Τμήμα Οικονομικών Επιστημών Πανεπιστημίου Θεσσαλίας

Τμήμα Πολιτικών Μηχανικών Πανεπιστημίου Θεσσαλίας

Τμήμα Φυσικής Διεθνούς Πανεπιστημίου Ελλάδος

Μεταπτυχιακή Διατριβή

Ανάλυση επενδυτικών "συναισθημάτων" και ρίσκου με βάση την εξόρυξη δεδομένων, την επεξεργασία φυσικών γλωσσών και τα κοινωνικά δίκτυα

Αλεξανδρίδης Αναστάσιος

ΑΕΜ : 00033

Επιβλέπων Καθηγητής

Παπαδάμου Στέφανος

Βόλος 2021

University of Thessaly

Department of Economics

Department of Civil Engineering

International Hellenic University

Department of Physics

Postgraduate Thesis

Sentiment & risk Analysis of Financial investments based on data Mining, Natural Language Processing, and Social Networks.

Anastasios Alexandridis

Supervisor

Papadamou Stefanos

Volos 2021

Πανεπιστήμιο Θεσσαλίας

Τμήμα Οικονομικών Επιστημών

Τμήμα Πολιτικών Μηχανικών

Διεθνές Πανεπιστήμιο Ελλάδος

Τμήμα Φυσικής

Διπλωματική Διατριβή

Ανάλυση επενδυτικών "συναισθημάτων" και ρίσκου με βάση την εξόρυξη δεδομένων, την επεξεργασία φυσικών γλωσσών και τα κοινωνικά δίκτυα

Αλεξανδρίδης Αναστάσιος

Επιβλέπων Καθηγητής

Παπαδάμου Στέφανος

Βόλος 2021

Αυτή η εργασία αφιερώνεται στα τμήματα Οικονομικών Επιστημών, Πολιτικών Μηχανικών, Φυσικής, καθώς και στο εκπαιδευτικό προσωπικό της Οικονομικής Φυσικής για το κοινωνικό έργο που επιτελεί. Τους ευχαριστώ για την τεχνογνωσία που μου προσέφεραν.

Σύνοψη

Είναι γνωστό ότι η συμπεριφορική χρηματοοικονομική είναι ένα σημαντικό ζήτημα των σύγχρονων αγορών. Οι επενδυτικές αποφάσεις επηρεάζονται από την δημοσιότητα των χρηματοοικονομικών άρθρων. Είναι εμφανές πως υπάρχει μία σύνδεση μεταξύ του συναισθήματος το οποίο εκφράζεται από τον Τύπο και της διαδικασίας λήψης αποφάσεων των επενδυτών. Σύμφωνα με την προαναφερθείσα σύνδεση, αυτή η εργασία προσπαθεί να ανακαλύψει την σχέση μεταξύ της μεταβλητότητας η οποία επηρεάζει την τιμή των μετοχών και το συναισθηματικό περιβάλλον του χρηματοοικονομικού Τύπου, ο οποίος αντιπροσωπεύει τον επενδυτικό κόσμο. Έχοντας ως γνώμονα την εύρεση μίας τέτοιου είδους σχέσης, ένας νέος δείκτης δημιουργήθηκε ο οποίος ποσοτικοποιεί το συναίσθημα της λύπης (Sadness) που εμπεριέχεται στα χρηματιστηριακά άρθρα. Αυτός ο δείκτης είναι το προϊόν των metadata, που πηγάζουν από την επεξεργασία της φυσικής γλώσσας (Αγγλική), ενός συνόλου άρθρων με εξαιρετικά υψηλή δημοτικότητα. Το εργαλείο που χρησιμοποιείται για την επεξεργασία των φυσικών γλωσσών είναι το Volume Language understanding IBM Watson. Για την εύρεση οικονομετρικής σχέσης μεταξύ του δείκτη Sadness και του Down Jones εφαρμόστηκε ένα Vector Error Correction Model με την βοήθεια του οικονομετρικού πακέτου Eviews. Τα αποτελέσματα έδειξαν πως ο δείκτης Sadness συμπεριφέρεται σχεδόν το ίδιο με τον VIX.

Abstract

It is well known that behavioral finance is a key aspect of the modern markets. The investment decisions are affected by the public gaze of the financial articles since it seems to be a linkage between the emotion which is expressed by the press and the decision-making process of the investors. Under the scope of the above-mentioned linkage this paperwork attempts to discover the relationship between the volatility which affects the value of the stocks and the emotional environment of the financial press that represent the investors' world. In order to discover such a relationship a new index or timeseries is formed which quantifies the sentiment of sadness that lies in the financial articles. This new index is the product of the metadata that sourcing from the Natural Language Processing and emotional analysis of a corpus of articles with extremely high publicity. The tool for the NLP and the sentiment analysis is the Volume Language understanding IBM Watson. For the econometric relationship between the timeseries "Sadness" and the Down Jones Index a Vector Error Correction Model has been applied on the econometric package Eviews. The results indicate that the sentiment index of "Sadness" behaves almost like the Volatility Index VIX.

Λέξεις Κλειδιά

Behavioral economics

Natural Language Processing – NLP

machine learning

supervised learning

unsupervised learning

reinforcement learning

volatility

Volatility Index VIX

Sadness

Down Jones Index DJI

Natural Language based Financial Forecasting NLFF

emotion recognition

DeepQA

corpus of knowledge

curating the context

weighted evidence scores

Google Trends

Vector Error Correction Model

Περιεχόμενα

1.Εισαγωγή

Στην σύγχρονη κοινωνία, όπου τα δεδομένα, το διαδίκτυο και τα μέσα κοινωνικής δικτύωσης είναι βασικό κομμάτι της καθημερινής ζωής, η συμπεριφορά των κοινωνικών μονάδων και ομάδων μπορεί να επηρεαστεί από τον σύγχρονο τύπο. Η συμπεριφορική και η μαθηματική της ερμηνεία είναι ένα τεράστιο κεφάλαιο της διαδικτυακής πραγματικότητας, το οποίο δεν θα μπορούσε να αφήσει ανεπηρέαστη την οικονομική επιστήμη και τον κόσμο των επενδυτών.

Στην οικονομική επιστήμη η έννοια της συμπεριφορικής οικονομικής (Behavioral economics) και κατ' επέκταση της χρηματοοικονομικής θεωρίας είναι ένας σχετικά καινούργιος τομέας, που μόλις τα τελευταία 20 χρόνια κέρδισε έδαφος στο διεθνές στερέωμα. Ο σκοπός της συμπεριφορικής οικονομικής θεωρίας μελετά τις επιπτώσεις διαφόρων ψυχολογικών, κοινωνικών αλλά και συναισθηματικών παραγόντων στις ατομικές αλλά και στις συλλογικές αποφάσεις (π.χ. οργανισμών) οι οποίες διαφοροποιούνται από την κλασική οικονομική θεωρία (Lin, 2011).

Το 1979 οι Kahneman και Tversky δημοσίευσαν ένα άρθρο με τίτλο "Prospect theory : An Analysis of Decision Under Risk" (Kahneman, 1979) το οποίο βασίστηκε στην γνωστική ψυχολογία για να επεξηγήσει τα σημεία που αφορούν την λήψη οικονομικών αποφάσεων, κάτι που δεν είναι απόλυτα εφικτό από την κλασική οικονομική θεωρία (Lin, 2011).Για την παραπάνω εργασία οι ερευνητές κέρδισαν το βραβείο Νόμπελ Οικονομικών το 2002.

Σήμερα είναι γνωστό από την συμπεριφορική χρηματοοικονομική ότι η συμπεριφορά των επενδυτών είναι ένας από τους σημαντικότερους παράγοντες, και για πολλούς αναλυτές ίσως ο σημαντικότερος στην αύξηση του ρίσκου και της μεταβλητότητας των αγορών.

Το ερώτημα όμως παραμένει. Πώς μπορούμε να ορίσουμε και να ποσοτικοποιήσουμε τους συμπεριφορικούς παράγοντες, ιδιαίτερα σε έναν σύγχρονο ψηφιακό κόσμο; Η απάντηση μπορεί να δοθεί από έναν άλλο νέο και εξελισσόμενο τομέα, αυτόν της τέχνης των αλγορίθμων συμπεριφορικής ανάλυσης που βασίζονται σε σύγχρονες εφαρμογές την τεχνητής Νοημοσύνης και της επεξεργασίας των φυσικών γλωσσών (Xing, 2018).

Η επεξεργασία των φυσικών γλωσσών (Natural Language Processing – NLP) είναι ένα διεπιστημονικό πεδίο που συνδυάζει την γλωσσολογία, την επιστήμη υπολογιστών και την τεχνητή νοημοσύνη (Chowdhary, 2020).

Αν και η έννοια των ευφυών συστημάτων προτάθηκε αρχικά από τον Alan Turing το 1950 (Turing, 1950), μόλις τα τελευταία 20 χρόνια εφαρμόζονται ευρέος τεχνικές μηχανικής μάθησης (machine learning) και όχι αλγόριθμοι ωμής βίας (brutal force) σε πολύπλοκα προβλήματα. Αυτό συμβαίνει, διότι πριν δεν υπήρχε το κατάλληλο υλικό μέρος (hardware), άρα και αρκετή υπολογιστική ισχύς για να εφαρμοστούν συγκεκριμένες μαθηματικές θεωρίες μηχανικής μάθησης, όπως λ.χ. η Support Vector Machine.

Υπάρχουν τρείς βασικές κατηγορίες μηχανικής μάθησης. Η πρώτη είναι η επιτηρούμενη μάθηση (supervised learning), όπου εισάγονται δεδομένα στο σύστημα υπό την μορφή παραδειγμάτων, καθώς και τα επιθυμητά αποτελέσματα από έναν εκπαιδευτή. Η δεύτερη είναι η μη επιτηρουμένη μάθηση (unsupervised learning) στην οποία δεν δίνονται παραδείγματα και πρέπει από μόνο του το σύστημα να βρει τα μοτίβα (patterns) που ¨κρύβονται¨ στα δεδομένα. Τέλος έχουμε την ενισχυτική μάθηση (reinforcement learning), όπου υπάρχει ένας στόχος (π.χ. ένας ψηφιακός παίκτης ενός ηλεκτρονικού παιχνιδιού), ο οποίος πρέπει να επιτευχθεί μέσα σε ένα δυναμικό περιβάλλον (Bishop, 2006).

Με γνώμονα τις προαναφερθείσες εξελίξεις στον κόσμο της πληροφορίας η εν λόγω εργασία προχωρά στην μελέτη της συμπεριφοράς των επενδυτών που είναι σχετική με την μεταβλητότητα (volatility) και την συναισθηματική κατάσταση της κοινής γνώμης στον χρηματοοικονομικό κόσμο. Για να οριστεί ποσοτικά η συναισθηματική κατάσταση των επενδυτών, δημιουργούνται metadata από κείμενα και άρθρα που έχουν δημοσιευθεί στον τύπο και στο twitter, χρησιμοποιώντας ένα προηγμένο εργαλείο επεξεργασίας φυσικών γλωσσών, το Natural Volume Language understanding IBM Watson, που είναι σε θέση να παραγάγει ως metadata τα βασικά συναισθήματα που χαρακτηρίζουν ένα κείμενο.

Το ερωτήματα αυτής της εργασίας είναι :

- Κατά πόσο επηρεάζει την μεταβλητότητα (volatility) η συναισθηματική διάσταση και κατά πόσο συσχετίζεται με τον δείκτη της μεταβλητότητας;
- Υπάρχει σχέση του χρηματιστηριακού δείκτη Down Jones και των συναισθημάτων που εκφράζουν τα άρθρα του χρηματιστηριακού τύπου;
- Μπορεί να δημιουργηθεί ένας δείκτης που εκφράζει το συναίσθημα του τύπου και να λειτουργήσει όπως ένας κλασικός δείκτης μεταβλητότητας (Volatility Index VIX);

Για να απαντηθούν τα παραπάνω ερωτήματα δημιουργήσαμε μία χρονοσειρά προερχόμενη από τα metadata του συναισθήματος της λύπης (Sadness), το οποίο ήταν και το κυρίαρχο συναίσθημα στον χρηματιστηριακό τύπο των ΗΠΑ κατά την πρώτη περίοδο της έξαρσης του Covid. Η χρονική περίοδος των άρθρων που επιλέχθηκαν ήταν από 03/02/2020 έως 28/05/2020. Αυτή την χρονοσειρά την χρησιμοποιήσαμε ως δείκτη μεταβλητότητας (VIX) σε ένα μοντέλο VECM για να δούμε το πώς και αν επηρεάζει (long / short run equilibrium) τον δείκτη Down Jones.

Το αποτέλεσμα ήταν ότι μετά από πάρα πολλές δοκιμές χρονοσειρών, που προέρχονται από διαφορετικά άρθρα, κάποιες λειτουργήσαν όπως και ο VIX. Αυτή με τα καλύτερα αποτελέσματα χρησιμοποιείται και στην εν λόγω εργασία.

2.Ανασκόπηση της βιβλιογραφίας

2.1 Αλγόριθμοι Συμπεριφοράς

Η έννοια των συμπεριφορικών αλγορίθμων είναι η αυτόματη αναγνώριση και "κατανόηση" της ανθρώπινης συμπεριφοράς σε υπολογιστικά και ευφυή περιβάλλοντα (Ramon, 2019).Η ανάλυση των εν λόγω συστημάτων βασίζεται σε δεδομένα τα οποία προέρχονται από εκφράσεις του προσώπου και κινήσεις του σώματος, ηχητικά σήματα (ομιλία) και κείμενα από φυσικές γλώσσες.

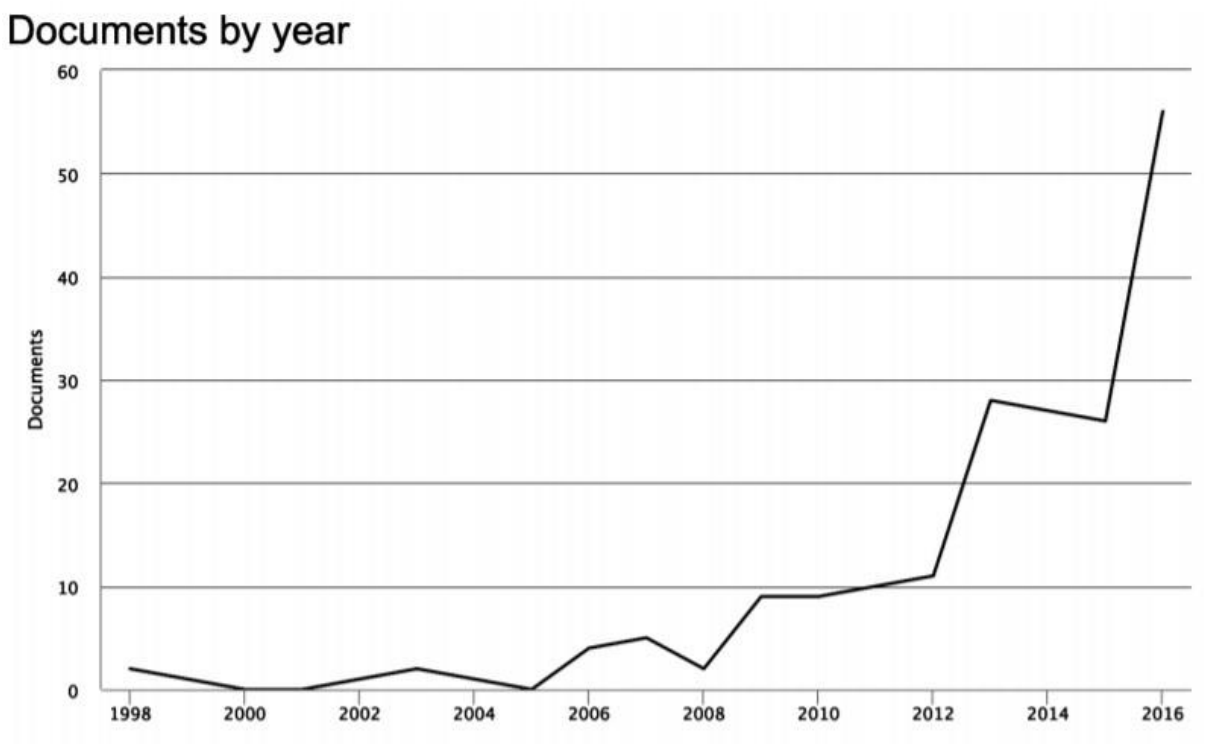
Σύμφωνα με τον (Mao, 2015) ένα καλό μοντέλο συμπεριφοράς δεν αξιολογεί ένα συγκεκριμένο σύστημα αλλά έχει ως σκοπό τον ευρύτερο στόχο να μελετηθεί η ανθρώπινη συμπεριφορά αναπτύσσοντας γενικευμένα μοντέλα σε διαφορετικά πεδία.

2.2 Υπολογιστική ψυχολογία

Σύμφωνα με τον (Sun R. , 2008) η υπολογιστική ψυχολογία έχει ως σκοπό την εξερεύνηση του νοήματος της γνωστικής λειτουργίας (συμπεριλαμβανομένων των κινήτρων, των συναισθημάτων, της αντίληψης και διαφόρων γνωστικών λειτουργιών) μέσω της κατανόησης αυτών από συγκεκριμένα υπολογιστικά μοντέλα, τα οποία είναι αλγοριθμικές διαδικασίες. Οι μαθηματικές έννοιες που εμπεριέχονται (π.χ. εξισώσεις, σχέσεις, οντότητες) προσδιορίζονται από μοντέλα τα οποία έχουν νοητική/εννοιολογική διάσταση. Στην περίπτωσή μας πρόκειται για την ποσοτικοποίηση των βασικών συναισθημάτων, και ιδιαίτερα της λύπης (Sadness), τα οποία εκφράζονται μέσα από τα άρθρα διαφόρων διάσημων αρθρογράφων του χρηματοοικονομικού κόσμου.

2.3 Συναφές έργο

Στην βιβλιογραφία και στον ερευνητικό χώρο έχει δημιουργηθεί ένα νέο πεδίο, αυτό της χρηματοοικονομικής πρόβλεψης, βασιζόμενης στην φυσική γλώσσα (Natural Language based Financial Forecasting (NLFF) (Xing, 2018), η οποία βρίσκει εφαρμογές κυρίως στην χρηματιστηριακή αγορά. Η γραφική αναπαράσταση των σχετικών άρθρων που έχουν δει το φως της δημοσιότητας τα τελευταία 20 χρόνια απεικονίζεται στο γράφημα 1 .



Γράφημα 1 - Ο αριθμός των άρθρων που είναι σχετικός με την NLFF, πηγή: (Xing, 2018)

Για την αναπαράσταση χρηματοοικονομικών δεδομένων κειμένου (textual financial data) ως βασικά χαρακτηριστικά (features), τα οποία εύκολα ένας υπολογιστής μπορεί να επεξεργαστεί, οι περισσότερες ερευνητικές εργασίες για NLFF χρησιμοποιούν την τεχνική του σάκου των λέξεων (bag of words). Σύμφωνα με την τεχνική του σάκου των λέξεων κάθε κείμενο το οποίο απαρτίζεται από μικρές προτάσεις αναπαρίσταται ως ένας σάκος (πολύ-σύνολο) λέξεων. Βάσει της τεχνικής του σάκου των λέξεων εφαρμόζεται η μέθοδος ¨vectorization¨, η οποία αποτελείται από τρεις βασικές τεχνικές “tokenization”, “counting” και “normalization”. Το κείμενο διαιρείται σε

λεξικογραφικές μονάδες (tokens). Τα δεδομένα ή λεξικογραφικές μονάδες μετατρέπονται σε αριθμητικά ανύσματα (numerical vectors) με βάση την συχνότητα με την οποία εμφανίζονται στο εν λόγω κείμενο (counting). Τέλος εφαρμόζεται η κανονικοποίηση (normalization) (Zhang, 2010).

Μερικές φορές οι προαναφερθείσες υπολογιστικές μέθοδοι μπορούν να χρησιμοποιηθούν μαζί με τεχνικούς χρηματοοικονομικούς δείκτες, όπως ο κινητός μέσος (Moving Average Converge Divergence – MACD). Ένας άλλος τομέας εφαρμογών είναι διάφοροι μακροοικονομικοί δείκτες οι οποίοι λόγω της αδόμητης φύσης τους είναι κατάλληλοι για διάφορες NLP τεχνικές. Τέλος αρκετά εννοιολογικά χαρακτηριστικά και συναισθηματικές πληροφορίες αντλούνται από διάφορα νέα του χρηματοοικονομικού κόσμου (Xing, 2018). Η συναισθηματική ανάλυση χαρακτηρίζεται ως ένα πολυσύνθετο πρόβλημα (suitcase research problem) το οποίο έχει να αντιμετωπίσει πολλά θέματα της επεξεργασίας φυσικών γλωσσών. Για αυτό χαρακτηρίζεται ως και μία μεγάλη βαλίτσα, όπως άλλωστε και η αναγνώριση συναισθημάτων (emotion recognition) και η εξόρυξη γνώμης (opinion mining) (Cambria, 2017).

Ο (Butler, 2009) χρησιμοποίησε δύο βασικές τεχνικές συσταδοποίησης (n-gram modeling ή bag of words και Support Vector Machine) για την επεξεργασία φυσικών γλωσσών με σκοπό την ανάλυση εταιρικών ετήσιων αναφορών (corporate annual reports), ώστε για να προβλέψουν την απόδοση της μετοχής για τον επόμενο χρόνο. Παρόλη την ικανοποιητική απόδοση των μοντέλων, δεν ήταν δυνατόν να δημιουργήσουν ένα χαρτοφυλάκιο επενδύσεων με διασπορά (spread) του ρίσκου λόγω της μεταβλητότητας που παρουσίαζε η αγορά από την κρίση του 2008. Τέλος οι μελετητές τονίζουν πως απαιτείται μεγαλύτερος αριθμός εταιρειών από αυτόν του S&P 500 για πιο αποδοτικές δοκιμές, επειδή το σύνολο των ετήσιων αναφορών, που είναι και το σύνολο δεδομένων, δεν είναι αρκετά ικανοποιητικό.

Ο (Groth, 2011) χρησιμοποιώντας NLP τεχνικές συγκρίνει διάφορες ιστορίες οι οποίες είναι σχετικές με την χρηματιστηριακή αγορά και με τις τιμές των μετοχών. Η τεχνικές μηχανικής μάθησης που χρησιμοποιήθηκαν είναι Naïve Bayes, K-nearest Neighbor, Neural Networks και Support Vector Machine με σκοπό τον εντοπισμό μοτίβων τα οποία θα μπορούσαν να εξηγήσουν την αύξηση της μεταβλητότητας (Volatility) και του ρίσκου. Τα αποτελέσματα έδειξαν πως τα αδόμητα δεδομένα κειμένου είναι πολύτιμη πηγή πληροφορίας σχετικά με το ρίσκο των αγορών και προτείνουν μια εξόρυξη δεδομένων κειμένου σε καθημερινή βάση. Τέλος οι Knn,NNet και

SVM αποδίδουν καλύτερα από την Naïve Bayes. Στο ίδιο μήκος κύματος κινείται και ο (Hagenau, 2013) βελτιώνοντας την απόδοση των τεχνικών συσταδοποίησης κειμένων με χρηματοοικονομικά νέα και αναφέρει πως η ακρίβεια για την πορεία των μετοχών βάσει του χρηματοοικονομικού Τύπου υπερβαίνει το 58%.

Στο άρθρο (Sun B. C., 2014) δόθηκε μία μέθοδος προ-επεξεργασίας για την αφαίρεση του θορύβου από χρηματοοικονομικά κείμενα (π.χ. άρθρα) και την διαμόρφωση τους, ώστε να είναι χρήσιμα για περαιτέρω εξαγωγή μετά-δεδομένων σχετικά με τα συναισθήματα των επενδυτών. Έξι βήματα επεξεργασίας φυσικών γλωσσών χρησιμοποιήθηκαν, όπως συντακτικοί και σημειολογικοί αλγόριθμοι αναίρεσης (negotiation handling algorithms) για την μείωση του θορύβου. Τα αποτελέσματά τους έδειξαν καλύτερη ικανότητα ταξινόμησης από αντίστοιχα συστήματα.

Στο έργο (Alvim, 2010) χρησιμοποιήθηκε η τεχνική του σάκου των λέξεων (bag of words) για την δημιουργία μιας μεθόδου προ-επεξεργασίας (pre-processing) με σκοπό την βελτίωση της ακρίβειας στην συναισθηματική ταξινόμηση (sentiment classification). Για την εξαγωγή των κύριων χαρακτηριστικών (features) εφαρμόζεται ένας καθοδηγούμενος μαθησιακός αλγόριθμος μετασχηματισμού εντροπίας (Entropy Guided Transformation Learning Algorithm) για να εξάγει τα απαιτούμενα χαρακτηριστικά. Για την συναισθηματική ταξινόμηση (sentiment classification) χρησιμοποιήθηκαν οι Support Vector Machines και ο Naïve Bayes. Το πλήθος δείγματος ήταν 1500 άρθρα εφημερίδων σχετικά με την εταιρεία Petrobras της Πορτογαλίας. Η προσέγγιση της επεξεργασίας της φυσικής γλώσσας βελτίωσε ελάχιστα την ακρίβεια της συναισθηματικής ταξινόμησης (sentiment classification).

Στο έργο (Luccioni, 2019) χρησιμοποιούνται τα κείμενα που φανερώνουν τον κίνδυνο και την αβεβαιότητα που μπορούν να επηρεάσουν τις λειτουργίες ή την οικονομική θέση των εταιρειών, όπως απαιτείται από το νομικό σύστημα των Η.Π.Α., με σκοπό να εντοπίσουν ποιες εταιρείες γνωστοποιούν τους κινδύνους (climate risks) και ποιες όχι. Για αυτό τον λόγο χρησιμοποιήθηκε η επεξεργασία φυσικών γλωσσών για αναφορές τύπου 10-Κ. Καταλήγουν στο συμπέρασμα ότι απαιτείται ο συνδυασμός ανθρώπου και επεξεργασίας φυσικών γλωσσών από σύγχρονα συστήματα για να υπάρξει ανάλυση σε βάθος.

Στο έργο (Mishev, 2020) τονίζεται η σπουδαιότητα της ποσοτικοποίησης των συναισθημάτων στην χρηματοοικονομική αγορά. Ακόμη επισημαίνεται πως τα μοντέλα συναισθηματικής ανάλυσης γενικού σκοπού δεν είναι αποτελεσματικά στην ανάλυση οικονομικών άρθρων λόγω της γενικής τους φύσης και της αδυναμίας κατάλληλης χρήσης χαρακτηριστικών λέξεων (labels) σχετικών με το αντικείμενο. Μετά από εκατοντάδες πειραματικές διαδικασίες και δοκιμές μηχανών συναισθηματικής ταξινόμησης ο συνδυασμός ανάπτυξης και χρήσης ειδικών λεξικών με την συναισθηματική ανάλυση κρίνεται πολύ πιο αποδοτικός λόγω καλύτερης εκπαίδευσης και ρύθμισης του συναισθηματικού μοντέλου ταξινόμησης.

Στο έργο τους οι (Khant, 2018) προτείνουν μία μεθοδολογία για την επίδραση του Τύπου που έχουν τα άρθρα του Τύπου στις επενδύσεις χρησιμοποιώντας τον Naïve Bayes, ο οποίος μπορεί να χρησιμοποιηθεί στην χρηματοοικονομική ορολογία. Ακόμη τονίζεται η σπουδαιότητα της τεχνητής νοημοσύνης και της επεξεργασίας των φυσικών γλωσσών. Το εύρημα είναι πως η χρήση φίλτρων και λέξεων κλειδιών βελτιώνει την απόδοση των μοντέλων συναισθηματικής ταξινόμησης.

Στο κείμενο (Azzi, 2019) παρουσιάζεται το Sentiment Boundary Detection που χρησιμοποιείται σε θορυβώδη κείμενα χρηματοοικονομικής στο πρώτο workshop για χρηματοοικονομική τεχνολογία και επεξεργασία φυσικών γλωσσών (Financial Technology and Natural Language Processing). Τις καλύτερες επιδόσεις είχαν το Γαλλικό Fin SBD με απόδοση 0,92 και το Αγγλικό με 0,885.

3. Εργαλεία

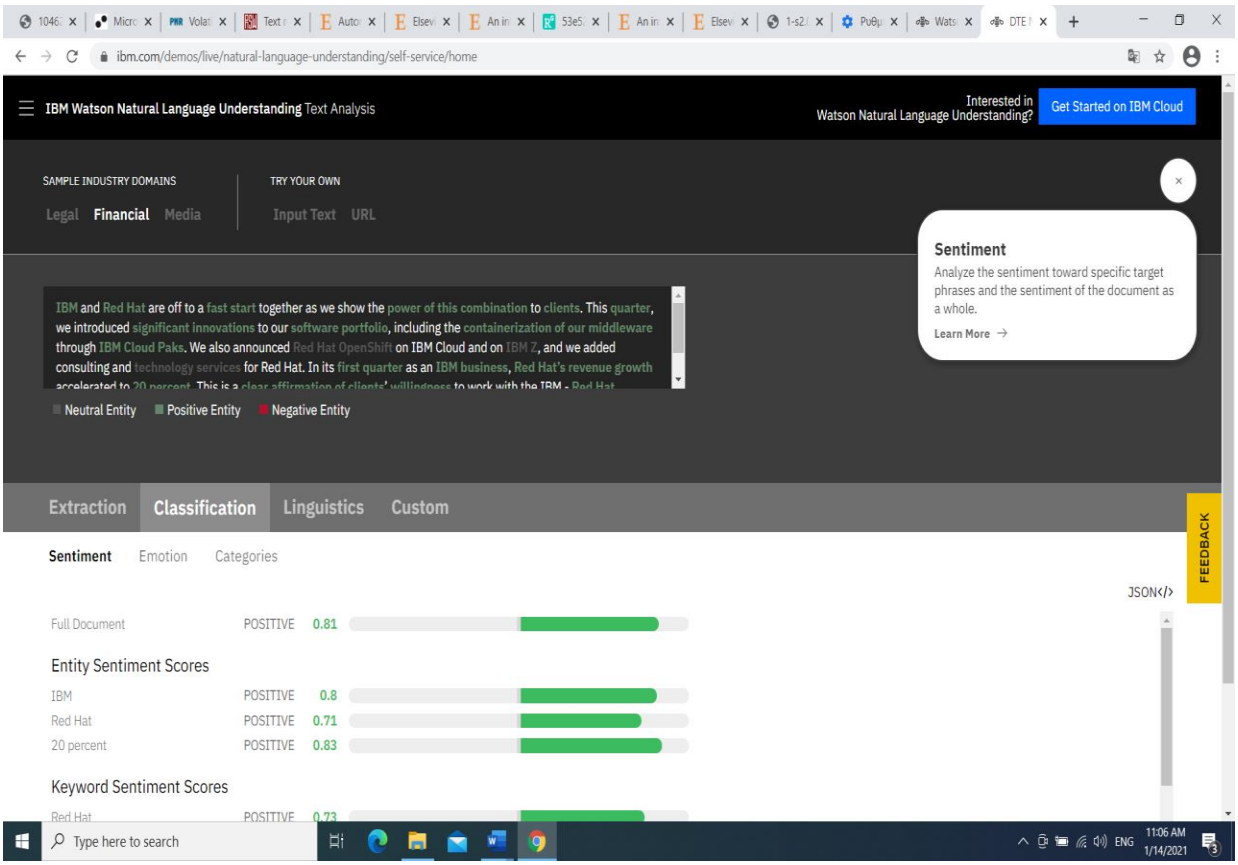
3.1 IBM Watson

Ο Watson είναι μια υπολογιστική πλατφόρμα ερωτήσεων/απαντήσεων (DeepQA), η οποία αναπτύχθηκε από το ερευνητικό κέντρο της IBM. Δεν είναι μια υπολογιστική μηχανή που λειτουργεί με μεθοδεύσεις ωμής βίας (brutal force) αλλά με μοντέλα μηχανικής μάθησης και μια από τις πιο εξελιγμένες μορφές cognitive computing. Η απλοϊκή επεξήγηση της μηχανής ακολουθεί την λογική αλληλουχία που ακολουθεί και ο άνθρωπος. Πρώτον παρακολουθεί τα στοιχεία και τα δεδομένα που διαθέτει. Δεύτερον αναπτύσσει μία υπόθεση βάσει των γνωστικών δεδομένων του. Τρίτον αξιολογεί ποια υπόθεση είναι σωστή και ποια λανθασμένη, και τέλος αποφασίζει την καλύτερη διαθέσιμη επιλογή βάσει μετρητικών εργαλείων. Πιο συγκεκριμένα δεν αναζητά συνώνυμα και λέξεις κλειδιά, όπως μία τυπική μηχανή αναζήτησης, αλλά ερμηνεύει το κείμενο και την σημειολογική του διάσταση αναλύοντας την γραμματική και την δομή του. Μπορεί να λειτουργήσει με αδόμητα δεδομένα, όπως κείμενα φυσικών γλωσσών.

Η συγκεκριμένη πλατφόρμα DeepQA, όταν εργάζεται σε ένα συγκεκριμένο πεδίο μαθαίνει την γλώσσα γύρω από τον εν λόγω τομέα με την χρήση μηχανικών μοντέλων. Με την βοήθεια ενός ειδικού συγκεντρώνει τον σχετικό βιβλιογραφικό σωρό γνώσης (corpus of knowledge) που απαιτείται για ένα συγκεκριμένο πεδίο και αποτελείται από σχετικά κείμενα (π.χ. επιστημονικά άρθρα). Τα εν λόγω δεδομένα φιλτράρονται με την βοήθεια ενός ειδικού στο συγκεκριμένο πεδίο. Η διαδικασία ονομάζεται curating the context, και η πιθανή μετάφραση είναι υπεφημερία του περιεχομένου. Μετά προχωράμε στην διαδικασία, η οποία είναι γνωστή ως ingesting the corpus, κατά την οποία ο Watson δημιουργεί ένα γράφο γνώσης (knowledge graph) για την δημιουργία σαφέστερων ερωτήσεων. Στην επόμενη φάση ξεκινάει η διαδικασία εκπαίδευσης των αλγορίθμων μηχανικής μάθησης του συστήματος από έναν ειδικό ο οποίος φορτώνει δεδομένα εκπαίδευσης στην μορφή ερωτήσεων/απαντήσεων με σκοπό την εκπαίδευση του συστήματος σε γλωσσικά μοτίβα του συγκεκριμένου πεδίου. Η διαδικασία της εκπαίδευσης είναι συνεχόμενη με σκοπό την βελτίωση του συστήματος. Το σύστημα είναι έτοιμο να δεχτεί ερωτήσεις. Αναγνωρίζει τα μέρη του γραπτού λόγου για ένα συγκεκριμένο ερώτημα, δημιουργεί υποθέσεις και αναζητά δεδομένα για να υποστηρίξει η να απορρίψει την κάθε υπόθεση. Κάθε ¨εδάφιο¨ δεδομένων λαμβάνει μία

βαθμολογία βασισμένη σε στατιστικά μοντέλα, επομένως το κάθε στοιχείο έχει ένα δείκτη βαρύτητας (weighted evidence scores). Τα στοιχεία με τις μεγαλύτερες βαρύτητες χρησιμοποιούνται για να ανακαλύψει μοτίβα που μεταφράζονται σε γνώση ή βαθμολογία συναισθήματος χρησιμοποιώντας αλγορίθμους συναισθηματικής ανάλυσης.

Όπως προαναφέρθηκε, ο Watson είναι μια deep QA πλατφόρμα. Για να γίνει εφικτή η ανάπτυξή της, η IBM συναρμολόγησε ένα υπολογιστικό cluster από πολύ-πύρινους Server με πολύ μεγάλη κύρια μνήμη, το οποίο ονομάστηκε DeepQA cluster και είναι σε θέση να εκτελέσει πολύ μεγάλης κλίμακας πειράματα (ερωτήσεων/απαντήσεων). Για την πειραματική διαδικασία αναπτύχθηκε ένα μετρητικό εργαλείο καταχώρισης πειραμάτων (Watson Error Analysis Tool – WEAT) που είναι σε θέση να μετρήσει την απόδοση των εν λόγω πειραμάτων και χρησιμοποιεί ως εργαλείο διαχείρισης δεδομένων (π.χ. διακομιστής βάσης δεδομένων με υποστήριξη JSON και XML). Μέσω αυτής της μεθοδολογίας, η οποία ονομάστηκε AdaptWatson, έγινε εφικτή η ενσωμάτωση των απαιτούμενων αλγοριθμικών τεχνοτροπιών μηχανικής εκμάθησης, έτσι ώστε να επέλθει το τελικό αποτέλεσμα. Ο Watson χρησιμοποιεί ως μηχανή επεξεργασίας φυσικής γλώσσας τον πυρήνα NLPcore του πανεπιστημίου Στάνφορντ (Brown, 2013). Αργότερα προστέθηκαν και τα εργαλεία συναισθηματικής ανάλυσης και μέτρησης συναισθημάτων, μιας και η μηχανή φέρει δυνατότητες ¨γνωστικής λειτουργίας¨ και ανάλυσης (cognition). Ένα κλασικό εργαλείο ανάλυσης είναι και το Natural Volume Language understanding το οποίο μπορεί να παραγάγει metadata από μη δομημένα κείμενα φυσικής γλώσσας, όπως τα βασικά συναισθήματα που χαρακτηρίζουν το κείμενο υπό μελέτη.



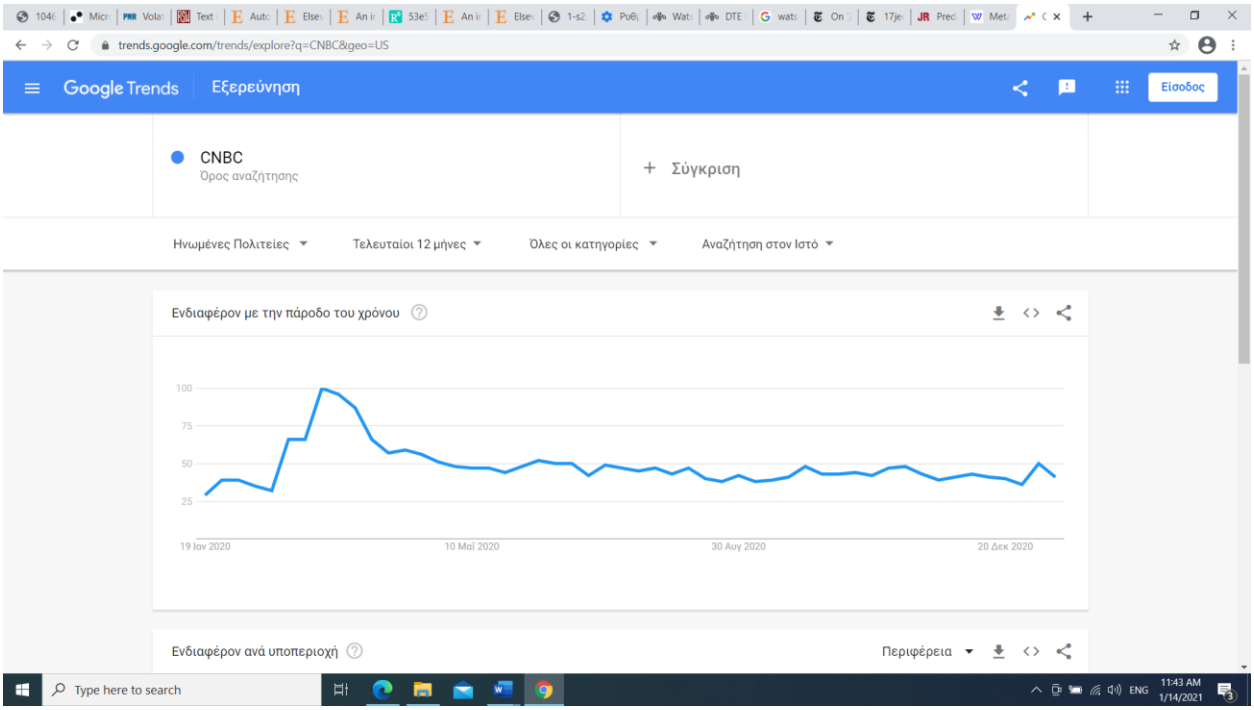
Φωτογραφία 1 – Το περιβάλλον Natural Volume Language understanding

Ο αρχικός σχεδιασμός του Watson είναι στην ουσία και τρόπος δοκιμής της εν λόγω υπολογιστικής μηχανής και πρόκειται για έναν ψηφιακό παίκτη του τηλεπαιχνιδιού Jeopardy. Τον Φλεβάρη του 2013 ο Watson πήρε μέρος σε ένα πολύ διαφορετικό παιχνίδι Jeopardy, το Jeopardy IBM Challenge, και κέρδισε τους δύο καλύτερους παίκτες όλων των εποχών στην ιστορία του παιχνιδιού.

Φωτογραφία 2 -Στιγμιότυπο από το παιχνίδι Jeopardy IBM Challenge

3.2 Google Trends

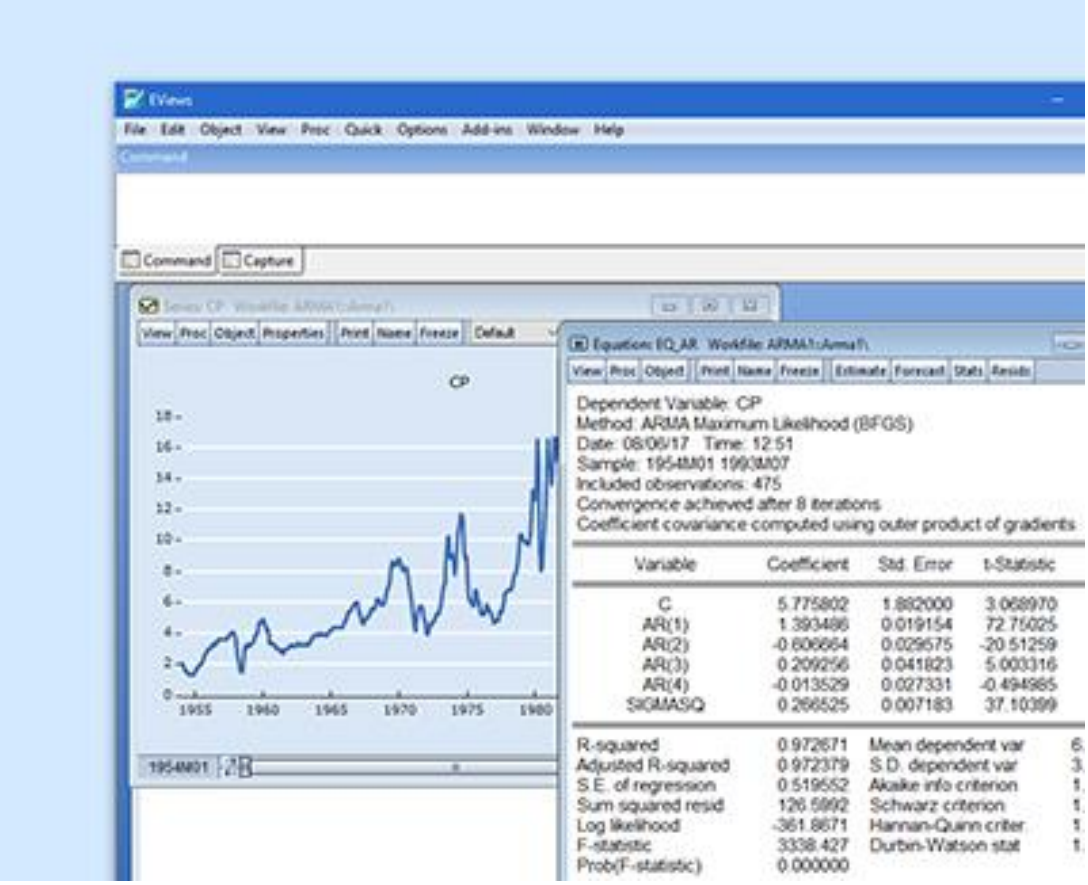
Για την μέτρηση της δημοτικότητας των άρθρων χρησιμοποιείται το εργαλείο Google Trends το οποίο αναλύει την δημοτικότητα κορυφαίων αναζητήσεων στην μηχανή αναζήτησης Google. Είναι μία πολύ ακριβής μηχανή και έχει χρησιμοποιηθεί σε αναρίθμητες έρευνες και δημοσκοπήσεις, όπως η (Lui, 2011) που μελετάει τις δυνατότητες πρόβλεψης του εργαλείου σε εκλογικές αναμετρήσεις, όπως αυτής του 2010 στις Η.Π.Α.



Φωτογραφία 3 - Η εφαρμογή Google Trends

3.3 Eviews

Για τις οικονομετρικές αναλύσεις χρησιμοποιείται το οικονομετρικό πακέτο Eviews. Πρόκειται για ένα ολοκληρωμένο και αυτοματοποιημένο πακέτο οικονομετρικών εφαρμογών με φιλικό περιβάλλον και διεπαφή χρήστη που δεν απαιτεί την συγγραφή κώδικα.



Φωτογραφία 4- Το περιβάλλον του Eviews

4. Μεθοδολογία

4.1 Η δημιουργία της χρονοσειράς Sadness

Σκοπός της σειράς Sadness είναι η ποσοτικοποίηση του συναισθήματος της λύπης σε διάφορα διάσημα άρθρα χρηματοοικονομικής ανάλυσης σχετικά με τον Down Jones. Για να επιτευχθεί αυτή η ποσοτικοποίηση, χρησιμοποιείται το εργαλείο IBM Watson Natural Language Understanding (https://www.ibm.com/cloud/watson-natural-language-understanding) το οποίο δίνει αριθμητική ένδειξη σε ποσοστιαία βάση των βασικών συναισθημάτων της λύπης, της χαράς, του φόβου και της απέχθειας που εκφράζει ένα κείμενο. Αυτό το ποσοστό είναι και το βασικό συστατικό δημιουργίας του συναισθηματικού μας δείκτη μεταβλητότητας.

Παρακάτω παρουσιάζεται η γενική μεθοδολογία που ακολουθείται για την δημιουργία της χρονοσειράς Sadness. Είναι σημαντικό να ειπωθεί ότι η μεθοδολογία δεν είναι απόλυτη, επομένως χρειάζεται πολύ μεγάλη εμπειρία με αντίστοιχα άρθρα και παρά πολλές επαναλήψεις του πειράματος μέχρι να υπάρξει μία λειτουργική χρονοσειρά, π.χ. για να πετύχουμε την χρονοσειρά που χρησιμοποιείται στην εν λόγω εργασία και να δουλεύει ικανοποιητικά, έτσι ώστε να εφαρμόζεται το VECM και τα κριτήριά του, δοκιμάστηκαν εκατοντάδες χρονοσειρές και συνδυασμοί άρθρων. Τέλος αξίζει να σημειωθεί ότι η έλλειψη ακρίβειας ίσως οφείλεται και στο ότι χρησιμοποιήθηκε η Demo έκδοση του IBM Watson Natural Language Understanding η οποία δεν δίνει καμία πρόσβαση σε ρυθμίσεις και στα features του εργαλείου, έτσι ώστε να μπορούμε να θέσουμε περισσότερους κανόνες για να πετύχουμε μία πιο σταθερή μεθοδολογία.

1ο Βήμα: Αναζήτηση άρθρων μέσω μηχανής αναζήτησης Google

2ο Βήμα: Έλεγχος του αρθρογράφου και του εκδοτικού φορέα για υψηλή δημοτικότητα μέσω του Google Trends για την περίοδο της χρονοσειράς.

3ο Βήμα: Μετάβαση του κειμένου στην μηχανή IBM Watson Natural Language Understanding και ανάλυση.

4ο Βήμα: Επανάληψη της παραπάνω διαδικασίας μέχρι να συγκεντρωθούν αρκετά άρθρα (3 έως 10, ανάλογα την διαθεσιμότητα) και υπολογίζουμε τον Μέσο όρο των αποτελεσμάτων.

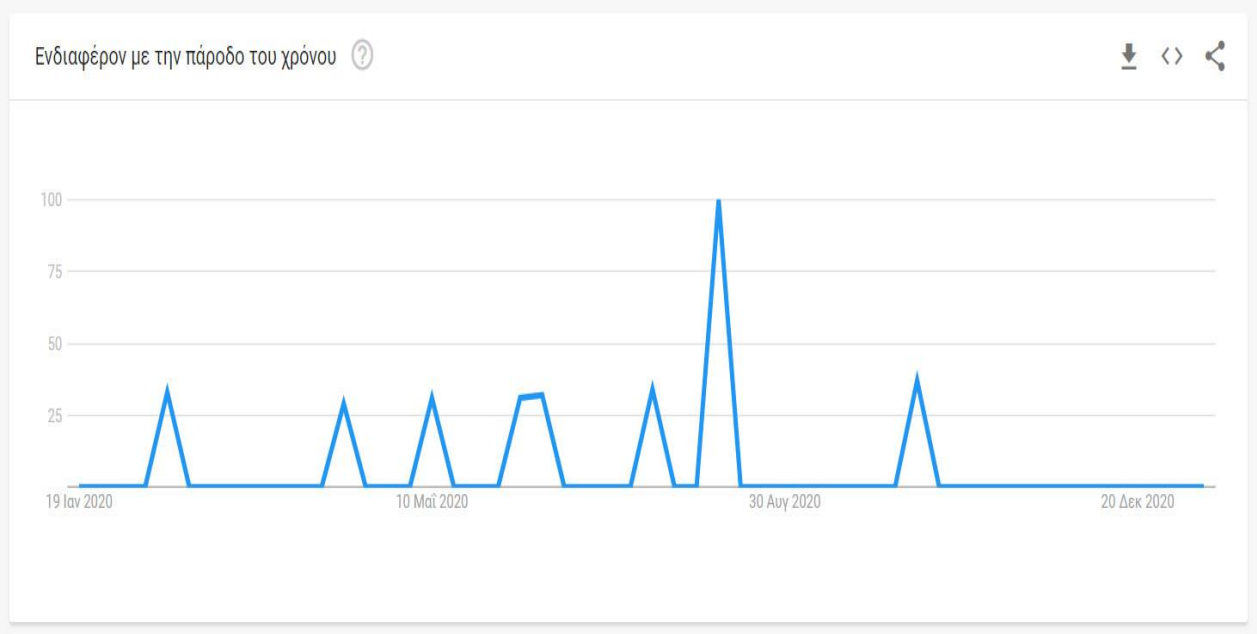
5ο Βήμα: Σύγκριση των αποτελεσμάτων με προηγούμενες τιμές του VIX για να αποφύγουμε τις ακραίες τιμές (outliers).

Η παραπάνω χρονοσειρά δοκιμάζεται αν πληροί τα κριτήρια του VECM. Αν δεν πληροί τα κριτήρια του VECM (π.χ. Johansen cointegration test), η διαδικασία επαναλαμβάνεται με αλλαγές κάποιων άρθρων που δεν έχουμε επιλέξει μέχρι να βρεθεί η χρονοσειρά που θα πληροί τα εν λόγω κριτήρια. Πρόκειται για μία διαδικασία δοκιμών (trial and error process).

Παρακάτω δίνεται ένα ενδεικτικό παράδειγμα ενός σημείου της χρονοσειράς που χρησιμοποιείται στην εν λόγω εργασία. Η χρονική στιγμή που επιλέχθηκε είναι 11/03/2020.

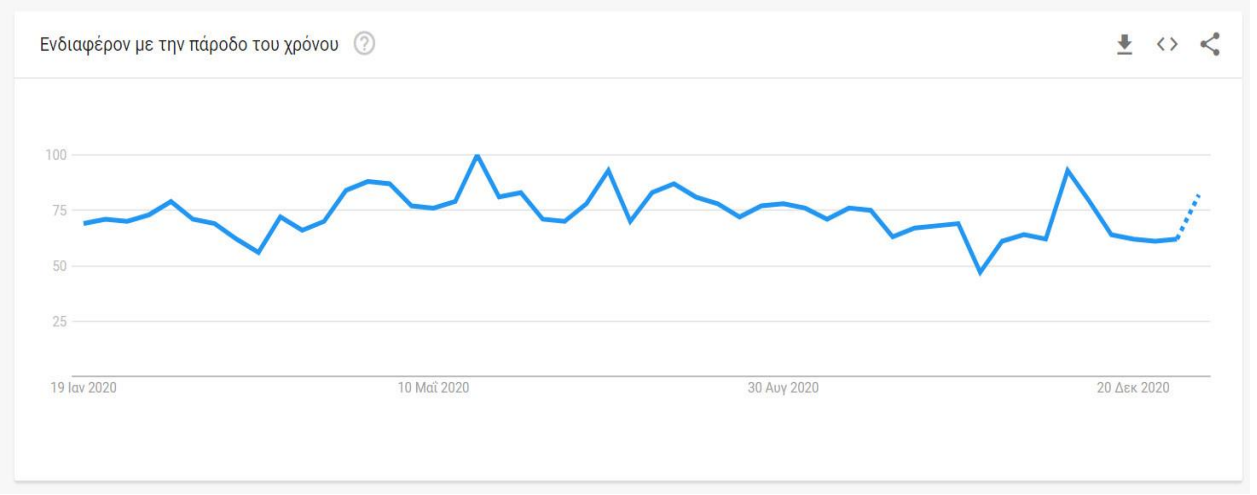
Βήμα 1ο : Έχουμε βρει το άρθρο που μας ενδιαφέρει και έχει τίτλο ¨Bears On The Prowl: Dow Jones Industrial Average Now Down 20% From Highs After Another Rough Day On Street.¨

Βήμα 2ο : Ελέγχουμε την δημοτικότητα του συγγραφέα (JJ Kinahan). Είναι εμφανές από τα αποτελέσματα του google Trends ότι για το διάστημα που μας ενδιαφέρει η δημοτικότητά του είναι υψηλή.



Φωτογραφία 5- Αποτελέσματα Google Trends του συγγραφέα (JJ Kinahan).

Κυρίως μας ενδιαφέρει η δημοτικότητα του εκδοτικού φορέα, που στην προκειμένη περίπτωση είναι το Forbes και η οποία, όπως φαίνεται από τα αποτελέσματα, είναι αρκετά υψηλή.



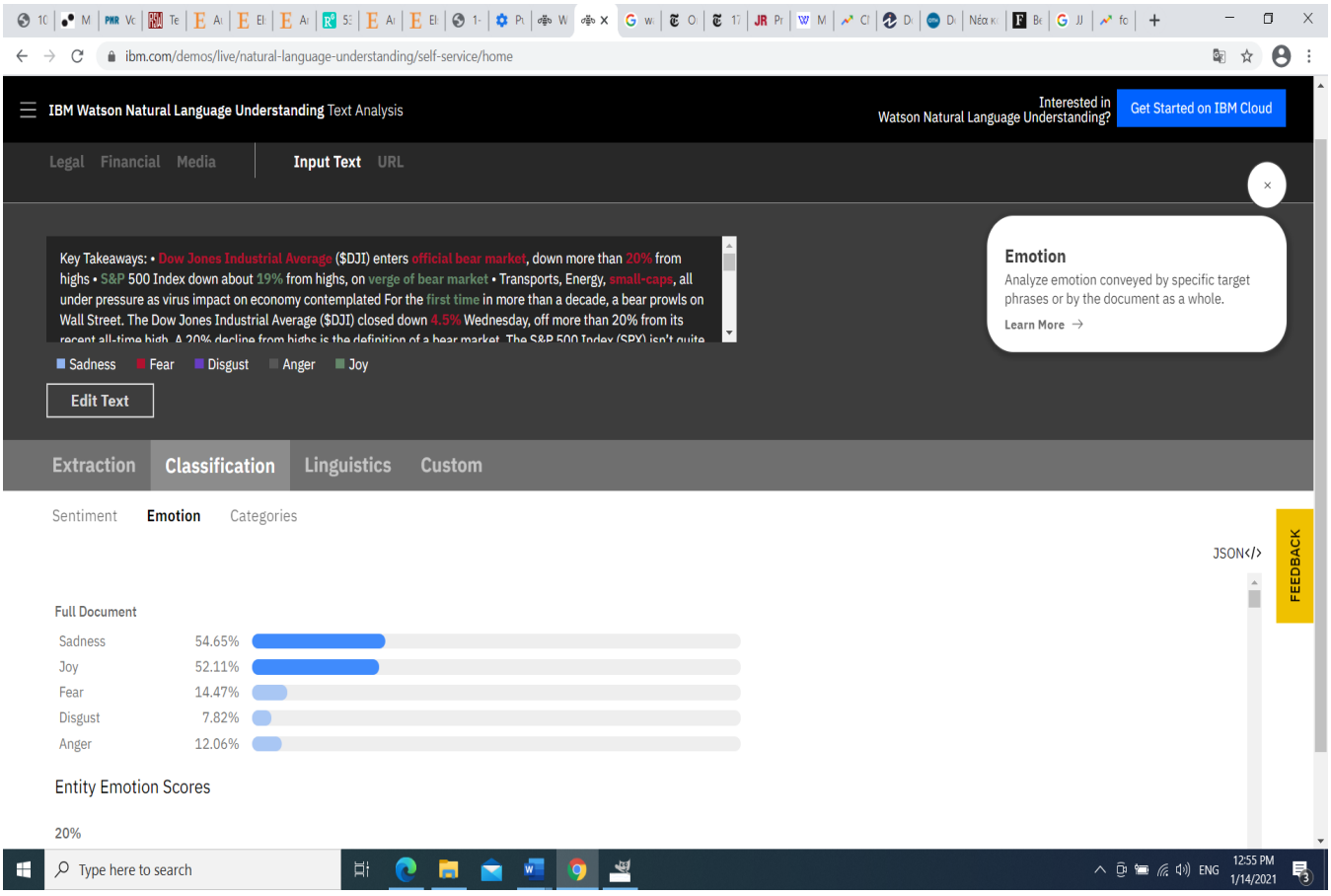
Φωτογραφία 6- Αποτελέσματα Google Trends για το Forbes

Αυτό μας δίνει το πράσινο φως, για να συνεχίσουμε το πείραμα και να πάμε στο 3ο βήμα.

Βήμα 3ο : Αντιγράφουμε το κύριο κείμενο του άρθρου για να αποφύγουμε τις διαφημίσεις και άλλα ανεπιθύμητα μέρη και το εισάγουμε στην IBM Watson Natural Volume Language understanding.

Απόσπασμα από το άρθρο

¨For the first time in more than a decade, a bear prowls on Wall Street. The Dow Jones Industrial Average ($DJI) closed down 4.5% Wednesday, off more than 20% from its recent all-time high. A 20% decline from highs is the definition of a bear market.¨



Φωτογραφία 6- Τα αποτελέσματα του IBM Watson Natural Volume Language understanding σχετικά με το συναίσθημα.

Βήμα 4ο : Επαναλαμβάνουμε την διαδικασία και βγάζουμε τον μέσο όρο των άρθρων.

Sadness	Articles
54.65	https://www.forbes.com/sites/jjkinahan/2020/03/11/bears-on-the-prowl-dow-jones-industrial-average-now-down-20-from-highs-after-another-rough-day-on-street/
62.4	https://www.wsj.com/articles/global-markets-calmer-after-two-hectic-days-11583899913
47.6	https://www.nbcnews.com/business/markets/dow-hits-bear-market-plunging-more-1-600-points-n1155696
54.88	Μέσος Όρος

Πίνακας 1 - Αποτελέσματα του IBM Watson Natural Language Understanding

Σημείωση: Ο VIX στις 11/03/2020 ήταν 53.9

Βήμα 5ο : Εφόσον ολοκληρωθεί η χρονοσειρά, ελέγχεται για correlation με τον DJI και τον VIX.

Προσοχή : Αν η σειρά δεν πληροί τα κριτήρια του VECM, ξαναφτιάχνουμε καινούργια αλλάζοντας κάποια άρθρα και επαναλαμβάνουμε την διαδικασία μέχρι να λειτουργήσει.

4.2 Ο Volatility Index – VIX

Ο VIX είναι ένας δείκτης, όπως ο Down Jones Industrial Average (DJIA), ο οποίος υπολογίζεται σε πραγματικό χρόνο. Πρωτοπαρουσιάστηκε το 1993 από τον Whaley (1993) και είχε ως σκοπό να μετρήσει την μεταβλητότητα σε βραχύ χρονικό διάστημα (short term). Αρχικά υπολογίστηκε για τον S&P 100 και σχεδιάστηκε ως ένας ενεργός δείκτης για την αγορά συμβολαίων δικαιωμάτων προαίρεσης (options). Ως υψηλή μεταβλητότητα ορίζουμε τις ακραίες αλλαγές τιμής ενός χρηματοοικονομικού προϊόντος σε βραχύ χρονικό (short term) διάστημα. Η βασική έννοια είναι ο καθορισμός της σταθεράς β η οποία συγκρίνει την μεταβλητότητα μιας μετοχής έναντι ενός πιο διευρυμένου ορόσημου δείκτη (Whaley, 2009).

$\beta = 1$: Η μεταβλητότητα είναι ίση με τον ορόσημο δείκτη (benchmark index).

$\beta < 1$: Η μεταβλητότητα είναι χαμηλότερη από τον ορόσημο δείκτη.

β > 1 : Η μεταβλητότητα είναι μεγαλύτερη από τον ορόσημο δείκτη.

Βάσει του CBOE η φόρμουλα υπολογισμού του δείκτη είναι :

$$\sigma^2 = \frac{2}{T}\sum_i \frac{\Delta K_i}{K_i^2}\, e^{RT} Q(K_i) - \frac{1}{T}\left[\frac{F}{K_0} - 1\right]^2 (1)$$

4.3 Ο δείκτης Down Jones DJI

Ο δείκτης Down Jones είναι ο βεβαρημένος δείκτης (weighted index) της τιμής των μετοχών των 30 μεγαλύτερων εταιρειών του χρηματιστηρίου NYSE, οι οποίες έχουν δημόσια συναλλαγή (publicly-traded) και θεωρούνται ορόσημο για την συμπεριφορά ολόκληρης της αγοράς. Οι εν λόγω εταιρείες είναι blue chip και χαρακτηρίζονται από μεγάλη κεφαλαιακή επάρκεια (large market capitalizations) και με σταθερό ιστορικό κερδοφορίας.

Η φόρμουλα υπολογισμού του είναι :

DJIA Price = SUM (Component stock prices) / Dow Divisor (2)

Ο δείκτης Dow Division στην φόρμουλα 2 είναι ένας αριθμητικός δείκτης που συμβάλλει στην ιστορική συνεκτικότητα του Down Jones και συντελεί στην ακρίβειά του. Συστάθηκε για πρώτη φορά το 1986 και πήρε το όνομά του από τον Charles Dow, που ήταν και ένας από τους εμπνευστές του Down Jones.

4.4 Το μοντέλο VECM

Σε αυτό το σημείο παρουσιάζεται η μεθοδολογία του μοντέλου Vector Error Correction Model. Ο λόγος που επιλέχθηκε το VECM είναι γιατί μπορεί να υποδηλώσει αν και κατά πόσο μία εξαρτημένη μεταβλητή X_t επηρεάζεται από το ιστορικό της παρελθόν X_{t-n} , καθώς και από μία άλλη επεξηγηματική μεταβλητή (Error Correction Model) ή μεταβλητές (Vector Error Correction Model) στην γενικευμένη του μορφή. Τέλος το VECM είναι σε θέση να μας δώσει τον χρόνο προσαρμογής της εξαρτημένης, τον οποίο ονομάζουμε error correction term.

4.4.1. Η αυτοπαλινδρόμηση και μοντελοποίηση διόρθωσης σφάλματος (Error correction Modeling)

Εφόσον το Vector Autoregresion (VAR) είναι η γενικευμένη μορφή του Autoregrassion (AR), πρώτα δίνεται η επεξήγηση του αυτοπαλίνδρομου μοντέλου (AR) με σκοπό την καλύτερη κατανόηση. Ένα AR μοντέλο είναι μία εξίσωση κατά την οποία η εξαρτημένη μεταβλητή (δηλ. έξοδος του συστήματος) εξαρτάται από τις ιστορικές της τιμές και ένα στοχαστικό μέρος το οποίο είναι γνωστό ως error term ή κατάλοιπα και αντιπροσωπεύει τον λευκό θόρυβο που εμπεριέχεται στην χρονοσειρά.

Βάσει του (Sims, 1980) ένα VAR θεωρείται ένα σύστημα εξισώσεων το οποίο γενικεύει την ιδέα των AR μοντέλων. Περιγράφεται ως η σχέση μεταξύ της εξαρτημένης μεταβλητής η οποία εκτιμάται από τις τιμές της με υστέρηση (δηλ. τρέχουσες και ιστορικές) και τις εναπομείνασες n-1 μεταβλητές. Περαιτέρω θα πρέπει να τονίσουμε ότι το VAR είναι μια στοχαστική διαδικασία παρά μια ντετερμινιστική, κάτι που θα μπορούσε να δημιουργήσει πρόβλημα στην αξιοπιστία των εκτιμώμενων αποτελεσμάτων.

Απλοποιώντας την επεξήγηση του (Sims, 1980) και τον δοθέντα φορμαλισμό, το Var είναι σε θέση να ανακαλύψει και να εντοπίσει την γραμμική δια-εξάρτηση (interdependence) μεταξύ διαφόρων χρονοσειρών καθώς και τις δυναμικές (dynamics) της κάθε μίας ξεχωριστά. Το πλεονέκτημά του είναι πως μπορεί να προσεγγίσει με πιο ολοκληρωμένο τρόπο ένα οικονομετρικό πρόβλημα από ένα AR, διότι έχει την δυνατότητα να εντοπίζει τα χαρακτηριστικά των δεδομένων από γραμμικά συνδεδεμένες χρονοσειρές.

4.4.2 Η έννοια της στατικότητας και τα τέστ μοναδιαίας ρίζας

Υπάρχει ένας διχασμός στην ακαδημαϊκή κοινότητα σχετικά με την ιδέα της στατικότητας Η μία πλευρά υποστηρίζει πως η ιδέα της στατικότητας είναι απαραίτητη για την εφαρμογή παλινδρομήσεων, ειδικά σε μερικές μακροοικονομικές εφαρμογές (Hachemeister, 1975). Η άλλη πλευρά που αποτελείται από τους κλασικούς μελετητές, όπως ο (Granger, Spurious regressions in econometrics., 1974), υποστηρίζει πως η τάση σε μια σειρά δημιουργεί αμφίβολα αποτελέσματα, επειδή υπερισχύει της παλινδρομικής διαδικασίας. Βέβαια, υπάρχουν διάφορες τεχνικές

αφαίρεσης της τάσης, όπως η διαφορά n-τάξης (π.χ. πρώτες, δεύτερες διαφορές) καθώς και οι λογαριθμικές διαφορές.

Η στατικότητα σε μια χρονοσειρά εξασφαλίζεται με τον έλεγχο ύπαρξης μοναδιαίας ρίζας μέσω κάποιου τεστ ελέγχου μοναδιαίας ρίζας. Σε αυτή την εργασία χρησιμοποιείται το επαυξημένο Dickey and Fuller test (Cheung, 1995), διότι είναι και το πιο κατάλληλο στην περίπτωσή μας βάσει της βιβλιογραφίας.

Για να γίνει κατανοητό το επαυξημένο Dickey & Fuller test (Augmented Dickey and Fuller) ο απλός έλεγχος Dickey & Fuller επεξηγείται. Ο εν λόγω έλεγχος έχει μια AR διαδικασία

$X_t = a + \rho X_t - 1 + e_t$ (3)

Αν a=0, τότε έχουμε τυχαίο περίπατο (random walk)
Αν $a \neq 0$, τότε έχουμε μία στοχαστική τάση ή τυχαίο περίπατο με τάση.
Εφόσον η παραπάνω εξίσωση δεν είναι πολύ βολική, την ξαναγράφουμε ως :

$X_t - X_t - 1 = a + (\rho - 1)X_{t-1} + e_t \leftrightarrow (\delta = \rho - 1)\left(\Delta_{x_t} X_t - X_{t-1}\right), \Delta_{x_t} = a + \delta X_{t-1} + e_t (4)$

Η υπόθεση ορίζεται ως :

H0: Η χρονοσειρά δεν είναι στάσιμη, ρ=1,δ=0, $\delta = \delta$ ή $t \geq DF\ critical$, $\Delta X_t = a + 0 + e_t$
H1: Η χρονοσειρά είναι στάσιμη, $\delta = \delta\ or\ t < DF\ critical$, $\Delta X_t = a + \delta\ X_t - 1 + e_t$
Σημείωση : η DF critical προέρχεται από την μορφοποίηση της κατανομής t από τους Dickey and Fuller.

Γενικεύοντας τον έλεγχο Dickey and Fuller, έχουμε το Augmented Dickey and Fuller test που εφαρμόζεται σε οποιαδήποτε AR(i) διαδικασία.

$\Delta y_t = a + \delta y_{t-1} + \sum_{i=1}^{n} \beta_i \Delta y_{t-i} + e_t$ (5)

4.4.3 Cointegration test

Η συνολοκλήρωση των εμπλεκόμενων μεταβλητών είναι απαραίτητη, εφόσον απαιτείται ο εντοπισμός της συγχρονισμένης κίνησης (co-movement) ο οποίος χαρακτηρίζει τις αυξομειώσεις σε βραχύ χρονικό διάστημα (short-term termoils) και οι οποίες διορθώνονται στο μακρύ χρονικό διάστημα (Bahmani-Oskooee, 1994). Η ιδέα προτάθηκε αρχικά από τον (Granger, Spurious

regressions in econometrics, 1981). Υπενθυμίζουμε ότι η συνολοκλήρωση είναι μια στατιστική ιδιότητα των μεταβλητών μιας χρονοσειράς με τάξη ολοκλήρωσης n, η οποία ακολουθείται από έναν γραμμικό συνδυασμό των ομαδοποιημένων μεταβλητών οι οποίες ολοκληρώνονται με τάξη μικρότερη του n. Καλό θα ήταν να λάβουμε υπ' όψη ότι πολλές χρονοσειρές χαρακτηρίζονται από στοχαστικές τάσεις.

Πολλά τεστ συνολοκλήρωσης έχουν αναπτυχθεί, όπως αυτό του κατωφλίου από τους (Balke, 1997). Η εν λόγω εργασία χρησιμοποιεί το Johansen cointegration test (Johansen, 1990). Το τεστ έχει δύο μεθόδους, την trace και αυτή της ιδιοτιμής (Hänninen, 2012).

Η trace έχει :

- Null hypothesis Ho: Ο αριθμός των συνολοκληρωμένων διανυσμάτων r, τα οποία είναι LR < n, LR = 1,2,…,n LR ο μέγιστος αριθμός του r.

- Εναλλακτική H_1: Ο αριθμός των συνολοκληρωμένων διανυσμάτων, που είναι LR = n

Ο τύπος είναι : $LR_{tr}(r/n) = -T * \Sigma_n \log(1-\lambda hat)$ (6)

Η ιδιοτιμή έχει:

- H0 : Το r είναι LR < n , LR = 1,2,…,n-1 (LR η μέγιστη ιδιοτιμή)
- H1: LR = r + 1

4.4.4. Error correction Model

$\Delta y_t = \delta_0 + \delta_1 \Delta X_t - u_t$ (7) Η χρονοσειρά είναι στατική

Στην περίπτωση που το y_t και x_t συνολοκληρώνονται, υπάρχει μία ισοζύγια τιμή η οποία είναι γραμμικός συνδυασμός του $y_E = a + \beta X^E$, οπότε μπορούμε να βγάλουμε ένα λογικό συμπέρασμα για την μακρά σχέση του X και του Y.

Υπάρχει η πιθανότητα το $Y_t = c + \delta$ να είναι διαφορετικό από την τιμή ισοζυγίου (equilibrium value)

$y_t = c + \delta_i X_t - \delta 2 X_{t-1} + \mu y_{t-1} - u_t$ (8)

Η εν λόγω εξίσωση υποδηλώνει ότι το Y_t έχει μία χρονική υστέρηση στις αντιδράσεις του X_t και μ είναι ο βαθμός της. Όπως μπορούμε να δούμε, δεν υπάρχει οικονομικό περιεχόμενο. Αν το X και το Y είναι στάσιμα, έχουμε μια ψευδή (spurious) παλινδρόμηση λόγω του ότι υπάρχει ένα στατιστικά σημαντικό δ_1. Εφόσον θέλουμε να βρούμε οικονομικές σχέσεις χωρίς spurious παλινδρόμηση, χρησιμοποιούμε το μοντέλο error correction.

$$y_t - y_t - 1 = c + \delta_1 X_t - \delta 2 X_{t-1} - (1-\mu) y_{t-1} + \mu y_{t-1} - u_t \text{ (9)}$$

Εφόσον το $\delta_1 X_t$ δεν είναι στάσιμο, τότε παίρνουμε τις πρώτες διαφορές

$$\Delta yt = c + \delta 1 Xt - \delta 1 Xt - 1 + \delta 2 Xt + \delta 2 Xt - 1 - (1-\mu)\, yt - 1 + ut = c + \delta 1 \Delta Xt - \lambda\,(yt - 1 - \alpha - \beta Xt - 1) + ut$$

$$\lambda = 1-\mu\ ,\ \beta = (\delta 1 + \delta 2)/(1-\mu) \text{ (10)}$$

$$\Delta y_t = c + \delta_1 X_t - \delta_1 X_{t-1} + \delta_2 X_t + \delta_1 X_{t-1} - (1-\mu) y_{t-1} + u_t = c + \delta_1 X_{t-\lambda} - \lambda(y_t - 1 - \alpha - \beta X_{t-1}) + u_t \text{ (11)}$$

$$\lambda = 1 - \mu, \beta = (\delta_1 + \delta_2)/(1-\mu)$$

Εφόσον η αλλαγή στο y_t θα είναι σημαντικά αρνητική, διορθώνουμε το σφάλμα της τελευταίας περιόδου και προσαρμόζουμε την τιμή ισοζυγίου του y. Το μοντέλο έχει επίπτωση στην μακροχρόνια (long run) σχέση μεταξύ του Y και του X, επομένως μπορεί να χρησιμοποιηθεί για οικονομετρικές εφαρμογές.

Το Vector Error Correction Model είναι η γενικευμένη μορφή του ECM

$$\text{VECM: } \Delta y_t = \beta_0 + \sum_{i=1}^{n} \beta_i \Delta y_{t-i} + \sum_{i=0}^{n} \delta_i \Delta x_{t-i} + \varphi z_{t-1} + \mu_t \text{ (12)}$$

4.4.5 Το Wald Test

Το Walt test βασίζεται στην βεβαρημένη (weighted) απόσταση μεταξύ των εκτιμώμενων τιμών και των υποθετικών, που είναι και η μηδενική υπόθεση. Με τον όρο βεβαρημένη απόσταση (weighted distance) καθορίζεται η επαλήθευση του τεστ. Μια μεγάλη βεβαρημένη απόσταση υποδεικνύει ότι η μηδενική υπόθεση δεν είναι έγκυρη (Fahrmeir, 2013).

5. Παρουσίαση Δεδομένων, αναλύσεις και ευρήματα

Σ' αυτό το κεφάλαιο παρουσιάζονται τα δεδομένα και οι αναλύσεις τους. Πιο συγκεκριμένα οι τρεις δείκτες (DJI, SADNESS,VIX) είναι το βασικό στοιχείο μελέτης για την χρονική περίοδο 03/02/2020-28/05/2020, δηλ. την χρονική περίοδο κατά την οποία υπήρξε η πρώτη παγκόσμια επιδημιολογική έξαρση του ιού COVID, η οποία ήταν και η κύρια αιτία της τεράστιας χρηματιστηριακής πτώσης και του DJI. Ακολουθεί η βηματική εφαρμογή του μοντέλου VECM χρησιμοποιώντας την μεταβλητή Sadness ως δείκτη μεταβλητότητας (volatility) και τον DJI ως την εξαρτημένη μεταβλητή. Τέλος το VECM εφαρμόζεται και με την μεταβλητή VIX, έτσι ώστε να υπάρχει μέτρο σύγκρισης μεταξύ των δύο δεικτών.

5.1 Οι χρηματιστηριακοί δείκτες

Οι δύο χρηματιστηριακοί δείκτες, που εξετάζονται, είναι πρώτον ο DJI, όπως δίνεται και από τα ιστορικά δεδομένα (πηγή : yahoo finance) και δεύτερον ο δείκτης μεταβλητότητας (volatility) Sadness που δημιουργήσαμε παραπάνω. Υπενθυμίζουμε ότι ο δείκτης μας υποδεικνύει την κοινωνική κυκλοθυμία του χρηματιστηριακού κόσμου ή την συναισθηματική αστάθεια (emotional instability). Τέλος ό δείκτης μεταβλητότητας VIX (πηγή : yahoo finance) χρησιμοποιείται ως μέτρο σύγκρισης. Τα δεδομένα δίνονται στον πίνακα 2.

Πίνακας δεδομένων

Date	DJI	SADNESS	VIX
2/3/2020	28399.81	19.3	17.97
2/4/2020	28807.63	17.05	16.05
2/5/2020	29290.85	16	15.15
2/6/2020	29379.77	15.74	14.96
2/7/2020	29102.51	16.62	15.47
2/10/2020	29276.82	16.01	15.04
2/11/2020	29276.34	16.3	15.18
2/12/2020	29551.42	14.3	13.74
2/13/2020	29423.31	15.25	14.15
2/14/2020	29398.08	14.51	13.68
2/18/2020	29232.19	15.9	14.83
2/19/2020	29348.03	15.41	14.38

2/20/2020	29219.98	16.62	15.56
2/21/2020	28992.41	19.2	17.08
2/24/2020	27960.8	28.21	25.03
2/25/2020	27081.36	30.62	27.85
2/26/2020	26957.59	30.7	27.56
2/27/2020	25766.64	42.31	39.16
2/28/2020	25409.36	45	40.11
3/2/2020	26703.32	38.2	33.42
3/3/2020	25917.41	40.8	36.82
3/4/2020	27090.86	35.01	31.99
3/5/2020	26121.28	44.32	39.62
3/6/2020	25864.78	46.72	41.94
3/9/2020	23851.02	58.61	54.46
3/10/2020	25018.16	52.3	47.3
3/11/2020	23553.22	54.88	53.9
3/12/2020	21200.62	72.3	75.47
3/13/2020	23185.62	61.2	57.83
3/16/2020	20188.52	80	82.69
3/17/2020	21237.38	72.3	75.91
3/18/2020	19898.92	74.1	76.45
3/19/2020	20087.19	70.3	72
3/20/2020	19173.98	65.1	66.04
3/23/2020	18591.93	59.32	61.59
3/24/2020	20704.91	59.84	61.67
3/25/2020	21200.55	61.2	63.95
3/26/2020	22552.17	60.3	61
3/27/2020	21636.78	63.21	65.54
3/30/2020	22327.48	56.2	57.08
3/31/2020	21917.16	52.8	53.54
4/1/2020	20943.51	56.7	57.06
4/2/2020	21413.44	51.1	50.91
4/3/2020	21052.53	44.21	46.8
4/6/2020	22679.99	42.3	45.24
4/7/2020	22653.86	45.1	46.7
4/8/2020	23433.57	41.8	43.35
4/9/2020	23719.37	39.74	41.67
4/13/2020	23390.77	39.52	41.17
4/14/2020	23949.76	35.84	37.76
4/15/2020	23504.35	38.27	40.84
4/16/2020	23537.68	39.2	40.11
4/17/2020	24242.49	36.87	38.15
4/20/2020	23650.44	41.8	43.83
4/21/2020	23018.88	42.11	45.41

4/22/2020	23475.82	39.2	41.98
4/23/2020	23515.26	38.74	41.38
4/24/2020	23775.27	33.86	35.93
4/27/2020	24133.78	31.91	33.29
4/28/2020	24101.55	33.57	33.57
4/29/2020	24633.86	29.4	31.23
4/30/2020	24345.72	32.16	34.15
5/1/2020	23723.69	35.02	37.19
5/4/2020	23749.76	33.45	35.97
5/5/2020	23883.09	31.91	33.61
5/6/2020	23664.64	31.87	34.12
5/7/2020	23875.89	29.92	31.44
5/8/2020	24331.32	25.41	27.98
5/11/2020	24221.99	25.02	27.57
5/12/2020	23764.78	31.47	33.04
5/13/2020	23247.97	33.16	35.28
5/14/2020	23625.34	30.25	32.61
5/15/2020	23685.42	29.92	31.89
5/18/2020	24597.37	27.84	29.3
5/19/2020	24206.86	28.23	30.53
5/20/2020	24575.9	25.86	27.99
5/21/2020	24474.12	26.12	29.53
5/22/2020	24465.16	26.43	28.16
5/26/2020	24995.11	26.29	28.01
5/27/2020	25548.27	26.62	27.62
5/28/2020	25400.64	27.1	28.59

Πίνακας 2 – Τα δεδομένα του του μοντέλου VECM

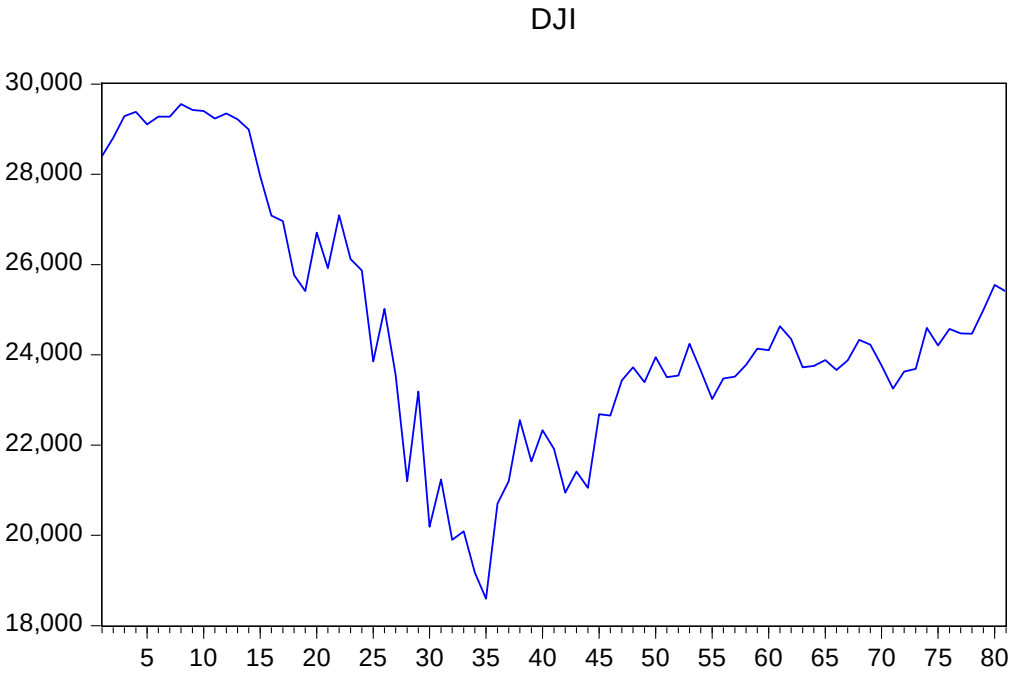
Οι κύριοι εκδοτικοί φορείς που χρησιμοποιήθηκαν για τον κύριο όγκο των άρθρων δίνονται στον πίνακα 3.

Εκδοτικοί φορείς	
CNN Business	https://edition.cnn.com/BUSINESS
CNBC News	https://www.cnbc.com/world/?region=world
Wall Street Journal	https://www.wsj.com/
Forbes	https://www.forbes.com/?sh=16d50cad2254
Barrons	https://www.barrons.com/
Investors Business Daily	https://www.investors.com/
Markets Insider	https://markets.businessinsider.com/

Πίνακας 3 : Οι βασικοί εκδοτικοί φορείς που χρησιμοποιήθηκαν

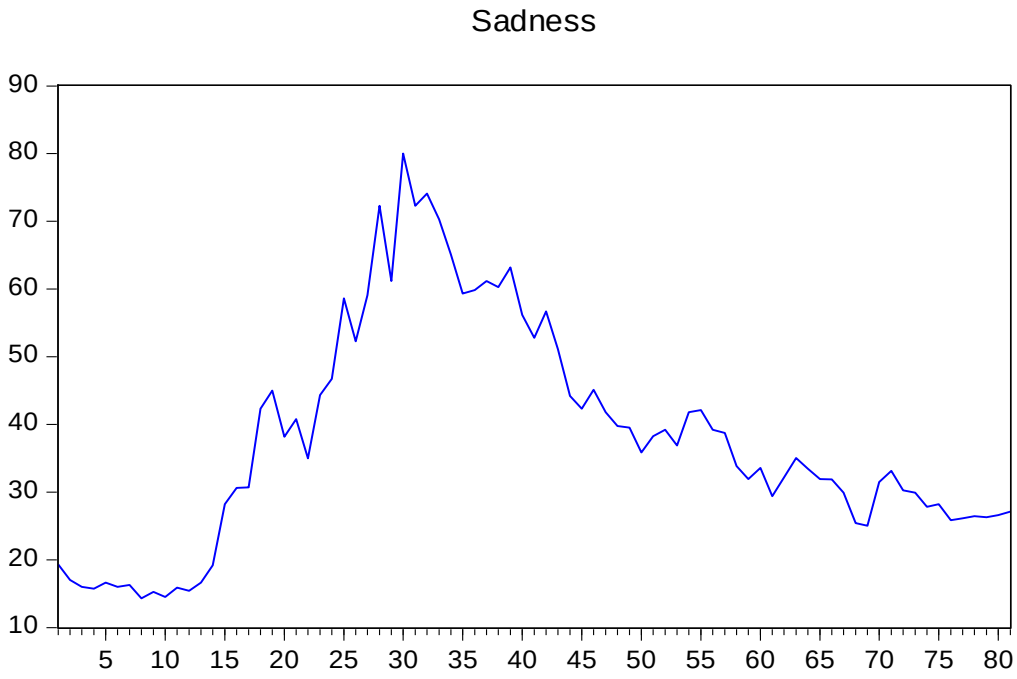
5.2 Διερευνητική μέθοδος ανάλυσης δεδομένων (Exploratory Data Analysis)

Εδώ παρουσιάζεται η γραφική απεικόνιση των εμπλεκόμενων μεταβλητών και δίνονται κάποια εμπειροτεχνικά συμπεράσματα σύγκρισης. Η EDA μέθοδος είναι το πρώτο βήμα σε κάθε ποσοτική ανάλυση που περιέχει δεδομένα.



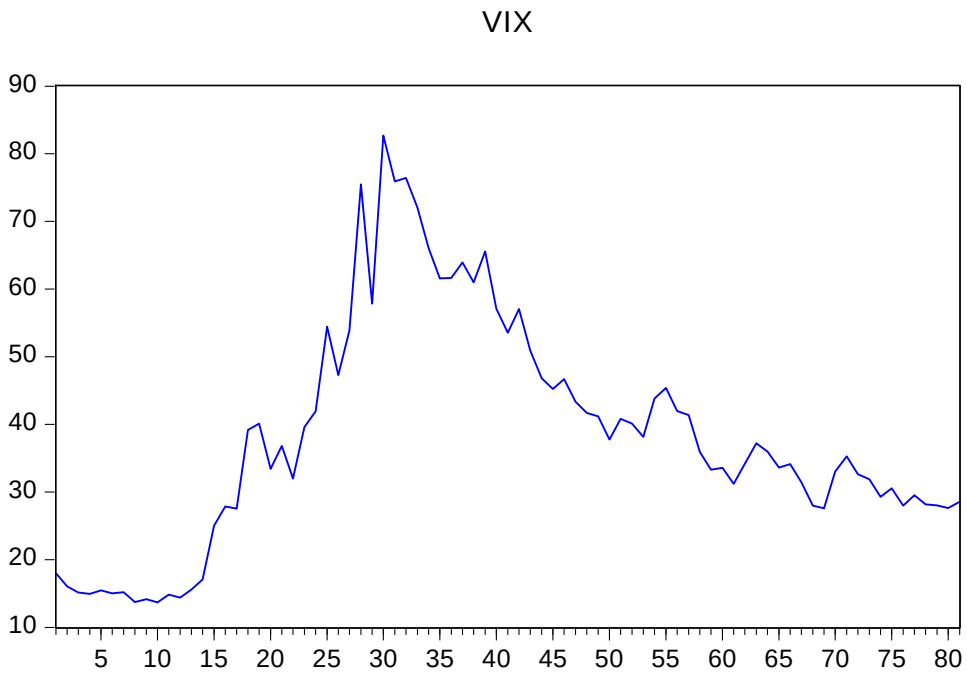
Γράφημα 3 – Ο δείκτης DJI

Στο γράφημα 3 είναι εμφανέστατη η μεγάλη πτώση του Down Jones λόγω του COVID και η σταδιακή αποκατάστασή του σε 80 ημέρες. Μέχρι την 35η ημέρα η πτώση κορυφώνεται.



Γράφημα 4 – Ο δείκτης Sadness δείχνει το συναίσθημα της λύπης που εμπεριέχεται στα επιλεγμένα χρηματιστηριακά άρθρα.

Το ίδιο ανεστραμμένο μοτίβο ακολουθεί ο δείκτης λύπης που δημιουργήσαμε (γράφημα 4), δηλ. έχει μέγιστο στις 28 ημέρες. Είναι εμφανές ότι ο δείκτης λύπης προηγείται του DJI κάτι που υποδεικνύει πώς μπορεί να εξελιχθεί ένα σημαντικό εργαλείο χρηματιστηριακής πρόγνωσης. Η συσχέτιση μεταξύ των δύο δεικτών είναι -0.82 και για τον υπολογισμό της χρησιμοποιήθηκε ο δείκτης pearson correlation και η μέθοδός του.



Γράφημα 5 – Ο δείκτης VIX

Ο VIX (γράφημα 5) κορυφώνεται στις 30 ημέρες και ακολουθεί ακριβώς το μοτίβο του δείκτη Sadness. Είναι εμφανές ότι στόχος μας είναι η χρήση του δείκτη SADNESS ως VIX.

5.3 Το μοντέλο Vector Error correction VECM

Το εν λόγω μοντέλο εφαρμόζεται για να διερευνήσουμε οικονομετρικά την σχέση μεταξύ του DJI και του δείκτη SADNESS που δημιουργήσαμε. Τέλος το μοντέλο VECM θα εφαρμοστεί και με τον VIX για να υπάρχει ένα μέτρο σύγκρισης και αναφοράς σε ένα δείκτη, ο οποίος θεωρείται ορόσημο στην μελέτη της μεταβλητότητας. Εν συντομία τα βήματα της μεθοδολογίας είναι :

- Βήμα 1ο : Διασφάλιση της στασιμότητας της χρονοσειράς
- Βήμα 2ο : Καθορισμός του βέλτιστου μεγέθους υστέρησης (lag length p) που θα χρησιμοποιηθεί
- Βήμα 3ο : Εφαρμογή του τεστ συνολοκλήρωσης (cointegration) Johansen με βάση το βέλτιστο μέγεθος υστέρησης (optimal lag order).
- Βήμα 4ο (i) : Αν δεν υπάρχει εξίσωση συνολοκλήρωσης, τότε θα εφαρμοστεί ένα αυτοπαλίνδρομο μοντέλο τύπου VAR (Vector Autoregression).

- Βήμα 4ο (ii) : Αν υπάρχει εξίσωση συνολοκλήρωσης, τότε θα προχωρήσουμε στην εφαρμογή του VECM για τον βέλτιστο αριθμό υστέρησης.
- Βήμα 5ο : Εφαρμογή των διαγνωστικών τεστ, όπως το Walt test.

5.3.1. Η στασιμότητα των χρονοσειρών DJI, SADNESS, VIX

Για την διασφάλιση της στασιμότητας θα εφαρμοστεί ο επαυξημένος έλεγχος Dickey and Fuller (Augmented Dickey and Fuller test) ο οποίος έχει στατιστικές υποθέσεις H_0 null and alternative H_1:

H_0 : Η χρονοσειρά έχει μοναδιαία ρίζα ($\delta = 0$)
H_1 : Η χρονοσειρά δεν έχει μοναδιαία ρίζα ($\delta \neq 0$)

Στην περίπτωσή μας θα εφαρμόσουμε τον έλεγχο Dickey & Fuller για τις πρώτες διαφορές, διότι υπάρχει τάση στις εν λόγω χρονοσειρές, όπως είναι εμφανές από την EDA. Το επιλεχθέν διάστημα εμπιστοσύνης είναι 95% βασιζόμενο στην κατηγορία t-statistics. Αν δεν διασφαλίσουμε την στασιμότητα της χρονοσειράς, δεν γίνεται να χρησιμοποιηθεί στην εφαρμογή του VECM. Τέλος, όσον αφορά την βέλτιστη υστέρηση, θα χρησιμοποιηθεί το κριτήριο πληροφορίας Schwarz.

5.3.1.1 Η χρονοσειρά του DJI

Null Hypothesis: D(DJI) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=11)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-12.63441	0.0001
Test critical values:	1% level	-3.515536	
	5% level	-2.898623	
	10% level	-2.586605	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DJI,2)
Method: Least Squares
Date: 01/11/21 Time: 20:02
Sample (adjusted): 3 81
Included observations: 79 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(DJI(-1))	-1.347539	0.106656	-12.63441	0.0000
C	-55.67104	90.80818	-0.613062	0.5416

R-squared	0.674595	Mean dependent var	-7.031003
Adjusted R-squared	0.670369	S.D. dependent var	1404.539
S.E. of regression	806.3951	Akaike info criterion	16.24802
Sum squared resid	50071020	Schwarz criterion	16.30800
Log likelihood	-639.7966	Hannan-Quinn criter.	16.27205
F-statistic	159.6282	Durbin-Watson stat	1.855477
Prob(F-statistic)	0.000000		

Πίνακας 3- Έλεγχος μοναδιαίας ρίζας DJI

Όπως είναι εμφανές στον πίνακα 3, η πιθανότητα είναι 0, άρα βρισκόμαστε εντός του διαστήματος εμπιστοσύνης 95%, και αυτό φαίνεται και από την τιμή του t-Statistics (-12.63441). Επομένως η χρονοσειρά μπορεί να χρησιμοποιηθεί.

5.3.1.2 Η χρονοσειρά Sadness

Null Hypothesis: D(SADNESS) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=11)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-11.44682	0.0001
Test critical values:	1% level	-3.515536	
	5% level	-2.898623	
	10% level	-2.586605	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(SADNESS,2)
Method: Least Squares
Date: 01/11/21 Time: 20:09
Sample (adjusted): 3 81
Included observations: 79 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(SADNESS(-1))	-1.258254	0.109922	-11.44682	0.0000
C	0.151145	0.532603	0.283785	0.7773

R-squared	0.629860	Mean dependent var	0.034557
Adjusted R-squared	0.625053	S.D. dependent var	7.729530
S.E. of regression	4.733014	Akaike info criterion	5.971992
Sum squared resid	1724.909	Schwarz criterion	6.031978
Log likelihood	-233.8937	Hannan-Quinn criter.	5.996024
F-statistic	131.0296	Durbin-Watson stat	1.934228
Prob(F-statistic)	0.000000		

Πίνακας 4- Έλεγχος μοναδιαίας ρίζας Sadness

Όπως είναι εμφανές στον πίνακα 4, η πιθανότητα είναι 0.0001, άρα βρισκόμαστε μέσα στο διάστημα εμπιστοσύνης 95%.Η τιμή t-Statistics είναι(-11.44682). Επομένως η χρονοσειρά μπορεί να χρησιμοποιηθεί.

5.3.1.3 . Η χρονοσειρά VIX

Null Hypothesis: D(VIX) has a unit root
Exogenous: Constant
Lag Length: 0 (Automatic - based on SIC, maxlag=11)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-13.01670	0.0001
Test critical values:	1% level	-3.515536	
	5% level	-2.898623	
	10% level	-2.586605	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(VIX,2)
Method: Least Squares
Date: 01/11/21 Time: 20:12
Sample (adjusted): 3 81
Included observations: 79 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(VIX(-1))	-1.374391	0.105587	-13.01670	0.0000
C	0.204467	0.600508	0.340490	0.7344

R-squared	0.687544	Mean dependent var	0.036582
Adjusted R-squared	0.683486	S.D. dependent var	9.484964
S.E. of regression	5.336202	Akaike info criterion	6.211896
Sum squared resid	2192.579	Schwarz criterion	6.271882
Log likelihood	-243.3699	Hannan-Quinn criter.	6.235928
F-statistic	169.4345	Durbin-Watson stat	1.923805
Prob(F-statistic)	0.000000		

Πίνακας 5- Έλεγχος μοναδιαίας ρίζας VIX

Όπως είναι εμφανές στον πίνακα 5, η πιθανότητα είναι 0.0001, άρα είμαστε μέσα στο διάστημα εμπιστοσύνης 95%. Η τιμή t-Statistics είναι (-13.01670). Επομένως η χρονοσειρά μπορεί να χρησιμοποιηθεί.

5.3.2 Η βέλτιστη χρονική υστέρηση (optimal lag length p)

Για να καθορίσουμε την βέλτιστη χρονική υστέρηση, θα εφαρμόσουμε ένα μοντέλο VAR και θα ελέγξουμε τον μέγιστο αριθμό αυτής με την χρήση διαφόρων κριτηρίων.

5.3.2.1 Οι χρονοσειρές DJI και Sadness

Εφαρμόζουμε το VAR με εξαρτημένη την μεταβλητή DJI και έπειτα προχωράμε στην εφαρμογή των κριτηρίων.

VAR Lag Order Selection Criteria
Endogenous variables: DJI SADNESS
Exogenous variables: C
Date: 01/11/21 Time: 20:20
Sample: 1 81
Included observations: 74

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-955.4370	NA	5.93e+08	25.87667	25.93895	25.90152
1	-787.9821	321.3324	7154910.	21.45897	21.64579	21.53350
2	-777.3851	19.76186*	5988414.*	21.28068*	21.59204*	21.40488*
3	-776.0871	2.350404	6446751.	21.35371	21.78961	21.52759
4	-772.7389	5.882027	6569760.	21.37132	21.93177	21.59489
5	-768.4150	7.362321	6525655.	21.36257	22.04756	21.63582
6	-766.4223	3.285217	6909663.	21.41682	22.22636	21.73975
7	-762.5185	6.225080	6955553.	21.41942	22.35350	21.79203

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Πίνακας 6 - Εφαρμογή κριτηρίων

Όπως είναι εμφανές από τα αποτελέσματα του πίνακα 6, έχουμε ομοφωνία στο ότι η βέλτιστη υστέρηση (lag) είναι 2.

5.3.2.2 Οι χρονοσειρές DJI και VIX

VAR Lag Order Selection Criteria
Endogenous variables: DJI VIX
Exogenous variables: C
Date: 01/11/21 Time: 20:28
Sample: 1 81
Included observations: 74

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-944.3511	NA	4.40e+08	25.57706	25.63933	25.60190
1	-796.9336	282.8822	9113276.	21.70091	21.88772	21.77543
2	-785.0380	22.18375	7364420.	21.48751	21.79887*	21.61172*
3	-783.9202	2.024030	7966787.	21.56541	22.00132	21.73930
4	-778.0860	10.24926	7591233.	21.51584	22.07629	21.73941
5	-769.2306	15.07818*	6671096.*	21.38461*	22.06960	21.65786
6	-766.7406	4.105059	6969361.	21.42542	22.23496	21.74836
7	-762.9507	6.043382	7037286.	21.43110	22.36518	21.80372

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Πίνακας 7 - Εφαρμογή κριτηρίων

Στην περίπτωση του DJI και του VIX έχουμε διαφορετικά αποτελέσματα (πίνακας 7). Στην εν λόγω περίπτωση θα επιλέξουμε το μικρότερο lag, όπως αυτό προτείνεται από τα Schwarz information criterion και Hannan-Quinn information criterion.

5.3.3 Εφαρμογή του Johansen cointegration test

Στο εν λόγω σημείο θα εφαρμοστεί το Johansen cointegration test για lag =2 και στις δύο περιπτώσεις.

5.3.3.1 Johansen cointegration test για DJI και Sadness

Date: 01/11/21 Time: 21:24
Sample (adjusted): 4 81
Included observations: 78 after adjustments
Trend assumption: Linear deterministic trend
Series: DJI SADNESS
Lags interval (in first differences): 1 to 2

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.165884	17.50128	15.49471	0.0246
At most 1	0.042082	3.353445	3.841466	0.0671

Trace test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None	0.165884	14.14784	14.26460	0.0522
At most 1	0.042082	3.353445	3.841466	0.0671

Max-eigenvalue test indicates no cointegration at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegrating Coefficients (normalized by b'*S11*b=I):

DJI	SADNESS
0.000748	0.100779
-5.87E-05	-0.072382

Unrestricted Adjustment Coefficients (alpha):

D(DJI)	-304.9647	39.89904
D(SADNESS)	1.593661	0.415551

1 Cointegrating Equation(s): Log likelihood -817.2969

Normalized cointegrating coefficients (standard error in parentheses)

DJI	SADNESS
1.000000	134.6902
	(22.7032)

Adjustment coefficients (standard error in parentheses)

D(DJI)	-0.228182
	(0.06269)
D(SADNESS)	0.001192
	(0.00036)

Πίνακας 8 - Johansen cointegration test για DJI και Sadness

Όπως μπορούμε να δούμε από τον πίνακα 8, έχουμε ένα cointegrating equilibrium μέσα στα όρια εμπιστοσύνης 95% με p-value 2,46%. Επομένως μπορούμε να προχωρήσουμε κανονικά στο VECM για τους δείκτες DJI και Sadness.

5.3.3.2 Johansen cointegration test για DJI και VIX

Date: 01/11/21 Time: 21:40
Sample (adjusted): 4 81
Included observations: 78 after adjustments
Trend assumption: Linear deterministic trend
Series: DJI VIX
Lags interval (in first differences): 1 to 2

Unrestricted Cointegration Rank Test (Trace)

Hypothesized	Trace	0.05

No. of CE(s)	Eigenvalue	Statistic	Critical Value	Prob.**
None *	0.160219	16.12379	15.49471	0.0402
At most 1	0.031591	2.503881	3.841466	0.1136

Trace test indicates 1 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None	0.160219	13.61991	14.26460	0.0630
At most 1	0.031591	2.503881	3.841466	0.1136

Max-eigenvalue test indicates no cointegration at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegrating Coefficients (normalized by b'*S11*b=I):

DJI	VIX
0.000892	0.125918
-7.29E-05	-0.072456

Unrestricted Adjustment Coefficients (alpha):

D(DJI)	-304.2447	-9.242061
D(VIX)	1.517802	0.632329

1 Cointegrating Equation(s): Log likelihood -825.0364

Normalized cointegrating coefficients (standard error in parentheses)

DJI	VIX
1.000000	141.2026
	(18.7436)

Adjustment coefficients (standard error in parentheses)

D(DJI)	-0.271312
	(0.07341)
D(VIX)	0.001354
	(0.00052)

Πίνακας 9 - Johansen cointegration test για DJI και VIX

Όπως μπορούμε να δούμε από τον πίνακα 9, έχουμε ένα cointegrating equilibrium μέσα στα όρια εμπιστοσύνης 95% με οριακό p-value 4.02%. Επομένως μπορούμε να προχωρήσουμε κανονικά στο VECM για τους δείκτες DJI και VIX.

5.4 Η εφαρμογή του VECM

Σ' αυτό το σημείο θα εφαρμόσουμε το VECM και θα επεξηγήσουμε τα αποτελέσματά του. Όπως προαναφέραμε, θα εφαρμόσουμε το VECM για δύο ομάδες χρονοσειρών (DJI-Sadness, DJI-VIX).

5.4.1 VECM για DJI-Sadness

Vector Error Correction Estimates
Date: 01/11/21 Time: 21:47
Sample (adjusted): 4 81
Included observations: 78 after adjustments
Standard errors in ( ) & t-statistics in [ ]

Cointegrating Eq:	CointEq1	
DJI(-1)	1.000000	
SADNESS(-1)	134.6902	
	(22.7032)	
	[5.93265]	
C	-29640.17	
Error Correction:	**D(DJI)**	**D(SADNESS)**
CointEq1	-0.228182	0.001192
	(0.06269)	(0.00036)
	[-3.63968]	[3.29193]
D(DJI(-1))	-0.230683	0.002393
	(0.15794)	(0.00091)
	[-1.46054]	[2.62219]
D(DJI(-2))	0.128144	-0.000820
	(0.16642)	(0.00096)
	[0.77001]	[-0.85317]
D(SADNESS(-1))	30.49225	-0.027731
	(30.0003)	(0.17333)
	[1.01640]	[-0.15998]
D(SADNESS(-2))	5.485368	-0.164663
	(29.3450)	(0.16955)
	[0.18693]	[-0.97118]
C	-58.15321	0.224650
	(84.1610)	(0.48626)
	[-0.69098]	[0.46199]
R-squared	0.304567	0.287452
Adj. R-squared	0.256273	0.237970
Sum sq. resids	39427474	1316.188
S.E. equation	740.0026	4.275557
F-statistic	6.306536	5.809177
Log likelihood	-622.8745	-220.8829

Akaike AIC	16.12499	5.817509
Schwarz SC	16.30627	5.998795
Mean dependent	-49.87447	0.142308
S.D. dependent	858.0777	4.897864

Determinant resid covariance (dof adj.)	5079187.
Determinant resid covariance	4327828.
Log likelihood	-817.2969
Akaike information criterion	21.31530
Schwarz criterion	21.73830

Πίνακας 10 - Αποτελέσματα VECM για DJI και Sadness

Το μοντέλο VECM για την DJI ως εξαρτημένη μεταβλητή (target variable) βάσει των αποτελεσμάτων του πίνακα 10.

$$\Delta DJI_t = -0.228182\ ect_{t-1} - 0.230683\Delta DJI_{t-1} + 0.128144DJI_{t-2} + 3049225\Delta Sadness_{t-1} + 5.485368\Delta Sadness_{t-2} - 58.15321$$

Σημείωση : ο όρος φ είναι ο ect_{t-1} (error correction term) ο οποίος υποδηλώνει την ταχύτητα προσαρμογής (speed adjustment). Το ύψος της διόρθωσης είναι 2,22% για το long run equilibrium.

Η εξίσωση συνολοκλήρωσης (cointegration equation)

$$ect_{t-1} = 1.0000DJI_{t-1} + 134.6902Sadness_{t-1} - 29640.17$$

Η σχέση του δείκτη Sadness με τον δείκτη DJI είναι εμφανής και ξεκάθαρη στην παραπάνω εξίσωση για την μακρά χρονική περίοδο (long-run period). Το ερώτημα όμως είναι τι γίνεται με την άμεση χρονική περίοδο (short-run relationship). Για αυτό τον λόγο θα προχωρήσουμε στο διαγνωστικό έλεγχο Walt.

Dependent Variable: D(DJI)
Method: Least Squares (Gauss-Newton / Marquardt steps)
Date: 01/11/21 Time: 22:53
Sample (adjusted): 4 81
Included observations: 78 after adjustments
D(DJI) = C(1)*(DJI(-1) + 134.690234215*SADNESS(-1) - 29640.1674209) + C(2)*D(DJI(-1)) + C(3)*D(DJI(-2)) + C(4) *D(SADNESS(-1)) + C(5)*D(SADNESS(-2)) + C(6)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-0.228182	0.062693	-3.639684	0.0005
C(2)	-0.230683	0.157944	-1.460543	0.1485
C(3)	0.128144	0.166418	0.770015	0.4438
C(4)	30.49225	30.00033	1.016397	0.3128
C(5)	5.485368	29.34502	0.186927	0.8522
C(6)	-58.15321	84.16098	-0.690976	0.4918

R-squared	0.304567	Mean dependent var	-49.87447

Adjusted R-squared	0.256273	S.D. dependent var	858.0777
S.E. of regression	740.0026	Akaike info criterion	16.12499
Sum squared resid	39427474	Schwarz criterion	16.30627
Log likelihood	-622.8745	Hannan-Quinn criter.	16.19756
F-statistic	6.306536	Durbin-Watson stat	2.055767
Prob(F-statistic)	0.000066		

Πίνακας 11 - Εκτίμηση εξίσωσης VECM με LS μέθοδο

Η μεταβλητή φ (C1=-0,228182 στον πίνακα 11) έχει οικονομική ερμηνεία λόγω της στατιστικής σημαντικότητας. Το αρνητικό πρόσημο δίνει την φορά της διόρθωσης στο VECM και υποδηλώνει ότι υπάρχει αιτιότητα (negative long run causality) από την επεξηγηματική μεταβλητή (Sadness) ως προς την εξαρτημένη (DJI). Με απλά λόγια, όπως είδαμε και από την cointegration equation η Sadness, έχει αρνητική σχέση με τον DJI. Θα πρέπει να σημειωθεί ότι οι σταθερές θετικού πρόσημου για τον όρο φ δεν είναι θεμιτό αποτέλεσμα, γιατί δεν έχουμε σύγκλιση σε βάθος χρόνου (long run converge).

Σ' αυτό το στάδιο μας ενδιαφέρουν οι μεταβλητές C(4) και C(5) για το Walt test, διότι αφορούν την μεταβλητή Sadness και μπορούμε να δούμε αν έχουν κάποιου είδους αιτιότητα σε βραχύ χρονικό διάστημα (short run) στην εξαρτημένη, που είναι ο DJI. Άρα για το Walt test έχουμε C(4)=C(5)=0.

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	0.537153	(2, 72)	0.5867
Chi-square	1.074306	2	0.5844

Null Hypothesis: C(4)=C(5)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(4)	30.49225	30.00033
C(5)	5.485368	29.34502

Restrictions are linear in coefficients.

Πίνακας 12 - Αποτελέσματα Walt test

Είναι εμφανές ότι λόγω της p-value για την κατανομή Chi-Square (p-value = 0.5844 δείτε πίνακα 12) δεν μπορούμε να αποκλείσουμε την μηδενική υπόθεση, επομένως δεν υπάρχουν στοιχεία αιτιότητας για βραχύ χρονικό διάστημα (short-run causality C(4)=C(5)=0) από την Sadness στον

DJI. Για να επιβεβαιώσουμε την σταθερότητα του μοντέλου μας θα προχωρήσουμε σε διαγνωστικό τεστ καταλοίπων LM (residual diagnostics Breusch-Godfrey test) για την ύπαρξη συσχέτισης (serial correlation).

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	2.103382	Prob. F(2,70)	0.1297
Obs*R-squared	4.421801	Prob. Chi-Square(2)	0.1096

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 01/11/21 Time: 23:27
Sample: 4 81
Included observations: 78
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.083105	0.159722	0.520309	0.6045
C(2)	0.469544	0.466465	1.006600	0.3176
C(3)	-0.117820	0.301045	-0.391369	0.6967
C(4)	-19.03538	35.24973	-0.540015	0.5909
C(5)	-19.59565	36.00580	-0.544236	0.5880
C(6)	21.36188	88.93616	0.240193	0.8109
RESID(-1)	-0.605922	0.607203	-0.997890	0.3218
RESID(-2)	0.375285	0.260524	1.440499	0.1542

R-squared	0.056690	Mean dependent var	3.32E-14
Adjusted R-squared	-0.037641	S.D. dependent var	715.5733
S.E. of regression	728.9164	Akaike info criterion	16.11791
Sum squared resid	37192340	Schwarz criterion	16.35962
Log likelihood	-620.5985	Hannan-Quinn criter.	16.21467
F-statistic	0.600966	Durbin-Watson stat	1.971125
Prob(F-statistic)	0.753069		

Πίνακας 13 - Αποτελέσματα Godfrey test

Η μηδενική υπόθεση, πως δεν υπάρχει συσχέτιση, δεν μπορεί να απορριφθεί, γιατί βάσει των αποτελεσμάτων (πίνακας 13) για την κατανομή Chi-Square η p-value = 0.10 >0.05.

5.4.2 VECM για DJI-VIX

Ό,τι εφαρμόσαμε για την ομάδα DJI-Sadness θα το εφαρμόσουμε και για την ομάδα DJI-VIX για να έχουμε ένα μέτρο σύγκρισης και να δούμε αν όντως η χρονοσειρά Sadness συμπεριφέρεται σαν τον VIX.

Vector Error Correction Estimates
Date: 01/12/21 Time: 01:18
Sample (adjusted): 4 81
Included observations: 78 after adjustments
Standard errors in ( ) & t-statistics in [ ]

Cointegrating Eq:	CointEq1	
DJI(-1)	1.000000	
VIX(-1)	141.2026 (18.7436) [7.53338]	
C	-29948.31	
Error Correction:	**D(DJI)**	**D(VIX)**
CointEq1	-0.271312 (0.07341) [-3.69601]	0.001354 (0.00052) [2.58974]
D(DJI(-1))	-0.043770 (0.16314) [-0.26829]	0.001272 (0.00116) [1.09529]
D(DJI(-2))	0.065211 (0.16563) [0.39373]	0.000255 (0.00118) [0.21610]
D(VIX(-1))	63.20381 (26.0211) [2.42894]	-0.331320 (0.18527) [-1.78835]
D(VIX(-2))	-2.460880 (26.3721) [-0.09331]	-0.021805 (0.18776) [-0.11613]
C	-57.91547 (82.6571) [-0.70067]	0.288546 (0.58850) [0.49030]
R-squared	0.328784	0.243345
Adj. R-squared	0.282172	0.190799
Sum sq. resids	38054493	1929.058
S.E. equation	727.0039	5.176144
F-statistic	7.053614	4.631129
Log likelihood	-621.4922	-235.7922
Akaike AIC	16.08954	6.199801
Schwarz SC	16.27083	6.381087

Mean dependent	-49.87447	0.172308
S.D. dependent	858.0777	5.754111
Determinant resid covariance (dof adj.)		6194121.
Determinant resid covariance		5277831.
Log likelihood		-825.0364
Akaike information criterion		21.51375
Schwarz criterion		21.93675

Πίνακας 14 – Αποτελέσματα μοντέλου VIX και DJI

Το μοντέλο VECM για την DJI ως εξαρτημένη μεταβλητή (target variable) βάσει των αποτελεσμάτων του πίνακα 14.

$$\Delta DJI_t = -0.271312\ ect_{t-1} - 0.043770\Delta DJI_{t-1} + 0.065211 DJI_{t-2} + 63.20381\Delta VIX_{t-1} - 2.460880\Delta VIX_{t-2} - 57.91547$$

Σημείωση : ο όρος φ είναι ο ect_{t-1} (error correction term) ο οποίος υποδηλώνει την ταχύτητα προσαρμογής (speed adjustment). Το ύψος της διόρθωσης είναι 2,27% για το long run equilibrium.

Η εξίσωση συνολοκλήρωσης (cointegration equation)

$$ect_{t-1} = 1.0000 DJI_{t-1} + 141.2026 VIX_{t-1} - 29948.31$$

Η σχέση του δείκτη VIX με τον δείκτη DJI είναι εμφανής και ξεκάθαρη στην παραπάνω εξίσωση για την μακρά χρονική περίοδο (long-run period). Σ' αυτή την φάση θα εξετάσουμε την βραχεία χρονική σχέση (short run) του VIX με τον DJI με την χρήση του Walt test.

Dependent Variable: D(DJI)
Method: Least Squares (Gauss-Newton / Marquardt steps)
Date: 01/12/21 Time: 01:28
Sample (adjusted): 4 81
Included observations: 78 after adjustments
D(DJI) = C(1)*(DJI(-1) + 141.202639712*VIX(-1) - 29948.3145347) + C(2)*D(DJI(-1)) + C(3)*D(DJI(-2)) + C(4)*D(VIX(-1)) + C(5)*D(VIX(-2)) + C(6)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-0.271312	0.073407	-3.696014	0.0004
C(2)	-0.043770	0.163144	-0.268291	0.7892
C(3)	0.065211	0.165625	0.393729	0.6949
C(4)	63.20381	26.02114	2.428941	0.0176
C(5)	-2.460880	26.37205	-0.093314	0.9259
C(6)	-57.91547	82.65710	-0.700671	0.4858
R-squared	0.328784	Mean dependent var		-49.87447

Adjusted R-squared	0.282172	S.D. dependent var	858.0777
S.E. of regression	727.0039	Akaike info criterion	16.08954
Sum squared resid	38054493	Schwarz criterion	16.27083
Log likelihood	-621.4922	Hannan-Quinn criter.	16.16212
F-statistic	7.053614	Durbin-Watson stat	2.050309
Prob(F-statistic)	0.000020		

Πίνακας 15 – Εκτίμηση εξίσωσης VECM με LS μέθοδο

Και σ' αυτή την περίπτωση η μεταβλητή φ (C1=-0,271312 στον πίνακα 15) έχει οικονομική ερμηνεία λόγω της στατιστικής σημαντικότητας. Το αρνητικό πρόσημο δίνει την φορά της διόρθωσης στο VECM και υποδηλώνει ότι υπάρχει αιτιότητα (negative long run causality) από την επεξηγηματική μεταβλητή (VIX) ως προς την εξαρτημένη (DJI).

Σ' αυτό το στάδιο, όπως και με την Sadness, μας ενδιαφέρουν οι μεταβλητές C(4) και C(5) για το Walt test, διότι αφορούν την μεταβλητή VIX και μπορούμε να δούμε αν έχουν κάποιου είδους αιτιότητα σε βραχύ χρονικό διάστημα (short run) στην εξαρτημένη, που είναι ο DJI. Άρα για το Walt έχουμε C(4)=C(5)=0.

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	3.009756	(2, 72)	0.0555
Chi-square	6.019511	2	0.0493

Null Hypothesis: C(4)=C(5)=0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(4)	63.20381	26.02114
C(5)	-2.460880	26.37205

Restrictions are linear in coefficients.

Πίνακας 16 - Αποτελέσματα Walt test

Σε αντίθεση με την χρονοσειρά της Sadness, στην περίπτωση του VIX μπορούμε να απορρίψουμε την μηδενική υπόθεση, εφόσον η p-value της Chi-square είναι 4,93%<5%, επομένως έχουμε στοιχεία αιτιότητας για βραχύ χρονικό διάστημα (short-run causality C(4)=C(5)=0) από την VIX στον DJI. Όπως και στην προηγούμενη περίπτωση, θα εκτελέσουμε και πάλι το Breusch-Godfrey test για την συσχέτιση.

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	1.882596	Prob. F(2,70)	0.1598
Obs*R-squared	3.981350	Prob. Chi-Square(2)	0.1366

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 01/12/21 Time: 01:37
Sample: 4 81
Included observations: 78
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.158508	0.201990	0.784730	0.4353
C(2)	0.521970	0.478295	1.091313	0.2789
C(3)	-0.143201	0.222131	-0.644669	0.5212
C(4)	-32.27039	38.32485	-0.842023	0.4026
C(5)	-44.78308	51.51351	-0.869346	0.3876
C(6)	28.76550	88.33046	0.325658	0.7457
RESID(-1)	-0.754879	0.669683	-1.127219	0.2635
RESID(-2)	0.298478	0.233802	1.276629	0.2060

R-squared	0.051043	Mean dependent var	1.38E-14
Adjusted R-squared	-0.043853	S.D. dependent var	703.0037
S.E. of regression	718.2526	Akaike info criterion	16.08843
Sum squared resid	36112080	Schwarz criterion	16.33015
Log likelihood	-619.4489	Hannan-Quinn criter.	16.18520
F-statistic	0.537885	Durbin-Watson stat	2.018808
Prob(F-statistic)	0.802899		

Πίνακας 17 - Αποτελέσματα Godfrey test

Η μηδενική υπόθεση, πως δεν υπάρχει συσχέτιση, δεν μπορεί να απορριφθεί, γιατί βάσει των αποτελεσμάτων (πίνακας 17) για την κατανομή Chi-Square η p-value = 0.13 >0.05.

6. Συμπεράσματα

Εφόσον δημιουργήσαμε την χρονοσειρά Sadness χρησιμοποιώντας εργαλεία με τεχνολογία επεξεργασίας φυσικών γλωσσών και συναισθηματικής ανάλυσης, ελέγξαμε μέσω εργαλείων, όπως το Google Trends, ότι τα άρθρα των αρθρογράφων για τον Down Jones και οι εκδοτικοί ηλεκτρονικοί φορείς έχουν μεγάλη αποδοχή από το ευρύτερο κοινό. Επίσης διασφαλίσαμε ότι υπάρχει υψηλή συσχέτιση της Sadness με τον δείκτη DJI και τον VIX και στην συνέχεια προχωρήσαμε στην εκτέλεση ενός VECM για δύο ομάδες χρονοσειρών (DJI-Sadness & DJI-VIX). Το αποτέλεσμα έδειξε ότι ο Sadness συμπεριφέρεται ακριβώς όπως ο VIX στο μοντέλο VECM με ελάχιστες διαφορές και έχει άμεση σχέση και επίδραση στην εξαρτημένη μεταβλητή DJI. Επομένως ο Sadness μπορεί να χρησιμοποιηθεί ως ένας συναισθηματικός δείκτης μεταβλητότητας (Sentiment Volatility Index) όπως και ο κλασικός VIX. Γίνεται άμεσα αντιληπτό ότι η συναισθηματική διάσταση των άρθρων μπορεί να ποσοτικοποιηθεί και συσχετίζεται με τους χρηματιστηριακούς δείκτες Down Jones και VIX. Τέλος είναι εμφανές ότι η συναισθηματική διάσταση του τύπου καθορίζει σε μεγάλο βαθμό την κρίση της κοινής γνώμης των επενδυτών, η οποία με την σειρά της επηρεάζει σημαντικά την μεταβλητότητα του Down Jones. Για τους προαναφερθέντες λόγους η σύγχρονη χρηματοοικονομική κοινότητα πρέπει να προχωρήσει σε περαιτέρω διερεύνηση των επενδυτικών συναισθημάτων και ποσοτικοποίηση αυτών χρησιμοποιώντας σύγχρονα εργαλεία συναισθηματικής ανάλυσης.

7. Αναφορές

Alvim, V. (2010). Sentiment of Financial News: A Natural Language Processing Approach. *1st Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology*, (σ. 16). Buenos Aires .

Azzi, B. (2019). The FinSBD-2019 Shared Task: Sentence boundary detection in PDF Noisy text. *Proceedings of the First Workshop on Financial Technology and Natural Language Processing.*

Bahmani-Oskooee. (1994). hort-run versus long-run effects of devaluation: error-correction modeling and cointegration. *Eastern Economic Journal*, 453-464.

Balke, F. (1997). Threshold cointegration. International economic review. *International economic review*, 627-645.

Bishop. (2006). *Pattern Recognition and Machine Learning.* Springer Science+Business Media.

Brown, E. m.-F. (2013). IBM Research report. *IBM Research Division* .

Butler, K. (2009). Financial Forecasting Using Character N-Gram Analysis and Readability Scores of Annual Reports. *Canadian Conference on Artificial Intelligence*, 39-51.

Cambria, P. G. (2017). Sentiment Analysis Is a Big Suitcase. *IEEE Intelligent Systems* , 74-80.

Cheung, L. (1995). Lag order and critical values of the augmented Dickey–Fuller test. *Journal of Business & Economic Statistics*, 277-280.

Chowdhary. (2020). Natural language processing. *undamentals of Artificial Intelligence*.

Fahrmeir. (2013). Regression. *Models, Methods and Applications*.

Granger. (1974). Spurious regressions in econometrics. *A Companion of Theoretical Econometrics*, 557-61.

Granger. (1981). Spurious regressions in econometrics. *A Companion of Theoretical Econometrics*, 557-61.

Groth, M. (2011). An intraday market risk management approach based on textual analysis. *Decision Support Systems*, 680-691.

Hachemeister. (1975). Credibility for regression models with application to trend. *redibility: Theory and Applications*, 307-48.

Hagenau, l. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*.

Hänninen. (2012). Modern Time Series Analysis in Forest Products Markets. *Springer*.

Johansen, J. (1990). Maximum likelihood estimation and inference on cointegration—with applications to the demand for money. *Oxford Bulletin of Economics and statistics*, 169-210.

Kahneman, T. (1979). Prospect theory: An analysis of decision under risk. *Handbook of the fundamentals of financial decision making*, 99-127.

Khant, M. (2018). Analysis of Financial News Using Natural Language Processing and Artificial Intelligence. *International Conference On Business Innovation*, (σ. 176). Colombo.

Lin, T. C. (2011). A Behavioral Framework for Securities Risk. *SSRN*, 54.

Luccioni, P. (2019). Using natural language processing to analyze financial climate disclosures. *Proceedings of the 36th International Conference on Machine Learning.* California.

Lui, M. (2011). ON THE PREDICTABILITY OF THE U.S. ELECTIONS. *wellesley*.

Mao, Q. (2015). Experimental Studies of Human Behavior in Social Computing. *Doctoral dissertation, Harvard University*.

Mishev, G. (2020). Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE Access*, 21.

Ramon, M. P. (2019). Counterfactual explanation algorithms for behavioral and textual data. *arXiv preprint arXiv:1912.01819*.

Sims. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, 1-48.

Sun, B. C. (2014). Pre-processingOnline Financial Text for Sentiment Classification: A Natural Language Processing Approach. *In 2014 IEEE Conference on Computational Intelligence for Financial Engineering & Economics*.

Sun, R. (2008). The Cambridge Handbook of Computational Psychology. *Cambridge University Press* .

Turing. (1950). COMPUTING MACHINERY AND INTELLIGENCE. *Mind 49*.

Whaley. (2009). Understanding VIX. *The Journal of Portfolio Management*.

Xing, C. (2018). Natural Language Based Financial Forecasting: A Survey. *Artificial Intelligence Review*, 49-73.

Zhang, J. (2010). Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 43-52.