ΣΤΡΑΤΙΩΤΙΚΗ ΣΧΟΛΗ ΕΥΕΛΠΙΔΩΝ
Τμήμα Στρατιωτικών Επιστημών

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
Σχολή Μηχανικών Παραγωγής & Διοίκησης

ΔΙΙΔΡΥΜΑΤΙΚΟ ΔΙΑΤΜΗΜΑΤΙΚΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«Εφαρμοσμένη Επιχειρησιακή Έρευνα & Ανάλυση»

Μελέτη Δεδομένων Κοινωνικών Δικτύων

Νικόλαος Κρυστάλλης

Αθήνα, Ιανουάριος 2021

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κο Νικόλαο Δούκα για την επίβλεψή του και τις συμβουλές του, οι οποίες συνέβαλαν καθοριστικά στην ολοκλήρωση της παρούσας διπλωματικής εργασίας. Ευχαριστώ θερμά όλους τους καθηγητές μου για τις γνώσεις που μου μετέδωσαν.
Οφείλω, επίσης, ένα μεγάλο ευχαριστώ στην οικογένεια μου που είναι πάντα δίπλα μου και σε όλους τους φίλους μου.

Η Μεταπτυχιακή Διατριβή του Νικολάου Κρυστάλλη εγκρίνεται:
12/2/2021

Τριμελής Εξεταστική Επιτροπή

Αναπλ. Καθηγητής **Δούκας Νικόλαος** (ΣΣΕ) (Επιβλέπων)

Καθηγητής Νικόλαος Ματσατσίνης (Πολυτεχνείο Κρήτης) .....

Αναπλ. Καθηγητής Νικόλαος Μπάρδης (ΣΣΕ)

Στην οικογένεια μου

Μελέτη Δεδομένων Κοινωνικών Δικτύων

Σημαντικοί όροι: Κοινωνικά Δίκτυα, Ανάλυση Συναισθήματος, Twitter

Περίληψη

Η χρήση και δημιουργία εργαλείων για την επεξεργασία φυσικής γλώσσας έχει επιταχυνθεί τα τελευταία χρόνια καθώς ο όγκος των δεδομένων που μπορεί να συλλεχθεί είναι σημαντικά μεγαλύτερος ενώ και οι δυνατότητες για επεξεργασία σε υλικοτεχνικό και λογισμικό επίπεδο έχουν αυξηθεί.

Η παρούσα εργασία παραθέτει όλη τη διαδικασία που ακολουθήθηκε για την επίτευξη της συναισθηματικής ανάλυσης. Αρχικά από τη διασύνδεση με την κοινωνική πλατφόρμα, τη συλλογή και κανονικοποίηση των δεδομένων και την εξαγωγή και οπτικοποίηση των αποτελεσμάτων. Ο κώδικας που χρησιμοποιήθηκε για την πραγματοποίηση των παραπάνω διαδικασιών είναι αναρτημένη στο Διαδίκτυο.

Η παρούσα εργασία εκπονήθηκε έχοντας ως στόχο την ανάλυση δεδομένων που προέρχονται από το κοινωνικό δίκτυο Twitter και κυρίως την ανάλυση συναισθήματος (Sentiment Analysis) σύμφωνα με τα δεδομένα του συγκεκριμένου κοινωνικού δικτύου. Αφορά το πεδίο της Επεξεργασίας Φυσικής Γλώσσας εφόσον πραγματοποιείται ανάλυση γραπτών κειμένων με στόχο την εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τους χρήστες που τα δημιούργησαν. Η βασική ανάλυση που διεξάχθηκε στο πλαίσιο της εργασίας είναι η συναισθηματική ανάλυση βασισμένη σε συναισθηματικό λεξικό για την Ελληνική γλώσσα.

Study of Social Media Data

Keywords: Social Media, Sentiment Analysis, Twitter

Abstract

The use and creation of tools for natural language processing has accelerated in recent years as the amount of data that can be collected is significantly greater and the potential for processing at hardware and software level has increased.

This dissertation lists the entire process followed to achieve emotional analysis. Initially from interconnection with the social platform, the collection and normalization of data and the extraction and visualization of results. The code used to carry out the above procedures is posted on the Internet.

The purpose of this dissertation is to analyze data from the social network Twitter and to analyse especially Sentiment Analysis according to the data of the specific social network. It concerns the field of Natural Language Processing if written texts are analyzed in order to draw useful conclusions about the users who created them. The basic analysis carried out in the context of the work is the emotional analysis based on an emotional dictionary for the Greek language.

Περιεχόμενα

Εισαγωγή

Τα τελευταία χρόνια, η δημοτικότητα των κοινωνικών δικτύων έχει αυξηθεί δραματικά, με όλο και περισσότερους χρήστες να μοιράζονται τις στιγμές τους αλλά και πληροφορίες μέσα από διαφορετικές πλατφόρμες. Η υπηρεσία κοινωνικής δικτύωσης microblogging (SNS), έχει γίνει μια δημοφιλής πλατφόρμα επικοινωνίας για εκατομμύρια χρήστες παγκοσμίως.

Οι Εταιρείες χρησιμοποιούν τα κοινωνικά δίκτυα για να προωθήσουν τα προϊόντα τους, επαγγελματίες συντηρούν δημόσια προφίλ για να δικτυώνονται με περισσότερους πελάτες ή και συναδέλφους τους, ενώ και απλοί χρήστες συζητούν για οποιοδήποτε θέμα τους ενδιαφέρει. Αυτή η δημοτικότητα έχει κάνει το Twitter έναν καθολικό μηχανισμό μέσω του οποίου το ψηφιακό περιεχόμενο μπορεί να παραδοθεί και να αναδιανεμηθεί από δεκάδες χρήστες σε δευτερόλεπτα. Οι συνέπειες μιας τέτοιας διάχυσης μπορεί να είναι τεράστιες. Ως εκ τούτου, λόγω της ευρείας χρήσης του, το Twitter έχει γίνει βασικό ζήτημα για πολλούς μελετητές, δημόσιους υπεύθυνους λήψης αποφάσεων υπεύθυνους, χάραξης πολιτικής, και παγκόσμιους διαχειριστές επιχειρήσεων (Aladwani, 2015).

Σκοπός της παρούσας εργασίας είναι η προσέγγιση της πλατφόρμας κοινωνικής δικτύωσης του Twitter, μέσω συγκομιδής δεδομένων και εφαρμογής τεχνικών επεξεργασίας φυσικής γλώσσας (Natural Language Processing) προκειμένου να πραγματοποιηθεί ανάλυση συναισθήματος. Τα βήματα τα οποία ακολουθούνται είναι:

- Η εγγραφή και Είσοδος στην πλατφόρμα ως προγραμματιστής
- Η αποστολή ερωτημάτων στους εξυπηρετητές (servers) με συγκεκριμένη δομή τα οποία είναι εφοδιασμένα με τα απαραίτητα διαπιστευτήρια.
- Η άντληση και η αποθήκευση των Δεδομένων
- Ο καθαρισμός και η Επεξεργασία Δεδομένων
- Η Εφαρμογή τεχνικών Επεξεργασίας Φυσικής Γλώσσας και η
- Οπτικοποίηση των Συμπερασμάτων

Η διπλωματική εργασία διαρθρώνεται ως εξής: στο πρώτο κεφάλαιο αναλύεται η έννοια των κοινωνικών δικτύων, όπως είναι του Twitter, η έννοια της ανάλυση συναισθήματος και πραγματοποιείται μια ιστορική αναδρομή. Στο δεύτερο κεφάλαιο αναφέρονται οι στόχοι της εργασία και η θεωρητική ανάλυση των δεδομένων, όπως είναι η συλλογή και η αποθήκευση δεδομένων. Στο τρίτο κεφάλαιο γίνεται η ανάλυση των Τεχνικών Ταιριάσματος Νήματος και τέλος στο τέταρτο κεφάλαιο έχουμε την παρουσίαση των αποτελεσμάτων και την αξιολόγησή τους.

ΚΕΦΑΛΑΙΟ 1

1.1 Κοινωνικά Δίκτυα - Twitter

Τα τελευταία χρόνια τα κοινωνικά δίκτυα έχουν επιφέρει σημαντικές αλλαγές στον τρόπο με τον οποίο ενημερώνονται αλλά και επικοινωνούν οι άνθρωποι. Για τον λόγο αυτό δίνεται ιδιαίτερη έμφαση στην ανάπτυξη των κοινωνικών δικτύων και στην αξιοποίηση των δεδομένων που παράγονται μέσα από αυτά.

Το μεγαλύτερο ενδιαφέρον το έχει συγκεντρώσει το Twitter[1], το οποίο είναι ένα σύστημα microblogging που επιτρέπει την αποστολή και λήψη σύντομων δημοσιεύσεων που ονομάζονται tweets. Δημιουργήθηκε το Μάρτιο του 2006 και μέσα σε μικρό χρονικό διάστημα προσέλκυσε πολύ μεγάλο αριθμό χρηστών. Πλέον εξυπηρετεί περισσότερους από 310 εκατομμύρια ενεργούς χρήστες το μήνα.

Τα tweets μπορούν να έχουν μήκος έως 280 χαρακτήρες και μπορούν να περιλαμβάνουν συνδέσμους προς σχετικούς ιστότοπους. Αυτό ωθεί τους χρήστες να γράφουν πιο περιεκτικά μηνύματα. Επίσης είναι δυνατό να πραγματοποιούνται συζητήσεις ανάμεσα σε πολλούς χρήστες, οι οποίες είναι ορατές σε όσους τους ακολουθούν. Οι χρήστες του Twitter ακολουθούν άλλους χρήστες. Κάποιος μπορεί να επιλέξει να ακολουθήσει άτομα και οργανισμούς με παρόμοια ακαδημαϊκά και προσωπικά ενδιαφέροντα. Το Twitter είναι το πιο δημοφιλές μέσο microblogging με περισσότερους από 300 εκατομμύρια χρήστες και γίνεται όλο και πιο δημοφιλές στους ακαδημαϊκούς καθώς και στους μαθητές, στους υπεύθυνους χάραξης πολιτικής, στους πολιτικούς και στο ευρύ κοινό.[2]

Πλέον, η διείσδυση του μέσου είναι πολύ μεγάλη στην κοινωνία και είναι ένα πολύ σημαντικό εργαλείο για τις εταιρίες αλλά και τα πολιτικά κόμματα που

[1] https://twitter.com

[2] https://esrc.ukri.org/research/impact-toolkit/social-media/twitter/what-is-twitter/

δραστηριοποιούνται στο Twitter για να προσεγγίσουν μεγάλο κομμάτι του πληθυσμού και να βελτιώσουν την εικόνα τους. Πολλοί υποψήφιοι, κόμματα, δημοσιογράφοι και ένα σταθερά αυξανόμενο μερίδιο του κοινού χρησιμοποιούν το Twitter για να σχολιάσουν, να αλληλοεπιδράσουν και να ερευνήσουν τις δημόσιες αντιδράσεις στην πολιτική (Jungherr, 2016). Ένα χαρακτηριστικό παράδειγμα είναι οι προεδρικές εκλογές στις Η.Π.Α. το 2016, όπου η ημέρα των Εκλογών ήταν μια υπενθύμιση της επιρροής του Twitter στα μέσα ενημέρωσης και της διανομής πληροφοριών και σύμφωνα με στοιχεία μέχρι τις 10 μ.μ. είχαν σταλεί 40 εκατομμύρια μηνύματα σχετικά με τις εκλογές, υπερβαίνοντας τα 31 εκατομμύρια που στάλθηκαν την Ημέρα των Εκλογών 2012 (Isaac and Sydney, 2016).

Η γρήγορη φύση των tweets σημαίνει ότι το Twitter χρησιμοποιείται ευρέως από χρήστες smartphone που δεν θέλουν να διαβάσουν κείμενα μεγάλου περιεχομένου στην οθόνη. Επιπλέον, το Twitter επιτρέπει την προσέγγιση μεγάλου αριθμού ατόμων γρήγορα μέσω tweets και retweets, την ενημέρωση για τις τελευταίες ειδήσεις και την εύκολη προώθηση της έρευνά και της δουλειάς άλλων εμπειρογνωμόνων στον ίδιο τομέα. Για παράδειγμα κάποιος μπορεί να συμμετέχει σε συζητήσεις για εκδηλώσεις, όπως είναι ένα συνέδριο που δεν μπορεί να παρακολουθήσει αυτοπροσώπως.

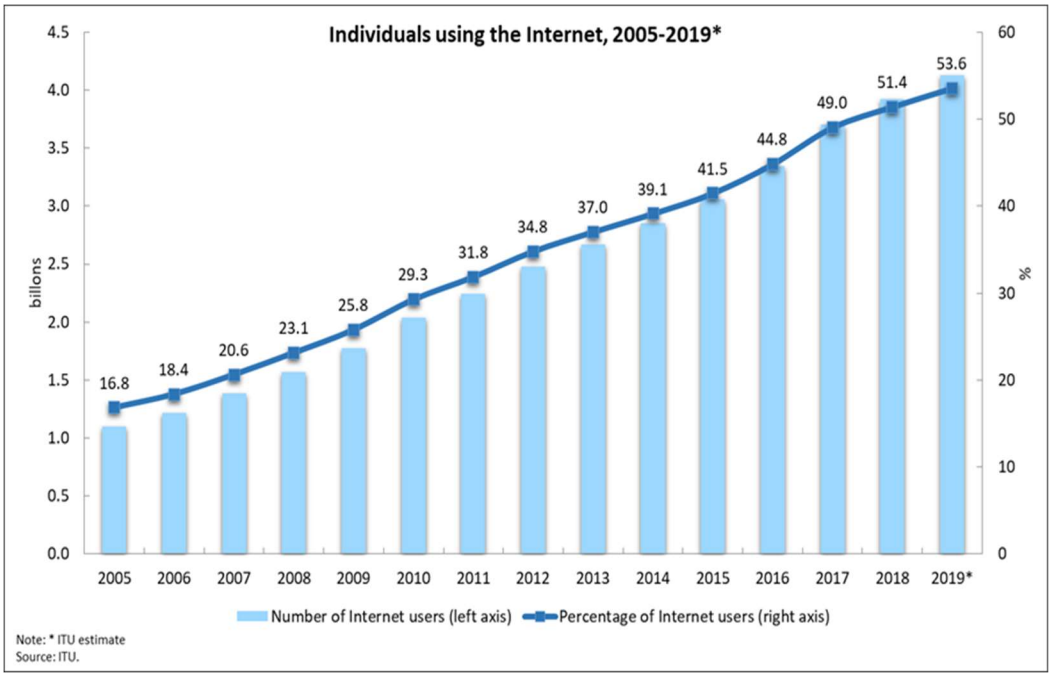
Εικόνα 1: Λογότυπο Twitter

Πηγή: https://en.wikipedia.org/wiki/Twitter

1.2 Στατιστικά Στοιχεία Κοινωνικών Δικτύων - Twitter

Το διαδίκτυο αποτελεί, σήμερα, μια παγκόσμια πηγή πληροφόρησης για εκατομμύρια χρήστες και εκτιμάτε ότι στο τέλος του 2019, το 53,6% του παγκόσμιου πληθυσμού, ή 4,1 δισεκατομμύρια άνθρωποι, χρησιμοποιούσαν το Internet για κάθε μορφής δραστηριότητα όπως είναι η εργασία, η εκπαίδευση ή οι μετακινήσεις (Πίνακας 1).

Πίνακας 1 : Χρήση Internet 2005 -2019



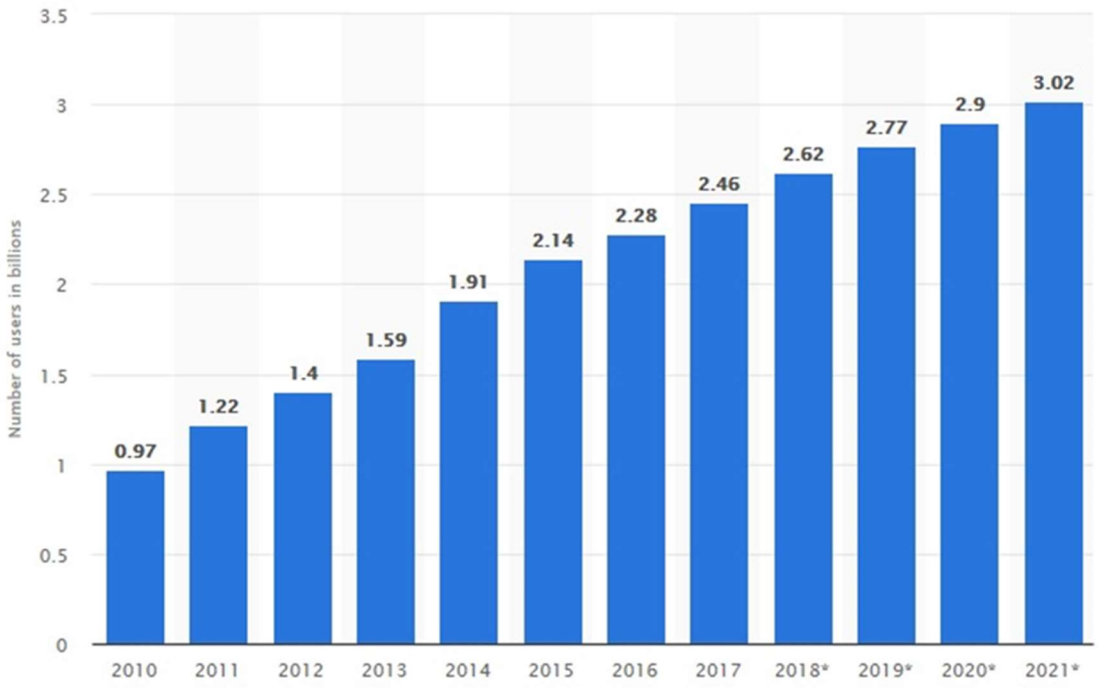
Πηγή:

https://public.tableau.com/profile/ituint#!/vizhome/ITUIndividualsusingInternetv2/IndividualsusingInternet2005-2019

Αντίστοιχα, η χρήση κοινωνικών μέσων είναι μια από τις πιο δημοφιλείς διαδικτυακές δραστηριότητες. Το 2018, περίπου 2,62 δισεκατομμύρια άνθρωποι χρησιμοποίησαν τα κοινωνικά μέσα παγκοσμίως, ένας αριθμός που προβλέπεται να αυξηθεί σε περίπου 3,02 δισεκατομμύρια το 2021. Στην πραγματικότητα, το

μεγαλύτερο μέρος της παγκόσμιας ανάπτυξης των κοινωνικών μέσων οφείλεται στην αυξανόμενη χρήση των κινητών συσκευών (Πίνακας 2).

Πίνακας 2: Πλήθος χρηστών Κοινωνικών Δικτύων παγκοσμίως

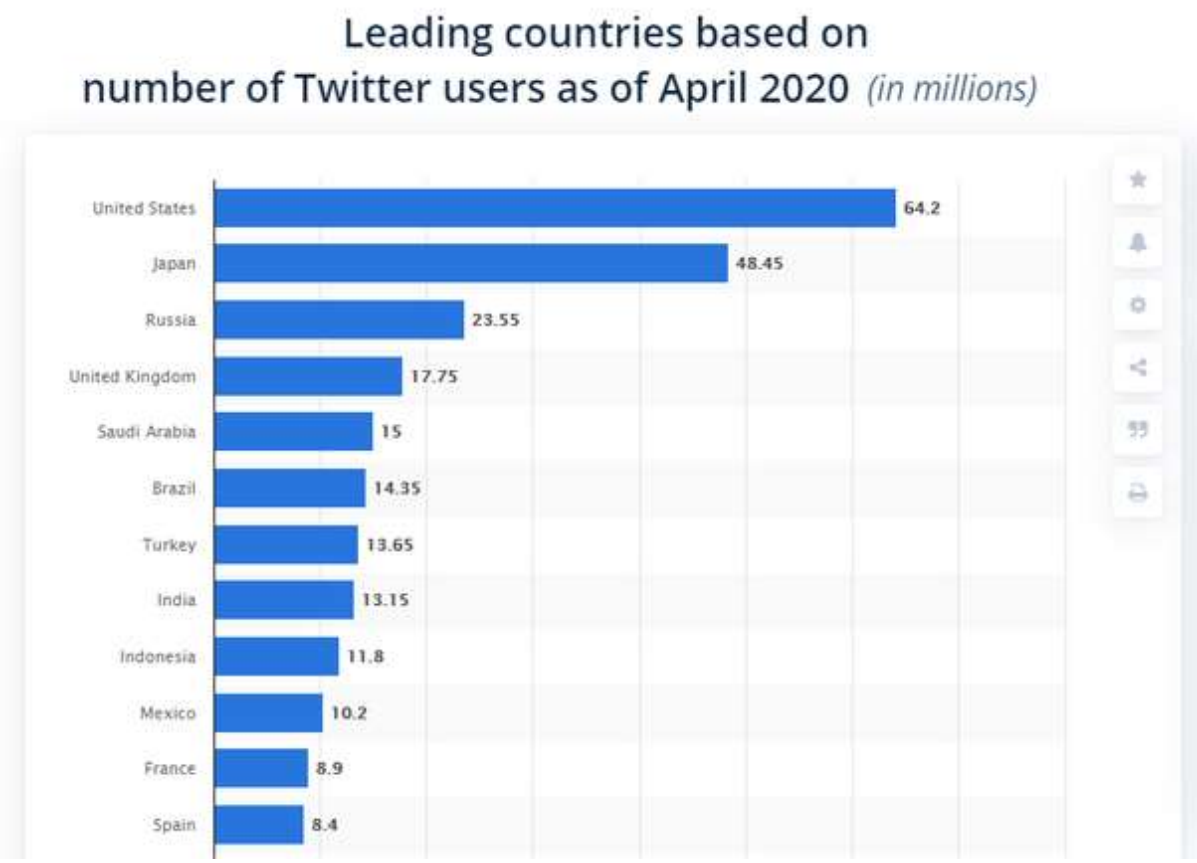


Πηγή: https://ec.europa.eu/knowledge4policy/visualisation/number-social-media-users-worldwide-2010-17-forecasts-2021_en

Όσον αφορά, το κοινωνικό δίκτυο Twitter είναι ιδιαίτερα δημοφιλές στις Ηνωμένες Πολιτείες, όπου από τον Απρίλιο του 2020, η υπηρεσία microblogging είχε προσέγγιση χρηστών 64,2 εκατομμυρίων χρηστών. Η Ιαπωνία και η Ρωσία κατατάχθηκαν στην δεύτερη και τρίτη θέση με 48,45 και 23,55 εκατομμύρια χρήστες αντίστοιχα. Από το τέταρτο τρίμηνο του 2019, το Twitter είχε 152 εκατομμύρια καθημερινά ενεργούς χρήστες με δυνατότητα δημιουργίας εσόδων παγκοσμίως. Επίσης, υπολογίζεται ότι ο μέσος χρόνος ανά ημέρα που ξοδεύεται στα μέσα κοινωνικής δικτύωσης ποικίλλει σημαντικά ανά χώρα και για παράδειγμα

οι χρήστες στις Η.Π.Α. θεωρείται ότι αφιέρωσαν περίπου μία ώρα και 57 λεπτά χρησιμοποιώντας κοινωνικά μέσα κάθε μέρα (Πίνακας 3).

Πίνακας 3: Οι χώρες με τους περισσότερους χρήστες Twitter το 2020

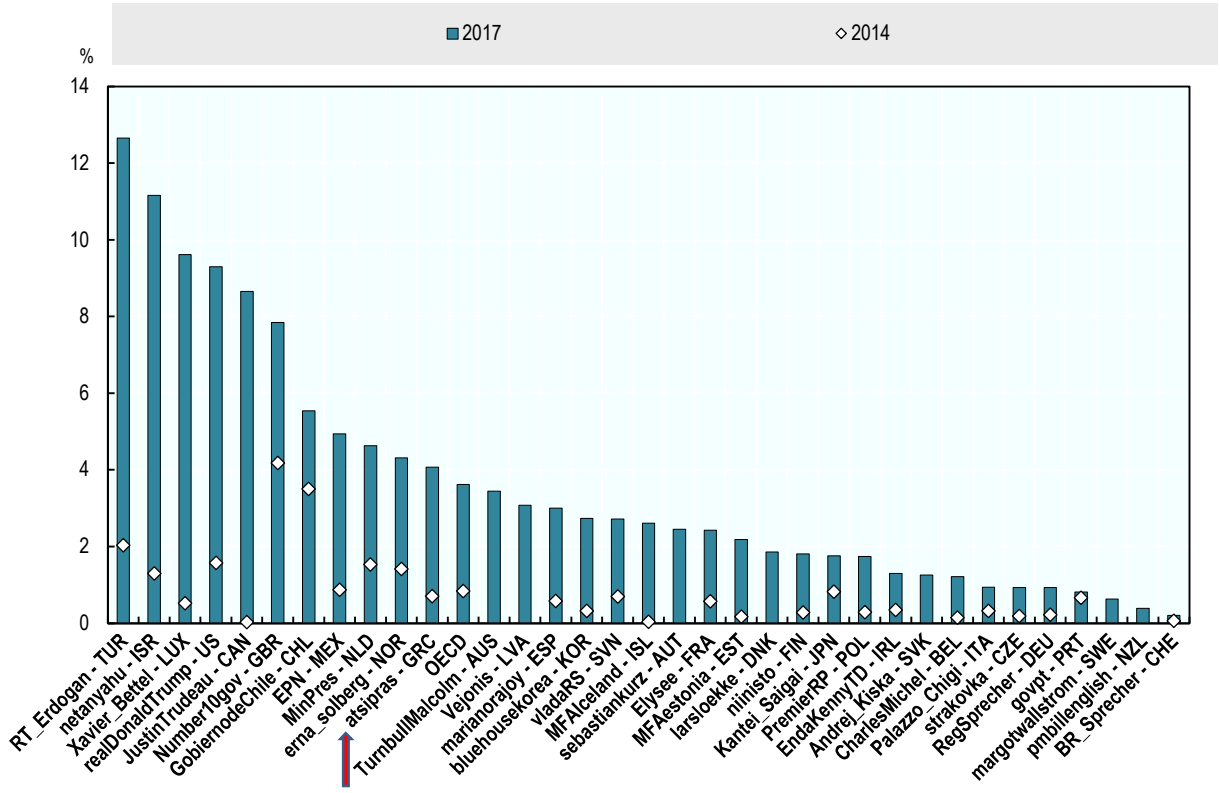


Πηγή: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Το Twitter έχει γίνει όλο και πιο σημαντικό εργαλείο και στην εσωτερική και διεθνή πολιτική και αποτελεί πλέον ένα ευρέως χρησιμοποιούμενο εργαλείο από κυβερνητικούς αξιωματούχους προκειμένου να επικοινωνούν απευθείας με τους πολίτες. Η αύξηση της χρήσης του από κυβερνήσεις και η χρήση αυτού του είδους η επικοινωνία από τους πολίτες υπήρξε αξιοσημείωτη τα τελευταία χρόνια. Πλέον, έχει γίνει ένας τρόπος προώθησης πολιτικών και αλληλεπίδρασης με πολίτες και άλλους αξιωματούχους, και οι περισσότεροι παγκόσμιοι ηγέτες και τα υπουργεία

εξωτερικών έχουν έναν επίσημο λογαριασμό Twitter. Ο τωρινός πρόεδρος των ΗΠΑ Ντόναλντ Τραμπ είναι γνωστός ως χρήστης του Twitter. Ο αριθμός των οπαδών ως μερίδιο του συνολικού πληθυσμού αυξήθηκε σημαντικά σε όλες σχεδόν τις χώρες και στο παρακάτω διάγραμμα παρουσιάζονται οι κυβερνητικοί αξιωματούχοι με τον υψηλότερο αριθμό σε ακολούθους (Followers), τον Μάιο του 2017 (Πίνακας 4).

Πίνακας 4: Οι περισσότεροι κυβερνητικοί αξιωματούχοι που ακολουθήθηκαν στο Twitter, 2017



Πηγή: OECD Digital Economy Outlook, 2017, OECD (2017), "Most followed government officials on Twitter, 2017: Followers as a percentage of total population", in *Trends*, OECD Publishing, Paris, https://doi.org/10.1787/9789264276284-graph97-en.

Χαρακτηριστικό παράδειγμα παρακολούθησης μέσων κοινωνικής δικτύωσης

Το βράδυ της 9ης Απριλίου, 2017, οι United Airlines[3] αφαίρεσαν βίαια έναν επιβάτη από μια πτήση με υπερβολικές κρατήσεις. Το περιστατικό με τον επιβάτη μαγνητοσκοπήθηκε από άλλους επιβάτες στα κινητά τηλέφωνα τους και δημοσιεύτηκε αμέσως. Ένα τέτοιο βίντεο, δημοσιεύτηκε στο Facebook [4] , μοιράστηκε περισσότερες από 87.000 φορές και προβλήθηκε 6.800.000 φορές μέχρι τις 6 μ.μ. τη Δευτέρα, μόλις 24 ώρες αργότερα.

Το φιάσκο μεγεθύνθηκε από την απορριπτική αντίδραση της εταιρείας. Τη Δευτέρα το απόγευμα, κοινοποίησαν στο Twitter μια δήλωση από τον διευθύνοντα σύμβουλο προκειμένου να απολογηθεί για το "ότι πρέπει να ξανά φιλοξενήσει πελάτες".

Ακριβώς το είδος της καταστροφής των δημοσίων σχέσεων που δεν επιθυμεί μια εταιρεία. Αποτελεί επίσης ένα εξαιρετικό παράδειγμα του γιατί ενδιαφέρει όχι μόνο για το αν οι άνθρωποι μιλούν για μια Εταιρεία ή ένα Προϊόν, αλλά το πώς μιλούν γι' αυτό. Περισσότερες αναφορές δεν είναι ίσες με θετικές αναφορές.

Στη σύγχρονη εποχή, τα brands όλων των σχημάτων και μεγεθών έχουν σημαντικές αλληλεπιδράσεις με τους πελάτες, τους υποψήφιους χρήστες, ακόμα και τον ανταγωνισμό τους σε κοινωνικά δίκτυα όπως το Facebook, το Twitter και το Instagram[5]. Τα περισσότερα τμήματα μάρκετινγκ έχουν ήδη συντονιστεί σε ηλεκτρονικές αναφορές όσον αφορά τον όγκο – μετρούν περισσότερες συνομιλίες ως περισσότερη αναγνωρισιμότητα του brand. Σήμερα, όμως, υπάρχει η δυνατότητα να πραγματοποιηθεί ένα βήμα βαθύτερα. Χρησιμοποιώντας ανάλυση

[3] https://en.wikipedia.org/wiki/United_Express_Flight_3411_incident

[4] https://www.facebook.com

[5] https://www.instagram.com

συναισθήματος στα μέσα κοινωνικής δικτύωσης, αντλούνται απίστευτες πληροφορίες για την ποιότητα της συνομιλίας που συμβαίνει γύρω από ένα brand.

1.3 Η Ανάλυση Συναισθήματος

Συναισθηματική Ανάλυση

Η ανάλυση συναισθημάτων είναι η αυτοματοποιημένη διαδικασία ανάλυσης δεδομένων κειμένου και ταξινόμησης γνωμών ως αρνητικών, θετικών ή ουδέτερων [6] . Οι επιχειρήσεις μπορούν να χρησιμοποιήσουν την ανάλυση συναισθημάτων για να παρακολουθήσουν τη φήμη της επωνυμίας στα μέσα κοινωνικής δικτύωσης και να κατανοήσουν τι αρέσει στους πελάτες και τι αντιπαθούν σχετικά με το προϊόν ή την υπηρεσία τους.
Συνήθως, εκτός από την αναγνώριση της γνώμης, αυτά τα συστήματα εξάγουν χαρακτηριστικά της έκφρασης, όπως είναι:

- Πόλωση: Εάν ο ομιλητής εκφράζει θετική ή αρνητική γνώμη,
- Θέμα στο οποίο μπορεί να αναφέρεται ο χρήστης.
- Υποκείμενο: το πρόσωπο ή η οντότητα που εκφράζει τη γνώμη.

Η ανάλυση συναισθημάτων είναι ένα θέμα μεγάλου ενδιαφέροντος και ανάπτυξης, καθώς έχει πολλές πρακτικές εφαρμογές. Δεδομένου ότι οι δημόσιες και ιδιωτικές διαθέσιμες πληροφορίες μέσω του Διαδικτύου αυξάνονται συνεχώς, ένας μεγάλος αριθμός κειμένων που εκφράζουν γνώμες είναι διαθέσιμα σε τοποθεσίες αξιολόγησης προϊόντων ή υπηρεσιών, φόρουμ, ιστολόγια και μέσα κοινωνικής δικτύωσης.

Με τη βοήθεια των συστημάτων ανάλυσης συναισθημάτων, αυτές οι μη δομημένες πληροφορίες θα μπορούσαν να μετατραπούν αυτόματα σε δομημένα δεδομένα

[6] https://en.wikipedia.org/wiki/Sentiment_analysis

δημόσιων γνωμών σχετικά με προϊόντα, υπηρεσίες, επωνυμίες, πολιτική ή οποιοδήποτε θέμα για το οποίο οι άνθρωποι μπορούν να εκφράσουν απόψεις. Αυτά τα δεδομένα μπορεί να είναι πολύ χρήσιμα για εμπορικές εφαρμογές, όπως η ανάλυση μάρκετινγκ, οι δημόσιες σχέσεις, οι κριτικές προϊόντων, η καθαρή βαθμολόγηση διοργανωτή, τα σχόλια προϊόντων και η εξυπηρέτηση πελατών.

Εκτιμάται ότι το 90% των παγκόσμιων δεδομένων είναι μη δομημένα και δεν οργανώνονται με προκαθορισμένο τρόπο[7]. Τα περισσότερα από αυτά προέρχονται από δεδομένα κειμένου, όπως μηνύματα ηλεκτρονικού ταχυδρομείου, δελτία υποστήριξης, συνομιλίες, κοινωνικά μέσα, έρευνες, άρθρα και έγγραφα. Αυτά τα κείμενα είναι συνήθως δύσκολα, χρονοβόρα και ακριβά για την ανάλυση, την κατανόηση και την ταξινόμηση.

Ιστορική Αναδρομή

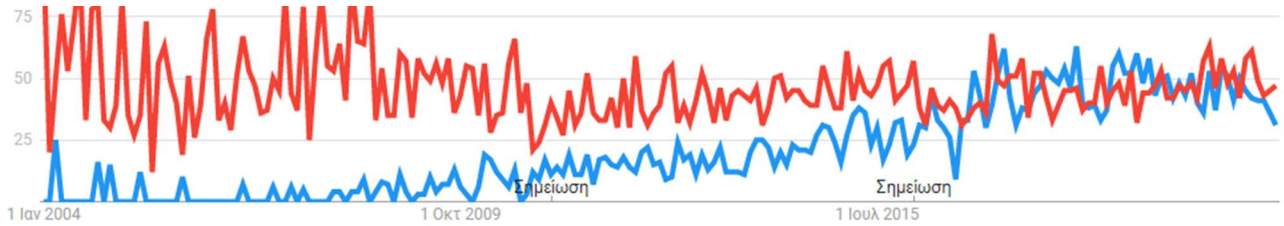
Οι ρίζες της ανάλυσης συναισθημάτων στις μελέτες σχετικά με την ανάλυση της κοινής γνώμης εμφανίζονται στις αρχές του 20ου αιώνα και στην ανάλυση υποκειμενικότητας του κειμένου που διενεργήθηκε από την Κοινότητα της Γλωσσολογίας στης δεκαετία του 1990. Ωστόσο, το ξέσπασμα της ανάλυσης συναισθήματος που βασίζεται σε υπολογιστή προέκυψε μόνο με τη διαθεσιμότητα υποκειμενικών κειμένων στο διαδίκτυο. Κατά συνέπεια, το 99% των δημοσιεύσεων δημοσιεύθηκε μετά το 2004. Οι δημοσιεύσεις που αφορούν ανάλυση συναισθήματος είναι διάσπαρτες σε πολλαπλούς τόπους δημοσιεύσεων, και ο συνολικός αριθμός δημοσιεύσεων στους 15 κορυφαίους τόπους αντιπροσωπεύει μόνο περίπου το 30% των δημοσιεύσεων συνολικά. Τα τελευταία χρόνια, η ανάλυση συναισθήματος έχει μετατοπιστεί από την ανάλυση των διαδικτυακών κριτικών προϊόντων σε κείμενα κοινωνικής δικτύωσης από το Twitter

[7] https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/?sh=d0192d615f64

και το Facebook. Πολλά θέματα πέρα από τις χρηματιστηριακές αγορές, οι εκλογές, οι καταστροφές, η ιατρική, η μηχανική λογισμικού και η διαδικτυακής παρενόχλησης (cyber bullying) επεκτείνουν την αξιολόγηση του συναισθήματος.

Έχει παρατηρηθεί μια τεράστια αύξηση του αριθμού των δημοσιεύσεων που επικεντρώνονται στην ανάλυση συναισθημάτων και στην εξόρυξη απόψεων κατά τα τελευταία έτη. Σύμφωνα με τα δεδομένα, σχεδόν 7.000 έγγραφα αυτού του θέματος έχουν καταχωρηθεί και, το πιο ενδιαφέρον, το 99% των άρθρων έχουν εμφανιστεί μετά το 2004. Η ανάλυση συναισθήματος αποτελεί έναν από τους ταχύτερα αναπτυσσόμενους τομείς έρευνας. Ο Πίνακας 5 δείχνει την αύξηση των αναζητήσεων που έγιναν με μια συμβολοσειρά αναζήτησης "ανάλυση συναισθήματος" και "σχόλια πελατών" στις Ηνωμένες Πολιτείες Αμερικής στη μηχανή Google Search. Παρατηρήθηκε ότι οι πρώτες ακαδημαϊκές μελέτες που μετρούν τις δημόσιες δηλώσεις είναι κατά τη διάρκεια και μετά το 2ο Παγκόσμιο Πόλεμο και τα κίνητρά τους είναι άκρως πολιτικής φύσεως. Η σύγχρονης ανάλυση συναισθήματος συνέβη μόνο στα μέσα της δεκαετίας του 2000, και επικεντρώθηκε στις κριτικές προϊόντων που διατίθενται στο διαδίκτυο, π.χ. IMdb, Yelp. Έκτοτε, η χρήση της ανάλυσης συναισθημάτων έχει φθάσει σε πολλούς άλλους τομείς, όπως η πρόβλεψη των χρηματοπιστωτικών αγορών και οι αντιδράσεις σε τρομοκρατικές επιθέσεις. Επιπλέον, η έρευνα για την ανάλυση συναισθήματος και η επεξεργασία της φυσικής γλώσσας έχει αντιμετωπίσει πολλά προβλήματα που συμβάλλουν στη δυνατότητα εφαρμογής της ανάλυσης συναισθημάτων για την ανίχνευση ειρωνείας και την πολύγλωσση υποστήριξη. Επιπροσθέτως, όσον αφορά τα συναισθήματα, οι προσπάθειες προχωρούν από την απλή ανίχνευση της Πόλωσης σε πιο σύνθετες αποχρώσεις των συναισθημάτων και τη διαφοροποίηση των αρνητικών συναισθημάτων όπως ο θυμός και η θλίψη (Mäntylä, et al., 2016) .

Πίνακας 5: Δεδομένα που δείχνουν τη σχετική δημοτικότητα των ερωτημάτων αναζήτησης "ανάλυση συναισθήματος" και "σχόλια πελατών" στις Ηνωμένες Πολιτείες Αμερικής



Πηγή: Google – Trends

https://trends.google.com/trends/explore?date=all&geo=US&q=sentiment%20analysis,customer%20feedback)

Πλεονεκτήματα της ανάλυσης συναισθήματος

Τα συστήματα ανάλυσης συναισθημάτων επιτρέπουν στις εταιρείες να κάνουν αίσθηση αυτής της θάλασσας μη δομημένου κειμένου αυτοματοποιώντας τις επιχειρηματικές διαδικασίες, δημιουργώντας χρήσιμες πληροφορίες και εξοικονομώντας ώρες μη αυτόματης επεξεργασίας δεδομένων, με άλλα λόγια, καθιστώντας τις ομάδες πιο αποτελεσματικές. Μερικά από τα πλεονεκτήματα της ανάλυσης συναισθήματος είναι τα εξής:

i. Δυνατότητα κλιμάκωσης

Η ανάλυση συναισθημάτων επιτρέπει την επεξεργασία δεδομένων σε κλίμακα με αποδοτικό και οικονομικά αποδοτικό τρόπο.

ii. Ανάλυση σε πραγματικό χρόνο

Η ανάλυση συναισθημάτων μπορεί να χρησιμοποιηθεί για τον εντοπισμό κρίσιμων πληροφοριών που επιτρέπουν την ευαισθητοποίηση της κατάστασης κατά τη διάρκεια συγκεκριμένων σεναρίων σε πραγματικό χρόνο. Υπάρχει κρίση δημοσίων σχέσεων στα μέσα κοινωνικής δικτύωσης που θα εκραγεί; Ένας θυμωμένος πελάτης που πρόκειται να δημιουργήσει προβλήματα; Ένα σύστημα ανάλυσης

συναισθήματος μπορεί να μας βοηθήσει να εντοπίσουμε αμέσως αυτά τα είδη καταστάσεων και να αναλάβουν δράση.

iii. Σαφή κριτήρια

Οι άνθρωποι δεν τηρούν σαφή κριτήρια για την αξιολόγηση του αισθήματος ενός κειμένου. Εκτιμάται ότι διαφορετικοί άνθρωποι συμφωνούν μόνο σε ποσοστό 60-65% όταν κρίνουν το συναίσθημα για ένα συγκεκριμένο κομμάτι κειμένου. Είναι ένα υποκειμενικό έργο που επηρεάζεται σε μεγάλο βαθμό από προσωπικές εμπειρίες, σκέψεις και πεποιθήσεις. Χρησιμοποιώντας ένα συγκεντρωτικό σύστημα ανάλυσης συναισθημάτων, οι εταιρείες μπορούν να εφαρμόσουν τα ίδια κριτήρια σε όλα τα δεδομένα τους. Αυτό συμβάλλει στη μείωση των σφαλμάτων και στη βελτίωση της συνέπειας των δεδομένων.

Επιπλέον, η ανάλυση συναισθήματος είναι χρήσιμη στην παρακολούθηση των μέσων κοινωνικής δικτύωσης, επειδή βοηθά να πραγματοποιηθούν τα παρακάτω:

- *Προτεραιότητα δράσης:* Η ανάλυση συναισθήματος επιτρέπει το εύκολο φιλτράρισμα στις μη αναγνωσμένες αναφορές με θετικά ή και αρνητικά σχόλια, δίνοντας την ευχέρεια για την ιεράρχηση των δράσεων που πρέπει να γίνουν.
- *Συντονισμός σε μια συγκεκριμένη χρονική στιγμή* – δηλαδή εάν υπάρχει προβάδισμα στην καμπάνια ενός προϊόντος ή την ημέρα που το αντίστοιχο προϊόν ενός ανταγωνιστή είχε πτώση σε επίπεδο δημόσιων αναφορών.
- *Ανάλυση του ανταγωνιστή*. Η χρήση του εργαλείου της συναισθηματικής ανάλυσης βοηθά στο να παρατηρηθεί εάν υπάρχει μια αρνητική απόκριση σε μια συγκεκριμένη λειτουργία του νέου προϊόντος του ανταγωνιστή.

Όσον αφορά την παρακολούθηση επωνυμίας, η ανάλυση συναισθημάτων είναι χρήσιμη επειδή βοηθά :

- Στην κατανόηση του πώς η φήμη της επωνυμίας σας εξελίσσεται με την πάροδο του χρόνου.

- Στην εξερεύνηση του ανταγωνισμού και πώς η φήμη του εξελίσσεται με την πάροδο του χρόνου.
- Στον εντοπισμό πιθανών κρίσεων δημοσίων σχέσεων. Δίνοντας προτεραιότητα στις πυρκαγιές που πρέπει να τεθούν αμέσως σε έλεγχο και τις αναφορές που μπορούν να περιμένουν.
- Στον συντονισμό σε μια συγκεκριμένη χρονική στιγμή. Ελέγχοντας τις αναφορές μια συγκεκριμένης ημέρας ή ενός μόνο προϊόντος.

Χρήση της ανάλυση συναισθήματος;

- ✓ Ανάλυση των δημοσιεύσεων tweets για ένα χρονικό διάστημα, για να δείτε το συναίσθημα ενός συγκεκριμένου ακροατηρίου.
- ✓ Εκτέλεση της ανάλυσης συναισθημάτων σε όλα τα μέσα κοινωνικής δικτύωσης που αναφέρονται στο brand που ενδιαφέρει και κατηγοριοποιήστε αυτόματα με επείγοντα χαρακτήρα.
- ✓ Αυτόματη ειδοποίηση της αντίστοιχης ομάδας της εταιρείας ή του οργανισμού με τις αναφορές που γίνονται στο διαδίκτυο γύρω από το πεδίο ενδιαφέροντος της.
- ✓ Αυτοματοποίηση οποιαδήποτε ή όλων αυτών των διεργασιών.
- ✓ Χρήση αναλύσεων για την απόκτηση καλύτερης εικόνας για το τι συμβαίνει στα κανάλια των μέσων κοινωνικής δικτύωσης.

Κεφάλαιο 2

2.1 Στόχοι της εργασίας

Βασικός στόχος της εργασίας είναι από την μελέτη κειμένων – δημοσιεύσεων που προέρχονται από τα κοινωνικά δίκτυα (Twitter[8]) να εξάγουμε συμπεράσματα σχετικά με τα συναισθήματα των χρηστών. Το βασικό εργαλείο που θα χρησιμοποιήσουμε είναι η εφαρμογή Ελληνικού συναισθηματικού λεξικού (Greek Sentiment Lexicon[9]) στα προαναφερθέντα κείμενα και θα αναλύσουμε τα δεδομένα για την εξαγωγή συμπερασμάτων.

Το Διαδίκτυο εξελίσσεται και μαζί και οι υπηρεσίες που προσφέρουν οι ιστοσελίδες καθώς και οι πλατφόρμες των κοινωνικών δικτύων. Στην παρούσα εργασία η εξόρυξη και συλλογή δεδομένων πραγματοποιείται με βάση το Twitter και την δυνατότητα που δίνεται μέσα από τη διασύνδεση προγραμματισμού εφαρμογών API (Application Programming Interface). Αρχικά το API είναι ένα πρωτόκολλο διασύνδεσης ή επικοινωνίας μεταξύ του πελάτη (χρήστη Η/Υ) και του εξυπηρετητή της πλατφόρμας (server) που αποσκοπεί στην απλούστευση της δημιουργίας λογισμικού (software) από την πλευρά του χρήστη. Περιγράφεται ως ένα συμβόλαιο κατά το οποίο η αποστολή αιτημάτων (requests) με συγκεκριμένη δομή από τον χρήστη θα επιστρέψει μια απάντηση του εξυπηρετητή σε συγκεκριμένη μορφή ή θα ξεκινήσει μια καθορισμένη ενέργεια. Μια πρακτική η οποία είναι αρκετά διαδεδομένη σχετικά με την άντληση συμπερασμάτων και αναλυτικών στοιχείων (analytics) για τα κοινωνικά δίκτυα. Προκειμένου να υπάρξει πρόσβαση στα στοιχεία του Twitter υπήρξε αίτηση για άδεια προγραμματιστή από την πλατφόρμα η οποία δόθηκε για εκπαιδευτικούς σκοπούς.

---

[8] Twitter: https://twitter.com/

[9] Greek Sentiment Lexicon: https://github.com/MKLab-ITI/greek-sentiment-lexicon

Η επικοινωνία και αυθεντικοποίησης μεταξύ Twitter και χρήστη έγινε βάσει των προδιαγραφών του OAuth 2.0 καθότι αποτελεί μια μέθοδος εξουσιοδότησης για την παροχή πρόσβασης σε πόρους μέσω του πρωτοκόλλου HTTP. Μπορεί να χρησιμοποιηθεί για την έγκριση διαφόρων εφαρμογών ή χειροκίνητης πρόσβασης του χρήστη.

Η ανάλυση των δεδομένων έγινε στο Ολοκληρωμένο Προγραμματιστικό Περιβάλλον IDE (Integrated Developer Environment) του RStudio. Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε είναι η R.
Η γλώσσα R χρησιμοποιείται ευρέως για την ανάλυση δεδομένων την εξόρυξη δεδομένων και την ανάπτυξη στατιστικού λογισμικού. Χαρακτηρίζεται ως γλώσσα η οποία είναι προσανατολισμένη στα δεδομένα και μεγάλο μέρος των βιβλιοθηκών (libraries) οι οποίες έχουν αναπτυχθεί είναι συνυφασμένες με τη μελέτη δεδομένων και οπτικοποίηση αποτελεσμάτων.

Η ανάπτυξη λογισμικού γίνεται μέσω των αρχείων που περιέχουν κομμάτια του κώδικα (Scripts) συνεπώς θεωρείται επιβεβλημένη η ανάγκη για παρακολούθηση των αλλαγών που γίνονται . Για την παρακολούθηση και την τροποποίηση του κώδικα χρησιμοποιήθηκε η πλατφόρμα GitHub η οποία παρέχει την εφαρμογή GitHub Desktop με την οποία ελέγχονται οι αλλαγές που πραγματοποιούνται (version controlling).

Η εργασία στο πλαίσιο της Πληροφορική κατατάσσεται στη θεματική ενότητα της Επεξεργασίας Φυσικής Γλώσσας NPL (Natural Language Processing). Η επεξεργασία της φυσικής γλώσσας (NLP) είναι ένα υποπεδίο της Γλωσσολογίας, της επιστήμης των υπολογιστών, της μηχανικής πληροφοριών και της τεχνητής νοημοσύνης που αφορά τις αλληλεπιδράσεις μεταξύ υπολογιστών και ανθρώπινων (φυσικών) γλωσσών, ιδίως για να αναλύσει μεγάλες ποσότητες δεδομένων φυσικής γλώσσας.

Σε συνδυασμό με την R και την πληθώρα των βιβλιοθηκών που αφορούν την Επεξεργασία Φυσικής Γλώσσας ήταν εφικτή η ανάλυση δεδομένων και η

παρουσίαση συγκεντρωτικών αποτελεσμάτων που ακολουθούν στις επόμενες σελίδες.

2.2 Σχεδιασμός και Εφαρμογή

Ως πρώτο βήμα για την προσπάθεια κατανόησης με συστηματικό και επιστημονικό τρόπο τις συναισθηματικές διακυμάνσεις που εμφανίζονται στο Twitter ήταν η συλλογή δεδομένων από την πλατφόρμα και ο «καθαρισμός» των δεδομένων. Ωστόσο έπρεπε να απαντηθούν επιπλέον τα κάτωθι ερωτήματα:

- Ποιος θα είναι ο τρόπος συλλογής και αποθήκευσης των δεδομένων;
- Ποια η δομή των αιτημάτων για τη συλλογή tweets;
- Ποιες μεθόδους θα χρησιμοποιούσαμε για τον «καθαρισμό» τους;
- Πως θα πραγματοποιούσαμε τη συναισθηματική ανάλυση;
- Ποια τα κριτήρια για την εξαγωγή αποτελεσμάτων;
- Οπτικοποίηση των συμπερασμάτων

Συλλογή Δεδομένων, Αποθήκευση Δεδομένων, Κριτήρια Αιτημάτων και Μέθοδοι Καθαρισμού

Σύμφωνα και με την Εισαγωγή της παρούσας εργασίας ο τρόπος συλλογής των tweets πραγματοποιείται σε συνεργασία με το Twitter μέσω την αποστολή δομημένων αιτημάτων στους εξυπηρετητές του βάσει του πρωτοκόλλου API (Application Platform Interface) το οποίο στην προκειμένη περίπτωση είναι OAuth 2.0 .

Για την αποθήκευση των δεδομένων επιλέχθηκαν δύο τύποι αρχείων. Ο πρώτος είναι τα αρχεία CSV (Comma Separated Value) με κατάληξη .csv και ο δεύτερος τα αρχεία κειμένου (TEXT) με κατάληξη .txt .

Η εργασία διεκπεραιώνεται για κείμενα με βάσει την Ελληνική γλώσσα, με καθημερινή συλλογή και από το κέντρο της Αθήνας με ακτίνα 10 μιλίων. Η επιλογή της γλώσσας υπήρξε χρονοβόρα, καθώς υπάρχουν εκτενή άρθρα και βάσεις (Adam Tsakalidis, 2018) δεδομένων με έτοιμα κείμενα τα οποία επικεντρώνονται στην Αγγλική Γλώσσα και έτοιμα snippets (κομμάτια κώδικα) που επίσης έχουν ως στόχο τη μελέτη και κατανόηση Αγγλικών tweets και όχι αντίστοιχος όγκος για την Ελληνική γλώσσα.

Πέραν της συλλογής και αποθήκευσης των δεδομένων επί καθημερινής βάσης το αμέσως επόμενο βήμα είναι εάν για το σκοπό της εργασίας ήταν αναγκαίο να επεξεργαστούμε αυτούσια τα tweets; Παρακάτω αναλύεται πως ακριβώς γίνεται ο «καθαρισμός» τους βάσει της μεθόδου που επιλέχθηκε για να χρησιμοποιηθεί για τη συναισθηματική ανάλυση κειμένου.

Μέθοδος Συναισθηματικής Ανάλυσης, Κριτήρια Εξαγωγής Συμπερασμάτων και Οπτικοποίηση Αποτελεσμάτων

Η μέθοδος που επιλέχθηκε είναι η συναισθηματική ανάλυση με βάση Ελληνικό Λεξικό Συναισθημάτων στο οποίο υπάρχει αντιστοίχιση λέξεων με συναίσθημα. Στο οποίο κριτήρια αποτελούν:

1. Η θέση της λέξης εντός του κειμένου.
2. Η σειρά των λέξεων που αντιστοιχίζονται.

Μετά την ολοκληρωμένη επεξεργασία και ανάλυση των tweets στο τέλος της ημέρας υπάρχει η τελική αποτύπωση σε πίνακα, των αποτελεσμάτων για κάθε συναίσθημα. Τα αποτελέσματα που έχουν εξαχθεί από κάθε ημέρα έχουν συμπληρωθεί σε ένα αρχείο excel (.xlsx) λειτουργώντας ως σύνολο δεδομένων (dataset) για τη στατιστική ανάλυση.

Η δυνατότητα που παρέχεται από το πρόγραμμα Microsoft Excel αλλά και από το R-Studio είναι η παραγωγή γραφημάτων όπου απεικονίζονται χρονοσειρές των συναισθημάτων κάθε ημέρας, ο αριθμός των tweets που επεξεργασθήκαμε και η ημερομηνία κατά την οποία συλλέχθηκαν.

2.2.1 Συλλογή Δεδομένων- Application Platform Interface (API) – Open Authentication (OAuth)

Ο ανοιχτός έλεγχος ταυτότητας (OAuth) είναι ένα ανοικτό πρότυπο για έλεγχο ταυτότητας, που εγκρίνεται από το Twitter για την παροχή πρόσβασης σε προστατευμένες πληροφορίες. Οι κωδικοί πρόσβασης είναι ευάλωτοι στην κλοπή και η OAuth παρέχει μια ασφαλέστερη εναλλακτική προσέγγιση με τη βοήθεια μιας χειραψίας τριών κατευθύνσεων. Επίσης, βελτιώνει την εμπιστοσύνη του χρήστη στην εφαρμογή καθώς ο κωδικός πρόσβασης του χρήστη για το Twitter δεν μοιράζεται ποτέ με εφαρμογές τρίτων κατασκευαστών. Ο έλεγχος ταυτότητας των αιτήσεων API στο Twitter πραγματοποιείται με χρήση του OAuth.

Παρακάτω συνοψίζονται τα βήματα που εμπλέκονται στη χρήση του Oai για πρόσβαση στο Twitter API. Τα API του Twitter μπορούν να προσπελαστούν μόνο από εφαρμογές. Παρακάτω βλέπουμε λεπτομερώς το βήματα για την πραγματοποίηση κλήσης API από μια εφαρμογή Twitter χρησιμοποιώντας το OAuth:

1. Οι εφαρμογές είναι γνωστές ως καταναλωτές και απαιτούνται όλες οι εφαρμογές να εγγραφούν στο Twitter. Μέσω αυτής της διαδικασίας, η εφαρμογή εκδίδει ένα κλειδί (token) και μυστικό (secret) καταναλωτή το οποίο η εφαρμογή πρέπει να προχωρήσει σε πιστοποίηση με το Twitter.
2. Η εφαρμογή χρησιμοποιεί το κλειδί και το μυστικό του καταναλωτή για τη δημιουργία μια μοναδικής σύνδεσης στην οποία απευθύνεται ο χρήστης για

έλεγχο ταυτότητας. Ο χρήστης εξουσιοδοτεί την εφαρμογή, πιστοποιώντας τον εαυτό του στο Twitter. Το Twitter επαληθεύει την ταυτότητα του χρήστη και εκδίδει έναν ελεγκτή OAuth που ονομάζεται επίσης PIN.

3. Ο χρήστης παρέχει αυτό το PIN στην εφαρμογή. Η εφαρμογή χρησιμοποιεί το PIN για να ζητήσετε ένα "διακριτικό πρόσβασης" και "πρόσβαση μυστικό" μοναδικό για το χρήστη.

Αιτήματα για τη συλλογή Tweets

Διεκπεραιώνοντας τη διαδικασία εγγραφής της εφαρμογής (στην προκειμένη περίπτωση αναφέρεται ως Under Lens) στο API του Twitter και έχοντας τα διαπιστευτήρια (credentials) για τη σύνδεση με τους εξυπηρετητές (servers) της πλατφόρμας προχωρήσαμε στη διαδικασία συλλογής των tweets. Τα κριτήρια με τα οποία δομήθηκαν τα αιτήματα είναι τα εξής:

i. Η περιοχή στην οποία αναζητήθηκαν δημοσιεύσεις είναι το κέντρο της Αθήνας με ακτίνα 10 μιλίων.
ii. Η γλώσσα στην οποία ήταν γραμμένες οι δημοσιεύσεις είναι τα Ελληνικά
iii. Ο αριθμός των tweets βάσει και των περιορισμών του API ήταν 18000 tweets ημερησίως.

RStudio και Twitter

Στο Ολοκληρωμένο Προγραμματιστικό Περιβάλλον IDE (Integrated Development Environment) πραγματοποιήθηκαν όλες οι παραπάνω διεπαφές με τους εξυπηρετητές το RStudio. Εκτός των πακέτων συναρτήσεων που προορίζονται για την επεξεργασία φυσικής γλώσσας NLP (Natural Language Process), στις βιβλιοθήκες που διατίθενται από την κοινότητα του RStudio υπάρχουν και δύο πακέτα το twitteR και το rtweet. Στην παρούσα εργασία το πακέτο το οποίο προτιμήθηκε μετά από δοκιμές είναι το rtweet.

Το rtweet παρέχει στους χρήστες μια σειρά λειτουργιών που έχουν σχεδιαστεί για την εξαγωγή δεδομένων από το Twitter's REST και τα API ροής.

Έχει τρεις κύριους στόχους:

- Διατύπωση και αποστολή αιτημάτων στο υπόλοιπο του Twitter και API ροής.
- Ανακτήστε και επαναλάβετε τα επιστρεφόμενα δεδομένα.
- Αντιπαραβάλει δεδομένα σε τακτοποιημένες δομές.

Η πρώτη συνάρτηση που θέτουμε σε εφαρμογή είναι η create_token() η οποία καθορίζει εξ' αρχής τα διαπιστευτήρια μας για τη δημιουργία ασφαλούς καναλιού επικοινωνίας και μετάδοσης δεδομένων με το Twitter.

Όπως έχει ήδη αναφερθεί η συλλογή των tweets έχει τις εξής παραμέτρους:

- Γεωγραφικά : κέντρο Αθήνας με ακτίνα 10 μιλίων
- Γλώσσα : Ελληνικά
- Αριθμός : 18.000 tweets
- Συχνότητα : Ημερήσια

Δίνεται το παράδειγμα της συνάρτησης search_tweets με τις παραπάνω παραμέτρους που αποθηκεύεται στη μεταβλητή tweets :

```
> tweets <- rtweet::search_tweets('lang:el' , n = 18000, retryonratelimit = TRUE, include_rts = FALSE , geocode = "37.98381,23.72754,10mi", since = since, until = until )
```

2.2.2 Αποθήκευση Δεδομένων - Αλγόριθμοι «καθαρισμού» και δόμησης δεδομένων

Μετά το πέρας της συλλογής Δεδομένων συνολικά για την ημέρα με τις παραμέτρους που έχουν οριστεί παραπάνω προχωράμε στην αποθήκευση τους. Οι πληροφορίες που αποθηκεύτηκαν είναι:

- Κείμενο των Tweets
- Χρονοσφραγίδες (Timestamps)

Οι τύποι αρχείων που χρησιμοποιήθηκαν είναι .txt και .csv .

Το Κείμενο των Tweets που αποθηκεύτηκαν με βάση την ημερομηνία συλλογής χρησιμοποιήθηκαν για την διεξαγωγή συναισθηματικής ανάλυσης ενώ οι Χρονοσφραγίδες των Tweets συλλέχθηκαν για την επαλήθευση της ημερομηνίας και ώρας συλλογής.

Μετά τη συλλογή και αποθήκευση των δεδομένων το επόμενο βήμα είναι η προ-επεξεργασία και «καθαρισμός» των δεδομένων προκειμένου να αρχίσει το στάδιο της ανάλυσης και εξαγωγής συμπερασμάτων. Τα προβλήματα που προέκυψαν κατά τη διάρκεια διαχείρισης των δεδομένων είναι τα παρακάτω:

- Η δομή με την οποία θα αποθηκευόντουσαν τα δεδομένα προκειμένου να είναι προσβάσιμα στις συναρτήσεις των πακέτων που θα χρησιμοποιήσουμε.
- Η παραγωγή Ελληνικών Λέξεων Διακοπής (Greek Stop Words) οι οποίες είναι λέξεις που φιλτράρονται πριν από την επεξεργασία δεδομένων.
- Η υιοθέτηση κατάλληλων τεχνικών για την ανάλυση που βασίζεται στη μέθοδο του Λεξικού Συναισθήματος.
- Η έλλειψη πακέτου Επεξεργασίας Φυσικής Γλώσσας στην προγραμματιστική γλώσσα R για την Ελληνική Γλώσσα.
- Η αλλαγή κωδικοποίησης εάν χρειαζόταν σε tweets από ISO-8859-7 σε UTF-8 ή και αντιστρόφως.

Κεφάλαιο 3

3.1 Συναισθηματικό Λεξικό

Στην παρούσα εργασίας χρησιμοποιήθηκε το Greek Sentiment Lexicon (Adam Tsakalidis, 2018) το οποίο αναπτύχθηκε στο πλαίσιο του Social Sensor ένα πρόγραμμα που χρηματοδοτήθηκε από την Ευρωπαϊκή Ένωση με στόχο τη συλλογή δεδομένων από προκειμένου να εντοπίζονται οι τάσεις , τα γεγονότα , και ενδιαφέροντα περιεχόμενα από τα κοινωνικά δίκτυα.

Το λεξικό - βασισμένο στη μέθοδο εξαγωγής χαρακτηριστικών βασίζεται στην ύπαρξη ενός λεξικού συναισθήματος. Συνήθως, ένα λεξικό συναισθήματος αποτελείται από ένα σύνολο όρων σε μια συγκεκριμένη γλώσσα, που φέρουν κάποιο είδος συναισθηματικού βάρους, σχολιασμένη κατά μήκος μιας σειράς διαστάσεων. Ο αριθμός των διαστάσεων (συναισθήματα) είναι εξαρτώμενο από το λεξικό, ενώ, για κάθε διάσταση, ένας δεδομένος όρος μπορεί να σημειωθεί είτε με δυαδικό τρόπο (π.χ. ο όρος χαρακτηρίζεται από το συναίσθημα θυμού ή όχι) , ή με τη χρήση μιας συγκεκριμένης κλίμακας εκτίμησης. Οι όροι μπορούν επίσης να σχολιασθούν σχετικά με την υποκειμενικότητα τους, δηλαδή η ταξινόμηση ενός εγγράφου γίνεται βάσει είτε της υποκειμενικότητας είτε της αντικειμενικότητάς ή/και της Πόλωσης – που προσπαθεί να απαντήσει στο ερώτημα εάν ένα έγγραφο είναι θετικό, αρνητικό ή ουδέτερο. Ο σχολιασμός μπορεί να επιτευχθεί είτε χειροκίνητα (εμπειρογνώμονες) , ή αυτόματα (μηχανή εφαρμογή εκμάθησης σχετικά με το συναίσθημα – σε προσημειωμένα έγγραφα).

Δομή Συναισθηματικού Λεξικού

Κάθε γραμμή του λεξιλογίου διαχωρίζεται με στηλοθέτη και αντιστοιχεί στα ακόλουθα πεδία:

- Όρος: Ελληνικός όρος στον οποίο αναφέρονται οι ακόλουθοι σχολιασμοί. Στις περιπτώσεις των επιθέτων, και τα τρία γένη (αρσενικό, θηλυκό, ουδέτερο) συνεπάγονται με την παροχή των επιθημάτων (-ω-ο). Σε ορισμένες περιπτώσεις, οι όροι αναφέρονται σε συστατικά των μεγαλύτερων λέξεων. Οι όροι αυτοί τελειώνουν με παύλα (-).
- POS_x: Τμήμα ετικέτας ομιλίας που εκχωρείται από το σχολιαστή x (x=1,2,3,4). Σε μερικές περίπλοκες περιπτώσεις, δεν υπάρχει καμία ομόφωνη συμφωνία σχολιασμού για το POS ενός όρου.
- Υποκειμενικότητα_x: Κλάση υποκειμενικότητας που εκχωρείται από το σχολιαστή x. Οι πιθανές τιμές είναι στόχος (OBJ), αδύναμα υποκειμενική (SUBJ-) και έντονα υποκειμενική (SUBJ+).
- Πόλωση _x: Κατηγορία Πόλωσης που εκχωρείται από το σχολιαστή x. Οι πιθανές τιμές είναι Θετικές (POS), Αρνητικές (NEG), Τόσο θετικές όσο και αρνητικές (BOTH) και Ουδέτερη (N/A).
- Θυμός_x: Βαθμολογία θυμού ανατεθεί από το σχολιαστή x. Οι πιθανές τιμές είναι μεταξύ 1 και 5 και N/A.
- Απέχθεια_x: Βαθμολογία «Απέχθεια» που αποδίδεται από το σχολιαστή x. Οι πιθανές τιμές είναι μεταξύ 1 και 5 και N/A.
- Φόβος_x: Βαθμολογία Φόβου ανατεθεί από το σχολιαστή x. Πιθανές τιμές είναι μεταξύ 1 και 5 και N/A.
- Ευτυχία_x: σκορ Ευτυχίας ανατεθεί από το σχολιαστή x. Πιθανές τιμές είναι μεταξύ 1 και 5 και N/A.
- Θλίψη_x: Βαθμολογία Θλίψης αποδίδεται από το σχολιαστή x. Οι πιθανές τιμές είναι μεταξύ 1 και 5 και N/A.

- Έκπληξη_x: Βαθμολογία Έκπληξης ανατεθεί από το σχολιαστή x. Πιθανές τιμές είναι μεταξύ 1 και 5 και N/A.
- Πρόσθετο_x: Άλλοι όροι που ο σχολιαστής x θεωρεί ως συναφής με αυτόν τον όρο.
- Σχόλιο_x: σχόλια για να διευκρινίσει τη χρήση του όρου.

Η τιμή N/A (μη εφαρμόσιμη) χρησιμοποιείται σε περιπτώσεις που ο σχολιασμός θεωρεί ότι καμία τιμή δεν είναι κατάλληλη για τον όρο ή σε περιπτώσεις όπου δεν ήταν βέβαιοι για την κατάλληλη τιμή.

Τεκμηρίωση Συναισθηματικού Λεξικού

Το λεξικό πάνω στο οποίο βασίσθηκε το Greek Sentiment Lexicon είναι του Τριανταφυλλίδη[10] (Τριανταφυλλίδης, 1998) του οποίου η ηλεκτρονική έκδοση παρέχει ετικέτες σχετικά με τα μέρη του λόγου, παραδείγματα, ετυμολογία κ.α. με τέτοιο τρόπο ώστε να είναι δημοφιλές για εργασίες Επεξεργασίας Φυσικής Γλώσσας για τα Ελληνικά. Από εκεί έγινε εξαγωγή των λέξεων που είχαν τις ετικέτες ειρωνεία , χλευασμό ή χυδαίες εκφράσεις. Αρχικά 4 σχολιαστές από όπου προέρχονταν 2 από την Πληροφορική και 2 από τη Γλωσσολογία έπαιρναν τα λήμματα και σχολίαζαν με βάση την αρνητική , την θετική ή την ουδέτερη φύση τους με κλίμακα από το 1 έως το 5 βασιζόμενοι στα6 συναισθήματα όπως είχαν διατυπωθεί από τον Ekman Θυμός, Απέχθεια, Φόβος, Ευτυχία, Θλίψη και Έκπληξη (Ekman, 1992). Στη συνέχεια, καταργήθηκαν οι λέξεις για τις οποίες υπήρχε μια βαθμολογία υποκειμενικότητας που έλειπε για περισσότερους από έναν σχολιαστή, μειώνοντας το λεξικό σε 2260 λέξεις. Διορθώθηκαν οι λίγες καταχωρήσεις που κρίθηκαν αντικειμενικές αλλά είχαν μη μηδενική Πόλωση ή συναισθηματική βαθμολογία, μετατρέποντας τις θετικές και αρνητικές βαθμολογίες σε 0 και τα αποτελέσματα

[10] https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/index.html

των συναισθημάτων σε 1 (δηλαδή, η ελάχιστη επιτρεπόμενη βαθμολογία τους), δεδομένου ότι οι συμμετοχές αυτές κρίθηκαν εσφαλμένα από το σχολιαστή, καθώς δεν ήταν σύμφωνα με τις οδηγίες σχολιασμού. Επίσης, μετατράπηκαν οι βαθμολογίες υποκειμενικότητας σε τρεις τιμές: 0 για το στόχο, 0 για την υποκειμενικότητα και 1 για έντονα υποκειμενική. Τέλος υπολογίσθηκε ο μέσος όρος για το υποκειμενικό, θετικό, αρνητικό και οι έξι βαθμολογίες συναισθημάτων, όπως παρέχονται από τους σχολιαστές. Η συμφωνία των σχολιασμού εμφανίζεται στους πίνακες 1 και 2. Μετράμε τη συμφωνία με βάση την Cohen Kappa coefficient[11] για τις θετικές και αρνητικές διαστάσεις, δεδομένου ότι αυτές οι δύο διαφορετικές κλάσεις· για τα υπόλοιπα, μετρήθηκε η συμφωνία Pearson Correlation [12]. Παρατηρούμε μια δίκαιη συμφωνία (.40–.60) στις περισσότερες περιπτώσεις, με εξαίρεση το συναίσθημα της έκπληξης . Ο λόγος πίσω από αυτό είναι πιθανώς η φύση της έκπληξη, το οποίο, σε αντίθεση με τα υπόλοιπα, μπορεί να εκφραστεί τόσο σε θετικό όσο και σε αρνητικό τρόπο, αμφισβητώντας έτσι τους σχολιαστές.

Πίνακας 6: Συμφωνία των σχολιασμών για την υποκειμενικότητα (συσχέτιση Pearson), θετική και αρνητική (Κάππα), αντίστοιχα.

(a) Subjectivity				(b) Positive				(c) Negative			
	#2	#3	#4		#2	#3	#4		#2	#3	#4
#1	.47	.90	.77	#1	.40	.82	.51	#1	.28	.85	.45
#2		.45	.59	#2		.38	.45	#2		.31	.42
#3			.60	#3			.53	#3			.47

[11] https://en.wikipedia.org/wiki/Cohen%27s_kappa

[12] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Πίνακας 7: Συμφωνία των σχολιαστών (συσχέτιση Pearson) για τα έξι Συναισθήματα

(a) Anger				(b) Disgust				(c) Fear			
	#2	#3	#4		#2	#3	#4		#2	#3	#4
#1	.28	.68	.55	#1	.47	.74	.57	#1	.37	.60	.35
#2		.34	.39	#2		.45	.53	#2		.41	.28
#3			.58	#3			.56	#3			.46
(d) Happy				(e) Sad				(f) Surprise			
	#2	#3	#4		#2	#3	#4		#2	#3	#4
#1	.42	.83	.62	#1	.40	.59	.47	#1	.18	.50	.17
#2		.40	.53	#2		.39	.46	#2		.18	.40
#3			.62	#3			.53	#3			.20

Η επιπλέον διόρθωση των συναισθηματικών αξιών των λέξεων έγινε από δεδομένα που συλλέχθηκαν από το Twitter και περιγράφονται στο άρθρο «Building and evaluating resources for sentiment analysis in the Greek language» (Adam Tsakalidis, 2018).

3.1 Αλγόριθμοι Ταιριάσματος Λέξεων (String Matching - Metrics)

Μετρήσεις Συμβολοσειράς

Οι μετρήσεις συμβολοσειράς (String Metrics) είναι τρόποι ποσοτικοποίησης της ομοιότητας μεταξύ δύο πεπερασμένων ακολουθιών χαρακτήρων, συνήθως δεδομένων κειμένου. Με τα χρόνια, πολλές μέθοδοι μέτρησης έχουν αναπτυχθεί. Οι Μετρικές είναι βασισμένες σε μια μαθηματική κατανόηση του συνόλου όλων των λέξεων που μπορούν να αποτελούνται από ένα πεπερασμένο αλφάβητο, άλλα είναι βασισμένα σε αρχές ευριστικής (Heuristic) , όπως πώς το κείμενο ηχεί όταν προφέρεται από έναν εγγενή αγγλικό ομιλητή.

Οι όροι “sting metrics” και “string distance” χρησιμοποιούνται περισσότερο ή λιγότερο εναλλάξ στη βιβλιογραφία. Από μαθηματική άποψη, οι μετρήσεις συμβολοσειρών συχνά δεν υπακούν στις απαιτήσεις που απαιτούνται συνήθως από μια συνάρτηση απόστασης. Για παράδειγμα, δεν είναι αληθές ότι για όλες τις μετρήσεις συμβολοσειράς η απόσταση 0 σημαίνει ότι δύο νήματα είναι ίδια (π.χ. στην απόσταση q-grams). Παρ 'όλα αυτά, οι μετρήσεις νημάτων είναι πολύ χρήσιμες στην πράξη και έχουν πολλές εφαρμογές.

Η μετρική που μπορούμε να επιλέξουμε για μια εφαρμογή εξαρτάται αρκετά από τη φύση του κειμένου και την αιτία των διαστάσεων μεταξύ των λέξεων κειμένου που μετρούμε.

Η φύση της Ελληνικής Γλώσσας λόγων των πολλών καταλήξεων μας ώθησε να χρησιμοποιήσουμε περισσότερες από μια μετρικές λέξεων προκειμένου να ελέγξουμε τις επιδόσεις τους.

Hamming Distance

Στην Πληροφορική, η Hamming απόσταση μεταξύ δύο strings ίσου μήκους είναι ο αριθμός των θέσεων στις οποίες τα αντίστοιχα σύμβολα είναι διαφορετικά. Πιο απλά, μετρά τον ελάχιστο αριθμό αντικαταστάσεων που απαιτούνται για την αλλαγή μιας συμβολοσειράς στην άλλη ή τον ελάχιστο αριθμό σφαλμάτων που θα μπορούσαν να έχουν μετατρέψει τη μία συμβολοσειρά στην άλλη. Σε ένα γενικότερο πλαίσιο, η απόσταση της αιώρησης είναι μία από τις διάφορες μετρήσεις συμβολοσειρών για τη μέτρηση της απόστασης επεξεργασίας μεταξύ δύο ακολουθιών.

Παράδειγμα :

Η απόσταση Hamming για τα παρακάτω:

- "**Νikος**" και "**Νάκος**" είναι 1.
- "**karolin**" και "**kerstin**" είναι 3.

- **1011101** και **1001001** είναι 2.
- **2173896** και **2233796** είναι 3.

Στην εργασία χρησιμοποιήθηκε η απόσταση Hamming με μηδενικό περιθώριο λάθους.

Πλεονέκτημα:

- ❖ Ακριβής αντιστοίχιση

Μειονέκτημα:

- ❖ Λέξεις έμεναν εκτός αντιστοίχισης λόγω μικρών αλλαγών π.χ. ορθογραφικών λαθών

Απόσταση Jaro - Winkler

Στην Πληροφορική και τη Στατιστική, η απόσταση Jaro-Winkler είναι μια Μετρική Νήματος η οποία μετρά μια απόσταση μεταξύ δύο ακολουθιών. Είναι μια παραλλαγή που προτάθηκε το 1990 από William E. Winkler της απόστασης Jaro (1989, Matthew A. Jaro).

Η απόσταση Jaro-Winkler χρησιμοποιεί μια προκαθορισμένη κλίμακα p που δίνει πιο ευνοϊκές αξιολογήσεις σε συμβολοσειρές που ταιριάζουν από την αρχή για ένα σύνολο προθέματος μήκους l .

Όσο χαμηλότερη είναι η απόσταση Jaro-Winkler για δύο νήματα τόσο πιο όμοια είναι τα νήματα . Το σκορ είναι κανονικοποιημένο, με το 0 να ισοδυναμεί με καμία ομοιότητα και 1 να είναι μια ακριβής αντιστοιχία. Η ομοιότητα Jaro-Winkler είναι συμπληρωματική (1 – Απόσταση Jaro-Winkler).

Αν και συχνά αναφέρεται ως μετρική απόσταση, η απόσταση Jaro-Winkler δεν είναι μια μετρική με τη μαθηματική έννοια του όρου αυτού, επειδή δεν έχει την ιδιότητα της τριγωνικής ανισότητας.

Jaro Ομοιότητα:

Η ομοιότητα Jaro sim_j δύο δοθέντων νημάτων κειμένου $S_1 και S_2$ είναι :

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

Όπου:

- Το $|s_i|$ είναι το μέγεθος του νήματος S_i
- m είναι ο αριθμός των χαρακτήρων που συμφωνούν στα 2 νήματα
- t είναι ο αριθμός των μεταθέσεων κατά το ήμισυ.

Δύο χαρακτήρες από το S_1και S_2 αντίστοιχα, θεωρούνται ότι ταιριάζουν μόνο εάν είναι τα ίδια και όχι μακρύτερα από $\left\lfloor \frac{max(|s_1|,|s_2|)}{2} \right\rfloor - 1$

Κάθε χαρακτήρας του S_1 συγκρίνεται με όλους τους χαρακτήρες του S_2. Μετά τη σύγκριση προκύπτει ο αριθμός «ταιριασμάτων» μεταξύ των 2 νημάτων (αλλά σε διαφορετική σειρά) διαιρούμενος με το 2 είναι ο αριθμός των μεταθέσεων. Για παράδειγμα συγκρίνοντας το CRATE με το TRACE , μόνο τα 'R' 'A' 'E' είναι οι χαρακτήρες που ταιριάζουν, συνεπώς το m = 3. Παρόλο που το 'C', 'T' εμφανίζονται και στα 2 νήματα κειμένου είναι μακρύτερα από την τιμή 1 (η οποία είναι αποτέλεσμα του $\left\lfloor \frac{|5|}{2} \right\rfloor - 1$). Άρα το t = 0 . Στο DwAyNE εναντίον DuANE, τα ίδια γράμματα είναι ήδη στην ίδια σειρά D-A-N-E, οπότε δεν χρειάζονται μεταφορά.

Οι παραπάνω μετρικές χρησιμοποιήθηκαν κατά τη διάρκεια της αντιστοίχισης των λέξεων από τα tweets με τους όρους του Greek Sentiment Lexicon προκειμένου να εξαχθούν οι τιμές των συναισθημάτων.

Πλεονέκτημα:

- Ακριβής αντιστοίχιση
- Αντιστοίχιση και λέξεων που εμφανίζουν μεταθέσεις με βάσει τον τύπο Jaro – Winkler

Μειονέκτημα:

- Αντιστοίχιση λάθος λέξεων

Κεφάλαιο 4

Επεξεργασία Δεδομένων

Μετά την εξόρυξη δεδομένων και την αποθήκευση τους σε αρχεία όπως αναφέρονται στο Κεφάλαιο 2 αρχίζει η διαδικασία επεξεργασίας των κειμένων.

Το βασικό αρχείο κώδικα που αφορά το κομμάτι της επεξεργασίας και της τελικής ανάλυσης των κειμένων βρίσκεται στο Παράρτημα Α κάτω από τον τίτλο Sentiment Analysis και αφορά όλα τα παρακάτω βήματα

Τύποι Δεδομένων σε R:

- Διανύσματα = Arrays
- Λίστες = Lists
- Πλαίσια Δεδομένων = Data Frames
- Πίνακες = Matrices

Έχοντας πλέον ένα σύνολο δεδομένων τα κείμενα που υπάρχουν μέσα στα tweets γίνεται εισαγωγή από αρχεία κειμένου τύπου .txt και έπειτα διενεργούνται οι διαδικασίες μορφοποίησης τους.

Στο επίπεδο του μετασχηματισμού της ανίχνευσης και αλλαγής κωδικοποίησης από ISO 8859-7 σε UTF-8 έγινε χρήση των πακέτων *stringi* και *base*.

Στη συνέχεια διαχωρίζουμε τις λέξεις σε κάθε νήμα και τις αποθηκεύουμε σε νέες μεταβλητές με χρήση του πακέτου *tokenizers.*

Οι παράμετροι που εμφανίζονται στη συνάρτηση tokenize_words έχουν τις εξής σημασίες:

1. Data: Τα tweets της ημέρας.

2. Lowercase = TRUE: όλα τα σύμβολα είτε σε Κεφαλαία είτε σε μικρά γράμματα μετατρέπονται σε μικρά γράμματα.
3. Stopwords = stop_words: είναι το σύνολο των λέξεων που αφαιρούνται καθώς δεν έχουν σημασία όπως άρθρα, συζευκτικοί ή διαζευκτικοί σύνδεσμοι, τοπικές ή δηκτικές αντωνυμίες τόσο στα Ελληνικά όσο και στα Αγγλικά.
4. Strip_puct = TRUE: Η αφαίρεση σημείων στίξης.
5. Stip_numeric = TRUE: Η αφαίρεση των αριθμών

Ωστόσο μια επιπλέον διαφορά των Ελληνικών έναντι των Αγγλικών είναι η χρήση τόνων ή διαλυτικών με αποτέλεσμα όλα τα φωνήεντα να προσαρμοστούν ως φωνήεντα χωρίς τόνους για να είναι δυνατή η επεξεργασία τους.
Έχοντας εκτελέσει τα παραπάνω έχει ολοκληρωθεί ο καθαρισμός και η προ-επεξεργασία των δεδομένων μας.

Τα βασικά στάδια για την επεξεργασία έχουν ως εξής:

a. Εισαγωγή δεδομένων από κείμενο σε κατάλληλο τύπο δεδομένων
b. Αφαίρεση κενών γραμμών.
c. Αλλαγή της κωδικοποίησης από ISO-8859 σε UTF-8 με αποθήκευση σε νέο αρχείο.
d. Αποθήκευση κάθε tweet σε ξεχωριστό διάνυσμα
e. Μετατροπή των κεφαλαίων σε μικρά γράμματα.
f. Αντικατάσταση των τονισμένων φωνηέντων με τα αντίστοιχα μη τονισμένα φωνήεντα πχ. Ά->α , ή->η ,ϊ->ι κοκ
g. Αφαίρεση των Ελληνικών Λέξεων Διακοπής
h. Εύρεση και αφαίρεση URL και Emojis

Συναισθηματική Ανάλυση

Για την αξιοποίηση του Ελληνικού συναισθηματικού λεξικού προβήκαμε σε κάποιες διορθώσεις όπως οι παρακάτω :

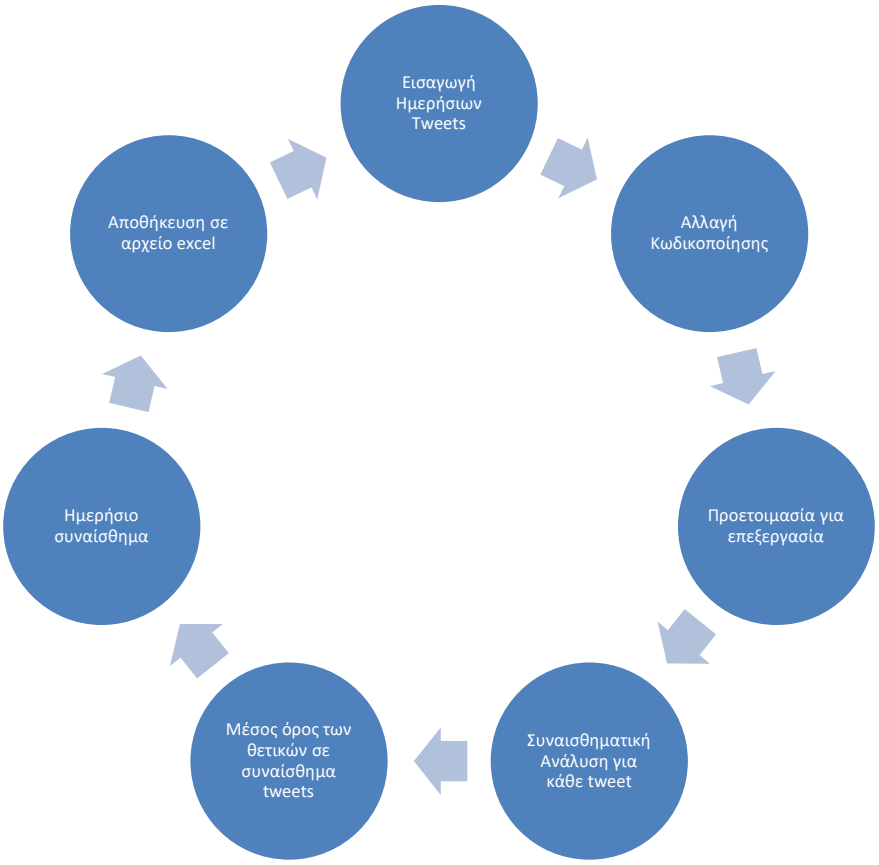
- Στην αντικατάσταση των N/A με μηδέν.
- Στην αντικατάσταση των NEG με -1 και των POS με +1

- Στην αντικατάσταση των τιμών ΒΟΤΗ στις στήλες με την Πόλωση (Polarity) αναλόγως του μέσου όρου των αριθμητικών τιμών των υπολοίπων κελιών της γραμμής σε περίπτωση θετικού μέσου όρου τα ΒΟΤΗ μετατρέπονταν σε 1 ενώ σε περίπτωση αρνητικού μέσου όρου σε -1 Παράδειγμα:
  - ΒΟΤΗ | 0 | ΒΟΤΗ | -1 → μέσος όρος γραμμής αριθμητικών στοιχείων είναι -0,5 νέα γραμμή
  - -1 | 0 | -1 |-1

Βήματα συναισθηματικής ανάλυσης:

- ✓ Εισαγωγή των Ημερήσιων tweets
- ✓ Δημιουργία πίνακα emotions διαστάσεων n*7 (για n - λέξεις με 6 συναισθήματα συν την Πόλωση)
- ✓ Έλεγχος αντιστοίχισης λέξεων με βάση τις μετρικές λέξεων (Hamming – Jaro Winkler)
- ✓ Εφόσον υπήρχε ταύτιση και τιμή μεγαλύτερη ίση του 3 αναφορικά με τα συναισθήματα συμπληρωνόταν στο προαναφερθέν πίνακα. Ο λόγος που αποθηκεύαμε μόνο τα συναισθήματα με τιμή πάνω από 3 είναι για να έχουμε ένα πιο ενισχυμένο 'σήμα'.
- ✓ Συγκέντρωση του μέσου όρου κατά στήλη για κάθε πίνακα για κάθε tweet στον πίνακα all_sentiments
- ✓ Συγκέντρωση των γραμμών του πίνακα all_sentiments που είχαν μόνο συναίσθημα (δλδ διαγραφή μηδενικών γραμμών) στον πίνακα only_sentiments

Πίνακας 8: Κύκλος εργασιών



Παράδειγμα

Στις παρακάτω σελίδες ακολουθούν εικόνες σχετικά με την επεξεργασία των Tweets που συλλέχθηκαν στις 20.12.2020 και χρησιμοποιείται η μετρική απόσταση Hamming :

Πίνακας 9: Παράδειγμα ανάλυσης Tweet

```
> data[10]
[1] "Εμάς τους καταναλωτέςπελατες τι μας ωφελεί πόσοι λύκοι κάνουν παρέα"
> clean_tokens[10]
[[1]]
[1] "εμας"              "καταναλωτεςπελατες" "ωφελει"            "ποσοι"             "λυκοι"
[6] "κανουν"            "παρεα"

> o
[1]    0    0    0    0    0    0 1570
> length(o)
[1] 7
> Greek_Lexicon[1570,1]
[1] "παρέα"
> emotions
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    5    0    3
```

Ακολουθώντας την παραπάνω διαδικασία για όλα τα tweets καταλήγουμε στον παρακάτω πίνακα όπου εμφανίζονται όλα τα συναισθήματα:

Πίνακας 10: Συνολικά Συναισθήματα για τις 20.12.2020

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
1	0.0	0.0	0	0.0	0.0	0.0
2	0.0	0.0	0	0.0	0.0	0.0
3	0.0	0.0	0	0.0	0.0	0.0
4	0.0	0.0	0	0.0	0.0	0.0
5	0.0	0.0	0	0.0	0.0	0.0
6	0.0	0.0	0	0.0	0.0	0.0
7	0.0	0.0	0	0.0	0.0	0.0
8	0.0	0.0	0	0.0	0.0	0.0
9	0.0	0.0	0	0.0	0.0	0.0
10	0.0	0.0	0	5.0	0.0	3.0
11	0.0	0.0	0	0.0	0.0	0.0
12	0.0	0.0	0	0.0	0.0	0.0
13	0.0	0.0	0	0.0	0.0	0.0
14	0.0	0.0	0	0.0	0.0	0.0
15	0.0	0.0	0	0.0	0.0	0.0
16	0.0	0.0	0	5.0	0.0	0.0
17	0.0	0.0	0	0.0	0.0	0.0
18	0.0	0.0	0	0.0	0.0	0.0
19	0.0	0.0	0	0.0	0.0	0.0
20	2.0	0.0	0	1.5	0.0	2.0
21	0.0	0.0	0	0.0	0.0	0.0
22	0.0	0.0	0	4.0	0.0	3.0
23	0.0	0.0	0	0.0	0.0	0.0
24	0.0	0.0	0	0.0	0.0	0.0
25	0.0	0.0	0	0.0	0.0	0.0
26	0.0	0.0	0	0.0	0.0	0.0
27	0.0	0.0	0	0.0	0.0	0.0

Ο πίνακας ωστόσο που χρησιμοποιούμε για την αποτύπωση των συναισθημάτων αφορά μόνο τις μη μηδενικές τιμές και είναι ο παρακάτω:

Πίνακας 11: Διάθεση της ημέρας 20.12.2020

	V1	V2	V3	V4	V5	V6
1	0.00	0.000000	0.00	5.000000	0.00	3.000000
2	0.00	0.000000	0.00	5.000000	0.00	0.000000
3	2.00	0.000000	0.00	1.500000	0.00	2.000000
4	0.00	0.000000	0.00	4.000000	0.00	3.000000
5	0.00	0.000000	0.00	4.000000	0.00	4.000000
6	0.00	3.000000	0.00	3.000000	0.00	3.000000
7	0.00	0.000000	0.00	3.000000	0.00	0.000000
8	0.00	0.000000	0.00	0.000000	4.00	4.000000
9	0.00	0.000000	0.00	4.000000	0.00	4.000000
10	0.00	0.000000	0.00	2.500000	0.00	1.500000
11	0.00	0.000000	0.00	1.500000	0.00	0.000000
12	0.00	0.000000	0.00	4.000000	0.00	4.000000
13	0.00	0.000000	0.00	3.000000	0.00	0.000000
14	4.00	0.000000	3.00	0.000000	0.00	4.000000
15	3.00	0.000000	0.00	0.000000	3.00	4.000000
16	0.00	0.000000	0.00	1.500000	0.00	0.000000
17	1.50	0.000000	0.00	0.000000	1.50	2.000000
18	0.00	0.000000	0.00	1.000000	0.00	0.000000
19	0.00	0.000000	0.00	1.500000	0.00	0.000000
20	1.00	0.000000	0.00	0.000000	0.00	1.000000
21	3.00	3.000000	3.00	3.000000	3.00	4.000000
22	3.00	0.000000	0.00	0.000000	3.00	4.000000
23	0.00	0.000000	0.00	5.000000	0.00	3.000000
24	0.00	0.000000	0.00	1.500000	0.00	0.000000
25	0.00	0.000000	0.00	3.000000	0.00	4.000000
26	3.00	0.000000	0.00	0.000000	3.00	4.000000
27	3.00	3.000000	0.00	3.000000	0.00	0.000000
28	0.00	3.000000	0.00	3.000000	0.00	3.000000
29	0.00	0.000000	0.00	5.000000	0.00	3.000000
30	0.00	0.000000	0.00	5.000000	0.00	0.000000
31	0.00	0.000000	0.00	5.000000	0.00	5.000000
32	3.00	3.000000	0.00	0.000000	0.00	3.000000

Και το τελικό συναίσθημα αποδίδεται ως ο μέσος όρος των στηλών από τον πίνακα που περιέχει μόνο τα tweets που περιέχουν «συναίσθημα».

```
> mood_of_the_day_only_sentiments
[1] 1.0198322 0.7953089 0.4011442 2.1534325 0.3636918 2.6817696
```

Οπτικοποίηση και Ταυτότητα Δεδομένων

Στο τελικό αρχείο excel "mood_of_twitter_revised.xlsx" υπάρχει ο πίνακας με τις τελικές τιμές συναισθήματος για κάθε ημέρα για το σύνολο του χρονικού διαστήματος για το οποίο συλλέγονταν τα δεδομένα. '

Στην επιλογή που δινόταν για τα γραφήματα μέσω του Microsoft Excel υπήρχε η δυνατότητα των χρονοσειρών ωστόσο επιλέξαμε να χρησιμοποιήσουμε "smooth functions".

Μέσω της R εισήγαμε τον τελικό πίνακα σε ένα "data frame" με την ονομασία data_sentiment. Όπου στα παρακάτω γραφήματα ο άξονας x είναι οι Ημερομηνίες και στον άξονα y είναι οι τιμές του κάθε συναισθήματος από 1 έως 5.

Το "smooth function" είναι η Τοπική πολυωνυμική παλινδρόμηση (Local Polynomial Regression[13]).

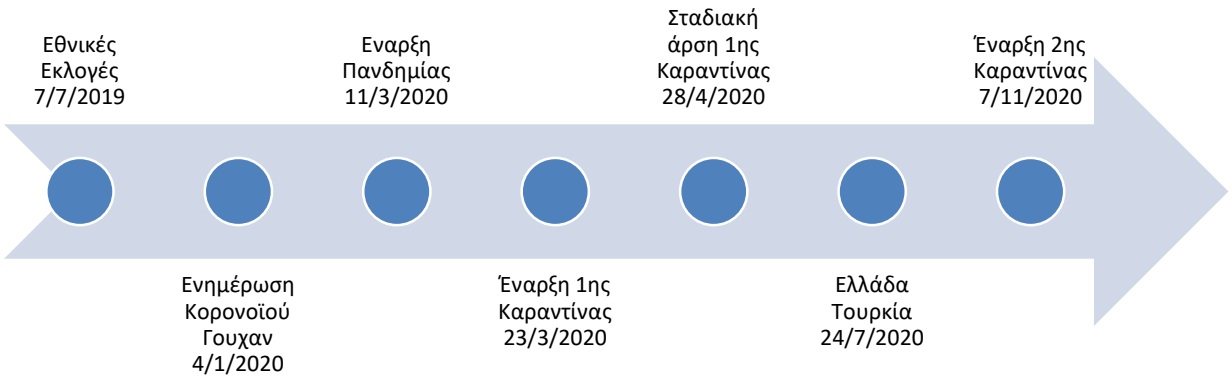
Πίνακας 12: Ταυτότητα Δεδομένων από Twitter

Data Source	Twitter
No of Tweets	*9.365.343*
Start Date	*1/6/2019*
End Date	*26/1/2021*
Missing Dates	*"2019-07-12"*
	"2019-07-13"
	"2019-07-14"
	"2019-07-15"
	"2019-07-16"
	"2020-01-01"
	"2020-05-31"
	"2020-06-04"
	"2020-07-02"
	"2021-01-03"

[13] Local Polynomial Regression: https://en.wikipedia.org/wiki/Local_regression

Συλλογή Δεδομένων - Γεγονότα

Η συλλογή των δεδομένων διήρκεσε από τον Ιούνιο του 2019 έως και το Νοέμβριο του 2020.



Εικόνα 2: Χρονοσειρά γεγονότων που έλαβαν χώρα στην Ελλάδα

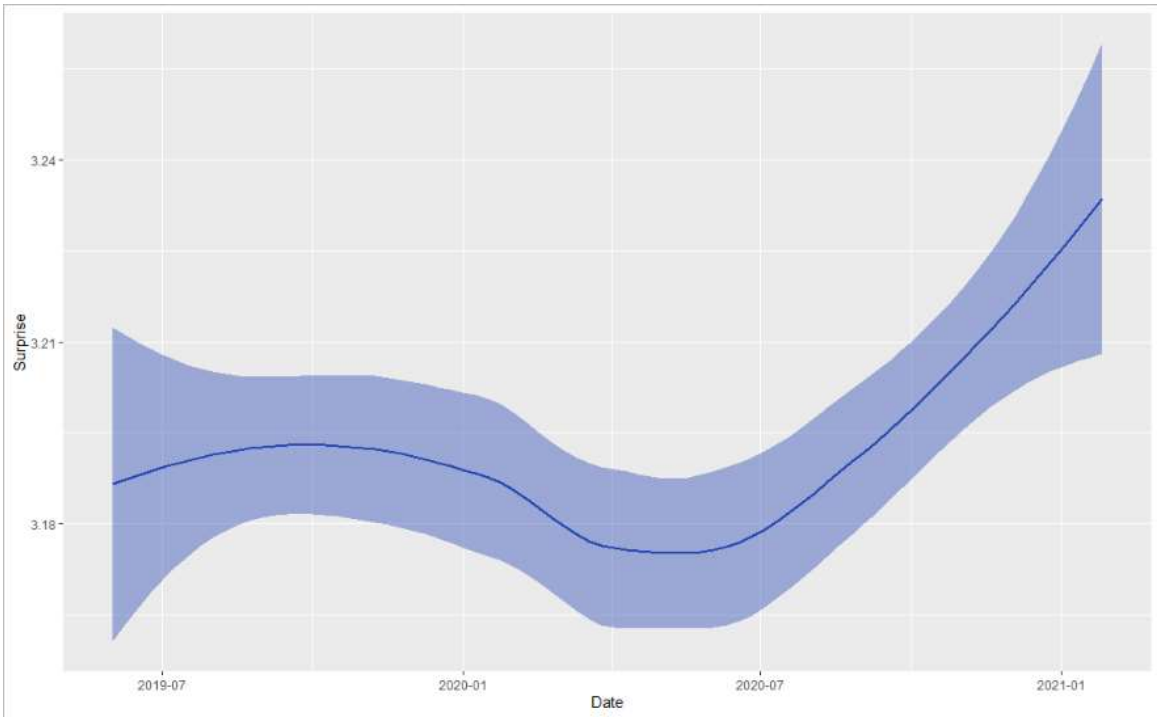
Παρουσίαση των Αποτελεσμάτων και Αξιολόγησή τους

Παρουσίαση Γραφημάτων με Συναισθήματα χρησιμοποιώντας την απόσταση Hamming

Τα παρακάτω γραφήματα έχουν χρησιμοποιήσει τα δεδομένα από το excel αρχείο "mood_of_twitter_revised.xlsx" τα οποία αποτελούν και το τελικό αρχείο στο αποθηκευόταν το συναίσθημα της κάθε ημέρας.

Στα παρακάτω γραφήματα φαίνεται η βασική πορεία των συναισθημάτων με την πιο σκούρα γραμμή ενώ το μέγεθος της ζώνης γύρω της υποδεικνύει εάν οι γειτονικές τιμές είχαν μεγάλη η μικρή διακύμανση πχ. τα συναισθήματα Θυμού , Χαράς, Έντονης Δυσαρέσκειας έχουν μικρή διακύμανση δλδ μικρή ζώνη ενώ τα

συναισθήματα του Φόβου, της Έκπληξης και της Θλίψης έχουν μεγάλη διακύμανση ως εκ τούτου και μεγαλύτερη ζώνη.

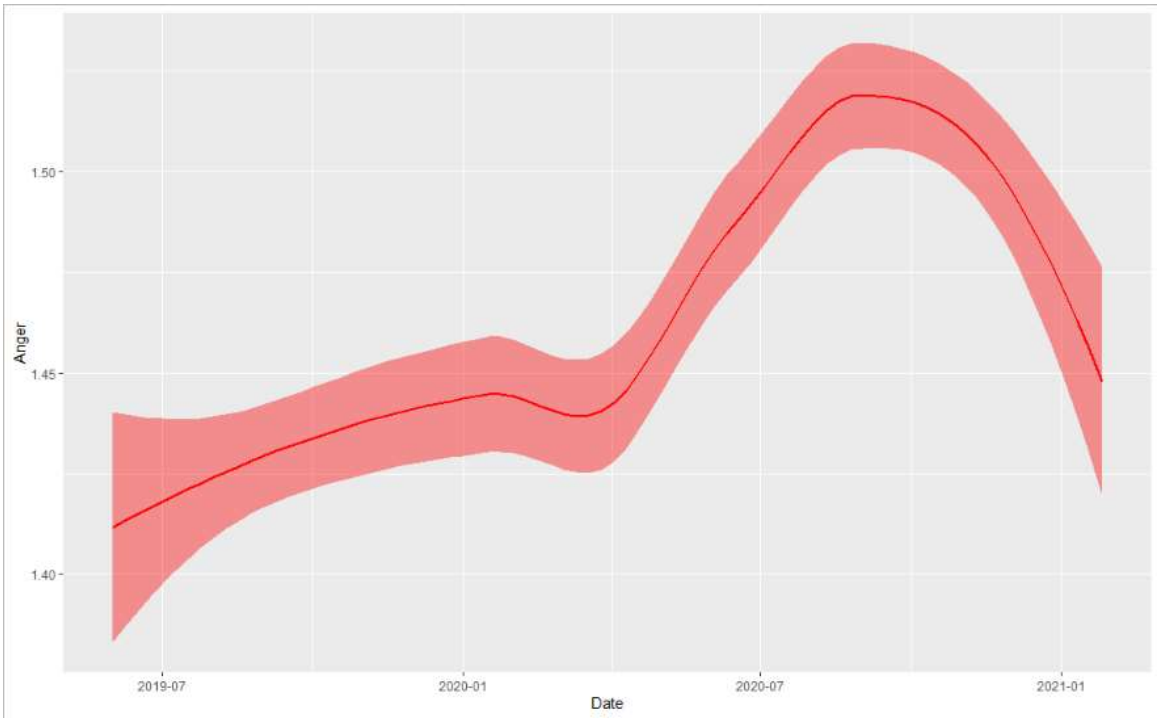


Εικόνα 3: Συναίσθημα Έκπληξης (Hamming)

Το συναίσθημα της έκπληξης όπως απεικονίζεται στο γράφημα ακολουθεί μια σταθερή πορεία από τις Εθνικές Εκλογές τον Ιούλιο του 2019 έως και τις πρώτες ανακοινώσεις για τον COVID-19 τον Ιανουάριο του 2020. Μετά τον Ιανουάριο του 2020 υπάρχει άνοδος έως και τον Ιούλιο του 2020 όπου πλέον και η πανδημία έχει μπει στην καθημερινότητα μας.

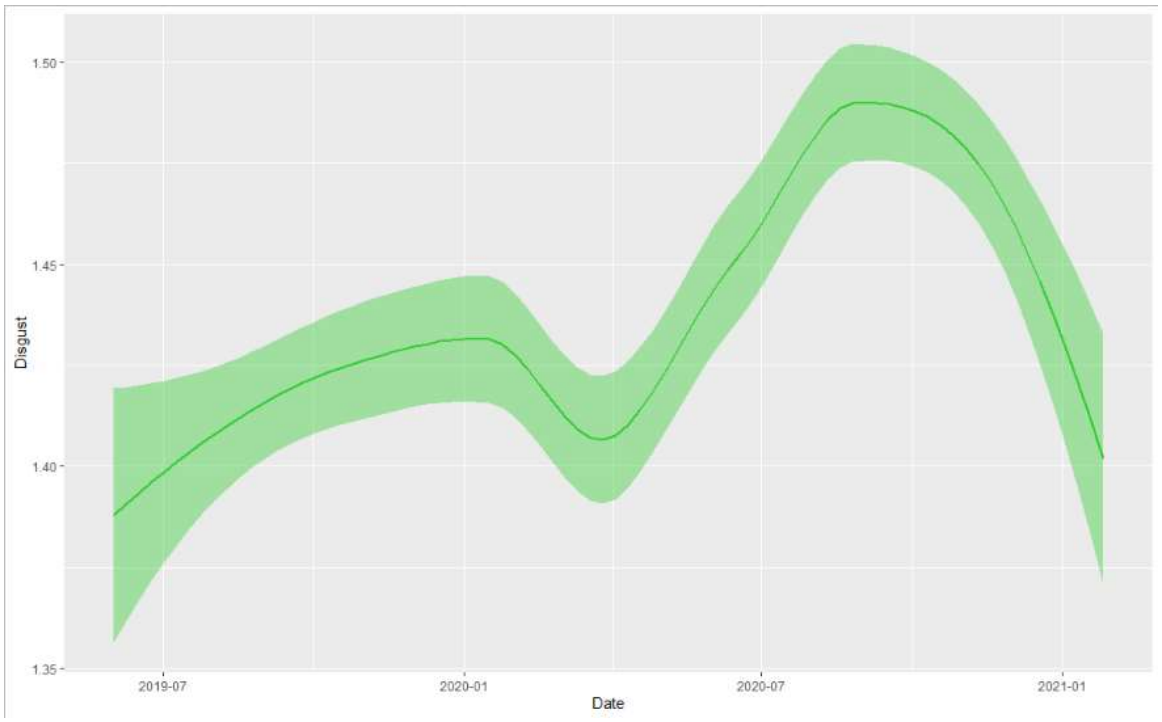
Το συναίσθημα του Θυμού δείχνει μια στασιμότητα με ελαφρά ανοδική απορία από τον Ιούλιο του 2019 έως και τον Μάρτιο του 2020 (1η Καραντίνα) από εκεί κλιμακώνεται σημαντικά έως Οκτώβριο με

Νοέμβριο του 2020 και .



Εικόνα 4: Συναίσθημα Θυμού (Hamming)

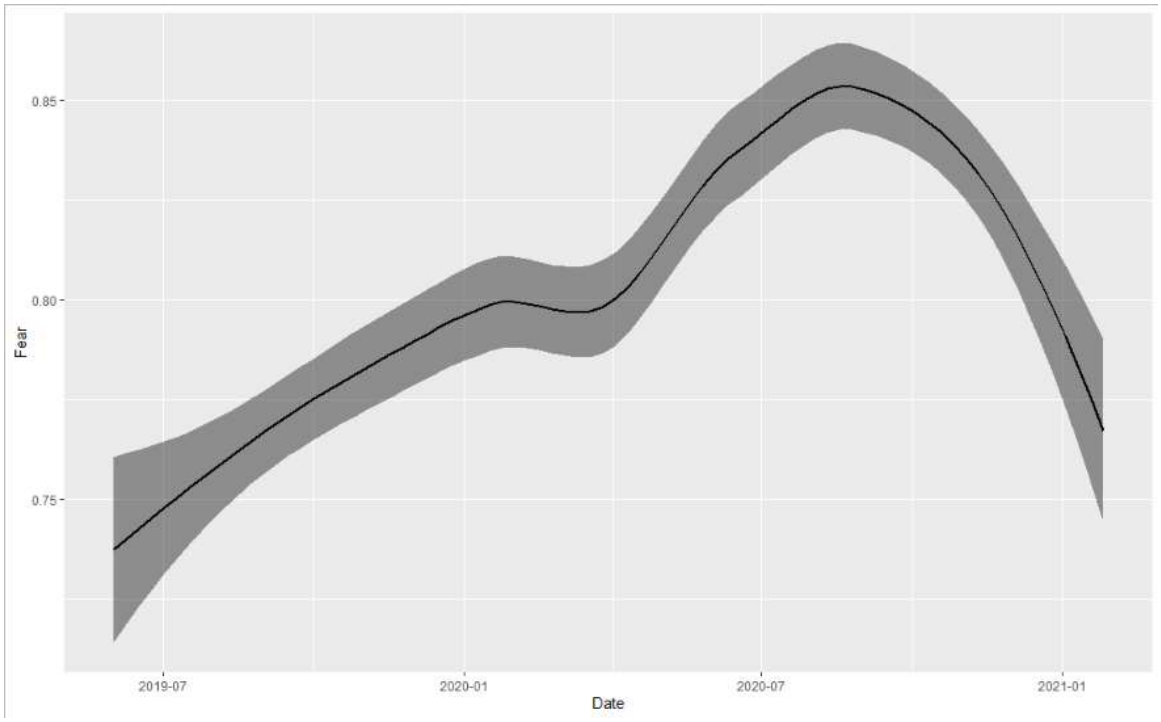
Το συναίσθημα του Θυμού δείχνει μια στασιμότητα με ελαφρά ανοδική απορία από τον Ιούλιο του 2019 έως και τον Μάρτιο του 2020 (1η Καραντίνα) από εκεί κλιμακώνεται σημαντικά έως τον Οκτώβριο και Νοέμβριο του 2020 και .



Εικόνα 5: Συναίσθημα Έντονης Δυσαρέσκειας (Hamming)

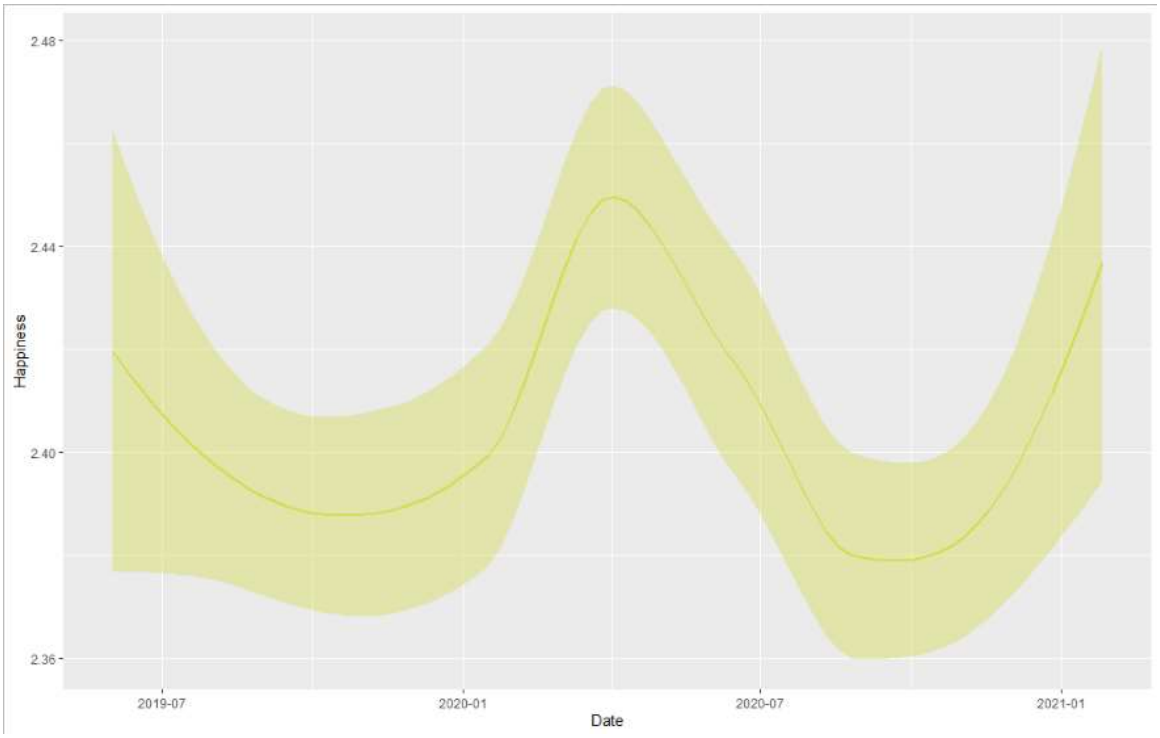
Το συναίσθημα της δυσαρέσκειας ακολουθεί μια ήπια άνοδο από τον Ιούλιο του 2019 έως και τον Ιανουάριο του 2020 μετά διαπιστώνουμε σημαντική πτώση μέχρι τον Απρίλιο του 2020 και από την άρση της 1ης Καραντίνας ακολουθεί ανοδική πορεία έως το Νοέμβριο του 2020 από όπου αρχίζει να αποκλιμακώνεται έως τον Ιανουάριο του 2021.

Εικόνα 6: Συναίσθημα Φόβου (Hamming)



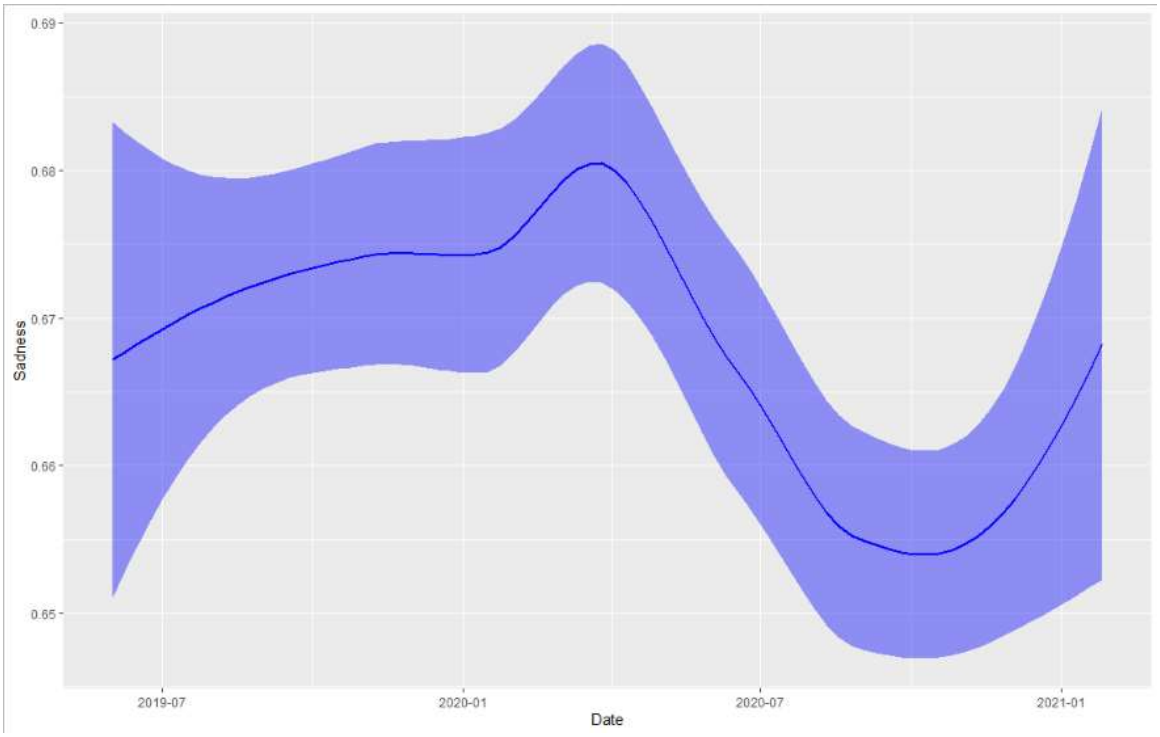
Ο φόβος όπως αποτυπώνεται στο παραπάνω γράφημα επίσης από τον Ιούλιο του 2019 είναι σταθερά ανοδικός έως τον Ιανουάριο του 2020 όπου υπάρχουν και οι πρώτες ανακοινώσεις από τον Παγκόσμιο Οργανισμό Υγείας για κήρυξη του πλανήτη σε καθεστώς πανδημίας όπου δείχνει στάσιμος έως τον Απρίλιου του 2020 και συνεχίζει να αυξάνεται με αποκορύφωμα τα μέσα Ιουλίου και Αύγουστο του 2020 φτάνει στη μέγιστη τιμή του. Από το Σεπτέμβριο του 2020 έως και τον Ιανουάριο του 2021 είναι σε σταθερά πτωτική τάση.

Εικόνα 7: Συναίσθημα Χαράς (Hamming)



Το συναίσθημα της χαράς έχει πτωτική τάση έως τον Ιανουάριο του 2020 έπειτα υπάρχει απότομη αύξηση με μέγιστο σημείο τον Απρίλιο του 2020 μετά συνεχίζει πτωτικά έως τον Οκτώβριο του 2020 και αυξάνεται ξανά από το Νοέμβριο του 2020 έως και τον Ιανουάριο του 2021.

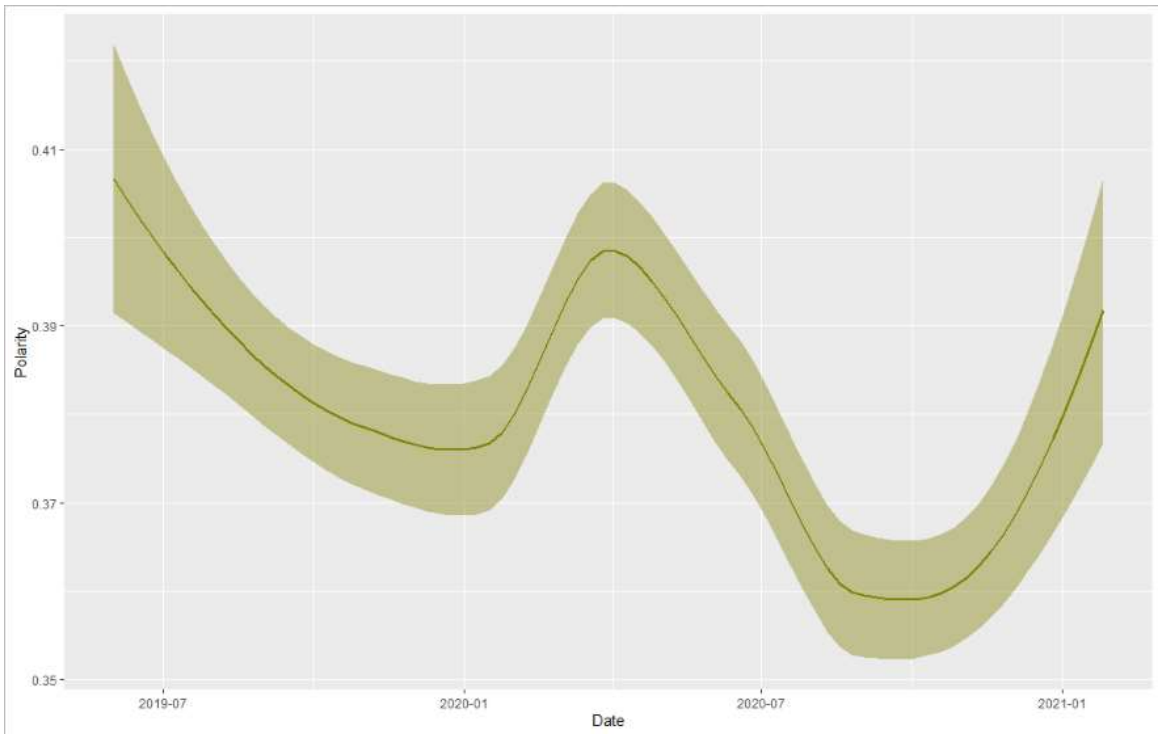
Εικόνα 8: Συναίσθημα Θλίψης (Hamming)



Το συναίσθημα της θλίψης σημειώνει τις περισσότερες σημαντικές μεταβολές. Από τον Ιούνιο 2019 έως τον Ιανουάριο του 2020 είναι σταθερή με ελαφρά ανοδική πορεία με αποκορύφωμα τον Απρίλιο του 2020. Μετά όμως μειώνεται απότομα έως τον Οκτώβριο του 2020 και αυξάνεται από την έναρξη της 2ης Καραντίνας το Νοέμβριο του 2020 έως και τον Ιανουάριο του 2021.

Πόλωση (Hamming) :

Επίσης υπάρχει και μέτρηση της Πόλωση με αρνητική Πόλωση το -1 και θετική Πόλωση το +1



Πίνακας 13: Πόλωση Hamming

Στο παραπάνω γράφημα υπάρχει πτώση έως τον Ιανουάριο του 2020 απότομη αύξηση έως τον Απρίλιο του 2020 και φτάνει στο ελάχιστο σημείο την περίοδο Σεπτεμβρίου με Νοέμβριου του 2020 από όπου αρχίζει και ξανά αυξάνεται έως τον Ιανουάριο του 2021.

Συσχέτιση Συναισθημάτων και Πόλωσης σε Hamming το έτος 2020:

Στον παρακάτω πίνακα παρουσιάζονται οι συντελεστές συσχέτισης των μεταβλητών που έχουν χρησιμοποιηθεί στην παραπάνω ανάλυση και αφορούν το έτος 2020. Σύμφωνα με τα αποτελέσματα είναι προφανές ότι οι περισσότερες τιμές των συντελεστών είναι μικρές και μέτριας συσχέτισης. Εμφανίζεται ισχυρή αρνητική συσχέτισή μεταξύ Θυμού, Χαράς και Έντονης Δυσαρέσκειας καθώς και Πόλωσης με Θυμό , Έντονη Δυσαρέσκειας και Φόβου (ανοιχτό κόκκινο χρώμα). Καθώς υπάρχει επίσης ισχυρή θετική συσχέτιση Θυμού , Έντονης Δυσαρέσκειας

και Έντονης Δυσαρέσκειας με Φόβο καθώς και Πόλωσης με Χαράς (ανοιχτό πράσινο χρώμα)

Πίνακας 14: Συσχέτιση Pearson για συναισθήματα 2020

	Anger	*Disgust*	*Fear*	*Happiness*	*Sadness*	*Surprise*	*Polarity*
Anger	1						
Disgust	0,835699	1					
Fear	0,694913	0,78279	1				
Happiness	-0,77062	-0,82577	-0,66883	1			
Sadness	0,127295	0,023319	0,101274	-0,154	1		
Surprise	-0,32462	-0,22923	-0,15962	0,513094	-0,25844	1	
Polarity	-0,81253	-0,90883	-0,78562	0,90115	0,035762	0,243606	1

Πίνακας 15 : Περιγραφική Στατιστική Συναισθημάτων Hamming 2020

Anger		*Disgust*		*Fear*	
Μέσος	1,48423012	*Μέσος*	1,45653568	*Μέσος*	
Τυπικό σφάλμα	0,00541631	*Τυπικό σφάλμα*	0,00598661	*Τυπικό σφάλμα*	
Διάμεσος	1,48216907	*Διάμεσος*	1,44865956	*Διάμεσος*	
Επικρατούσα τιμή	1,46455575	*Επικρατούσα τιμή*	1,3825164	*Επικρατούσα τιμή*	
Μέση απόκλιση τετραγώνου	0,10305237	*Μέση απόκλιση τετραγώνου*	0,11390308	*Μέση απόκλιση τετραγώνου*	
Διακύμανση	0,01061979	*Διακύμανση*	0,01297391	*Διακύμανση*	
Κύρτωση	0,99245219	*Κύρτωση*	0,88397861	*Κύρτωση*	
Ασυμμετρία	-0,05576526	*Ασυμμετρία*	0,23163384	*Ασυμμετρία*	
Εύρος	0,66198667	*Εύρος*	0,81979447	*Εύρος*	
Ελάχιστο	1,16030567	*Ελάχιστο*	1,09419805	*Ελάχιστο*	
Μέγιστο	1,82229234	*Μέγιστο*	1,91399252	*Μέγιστο*	
Άθροισμα	537,291304	*Άθροισμα*	527,265916	*Άθροισμα*	
Πλήθος	362	*Πλήθος*	362	*Πλήθος*	

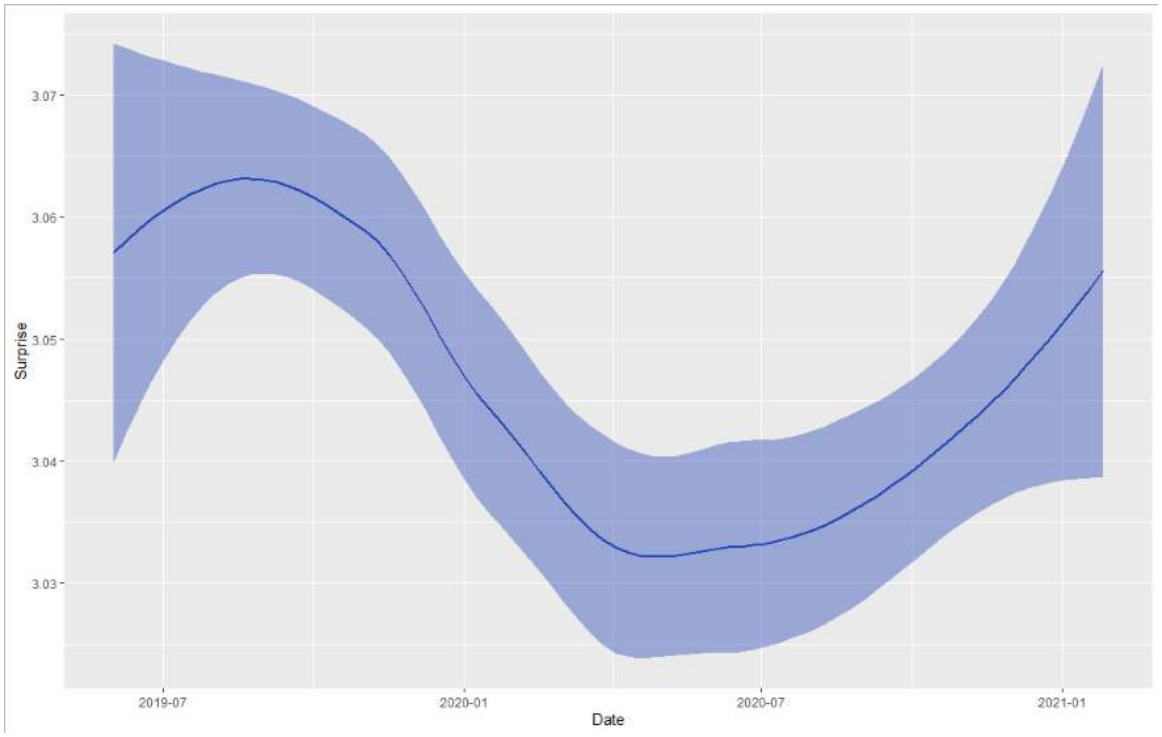
Happiness		*Sadness*		*Surprise*	
Μέσος	2,40731442	*Μέσος*	0,66388979	*Μέσος*	3,19218424
Τυπικό σφάλμα	0,00785958	*Τυπικό σφάλμα*	0,00282367	*Τυπικό σφάλμα*	0,00448109
Διάμεσος	2,4050815	*Διάμεσος*	0,66537192	*Διάμεσος*	3,18310768
Επικρατούσα τιμή	2,50729875	*Επικρατούσα τιμή*	0,61439475	*Επικρατούσα τιμή*	3,21293977
Μέση απόκλιση τετραγώνου	0,14953877	*Μέση απόκλιση τετραγώνου*	0,05372393	*Μέση απόκλιση τετραγώνου*	0,08525848
Διακύμανση	0,02236184	*Διακύμανση*	0,00288626	*Διακύμανση*	0,00726901
Κύρτωση	2,28318734	*Κύρτωση*	0,25801867	*Κύρτωση*	2,46223822
Ασυμμετρία	0,8350405	*Ασυμμετρία*	0,19549006	*Ασυμμετρία*	0,96570528
Εύρος	1,0205349	*Εύρος*	0,35291167	*Εύρος*	0,62888231
Ελάχιστο	1,95079365	*Ελάχιστο*	0,52279539	*Ελάχιστο*	2,90928393
Μέγιστο	2,97132855	*Μέγιστο*	0,87570706	*Μέγιστο*	3,53816624

Άθροισμα	871,44782	*Άθροισμα*	240,328102	*Άθροισμα*	1155,57069
Πλήθος	362	*Πλήθος*	362	*Πλήθος*	362

Πίνακας 16 : Περιγραφική Στατιστική Πόλωσης Hamming 2020

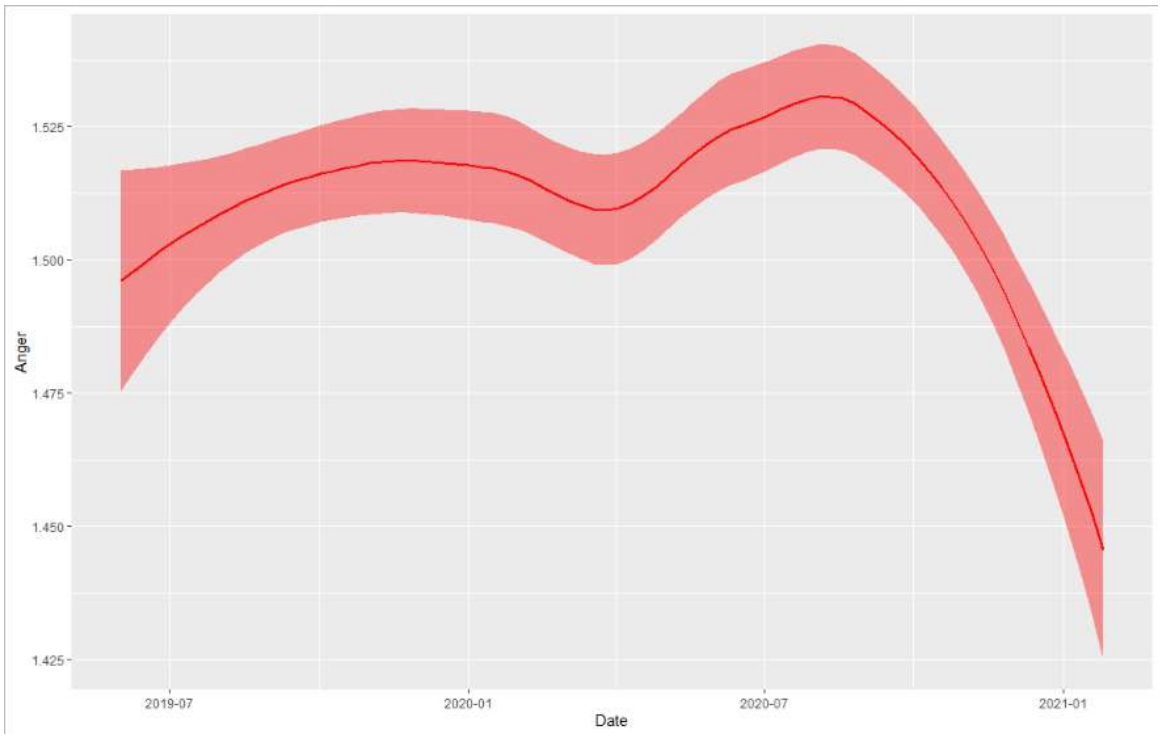
Polarity	
Μέσος	0,37382586
Τυπικό σφάλμα	0,00295675
Διάμεσος	0,37488875
Επικρατούσα τιμή	0,40560525
Μέση απόκλιση τετραγώνου	0,05625603
Διακύμανση	0,00316474
Κύρτωση	0,92885558
Ασυμμετρία	-0,20351756
Εύρος	0,3859078
Ελάχιστο	0,16227306
Μέγιστο	0,54818087
Άθροισμα	135,32496
Πλήθος	362

Παρουσίαση Γραφημάτων με Συναισθήματα χρησιμοποιώντας την απόσταση Jaro – Winkler:



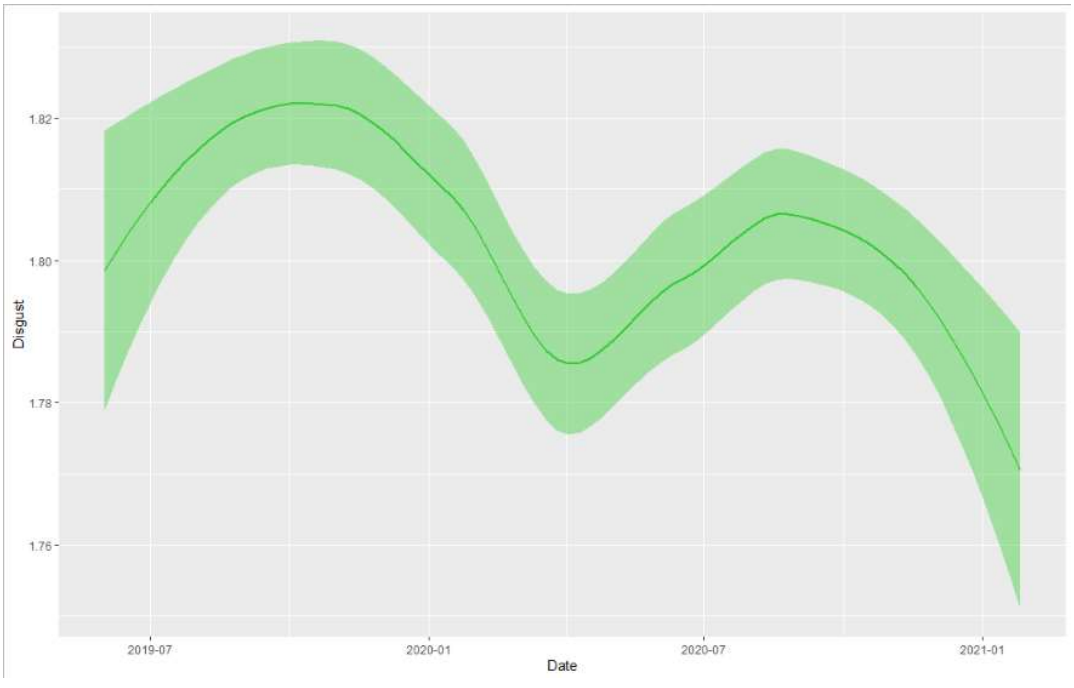
Εικόνα 9 : Συναίσθημα Έκπληξης (Jaro – Winkler)

Το συναίσθημα της Έκπληξης ακολουθεί μια ελαφρώς ανοδική πορεία από τον Ιούνιο του 2019 έως και τον Οκτώβριο του 2019 και μετά μειώνεται σταθερά έως τον Μάρτιο του 2020 από όπου αυξάνεται σταθερά έως τον Ιανουάριο του 2021.



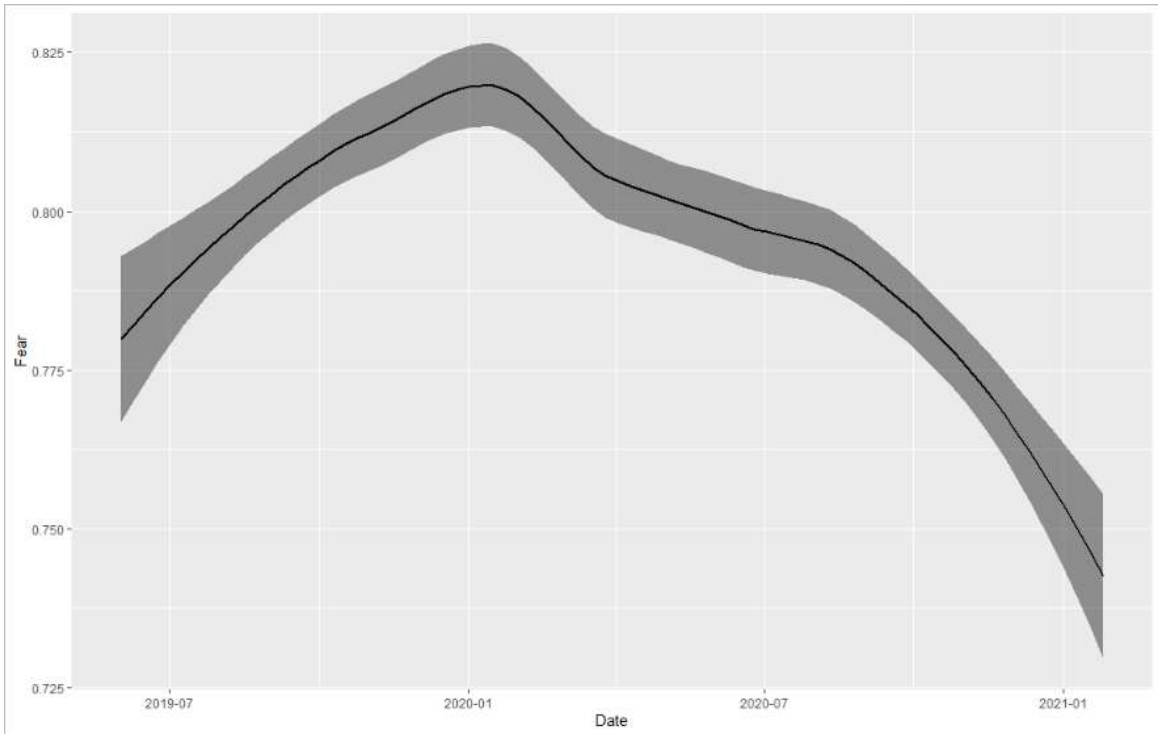
Εικόνα 10: Συναίσθημα Θυμού (Jaro – Winkler)

Το συναίσθημα του θυμού ακολουθεί σταθερή ήπια ανοδική πορεία από τον Ιούνιο του 2019 έως και τον Ιανουάριο του 2021 με μια ελαφριά κάμψη τον Απρίλιο του 2020 όπου ανεβαίνει με μεγαλύτερη κλίση έως τον Σεπτέμβριο του 2020 όπου βρίσκεται στο υψηλότερο σημείο και συνεχίζει να μειώνεται έως τον Ιανουάριο του 2021.



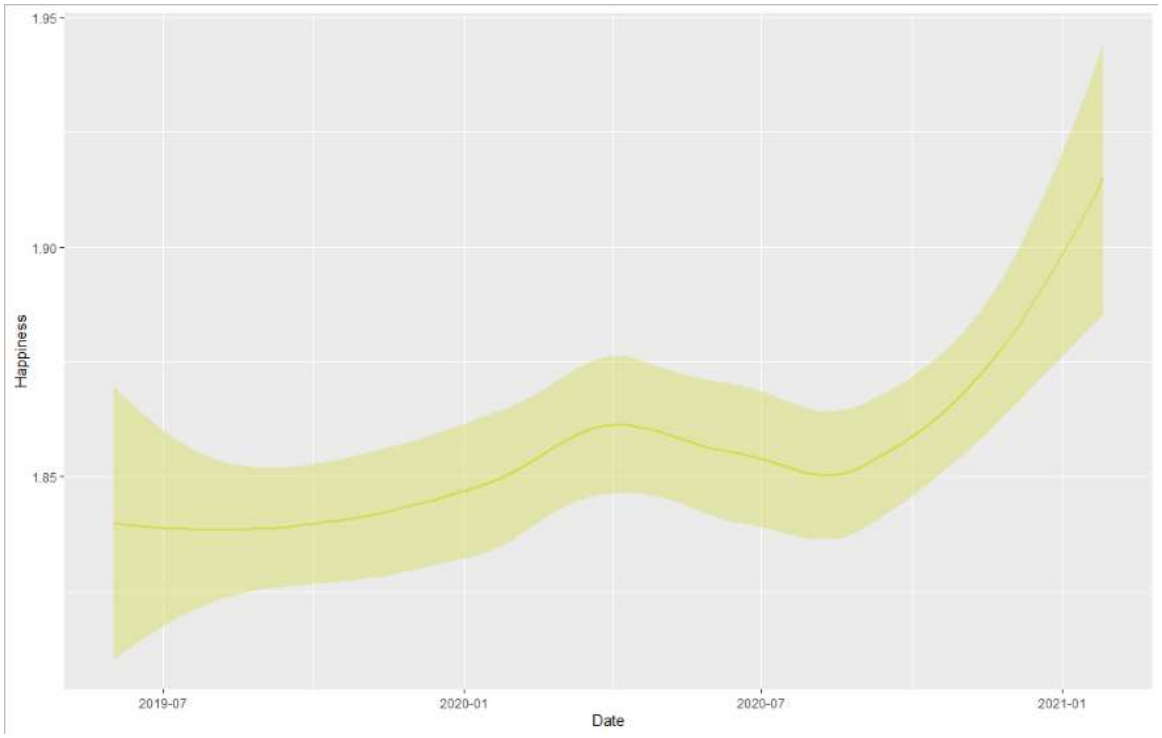
Εικόνα 11 : Συναίσθημα Έντονης Δυσαρέσκειας (Jaro – Winkler)

Το συναίσθημα της Έντονης Δυσαρέσκειας αυξάνεται από τον Ιούνιου του 2019 έως τον Οκτώβριο του 2020 από εκεί και έπειτα μειώνεται έως τον Απρίλιο του 2020. Από όπου αυξάνεται έως τον Οκτώβριο του 2020 και συνεχίζει μειούμενο έως σήμερα.



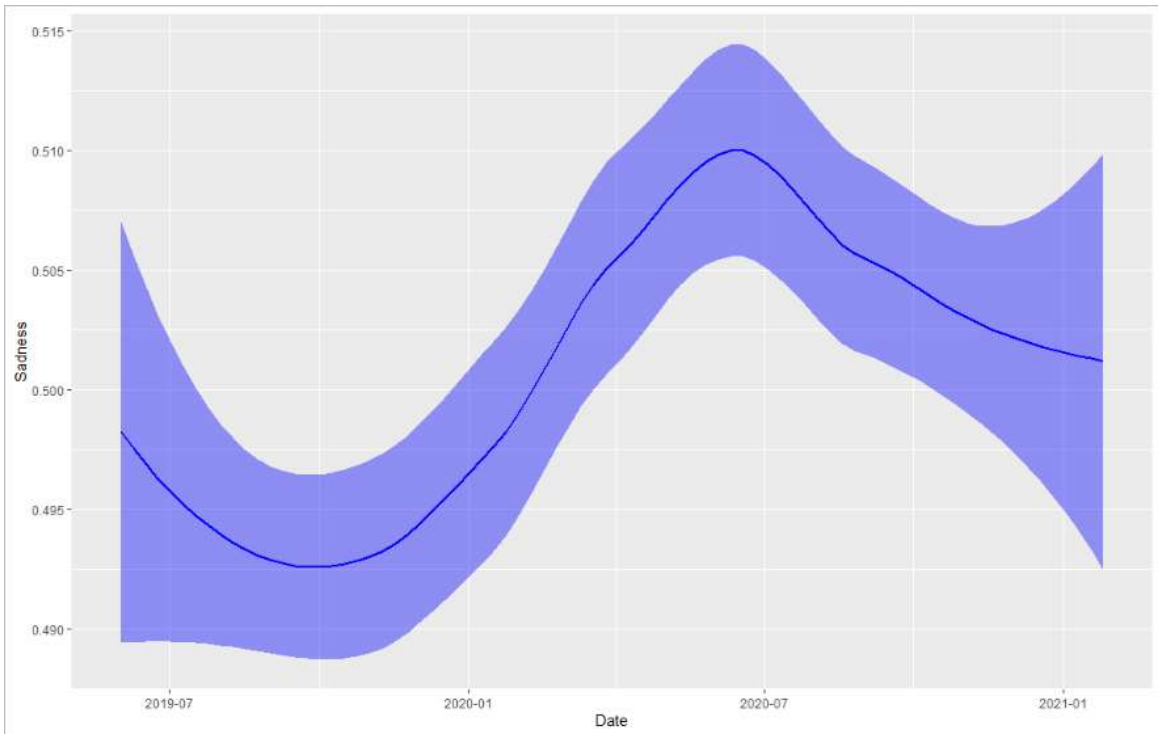
Εικόνα 12 : Συναίσθημα Φόβου (Jaro – Winkler)

Το αίσθημα του φόβου αυξάνεται το χρονικό διάστημα Ιουνίου 2019 έως και Ιανουάριο του 2020 έπειτα σημειώνει σημαντική πτώση έως και σήμερα.



Εικόνα 13 : Συναίσθημα Χαράς (Jaro – Winkler)

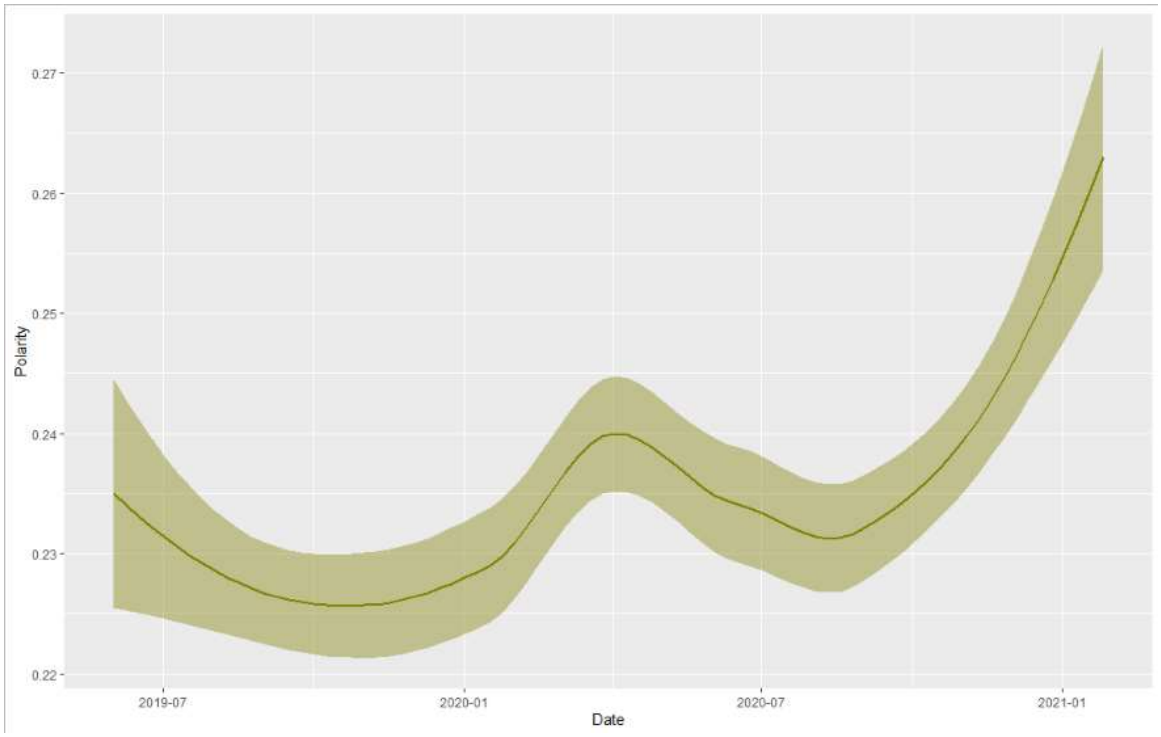
Το συναίσθημα της Χαράς είναι σταθερό από τον Ιούνιο του 2019 έως και τον Απρίλιο του 2020 αυξάνεται έως τον Απρίλιο του 2020 και μετά μειώνεται έως και Αύγουστο του 2020 και αυξάνεται έως τον Ιανουάριο του 2021.



Εικόνα 14 : Συναίσθημα Θλίψης (Jaro – Winkler)

Το συναίσθημα της Θλίψης είναι μειωμένο από τον Ιούλιο του 2019 έως και τον Οκτώβριο του 2019. Αυξάνεται με την έλευση της πανδημίας με αποκορύφωμα τον Ιούλιο του 2020 και μετά μειώνεται σταθερά έως και σήμερα

Πόλωση (Jaro - Winkler) :



Πίνακας 17: Πόλωση (Jaro - Winkler)

Η Πόλωση με βάση τη μετρική Jaro Winkler εμφανίζει πτώση έως τον Ιανουάριο του 2020 από όπου υπάρχει αύξηση έως τον Απρίλιο του 2020 μετά πτώση έως τον Οκτώβριο του 2020 και ξανά αύξηση έως τον Ιανουάριο του 2021.

Συσχέτιση Συναισθημάτων και Πόλωσης σε Jaro – Winkler το έτος 2020:

Στον παρακάτω πίνακα παρουσιάζονται οι συντελεστές συσχέτισης συναισθημάτων σε Jaro – Winkler που χρησιμοποιήθηκαν στην συγκεκριμένη ανάλυση. Σύμφωνα με τα αποτελέσματα τιμές Ισχυρή συσχέτιση εμφανίζουν το συναίσθημα της Χαράς και της Πόλωσης (ανοιχτό πράσινο χρώμα) και ισχυρή αρνητική συσχέτιση το συναίσθημα της Χαράς με το Θυμό και την Έντονη Δυσαρέσκεια ενώ και η Πόλωση με τον Θυμό (ανοιχτό κόκκινο χρώμα).

Πίνακας 18: Συσχέτιση Pearson Jaro Winkler 2020

	Anger	*Disgust*	*Fear*	*Happiness*	*Sadness*	*Surprise*	*Polarity*
Anger	1						
Disgust	0,607153	1					
Fear	0,519687	0,51056	1				
Happiness	-0,72913	-0,7064	-0,56176	1			
Sadness	0,208222	-0,04143	0,270946	-0,10333	1		
Surprise	-0,3727	0,30923	-0,2598	0,696483	-0,07178	1	
Polarity	-0,81309	-0,67629	-0,6543	0,824349	-0,25467	0,382826	1

Πίνακας 19: Περιγραφική Στατιστική Συναισθημάτων Jaro Winkler 2020

Anger		*Disgust*		*Fear*	
Μέσος	1,5153127	*Μέσος*	1,79944322	*Μέσος*	0,79561615
Τυπικό σφάλμα	0,00366027	*Τυπικό σφάλμα*	0,00340784	*Τυπικό σφάλμα*	0,00235777
Διάμεσος	1,51664499	*Διάμεσος*	1,80031135	*Διάμεσος*	0,79537812
Επικρατούσα τιμή	1,54627805	*Επικρατούσα τιμή*	1,76145176	*Επικρατούσα τιμή*	0,78206699
Μέση απόκλιση τετραγώνου	0,06964134	*Μέση απόκλιση τετραγώνου*	0,06483866	*Μέση απόκλιση τετραγώνου*	0,04485969
Διακύμανση	0,00484992	*Διακύμανση*	0,00420405	*Διακύμανση*	0,00201239
Κύρτωση	13,9635763	*Κύρτωση*	4,19944033	*Κύρτωση*	3,26852508
Ασυμμετρία	-0,84046214	*Ασυμμετρία*	-0,55677715	*Ασυμμετρία*	-0,5340348
Εύρος	0,93294068	*Εύρος*	0,62495768	*Εύρος*	0,37071904
Ελάχιστο	1,03583479	*Ελάχιστο*	1,46244389	*Ελάχιστο*	0,5652165
Μέγιστο	1,96877547	*Μέγιστο*	2,08740157	*Μέγιστο*	0,93593554
Άθροισμα	548,543199	*Άθροισμα*	651,398445	*Άθροισμα*	288,013047
Πλήθος	362	*Πλήθος*	362	*Πλήθος*	362
Happiness		*Sadness*		*Surprise*	
Μέσος	1,85893407	*Μέσος*	0,50282085	*Μέσος*	3,03848862
Τυπικό σφάλμα	0,00533019	*Τυπικό σφάλμα*	0,00150429	*Τυπικό σφάλμα*	0,00302053
Διάμεσος	1,84442614	*Διάμεσος*	0,5017774	*Διάμεσος*	3,03691263
Επικρατούσα τιμή	1,94964066	*Επικρατούσα τιμή*	0,48496798	*Επικρατούσα τιμή*	3,10273484
Μέση απόκλιση τετραγώνου	0,1014138	*Μέση απόκλιση τετραγώνου*	0,02862114	*Μέση απόκλιση τετραγώνου*	0,05746942
Διακύμανση	0,01028476	*Διακύμανση*	0,00081917	*Διακύμανση*	0,00330273
Κύρτωση	10,1447924	*Κύρτωση*	8,95256352	*Κύρτωση*	4,35362804
Ασυμμετρία	2,32863428	*Ασυμμετρία*	0,7845546	*Ασυμμετρία*	0,95383609
Εύρος	0,85865922	*Εύρος*	0,33795507	*Εύρος*	0,48794594
Ελάχιστο	1,63369421	*Ελάχιστο*	0,37107789	*Ελάχιστο*	2,87136427
Μέγιστο	2,49235343	*Μέγιστο*	0,70903295	*Μέγιστο*	3,3593102
Άθροισμα	672,934135	*Άθροισμα*	182,021146	*Άθροισμα*	1099,93288

Πλήθος	362	*Πλήθος*	362	*Πλήθος*	362

Πίνακας 20: Πόλωση Jaro - Winkler 2020

Polarity	
Μέσος	0,236201274
Τυπικό σφάλμα	0,001733574
Διάμεσος	0,235389314
Επικρατούσα τιμή	0,252338293
Μέση απόκλιση τετραγώνου	0,032983495
Διακύμανση	0,001087911
Κύρτωση	13,97967503
Ασυμμετρία	1,643208961
Εύρος	0,40507531
Ελάχιστο	0,069974729
Μέγιστο	0,475050039
Άθροισμα	85,50486111
Πλήθος	362

Συμπεράσματα και προτάσεις για Μελλοντική Έρευνα

Στο πλαίσιο της παρούσας διπλωματικής εργασίας εξετάστηκε το ζήτημα της προσέγγισης της πλατφόρμας κοινωνικής δικτύωσης Twitter, μέσω συγκομιδής δεδομένων και εφαρμογής τεχνικών επεξεργασίας φυσικής γλώσσας (Natural Language Processing) προκειμένου να πραγματοποιηθεί ανάλυση συναισθήματος.

Όσον αφορά τα αποτελέσματα της έρευνας παραθέτουμε τα θετικά συμπεράσματα παρακάτω:

- ✓ η συστηματική συλλογή δεδομένων και η επεξεργασία τους με βάση το Συναισθηματικό λεξικό μας δίνει την εικόνα των συναισθημάτων που επικράτησαν στο Ελληνικά tweets για την περίοδο 6/2019 έως 11/2020
- ✓ η δυνατότητα να διερευνήσουμε σε βάθος στον τομέα σχετικά με την ανάλυση κειμένων και συγκεκριμένα με την συναισθηματική ανάλυση και
- ✓ γίνεται αντιληπτό ότι η ιδιαιτερότητα της Ελληνικής γλώσσας μας ωθεί στο να εξετάσουμε περισσότερες επιλογές για τις συγκριτικές αποστάσεις των λέξεων

Αναφορικά με τις αρνητικές πτυχές της εργασίας:

- × Δεν χρησιμοποιήθηκε μοντέλο σχετικά με τα μέρη του λόγου αναφορικά με Υποκείμενα , Ρήματα , Αντικείμενα , Ουσιαστικά , Επίθετα παρά μόνο η αντιστοίχιση λέξεων και συλλογή των αντίστοιχων συναισθηματικών αξιών
- × Ακόμα και με τη χρήση διαφορετικών μετρικών αποστάσεων ανάμεσα σε λέξεις δεν ταυτοποιήθηκαν ακριβώς όλες οι λέξεις και βασικές αιτίες τις καταλήξεις και η έλλειψη ορθογραφίας των χρηστών

Το πεδίο των μέσων κοινωνικής δικτύωσης αποτελεί ένα καινούριο αντικείμενο μελέτης με πολλές προεκτάσεις και οι έρευνες που υπάρχουν για αυτό είναι σχετικά περιορισμένες. Με την έρευνα που πραγματοποιήθηκε στα πλαίσια της

συγκεκριμένης εργασίας θα μπορούσαν να γίνουν ορισμένες προτάσεις για μελλοντική έρευνα όπως είναι:

i. Η Εισαγωγή των emojis ως παράμετρο της ανάλυσης κειμένων
ii. Η Χρήση μοντέλων βαθιάς εκμάθησης για την παραγωγή αξιόπιστων μοντέλων συναισθηματικής ανάλυσης και
iii. Η Παραγωγή απόστασης που θα είναι πιο κατάλληλη για γλώσσες, όπως είναι η Ελληνική.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Aladwani Adel M., 2015 Facilitators, characteristics, and impacts of Twitter use: Theoretical analysis and empirical illustration, International Journal of Information Management 35 (2015) 15–25
2. Isaac Mike και Sydney Ember (2016). "For Election Day Influence, Twitter Ruled Social Media". The New York Times https://www.nytimes.com/2016/11/09/technology/for-election-day-chatter-twitter-ruled-social-media.html
3. Jungherr, Andreas (2016). "Twitter Use in Election Campaigns: A Systematic Literature Review". Journal of Information Technology & Politics, Volume 13 Issue 172–91, doi: 10.1080/19331681.2015.1132401

Αναφορές

Adam Tsakalidis, S. P. R. V. K. I. C. B. A. I. C. M. L. &. Y. K., 2018. Building and evaluating resources for sentiment analysis in the Greek language. *Springer - Lang Resources & Evaluation*, 1 December, p. 1021–1044.

Aladwani, A. M., 2015. Facilitators, characteristics and impacts of Twitter use: Theoretical Analysis and Emperical illustrations. *International Journal of Information Management*, pp. 15-25.

Ekman, P., 1992. An Argument for Basic emotions. *Cognition and Emotion*, pp. 169-200.

Mäntylä, M., Graziotin, D. & Kuutila, M., 2016. The Evolution of Sentiment Analysis - A Review of Research Topics. *Computer Science Review*, 05 12.

Τριανταφυλλίδης, Ί., 1998. ΛΕΞΙΚΟ ΤΗΣ ΚΟΙΝΗΣ ΝΕΟΕΛΛΗΝΙΚΗΣ.

Διαδίκτυο

1. https://esrc.ukri.org/research/impact-toolkit/social-media/twitter/what-is-twitter/
2. https://en.wikipedia.org/wiki/Twitter
3. https://ec.europa.eu/knowledge4policy/visualisation/number-social-media-users-worldwide-2010-17-forecasts-2021_en
4. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
5. https://trends.google.com/trends/explore?date=all&geo=US&q=sentiment%20analysis
6. https://link.springer.com/article/10.1007/s10579-018-9420-4
7. https://monkeylearn.com/sentiment-analysis/
8. https://twitter.com
9. https://wikipedia.com
10. https://instagram.com
11. https://facebook.com
12. https://en.wikipedia.org/wiki/United_Express_Flight_3411_incident
13. https://en.wikipedia.org/wiki/Sentiment_analysis
14. https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/?sh=d0192d615f64
15. https://github.com/MKLab-ITI/greek-sentiment-lexicon
16. https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/index.html
17. https://en.wikipedia.org/wiki/Cohen%27s_kappa
18. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
19. https://en.wikipedia.org/wiki/Local_regression

Παράρτημα Α

Παρακάτω παρατίθεται ο κώδικας σε R:

Mining Tweets:

```
#Mining Twitter by Location
#Installing rtweet via GitHub
#library(devtools)
#devtools::install_github("mkearney/rtweet")
#install.packages("rworldmap")
#using the right libraries
library(rtweet)
#Autorization for Twitter's  API
## access token method: create token and save it as an environment
variable
create_token(
  app = "UnderLens",
#Coordinates 10 Greek Cities
Cities<- c("Athens", "Thessaloniki", "Patras", "Heraklion",
"Larissa","Volos", "Rhodes", "Ioannina","Chania","Agrinio","Xanthi")
Latitude <-
c(37.983810,40.640064,38.246639,35.338734,39.638779,39.360519,36.434
052,39.674530,35.513828,38.624474,41.1349)
Longtitude <-
c(23.727539,22.944420,21.734573,25.144213,22.415979,22.945320,28.217
638,20.840210,24.018038,21.409594,24.8880)
Top10<- data.frame(Cities, Latitude,Longtitude)
a<-list.files(path = paste(getwd(),"/Tweets TXT", sep = ""))
last_day<-a[[length(a)-1]]
cat("The last tweet file is:", last_day)
print(Top10$Cities)
print("At first type No of City , pick days(<=7) and radius in miles
metric")
u<-scan()
```

```
coords<-
paste(Top10$Latitude[u[1]],Top10$Longtitude[u[1]],paste(u[3],"mi",se
p = ""),sep = ",")
print(coords)
#User is picking his days
  if(u[2]==1){cat("In about 1 minute the proccedure will be
completed")
  } else {
    cat("Estimated Time for Accomplishment:", (u[2]-1)*13+1,
"minutes",Sys.Date(), sep = " ")
  }
  for(i in 1:as.integer(u[2])){
    #Specifying dates of mining
    since<-Sys.Date()-u[2]+i-1
    until <- since+1
    #Mining Tweets
    tweets <- rtweet::search_tweets('lang:el' ,  n = 18000,
retryonratelimit = TRUE, include_rts = FALSE , geocode = coords,
since = since, until = until )
    #Storing Tweets
    date()
    since<-gsub("-",".",since)
    write(tweets[[5]], file =
paste(since,as.character(Top10$Cities[u[1]]), ".txt", sep = ""))
    write.csv2(tweets[[3]] , file =
paste("timestamps_",since,as.character(Top10$Cities[u[1]]), ".csv",
sep = ""))
    write.csv2(tweets[[5]] , file = paste(since,
as.character(Top10$Cities[u[1]]),".csv", sep = ""))
  }
rm(u)
```

Sentiment Analysis Script:

```
#Loading Libraries

library(tokenizers)

library(readxl)

library(readr)

library(stringr)

library(stringdist)

library(tm)

library(bpa)

library(openxlsx)

library(dplyr)

#Sentiment Analysis

#Import Sentiment Lexicon

Greek_Lexicon <- read_excel("Greek Sentiment Lexicons/Fixed_Greek_Lexicon.xlsx")

#Greek_Lexicon <- read_excel("C:/Users/nickr/Desktop/Projects/R_Twitter/Twitter/Greek Sentiment Lexicons/Greek_Lexicon.xlsx")

#View(Greek_Lexicon)

Greek_Lexicon<-as.data.frame(Greek_Lexicon)

Greek_Lexicon$Anger1<-as.numeric(as.character(Greek_Lexicon$Anger1))

Greek_Lexicon$Anger2<-as.numeric(as.character(Greek_Lexicon$Anger2))

Greek_Lexicon$Anger3<-as.numeric(as.character(Greek_Lexicon$Anger3))

Greek_Lexicon$Anger4<-as.numeric(as.character(Greek_Lexicon$Anger4))
```

```
Greek_Lexicon$Disgust1<-
as.numeric(as.character(Greek_Lexicon$Disgust1))

Greek_Lexicon$Disgust2<-
as.numeric(as.character(Greek_Lexicon$Disgust2))

Greek_Lexicon$Disgust3<-
as.numeric(as.character(Greek_Lexicon$Disgust3))

Greek_Lexicon$Disgust4<-
as.numeric(as.character(Greek_Lexicon$Disgust4))

Greek_Lexicon$Fear1<-as.numeric(as.character(Greek_Lexicon$Fear1))

Greek_Lexicon$Fear2<-as.numeric(as.character(Greek_Lexicon$Fear2))

Greek_Lexicon$Fear3<-as.numeric(as.character(Greek_Lexicon$Fear3))

Greek_Lexicon$Fear4<-as.numeric(as.character(Greek_Lexicon$Fear4))

Greek_Lexicon$Happiness1<-
as.numeric(as.character(Greek_Lexicon$Happiness1))

Greek_Lexicon$Happiness2<-
as.numeric(as.character(Greek_Lexicon$Happiness2))

Greek_Lexicon$Happiness3<-
as.numeric(as.character(Greek_Lexicon$Happiness3))

Greek_Lexicon$Happiness4<-
as.numeric(as.character(Greek_Lexicon$Happiness4))

Greek_Lexicon$Sadness1<-
as.numeric(as.character(Greek_Lexicon$Sadness1))

Greek_Lexicon$Sadness2<-
as.numeric(as.character(Greek_Lexicon$Sadness2))

Greek_Lexicon$Sadness3<-
as.numeric(as.character(Greek_Lexicon$Sadness3))

Greek_Lexicon$Sadness4<-
as.numeric(as.character(Greek_Lexicon$Sadness4))
```

```
Greek_Lexicon$Surprise1<-
as.numeric(as.character(Greek_Lexicon$Surprise1))

Greek_Lexicon$Surprise2<-
as.numeric(as.character(Greek_Lexicon$Surprise2))

Greek_Lexicon$Surprise3<-
as.numeric(as.character(Greek_Lexicon$Surprise3))

Greek_Lexicon$Surprise4<-
as.numeric(as.character(Greek_Lexicon$Surprise4))

#View(Greek_Lexicon)

#Cleaning Vowels Greek_Lexicon

grclean1<-c()

grclean2<-c()

grclean3<-c()

grclean4<-c()

grclean5<-c()

grclean6<-c()

grclean7<-c()

grclean8<-c()

grclean9<-c()

grclean10<-c()

for (k in 1:length(Greek_Lexicon[[1]])) {

  grclean1[[k]]<-gsub("ά","α",Greek_Lexicon[[1]][k])

  grclean2[[k]]<-gsub("έ","ε",grclean1[[k]])

  grclean3[[k]]<-gsub("ή","η",grclean2[[k]])

  grclean4[[k]]<-gsub("ί","ι",grclean3[[k]])
```

```
  grclean5[[k]]<-gsub("ό","ο",grclean4[[k]])
  grclean6[[k]]<-gsub("ύ","υ",grclean5[[k]])
  grclean7[[k]]<-gsub("ώ","ω",grclean6[[k]])
  grclean8[[k]]<-gsub("ϋ","υ",grclean7[[k]])
  grclean9[[k]]<-gsub("ϊ","ι",grclean8[[k]])
  grclean10[[k]]<-gsub("ΐ","ι",grclean9[[k]])
}
#View(Greek_Lexicon)
#Loading Data
#txt or csv
#data_csv<-read.csv(file.choose(),sep=",",skipNul = TRUE)
fileslist <- choose.files()
fileslist <-sort(fileslist)
mood<-matrix(nrow = length(fileslist), ncol = 7)

for(e in 1:length(fileslist)){
  data<-read_lines(fileslist[e], skip_empty_rows = TRUE, progress =
TRUE)
  #Checking encoding of the document
  encoding_check<-stringi::stri_enc_detect(data)
  if(encoding_check[[1]][1,1]!="UTF-8"                              ||
encoding_check[[2]][1,1]!="UTF-8" ){
    #Changing encoding from ISO-8859-7 to UTF-8
    c<-encoding_check[[1]][1,1]
    #Renaming the file
```

```
    l<-substr(fileslist[e],nchar(fileslist[e])-
19,nchar(fileslist[e]))

    b<-paste("fixed_encoding_UTF_8_",l,sep = "")

    file.create(b)

    #Fixing the encoding and writing a new file

    writeLines(iconv(readLines(fileslist[e]), from ="ISO-8859-7" , to
= "UTF8"),file(b, encoding="UTF-8"))

    data<-read_lines(b, skip_empty_rows = TRUE, progress = TRUE)

    writeLines(data,file(b))

  } else if (encoding_check == ""){

    next}

  #Setting Stop Words and Text cleaning

  greek_stop_words<-
c("εκει","https","άλλους","άλλο","καν","ειστε","άλλη","κάποια","πάνω
","κάτω","t.co","u","0001f92a","εχεις","αλλα","άλλα","τι","κατά","γι
ατι","γιατί","αλλά","ως","μέσα","ειχε","όπως","όλο","ο","α","β","γ",
"δ","ε","ζ","η","θ","ι","κ","λ","μ","ν","ξ","ο","π","ρ","σ","τ","υ",
"φ","χ","ψ","ω","a","b","c","d","e","f","να","ναι","μας","τετοιες","
ήταν","ηταν","αυτο","ας","εγω","εχει","ή","η","εκεί","και","λίγο","λ
ιγο","πάλι","μονο","απ","μόνο","αυτά","αυτή","αυτα","αυτη","εγώ","ού
τε","υπάρχει","-
","κάνει","στους","κάθε","πρέπει","τώρα","λέει","όχι","ήταν","amp","
δύο","σαν","το","να","για","του","είναι","ειναι","στις","έχω","μετά"
,"μη","κάτι","είσαι","πολύ","σήμερα","καλημέρα","όλα","ολα","όλοι","
ολοι","όλες","ολες","πολυ","πολλή","πολλά","πολλη","πολλα","την","με
","του","της","τα","που"        ,        "στο","είναι",        "θα",
"τον","σε","από","απο", "μου","στην","οι", "τους","μας","τη", "των",
"στη","στα","τις",      "ότι","οτι",      "σου","στον","αλλά","μια",
"τι","αν","σας","έχει","ένα","αυτό","δε","όταν","κι",        "γιατί",
"πως","πιο", "μην", "έχουν", "ρε","μόνο")
```

```
  stop_words          <-          append(greek_stop_words          ,
stopwords::stopwords(language = "en"))

  #removing urls

  url_pattern          <-          "http[s]?://(?:[a-zA-Z]|[0-9]|[$-
_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+"

  data<-gsub(pattern = url_pattern,"",data)

  #removing english strings

  data<-gsub("[a-z]|[A-Z]","",data)

  #removing emojis

  data<-gsub("<U|[$-_@.&+]|[0-9]|F|[0-9]>","",data)

  #Pops up a window to choose the txt file we want

  #Tokenization techniques

  tokens <-tokenize_words(data , lowercase = TRUE, stopwords =
stop_words , strip_punct = TRUE , strip_numeric = TRUE)

  #Cleaning vowels tokens

  clean1<-c()

  clean2<-c()

  clean3<-c()

  clean4<-c()

  clean5<-c()

  clean6<-c()

  clean7<-c()

  clean8<-c()

  clean9<-c()

  clean10<-c()
```

```
  clean1<-gsub("ά","α",unlist(tokens))#  We  unlist  the  tokens  to
"clean"
  clean2<-gsub("έ","ε",clean1)
  clean3<-gsub("ή","η",clean2)
  clean4<-gsub("ί","ι",clean3)
  clean5<-gsub("ό","ο",clean4)
  clean6<-gsub("ύ","υ",clean5)
  clean7<-gsub("ώ","ω",clean6)
  clean8<-gsub("ϋ","υ",clean7)
  clean9<-gsub("ϊ","ι",clean8)
  clean10<-gsub("ΐ","ι",clean9)
  count<- 0
  clean_tokens<-tokens
  for(f in 1:length(clean_tokens)){
    for(g in 1:length(clean_tokens[[f]])){
      count<-count+1
      clean_tokens[[f]][g]<-clean10[count]
    }
  }
  #General
  #Setting Up Variables for Outputs of the code
  all_sentiments<-matrix(nrow = length(clean_tokens), ncol = 7)
  o<-c()
  s<-0
```

```
  z<-c()
  a<-0
  #Select the String Matching Technique
  str_match_meth<-"jw"
  #Starting to string match for each tweet's tokens and assigning sentiment values according to Sentiment Lexi
  for (i in 1:length((clean_tokens)))   {
    o<-c()
    s<-0
    z<-c()
    a<-0
    #Pick a tweet
    y<-i
    for (p in 1:length(clean_tokens[[y]])) {
      a<-nchar(clean_tokens[[y]][p])
      if(is.na(a)!=TRUE){
        #String Matching Function amatch() returns the position of word in grclean10 which is the First Column of Greek Sentiment Lexi
        o[p]<-amatch(clean_tokens[[y]][p],   grclean10,   method   = str_match_meth, nomatch = 0)
      }
    }
    #Controlling the matches if there are null for each word
    m<-c()
    m<-which(o!=0,arr.ind = TRUE )
```

```
if(length(m)==0){
  emotions <-matrix(0L,nrow=1, ncol=7)
} else {
  emotions<-matrix(0L,nrow= length(m),ncol = 7)
  for (j in 1:length(m)) {
    t<-o[m[j]]
    emotions[j,1]<-max(Greek_Lexicon[t,14:17])
    emotions[j,2]<-max(Greek_Lexicon[t,18:21])
    emotions[j,3]<-max(Greek_Lexicon[t,22:25])
    emotions[j,4]<-max(Greek_Lexicon[t,26:29])
    emotions[j,5]<-max(Greek_Lexicon[t,30:33])
    emotions[j,6]<-max(Greek_Lexicon[t,34:37])
    emotions[j,7]<-max(Greek_Lexicon[t,10:13])
    emotions[is.na(emotions)]<-0
  }
  for(h in 1:length(m)){
    for(v in 1:6){
      if (emotions[h,v]<3&is.na(emotions[h,v])!=TRUE){
        emotions[h,v]<-0
      } else {
        next
    }
  }
}
```

```
    emotions[is.na(emotions)]<-0

    only_emotions<-emotions[rowSums(is.na(emotions))                != ncol(emotions),]

      if(length(only_emotions)==7){all_sentiments[y,]<- only_emotions

  } else {

    all_sentiments[y,]<-c(colMeans(only_emotions))

  }

    }

  }

  column_names<- c("Anger","Disgust","Fear","Happiness","Sadness","Surprise","Polarit y")

  neutral_rows<-which(all_sentiments==0 , arr.ind = TRUE)

  #Mood of the Day including empty rows

  mood_of_the_day<-c(colMeans(all_sentiments))

  #Mood of the Day excluding empty rows

  all_sentiments[is.na(all_sentiments)]<-0

  all_sentiments[all_sentiments == 0]<-NA

  only_sentiments<-all_sentiments[rowSums(is.na(all_sentiments))   != ncol(all_sentiments),]

  only_sentiments[is.na(only_sentiments)]<-0

  mood_of_the_day_only_sentiments<-matrix(colMeans(only_sentiments), nrow = 1, ncol =7)

  mood[e,]<-mood_of_the_day_only_sentiments
```

```
  mood<-as.data.frame(mood)
  names(mood)<-column_names
  all_sentiments<-(as.data.frame(all_sentiments))
  names(all_sentiments)<-column_names
  all_sentiments[is.na(all_sentiments)]<-0
}
title<-paste("data_",str_match_meth,"revised.csv",sep = "")
write.csv2(mood, file = title)
#saveWorkbook(wb = hi,file = excelName)
#excelName<-paste(Sys.Date(),"moodTwitter.xlsx",sep = "_")
#hi<-createWorkbook()
#addWorksheet(hi, l)
#addWorksheet(hi, paste("data",l,sep = "_"))
#writeData(hi, sheet = l , x =all_sentiments)
#writeData(hi, sheet = paste("data",l,sep = "_") , x
=as.data.frame(data))
```

Finding Word Frequency:

```
#Goal is to find stop words / most frequent words without any impact
to sentiment analysis
library(tidytext)
library(tidyverse)
library(harrypotter)
library(tokenizers)
```

```
library(readxl)

library(readr)

library(stringr)

library(stringdist)

library(readxl)

library(openxlsx)

fileslist <- choose.files()

fileslist <-sort(fileslist)

mood<-matrix(nrow = length(fileslist), ncol = 6)

greek_vowels <-function(text){

  # We unlist the tokens to "clean"

  clean1<-gsub("ά","α",unlist(text))

  clean2<-gsub("έ","ε",clean1)

  clean3<-gsub("ή","η",clean2)

  clean4<-gsub("ί","ι",clean3)

  clean5<-gsub("ό","ο",clean4)

  clean6<-gsub("ύ","υ",clean5)

  clean7<-gsub("ώ","ω",clean6)

  clean8<-gsub("ϋ","υ",clean7)

  clean9<-gsub("ϊ","ι",clean8)

  clean10<-gsub("ΐ","ι",clean9)

}

for (i in fileslist) {
```

```
  data<-read_lines(fileslist[i], skip_empty_rows = TRUE, progress =
TRUE)

  data<-iconv(readLines(fileslist[i]), from ="ISO-8859-7" , to =
"UTF8")

  # Removing URLs , Emotes and English Letters

  url_pattern          <-          "http[s]?://(?:[a-zA-Z]|[0-9]|[$-
_@.&+]|[!*\\(\\),]|(?:%[0-9a-fA-F][0-9a-fA-F]))+"

  clean_data1<-gsub(pattern = url_pattern,"",data)

  clean_data2<-gsub("[a-z]|[A-Z]","",clean_data1)

  data<-gsub("<U|[$-_@.&+]|[0-9]|F|[0-9]>","",clean_data2)

  tokens <-tokenize_words(data , lowercase = TRUE, stopwords =
stop_words , strip_punct = TRUE , strip_numeric = TRUE)

  cltokens<-greek_vowels(tokens)

  Greek_Lexicon        <-        read_excel("Greek        Sentiment
Lexicons/Greek_Lexicon.xlsx")

  freq_table<-sort(table(cltokens), decreasing=T)

}
```

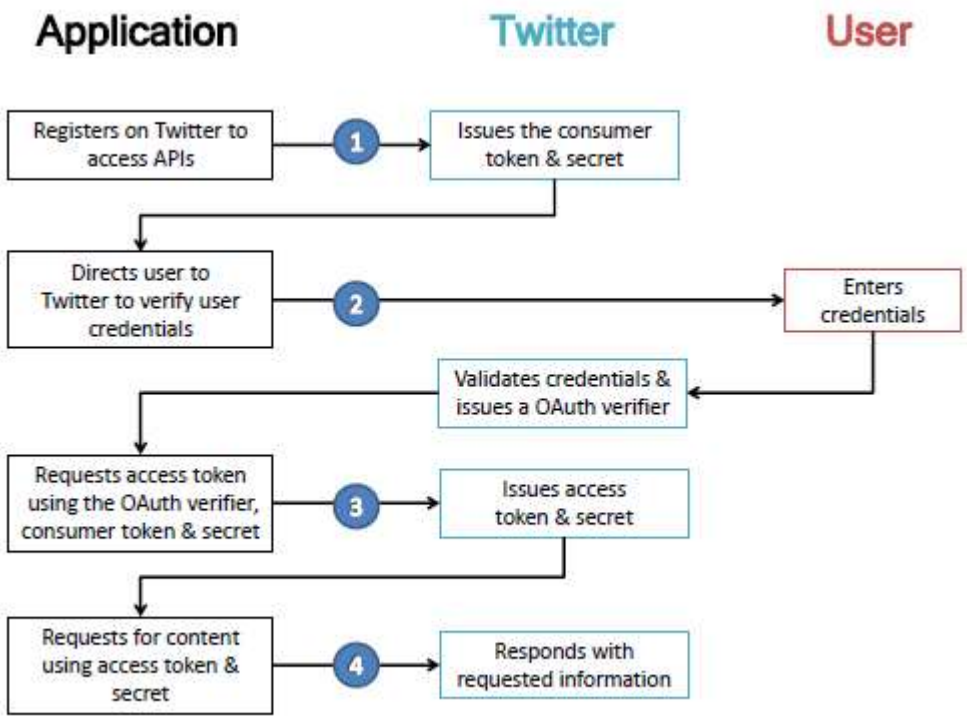
Data Visualization:

```
library(readxl)
library(ggplot2)
library(tidyr)
library(dplyr)
mood_of_twitter <- read_excel("mood_of_twitter_revised.xlsx")
#View(mood_of_twitter)
#Setting the data
data_sentiment<-mood_of_twitter
colors<-data.frame("#FF0000","#32CD32","#000000",
"#d1dd4a","#0000FF","#2243B6","#808000")
names(colors)<-c("Anger","Disgust","Fear","Happiness", "Sadness",
"Surprise","Polarity")
#Missing Dates
d <- as.Date(data_sentiment$Date)
```

```
date_range <- seq(min(d), max(d), by = 1)
missing_dates<- date_range[!date_range %in% d]
#Plot timeline data series with smoothing
p <- ggplot(data = data_sentiment , aes(x = Date, y = Disgust))
      geom_line(color = colors$Disgust, size = 1)
p + stat_smooth(color = colors$Disgust, fill = colors$Disgust,method
= "loess")
```

Παράρτημα Β

Πίνακας 2.1: Διάγραμμα Ροής OAuth



Πηγή:

Η ιστοσελίδα στην απευθύνεται ένας χρήστης για τη δημιουργία εφαρμογής και την απόκτηση πιστοποιητικών πρόσβασης είναι η https://developer.twitter.com

id	Greek Stop Words
1	εκει
2	https
3	άλλους
4	άλλο
5	άλλη
6	κάποια
7	πάνω
8	κάτω
9	t.co
10	u
11	0001f92a
12	εχεις
13	αλλα
14	άλλα
15	τι
16	κατά
17	γιατι
18	γιατί
19	αλλά
45	υ
46	φ
47	χ
48	ψ
49	ω
50	a
51	b
52	c
53	d
54	e
55	f
56	να
57	ναι
58	μας
59	τετοιες
60	ήταν
61	ηταν
62	αυτο
63	ας
64	εγω

20	ως	65	εχει
21	μέσα	66	ή
22	ειχε	67	η
23	όπως	68	εκεί
24	όλο	69	και
25	ο	70	λίγο
26	α	71	λιγο
27	β	72	πάλι
28	γ	73	μονο
29	δ	74	απ
30	ε	75	μόνο
31	ζ	76	αυτά
32	η	77	αυτή
33	θ	78	αυτα
34	ι	79	αυτη
35	κ	80	εγώ
36	λ	81	ούτε
37	μ	82	υπάρχει
38	ν	83	-
39	ξ	84	κάνει
40	ο	85	στους
41	π	86	κάθε

42	ρ	87	πρέπει
43	σ	88	τώρα
44	τ	89	λέει
		90	όχι

91	ήταν	116	πολυ
92	amp	117	πολλή
93	δύο	118	πολλά
94	σαν	119	πολλη
95	το	120	πολλα
96	να	121	την
97	για	122	με
98	του	123	του
99	είναι	124	της
100	ειναι	125	τα

101	στις	126	που
102	έχω	127	δεν
103	μετά	128	στο
104	μη	129	είναι
105	κάτι	130	θα
106	είσαι	131	τον
107	πολύ	132	σε
108	σήμερα	133	από
109	καλημέρα	134	απο
110	όλα	135	μου
111	ολα	136	στην
112	όλοι	137	οι
113	ολοι	138	τους
114	όλες	139	μας
115	ολες	140	τη
141	των	167	αααα
142	στη	168	ανω
143	στα	169	ααα
144	τις	170	αυτης
145	ότι	171	αυτου
146	οτι	172	γιαυτο

147 σου
148 στον
149 αλλά
150 μια
151 τι
152 αν
153 σας
154 έχει
155 ένα
156 αυτό
157 δε
158 όταν
159 κι
160 γιατί
161 πως
162 πιο
163 μην
164 έχουν
165 ρε
166 μόνο
173 γτ
174 δικος
175 καθενα
176 κτλπ
177 λα
178 μιλ
179 πο
180 σεις
181 στην
182 τους
183 τες
184 χαχαχαχαχα
185 χουμε