ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΙΔΙΚΕΥΣΗ: ΠΛΗΡΟΦΟΡΙΚΗ ΤΗΣ ΥΓΕΙΑΣ

# ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ (ΑΝΑΚΤΗΣΗ ΚΑΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ) ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ ΚΟΙΝΩΝΙΚΟ ΔΙΚΤΥΟ TWITTER ΑΝΑΦΟΡΙΚΑ ΜΕ ΤΗΝ ΨΥΧΟΛΟΓΙΑ ΤΟΥ ΠΛΗΘΥΣΜΟΥ ΣΤΗ ΠΑΝΔΗΜΙΑ COVID-19

**ΑΦΡΟΔΙΤΗΣ Ν. ΚΑΤΙΚΑ**

ΜΗΧΑΝΙΚΟΥ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΗΝΑ 2023

# ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ (ΑΝΑΚΤΗΣΗ ΚΑΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ) ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ ΚΟΙΝΩΝΙΚΟ ΔΙΚΤΥΟ TWITTER ΑΝΑΦΟΡΙΚΑ ΜΕ ΤΗΝ ΨΥΧΟΛΟΓΙΑ ΤΟΥ ΠΛΗΘΥΣΜΟΥ ΣΤΗ ΠΑΝΔΗΜΙΑ COVID-19

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ:

«ΟΡΓΑΝΩΣΗ ΚΑΙ ΔΙΟΙΚΗΣΗ ΥΠΗΡΕΣΙΩΝ ΥΓΕΙΑΣ – ΠΛΗΡΟΦΟΡΙΚΗ ΤΗΣ ΥΓΕΙΑΣ»

ΕΙΔΙΚΕΥΣΗ: ΠΛΗΡΟΦΟΡΙΚΗ ΤΗΣ ΥΓΕΙΑΣ

# ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ (ΑΝΑΚΤΗΣΗ ΚΑΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ) ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ ΚΟΙΝΩΝΙΚΟ ΔΙΚΤΥΟ TWITTER ΑΝΑΦΟΡΙΚΑ ΜΕ ΤΗΝ ΨΥΧΟΛΟΓΙΑ ΤΟΥ ΠΛΗΘΥΣΜΟΥ ΣΤΗ ΠΑΝΔΗΜΙΑ COVID-19

**ΑΦΡΟΔΙΤΗΣ Ν. ΚΑΤΙΚΑ**

ΜΗΧΑΝΙΚΟΥ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΘΗΝΑ 2023

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

ΚΑΘΗΓΗΤΡΙΑ ΦΛΩΡΑ ΜΑΛΑΜΑΤΕΝΙΟΥ (ΕΠΙΒΛΕΠΟΥΣΑ)
ΔΡ. ΒΑΣΙΛΙΚΗ ΚΟΥΦΗ, ΕΔΙΠ
ΔΡ. ΕΜΜΑΝΟΥΗΛ ΖΟΥΛΙΑΣ, ΕΔΙΠ

**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΤΜΗΜΑ ΝΟΣΗΛΕΥΤΙΚΗΣ**

ΔΙΙΔΡΥΜΑΤΙΚΟ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ:

ΕΙΔΙΚΕΥΣΗ: ΠΛΗΡΟΦΟΡΙΚΗ ΤΗΣ ΥΓΕΙΑΣ

**ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ**

# ΕΦΑΡΜΟΓΗ ΓΙΑ ΤΗΝ ΕΞΟΡΥΞΗ (ΑΝΑΚΤΗΣΗ ΚΑΙ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ) ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ ΚΟΙΝΩΝΙΚΟ ΔΙΚΤΥΟ TWITTER ΑΝΑΦΟΡΙΚΑ ΜΕ ΤΗΝ ΨΥΧΟΛΟΓΙΑ ΤΟΥ ΠΛΗΘΥΣΜΟΥ ΣΤΗ ΠΑΝΔΗΜΙΑ COVID-19

ΤΗΣ ΑΦΡΟΔΙΤΗΣ Ν. ΚΑΤΙΚΑ

## ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία, που εκπονείται στο πλαίσιο του Διιδρυματικού Προγράμματος Μεταπτυχιακών Σπουδών «Οργάνωση και Διοίκηση Υπηρεσιών Υγείας – Πληροφορική της Υγείας», πραγματεύεται την εξόρυξη γνώσης απο το Twitter σχετικά με το Long Covid. Αρχικά, γίνεται αναφορά στα μεγάλα δεδομένα στην Υγεία και την επεξεργασία φυσικής γλώσσας. Στη συνέχεια αναφέρονται περιπτώσεις όπου η επεξεργασία φυσικής γλώσσας εξήγαγε χρήσιμα συμπεράσματα για τη δημόσια υγεία την περίοδο του Covid-19. Ακολουθεί η μελέτη περίπτωσης όπου αναλύθηκαν tweets στα ελληνικά που αναφέρονταν στο Long Covid. Μετά την προεπεξεργασία των δεδομένων βρέθηκαν τα κύρια θέματα στα οποία αναφέρονται οι χρήστες και έγινε ανάλυση συναισθήματος. Η προεπεξεργασία και η ανάλυση των αποτελεσμάτων έγινε σε Python με το εργαλείο Jupyter Notebook . Τα αποτελέσματα έδειξαν τα ακόλουθα κύρια θέματα: οι πολίτες συζητούν τις επιπτώσεις του Long Covid, τις επιπτώσεις του Long Covid σε συγκεκριμένες ομάδες πληθυσμού όπως τα παιδιά και τέλος για τρόπους αντιμετώπισης όπως τα εμβόλια. Το 59% των tweets είχε αρνητικό συναίσθημα ενώ τα υπόλοιπα είχαν θετικό ή ουδέτερο συναίσθημα.

**Λέξεις-κλειδιά:** Μεγάλα δεδομένα, Twitter, Επεξεργασία Φυσικής Γλώσσας, Ανάλυση Συναισθήματος, Covid-19, πανδημία

**NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS**

**FACULTY OF NURSING**

INTERUNIVERSITY POSTGRADUATE PROGRAM IN HEALTH CARE
MANAGEMENT AND HEALTH CARE INFORMATICS
SPECIALIZATION: HEALTH INFORMATICS

**DISSERTATION**

**MINING TWITTER DATA TO UNDERSTAND USERS' ATTITUDE TO COVID-19**

BY AFRODITI N. KATIKA

## SUMMARY

This thesis, prepared in the framework of the InterUniversity Postgraduate Program "Health Care Management – Health Informatics" discusses Greek speaking users' attitude with regards to Long Covid. Initially there is a mention of big data in healthcare and natural language processing. Afterwards, there is a list of use cases where natural language processing provided important insights for public health over the pandemic of Covid-19. Then the main use case of this thesis is presented. Twitter data in the Greek language referring to Long Covid were mined, processed and analysed to model main topics of discussion and analyse sentiment. Pre-processing, analysis and results were performed in Python using Jupyter Notebook tool. Results highlighted the following discussion topics: Greek-speaking users discuss Long Covid effects, Long Covid effects in specific population groups like children and vaccines. 59% of analysed tweets conveyed a negative sentiment while the rest had positive or neutral sentiment.

**Keywords**: Big Data, Twitter, Natural Language Processing, NLP, Sentiment Analysis, Covid-19, Pandemic

**ΕΥΧΑΡΙΣΤΙΕΣ**

Ευχαριστώ πολύ τους καθηγητές, το διοικητικό προσωπικό και τους συμφοιτητές μου στο ΔΠΜΣ για τη συνεργασία και τις γνώσεις που αποκόμισα κατά τη διάρκεια του μεταπτυχιακού προγράμματος. Ευχαριστώ ιδιαιτέρως την επιβλέπουσα καθηγήτριά μου κα Φλώρα Μαλαματένιου για την πολύτιμη βοήθεια και καθοδήγηση κατά τη διάρκεια εκπόνησης της διπλωματικής εργασίας, καθώς και τα μέλη της εξεταστικής επιτροπής για την αξιολόγησή της.

Ευχαριστώ με όλη μου την καρδιά το συζυγό μου Αντώνη και τον γιο μου Δημήτρη. Χωρίς την στήριξη τους, αμφιβάλλω οτι θα τα κατάφερνα. Τέλος ευχαριστώ τους γονείς μου και την αδελφή μου που όλα αυτά τα χρόνια με τη στάση τους μου δείχνουν την αγάπη για την προσφορά, τη δια βίου μάθηση και τη διαρκή εξέλιξη.

Πίνακας Περιεχομένων

Λιστα Εικόνων

# Εισαγωγή

Η εξέλιξη της τεχνολογίας και η ψηφιοποίηση υπηρεσιών έχει διεισδύσει και στο χώρο της υγείας με αποτέλεσμα την παραγωγή μεγάλων δεδομένων που μπορούν να αναλυθούν για να παραχθεί πολύτιμη γνώση σε κυβερνητικούς φορείς και φορείς δημόσιας υγείας. Μια απο αυτές τις πηγές μεγάλων δεδομένων είναι τα κοινωνικά δίκτυα όπου ο πληθυσμός εκφράζει τη γνώμη του και το συναίσθημα του για διάφορα θέματα της επικαιρότητας.

Μέσα από τεχνικές όπως η επεξεργασία φυσικής γλώσσας και η ανάλυση συναισθήματος, υπάρχει μια πρωτοφανής ευκαιρία για τη δημόσια υγεία να επεξεργαστεί τα μηνύματα αυτά σε πραγματικό χρόνο ώστε είτε να βελτιώσει τους υπολογισμούς για τον επιπολασμό μιας ασθένειας είτε να εντοπίσει δημοφιλή συμπτώματα είτε να κατανοήσει καλύτερα την ψυχολογία του πληθυσμού.

Η πανδημία του Covid-19 η οποία ξέσπασε το 2020 δημιούργησε μια πρωτόγνωρη κατάσταση παγκοσμίως όπου οι κυβερνήσεις κλήθηκαν να πάρουν μέτρα περιορισμού εξάπλωσης της πανδημίας όπως η παραμονή στο σπίτι και η μάσκα. Λίγους μήνες μετά παρατηρήθηκε το φαινόμενο του Long Covid δηλαδή ασθενείς που ανέρρωναν απο Covid-19 αλλα η αποκατάσταση τους αργούσε ή ακόμη χειρότερα εμφάνιζαν κάτι καινούριο.

Σκοπός της παρούσας διπλωματικής εργασίας είναι η διερεύνηση των κύριων θεμάτων που συζητούν οι Έλληνες στο Twitter σχετικά με το Long Covid καθώς και η ανάλυση των συναισθημάτων τους ως προς αυτή την πρωτόγνωρη και σχετικά αχαρτογράφητη ασθένεια που είναι εν εξελίξει. Προς αυτή την κατεύθυνση, εξήχθησαν ελληνικά tweets που αναφέρονταν στο Long Covid κατά τη διάρκεια του 2022. Στη συνέχεια επεξεργάστηκαν και καθαρίστηκαν ώστε να είναι έτοιμα για να αναδειχθούν τα κύρια θέματα και το συναίσθημα που κουβαλούν.

Παρακάτω παρουσιάζονται συνοπτικά τα περιεχόμενα των κεφαλαίων της παρούσας εργασίας.

Στο 1ο κεφάλαιο γίνεται αναφορά στα μεγάλα δεδομένα και τα είδη μεγάλων δεδομένων που παράγονται στο χώρο της υγείας. Επίσης δίνεται ο ορισμός των κοινωνικών δικτύων που είναι μια ακόμη πηγή παραγωγής μεγάλων δεδομένων και εξηγείται το γιατί τα κοινωνικά δίκτυα εν δυνάμει αφορούν τη δημόσια υγεία.

Το 2ο κεφάλαιο ξεκινά με την παρουσίαση της διαδικασίας εξόρυξης γνώσης απο τα μεγάλα δεδομένα. Στη συνέχεια γίνεται αναφορά στην επεξεργασία φυσικής γλώσσας και την υποκατηγορίας της ανάλυση συναισθήματος και αναλύονται διάφορες τεχνικές εξόρυξης γνώσης στις τρεις κύριες κατηγορίες: λεξικά συναισθημάτων, μηχανική μάθηση και υβριδικές μέθοδοι.

Στο 3ο κεφάλαιο παρουσιάζονται περιπτώσεις εξόρυξης γνώσης και επεξεργασίας φυσικής γλώσσας στο χώρο της υγείας.

Στο 4ο κεφάλαιο παρουσιάζονται περιπτώσεις εξόρυξης γνώσης και επεξεργασίας φυσικής γλώσσας συγκεκριμένα την περίοδο του Covid-19 καθώς γίνεται και μια αναφορά ανάλογων εργασιών στην Ελλάδα. Οι κύριες περιπτώσεις αφορούν στην ανάλυση συναισθήματος ως προς τα εμβόλια κατα του Covid-19, το Long Covid και την γενικότερη ψυχολογία του πλήθους.

Στο 5ο κεφάλαιο παρουσιάζεται η μελέτη περίπτωσης με απώτερο στόχο την ανάλυση συναισθήματος των Ελλήνων ως προς το Long Covid που μέχρι και αυτή τη χρονική στιγμή δεν καλυφθεί βιβλιογραφικά εκτεταμένα. Αρχικά, περιγράφεται η διαδικασία εξαγωγής δεδομένων απο το Twitter καθώς και η προεπεξεργασία τους. Στη συνέχεια, εφαρμόζονται δύο μοντέλα που στοχεύουν στην ανάδειξη κύριων θεμάτων συζήτησης (Latent Dirichlet Allocation και Gibbs Sampling Dirichlet Multinomial Mixture). Τέλος εφαρμόζεται το Greek-BERT προκειμένου να γίνει η ανάλυση των συναισθημάτων και παρουσιάζονται τα αποτελέσματα.

Τέλος, καταγράφονται τα συμπεράσματα και διατυπώνονται προτάσεις για μελλοντικές επεκτάσεις αυτής της εργασίας.

# Κεφάλαιο 1 Μεγάλα Δεδομένα

## 1.1 Δεδομένα

Με τον όρο Δεδομένο (data) ορίζεται οποιαδήποτε μοναδική παρατήρηση ή γεγονός. Ετυμολογικά η λέξη "δεδομένα" προέρχεται από την λατινική λέξη datum κι αφορά στη συλλογή στοιχείων, αριθμών, παρατηρήσεων, κ.α. που αν επεξεργαστούν, οργανωθούν, ερμηνευτούν κι αναλυθούν κατάλληλα απο έναν υπολογιστή παράγουν πολύτιμες "πληροφορίες" [1]. Όσο περισσότερη πληροφόρηση έχουμε, τόσο καλύτερη γίνεται η λήψη αποφάσεων. Εξ ’ορισμού λοιπόν είναι σαφές ότι η αποθήκευση δεδομένων έχει ως στόχο την παροχή πληροφορίας και την διευκόλυνση αποφάσεων στον επαγγελματία.

Τα δεδομένα χωρίζονται σε δομημένα (structured data) και μη δομημένα (un-structured data). Τα μη δομημένα δεδομένα είναι πιο δύσκολο να επεξεργαστούν καθώς δεν υπάρχει ορισμένο σχήμα (schema) και άρα δεν υπάρχει κατανόηση του τι είναι το περιεχόμενο. Πλέον η πλειονότητα των αποθηκευμένων δεδομένων αφορά μη δομημένα δεδομένα. Παραδείγματα δομημένων δεδομένων αποτελούν τα πεδία μιας φόρμας όπου ο χρήστης καλείται να επιλέξει μεταξύ συγκεκριμένων τιμών. Παράδειγμα μη δομημένων δεδομένων είναι η διάγνωση που θα γράψει ενας ιατρός σε ελεύθερο κείμενο.

## 1.2 Μεγάλα Δεδομένα (Big Data)

Απο τη στιγμή που αναπτύχθηκαν τεχνολογίες που καθιστούσαν εφικτή την αποθήκευση και διαχείριση τεράστιου όγκου δεδομένων, εισήχθη η έννοια των Μεγάλων Δεδομένων. Ως μεγάλα δεδομένα ορίζονται εξαιρετικά μεγάλα σύνολα δεδομένων που τα συμβατικά πληροφοριακά συστήματα δε μπορούν να τα διαχειριστούν και ως εκ τούτου απαιτούν νέες τεχνικές διαχείρισης ώστε να εξαχθούν χρήσιμες πληροφορίες και συμπεράσματα. [2]

Η σημασία των δεδομένων μέσα σ’ ένα σύγχρονο επιχειρηματικό ή κοινωνικό πλαίσιο είναι τόσο μεγάλη που συγκαταλέγονται πλέον στον κατάλογο των άυλων περιουσιακών στοιχείων των οργανισμών κι αποτελούν σημαντικό κεφάλαιο τους. [3]

Τα μεγάλα δεδομένα αυξάνονται σε τρεις διαφορετικές συνιστώσες: όγκος, ταχύτητα και ποικιλία (γνωστά και ως 3Vs: volume, velocity and variety). Με όγκο εννοούμε οτι η ποσότητα των δεδομένων που αποθηκεύονται έχει αυξηθεί. Ο λόγος που έχουν αυξηθεί είναι η αύξηση των υπηρεσιών που έχουν ψηφιοποιηθεί, η ελεύθερη πρόσβαση σε καινούριες εφαρμογές αλλά και η αύξηση της ακρίβειας των αισθητήρων που παράγουν περισσότερες

τιμές. Η ταχύτητα αφορά τον ρυθμό αποθήκευσης δεδομένων δηλαδή πόσο γρήγορα εισέρχονται δεδομένα στη βάση, πόσο γρήγορα επεξεργάζονται αλλά και πόσο γρήγορα αντικαθίστανται. Τέλος η ποικιλία αφορά τους διαφορετικούς τύπους δεδομένων που πλέον μπορεί να είναι δεδομένα βάσεων, βίντεο, ήχος, ελεύθερο κείμενο κλπ. [4]

Αυτές οι τρεις συνιστώσες προσθέτουν πολυπλοκότητα στην διαχείριση και επεξεργασία των δεδομένων. Δυστυχώς, οι τεχνολογίες αποθήκευσης δεδομένων έχουν διεισδύσει στους τομείς υπηρεσιών πιο γρήγορα απο τις τεχνολογίες επεξεργασίας τους με αποτέλεσμα σε σημαντικά μεγάλο βαθμό ο τεράστιος όγκος των δεδομένων που συλλέγονται, στις περισσότερες περιπτώσεις, παραμένουν ανεκμετάλλευτα καταλήγοντας σε ενα τάφο δεδομένων (“data tomb”) [5].

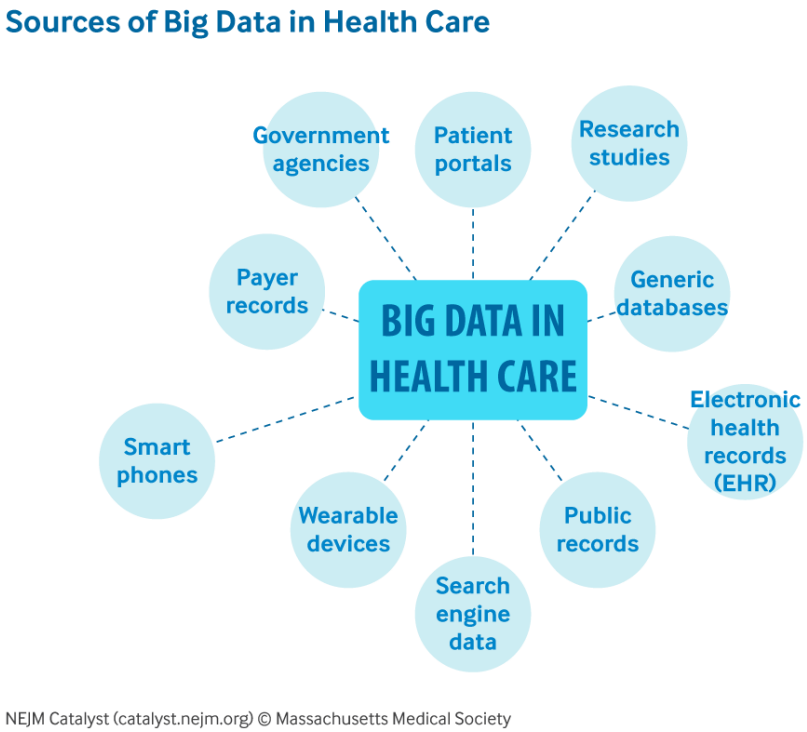
Επιπροσθέτως τα τελευταία χρόνια έχουν προστεθεί και επιπλέον συνιστώσες που χαρακτηρίζουν τα μεγάλα δεδομένα: εγκυρότητα, αξία και μεταβλητότητα (veracity, value and variability). Η εγκυρότητα αφορά το πόσο αξιόπιστα και καθαρά είναι τα δεδομένα μας. Πολλές φορές τα δεδομένα έχουν πολύ θόρυβο (πχ ακραίες τιμές που επηρεάζουν τα αξιόπιστα συμπεράσματα) ή πολλές ανωμαλίες απαιτώντας μεγάλη επεξεργασία για να καθαριστούν. Η αξία είναι ίσως η πιο σημαντική παράμετρος που αφορά το ποσοτικό πλεονέκτημα που προσφέρει η εξόρυξη γνώσης απο τα δεδομένα σε κάποιο οργανισμό. Αν τα δεδομένα ενός οργανισμού αποθηκεύονται αλλά δεν χρησιμοποιούνται για να βελτιωθούν οι αποφάσεις και τα προϊόντα ενός οργανισμού τότε τα δεδομένα αυτά είναι σχεδόν άχρηστα. Η μεταβλητότητα αφορά την αλλαγή των δεδομένων, τις προσθήκες νέων τιμών ή τη μεταβολή των υπαρχουσών κάτι που μπορεί να επηρεάσει τη μοντελοποίηση των δεδομένων και τις υποθέσεις μας. [6]

### 1.3 Μεγάλα Δεδομένα στην Υγεία

Όπως οι άλλοι τομείς υπηρεσιών ετσι και ο τομέας της Υγείας παράγει εναν απίστευτα μεγάλο αριθμό δεδομένων. Μέχρι τις αρχές του 21$^{ου}$ αιώνα η αποθήκευση αυτών των δεδομένων απο το ιατρικό προσωπικό γινόταν κυρίως χειρόγραφα οπότε η εξαγωγή γνώσης ήταν δύσκολη μιας που ο αριθμός των δεδομένων ήταν απομονωμένος. Η πρακτική του χειρόγραφου ιατρικού αρχείου είναι αρκετά παλιά και υπάρχουν εγγραφές σε παπύρους στην Αίγυπτο το 1600π.χ. [2]

Πλέον τα δεδομένα αποθηκεύονται ψηφιακά. Κάποιες σημαντικές κατηγορίες δεδομένων είναι [2]:

- Ηλεκτρονικοί φάκελοι υγείας (electronic health records – EHR). Τα συστήματα αυτά περιλαμβάνουν παρελθοντικές και τωρινές πληροφορίες σωματικής και ψυχικής υγείας και ασθενειών για κάθε άτομο και είναι διαθέσιμα για την παροχή ιατρικών υπηρεσιών και την πιο ενημερωμένη λήψη αποφάσεων . Αυτή η ευρεία κατηγορία περιλαμβάνει:
    - τους ηλεκτρονικούς φακέλους ασθενών (πχ. διαγνώσεις, ιατρικές εικόνες, αποτελέσματα ιατρικών πράξεων ενός νοσοκομείου ή ενός διαγνωστικού κέντρου)
    - το ιστορικό ενος ατόμου (πχ. αλλεργίες, δημογραφικά χαρακτηριστικά)
    - ιστορικό φαρμάκων και συνταγογραφήσεων
    - δεδομένα που παράγουν και αποθηκεύουν ασφαλιστικοί φορείς (δημόσιοι και ιδιωτικοί)
    - δεδομένα που προκύπτουν απο έξυπνες συσκευές παρακολούθησης διασυνδεδεμένες με ένα ιατρικό κέντρο ή ιατρικό προσωπικό
- Δεδομένα που προκύπτουν σε έναν οργανισμό υγείας όπως ο μέσος χρόνος αναμονής στα επείγοντα, οι βάρδιες του προσωπικού, τα κόστη και τα έσοδα
- Κοινωνικά, δημογραφικά και περιβαλλοντικά δεδομένα που χαρακτηρίζουν ένα σύνολο πληθυσμού και την περιοχή που μένουν
- Δεδομένα που αφορούν το γονιδίωμα ενός οργανισμού και χρησιμοποιούνται στους τομείς της βιοιατρικής έρευνας και μοριακής ιατρικής με στόχο τις εξατομικευμένες θεραπείες
- Δεδομένα που προκύπτουν απο έξυπνες συσκευές (internet of things) και σένσορες πχ. μέτρηση σφυγμών, βημάτων κλπ. Οι τεχνολογίες λήψης, αποθήκευσης και μετάδοσης δεδομένων μέσω διαδικτύου έχουν κάνει δυνατή την παρακολούθηση ασθενών με χρόνιες παθήσεις εξ’ αποστάσεως (πχ. Απ το σπίτι τους ή κάποιο οίκο ευγηρίας).




*Εικόνα 1: Πηγές μεγάλων δεδομένων στην υγεία* [7]

1.3.1 Τομείς εφαρμογών στην υγεία που παράγουν ή χρησιμοποιούν μεγάλα δεδομένα

Τα τελευταία χρόνια έχουν αναπτυχθεί πολλές εφαρμογές που είτε παράγουν είτε χρησιμοποιούν μεγάλα δεδομένα σε μια απο τις παραπάνω μορφές.

Στην Ευρωπαϊκή Ένωση, οι εφαρμογές χρήσης των μεγάλων δεδομένων είναι ενα σημαντικό θέμα που καλείται να συμβάλει στην επίλυση του παρακάτω προβλήματος: ο μέσος όρος ζωής των πολιτών της Ευρωπαϊκής Ένωσης επιμηκύνεται ως εκ τούτου όλο και περισσότεροι άνθρωποι ζουν με χρόνια νοσήματα ή και πολλαπλά νοσήματα επιβαρύνοντας το σύστημα υγείας [8] .

Ο πρώτος τομέας αφορά εφαρμογές που παρακολουθούν τη μέση υγεία των πολιτών και προσπαθούν να τη βελτιώσουν. Παραδείγματα τέτοιων εφαρμογών παρακολουθούν και αποθηκεύουν τα ημερήσια βήματα των χρηστών ενθαρρύνοντας τους να αθληθούν περισσότερο, παρομοίως παρακολουθούν τη διατροφή, τις συνθήκες ύπνου ή τους παλμούς. Τα δεδομένα αυτά είτε οπτικοποιούνται για να καταλάβει ο ίδιος ο χρήστης τις συνήθειες του και να τις βελτιώσει είτε μπορεί να τα διαμοιραστεί με τον προσωπικό του ιατρό. [7]

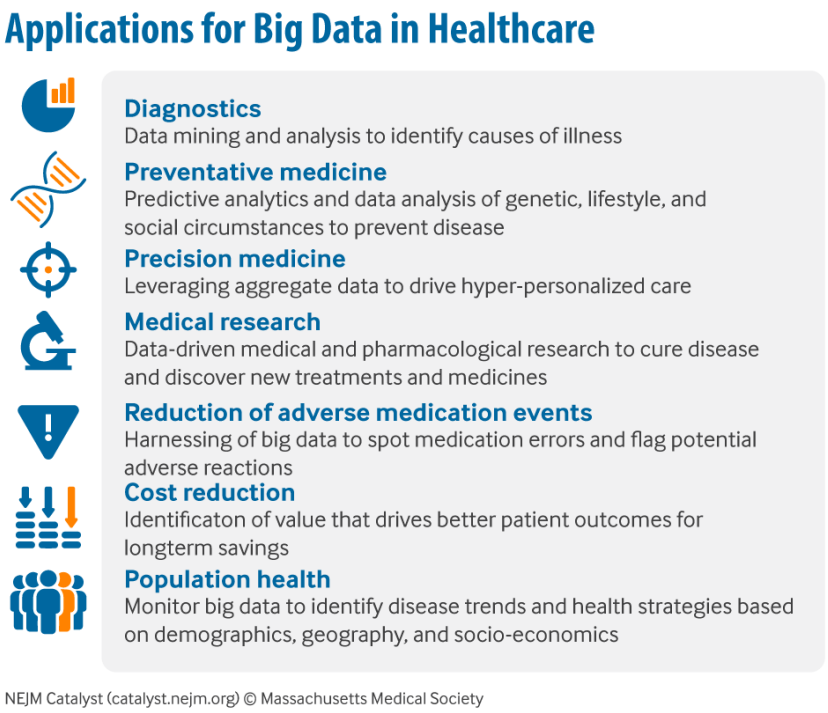
Ο δεύτερος τομέας αφορά εφαρμογές που ψηφιοποιούν διαδικασίες και αρχεία που μέχρι πρότινος γίνονταν χειρόγραφες και δια ζώσης με στόχο την ευκολότερη και πιο συστηματική

πρόσβαση σε αυτά τα δεδομένα. Σε αυτή την κατηγορία ανήκει ο ηλεκτρονικός φάκελος ασθενούς με καταχώρηση ιατρικών πράξεων, συνταγογραφήσεων και αποτελεσμάτων εξετάσεων ή διαγνώσεων που μπορεί να προσπελαστεί απο τον ίδιο τον ασθενή αλλα και απο διαφορετικό ιατρικό προσωπικό [7]. Εδώ ανήκουν και εφαρμογές που επιτρέπουν σε ασθενείς σε απομονωμένες περιοχές ή με κινητικά προβλήματα να παρακολουθούν τα συμπτώματα τους και να επικοινωνούν με ιατρικό προσωπικό εξ' αποστάσεως.

Ο τρίτος τομέας αφορά εφαρμογές που χρησιμοποιούν μεγάλα δεδομένα για να παρθούν αποφάσεις που αυξάνουν την ποιότητα των ιατρικών υπηρεσιών. Εδώ ανήκουν εφαρμογές που χρησιμοποιούν μεγάλα δεδομένα είτε για να αυξηθεί η ακρίβεια της διάγνωσης του ασθενούς είτε για να γίνει πιο γρήγορη. Επίσης εδώ ανήκουν εφαρμογές που στοχεύουν στη βελτίωση των αποφάσεων μέσα σε εναν ιατρικό οργανισμό ώστε ο οργανισμός να γίνει πιο αποτελεσματικός οσον αφορά στο κόστος και το περιβαλλοντικό του αποτύπωμα πχ να μειωθούν τα κόστη χωρίς να μειωθεί η ποιότητα των υπηρεσιών ή να εντοπίζονται οι μερίδες του πληθυσμού που είναι πιο ευάλωτες προσφέροντας τους προληπτικές εξετάσεις και μειώνοντας μακροπρόθεσμα την ανάγκη για νοσηλεία. Τέλος εδώ ανήκουν και εφαρμογές στο χώρο της έρευνας οι οποίες χρησιμοποιώντας ιστορικά δεδομένα επικεντρώνονται σε νέα φάρμακα και τομείς που έχουν περισσότερες πιθανότητες επιτυχίας [7].

Μια γνωστή εφαρμογή του τρίτου τομέα ήταν η συνεργασία της Intel και ενός γαλλικού πανεπιστημιακού νοσοκομείου ώστε με βάση τα ιστορικά δεδομένα να προβλέπονται οι αριθμοί των ασθενών που θα επισκέπτονται τα επείγοντα τις επόμενες 15 μέρες. Με αυτή τη γνώση, οι υπεύθυνοι του νοσοκομείου μπορούν να προγραμματίσουν πιο αποτελεσματικά τις βάρδιες του ιατρικού προσωπικού ώστε να μειώσουν τους χρόνους αναμονής των ασθενών. [9]

Ο τέταρτος τομέας στοχεύει στην ανάπτυξη εξατομικευμένων θεραπειών για τον ασθενή βασιζόμενες σε πληροφορίες που δεν ήταν παλιότερα διαθέσιμες όπως το γονιδίωμα του ή ο τρόπος που μεταβολίζει τροφές. [7] Εναλλακτικά εδώ ανήκουν και εφαρμογές που προβλέπουν εξατομικευμένο ρίσκο ενός πολίτη να νοσήσει απο κάποια ασθένεια εντοπίζοντας παράγοντες κινδύνου πάνω του. [8]




*Εικόνα 2: Εφαρμογές μεγάλων δεδομένων στην υγεία* [7]

### 1.3.2 Προκλήσεις στα μεγάλα δεδομένα στην υγεία

Μια κύρια πρόκληση είναι οτι ο όγκος των δεδομένων που παράγεται στο χώρο της υγείας είναι αρκετά μεγάλος. Μέχρι πρόσφατα, οι δομές υγείας αποθήκευαν τα δεδομένα μέσα στην ίδια τη δομή (πχ. Δωμάτια με servers). Λόγω του όγκου όμως η ταχύτητα επεξεργασίας καθώς και η αξιοπιστίας τους καθιστά την αποθήκευση τους εντός του νοσοκομείου δύσκολη. Η εναλλακτική είναι η αποθήκευση των δεδομένων σε παρόχους cloud το οποίο όμως ενέχει κινδύνους ασφάλειας και προστασίας δεδομένων. Ήδη έχουν υπάρξει πολλές περιπτώσεις διαρροής προσωπικών δεδομένων υγείας.

Η δεύτερη πρόκληση ξεκινά απ το γεγονός οτι τα δεδομένα στο χώρο της υγείας είναι αρκετά ετερογενή. Οταν μιλάμε για ετερογενή δεδομένα καταρχήν εννοούμε τον τύπο τους. Πχ. Τα δεδομένα μπορεί να είναι ιατρικές εικόνες, βίντεο υπερήχων, αποτελέσματα αιματολογικών εξετάσεων ή ελεύθερο κείμενο στην περίπτωση μιας διάγνωσης. Η ετερογένεια αναφέρεται και στο γεγονός οτι προέρχονται απο διαφορετικές πηγές (δημόσια νοσοκομεία, ιδιωτικά νοσοκομεία, προϊόντα ιδιωτικών εταιριών, πανεπιστημιακά ιδρύματα) οι οποίες πρέπει να συμφωνήσουν εκ των προτέρων σε μια μορφή κοινή. Τέλος η ετερογένεια οφείλεται και στο γεγονός οτι κάποια απο αυτά τα δεδομένα συμπληρώνονται με ανθρώπινο χέρι με αποτέλεσμα να καταγράφονται με λάθη ή μη συστηματικό τρόπο. Όλα αυτά σημαίνουν οτι ο

διαμοιρασμός των δεδομένων δεν είναι εύκολος μεταξύ οργανισμών. Επιπλέον λόγω της ετερογένειας προκειμένου να χρησιμοποιηθούν τα δεδομένα αυτά πρέπει κάποιος να τα καθαρίσει, να τα επεξεργαστεί και να τα κατηγοριοποιήσει [2].

### 1.4 Κοινωνικα Δίκτυα

Ως κοινωνικά δίκτυα ορίζονται εφαρμογές και πλατφόρμες που επιτρέπουν στους ανθρώπους να επικοινωνούν και να μοιράζονται πληροφορίες στο διαδίκτυο με τη χρήση υπολογιστή ή έξυπνου κινητού [10]. Η επικοινωνία μπορεί να είναι δημόσια ή ιδιωτική και οι τύποι των δεδομένων που διαμοιράζονται ποικίλλουν: ελεύθερο κείμενο, φωτογραφίες, βίντεο ή ηχητικά μηνύματα.

Αρχικά τα κοινωνικά δίκτυα χρησιμοποιούνταν για επικοινωνία με φίλους και επαγγελματικούς συνεργάτες αλλα σταδιακά έγιναν και μια κύρια πηγή ενημέρωσης. Επίσης αποτελούν ένα μέσο έκφρασης απόψεων και συναισθημάτων. Χάρη στην ταχύτητα διαμοιρασμού (και την έλλειψη ελέγχου που εφαρμόζουν τα παραδοσιακά μέσα ενημέρωσης) μια είδηση ή άποψη μπορεί να ταξιδέψει πολύ γρήγορα γι αυτό και τα μέσα κοινωνικής δικτύωσης προτιμούνται ειδικά για κάποιον που θέλει να μάθει τι συμβαίνει αυτή τη στιγμή.

Το Twitter είναι ένα απο τα πιο δημοφιλή μέσα κοινωνικής δικτύωσης όπου οι χρήστες εγγράφονται δωρεάν και ανήκει στην κατηγορία microblogging. Εμφανίστηκε τον Ιούλιο του 2006. Στην αρχή οι χρήστες μοιράζονταν μηνύματα των 140 χαρακτήρων αλλα από τον Νοέμβριο του 2016 το όριο αυξήθηκε στους 280 χαρακτήρες. Το Twitter αποτελεί το πιο άμεσο μέσο έκφρασης απόψεων ειδικά για θέματα επικαιρότητας και αιχμής επηρεάζοντας γρήγορα και την ‘κοινή γνώμη’ καθώς τις περισσότερες φορές τα μηνύματα είναι προσβάσιμα και σε άτομα που δεν ανήκουν στο δίκτυο του χρήστη ή δεν ειναι καν εγγεγραμμένα στην πλατφόρμα.

Τα δεδομένα που παράγει το Twitter ανήκουν στην κατηγορία των μεγάλων δεδομένων και αποτελούνται απο δομημένα και μη δομημένα δεδομένα. Κάθε tweet περιλαμβάνει συγκεκριμένα δομημένα πεδία όπως πχ το όνομα του χρήστη, την περιοχή, τη γλώσσα, την ημερομηνία δημοσίευσης αλλα το κύριο μήνυμα του tweet είναι ελεύθερο κείμενο του χρήστη αρα και μη δομημένο.

Ο τεράστιος όγκος των δεδομένων που είναι διαθέσιμος παγκοσμίως μέσω του Twitter, θα μπορούσε να αξιοποιηθεί και να αποτελέσει την πρώτη ύλη για ανάλυση των απόψεων της κοινωνίας σε πραγματικό χρόνο.

### 1.4.1 Κοινωνικά δίκτυα και υγεία

Τα τελευταία χρόνια έχει αυξηθεί πολύ η διαθεσιμότητα δεδομένων στα κοινωνικά δίκτυα που αφορούν την υγεία. Έχει βρεθεί οτι οι άνθρωποι προτιμούν να μοιράζονται εμπειρίες που αφορούν την υγεία τους με φίλους τους παρά με ιατρικό προσωπικό [11]. Η ανωτέρω μελέτη βρήκε επίσης οτι ειδικά όταν αφορά ευαίσθητες ασθένειες όπως καρκίνος του μαστού, aids, καρκίνος του προστάτη οι ασθενείς είναι πιο πιθανό να αναζητήσουν άλλους ομοιοπαθείς online απο όταν πάσχουν απο κάτι εξίσου επικίνδυνο αλλά λιγότερο ταμπού πχ. καρδιακές παθήσεις.

Επιπλέον κάποιες φορές για μικρά προβλήματα υγείας ο άνθρωπος μπορεί να αναζητήσει συμβουλές πρώτα στον φιλικό και κοινωνικό του κύκλο χωρίς να απευθυνθεί σε επαγγελματία υγείας (πχ. Facebook groups συγκεκριμένου σκοπού).

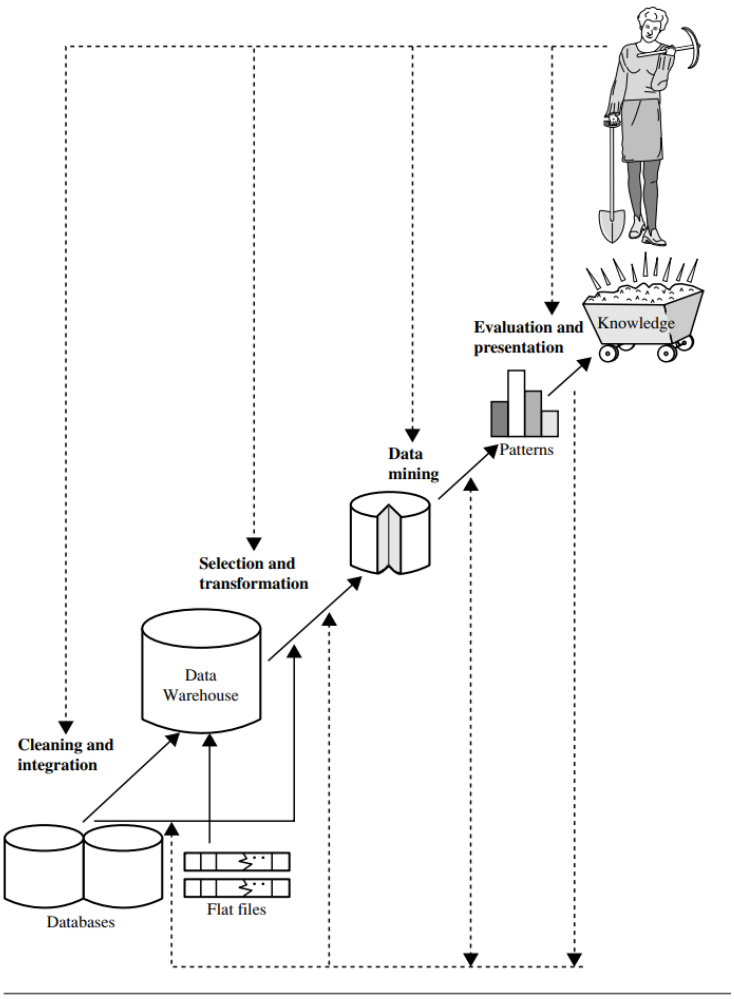
# Κεφάλαιο 2 Εξόρυξη Δεδομένων

## 2.1 Ορισμός

Η εξόρυξη δεδομένων (data mining) είναι η διαδικασία μέσω της οποίας επεξεργαζόμαστε μεγάλο αριθμό δεδομένων με σκοπό την εύρεση μοτίβων και την ανακάλυψη γνώσης [5]. Ουσιαστικά ο σκοπός της εξόρυξης δεδομένων είναι η ανακάλυψη της γνώσης οπότε ένας εναλλακτικός ορισμός για αυτή τη διαδικασία είναι: ανακάλυψη γνώσης απο δεδομένα (knowledge discovery from data – KDD). Για λόγους απλότητας όμως έχει επικρατήσει ευρέως ο όρος εξόρυξη ο οποίος κυριολεκτικά αναφέρεται σε ένα απο τα βήματα της διαδικασίας [5].

Η εξόρυξη δεδομένων ακολουθεί επαναληπτικά τα παρακάτω βήματα όπως φαίνεται και στην Εικόνα 3:

- Συλλογή δεδομένων (selection). Στο στάδιο αυτό μελετάται ο τομέας που αφορά το ερευνητικό ερώτημα και οι διαθέσιμες πηγές δεδομένων. Αφου βρεθούν κατάλληλες πηγές, τα δεδομένα συλλέγονται.
- Επεξεργασία δεδομένων (processing). Στο στάδιο αυτό τα δεδομένα καθαρίζονται ωστε να απαλειφθούν οι προβληματικές ή ελλειπείς εγγραφές. Παραδείγματα tasks επεξεργασίας: απαλοιφή emojis, κοινών λέξεων και τόνων.
- Μετασχηματισμός δεδομένων (transformation). Στο στάδιο αυτό εντοπίζονται τα χρήσιμα πεδία και μετασχηματίζονται όπου χρειάζεται για να ειναι πιο εύκολη, ουσιαστική ή ομογενοποιημένη η ανακάλυψη της γνώσης (για παράδειγμα κάποια πεδία μπορεί να γίνουν ενα, κάποια πεδία να αλλάξουν τύπο τιμών κλπ)
- Εξόρυξη δεδομένων (data mining). Στο στάδιο αυτό επιλέγονται οι κατάλληλοι μέθοδοι και αλγόριθμοι εξόρυξης και εφαρμόζονται για την εύρεση μοτίβων
- Διερμηνεία/Αξιολόγηση/Οπτικοποίηση (interpretation/evaluation/visualisation). Στο στάδιο αυτό γίνεται η παρουσίαση των αποτελεσμάτων προκειμένου να αξιολογηθούν




*Εικόνα 3: Στάδια Εξόρυξης Γνώσης* [5]

Η εξόρυξη δεδομένων στο χώρο της υγείας εφαρμόζεται σε περιπτώσεις που ομαδοποιούνται σε τέσσερις κατηγορίες: περιγραφική ανάλυση δεδομένων (descriptive), διαγνωστική αναλυτική (diagnostic), προγνωστική αναλυτική (predictive), καθοδηγητική αναλυτική (prescriptive). [2]

Η κατηγορία περιγραφικής αναλυτικής επικεντρώνεται στην περιγραφή της ιατρικής κατάστασης που αντιμετώπισε ο ασθενής και πλήρους σχολιασμού αυτής συνδυάζοντας δεδομένα απο διαφορετικές πηγές. [2]

Στην κατηγορία της διαγνωστικής αναλυτικής γίνεται διάγνωση της τωρινής κατάστασης με βάσει υπάρχοντες παράγοντες και αποτελέσματα ιατρικών πράξεων. Εξηγούνται λόγοι και οι παράγοντες που οδήγησαν σε συγκεκριμένα γεγονότα. [2]

Στην κατηγορίας της προγνωστικής αναλυτικής, γίνεται πρόγνωση της μελλοντικής κατάστασης χρησιμοποιώντας trends και πιθανότητες. Είναι μια ιδιαίτερα χρήσιμη

κατηγορία ειδικά όταν θέλουμε να προβλέψουμε την πιθανότητα ο ασθενής να εμφανίσει επιπλοκές. Εδώ ανήκει και η ευρύτερη περίπτωση της πρόβλεψης στη δημόσια υγεία (κοιτώντας δεδομένα που αφορούν πληθυσμούς πχ η πρόβλεψη επιδημιών και οι προειδοποιήσεις οταν τα πρώτα σήματα εμφανιστούν). [2]

Η τελευταία κατηγορία, αυτή της καθοδηγητικής αναλυτικής αφορά περιπτώσεις όπου προτείνεται σχέδιο δράσης στους επαγγελματίες υγείας για την καλύτερη δυνατή απόφαση. Για παράδειγμα, ο επαγγελματίας υγείας μπορεί να λάβει σύσταση να αποφύγει να δώσει μια συγκεκριμένη θεραπεία στον ασθενή λόγω παρατηρούμενων παρενεργειών στον ασθενή και προβλεπόμενων απο το σύστημα επιπλοκών. [2]

Πρόσφατα, κάποιοι ερευνητές έχουν προσθέσει την κατηγορία της γνωσιακής αναλυτικής (cognitive) όπου χρησιμοποιώντας μηχανική μάθηση η εφαρμογή μαθαίνει και αυτοβελτιώνεται και αυτοεξελίσεται. Σε αυτή την κατηγορία μειώνεται η ανάγκη των καλών δεδομένων. [2]

2.2 Επεξεργασία Φυσικής γλωσσας

Η επεξεργασία φυσικής γλώσσας (natural language processing – NLP) είναι ένας κλάδος της επιστήμης των υπολογιστών, πιο συγκεκριμένα της τεχνητής νοημοσύνης που ασχολείται με το να κάνει τον υπολογιστή να κατανοεί γραπτό και προφορικό λόγο όπως τον κατανοούν και οι άνθρωποι [12].

Ο κλάδος της επεξεργασίας φυσικής γλώσσας συνδυάζει υπολογιστική γλωσσολογία (τη βασισμένη σε κανόνες κωδικοποίηση της γλώσσας) με τεχνητή νοημοσύνη ώστε να κάνουν τον υπολογιστή να κατανοήσει τον ανθρώπινο λόγο και το πλήρες ‘νόημα’ αυτού συμπεριλαμβανομένου της διάθεσης/κίνητρο του συγγραφέα ή ομιλητή καθώς και το συναισθηματικό πρόσημο του λόγου [12] .

Κάποιες απο τις προκλήσεις που καλούνται να αντιμετωπίσουν οι ερευνητές και προγραμματιστές όταν αναπτυσουν εφαρμογές που βασίζονται σε επεξεργασία φυσικής γλώσσας είναι το γεγονός οτι η ανθρωπινη γλώσσα παρουσιάζει πολλά ομώνυμα, ομόφωνα, ιδιωματισμούς, μεταφορές, σαρκασμούς καθώς και σε πολλές περιπτώσεις γραμματικά και συντακτικά λάθη. Κάποια παραδείγματα:

- ‘πιάνω το αστείο’ και ‘πιάνω ενα μήλο’. Η πρώτη φράση χρησιμοποιεί το ρήμα ΄πιάνω΄μεταφορικά

- ‘κράτα μικρό καλάθι’. Είναι μια ιδιωματική φράση που σημαίνει ‘κράτα μικρές προσδοκίες΄
- ΄βάνω΄ αντί για ΄βάζω΄

Η επεξεργασία φυσικής γλώσσας είναι πίσω απο εφαρμογές που μεταφράζουν κείμενα απο μια γλώσσα σε μία άλλη, εφαρμογές που ακολουθούν ανθρώπινες εντολές μέσω ομιλίας και εφαρμογές που συνοψίζουν μεγάλα κείμενα σε πραγματικό χρόνο.

Κύριοι τομείς χρήσης επεξεργασίας φυσικής γλώσσας είναι [12]:

- Ανίχνευση κακόβουλων (spam) e-mail. Υπάρχουν εφαρμογές που ανιχνεύουν την χρήση υπερβολικών οικονομικών όρων, επιθετικές λέξεις και εκφράσεις, δημιουργία αίσθησης βιασύνης, γραμματικά λάθη και ορθογραφικά λάθη σε κύρια ονόματα και ονόματα εταιρειών
- Εξυπηρέτηση πελατών. Υπάρχουν εφαρμογές που μπορούν να συνομιλήσουν με έναν πελάτη και να δώσουν λύσεις σε εύκολα ερωτήματα του ή να ανακατευθύνουν το ερώτημα του σε έναν υπάλληλο. Οι εφαρμογές αυτές αναγνωρίζουν τον ανθρώπινο λόγο σε πραγματικό χρόνο και είναι ικανές να απαντήσουν με τον ίδιο τρόπο.
- Ανάλυση συναισθήματος. Για αυτή την κατηγορία που είναι και η έμφαση της εργασίας γίνεται αναφορά σε ξεχωριστή ενότητα παρακάτω
- Σύνοψη μεγάλων κειμένων. Είτε προς χρήση σε ερευνητικές βάσεις δεδομένων είτε προς κατανάλωση απο αναγνώστες που θέλουν να κατανοήσουν πολλά κείμενα σε λίγο χρόνο
- Ψηφιακοί βοηθοί πχ. Siri και Alexa

## 2.3 Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος (sentiment analysis or opinion mining) στηρίζεται στην επεξεργασία φυσικής γλώσσας και είναι μια μέθοδος που επιτρέπει την ανάθεση συναισθηματικού προσήμου σε ελεύθερο κείμενο. Το πρόσημο μπορεί να είναι θετικό, αρνητικό ή ουδέτερο ή θα μπορούσε να είναι και πιο συγκεκριμένο πχ. θυμός, φόβος, χαρά.

Η ανάλυση συναισθήματος εμφανίστηκε το 2004 εξερευνώντας αν το συναίσθημα είναι θετικό ή αρνητικό. Στη συνέχεια τα συναισθήματα που μπορούσαν να ανιχνευτούν

διευρύνθηκαν (φόβος, λύπη, έκπληξη κλπ) και προστέθηκε και η ένταση του συναισθήματος (πχ. Πολυ αρνητικό συναίσθημα) [13].

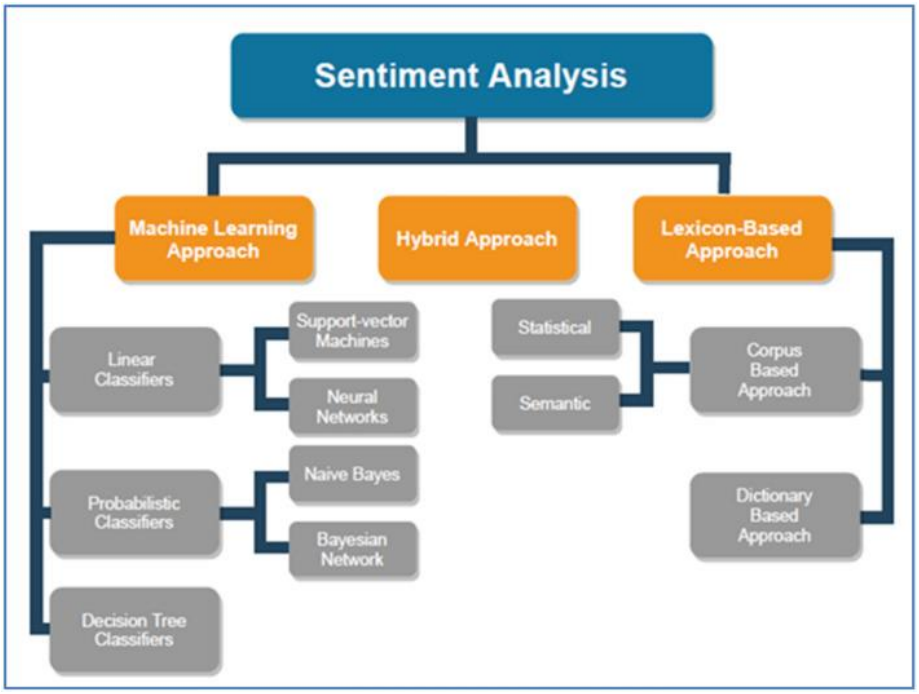
Η ανάλυση συναισθήματος στα κοινωνικά δίκτυα αποτελεί ενα διαδεδομένο ερευνητικό θέμα. Είναι δημοφιλές στον τομέα του επιχειρησιακού μάρκετινγκ όπου σφυγμομετρούν την άποψη του κοινού προς το προιον που προωθούν. [14] Μια ακόμη σημαντική χρήση είναι η σφυγμομέτρηση της κοινής γνώμης για πολιτικούς λόγους (πχ επερχόμενες εκλογές) ή για μια συγκεκριμένη εταιρεία. Η ανάλυση συναισθήματος μπορεί να χρησιμοποιηθεί για να εξαχθούν συμπεράσματα για την γνώμη των χρηστών για μια ταινία ή ακομη και για μια συνταγή μαγειρικής.

Πέρα απο ανάλυση συναισθήματος του συνολικού μηνύματος, μια εφαρμογή ανάλυσης συναισθήματος μπορεί να αναθέσει συναίσθημα σε διαφορετικές λέξεις ή φράσεις του μηνύματος ή μπορεί να επισημάνει διαφορετικές θεματικές κατηγορίες σε ένα μήνυμα και να αποδώσει διαφορετικό συναίσθημα σε καθεμία απο αυτές (aspect-based sentiment analysis). [15] Για παράδειγμα το μήνυμα: ‘ο ιατρός μου είναι ακριβός αλλά πάντα στην ώρα του’ έχει διπλή σημασία με διπλό συναίσθημα.

### 2.4 Τεχνικές Επεξεργασίας Φυσικής Γλώσσας

Στην επεξεργασία φυσικής γλώσσας υπάρχουν τρεις βασικές κατηγορίες τεχνικών όπως φαίνεται κ στην Εικόνα 4:

- με βάση λεξικά και καθορισμό κανόνων
- με χρήση μηχανικής μάθησης
- και η υβριδική κατηγορία που χρησιμοποιεί σειριακά λεξικό για να αναθέσει πρόσημο και στη συνέχεια χρησιμοποιεί αυτό σαν μια παραπάνω παράμετρο στο μοντέλο μηχανικής μάθησης.




*Εικόνα 4: Τεχνικές Ανάλυσης Συναισθήματος* [16]

### 2.4.1 Λεξικά Συναισθημάτων

Η ανάλυση συναισθήματος με λεξικό χρειάζεται μικρή προετοιμασία και η ιδέα είναι αρκετά απλή. Στην πιο απλοϊκή της μορφή, μπορούμε να βρούμε το συναίσθημα μιάς πρότασης μετρώντας τον αριθμό των φορών που εμφανίζεται η λέξη ΄λυπη΄σε ενα tweet.

Αυτή η μέθοδος λειτουργεί όπως είναι (out-of-the-box) σε πολλές περιπτώσεις και μπορεί να είναι λειτουργική με μικρή προεργασία.

Όμως αν κάποιος θέλει να επενδύσει χρόνο για να αυξήσει την ακρίβεια της, τότε το λεξικό χρειάζεται αρκετή προσπάθεια με το χέρι για να συντηρηθεί. Ένα επιπλέον μειονέκτημα ειναι οτι η ανάλυση συναισθήματος με λεξικά δεν αντιμετωπίζει καλά την ‘φιλική’ ανεπίσημη χρήση της γλώσσας και δεν εξετάζει το πρόσημο στα πλαίσια όλου του κειμένου αλλά κάθε φράση ξεχωριστά. Για παράδειγμα στη φράση ‘είμαι λιώμα΄ είναι δύσκολο ένα λεξικό να διακρίνει αν η λέξη ΄λιωμα’ έχει θετικό ή αρνητικό πρόσημο μιας που στην παρούσα φράση χρησιμοποιείται μεταφορικά και ανάλογα με το υπόλοιπο κείμενο θα μπορούσε να έχει και τα δύο πρόσημα. [17]

Οι πιο γνωστές τεχνικές ανάλυσης συναισθήματος με χρήση λεξικών είναι οι Textblob, Vader, SentiStrength, Senti Word Net, Linguistic Inquiry Word Count (LIWC) και Affective Norms for English Words (ANEW) [18].

Αναφέρονται ενδεικτικά οι μέθοδοι Vader και Textblob.

#### 2.4.1.1 VADER

Η μέθοδος Vader (Valence Aware Dictionary and Sentiment Reasoner) είναι βασισμένη σε λεξικό και κανόνες και χρησιμοποιείται ιδιαίτερα στην ανάλυση συναισθήματος απο κοινωνικά δίκτυα. Εφευρέθηκε το 2014. [17]

Η μέθοδος χρησιμοποιεί μια λίστα απο λέξεις που λαμβάνουν πρόσημο είτε θετικό είτε αρνητικό ανάλογα με τη σημασιολογική τους προέλευση ώστε να υπολογιστεί το συναίσθημα όλου του κειμένου. [17]

Το αποτέλεσμα της εφαρμογής Vader σε μια δοθείσα πρόταση είναι οι πιθανότητες η πρόταση αυτή να είναι: θετική, αρνητική και ουδέτερη.

Για παράδειγμα: «Περνάω σούπερ»

- Θετικό πρόσημο: 99%
- Ουδέτερο πρόσημο: 0%
- Αρνητικό πρόσημο: 1% [17]

#### 2.4.1.2 TextBlob

Η βιβλιοθήκη Textblob χρησιμοποιείται στην Python για την επεξεργασία κειμένου. Παρέχει API και χρησιμοποιείται για ποικιλία εργασιών στην επεξεργασία φυσικής γλώσσας, μια εκ τις οποίες είναι και η ανάλυση συναισθήματος. [17]

Η βιβλιοθήκη επιστρέφει τιμές για δυο (2) μεταβλητές για κάθε δοθείσα πρόταση:

- Πόλωση (Polarity), οι τιμές είναι ανάμεσα στο -1 και 1. -1 δηλώνει αρνητικό συναίσθημα και +1 δηλώνει θετικό συναίσθημα
- Υποκειμενικότητα (Subjectivity), οι τιμές είναι ανάμεσα το 0 και 1. Οι υποκειμενικές προτάσεις συνήθως αφορούν προσωπική γνώμη συναίσθημα ή κρίση [17]

### 2.4.2 Αλγόριθμοι Μηχανικής Μάθησης

Η μηχανική μάθηση ως υποκατηγορία και μέθοδος εξόρυξης δεδομένων ασχολείται με τη μελέτη αλγορίθμων οι οποίοι βελτιώνουν την συμπεριφορά τους σε κάποια εργασία που τους έχει ανατεθεί χρησιμοποιώντας την εμπειρία τους ως εκείνη τη στιγμή. [19]

Η μηχανική μάθηση χωρίζεται σε δύο (2) κατηγορίες: στη μηχανική μάθηση υπο επίβλεψη γνωστή και ως supervised learning και στη μηχανική μάθηση χωρίς επίβλεψη γνωστή και ως unsupervised learning.

Στη μάθηση υπο επίβλεψη ο αλγόριθμος προβλέπει ένα αποτέλεσμα ή ταξινομεί σε ομάδες μαθαίνοντας απο δεδομένα που έχουν ήδη κατηγοριοποιηθεί (labelled) ως προς την εργασία που καλείται ο αλγόριθμος να επιτελέσει. Για παράδειγμα δίνοντας στον αλγόριθμο έναν μεγάλο αριθμό απο emails που έχουν ήδη κατηγοριοποιηθεί για το αν ειναι κακόβουλα ή οχι, ο αλγόριθμος εκπαιδεύεται στα χαρακτηριστικά των emails κ μπορεί να προβλέψει αν είναι κακόβουλα. Λόγω του οτι είναι γνωστό το αν είναι κακόβουλα ή όχι ο αλγόριθμος βελτιώνεται εξετάζοντας την ποιότητα των προβλέψεων και ξαναπροσπαθώντας.

Οι πιο γνωστοί αλγόριθμοι στη μάθηση υπο επίβλεψη που χρησιμοποιούνται ευρέως σε προβλήματα ανάλυσης συναισθήματος είναι ο Random Forest (RF), ο Naïve Bayes (NB) και ο Support Vector Machine (SVM).

Στη μάθηση χωρίς επίβλεψη, ο αλγόριθμος καλείται να ομαδοποιήσει τα δεδομένα χωρίς να ξέρει εκ των προτέρων ποιές είναι οι ομάδες. Τα δεδομένα εισόδου δηλαδή δεν έχουν κατηγοριοποιηθεί. Αυτή είναι και η βασική διαφορά των δυο κατηγοριών μηχανικής μάθησης. Αυτό σημαίνει οτι μπορούμε να χρησιμοποιήσουμε αυτή τη μέθοδο έχοντας αποτελέσματα πιο γρήγορα όμως ιστορικά η ακρίβεια αυτών των αλγορίθμων είναι χαμηλότερη.

Οι βασικές εργασίες που επιτελεί ένας αλγόριθμος χωρίς επίβλεψη είναι:

- Η ομαδοποίηση (clustering) όπου τα δεδομένα ομαδοποιούνται με βάση κοινά χαρακτηριστικά. Μια περίπτωση χρήσης εδώ είναι στο marketing όπου οι πελάτες ενός καταστήματος ομαδοποιούνται σε ομάδες όπως: «αγαπάνε το κατάστημα», «θέλουν κίνητρο για ν αγοράσουν, επισκέπτονται πολλά καταστήματα», «αγοράζουν συγκεκριμένα πράγματα» κλπ
- Η συσχέτιση όπου δημιουργούνται σχέσεις μεταξύ δεδομένων. Μια περίπτωση χρήσης εδώ είναι στις προτάσεις αγοράς πχ. Οι πελάτες που αγοράζουν παμπερς, αγοράζουν και βρεφικά γάλατα
- Τέλος η μείωση παραμέτρων. Εδώ μια περίπτωση χρήσης είναι στην προεπεξεργασία δεδομένων όπου η μάθηση χωρις επίβλεψη μπορεί να μειώσει τα

χαρακτηριστικά εισόδου όταν αυτά έχουν υψηλή συσχέτιση μεταξύ τους ώστε να μειώσει τον «θόρυβο»

Ο πιο γνωστός αλγόριθμος μηχανικής μάθησης χωρίς επίβλεψη που χρησιμοποιείται σε προβλήματα ανάλυσης συναισθήματος είναι η ανάδειξη και μοντελοποιηση θεμάτων (topic modelling) με χρήση διάφορων μοντέλων όπως Latent Dirichlet Allocation (LDA) ή Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM).

Η μηχανική μάθηση (machine leaning) παρουσιάζει σημαντικές αποκλίσεις απόδοσης ανάλογα με την περίπτωση ανάλυσης συναισθήματος (σε κάποιες τα πάει πολύ καλά, σε κάποιες χάλια), είναι εξαρτώμενη απο το μέγεθος του συνόλου δεδομένων (dataset) και τέλος δεν είναι πολύ αποτελεσματική σε δεδομένα που παρουσιάζουν πολύ έντονο αρνητικό συναίσθημα [18]. Πολλές φορές είναι επεξεργαστικά πολύ ακριβές μέθοδοι. Τέλος πολλές μέθοδοι μηχανικής μάθησης λειτουργούν ως μαύρο κουτί (black-box) που σημαίνει οτι είναι δύσκολο στον άνθρωπο να εξηγήσει γιατί πάρθηκε μια απόφαση και ως εκ τούτου είναι πιο δύσκολο να γενικευτούν, να τροποποιηθούν ή να χρησιμοποιηθούν ως έχουν σε διαφορετικές περιπτώσεις χρήσης.

Τα τελευταία χρόνια χρησιμοποιούνται τεχνικές βαθιάς μάθησης (deep learning) που αποτελεί υποκατηγορία της Μηχανικής Μάθησης. Η βαθιά μάθηση πλεονεκτεί στο να αναγνωρίζει τις συντακτικές (αλλεργία, αλλεργικός, αλλεργιογόνο) και σημασιολογικές εξαρτήσεις μεταξύ λέξεων. Επιπλέον στη βαθιά μάθηση δε χρειάζεται να επενδυθεί πολύς χρόνος για να επιλεχθούν τα σημαντικά χαρακτηριστικά κατί που συνήθως γίνεται απο τον ίδιο τον ερευνητή. Τα πιο σημαντικά χαρακτηριστικά επιλέγονται αυτόματα κατα τη διάρκεια της εκπαίδευσης του μοντέλου απο ολόκληρο το σετ δεδομένων [20].

#### 2.4.2.1 Αλγόριθμος Random Forest

Όπως υπονοεί και το όνομα του αποτελείται από ένα μεγάλο αριθμό μεμονωμένων δέντρων απόφασης που λειτουργούν ως σύνολο, το πόσα δέντρα θα κατασκευάσει του το λέμε εμείς στο μοντέλο μας με βάση το νούμερο των n_estimators. Κάθε δέντρο απόφασης δίνει μια προτεινόμενη τιμή της εξαρτημένης μεταβλητής, όπως φαίνεται στη παρακάτω εικόνα.
Από το σύνολο των δέντρων αποφάσεων κάθε πιθανή τιμή της εξαρτημένης μεταβλητής, έχει προταθεί συγκεκριμένες φορές από το σύνολο των δέντρων αποφάσεων, που αποτελεί και τη βάση ταξινόμησης με βάση τον αλγόριθμο random forest. Επί της ουσίας αυτή η εικόνα παρουσιάζει μια απεικόνιση ενός ταξινομητή μεμονωμένων δέντρων απόφασης. Κάθε μεμονωμένο δέντρο στο τυχαίο δάσος ξετυλίγει μια πρόβλεψη και η πρόβλεψη με τις

περισσότερες ψήφους γίνεται η πρόβλεψη του μοντέλου μας. Ο αλγόριθμος λειτουργεί με την λογική, ότι πολλοί εκτιμητές υπερβολικής τοποθέτησης(overfitting estimators) μπορούν να συνδυαστούν για να μειώσουν την επίδραση αυτής της υπερβολικής τοποθέτησης, αυτή η λογική βασίζεται σε μια μέθοδο συνόλου που ονομάζεται Bagging. Η μέθοδος Bagging χρησιμοποιεί ένα σύνολο παράλληλων εκτιμητών, καθένας από τους οποίους ταιριάζει υπερβολικά στα δεδομένα και υπολογίζει τα αποτελέσματα για να βρει μια καλύτερη ταξινόμηση. Βασικό πλεονέκτημα του Random Forest είναι η ικανότητά του να διαχειρίζεται αποτελεσματικά μεγάλο αριθμό ανεξάρτητων μεταβλητών ενώ αντίστοιχα έχει χαμηλό απαιτούμενο χρόνο εκτέλεσής, οπότε είναι γρήγορος. Ενώ ένα αρκετά σημαντικό μειονέκτημα των random forests είναι το γεγονός ότι είναι πολύ απαιτητικά όσο αφορά τον χρόνο και υπολογιστικές δυνατότητες (computer resources). Θεωρείται ένα μοντέλο καθαρά «μαύρου κουτιού» (black box model). Το n_estimators αφορά το πόσα δέντρα (random trees) θα έχει το μοντέλο, by default = 100, όπως και στον KNN η παράμετρος random_state επιτρέπει τον έλεγχο των αποτελεσμάτων, παρόλο που ο αλγόριθμος επαναλαμβάνεται πολλές φορές χρησιμοποιώντας τυχαίες επιλογές χαρακτηριστικών και δειγμάτων. Το max_depth αφορά το μέγιστο βάθος του δέντρου. Εάν δεν υπάρχει, τότε οι κόμβοι επεκτείνονται έως ότου όλα τα φύλλα να μη μπορούν να επεκταθούν άλλο ή έως ότου όλα τα φύλλα περιέχουν λιγότερα από δείγματα min_samples_split samples. Η επιλογή max_features αφορά τον αριθμό των χαρακτηριστικών που πρέπει να λάβει υπόψη ο αλγόριθμος, παίρνει τιμές {“auto”, “sqrt”, “log2”} και το default είναι το “auto”. Το min_samples leaf αφορά το χαμηλότερο πλήθος των φύλλων του δέντρου. Το μέγεθος του δευτερεύοντος δείγματος ελέγχεται με την παράμετρο max_samples εάν bootstrap = True (προεπιλογή), διαφορετικά χρησιμοποιείται ολόκληρο το σύνολο δεδομένων για τη δημιουργία κάθε δέντρου. Το n_jobs αφορά τον αριθμό των jobs που τρέχουν παράλληλα.

#### 2.4.2.2 Αλγόριθμος Naïve Bayes

Το βασικό χαρακτηριστικό αυτού του αλγορίθμου είναι οτι υποθέτει ότι όλες οι παράμετροι εισόδου (features) είναι ανεξάρτητες η μία απο την άλλη. Σε αυτό το χαρακτηριστικό οφείλει και την πρώτη λέξη του ονόματος του – naïve (αφελής) [21].

Τα πλεονεκτήματα αυτού του αλγορίθμου είναι:

- Η ταχύτητα υπολογισμών που οφείλεται στην αφελή υπόθεση ότι οι παράμετροι εισόδου είναι ανεξάρτητες η μία απ την άλλη. Κάποιες φορές προτιμάται η ταχύτητα απο την ακρίβεια

- Ο αλγόριθμος αυτός λειτουργεί πολύ καλά σε περιπτώσεις χρήσης όπου τα δεδομένα έχουν πάρα πολλές παραμέτρους εισόδου, όπως συμβαίνει στην κατηγοριοποίηση κειμένου στα spam email.

Το μειονέκτημα αυτού του αλγορίθμου είναι η αφελής υπόθεση που αναφέρθηκε παραπάνω μιάς που σπάνια ισχύει στην πραγματική ζωή. [21]

#### 2.4.2.3 Topic Modelling

Η Μοντελοποίηση Θεμάτων (Topic Modelling) είναι μια μέθοδος που ανήκει στην μηχανική μάθηση χωρίς επίβλεψη. Η μέθοδος αυτή παρατηρεί λέξεις και έγγραφα και αναδεικνύει τα κύρια θέματα (‘topics’) ακόμη και όταν εμείς δεν ξέρουμε τι ψάχνουμε.

Κάθε έγγραφο είναι μια συλλογή απο λέξεις (bag of words), ως εκ τούτου η σειρά των λέξεων ή η γραμματική υπόσταση των λέξεων δε μας απασχολεί. Όμως οι λέξεις υφίστανται προεπεξεργασία ώστε να κρατιέται μόνο η ρίζα της λέξης, συνεπώς η ομαδοποίηση και ανάδειξη θεμάτων διευκολύνεται. [22] Για παράδειγμα, οι λέξεις ‘barking’, ‘bark’, ‘barked’ (γαβγίζω) θα ομαδοποιούνταν σε ‘bark’.

Δυο δημοφιλείς τεχνικές κράτησης της ρίζας είναι η λημματοποίηση (lemmatization προκύπτει απο το lemma = λημμα) και η αποκοπή καταλήξεων (stemming απο το stem = μίσχος). [23]

Και οι δυο τεχνικές έχουν τον ίδιο σκοπό: να ανάγουν τις λέξεις σε μια κοινή βάση αφαιρώντας γράμματα απο το τέλος της λέξης ωστε οι λέξεις να μπορούν να θεωρηθούν παρόμοιες στα πλαίσια μιας ανάλυσης δεδομένων. Έτσι το λεξιλόγιο των δεδομένων μας μικραίνει και είναι πιο εύκολο να επεξεργαστεί και αφαιρείται ο θόρυβος.

Μια γνωστή χρήση αυτών των τεχνικών είναι στις λέξεις που εισάγουμε και τα αποτελέσματα που επιστρέφει μια μηχανή αναζήτησης πχ google.

Και οι δυο τεχνικές είναι διαθέσιμες στο Natural Language ToolKit (nltk) της Python.

Όμως οι δυο τεχνικές έχουν ουσιαστικές διαφορές.

Στην αποκοπή καταλήξεων (stemming) γίνεται μια ωμή αποκοπή των τελευταίων γραμμάτων μιας λέξης και κρατείται η βάση (‘ριζα’). Αυτή η τεχνική είναι πιο γρήγορη, λειτουργεί με εφαρμογή κανόνων και είναι αποτελεσματική στο 80% των περιπτώσεων. Ο πιο γνωστός αλγόριθμος για αποκοπή καταλήξεων είναι ο αλγόριθμος του Porter [23].

Στη λημματοποίηση εξετάζονται οι διαφορετικές μορφές μιας λέξης και κρατιέται μόνο η ετυμολογική/σημασιολογική ρίζα της λέξης ώστε να ομαδοποιηθούν λέξεις που έχουν το ίδιο νόημα αλλά εκφράζονται με διαφορετικούς χρόνους ρήματος ή ως ουσιαστικό (‘λημμα’). Είναι μια τεχνική που απαιτεί χρήση λεξικών ώστε να βρεθεί η σημασιολογική και μορφολογική ρίζα κάθε λέξης. Στα πλεονεκτήματα της τείνει να είναι πιο αποτελεσματική απο την αποκοπή καταλήξεων ειδικά σε γλώσσες όπως τα ελληνικά ή τα ισπανικά με σημαντική μορφολογία.

Παράδειγμα χρήσης: και για τις δύο λέξεις ‘δίνω’ και ‘έδωσα’, η ρίζα θα είναι ‘δίνω’.

Το αποτέλεσμα της μοντελοποίησης θεμάτων είναι ένας αριθμός προκαθορισμένων ομάδων (clusters) που περιέχουν λέξεις που συνδέονται μεταξύ τους. Για κάθε λέξη υπολογίζεται η πιθανότητα να περιλαμβάνεται σε ένα θέμα και για κάθε έγγραφο υπολογίζεται η πιθανότητα να αναφέρεται/περιέχει κάθε θέμα. Το μειονέκτημα του αλγορίθμου είναι οτι πρέπει ο ερευνητής να εξετάσει αυτές τις ομάδες και να αποφασίσει αν η ομάδα έχει νόημα και ποιο είναι το κοινό χαρακτηριστικό τους (theme).

Η μέθοδος αυτή χρησιμοποιείται ευρέως στην ανάλυση δεδομένων απο κοινωνικά δίκτυα. Μια περίπτωση χρήσης είναι στον επιχειρηματικό τομέα όπου μια εταιρεία μπορεί να δει τα κυρια θέματα συζήτησης στα οποία αναφέρεται το όνομα της. Μια δεύτερη περίπτωση χρήσης ειναι στα τμήματα εξυπηρέτησης πελατών όπου η εταιρεία μπορεί να οργανώσει τα μηνύματα των πελατών και να καταλάβει τα κύρια προβλήματα του προιόντος. Έχει χρησιμοποιηθεί και στον ιατρικό τομέα για να εξάγει συμπεράσματα για τη γνώμη του κοινού ως προς συγκεκριμένα εμβόλια για τον HPV και την γρίπη.

2.4.2.3.1 Latent Dirichlet Allocation

Η μέθοδος Latent Dirichlet Allocation (LDA) αποτελεί την πιο γνωστή μέθοδο για μοντελοποίηση θεμάτων, εξού και η πρώτη λέξη του ονόματος αυτής της τεχνικής είναι latent – δηλαδή κρυμμένο.

Καθώς η μέθοδος ανήκει στην κατηγορία της μάθησης χωρίς επίβλεψη, ο αριθμός των κύριων θεμάτων προ-καθορίζεται. Δεν υπάρχει ιδανικός αριθμός θεμάτων και είναι αποτέλεσμα πειραματισμών. Σε γενικές γραμμές εξαρτάται απο τη χρήση των θεμάτων αργότερα και απο το εύρος των εγγράφων. Αν για παράδειγμα, χρειάζεται να επιβεβαιωθούν τα αποτελέσματα απο άνθρωπο τότε ένας μικρότερος αριθμός θεμάτων προτιμάται γιατί είναι πιο διακριτά τα θέματα πχ. πέντε (5) θέματα με τον κίνδυνο όμως οι λέξεις να είναι γενικές. Απο την άλλη μεριά αν το σύνολο των εγγράφων είναι αρκετά αχανές για παράδειγμα

θέματα συζήτησης στο twitter τότε είναι πιθανό να επιλεγεί ένας αριθμός θεμάτων μεγάλος πχ 60-70. Αυτό όμως ενέχει το κίνδυνο κάποια θέματα να είναι υπο-θέματα κάποιου άλλου και αρκετές λέξεις να επαναλαμβάνονται [24].

Η μέθοδος Hierarchical Dirichlet Process (HDP) είναι μια επέκταση του LDA με τη διαφορά οτι η μέθοδος αυτη μπορεί να ανακαλύψει τον αριθμό των θεμάτων αυτόματα χωρίς ο αριθμός αυτός να πρέπει να έχει προ-αποφασιστεί απο τον ερευνητή.

**Δείκτες επιτυχίας (success metrics)**

Δυο ποιοτικοί δείκτες που κρίνουν την επιτυχία των αποφάσεων του μοντέλου είναι η μέθοδος της λέξης ‘εισβολεα’ (‘word intrusion’) και του θέματος εισβολέα [25].

Στη μέθοδο της λέξης εισβολέα, επιλέγουμε τις λέξεις με υψηλότερες πιθανότητες του κάθε θέματος (πχ για ενα θέμα που φαίνεται να αφορά φρούτα, οι λέξεις με τις υψηλότερες πιθανότητες είναι: μήλο, μανταρίνι, πορτοκάλι, φράουλα) και ανακατεύουμε εκει μέσα μια λέξη που έχει πολύ υψηλές πιθανότητες να ανήκει σε σε διαφορετικό θέμα και έχει χαμηλή πιθανότητα για αυτό το θέμα (πχ για το προηγούμενο παράδειγμα μια τέτοια λέξη θα ήταν: πιπέρι). Έπειτα ζητάμε απο ενα σύνολο απο χρήστες να ξεχωρίσουν σε κάθε θέμα τη λέξη που δεν ανήκει εκεί και είναι ο εισβολέας. Εάν οι χρηστές ξεχωρίσουν εύκολα τη λέξη εισβολέα (πχ η πλειοψηφία τους τη ξεχωρίσει) σημαίνει οτι όντως οι υπόλοιπες λέξεις ‘δένουν’ μεταξύ τους και υπάρχει ένα κοινό θέμα. Εάν η επιλογή της λέξης εισβολέα γίνει κατά τύχη και οι γνώμες διίστανται τότε αυτό είναι ένδειξη οτι οι λέξεις που αποτελούσαν το θέμα δεν είχαν κάποιον εύληπτο δεσμό. [25]

Στη μέθοδο του θέματος εισβολέα, δοκιμάζουμε κάτι αντίστοιχο με την προηγούμενη μέθοδο αλλά αυτή τη φορά κρίνουμε τις πιθανότητες που έχουν δοθεί σε ένα έγγραφο (ή tweet) να ανήκει σε κάποιο θέμα. Για το πείραμα επιλέγονται κάποια έγγραφα και τρία απο τα πιο πιθανά θέματα τους. Σε κάθε σύνολο τοπ θεμάτων, προστίθεται ακόμη ένα θέμα που με βάση τις αποφάσεις του μοντέλου έχει χαμηλές πιθανότητες να αφορά το κάθε έγγραφο. Παρομοίως με την προηγούμενη μέθοδο ο χρήστης προσπαθεί να απομονώσει το ‘ασχετο’ θεμα που έχει εισβάλει στη λίστα. Για το κάθε έγγραφο βλέπει τον τίτλο και κάποιες εισαγωγικές προτάσεις. Ο ερευνητής καταγράφει πόσες φορές οι χρήστες απομόνωσαν όντως το θέμα εισβολέα [25].

Για αυτές τις ποιοτικές μεθόδους ελέγχου και δεδομένου οτι το θέμα της έρευνας είναι κατανοητό, οι χρήστες μπορούν να είναι και απο το Mechanical Turk της Amazon [25] το

οποίο επιτρέπει σε ερευνητές και εταιρείες πρόσβαση σε αρκετούς χρήστες με ενα μικρό αντίτιμο (https://www.mturk.com).

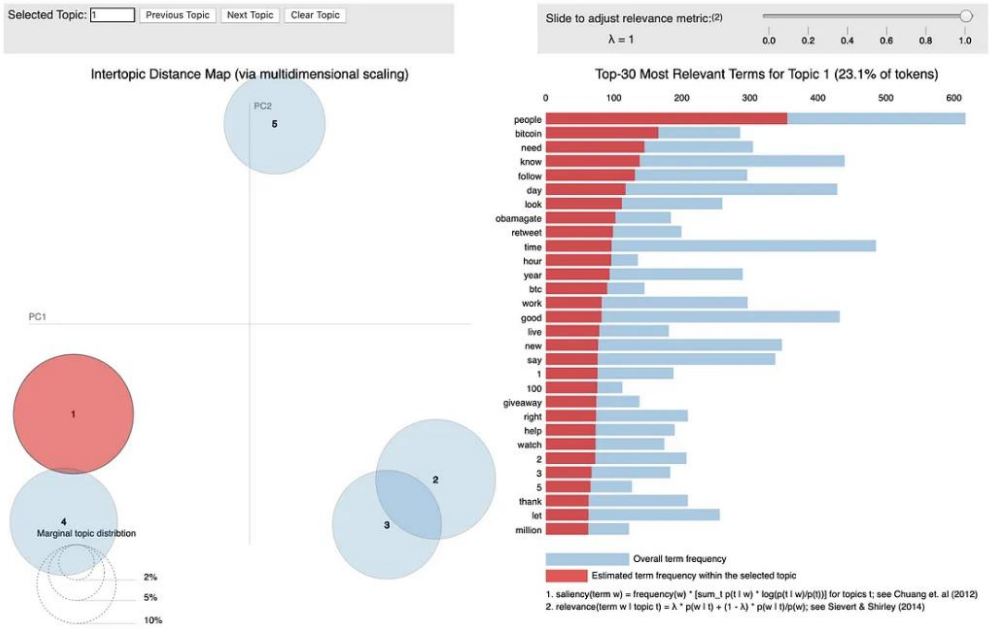
Υπάρχουν και δυο ποσοτικοί δείκτες επιτυχίας [26].

Ο πρώτος είναι η σύγχυση (perplexity) που αφορά στη δύναμη πρόβλεψης του μοντέλου σχετικά με τα έγγραφα. Εάν οι περισσότερες απο τις λέξεις σε ένα έγγραφο είναι απο τις πιο δημοφιλείς λέξεις ενός θέματος, τότε η πρόβλεψη για την ανάθεση θέματος είναι πιο ακριβής. Οσο χαμηλότερος ο δείκτης αυτός τόσο καλύτερη η δύναμη πρόβλεψης του μοντέλου.

Ο δεύτερος ποσοτικός δείκτης επιτυχίας είναι το σκορ συνοχής (coherence score) το οποίο δείχνει ποσο ερμηνεύσιμα είναι τα θεματα στο ανθρώπινο μάτι. Σε αυτή την περίπτωση τα θέματα αντιπροσωπεύονται σαν τις τοπ Ν λέξεις με τη μεγαλύτερη πιθανότητα να ανήκουν σε αυτό το θέμα. Συνοπτικά, το σκορ συνοχής μετράει πόσο ‘παρόμοιες΄οι λέξεις ενός θέματος είναι μεταξύ τους. Όσο πιο μεγάλη πιθανότητα υπάρχει να εμφανιστούν οι πιο δημοφιλείς λέξεις ενός θέματος ταυτόχρονα σε ένα θέμα τοσο καλύτερη είναι η κατηγοριοποίηση του μοντέλου.

Πέρα απο τους ποιοτικούς και ποσοτικούς δείκτες, είναι χρήσιμη και η οπτικοποίηση των αποτελεσμάτων.

Ένας συνηθισμένος τρόπος αναπαράστασης των θεμάτων και των δεδομένων είναι με φούσκες – κάθε φουσκα εκπροσωπεί ένα θέμα. Αν τα θέματα έχουν μοντελοποιηθεί με ιδανικό τρόπο, οι φούσκες απλώνονται συμμετρικά στους άξονες ώστε να καλύπτουν το σύνολο των δεδομένων, έχουν παρόμοιο μέγεθος και δεν επικαλύπτονται (διακριτά θέματα) [24].




*Εικόνα 5: Τεχνικές Οπτικοποίησης κύριων θεμάτων* [24]

2.4.2.3.2 Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM)

Ο Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) είναι ακόμη ένα μοντέλο στην κατηγορία της μηχανικής μάθησης χωρίς επίβλεψη και συγκεκριμένα στη μοντελοποίηση θεμάτων όπως και ο Latent Dirichlet Allocation που είδαμε στην προηγούμενη ενότητα.

Παρομοίως με τον LDA, ο GSDMM μπορει να αναδείξει τα θέματα που υπάρχουν στα έγγραφα λαμβάνοντας ως είσοδο μόνο τα ίδια τα έγγραφα.

Οι κύριες διαφορές του με τον LDA είναι:

- Στον GSDMM κάθε έγγραφο (tweet) αναφέρεται σε ένα θέμα μόνο ενώ στον LDA μπορεί να αναφέρεται σε περισσότερα
- Ο GSDMM ενδείκνυται για μοντελοποίηση θεμάτων μικρών κειμένων πχ 60 λέξεις και γι αυτό έχει χρησιμοποιηθεί στο παρελθόν στην εξόρυξη γνώσης απο tweets ή online κριτικές ταινιών [27]

Ο GSDMM απαιτεί δύο (2) εισόδους: τα έγγραφα και το μήκος του λεξικού. Όπως επίσης ορίζονται και κάποιες παράμετροι: ο επιθυμητός αριθμός θεμάτων K και ο αριθμός επαναλήψεων x. Το μοντέλο ξεκινάει και τρέχει για x επαναλήψεις προσπαθώντας να ταξινομήσει τα έγγραφα σε λιγότερες απο K ομάδες (clusters).

Οι ερευνητές Yin και Wang [27] που μελέτησαν τον αλγόριθμο πρότειναν την αναλογία του με την ακόλουθη διαδικασία γνωστή και ως Movie Group Process (MGP). Φανταζόμαστε τα έγγραφα ως μαθητές σε μια συζήτηση για ταινίες και τις λέξεις των εγγράφων ως τις ταινίες που έχει παρακολουθήσει ο μαθητής. Το πρόβλημα ομαδοποίησης είναι ανάλογο με την προσπάθεια να ταξινομηθούν οι μαθητές σε γκρουπς ανάλογα με τα κοινά τους ενδιαφέροντα στις ταινίες έτσι ώστε οι μαθητές με κοινά ενδιαφέροντα να βρίσκονται στο ίδιο τραπέζι και οι μαθητές σε διαφορετικά τραπέζια να έχουν διαφορετικά ενδιαφέροντα. Στην αρχή ταξινομούμε τους μαθητές τυχαία στα τραπέζια και μετά ζητάμε απο κάθε μαθητή να ξαναδιαλέξει τραπέζι με τη σειρά με δύο κανόνες 1. Να διαλέξει ένα τραπέζι με περισσότερους μαθητές 2. Να διαλέξει ένα τραπέζι του οποίου οι μαθητές που είναι εκεί έχουν παρόμοια ενδιαφέροντα στις ταινίες με αυτόν. Όπως συνεχίζει η διαδικασία, κάποια τραπέζια μεγαλώνουν ενώ άλλα εξαφανίζονται. Στο τέλος μόνο κάποια απο τα τραπέζια θα έχουν μαθητές και οι μαθητές σε κάθε τραπέζι θα μοιράζονται κοινά ενδιαφέροντα ταινιών.

Η Εικόνα 6 συγκρίνει LDA και GSDMM σε συγκεκριμένα σημεία όσον αφορά την ανάλυση σύντομων κειμένων όπως τα tweets. Ο GSDMM επιβεβαιωνει την ακρίβεια του όμως είναι πιο αργός καθώς χρειάζεται αρκετές επαναλήψεις για μια σωστή ταξινόμηση κατι το οποίο τον καθιστά ακριβό και ακατάλληλο για τεράστια σύνολα δεδομένων. Δεν υπάρχουν πολλοί έτοιμοι τρόποι (πχ δείκτες επιτυχίας, αναπαραστάσεις) για να κριθεί αποτελεσματικά παρόλο που οι ίδιοι δείκτες με τον LDA είναι εφαρμόσιμοι (πχ. coherence model) [28].

| | LDA | GSDMM |
|---|---|---|
| Topic-Clustering Accuracy | − | + |
| Model Training Speed | + | ✗ |
| Ability to handle large datasets | + | − |
| Visualisation / Interpretability tools | + | − |

*Εικόνα 6: Προτερήματα και Μειονεκτήματα μεταξύ των τεχνικών LDA και GSDMM* [28]

### 2.4.2.4 Μετασχηματιστές (Transformers)

Ο μετασχηματιστής είναι μια τεχνολογία της μηχανικής μάθησης, συγκεκριμένα νευρωνικών δικτύων και πιο συγκεκριμένα μια εξέλιξη αυτών που ονομάζεται βαθιά μάθηση (deep learning.) Αναπτύχθηκε αρχικά απο ερευνητές του Google Brain το 2017 [29] εκπαιδευόμενο σε 40 χιλιάδες προτάσεις της εφημερίδας Wall Street Journal (WSJ) με αρχική περίπτωση

χρήσης την μετάφραση απο τα αγγλικά στα γερμανικά όπου και είχε καλύτερη απόδοση απο τα state-of-the-art συστήματα μέχρι τότε.

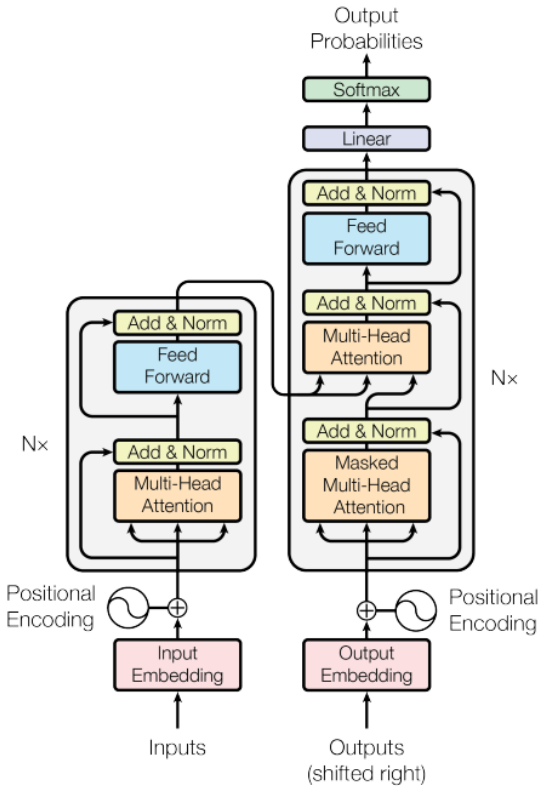
Απο τότε χρησιμοποιείται στην επεξεργασία φυσικής γλώσσας (NLP) με τις ακόλουθες περιπτώσεις χρήσης:

- Κατηγοριοποίηση κειμένου σε θέματα
- Εξόρυξη γνώσης
- Ερωτήσεις και απαντήσεις
- Αναγνώριση εικόνων
- Ανάλυση συναισθήματος

Πέρα απο ελεύθερο κείμενο σαν είσοδο, ενας μετασχηματιστής (transformer) μπορεί να χρησιμοποιήσει σαν είσοδο προφορικό λόγο, εικόνες, δομημένο κείμενο και σήματα.

Ένας μετασχηματιστής σχεδιάστηκε για να βοηθήσει τους υπολογιστές να κατανοήσουν το νόημα των λέξεων σε μια πρόταση αναλύοντας τα περιβάλλοντα στοιχεία στα οποία εμφανίζονται. Αυτό γίνεται διαιρώντας μια πρόταση σε μικρότερα τμήματα που ονομάζονται τεμάχια και στη συνέχεια συγκρίνοντας κάθε τεμάχιο με τα υπόλοιπα τεμάχια στην πρόταση.

Η σύγκριση γίνεται με τη χρήση ενός μοντέλου που περιλαμβάνει δύο βασικά στοιχεία: τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder). Ο κωδικοποιητής (multiple layers of encoding) αναλύει την είσοδο και παράγει μια αναπαράσταση της εισόδου, ενώ ο αποκωδικοποιητής (multiple layers of decoding) χρησιμοποιεί αυτήν την αναπαράσταση για να παράξει μια έξοδο. [29]




*Εικόνα 7: Αρχιτεκτονική ενός μετασχηματιστή* [29]

Το γεγονός ότι η εκπαίδευση του συστήματος μπορεί να σπάσει σε παράλληλα tasks επιτρέπει το σύστημα να εκπαιδευτεί σε πολύ μεγάλα σύνολα δεδομένων. Αυτό οδήγησε στην ανάπτυξη συστημάτων που έχουν ήδη εκπαιδευτεί σε τεράστια σύνολα δεδομένων όπως η Wikipedia και μπορούν να παραμετροποιηθούν για συγκεκριμένες περιπτώσεις χρήσεις. Τα δυο πιο γνωστά τέτοια συστήματα είναι το BERT και το GPT.

Η τεχνολογία των μετασχηματιστών έχει ήδη χρησιμοποιηθεί και στον τομέα της υγείας:

- Αναγνώριση ψηφιακών αρχείων ασθενούς και κατηγοριοποίηση τους ετσι ωστε το προσωπικό υγείας να μπορεί να βρει αυτό που ψάχνει πιο γρήγορα
- Μηχανική μετάφραση με σκοπό να βοηθήσει την επικοινωνία ασθενών που επισκέπτονται νοσοκομείο διαφορετικής χώρας. Η μηχανική μετάφραση χρησιμοποιείται για να μεταφράσει τις εξετάσεις και τις διαγνώσεις τους
- Ιατρικές ερωτήσεις και απαντήσεις: σε αυτή την περίπτωση το προσωπικό υγείας μπορεί να βρει απαντήσεις σε ερωτήσεις πολύ πιο γρήγορα
- Ανάλυση ιατρικών εικόνων ώστε να εντοπιστούν οι περιοχές των εικόνων με ιατρικό και διαγνωστικό ενδιαφέρον

2.4.2.4.1 BERT

Το μοντελο BERT (Bidirectional Encoder Representations from Transformers) είναι υποκατηγορία των μετασχηματιστών και κατασκευάστηκε απο τη Google το 2018 [30]. Η χρήση του είναι ελεύθερη.

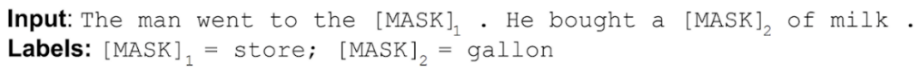
Η ανάπτυξη του μοντέλου αποτελείται απο δύο στάδια: η εκπαίδευση του μοντέλου (pre-training) και η παραμετροποιηση του μοντελου με στόχο τη βέλτιστη επίδοση (fine-tuning).

Αυτό το μοντέλο έχει εκπαιδευτεί σε μεγάλο όγκο κειμένων (για παράδειγμα το γνωστό pre-trained BERT έχει εκπαιδευτεί στη Wikipedia) ώστε να κατανοήσει τη φυσική γλώσσα και να εκτελέσει διάφορες εργασίες που σχετίζονται με την επεξεργασία φυσικής γλώσσας (NLP), όπως η αναγνώριση ονομάτων προσώπων, η κατηγοριοποίηση κειμένων και η ανάλυση συναισθημάτων.

Το BERT αναπαριστά την κάθε λέξη λαμβάνοντας υπόψη το νόημα της πρότασης. Για κάθε λέξη κοιτάει τι προηγείται στην πρόταση και τι έπεται. Για παράδειγμα στην πρόταση ‘I accessed the bank account’, σε ενα μοντέλο επεξεργασίας φυσικής γλώσσας όπως είναι το word2vec η λέξη bank θα είχε την ίδια αναπαράσταση ανεξαρτήτως του νοήματος της πρότασης. Δηλαδή το μοντέλο word2vec θα έδινε την ιδια αναπαράσταση σε αυτή τη λέξη είτε την έβρισκε σε αυτή την πρόταση είτε την έβρισκε στην πρόταση ‘river bank’ (όχθη ποταμού). Αντίθετα, το BERT δίνει διαφορετική αναπάρασταση σε αυτή τη λέξη ανάλογα με την πρόταση και μάλιστα ακόμη και στην ίδια πρόταση θα έδινε διαφορετική αναπαράσταση στην πρόταση ‘I accessed the bank’ και διαφορετική στην πρόταση ‘I accessed the bank account’ διότι κοιτάει και προς τις δυο κατευθύνσεις της λέξης μέσα στην πρόταση (bidirectional) [31].

Στην εκπαίδευση του BERT υπάρχουν δυο πολυ σημαντικές δοκιμασίες:

- Στην πρώτη δοκιμασία αφαιρούνται λέξεις απο προτάσεις (αντικαθιστούνται απο ένα ειδικό token [MASK]) και το BERT καλείται να προβλέψει ποια λέξη λείπει απο τις προτάσεις έχοντας διαβάσει όλη την ακολουθία απο λέξεις. Αυτή η δοκιμασία είναι γνωστή και ως masked language modelling.
- Στη δεύτερη δοκιμασία το BERT καλείται να διαβάσει δυο προτάσεις και ν’ αποφανθεί αν η δεύτερη πρόταση είναι πιθανό να έπεται δεδομένου του νοήματος της πρώτης. Αυτη η δοκιμασία είναι γνωστή και ως next sentence prediction (NSP).




*Εικόνα 8: Η πρώτη δοκιμασία του BERT* [31]


*Εικόνα 9: Η δεύτερη δοκιμασία του BERT* [31]

Το BERT έχει χρησιμοποιηθεί ευρέως στον βιοιατρικό τομέα και γνωστές παραλλαγές του εκπαιδευμένες σε ιατρικά κείμενα είναι το BioBERT και PubMedBERT [32]. Το BioBERT έχει εκπαιδευτεί σε περιλήψεις (abstracts) του PubMed και στο πλήρες κείμενο των PMC άρθρων και ως εκ τούτου έχει μεγαλύτερη ακρίβεια σε βιοιατρικά tasks [33]. Το σύνολο των δεδομένων πάνω στο οποίο εκπαιδεύτηκε το BioBERT φαίνεται στην Εικόνα 10:

| Corpus | Number of words | Domain |
|---|---|---|
| English Wikipedia | 2.5B | General |
| BooksCorpus | 0.8B | General |
| PubMed Abstracts | 4.5B | Biomedical |
| PMC Full-text articles | 13.5B | Biomedical |

*Εικόνα 10: Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση του BioBert* [33]

2.4.2.4.2 Greek BERT

Το Greek BERT είναι μια παραλλαγή του BERT που έχει εκπαιδευτεί εξολοκλήρου σε ελληνικά κείμενα [34], στο ελληνικό κομμάτι της Wikipedia, του Europarl και OSCAR. Το Greek BERT αποτελείται από έναν κωδικοποιητή που είναι εκπαιδευμένος να αναλύει τη σημασία των λέξεων στα ελληνικά κείμενα και να δημιουργεί μια αναπαράσταση του κειμένου σε μια χωρική αναφορικότητα, ώστε να μπορεί να χρησιμοποιηθεί για διάφορες εργασίες της επεξεργασίας φυσικής γλώσσας.

Υπάρχουν και άλλα μοντέλα BERT που προσπαθούν να επιτύχουν ακρίβεια σε πολλαπλές γλώσσες όπως το M-BERT που έχει εκπαιδευτεί σε 100 διαφορετικές γλώσσες και το XML. Όμως λόγω του δύσκολου και σχετικά σπάνιου αλφάβητου και λέξεων στην ελληνική γλώσα δεν έχουν μεγάλη ακρίβεια στα ελληνικά [34] καθώς πολύ λίγες λέξεις του συνολικού corpus και στις δύο περιπτώσεις αφορούν τα ελληνικά (1-2%).

| Corpus | Size (GB) | Training pairs (M) | Tokens (B) |
|---|---|---|---|
| Wikipedia | 0.73 | 0.28 | 0.08 |
| Europarl | 0.38 | 0.14 | 0.04 |
| OSCAR | 27.0 | 10.26 | 2.92 |
| Total | 29.21 | 10.68 | 3.04 |

*Εικόνα 11: Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση του Greek BERT* [34]

2.4.2.4.3 GPT

Η δεύτερη πιο γνωστή κατηγορία μετασχηματιστών είναι το Generative Pre-trained Transformer (GPT).

Σε σχέση με το BERT που ενδείκνυται για εργασίες επεξεργασίας φυσικής γλώσσας που σχετίζονται με την κατηγοριοποίηση των εγγράφων, το GPT ενδείκυται για εργασίες που παράγουν γλώσσα [32].

Μια γνωστή παραλλαγή του GPT στον ιατρικό τομέα είναι το BioGPT το οποίο έχει εκπαιδευτεί σε βιοιατρική βιβλιογραφία (15 εκατομμύρια abstracts του PubMed) και δεδομένου αυτού είχε καλύτερη απόδοση σε συγκεκριμένες βιοιατρικές εργασίες επεξεργασίας φυσικής γλώσσας [32].

Στην Εικόνα 12 βλέπουμε την απόδοση αρκετών μετασχηματιστών που έχουν αναφερθεί εώς τώρα στην εργασία των ερωτήσεων και απαντήσεων (PubMedQA).

| Model | Accuracy |
|---|---|
| PubMedBERT [9] | 55.8 |
| BioELECTRa [28] | 64.2 |
| BioLinkBERT$_{base}$ [29] | 70.2 |
| BioLinkBERT$_{large}$ [29] | 72.2 |
| GPT-2$_{medium}$ | 75.0 |
| BioGPT | **78.2** |

*Εικόνα 12: Αποτελέσματα μοντέλων σε εργασία ερωτήσεων και απαντήσεων (PubMedQA)* [32]

Πρόσφατα άνοιξε στο κοινό το chatGPT το οποίο αναπτύχθηκε απ τον οργανισμό OpenAI και έχει εκπαιδευτεί σε τεράστιο όγκο δεδομένων που περιλαμβάνει βιβλία, άρθρα και websites.

### 2.4.3 Υβριδικές τεχνικές

Η κύρια πρόκληση βρίσκεται στο γεγονός οτι ο κύριος όγκος δεδομένων ειναι μη δομημενος. Μια άλλη πρόκληση είναι οτι οι κατηγορίες που αναφέρθηκαν παραπάνω έχουν διάφορα μειονεκτήματα όταν χρησιμοποιούνται ξεχωριστά. Γι αυτό το λόγο στην ερευνητική κοινότητα συνήθως χρησιμοποιούνται συνδιαστικά (υβριδικό μοντέλο). Όταν χρησιμοποιούνται μαζί τα αποτελέσματα έχουν βρεθεί πιο ακριβή και πιο σταθερά [18].

# Κεφάλαιο 3 Εξόρυξη γνώσης απο κοινωνικά δίκτυα στον τομέα της υγείας

## 3.1 Παρακολούθηση δημόσιας υγείας

Τα τελευταία χρόνια υπάρχει ενδιαφέρον στον τομέα της υγείας στην εξόρυξη γνώσης απο κοινωνικά δίκτυα για πολλαπλούς σκοπούς.

Κάθε φορά που ένας πολίτης θα λάβει υπηρεσίες υγείας, θα δημιουργήσει μια γνώμη και ένα συναίσθημα για αυτην την υπηρεσία. Η επεξεργασία φυσικής γλώσσας και η ανάλυση συναισθήματος έχουν καταστήσει δυνατόν να συλλεχθούν και να επεξεργαστούν αυτές οι γνώμες και τα συναισθήματα ωστε να εξαχθούν ολιστικά συμπεράσματα.

Η ανάλυση συναισθήματος σε μηνύματα που αφορούν ιατρικά θέματα είναι ενα πολυ-συνθετο πρόβλημα. Ο χρήστης μπορεί να υποφέρει απο ένα σύνδρομο το οποίο του προκαλεί δυσαρέσκεια παρόλο που μπορεί να έμεινε ευχαριστημένος απο την τελευταία του εξέταση. Αυτά είναι δύσκολο να διαχωριστούν αν συνυπάρχουν σε ενα μήνυμα και να αποδοθεί ένα συναίσθημα για όλο το μήνυμα.

Ένας σημαντικός λογος είναι η παρακολούθησης της δημόσιας υγείας. Η επιστήμη της διανομής και μετα-προσδιορισμού της πληροφορίας σε ψηφιακά μέσα με απότερο σκοπό να βελτιωθεί η δημόσια υγεία και πολιτική της δημόσιας υγείας ονομάζεται επιδημιολογία της πληροφόρησης (infodemiology) [35]. Ετυμολογικά η λέξη αποτελεί συνδυασμό των λέξεων πληροφορία (information) και επιδημιολογία (epidemiology). Επιδημιολογία ειναι η επιστήμη που ασχολείται με την κατανομή και εξέλιξη νοσημάτων στον ανθρώπινο πληθυσμό και των παραγόντων που τα επηρεάζουν. Παραδοσιακά τα εργαλεία που χρησιμοποιούνταν για τη συλλογή πληροφορίας στην επιδημιολογία ήταν οι έρευνες, τα ιατρικά αρχεία και οι μελέτες πληθυσμών. Δυστυχώς η εξαγωγή γνώσης με αυτά τα μέσα σχεδόν ποτέ δεν ήταν σε πραγματικό χρόνο. Αντίθετα στην επιδημιολογία της πληροφορίας τα δεδομένα αυτά συλλέγονται σχεδόν σε πραγματικό χρόνο. Τέτοια παραδείγματα είναι:

- Ανάλυση αναζητήσεων απο μηχανές αναζήτησης για να προβλεφθεί η αύξηση μιας ασθένειας (πχ. Γρίπη)
- Η παρακολούθηση των tweets στο twitter προκειμένου να φανούν trends σε αύξηση συνδρόμων και άλλων ασθενειών

- Η παρακολούθηση των ιατρικών δημοσιεύσεων στο διαδίκτυο πχ ιστότοποι που προωθούν τον αντι-εμβολιασμό

Γενικότερα η ανάλυση του πως οι άνθρωποι ψάχνουν και περιηγούνται στο διαδίκτυο για πληροφορίες που σχετίζονται με την υγεία καθώς και το πως επικοινωνούν και μοιράζονται αυτες τις πληροφορίες είναι χρήσιμη γνώση στην παρακολούθηση πληθυσμών για λόγους δημόσιας υγείας

Ερευνητές έδειξαν οτι οι άνθρωποι συμβουλεύονται το διαδίκτυο μια εβδομάδα πριν επικοινωνήσουν με το γιατρό τους και το 2009 ερευνητές που σχετίζονται με την εταιρεία Google έδειξαν πως οι αναζητήσεις στη μηχανή αναζητήσεις της Google μπορούν να προβλέψουν ξεσπάσματα της εποχιακής γρίπης στις Ηνωμένες Πολιτείες [35].

Αυτή τη στιγμή δεν υπάρχει έτοιμος μηχανισμός που να εντοπίζει σε πραγματικό χρόνο τον επιπολασμό και τη σοβαρότητα ασθενειών που επηρεάζουν μεγάλο ποσοστό του πληθυσμού.

Το 2019, ερευνητές πρότειναν μια μεθοδο βασισμένη στα νευρωνικά δίκτυα και τη βαθιά μάθηση για συγχρονική παρακολούθηση μέσω Twitter του πληθυσμου της Αυστραλίας που υποφέρει απο εποχιακή αλλεργεία (pollen allergy), που εκτιμάται οτι επηρεάζει περίπου το 20% των Αυστραλών. Ταυτόχρονα έχει βρεθεί οτι σε ασθένειες όπως η εποχιακή αλλεργεία είναι δύσκολο να υπολογιστεί το ποσοστό των ανθρώπων που υποφέρει απο αυτή χρησιμοποιώντας επίσημες πηγές δεδομένων μιας που οι περισσότεροι άνθρωποι αγοράζουν θεραπείες (πχ. Αντισταμινικά) χωρίς να επισκεφτούν ιατρικό προσωπικό. Χρησιμοποίησαν το μοντέλο attention που μπορεί να συσχετίσει λέξεις, αναθέτοντας τους διαφορετικό βάρος, αυξάνοντας την ακρίβεια των αποτέλεσμάτων. Βασίστηκαν σε tweets ανθρώπων που μιλούσαν για την πάθηση τους και τη θεραπεία τους και τα αποτελέσματα της μελέτης τους έδειξε οτι θα μπορούσε να αναπτυχθεί μηχανισμός που να χρησιμοποιείται απο οργανισμούς δημόσιας υγείας [20].

### 3.2 Βελτίωση ποιότητας υγειονομικών υπηρεσιών

Το να μπορείς να κατανοήσεις την εμπειρία των ασθενών μετά απο επίσκεψη τους σε υγειονομική δομή είναι πολύ σημαντικό στη διαδικασία παροχής φροντίδας και κρίσιμο στην προσπάθεια της δομής να βελτιώσει τις υπηρεσίες της και να κινητοποιήσει το προσωπικό να δουλεύει πιο αποτελεσματικά. Επιπλέον η επεξεργασία φυσικής γλώσσας και η ανάλυση συναισθήματος σε πραγματικό χρόνο μπορεί να επιτρέψει σε μια δομή υγείας να απαντήσει σε επείγουσες αρνητικές κριτικές με προτεραιότητα προκειμένου να βελτιώσει τη σχέση της με αυτούς τους ασθενείς. [36]

Η κύρια μέθοδος απόκτησης αυτής της πληροφορίας είναι μέσω ερωτηματολογίων που διανέμονται περιοδικά. Όμως η μηχανική μάθηση και η ανάλυση συναισθήματος προσφέρουν μια πιο γρήγορη και φθηνή εναλλακτική συγκρίσιμης ακρίβειας.

Πηγές συλλογής δεδομένων εδώ είναι οι κριτικές της δομής στο διαδίκτυο, online ερωτηματολόγια και αναφορές στη δομή σε διάφορα websites και portals.

Ερευνητές εφάρμοσαν τεχνικές μηχανικής μάθησης χρησιμοποιώντας το λογισμικό εξόρυξης δεδομένων Weka σε 6412 σχόλια δημοσιευμένα το 2010 στο website του εθνικου συστήματος υγείας της Αγγλίας (English National Health Service) και σύγκριναν τα αποτελέσματα τους με τα αποτελέσματα του ερωτηματολογίου που συμπληρώνεται εγγράφως σε 161 νοσοκομεία ενηλίκων στην Αγγλία. Υπήρχε συμφωνία 81% στα αποτελέσματα που σχετίζονταν με την καθαριότητα της δομής, συμφωνία 84% στον δείκτη "μου συμπεριφέρθηκαν με αξιοπρέπεια" και 89% συμφωνία στο πόσοι θα συνέστηναν το νοσοκομείο σε κάποιον αλλο. Παρατήρησαν μια μέτρια συσχέτιση μεταξύ των προβλέψεων του αλγορίθμου και των αποτελεσμάτων της έρευνας (Spearman rho 0.37-0.51, $P<.001$ και για τους τρεις δείκτες). [37]

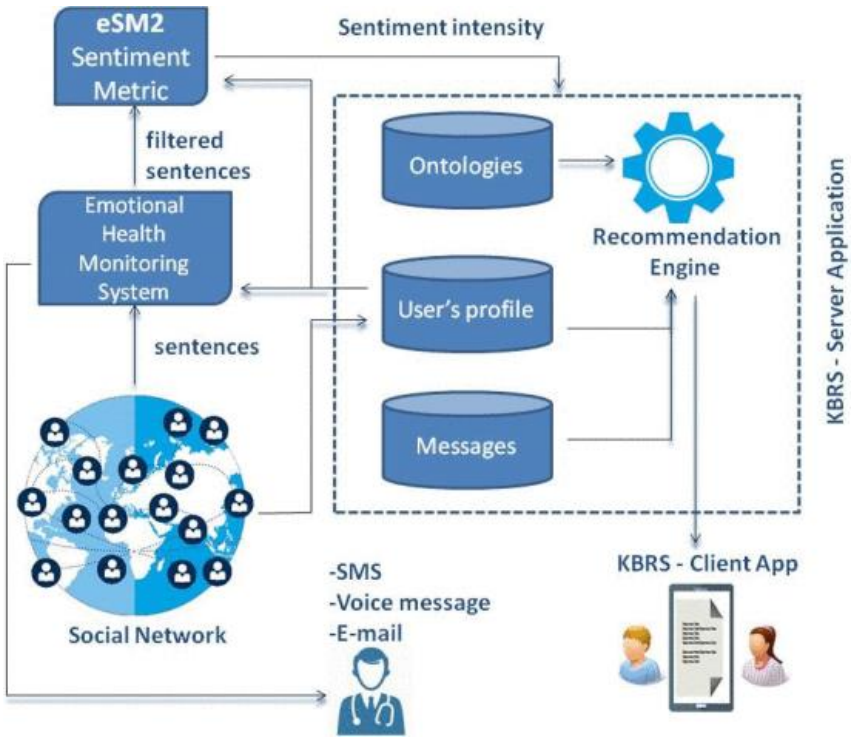
Καθώς τα σχόλια που ανέλυσαν ήταν μέρος ευρύτερου συνόλου ερωτήσεων μπόρεσαν να ελέγξουν το κατα πόσο ταύτισαν το συναίσθημα που ανάθεσαν με τις απαντήσεις στις αντίστοιχες ερωτήσεις κλειστού τύπου (που είναι ευκολο να ερμηνευθούν απο μια μηχανή) και έτσι δε χρειάστηκε κάποιος ερευνητής να αναθέσει συναισθηματικό πρόσημο με το χέρι στα δεδομένα εκπαίδευσης του αλγορίθμου. [37]

Απο την άλλη μεριά η πρόκληση ήταν οτι τα online σχόλια συνήθως είναι αρκετά πολωμένα (είτε θετικά είτε αρνητικά). Επιπλέον κάποια σχόλια ασθενών που περιέχουν ειρωνία, σαρκασμό ή χιούμορ δεν μπορούν να ανιχνευθούν εύκολα. Τέλος υπήρχαν αρκετές φράσεις που θα μπορούσαν να χρησιμοποιηθούν σε θετικό και αρνητικό περιεχόμενο. [37]

### 3.3 Καλύτερη κατηγοριοποίηση της έντασης μιας ασθένειας

Δεδομένα απο τα κοινωνικά δίκτυα μπορούν να βοηθήσουν στο να δείξουν τη συναισθηματική κατάσταση και συσσωρευμένο στρες ενός ασθενούς – και οι δύο καταστάσεις θεωρούνται επιβαρυντικές για μια ασθένεια. Ερευνητές ανέπτυξαν σύστημα μηχανικής μάθησης που αναγνωρίζει τα επίπεδα κατάθλιψης και άγχους ενός χρήστη με βάση τα tweets του και μπορεί να στείλει μια σειρά απο προειδοποιητικά μηνύματα (recommender system). Το σύστημα χρησιμοποιεί υπάρχοντα σημάδια για άγχος και κατάθλιψη πχ σύντομες φρασεις, χρήση του 'εγω'. Το σύστημα λαμβάνει ως δεδομένα

εισόδου τα tweets του χρήστη και κατηγοριοποιεί το συναίσθημα τους έχοντας λάβει υπόψη το φύλο, ηλικία και τοποθεσία του χρήστη. Φράσεις που σχετίζονται με το θυμό, την απέχθεια και την έκπληξη κατηγοριοποιούνται στο στρες ενώ φράσεις που εκπέμπουν φόβο ή λύπη κατηγοριοποιούνται στην κατάθλιψη. [38]




*Εικόνα 13: Αρχιτεκτονική συστήματος ανίχνευσης κατάθλιψης απο tweets* [38]

Ερευνητές ανέπτυξαν ένα μοντέλο μηχανικής μάθησης που υπολογίζει τον κίνδυνο ενος χρήστη να αναπτύξει διαβήτη τύπου 2 με βάση τα tweets τους. Οι χρήστες που συμμετείχαν συμπλήρωσαν πρώτα το επίσημο ερωτηματολόγιο υπολογισμού του κινδύνου για διαβήτη τύπου 2 (Type 2 Diabetes Mellitus -T2DM) και στη συνέχεια ερωτήθηκαν για το προφίλ τους στο Twitter. Οι ερευνητές δημιούργησαν ένα δικό τους λεξικό ενσωματώνοντας λέξεις και hashtags που αφορούν την άθληση όπως #5k (αγώνας 5 χιλιομέτρων), φαγητά, ονόματα εστιατορίων και άλλες λέξεις που σχετίζονται με την παχυσαρκία. Παρόλο που η ύπαρξη λέξεων απο το λεξικό μέσα στα tweets δεν ήταν συχνή  και παράλληλα υπάρχουν πολλές κρυμμένες παράμετροι που επηρεάζουν τον υπολογισμό αυτού του κινδύνου (όπως γενετικοί παράγοντες), οι ερευνητές υπολόγισαν οτι αν μια βελτιωμένη έκδοση του μοντέλου τους δοκιμάζοταν, θα αναγνώριζε 16000 διαβητικούς αμερικανούς και 140000 προ-διαβητικους αμερικάνους που είναι αυτή τη στιγμή αδιάγνωστοι [39].

| *Correct Label* | *Predicted Label* | *Relevance* |
|---|---|---|
| less-risk | less-risk | chicken waffles tea reading… |
| less-risk | at-risk | cake food starving sit sit heart… |
| at-risk | less-risk | bacon run cup pack writing rolls parkour… |
| at-risk | at-risk | catfish peanut butter pie picnic bland pop… |

*Εικόνα 14: Λέξεις απο tweets κατηγοριοποιημένες είτε σε χαμηλό ή υψηλό κίνδυνο* [39]

Επιπροσθέτως οι χρήστες των κοινωνικών δικτύων μπορεί να σχολιάσουν τη γνώμη τους για ένα εμβόλιο ή τις παρενέργειες ενός φαρμάκου.

Μελετητές χρησιμοποίησαν δείγμα απο 9581 tweets που αναφέροταν σε εμβόλια το 2019. Την ίδια περίοδο υπήρχε έξαρση ιλαράς (measles) στις Ηνωμένες Πολιτείες. Η ανάλυση συναισθήματος και η τεχνική TF-IDF έδειξαν οτι η πλειοψηφία των tweets (77%) αναφέρονταν στην αναζήτηση καλύτερων εμβολίων ενώ τα υπόλοιπα έδειξαν ανησυχία για την έξαρση ιλαράς και debate μεταξύ υπερμάχων και πολέμιων των εμβολίων. [40]

### 3.4 Εφαρμογές που συνομιλούν με χρήστες

Το επόμενο βήμα στην επεξεργασία φυσικής γλώσσας μετα την αυτόματη διάγνωση ενός προβλήματος είναι οι εφαρμογές που συνομιλούν με τον χρήστη σε πραγματικό χρόνο χρησιμοποιώντας φυσική γλώσσα.

Για παράδειγμα οι ασθενείς μπορούν να κανονίσουν τα ραντεβού τους, να λάβουν υπενθυμίσεις, να μάθουν τα αποτελέσματα των εξετάσεων και να δώσουν τη γνώμη τους 24 ωρες το 24ωρο. Με βάση τις συνομιλίες με την εφαρμογή, η δομή υγείας μπορεί να βελτιωθεί περαιτέρω. [15]

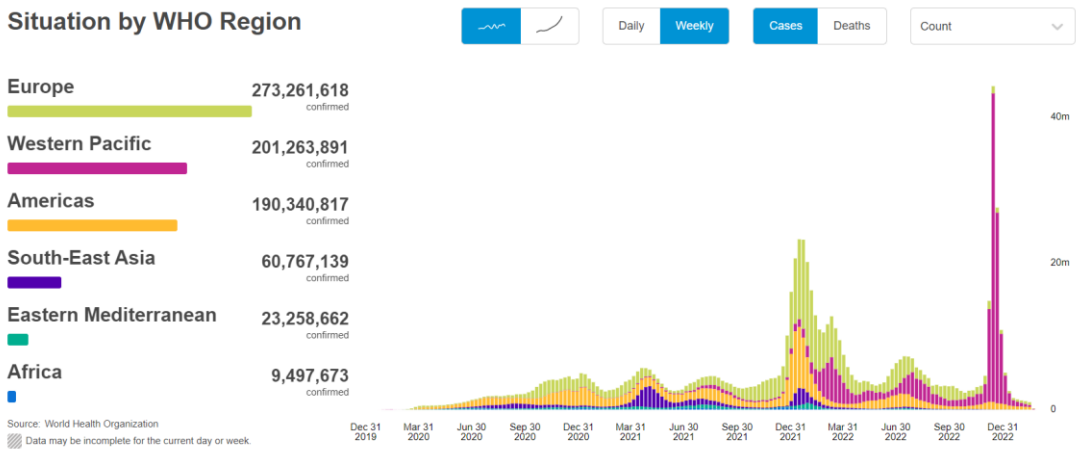
# Κεφάλαιο 4 Επεξεργασία φυσικής γλώσσας στο Covid-19

## 4.1 Covid-19

Η επιδημία του Covid-19 έχει επηρεάσει τις ζωές όλων παγκοσμίως απο το Μάρτιο του 2020 που ο ΠΟΥ κήρυξε την αρχή της επιδημιας. Το Covid-19 προκαλείται απο τον ιο SARS-CoV-2 και είναι μια εξαιρετικά μεταδοτική ασθένεια. Τα πιο συνηθισμένα συμπτώματα περιλαμβάνουν βήχα, πονόλαιμο, πυρετό και απώλεια γεύσης. Οι πιο πολλοί άνθρωποι που προσβάλλονται έρχονται αντιμέτωποι με μια ήπια εώς μέτρια ίωση. Κάποιοι όμως χρήζουν νοσηλείας. Μέχρι αυτή τη στιγμή έχουν υπάρξει 758 εκατομμύρια κρούσματα και 6.9 εκατομμύρια θάνατοι. [41]

Πολλά κράτη πήραν μέτρα μέσα σε αυτα τα 2.5 χρόνια για να περιορίσουν την εξάπλωση του ιού όπως εμφανίζονταν καινούριες μεταλλάξεις και ο αριθμός των κρουσμάτων αυξανόταν. Τα μέτρα περιλάμβαναν απαγόρευση μετακινήσεων και συναθροίσεων (‘lockdown’), απαγόρευση διασυνοριακών μετακινήσεων, επιβολή συγκεκριμένων αποστάσεων σε χώρους συνωστισμού και συστηματική χρήση μάσκας.

Τα πρώτα εμβόλια εγκρίθηκαν για χρήση εκτός κλινικών δοκιμών τον Δεκέμβριο του 2020 και μέχρι τον Μάρτιο του 2023 είχαν εμβολιαστεί με τουλάχιστον μια δόση πανω απο 5 δισεκατομμύρια άνθρωποι παγκοσμίως. [41]




*Εικόνα 15: Κρούσματα Covid-19 παγκοσμίως* [41]

## 4.2 Επεξεργασία φυσικής γλώσσας στα κοινωνικά δίκτυα κατα τη διάρκεια του Covid-19

Η πανδημία του Covid-19 καθήλωσε το μεγαλύτερο μέρος του πληθυσμου μέσα στο σπίτι του για πολύ μεγάλο διάστημα.

Τα μέτρα καθώς και τα εμβόλια έχουν προκαλέσει αφορμή για πολλές συζητήσεις και διχασμό απόψεων παγκοσμίως. Πολλοί ήταν αρνητικοί στην κάθετη επιβολή των μέτρων και σκεπτικοί απέναντι στον εμβολιασμό – γενικότερα υπήρξαν πολλές αντιδράσεις στα κοινωνικά δίκτυα.

Ερευνητές απέδειξαν οτι τα κοινωνικά δίκτυα μπορούν να χρησιμοποιηθούν για την παρακολούθηση της δημόσιας υγείας μιας που έδειξαν απο την ερευνα τους οτι πολλα συμπτώματα του Covid-19 παρουσιάστηκαν σε tweets πριν ανακοινωθούν επίσημα απο τον φορέα CDC στις Ηνωμενες Πολιτείες της Αμερικής [42].
Παρόλο που τα κοινωνικα δίκτυα αποτελούν έναν πυλώνα ενημέρωσης, κατα τη διάρκεια της πανδημίας παρατηρήθηκαν φαινόμενα τοξικού infodemics (δηλαδή διάχυση πληροφορίας αμφιβόλου ποιότητας). Για παράδειγμα ειδήσεις που σχετίζουν τα εμβόλια κατά του ιού Covid-19 με επιθυμία των κυβερνήσεων να ελέγχουν τους πολιτες ή συμβουλές καταπολέμησης του ιού με σκόρδο.

Μια μελέτη των αποτελεσμάτων αναζήτησεων στο Pubmed και στο Google Scholar με λέξεις κλειδιά ‘sentiment analysis’, ‘natural language processing’, ‘social media’, ‘Covid-19’ δείχνει τρεις θεματικές στις έρευνες που χρησιμοποίησαν την ανάλυση συναισθήματος με δεδομένα της περιόδου Covid-19.

- Αντιδράσεις στα εμβόλια
- Ψυχολογία του πληθυσμού
- Long Covid

### 4.2.1 Covid-19 και αντίδρασεις στα εμβόλια

Οι περισσότερες έρευνες με γενικότερη θεματική την πανδημία του Covid-19 και την ανάλυση συναισθήματος εστίασαν στη γνώμη και αντίδραση των πληθυσμών πάνω στα εμβόλια.

Ερευνητές ανέλυσαν 31,100 tweets χρηστών στην Αυστραλία που περιείχαν λέξεις σχετικές με τα εμβόλια του Covid-19 κατα τη διάρκεια μεταξύ Ιανουαρίου και Οκτωβριου του 2020. Οπτικοποίησαν σύννεφα λέξεων που εμφανίζονται συχνά και ανέπτυξαν ενα Latent Dirichlet Allocation (LDA) μοντελο μηχανικής μάθησης. Τέλος εφάμοσαν ανάλυση συναισθήματος με το πακέτο της R syuzhet αναθέτοντας σε κάθε tweet θετικό ή αρνητικό συναισθηματικό πρόσημο και ένα απο τα ακόλουθα οκτώ (8) συναισθήματα (θυμός, φόβος, ανυπομονησία, εμπιστοσύνη, έκπληξη, θλίψη, ευτυχία και αηδία). Βρήκαν τρεις κύριες θεματικές στα tweets: 1. Γνώμες σχετικές με το Covid-19 και τα εμβόλια 2. Εκφράσεις υπέρ μέτρων που βοηθούν στον περιορισμό της διασποράς του ιού 3. Αναλήθειες και παράπονα για τον έλεγχο που έχει επιφέρει ο Covid-19. Σχεδόν δύο τρίτα των tweets ήταν θετικά ως προς τον εμβολιασμό αλλά το επίπεδο της θετικότητας δε φάνηκε αρκετό για να επιτύχει την εμβολιαστική κάλυψη και ανοσοποίηση που χρείαζεται η χώρα. Η σύσταση των ερευνητών προς την κυβέρνηση της Αυσταλίας ήταν να μελετήσει τη δημόσια γνώμη και το συναίσθημα ως προς τα εμβόλια και να χρησιμοποίησει τη γνώση αυτή σε μια αποτελεσματική προωθητική καμπάνια εμβολιασμού. [43]

Μια διαφορετική ομάδα ερευνητών ανέλυσε 75665 tweets με θέμα τα εμβόλια κατά του Covid-19 χρησιμοποιώντας το μοντέλο Latent Dirichlet Allocation (LDA) και τη μέθοδο VADER και κατέληξαν οτι ο αριθμος των θετικών tweets ήταν διπλάσιος απ τα αρνητικά. Επίσης παρατήρησαν οτι κάποια εμβόλια είχαν περισσοτερα αρνητικά σχόλια απο θετικά σε συγκεκριμένες χώρες πχ. Το Sputnik V στις Ηνωμενες Πολιτειες και το Sinovac στην Αγγλια και τον Καναδα. Στα θετικά tweets οι χρήστες ήταν ευγνώμονες για τη δυνατότητα εμβολιασμού ενώ στα αρνητικά οι χρήστες παραπονιούνταν για τις παρενέργειες του εμβολίου όπως πυρετός, πόνος στο χέρι κλπ. [44]

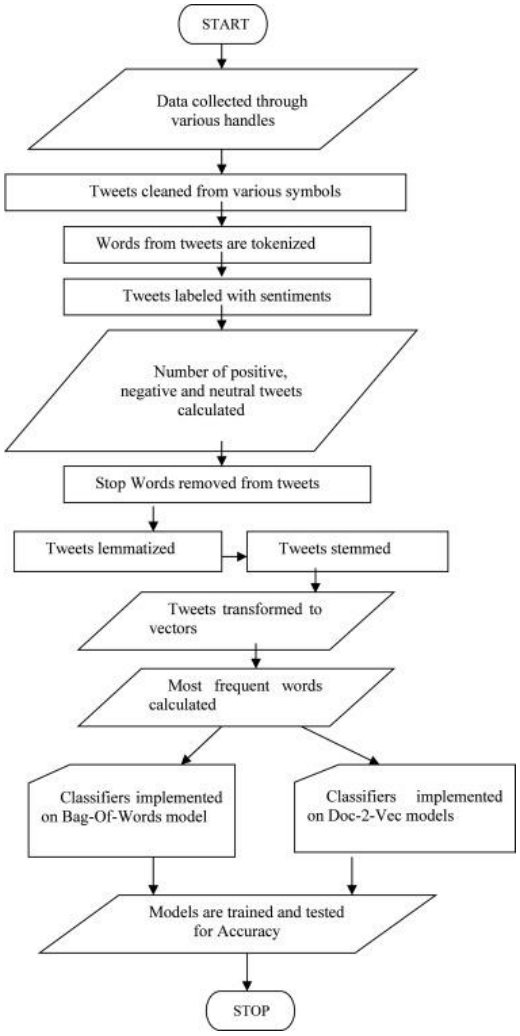
### 4.2.2 Covid-19 και ψυχολογία του πληθυσμού

Απεδείχθη οτι μετά τα αυστηρά μέτρα κατά της πανδημίας τα περιστατικά κατάθλιψης αυξήθηκαν σε συγκεκριμένη περιοχή της Αυστραλίας. Σκοπός της έρευνας ήταν να επιστήσει την προσοχή στις κυβερνήσεις να λαμβάνουν υπόψη την κατάθλιψη και τα κοινωνικά δίκτυα όταν σχεδιάζουν μέτρα κατά των επιδημιών. Η έρευνα ανακάλυψε διάφορα μοτίβα χρήσης του Twitter από χρήστες με κατάθλιψη όπως η δραστηριότητα αναρτήσεων μεταξύ 11μ και 6πμ - βραδινές ώρες ή η κατεξοχήν χρήση του πρώτου ενικού (‘εγώ’) [45]. Η έρευνα βασίστηκε στην τεχνική TF-IDF (term frequency – inverse document frequency) μια συνηθισμένη στατιστική μέθοδο στην επεξεργασία φυσικής γλώσσας η οποία μετράει τη

σημαντικότητα μιας λέξης μέσα στο κείμενο κρατώντας τη σύνδεση με άλλα κείμενα (‘corpus’). Η TF-IDF πολλαπλασιάζει δύο νούμερα: τον αριθμό εμφανίσεων της λέξης αυτής στο συγκεκριμένο κείμενο (term frequency) και τον λογάριθμο που προκύπτει αν διαιρεθεί ο συνολικός αριθμός κειμένων με τα κείμενα στα οποια εμφανίζεται ο ζητούμενος όρος (inverse document frequency) [46].

Στην ίδια θεματική, ομάδα ερευνητών ανέλυσαν tweets που αναφέρονταν στο Covid-19 και την ψυχική υγεία στην Αμερική. Χρησιμοποιώντας τις τεχνικές LDA και βαθιάς μαθησης βρήκαν θετική συσχέτιση μεταξύ ερωτήσεων για την ψυχική υγεία και την πανδημία. Τα κύρια θέματα που ανέδειξε η τεχνική LDA ηταν ‘stay-at-home’, ‘death poll’, ‘politics and policy’. Μεταξύ των χρηστών που πόσταραν σχετικά tweets το πιο δημοφιλές γκρουπ ήταν οι λευκοί ανδρες 30-49 ετών [47].

Μια άλλη μελέτη ανέλυσε δύο ειδών tweets που αναρτήθηκαν στο πρώτο κύμα της πανδημίας (Δεκέμβριο του 2019 με Μάιο του 2020) και έδειξε οτι τα tweets που αναρτούνται για πρώτη φορά απο τους χρήστες έχουν κυρίως θετικό ή ουδέτερο συναισθηματικό πρόσημο όμως τα re-tweets δηλαδή τα tweets που αναπαράγονται απο διαφορετικούς χρήστες έχουν κυρίως αρνητικό συναίσθημα, φαίνεται δηλαδη πιο εύκολο να αναπαράξεις αρνητικά tweets [48] .

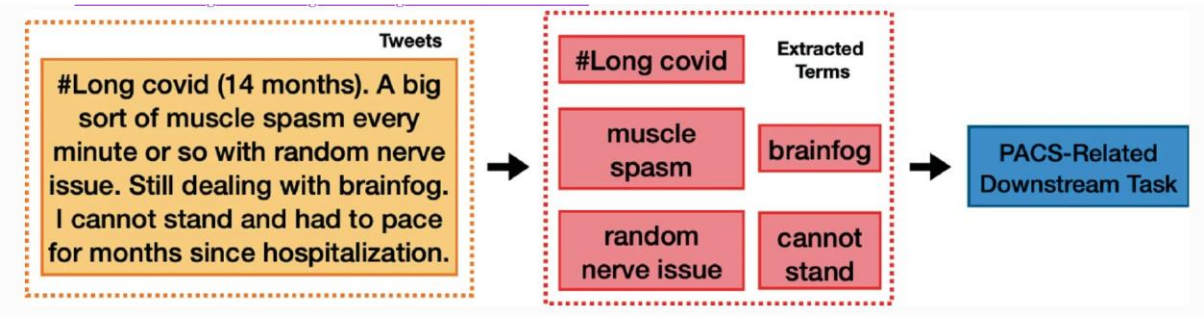



*Εικόνα 16: Μέθοδος εξόρυξης γνώσης απο tweets* [48]

### 4.2.3 Long Covid

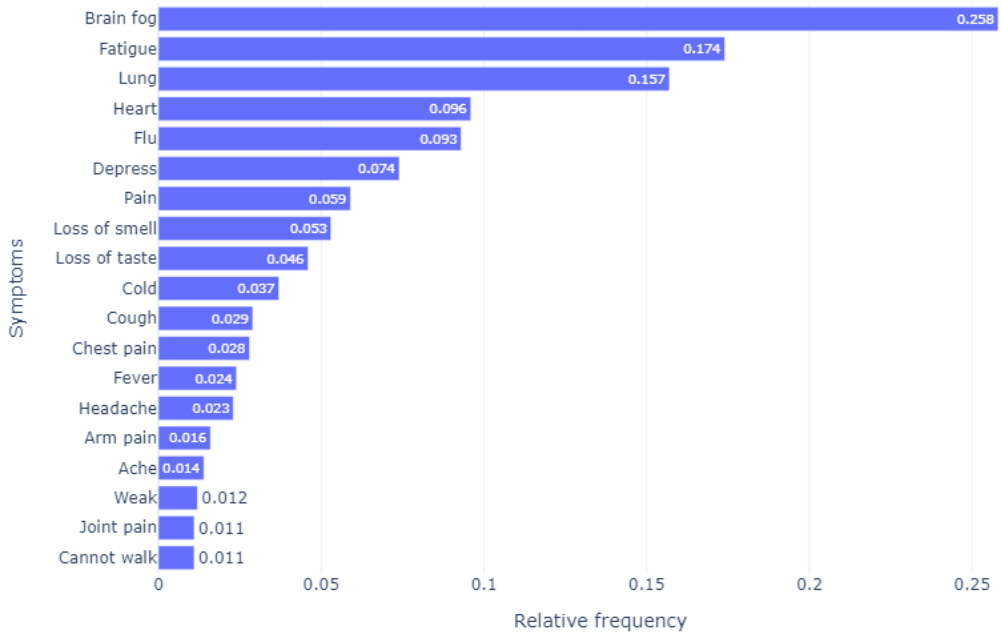
Το Long Covid είναι ενα σύνολο απο διαφορετικά συμπτώματα που επιμένουν μετά απο ανάρρωση απο Covid-19. Τα συμπτώματα επιμένουν ακόμη και εβδομάδες ή μήνες. Αυτή τη στιγμή δεν ειναι γνωστό ποιοί παράγοντες αυξάνουν την πιθανότητα για Long Covid όπως επίσης δεν είναι γνωστό το αποτέλεσμα αυτού του συνδρόμου στη δημόσια υγεία. Η ανεπάρκεια της έρευνας σε αυτό το σύνδρομο κάνουν απαραίτητη την εξερεύνηση των συναισθημάτων του πληθυσμού γύρω απο αυτό το θέμα.

Τα κοινωνικά δίκτυα αποτελούν μια καινοτόμα πηγή για φορείς δημόσιας υγείας να εντοπίζουν σε πραγματικό χρόνο καινούρια συμπτώματα που μπορεί να αποτελούν συμπτώματα του Long Covid. [49]




*Εικόνα 17: Διαδικασία εξόρυξης γνώσης συμπτωμάτων απο tweets που αναφέρονται στο Long Covid* [49]

Ανάλυση σε tweets για το Long Covid που αναρτηθηκαν μεταξυ του Μαιου του 2020 και του Δεκεμβριου του 2021 ανίχνευσε τα συμπτώματα του Long Covid αλλά επιπλέον προέβλεψε και τα πιο πιθανά σχετιζόμενα συμπτώματα. Συγκεκριμένα η θολούρα, η κούραση και προβλήματα με την αναπνοή/πνεύμονες ήταν τα τρία πιο συνηθισμένα συμπτώματα που εντόπισε η ανάλυση. Ακολουθούμενα απο καρδιακά προβλήματα, συμπτώματα γρίπης και κατάθλιψη. Τα αποτελέσματα έδειξαν με 77% πεποίθηση οτι οι ασθενείς με προβλήματα αναπνοής και απώλεια γεύσης είναι πιθανό να έχουν και απώλεια όσφρησης. [50]




*Εικόνα 18: Σχετική συχνότητα συμπτωμάτων σε ασθενείς με Long Covid* [50]

Ερευνητές ανέλυσαν ~62000 tweets με αναφορά στο #longCovid (και παρεμφερή hashtags) κατά τη χρονική περίοδο 25 Μαρτιου – 1η Απριλιου 2022. Μελέτησαν τις πιο κοινές λέξεις και εφάρμοσαν ανάλυση συναισθήματος και μοντελοποίηση θεμάτων (topic modelling). Χρησιμοποίησαν Επεξεργασία φυσικής γλώσσας και LDA για να αναδείξουν τα πιο δημοφιλή θέματα και να δημιουργήσουν ομάδες. Τα συχνότερα συναισθήματα ήταν αυτο της

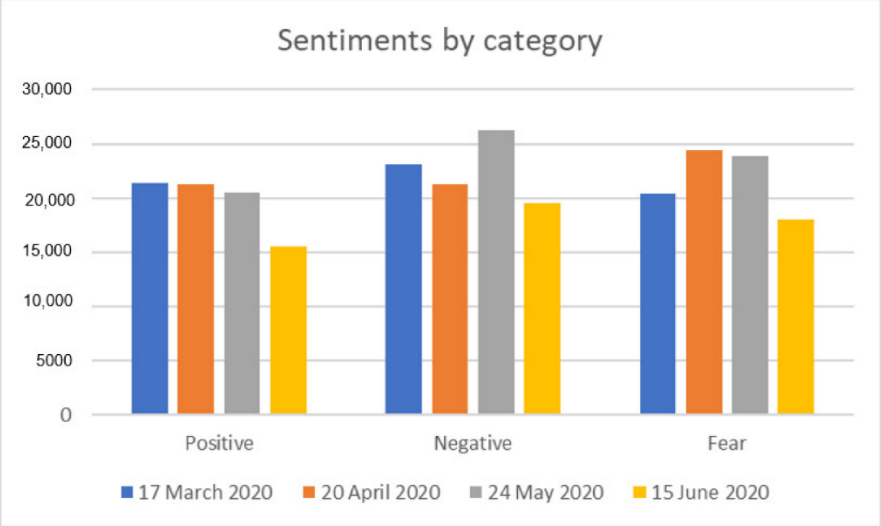
εμπιστοσύνης (11.68%), του φόβου (11.26%) και της λύπης (9.76%). Τα συναισθήματα συνδέονταν με ανησυχια για: επιμόλυνση, πανδημία, χρόνια αναπηρία και κρατικές οδηγίες. Η ανάλυση συναισθήματος έδειξε οτι οι άνθρωποι έχουν μοιρασμένη θετική (19.9%) και αρνητική (18.39%) διάθεση απέναντι στο Long Covid [51].

Στις περισσότερες έρευνες έως τώρα τα κύρια κοινωνικά δίκτυα που προτιμούνται είναι το Twitter και το εκ Κινας ορμωμενο Weibo λόγω των εξης πλεονεκτημάτων:

- Πολλοί χρήστες και πολλά δεδομένα
- Το API του που προσφέρει έναν εύκολο τρόπο για πρόσβαση στα tweets
- Την επιλογή να προβάλεις tweets σε συγκεριμένη γλώσσα ή γεωγραφική περιοχή (κοντά σου)

4.3 Επεξεργασία φυσικής γλώσσας στην Ελλάδα την περίοδο του covid-19

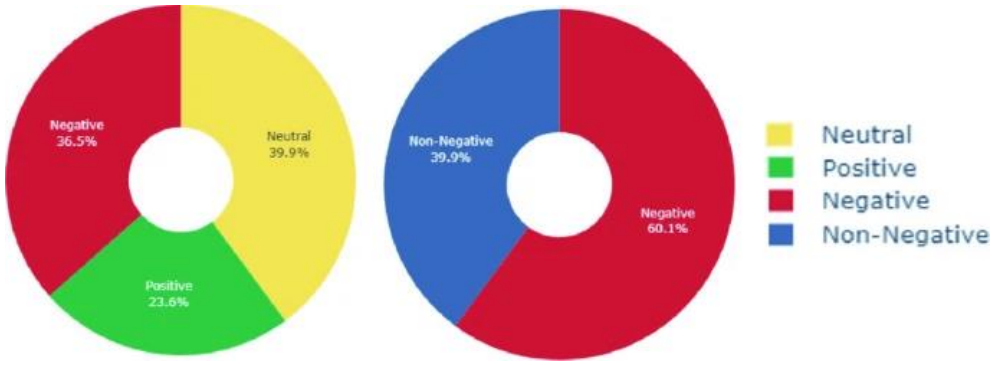
Εξίσου πολλή και η ελληνική δραστηριότητα στο Twitter κατά τη διάρκεια της πανδημίας. Το πρώτο επιβεβαιωμένο κρούσμα στην Ελλάδα εμφανίστηκε στις 6 Φεβρουαριου του 2020. Στο πρώτο κύμα της πανδημίας, το αρχικό κύριο συναίσθημα ήταν αυτό της έκπληξης ενώ το συναισθήμα του άγχους αυξήθηκε τον Απρίλιο και Μαιο του 2020 και μειώθηκε τον Ιούνιο του 2020 [52,53]. Τα συναισθήματα που αναδείχθηκαν ήταν παρεμφερή με έρευνες σε άλλες χώρες ειδικά ο φόβος του θανάτου. Χρησιμοποιήθηκαν τα έξι (6) βασικά συναισθήματα όπως αυτά έχουν οριστεί απ τον Paul Ekman.




*Εικόνα 19: Συναισθήματα σε ελληνικά tweets που αφορούν το Covid-19* [52]

Οι Έλληνες είχαν έντονες αντιδράσεις online ως προς τα εμβόλια. Το 60.1% των ελληνικών tweets μεταξύ Μαίου και Νοεμβριου 2021 που αναφέρονταν στον εμβολιασμό ήταν

αρνητικά. Ενώ στην ίδια έρευνα μόνο το 36.5% των tweets στα αγγλικά ήταν αρνητικά. Φαίνεται δηλαδή οτι οι Έλληνες ήταν πιο αρνητικά προσκείμενοι στον εμβολιασμό κατά του Covid-19 [54].




*Εικόνα 20: Συναισθήματα σε ελληνικά tweets που αφορούν τα εμβόλια κατά του Covid-19* [54]

Όσον αφορά στο Long Covid, πραγματοποιήθηκε ανάλυση σε ερωτηματολόγια απο 208 ασθενείς με Long Covid στην Ελλάδα. Στην πειοψηφία τους οι ασθενείς (68.8%) δε χρειάστηκαν νοσηλεία και ανάφεραν (66.8%) οτι τα συμπτώματα τους επιμένουν για περισσότερο απο έξι μήνες. [55]

# Κεφάλαιο 5 Μελέτη Περίπτωσης

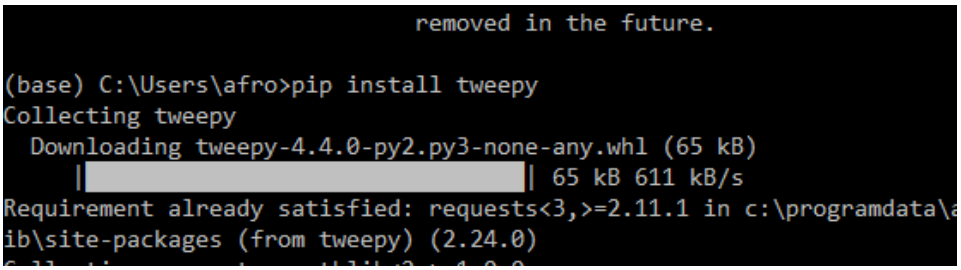
## 5.1 Συλλογή δεδομένων απ'το Twitter

Στόχος της διπλωματικής εργασίας είναι η εξαγωγή συμπερασμάτων για τα συναισθήματα του πληθυσμού στον ελλαδικό χώρο ως προς το Long Covid που φαίνεται οτι δεν έχει καλυφθεί ενδελεχώς ερευνητικά ως περίπτωση τόσο στην Ελλάδα όσο και στο εξωτερικό.

Για τη συλλογή δεδομένων προτιμήθηκε το Twitter [56] καθώς είναι ένα πολύ δημοφιλές μέσο κοινωνικής δικτύωσης στην Ελλάδα.

Αποκτώντας την ερευνητική άδεια που παρέχει το Twitter είναι διαθέσιμη η εξαγωγή 10 εκατομυρίων tweets το μήνα. Η ιστορικότητα αυτών ξεκινάει απο το 2006 οποτε καλύπτει σίγουρα όλο το διάστημα απ το 2020 εως σήμερα.

Χρησιμοποιήθηκε η βιβλιοθηκη tweepy προκειμένου να αποκτήθεί πρόσβαση στο Twitter API μετά απο εισαγωγή των προσωπικών κωδικών.




*Εικόνα 21: Εγκατάσταση Tweepy*

Εξήχθησαν tweets με hashtag #longCovid ή που περιείχαν τις λέξεις 'Long Covid' στην ελληνικη γλωσσα κατα τη διάρκεια του ετους 2022. Δημιουργήθηκε ένα αρχικό σύνολο 1000 tweets το οποίο και αποθηκεύτηκε σε ενα αρχείο excel.

Τα αποτελέσματα του Twitter API εμφανίζονται με πεπερασμένο αριθμό tweets ανά σελίδα οπότε χρησιμοποιήθηκε η συνάρτηση Paginator για να ληφθούν όλα τα διαθέσιμα tweets θέτοντας ως ανώτατο όριο το 1000.

```
import tweepy
import pandas as pd

client = tweepy.Client(bearer_token='AAAAAAAAAAAAAAAAAAAAAP4QXgEAAAAA031phEFaTS2CG1LVAG0G0e0KXQo%3DctXNcvI

# Replace with your own search query
query = '#longcovid -is:retweet lang:el'

# Replace with time period of your choice
start_time = '2022-01-01T00:00:00Z'

# Replace with time period of your choice
end_time = '2022-12-31T00:00:00Z'

tweets = tweepy.Paginator(client.search_all_tweets, query=query,
                                tweet_fields=['context_annotations', 'created_at'],start_time=start_time
                                end_time=end_time, max_results=100).flatten(limit=1000)
# We create a pandas dataframe as follows:
data = pd.DataFrame(data=[tweet.text for tweet in tweets], columns=['Tweets'])
result = data.head(10)
print(result)
data.to_excel('tweets_long_covid.xls', index = False)
```

*Εικόνα 22: Εξαγωγή δεδομένων απ το Twitter*

| | |
|---|---|
| 13 | #LongCovid #ανοσμία #όσφρηση https://t.co/TBAfQqMRuX |
| 14 | Εγώ ξέρω ήδη πως θα ξεκινήσει το 2023, με φυσιοθεραπεία για #LongCovid ραντεβού κλείστηκε, και το 2022 θα κλείσει με εξετάσεις αίματος. Καλά πήγε και αυτό. |
| | Μία γνωστή μου διαπίστωσε προβλήματα ακοής, πήγε στον ΩΡΛ & της είπε ότι έχει χάσει το 50% της ακοής της μετά από νόσηση με Covid. Θα προσμετρηθεί στις περιπτώσεις #LongCovid; Ποιος ξέρει; Υπάρχει άλλωστε πραγματική επιδημιολογική επιτήρηση στη χώρα; |
| 15 | #CovidIsNotOver |
| | #LongCovid Πολύ γουστάρω τη φάση. Πλέον μετά απο 3 χρόνια ιατρικού και όχι μόνο φασισμού, στο τουιτερ έχουν μείνει ενεργοί μόνο 3-4 ντόπιοι σκιτζήδες που αναπαράγουν την |
| 16 | δολοφονική προπαγάνδα τους και συνήθως αναλώνονται σε μεταξύ τους κουβεντούλα βλογωντας τα γένια τους. (1) |
| 17 | Στο μικροσκόπιο ελληνικών μελετών η long CoViD https://t.co/dyu4KRvfFV #LongCovid #κορονοιος #ΑΠΘ @Auth_University #ΑΧΕΠΑ |
| 18 | Γκίκας Μαγιορκίνης: «Ο κορωνοϊός θα γίνει ενδημικός, ελπίδα τα εμβόλια που πιάνουν όλες τις παραλλαγές» https://t.co/oEtp3gQlwn #longcovid #gkikasmagiorkinis #gripi #emvolia |
| 19 | Να ρωτήσω, η Coca Cola Light είχε πάντα γεύση αφρού ξυρίσματος, ή έχω πάθει εγώ παροσμία από το #LongCovid; |
| 20 | Όταν οι συνεπείς επιστήμονες αναγκάζονται να συζητήσουν θέματα προπαγάνδας, "κάτι δεν πάει καθόλου καλά στο βασίλειο της Δανιμαρκίας"! Όταν προσπαθείς να τα βάλεις με την (βολική για το λαό) προπαγάνδα της εξουσίας, είναι πολύ άνισος ο αγώνας. #Covid_19 #CovidIsNotOver #LongCovid https://t.co/zEtyKz6bXZ |
| | 130 νεκροί την βδομαδα, 34.614 σύνολο. |
| | Τι λέτε, θα ξανανικήσουμε τον #κορονοιο; Κ αν ναι, θα είναι η 7η; Η 8η; Ποια τελοσπαντων φορά που θα κερδίσουμε τον #κορωνοιο; Με #LongCovid, ειδικά στα παιδιά που κολλανε συνεχως, τι κανουμε; |
| 21 | #COVID19 #CovidIsNotOver #covid19_gr https://t.co/O67bq0jld2 |
| | 1/ |
| | Covid update |
| | τα άσχημα νέα:νεκροψίες σε μη εμβολιασμένους επιβεβαιώνουν ότι ο ιός εξαπλώθηκε σε όλο το σώμα και παρέμεινε στον ιστούς για μήνες, ακόμη κ στον εγκέφαλο, ενισχύοντας την έρευνα για #LongCovid |
| | https://t.co/pv6vsnyqg3 |

*Εικόνα 23: Δείγμα δεδομένων μετά τη συλλογή*

## 5.2 Προ επεξεργασία - Καθαρισμός Δεδομένων

Το πρώτο στάδιο στην εξόρυξη γνώσης είναι η προ-επεξεργασία των δεδομένων με σκοπό να αφαιρεθούν προφανή λάθη και άλλες ασυνέπειες που ‘βρωμίζουν’ το σύνολο των δεδομένων και θα μπερδέψουν τους αλγορίθμους αργότερα.

Πολλοί ερευνητές χρησιμοποιούν σε αυτό το βήμα μια μέθοδο που δημιουργεί σύννεφο λέξεων (wordcloud) [44]. Το wordcloud είναι μια συνήθης εύκολη πρακτική με την οποία ξεκινούν οι ερευνητές καθότι αποκαλύπτει τις πιο συχνές λέξεις των tweets και παρέχει μια επιβεβαίωση στις επόμενες τεχνικές. Με αυτή την κίνηση πετυχαίνουν δύο στόχους:

- Διαπιστώνουν κοινά λάθη στα δεδομένα

- Συγκρίνουν και επιβεβαιώνουν τα αποτελέσματα και τη γνώση που παράχθηκε σε επόμενα στάδια της εργασίας

Στα πλαισια αυτής της εργασίας χρησιμοποιήθηκε η μέθοδος Wordcloud της Python. Το σύννεφο λέξεων (wordcloud) χωρίς καμία προεπεξεργασία είναι αυτό που φαίνεται στην Εικόνα 24:


*Εικόνα 24: Σύννεφο λέξεων χωρίς προεπεξεργασία*

Η πρωτη απόπειρα για να βρεθούν οι πιο συχνά εμφανιζόμενες λέξεις στα tweets (wordcloud) αποκαλύπτει links (εμφάνιση https), λέξεις των οποίων η ύπαρξη μέσα στο tweet ήταν απαραίτητη προκειμένου να είναι κομμάτι του συνόλου που συλλέχθη και πολλές λέξεις που δεν έχουν ιδιαίτερο νόημα ή συναισθηματικό πρόσημο, γνωστές και ως stopwords. Αυτές οι λέξεις είναι συνήθως άρθρα, κάποια επιρρήματα, το ρήμα ‘είμαι’ κλπ. Τα stopwords αφαιρούνται συνήθως απο το σύνολο δεδομένων στο βήμα του καθαρισμού καθώς αυτές οι λέξεις θα μπέρδευαν τους αλγορίθμους λόγω της υψηλής συχνότητας εμφάνισης τους.

Σε αυτό το στάδιο εφαρμόστηκαν στο σύνολο των tweets οι ακόλουθες αφαιρέσεις [43,44]:

- Links και τα Usernames
- οι λέξεις ‘Long Covid’, ‘long’, ‘Covid’ και #longCovid που χρησιμοποιήθηκαν στην αναζήτηση και άρα λογικό είναι να εμφανίζονται σε κάθε tweet

- οι λέξεις κορονοιος και πανδημία και οι εναλλακτικοί τρόποι γραφής τους που ομοίως εμφανίζονταν πολύ συχνά χωρίς να προσφέρουν κάτι καινούριο στο νόημα του tweet
- λέξεις με μέγεθος μικρότερο των 4 χαρακτήρων πχ. Εδώ, amp. Ήδη, μας, νεα
- οι ελληνικές λέξεις που θεωρούνται stopwords πχ εγώ, είναι, που χρησιμοποιώντας και συνδυάζοντας το αρχείο stopwords-iso για την ελληνικη γλώσσα [57] και τη βιβλιοθήκη nltk που περιέχει μια λίστα απο ελληνικά stopwords. Η βιβλιοθήκη nltk παρέχει stopwords λιστες σε πολλές γλώσσες [58]
- τα σημεία στίξης
- οι διπλότυπες εγγραφές αφού είχαν υποστεί την ανωτέρω επεξεργασία

```python
# Importing modules
import pandas as pd
from wordcloud import WordCloud
import re
import numpy
import unicodedata

def remove_usernames_links(tweets):
    tweets = re.sub('@[^\s]+','',tweets)
    tweets = re.sub('http[^\s]+','',tweets)
    tweets = tweets.lower()
    tweets = re.sub(r"long covid", "", tweets)
    tweets = re.sub(r"long", "", tweets)
    tweets = re.sub(r"covid", "", tweets)
    tweets = re.sub(r"\b\w{1,3}\b", "", tweets)
    tweets = re.sub(r'[^\w\s]', '', tweets)

    return tweets

def strip_accents_and_lower(s):
     return ''.join(c for c in unicodedata.normalize('NFD', s)
                    if unicodedata.category(c) != 'Mn').lower()

# Read data into dataframe
tweets = pd.read_excel('C:/Users/afro/Desktop/health_informatics/thesis/tweets_long_covid_hashtag_text.xls')

# Print head + number of tweets
result = tweets.head(10)
print(result)
print(len(tweets.index))

#clean usernames, links, hashtag
#lowercase

tweets['Tweets'] = tweets['Tweets'].apply(remove_usernames_links)
tweets['Tweets'] = tweets['Tweets'].apply(strip_accents_and_lower)

#remove duplicates
tweets = tweets.drop_duplicates()
```

*Εικόνα 25: Προεπεξεργασία δεδομένων*

Μια ιδιομορφία της ελληνικής γλώσσας είναι οι τόνοι. Οι λέξεις με τόνους και οι ίδιες λέξεις χωρίς τους τόνους δεν αναγνωρίζονταν απ το πρόγραμμα ως ταυτόσημες. Γι αυτό το λόγο από τις λέξεις αφαιρέθηκαν οι τόνοι και τα γράμματα μετατράπηκαν σε πεζά. Στη συνέχεια συγκρίθηκαν με τα αρχεία που περιείχαν τα stopwords.

```
#remove stopwords

#1 using iso source
stopwordsel = open('stopwords-el.txt', 'r', encoding = "utf-8")

#read text file into list
stopwords_list = stopwordsel.read().split('\n')

#2 using nltk corpus
from nltk.corpus import stopwords
#from nltk.tokenize import word_tokenize


stopwords = set(stopwords_list + stopwords.words('greek'))
stopwords_no_accents_lower = set([strip_accents_and_lower(s) for s in stopwords])


#remove stopwords function
tweets['Tweets'] = tweets['Tweets'].apply(
    lambda x: ' '.join([word for word in x.split() if strip_accents_and_lower(word) not in stopwords_no_accents_lower])
)

result = tweets['Tweets'].head(10)
print(result)
print(len(tweets.index))
tweets['Tweets'].to_excel('tweets_after_rmv_stopwords.xls', index = False)
```

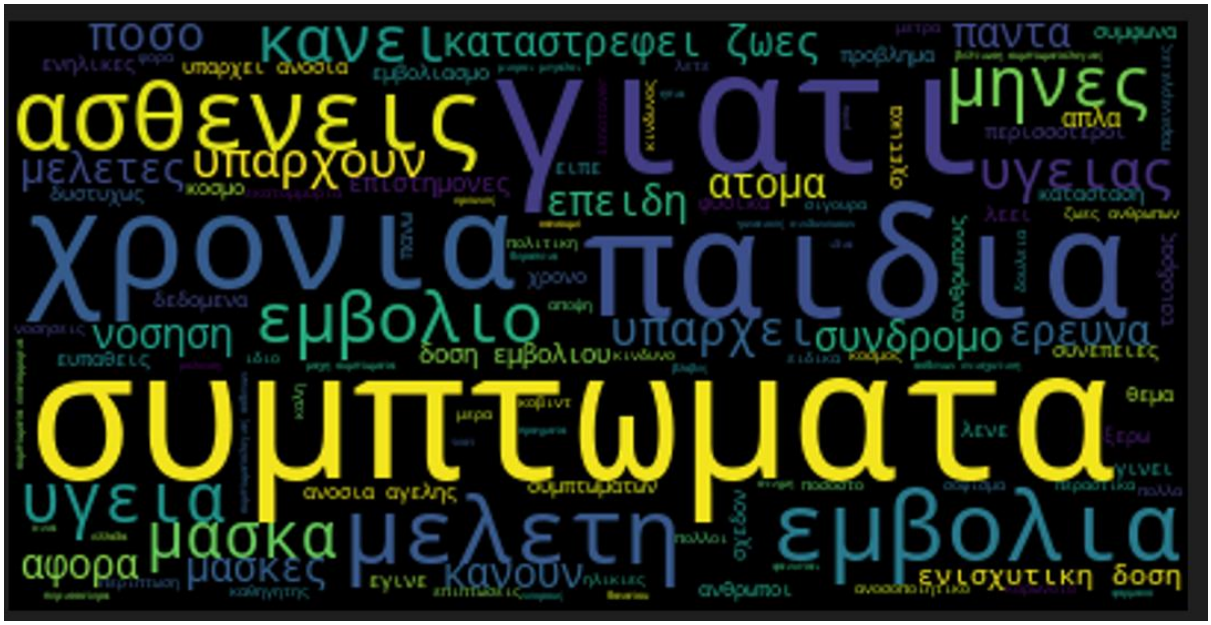
*Εικόνα 26: Προεπεξεργασία σχετική με την ελληνική γλώσσα*

Εξετάζονται οι δέκα πρώτες εγγραφές μετά τη διαδοχική προεπεξεργασία όπως αυτές φαίνονται στην Εικόνα 27.

```
                                              Tweets
0  2023.\n\nΗ χρονιά εξαρσης σε αφορητο βαθμο της...
1  @Zebra_Delta Δεν ισχύει αυτό τι λες; Βιολογικά...
2  @AlbertHitter Ε τότε για τα εισπνεομενα που δε...
3  @poirazis @TsioutisC Την ημέρα που εκατάλαβεν ...
4  Long covid σε ανεμβολίατσους που επαναμολύνοντ...
5  @tklikas @TinaTs2 Τα πάντα ξέρουμε...δεν έχουμ...
6  @tklikas @Sandbird_ Έτρεξαν και κάποιοι άλλοι ...
7  @pepe_g12 ιατρείο long COVID Ευαγγελισμού, ότα...
8  @pepe_g12 Οι τελευταίες σουηδικες μελέτες δίνο...
9  Στη σύσταση ειδικής επιστημονικής επιτροπής Lo...
1000
                                              Tweets
0  2023\n\n χρονια εξαρσης  αφορητο βαθμο  ξαφνικ...
1    ισχυει αυτο   βιολογικα παραλογο  μεταλλαξει...
2    τοτε   εισπνεομενα   βρισκω φταινε  πνευμονο...
3           ημερα  εκαταλαβεν    placebo  παθαινεις
4                   ανεμβολιατσους  επαναμολυνονται
5     παντα ξερουμε εχουμε χασει ουτε  ενημερωση ...
6    ετρεξαν  καποιοι αλλοι  στην συνεχεια προετρ...
7   ιατρειο  ευαγγελισμου οταν περασα  φορα  εκαν...
8    τελευταιες σουηδικες μελετες δινουν ποσοστο ...
9   συσταση ειδικης επιστημονικης επιτροπης  προχ...
951
0    2023 χρονια εξαρσης αφορητο βαθμο ξαφνικιτιδας...
1    ισχυει βιολογικα παραλογο μεταλλαξεις προκυπτο...
2    εισπνεομενα βρισκω φταινε πνευμονολογοι γραφου...
3                   ημερα εκαταλαβεν placebo παθαινεις
4                       ανεμβολιατσους επαναμολυνονται
5    παντα ξερουμε χασει ενημερωση tessy νοσηλειες ...
6    ετρεξαν συνεχεια προετρεψαν κανουν booster δηθ...
7    ιατρειο ευαγγελισμου περασα φορα εκανα συνελθω...
8    τελευταιες σουηδικες μελετες δινουν ποσοστο νε...
9    συσταση ειδικης επιστημονικης επιτροπης προχωρ...
Name: Tweets, dtype: object
951
```

*Εικόνα 27: Δείγμα δεδομένων μετά την προεπεξεργασία*

Η δεύτερη απόπειρα για ένα σύννεφο λέξεων μετα την προεπεξεργασία των δεδομένων είναι πολύ πιο αξιόπιστη όπως φαίνεται στην Εικόνα 28 και αναλύεται στο Κεφάλαιο 6.




*Εικόνα 28: Σύννεφο λέξεων μετα την προεπεξεργασία*

5.3 Μετασχηματισμός Δεδομένων

Το επόμενο βήμα είναι αυτό του μετασχηματισμού και της κανονικοποίησης (normalisation) των δεδομένων.

Προκειμένου να χρησιμοποιηθεί η τεχνική της μοντελοποίησης θεμάτων (topic modelling), η λημματοποίηση (lemmatization) είναι απαραίτητο βήμα της προεργασίας.

Χρησιμοποιήθηκε το pipeline Spacy το οποιο υποστηρίζει και την ελληνική γλώσσα και παρέχει λειτουργικότητα για λημματοποίηση. [59]

Παρακάτω στην Εικόνα 29 παρατίθεται παράδειγμα ενός tweet μετα τη λημματοποίηση. Παρατηρείται οτι η αναγωγή στο λήμμα της λέξης επιδέχεται βελτίωσης.

```
import pandas as pd
import spacy

nlp = spacy.load('el_core_news_sm')
for token in nlp('ισχύει λες; βιολογικά παράλογο. μεταλλάξεις προκύπτουν ιος εξαπλώνεται τρελός περιορίζεται. αποστείρωση μιλ
    print(f"{token.text:<6}--> {token.lemma_}", end='\n')

# Read data into dataframe
#tweets = pd.read_excel('C:/Users/afro/Desktop/health_informatics/thesis/tweets_after_rmv_stopwords.xls')
```

```
ισχύει--> ισχύω
λες   --> λες
;     --> ;
βιολογικά--> βιολογικά
παράλογο--> παράλογος
.     --> .
μεταλλάξεις--> μεταλλάξεις
προκύπτουν--> προκύπτω
ιος   --> ιος
εξαπλώνεται--> εξαπλώνεται
τρελός--> τρελόςς
περιορίζεται--> περιορίζω
.     --> .
αποστείρωση--> αποστείρωση
μιλάς --> μιλάς
εναλλακτική--> εναλλακτικός
θάνατος--> θάνατος
;     --> ;
ντροπή--> ντροπή
!     --> !
```

*Εικόνα 29: Παράδειγμα tweet μετά απο λημματοποίηση*

```
def lemmatize(text, nlp):
   doc = nlp(text)
   lemmatized_text = []
   for token in doc:
       lemmatized_text.append(token.lemma_)
   return " ".join(lemmatized_text)
#testing the function on a single sample for explaination
#print(lemmatize('Reading NLP blog is fun.' ,nlp ))
#Performing lemmatization on every row
tweets_df.Tweets=tweets_df.Tweets.apply(lambda x:lemmatize(x,nlp))
result = tweets_df.head(10)
print(result)
```

*Εικόνα 30: Υλοποίηση λημματοποίησης*

Η σύγκριση των δέκα πρωτων tweets πριν και μετα τη λημματοποίηση φαίνεται στην Εικόνα 31.

| | Tweets |
|---|---|
| 0 | 2023 χρονια εξαρσης αφορητο βαθμο ξαφνικιτιδας... |
| 1 | ισχυει βιολογικα παραλογο μεταλλαξεις προκυπτο... |
| 2 | εισπνεομενα βρισκω φταινε πνευμονολογοι γραφου... |
| 3 | ημερα εκαταλαβεν placebo παθαινεις |
| 4 | ανεμβολιατσους επαναμολυνονται |
| 5 | παντα ξερουμε χασει ενημερωση tessy νοσηλειες ... |
| 6 | ετρεξαν συνεχεια προετρεψαν κανουν booster δηθ... |
| 7 | ιατρειο ευαγγελισμου περασα φορα εκανα συνελθω... |
| 8 | τελευταιες σουηδικες μελετες δινουν ποσοστο νε... |
| 9 | συσταση ειδικης επιστημονικης επιτροπης προχωρ... |

| | Tweets |
|---|---|
| 0 | 2023 χρονος εξαρση αφορητος βαθμο ξαφνικιτιδα ... |
| 1 | ισχυει βιολογικα παραλογος μεταλλαξεις προκυπτ... |
| 2 | εισπνεομενος βρισκω φταινε πνευμονολογος γραφω... |
| 3 | ημερα εκαταλαβον placebo παθαινεις |
| 4 | ανεμβολιατσος επαναμολυνω |
| 5 | παντα ξερουμε χασει ενημερωση tessy νοσηλειος ... |
| 6 | ετρεξαν συνεχεια προετρεψαν κανω booster δηθεν... |
| 7 | ιατρειος ευαγγελισμος περασα φορο εκανα συνελθ... |
| 8 | τελευταιος σουηδικες μελετα δινω ποσοστος νεος... |
| 9 | συσταση ειδικη επιστημονικη επιτροπης προχωρω ... |

*Εικόνα 31: Δείγμα tweets μετα τη λημματοποίηση*

Τέλος παράγεται καινούριο σύννεφο λέξεων για να παρατηρηθεί πως αλλάζουν οι πιο δημοφιλείς λέξεις μετά τη λημματοποίηση.


*Εικόνα 32: Σύννεφο λέξεων μετά την λημματοποίηση*

## 5.4 Εξόρυξη γνώσης

Στο στάδιο της εξόρυξης γνώσης, το πρώτο που αναζητήσαμε είναι η ανάδειξη των κυριότερων θεμάτων συζήτησης μεσα στα tweets το οποίο και συγκρίναμε με τις δημοφιλείς λέξεις που έβγαλε η μέθοδος του σύννεφου λέξεων στο προηγουμενο στάδιο.

### 5.4.1 Μοντελοποίηση θεμάτων

Εφαρμόσαμε μοντελοποίηση θεμάτων το οποίο εχει χρησιμοποιηθεί αρκετά σε συναφείς έρευνες του εξωτερικού και συγκεκριμένα το μοντέλο LDA (Latent Dirichlet Allocation) που εχει αναλυθεί και στο προηγούμενο κεφάλαιο [22,60,61].

Το Gensim ειναι μια δωρεάν, open-source βιβλιοθήκη της Python που υλοποιεί μοντελοποίηση θεμάτων (https://pypi.org/project/gensim/).

Το πρώτο βήμα είναι η δημιουργία λεξικού και συνόλου δεδομένων σε μορφή που να μπορεί να αναγνωριστεί και χρησιμοποιηθεί για μοντελοποίηση θεμάτων. [61]

Η εφαρμογή του αλγορίθμου απαιτεί τη δημιουργία ενός index (λεξικό) που αντιστοιχεί κάθε λέξη που εμφανίζεται στο σύνολο των tweets με ένα μοναδικό id. Το index αυτό στη genism ονομάζεται Dictionary. Από αυτό, αφαιρέθηκαν οι λέξεις που είτε εμφανίζονται πολυ λίγες φορές είτε εμφανίζονται πολλές.

Οι ορισμοί που δώσαμε είναι:

- Λίγες φορες: οι λέξεις εμφανίζονται σε λιγότερα απο 2 tweets
- Πολλές φορές: οι λέξεις εμφανίζονται σε περισσότερα απο 99% των tweets

Επίσης, απαιτείται η αναπαράσταση του συνόλου των tweets με τη μορφή BoW (Bag of Words). Το σύνολο δεδομένων στη μορφή αυτή ονομάζεται πλέον Corpus στη μέθοδο Gensim.

```
#converting to tokens
def generate_tokens(tweet):
    words=[]
    for word in tweet.split(" "):
    # using the if condition because we introduced extra spaces during text cleaning
        if word!="":
            words.append(word)
    return words
#storing the generated tokens in a new column named 'words'
tweets_df['tokens']=tweets_df.Tweets.apply(generate_tokens)

def create_dictionary(words):
    return corpora.Dictionary(words)
#passing the dataframe column having tokens as the argument
id2word=create_dictionary(tweets_df.tokens)
print(id2word)
```

*Εικόνα 33: Αναγωγή των λέξεων σε tokens*

Δεν υπάρχει ιδανική απάντηση για το πόσα θέματα να επιλεγούν. Στη βιβλιογραφία είναι εξίσου συνηθισμένο είτε ένα νούμερο να δίνεται απο τον ερευνητή είτε το νούμερο να

επιλέγεται μετά απο σειρά πειραμάτων που αυξάνουν τον αριθμό των θεμάτων μεχρι να βρεθεί ο ιδανικός αριθμός. Δεδομένου ότι επιθυμούμε την ερμηνεία των θεμάτων μετά την εξαγωγή τους επιλέξαμε ένα μικρό αριθμό απο θέματα, πέντε (5).

```
id2word = Dictionary(tweets_df['tokens'])
print(len(id2word))

id2word.filter_extremes(no_below=2, no_above=.99)
print(len(id2word))

corpus = [id2word.doc2bow(d) for d in tweets_df['tokens']]

base_model = LdaMulticore(corpus=corpus, num_topics=9,id2word=id2word, workers=12, passes=5)

words = [re.findall(r'"([^"]*)"',t[1]) for t in base_model.print_topics()]


# Create Topics
topics = [' '.join(t[0:5]) for t in words]


# Getting the topics
for id, t in enumerate(topics):
    print(f"------ Topic {id} ------")
    print(t, end="\n\n")
```

*Εικόνα 34: Μέρος της υλοποίησης LDA*

Τα αποτελέσματα του αλγορίθμου LDA φαίνονται στην Εικόνα 35.

```
------ Topic 0 ------
υπάρχω γιατί όσφρηση μελέτη χιλιάδα

------ Topic 1 ------
μήνας κάνω ασθενής τσιόδρα πολύς

------ Topic 2 ------
συμπτώματα παιδί συμπτώμα πολύς άνθρωπος

------ Topic 3 ------
χρόνος υγεία μάσκα μεγάλος γίνομαι

------ Topic 4 ------
εμβόλιο υπάρχω ανοσία αγέλη τσακρής
```

*Εικόνα 35: Αποτελέσματα μοντέλου LDA*

Για τα θέματα που αναδείχτηκαν, υπολογίζουμε δυο ποσοτικούς δείκτες επιτυχίας:

- Σύγχυση (perplexity)
- Συνοχή (coherence)

```
base_perplexity = base_model.log_perplexity(corpus)
print('\nPerplexity: ', base_perplexity)

# Compute Coherence Score
coherence_model = CoherenceModel(model=base_model, texts=tweets_df['tokens'],
                                 dictionary=id2word, coherence='c_v')
coherence_lda_model_base = coherence_model.get_coherence()
print('\nCoherence Score: ', coherence_lda_model_base)
```

*Εικόνα 36: Υπολογισμός σκορ συνεκτικότητας*

```
Perplexity:  -7.187703128961239

Coherence Score:  0.4009843681195341
```

*Εικόνα 37: Πρώτα αποτελέσματα σκορ συνεκτικότητας για πέντε θέματα*

Το σκορ συνοχής που υπολογίστηκε ειναι στην κατηγορία του χαμηλού (όχι όμως και άσχημου που θα ήταν αν το σκορ έπεφτε απο 0.3 και κάτω).

Προκειμένου να βρούμε τον καλύτερο δυνατό αριθμό θεμάτων, δημιουργήσαμε μια συνάρτηση που δοκιμάζει όλους τους αριθμούς θεμάτων απο 2 εως το 19 και εκτυπωνει το σκορ συνοχής τους.

```
def compute_coherence_values(dictionary, corpus, texts, limit, start=2, step=3):

    coherence_values_topic = []
    model_list_topic = []
    for num_topics in range(start, limit, step):
        model = LdaMulticore(corpus=corpus, num_topics=num_topics, id2word=id2word)
        model_list_topic.append(model)
        coherencemodel = CoherenceModel(model=model, texts=texts, dictionary=dictionary, coherence='c_v')
        coherence_values_topic.append(coherencemodel.get_coherence())
        print('\n number of topics and respective coherence: ', num_topics, coherence_values_topic)

    return model_list_topic, coherence_values_topic


model_list_topic, coherence_values_topic = compute_coherence_values(dictionary=id2word,
                                                    corpus=corpus,
                                                    texts=tweets_df['tokens'],
                                                    start=2, limit=20, step=1)
```

*Εικόνα 38: Εύρεση βέλτιστου αριθμού θεμάτων*

Παρατηρούμε οτι το coherence value αυξάνεται λίγο όχι όμως δραματικά. Η καλύτερη επίδοση είναι στα 19 θέματα όπου το coherence value ειναι 0.517.

| Number of Topics | Coherence Score |
|---|---|
| 2 | 0.425 |
| 3 | 0.427 |

| | |
|---|---|
| 4 | 0.434 |
| 5 | 0.458 |
| 6 | 0.454 |
| 7 | 0.458 |
| 8 | 0.458 |
| 9 | 0.476 |
| 10 | 0.478 |
| 11 | 0.488 |
| 12 | 0.488 |
| 13 | 0.488 |
| 14 | 0.510 |
| 15 | 0.498 |
| 16 | 0.492 |
| 17 | 0.504 |
| 18 | 0.510 |
| 19 | 0.517 |

*Πίνακας 1: Σκορ συνεκτικότητας θεμάτων 2-19*

Επίσης πειραματιστήκαμε με διαφορετικές τιμές για τη μεταβλητή passes (5,20) η οποία όμως δεν έδειξε ουσιαστικές διαφορές στην επίδοση του μοντέλου.

Άλλες παράμετροι με τις οποίες θα μπορούσαμε να πειραματιστούμε είναι οι alpha, beta, random state, learning decay και iteration.

Στη συνέχεια δοκιμάσαμε μια δεύτερη τεχνική μοντελοποίησης θεμάτων η οποία προτείνεται για μικρά κείμενα όπως τα tweets, τον GSDMM.

Παρόλο που ο GSDMM προτείνεται συχνά στη βιβλιογραφία για έγγραφα που αποτελούνται απο λιγότερες απο 60 λέξεις όπως τα tweets, δεν υπάρχουν πολλές βιβλιοθήκες που να τον υλοποιούν με εύκολο τρόπο. Στα πλαίσια αυτής της εργασίας χρησιμοποιήθηκε η υλοποίηση του αλγορίθμου απο τη σελίδα Github rwalk/gsdmm[62].

```
np.random.seed(0)
mgp = MovieGroupProcess(K=6, alpha=0.01, beta=0.01, n_iters=5)

vocab = set(x for review in tweets_df['tokens'] for x in review)
n_terms = len(vocab)
model = mgp.fit(tweets_df['tokens'], n_terms)

def top_words(cluster_word_distribution, top_cluster, values):
    for cluster in top_cluster:
        sort_dicts =sorted(mgp.cluster_word_distribution[cluster].items(), key=lambda k: k[1], reverse=True)[:values]
        print("\nCluster %s : %s"%(cluster,sort_dicts))

doc_count = np.array(mgp.cluster_doc_count)
print('Number of documents per topic :', doc_count)

# topics sorted by the number of document they are allocated to
top_index = doc_count.argsort()[-10:][::-1]
print('\nMost important clusters (by number of docs inside):', top_index)
# show the top 5 words in term frequency for each cluster
top_words(mgp.cluster_word_distribution, top_index, 10)
```

*Εικόνα 39: Υλοποίηση GSDMM*

```
In stage 59: transferred 58 clusters with 5 clusters populated
Number of documents per topic : [167 166 221 196 201]

Most important clusters (by number of docs inside): [2 4 3 0 1]

Cluster 2 : [('γιατι', 27), ('υπαρχω', 23), ('συμπτωματα', 19), ('εμβολιο', 19), ('εμβολιασμενος', 17)]

Cluster 4 : [('συμπτωματα', 46), ('εμβολιος', 32), ('παιδια', 24), ('υγεια', 21), ('δοση', 21)]

Cluster 3 : [('συμπτωματα', 30), ('ανθρωπος', 25), ('παιδια', 15), ('κινδυνευω', 15), ('προβλημα', 14)]

Cluster 0 : [('τσιοδρα', 28), ('ζωεςς', 25), ('ανοσια', 20), ('καταστρεφει', 19), ('τσακρης', 18)]

Cluster 1 : [('υγεια', 31), ('γιατι', 15), ('κανω', 14), ('υπαρχω', 14), ('μασκα', 11)]
Coherence Score for GSDMM: 0.4000753454766127
```

*Εικόνα 40: Αποτελέσματα GSDMM*

### 5.4.2 Ανάλυση Συναισθήματος

Για την ανάλυση συναισθήματος, πειραματιστήκαμε με το ήδη εκπαιδευμένο (pre-trained) μοντέλο Greek-BERT το οποίο με βάση τη βιβλιογραφία περιμένουμε να έχει την καλύτερη απόδοση στην ελληνική γλώσσα μιας που έχει εκπαιδευτεί με δεδομένα της ελληνικής γλώσσας.

Για την ανάλυση συναισθήματος εφαρμόσαμε λιγότερη προεπεξεργασία συγκριτικά με τη μοντελοποίηση θεμάτων. Συγκεκριμένα:

- Αφαιρέθηκαν usernames, links και διπλότυπες εγγραφές
- Επίσης αφαιρέθηκαν οι τόνοι, τα κενά απ την αρχή και το τέλος και μετατράπηκαν οι χαρακτήρες σε πεζοί

Ο λόγος που εφαρμόστηκε λιγότερη προ-επεξεργασία είναι για να μη χαθεί το νόημα της προτασης [63].

```
def remove_usernames_links(tweets):
    tweets = re.sub('@[^\s]+',"",tweets)
    tweets = re.sub('http[^\s]+',"",tweets)
    tweets = tweets.lower()
    tweets = tweets.strip()
#     tweets = re.sub(r"long covid","", tweets)
#     tweets = re.sub(r"long","", tweets)
#     tweets = re.sub(r"covid","", tweets)
#     tweets = re.sub(r"\b\w{1,3}\b","", tweets)
#     tweets = re.sub(r'[^\w\s]','', tweets)
    return tweets

def strip_accents_and_lower(s):
     return ''.join(c for c in unicodedata.normalize('NFD', s)
                    if unicodedata.category(c) != 'Mn').lower()
```

*Εικόνα 41: Προεπεξεργασία για ανάλυση συναισθήματος*

Εγκαταστησαμε τη βιβλιοθηκη transformers και torch και εισαγαμε το μοντελο.

```
from transformers import AutoTokenizer, AutoModel

tokenizer = AutoTokenizer.from_pretrained("nlpaueb/bert-base-greek-uncased-v1")
model = AutoModel.from_pretrained("nlpaueb/bert-base-greek-uncased-v1")
```

*Εικόνα 42: Φόρτωση μοντέλου*

Στη συνέχεια μετατρέψαμε τα tweets του συνόλου δεδομένων σε tokens.

Το πρόβλημα που καλούμαστε να λύσουμε είναι οτι ένα μοντέλο μηχανικής μάθησης χρειάζεται ενα σετ δεδομένων εκπαίδευσης όπου το συναίσθημα να έχει δοθεί ήδη έτσι ωστε το μοντέλο να μάθει απ τα δεδομένα εκπαίδευσης και να μπορέσει μετά να προβλέψει το συναίσθημα και για το σετ δεδομένων test.

Αυτό στην περίπτωση μας δεν ήταν δυνατό καθώς τα tweets που αποκτήσαμε απ το Twitter δεν έχουν ήδη ανατεθειμένο συναίσθημα και επίσης ο ογκος τους ηταν σχετικά μικρός για να αφιερωθεί ένα μέρος τους στην εκπαίδευση.

Οπότε εκπαιδεύσαμε το μοντέλο σε κριτικές που αφορούν το Σκρουτζ που είχαν ήδη συναισθηματικό πρόσημο [64] όμως η πρόβλεψη αφορούσε τα tweets του Long Covid που έχουμε κατεβάσει.

Το πλεονέκτημα είναι οτι έχουμε πλέον ένα μεγάλο σύνολο δεδομένων εκπαίδευσης που αποτελείται απο 6552 κριτικές στην ελληνική γλώσσα.

Το μειονέκτημα είναι οτι το σύνολο δεδομένων έχει χρησιμοποιησει δύο (2) συναισθήματα μόνο: θετικό και αρνητικό και ως εκ τουτου προέβλεψε μεταξύ αυτών των δύο

συναισθημάτων. Το σύνολο δεδομένων εκπαίδευσης περιέχει 3276 θετικές και 3276 αρνητικές κριτικές – είναι αρκετά ισορροπημένο.

Ένα δείγμα των δεδομένων φαίνεται στην Εικόνα 43.

```
(5896, 3)
(656, 3)
        id                                          DATA_COLUMN  LABEL_COLUMN
3626  3626    παρηγγειλα μεσα Ιουλιου ενα κλιματιστικο και...             0
3564  3564    Εχω κανει παραγγελια 26/11 και πληρωσα ηλεκτ...             0
2811  2811    Απλα απαραδεκτοι!!!   Η παραγγελεια εγινε τη...             0
1461  1461    Το συγκεκριμενο καταστημα ειναι πραγματικα π...             1
5237  5237    17-11-2020 Πλήρωσα και παρέλαβα μετά από δυο...             0
...    ...                                                ...           ...
5616  5616    Έκανα μια παραγγελία 02/04/2020 ένα Στέγαστρ...             0
5626  5626    Δεν είναι η πρώτη φορά που αντιμετωπίζω πρόβ...             0
4552  4552    Ικανοποιητική Εξυπηρέτηση   Γρήγορη Παραλαβή...             1
5198  5198    Δεν υπάρχουν λόγια για να περιγράψω το συγκε...             0
3823  3823    Καλώ σήμερα 23/4/2020 για να μάθω την εξέλιξ...             0
```

*Εικόνα 43: Δείγμα δεδομένων εκπαιδευσης*

Η ακρίβεια του μοντέλου στα δεδομένα του test είναι αρκετά καλή.

```
[313, 15]
[10, 318]
Accuracy: 96.19%
Precision: 96.90%
Recall: 95.43%
F1-score: 96.16%
```

*Εικόνα 44: Ακρίβεια μοντέλου σε δεδομένα τεστ απο το σύνολο δεδομένων εκπαίδευσης*

Στη συνέχεια το μοντέλο κλήθηκε να προβλέψει το συναίσθημα στα tweets που αφορούν το Long Covid και ένα δείγμα των αποτελεσμάτων φαίνεται στην Εικόνα 45.

| Tweets | Sentiment |
|---|---|
| 2023.η χρονια εξαρσης σε αφορητο βαθμο της ξαφνικιτιδας, λο | 0 |
| δεν ισχυει αυτο τι λες; βιολογικα παραλογο. οι μεταλλαξεις προ | 0 |
| ε τοτε για τα εισπνεομενα που δε βρισκω, φταινε οι πνευμονολο | 0 |
| την ημερα που εκαταλαβεν οτι με το placebo δεν παθαινεις long | 1 |
| long covid σε ανεμβολιατσους που επαναμολυνονται | 1 |
| τα παντα ξερουμε...δεν εχουμε χασει ουτε μια ενημερωση στο t | 0 |
| ετρεξαν και καποιοι αλλοι που στην συνεχεια προετρεψαν τους | 0 |
| ιατρειο long covid ευαγγελισμου, οταν περασα 1η φορα κι εκαν | 0 |
| οι τελευταιες σουηδικες μελετες δινουν ποσοστο 1% long covid | 0 |
| στη συσταση ειδικης επιστημονικης επιτροπης long covid προχω | 1 |
| αδερφε δεν πας καλα . σε τι σου φταινε οι υπολοιποι αν εσυ εχ | 0 |
| στη συσταση ειδικης επιστημονικης επιτροπης long covid προχω | 1 |
| περαστικα γρηγορα ευχομαι! υπαρχουν ιατρεια long covid σε αρ | 1 |

*Εικόνα 45: Δείγμα Αποτελεσμάτων πρόβλεψης συναισθήματος στα Tweets*

Παρατηρούμε οτι τα περισσότερα tweets έχουν κατηγοριοποιηθεί με αρνητικό συναίσθημα το οποίο ίσως αναμένεται δεδομένου οτι το θέμα Long Covid δεν ειναι ευχάριστο.

| Sentiment | Count |
|---|---|
| Negative | 555 |
| Positive | 382 |

*Πίνακας 2: Αποτελέσματα Ανάλυσης Συναισθήματος*

Συγκρίναμε τα αποτελέσματα του μοντέλου με συναίσθημα που ανατέθηκε απο άνθρωπο και το μοντέλο είχε ακρίβεια 94%.

| Results | Count | Percentage |
|---|---|---|
| Human: Positive, BERT: Positive | 169 | 0.370 |
| Human: Positive, BERT: Negative | 2 | 0.004 |
| Human: Negative, BERT: Negative | 261 | 0.571 |
| Human: Negative, BERT: Positive | 25 | 0.055 |
| | 457 | |

*Πίνακας 3: Σύγκριση επίδοσης ανθρώπου και BERT*

Η μόνη παρατήρηση είναι οτι tweets με ουδέτερο πρόσημο έπρεπε να μπουν σε μια απ τις δύο κατηγορίες καθώς το μοντέλο που χρησιμοποιήθηκε είχε εκπαιδευτεί ώστε να

κατηγοριοποιεί σε δύο συναισθήματα (αρνητικό, θετικό) και στην πλειοψηφία τους, όπως παρατηρήσαμε κατόπιν στο δείγμα, έχουν κατηγοριοποιηθεί στο θετικό συναίσθημα.

# Κεφάλαιο 6 Συμπεράσματα – Μελλοντικές Επεκτάσεις

## 6.1 Συμπεράσματα

Ανακτήθηκαν 1000 tweets στην ελληνική γλώσσα τα οποία αναρτήθηκαν κατα τη διάρκεια του 2022. Προυπόθεση για να συμμετάσχουν στο σύνολο δεδομένων ήταν να περιέχουν τη φράση Long Covid ή το hashtag #longCovid. Μετά την προ επεξεργασία και την αφαίρεση διπλότυπων εγγραφών, το σύνολο των δεδομένων μειώθηκε σε 937.

Οι πιο συχνά εμφανιζόμενες λέξεις στο σύνολο των δεδομένων είναι: ΄συμπτώματα’, ‘παιδιά’, ‘μελέτη’, ‘εμβόλια’, ΄χρόνια’, άσθενείς’. Μετά απο μια οπτική εξέταση των tweets και με βάση το wordcloud καταλαβαίνουμε οτι υπάρχει ενα άγχος για τα συμπτώματα του Long Covid και τη χρονική τους διάρκεια. Επισης υπάρχει ανησυχία ως προς τις επιπτώσεις του Long Covid σε μερίδες του πληθυσμού όπως τα παιδιά. Τέλος συζητάται η θετική ή αρνητική επίδραση των εμβολίων στο Long Covid.

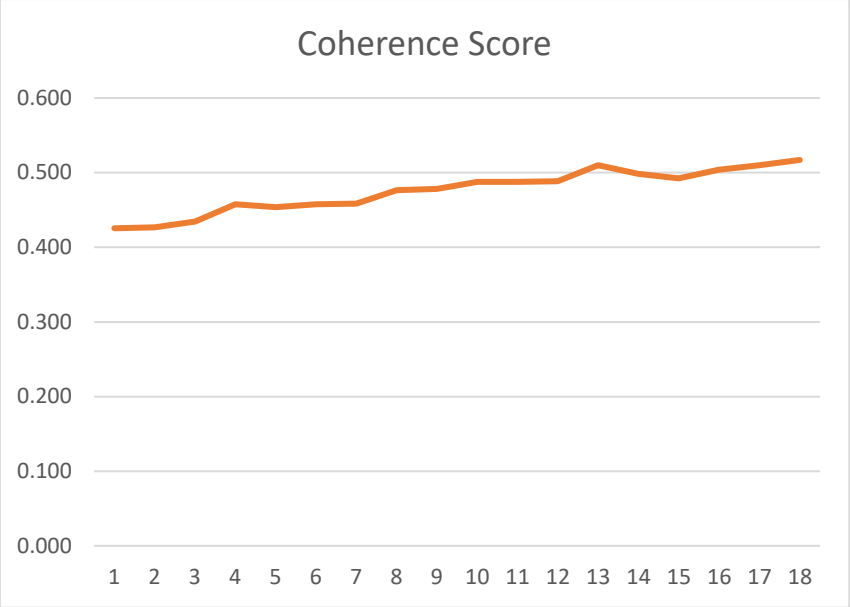

*Εικόνα 46: Τελικό σύννεφο λέξεων*

Στη συνέχεια εφαρμόστηκαν δυο διαφορετικές τεχνικές μοντελοποίησης θεμάτων για να αναδειχθούν τα θέματα συζήτησης γύρω απ το Long Covid.

Στον LDA, δοκιμάστηκαν επαναληπτικά διαφορετικοί αριθμοί θεμάτων και βαθμολογήθηκαν σε σχέση με τη συνοχή τους. Βλέπουμε ότι μεταξύ 4 και 19 θέματων δεν υπάρχει ουσιαστική βελτίωση στο σκορ συνοχής το οποίο κυμαίνεται μεταξύ 0.43-0.52.

| Number of Topics | Coherence Score |
|---|---|
| 2 | 0.425 |
| 3 | 0.427 |
| 4 | 0.434 |
| 5 | 0.458 |
| 6 | 0.454 |
| 7 | 0.458 |
| 8 | 0.458 |
| 9 | 0.476 |
| 10 | 0.478 |
| 11 | 0.488 |
| 12 | 0.488 |
| 13 | 0.488 |
| 14 | 0.510 |
| 15 | 0.498 |
| 16 | 0.492 |
| 17 | 0.504 |
| 18 | 0.510 |
| 19 | 0.517 |

*Πίνακας 4: Σκορ συνεκτικότητας για θέματα 2-19*




*Εικόνα 47: Σκορ συνεκτικότητας ανά αριθμό θεμάτων*

Οπότε στο πλαίσιο αυτής της εργασίας προτιμήθηκε ένας μικρότερος αριθμός θεμάτων ούτως ώστε να γίνει και ποιοτική ανάλυση των αποτελεσμάτων. Αναδείχτηκαν πέντε (5) θέματα. Παρατηρούμε οτι υπάρχει επικάλυψη θεμάτων πχ το θέμα ένα (1) με το θεμα τρία (3) είναι αρκετά παρόμοια.

Το πρώτο θέμα αφορά στην υγεία και τον χρόνο που παίρνει το Long Covid για να θεραπευτεί. Το δεύτερο θέμα αφορά στα συμπτώματα του Long Covid. Το τρίτο θέμα είναι αρκετά παρόμοιο με το πρώτο με έμφαση στη διάρκεια των συμπτωματων. Το τεταρτο θεμα αναφερεται σε ευπαθείς ομαδες οπως τα παιδια και το πεμπτο θέμα αναφέρεται στα εμβόλια

| | Topic # 01 | Topic # 02 | Topic # 03 | Topic # 04 | Topic # 05 |
|---|---|---|---|---|---|
| 0 | υγεία | άνθρωπος | υγεία | εμβόλιο | εμβόλιο |
| 1 | γιατί | κάνω | ασθενής | παιδί | πολύς |
| 2 | υπάρχω | χρόνος | κάνω | μήνας | συμπτώματα |
| 3 | μήνας | συμπτώμα | χρόνος | κάνω | γιατί |
| 4 | χρόνος | κινδυνεύω | υπάρχω | μάσκα | κίνδυνος |

*Εικόνα 48: Τελική μορφή θεμάτων LDA*

Όσον αφορά στον GSDMM, προέκυψαν πεντε (5) θέματα και είναι σχετικά ισοκατανεμημένα. Παρομοίως με τον LDA τα θέματα αφορούν:

- Πρώτο θεμα: θετικές και αρνητικές επιπτώσεις των εμβολίων στο Long Covid
- Δεύτερο θέμα: επιπτώσεις στα παιδιά

- Τρίτο θέμα: σχετικά παρεμφερές με το δεύτερο
- Τέταρτο θεμα: αναφορά σε δημόσια πρόσωπα της Long Covid επικαιρότητας
- Πέμπτο θέμα: η αποτελεσματικότητα της μασκας

```
In stage 59: transferred 58 clusters with 5 clusters populated
Number of documents per topic : [167 166 221 196 201]

Most important clusters (by number of docs inside): [2 4 3 0 1]

Cluster 2 : [('γιατι', 27), ('υπαρχω', 23), ('συμπτωματα', 19), ('εμβολιο', 19), ('εμβολιασμενος', 17)]

Cluster 4 : [('συμπτωματα', 46), ('εμβολιος', 32), ('παιδια', 24), ('υγεια', 21), ('δοση', 21)]

Cluster 3 : [('συμπτωματα', 30), ('ανθρωπος', 25), ('παιδια', 15), ('κινδυνευω', 15), ('προβλημα', 14)]

Cluster 0 : [('τσιοδρα', 28), ('ζωεςς', 25), ('ανοσια', 20), ('καταστρεφει', 19), ('τσακρης', 18)]

Cluster 1 : [('υγεια', 31), ('γιατι', 15), ('κανω', 14), ('υπαρχω', 14), ('μασκα', 11)]
Coherence Score for GSDMM: 0.4000753454766127
```

*Εικόνα 49: Τελική μορφή θεμάτων GSDMM*

Το coherence score του GSDMM είναι 0.40.

Οπότε συνολικά αν συγκρίνουμε τις δυο μεθόδους μοντελοποίησης θεμάτων που εφαρμόστηκαν στα δεδομένα μας ως προς τον δείκτη συνοχής παρατηρούμε οτι ο LDA είχε καλύτερη απόδοση.

| | LDA | GSDMM |
|---|---|---|
| Coherence Score | Max. 0.47 | Max. 0.40 |

*Πίνακας 5: Σύγκριση LDA & GSDMM*

Στη συνέχεια εφαρμόστηκε ανάλυση συναισθήματος χρησιμοποιώντας το μοντέλο Greek-BERT. Παρόλο που το μοντέλο αυτό έχει παρουσιαστεί έδω και δύο χρόνια, προκαλεί εντύπωση το γεγονός οτι δεν βρέθηκαν πολλά σύνολα δεδομένων εκπαίδευσης. Η εκπαίδευση έγινε πάνω σε κριτικές του Skroutz και στη συνέχεια η πρόβλεψη έγινε στο δικο μας σύνολο δεδομένωνμε 937 tweets που αφορούν το Long Covid.

Παρατηρούμε οτι τα περισσότερα tweets έχουν κατηγοριοποιηθεί με αρνητικό συναίσθημα το οποίο ίσως αναμένεται δεδομένου οτι το θέμα Long Covid δεν ειναι ευχάριστο. Τα αποτελέσματα της ανάλυσης συναισθήματος εμφανίζουν συνάφεια στα ποσοστά με την προηγούμενη μελέτη σε ελληνικά tweets που αφορούσε το Covid-19 και τις αντιδράσεις στα εμβόλια [54].

| **Sentiment** | **Count** | **%** |
|---|---|---|

| | | |
|---|---|---|
| Negative | 555 | 59.2% |
| Positive or Neutral | 382 | 40.7% |

*Πίνακας 6: Αποτελέσματα ανάλυσης συναισθήματος*

Συγκρίναμε τα αποτελέσματα του μοντέλου στα μισά περίπου tweets με συναίσθημα που ανατέθηκε απο άνθρωπο και το μοντέλο είχε ακρίβεια 94% το οποίο είναι αρκετά υψηλό.

| **Results** | **Count** | **Percentage** |
|---|---|---|
| Human: Positive, BERT: Positive | 169 | 0.370 |
| Human: Positive, BERT: Negative | 2 | 0.004 |
| Human: Negative, BERT: Negative | 261 | 0.571 |
| Human: Negative, BERT: Positive | 25 | 0.055 |
| | 457 | |

*Πίνακας 7: Σύγκριση επίδοσης ανθρώπου και Greek-BERT*

## 6.2 Μελλοντικές Επεκτάσεις

Η εργασία αυτή έλαβε ορισμένες υποθέσεις και επιδεχεται μελλοντικών βελτιωμένων επεκτάσεων.

Καταρχήν στη συλλογή των δεδομένων επιλέχθηκε η ελληνική γλώσσα αλλά δεν εφαρμόστηκε περιορισμός στη χώρα προέλευσης, κάτι που θα ήταν σκόπιμο για να υπάρχει περισσότερη εμπιστοσύνη οτι τα tweets αντιπροσωπεύουν τον ελλαδικό χώρο (και όχι ας πούμε ομογενείς σε διαφορετικές χώρες) προκειμένου να μη μειωθεί δραματικά ο αριθμός των tweets. Η χώρα προέλευσης δεν συμπληρώνεται πάντα.

Επίσης λόγω του οτι η έρευνα έγινε για tweets στα ελληνικά υπάρχουν περιορισμένες βιβλιοθήκες προσαρμοσμένες στην ελληνική γλώσσα με αποτέλεσμα το στάδιο της προ-επεξεργασίας να μην είναι ιδανικό. Ο τονισμός των λέξεων καθώς και η πλούσια μορφολογία της ελληνικής γλώσσας προκάλεσαν επιπρόσθετες δυσκολίες στην προεπεξεργασία και τη λημματοποίηση. Υπήρξαν διάφορα λάθη που εντοπίζονται απο το ανθρώπινο μάτι αλλά όχι απο τη μηχανή. Θα ειχε ενδιαφέρον να επενδυθεί χρόνος για να κατασκευαστεί ενα pipeline και lemmatizer συγκεκριμένα για τέτοιες εργασίες επεξεργασίας φυσικής γλώσσας και εξόρυξης γνώσης στα ελληνικά.

Στα πλαίσια αυτής της εργασίας χρησιμοποιήθηκαν δύο αρκετά διαδεδομένοι αλγόριθμοι για μοντελοποίηση θεμάτων, ο Latent Dirichlet Allocation (LDA) και ο GSDMM. Θα είχε ενδιαφέρον να δοκιμαστούν περισσότεροι πειραματισμοί για βελτιστοποίηση των αποτελεσμάτων και στους δυο αλγορίθμους. Καθώς και διαφορετικοί αλγόριθμοι μοντελοποίησης θεμάτων όπως ο HDP ο οποίος αναδεικνύει τα θέματα και τον αριθμό τους.

Όσον αφορά στην ανάλυση συναισθήματος, μελλοντικές επεκτάσεις μπορούν να αφορούν στην εκπαίδευση μοντέλων πάνω σε ελληνικά tweets και συγκεκριμένα ελληνικά tweets που αφορούν την υγεία καθώς το μοντέλο αυτής της εργασίας εκπαιδεύτηκε σε κριτικές e-commerce site (Skroutz). Θα άξιζε τον κόπο να επενδυθεί χρόνος για να δημιουργηθούν περισσότερα σύνολα δεδομένων εκπαίδευσης με εύρος συναισθήματα το οποίο σίγουρα θα αυξήσει τις εργασίες στον τομέα της ανάλυσης συναισθήματος.

# Βιβλιογραφία

1. DATA | meaning [Internet]. Cambridge English Dict. [cited 2022 Jun 22]. Available from: https://dictionary.cambridge.org/dictionary/english/data

2. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data [Internet]. SpringerOpen; 2019 [cited 2022 Jun 23];6:1–25. Available from: https://link.springer.com/articles/10.1186/s40537-019-0217-0

3. Μαλλίδη Κ. Εξόρυξη Γνώσης από το Twitter με σκοπό την Ανάλυση Συναισθήματος σχετικά με τον Covid-19 [Internet]. Ελληνικό Ανοικτό Πανεπιστήμιο / Hellenic Open University; 2021 [cited 2022 Dec 20]. Available from: https://apothesis.eap.gr/handle/repo/52129

4. Laney D. 3D data management: controlling data volume, velocity, and variety [Internet]. Appl. Deliv. Strateg. META Group Inc; 2001. Available from: https://studylib.net/doc/8647594/3d-data-management--controlling-data-volume--velocity--an...

5. Han J, Kamber M, Pei J. Data Mining. Concepts and Techniques [Internet]. 3rd ed. Morgan Kaufmann Books - Elsevier; 2011 [cited 2022 Jun 22]. Available from: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

6. Botelho B, Bigelow S. What is Big Data and Why is it Important? [Internet]. TechTarget. 2022 [cited 2023 Mar 11]. Available from: https://www.techtarget.com/searchdatamanagement/definition/big-data

7. Healthcare Big Data and the Promise of Value-Based Care [Internet]. NEJM Catal. 2018 [cited 2023 Mar 11]. Available from: https://catalyst.nejm.org/doi/full/10.1056/CAT.18.0290

8. Pastorino R, De Vito C, Migliara G, Glocker K, Binenbaum I, Ricciardi W, et al. Benefits and challenges of Big Data in healthcare: an overview of the European initiatives. Eur J Public Health [Internet]. Oxford University Press; 2019 [cited 2023 Mar 11];29:23. Available from: /pmc/articles/PMC6859509/

9. Ambert K, Beaune S, Paré A, Chaibi A, Briard L, Bhattacharjee A, et al. French Hospital

Uses Trusted Analytics Platform to Predict Emergency Department Visits and Hospital Admissions [Internet]. Available from: https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/french-hospital-analytics-predict-admissions-paper.pdf

10. SOCIAL MEDIA | meaning [Internet]. Cambridge English Dict. [cited 2022 Jun 23]. Available from: https://dictionary.cambridge.org/dictionary/english/social-media

11. Davison KP, Pennebaker JW, Dickerson SS. Who talks? The social psychology of illness support groups. Am Psychol [Internet]. American Psychological Association Inc.; 2000 [cited 2022 Dec 20];55:205–17. Available from: https://psycnet.apa.org/doi/10.1037/0003-066X.55.2.205

12. What is Natural Language Processing? [Internet]. IBM. [cited 2022 Jun 28]. Available from: https://www.ibm.com/cloud/learn/natural-language-processing

13. Hu M, Liu B. Mining and summarizing customer reviews. KDD-2004 - Proc Tenth ACM SIGKDD Int Conf Knowl Discov Data Min [Internet]. Association for Computing Machinery; 2004 [cited 2023 Mar 13];168–77. Available from: https://dl.acm.org/doi/10.1145/1014052.1014073

14. Gohil S, Vuik S, Darzi A. Sentiment Analysis of Health Care Tweets: Review of the Methods Used. JMIR Public Heal Surveill [Internet]. JMIR Publications Inc.; 2018 [cited 2022 Jun 21];4. Available from: /pmc/articles/PMC5938573/

15. Subramaniam A. Patient Sentiment Analysis in Healthcare [Internet]. KANINI. [cited 2023 Mar 12]. Available from: https://kanini.com/blog/ai/patient-sentiment-analysis-in-healthcare/

16. Sarirete A. Sentiment analysis tracking of COVID-19 vaccine through tweets. J Ambient Intell Humaniz Comput [Internet]. Nature Publishing Group; 2022 [cited 2022 Jun 21];1:1. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8966855/

17. ES S. Sentiment Analysis in Python: TextBlob vs Vader Sentiment vs Flair vs Building It From Scratch [Internet]. neptune.ai. 2023 [cited 2022 Nov 20]. Available from: https://neptune.ai/blog/sentiment-analysis-python-textblob-vs-vader-vs-flair

18. Shofiya C, Abidi S. Sentiment Analysis on COVID-19-Related Social Distancing in Canada Using Twitter Data. Int J Environ Res Public Health [Internet]. Int J Environ Res

Public Health; 2021 [cited 2022 Jun 21];18. Available from: https://pubmed.ncbi.nlm.nih.gov/34204907/

19. Georgouli A, Γεωργούλη Α. Μηχανική Μάθηση. 2015 [cited 2022 Dec 21]. Available from: http://repository.kallipos.gr/handle/11419/3382

20. Du J, Michalska S, Subramani S, Wang H, Zhang Y. Neural attention with character embeddings for hay fever detection from twitter. Heal Inf Sci Syst [Internet]. Springer International Publishing; 2019 [cited 2022 Jun 24];7:1–7. Available from: https://link.springer.com/article/10.1007/s13755-019-0084-2

21. Yıldırım S. Naive Bayes Classifier — Explained [Internet]. Towar. Data Sci. 2020 [cited 2022 Dec 21]. Available from: https://towardsdatascience.com/naive-bayes-classifier-explained-50f9723571ed

22. Kulshrestha R. A Beginner's Guide to Latent Dirichlet Allocation(LDA) [Internet]. Towar. Data Sci. 2019 [cited 2022 Dec 21]. Available from: https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

23. Stemming and lemmatization [Internet]. [cited 2023 Jan 30]. Available from: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

24. Azad A. Twitter Topic Modeling [Internet]. Towar. Data Sci. 2020 [cited 2023 Feb 13]. Available from: https://towardsdatascience.com/twitter-topic-modeling-e0e3315b12e2

25. Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM. Reading Tea Leaves: How Humans Interpret Topic Models. Adv Neural Inf Process Syst 22 [Internet]. 2009 [cited 2023 Feb 12]. p. 288–96. Available from: https://papers.nips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

26. Gan J, Qi Y. Selection of the Optimal Number of Topics for LDA Topic Model—Taking Patent Policy Analysis as an Example. Entropy [Internet]. Multidisciplinary Digital Publishing Institute (MDPI); 2021 [cited 2023 Feb 14];23. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8534395/

27. Yin J, Wang J. A Dirichlet multinomial mixture model-based approach for short text clustering. Proc ACM SIGKDD Int Conf Knowl Discov Data Min [Internet]. New York: Association for Computing Machinery; 2014 [cited 2023 Feb 11]. p. 233–42. Available from: https://dl.acm.org/doi/10.1145/2623330.2623715

28. Pelgrim R. Short-Text Topic Modelling: LDA vs GSDMM [Internet]. Towar. Data Sci. 2021 [cited 2023 Feb 16]. Available from: https://towardsdatascience.com/short-text-topic-modelling-lda-vs-gsdmm-20f1db742e14

29. Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. Attention Is All You Need. 31st Conf Neural Inf Process Syst (NIPS 2017) [Internet]. 2017 [cited 2023 Feb 18]. Available from: https://arxiv.org/abs/1706.03762

30. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019 - 2019 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol - Proc Conf [Internet]. Association for Computational Linguistics (ACL); 2018 [cited 2023 Feb 6];1:4171–86. Available from: https://arxiv.org/abs/1810.04805v2

31. Devlin J, Chang M-W. Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing – Google AI Blog [Internet]. Google Res. 2018 [cited 2023 Feb 18]. Available from: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

32. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform [Internet]. Oxford Academic; 2022 [cited 2023 Feb 18];23. Available from: https://academic.oup.com/bib/article/23/6/bbac409/6713511

33. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics [Internet]. Oxford University Press; 2019 [cited 2023 Feb 18];36:1234–40. Available from: http://arxiv.org/abs/1901.08746

34. Koutsikakis J, Chalkidis I, Malakasiotis P. GREEK-BERT: The Greeks visiting Sesame Street. 11th Hell Conf Artif Intell [Internet]. 2020 [cited 2023 Feb 4]. p. 110–7. Available from: https://doi.org/10.1145/3411408.3411440

35. Eysenbach G. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. J Med Internet Res 2009 [Internet]. Journal of Medical Internet Research; 2009 [cited 2022 Jul 19];11. Available from: https://www.jmir.org/2009/1/e11

36. How Crucial is Sentiment Analysis in Healthcare - Blog [Internet]. Pract. Build. 2022

[cited 2023 Mar 12]. Available from: https://www.practicebuilders.com/blog/sentiment-analysis-in-healthcare/

37. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. J Med Internet Res 2013 [Internet]. Journal of Medical Internet Research; 2013 [cited 2022 Jul 24];15. Available from: https://www.jmir.org/2013/11/e239

38. Rosa RL, Schwartz GM, Ruggiero WV, Rodriguez DZ. A Knowledge-Based Recommendation System That Includes Sentiment Analysis and Deep Learning. IEEE Trans Ind Informatics. IEEE Computer Society; 2019;15:2124–35.

39. Bell D, Laparra E, Kousik A, Ishihara T, Surdeanu M, Kobourov S. Detecting Diabetes Risk from Social Media Activity. Proc 9th Int Work Heal Text Min Inf Anal (LOUHI 2018) [Internet]. 2018 [cited 2023 Mar 12]. p. 1–11. Available from: https://aclanthology.org/W18-5601/

40. Raghupathi V, Ren J, Raghupathi W. Studying Public Perception about Vaccination: A Sentiment Analysis of Tweets. Int J Environ Res Public Heal 2020, Vol 17, Page 3464 [Internet]. Multidisciplinary Digital Publishing Institute; 2020 [cited 2023 Mar 14];17:3464. Available from: https://www.mdpi.com/1660-4601/17/10/3464/htm

41. WHO Coronavirus (COVID-19) Dashboard [Internet]. [cited 2022 Jun 24]. Available from: https://covid19.who.int/

42. Guo JW, Radloff CL, Wawrzynski SE, Cloyes KG. Mining twitter to explore the emergence of COVID-19 symptoms. Public Health Nurs [Internet]. John Wiley & Sons, Ltd; 2020 [cited 2022 Dec 21];37:934–40. Available from: https://onlinelibrary.wiley.com/doi/full/10.1111/phn.12809

43. Kwok S, Vadde SK, Wang G. Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. J Med Internet Res 2021 [Internet]. 2021 [cited 2022 Aug 18];23. Available from: https://www.jmir.org/2021/5/e26953

44. Yin H, Song X, Yang S, Li J. Sentiment analysis and topic modeling for COVID-19 vaccine discussions. World Wide Web [Internet]. Springer; 2022 [cited 2023 Mar 2];25:1067–83. Available from: https://doi.org/10.1007/s11280-022-01029-y

45. Jianlong Zhou HZSYJSXGCF. Detecting Community Depression Dynamics Due to COVID-19 Pandemic in Australia. IEEE Trans Comput Soc Syst [Internet]. 2021 [cited 2022 Dec 10];8:958–67. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9325873

46. Stecanella B. Understanding TF-ID: A Simple Introduction [Internet]. MonkeyLearn Blog. 2019 [cited 2022 Dec 16]. Available from: https://monkeylearn.com/blog/what-is-tf-idf/

47. Zhang S, Sun L, Zhang D, Li P, Liu Y, Anand A, et al. The COVID-19 Pandemic and Mental Health Concerns on Twitter in the United States. Heal Data Sci [Internet]. 2022 [cited 2022 Dec 11];2022. Available from: https://doi.org/10.34133/2022/9758408

48. Chakraborty K, Bhatia S, Bhattacharyya S, Platos J, Bag R, Hassanien AE. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. Appl Soft Comput [Internet]. Elsevier; 2020 [cited 2023 Mar 16];97:106754. Available from: https://doi.org/10.1016/j.asoc.2020.106754

49. Bhambhoria R, Saab J, Uppal S, Li X, Yakimovich A, Bhatti J, et al. Towards Providing Clinical Insights on Long Covid from Twitter Data. Stud Comput Intell [Internet]. Springer Science and Business Media Deutschland GmbH; 2023 [cited 2022 Dec 16];1060:267–78. Available from: https://link.springer.com/chapter/10.1007/978-3-031-14771-5_19

50. Matharaarachchi S, Domaratzki M, Katz A, Muthukumarana S. Discovering Long COVID Symptom Patterns: Association Rule Mining and Sentiment Analysis in Social Media Tweets. JMIR Form Res 2022;6(9)e37984 [Internet]. JMIR Formative Research; 2022 [cited 2023 Mar 16];6:e37984. Available from: https://formative.jmir.org/2022/9/e37984

51. Awoyemi T, Ebili U, Olusanya A, Ogunniyi KE, Adejumo A V. Twitter Sentiment Analysis of Long COVID Syndrome. Cureus [Internet]. Cureus Inc.; 2022 [cited 2022 Dec 21];14. Available from: /pmc/articles/PMC9278796/

52. Kydros D, Argyropoulou M, Vrana V. A Content and Sentiment Analysis of Greek Tweets during the Pandemic. Sustain 2021 [Internet]. Multidisciplinary Digital Publishing Institute; 2021 [cited 2021 Nov 20];13:6150. Available from: https://www.mdpi.com/2071-1050/13/11/6150/htm

53. Geronikolou S, Drosatos G, Chrousos G. Emotional Analysis of Twitter Posts During the

First Phase of the COVID-19 Pandemic in Greece: Infoveillance Study. JMIR Form Res 2021 [Internet]. JMIR Formative Research; 2021 [cited 2022 Dec 18];5. Available from: https://formative.jmir.org/2021/9/e27741

54. Kapoteli E, Koukaras P, Tjortjis C. Social Media Sentiment Analysis Related to COVID-19 Vaccines: Case Studies in English and Greek Language. IFIP Adv Inf Commun Technol [Internet]. Springer Science and Business Media Deutschland GmbH; 2022 [cited 2022 Dec 18];647:360–72. Available from: https://link.springer.com/chapter/10.1007/978-3-031-08337-2_30

55. Katsarou MS, Iasonidou E, Osarogue A, Kalafatis E, Stefanatou M, Pappa S, et al. The Greek Collaborative Long COVID Study: Non-Hospitalized and Hospitalized Patients Share Similar Symptom Patterns. J Pers Med 2022, Vol 12, Page 987 [Internet]. Multidisciplinary Digital Publishing Institute; 2022 [cited 2023 Mar 19];12:987. Available from: https://www.mdpi.com/2075-4426/12/6/987/htm

56. Use Cases, Tutorials, & Documentation | Twitter Developer Platform [Internet]. [cited 2023 Mar 30]. Available from: https://developer.twitter.com/en

57. GitHub - stopwords-iso/stopwords-el: Greek stopwords collection [Internet]. [cited 2023 Jan 14]. Available from: https://github.com/stopwords-iso/stopwords-el

58. NLTK :: Natural Language Toolkit [Internet]. [cited 2023 Mar 19]. Available from: https://www.nltk.org/

59. spaCy Usage Documentation [Internet]. [cited 2023 Feb 5]. Available from: https://spacy.io/usage/linguistic-features#lemmatization

60. Kwok SWH, Vadde SK, Wang G. Tweet Topics and Sentiments Relating to COVID-19 Vaccination Among Australian Twitter Users: Machine Learning Analysis. J Med Internet Res 2021 [Internet]. Journal of Medical Internet Research; 2021 [cited 2022 Jun 24];23. Available from: https://www.jmir.org/2021/5/e26953

61. Singh Tanwar G. Topic Modeling with Latent Dirichlet Allocation (LDA) on the Tweets Mentioning Elon Musk: Part I [Internet]. MLearning.ai. 2022 [cited 2023 Feb 5]. Available from: https://medium.com/mlearning-ai/topic-modelling-with-lda-on-the-tweets-mentioning-elon-musk-687076a2c86b

62. Walker R. rwalk/gsdmm: GSDMM: Short text clustering [Internet]. Github; [cited 2023

Feb 15]. Available from: https://github.com/rwalk/gsdmm

63. Chalkidis I. nlpaueb/bert-base-greek-uncased-v1 [Internet]. Hugging Face. [cited 2023 Apr 1]. Available from: https://huggingface.co/nlpaueb/bert-base-greek-uncased-v1

64. Skroutz Sentiment Analysis with BERT (Greek) [Internet]. Kaggle. [cited 2023 Mar 19]. Available from: https://www.kaggle.com/code/nikosfragkis/skroutz-sentiment-analysis-with-bert-greek/notebook