμόζεται και πάλι η

διαδικασία της γραμμικής αναζήτησης, ελαχιστοποιώντας το μέσο τετραγωνικό σφάλμα (MSE).

Λόγω της θετικής προκατάληψης που περιέχει το μοντέλο εκθετικής τάσης χρησιμοποιείται σε ορισμένες μόνο ειδικές περιπτώσεις, όπως η εισαγωγή ενός προϊόντος στην αγορά. Θετική προκατάληψη εντοπίζεται και στα μοντέλα γραμμικής τάσης. Γι' αυτό το λόγο τα μοντέλα φθίνουσας τάσης τυγχάνουν μεγάλης αποδοχής ιδιαίτερα για προβλέψεις μεγάλου χρονικού ορίζοντα. Εμπειρικά αποτελέσματα φαίνεται να δικαιολογούν την επιλογή αυτή.

6.4.3 Αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου (μέθοδος ARIMA)

Τα αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου, ανήκουν στα στοχαστικά μαθηματικά μοντέλα και με την βοήθειά τους μπορούμε να περιγράψουμε την διαχρονική εξέλιξη φυσικών μεγεθών, που εξαρτώνται από μη ντετερμινιστικούς παράγοντες. Είναι αρκετά διαδεδομένα, και ειδικά σε περιπτώσεις που εμπεριέχονται φυσικά μεγέθη, τα οποία δεν τα γνωρίζουμε απόλυτα, και επιπλέον όταν δεν γνωρίζουμε τους παράγοντες οι οποίοι τα επηρεάζουν. Τα στοχαστικά αυτά μοντέλα περιέχουν τον τυχαίο παράγοντα, τις τιμές του μεγέθους οι οποίες εμφανίστηκαν σε παρελθοντικές χρονικές στιγμές και μπορεί και κάποιους επιπλέον στοχαστικούς παράγοντες. Πιο συγκεκριμένα, με την χρήση αυτών των μοντέλων, δυνάμεθα να υπολογίσουμε την πιθανότητα ή την τιμή του μεγέθους που εξετάζουμε να βρίσκεται σε ένα συγκεκριμένο διάστημα. Ως απόρροια όλων των παραπάνω, μπορούμε να αντιληφθούμε ότι τα Αυτοπαλινδρομικά μοντέλα κινητού μέσου όρου είναι αρκετά αποτελεσματικά κυρίως σε βραχυπρόθεσμες προβλέψεις, εφόσον δίνουν μεγαλύτερη έμφαση στις πιο πρόσφατες παρελθοντικές παρατηρήσεις. Βασική προϋπόθεση για την καλύτερα αποτελέσματα στα εξής μοντέλων είναι να εφαρμόζονται σε χρονοσειρές οι οποίες είναι στάσιμες και διακριτές. Διακριτές είναι οι χρονοσειρές που όλες οι παρατηρήσεις τους έχουν ληφθεί σε χρονικές στιγμές που ισαπέχουν μεταξύ τους, ενώ στάσιμες θεωρούνται αυτές που η μέση τιμή, η διακύμανσή τους και η συνάρτηση αυτοσυσχέτισής τους είναι σταθερές σε όλη την διάρκεια του χρόνου.

6.5 Σφάλματα

Αναλύοντας τα μοντέλα προβλέψεων γίνεται εμφανής η ανάγκη αξιολόγησης των αποτελεσμάτων, καθώς και η επιλογή του κατάλληλου μοντέλου για κάθε χρονοσειρά δεδομένων. Ως μέτρο ακρίβειας και καλής προσαρμογής του εκάστοτε μοντέλου χρησιμοποιούνται κάποιοι στατιστικοί δείκτες ακρίβειας προβλέψεων. Έχοντας ήδη ορίσει το στατιστικό σφάλμα ως τη διαφορά της πραγματικής τιμής από την προβλεπόμενη για μία περίοδο ($e_t = Y_t - F_t$) παραθέτονται στη συνέχεια οι ορισμοί των κυριότερων τύπων σφαλμάτων.

- Απόλυτο σφάλμα (Absolute Error)

$$e_t = |Y_t - F_t|$$

- Μέσο σφάλμα (Mean Error): Συχνά ο δείκτης αναφέρεται και ως bias καθώς αν παίρνει θετικές τιμές, δηλώνει απαισιοδοξία, ενώ αντίθετα αν παίρνει αρνητικές, δηλώνει αισιοδοξία. Αυτό συμβαίνει διότι οι προβλέψεις ήταν κατά μέσο όρο μικρότερες, και μεγαλύτερες αντίστοιχα, από τις πραγματικές τιμές. Όσο δε ο δείκτης βρίσκεται κοντά στο μηδέν τότε αντιλαμβάνεται κανείς ότι τα σφάλματα είναι τυχαία και όχι συστηματικά.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)$$

- Μέσο απόλυτο σφάλμα (Mean Absolute Error): Ο δείκτης αυτός δηλώνει ένα μέσο μέτρο αστοχίας της πρόβλεψης αγνοώντας την κατεύθυνσή της. Όσο πιο μεγάλη τιμή παίρνει, τόσο μικρότερη είναι η ακρίβεια της πρόβλεψης που εφαρμόστηκε.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - F_i|$$

• Μέσο τετραγωνικό σφάλμα (Mean Squared Error): Όπως το μέσο απόλυτο σφάλμα, έτσι και αυτό, υπολογίζει την ακρίβεια της πρόβλεψης που εφαρμόστηκε. Η διαφορά έγκειται στο γεγονός ότι δίνεται βάρος στα μεγάλα σφάλματα σε σχέση με τα μικρότερα. Η χρήση του δείκτη αυτού χρησιμεύει στη βελτιστοποίηση των παραμέτρων εξομάλυνσης.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2$$

• Ρίζα μέσου τετραγωνικού σφάλματος (Root Mean Squared Error): Έχει τις ιδιότητες του μέσου τετραγωνικού σφάλματος, με τη διαφορά ότι εκφράζεται στις μονάδες της αρχικής χρονοσειράς.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - F_i)^2}$$

• Μέσο απόλυτο ποσοστιαίο σφάλμα (Mean Absolute Percentage Error): Αυτός ο δείκτης σφάλματος είναι και αυτός ο οποίος θα χρησιμοποιηθεί και στο πείραμα στο οποίο έπεται στην παρούσα διπλωματική εργασία. Αποτελεί ίσως το πιο χρήσιμο μέτρο για την σύγκριση μεθόδων πρόβλεψης. Είναι πολύ χρήσιμο εργαλείο όταν οι πραγματικές τιμές είναι ιδιαίτερα υψηλές όπως στο πείραμά μας. Τέλος, λαμβάνει τιμές μεγαλύτερες του μηδενός σε ποσοστιαία μορφή και όσο πιο κοντά βρίσκεται στο μηδέν, τόσο πιο αποδοτική είναι η μέθοδος πρόβλεψης που χρησιμοποιήθηκε.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - F_i}{Y_i}\right| \cdot 100 \ \ (\%)$$

• Συμμετρικό μέσο απόλυτο ποσοστιαίο σφάλμα (Symmetric Mean Absolute Percentage Error): Αποτελεί μια παραλλαγή του MAPE στην οποία το απόλυτο του σφάλματος δε διαιρείται απλώς με την πραγματική τιμή αλλά με το ημιάθροισμα της πραγματικής τιμής και της πρόβλεψης. Υπολογίζεται από τον τύπο:

$$sMAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - F_i}{(\frac{Y_i + F_i}{2})}\right| \cdot 100 \ (\%)$$

• Μέσο απόλυτο κανονικοποιημένο σφάλμα (Mean Absolute Scaled Error): Προτάθηκε από τους Hyndman και Koehler(2006) για την αντιμετώπιση των περιπτώσεων απροσδιοριστίας των δεικτών MAPE και sMAPE αλλά και για να δοθεί η ίδια βαρύτητα στα μικρά και τα μεγάλα σφάλματα. Θυμίζει το μέσο απόλυτο σφάλμα είναι όμως κανονικοποιημένο με τη μέση τιμή των διαφορών πρώτου βαθμού της χρονοσειράς.

$$MAsE = \frac{\frac{1}{n}\sum_{i=1}^{n}|Y_i - F_i|}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}$$

6.6 Επιλογή της κατάλληλης μεθόδου πρόβλεψης

Η επιλογή της κατάλληλης μεθόδου, η οποία θα χρησιμοποιηθεί για την παραγωγή προβλέψεων, δεν είναι μια εύκολη διαδικασία που μπορεί να γίνει από απλούς χρήστες ενός προγράμματος. Για την επιλογή της βέλτιστης μεθόδου πρόβλεψης μπορούμε να εξετάσουμε ορισμένους παράγοντες που επηρεάζουν το βαθμό εφαρμογής μιας μεθόδου και ως εκ τούτου τα αποτελέσματά της. Οι κυριότεροι λοιπόν παράγοντες είναι:

- **Χρονικός ορίζοντας**

Βασικό κριτήριο επιλογής μιας μεθόδου πρόβλεψης αποτελεί το χρονικό διάστημα στο μέλλον στο οποίο θα αναφέρεται η πρόβλεψη. Οι ποιοτικές μέθοδοι χρησιμοποιούνται περισσότερο για μακροπρόθεσμες προβλέψεις ενώ οι ποσοτικές μέθοδοι χρησιμοποιούνται περισσότερο για μεσοπρόθεσμες και βραχυπρόθεσμες προβλέψεις. Επίσης σημαντικό στοιχείο είναι και το πλήθος των περιόδων για το οποίο απαιτείται

πρόβλεψη. Ορισμένες τεχνικές είναι κατάλληλες για προβλέψεις που αντιστοιχούν σε 1 ή 2 περιόδους μετά από την πιο πρόσφατη παρατήρηση, ενώ άλλες σε περισσότερες.

- **Πρότυπο συμπεριφοράς των δεδομένων**

Βασική προϋπόθεση στην πλειοψηφία των μεθόδων πρόβλεψης είναι η αναγνώριση του προτύπου συμπεριφοράς των δεδομένων πάνω στο οποίο θα στηριχθεί η πρόβλεψη. Τα τέσσερα βασικά πρότυπα συμπεριφοράς που συχνά εμφανίζονται στις χρονοσειρές και τις περισσότερες φορές συνυπάρχουν είναι το σταθερό πρότυπο, το πρότυπο της τάσης, το εποχιακό και το κυκλικό πρότυπο. Είναι σημαντικό, λόγω του ότι η ικανότητα των διαφόρων μεθόδων να παράγουν αξιόπιστες προβλέψεις για διαφορετικά πρότυπα δεδομένων ποικίλλει, η μέθοδος που θα επιλεγεί να είναι κατάλληλη για το συγκεκριμένο πρότυπο.

- **Κόστος**

Το κόστος μιας μεθόδου πρόβλεψης καθορίζεται από τον όγκο των δεδομένων που απαιτεί η μέθοδος και από την πολυπλοκότητα της εφαρμογής της.

- **Αξιοπιστία**

Η αξιοπιστία είναι στενά συνδεδεμένη με το επίπεδο λεπτομέρειας που απαιτείται σε μια πρόβλεψη. Σε ορισμένες περιπτώσεις ένα ποσοστό ακρίβειας ±10% θεωρείται ικανοποιητικό, ενώ σε άλλες έστω και 1 διακύμανση της τάξης του ±5% μπορεί να αποδειχτεί καταστροφική.

- **Απλότητα και ευκολία εφαρμογής**

Έχει αποδειχτεί στην πράξη ότι προτιμώνται μέθοδοι που είναι κατανοητές και εύκολες στην εφαρμογή τους.

ΚΕΦΑΛΑΙΟ 7: Πείραμα

7.1 Εισαγωγή

Ένας από τους στόχους της παρούσας διπλωματικής εργασίας είναι η μελέτη στην αντικτύπου των δημοσιευμένων νέων στις κινήσεις των μετοχών. Από τη βιβλιογραφία συμπεραίνεται ότι η αξία των μετοχών, και των χρηματοοικονομικών προϊόντων εν γένει, συμπεριλαμβάνουν μέσα τους το στοιχείο του κλίματος της αγοράς που τις περιβάλλει. Το sentiment των ειδήσεων καθορίζει σε σημαντικό βαθμό τις κινήσεις των επενδυτών και τη λήψη των αποφάσεών τους. Πιο συγκεκριμένα η επιλογή των επενδυτικών στόχων βασίζεται στην επεξεργασία ενός μεγάλου πλήθους πληροφοριών. Τα fundamentals των εταιριών και η τεχνική ανάλυση των δεικτών των μετοχών τους είναι κυρίαρχα στην προεργασία των ανθρώπων της αγοράς για τις επενδυτικές τους κινήσεις ιδίως όταν πρόκειται για προσδοκία μακροπρόθεσμων κερδών. Όμως οι παίκτες του χρηματοοικονομικού γίγνεσθαι δεν υποτιμούν σε καμία περίπτωση την αίσθηση της αγοράς περί των μετοχικών και όχι μόνο. Άλλωστε γνωρίζουν καλά ότι η οικονομία και οι κατευθύνσεις των χρηματιστηρίων επηρεάζονται σε καθοριστικό βαθμό από τις ειδήσεις που μεταδίδονται με ασύλληπτη ταχύτητα σε όλα τα μήκη και πλάτη του πλανήτη αλλά και την ψυχολογία που αυτές διαμορφώνουν.

Στην παρούσα διπλωματική διατριβή προτείνεται για το μέλλον η δημιουργία ενός ολοκληρωμένου προγράμματος το οποίο να συγκεντρώνει την ειδησεογραφία που αφορά μία πιθανή επενδυτική κίνηση. Μετά την επεξεργασία της να εξάγει με ποσοτικοποιημένο τρόπο το συνολικό sentiment της αγοράς ως προς αυτήν την επένδυση και να συμβάλει καθοριστικά στη λήψη αποφάσεων. Προϋπόθεση όμως για να αρχίσει η υλοποίηση ενός τέτοιου μεγάλου εγχειρήματος είναι η δημιουργία μικρότερης κλίμακας πειραμάτων των οποίων η σύνθεση να οδηγήσει τελικά στην πραγματοποίηση του μεγάλου, όπως γίνεται σε αυτήν την εργασία.

Επιχειρείται με την υλοποίηση της παρακάτω διάταξης να μετρηθεί το αν για αρχή μία αξιόπιστη οικονομική είδηση η οποία έχει υποστεί επεξεργασία μέσω ανάλυσης του sentiment μπορεί να συμμετάσχει στα κέρδη ή τις ζημίες ενός επενδυτή, είτε ακόμα και σε ένα επίπεδο παραπάνω να αυξήσει τα προσδοκόμενα κέρδη ή προστατεύσει τον επενδυτή και να μετριάσει τις πιθανές απώλειες που θα είχε. Επίσης, εξετάζεται το αν μία είδηση υποβαλλόμενη σε μία τέτοια επεξεργασία μπορεί να λειτουργήσει συνθετικά και να βοηθήσει τα αποτελέσματα που εξάγονται από αριθμητικές μεθόδους

προβλέψεων και να μετριάσει το σφάλμα που προκύπτει από την τυχαιότητα των τιμών των μετοχών.

7.2 Δομή του Πειράματος

7.2.1 Περίληψη

Η διενέργεια του πειράματος έχει σκοπό την αποτύπωση του financial sentiment από δημοσιευμένες ειδήσεις στις τιμές και τις κινήσεις των μετοχών. Για το σκοπό αυτό επιλέχθηκαν τρεις μετοχικοί τίτλοι προς επεξεργασία. Στη συνέχεια καταγράφηκαν οι τιμές κλεισίματος του κάθε τίτλου για ένα χρονικό διάστημα τεσσάρων μηνών. Κατόπιν, αντλήθηκαν ειδήσεις που δημοσιεύθηκαν και διαβάστηκαν από ένα πολύ μεγάλο μέρος των επενδυτών στην παγκόσμια χρηματιστηριακή αγορά και αφορούν τις συγκεκριμένες εταιρίες. Στις τιμές κλεισίματος εφαρμόστηκαν μαθηματικές απλές μέθοδοι πρόβλεψης, αλλά και κριτική πρόβλεψη μέσω ανάλυσης των δημοσιευμένων νέων που συγκεντρώθηκαν. Το πείραμα μετατρέπει τον ερευνητή σε επενδυτή με κεφάλαιο της τάξεως των $10,000 και του ζητά να το επενδύσει βασισμένος στις παραπάνω προβλέψεις.

7.2.2 Χρονικό Διάστημα Συλλογής Πληροφοριών

Το χρονικό διάστημα συλλογής των πληροφοριών επιλέχθηκε να είναι τέσσερεις μήνες και συγκεκριμένα από 4 Φεβρουαρίου 2015 έως 1 Ιουνίου 2015. Η επιλογή αυτή στηρίχθηκε στο γεγονός ότι το διάστημα οφείλει να είναι αρκετό για να υπάρχουν συνολικά 239 τιμές κλεισίματος για τις τρεις μετοχές, άρα οι αριθμητικές μέθοδοι να έχουν την αξιοπιστία που απαιτείται, αλλά και να βρίσκονται χρονικά κοντά στις ημέρες διεξαγωγής του πειράματος ώστε το σύνολο των ειδήσεων που μπορούν να αντληθούν να βρεθούν εύκολα.

7.2.3 Επιλογή των Μετοχικών Τίτλων

Για την αντιπροσωπευτικότητα του πειράματος επιλέχθηκαν μετοχές που να συγκεντρώνουν τα εξής σημαντικά χαρακτηριστικά:

Α. Να μην επηρεάζονται από την ελληνική κρίση την περίοδο που επιλέχθηκε. Συνεπώς οι εταιρίες που επιλέχθηκαν έπρεπε να είναι ενταγμένες σε χρηματιστήριο εκτός Ευρώπης αλλά και να μην είναι ευρωπαϊκές.

Β. Να ανήκουν σε επιχειρήσεις των οποίων τα προϊόντα να μην επηρεάζονται από την ελληνική κρίση έτσι ώστε οι τιμές των μετοχών τους να μην κινούνται ανάλογα με την έκβαση της ελληνικής κρίσης.

Γ. Οι ειδήσεις που αφορούν τις συγκεκριμένες μετοχές να είναι στην συντριπτική τους πλειονότητα στα αγγλικά διότι το σύστημα ανίχνευσης του sentiment χρησιμοποιεί την αγγλική γλώσσα.

Δ. Να έχουν υψηλή κεφαλαιοποίηση, τζίρο και να αξιολογούνται με καλή πιστοληπτική ικανότητα από τους διεθνής οίκους ώστε να επιδεικνύουν μία σχετική σταθερότητα.

Ε. Τα δημοσιευμένα νέα που υπάρχουν στην αγορά και κυρίως στο Διαδίκτυο να είναι εξειδικευμένα για τις συγκεκριμένες μετοχές, ώστε να μη γίνεται παρανόηση ως προς την ανάλυση του sentiment.

ΣΤ. Να μπορούν να συλλεχθούν εύκολα οι ειδήσεις.

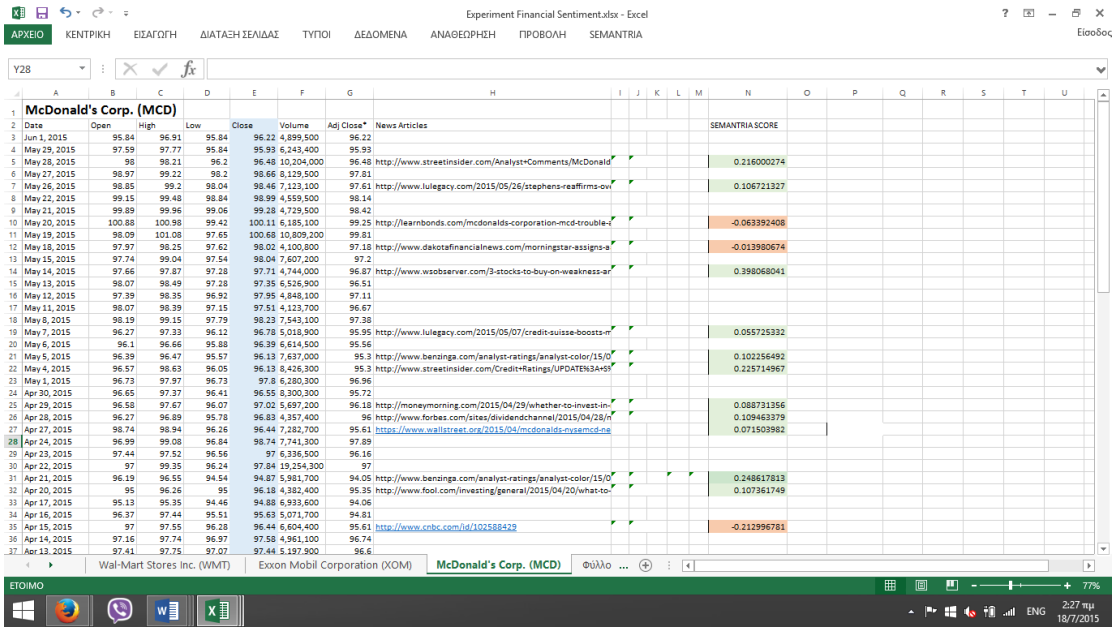
Οι επιχειρήσεις των οποίων τις μετοχές επιλέχθηκαν για το παρών πείραμα είναι η **McDonald's Corp. (MCD),** η **Exxon Mobil Corporation (XOM)** και η **Wal-Mart Stores Inc. (WMT).** Όλες εδρεύουν στην Αμερική και οι μετοχικοί τίτλοι τους διαπραγματεύονται στο χρηματιστήριο της Νέας Υόρκης και βρίσκονται στον δείκτη S&P 500. Τα παραπάνω κριτήρια πληρούνται και στις τρεις.

7.2.4 Άντληση Πληροφοριών - Ειδήσεων

Οι επενδυτές και τα τραπεζικά-χρηματιστηριακά ιδρύματα ανά τον κόσμο χρησιμοποιούν την πλατφόρμα της Bloomberg για να αντλήσουν κάθε είδους πληροφορία για τα χρεόγραφα που διαπραγματεύονται σε όλον τον πλανήτη, όπως

αναφέρθηκε και σε προηγούμενο κεφάλαιο. Υπάρχει στη συγκεκριμένη πλατφόρμα ειδική σελίδα με την παροχή της ειδησεογραφίας που αφορά την εκάστοτε μετοχή. Το newsfeed της συγκεκριμένης πλατφόρμας χρησιμοποιήθηκε για την άντληση των υπό επεξεργασία ειδήσεων. Για κάθε μετοχή επιλέχθηκε διαφορετικό πλήθος ειδήσεων ώστε να εξεταστεί το αν επηρεάζει και το πλήθος των δημοσιευμένων νέων στην ανάλυση των τιμών των μετοχών. Συγκεκριμένα για την WMT έγινε η επιλογή 1 είδησης ανά εβδομάδα, για την XOM 1,5 ειδήσεις, ενώ για την MCD 2 ειδήσεις ανά εβδομάδα. Η επιλογή έγινε τυχαία με μοναδικό κριτήριο να αφορά η είδηση τη συγκεκριμένη μετοχή και να είναι συγχρόνως δημοσιευμένη στο Διαδίκτυο ως ανεξάρτητο άρθρο. Επειδή δεν υπήρχε στη διάθεση του πειράματος ένα άρτιο σύστημα φιλτραρίσματος του περιεχομένου το οποίο να είχε την ικανότητα να απορρίπτει άρθρα ή ειδήσεις με εξαιρετικά πολύ θόρυβο ή νέα τα οποία δε θα είχαν επιρροή σε έναν επενδυτή για να λάβει κάποια απόφαση, διασφαλίστηκε εμπειρικά ότι όλα τα ειδησεογραφικά δεδομένα είχαν τη δυνατότητα να προκαλέσουν συναισθήματα στον αναγνώστη τέτοια ώστε να μπορούν να παράξουν κριτική πρόβλεψη.

Στον πίνακα 7.1 φαίνεται ένα μέρος των δεδομένων της μετοχής της MCD όπως ταξινομήθηκαν. Διακρίνονται οι ημερομηνίες, οι διάφορες τιμές της μετοχής, ο τζίρος που πραγματοποιήθηκε ανά ημέρα καθώς και οι διαδικτυακές διευθύνσεις των ειδήσεων που χρησιμοποιήθηκαν για την εξαγωγή του financial sentiment.



McDonald's Corp. (MCD)								
Date	Open	High	Low	Close	Volume	Adj Close*	News Articles	SEMANTRIA SCORE
Jun 1, 2015	95.84	96.91	95.84	96.22	4,899,500	96.22		
May 29, 2015	97.59	97.77	95.84	95.93	6,243,400	95.93		
May 28, 2015	98	98.21	96.2	96.48	10,204,000	96.48	http://www.streetinsider.com/Analyst+Comments/McDonald	0.216000274
May 27, 2015	98.97	99.22	98.2	98.66	8,129,500	97.81		
May 26, 2015	98.85	99.2	98.04	98.46	7,123,100	97.61	http://www.lulegacy.com/2015/05/26/stephens-reaffirms-ov	0.106721327
May 22, 2015	99.15	99.48	98.84	98.99	4,559,500	98.14		
May 21, 2015	99.89	99.96	99.06	99.28	4,729,500	98.42		
May 20, 2015	100.88	100.98	99.42	100.11	6,185,100	99.25	http://learnbonds.com/mcdonalds-corporation-mcd-trouble-	-0.063392408
May 19, 2015	98.09	101.08	97.65	100.68	10,809,200	99.81		
May 18, 2015	97.97	98.25	97.62	98.02	4,100,800	97.18	http://www.dakotafinancialnews.com/morningstar-assigns-a	-0.013980674
May 15, 2015	97.74	99.04	97.54	98.04	7,607,200	97.2		
May 14, 2015	97.66	97.87	97.28	97.71	4,744,000	96.87	http://www.wsobserver.com/3-stocks-to-buy-on-weakness-ar	0.398068041
May 13, 2015	98.07	98.49	97.28	97.35	6,526,900	96.51		
May 12, 2015	97.39	98.35	96.92	97.95	4,848,100	97.11		
May 11, 2015	98.07	98.39	97.15	97.51	4,123,700	96.67		
May 8, 2015	98.19	99.15	97.79	98.23	7,543,100	97.38		
May 7, 2015	96.27	97.33	96.12	96.78	5,018,900	95.95	http://www.lulegacy.com/2015/05/07/credit-suisse-boosts-m	0.055725332
May 6, 2015	96.1	96.66	95.88	96.39	6,614,500	95.56		
May 5, 2015	96.39	96.47	95.57	96.13	7,637,000	95.3	http://www.benzinga.com/analyst-ratings/analyst-color/15/0	0.102256492
May 4, 2015	96.57	98.63	96.05	96.13	8,426,300	95.3	http://www.streetinsider.com/Credit+Ratings/UPDATE%3A+St	0.225714967
May 1, 2015	96.73	97.97	96.73	97.8	6,280,300	96.96		
Apr 30, 2015	96.65	97.37	96.41	96.55	8,300,300	95.72		
Apr 29, 2015	96.58	97.67	96.07	97.02	5,697,200	96.18	http://moneymorning.com/2015/04/29/whether-to-invest-in-	0.088731356
Apr 28, 2015	96.27	96.89	95.78	96.83	4,357,400	96	http://www.forbes.com/sites/dividendchannel/2015/04/28/n	0.109463379
Apr 27, 2015	98.74	98.94	96.26	96.44	7,282,700	95.61	https://www.wallstreet.org/2015/04/mcdonalds-nysemcd-ne	0.071503982
Apr 24, 2015	96.99	99.08	96.84	98.74	7,741,300	97.89		
Apr 23, 2015	97.44	97.52	96.56	97	6,336,500	96.16		
Apr 22, 2015	97	99.35	96.24	97.84	19,254,300	97		
Apr 21, 2015	96.19	96.55	94.54	94.87	5,981,700	94.05	http://www.benzinga.com/analyst-ratings/analyst-color/15/0	0.148617813
Apr 20, 2015	95	96.26	95	96.18	4,382,400	95.35	http://www.fool.com/investing/general/2015/04/20/what-to	0.107361749
Apr 17, 2015	95.13	95.35	94.46	94.88	6,933,600	94.06		
Apr 16, 2015	96.37	97.44	95.51	95.63	5,071,700	94.81		
Apr 15, 2015	97	97.55	96.28	96.44	6,604,400	95.61	http://www.cnbc.com/id/102588429	-0.212996781
Apr 14, 2015	97.16	97.74	96.97	97.58	4,961,100	96.74		
Apr 13, 2015	97.41	97.75	97.07	97.44	5,197,900	96.6		

Πίνακας 7.1: Πίνακας Δεδομένων Πειραματικής Διάταξης

7.3 Sentiment Analysis

Το σύστημα που χρησιμοποιήθηκε ονομάζεται SEMANTRIA (https://semantria.com/). Η εταιρία LEXALYTICS εδρεύει στη Βοστώνη των Ηνωμένων Πολιτειών, ειδικεύεται στο text και sentiment analysis και δημιούργησε το σύστημα SEMANTRIA το οποίο είναι ένα πρόγραμμα που βασίζεται στο EXCEL της Microsoft Corp. και χρησιμοποιείται ευρέως από τις επιχειρήσεις για την ανάλυση κειμένων στις σελίδες κοινωνικής δικτύωσης, σε έρευνες αγοράς αλλά και σε αξιολογήσεις προϊόντων. Η ανάλυση των κειμένων αυτών γίνεται σε επίπεδο λεξιλογικό (text) και συναισθηματικό (sentiment) ώστε να εξάγεται με ποσοτικοποιημένο τρόπο η φυσική γλώσσα που δίνεται ως ανάδραση (feedback) στις επιχειρήσεις για την αξιολόγηση των προϊόντων και των υπηρεσιών που παρέχουν. Η απουσία ενός ειδικού, εκπαιδευμένου συστήματος αξιολόγησης και ποσοτικοποίησης του FINANCIAL sentiment από τα δημοσιευμένα νέα, οδήγησε στην επιλογή αυτή καθώς το σύστημα αυτό αποτελεί ότι πιο ειδικευμένο και εγγυές στο υπό αξιολόγηση sentiment.

Το SEMANTRIA λαμβάνει το κείμενο της κάθε είδησης που εισάγουμε και εξάγει έναν αριθμό με όρια από το 1 έως το -1, τέτοιον ώστε στην περίπτωση που είναι θετικός το κείμενο που κλήθηκε να επεξεργαστεί αξιολογείται ως θετικό και άρα η ανάγνωσή του παροτρύνει τον αναγνώστη να επιλέξει τη μετοχή προς επένδυση, ενώ σε αντίθετη περίπτωση αν ο αριθμός είναι αρνητικός τότε αξιολογεί ότι η τιμή της συγκεκριμένης μετοχής θα μειωθεί και άρα συνιστά την πώλησή της.

Επειδή ακριβώς το σύστημα αυτό δεν είναι ειδικευμένο στην αξιολόγηση κειμένων της χρηματιστηριακής αγοράς, ενέχει και το ανάλογο σφάλμα.

Σε ένα λογιστικό φύλλο εισήχθησαν οι ημερομηνίες και οι τιμές κλεισίματος της κάθε μετοχής όπως φαίνεται στον πίνακα 7.1.

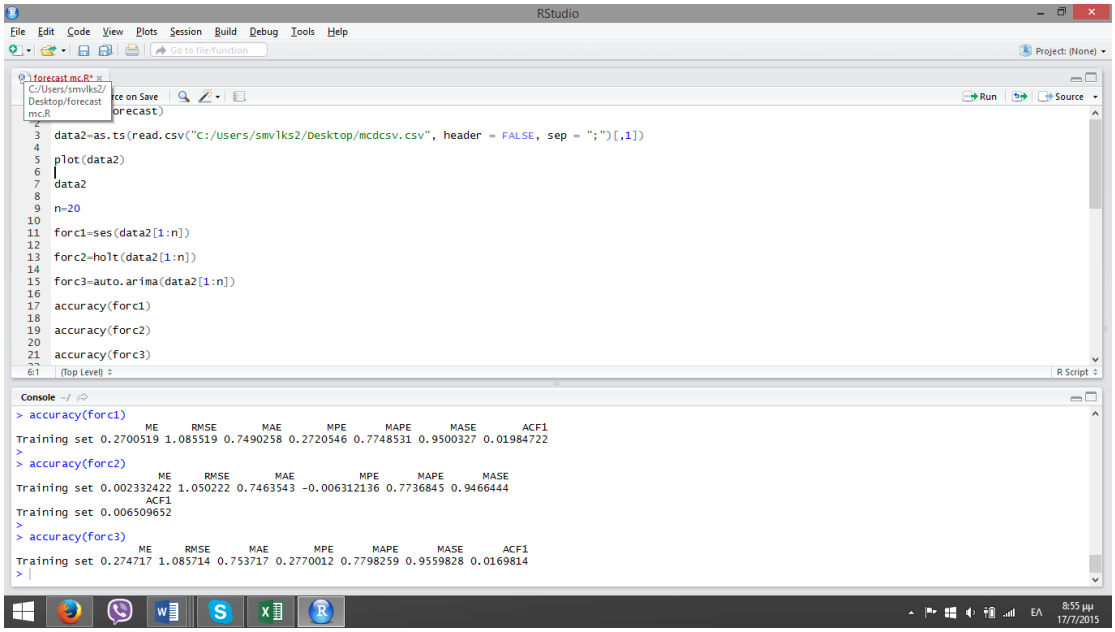
Κατόπιν αντιστοιχήθηκε η κάθε είδηση και η αξιολόγησή της, δηλαδή το νούμερο που προκύπτει από την ανάλυση του προγράμματος SEMANTRIA, για κάθε ημέρα, όπως και η ένδειξη buy ή sell με γνώμονα το αν ο αριθμός αυτός είναι θετικός ή αρνητικός αντίστοιχα.

7.4 Απλές Μέθοδοι Προβλέψεων

Το πρόγραμμα που χρησιμοποιήθηκε για την εφαρμογή των απλών μεθόδων προβλέψεων ονομάζεται RStudio το οποίο είναι ένα πρόγραμμα ελεύθερου λογισμικού με ενσωματωμένες τις τρεις μεθόδους προβλέψεων που θα χρησιμοποιήσαμε. Συγκεκριμένα την SES (Απλή Εκθετική Εξομάλυνση), την HOLT (εκθετική εξομάλυνση με προσαρμογή στην τάση) και την ARIMA (ολοκληρωμένα αυτοπαλίνδρομα υποδείγματα κινητού μέσου όρου).

Για την εφαρμογή των παραπάνω μεθόδων κρίθηκε αναγκαίο να διευρυνθεί το δείγμα των τιμών των κλεισιμάτων των μετοχών από την 3η Μαίου 2015 ως την 4η Φεβρουαρίου 2015 ώστε να υπάρχουν αρκετά δεδομένα, συγκεκριμένα είκοσι τιμές, ώστε να δημιουργηθεί και η πρώτη χρονοσειρά.

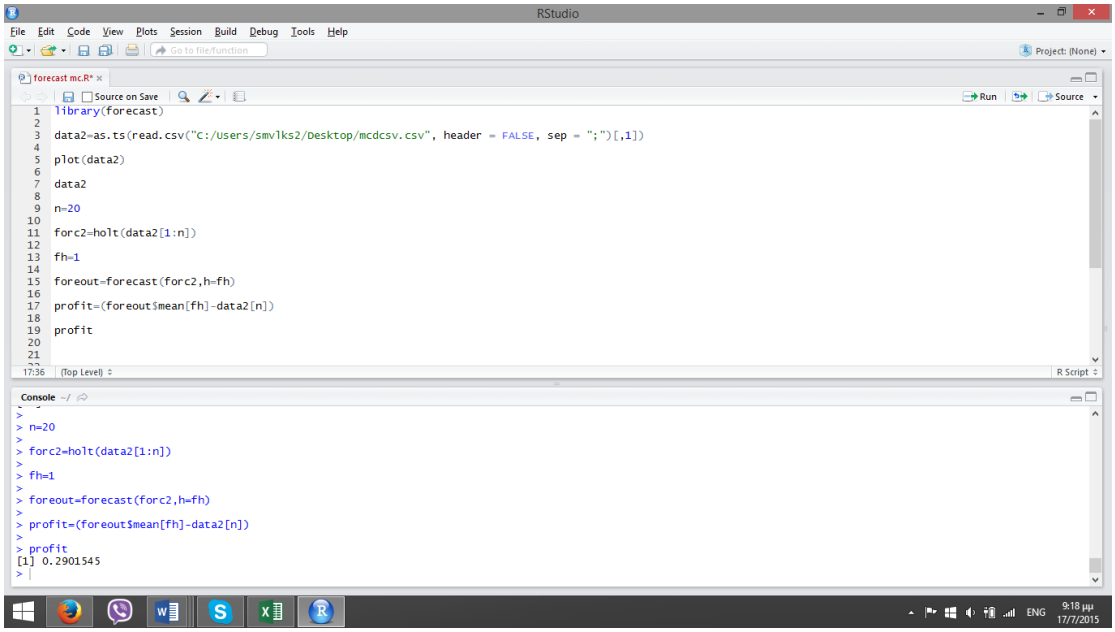
Δημιουργήθηκε για το σκοπό αυτό και με τη βοήθεια του προγράμματος ένας αλγόριθμος ο οποίος στην αρχή αντλούσε τα δεδομένα των τιμών των κλεισιμάτων και των τριών μετοχών και τις μετέτρεπε σε χρονοσειρές μήκους 20 η πρώτη και με μεταβλητό μέγεθος ύστερα, ανάλογα με το πότε προκύπτει είδηση. Για παράδειγμα, για τη μετοχή της MCD η πρώτη χρονοσειρά είχε μήκος 20, η δεύτερη 25, η τρίτη 31, η τέταρτη 31, η πέμπτη 37 και ούτω καθεξής. Στις 55 χρονοσειρές που δημιουργήθηκαν εφαρμόστηκαν και οι τρεις μέθοδοι πρόβλεψης που προαναφέρθηκαν και με τη χρήση της εντολής accuracy αποφανθήκαμε για το ποια μέθοδος εφαρμόζει καλύτερα στην κάθε χρονοσειρά. Η αξιολόγηση έγινε βάσει του από ποιο μοντέλο πρόβλεψης πρόκυπτε το μικρότερο MAPE (Μέσο Απόλυτο Ποσοστιαίο Σφάλμα).



Εικόνα 7.2: Υπολογισμός accuracy της πρώτης χρονοσειράς για την MCD

Στην εικόνα 7.2 παρουσιάζεται ο αλγόριθμος στον οποίο εισήχθησαν τα δεδομένα της πρώτης χρονοσειράς με μήκος n=20 για τον μετοχικό τίτλο της MCD. Ο αλγόρυθμος εφήρμοσε στη συνέχεια τις τρεις μεθόδους των προβλέψεων που ήδη υπάρχουν ως συναρτήσεις στην βιβλιοθήκη «forecast» και τις αποθήκευσε σε τρεις διαφορετικές μεταβλητές forc1, forc2 και forc3. Με την εντολή «accuracy» υπολογίστηκε ξεχωριστά το σφάλμα των τριών μεθόδων και παρουσιάστηκε στην κονσόλα. Στην περίπτωση της πρώτης χρονοσειράς με μήκος n=20 η μέθοδος πρόβλεψης που επιλέγεται είναι η Holt (Μοντέλο Γραμμικής Τάσης) διότι το μέσο απόλυτο ποσοστιαίο σφάλμα είναι μικρότερο.

Για την εξαγωγή της πρόβλεψης κέρδους η ζημίας χρησιμοποιήθηκε στην κάθε χρονοσειρά το κατάλληλο μοντέλο πρόβλεψης και με ορίζοντα την επόμενη ημέρα. Υπολογίστηκε με αυτόν τον τρόπο το ενδεχόμενο κέρδος ή ζημία που θα προέκυπτε την επομένη, από την αφαίρεση της πρόβλεψης από την τελευταία τιμή κλεισίματος. Αν το αποτέλεσμα ήταν θετικό τότε η πρόβλεψη καταδείκνυε κέρδος και σε αντίθετη περίπτωση αν το αποτέλεσμα ήταν αρνητικό καταδείκνυε ζημία.



Εικόνα 7.3: Υπολογισμός κέρδους για την πρώτη χρονοσειρά της MCD με τη μέθοδο Holt

Στην εικόνα 7.3 παρουσιάζεται ο αλγόριθμος στον οποίο εισήχθησαν τα δεδομένα της πρώτης χρονοσειράς με μήκος n=20 για τον μετοχικό τίτλο της MCD. Ο αλγόρυθμος εφαρμόζει στη χρονοσειρά τη μέθοδο πρόβλεψης Holt διότι ήταν αυτή με το μικρότερο μέσο απόλυτο ποσοστιαίο σφάλμα για αυτήν. Υπολογίζεται με την εντολή «forecast» και ορίζοντα 1, μέσω της μεταβλητής fh=1, την τιμή που προβλέπει η μέθοδος ότι θα πάρει η μετοχή την επόμενη ημέρα. Στη συνέχεια η τιμή αυτή συγκρίνεται με την τιμή της μετοχής της προηγούμενης ημέρας, δηλαδή την τελευταία τιμή της υπό επεξεργασίας χρονοσειράς και παρουσιάζεται στην κονσόλα. Στην περίπτωση αυτή η τιμή είναι θετική συνεπώς την επόμενη ημέρα προσδοκάται άνοδο της τιμής της μετοχής.

Στο ίδιο λογιστικό φύλλο καταγράφτηκε για κάθε ημέρα είδησης ανάλογα με το ποσό που προβλεπόταν ως κέδρος ή ζημία η ένδειξη buy ή sell για θετικά και αρνητικά αποτελέσματα όπως δείχνει η εικόνα 7.3.

7.5 Υβριδικό Σύστημα

Η τρίτη μέθοδος που αποφασίστηκε να ακολουθηθεί είναι ο συνδυασμός της κριτικής πρόβλεψης που προέκυπτε από την ανάλυση των δημοσιευμένων ειδήσεων και των απλών μεθόδων πρόβλεψης.

Επειδή σε πολλές περιπτώσεις παρατηρήθηκε ότι τα αποτελέσματα των παραπάνω προβλέψεων ήταν αντικρουόμενα αποφασίστηκε ότι οφείλαμε να εισάγουμε πέραν των επιλογών buy και sell, την επιλογή hold που σημαίνει ότι προτιμάται να μη γίνει κάποια επενδυτική κίνηση.

Δημιουργήθηκε μία κανονικοποιημένη κλίμακα βαθμονόμησης του αποτελέσματος της αξιολόγησης των δύο παραπάνω μεθόδων. Η κλίμακα αυτή λάμβανε τιμές από το -3 έως το 3.

Στην περίπτωση της πρόβλεψης μέσω του sentiment analysis, αφού οι τιμές που εμφάνιζε το σύστημα SEMANTRIA κυμαίνονταν από -1 έως 1 η κλίμακα βαθμονόμησης είχε ως εξής:

Αν η τιμή βρισκόταν στο διάστημα μεταξύ	λάμβανε η πρόβλεψη την τιμή
[-1 , -0,72)	-3
[-0,72 , -0,43)	-2
[-0,43 , -0,14)	-1
[-0,14 , 0,14]	0
(0,14 , 0,43]	1
(0,43 , 0,72]	2
(0,72 , 1]	3

Πίνακας 7.4: Πίνακας βαθμονόμησης του financial sentiment για το SEMANTRIA

Στην περίπτωση της πρόβλεψης μέσω των απλών μεθόδων προβλέψεων οι τιμές που εμφάνιζε η κάθε χρονοσειρά κανονικοποιήθηκε με διαφορετικά διαστήματα για κάθε μετοχή αφού τα αποτελέσματα ήταν διαφορετικά. Συγκεκριμένα:

MCD

Αν η τιμή βρισκόταν στο διάστημα μεταξύ	λάμβανε η πρόβλεψη την τιμή
[-0.5 , -0,36)	-3
[-0,36 , -0,21)	-2
[-0,21 , -0,07)	-1
[-0,07 , 0,07]	0
(0,07 , 0,14]	1

(0,14 , 0,36]	2
(0,36 , 0.5]	3

Πίνακας 7.5: Πίνακας βαθμονόμησης των μαθηματικών αποτελεσμάτων πρόβλεψης για την MCD

XOM

Αν η τιμή βρισκόταν στο διάστημα μεταξύ	λάμβανε η πρόβλεψη την τιμή
[-0.42 , -0,715)	-3
[-0,715 , -0,429)	-2
[-0,429 , -0,143)	-1
[-0,143 , 0,143]	0
(0,143 , 0,429]	1
(0,429 , 0,715]	2
(0,715 , -0.42]	3

Πίνακας 7.6: Πίνακας βαθμονόμησης των μαθηματικών αποτελεσμάτων πρόβλεψης για την XOM

WMT

Αν η τιμή βρισκόταν στο διάστημα μεταξύ	λάμβανε η πρόβλεψη την τιμή
[-0.96 , -0,69)	-3
[-0,69 , -0,41)	-2
[-0,41 , -0,14)	-1
[-0,14 , 0,14]	0
(0,14 , 0,41]	1
(0,41 , 0,69]	2
(0,69 , 0.96]	3

Πίνακας 7.7: Πίνακας βαθμονόμησης των μαθηματικών αποτελεσμάτων πρόβλεψης για την WMT

Στο ίδιο λογιστικό φύλλο εισήχθη ο μέσος όρος των δύο τιμών ο οποίος κυμαινόταν από -3 έως 3 και δίπλα η ένδειξη buy αν το αποτέλεσμα προέκυπτε θετικό, sell αν το

αποτέλεσμα ήταν αρνητικό και hold αν το αποτέλεσμα ήταν 0, δηλαδή πλήρως αντικρουόμενο, όπως φαίνεται στον Πίνακα 7.8.

Όπως φαίνεται στον Πίνακα 7.8 ανάλογα με τα αποτελέσματα των τριών παραπάνω μεθόδων ο επενδυτής οδηγήθηκε και στις κατάλληλες αποφάσεις.

7.6 Μετρήσεις

Το πείραμα υποδεικνύει ότι στην περίοδο αυτή και για τις ημέρες για τις οποίες υπάρχει ειδησεογραφία οφείλουμε ως επενδυτές να επενδύσουμε το ποσό των $10,000 για την αγοραπωλησία των τριών προς εξέταση μετοχών λαμβάνοντας υπόψιν τις υποδείξεις buy, sell ή hold των τριών αναλύσεων, sentiment analysis, math analysis και hybrid analysis.

Τα αποτελέσματα των αναλύσεων σε πολλές περιπτώσεις δεν εναλλάσονται και παρατηρούνται σε πολλές περιπτώσεις ότι σε συνεχόμενες μέρες ειδησεογραφίας υπάρχει ίδια η εντολή, είτε buy είτε sell. Η παραδοχή που έχει γίνει είναι ότι προφανώς δεν επανεπενδύεται με κάποιο τρόπο το κεφάλαιο αλλά στην περίπτωση που υπάρχει ίδια εντολή την ημέρα της ειδησεογραφίας με την προηγούμενη, τότε αυτή αυτόματα μεταφράζεται ως hold, δηλαδή δεν πραγματοποιείται κάποια κίνηση.

Όπως φαίνεται καθαρά στον Πίνακα 7.8 που αφορά την μετοχή της MCD το πείραμα χωρίστηκε σε τρείς ξεχωριστές φάσεις.

Στην πρώτη ο επενδυτής λαμβάνει αποφάσεις επηρεασμένος μόνο από τα αποτελέσματα της εφαρμογής των μαθηματικών μοντέλων πρόβλεψης. Συνεπώς, την 3^{η} Μαρτίου 2015, επενδύει σε πρώτη φάσητο κεφάλαιό του και αποκτά 100 μετοχές της εταιρίας όπως φαίνεται και στη στήλη της εν λόγω απόφασης, ενώ το υπόλοιπο του κεφαλαίου προσδιορίζεται στα $26 όπως φαίνεται στη στήλη του Κεφαλαίου. Η επόμενη στιγμή που πρέπει να λάβει μια απόφαση λαμβάνει χώρα την 10^{η} Μαρτίου 2015. Το αποτέλεσμα των απλών μεθόδων πρόβλεψης προτρέπει τον επενδυτή να αγοράσει. Επειδή όμως έχει ήδη επενδύσει το κεφάλαιό του, η εντολή αυτή μετατρέπεται σε hold όπως εξηγείται παραπάνω, με αποτέλεσμα να διατηρείται το χαρτοφυλάκιό του ίδιο. Την αμέσως επόμενη ειδησεογραφική ημέρα (18 Μαρτίου 2015) η μαθηματική μέθοδος έχει βγάλει ως αποτέλεσμα αρνητικό αριθμό και έτσι η απόφαση του επενδυτή μετατρέπεται σε sell. Αποτέλεσμα είναι να ρευστοποιήσει τις

μετοχές που κατέχει και να έχει έτσι ρευστά διαθέσιμα $9.726,00. Η διαδικασία αυτή επαναλαμβάνεται μέχρι την 1η Ιουνίου 2015, η οποία είναι η καταληκτική ημερομηνία, στην οποία ο επενδυτής πουλάει όλο το χαρτοφυλάκιο που διαθέτει αποκομίζοντας $9.527,44.

Στη δεύτερη φάση ο επενδυτής ακολουθεί την ίδια διαδικασία, μόνο που αυτή τη φορά τα ερεθίσματα της απόφασής του, λαμβάνονται από την αξιολόγηση της επιλεγμένης ειδησεογραφίας, για κάθε ημέρα. Κατά αυτόν τον τρόπο σε διαφορετικές ημέρες επιλέγει να είτε να επενδύσει είτε να ρευστοποιεί το χαρτοφυλάκιό του όπως καταδεικνύεται στις κατάλληλες στήλες του Πίνακα 8.7 για το Financial Sentiment για τον μετοχικό τίτλο της MCD. Η διαδικασία αυτή καταλήγει επίσης την 1η Ιουνίου 2015 και καταγράφεται ότι η αξία των μετοχών του είναι $9.709,45.

Στην τρίτη και τελευταία φάση ο επενδυτής ακολουθεί παρόμοια διαδικασία με τη διαφορά ότι η απόφασή του εξαρτάται τόσο από το μαθηματικό μοντέλο όσο και από την ανάλυση του Financial Sentiment. Όταν δε, οι δύο προβλέψεις είναι απόλυτα αντικρουόμενες ή προβλέπουν μικρές μεταβολές τότε η επιλογή που προτείνεται στον επενδυτή είναι να μην πραγματοποιήσει κάποια κίνηση, είτε αγοράζοντας είτε ρευστοποιώντας. Με άλλα λόγια η ανάλυση του δίνει την εντολή hold. Το τέλος της διαδικασίας την 1η Ιουνίου για την μετοχή της MCD όπως φαίνεται στον Πίνακα 7.8 το κεφάλαιό μας έχει αξία $9.575,24.

Μετά τις τρεις βασικές φάσεις υπολογίζεται η αξία του χαρτοφυλακίου που θα είχε ο επενδυτής αν αγόραζε μετοχές την πρώτη ημέρα και δίχως να προβεί σε κινήσεις και να πάρει αποφάσεις απλά ρευστοποιούσε την τελευταία.

Το πείραμα ολοκληρώνεται με τον υπολογισμό της ποσοστιαία μεταβολή του υπενδεδυμένου ποσού για τις τέσσερεις περιπτώσεις.

7.7 Αποτελέσματα

Από την επένδυση του κεφαλαίου των $10,000 την 1η Ιουνίου 2015 προκύψανε τα εξής αποτελέσματα για τις τρείς μετοχές:

$	ΜΑΘΗΜΑΤΙΚΗ	SENTIMENT	ΥΒΡΙΔΙΚΗ	ΧΩΡΙΣ ΚΙΝΗΣΕΙΣ
MCD	9527.44	**9709.45**	9575.24	9648.00
XOM	9760.32	**10355.98**	9944.14	10132.09
WMT	**9748.45**	9221.72	9608.71	9050.15

Πίνακας 7.9: Παρουσίαση Συνολικών Αποτελεσμάτων Επενδυτικής Διαδικασίας

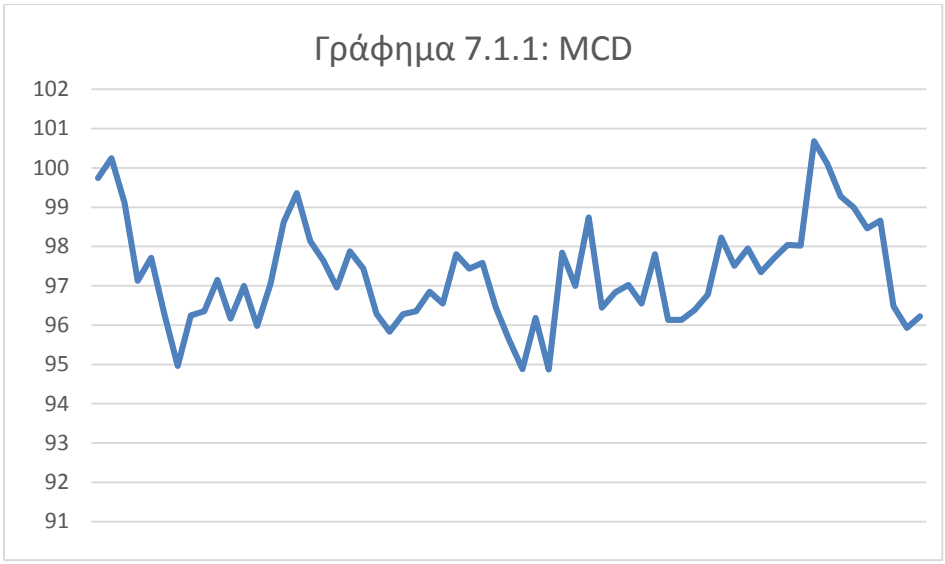
MCD:

Στο γράφημα 7.1.1 παρατηρείται η περίοδος που γινόταν η επένδυση του ποσού των $10,000 για τη μετοχή της MCD. Όπως φαίνεται και στον πίνακα 7.11 σε περίπτωση που ο επενδυτής αγόραζε στην αρχική ημερομηνία και πούλαγε τις μετοχές που απέκτησε την τελευταία ημέρα δίχως να κάνει κάποια άλλη επενδυτική κίνηση, η απώλειά του υπολογίζεται στο 3,52%.

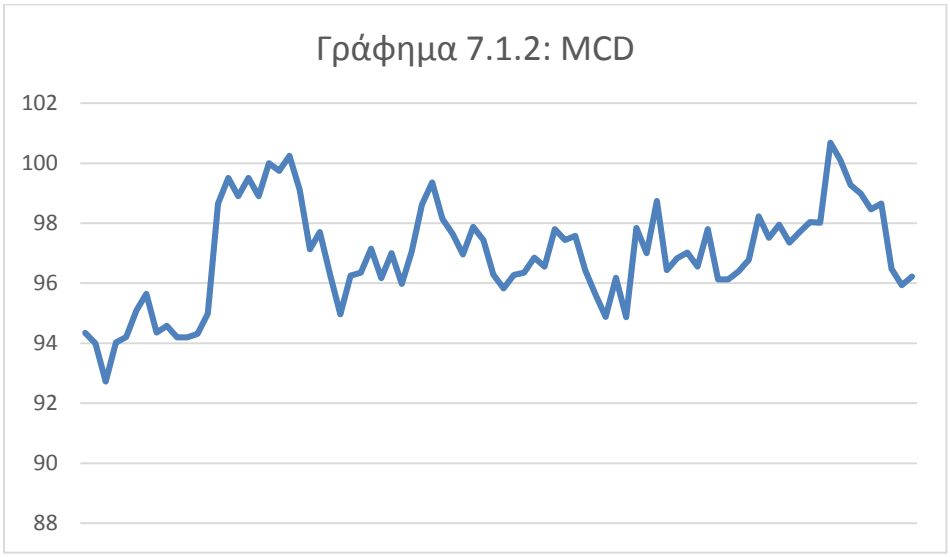
Στο γράφημα 7.1.2 παρατηρείται η περίοδος που αναλύθηκε μέσω των απλών μεθόδων πρόβλεψης. Αυτή διακρίνεται από ένα μεγάλο volatility (μεταβλητότητα). Αυτός είναι και ο λόγος για τον οποίο στα αποτελέσματα η αυστηρά μαθηματική μέθοδος αποδείχθηκε περισσότερο επιβλαβής για τον επενδυτή. Η απώλειά του υπολογίζεται στο 4,73%.

Το παραπάνω γεγονός διατηρεί σε χαμηλά αν και σε καλύτερα επίπεδα την υβριδική μέθοδο με αποτέλεσμα η απώλεια να υπολογίζεται στο 4,25%.

Η εφαρμογή της μεθόδου της ανάλυσης του financial sentiment αποδεικνύεται ότι περιορίζει κατά πολύ τις απώλειες που θα είχε ένας επενδυτής και καταγράφει ζημίες μόλις 2,91%



Γράφημα 7.1.1: Χρονοσειρά MCD



Γράφημα 7.1.2: Διευρυμένη Χρονοσειρά MCD

XOM:

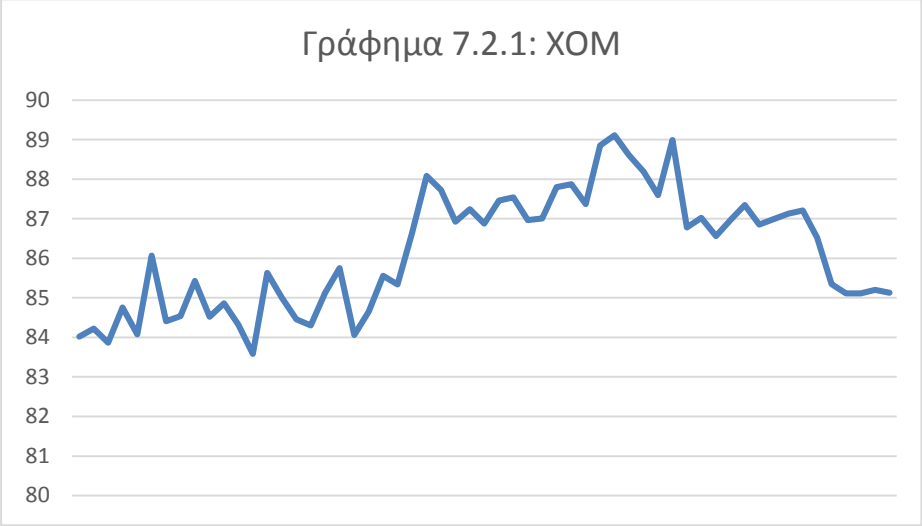
Στο Γράφημα 7.2.1 παρατηρείται η περίοδος που γινόταν η επένδυση του ποσού των $10,000 για τη μετοχή της XOM. Όπως φαίνεται και στον πίνακα 7.11 σε περίπτωση που ο επενδυτής αγόραζε στην αρχική ημερομηνία και πούλαγε τις μετοχές που απέκτησε την τελευταία ημέρα δίχως να κάνει κάποια άλλη επενδυτική κίνηση, το κέρδος του υπολογίζεται στο 1,32%.

Στο Γράφημα 7.2.2 παρατηρείται η περίοδος που αναλύθηκε μέσω των απλών μεθόδων πρόβλεψης. Αυτή διακρίνεται επίσης από ένα μεγάλο volatility (μεταβλητότητα). Αυτός είναι και ο λόγος για τον οποίο στα αποτελέσματα η αυστηρά μαθηματική

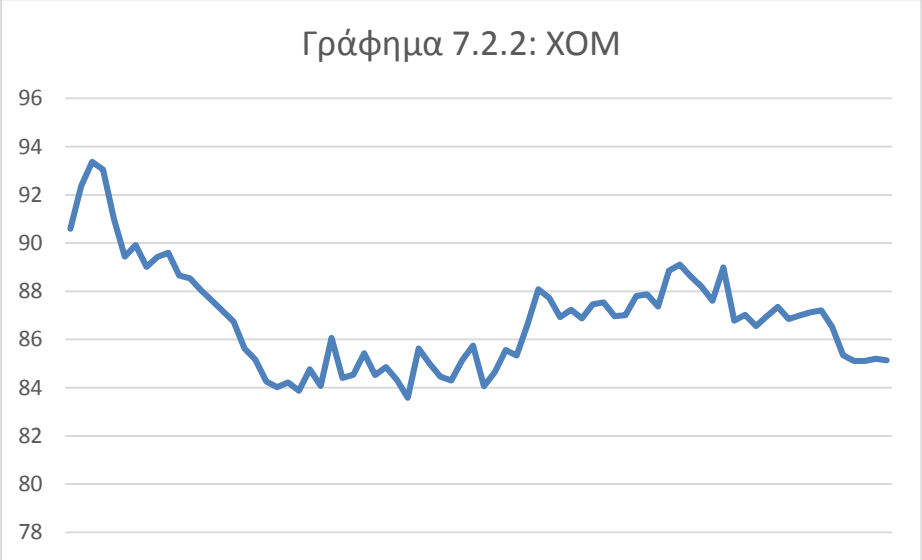
μέθοδος αποδείχθηκε επιβλαβής για τον επενδυτή παρόλο το γεγονός ότι θα είχε κέρδος. Η απώλειά του υπολογίζεται στο 2,40%.

Το παραπάνω γεγονός επηρεάζει αρνητικά, αν και βρίσκεται σε καλύτερα επίπεδα, την υβριδική μέθοδο με αποτέλεσμα η απώλεια να περιορίζεται στο 0,56%.

Η εφαρμογή της μεθόδου της ανάλυσης του financial sentiment αποδεικνύεται καλύτερη όχι μόνο διότι δεν καταγράφει απώλειες αλλά διότι αυξάνει το κέρδος που θα είχε ο επενδυτής. Με αυτόν τον τρόπο καταγράφονται κέρδη της τάξεως του 3,56%.



Γράφημα 7.2.1: Χρονοσειρά XOM



Γράφημα 7.3.2: Διευρυμένη Χρονοσειρά XOM

WMT:

Στο Γράφημα 7.3.1 παρατηρείται η περίοδος που γινόταν η επένδυση του ποσού των $10,000 για την μετοχή της WMT. Όπως φαίνεται και στον πίνακα 7.11 σε περίπτωση

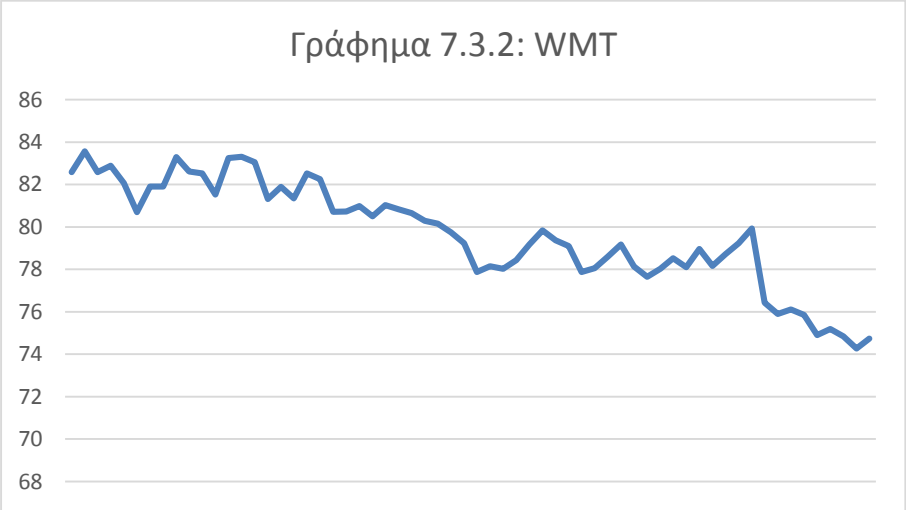
που ο επενδυτής αγόραζε στην αρχική ημερομηνία και πούλαγε τις μετοχές που απέκτησε την τελευταία ημέρα δίχως να κάνει κάποια άλλη επενδυτική κίνηση, η ζημία του υπολογίζεται στο 9,50%.

Στο Γράφημα 7.1.2 παρατηρείται η περίοδος που αναλύθηκε μέσω των απλών μεθόδων πρόβλεψης. Αυτή διακρίνεται επίσης από σαφή πτωτική τάση. Αυτός είναι και ο λόγος για τον οποίο στα αποτελέσματα η αυστηρά μαθηματική μέθοδος περιορίζει σε μεγάλο βαθμό τις απώλειες του επενδυτή. Η απώλειά του υπολογίζεται μόλις 2,52%.

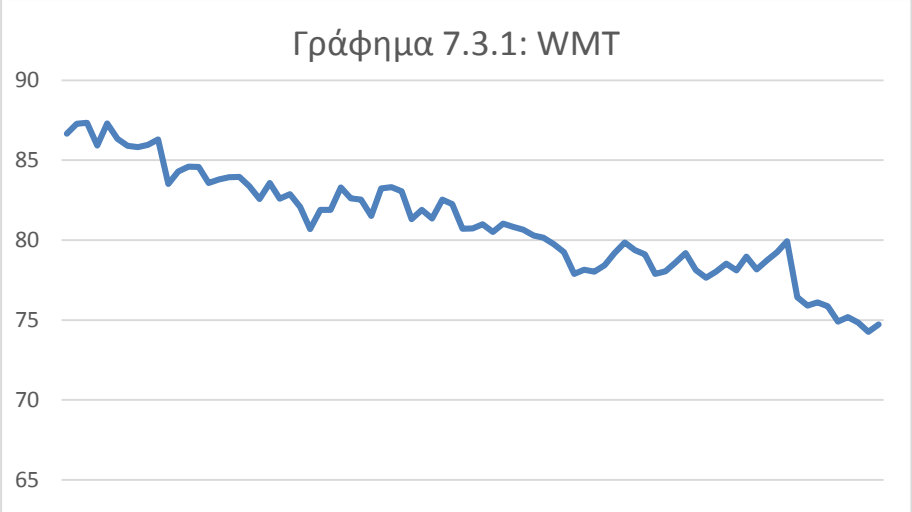
Η εφαρμογή της μεθόδου ανάλυσης του financial sentiment αποδεικνύεται ναι μεν χειρότερη από την εφαρμογή των απλών μεθόδων πρόβλεψης αλλά καλύτερη από μία δίχως ανάλυση επένδυση. Με αυτόν τον τρόπο περιορίστηκε η ζημία και καταγράφηκαν απώλειες της τάξεως του 7,78%.

Το παραπάνω γεγονός επηρεάζει αρνητικά, αν και βρίσκεται σε πολύ καλύτερα επίπεδα, την υβριδική μέθοδο με αποτέλεσμα η απώλεια να περιορίζεται στο 3,91%.

Αξίζει να σημειωθεί ότι για την μετοχή της WMT χρησιμοποιήθηκαν τα λιγότερα ειδησεογραφικά δεδομένα.



Γράφημα 7.3.1: Χρονοσειρά WMT



Γράφημα 7.3.2: Διευρυμένη Χρονοσειρά WMT

7.8 Συμπεράσματα Πειράματος

Οι παίκτες στην χρηματοοικονομική αγορά δεν έχουν σκοπό μόνο την εξασφάλιση του κέρδους των επενδύσεών τους αλλά και τον περιορισμό των απωλειών τους σε περίπτωση που τα χρεόγραφα τα οποία έχουν εντάξει στο χαρτοφυλάκιό τους καταγράψουν ζημίες. Συνεπώς, οι μέθοδοι πρόβλεψης δεν έχουν στόχο μόνο την μεγιστοποίηση του κέρδους αλλά και την κάλυψη από ενδεχόμενες απώλειες.

Όταν η μετοχικοί τίτλοι διακρίνονται από μεγάλη μεταβλητότητα, οι απλές μαθηματικές μέθοδοι προβλέψεων δυσκολεύονται να εξάγουν ασφαλή συμπεράσματα. Στην περίπτωση που καταγράφεται σαφής ανοδική ή καθοδική τάση, τα μαθηματικά μοντέλα αποδεικνύονται χρήσιμα και αξιόπιστα. Δυστυχώς όμως, δεν γνωρίζει κανείς με βεβαιότητα το σημείο στο οποίο αυτή η τάση ανατρέπεται και ούτε μπορεί εύκολα να την προβλέψει. Το συμπέρασμα αυτό καταγράφτηκε στο γεγονός ότι η μία εκ των τριών μετοχών είχε σαφή πτωτική τάση και το μαθηματικό μοντέλο περιόρισε σημαντικά τις απώλειες. Στις άλλες δύο περιπτώσεις οι τάσεις εναλλάσσονταν και για αυτόν τον λόγο ένα προσδοκώμενο κέρδος μετατράπηκε σε ζημία, ενώ μία πτώση δεν έγινε δυνατό να αποσοβηθεί σε ικανοποιητικό βαθμό.

Η εφαρμογή πρόβλεψης της κίνησης των μετοχών μέσω την ανάλυσης και ποσοτικοποίησης του financial sentiment στο πείραμα παρατηρήθηκε ότι, σε όλες τις περιπτώσεις, επέτυχε το σκοπό της. Στην περίπτωση πτώσης της τιμής των χρεογράφων η απώλειες περιορίστηκαν αλλά και σε περίοδο ανόδου τα κέρδη ήταν μεγαλύτερα από τα προσδοκώμενα. Στο παρόν πείραμα, για την κάθε μετοχή

χρησιμοποιήθηκε διαφορετικός αριθμός ειδήσεων, παρά το γεγονός οτι το χρονικό πλαίσιο ήταν το ίδιο. Καταδεικνύεται, με τον τρόπο αυτό, ότι στην μετοχή όπου χρησιμοποιήθηκε μικρότερο πλήθος ειδήσεων, ο περιορισμός της απώλειας ήταν μικρότερος.

Ο συνδυασμός των δύο προβλέψεων, δηλαδή το υβριδικό μοντέλο, δεν στάθηκε ικανό να δώσει καλύτερα αποτελέσματα, τουλάχιστον στο παραπάνω πείραμα.

Το γενικό συμπέρασμα που εξάγεται από την εφαρμογή του συγκεκριμένου πειράματος είναι ότι το financial sentiment όχι μόνο επηρεάζει τις τιμές των μετοχών αλλά έχει την δυνατότητα να προβλέψει σε ένα βαθμό και την πορεία των τιμών των μετοχών. Με αυτόν τον τρόπο, ο εκάστοτε επενδυτής μπορεί να αυξήσει την κερδοφορία του ή να περιορίσει την απώλειά του.

ΚΕΦΑΛΑΙΟ 8: ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

Η επιστήμη των προβλέψεων κατέχει κυρίαρχη θέση στη σύγχρονη χρηματοοικονομική αγορά. Τόσο οι στατιστικές όσο και οι κριτικές προβλέψεις λειτουργούν συνδυαστικά η μία με τη βοήθεια των υπολογιστών και η άλλη με στο μυαλό των επενδυτών. Μόλις την τελευταία δεκαπενταετία οι υπολογιστές έχουν αρχίσει να συμμετέχουν στο judgmental forecasting με τόσο ενεργό τρόπο. Η ιδέα του συνδυασμού της χρήσης της υπολογιστικής γλωσσολογίας με τα οικονομικά δημοσιεύματα είναι δε ακόμα πιο πρόσφατη.

8.1 Πειραματικές προεκτάσεις

Στην παρούσα διπλωματική εργασία πραγματοποιήθηκε ένα πείραμα το οποίο κατέδειξε ότι η χρήση προγραμμάτων ερμηνείας του συναισθήματος που εκπέμπει μία οικονομική είδηση είναι ικανή να ωφελήσει έναν επενδυτή στο να καταγράψει κέρδη στην επενδυτική επιλογή που θα κάνει, αλλά και να προστατευτεί από τυχούσες άστοχες επιλογές μειώνοντας τη ζημία του. Φυσικά το παραπάνω πείραμα δύναται να επαναληφθεί αλλάζοντας αρκετές παραμέτρους έτσι ώστε να αντλήσει κανείς παραπάνω συμπεράσματα.

- Επιλογή επενδυτικών προϊόντων

Στο πείραμα που πραγματοποιήθηκε έγινε επιλογή τριών μετοχικών τίτλων της αμερικάνικης χρηματαγοράς. Όμως όπως γίνεται εύκολα αντιληπτό οι μετοχές ενώ είναι ευρέως διαδεδομένοι επενδυτικοί στόχοι, δεν είναι και οι μοναδικοί. Η αγορά των ομολόγων είναι επίσης ένας τεράστιος τομέας στο οποίο επιδρά καθοριστικά το financial sentiment. Ακόμα, τα commodities όπως το πετρέλαιο και ο χρυσός έχουν τη δική τους θέση στην αγορά και καταλαμβάνουν ιδιαίτερη θέση στα επενδυτικά ή και αποταμιευτικά χαρτοφυλάκια. Επίσης πολλά δημοσιεύματα δεν ασχολούνται με μεμονωμένα χρηματοοικονομικά προϊόντα αλλά με σύνολα προϊόντων όπως για παράδειγμα τις επιχειρήσεις του ασφαλιστικού κλάδου, ή ακόμα και με ολόκληρους χρηματιστηριακούς δείκτες. Συνεπώς το πείραμα θα μπορούσε να έχει προεκτάσεις σε άλλα οικονομικά προϊόντα, σε ένα χαρτοφυλάκιο πολλών διαφορετικών οικονομικών ειδών ή ακόμα και σε χρηματιστηριακούς δείκτες.

- Πλήθος δεδομένων

Για την εκτέλεση του πειράματός μας χρησιμοποιήθηκαν τρείς μετοχικοί τίτλοι σε μία χρονική περίοδο τεσσάρων μηνών. Μία ενδιαφέρουσα προέκταση λοιπόν θα ήταν η χρήση μεγαλύτερου αριθμού μετοχών ή χρεογράφων εν γένει αλλά και μεγαλύτερο εύρος των παρατηρήσεων είτε πρόκειται για τιμές κλεισίματος είτε για πλήθος ειδήσεων ανά ημέρα.

- Χρονικός ορίζοντας

Όπως έχει προαναφερθεί το χρονικό διάστημα στο οποίο καταγράφηκαν οι παρατηρήσεις και αναλύθηκαν τα δεδομένα ήταν οι τέσσερεις μήνες, κυρίως για πρακτικούς λογούς. Η επέκταση του χρονικού διαστήματος πέραν των τεσσάρων μηνών θα μπορούσε να προδώσει μεγαλύτερη ακρίβεια στις παρατηρήσεις και να προσεγγίσει περισσότερο την πρακτική των επενδυτών που προσβλέπουν στο μακροπρόθεσμο κέρδος.

- Επενδυτική πρακτική

Το ποσό που καλείται να επενδύσει κάποιος μπορεί να είναι συγκεκριμένο αλλά το πώς θα το επενδύσει μπορεί να διαφέρει. Στο πείραμά επιλέχθηκε να επενδύεται κάθε φορά όλο το ποσό που είχε στη διάθεσή του ο επενδυτής ενώ μία ενδιαφέρουσα διαφοροποίηση θα ήταν να υπάρχει ένας διαφορετικός τρόπος τοποθέτησης των χρημάτων ανάλογα με τα αποτελέσματα του financial sentiment analysis. Θα μπορούσε για παράδειγμα ο επενδυτής να δρα πιο συντηρητικά και να κάνει χρήση μέρος μόνο των χρημάτων που του είναι διαθέσιμα ή να ρευστοποιεί μέρος του χαρτοφυλακίου όταν του υποδεικνύεται ότι πρέπει να πουλήσει.

- Προγράμματα ανάλυσης sentiment

Οι αλγοριθμικές εφαρμογές που αναλύουν και ποσοτικοποιούν το συναίσθημα των ειδήσεων διαφέρουν μεταξύ τους. Η χρήση διαφορετικών λεξιλογίων ή διαφορετικών βαρών στις ήδη υπάρχουσες λέξεις θα είχαν τη δυνατότητα να εξάγουν πιο ασφαλή αποτελέσματα. Η σύγκριση των προγραμμάτων αυτών ή ο συνδυασμός τους θα μπορούσε να προεκτείνει την πειραματική διαδικασία προς όφελος των συμπερασμάτων που θα προκύψουν.

- Μαθηματικές μέθοδοι

Τέλος, για το συγκεκριμένο πείραμα τις μεθόδους προβλέψεων βάσει του μικρότερου μέσου απολύτου ποσοστιαίου σφάλματος μεταξύ των SES, Holt και ARIMA. Προφανώς υπάρχουν στη βιβλιογραφία και την πρακτική και άλλα μοντέλα πρόβλεψης

που θα μπορούσε να ελεγχθεί το αν η συμβολή τους θα ήταν επικερδής ή επιζήμια στα αποτελέσματα των επενδύσεών που έγιναν.

8.2 Συστημικές Προεκτάσεις

Η διπλωματική αυτή διατριβή, όπως έχει προαναφερθεί, εντάσσεται σε ένα γενικότερο και μεγαλύτερο εγχείρημα, τη δημιουργία ενός πρότυπου συστήματος συλλογής πληροφοριών, ειδήσεων και απόψεων γύρω από την οικονομία και τα χρηματοοικονομικά προϊόντα, την ανάλυσή τους, βάσει του οικονομικού συναισθήματος που περιέχουν και την εξαγωγή ενός ποσοστού «έγκρισης» ή «απόρριψης» μιας ενδεχόμενης επένδυσης. Η μελλοντική πρόκληση λοιπόν, αφορά φυσικά την αυτοματοποίηση των διαδικασιών που αναλύθηκαν σε αυτήν την εργασία με τη χρήση των γλωσσών του προγραμματισμού σε συνδυασμό με τις νέες εξελίξεις στον τομέα του Natural Language Processing.

Η συν-λειτουργία ενός πλήθους λεξικών αναγνώρισης των συναισθηματικών λέξεων και φράσεων με λεξικά οικονομικών όρων και η προέκταση της αναλυτικής διαδικασίας αποτελεί επίσης πρόκληση για το μέλλον.

Επίσης το sentiment analysis μπορεί να επεκταθεί στην αγορά του συναλλάγματος. Οι ανακοινώσεις των κεντρικών τραπεζών για τα επιτόκια και τα αποθεματικά, αλλά και οι προβλέψεις που κάνουν οι αναλυτές έχουν τη δυνατότητα να προβλέψουν τόσο τιμές εθνικών χρηματιστηριακών δεικτών αλλά και να διαμορφώσουν ολόκληρα χαρτοφυλάκια.

Τέλος, η μελέτη για το financial sentiment έχει τη δυνατότητα να επεκταθεί στην πρόβλεψη των μεγεθών των εθνικών οικονομιών. Οι αναλύσεις της πολιτικής οικονομίας των κρατών και οι κρίσεις που εκφράζονται έχουν τη δυνατότητα να επηρεάσουν τα εθνικά ομόλογα και τα παράγωγά τους. Άλλωστε ο δανεισμός των κρατών αποτελεί βασικό στοιχείο των εθνικών οικονομιών και η αναζήτηση του φθηνού δανεισμού καθορίζει σημαντικά την πορεία των αναπτυξιακών ή υφεσιακών στοιχείων κάθε οικονομίας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Ασημακόπουλος Β., Πετρόπουλος Φ. 2011. *Επιχειρησιακές Προβλέψεις*, Αθήνα.

2. Abbasi, A., H. Chen and A. Salem 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems* 26(3).

3. Ahmad A., Hsinchun Chen, and Arab Salem. 2008. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. ACM Transactions on Information Systems (TOIS) , 26:No. 3, Article 12, June.

4. Antweiler, W. and M. Frank 2004. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance* 59(3): 1259-1294.

5. Baker, M. and J. Wurgler, 2007: Investor sentiment in the stock market. Journal of Economic Perspectives, 21(2), 129–152

6. Benamara, F., C. Cesarano, A. Picariello, D. Reforgiato, and VS Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In Proceedings of International Conference on Weblogs anδ Social Media, ICWSM,Boulder,CO

7. Choi Y. and C. Cardie. 2008. "Learning with compositional semantics as structural inference for subsentential sentiment analysis", In Proc. of EMNLP

8. Coval, J.D., and Shumway, T.: Is sound just noise? Journal of Finance, 56(5) 1887-191

9. Davis, A., J. Piger and L. Sedor 2006. Beyond the numbers: an analysis of optimistic and pessimistic language in earnings press releases. Technical Report. Federal Reserve Bank of St. Louis.

10. Devitt, A. and K. Ahmad 2007. Sentiment Polarity Identification in Financial News: A Cohesion-Based Approach. *Association of Computational*

Linguistics, Prague, Czech Republic

11. Ekman P. and W. V. Friesen. 1971. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17:124–129

12. Ghose, A., P. Ipeirotis and S. Arun 2007. Opinion Mining Using Econometrics: A Case Study on Reputation Systems. *Association of Computational Linguistics*, Prague, Czech Republic.

13. Gidofalvi, G. 2001. Using News Articles to Predict Stock Price Movements. Department of Computer Science and Engineering. University of California, San Diego.

14. Hyndman, R.J., Athanasopoulos (2014) Forecasting: principles and practice, OTexts: Melbourne, Australia. http://www.otexts.org/fpp.

15. Hyndman, R.J. and Khandakar, Y. (2008) "Automatic time series forecasting: The forecast package for R", Journal of Statistical Software, 26(3).

16. Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D. (2008) Forecasting with exponential smoothing: the state space approach, Springer-Verlag: New York. http://www.exponentialsmoothing.net.

17. Grefenstette, G., Y. Qu, J. Shanahan and D. Evans 2004. Coupling Niche Browsers and Afect Analysis for an Opinion Mining Application. *7th International Conference on "Recherche d'Information Assistee par Ordinateur"*, Avignon, France.

18. Hatzivassiloglou, Vasileios and Kathleen McKeown. 1997. Predicting the semantic orientation of adjectives. In Proceedings of 35th Meeting of the Association for Computational Linguistics, pages 174–181, Madrid.

19. Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of KDD '04, the ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–--177, Seattle, US. ACM Press.

20. Kim, Soo-Min and E. Hovy. 2004.Determining the sentiment of opinions. In Proceedings of COLING 2004, pages 1367–1373, Geneva

21. Kumar, A., and Ch. Lee. (2006) "Retail Investor Sentiment and Return Comovement," Journal of Finance.

22. Loughran, T. & McDonald, B. (2011). When is a liability not a liability? Textual Analysis, Dictionaries and 10-Ks. The Journal of Finance, 66(1), 35-66.

23. Mishne, G. 2005. Experiments with Mood Classification in Blog Posts. *SIGIR*, Salvador, Brazil.

24. Mitchell, M.L., J. Harold Mulherin.J.H. 1994: The impact of public information on the stock market. Journal of Finance, 49(3):-95

25. Osgood, C.E. George J. Suci, and Percy H. Tannenbaum. 1957. *The Measurement of meaning*. University of Illinois Press, Chicago, Ill.

26. Pang, B., L. Lee and S. Vaithyanathan 2002. Thumbs up?: Sentiment classification using machine learning techniques. *Association for Computational Linguistics*, Philadelphia, PA.

27. Schumaker, R. P. and H. Chen 2006. Textual Analysis of Stock Market Prediction Using Financial News Articles. *Americas Conference on Information Systems*, Acapulco, Mexico.

28. Sekine, S. and C. Nobata 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Language Resources and Evaluation Conference*, Lisbon, Portugal.

29. Statman, M., 1999: Behaviorial finance: Past battles and future engagements. Financial Analysts Journal, 55 (6), pp. 18–27.

30. Stone, Philip J., Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. The General Inquirer: A Computer Approach to Content Analysis. MIT Press, Cambridge, MA

31. Taboada, Maite, Caroline Anthony, and Kimberly Voll. 2006. Creating semantic orientation dictionaries. In Proceedings of 5th International Conference on Language Resources and Evaluation (LREC), pages 427–432, Genoa

32. Tetlock, P. 2007. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance* 62(3): 1139-1168.

33. Tetlock, P. C., Saar-Tsechansky, M., Macskassy, S. (2008): More than words: Quantifying language to measure _rms' fundamentals. Journal of Finance 63(3), 1437-1467.

34. Turney, P. 2002. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", In Proc. of ACL, pages 417–424.

35. Turney, P. and M. Littman. 2003.Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21(4):315–346

36. Young, L. & Soroka, S. (April 2011). Affective news: The automated coding of sentiment in political texts, forthcoming in Political Communication.

37. Valitutti, A., C. Strapparava, and O. Stock. 2004. Developing affective lexical resources. *PsychNology Journal*, 2(1):61–83.

38. Wiebe, J., T. Wilson and M. Bell 2001. Identifying Collocations for Recognizing Opinions. *Association for Computational Linguistics*, Toulouse, France.

39. Wiebe, J., T. Wilson, R. Bruce, M. Bell and M. Martin 2004. Learning Subjective Language. *Computational Linguistics* 30(3): 277-308.

40. Wilson, T., Wiebe, J., and Hofmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technologies Conference/ Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, CA.