

Ατομική Διπλωματική Εργασία

**ΜΕΤΡΗΣΗ ΤΗΣ ΠΟΛΩΣΗΣ ΣΕ ΚΟΙΝΩΝΙΚΟ ΠΛΑΙΣΙΟ ΕΝ
ΜΕΣΩ ΠΑΝΔΗΜΙΑΣ ΤΟΥ ΚΟΡΟΝΟΪΟΥ, ΜΕΣΑ ΑΠΟ ΤΑ
ΚΟΙΝΩΝΙΚΑ ΔΙΚΤΥΑ**

Κωνσταντίνος Γεωργίου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ



ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

**Μέτρηση της πόλωσης σε κοινωνικό πλαίσιο εν μέσω πανδημίας του κορονοϊού,
μέσα από τα κοινωνικά δίκτυα**

Κωνσταντίνος Γεωργίου

Επιβλέπων Καθηγητής

Δρ. Γιώργος Πάλλης

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

Μάιος 2022

Ευχαριστίες

Θα ήθελα να ευχαριστήσω κάποια συγκεκριμένα άτομα τα οποία συνέβαλαν σε ένα σημαντικό βαθμό έτσι ώστε να ολοκληρωθούν με επιτυχία οι στόχοι της ατομικής διπλωματικής μου εργασίας.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον επιβλέπων καθηγητή μου Δρ. Γιώργο Πάλλη, Αναπληρωτής Καθηγητής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου, διότι ήταν αυτός που μου έδωσε την ευκαιρία να ασχοληθώ με ένα θέμα που είναι τόσο σημαντικό και παράλληλα επίκαιρο. Οι γνώσεις και ο ενθουσιασμός του για αυτό που κάνει ήταν τα σημεία που με κέρδισαν για να δουλέψω μαζί του. Η συμβολή του καθ' όλη την διάρκεια της ατομικής διπλωματικής μου εργασίας ήταν καίριας σημασίας και ανταποκρινόταν άμεσα στα οποιαδήποτε αιτήματα μου.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω τον υποψήφιο Δρ. Δημήτρη Πασχαλίδη, ερευνητής στο Τμήμα Πληροφορικής του Πανεπιστημίου Κύπρου. Με τη βοήθειά του ήμουν σε θέση όχι μόνο να καταλάβω σημαντικές έννοιες αλλά και να διεκπεραιώσω εργασίες κρίσιμες έτσι ώστε να φέρω εις πέρας την διπλωματική μου εργασία.

Περίληψη

Κανείς δεν μπορεί να διαψεύσει το γεγονός ότι η πανδημία του κορονοϊού ταλανίζει την ανθρωπότητα τα τελευταία δύο χρόνια. Εκτός αυτού, η τεχνολογία έχει κάνει μεγάλα βήματα και έχει αναπτυχθεί σε τέτοιο βαθμό που οι άνθρωποι μπορούν να επικοινωνούν μεταξύ τους σε καθημερινή βάση εύκολα και αποδοτικά. Οι δύο πιο πάνω παράγοντες, δηλαδή η πανδημία του κορονοϊού καθώς και η ανάπτυξη της τεχνολογίας, έχουν μετατρέψει τα κοινωνικά δίκτυα, όπως είναι για παράδειγμα το Facebook, σε μικρογραφία του πραγματικού κόσμου. Με λίγα λόγια, οι άνθρωποι νιώθουν άνετα να συνδέονται στα μέσα κοινωνικής δικτύωσης και να μοιράζονται απόψεις, σκέψεις και συναισθήματα με άλλους χρήστες.

Έχοντας αυτά κατά νου, σκοπός της ατομικής διπλωματικής μου εργασίας ήταν η εφαρμογή υπολογιστικών μεθόδων έτσι ώστε να γίνει ταυτοποίηση των διαφόρων θεμάτων που συζητιούνταν μέσω των κοινωνικών δικτύων και είχαν ως επίκεντρο την πανδημία. Έτσι, χρησιμοποίησα τεχνικές και αλγόριθμους για κατασκευή και ανάλυση γράφου, επεξεργασίας της φυσικής γλώσσας, ανάλυσης δεδομένων καθώς και μηχανική μάθηση για εκπαίδευση μοντέλων. Πιο συγκεκριμένα, το αρχικό στάδιο ήταν η συλλογή δεδομένων από την πλατφόρμα Crowdtangle. Ακολούθησε η κατασκευή γράφου με τέτοιο τρόπο που να αντιπροσωπεύει το δίκτυο αλληλεπίδρασης των διαφόρων σελίδων και γκρουπ στο Facebook. Στη συνέχεια, βρέθηκε η πόλωση μεταξύ κόμβων του γράφου με τρόπο που συστήθηκε από συγγράμματα και τέλος εφαρμόστηκε η μοντελοποίηση θεμάτων που είχε σαν αποτέλεσμα την εμφάνιση των θεμάτων που συζητιούνταν στην περίοδο της πανδημίας. Για να ολοκληρωθεί η έρευνα και να επιβεβαιωθεί η ακρίβεια των αποτελεσμάτων, έγινε τόσο ποσοτική όσο και ποιοτική ανάλυση των αποτελεσμάτων. Συνεπώς, βρέθηκε τόσο η πόλωση σε επίπεδο γράφου όσο και σε επίπεδο συγκεκριμένων θεμάτων συζήτησης.

Περιεχόμενα

Κεφάλαιο 1 Εισαγωγή	1
1.1 Γενική εισαγωγή	1
1.2 Ορολογία.....	3
1.3 Σημαντικά βήματα για την επίτευξη του τελικού στόχου	4
1.4 Προβλήματα που κλήθηκα να αντιμετωπίσω	5
1.5 Περίγραμμα περιεχομένων	5
 Κεφάλαιο 2 Υπόβαθρο και Σχετική δουλειά	7
2.1 Υπόβαθρο (Προκαταρκτικό Στάδιο)	7
2.1.1 Τεχνική εξερεύνησης των δεδομένων (Exploratory Data Analysis)	8
2.1.2 BERT	8
2.1.3 Τεχνική ομαδοποίησης (Clustering)	9
2.2 Αναγνώριση και Μέτρηση της Πόλωσης	10
 Κεφάλαιο 3 Μεθοδολογία	13
3.1 Αρχιτεκτονικό διάγραμμα.....	13
3.2 Επεξήγηση των σταδίων που απεικονίζονται στο διάγραμμα	14
3.2.1 Συλλογή των Δεδομένων	14
3.2.2 Δόμηση του Γράφου	14
3.2.3 Κατάτμηση του Γράφου	15
3.2.4 Εξαγωγή της Πόλωσης	18
3.2.5 Εξαγωγή των Διπόλων.....	19
3.2.6 Τελική Ανάλυση	20
3.3 Λεπτομέρειες υλοποίησης	21
 Κεφάλαιο 4 Πειράματα και Αποτελέσματα.....	23
4.1 Περιγραφή του συνόλου των δεδομένων.....	23
4.2 Πείραμα.....	26
4.3 Αποτελέσματα.....	30

4.3.1 Κοινότητα 1	30
4.3.2 Κοινότητα 2	33
4.3.3 Κοινότητα 4	34
4.3.4 Κοινότητα 5	36
4.3.5 Κοινότητα 9	39
4.3.6 Κοινότητα 24	41
4.3.7 Κοινότητα 45	42
 Κεφάλαιο 5 Συμπεράσματα	43
5.1 Συμπεράσματα διπόλων	44
5.1.1 Δίπολο (1,4)	44
5.1.2 Δίπολο (5,45)	47
5.1.3 Δίπολο (1,5)	50
5.1.4 Δίπολο (5,9)	54
5.2 Ποσοτική έρευνα	55
5.3 Μελλοντική έρευνα.....	56
 Βιβλιογραφία	58
 Παράρτημα Α	Α-1

Κεφάλαιο 1

Εισαγωγή

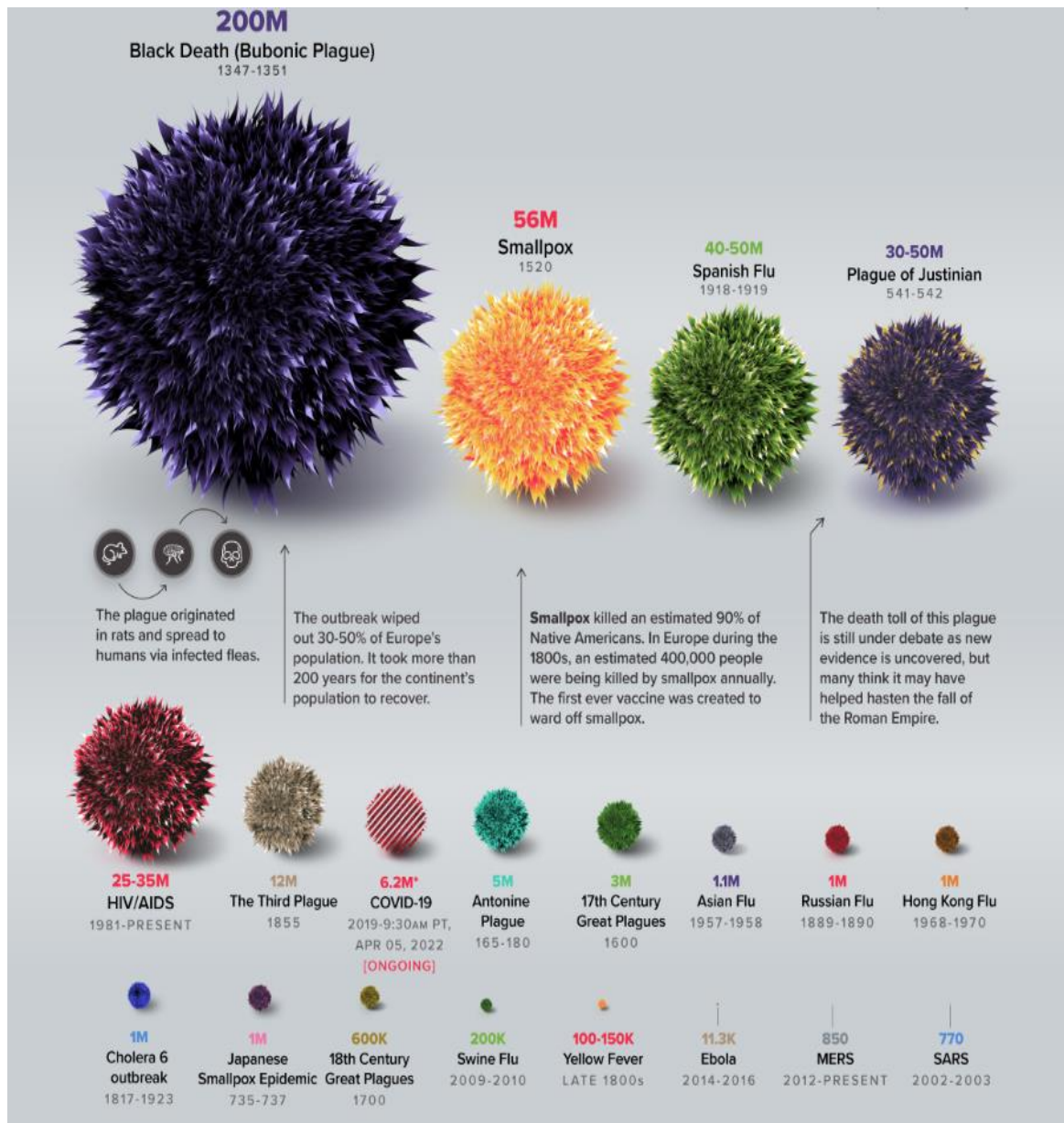
1.1 Γενική εισαγωγή	1
1.2 Ορολογία.....	3
1.3 Σημαντικά βήματα για την επίτευξη του τελικού στόχου	4
1.4 Προβλήματα που κλήθηκα να αντιμετωπίσω	5
1.5 Περίγραμμα περιεχομένων	5

1.1 Γενική εισαγωγή

Αναμφίβολα, η πανδημία του κορονοϊού ήταν το θέμα το οποίο κατείχε την προσοχή των ανθρώπων ανά το παγκόσμιο τα τελευταία χρόνια, κάτι το οποίο είναι αναμενόμενο αφού φαίνεται να είχε σημαντική επίδραση στο ανθρώπινο είδος. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, έχουν καταγραφεί τουλάχιστον 516 εκατομμύρια θετικά περιστατικά κορονοϊού ενώ έχουν σημειωθεί περίπου 6,2 εκατομμύρια θάνατοι [1]. Επιπλέον στατιστικά στοιχεία τα οποία παραθέτει το IHME (Institute for Health Metrics and Evaluation) δείχνουν ότι τουλάχιστον 7 εκατομμύρια θάνατοι οφείλονται στην νόσο του κορονοϊού [2], ενώ σύμφωνα με ένα μοντέλο μηχανικής μάθησης που φτιάχτηκε από το ίδιο ινστιτούτο υπολογίζει ότι από τις αρχές του 2020 οι θάνατοι λόγω κορονοϊού έχουν ξεπεράσει κατά πολύ περισσότερο αυτόν τον αριθμό. Συγκεκριμένα, το μοντέλο υποδεικνύει ότι περίπου 13 εκατομμύρια θάνατοι έχουν προκληθεί από την πανδημία του κορονοϊού σε παγκόσμιο επίπεδο [3].

Παράλληλα, συγκρίνοντας τη συγκεκριμένη πανδημία με άλλες πανδημίες του παρελθόντος φαίνεται ότι ήταν ανάμεσα στις πιο επικίνδυνες για τον άνθρωπο.

Παρατηρώντας την πιο κάτω εικόνα, μπορούμε να διαπιστώσουμε ότι όντως βρίσκεται στις οκτώ πιο απειλητικές μέσα στο πέρασμα του χρόνου λαμβάνοντας υπόψη τον αριθμό των θανάτων ανά το παγκόσμιο και, επίσης, στις δύο πιο απειλητικές που συνεχίζουν να αποτελούν σοβαρό κίνδυνο μέχρι και σήμερα.



Σχήμα 1.1: Εικόνα που απεικονίζει τις πανδημίες τις οποίες το ανθρώπινο είδος κλήθηκε να αντιμετωπίσει, καθώς και η σύγκριση της πανδημίας του κορονοϊού με τις υπόλοιπες [4].

Την ίδια στιγμή που ο άνθρωπος καλείται να ανταπεξέλθει σε αυτές τις δύσκολες συνθήκες, παρατηρείται συνεχής πρόοδος και ανάπτυξη στον τομέα της τεχνολογίας. Οι άνθρωποι είναι σε θέση να επικοινωνούν εύκολα και αποδοτικά με τους συνανθρώπους τους και το διαδίκτυο γεμίζει με δεδομένα και πληροφορίες. Πολλά από αυτά αφορούν την κατάσταση της πανδημίας αφού σε μέσα κοινωνικής δικτύωσης, όπως είναι οι πλατφόρμες του Facebook και του Twitter, οι άνθρωποι εκμεταλλεύονται την ευκαιρία της εύκολης και αποτελεσματικής επικοινωνίας για να ανταλλάξουν απόψεις και να συζητήσουν θέματα που εξελίσσονται γύρω από την πανδημία του κορονοϊού.

Λαμβάνοντας υπόψη ότι πρόκειται για μια κατάσταση που επηρεάζει τις ζωές των ανθρώπων ανά το παγκόσμιο αλλά και το γεγονός ότι μας δίνετε η δυνατότητα να αναλύσουμε δεδομένα που βρίσκονται στο διαδίκτυο, σκοπός μας είναι να μετατρέψουμε αυτές τις πληροφορίες σε γνώση. Πιο συγκεκριμένα, η διπλωματική μου εργασία έχει ως απώτερο στόχο τη δημιουργία ενός συστήματος το οποίο αφού πάρει ως είσοδο κάποια δεδομένα από τα κοινωνικά δίκτυα εν καιρώ πανδημίας, να εφαρμόζει υπολογιστικές μεθόδους για να καταφέρει την ταυτοποίηση των διαφόρων θεμάτων που συζητιούνταν.

1.2 Ορολογία

Λογαριασμός: στην έρευνα αυτή ορίζεται ως η σελίδα ή το γκρουπ στην πλατφόρμα του Facebook. Οι λογαριασμοί προκύπτουν από την συλλογή των δεδομένων που κάνουμε από την σελίδα Crowdtangle API απαντώντας το ερώτημα “ποιος έκανε την συγκεκριμένη δημοσίευση”.

Κοινότητες: οι ομάδες λογαριασμών. Αφού γίνει η υλοποίηση του γράφου και εφαρμόσουμε την τεχνική της ομαδοποίησης, οι λογαριασμοί τοποθετούνται σε ομάδες που μεγιστοποιούν την μετρική modularity. Άρα οι κοινότητες υποδηλώνουν μια σχέση σε επίπεδο γράφου.

Κοινότητες δημοσιεύσεων: οι ομάδες δημοσιεύσεων σε μια κοινότητα. Με την τεχνική της ομαδοποίησης, οι δημοσιεύσεις κάθε κοινότητας χωρίζονται σε ομάδες με τέτοιο τρόπο που δημοσιεύσεις που βρίσκονται στην ίδια ομάδα σημαίνει ότι είναι παρόμοιες και έχουν το ίδιο νοηματικό πλαίσιο. Μια κοινότητα μπορεί να έχει πολλές κοινότητες δημοσιεύσεων αλλά κάθε δημοσίευση ανήκει σε μια μόνο κοινότητα δημοσιεύσεων (που και αυτή με την σειρά της ανήκει μόνο σε μια κοινότητα). Άρα οι κοινότητες δημοσιεύσεων υποδηλώνουν μια σχέση σε επίπεδο κειμένου.

Δίπολα: ζευγάρι κοινοτήτων. Κάθε δίπολο προβάλλει κάποιες πληροφορίες γι' αυτό το ζευγάρι όπως είναι η τιμή της πόλωσης που έχει σε επίπεδο γράφου και ο αριθμός των συνδέσεων του.

Επιλεγμένες κοινότητες: οι κοινότητες η οποίες επιλέχθηκαν με σκοπό να μελετηθούν σε βάθος. Αυτό γίνεται αφού επιλέξουμε συγκεκριμένα δίπολα που θέλουμε να αναλύσουμε και πρώτα παίρνουμε τις κοινότητες τους για να βγάλουμε τα πρώτα μας συμπεράσματα.

1.3 Σημαντικά βήματα για την επίτευξη του τελικού στόχου

Για την επιτυχή ολοκλήρωση της ατομικής διπλωματικής μου εργασίας υπήρχαν κάποια βήματα «κλειδιά» στα οποία αξίζει να αναφερθούμε και που περιγράφονται με λεπτομέρεια στο τρίτο και τέταρτο κεφάλαιο. Στα πρώτα στάδια της έρευνας, έπρεπε να παρθεί μια απόφαση όσο αφορά το κτίσιμο και τη χρήση μιας δομής δεδομένων που να μας επιτρέπει να δούμε την αλληλεπίδραση που έχουν τα δεδομένα μεταξύ τους. Έτσι, το πρώτο κρίσιμο βήμα είχε να κάνει με το κτίσιμο του γράφου που στην ουσία θα ήταν η βάση όλης της έρευνας. Το επόμενο βήμα «κλειδί» έχει να κάνει με την διαχείριση των δεδομένων και πιο συγκεκριμένα με την προ-επεξεργασία και τον καθαρισμό τους. Αυτό ήταν αρκετά σημαντικό έτσι ώστε να μειώσουμε τον «θόρυβο» των δεδομένων και να έχουμε όσο το δυνατό γίνεται καλύτερα αποτελέσματα, ειδικά σε περιπτώσεις που γινόταν χρήση αλγορίθμων και μοντέλων μηχανικής μάθησης όπως το framework μηχανικής μάθησης BERT. Τέλος, μέσω της ποσοτικής και ποιοτικής ανάλυσης διασφαλίσαμε την εγκυρότητα και την ορθότητα των τελικών αποτελεσμάτων που παράξαμε.

1.4 Προβλήματα που κλήθηκα να αντιμετωπίσω

Τα περισσότερα αδιέξοδα που βρήκα είχαν να κάνουν με τον μεγάλο όγκο των δεδομένων. Κάποια παραδείγματα από αυτού του είδους προβλήματα ήταν η συλλογή, η οργάνωση και η επεξεργασία των δεδομένων. Επιπλέον, αυτές οι διεργασίες επηρέαζαν σε μεγάλο κομμάτι την ταχύτητα εκτέλεσης του κώδικα. Για την επίλυση αυτού του προβλήματος χρησιμοποιήθηκαν διάφορες λύσεις. Η πιο συχνή λύση ήταν αυτή της μετατροπής του σειριακού κώδικα σε παράλληλο κώδικα με την χρήση νημάτων. Με αυτόν τον τρόπο, γινόταν μέγιστη αξιοποίηση των πολυεπεξεργαστών και συνεπώς μειωνόταν ο χρόνος διεκπεραίωσης της συγκεκριμένης εργασίας. Ακόμα μια μέθοδος που χρησιμοποιήθηκε ήταν η επεξεργασία παρτίδων (batch processing) και η χρήση ενδιάμεσων αποτελεσμάτων. Η μέθοδος αυτή εφαρμόζοταν κυρίως για εργασίες που ήταν αρκετά μεγάλες και θέλαμε να μην χάσουμε όλα τα ενδιάμεσα αποτελέσματα αν κάτι δεν πήγαινε όπως αναμενόταν. Τέλος χρησιμοποιήθηκε και το Google Colab, ένα προϊόν της Google που επιτρέπει σε οποιοδήποτε να γράψει κώδικα σε Python και να τον εκτελέσει στον browser με CPU, GPU ή TPU.

1.5 Περίγραμμα περιεχομένων

Η παρούσα διπλωματική εργασία αποτελείται συνολικά από πέντε κεφάλαια. Αυτά ονομάζονται ως εξής: «Εισαγωγή», «Υπόβαθρο και Σχετική δουλειά», «Μεθοδολογία», «Πειράματα και Αποτελέσματα» και «Συμπεράσματα».

Στο πρώτο κεφάλαιο «Εισαγωγή» δίνεται μια γενική εικόνα γύρω από την κατάσταση που επικρατεί με επίκεντρο την πανδημία. Αναφέρονται, επίσης, κάποια βήματα τα οποία είχαν σημαντικό ρόλο στην έρευνα καθώς και ορισμένα προβλήματα τα οποία έπρεπε να αντιμετωπίσουμε. Σε αυτό το κεφάλαιο υπάρχει και το υποκεφάλαιο με όνομα «Ορολογία» που έχει την χρήση λεξικού εξηγώντας συνοπτικά λέξεις και εκφράσεις που χρησιμοποιούνται και έχουν καθοριστική σημασία για την κατανόηση της έρευνας.

Το δεύτερο κεφάλαιο με τίτλο «Υπόβαθρο και Σχετική δουλειά» έχει ως στόχο να υποδείξει το προκαταρκτικό στάδιο της διπλωματικής εργασίας που αποτελείται από διάφορες τεχνικές. Αναφέρονται και κάποιες άλλες δουλείες και άρθρα που σχετίζονται με θέματα παραπλήσια του δικού μας.

Στο κεφάλαιο «Μεθοδολογία», σκιαγραφείται σε υψηλό επίπεδο η εργασία που έχει γίνει και έτσι βλέπουμε την ακολουθία των βημάτων για την διεκπεραίωση του τελικού στόχου. Σε αυτό το κεφάλαιο, αναφέρονται και κάποιες λεπτομέρειες που έχουν να κάνουν με την υλοποίηση όπως είναι η προγραμματιστική γλώσσα, τα εργαλεία και οι διάφορες ενότητες που χρησιμοποιήθηκαν.

Όσο αφορά το κεφάλαιο «Πειράματα και Αποτελέσματα», αυτό είναι ο πυρήνας όλης της έρευνας. Περιγράφονται όλες οι λεπτομέρειες από την αρχή μέχρι το τέλος για κάθε στάδιο. Είναι το κεφάλαιο το οποίο μιλά για την συλλογή των δεδομένων και το πείραμα που έγινε. Τέλος γίνεται αναφορά σε αποτελέσματα από τις κοινότητες που βρήκαμε, δηλαδή τα θέματα που συζητιόνταν ανά κοινότητα.

Το τελευταίο κεφάλαιο με τίτλο «Συμπεράσματα», είναι αυτό που παίρνει σαν βάση τα αποτελέσματα για κάθε κοινότητα που βρέθηκαν στο προηγούμενο κεφάλαιο έτσι ώστε να στηριχτούν επάνω τους τα συμπεράσματα. Τα συμπεράσματα, σε αντίθεση με τα αποτελέσματα, μιλούν για τις σχέσεις των διπόλων. Σε αυτό το κεφάλαιο, αναφέρουμε και τι θα μπορούσε να αλλάξει σε μελλοντική έρευνα για να βελτιώσουμε ορισμένα κομμάτια.

Κεφάλαιο 2

Υπόβαθρο και Σχετική δουλειά

2.1 Υπόβαθρο (Προκαταρκτικό Στάδιο)	7
2.1.1 Τεχνική εξερεύνησης των δεδομένων (Exploratory Data Analysis)	8
2.1.2 BERT	8
2.1.3 Τεχνική ομαδοποίησης (Clustering)	9
2.2 Αναγνώριση και Μέτρηση της Πόλωσης	10

2.1 Υπόβαθρο (Προκαταρκτικό Στάδιο)

Έχοντας υπόψη όσα ειπώθηκαν στα πλαίσια του κεφαλαίου της εισαγωγής, σκοπός αυτής της έρευνας είναι να μελετήσουμε τα δεδομένα στο διαδίκτυο για μια συγκεκριμένη χρονική περίοδο της πανδημίας. Όλο αυτό πηγάζει από το γεγονός ότι θεωρούμε, και μπορούμε να πούμε με βεβαιότητα, ότι τα μέσα κοινωνικής δικτύωσης είναι μια μικρογραφία του πραγματικού κόσμου. Με λίγα λόγια, οι άνθρωποι συμμερίζονται την άποψη ότι αυτές οι πλατφόρμες είναι ένα μέσο επικοινωνίας. Αυτό σημαίνει ότι νιώθουν άνετα να εκφράσουν τη γνώμη τους σε άλλους ανθρώπους και να αρχίσουν κάποιο διάλογο ή να έχουν κάποια διαφωνία. Άρα, αυτά τα ακατέργαστα δεδομένα μπορούν να μετατραπούν σε νέα γνώση για να κατανοήσουμε πως ένιωθαν οι άνθρωποι την περίοδο της πανδημίας αλλά και ποια ήταν τα θέματα τα οποία τους απασχολούσαν περισσότερο. Για την υλοποίηση του στόχου χρειαζόμαστε κάποιες τεχνικές όπως είναι η τεχνική της εξερεύνησης των δεδομένων (exploratory data analysis), η χρήση του framework BERT και την τεχνική της ομαδοποίησης (clustering).

2.1.1 Τεχνική εξερεύνησης των δεδομένων (Exploratory Data Analysis)

Η εξερεύνηση των δεδομένων είναι μια τεχνική η οποία μας δίνει την δυνατότητα να καταλάβουμε και να εξασφαλίσουμε μια πρώτη εικόνα για τα δεδομένα που πήραμε. Ανάλογα με τα δεδομένα που συλλέγει κάποιος μπορεί να διαφοροποιήσει τα βήματα που θα ακολουθήσει για να διεκπεραιώσει αυτό το κομμάτι δουλειάς. Για την συγκεκριμένη έρευνα, τρία ήταν τα βασικά βήματα που κάναμε. Το πρώτο βήμα είχε να κάνει με την επιλογή των δεδομένων. Αφού γνωρίζουμε ότι το σύνολο των δεδομένων είναι σε μορφή πίνακα όπου κάθε στήλη αντιπροσωπεύει μια κατηγορία δεδομένων και κάθε γραμμή έχει να κάνει με ένα αντικείμενο, διαλέγουμε τις στήλες που έχουν σημαντικές πληροφορίες για την έρευνα. Με λίγα λόγια, διερευνώντας τις διάφορες κατηγορίες πληροφοριών μπορούμε να γνωρίσουμε τα δεδομένα καλύτερα, με απώτερο σκοπό να δούμε τι έχουμε συλλέξει, τι θα μπορούσαμε να μελετήσουμε σε μελλοντική φάση και γενικά να κατασκευάσουμε ένα πλάνο για την καλύτερη αξιοποίησή τους. Το επόμενο στάδιο ήταν να οργανώσουμε και να επεξεργαστούμε τα δεδομένα. Άρα παίρνουμε τις στήλες τις οποίες χρειαζόμαστε και διορθώνουμε ή αφαιρούμε «ανωμαλίες» στα δεδομένα. Με τον όρο «ανωμαλίες» εννοούμε τιμές που είναι κενές (null values), λανθασμένες ή αντιπροσωπεύουν ακραίες τιμές (outliers). Μέρος αυτού του βήματος είναι τόσο η επεξεργασία κειμένου όσο και η μορφοποίηση μιας κοινής διάταξης όταν για παράδειγμα η στήλη αντιπροσωπεύει κάποια ημερομηνία ή ένα ηλεκτρονικό σύνδεσμο (λόγω URL shortener κάποιοι σύνδεσμοι μπορεί να είναι αυτούσιοι ή να έχουν περάσει από διαδικασία shortening). Τελευταίο βήμα είναι αυτό της συλλογής κάποιων πρώτων στατιστικών. Έτσι, μπορούμε να φτιάξουμε γραφικές παραστάσεις για να αυξήσουμε την διαφάνεια των δεδομένων.

2.1.2 BERT

Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα γλωσσικό μοντέλο το οποίο δημιουργήθηκε από ερευνητές του τμήματος Google AI Language. Χρησιμοποιεί τεχνολογία βαθιάς μάθησης Transformer για θέματα γύρω από το πεδίο της επεξεργασίας φυσικής γλώσσας (natural language processing) [5]. Όταν είχε δημοσιευτεί είχε γίνει μεγάλος σάλος καθώς είχε φοβερά καλές επιδόσεις σε

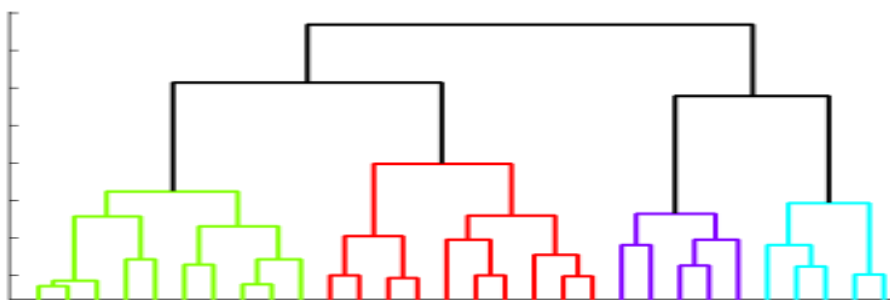
διάφορες εφαρμογές επεξεργασίας της φυσικής γλώσσας όπως είναι η απάντηση ερωτήσεων, η απόδοση συμπερασμάτων, και άλλα. Λόγω της διαφορετικής γλώσσας στο γραπτό λόγο, το BERT είχε εκπαιδευτεί με κείμενα από το Wikipedia με σκοπό να εξάγει το νόημα από μια πρόταση, μια παράγραφο και γενικά οποιοδήποτε κείμενο δίνεται σε αυτό ως είσοδο. Η καινοτομία που εμφανίζεται στο BERT είναι ότι γίνεται εκπαίδευση αμφίδρομων αναπαραστάσεων του κειμένου σε όλα τα επίπεδα του μοντέλου. Κάτι τελείως διαφορετικό από τις άλλες προσπάθειες που δοθέντος μιας πρότασης σάρωναν είτε μόνο προς τη μια κατεύθυνση (αριστερά προς δεξιά ή το αντίθετο) ή συνδύαζαν τις δύο κατευθύνσεις.

Όσο αφορά κάποιες λεπτομέρειες με τον τρόπο που το BERT μοντέλο δουλεύει είναι ότι χρησιμοποιεί δύο βήματα: το βήμα της προ-εκπαίδευσης και το βήμα Fine-Tuning. Όσο αφορά το πρώτο βήμα, χρησιμοποιείται μη επιβλεπόμενη μάθηση σε άγνωστα δεδομένα για να μπορέσει το μοντέλο να δημιουργήσει μια αναπαράσταση του κειμένου. Όσο αφορά το βήμα Fine-Tuning, χρησιμοποιείται επιβλεπόμενη μάθηση για να γίνει μια συγκεκριμένη διεργασία όπως είναι η απάντηση ερωτήσεων. Δηλαδή χρησιμοποιείται το ήδη υπάρχον μοντέλο και με μικρές αλλαγές, για παράδειγμα πρόσθεση ενός ακόμα επιπέδου εξόδου, μπορεί να πετυχαίνει εξίσου καλά αποτελέσματα και για άλλες διεργασίες που επιθυμούμε. Το βήμα της προ-εκπαίδευσης όμως χωρίζεται και αυτό με την σειρά του σε δύο άλλα βήματα που ονομάζονται Masked Language Model (MLM) και Next Sentence Prediction (NSP). Το MLM βήμα είναι υπεύθυνο για την αμφίδρομη εκπαίδευση του κειμένου. Συγκεκριμένα, το κείμενο χωρίζεται σε tokens και ένα μικρό υποσύνολο (περίπου το 15%) «κρύβονται» με σκοπό το μοντέλο να τα προβλέψει σωστά. Από την άλλη, το NSP βήμα δέχεται ζευγάρια από προτάσεις και προβλέπει εάν η επόμενη πρόταση είναι σωστή ή όχι. Έτσι πετυχαίνει την εκμάθηση των σχέσεων μεταξύ των προτάσεων. Το πιο ενδιαφέρον όμως είναι ότι το BERT εκτελεί ταυτόχρονα τα βήματα MLM και NSP.

2.1.3 Τεχνική ομαδοποίησης (Clustering)

Εξαιρετικής σημασίας είναι και η τεχνική της ομαδοποίησης [14]. Μέσω της συγκεκριμένης τεχνικής, υπήρχε η δυνατότητα να χωρίσουμε το σύνολο των δεδομένων σε πιο μικρές διακριτές ομάδες δεδομένων με τέτοιο τρόπο που το κάθε αντικείμενο να

υπάρχει μόνο σε μια ομάδα και να είναι πιο «κοντά» (η έννοια της απόστασης υποδηλώνει ομοιότητα, δηλαδή όμοια αντικείμενα βρίσκονται κοντά το ένα στο άλλο) σε αντικείμενα της ίδιας ομάδας παρά σε αντικείμενα άλλης ομάδας. Στην τεχνική της ομαδοποίησης πρέπει να ορίσουμε πως θα μετρούμε την ομοιότητα μεταξύ των δεδομένων ανάλογα με τον τύπο που έχουν. Για παράδειγμα, αν τα δεδομένα εκφράζονται ως διανύσματα τότε πρέπει να χρησιμοποιηθεί Cosine Distance, αν εκφράζονται σαν σύνολα τότε χρησιμοποιούμε Jaccard Distance ενώ αν είναι σημεία σε χώρο με συντεταγμένες χρησιμοποιούμε Euclidean Distance. Επίσης, μια άλλης παράμετρος που έπρεπε να ληφθεί υπόψιν ήταν η μέθοδος ομαδοποίησης. Για την συγκεκριμένη έρευνα χρησιμοποιήσαμε την Agglomerative, δηλαδή από κάτω προς τα πάνω. Άρα, στο πρώτο βήμα το κάθε αντικείμενο ήταν και μια ξεχωριστή ομάδα (αρχικός αριθμός ομάδων ήταν αυτός των αντικειμένων) και σε κάθε επόμενο βήμα όμοια αντικείμενα συγχωνεύονταν σε μια ομάδα. Η συνθήκη τερματισμού ενεργοποιείται όταν δεν γίνονται πλέον άλλες αλλαγές, δηλαδή όταν κανένα αντικείμενο δεν αλλάζει ομάδα. Στα πλαίσια της έρευνας, η συγκεκριμένη τεχνική εφαρμόστηκε δύο φορές. Η πρώτη φορά είχε να κάνει με την κατάτμηση του γράφου σε υποομάδες κόμβων και η δεύτερη φορά έγινε όταν έπρεπε να χωρίσουμε τις δημοσιεύσεις που είχαμε συλλέξει σε διάφορες ομάδες δημοσιεύσεων για να βρούμε το θέμα που συζητούσαν, το οποίο ήταν και το κοινό στοιχείο που τις ωθούσε στο να δημιουργήσουν μια ομάδα.



Σχήμα 2.1: Agglomerative Clustering

2.2 Αναγνώριση και Μέτρηση της Πόλωσης

Η ανάλυση των διαμαχών και των αμφισβητήσεων στα μέσα κοινωνικής δικτύωσης έχει προσελκύσει μεγάλη προσοχή και ένας σημαντικός αριθμός άρθρων έχει δώσει

ενδιαφέρουσες μελέτες. Έτσι κάποια από αυτά με βοήθησαν στο να κατανοήσω σημαντικές έννοιες όπως είναι η αναγνώριση και η ταυτοποίηση της πόλωσης. Από την άλλη, υπήρξαν και άρθρα χρήσιμα για την μέτρηση και ποσοτικοποίηση της πόλωσης.

Το άρθρο «Measuring Political Polarization: Twitter shows the two sides of Venezuela» [6] ήταν το πρώτο άρθρο που προσέγγισα καθώς ήθελα να πάρω την κεντρική ιδέα και έναν τρόπο που θα μπορούσε να πορευτεί κάποιος με παρόμοιο πλαίσιο. Το συγκεκριμένο άρθρο με βοήθησε στο να καταλάβω σημαντικές έννοιες όπως αυτή της πόλωσης, να δω την μεθοδολογία που ακολουθήθηκε αλλά και έναν τρόπο με τον οποίο μετριέται η πόλωση σε κοινωνικά δίκτυα. Παρόλ' αυτά, στο άρθρο υπήρχε η ανάγκη να μετρηθεί η ισχύς μιας άποψης στο κοινωνικό δίκτυο και έτσι να χωριστούν οι άνθρωποι που συμμετείχαν στο δίκτυο σε αυτούς που επηρεάζουν με την άποψη τους και αυτούς που επηρεάζονται, κάτι το οποίο δεν συμβαίνει στην δική μας έρευνα.

Σχετικά με την ταυτοποίηση πόλωσης, μελετήθηκε και το άρθρο «POLAR: A Holistic Framework for the Modelling of Polarization and Identification of Polarizing Topics in News Media» [7]. Στο άρθρο αυτό παρουσιάζεται ένα framework που επιτυγχάνει την ταυτοποίηση της πόλωσης και την εύρεση των θεμάτων που μιλούνται σε διάφορα άρθρα ειδήσεων. Για να γίνει αυτό εφικτό, το POLAR καινοτομεί στον τρόπο μοντελοποίησης της πόλωσης και χρησιμοποιεί μεθόδους επεξεργασίας φυσικής γλώσσας τελευταίας τεχνολογίας. Αποτελεί έναν σπουδαίο οδηγό για την δική μας έρευνα με την διαφορά ότι η πηγή των δεδομένων είναι κατά κύριο λόγο μια συλλογή από άρθρα ειδήσεων. Σε αντίθεση με την έρευνα αυτή που ο στόχος ήταν η μελέτη του αντίκτυπου της πανδημίας του κορονοϊού στα κοινωνικά δίκτυα, το POLAR αναζητούσε πηγές δεδομένων που να είναι πιο αξιόπιστες. Είναι γνωστό ότι τα άρθρα ειδήσεων είναι πιο μεγάλα, πιο περιγραφικά και γενικά εμπνέουν εμπιστοσύνη προς το περιεχόμενό τους σε σύγκριση με τις δημοσιεύσεις στα κοινωνικά δίκτυα όπου υπάρχει πολύς θόρυβος και πολλές φορές εντοπίζονται φαινόμενα διάδοσης ψευδών πληροφοριών.

Σε αντίθεση με τα παραπάνω, τα άρθρα «Quantifying Controversy on Social Media» [8] και «A Measure of Polarization on Social Media Networks Based on Community Boundaries» [9] ήταν βοηθητικά σε θέματα που αφορούσαν την διαδικασία της ποσοτικοποίησης της πόλωσης. Όσο αφορά το πρώτο άρθρο, αφού γίνει συλλογή

διαφόρων δεδομένων από την πλατφόρμα του Twitter για ένα οποιοδήποτε διάστημα, δημιουργείται ένα γενικό framework με στόχο τον εντοπισμό της αντιπαράθεσης (controversy) σε θέματα που δεν ανήκουν σε κάποιο συγκεκριμένο τομέα (όπως οικονομικό, πολιτικό και άλλα). Στο τέλος, δίνονται πολλές εναλλακτικές μέθοδοι για ποσοτικοποίηση της πόλωση σε ένα γράφο. Από την άλλη, στο δεύτερο άρθρο παρουσιάζεται ένας συγκεκριμένος και αποδοτικός τρόπος για την εύρεση της πόλωσης σε γράφους που αντιπροσωπεύουν κοινωνικά δίκτυα. Πιο συγκεκριμένα, σε αυτό το άρθρο γίνεται μια σύγκριση μεταξύ των δικτύων που εμφανίζουν πόλωση στο πλαίσιο τους και αυτών που δεν εμφανίζουν κάποια πόλωση με σκοπό να προταθεί μια καινοτόμα τεχνική που να είναι ακριβής στο να ξεχωρίζει αυτές τις περιπτώσεις. Η τεχνική ονομάζεται «Measuring Polarization on Community Boundaries» και χρησιμοποιείται και στην δική μας έρευνα γι' αυτό και αναπτύσσεται στο κεφάλαιο 3 «Μεθοδολογία».

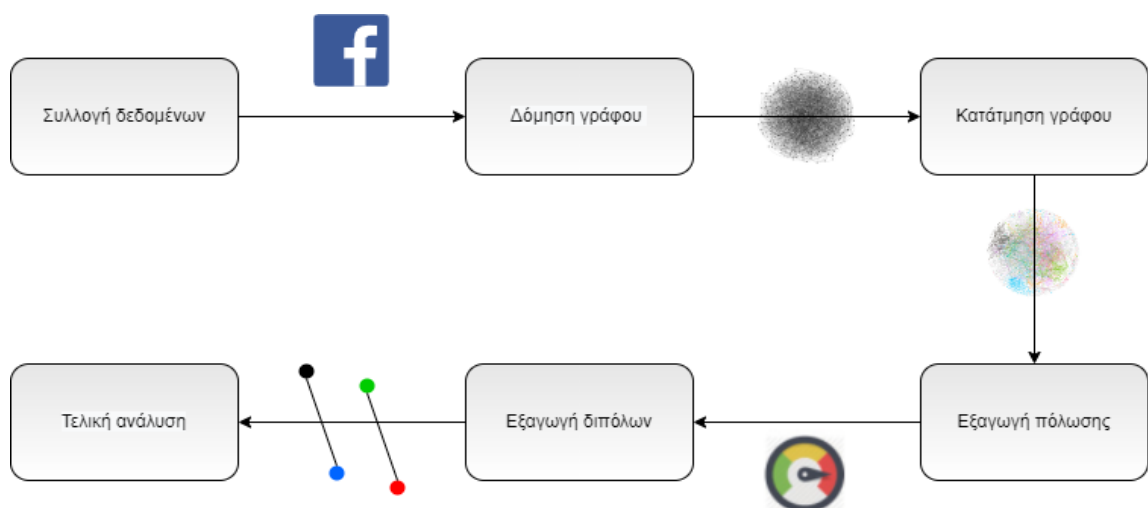
Τέλος, ένα ενδιαφέρον άρθρο για την πανδημία του κορονοϊού ήταν και το «As COVID-19 Emerged in U.S., Facebook Posts About It Appeared in a Wide Range of Public Pages, Groups» [10] που εκδόθηκε από το ερευνητικό κέντρο Pew. Σκοπός των ερευνητών ήταν να πάρουν δεδομένα που σχετίζονται με την πανδημία από τις 01/03/2020 (εμφάνιση δύο κρουσμάτων κορονοϊού στις Ηνωμένες Πολιτείες Αμερικής) μέχρι 31/03/2020 (περισσότερα από τέσσερις χιλιάδες κρούσματα και εντολές για να παραμείνουν οι κάτοικοι στα σπίτια τους στις περισσότερες πολιτείες) χρησιμοποιώντας την σελίδα Crowdtangle API. Έγινε ανάλυση σε επίπεδο περιεχομένου έτσι ώστε να εξακριβωθεί το θέμα το οποίο συζητούσαν οι άνθρωποι στις διάφορες σελίδες και γκρουπ, αν υπήρχε κάποια γεωγραφική τοποθεσία ως επίκεντρο καθώς και ποιος τύπος πηγής ήταν αυτός που οι άνθρωποι προτιμούσαν (όπως είναι σελίδες μέσω μαζικής ενημέρωσης, blogs, σελίδες κέντρων υγείας και άλλα). Όσο αφορά την δική μας έρευνα, ναι μεν γίνεται ανάλυση κειμένου για να βγάλουμε καινούρια πληροφορία όπως τα πιο πολυσυζητημένα θέματα, η εύρεση συναισθήματος και άλλα, όμως γίνεται με βάση τον μη κατευθυνόμενο γράφο και των κοινοτήτων που ομαδοποιούνται σε αυτόν.

Κεφάλαιο 3

Μεθοδολογία

3.1 Αρχιτεκτονικό διάγραμμα.....	13
3.2 Επεξήγηση των σταδίων που απεικονίζονται στο διάγραμμα	14
3.2.1 Συλλογή των Δεδομένων	14
3.2.2 Δόμηση του Γράφου	14
3.2.3 Κατάτμηση του Γράφου	15
3.2.4 Εξαγωγή της Πόλωσης	18
3.2.5 Εξαγωγή των Διπόλων.....	19
3.2.6 Τελική Ανάλυση	20
3.3 Λεπτομέρειες υλοποίησης	21

3.1 Αρχιτεκτονικό διάγραμμα



Σχήμα 3.1: Αρχιτεκτονικό διάγραμμα με τα στάδια εξεύρεσης πόλωσης στα κοινωνικά δίκτυα.

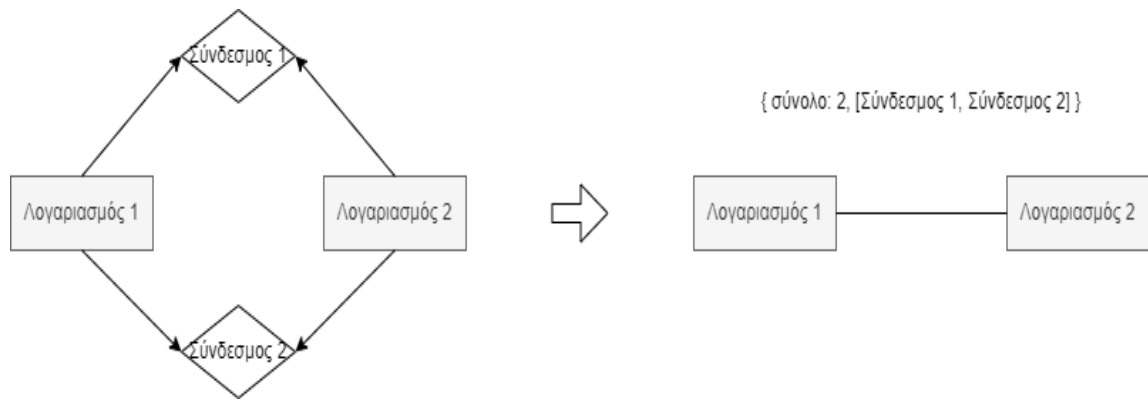
3.2 Επεξήγηση των σταδίων που απεικονίζονται στο διάγραμμα

3.2.1 Συλλογή των Δεδομένων

Ο σκοπός σε αυτό το στάδιο ήταν η συλλογή των δεδομένων που θα ήταν η βάση για να μετρήσουμε την πόλωση στα κοινωνικά δίκτυα. Με άλλα λόγια, η πηγή των δεδομένων θα μπορούσε να ήταν οποιαδήποτε. Έγινε έρευνα στα δεδομένα που προέρχονταν από τις πλατφόρμες του Facebook και του Twitter από τις ιστοσελίδες Crowdtangle API και Twitter API αντίστοιχα, όμως η διπλωματική εργασία κτίστηκε στα δεδομένα του Facebook καθώς υπήρχε μεγαλύτερος όγκος δεδομένων. Περισσότερες πληροφορίες σχετικά με το σύνολο των δεδομένων που συλλέχτηκαν δίνονται στο υποκεφάλαιο 4.1.

3.2.2 Δόμηση του Γράφου

Στο συγκεκριμένο στάδιο θέλαμε να χρησιμοποιήσουμε μια δομή δεδομένων που να μπορεί να αντικατοπτρίζει μια ολοκληρωμένη εικόνα των δεδομένων που συλλέξαμε. Η πρώτη προσπάθεια για να το πετύχουμε ήταν με την χρήση κατευθυνόμενου γράφου. Ο γράφος αυτός είχε κόμβους που υποδήλωναν είτε κάποιο χρήστη ή κάποιο σύνδεσμο και οι ακμές έδειχναν την σχέση μεταξύ της οντότητας και του URL που διαμοιραζόταν. Τα αποτελέσματα δεν ήταν ικανοποιητικά και έτσι μας ωθήσαν στην κατασκευή ενός μη κατευθυνόμενου γράφου. Τα νέα δεδομένα είναι ως εξής: ένας κόμβος αντιπροσώπευε ένα λογαριασμό στο Facebook (άρα θα μπορούσε να ήταν μια σελίδα ή ένα γκρουπ) και υπήρχε μια ακμή μεταξύ δύο κόμβων εάν και εφόσον αυτοί οι δύο κόμβοι «έδειχναν» σε τουλάχιστον έναν κοινό σύνδεσμο. Έτσι, για κάθε κόμβο ήταν γνωστοί οι σύνδεσμοι που μοιραζόταν με άλλους κόμβους καθώς και ποιοι κόμβοι ήταν αυτοί. Από την πλευρά των ακμών, κάθε μια ακμή περιείχε τις εξής πληροφορίες: ποιοι είναι οι κόμβοι στους οποίους ανήκει καθώς και πόσοι και ποιοι είναι οι κοινόι σύνδεσμοι που αντιπροσωπεύει.

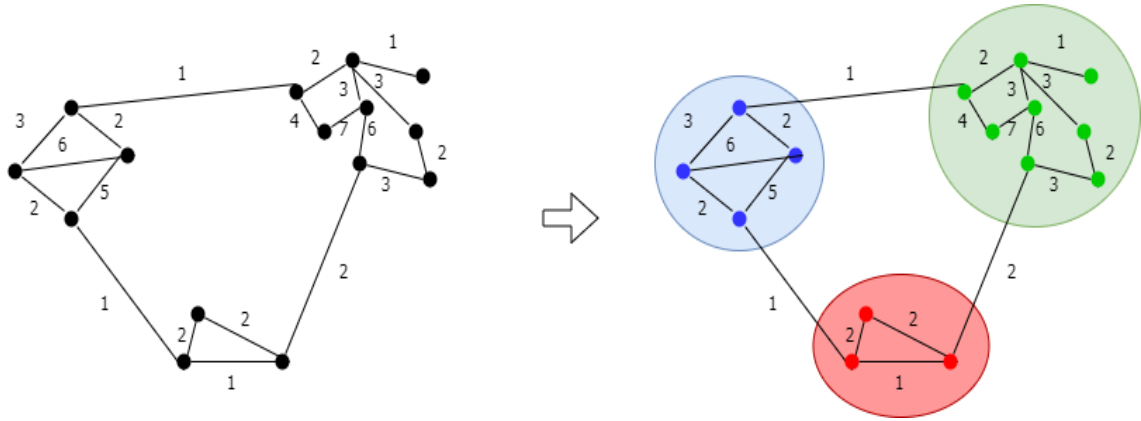


Σχήμα 3.2: Απεικόνιση της μετατροπής του κατευθυνόμενου γράφου σε μη κατευθυνόμενο γράφο.

3.2.3 Κατάτμηση του Γράφου

Στο τρίτο στάδιο το ζητούμενο ήταν ο διαχωρισμός του γράφου σε μικρότερα διακριτά τμήματα που θα αντιπροσωπεύουν κοινότητες. Στην ουσία, αυτός ήταν και ένας άλλος λόγος που επιλέχτηκε ο γράφος για να είναι η δομή δεδομένων στο προηγούμενο βήμα, αφού ξέρουμε ότι μια ιδιότητα των γράφων που αναπαριστούν κοινωνικά δίκτυα είναι η τοπικότητα (locality). Η τοπικότητα αναφέρεται στο γεγονός ότι οι ακμές των γράφων τείνουν να ομαδοποιούνται σε κοινότητες.

Έτσι, ο γράφος δόθηκε ως είσοδος σε ένα αλγόριθμο ο οποίος χρησιμοποιούσε διάφορες ευρετικές μεθόδους για να βρει την ιδανική κατάτμηση και η έξοδος του αλγορίθμου έδινε τα διάφορα διακριτά τμήματα τα οποία απαρτίζουν τον γράφο και μεγιστοποιούν το modularity. Διαισθητικά, όταν λέμε ιδανική κατάτμηση εννοούμε ότι θέλαμε η κατάτμηση του γράφου να έχει ψηλό δείκτη modularity για να υπάρχουν πυκνές συνδέσεις με ψηλή ένταση μεταξύ των κόμβων εντός των κοινοτήτων, αλλά αραιές συνδέσεις με χαμηλή ένταση μεταξύ των κόμβων σε διαφορετικές κοινότητες. Με πιο απλά λόγια, κόμβοι (σελίδες ή γκρουπ) που ενώνονται με πολλές ακμές (πυκνές συνδέσεις άρα πολλοί σύνδεσμοι) με ψηλή ένταση (κοινοί σύνδεσμοι) πρέπει να ανήκουν στην ίδια κοινότητα. Κάτι που πρέπει να σημειωθεί είναι ότι κάθε κόμβος μετά την κατάτμηση του γράφου ανήκει σε μια και μόνο μια κοινότητα και δεν υπάρχει καμία πιθανότητα κάποιος κόμβος να ανήκει σε περισσότερες από μια κοινότητες (disjoint and not overlapping communities).



Σχήμα 3.3: Απεικόνιση της κατάτμησης του γράφου.

Modularity είναι η μετρική η οποία δείχνει πόσο καλά σπάσαμε ένα γράφο σε υπογράφους και η τιμή αυτή κυμαίνεται στο διάστημα $[-1/2, 1]$. Για την σωστή εκτίμηση της μετρικής modularity, ο αλγόριθμος λαμβάνει υπόψη διάφορες πληροφορίες των ακμών που είναι εντός των υπογράφων καθώς και των ακμών που ενώνουν τους υπογράφους για να σχηματίσουν τον αρχικό γράφο. Κάποιες από τις κύριες πληροφορίες των ακμών που λαμβάνονται υπόψη είναι το σύνολο τους και η ένταση (ή βάρος) της κάθε ακμής εάν υπάρχει. Συγκεκριμένα, εφόσον έχουμε έναν γράφο με ακμές που έχουν βάρη, ο τύπος (1) δίνει την μετρική modularity του γράφου όπου:

- A_{ij} αντιπροσωπεύει το βάρος της ακμής που ενώνει τους κόμβους i και j ,
- k_i, k_j είναι το άθροισμα των βαρών των ακμών που συνδέονται στους κόμβους i και j αντίστοιχα,
- m είναι το άθροισμα όλων των βαρών των ακμών του γράφου,
- c_i, c_j είναι οι κοινότητες των κόμβων (υπογράφοι),
- δ είναι το Kronecker delta function ($\delta(x,y)=1$ αν $x=y$, αλλιώς 0).

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \quad \text{Τύπος (1)}$$

Για την υλοποίηση αυτού του σταδίου έγινε χρήση του αλγόριθμου Louvain. Είναι ένας αλγόριθμος μη επιβλεπόμενης μάθησης (δηλαδή παίρνει τον γράφο ως είσοδο

αλλά δεν ξέρει ποιες και πόσες είναι οι διάφορες κοινότητες που πρέπει να δώσει σαν έξοδο) και βρίσκει τις κοινότητες του γράφου. Αυτό γίνεται με ιεραρχικό τρόπο σε μη ορισμένο αριθμό βημάτων μέχρις ότου να βρει την μέγιστη τιμή modularity και να μην υπάρχουν άλλες αλλαγές που μπορούν να συμβούν.

Σε κάθε βήμα του αλγορίθμου υπάρχουν δύο φάσεις. Η πρώτη φάση ονομάζεται Modularity Optimization και έχει να κάνει με την εύρεση κοινοτήτων που έχουν τη μέγιστη τιμή modularity. Συγκεκριμένα, στην αρχή ο κάθε κόμβος ανήκει στην κοινότητα του εαυτού του (άρα αρχικά ο αριθμός των κοινοτήτων είναι ίσος με τον αριθμό των κόμβων), οπότε στην συνέχεια παίρνουμε τυχαία ένα-ένα κόμβο, έστω i , και τον αφαιρούμαι από την δική του κοινότητα και τον εντάσσουμε στην κοινότητα κάθε γειτονικού κόμβου, έστω αυτή ονομάζεται C . Αυτή η αλλαγή στην τιμή της μετρικής modularity δίνεται από τον τύπο (2) όπου:

- Σ_{in} είναι το άθροισμα των βαρών των ακμών της κοινότητας C ,
- Σ_{tot} είναι το άθροισμα των βαρών όλων των ακμών που «δείχνουν» στους κόμβους της κοινότητας C ,
- k_i είναι το άθροισμα των βαρών όλων των ακμών που «δείχνουν» στον κόμβο i ,
- $k_{i,in}$ είναι το άθροισμα των βαρών των ακμών από τον κόμβο i σε κάθε κόμβο στην κοινότητα C ,
- m είναι το άθροισμα όλων των βαρών των ακμών του γράφου.

Αφού υπολογιστεί αυτή η μεταβολή του modularity για όλες τις κοινότητες που προσπαθεί ένας κόμβος i να συνδεθεί, ο κόμβος i τοποθετείται στην κοινότητα που μεγιστοποιεί το modularity. Αυτή η διαδικασία γίνεται για όλους τους κόμβους και έπειτα είναι σειρά της δεύτερης φάσης για να έρθει στο προσκήνιο.

$$\Delta Q = \left[\frac{\Sigma_{in} + 2k_{i,in}}{2m} - \left(\frac{\Sigma_{tot} + k_i}{2m} \right)^2 \right] - \left[\frac{\Sigma_{in}}{2m} - \left(\frac{\Sigma_{tot}}{2m} \right)^2 - \left(\frac{k_i}{2m} \right)^2 \right] \quad \text{Τύπος (2)}$$

Η δεύτερη φάση ονομάζεται Community Aggregation. Σε αυτήν τη φάση η κάθε κοινότητα μετατρέπεται σε ένα νέο κόμβο που θα το αντιπροσωπεύει. Έτσι, κάθε καινούριος κόμβος έχει έναν ακέραιο αριθμό που συμβολίζει το σύνολο των ακμών που

είχαν μεταξύ τους οι κόμβοι της συγκεκριμένης κοινότητας (πριν μετατραπεί σε κόμβο) και κάθε ακμή μεταξύ των νέων κόμβων έχει έναν ακέραιο αριθμό που συμβολίζει το σύνολο των ακμών που είχαν μεταξύ τους οι δύο κοινότητες (πριν μετατραπούν σε κόμβους). Συμπερασματικά, είναι ένας άπληστος αλγόριθμος και ο χρόνος που χρειάζεται για να τρέξει είναι της τάξης $O(n \cdot \log n)$, όπου n είναι ο αριθμός όλων των κόμβων του γράφου.

3.2.4 Εξαγωγή της Πόλωσης

Αφού έγινε η εύρεση των μικρότερων κοινοτήτων που απαρτίζουν τον αρχικό γράφο, παίρνουμε όλους τους συνδυασμούς των δύο με σκοπό να σχηματίζουμε καινούριους γράφους που να αποτελούνται από ακριβώς δύο κοινότητες. Αυτό μας επιτρέπει να εφαρμόσουμε Community Boundary Polarization, όπως αναφέρεται από το άρθρο «A Measure of Polarization on Social Media Networks Based on Community Boundaries» [9], και έτσι το τι παίρνουμε σαν έξοδο από αυτό το στάδιο είναι η πόλωση για κάθε δίπολο, δηλαδή για κάθε ζευγάρι κοινοτήτων.

Όπως αναφέρθηκε, η τεχνική Community Boundary Polarization είναι αυτή που χρησιμοποιείται για να μας υποδείξει σε επίπεδο γράφου την πόλωση μεταξύ δύο κοινοτήτων που απαρτίζουν τον νέο γράφο που κτίστηκε. Βασίζεται στην ιδέα ότι οι κόμβοι και των δύο κοινοτήτων πρέπει να χωριστούν σε κόμβους που βρίσκονται στο εσωτερικό και σε κόμβους που βρίσκονται στα σύνορα. Επομένως, σε μια κοινότητα u το σύνολο των κόμβων των συνόρων καθορίζεται από δύο ιδιότητες:

1. Να έχουν τουλάχιστον μια ακμή που τους συνδέει σε ένα οποιοδήποτε κόμβο που ανήκει στην άλλη κοινότητα ($d_b(u)$)
2. Να έχουν τουλάχιστον μια ακμή που τους συνδέει σε ένα κόμβο που να μην ανήκει στην ίδια κοινότητα με αυτούς αλλά δεν ενώνεται με την άλλη κοινότητα ($d_i(u)$)

Από τον πιο κάτω τύπο (3) βρίσκουμε την πόλωση για την κοινότητα u . Αν ο αριθμός αυτός είναι θετικός ($P_u > 0$) τότε υποδηλώνει ότι υπάρχει πιθανότητα πόλωσης και ανταγωνισμού αφού οι κόμβοι της κοινότητας u τείνουν να «προτιμούν» συνδέσεις εντός της κοινότητας. Σε αντίθετη περίπτωση, αν ο αριθμός είναι αρνητικός ($P_u < 0$)

δείχνει ότι δεν υπάρχει πόλωση αφού οι κόμβοι της κοινότητας u τείνουν να «προτιμούν» συνδέσεις εκτός της κοινότητας τους.

$$P_v = \frac{d_i(v)}{d_b(v) + d_i(v)} - 0.5 \quad \text{Τύπος (3)}$$

Συνεχίζοντας με τον τύπο (4), όπου το B αναφέρεται στο σύνολο των κοινοτήτων του γράφου, αθροίζουμε τα αποτελέσματα που αφορούν τις πολώσεις των δύο κοινοτήτων και διαιρώντας μπορούμε να βρούμε την πόλωση για τον γράφο. Παρομοίως, αν ο αριθμός είναι θετικός ($P > 0$) τότε υποδηλώνει ότι υπάρχει πόλωση μεταξύ των κόμβων των δύο κοινοτήτων, ενώ αν ο αριθμός είναι αρνητικός ($P < 0$) δείχνει ότι δεν υπάρχει πόλωση αλλά συμφωνία μεταξύ των δύο κοινοτήτων.

$$P = \frac{1}{|B|} \sum_{v \in B} P_v \quad \text{Τύπος (4)}$$

3.2.5 Εξαγωγή των Διπόλων

Το ζητούμενο σε αυτήν την φάση είναι η εξαγωγή των κατάλληλων δίπολων. Δηλαδή, θέτουμε κάποια κριτήρια έτσι ώστε να επιλέξουμε συνδυασμούς κοινοτήτων για να κάνουμε περαιτέρω ανάλυση σε αυτά που συζητούν. Φυσικά το πρώτο κριτήριο είναι η πόλωση. Διαλέγουμε κάποια δίπολα τα οποία να έχουν ψηλό δείκτη πόλωσης και άλλα δίπολα που έχουν χαμηλό δείκτη πόλωσης έτσι ώστε να είμαστε σε θέση να μελετήσουμε την σύγκρουση αλλά και την ομοφωνία αντίστοιχα. Σημαντικό να αναφερθεί ότι δίπολα τα οποία έχουν κάποια πόλωση σε επίπεδο γράφου τείνουν να έχουν και πόλωση της ίδιας τάξης σε επίπεδο συζήτησης. Όμως, αυτό δεν συμβαίνει πάντα και δεν πρέπει να θεωρείται δεδομένο.

Επιπρόσθετα, ένα άλλο κριτήριο ήταν ο βαθμός διάδοσης ψευδών πληροφοριών όπως αυτά ορίζονται από την ιστοσελίδα Newsguard [11]. Η ιστοσελίδα αυτή είναι στατική

και παρουσιάζει ιστοσελίδες, ταξινομημένες ανά χώρα, που διασπείρουν ψευδείς πληροφορίες σχετικά με την πανδημία όπως είναι θέματα υγείας, θεωρίες συνωμοσίας και άλλα. Εκμεταλλευόμενοι αυτοί την ιστοσελίδα, οι κοινότητες του γράφου είχαν έναν θετικό αριθμό ο οποίος ήταν ανάλογος της διάδοσης ψευδών πληροφοριών από τις ιστοσελίδες που χρησιμοποιούν. Έτσι για το συγκεκριμένο κριτήριο, θέλαμε να προσεγγίσουμε συνδυασμούς κοινοτήτων οι οποίες να έχουν μια κοινότητα με χαμηλό δείκτη διάδοσης ψευδών πληροφοριών και για την άλλη κοινότητα να ισχύει το αντίθετο.

Τέλος, κριτήριο ήταν και ο αριθμός των συνδέσεων που υπάρχουν στο εσωτερικό των δίπολων. Με λίγα λόγια, έγινε επιλογή δίπολων που να έχουν μεγάλο αριθμό συνδέσεων μεταξύ τους.

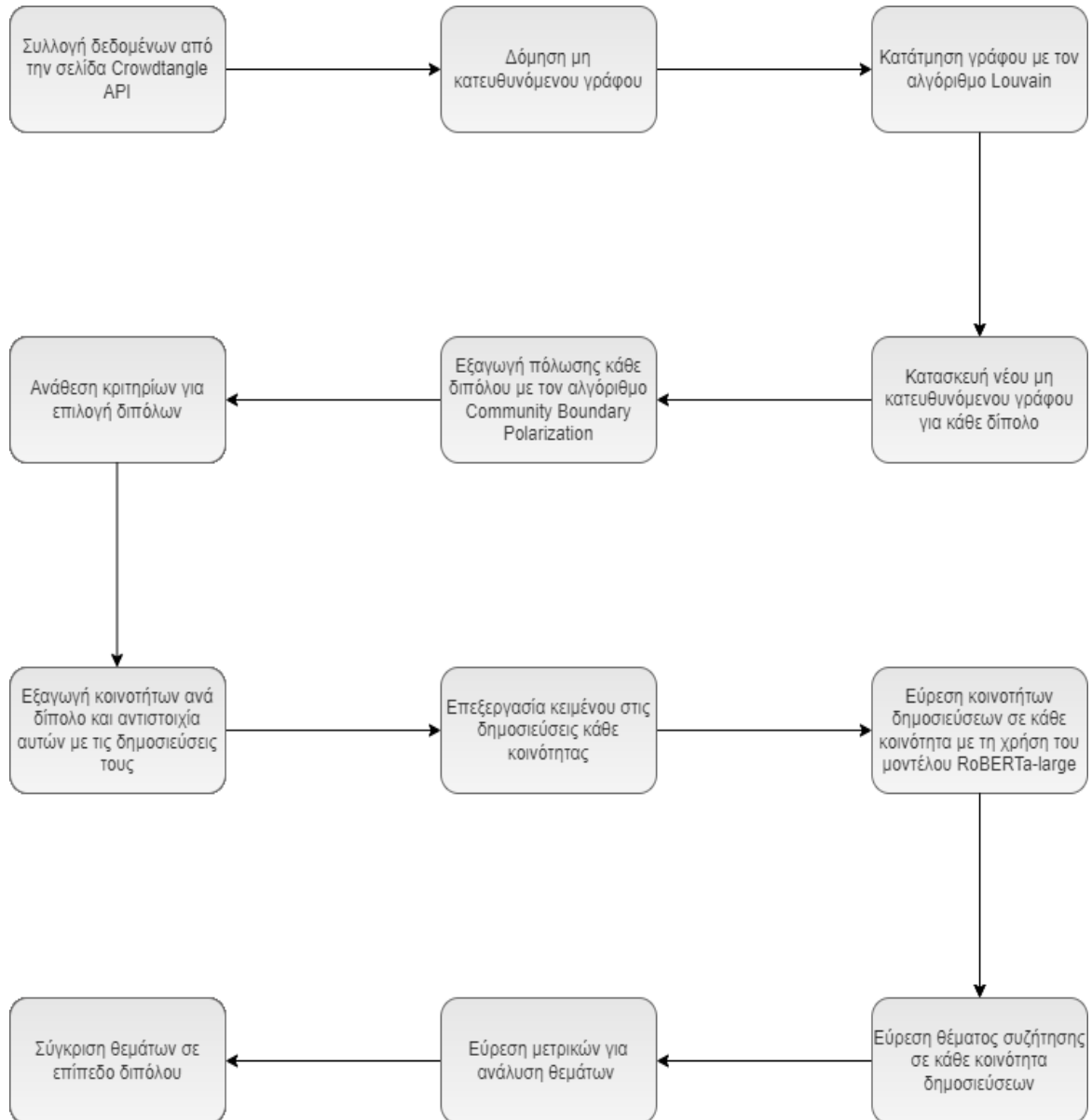
3.2.6 Τελική Ανάλυση

Όσο αφορά το βήμα της τελικής ανάλυσης, αυτό που γίνεται είναι να εφαρμόσουμε ένα πλάνο για μελέτη τόσο των δίπολων όσο και των κοινοτήτων του γράφου που τα απαρτίζουν σε επίπεδο κειμένου. Καταρχήν, παίρνουμε όλες τις κοινότητες από τα δίπολα και βρίσκουμε τα μηνύματα τα οποία μοιράζονται στην πλατφόρμα του Facebook και τα επεξεργαζόμαστε με τέτοιο τρόπο που να έχουμε μια αντιστοιχία για κάθε κοινότητα με μια λίστα με επεξεργασμένα μηνύματα.

Έπειτα, ομαδοποιούμε τα μηνύματα με απώτερο σκοπό να βρούμε κοινότητες από μηνύματα. Άρα, κάθε κοινότητα σε επίπεδο γράφου που εντοπίστηκε στο βήμα κατάτμησης του γράφου (βήμα 3) εξάγει πολλές κοινότητες μηνυμάτων σε επίπεδο κειμένου. Αυτό μας δίνει την δυνατότητα να παρατηρήσουμε πιο αφαιρετικά τα ομαδοποιημένα μηνύματα και άρα να εξάγουμε τουλάχιστον ένα θέμα από κάθε κοινότητα μηνυμάτων.

Προχωρώντας ακόμα ένα βήμα παραπέρα μπορούμε να μελετήσουμε σε σημαντικό βάθος τα θέματα που συζητιούνται τόσο ανά κοινότητα γράφου όσο και σε δίπολο παρατηρώντας κάποιες παραμέτρους. Μερικές παράμετροι που λήφθηκαν υπόψη στα

μηνύματα του κάθε θέματος ήταν τα συναισθήματα, το σθένος, η υποκειμενικότητα και η συμφωνία ή η σύγκρουση που έχουν.



Σχήμα 3.4: Αναλυτικό αρχιτεκτονικό διάγραμμα

3.3 Λεπτομέρειες υλοποίησης

Σχετικά με την υλοποίηση, η ατομική διπλωματική μου εργασία υλοποιήθηκε χρησιμοποιώντας την προγραμματιστική γλώσσα Python ενώ κάποια από τα περιβάλλοντα ανάπτυξης που χρησιμοποιήθηκαν ήταν το PyCharm και το Jupyter

Notebook. Ακόμα, σε κάθε στάδιο του πιο πάνω σκελετού έγινε χρήση διαφορετικών ενοτήτων (modules). Όπως αναφέρθηκε ήδη χρησιμοποιήθηκε και το Google Colab για κάποιες διεργασίες όπου ήταν πιο αποτελεσματικές με την χρήση GPU.

Για το κομμάτι της συλλογής των δεδομένων, υπήρχε η ανάγκη να υποδεχτούμε τα δεδομένα από το Crowdtangle API, να τα οργανώσουμε και να τα αποθηκεύσουμε. Για να διεκπεραιώσουμε αποδοτικά τις ενέργειες διαχείρισης δεδομένων χρειάστηκε η χρήση των ενοτήτων Json και Pickle.

Για το στάδιο που αφορά το κτίσιμο του γράφου καθώς και την κατάτμηση του σε τμήματα χρησιμοποιήθηκαν οι ενότητες Networkx και Community Louvain. Επίσης, για παρατήρηση του γράφου έγινε εικονική απεικόνιση χρησιμοποιώντας το εργαλείο Gephi.

Στο τελικό στάδιο που αναφέρεται στην ανάλυση κειμένου, η ενότητα BERT ήταν αυτή που βοήθησε στην ομαδοποίηση των μηνυμάτων. Σημαντικές ήταν και οι ενότητες TextBlob και Afinn για ανάλυση κειμένου. Εκμεταλλευτήκαμε επίσης και την ενότητα Matplotlib για αναπαράσταση γραφικών για τα τελικά μας συμπεράσματα.

Καθοριστικό ρόλο είχε η ενότητα Multiprocessing η οποία χρησιμοποιήθηκε σε σημεία τα οποία χρειάζονταν αρκετό χρόνο για να δώσουν κάποιο αποτέλεσμα. Έτσι με αυτή την ενότητα εκμεταλλευόμασταν την παράλληλη επεξεργασία μέσω νημάτων για να ολοκληρώσουμε μια διεργασία όσο το δυνατόν πιο γρήγορα.

Κεφάλαιο 4

Πειράματα και Αποτελέσματα

4.1 Περιγραφή του συνόλου των δεδομένων.....	23
4.2 Πείραμα	26
4.3 Αποτελέσματα.....	30
4.3.1 Κοινότητα 1	30
4.3.2 Κοινότητα 2	33
4.3.3 Κοινότητα 4	34
4.3.4 Κοινότητα 5	36
4.3.5 Κοινότητα 9	39
4.3.6 Κοινότητα 24	41
4.3.7 Κοινότητα 45	42

4.1 Περιγραφή του συνόλου των δεδομένων

Αρχικά, όσο αφορά την άντληση των δεδομένων, χρησιμοποιήθηκε η ιστοσελίδα Crowdtangle API [12]. Στην ουσία, η ιστοσελίδα αυτή δεν είναι τίποτα περισσότερο από ένα εργαλείο το οποίο αποθηκεύει περιεχόμενο που βρίσκεται στα μέσα κοινωνικής δικτύωσης, όπως είναι το Facebook, σε μια βάση δεδομένων. Με λίγα λόγια, κρατά στη βάση του δημοσιεύσεις μαζί με μεταπληροφορίες. Κάποιες από τις μεταπληροφορίες που κρατά είναι ο λογαριασμός που έκανε την εν λόγω δημοσίευση (σελίδα ή γκρουπ), το χρονικό πλαίσιο και πόσες αντιδράσεις είχε (δηλαδή υπάρχουν μετρητές για τα «likes», τα «comments» και άλλα). Εφόσον η δημοσίευση ενισχύεται με κάποιο βίντεο ή εικόνα, αποθηκεύονται επιπλέον μεταπληροφορίες όπως γίνεται στην περίπτωση ενός βίντεο που μπορούμε να πάρουμε και τον αριθμό των προβολών.

Παρόλ' αυτά, δεν παρέχονται καθόλου πληροφορίες που έχουν να κάνουν με δημογραφικά στοιχεία όπως είναι για παράδειγμα το όνομα του χρήστη που δήλωσε πως του αρέσει μια δημοσίευση ή ο τόπος που μένει. Λόγω του μεγάλου όγκου πληροφοριών που έχει, το Crowdtangle API είναι ένα βασικό εργαλείο για ερευνητές και εκδότες. Ειδικότερα, είναι χρήσιμο όταν θέλουμε να αναζητήσουμε και να αναλύσουμε περιεχόμενο που μοιάζει με ένα συγκεκριμένο θέμα ή προσπαθούμε να εντοπίσουμε διάχυση ψευδών ειδήσεων και παραπληροφόρησης στα μέσα κοινωνικής δικτύωσης.

Για τους σκοπούς της διπλωματικής εργασίας που έχει ως επίκεντρο την πανδημία του κορονοϊού διαμορφώσαμε την αναζήτηση που φαίνεται στην πιο κάτω εικόνα ανάλογα.

The image shows a three-step search interface for the Crowdtangle API.
STEP 1: Choose scope - A dropdown menu is set to 'Entire System'.
STEP 2: Filter the types of posts - Includes a search bar with the text 'Any words or phrases (e.g. Trump, Chance the Rapper)', a 'Post Type' dropdown set to 'All Posts', an 'Order By' dropdown set to 'Most Recent', a 'Limit Results' dropdown set to 'No limit', and two unchecked checkboxes for 'Branded Content Only' and 'Marked As Breaking Only'.
STEP 3: Pick your dates - A dropdown menu is set to 'Custom Date Range', with date pickers showing '04/21/2022' to '04/28/2022'.
A blue 'Fetch History' button is located at the bottom right.

Σχήμα 4.1: Στιγμιότυπο αναζήτησης δεδομένων στην σελίδα Crowdtangle API.

Πιο συγκεκριμένα, στο πεδίο «Search» τοποθετήσαμε τις λέξεις-κλειδιά της αναζήτησης κρατώντας την προεπιλεγμένη τιμή «All Posts» στο πεδίο «Post Type». Οι λέξεις-κλειδιά ήταν οι ακόλουθες: «covid», «corona», «coronavirus», «pandemic», «virus», «zeneca», «moderna», «pfizer», «antibody», «antibodies», «vaccine», «vaccines», «vax», «sars», «remdesivir», «hydroxychloroquine», «infect», «infections», «wuhan», «china», «chinese», «bill gates», «gates», «immune», «immunity», «mask», «masks», «face cover», «quarantine», «lockdown». Επίσης, στο πεδίο «Custom Date Range» τοποθετήσαμε την περίοδο αναζήτησης, δηλαδή από τις 21/01/2020 (1ο κρούσμα κορονοϊού στις Ηνωμένες Πολιτείες Αμερικής) μέχρι 21/06/2020.

Αφού έγινε το αίτημα και πήραμε τα δεδομένα σε αρχείο CSV (Comma Separated Values), είχαμε την απάντηση με τις εξής στήλες, όπως φαίνονται στον πιο κάτω πίνακα.

likeCount, shareCount, commentCount, loveCount, wowCount, hahaCount, sadCount, angryCount, thankfulCount, careCount	Δέκα μετρητές που αντιπροσωπεύουν την ποσότητα της κάθε αντίδρασης σε μια δημοσίευση.
account	Το όνομα του λογαριασμού που έχει κάνει την δημοσίευση.
type	Παίρνει τις τιμές facebook_page και facebook_group για να δείξει αν ο τύπος του συγκεκριμένου λογαριασμού είναι σελίδα ή γκρουπ.
verified	Παίρνει τις τιμές True και False για να δείξει κατά πόσο ο συγκεκριμένος λογαριασμός είναι αυτός που λέει ότι είναι για να πιστοποιηθεί η εγκυρότητα του.

post	Έχει την μορφή 'x y', όπου x και y είναι δύο μοναδικοί αριθμοί που αντιπροσωπεύουν την λογαριασμό και την δημοσίευση αντίστοιχα.
link	Σύνδεσμος που παραπέμπει στην εν λόγω δημοσίευση.
message	Η δημοσίευση αυτή καθ' αυτή (απλό κείμενο).

Πίνακας 4.2: Η ερμηνεία του περιεχομένου του CSV αρχείου ως απάντηση στην αναζήτηση των δεδομένων. Στην αριστερά πλευρά του πίνακα βρίσκονται όλες οι στήλες που υπήρχαν στο αρχείο, ενώ στην δεξιά πλευρά δίνεται η εξήγηση για το τι περιείχαν.

4.2 Πείραμα

Γι' αυτήν την έρευνα, οι κύριοι στόχοι είναι δύο: να βρεθεί τόσο η πόλωση σε επίπεδο γράφου με την χρήση διπόλων όσο και σε επίπεδο θεμάτων συζήτησης με μια σειρά τεχνικών ανάλυσης κειμένου. Συνεπώς, αφού πραγματοποιηθεί η συλλογή των δεδομένων, γίνονται οι απαραίτητες ενέργειες για να κτιστεί η βάση του πειράματος. Αρχικά, υλοποιείται ο μη κατευθυνόμενος γράφος. Όπως αναφέρθηκε και στο υποκεφάλαιο 3.2.2, ένας κόμβος αντιπροσωπεύει ένα λογαριασμό στο Facebook και υπάρχει μια ακμή μεταξύ δύο κόμβων αν και μόνο αν αυτοί οι δύο κόμβοι «δείχνουν» σε τουλάχιστον έναν κοινό σύνδεσμο. Ο γράφος αποτελείται από 10,930 κόμβους και 7,673,475 ακμές. Στην συνέχεια, με την βοήθεια του αλγόριθμου Louvain, γίνεται κατάτμηση του γράφου και αυτό έχει σαν έξοδο τις κοινότητες κόμβων που απαρτίζουν το γράφο. Πιο συγκεκριμένα, ο αλγόριθμος Louvain εντόπισε 191 διαφορετικές κοινότητες. Έπειτα, παίρνουμε όλους τους συνδυασμούς των δύο κοινοτήτων με σκοπό να σχηματίζουμε καινούριους γράφους και να εφαρμόσουμε Community Boundary Polarization για να υπολογιστεί η πόλωση για κάθε δίπολο.

Η εκκίνηση του πειράματος αρχίζει από την στιγμή που έγινε η εύρεση της πόλωσης για κάθε δίπολο. Έτσι, το επόμενο βήμα είναι η επιλογή των κατάλληλων δίπολων. Χρησιμοποιήθηκαν τα τρία κριτήρια για τα οποία έγινε αναφορά στο υποκεφάλαιο 3.2.5 αλλά συνδυάζοντάς τα με αποτέλεσμα να έχουμε δύο διαφορετικούς τρόπους επιλογής δίπολων. Με τον πρώτο τρόπο επιλέχτηκαν τέσσερα δίπολα που να έχουν μια κοινότητα με ψηλό δείκτη διάδοσης ψευδών πληροφοριών και την άλλη με χαμηλό δείκτη διάδοσης ψευδών πληροφοριών και να έχουν μεγάλο αριθμό συνδέσεων μεταξύ τους. Όπως φαίνεται στον πίνακα 4.3, οι κοινότητες 1 και 5 είχαν χαμηλό δείκτη διάδοσης ψευδών πληροφοριών, ενώ οι κοινότητες 4, 24 και 45 ψηλό. Ο δεύτερος τρόπος ήταν περίπου αντίστροφης λογικής αφού παρατηρήθηκε ότι στα δίπολα που παρήγαγε ο πρώτος τρόπος είχαν σχετικά χαμηλό αριθμό συνδέσεων. Έτσι, τα επόμενα τέσσερα δίπολα θέλαμε να έχουμε ψηλό αριθμό συνδέσεων χωρίς κατ' ανάγκη να μας νοιάζει ο βαθμός παραπληροφόρησης. Τόσο στον πρώτο τρόπο όσο και στον δεύτερο, από τα τέσσερα δίπολα που παρήγαγε ο κάθε ένας, τα δύο θέλαμε να έχουν θετική πόλωση για να δούμε τον ανταγωνισμό και τα άλλα δύο να έχουν αρνητική πόλωση για να παρατηρήσουμε την συμφωνία μεταξύ τους.

Δίπολα	Πόλωση	Αριθμός συνδέσεων
(1,4)	0.332	7,535
(5,45)	0.282	20,886
(5,24)	-0.017	30,012
(5,4)	-0.336	118,628
(1,9)	0.5	37,830
(1,5)	0.457	69,923
(5,9)	-0.411	742,912
(5,2)	-0.426	505,884

Πίνακας 4.3: Τα δίπολα που επιλέχτηκαν με τους δύο τρόπους. Τα πρώτα τέσσερα δίπολα ήταν αποτέλεσμα της εφαρμογής του πρώτου τρόπου και τα τέσσερα τελευταία που βγήκαν από τον δεύτερο τρόπο.

Παρατηρώντας τα δίπολα που επιλέχθηκαν εντοπίζονται επτά κοινότητες: 1, 2, 4, 5, 9, 24 και 45. Το επόμενο βήμα είναι να εμβαθύνουμε σε αυτές τις επιλεγμένες κοινότητες και το περιεχόμενο που έχουν. Έτσι, παίρνουμε τις δημοσιεύσεις για κάθε επιλεγμένη κοινότητα αφού ξέρουμε το περιεχόμενο που δημοσιεύει ο κάθε λογαριασμός αλλά και την κοινότητα στην οποία ανήκει. Έχοντας την αντιστοιχία κάθε επιλεγμένης κοινότητας με την λίστα των δημοσιεύσεων της, εφαρμόζουμε μια διαδικασία για να τις επεξεργαστούμε. Για κάθε δημοσίευση γίνονται τα ακόλουθα:

- Όλες οι λέξεις μετατρέπουν τα κεφαλαία τους γράμματα (εάν έχουν) σε πεζά έτσι ώστε να μην διαφέρουν οι λέξεις που είναι το ίδιο όπως για παράδειγμα Covid, covid, COVID,
- Γίνεται αφαίρεση των συνδέσμων (<https>, <http>, [www](http)), των αναφορών άλλων ατόμων (@user) και των hashes (#covid),
- Ελέγχεται κατά πόσο οι λέξεις τις συγκεκριμένης δημοσίευσης είναι της αγγλικής γλώσσας,
- Εύρεση των ονοματικών φράσεων (noun phrase) και δημιουργία των νέων δημοσιεύσεων.

Έπειτα, έχοντας επεξεργασμένες δημοσιεύσεις της κάθε επιλεγμένης κοινότητας, προχωρούμε στο βήμα εύρεσης θεμάτων συζήτησης για κάθε μια από αυτές. Αυτό μπορεί να υλοποιηθεί με το μοντέλο RoBERTa-large [13]. Το συγκεκριμένο μοντέλο υπάγεται στην ενότητα BERT που αυτόματα σημαίνει ότι δουλεύει με εκπαίδευση αμφίδρομων αναπαραστάσεων του κειμένου σε όλα τα επίπεδα του μοντέλου. Συγκεκριμένα, το μοντέλο RoBERTa-large έχει προ-εκπαιδευτεί με μια μεγάλη συλλογή από κείμενα στην αγγλική γλώσσα χρησιμοποιώντας την τεχνική MLM (Masked Language Model). Δίνοντας μια δημοσίευση ως είσοδο σε αυτό το μοντέλο το τι παίρνουμε σαν έξοδο είναι το embedding που το αντιπροσωπεύει. Δηλαδή, έναν πίνακα με 1024 θέσεις όπου σε κάθε θέση βρίσκεται ένας πραγματικός αριθμός και αυτός ο πίνακας είναι σε θέση να αναπαραστήσει το νόημα της πρότασης που δώσαμε σαν είσοδο.

Το ότι πλέον οι δημοσιεύσεις δεν είναι προτάσεις από λέξεις αλλά πίνακες με πραγματικούς αριθμούς (embeddings) σημαίνει ότι μπορούμε να εφαρμόσουμε ομαδοποίηση με Cosine Similarity και να φτιάξουμε κοινότητες από embeddings, που

με λίγα λόγια σημαίνει κοινότητες δημοσιεύσεων. Οι δύο παράμετροι που έπρεπε να οριστούν για την ομαδοποίηση των embeddings ήταν η παράμετρος `min_community_size` και η παράμετρος `threshold`. Η παράμετρος `min_community_size` προσδιορίζει τον ελάχιστο αριθμό στοιχείων που θα απαρτίζεται μια ομάδα, ενώ η παράμετρος `threshold` καθορίζει το κατώτατο όριο ομοιότητας με το οποίο δύο embeddings θα πρέπει να μοιάζουν για να είναι στην ίδια ομάδα. Μέσα από διάφορα πειράματα διαφάνηκε ότι η κατάλληλη τιμή για `min_community_size` είναι το 10 (δηλαδή η κάθε ομάδα να αποτελείται από τουλάχιστον 10 embeddings) και 0.4 για την παράμετρο `threshold` (για να ανήκουν 2 embeddings στην ίδια ομάδα να είναι κατά 40% τα ίδια). Είναι σημαντικό να αναφέρουμε ότι ένα θέμα που κληθήκαμε να αντιμετωπίσουμε σχετιζόταν με την δέσμευση μνήμης (CPU allocator error) αφού δεν ήταν αρκετή για να διεκπεραιώσει την συγκεκριμένη διεργασία. Για να λύσουμε αυτό το θέμα χρειάστηκε να αλλάξουμε την μέθοδο που ήταν υπεύθυνη για την ομαδοποίηση και να την φτιάξουμε με τέτοιο τρόπο που να προσεγγίζει την λύση εάν τα δεδομένα ήταν πολλά και δεν μπορούσε να χρησιμοποιηθεί η ακριβής μέθοδος. Αφού έγινε ομαδοποίηση των δημοσιεύσεων και ξέρουμε ότι για κάθε επιλεγμένη κοινότητα υπάρχουν πολλές κοινότητες δημοσιεύσεων μπορούμε τώρα για κάθε μια από αυτές να εξαγάγουμε το θέμα συζήτησης.

Το επόμενο βήμα είναι η επεξεργασία κειμένου στις δημοσιεύσεις κάθε θέματος. Έτσι πήραμε πέντε διαφορετικές μετρήσεις για να έχουμε μια πιο ολοκληρωμένη άποψη. Οι μετρήσεις ήταν:

- θετικότητα, έδειχνε εάν ένα κείμενο είχε θετικές λέξεις και έτσι πρόσφερε θετικά αποτελέσματα,
- υποκειμενικότητα, υποδείκνυε κατά πόσο μια δημοσίευση ήταν άποψη ενός χρήστη ή γεγονός,
- σθένος, για να δούμε πόση βαρύτητα είχαν οι λέξεις που χρησιμοποιούνταν,
- debate, φανερώνει την συζήτηση άρα ένας θετικός αριθμός έδειχνε συμφωνία ενώ ένας αρνητικός αριθμός έδειχνε διαφωνία
- τα συναισθήματα που προκαλεί.

4.3 Αποτελέσματα

Όπως ειπώθηκε στο προηγούμενο υποκεφάλαιο, οι επιλεγμένες κοινότητες έχουν πολλές κοινότητες δημοσιεύσεων και κάθε μια από αυτές μπορεί να μας δώσει το θέμα συζήτησης. Από τον πιο κάτω πίνακα, φαίνεται πόσες κοινότητες δημοσιεύσεων βρέθηκαν σε κάθε επιλεγμένη κοινότητα από την εφαρμογή του μοντέλου RoBERTa-large καθώς και τον συνολικό αριθμό των δημοσιεύσεων.

Κοινότητα	Συνολικός αριθμός δημοσιεύσεων	Πλήθος κοινοτήτων δημοσιεύσεων
1	139,015	212
2	133,454	169
4	5,109	36
5	112,326	127
9	112,287	135
24	12,700	86
45	1,261	10

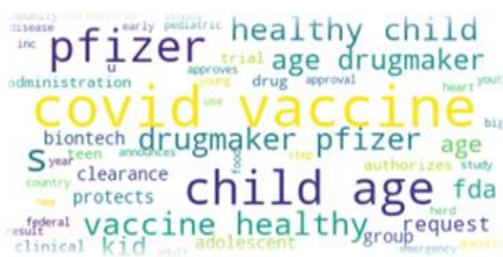
Πίνακας 4.4: Ανάλυση των δημοσιεύσεων ανά κοινότητα

4.3.1 Κοινότητα 1

Σύμφωνα με το πιο κάτω σχήμα τα θέματα τα οποία εντοπίστηκαν στην κοινότητα 1 είναι:

- Παροχή δωρεάν υπηρεσιών και αγαθών όπως είναι τα γλυκά σε άτομα με ιστορικό εμβολιασμού κατά του κορονοϊού. Κάποιες από τις δημοσιεύσεις περί δωρεάν αγαθών ήταν: «krispy kreme store free doughnut covid-19 vaccination record card», «krispy kreme valid covid-19 vaccination card doughnut offer good», «popular doughnut shop krispy kreme offering free donut covid-19 vaccine».

- Διαδικασία ελέγχου εμβολίου για άτομα σε νεαρή ηλικία. Μετά από παρατήρηση των επεξεργασμένων δημοσιεύσεων του θέματος, οι πιο συχνές εκφράσεις ήταν οι εξής: «drugmaker pfizer covid-19 vaccine healthy child age», «pfizer vaccine drugmaker pfizer covid-19 vaccine healthy child age», «early clinical trial pfizer 's covid-19 vaccine child age», «kid age pfizer covid-19 vaccine big step herd immunity».
- Σύγκριση μεταξύ των εμβολίων που υπάρχουν στην αγορά. Συγκεκριμένα, φαίνεται ότι η συγκεκριμένη κοινότητα κρίνει τα εμβόλια AstraZeneca και Johnson&Johnson λόγω του ότι προκαλούν σοβαρά θέματα υγείας (θρομβώσεις).



(α) Εμβόλιο σε παιδιά



(β) Δωρεάν προϊόντα



(γ) Εμβόλιο AstraZeneca



(δ) Εμβόλιο Johnson & Johnson

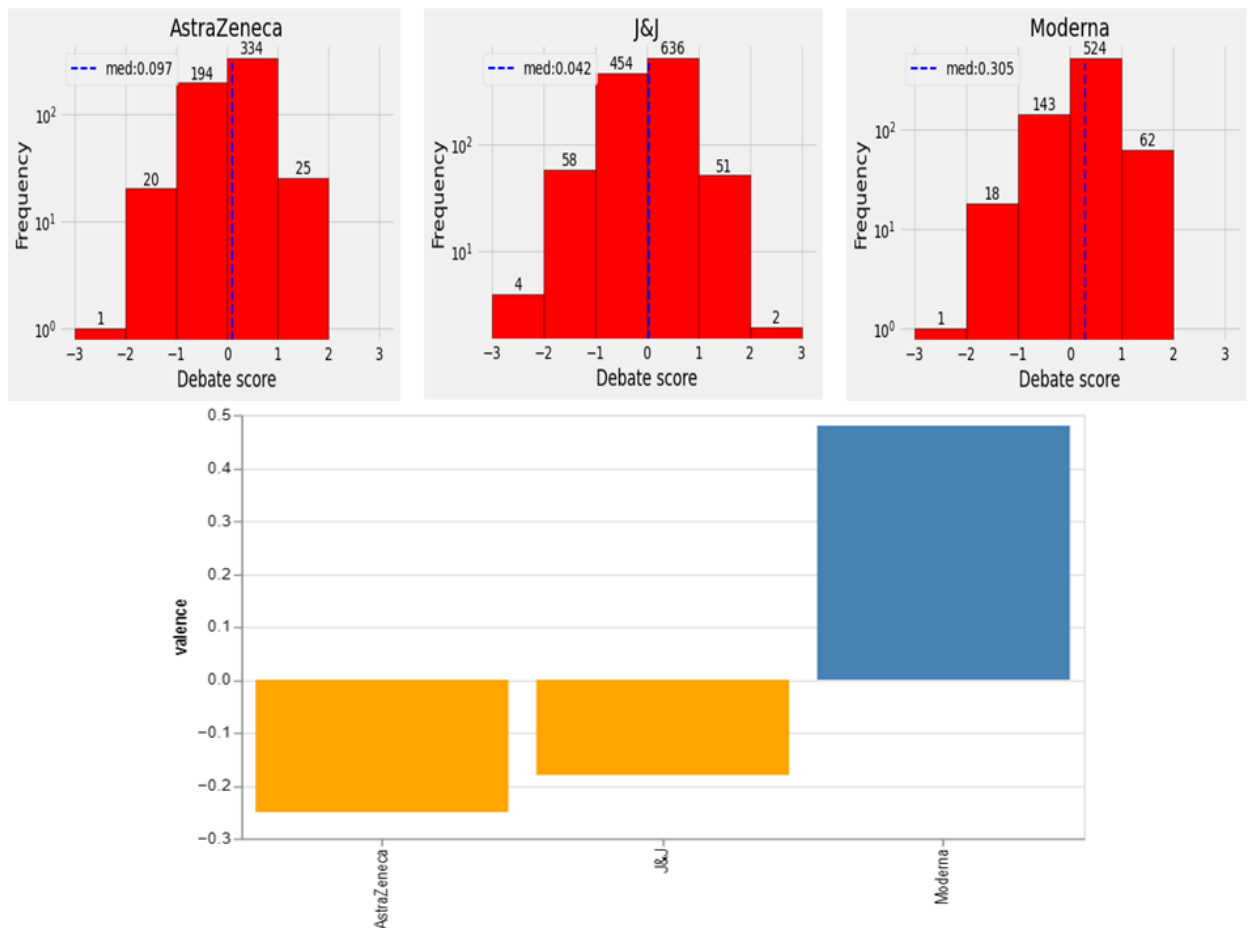


(ε) Εμβόλιο Moderna

Σχήμα 4.5: Θέματα συζήτησης κοινότητας 1

Όσο αφορά τα εμβόλια μπορούμε να πάρουμε περισσότερες πληροφορίες από τις πιο κάτω γραφικές. Σε αυτές συγκρίνονται τα δύο εμβόλια, τα οποία όπως αναφέρεται από τις δημοσιεύσεις προκαλούν θρομβώσεις, και το εμβόλιο Moderna το οποίο είχε θετικές

κριτικές. Τα ιστογράμματα στο πάνω μέρος φανερώνουν το debate score. Όπως διακρίνουμε τόσο το AstraZeneca όσο και το Johnson&Johnson εμβόλιο έχουν διάμεσο (median) κοντά στο 0 για την μετρική αυτή και έτσι δεν υπάρχει κάτι το οποίο να δείχνει με σιγουριά ότι πρόκειται για συμφωνία ή διαφωνία επί του θέματος. Παρόλ' αυτά, για το εμβόλιο Moderna έχει διάμεσος 0.3 που υποδεικνύει ομοφωνία μεταξύ των χρηστών της κοινότητας. Προχωρώντας στην γραφική που δείχνει το σθένος, στο ίδιο σχήμα, βλέπουμε ότι η βαρύτητα των λέξεων που συνοδεύουν τα εμβόλια AstraZeneca και Johnson&Johnson είναι αρνητικής σημασίας. Το αντίθετο συμβαίνει με τις λέξεις που συνοδεύουν το εμβόλιο Moderna. Έτσι μπορούμε να δούμε την επιρροή στις απόψεις των ανθρώπων όσο αφορά τα θέματα υγείας, που όπως αναφέρονται από τις δημοσιεύσεις, προκύπτουν μετά την δόση των δύο εμβολίων.



Σχήμα 4.6: Συζήτηση σχετικά με τα εμβόλια για τους χρήστες της κοινότητας 1

Κάποια επιπλέον θέματα που αφορούν την κοινότητα 1 και απεικονίζονται στο σχήμα A1 του Παραρτήματος Α είναι τα εξής:

- Διαδικασία εμβολιασμού, πιο συγκεκριμένα συζητιέται το εμβόλιο της Pfizer.
- Η χρήση της προστατευτικής μάσκας ως ένα ισχυρό μέτρο κατά τις διάδοσης της πανδημίας.
- Περιορισμοί σε ταξίδια είτε για λόγους αναψυχής ή για επαγγελματικούς. Οι χρήστες σε αυτήν την κοινότητα σχολιάζουν και την έκδοση πιστοποιητικού εμβολιασμού «covid passport» για τα ταξίδια.
- Οικονομικός τομέας της χώρας. Γίνεται λόγος για οικονομικό πλάνο ανάκαμψης αλλά και επιδόματα για στήριξη των ανέργων.
- Περιστατικό κορονοϊού σε κρουαζιερόπλοιο.

4.3.2 Κοινότητα 2

Η κοινότητα 2 δίνεται περιληπτικά αφού δεν επιλέχθηκε στην ανάλυση δίπολων. Τα θέματα τα οποία βρέθηκαν μέσα από συγκεκριμένες κοινότητες δημοσιεύσεων και απεικονίζονται στο σχήμα A2 του Παραρτήματος Α είναι τα εξής:

- Πειραματικές δοκιμές στους ανθρώπους με σκοπό την εξεύρεση αντιδότης απέναντι στην πανδημία του κορονοϊού.
- Διαδικασία εμβολιασμού. Πέρα από αυτό, συζητιούνται πιο συγκεκριμένα τα εμβόλια που βρίσκονται στην αγορά όπως Pfizer και AstraZeneca. Για το εμβόλιο AstraZeneca, η συγκεκριμένη κοινότητα κάνει λόγο για θρόμβους που προκαλούνται μετά από δόση του συγκεκριμένου εμβολίου.
- Απαγόρευση κυκλοφορίας και απομόνωση. Γίνεται συζήτηση ιδιαίτερα για την απομόνωση σε ξενοδοχεία.
- Περιορισμοί σε ταξίδια είτε για λόγους αναψυχής ή για επαγγελματικούς λόγους. Οι χρήστες σε αυτήν την κοινότητα σχολιάζουν και την έκδοση πιστοποιητικού εμβολιασμού «covid passport» για τα ταξίδια.
- Περιστατικό κορονοϊού σε κρουαζιερόπλοιο.

4.3.3 Κοινότητα 4

Όσο αφορά την κοινότητα 4, από το πιο κάτω σχήμα μπορούμε να διακρίνουμε τα θέματα:

- Νομιμοποίηση της κάνναβης. Όπως φαίνεται από τις δημοσιεύσεις αυτή ήταν μια πρόταση που τέθηκε στην Βουλή σαν μια δράση που στόχο έχει την ανάκαμψη της οικονομίας για την εποχή που θα ακολουθήσει την πανδημία. Κάποιες από τις δημοσιεύσεις σχετικές με αυτό το θέμα ήταν: «new york legalize marijuana 'soon economic recovery governor cuomo re», «new mexico state need economic relief marijuana legalization».
- Ανάπτυξη τεχνολογίας παράλληλα με την πανδημία.
- Ανάγκες των ανθρώπων. Αναφέρονται τα «θέλω» των ανθρώπων τα οποία έχουν στερηθεί κατά την περίοδο αυτή όπως «mental health need», «food shortage food bank economic hardship family need».
- Επίδραση της πανδημίας στην μουσική βιομηχανία και το πως κατάφερε να επιβιώσει.
- Περιβάλλον.



(α) Κάνναβη ως οικονομική λύση



(β) Τεχνολογία



(γ) Ανάγκες



(δ) Μουσική βιομηχανία



(ε) Περιβάλλον

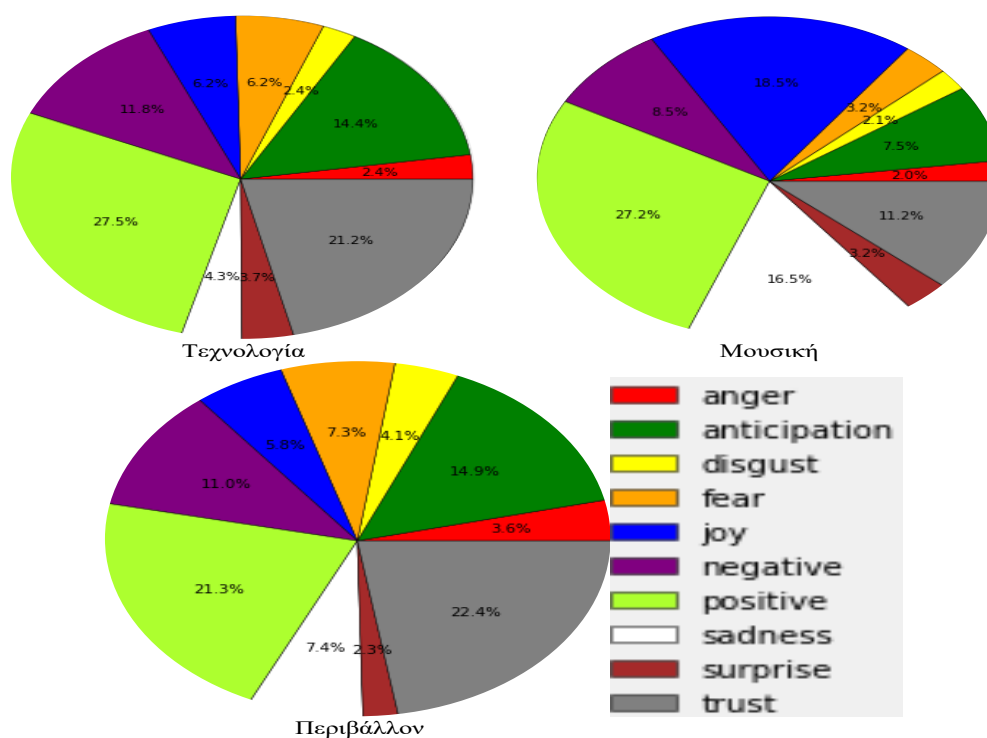
Σχήμα 4.7: Θέματα συζήτησης κοινότητας 4

Είναι σημαντικό να αναφερθεί ότι σε αυτήν την κοινότητα φαίνονται ξεκάθαρα κάποια «πλεονεκτήματα» που είχε η πανδημία, δηλαδή πως έκανε καλό στους ανθρώπους. Αυτό μπορούμε να το συμπεράνουμε από το πιο κάτω σχήμα που δείχνει τα συναισθήματα που αντανακλούν οι δημοσιεύσεις τις συγκεκριμένης κοινότητας για τα θέματα της τεχνολογίας, της μουσικής και του περιβάλλοντος. Τα συνολικά θετικά συναισθήματα (ανυπομονησία, ενθουσιασμός θετικότητα, έκπληξη και εμπιστοσύνη) για τα θέματα τεχνολογίας, μουσικής και περιβάλλοντος είναι 73%, 67,6% και 66,6% αντίστοιχα.

Όσο αφορά την τεχνολογία φαίνεται ότι οι άνθρωποι νιώθουν θετικά για την ανάπτυξη της και το πως εξελίσσεται κατά την διάρκεια της πανδημίας δείχνοντας αρκετή εμπιστοσύνη. Άλλο ένα χαρακτηριστικό είναι και η ανυπομονησία. Κάποιες από τις εκφράσεις μέσα από τις δημοσιεύσεις ήταν: «direct payment economy recovers basic need equitable solution», «remote worker workplace intelligence remote digital transformation», «discussion digital transformation business landscape business expert mitra innovation meritis business solution».

Όσο αφορά τη μουσική φαίνεται ότι οι άνθρωποι ναι μεν νιώθουν λύπη διότι δεν έγιναν συναυλίες που υπό κανονικές συνθήκες θα λάμβαναν χώρα, όμως υπάρχει ο ενθουσιασμός. Αυτό γιατί οι μουσικοί εφαρμόζουν το independent venue όπως φαίνεται και στο πιο πάνω σχήμα, δηλαδή να εμφανίζονται διαδικτυακά και να κρατούν συντροφιά στους ανθρώπους για τις δύσκολες μέρες που περνούν.

Όσο αφορά το περιβάλλον, η συγκεκριμένη κοινότητα αναφέρει ότι είναι εμφανές πως η πανδημία βοήθησε στο να κρατήσουμε το περιβάλλον καθαρό είτε μέσω απαγορεύσεων κυκλοφορίας και γενικά λόγω περιορισμένης μετακίνησης. Μέσα από τις δημοσιεύσεις οι άνθρωποι σχολιάζουν το γεγονός ότι ο αέρας είναι πιο καθαρός και μειώθηκαν αισθητά οι ρυθμοί παραγωγής διοξειδίου του άνθρακα. Από την γραφική παράσταση στο πιο κάτω σχήμα φαίνεται και μια νότα εμπιστοσύνης για την ανάπτυξη της και το πως εξελίσσεται κατά την διάρκεια της πανδημίας. Κάποιες από τις δημοσιεύσεις ήταν: «air quality», «clean air attack», «global lockdown air pollution clean air».



Σχήμα 4.8: Τεχνολογία – Μουσική – Περιβάλλον

Κάποια επιπλέον θέματα που αφορούν την κοινότητα 4 και απεικονίζονται στο σχήμα Α3 του Παραρτήματος Α είναι τα ακόλουθα:

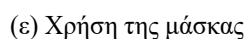
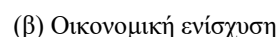
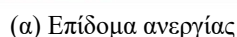
- Οικονομική ενίσχυση. Αυτό το θέμα αναφέρεται ειδικά προς τα άτομα νεαρής ηλικίας.
- Η χρήση της προστατευτικής μάσκας και πως αυτό το μέτρο έρχεται σε αντίθεση με την ανάγκη του ανθρώπου για κοινωνικοποίηση.
- Σχολείο και νέοι τρόποι διδασκαλίας.
- Εγκλήματα και απάτες με σκοπό κάποιοι να εξασφαλίσουν περισσότερη οικονομική βοήθεια από το κράτος.

4.3.4 Κοινότητα 5

Όσο αφορά την κοινότητα 5, τα θέματα τα οποία ξεχώρισαν και φαίνονται και στο πιο κάτω σχήμα ήταν:

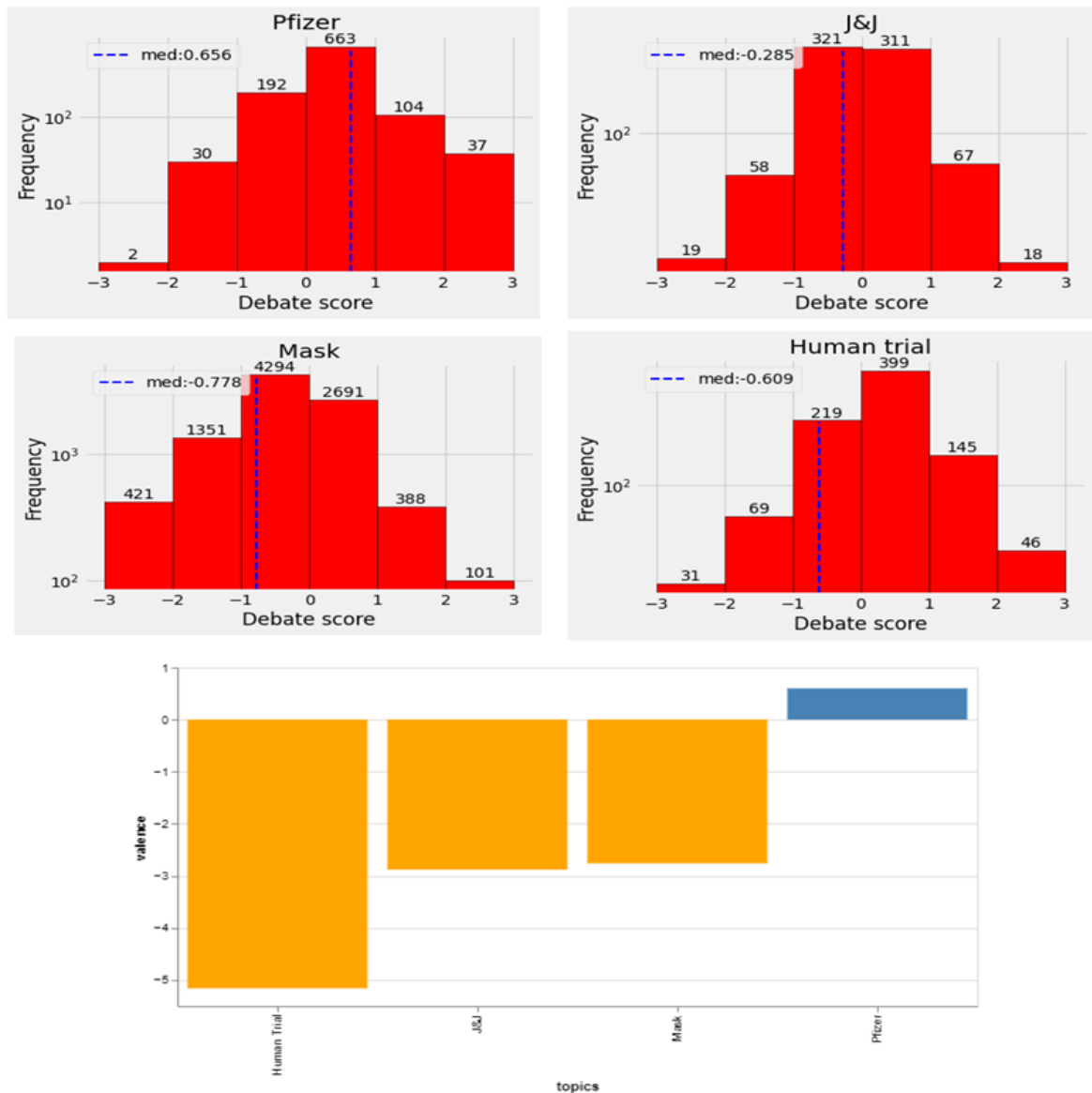
- Διαδικασία εμβολιασμού, πιο συγκεκριμένα συζητούνται τα εμβόλια που βρίσκονται στην αγορά όπως Pfizer και Johnson&Johnson. Για το εμβόλιο

- Χρήση μάσκας για προστασία από τον υό και έλεγχος εμβολίων κατά πόσο είναι ικανά και αποδοτικά.
- Εκλογές εν καιρώ πανδημίας. Οι χρήστες σχολιάζουν: «coronavirus outbreak large-scale outbreak reelection bid», «final week election day new coronavirus infection all-time high prominent issue».
- Επιδόματα ανεργίας και οικονομική ενίσχυση.



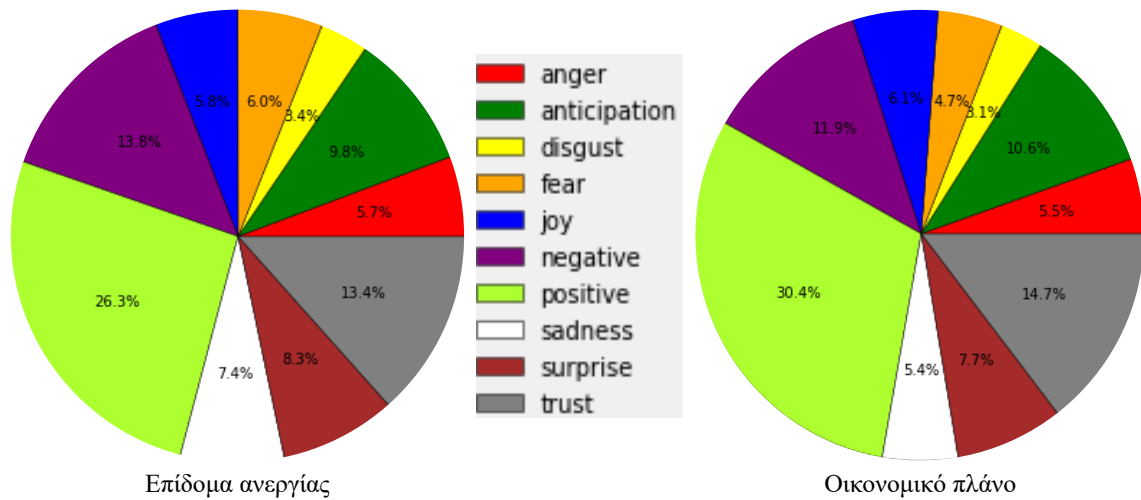
Η συγκεκριμένη κοινότητα φαίνεται να αντιδρά στον τρόπο με τον οποίο οι άνθρωποι αντιμετωπίζουν την πανδημία αφού δεν φαίνεται να ανέχονται τη χρήση της μάσκας και τις δοκιμές των εμβολίων που διεξάγονται με σκοπό οι ειδικοί να δουν πόσο αποτελεσματικά είναι απέναντι στον ιό. Η χρήση της μάσκας έχει debate score ίσο με -0.78 με το σθένος των λέξεων που την συνοδεύουν στις δημοσιεύσεις να είναι

αρνητικό (-2.7). Έτσι, το συμπέρασμα είναι ότι οι άνθρωποι δεν υποστηρίζουν τόσο το μέτρο αυτό και ότι υπάρχουν αρκετές διαφωνίες επί του θέματος. Κάτι ανάλογο συμβαίνει με τις δοκιμές των εμβολίων με την διαφορά ότι η βαρύτητα των λέξεων που χρησιμοποιούνται είναι ακόμα πιο αρνητική (σχεδόν -5). Αξιοσημείωτο το γεγονός ότι υπάρχει ένα χάσμα μεταξύ των εμβολίων Pfizer και Johnson&Johnson που πιθανόν να οφείλεται στις θρομβώσεις που αναφέρονται ότι προκαλεί το δεύτερο.



Σχήμα 4.10: Απόψεις για εμβόλια και μάσκα

Παρόλ' αυτά, οι στάση των χρηστών αυτής της κοινότητας φαίνεται να είναι τελείως διαφορετική σε θέματα που έχουν να κάνουν με τον οικονομικό τομέα. Το επίδομα ανεργίας και το οικονομικό πλάνο για ανάκαμψη της οικονομίας της χώρας κάνει την πλειοψηφία των ανθρώπων να είναι κυρίως θετικοί (26% και 30%) αλλά και πολλούς άλλους να έχουν και εμπιστοσύνη για το εγγύς μέλλον (13% και 14%).



Σχήμα 4.11: Συναισθήματα για επίδομα ανεργίας και οικονομική ενίσχυση

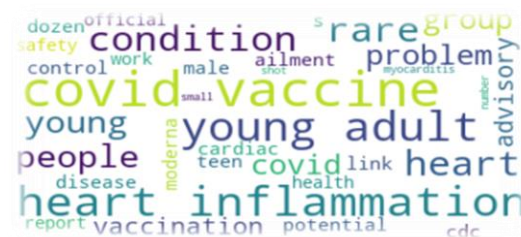
4.3.5 Κοινότητα 9

Βλέποντας το πιο κάτω σχήμα που απεικονίζει τα θέματα για την κοινότητα 9, μπορούμε να δούμε ότι τα πιο κύρια είναι:

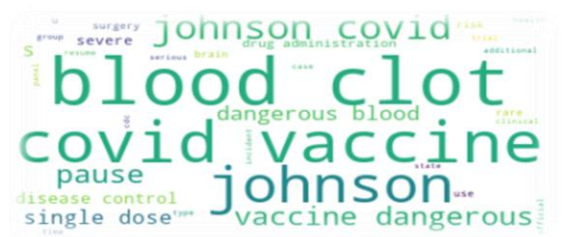
- Αρνητικές επιπτώσεις εμβολιασμού. Συγκεκριμένα, οι χρήστες αυτής της κοινότητας σχολιάζουν ότι τα εμβόλια Johnson&Johnson προκαλούν προβλήματα υγείας. Επίσης, γίνεται λόγος το γεγονός ότι τα εμβόλια είναι απειλή για τα άτομα νεαρής ηλικίας. Οι πλείστοι χρήστες αναφέρουν: «johnson covid-19 vaccine severe type blood clot», «johnson 's covid-19 vaccine rare blood clot disorder», «potential link covid-19 vaccine cardiac ailment young people», «young adult covid vaccine heart inflammation advisory group rare condition», «heart inflammation moderna s covid-19 vaccine male teen young adult».
- Πολιτικά θέματα. Κατηγορείται τόσο ο κυβερνήτης Cuomo ότι δεν ανέφερε σκόπιμα θανάτους που προκλήθηκαν λόγω κορονοϊού όσο και η πολιτικός

Pelosi για το γεγονός ότι προσπάθησε να συμπεριλάβει τη χρηματοδότηση των αμβλώσεων στο νομοσχέδιο για την καταπολέμηση του κορονοϊού («nancy pelosi s covid relief bill fund abortion», «pro-life campaign: halt abortions to free up resources for coronavirus fight»).

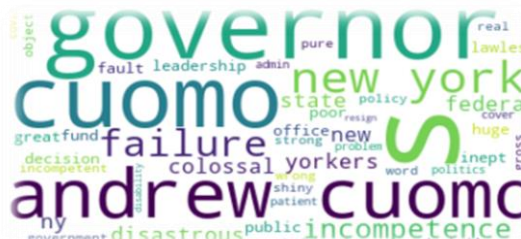
- Περιορισμοί όπως για παράδειγμα απαγόρευση κυκλοφορίας και έκδοση πιστοποιητικού εμβολιασμού για ταξίδια.



(α) Φλεγμονή λόγω εμβολίων σε νεαρούς



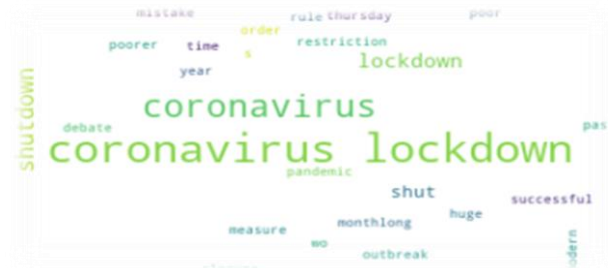
(β) Εμβόλιο Johnson&Johnson



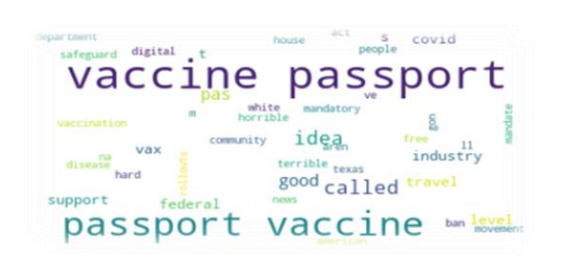
(γ) Κατηγορίες κατά Cuomo



(δ) Κατηγορίες κατά Pelosi



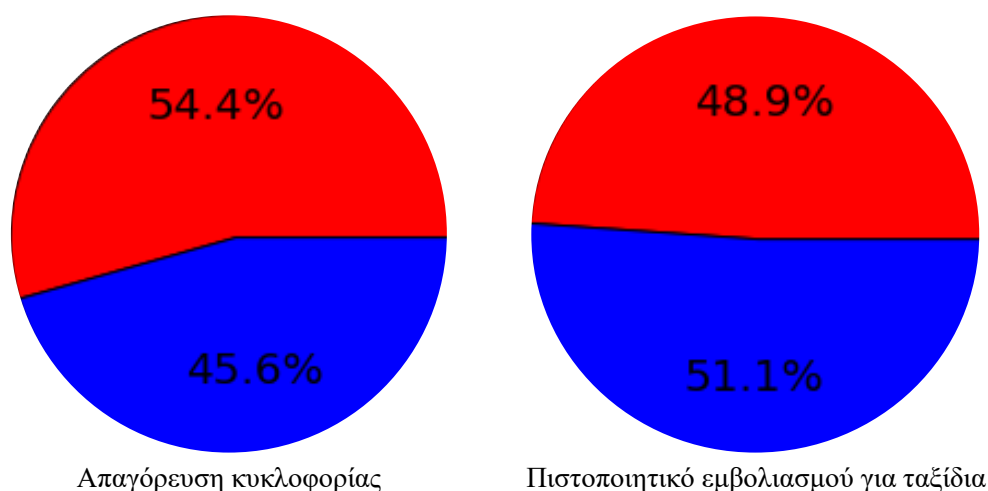
(ε) Απαγόρευση κυκλοφορίας



(ζ) Πιστοποιητικό εμβολιασμού

Σχήμα 4.12: Θέματα συζήτησης κοινότητας 9

Όσο αφορά τους περιορισμούς και πιο συγκεκριμένα αυτούς που έχουν να κάνουν με απαγόρευση κυκλοφορίας και πιστοποιητικό εμβολιασμού βλέπουμε ότι υπάρχει μια σύγχυση στα συναισθήματα των πολιτών. Μπορούμε να υποθέσουμε ότι κάποιοι υποστηρίζουν τα μέτρα ως ένα ισχυρό όπλο για την πιο αποτελεσματική αντιμετώπιση της πανδημίας ενώ άλλοι δεν θεωρούν πως είναι σωστή λύση αφού μπορεί να τους στοιχίσει σε διάφορους τομείς όπως είναι ο οικονομικός.



Σχήμα 4.13: Θετικά και αρνητικά συναισθήματα για την εφαρμογή περιορισμών

Κάποια επιπλέον θέματα που αφορούν την κοινότητα 9 και απεικονίζονται στο σχήμα A4 του Παραρτήματος Α είναι τα εξής:

- Οικονομική ενίσχυση για τους άνεργους πολίτες.
- Επιτάχυνση διαδικασίας εμβολιασμού και περισσότερες πληροφορίες για τα εμβόλια Pfizer και Moderna.
- Χρήση της μάσκας.
- Σχολείο και νέοι τρόποι διδασκαλίας.

4.3.6 Κοινότητα 24

Η κοινότητα 24 δίνεται περιληπτικά αφού δεν επιλέχθηκε στην ανάλυση δίπολων. Τα θέματα τα οποία βρέθηκαν μέσα από συγκεκριμένες κοινότητες δημοσιεύσεων και απεικονίζονται στο σχήμα A5 του Παραρτήματος Α είναι τα εξής:

- Ανάγκη για ανθρώπινη επαφή.
- Διάφοροι τρόποι αντιμετώπισης της πανδημίας. Αναφέρονται κυρίως περιορισμοί.
- Τρόποι προστασίας όπως είναι η χρήση αντισηπτικού και μάσκας.
- Συμπτώματα κορονοϊού που εμφανίζουν οι άνθρωποι αφού μολυνθούν.
- Θέματα γύρω από τα σχολεία και τους μαθητές, όπως είναι ο νέος τρόπος εκμάθησης και διάφορα δάνεια προς τους μαθητές για καταπολέμηση του υιού.
- Παραπληροφόρηση ειδήσεων που έχουν να κάνουν με την πανδημία.

4.3.7 Κοινότητα 45

Μπορούμε να διακρίνουμε κάποια θέματα της κοινότητας 45 από το πιο κάτω σχήμα όπως είναι:

- Εμβολιασμός.
- Περιστατικό κορονοϊού σε κρουαζιερόπλοιο.
- Μέτρα για καταπολέμηση του υιού όπως είναι η χρήση της μάσκας και η απαγόρευση κυκλοφορίας («tyrannical lockdown order», «due consideration ... people lockdown order», «need draconian lockdown», «country s mandatory lockdown», «lock people»).



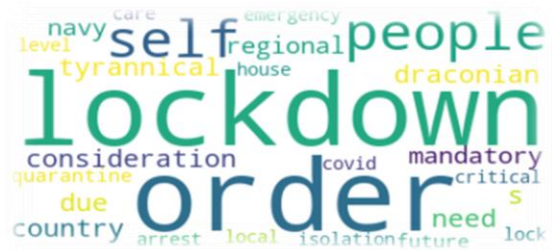
(α) Εμβολιασμός



(β) Χρήση της μάσκας



(γ) Περιστατικό κορονοϊού



(δ) Απαγόρευση κυκλοφορίας

Σχήμα 4.14: Θέματα συζήτησης κοινότητας 45

Κεφάλαιο 5

Συμπεράσματα

5.1	Συμπεράσματα διπόλων.....	44
5.1.1	Δίπολο (1,4)	44
5.1.2	Δίπολο (5,45)	47
5.1.3	Δίπολο (1,5)	50
5.1.4	Δίπολο (5,9)	54
5.2	Ποσοτική έρευνα	55
5.3	Μελλοντική έρευνα	56

Έχοντας αναλύσει τις κοινότητες και βρίσκοντας τα διάφορα θέματα που συζητούν οι χρήστες, σειρά έχει η διαδικασία εύρεσης των τελικών συμπερασμάτων τα οποία προκύπτουν από την ανάλυση των διπόλων. Για να γίνει αυτό εφικτό αλλά και για να εξασφαλιστεί η εγκυρότητα των συμπερασμάτων, γίνεται ποιοτική και ποσοτική έρευνα.

Όσο αφορά την ποιοτική έρευνα, σκοπός ήταν να γίνει κατανοητή η σχέση μεταξύ των κοινοτήτων ενός διπόλου και να μελετηθούν εις βάθος. Παρατηρήθηκε ότι σχεδόν όλες οι κοινότητες είχαν θέματα τα οποία να σχετίζονταν με τρεις συγκεκριμένους άξονες: τον άξονα υγείας, τον άξονα πρόληψης και διαχείρισης και τον άξονα οικονομίας. Ο άξονας υγείας εμπερικλείει τα θέματα όπως είναι ο εμβολιασμός και τα διάφορα είδη εμβολίων. Ο άξονας πρόληψης και διαχείρισης έχει να κάνει με τα διάφορα πρωτόκολλα και μέτρα που σκοπό έχουν την επιτυχή αντιμετώπιση της πανδημίας όπως είναι η χρήση της μάσκας, η απαγόρευση κυκλοφορίας, η έκδοση πιστοποιητικού εμβολιασμού για ταξίδια και οι νέες μέθοδοι διδασκαλίας (εξ αποστάσεως). Ο άξονας οικονομίας σχετίζεται με τα οικονομικά πλάνα, τα επιδόματα ανεργίας και γενικά

οτιδήποτε επηρεάζει την οικονομία. Άρα στην ποιοτική έρευνα, ερευνήσαμε την στάση των κοινοτήτων των διπόλων πρώτα στους τρεις άξονες και μετά ερευνήσαμε πιο αναλυτικά τα κοινά θέματα που συζητούσαν με απεικόνιση των στατιστικών σε γραφικές παραστάσεις.

Όσο αφορά την ποσοτική έρευνα, εφαρμόστηκε έτσι ώστε να επαληθεύσουμε τα συμπεράσματα που είχαμε. Υλοποιήθηκε με την χρήση του Mann-Whitney U Test. Με αυτόν τον τρόπο είμασταν σε θέση να καθορίσουμε εάν δύο ανεξάρτητα δείγματα έχουν την ίδια κατανομή ή όχι. Με λίγα λόγια, μπορούμε να αποδείξουμε με στατιστική ακρίβεια εάν δύο δείγματα έχουν αξιοσημείωτη διαφορά στις απόψεις όπου αυτό συνεπάγεται σε σύγκρουση ή η διαφορά είναι αμελητέα που αυτόματα σημαίνει συμφωνία.

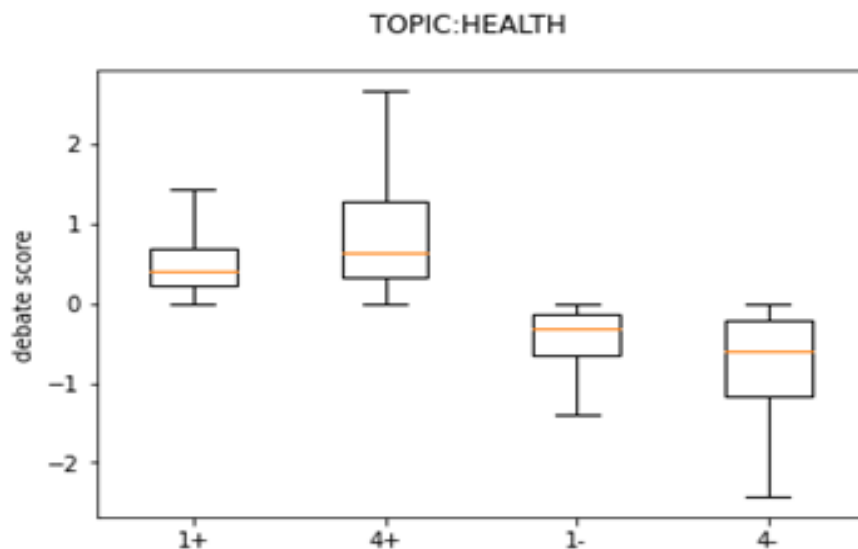
5.1 Συμπεράσματα διπόλων

5.1.1 Δίπολο (1,4)

Για τον άξονα υγείας, φαίνεται μέσω της γραφικής παράστασης που αναπαριστά την μετρική του debate score ότι δεν υπάρχει σαφής ένδειξη αν υπάρχει συμφωνία η διαφωνία σε κάποια κοινότητα. Αν πάρουμε για παράδειγμα την κοινότητα 1, βλέπουμε ότι οι θετικές τιμές έχουν διακύμανση μέχρι το 1.5 όμως αυτές αναιρούνται από τις αντίστοιχες αρνητικές τιμές που περίπου φτάνουν το -1.5 και έτσι με βάση αυτήν την γραφική δεν μπορούμε να συμπεράνουμε ότι τα θετικά υπερτερούν και ότι η κοινότητα φαίνεται να συμφωνεί σε θέματα γύρω από τον άξονα της κοινότητας. Το μόνο που μπορούμε να πούμε είναι ότι η διακύμανση των τιμών της κοινότητας 1 είναι μικρότερη από αυτή της κοινότητας 4. Συνεπώς, αυτό είναι ένα στοιχείο ότι οι απόψεις των χρηστών της κοινότητας 1 είναι πιο σταθερές και ισορροπημένες.

Παίρνοντας το 75% από το δείγμα των δύο κοινοτήτων (μέσω quantiles) είδαμε ότι η πλειοψηφία του debate score είναι 0.54 και 0.91 για τις κοινότητες 1 και 4 αντίστοιχα. Δηλαδή υπάρχει περισσότερη συμφωνία εντός της κοινότητας 4 αφού η τιμή είναι σχεδόν διπλάσια από την αντίστοιχη τιμή για την κοινότητα 1. Αυτό μπορεί να

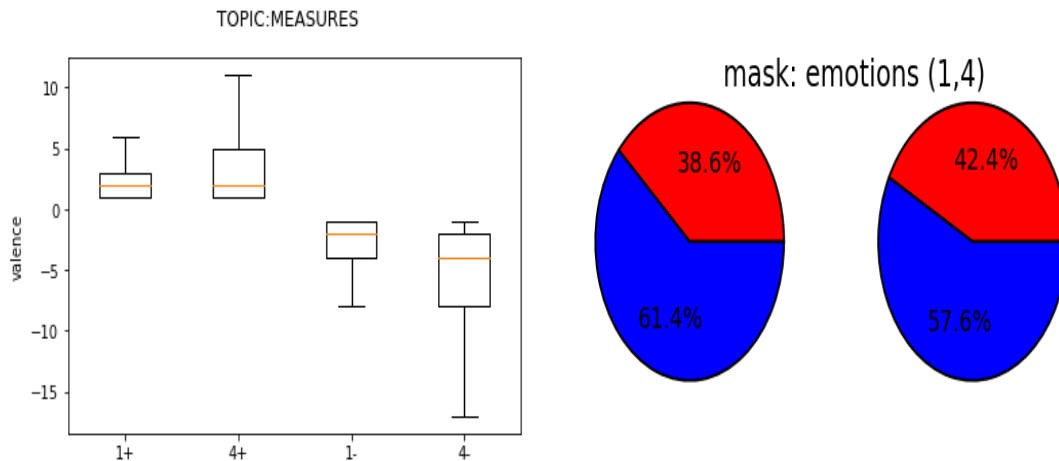
οφείλεται και στο γεγονός ότι για το θέμα των εμβολίων του άξονα της υγείας η κοινότητα 1 καταδίκασε τα εμβόλια Johnson&Johnson και AstraZeneca ότι προκαλούν θρομβώσεις.



Σχήμα 5.1: Άξονας υγείας για το δίπολο (1,4) με την μετρική debate score

Όσο για τον άξονα πρόληψης και διαχείρισης της πανδημίας, φαίνεται μέσω της πιο κάτω γραφικής παράστασης που βρίσκεται στα αριστερά και αναπαριστά την μετρική της βαρύτητας των λέξεων ότι η κοινότητα 4 έχει μεγαλύτερο αρνητικό σθένος. Με άλλα λόγια, η κοινότητα 4 χρησιμοποιεί λέξεις που είναι περισσότερο αρνητικές από την κοινότητα 1 όταν αναφέρεται σε θέματα που έχουν να κάνουν με τη χρήση της μάσκας, την απαγόρευση κυκλοφορίας και άλλα.

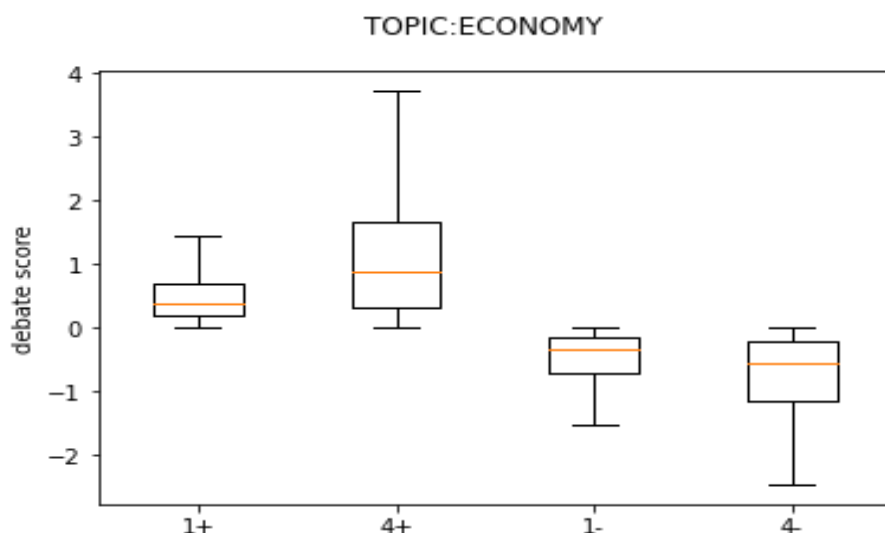
Παρατηρώντας την δεξιότερη γραφική παράσταση επαληθεύεται η πιο πάνω πρόταση αλλά τώρα σε επίπεδο συγκεκριμένου θέματος. Πιο αναλυτικά, τα δύο pie charts δείχνουν τα συναισθήματα των δύο κοινοτήτων (κοινότητα 1 στα αριστερά και κοινότητα 4 στα δεξιά). Παρόλο που η πλειοψηφία των συναισθημάτων είναι θετική και για τις δύο κοινότητες, εντούτοις βλέπουμε ότι η κοινότητα 1 εκφράζει λιγότερα αρνητικά συναισθήματα για την χρήση μάσκας από την κοινότητα 4. Ακόμα ένα στοιχείο που επικυρώνει την πιο πάνω θέση είναι ότι σε επίπεδο κειμένου η κοινότητα 1 χρησιμοποιεί λέξεις όπως «required», «mandate», ενώ η κοινότητα 4 χρησιμοποιεί λέξεις και εκφράσεις όπως «anti-mask movement», «no body language».



Σχήμα 5.2: Άξονας πρόληψης και διαχείρισης για το δίπολο (1,4)

Για τον άξονα οικονομίας, φαίνεται μέσω της γραφικής παράστασης που αναπαριστά την μετρική του debate score ότι γενικά η κοινότητα 4 έχει πιο πολλές και μεγάλες θετικές τιμές. Αυτό συνεπάγεται στο ότι οι χρήστες της συγκεκριμένης κοινότητας συμφωνούν σε μεγάλο βαθμό στα θέματα που υπάγονται κάτω από τον άξονα της οικονομίας.

Παίρνοντας το 75% από το δείγμα των δύο κοινοτήτων (μέσω quantiles) είδαμε ότι η πλειοψηφία του debate score είναι 0.43 και 1.20 για τις κοινότητες 1 και 4 αντίστοιχα, δηλαδή σχεδόν το τριπλάσιο. Αυτό μας οδηγεί στο συμπέρασμα ότι οι χρήστες που ανήκουν στην κοινότητα 4 συμφωνούν πολύ περισσότερο σε αυτά που συζητιούνται. Μπορούμε να επαληθεύσουμε όσα ειπώθηκαν και από τα δεδομένα σε επίπεδο κειμένου αφού η κοινότητα 1 μιλά συγκεκριμένα για επίδομα ανεργίας και οικονομικό πλάνο για επαναφορά της οικονομίας, ενώ η κοινότητα 4 εκτός από αυτά μιλά για νομιμοποίηση της κάνναβης, οικονομική υποστήριξη των παιδιών και γενικά για τρόπους που θα επιφέρουν την οικονομική ανάκαμψη.

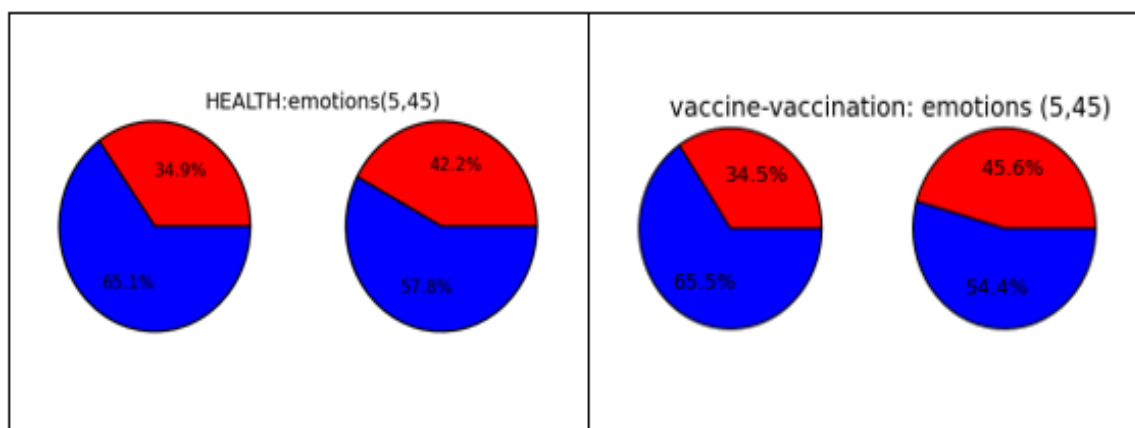


Σχήμα 5.3: Άξονας οικονομίας για το δίπολο (1,4)

5.1.2 Δίπολο (5,45)

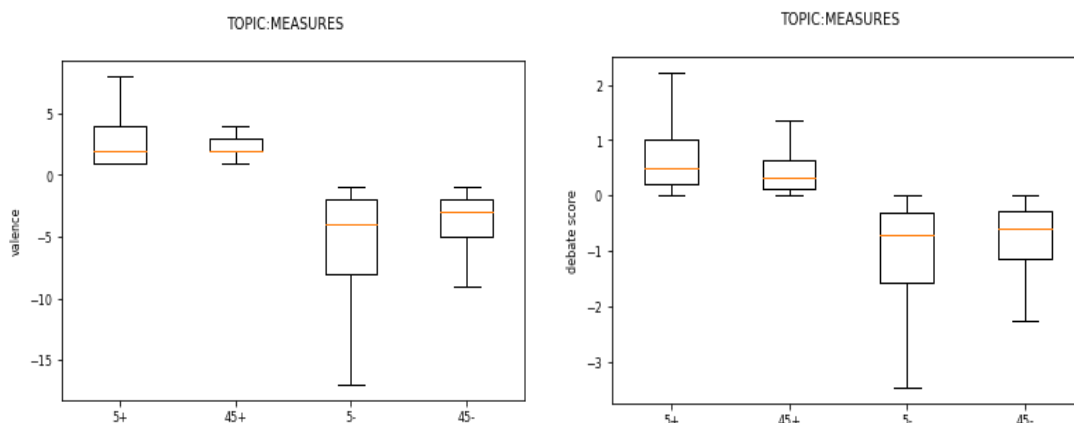
Για τον άξονα υγείας, μπορούμε να πούμε με βεβαιότητα ότι η κοινότητα 5 είναι πιο θετική από την κοινότητα 45. Μπορούμε να το διαπιστώσουμε από την αριστερή γραφική παράσταση στο πιο κάτω σχήμα που αναπαρίστανται τα συναισθήματα των δύο κοινοτήτων. Παρόλο που η πλειοψηφία των δημοσιεύσεων τους είναι θετικά προσκείμενες στα διάφορα θέματα υγείας όπως είναι η εφαρμογή εμβολιασμού και τα διάφορα εμβόλια, εντούτοις η κοινότητα 5 έχει λιγότερη απόκλιση μεταξύ θετικών και αρνητικών 65% έναντι 35% σε σχέση με την κοινότητα 45 όπου τα ποσοστά είναι 58% έναντι 42%. Άρα με λίγα λόγια υπάρχει ένα αρκετά σημαντικό ποσοστό 10% που χωρίζει τα θετικά συναισθήματα των δύο κοινοτήτων.

Συγκεκριμένα, θα μπορούσαμε να το δικαιολογήσουμε παρατηρώντας την γραφική στα δεξιά. Σε αυτήν σκιαγραφούνται και πάλι τα συναισθήματα αλλά αυτήν την φορά για ένα θέμα του άξονα υγείας που είναι ο εμβολιασμός. Βλέπουμε ότι το μοτίβο που ακολουθείται είναι το ίδιο με τα ποσοστά να είναι σχεδόν τα ίδια. Το μεγάλο ποσοστό 65% υπέρ των εμβολίων γίνεται εμφανές και από τις δημοσιεύσεις: «about 9/10 americans followed through on getting their 2nd dose of the vaccine. really encouraging news and a testament to our progress.», «more than three months into the u.s. vaccination drive, many of the numbers are encouraging.»



Σχήμα 5.4: Άξονας υγείας για το δίπολο (5,45)

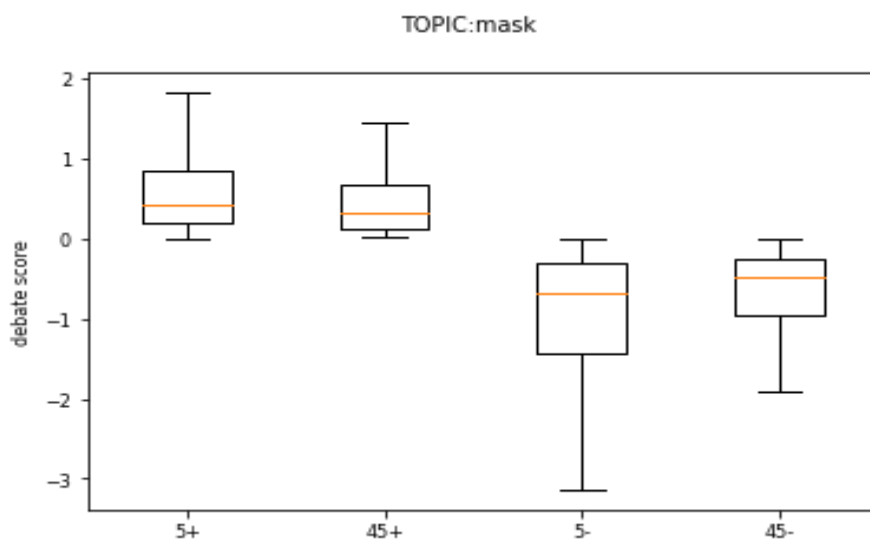
Όσο αφορά τον άξονα πρόληψης και διαχείρισης της πανδημίας, φαίνεται μέσω της πιο κάτω γραφικής παράστασης που βρίσκεται στα αριστερά και αναπαριστά την μετρική της βαρύτητας των λέξεων ότι η κοινότητα 5 έχει μεγαλύτερο αρνητικό σθένος. Παρόλο που και η κοινότητα 45 χρησιμοποιεί περισσότερο αρνητικές παρά θετικές λέξεις όταν αναφέρεται σε θέματα αυτού του άξονα που έχουν να κάνουν με τα μέτρα κατά τις πανδημίας (αφού οι αρνητικές τιμές φαίνονται να εκτείνονται περισσότερο από τις θετικές), για την κοινότητα 5 φαίνεται ξεκάθαρα ότι γίνεται σε μεγαλύτερο βαθμό. Στο ίδιο μοτίβο κινείται και η μετρική του debate score, όπως φαίνεται στην γραφική παράσταση του πιο κάτω σχήματος στα δεξιά, δείχνοντας ότι και στις δύο κοινότητες κυριαρχούν οι αρνητικές τιμές (ειδικά στην κοινότητα 5) που σημαίνει ότι υπάρχουν διαφωνίες και αντίθετες απόψεις για τα θέματα αυτά.



Σχήμα 5.5: Άξονας πρόληψης και διαχείρισης για το δίπολο (5,45)

Εμβαθύνοντας στην πιο πάνω υπόθεση, μπορούμε να κοιτάξουμε σε ένα από τα θέματα κάτω από τον άξονα πρόληψης και διαχείρισης. Συγκεκριμένα, το πιο κάτω σχήμα απεικονίζει το debate score όσο αφορά την χρήση της μάσκας.

Παρατηρώντας τις δημοσιεύσεις σε επίπεδο κειμένου μπορούμε και πάλι να διακρίνουμε ότι σε θέματα που αφορούν την μάσκα στην κοινότητα 5 φαίνεται να μην υπάρχει συμφωνία, όπως συμβαίνει και με την κοινότητα 45, με την διαφορά όμως ότι οι απόψεις είναι πιο έντονες. Κάποιες από αυτές ήταν: «wear your damn masks! our communities are the most impacted! and trump supporters will continue to follow his lead like a cult. we have to protect ourselves from the virus and idiots!», «ignorant, self-centered and selfish” is how one family describes people who won't wear masks in an obituary.». Στην αντίπερα όχθη: «mask gestapo.», «protesters in idaho staged a mask burning event at the state capital on saturday ... many protesters, often with children in tow, displayed signs reading “i will not self-suffocate.”»

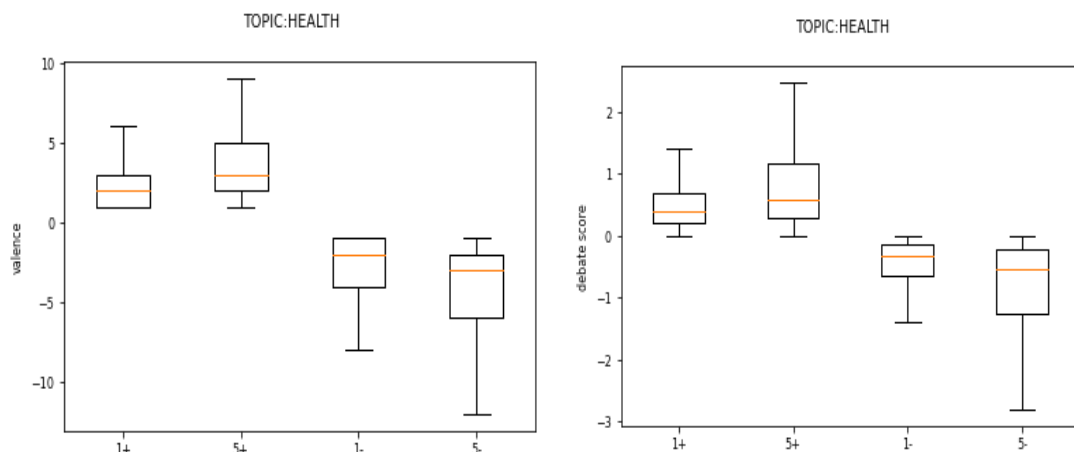


Σχήμα 5.6: Χρήση της μάσκας για το δίπολο (5,45)

Όσο αφορά τον τομέα της οικονομίας, δεν υπήρχαν θέματα τα οποία να είναι κοινά στις δύο κοινότητες. Η αιτία ήταν ότι στην κοινότητα 45 δεν υπήρχαν καθόλου θέματα που να έχουν να κάνουν με την οικονομία της χώρας.

5.1.3 Δίπολο (1,5)

Για τον άξονα υγείας, έχουμε και πάλι δύο γραφικές παραστάσεις, μια γραφική για το σθένος των λέξεων που χρησιμοποιούνται σε θέματα υγείας των δύο κοινοτήτων και άλλη μια που δείχνει το debate score. Όσο αφορά την κοινότητα 1, φαίνεται να υπάρχει μια ισορροπία εφόσον δεν μπορούμε να διακρίνουμε εάν το σθένος των λέξεων είναι περισσότερο αρνητικό ή θετικό και έτσι και το debate score ακολουθεί το ίδιο μοτίβο. Η κοινότητα 5, όμως, φαίνεται να χρησιμοποιεί λέξεις που έχουν περισσότερο αρνητική χροιά. Το debate score που αφορά την κοινότητα 5 δεν δείχνει ξεκάθαρα ότι κυριαρχούν οι θετικές τιμές έναντι των αρνητικών ή το αντίθετο αλλά διακρίνουμε γενικά να έχει μεγαλύτερο εύρος τιμών που δείχνει ότι δεν υπάρχει τόση σταθερότητα στις απόψεις.

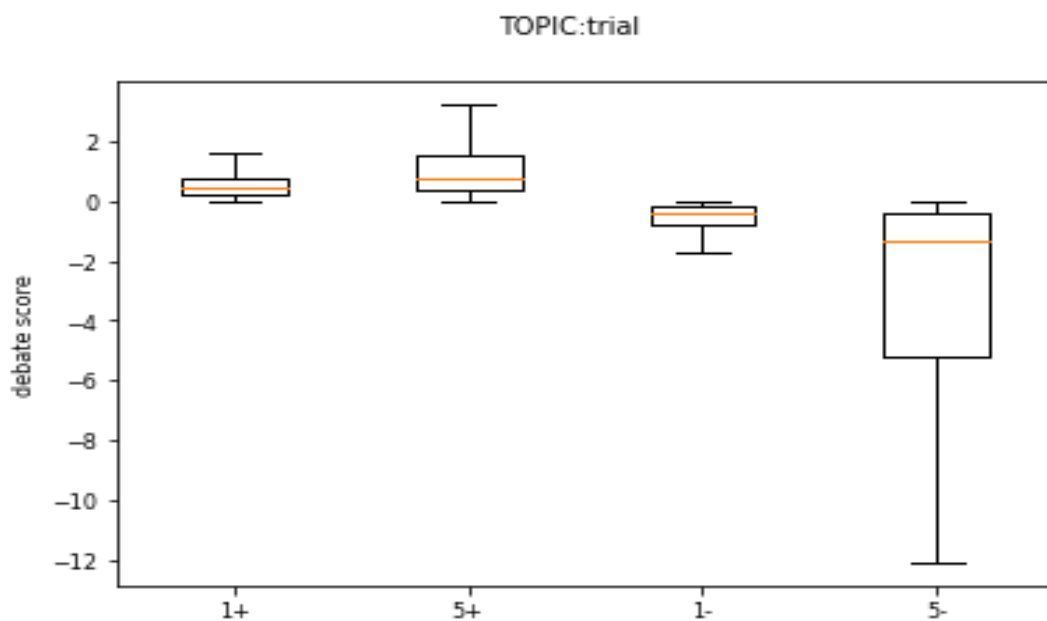


Σχήμα 5.7: Άξονας υγείας για το δίπολο (1,5)

Να υπενθυμίσουμε ότι και στις δύο κοινότητες βρέθηκαν δημοσιεύσεις όπου κατηγορούνται συγκεκριμένα εμβόλια για προβλήματα υγείας. Η κοινότητα 1 συζητούσε ότι θρομβώσεις προκαλούνται από τα εμβόλια AstraZeneca ενώ και τα δύο συμφωνούσαν στο ότι οι θρομβώσεις είναι αποτέλεσμα του εμβολίου Johnson&Johnson. Συγκεκριμένα, τα εμβόλια που συζητιούνταν από κοινού και τα ποσοστά υποστήριξης για τις κοινότητες 1 και 5 αντίστοιχα είναι τα εξής:

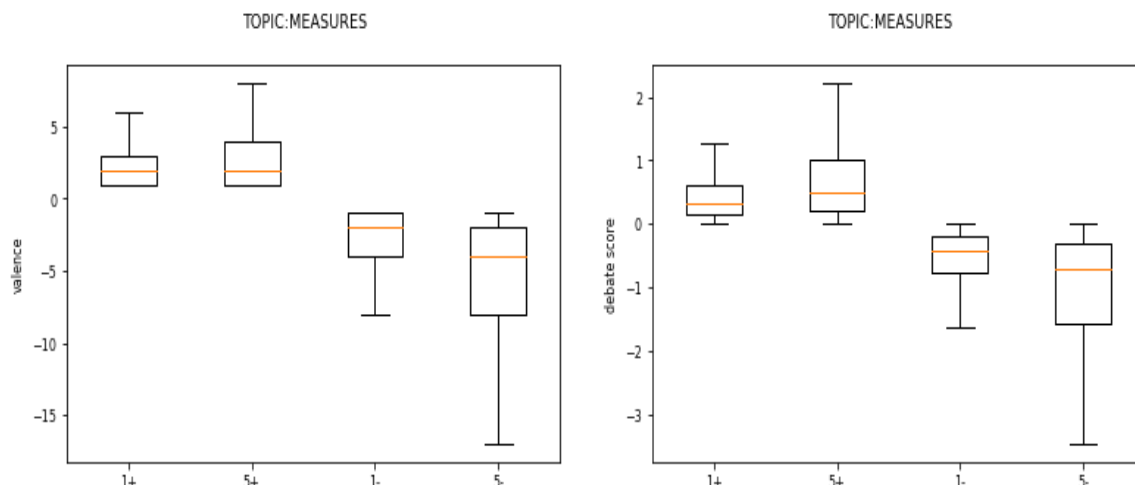
- Pfizer, με ποσοστό υποστήριξης 65% και 70%,
- Moderna, με ποσοστό υποστήριξης 64% και 70%
- Johnson&Johnson, με ποσοστό υποστήριξης 57% και 55%.

Ο λόγος που θα μπορούσε να δικαιολογήσει το γιατί υπάρχει τόση μεγάλη διαφορά στις απόψεις της κοινότητας 5 αλλά σχετική σταθερότητα στην κοινότητα 1 είναι το θέμα της δοκιμής των εμβολίων (human trial) το οποίο υπάγεται στον άξονα της υγείας. Στο πιο κάτω σχήμα διακρίνουμε και πάλι σχετική σταθερότητα στις απόψεις της κοινότητας 1. Κάτι το οποίο δεν ισχύει για την κοινότητα 5 αφού είναι φανερό ότι οι αρνητικές τιμές του debate score είναι πιο μεγάλες από τις αντίστοιχες θετικές και άρα υπάρχει μεγάλη διαφωνία όσο αφορά τις απόψεις γύρω από το θέμα της δοκιμής των εμβολίων στους ανθρώπους.



Σχήμα 5.8: Θέμα δοκιμής εμβολίων για το δίπολο (1,5)

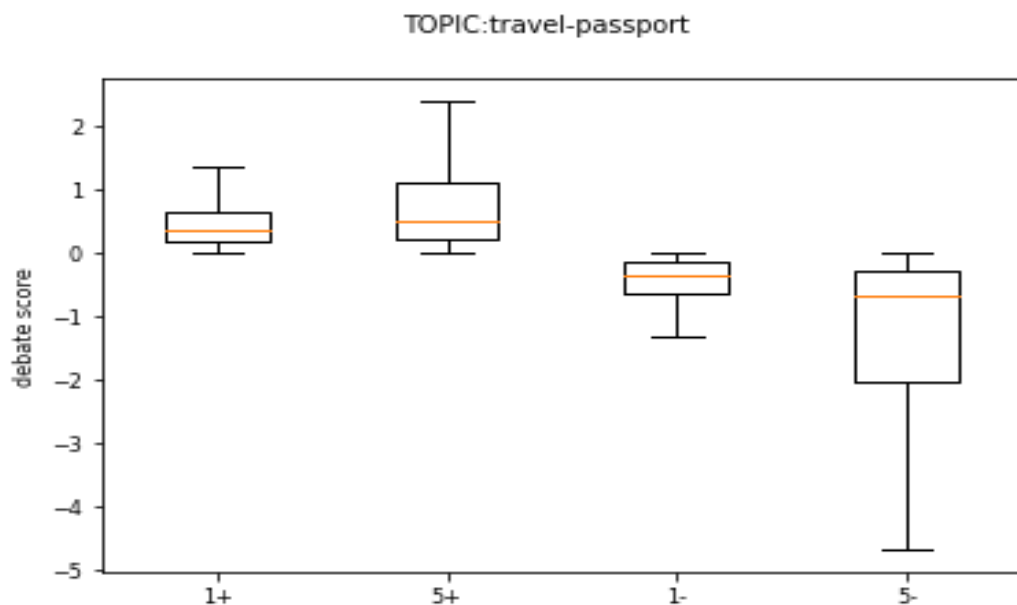
Όσο αφορά τον άξονα αντιμετώπισης της πανδημίας, έχουμε και πάλι δύο γραφικές παραστάσεις, μια που απεικονίζει το σθένος των λέξεων που χρησιμοποιούνται και άλλη μια που δείχνει το debate score. Όσο αφορά την κοινότητα 1, φαίνεται να υπερτερούν για λίγο οι αρνητικές τιμές και το debate score να είναι κάπως πιο ισορροπημένο. Η κοινότητα 5, όμως, φαίνεται ξεκάθαρα πως χρησιμοποιεί λέξεις με αρνητική σημασία και έτσι και είναι πιο εύκολο να διακρίνουμε ότι το debate score γέρνει προς τις αρνητικές τιμές που δείχνουν διαφωνία. Επίσης φαίνεται ότι υπάρχει μεγαλύτερο εύρος τιμών που δείχνει ότι δεν υπάρχει τόση σταθερότητα στις απόψεις.



Σχήμα 5.9: Άξονας μέτρων και πρωτοκόλλων για το δίπολο (1,5)

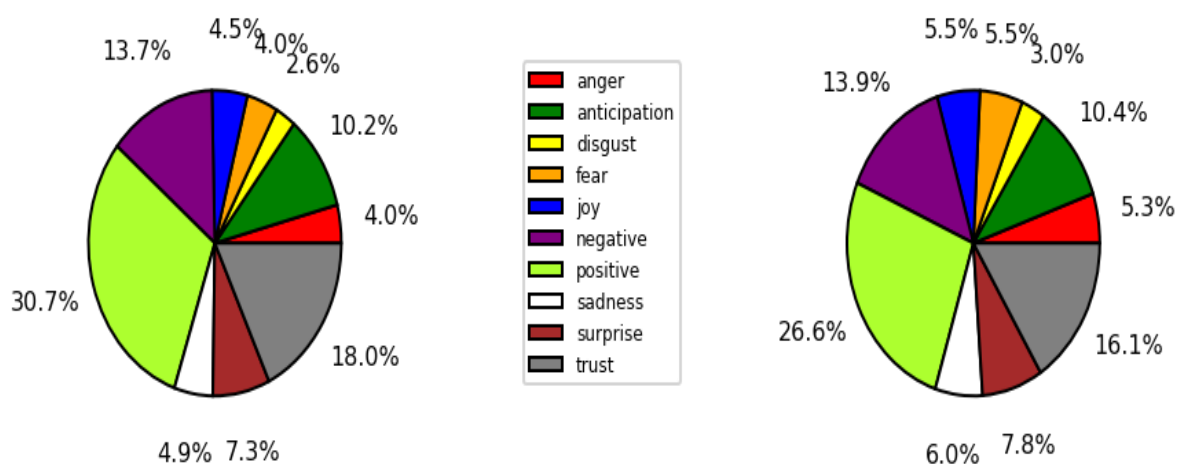
Κάτι το οποίο μπορεί να εξηγήσει γιατί υπάρχει τόσο μεγάλη διαφορά στις απόψεις της κοινότητας 5 και διαφωνία αλλά σχετική σταθερότητα στην κοινότητα 1 είναι το θέμα των ταξιδιών και της έκδοσης πιστοποιητικού εμβολιασμού για οποιοδήποτε θέλει να ταξιδέψει έξω από την χώρα του. Το θέμα αυτό είναι κάτω από την ομπρέλα του άξονα των πρωτοκόλλων ενάντια στην πανδημία. Σύμφωνα με την πιο κάτω γραφική παράσταση βλέπουμε και πάλι σχετική σταθερότητα στις απόψεις της κοινότητας 1. Κάτι το οποίο δεν συμβαίνει για την κοινότητα 5 αφού παρατηρείται ότι οι αρνητικές τιμές του debate score είναι πιο μεγάλες από τις αντίστοιχες θετικές και άρα υπάρχει μεγάλη διαφωνία όσο αφορά τις απόψεις γύρω από αυτό το θέμα.

Μπορούμε να επαληθεύσουμε όλα τα συμπεράσματα για τον συγκεκριμένο άξονα και από τα δεδομένα σε επίπεδο κειμένου. Η κοινότητα 5 μέσα από το κείμενο διχάζεται και εκφράζονται αντίθετες απόψεις αφού από την μια κάποιοι καταδικάζουν την έκδοση πιστοποιητικού εμβολιασμού ως υποχρεωτική συνθήκη για να κάνει κάποιος ταξίδι ενώ από την άλλη εκφράζεται το γεγονός ότι οι άνθρωποι θα έπρεπε να περιορίσουν τα ταξίδια καθώς συμβάλουν στην διασπορά του covid.



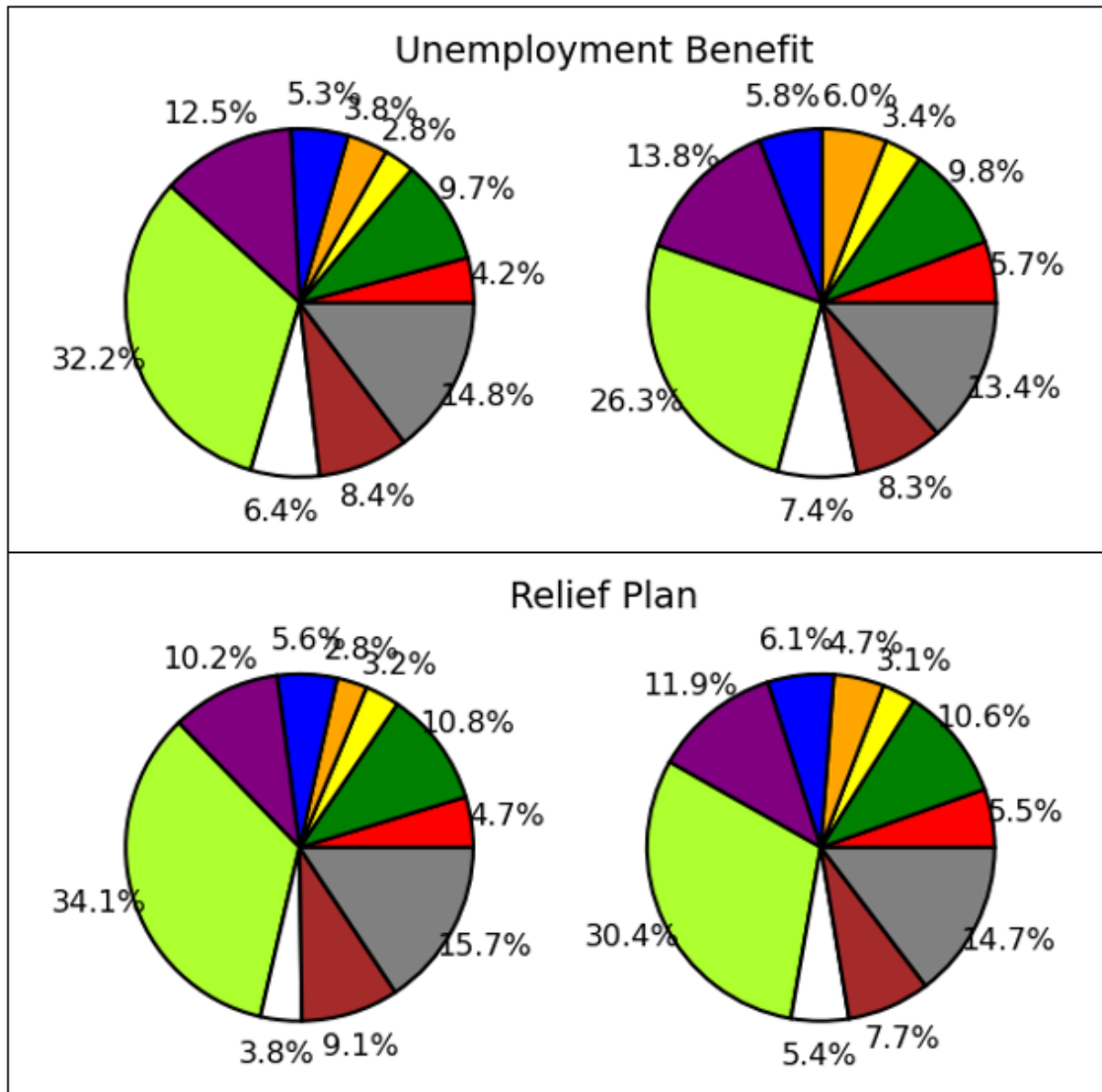
Σχήμα 5.10: Θέμα ταξιδιών και χρήσης πιστοποιητικού εμβολιασμού για την διεκπεραίωσή τους για το δίπολο (1,5)

Όσο αφορά τον άξονα της οικονομίας, έχουμε τις γραφικές των συναισθημάτων, στα αριστερά την γραφική της κοινότητας 1 και στα δεξιά την γραφική της κοινότητας 5. Για αυτόν τον άξονα δεν φαίνεται να υπάρχει κάποια διαφωνία αλλά οι δύο κοινότητες να κινούνται στο ίδιο μοτίβο. Και οι δύο δείχνουν ενθουσιασμό και εμπιστοσύνη με την κοινότητα 1 να δηλώνει ελαφρώς πιο θετικά συναισθήματα σαν σύνολο.



Σχήμα 5.11: Άξονας οικονομίας για το δίπολο (1,5)

Η πιο κάτω παρατήρηση οφείλεται στο γεγονός ότι και οι δύο κοινότητες σε επίπεδο κειμένου συζητούν για οικονομικό πλάνο το οποίο θα βοηθήσει την χώρα να ανακάμψει αλλά και σε επιδόματα για ανέργους.



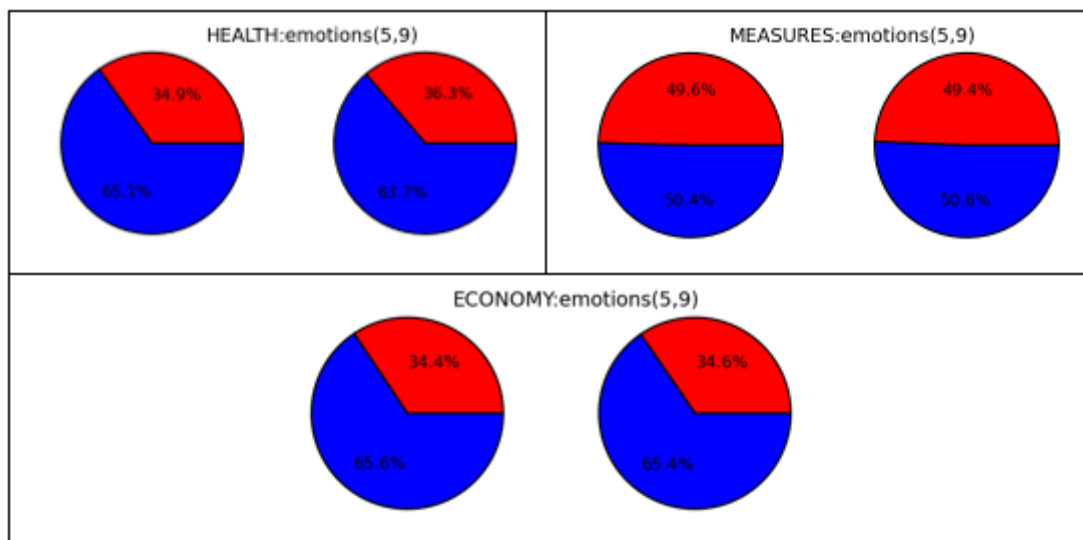
Σχήμα 5.12: Θέματα που σχετίζονται με την οικονομία για το δίπολο (1,5)

5.1.4 Δίπολο (5,9)

Το συγκεκριμένο δίπολο είναι το μόνο που σε σχέση με τα προηγούμενα τρία που έχουν αναλυθεί έχει αρνητική πόλωση με τιμή -0.411. Οπότε, θα ήταν μια καλή πρακτική να δούμε στους τομείς πως αυτό διαφέρει και δείχνει την ομοφωνία μεταξύ των δύο κοινοτήτων.

Κοιτάζοντας στα συναισθήματα των χρηστών των δύο κοινοτήτων αυτήν την φορά μπορούμε να δούμε ότι είναι πανομοιότυπα. Δηλαδή όσο αφορά τον τομέα υγείας, η δύο κοινότητες έχουν την ίδια στάση αφού το 65% των δημοσιεύσεων της κοινότητας 5 και το 64% της κοινότητας 9 έχουν θετικά σχόλια για θέματα τα οποία έχουν να κάνουν με τα διάφορα εμβόλια, τις δοκιμές των εμβολίων και γενικά την εφαρμογή εμβολιασμού και εκφράζουν θετικά συναισθήματα.

Όσο αφορά τον άξονα αντιμετώπισης της πανδημίας και οι δύο κοινότητες έχουν ένα ποσοστό της τάξης του 50% των θετικών συναισθημάτων. Παρομοίως, για τον οικονομικό τομέα φαίνεται να υπάρχει μεγάλη αισιοδοξία για την μελλοντική κατάσταση της οικονομία της χώρας αφού το ποσοστό και για τις δύο κυμαίνεται στο 65%.



Σχήμα 5.13: Τρεις άξονες για το δίπολο (5,9)

5.2 Ποσοτική έρευνα

Η ποσοτική έρευνα υπόδειξε ότι υπάρχει στατιστική διαφορά μεταξύ των τριών πρώτων διπόλων που περιγράφονται στο συγκεκριμένο κεφάλαιο αλλά όχι για το τέταρτο. Δηλαδή τα δίπολα (1,4) με τιμή πόλωσης 0.332, (5,45) με τιμή πόλωσης 0.282 και (1,5) με τιμή πόλωσης 0.457 σε επίπεδο γράφου όντως παρουσίαζαν πόλωση και σε επίπεδο

θεμάτων που συζητούσαν. Κάτι το οποίο δεν συμβαίνει με το δίπολο (5,9) με τιμή πόλωσης -0.411 που φάνηκε να μην έχει πόλωση ούτε σε επίπεδο κειμένου.

Ο τρόπος με τον οποίο εξετάστηκαν ήταν ο εξής: σε κάθε άξονα πήραμε τις τιμές των κοινοτήτων για κάθε μέτρηση (οι μετρήσεις ήταν το σθένος των λέξεων, το debate score και τα συναισθήματα) που εφαρμόσαμε και ελέγχουμε το p-value. Το p-value είναι η έξοδος από Mann-Whitney U Test που με αυτό μπορούμε να δούμε αν υπάρχει στατιστική σχέση μεταξύ των συνόλων που δώσαμε. Έτσι, αν το p-value ήταν μικρότερο ή ίσο του 5% (που αντιπροσωπεύει το ρίσκο λάθους, δηλαδή να συμπεράνουμε ότι υπάρχει στατιστική διαφορά χωρίς αυτό να ισχύει) αυτό υποδηλώνει ότι οι απόψεις στο δίπολο δεν είναι όμοιες και υπάρχει σύγκρουση. Σε αντίθετη περίπτωση, όπως φάνηκε στο τέταρτο δίπολο, όπου το p-value ήταν μεγαλύτερο από 0.05 αυτό σήμαινε ότι οι απόψεις είναι όμοιες και παρουσιάζεται συμφωνία.

5.3 Μελλοντική έρευνα

Όσο αφορά το κομμάτι της μελλοντικής έρευνας, θεωρώ ότι θα μπορούσαν να γίνουν αλλαγές σε κάποια στάδια όπως φαίνονται στο αρχιτεκτονικό διάγραμμα του κεφαλαίου 3. Με τον όρο αλλαγές εννοούμε ότι θα μπορούσε ένα στάδιο να ακολουθήσει εντελώς διαφορετική προσέγγιση ή απλά να μεταβάλουμε σε ένα μικρό βαθμό κάποια από τα στοιχεία που το συνθέτουν. Αυτό θα είχε ως αποτέλεσμα να πάρουμε και άλλες πληροφορίες όσο αφορά την πόλωση στα κοινωνικά δίκτυα εν καιρώ της πανδημίας.

Όσο αφορά το στάδιο της συλλογής δεδομένων, αυτό θα μπορούσε να μεταβληθεί με πολλούς τρόπους. Όπως προαναφέρθηκε, τα δεδομένα που συλλέγηκαν ήταν δεδομένα από την πλατφόρμα του Facebook για το χρονικό διάστημα από τις 21 Ιανουαρίου 2020 (πρώτο κρούσμα στις Ηνωμένες Πολιτείες Αμερικής) μέχρι τις 21 Ιουνίου 2021. Συνεπώς, σε μελλοντικό χρόνο θα μπορούσαμε να συλλέξουμε δεδομένα και από την πλατφόρμα του Twitter. Επίσης, θα μπορούσαμε να μεταβάλλουμε την παράμετρο του χρόνου των δεδομένων με το να διευρύνουμε το διάστημα που περιγράφουν τα δεδομένα ή να συλλέξουμε δεδομένα για ένα συγκεκριμένο χρονικό διάστημα όπως για παράδειγμα εφαρμογή του πρώτου περιορισμού απαγόρευσης κυκλοφορίας.

Επίσης, κάποιος θα μπορούσε να δώσει περισσότερη έμφαση στην διάδοση ψευδών πληροφοριών στην περίοδο της πανδημίας. Αυτό είχε γίνει στα αρχικά στάδια της έρευνας με την βοήθεια της λίστας από την ιστοσελίδα Newsguard. Όμως, δεν σημαίνει κατ' ανάγκη ότι οι ψευδείς πληροφορίες έρχονται πάντα από κακόβουλες ιστοσελίδες.

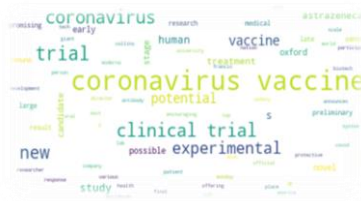
Τέλος, στα τελευταία στάδια της διπλωματικής μου εργασίας χρησιμοποιήθηκαν εργαλεία τα οποία είχαν επιλεγεί με κάποια κριτήρια, αφού πρώτα περάσαμε τα δεδομένα από διάφορες φάσεις καθαρισμού και προετοιμασίας. Στο μέλλον θα μπορούσαμε να εφαρμόσουμε και εναλλακτικές τεχνικές ανάλυσης δεδομένων για να μειώσουμε το θόρυβο όπως το να χρησιμοποιήσουμε μεθόδους τεχνητής νοημοσύνης αντί λεξικά.

Βιβλιογραφία

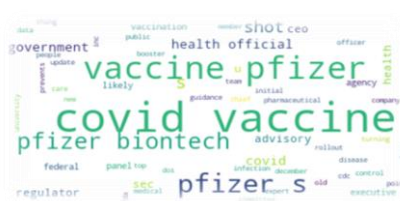
- [1] World Health Organization. WHO Coronavirus (COVID-19) Dashboard.
<https://covid19.who.int/>
- [2] IHME. COVID-19 Projections.
<https://covid19.healthdata.org/global/>
- [3] Adam D. (2022). The pandemic's true death toll: millions more than official counts.
<https://www.nature.com/articles/d41586-022-00104-8/>
- [4] Lepad N. (2020). Visualizing the History of Pandemics.
<https://www.visualcapitalist.com/history-of-pandemics-deadliest/>
- [5] Devlin J., Chang M. W., Lee K., Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
<https://arxiv.org/pdf/1810.04805.pdf>
- [6] Morales A. J., Borondo J., Losada J. C. (2015). Measuring Political Polarization: Twitter shows the two sides of Venezuela.
<https://aip.scitation.org/doi/pdf/10.1063/1.4913758>
- [7] Paschalides D., Pallis G., Dikaiakos M. D. (2021). POLAR: A Holistic Framework for the Modelling of Polarization and Identification of Polarizing Topics in News Media.
<https://dl.acm.org/doi/pdf/10.1145/3487351.3489443>
- [8] Garimella K., Morales G. F., Gionis A., Mathioudakis M. (2018). Quantifying Controversy on Social Media.
<https://dl.acm.org/doi/pdf/10.1145/3140565>

- [9] Guerra P. C., Cardie C. (2013). A Measure of Polarization on Social Media Networks Based on Community Boundaries.
<https://www.cs.cornell.edu/home/cardie/papers/ICWSM13-Polarization.pdf>
- [10] PEW RESEARCH CENTER. (2020). As COVID-19 Emerged in U.S., Facebook Posts About It Appeared in a Wide Range of Public Pages, Groups.
<https://www.pewresearch.org/journalism/2020/06/24/as-covid-19-emerged-in-u-s-facebook-posts-about-it-appeared-in-a-wide-range-of-public-pages-groups/>
- [11] Newsguard. Coronavirus Misinformation Tracking Center.
<http://web.archive.org/web/20200316022654/https://www.newsguardtech.com/coronavirus-misinformation-tracking-center/>
- [12] Crowdtangle API.
<https://help.crowdtangle.com/en/articles/1189612-crowdtangle-api>
- [13] Hugging Face. RoBERTa large model.
<https://huggingface.co/roberta-large>
- [14] Xu, D., Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms.
<https://link.springer.com/article/10.1007/s40745-015-0040-1>

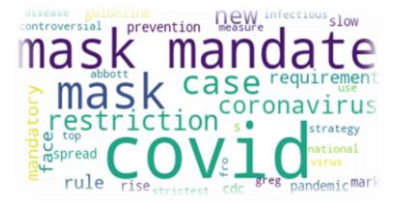
Παράρτημα Α



(α) Διαδικασία εμβολιασμού



(β) Εμβόλια Pfizer



(γ) Προστατευτική μάσκα



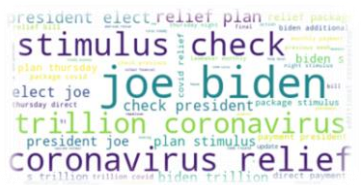
(δ) Περιστατικό κορονοϊού



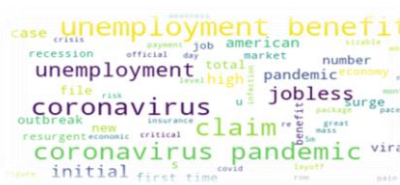
(ε) Πιστοποιητικό εμβολιασμού «covid passport»



(ζ) Περιορισμός μετακινήσεων

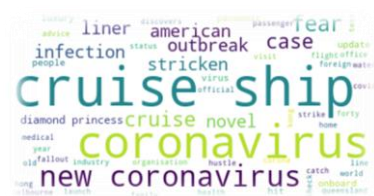
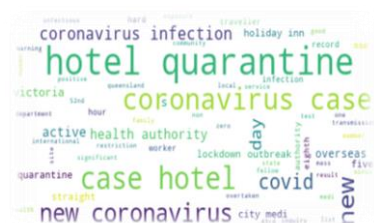
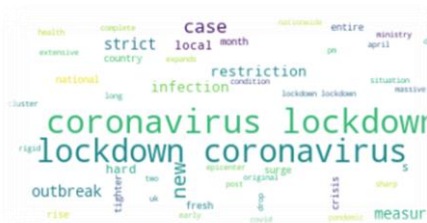
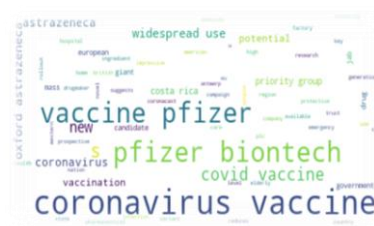
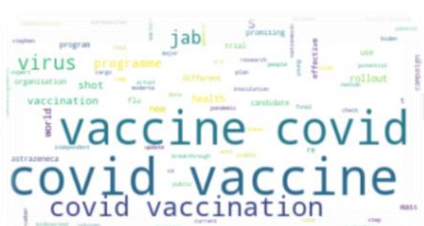
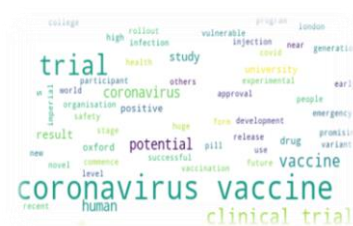


(η) Οικονομικό πλάνο



(θ) Επίδομα ανεργίας

Σχήμα Α1: Επιπρόσθετα θέματα για την κοινότητα 1.



Σχήμα Α2: Θέματα για την κοινότητα 2.



(α) Ανθρώπινη επαφή



(β) Προστατευτική μάσκα



(γ) Συμπτώματα



(δ) Επίδομα ανεργίας



(ε) Υγιεινή και προστασία



(ζ) Περιορισμοί



(η) Διαδικτυακά μαθήματα



(θ) Δάνειο προς μαθητές



(ι) Παραπληροφόρηση

Σχήμα Α5: Θέματα για την κοινότητα 24.