

# Μελέτη πανδημίας - αρνητικού ύφους έναντι των εμβολίων κατά του κορονοϊού στο κοινωνικό δίκτυο του Twitter με τεχνικές μηχανικής μάθησης



Capstone project

Δημήτριος Καρμούρης ΠΚΜ 202029

Επόπτης Καθηγήτρια Κάτια – Λήδα Κερμανίδου

Μέλος Εμμανουήλ Μαραγκουδάκης

Μέλος Φοίβος Μυλωνάς

# ΠΕΡΙΕΧΟΜΕΝΑ

|                                                    |    |
|----------------------------------------------------|----|
| Εισαγωγή .....                                     | 3  |
| Η Επιστήμη των Δεδομένων .....                     | 5  |
| Η πλατφόρμα του Twitter .....                      | 6  |
| Ρόλος twitter κατά τη διάρκεια της πανδημίας ..... | 8  |
| Μεθοδοι & αποτελέσματα .....                       | 8  |
| Προετοιμασία & αρχική αναλυση δεδομένων .....      | 8  |
| Επεξεργασία δεδομένων .....                        | 9  |
| Οπτικοποίηση των δεδομένων .....                   | 18 |
| Ετικετοσύννεφο – word cloud .....                  | 18 |
| Δημιουργία κανόνων συσχέτισης .....                | 21 |
| Στο σύνολο των δεδομένων .....                     | 21 |
| Στα παραδείγματα με αρνητικό ύφος.....             | 25 |
| Ομαδοποίηση – συσταδοποίηση (clustering) .....     | 27 |
| Δέντρα απόφασης .....                              | 28 |
| Random Forest .....                                | 29 |
| Κατηγοριοποίηση – Ταξινόμηση .....                 | 31 |
| Συμπεράσματα .....                                 | 35 |
| Βιβλιογραφία .....                                 | 36 |
| <br>                                               |    |
| Εικόνα 1. Twitter & χρήστης.....                   | 4  |
| Εικόνα 2. Big Data.....                            | 5  |
| Εικόνα 3. Σετ δεδομένων -αρχείο CSV- .....         | 9  |
| Εικόνα 4. Tweets - αριθμός χρηστών .....           | 11 |
| Εικόνα 5. Followers χρηστών .....                  | 12 |
| Εικόνα 6. Source - length.....                     | 13 |
| Εικόνα 7. Source – Compound - Sentiment.....       | 13 |
| Εικόνα 8. Location - Sum of Compound .....         | 15 |

|                                                                              |    |
|------------------------------------------------------------------------------|----|
| Εικόνα 9. Μήκος λέξεων tweets .....                                          | 16 |
| Εικόνα 10. Ημερομηνία - Compound .....                                       | 17 |
| Εικόνα 11. Συχνότητες λέξεων για τη δημιουργία ετικετοσύννεφου .....         | 18 |
| Εικόνα 12. Word cloud (ετικετοσύννεφο) .....                                 | 19 |
| Εικόνα 13. Κανόνες συσχέτισης .....                                          | 22 |
| Εικόνα 14. Κανόνες συσχέτισης αρνητικού ύφους .....                          | 26 |
| Εικόνα 15. Ομαδοποίηση (clustering) 3 συστάδων επάνω και 5 κάτω εικόνα ..... | 28 |
| Εικόνα 16. Δέντρο απόφασης (decision tree) .....                             | 29 |
| Εικόνα 17. Random Forest -Παρατίθενται 4 από τα 30 δέντρα .....              | 30 |
| Εικόνα 18. Random Forest – επιπλέον 2 από 30 δέντρα .....                    | 31 |
| Πίνακας 1. Association Rules – Confidence descending.....                    | 23 |
| Πίνακας 2. Association Rules - Support descending.....                       | 24 |
| Πίνακας 3. Association Rules - Lift descending.....                          | 25 |
| Πίνακας 4. Μετρική Accuracy διαφόρων μοντέλων. ....                          | 32 |
| Πίνακας 5. Confusion Matrix Gradient Boosted Trees .....                     | 32 |
| Πίνακας 6. Confusion Matrix Random Forest .....                              | 32 |
| Πίνακας 7. Confusion Matrix Decision Tree .....                              | 33 |
| Πίνακας 8. Confusion Matrix Fast Large Margin.....                           | 33 |
| Πίνακας 9. Confusion Matrix Generalized Linear Model .....                   | 33 |
| Πίνακας 10. Confusion Matrix Naive Bayes .....                               | 33 |



## ΕΙΣΑΓΩΓΗ

Στα τέλη του 2019 η ανθρωπότητα έρχεται αντιμέτωπη με έναν νέο ιό, μια νέα παγκόσμια πρωτοφανής για τη μέχρι πρότινος γνωστή ανθρώπινη ιστορία απειλή. Πολύ γρήγορα η εξάπλωσή του πήρε διαστάσεις πανδημίας όπως και χαρακτηρίστηκε το Μάρτιο του 2020, επισήμως από τον Παγκόσμιο Οργανισμό Υγείας (WHO-Director, 2020). Πρόκειται για τον κορονοϊό SARS-CoV-2 που προκαλεί την ασθένεια COVID-19 που είναι υπεύθυνη για την τρέχουσα πανδημία (Wikipedia, χ.χ.). Έκτοτε εξαπλώθηκε σε όλες τις χώρες του κόσμου απαριθμώντας συνεχώς νέες μολύνσεις με πρωτοφανείς ρυθμούς αύξησης των κρουσμάτων.

Από τα αρχικά στάδια εντοπισμού του ιού στην πόλη Γιουχάν της Κίνας παρατηρήθηκε μεταξύ άλλων και η προσπάθεια ενημέρωσης της -παγκόσμιας- κοινής γνώμης για τον άγνωστο μέχρι τότε ιό μέσω των μέσων κοινωνικής δικτύωσης. Σημαντική θέση ανάμεσα στα social media είχε (και έχει) το Twitter. Όλοι ίσως έχουμε ακούσει το παράδειγμα ενός ιατρού του Li Wenliang από το νοσοκομείο της πόλης Γιουχάν που είναι ο πρώτος γιατρός που αντιλήφθηκε τον νέο ιό και με τη χρήση των μέσων κοινωνικής δικτύωσης προσπάθησε να προειδοποιήσει τον κόσμο για την ύπαρξη αυτού του νέου ιού (Hegarty, 2020). Όπως θα δούμε και παρακάτω η πλατφόρμα του Twitter καταλαμβάνει εξέχουσα θέση ανάμεσα στα υπόλοιπα social media όσο αναφορά στη μετάδοση πληροφοριών και ειδήσεων.

Η παγκόσμια αντίδραση εναντίον του κορονοϊού εκδηλώθηκε με αλληπάλληλα lockdown για τον περιορισμό της μετάδοσης του καθώς δεν υπήρχε άλλο μέτρο αντιμετώπισής. Όταν στη φαρέτρα μας τις αρχές του 2021 προστέθηκε το όπλο των εμβολίων ενάντια στον ιό τότε παρατηρήθηκε μια έντονη δυσανεμία της κοινής γνώμης έναντι των εμβολίων και της υποχρεωτικότητας που προσπαθούσαν ορισμένες κυβερνήσεις να εφαρμόσουν.

Βέβαια σε όλο αυτό το διάστημα που μεσολάβησε ουδέποτε ελαττώθηκε η στροφή των ανθρώπων στα μέσα κοινωνικής δικτύωσης αλλά θα λέγαμε ότι γιγαντώθηκε. COVID-19 και εμβόλια ή εμβολιασμός είναι λέξεις κλειδιά που όλο αυτό το διάστημα συγκεντρώνουν τις περισσότερες αναρτήσεις καθώς και το ενδιαφέρον των χρηστών. Αν και αυτό έχει τα θετικά στοιχεία όπως την άμεση ενημέρωση και τη χάραξη κοινής στρατηγικής και πορείας ενάντια στον ιό δεν έλειψαν και οι προσπάθειες χειραγώγησης και παραπληροφόρησης της κοινής γνώμης (Eibensteiner F, 2021).

Μπορούμε κάλλιστα να ισχυριστούμε ότι είναι εμφανής η στροφή του κόσμου στα social media για να μοιραστεί πληροφόρηση ή και παραπληροφόρηση!

Πηγή έμπνευσης για τη συγκεκριμένη μελέτη στάθηκε μια παγκόσμια έρευνα σχετικά με ενδεχόμενη υιοθέτηση ενός εμβολίου κατά της COVID-19. Σύμφωνα με τα αποτελέσματα της ο πληθυσμός 19 χωρών, σε υψηλά ποσοστά θα ήταν δεκτικός να κάνει και θα έβλεπε θετικά ένα εμβόλιο εναντίον του κορονοϊού αν αυτό ήταν απολύτως ασφαλές. Ενδιαφέρον στα αποτελέσματα παρουσίασε το γεγονός ότι σε χώρες όπου οι πολίτες εμπιστεύονταν τις κυβερνήσεις και τα καθεστώτα υπήρχαν και τα υψηλότερα ποσοστά δεκτικότητας καθώς οι πολίτες ήταν πιο δεκτικοί σε υποχρεωτικότητα του εμβολιασμού ακόμα και αν αναγκάζονταν

από τον εργοδότη τους. Αν σκεφτούμε τη διαφορετικότητα των κοινωνιών τότε αντιλαμβανόμαστε τη σπουδαιότητα της συγκεκριμένης έρευνας καθώς βοηθά στο να γίνει πιο κατανοητό το γεγονός ότι κάθε κυβέρνηση και κάθε κοινωνία πρέπει να προσεγγίσει το ευαίσθητο αυτό θέμα της πανδημίας και του εμβολιασμού κατ' επέκταση, εντελώς διαφορετικά. Επιβάλλεται να βρει τη χρυσή τομή ανάμεσα σε μια κοινή συνισταμένη για την καταπολέμηση της πανδημίας και στην εξάλειψη των αρνητικών συναισθηματικών επιπτώσεων που φόρτισε ολόκληρη την ανθρωπότητα στην γνωστή ιστορία της (Lazarus JV, 2021).

Έχοντας λοιπόν κατά νου τα ανωτέρω, σε συνδυασμό με τη δημόσια «αντιπολίτευση» και τις επικριτικές τάσεις που δέχτηκε ο εμβολιασμός κατά του κορονοϊού, καθώς και τις πολλαπλές μελέτες που έχουν γίνει σχετικά με τα αρνητικά συναισθήματα που προκλήθηκαν από την πανδημία, έγινε μια προσπάθεια βαθύτερης κατανόησης των αιτιών αυτών. Συγκεκριμένα προσπαθήσαμε να εντοπίσουμε από που πηγάζουν τα αρνητικά συναισθήματα κατά του εμβολίου στην ελληνική κοινωνία παρά το γεγονός ότι κρίθηκε απαραίτητη η δημιουργία και η χρήση ενός εμβολίου για τον «έλεγχο» και κατευνασμό όλων των αρνητικών επιπτώσεων της πανδημίας. Το γνωστό σε όλους μας τείχος ανοσίας! Με αφορμή λοιπόν τη μεγάλη αβεβαιότητα που προκλήθηκε στην παγκόσμια κοινή γνώμη σχετικά με την ασφάλεια ή μη των εμβολίων κατά της COVID-19 που κυκλοφόρησαν το 2021 παρατηρήθηκε μια αρνητική τάση στο ευρύ κοινό σχετικά με τον εμβολιασμό. Όλο αυτό όπως ήταν αναμενόμενο πήρε μεγάλες διαστάσεις στα μέσα κοινωνικής δικτύωσης. Μιας και οι πεποιθήσεις και αντιλήψεις των περισσότερων επηρεάζονται σε μεγάλο βαθμό από τα μέσα κοινωνικής δικτύωσης (social media) μπορούμε να υποθέσουμε ότι ένα στιγμιότυπο μιας τέτοιας πλατφόρμας μπορεί να αποτελέσει μια εικόνα ενός κινούμενου στόχου.

Αντικείμενο της παρούσας εργασίας αποτελεί η λήψη ενός τέτοιου στιγμιότυπου όπου με τη βοήθεια αλγορίθμων μηχανικής μάθησης θα προσπαθήσουμε να το αναλύσουμε με σκοπό να εξάγουμε χρήσιμη πληροφορία για τον «κινούμενο στόχο» που στην περίπτωσή μας δεν είναι άλλος από το αρνητικό ύφος που επικρατεί στο κοινωνικό δίκτυο του Twitter σχετικά με τον εμβολιασμό.

**Λέξεις κλειδιά:** Twitter, εμβόλιο, παρενέργειες εμβολίου, εξόρυξη γνώσης, sentiment, μηχανική μάθηση, data analysis, εμβολιασμός, covid-19, RapidMiner Studio, αντιεμβολιαστές, social media.

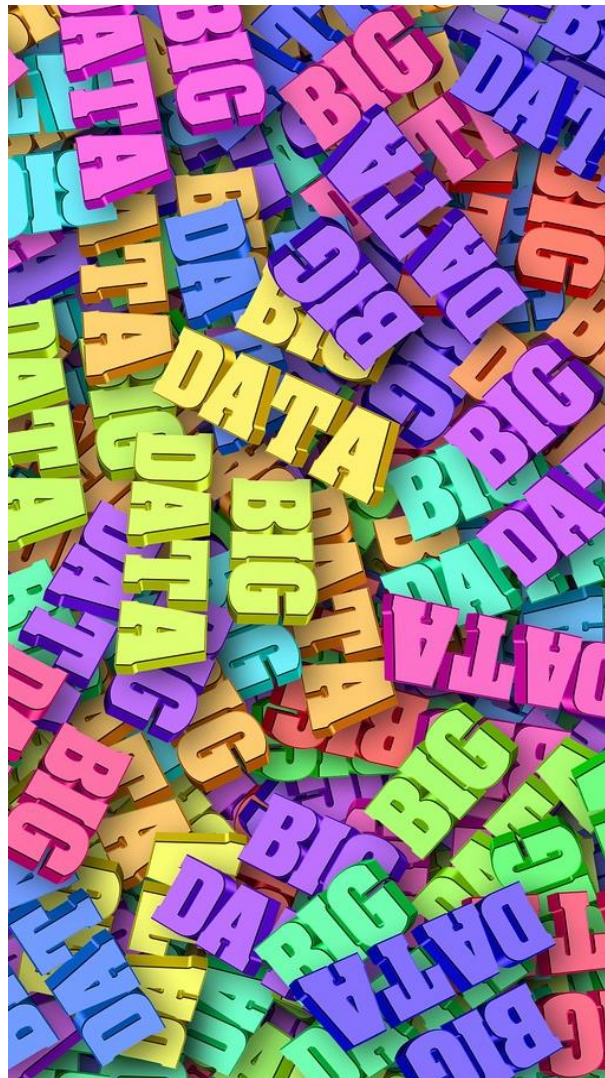


*Εικόνα 1. Twitter & χρήστης*

## Η ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Η Επιστήμη των Δεδομένων περιλαμβάνει τις τεχνολογίες και τεχνικές για την ανάλυση και επεξεργασία δεδομένων. Τα δεδομένα, τα οποία αναφέρονται ως το πετρέλαιο του 21ου αιώνα, είναι η βάση για τη Μηχανική Μάθηση. Είναι ο κλάδος της Τεχνητής Νοημοσύνης που πραγματεύεται τη μελέτη αλγορίθμων και τεχνικών που βελτιώνονται με αυτοματοποιημένο τρόπο χάρη στην αξιοποίηση δεδομένων τεράστιου όγκου που είναι αδύνατη, ή με οποιονδήποτε άλλο τρόπο επεξεργασία τους.

Η περιοχή της Επιστήμης Δεδομένων έχει ως αντικείμενο την εξαγωγή μη τετριμμένης και υποκρυπτόμενης γνώσης από μεγάλους όγκους δεδομένων (**data mining**). Κάνει εκτεταμένη χρήση αλγορίθμων, μηχανικής μάθησης και στατιστικής συμπερασματολογίας για την εξαγωγή γνώσης και προβλέψεων (IBM, 2021). Είναι απόρροια σημαντικών εξελίξεων σε δυο υποπεριοχές της πληροφορικής κατά τα τελευταία 15 χρόνια: Πρώτον εξαιτίας της σημαντική προόδου που σημειώθηκε σε αλγορίθμους και τεχνικές μηχανικής μάθησης και γενικότερα τεχνικές τεχνητής νοημοσύνης βασισμένες σε στατιστικές αρχές. Και δεύτερον ένεκα των ραγδαίων εξελίξεων στην περιοχή της διαχείρισης δεδομένων, που σε βελτίωση της ταχύτητας επεξεργασίας τεράστιων, ετερογενών, συνεχών μεταβαλλόμενων όγκων δεδομένων (Big Data). Οι σύγχρονες επεξεργαστικές δυνατότητες σε συνδυασμό με τους τεράστιους όγκους των δεδομένων οδήγησαν στις κατάλληλες προϋποθέσεις για την ανάπτυξη της συγκεκριμένης επιστήμης. (Επιστήμη Δεδομένων)



Εικόνα 2. Big Data

## Η ΠΛΑΤΦΟΡΜΑ ΤΟΥ TWITTER

*"Twitter is what's happening in the world and what people are talking about right now!"* Έχοντας αυτήν τη φράση κατά νου οι Jack Dorsey, Noah Glass, Biz Stone και Evan Williams το Μάρτιο του 2006 δημιούργησαν, τη σήμερα, δημοφιλή σε όλους μας, πλατφόρμα Twitter. Βέβαια ουδεμία σχέση είχε με την αρχική ιδέα της μητρικής εταιρείας "odeo" αλλά απ' ότι φαίνεται εκ των υστέρων η ιστορία τους δικαίωσε για την επιλογή τους αυτή. (Twitter-Inc., 2020). Το twitter αποτελεί τρανή απόδειξη ενός αποφθέγματος του George Orwell που λέει *"History is written by the winners"*.

Η ιδέα εμπνεύστηκε από την τεχνολογία του web-based instant messaging, που εμφανίστηκε στα τέλη του 20<sup>ου</sup> αιώνα και του ιδιαίτερος διαδεδομένου για την εποχή του text messaging (SMS) μέσω των δικτύων κινητής τηλεφωνίας. Το πάντρεμα αυτών των δύο σχετικά νέων τεχνολογιών δημιούργησε το Twitter. Με πιο απλά λόγια η ιδέα ήταν να μπορεί κάποιος να στείλει μέσω του Twitter ένα μήνυμα κειμένου με την ίδια ευκολία όπως ακριβώς στέλνει και ένα SMS. Γι' αυτό και υπήρξε εξαρχής περιορισμός των χαρακτήρων που μπορούν να περιέχονται σε ένα μήνυμα ενός χρήστη μέσω της πλατφόρμας, όπως ακριβώς αντίστοιχος περιορισμός ίσχυε και στα SMS (μέγιστο 140 χαρακτήρες). Βασική διαφοροποίηση αποτέλεσε το γεγονός ότι το μήνυμα μέσω της πλατφόρμας του Twitter μπορεί να διαβαστεί από περισσότερους του ενός παραλήπτες εν αντιθέση με τον έναν και μοναδικό παραλήπτη που θα είχε μέσω της συνήθους έως τότε αποστολής του από δίκτυο κινητής τηλεφωνίας. Σε αυτό κατά κύριο λόγο οφείλει και τη δημοφιλία του. (Wikipedia, n.d.)

Μέσω της πλατφόρμας του twitter οι χρήστες μπορούν να δημιουργούν, να μοιράζονται, να ανακαλύπτουν περιεχόμενο που συμβαίνει τώρα, με βάση τις θεματικές ενότητες, τις τάσεις και τα ενδιαφέροντα τους. Τελευταία στις δυνατότητες της πλατφόρμας έχουν προστεθεί και οι ζωντανές μεταδόσεις των χρηστών για το διαμοιρασμό περιεχομένου. Το Twitter λοιπόν, είναι μια πλατφόρμα κοινωνικής δικτύωσης και Microblogging όπου οι χρήστες μπορούν να δημοσιεύουν μικρά μηνύματα που λέγονται tweets, να αλληλοεπιδράσουν μεταξύ τους και να συνομιλούν σε άμεσο χρόνο. Η διαφορά του microblogging από το blogging έγκειται στο ότι το πρώτο επιτρέπει τη δημιουργία και ανταλλαγή περιεχομένου «μικρού μεγέθους» ανάμεσα στους χρήστες ενώ το δεύτερο είναι οι γνωστές σε όλους μας θεματικές ιστοσελίδες χρηστών όπου ο κάθε χρήστης μπορεί να γράψει ό,τι θέλει για το θέμα της σελίδας. Το twitter μέχρι και σήμερα παραμένει πιστό στο microblogging διατηρώντας μέγιστο αριθμό χαρακτήρων για κάθε tweet τους 280 χαρακτήρες. (About-Twitter, n.d.)

Σε αυτό το σημείο κρίνεται σκόπιμο να αναφέρουμε εν συντομία τον τρόπο που λειτουργεί η πλατφόρμα. Αυτό πέρα από τη δημοφιλία που του προσέδωσε θα μας βοηθήσει να κατανοήσουμε τους λόγους που επιλέχθηκε για τους σκοπούς της συγκεκριμένης εργασίας.

Αρχικά όλοι οι χρήστες του διαδικτύου έχουν πρόσβαση στην πλατφόρμα όπως αναφέρεται και στον ιστότοπο twitter.com. Το μόνο που χρειάζεται ένας χρήστης, από οποιοδήποτε μέρος της γης για να μπορεί να χρησιμοποιήσει το twitter είναι να έχει στη διάθεσή του σύνδεση στο

διαδίκτυο και μια υπολογιστική συσκευή (H/Y, smartphone κτλ). Επομένως όλοι οι χρήστες του διαδικτύου μπορούν να δουν και να διαβάσουν μηνύματα, αλλά μόνο οι εγγεγραμμένοι στην πλατφόρμα χρήστες, μπορούν να αλληλοεπιδράσουν με τα tweets μέσω αυτής (της πλατφόρμας) δηλαδή να δημοσιεύσουν μηνύματα.

Οι χρήστες επομένως έχουν τη δυνατότητα να εγγραφούν και αφού πιστοποιηθούν και δημιουργήσουν το προσωποποιημένο προφίλ τους να δημοσιεύουν μηνύματα (posts) μέσω της πλατφόρμας. Τα γνωστά tweets. Τα tweets μπορούν να περιέχουν εκτός από κείμενο και φωτογραφίες, βίντεο, ή και υπερσυνδέσμους (links). Τα συγκεκριμένα μηνύματα δεν μπορούν να επεξεργαστούν από τη στιγμή που θα δημοσιευτούν παρά μόνο να διαγραφούν.

Όπως έχουμε αναφέρει και παραπάνω τα tweets είναι δημόσια εξ' ορισμού και οι χρήστες που ακολουθούν προφίλ άλλων χρηστών μπορούν να δουν και να αλληλοεπιδράσουν *άμεσα* με τα tweets που δημοσιεύουν αυτοί οι χρήστες. Επίσης όποιος κάνει διαδικτυακή αναζήτηση με συγκεκριμένες λέξεις κλειδιά μπορεί να έχει πρόσβαση και σε ανάλογα μηνύματα στο twitter που περιέχουν τις συγκεκριμένες λέξεις κλειδιά. Ένα άλλο σημείο που χρειάζεται να γνωρίζουν οι χρήστες είναι ότι όλα τα tweets που δημοσιεύουν αποθηκεύονται και στο προφίλ τους. Δημοφιλή και εξαιρετικά διαδεδομένα κρίνονται και τα hashtags. Με τη χρήση τους σε λέξεις κλειδιά συγκεντρώνουν όλα tweets που σχετικά με τη λέξη κλειδί που έχει το hashtag.

Μια επιπλέον δυνατότητα που προσφέρει στους εγγεγραμμένους χρήστες είναι η επιλογή να εγγραφούν (subscribe) και να ακολουθούν άλλους εγγεγραμμένους χρήστες, το γνωστό σε όλους following. Αυτό διευκολύνει στην άμεση διάδοση των tweets καθώς βοηθά στην ενημέρωση των χρηστών που ακολουθούν να ενημερώνονται και να αλληλοεπιδρούν με τις ενημερώσεις (updates) του χρήστη που έχουν επιλέξει να ακολουθούν. Επιπροσθέτως μπορούν να αλληλοεπιδράσουν δημιουργώντας απάντηση (reply) σε tweets των άλλων χρηστών. Η απάντηση αποτελεί tweet και είναι πάντοτε δημόσια εν αντίθεση με τα άμεσα προσωπικά μηνύματα.

Όλα αυτά βέβαια υπόκεινται στους περιορισμούς που αναφέρονται στους όρους χρήσης που θέτει η εν λόγω πλατφόρμα. Οι εγγεγραμμένοι χρήστες δεσμεύονται κατά την εγγραφή τους με την υποχρέωση να σέβονται και να τηρούν καθ' όλη την διάρκεια που χρησιμοποιούν τη συγκεκριμένη πλατφόρμα τους όρους αυτούς, διαφορετικά γίνεται αναφορά και μπλοκάρισμα του λογαριασμού τους από τους διαχειριστές (administrators). (Twitter-Help-Center, n.d.)

Από μερικούς εκφράζεται η εσφαλμένη αντίληψη ότι δεν πρόκειται για ένα μέσο κοινωνικής δικτύωσης αλλά για ένα δίκτυο μετάδοσης πληροφορίας. Όπως αναφέραμε και παραπάνω δημιουργήθηκε με σκοπό το να μπορεί κάποιος να μοιραστεί νέα, πληροφορία, αντιλήψεις και λοιπά με περισσότερα άτομα και πιο άμεσα. Συνεπώς όπως πολύ εύστοχα αναφέρεται και στην ιστοσελίδα της πλατφόρμας *“Twitter is a way to learn about the world through another person's eyes”* (About-Twitter, χ.χ.).

## ΡΟΛΟΣ TWITTER ΚΑΤΑ ΤΗ ΔΙΑΡΚΕΙΑ ΤΗΣ ΠΑΝΔΗΜΙΑΣ

Από τα παραπάνω γίνεται εύκολα αντιληπτή η ικανότητα που έχει η συγκεκριμένη πλατφόρμα στο να μοιράζεται αμεσότητα και με μεγάλη απήχηση την πληροφορία που δημιουργείται από τους χρήστες αυτής καθώς και την ευκολία που παρέχει προς τους χρήστες για αυτό το σκοπό. Επίσης η χρήση του κατά τη διάρκεια της πανδημίας του κορονοϊού αυξήθηκε όπως σε όλον τον κόσμο έτσι και στην Ελλάδα.

Όπως είναι φυσικό λοιπόν από την αρχή της πανδημίας COVID-19 τα social media (Facebook, Instagram, blogs, Twitter, κλ) είχαν την τιμητική τους καθώς βοήθησαν να κατακλειστεί το διαδίκτυο με περιεχόμενο σχετικό με την πανδημία και τον ιό SARS-CoV-2 που την προκάλεσε. Μάλιστα η συγκεκριμένη πανδημία χαρακτηρίζεται από μερικούς και ως «first social media pandemic» καθώς προκάλεσε μια αντίστοιχου μεγέθους «διαδικτυακή» πανδημία στις πλατφόρμες των κοινωνικών δικτύων. Όλα αυτά έχει πολλαπλές αντίκτυπο στον τελικό χρήστη όπως ψυχολογικές επιπτώσεις, φόβος αντίσταση, αντίδραση κτλ.

Παρόλη την ύπαρξη επίσημων ιατρικών κοινοτήτων και οργανισμών που διεξήγαγαν έρευνες και δημοσίευαν τα αποτελέσματα τους, δεν έλειψαν προσπάθειες δημιουργίας παγκόσμιας υστερίας, παραπληροφόρησης και χειραγώγησης της κοινής γνώμης μέσω των social media. Αυτό θα μπορούσε να αποδοθεί στον πρωτοφανή βομβαρδισμό από πληροφορία που δέχεται ο μέσος χρήστης με αποτέλεσμα να δημιουργείται ένα τεράστιο χάσμα στην εξαγωγή γνώσης από αυτήν. Επίσης όπως είναι ήδη γνωστό στον κόσμο του διαδικτύου η παραπληροφόρηση και οι ψεύτικες ειδήσεις ταξιδεύουν πολύ πιο γρήγορα από τις αληθινές.

Ενώ όλες οι πλατφόρμες κοινωνικής δικτύωσης προσφέρουν πληροφορίες και ιατρικό περιεχόμενο, το twitter θα λέγαμε ότι κατέχει εξέχουσα θέση ανάμεσα τους διαδραματίζοντας ίσως το σπουδαιότερο παίκτη στο ρόλο της άμεσης διασποράς της πληροφορίας. Γι' αυτό και η υιοθέτηση του από τα μέλη των 7 παγκόσμιων ηγετών χωρών ως μέσο *άμεσης* ενημέρωσης σχετικά με την πορεία της πανδημίας και γενικά μια κρίση στον τομέα της υγείας. Μάλιστα δεν είναι ουκ ολίγα τα σχετικά tweets που έγιναν αμέσως viral με εμφανή τα σημάδια επήρειας στην κοινή γνώμη (Sohaib R Rufai, 2020). Εν μέσω πανδημίας θεωρείται υψίστης σημασίας η δυνατότητα άμεσου διαμοιρασμού πληροφορίας. Αυτό όπως θα δούμε μπορεί χαρακτηριστεί ως ένα βέλος στη φαρέτρα των δυνατοτήτων του. Βέβαια όπως ήταν φυσικό επακόλουθο δε διαθέτει μόνο προτερήματα αλλά μπορούν να εντοπιστούν και μελανές κηλίδες στην αμεσότητα αυτή που προσφέρει.

## ΜΕΘΟΔΟΙ & ΑΠΟΤΕΛΕΣΜΑΤΑ

### ΠΡΟΕΤΟΙΜΑΣΙΑ & ΑΡΧΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Για την υλοποίηση του παραπάνω θέματος έγινε χρήση έτοιμου σετ δεδομένων στην ιστοσελίδα [www.Kaggle.com](https://www.kaggle.com) με την ονομασία Greek labeled COVID-19 vaccine tweets. Το

αρχείο με τίτλο “greek\_tweets\_labeled.csv περιέχει σετ με δεδομένα από tweets στην ελληνική γλώσσα για την περίοδο από 13/7/2021 έως και 4/8/2021 από Έλληνες χρήστες με τοποθεσία την Ελλάδα (τουλάχιστον σε όσα αναφέρεται τοποθεσία). Περιέχει 25.093 examples και 11 features/attributes ένα εξ αυτών και το target value με την ονομασία label. τα attributes είναι τα ακόλουθα:

Username, Date, Followers, Source, Location, Tweet, Length, Compound, Label, J (κενή στήλη/attribute), K (κενή στήλη/attribute) και label με το target value που μας ενδιαφέρει. Αποτελεί την εντύπωση για το ύφος του κάθε tweet (sentiment) και παίρνει τρεις (3) τιμές: positive αν το ύφος του συγκεκριμένου tweet είναι θετικό, neutral αν το ύφος είναι αρνητικό και negative αν το ύφος είναι ουδέτερο δηλ. ούτε θετικό ούτε αρνητικό. Όπως φαίνεται και στις παρακάτω εικόνες.

```
1 Username,Date,Followers,Source,Location,Tweet,length,compound,label,,
2 CaeruleusCalva,4. August 2021,1088,Twitter for Android,Not So Ancient Greece,"οπας λενε και οι καθηγητες ωρλ τα ανεμβολιαστα
3 ζωα πεθαινουν απο κορωνοιο τα εμβολιασμενα ζωα δεν πεθαινουν απο κορωνοιο
4 συμπερασμα δεν υπαρχει εμβολιο για το να μην ειναι κανεις ζωο οργανισμος απο ζωικα κυτταρα τα οποια ο κορωνοιος εχει
5 στοχος",245,-0.8316,negative,,
6 kambanellas,4. August 2021,316,Twitter for Android,Cyprus," δυστυχως τα δεδομενα δεν αλλαζουν ειτε σου αρεσει ειτε οχι
7 δε θες εμβολιο προβλημα σου να μη θελεις να πιστεψεις δεδομενα επειδη εισαι σου καπνισε σε κανει προβληματικο
8 σορον αλλα δεν υπαρχει κατι αλλο",206,-0.6395,negative,,
9 em56464,4. August 2021,116,Twitter Web App,prefer not to say,χειροκροτουν τους πυροσβεστες μεχρι να σβησει η φωτια μετα
10 καταρες κ απειλες απολυσης αν δεν κανουν το εμβολιο,114,-0.34,negative,,
11 zetakosmatougk,4. August 2021,2955,Twitter for Android,greece,εχετε μισο λεπτο να σας μιλησω για το εμβολιο της μητσοτακη
12 γαμεισαι μητσοτακηπαραιτησου μητσοτακηκαθαρω φωτι...,113,-0.5849,negative,,
13 PeterParker_84,4. August 2021,149,Twitter Web App,Timbuktu Κάτω Αχαΐας,σκεψου ουτε οι ιδιοι δεν εμπιστευονται το εμβολιο
14 τους,55,-0.4023,negative,,
15 EvgenTzavaras,4. August 2021,1813,Twitter for Android,prefer not to say,"ημαρτον ξεστραβωθειτε ειστε και πιτσιρικάδες ασχημο
16 το κολλημα στην εγκεφαλικη σκατιλα απο τοσο νεαροι
17 το εμβολιο
18 μοντεψε με υπερβλακιτους πολλαπλασιους των σημερινων ποτε ηρθε ελλαδα και ποιοο θαρραλειο το κανανε ",217,-0.4804,negative,,
19 Faklana38,4. August 2021,1091,Twitter for Android,Bachelor τσαρδί,εχετε μισο λεπτο να σας μιλησω για το εμβολιο της μητσοτακη
20 γαμεισαι μητσοτακηπαραιτησου μητσοτακηκαθαρω φωτιες,113,-0.5423,negative,,
21 e_thessalia,4. August 2021,2200,E-ThessaliaAutoPosting,"Volos, Greece",το εμβολιο της αναμενεται να εξασφαλισει πληρη εγκριση
22 αρχες σεπτεμβριου,74,0.6705,positive,,
23 TheWolfpup,4. August 2021,2988,Twitter for Android,prefer not to say," α δηλαδη για το εμβολιο εχει μια σταση γιατι ειναι
γιατρος ενω για ολες τις μαρτυριες που ανεβαζει δεν ειναι
αγορι μου δεν εισαι ουτε για τα πανηγυρια",151,0.0772,positive,,
NemesisEfialtis,4. August 2021,110,Twitter Web App,Αθήνα,η πιθανοτητα να παθει κατι καποιος απο εμβολιο ειναι μικροτερη απο το
να παθει κατι απο κοβιντ ακομα και παιδια αυτα λενε τα δεδομενα ανοχη στα ψεκια τελος μορφωσου η βουλωνε δεξ και επιπλοκες
ακομα και σε νεους και θα κανεις κι αυτο το εμβολιο εργασια για το σπιτι,268,0,neutral,,
AlexGriva,4. August 2021,59,Twitter for Android,"Κέρκυρα, Ελλάς",σημερα ελεγα ποσο τυχερη νιωθω που στον κυκλο μου δεν εχω
ουτε εναν ανεμβολιαστομονο στη δουλεια με ρωτησε η αλλη αν εχω κανει το εμβολιο και της απαντησα ναι γιατι δεν ειμαι
ηλιθιαστη δεν το ειχε κανει@,205,0.4215,positive,,
Lunaticus_9,4. August 2021,41,Twitter Web App,prefer not to say,τωρα νομιζεις οτι εχεις εμβολιο γελανε και οι πετρες κολη
προστασια σου ευχομαι,83,0.8225,positive,,
malaganas,4. August 2021,17,Twitter for Android,prefer not to say,"εσεις κυριε το εχετε κανει το εμβολιο
21
22
23 λιμνη ευβοιας φωτιαβαρυνμπομη μητσοτακη γαμεισαι ",87,0,neutral,,
```

Εικόνα 3. Σετ δεδομένων (αρχείο CSV)

## ΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Θα ξεκινήσουμε από την ανάλυση και την επεξεργασία των πρωτογενών δεδομένων. Μια διαδικασία απαραίτητη, αποτελώντας, όπως θα δούμε και παρακάτω, το θεμέλιο λίθο οποιασδήποτε εφαρμογής Μηχανικής Μάθησης.

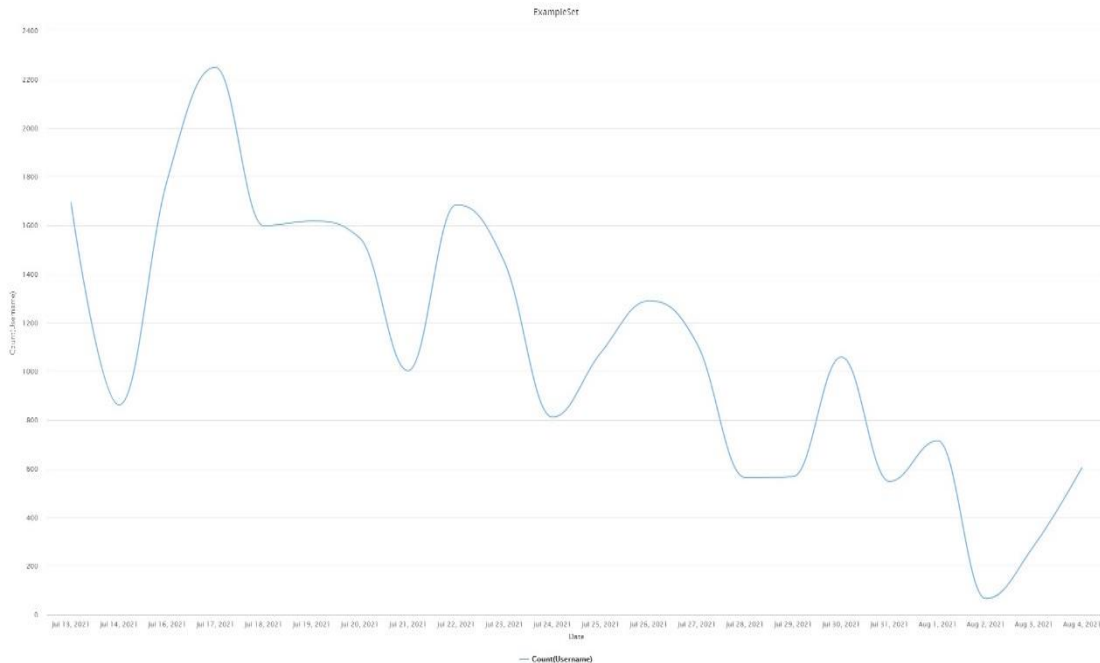
Από τα πιο κρίσιμα σημεία λοιπόν για την ανάλυση δεδομένων θεωρείται η προεπεξεργασία των δεδομένων. Ακολουθώντας λοιπόν τα βήματα της διαδικασίας εξόρυξης γνώσης σε αυτό το βήμα θα προβούμε στην προετοιμασία των δεδομένων μας. Καθαρίζουμε λοιπόν τα δεδομένα από σφάλματα, θόρυβο κλπ ώστε να είναι ποιοτικά.

Η σωστή διαμόρφωση του σετ δεδομένων που θα χρησιμοποιηθεί στη μηχανική μάθηση αποτελεί ίσως τη σπουδαιότερη διαδικασία για την εξαγωγή μη τετριμμένης, υποκρυπτόμενης, άγνωστης και εν δυνάμει χρήσιμης πληροφορίας από τα δεδομένα.

Με τη βοήθεια του RapidMiner Studio θα πραγματοποιήσουμε μια αρχική μελέτη και εν συνεχεία οπτικοποίηση των δεδομένων.

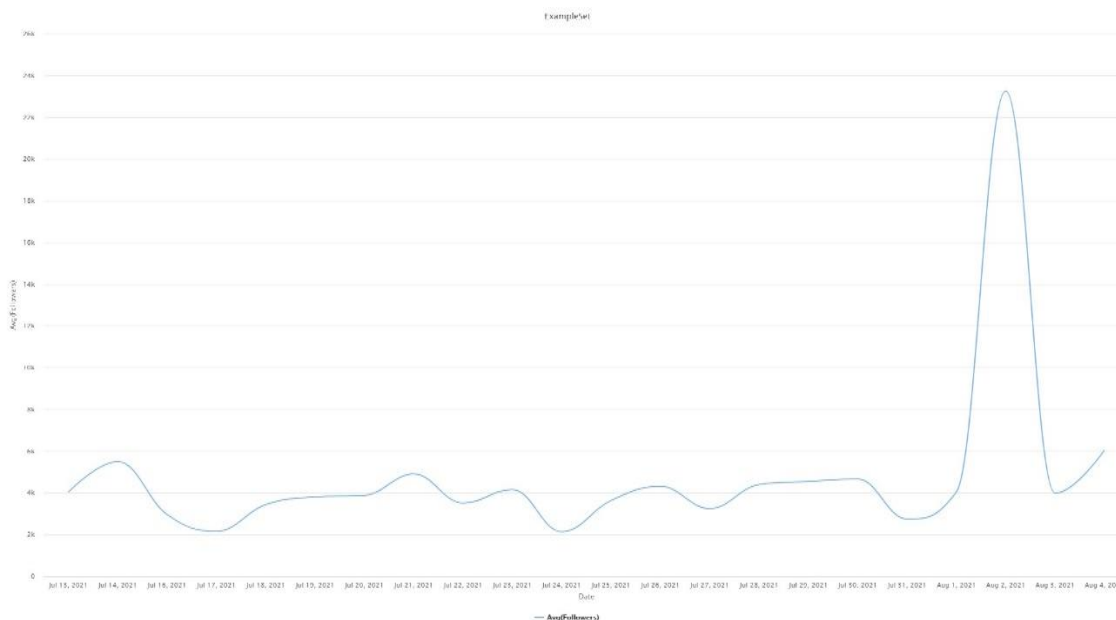
Εκτελώντας μια αρχική ανάλυση των δεδομένων που έχουμε στη διάθεσή μας παρατηρούμε τα εξής:

- ✓ Σε ορισμένους λογαριασμούς χρηστών οφείλεται η αποστολή των περισσότερων tweets με τα σχετικά tag κορονοϊό ή εμβόλιο κτλ. Ο πρώτος χρήστης έχει στείλει 119 διαφορετικά tweets ενώ ο δεύτερος 96. Ένα ενδιαφέρον διάγραμμα είναι αυτό που ακολουθεί στην Εικόνα 4. Στον οριζόντιο άξονα έχουμε την ημερομηνία και στον κάθετο τον αριθμό των χρηστών που δημοσίευσαν tweets. Βλέπουμε το μέγιστο του γραφήματος την 17 Ιουλίου 2021 με 2250 tweets χρηστών και το ελάχιστο την 2 Αυγούστου 2021 με 66 tweets χρηστών από το συγκεκριμένο example set. Ένας από τους λόγους είναι ότι την 17<sup>η</sup> Ιουλίου 2021 αποτέλεσε viral είδηση το τοπικό lockdown που εφαρμόστηκε αιφνίδια στη Μύκονο εξαιτίας της εκτόξευσης των κρουσμάτων κορονοϊού στο νησί ([iefimerida.gr](http://iefimerida.gr), 2021). Επίσης πολλά και ποικίλα σχόλια δημοσιεύτηκαν σχετικά με τον εμβολιασμό και το τι μέλλει γενέσθαι με τα επισιτιστικά επαγγέλματα καθώς και την υποχρεωτικότητα του εμβολίου και εν γένει του εμβολιασμού σε συγκεκριμένους κλάδους που ανακοινώθηκε ότι σχεδιάζει να προβεί η κυβέρνηση μετά την επιστροφή των αδειούχων από το μήνα Σεπτέμβριο και μετά. Στην αντίπερα όχθη την 2 Αυγούστου 2021 είχαμε τους ολυμπιακούς αγώνες στην επικαιρότητα καθώς και τη φυγή των αδειούχων του Αυγούστου για τις καλοκαιρινές τους διακοπές οπότε τα σχετικά tweet των χρηστών που αφορούσαν τον κορονοϊό ήταν εξαιρετικά μειωμένα.



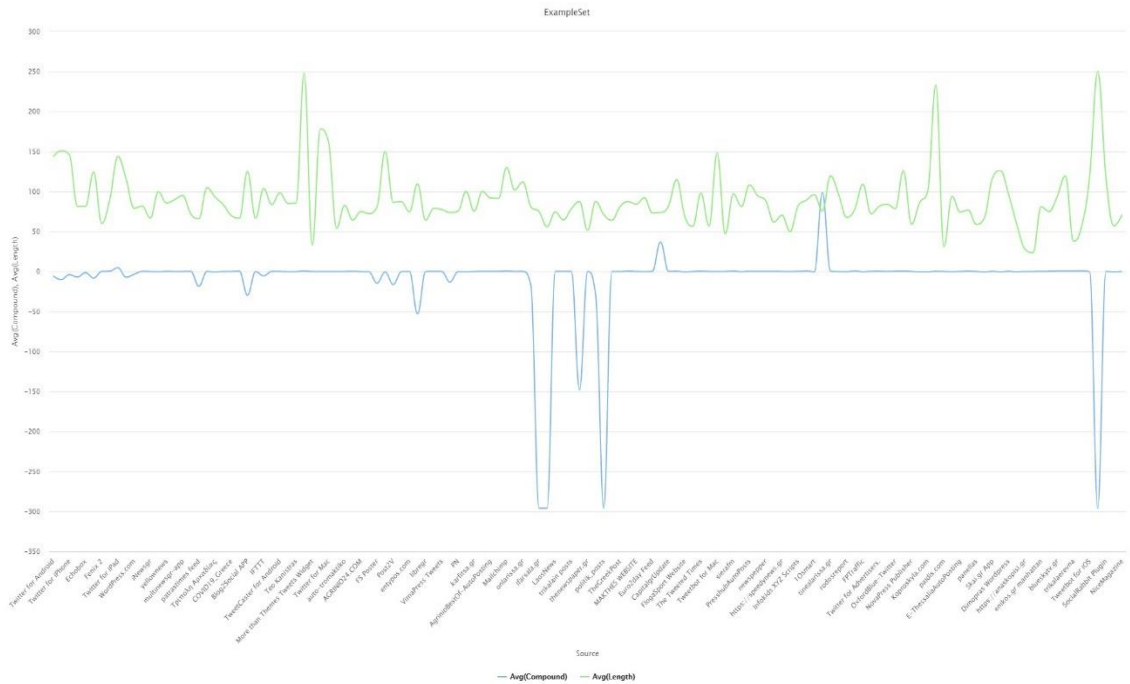
**Εικόνα 4. Tweets - αριθμός χρηστών**

- ✓ Στο attribute label βλέπουμε ότι στην πλειοψηφία (το μέγιστο) τους το ύψος σχετικά με τα εμβόλια και τον κορονοϊό είναι negative με 8.826 tweets, ακολουθεί το positive 8.050 και το neutral έρχεται τελευταίο (το ελάχιστο) με 7.325 tweets.
- ✓ Σχετικά με το attribute Date δηλαδή την ημερομηνία βλέπουμε την αρχική ημερομηνία 13/7/2021 και την τελική 4/8/2021 και υπάρχει στη διάθεσή μας και η χρονική διάρκεια που είναι 22 ημέρες.
- ✓ Επόμενο στη σειρά είναι το attribute Followers που μας δείχνει πληροφορίες για τον αριθμό των followers εκάστοτε χρήστη. Έχει στη διάθεση μας τη μέγιστη και την ελάχιστη τιμή αυτών και μας παραθέτει τον μέσο όρο αυτών. Στο ακόλουθο διάγραμμα στην Εικόνα 5 βλέπουμε στον οριζόντιο άξονα την ημερομηνία και στον κάθετο τον μέσο όρο των followers των χρηστών που στέλνουν tweets. Η πληροφορία που θα μπορούσαμε να αντλήσουμε από το συγκεκριμένο διάγραμμα είναι ότι από 13 Ιουλίου έως 1 Αυγούστου στέλνουν tweets χρήστες με μέσο από 2.000 έως 5.000 followers στους λογαριασμούς και στις 2 Αυγούστου βλέπουμε μια ξαφνική έκρηξη σε 23.000 followers. Ένα προφανές συμπέρασμα είναι ότι tweets σχετικά με τον εμβολιασμό απεστάλησαν από τους χρήστες με πολλούς ακόλουθους.

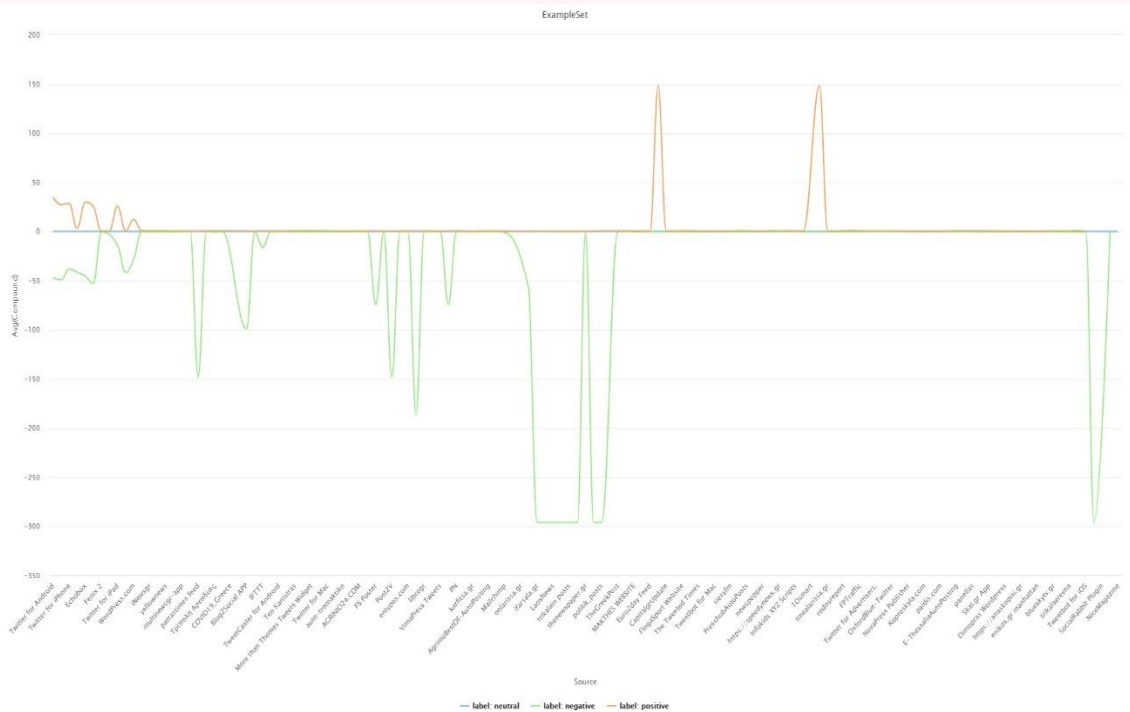


**Εικόνα 5. Followers χρηστών**

- ✓ Το attribute Source παρέχει πληροφορίες για την πηγή των tweet κάθε χρήστη. Βλέπουμε ότι τη μερίδα του λέοντος κατέχουν οι κινητές συσκευές καθώς το twitter for android έχει συνολικά χρησιμοποιηθεί από 11.734. Επίσης είναι προφανές ότι το twitter for android αφορά επί το πλείστον χρήστες - φυσικά πρόσωπα (δηλ. όχι επιχειρήσεις/νομικά πρόσωπα). Δεύτερο βλέπουμε ότι έρχεται το twitter web App με χρήση από 7907 χρήστες και ακολουθούν και άλλες τιμές από διάφορες άλλες πηγές όπως η ελάχιστη που αφορά την «Τρίπολη Αρκαδίας» με χρήση από έναν χρήστη από τοπικό ειδησεογραφικό πρακτορείο (νομικό πρόσωπο / οργανισμός) που ανέβασε το tweet με τη χρήση της συγκεκριμένης εφαρμογής. Παρακάτω ακολουθούν δυο διαγράμματα. Το πρώτο στην Εικόνα 6 έχει στον οριζόντιο άξονα τις πηγές (source) και στον κάθετο τις μέσες τιμές των compound και length. Το δεύτερο στην Εικόνα 7 έχει στον οριζόντιο άξονα τις πηγές (source) και στον κάθετο τη μέση τιμή του compound και απεικονίζει το αρνητικό, το θετικό και το ουδέτερο ύφος του κειμένου με πράσινο, πορτοκαλί και μπλε χρώμα αντίστοιχα. Ερμηνεύοντας τα διαγράμματα φαίνεται ότι οι χρήστες – φυσικά πρόσωπα που έχουν χρησιμοποιήσει για τις αναρτήσεις τους κινητές συσκευές (twitter for android) έχουν στην πλειοψηφία τους αρνητικό ύφος στις αναρτήσεις τους και τα tweets τους έχουν μέσο όρο μήκους (length) τους 150 χαρακτήρες. Ενώ tweets προερχόμενα από νομικά πρόσωπα – δημοσιογραφικούς οργανισμούς στην πλειοψηφία τους κατατάσσονται ως ουδέτερα με εξαίρεση ορισμένες ακραίες τιμές όπως πχ. Tweet από την πηγή source ifarsala.gr, politik\_posts και terranews που χαρακτηρίζονται για το αρνητικό ύφος των κειμένων τους ενώ larissanet.gr και famagustanews χαρακτηρίζονται για το θετικό ύφος τους.

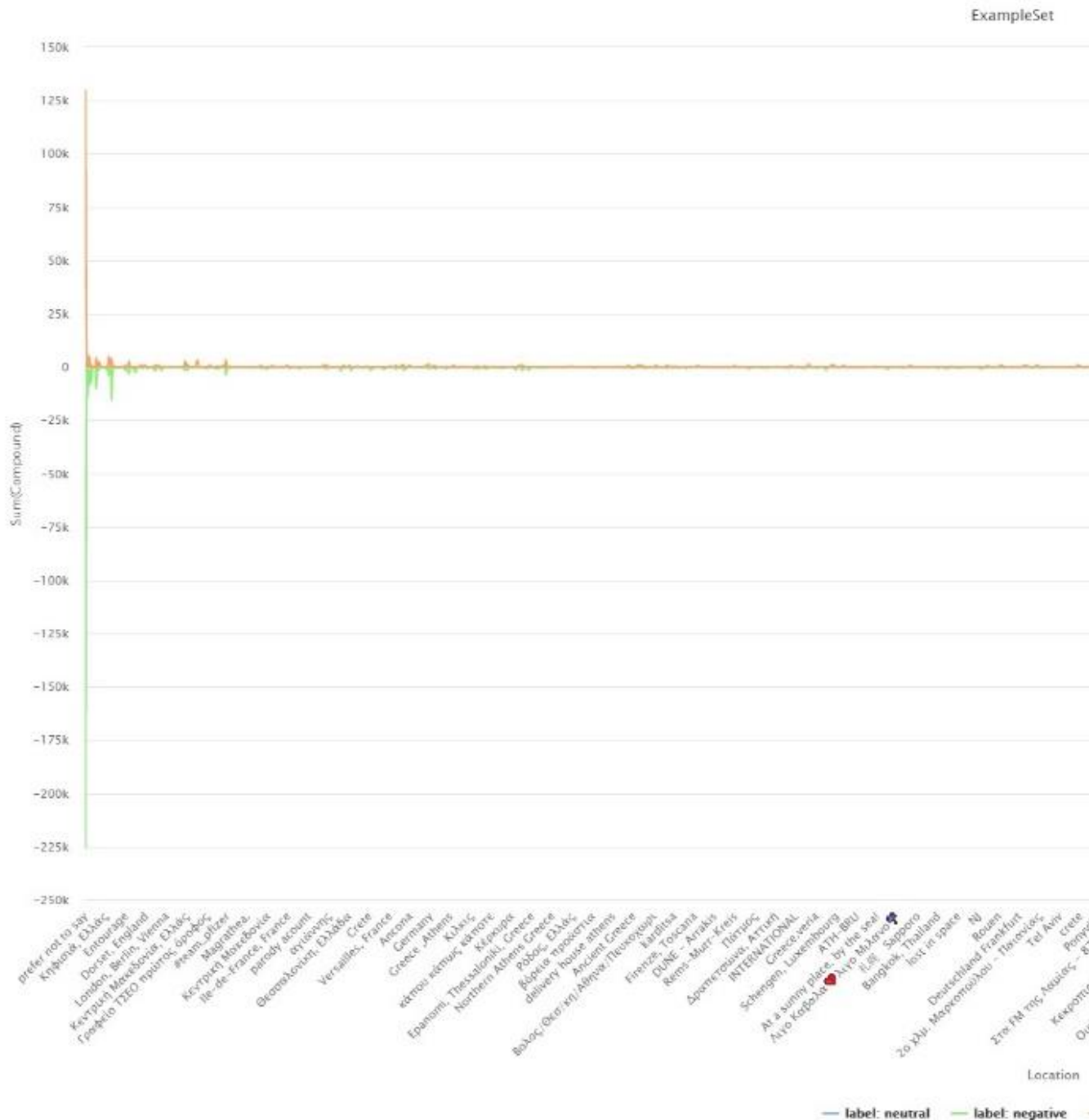


Εικόνα 6. Source - length



Εικόνα 7. Source – Compound - Sentiment

- ✓ Το attribute Location περιέχει πληροφορίες για την τοποθεσία των χρηστών που έχουν στείλει τα συγκεκριμένα tweets. Βλέπουμε πρώτη και με διαφορά την επιλογή “prefer not to say” με 12.473 χρήστες ακολουθούμενο από την τοποθεσία “Greece” με 864 τιμές και κλείνει με την τοποθεσία “MACEDONIA Greece” με τιμή ένα (1). Αυτό οφείλεται στις τελευταίες αλλαγές σχετικές με την προστασία προσωπικών δεδομένων όπου πλέον ο χρήστης διαθέτει την επιλογή να μην κοινοποιήσει την τοποθεσία του. Σε αυτήν την περίπτωση ο αλγόριθμος της πλατφόρμας του twitter να μην γνωρίζει την τοποθεσία του χρήστη αλλά δεν την κοινοποιεί αφού δεν έχει τη συγκατάθεσή αυτού. Το διάγραμμα στην Εικόνα 8 που ακολουθεί έχει στον οριζόντιο άξονα το attribute location και στον κάθετο το compound σε συνολικό άθροισμα (sum). Όπως επιβεβαιώνεται και στο διάγραμμα δεν μπορούμε να εξάγουμε χρήσιμη πληροφορία ως προς την τοποθεσία των χρηστών καθώς υπόκειται στον χρήστη αν θέλει να κοινοποιήσει την τοποθεσία του. Εν πάση περιπτώσει οι περισσότεροι έχουν ως επιλογή το prefer not to say και ελάχιστες εγγραφές υπάρχουν με πραγματικά δηλωμένες τοποθεσίες εξού και η τεράστια διασπορά στο διάγραμμα στην *Εικόνα 8*. Αυτό που φαίνεται ξεκάθαρα για άλλη μια φορά είναι το κυρίαρχο αρνητικό ύφος του συγκεκριμένου σετ δεδομένων καθώς στο κυρίαρχο παράδειγμα της εγγραφής prefer not to say έχουμε -226.061 αρνητικό άθροισμα του attribute compound ενώ το θετικό άθροισμα compound είναι +130.076 για την αντίστοιχη εγγραφή. Επίσης βλέπουμε ότι η επιλογή τοποθεσίας “prefer not to say” έχει τις ακραίες μέγιστες τιμές του διαγράμματος ενώ τα tweet που προέρχονται από άλλες τοποθεσίες τείνουν να είναι πιο ουδέτερα με ελάχιστη ίσως στα όρια του λάθους διαφοροποίηση θετικού – αρνητικού ύφους.

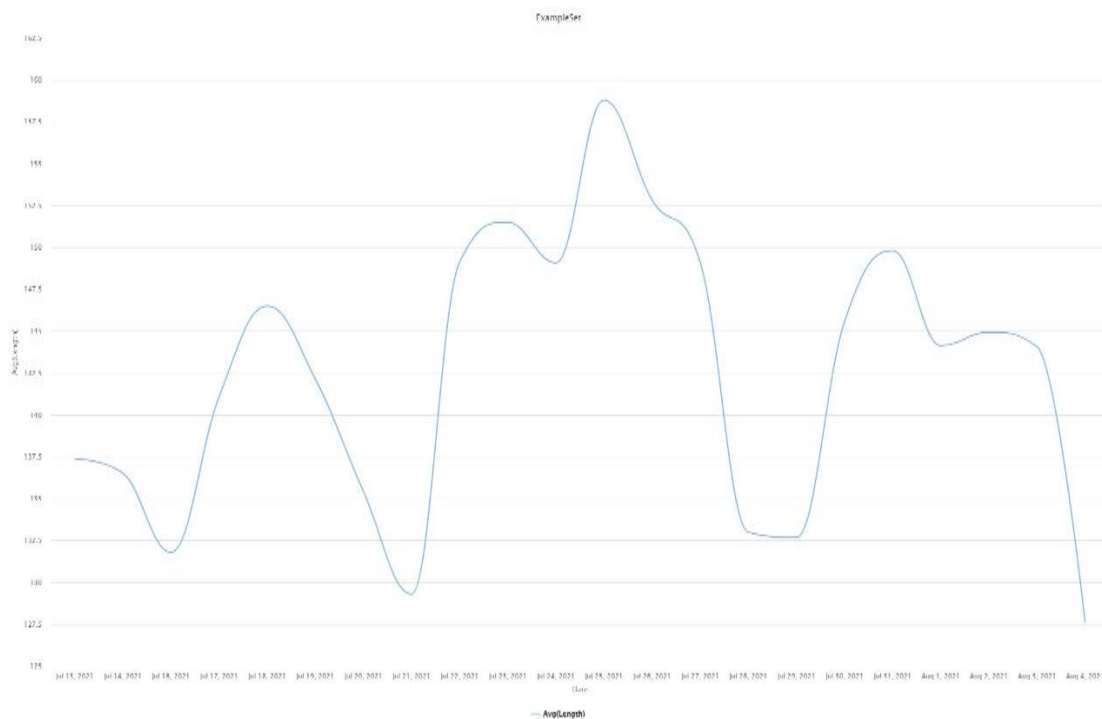


**Εικόνα 8. Location - Sum of Compound**

- ✓ Ακολουθούν τα attributes Cleanse\_tweet και lower\_gr\_final\_tweets που αποτελούνται από τα tweet κείμενα κάθε ενός από αυτά που έχουν στείλει οι χρήστες και είναι αυτά έχουμε ήδη προ-επεξεργαστεί στο excel. Ο λόγος ύπαρξης δύο ίδιων στηλών δεν

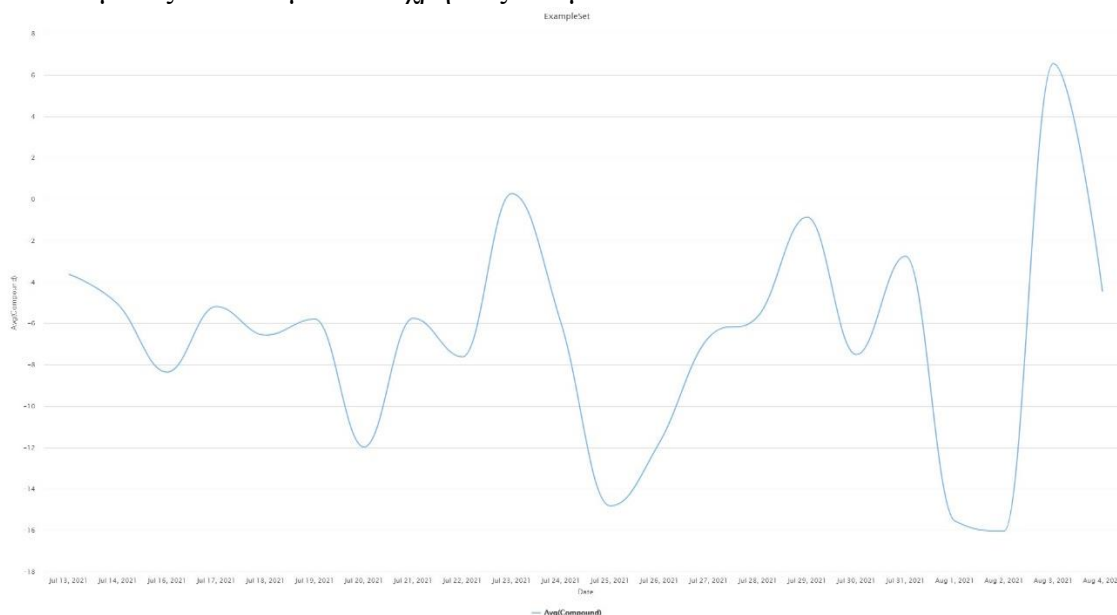
αποτελεί πρόβλημα στη μετέπειτα επεξεργασία τους με το RapidMiner Studio καθώς διαθέτει στη φαρέτρα του αρκετούς operators.

- ✓ Προ τελευταίο attribute είναι το Length και αφορά το μήκος λέξεων κάθε tweet. Μας παρέχει πληροφορία για το ελάχιστο που είναι μήκους δύο χαρακτήρων και το μέγιστο που είναι 282 χαρακτήρων με ένα μέσο όρο χαρακτήρων σε αυτό το σετ δεδομένων τους 142,59 χαρακτήρες. Στο ακόλουθο διάγραμμα στην Εικόνα 9 φαίνεται ότι το μέσο μέγεθος (σε χαρακτήρες) των tweets που αποστέλλονται από χρήστες έχει μια σημαντική αύξηση στις 26 Ιουλίου 2021 ενώ την 21<sup>η</sup> Ιουλίου, στις 27 και 28 Ιουλίου και στις 4 Αυγούστου παρατηρούνται οι μεγαλύτερες μειώσεις στο μέσο όρο των μεγεθών των tweet. Η αύξηση θα μπορούσαμε να υποθέσουμε ότι οφείλεται κυρίως στον έντονο σχολιασμό των μέτρων που ανακοινώθηκαν στα ΜΜΕ από την κυβέρνηση στις 26 Ιουλίου 2021 και αφορούσαν μέτρα «πίεσης» του πληθυσμού προς τον εμβολιασμό. Αντίθετα η μείωση του μέσου όρου μήκους των tweets στα τέλη Ιουλίου θα μπορούσε να αποδοθεί στην έναρξη των Ολυμπιακών αγώνων που έστρεψε τα βλέμματα και το ενδιαφέρον της ανθρωπότητας προς αυτήν την κατεύθυνση. Στις 4 Αυγούστου η πτώση του μέσου όρου ίσως μπορεί να αποδοθεί στην έλλειψη ενδιαφέροντος των χρηστών να σχολιάσουν κάτι σχετικό με τον εμβολιασμό επειδή στην συντριπτική πλειοψηφία τους φεύγουν για τις καλοκαιρινές διακοπές.



Εικόνα 9. Μήκος λέξεων tweets

- ✓ Κλείνουμε την πρωταρχική αυτή μελέτη στα δεδομένα μας με τελευταίο attribute το compound που αφορά συντελεστή βαρύτητας που χρησιμοποιήθηκε για το χαρακτηρισμό του ύφους των συγκεκριμένων tweets. Βλέπουμε ότι παίρνει τιμές από -959 έως 997 με μέσο όρο -6,750. Ο λόγος προφανώς είναι διότι στο συγκεκριμένο σετ δεδομένων κυριαρχεί το αρνητικό ύφος. Αυτό μας το επιβεβαιώνει και το παρακάτω γράφημα στην Εικόνα 10. Στον οριζόντιο άξονα είναι η ημερομηνία και στον κάθετο το Compound (μέση τιμή/ημερομηνία). Φαίνεται ξεκάθαρα και στο διάγραμμα το αρνητικό ύφος των tweet των χρηστών σχετικά με το εμβόλιο και τον εμβολιασμό. Τα δύο σημαντικότερα ελάχιστα τοποθετούνται χρονικά στις 25 Ιουλίου με -14,82 στις 1 Αυγούστου 2021 με -15,54 και στις 2 Αυγούστου 2021 με -16,05 ενώ η μέγιστη θετική μέση τιμή παρατηρείται στις 3 Αυγούστου 2021. Ερμηνεύοντας τις ακραίες παρατηρήσεις στις 25 Ιουλίου βλέπουμε ότι η κυβέρνηση στην προσπάθεια της να πιέσει για να εμβολιαστούν οι εργαζόμενοι λαμβάνει νέα μέτρα με μοχλό πίεσης την κατάργηση των μέχρι πρότινος δωρεάν τεστ κορονοϊού. Συγκεκριμένα ανακοινώνει ότι πλέον στην εστίαση και στον τουρισμό σταματούν τα δωρεάν τεστ κορονοϊού και πλέον θα τα επιβαρύνονται οι ίδιοι οι ενδιαφερόμενοι με αποτέλεσμα να «ξεσπάσει» θύελλα αρνητικών σχολίων εκείνη την ημέρα στο twitter. Στις 1 και 2 Αυγούστου 2021 επικρατεί ακραίος καύσωνας και η τάση φυγής των εκδρομέων του Αυγούστου μαζί με τη γενική ανησυχία για τη διασπορά του κορονοϊού στην ελληνική περιφέρεια. Αντιθέτως στις 3 Αυγούστου 2021 γίνονται σχετικές ανακοινώσεις σε ΜΜΕ που αφορούν την τρίτη δόση του εμβολίου έτσι ώστε να χτιστεί ανοσία και να περιοριστεί η υπερμεταδοτικότητα της μετάλλαξης Δέλτα στον πληθυσμό και δει στις ευπαθείς ομάδες. Βλέπουμε ότι οι χρήστες αντιμετώπισαν επί το πλείστον θετικά αυτά τα νέα.

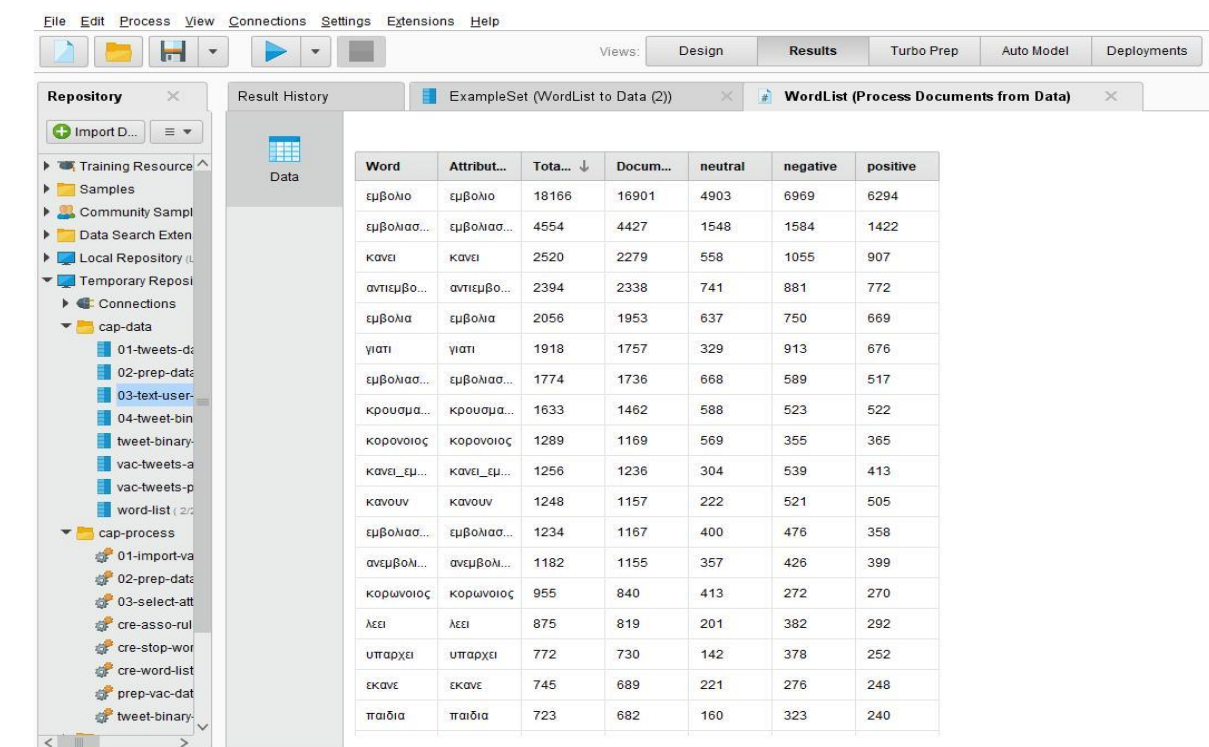


**Εικόνα 10. Ημερομηνία - Compound**

## ΟΠΤΙΚΟΠΟΙΗΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Στη συνέχεια ακολουθεί η οπτικοποίηση των δεδομένων με τη βοήθεια του RapidMiner Studio. Για την οπτικοποίηση επιλέξαμε να δημιουργήσουμε λίστα με λέξεις που επαναλαμβάνονται συχνά στο σετ δεδομένων μας (example set) και εν συνεχεία με τη βοήθεια μηχανικής μάθησης να εξάγουμε συμπεράσματα χρησιμοποιώντας κανόνες συσχέτισης. Το μεν πρώτο, δεν είναι κάτι άλλο από το γνωστό σε όλους μας word cloud ή ετικετοσύννεφο όπως μεταφράζεται στην ελληνική γλώσσα. Το δε δεύτερο, αφορά τη δημιουργία κανόνων συσχέτισης των λέξεων που εμφανίζονται σε μεγάλη συχνότητα στο σετ δεδομένων μας όπου με τη χρήση αλγορίθμων μηχανικής μάθησης θα μπορέσουμε να το εξάγουμε χρήσιμη για εμάς πληροφορία.

### ΕΤΙΚΕΤΟΣΥΝΝΕΦΟ – WORD CLOUD



| Word        | Attribut... | Tota... ↓ | Docum... | neutral | negative | positive |
|-------------|-------------|-----------|----------|---------|----------|----------|
| εμβολιο     | εμβολιο     | 18166     | 16901    | 4903    | 6969     | 6294     |
| εμβολιασ... | εμβολιασ... | 4554      | 4427     | 1548    | 1584     | 1422     |
| κανει       | κανει       | 2520      | 2279     | 558     | 1055     | 907      |
| αντιεμβο... | αντιεμβο... | 2394      | 2338     | 741     | 881      | 772      |
| εμβολια     | εμβολια     | 2056      | 1953     | 637     | 750      | 669      |
| γιατι       | γιατι       | 1918      | 1757     | 329     | 913      | 676      |
| εμβολιασ... | εμβολιασ... | 1774      | 1736     | 668     | 589      | 517      |
| κρουσμα...  | κρουσμα...  | 1633      | 1462     | 588     | 523      | 522      |
| κορωνοιος   | κορωνοιος   | 1289      | 1169     | 569     | 355      | 365      |
| κανει_εμ... | κανει_εμ... | 1256      | 1236     | 304     | 539      | 413      |
| κανουν      | κανουν      | 1248      | 1157     | 222     | 521      | 505      |
| εμβολιασ... | εμβολιασ... | 1234      | 1167     | 400     | 476      | 358      |
| ανεμβολι... | ανεμβολι... | 1182      | 1155     | 357     | 426      | 399      |
| κορωνοιος   | κορωνοιος   | 955       | 840      | 413     | 272      | 270      |
| λεει        | λεει        | 875       | 819      | 201     | 382      | 292      |
| υπαρχει     | υπαρχει     | 772       | 730      | 142     | 378      | 252      |
| εκανε       | εκανε       | 745       | 689      | 221     | 276      | 248      |
| παιδια      | παιδια      | 723       | 682      | 160     | 323      | 240      |

Εικόνα 11. Συχνότερες λέξεις για τη δημιουργία ετικετοσύννεφου



Μια σημαντική επισήμανση αφορά τον περιορισμό των λέξεων στο RapidMiner Studio για τη δημιουργία word cloud. Συγκεκριμένα υπάρχει περιορισμός έως 500 λέξεις για τη δημιουργία word cloud. Στην περίπτωση μας, επειδή ο operator Stem δεν υποστηρίζει την ελληνική γλώσσα δεν καταφέραμε να περιορίσουμε τα αποτελέσματά μας δηλ. τις παρόμοιες λέξεις που έχουν διαφορετικές καταλήξεις. Άμεση συνέπεια αυτού είναι ότι το διανυσματικό αρχείο κειμένου που δημιουργείται περιέχει χιλιάδες attributes που τα περισσότερα προέρχονται από ομοειδής λέξεις που προφανώς έχουν και την ίδια σημασία. Για να γίνει αντιληπτό το παραπάνω θα παραθέσουμε ένα παράδειγμα. Η λέξη «εμβόλιο», η λέξη «εμβολίων» και η λέξη «εμβόλια» έχουν το ίδιο νόημα -για τις ανάγκες της εργασίας μας- αλλά διαφορετικές καταλήξεις, με αποτέλεσμα ο αλγόριθμος να τις θεωρεί ξεχωριστές λέξεις και να δημιουργεί ξεχωριστά attributes για την κάθε μια από αυτές αδυνατώντας να τις κατατάξει κάτω από την ομπρέλα ενός attribute. Αυτό οδηγεί σε σημαντικές «εκπτώσεις» στην ποιότητα της εξαγόμενης πληροφορίας μιας και δημιουργεί ένα αχανές σετ δεδομένων με χιλιάδες πλεονάζοντα attributes που δεν προσδίδουν καμία χρήσιμη πληροφορία. Όπως είναι αναμενόμενο κάτι τέτοιο έχει άμεσο αντίκτυπο σε κάθε διαδικασία μηχανικής μάθησης με τη χρήση του εργαλείου RapidMiner που αφορά κείμενα στην ελληνική γλώσσα όπως και στην περίπτωση μας φυσικά.

Από το παραπάνω word cloud (ετικετοσύννεφο) είναι προφανές ότι οι λέξεις που «κυριαρχούν» στο σετ δεδομένων είναι οι ακόλουθες:

- εμβόλιο
- εμβολιασμένοι
- ανεμβολίαστοι
- κρούσματα,
- κάνει\_εμβόλιο,
- κορονοϊός.
- κάνουν,
- και αρκετές άλλες με μικρότερη συχνότητα εμφάνισης

Το ετικετοσύννεφο που δημιουργήθηκε με τη βοήθεια του RapidMiner Studio όπως φαίνεται στην *Εικόνα 12. Word cloud (ετικετοσύννεφο)* δεν απεικονίζει με ξεκάθαρο τρόπο, ή όπως τουλάχιστον θα περιμέναμε από ένα ετικετοσύννεφο να απεικονίζει, τις συχνότητες των λέξεων που φαίνονται στην *Εικόνα 11* γι' αυτό και παρατίθενται. Επίσης το RapidMiner Studio έχει περιορισμό 500 λέξεων για τη δημιουργία ενός ετικετοσύννεφου με αποτέλεσμα να εμφανίζει τις 500 λέξεις με τη μεγαλύτερη συχνότητα στην «ιδιόμορφη» αυτή απεικόνιση που διαθέτει.

## ΔΗΜΙΟΥΡΓΙΑ ΚΑΝΟΝΩΝ ΣΥΣΧΕΤΙΣΗΣ

### ΣΤΟ ΣΥΝΟΛΟ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Συνεχίζοντας θα περάσουμε στη δημιουργία κανόνων συσχέτισης για το συγκεκριμένο σετ δεδομένων. Στο σετ δεδομένων εκτελέσαμε επεξεργασία κειμένου και διανυσματοποίηση αυτού. Με τη συγκεκριμένη διαδικασία δημιουργήθηκαν χαρακτηριστικά attributes που είναι πλέον των 45.000 σε αριθμό. Το σκεπτικό είναι να μας οδηγήσει, αν γίνεται, σε μια αρχική εκτίμηση της πληροφορίας που «κρύβεται» και μπορούμε να εξάγουμε από τα δεδομένα έτσι ώστε να μας βοηθήσει περαιτέρω στη χάραξη των επόμενων βημάτων.

Η διαδικασία που ακολουθήθηκε ήταν, η δημιουργία ενός διανυσματοποιημένου αρχείου κειμένου με δυαδικές τιμές στα παραδείγματα, η μετατροπή τους σε πεζά, φιλτράρισμα λέξεων stopwords και δημιουργία 2-gram λέξεων<sup>1</sup>. Όπως αναφέρεται και προηγουμένως, στο ετικετοσύννεφο, το RapidMiner Studio δεν διαθέτει βιβλιοθήκη με ελληνικό λεξικό. Αυτό συνεπάγεται τα εξής δυο τινά. Πρώτον δεν είχαμε πρόσβαση στη διαδικασία stem με αποτέλεσμα να δημιουργείται ένα αχανές σετ χαρακτηριστικών με πάνω από 45.000 διανυσματοποιημένες λέξεις που οι περισσότερες έχουν την ίδια σημασία και δεύτερον χρειάστηκε να κατασκευάσουμε εμείς ένα αρχείο με stopwords το οποίο και χρησιμοποιήσαμε στη συνέχεια για τη δημιουργία κανόνων συσχέτισης στα χαρακτηριστικά (attributes). Όπως ήταν αναμενόμενο δεν καταφέραμε να φιλτράρουμε με υψηλά ποσοστά επιτυχίας όλες τις άχρηστες και χωρίς νόημα λέξεις με το δικό μας λεξικό – αρχείο με stopwords. Αυτό είναι εμφανές στη συνέχεια όπου μέσω της χρήσης αλγορίθμων όπως πχ Random Forest κάνουν την εμφάνισή τους λέξεις όπως «μέσω» κοκ χωρίς δηλαδή να έχουν καμία αξία για εμάς.

Στο διανυσματοποιημένο σετ δεδομένων με τη διαδικασία που περιεγράφηκε προηγουμένως και με τη βοήθεια των operators FP-Growth και Create Association Rules του RapidMiner Studio δημιουργούμε τους κανόνες συσχέτισης. Οι παράμετροι μου χρησιμοποιήθηκαν για τον operator “FP-Growth” είναι “input format → items in dummy coded columns” και “min support→0.02” ενώ για τον operator “Create Association Rules” είναι “criterion→confidence” και “min confidence→0.1”. Ο λόγος της χρήσης χαμηλών ποσοστών support και confidence οφείλεται κυρίως στην τεράστια διασπορά των attributes στο σετ δεδομένων μας. Στην περίπτωση επιλογής μεγαλύτερων ποσοστών, πέρα από το attribute «εμβόλιο» που «κατακλύζει» τα δεδομένα, δεν θα έβγαζε αποτέλεσμα ο αλγόριθμος συνεπώς και εμείς με τη σειρά μας δε θα μπορούσαμε να βρούμε κανόνες συσχέτισης στα δεδομένα αυτά επομένως ούτε να εξάγουμε πληροφορία.

---

<sup>1</sup> Επιλέχθηκε 2-gram με σκοπό να αποτυπωθούν εκφράσεις όπως πχ «παρενέργειες εμβολίων» κοκ

|                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
|-----------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | <b>AssociationRules</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|  | Association Rules                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|  | <pre> [εμβολιασμος] --&gt; [εμβολιασμος_εμβολιο] (confidence: 0.105) [εμβολιασμος] --&gt; [εμβολιασμος_εμβολιασμοι] (confidence: 0.106) [εμβολιασμος] --&gt; [εμβολιασμοι, εμβολιασμος_εμβολιασμοι] (confidence: 0.106) [εμβολιασμος] --&gt; [κορωνοιος] (confidence: 0.107) [εμβολιασμος] --&gt; [εμβολιασμενοι] (confidence: 0.107) [εμβολιασμος] --&gt; [ανεμβολιαστοι] (confidence: 0.114) [αντιεμβολιαστες] --&gt; [ανεμβολιαστοι] (confidence: 0.140) [αντιεμβολιαστες] --&gt; [συνταγμα] (confidence: 0.143) [εμβολιασμος] --&gt; [αντιεμβολιαστες] (confidence: 0.144) [αντιεμβολιαστες] --&gt; [εμβολιασμοι] (confidence: 0.156) [αντιεμβολιαστες] --&gt; [εμβολια] (confidence: 0.158) [εμβολια] --&gt; [κρουσματα] (confidence: 0.159) [εμβολιασμος] --&gt; [κρουσματα] (confidence: 0.164) [εμβολια] --&gt; [εμβολιασμενοι] (confidence: 0.166) [εμβολια] --&gt; [εμβολια_εμβολιασμος] (confidence: 0.173) [εμβολια] --&gt; [εμβολιασμος, εμβολια_εμβολιασμος] (confidence: 0.173) [εμβολιασμοι] --&gt; [εμβολιο_εμβολιασμοι] (confidence: 0.177) [εμβολια] --&gt; [ανεμβολιαστοι] (confidence: 0.187) [εμβολια] --&gt; [αντιεμβολιαστες] (confidence: 0.189) [εμβολια] --&gt; [εμβολιασμος, εμβολιασμοι] (confidence: 0.194) [κρουσματα] --&gt; [κορωνοιος] (confidence: 0.198) [εμβολιασμοι] --&gt; [αντιεμβολιαστες] (confidence: 0.210) [κρουσματα] --&gt; [εμβολια] (confidence: 0.213) [εμβολιασμος] --&gt; [εμβολιασμοι] (confidence: 0.215) [εμβολιασμοι] --&gt; [εμβολιασμος, εμβολια] (confidence: 0.218) [εμβολιασμος] --&gt; [εμβολια] (confidence: 0.224) [εμβολιασμοι] --&gt; [κρουσματα] (confidence: 0.233) [εμβολιασμοι] --&gt; [εμβολιασμοι_εμβολιο] (confidence: 0.242) [κορωνοιος] --&gt; [εμβολιασμος] (confidence: 0.257) [κρουσματα] --&gt; [κορωνοιος] (confidence: 0.267) [εμβολιασμοι] --&gt; [εμβολιασμος_εμβολιασμοι] (confidence: 0.270) [εμβολιασμοι] --&gt; [εμβολιασμος, εμβολιασμος_εμβολιασμοι] (confidence: 0.270) [αντιεμβολιαστες] --&gt; [εμβολιασμος] (confidence: 0.273) [εμβολια] --&gt; [εμβολιασμοι] (confidence: 0.274) </pre> |
|  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |

**Εικόνα 13. Κανόνες συσχέτισης**

Βασιζόμενοι στα αποτελέσματα των κανόνων συσχέτισης που έχουν τις υψηλότερες τιμές confidence και συγκεκριμένα ξεκινώντας από την τιμή 1 παρατηρούμε ότι δεν μπορούμε να εξάγουμε χρήσιμη για εμάς πληροφορία. Συγκεκριμένα στο συγκεκριμένο σετ δεδομένων τα tweets που περιέχουν τις λέξεις εμβόλιο, εμβολιασμός, εμβολιασμοί, κάνουν εμβόλιο, εμβόλια, κλ. τότε θα περιλαμβάνονται σε αυτά και οι λέξεις εμβολιασμός, έκανε, κάνω, εμβόλια εμβολιασμοί κ.α. Δηλαδή όλα τα item sets που έχουν confidence 1 (ένα) σημαίνει ότι ισχύει η υπόθεση (if) σε όλες τις τιμές της απόδοσης (then). Από τους συγκεκριμένους κανόνες με confidence 1 (ένα) δεν μπορούμε να εξάγουμε χρήσιμη πληροφορία πέρα από την επιβεβαίωση ότι το συγκεκριμένο example set περιέχει αναμφίβολα tweets σχετικά με τα εμβόλια και τον εμβολιασμό και τα κείμενα αυτών επί τω πλείστον συντάσσονται γύρω από τις συγκεκριμένες λέξεις κλειδιά. Με φθίνουσα ταξινόμηση του confidence συνεχίζουμε την αναζήτηση για ανακάλυψη κανόνων συσχέτισης όπου θα έχουν νόημα για εμάς. Μερικοί κανόνες που έχουν νόημα στην έρευνα μας είναι οι ακόλουθοι:

| Promises        | conclusion      | Confidence |
|-----------------|-----------------|------------|
| σύνταγμα        | αντιεμβολιαστές | 0.665      |
| εστίαση         | εμβολιασμός     | 0.620      |
| κορονοϊός       | εμβολιασμός     | 0.562      |
| κρούσματα       | εμβολιασμός     | 0.497      |
| ανεμβολίαστοι   | εμβολιασμός     | 0.436      |
| ανεμβολίαστοι   | εμβόλια         | 0.316      |
| ανεμβολίαστοι   | αντιεμβολιαστές | 0.284      |
| αντιεμβολιαστές | εμβολιασμός     | 0.273      |

*Πίνακας 1. Association Rules – Confidence descending*

Στους ανωτέρω κανόνες μπορεί η υπόθεση να καλύπτεται σε μικρότερο ποσοστό από την απόδοση και γι' αυτό να έχουν μικρότερες τιμές confidence αλλά είναι πιο λογικοί καθώς μπορεί να μας βοηθήσουν να εξάγουμε χρήσιμη πληροφορία.

Ερμηνεύοντας τους κανόνες που ξεχωρίσαμε παρατηρούμε ότι στο 66,5% των tweets όπου γίνεται αναφορά στη λέξη «σύνταγμα» τότε γίνεται αναφορά και στη λέξη «αντιεμβολιαστές». Αυτό συνεπάγεται ότι πολλοί χρήστες εκφράζουν τις ανησυχίες τους σχετικά με τα συνταγματικά τους δικαιώματα και τους αντιεμβολιαστές. Ίσως μια εξήγηση μπορεί να αποδοθεί στην υποχρεωτικότητα του εμβολισμού και κατά ποσό αυτό καταπατά συνταγματικά δικαιώματα. Ακόμη μια υπόθεση είναι ότι μια μερίδα κόσμου, αναφέρεται στις συνταγματικές ελευθερίες και αυτή η μερίδα είναι οι λεγόμενοι αντιεμβολιαστές. Επόμενο στη σειρά έχει να κάνει με την εστίαση και την υποχρεωτικότητα του εμβολιασμού με 62%. Όπως αναφέραμε και παραπάνω η ελληνική κυβέρνηση χρησιμοποιώντας ανακοίνωσε πως στην εστίαση είναι υποχρεωτικός ο εμβολιασμός και μόνο οι εμβολιασμένοι επιτρέπεται να εισέρχονται σε εσωτερικούς κλειστούς χώρους. Ο κορονοϊός και τα κρούσματα ακολουθούν με απόδοση τον εμβολιασμό με περίπου 50%. Αυτό αποδίδεται στην αναφορά που γίνεται από τους χρήστες σχετικά με κορονοϊό και κρούσματα και την άμεση συσχέτιση τους με τον εμβολιασμό. Λόγου χάρι πως μειώνονται ή αυξάνονται τα κρούσματα κορονοϊού και το ρόλο που παίζει ο εμβολιασμός σε όλο αυτό. Οι ανεμβολίαστοι ακολουθούν την υπόθεση με απόδοση εμβολιασμός, εμβόλια και αντιεμβολιαστές. Οι συζητήσεις που λαμβάνουν χώρα στο twitter και είναι σχετικές με ανεμβολίαστους τότε συνεπάγεται πως ακολουθεί η λέξη κλειδί εμβολιασμός με 43,6% ή εμβόλια με 31,6%, ή αντιεμβολιαστές με 28,4%. Αυτό σημαίνει ότι όπου γίνεται αναφορά σε ανεμβολίαστο ακολουθεί προτροπή για εμβολιασμό ή κάποια αναφορά σχετική με τα εμβόλια που κατά πάσα πιθανότητα είναι αρνητική όπως πχ παρενέργειες του εμβολίου (η αρνητικότητα τεκμαίρεται από την κυριαρχία του αρνητικού ύφους στο συγκεκριμένο σεντ δεδομένων). Επίσης σε ένα ποσοστό 27,3% ακολουθεί η λέξη αντιεμβολιαστές που ίσως συνεπάγεται την ταμπέλα που συνοδεύει τους συγκεκριμένους χρήστες που είναι ανεμβολίαστοι.

Παρατηρούμε ότι οι κανόνες με το υψηλότερο support  $> 0.03$  έχουν να κάνουν με τις λέξεις «εμβόλιο και εμβολιασμός» και δε μας βοηθούν στη εξαγωγή χρήσιμης για εμάς πληροφορίας. Οι κανόνες με τα item sets που ξεχωρίζουμε είναι:

| Promises        | conclusion      | Support |
|-----------------|-----------------|---------|
| Αντιεμβολιαστές | εμβολιασμός     | 0.026   |
| Ανεμβολίαστοι   | εμβολιασμός     | 0.021   |
| Εμβολιασμοί     | αντιεμβολιαστές | 0.015   |
| Σύνταγμα        | αντιεμβολιαστές | 0.014   |
| Ανεμβολίαστοι   | αντιεμβολιαστές | 0.014   |
| Εμβολιασμένοι   | ανεμβολίαστοι   | 0.013   |
| ανεμβολίαστοι   | εμβολιασμένοι   | 0.013   |

Πίνακας 2. Association Rules - Support descending

Επομένως με τη βοήθεια της μετρικής support παρατηρούμε ότι 2,6% των tweets περιέχουν ταυτόχρονα τις λέξεις «αντιεμβολιαστές, εμβολιασμός» στο σετ δεδομένων μας. 2,1% τις λέξεις «ανεμβολίαστοι, εμβολιασμός» κοκ. Συνολικά με την άθροιση των ανωτέρω παρατηρούμε ότι σε ένα συνολικό ποσοστό 11,6% του σετ δεδομένων γίνεται αναφορά στους ανεμβολίαστους, αντιεμβολιαστές, εμβολιασμό, σύνταγμα και στους εμβολιασμένους. Αυτές οι αναφορές στην πλειοψηφία τους τεκμαίρεται ότι διαθέτουν αρνητικό ύφος. Αυτό διότι σε όποιο tweet υπάρχουν τα συγκεκριμένα item set συνήθως αφορούν κάποιο «καυστικό» σχόλιο είτε έχει να κάνει με αντιεμβολιαστές είτε με αναφορά στο σύνταγμα είτε με ανεμβολίαστους. Αυτό το ποσοστό θα μπορούσαμε να υποθέσουμε ότι αφορά το «κίνημα» αυτών που έχουν αρνητική στάση σχετικά με το εμβόλιο και εκφράζουν την ανησυχία τους μέσα από τους λογαριασμούς τους στο twitter.

Με τη βοήθεια της μετρικής lift θα δούμε τη σημαντικότητα της συσχέτισης ανεξάρτητα από το support δηλ. τη σχέση ανάμεσα στην πιθανότητα ενός συγκεκριμένου κανόνα να συμβεί (απόδοση) σε σχέση με την αναμενόμενη πιθανότητα (αναμενόμενη απόδοση). Όσο μεγαλύτερη είναι η τιμή του lift τόσο μεγαλύτερη είναι η πιθανότητα η υπόθεση να οδηγήσει στην απόδοση. Άρα οι κανόνες με τις μεγαλύτερες τιμές lift έχουν μεγαλύτερες πιθανότητες να είναι αληθείς ενώ για τιμές lift=1 η υπόθεση και η απόδοση είναι ξένες μεταξύ τους. Επομένως όπως φαίνεται και στον Πίνακα 3 που ακολουθεί οι υψηλότερες τιμές της μετρικής lift ανήκουν σε item sets που δεν οδηγούν σε χρήσιμη για εμάς πληροφορία. Ξεχωρίζουμε τους παρακάτω κανόνες με item sets και τις αντίστοιχες τιμές lift:

| Promises | conclusion      | Lift  |
|----------|-----------------|-------|
| σύνταγμα | αντιεμβολιαστές | 6.878 |

|                 |                 |       |
|-----------------|-----------------|-------|
| εμβολιασμένοι   | ανεμβολίαστοι   | 5.851 |
| ανεμβολίαστοι   | εμβολιασμένοι   | 5.851 |
| ανεμβολίαστοι   | εμβόλια         | 3.915 |
| ανεμβολίαστοι   | αντιεμβολιαστές | 2.773 |
| ανεμβολίαστοι   | εμβολιασμός     | 2.385 |
| εμβολιασμοί     | αντιεμβολιαστές | 2.170 |
| αντιεμβολιαστές | εμβολιασμός     | 1.494 |

Πίνακας 3.Association Rules - Lift descending

Με την τελευταία μετρική του lift βλέπουμε πόσες φορές αναμένεται να ισχύει ταυτόχρονα η υπόθεση και η απόδοση στο συγκεκριμένο σετ δεδομένων με tweets συγκριτικά με την αναμενόμενη απόδοση. Ο πρώτος δηλ. «σύνταγμα→αντιεμβολιαστές» όπου μας επιβεβαιώνει ότι όταν γίνεται μνεία μέσα στα tweets σχετικά με το σύνταγμα τότε ενδέχεται να ακολουθήσει και αναφορά στους αντιεμβολιαστές. Ομοίως και για τους υπόλοιπους με μικρότερη πιθανότητα πηγαίνοντας προς τον τελευταίο καθώς είναι ταξινομημένοι με φθίνουσα σειρά.

Όπως ήταν αναμενόμενο ο τεράστιος όγκος των χαρακτηριστικών σε συνδυασμό με τους περιορισμούς της ελληνικής γλώσσας κυρίως στη χρήση του εργαλείου RapidMiner δε βοήθησε τα μέγιστα στην προσπάθεια μας. Παρόλα αυτά η προσέγγιση αυτή θεωρείται χρήσιμη καθώς πλέον έχουμε μια γενική εικόνα από πληροφορίες που χαρακτηρίζουν το τεράστιο αυτό σετ δεδομένων που έχουμε στη διάθεση μας, που σε διαφορετική περίπτωση, χωρίς δηλαδή τη βοήθεια της μηχανικής μάθησης, θα ήταν αδύνατο να εντοπίσουμε.

## ΣΤΑ ΠΑΡΑΔΕΙΓΜΑΤΑ ΜΕ ΑΡΝΗΤΙΚΟ ΥΦΟΣ

Διερευνώντας περαιτέρω και πιο στοχευμένα το δείγμα μας με σκοπό την εκμαίευση πληροφορίας από τα tweet με αρνητικό ύφος εκτελούμε σε αυτά μόνο διαδικασίες δημιουργίας κανόνων συσχέτισης.

Παρατηρούμε επίσης στις πρώτες θέσεις με confidence 1 τους συνδυασμούς {κάνει\_εμβόλιο → κάνει}, {εμβόλια → εμβολιασμοί}, κοκ που ομοίως με προηγουμένως δε μας βοηθούν στην εξαγωγή χρήσιμης πληροφορίας. Την προσοχή μας έχει κεντρίσει και πάλι ο κανόνας {σύνταγμα → αντιεμβολιαστές} με confidence 0,619 και {αντιεμβολιαστές → σύνταγμα} με confidence 0,137. Όπως αναφέρθηκε και εκτενέστερα στην προηγούμενη ενότητα αυτά τα αρνητικά tweet κάνουν αναφορές στις συνταγματικές ελευθερίες του ατόμου κάτι που απασχόλησε το δείγμα μας. Παρακάτω παραθέτω και τα αποτελέσματα από τη δημιουργία κανόνων συσχέτισης στα tweet με αρνητικό ύφος.

AssociationRules (Create Association Rules)

Show rules matching

all of these conclusions:

κανει

αντιεμβολιαστες

γιατι

εμβολια

εμβολιασμοι

κανει\_εμβολιο

κανουν

κρουσματα

εμβολιασμενοι

ανεμβολιαστοι

κανω

εκανε

κανουν\_εμβολιο

συνταγμα

αντιεμβολιαστες\_negative

εμβολιασμος\_εμβολιασμοι

εκανε\_εμβολιο

εμβολιασμοι\_εμβολιο

κανω\_εμβολιο

εμβολια\_εμβολιασμοι

εμβολια\_εμβολιασμος

εμβολιασμοι\_negative

Min. Criterion:

confidence

Min. Criterion Value:

| No. | Premises                 | Conclusion             | Support | Confiden... ↓ | LaPlace | Gain   | p-s   | Lift   | Convicti... |
|-----|--------------------------|------------------------|---------|---------------|---------|--------|-------|--------|-------------|
| 27  | κανει_εμβολιο            | κανει                  | 0.060   | 1             | 1       | -0.060 | 0.054 | 9.269  | ∞           |
| 28  | αντιεμβολιαστες_negative | αντιεμβολιαστες        | 0.021   | 1             | 1       | -0.021 | 0.019 | 10.277 | ∞           |
| 29  | εμβολια_εμβολιασμος      | εμβολια                | 0.012   | 1             | 1       | -0.012 | 0.011 | 12.446 | ∞           |
| 30  | εμβολιασμος_εμβολιασμοι  | εμβολιασμοι            | 0.019   | 1             | 1       | -0.019 | 0.018 | 15.176 | ∞           |
| 31  | εμβολιασμοι_εμβολιο      | εμβολιασμοι            | 0.016   | 1             | 1       | -0.016 | 0.015 | 15.176 | ∞           |
| 32  | εμβολιο_εμβολιασμοι      | εμβολιασμοι            | 0.013   | 1             | 1       | -0.013 | 0.012 | 15.176 | ∞           |
| 33  | εμβολιασμοι_negative     | εμβολιασμοι            | 0.012   | 1             | 1       | -0.012 | 0.011 | 15.176 | ∞           |
| 34  | κανουν_εμβολιο           | κανουν                 | 0.026   | 1             | 1       | -0.026 | 0.025 | 18.080 | ∞           |
| 35  | κανω_εμβολιο             | κανω                   | 0.014   | 1             | 1       | -0.014 | 0.013 | 34.058 | ∞           |
| 36  | εκανε_εμβολιο            | εκανε                  | 0.017   | 1             | 1       | -0.017 | 0.016 | 34.869 | ∞           |
| 26  | συνταγμα                 | αντιεμβολιαστες        | 0.013   | 0.619         | 0.992   | -0.030 | 0.011 | 6.362  | 2.370       |
| 25  | εκανε                    | εκανε_εμβολιο          | 0.017   | 0.583         | 0.988   | -0.041 | 0.016 | 34.869 | 2.360       |
| 24  | κανει                    | κανει_εμβολιο          | 0.060   | 0.556         | 0.957   | -0.156 | 0.054 | 9.269  | 2.117       |
| 23  | κανουν                   | κανουν_εμβολιο         | 0.026   | 0.477         | 0.973   | -0.084 | 0.025 | 18.080 | 1.863       |
| 22  | κανω                     | κανω_εμβολιο           | 0.014   | 0.473         | 0.985   | -0.045 | 0.013 | 34.058 | 1.871       |
| 21  | εμβολιασμοι              | εμβολιασμος_εμβολια... | 0.019   | 0.285         | 0.956   | -0.113 | 0.018 | 15.176 | 1.372       |
| 20  | εμβολιασμοι              | εμβολια                | 0.018   | 0.280         | 0.955   | -0.113 | 0.013 | 3.482  | 1.277       |

Data

Graph

Description

Annotations

[αντιεμβολιαστες] --> [εμβολια] (confidence: 0.139)

[γιατι] --> [κανει] (confidence: 0.144)

[αντιεμβολιαστες] --> [εμβολιασμοι] (confidence: 0.145)

[εμβολια] --> [εμβολια\_εμβολιασμος] (confidence: 0.150)

[εμβολια] --> [αντιεμβολιαστες] (confidence: 0.169)

[εμβολιασμοι] --> [εμβολιασμοι\_negative] (confidence: 0.180)

[εμβολιασμοι] --> [εμβολιο\_εμβολιασμοι] (confidence: 0.192)

[εμβολιασμοι] --> [αντιεμβολιαστες] (confidence: 0.214)

[εμβολιασμοι] --> [κρουσματα] (confidence: 0.216)

[αντιεμβολιαστες] --> [αντιεμβολιαστες\_negative] (confidence: 0.219)

[εμβολια] --> [εμβολιασμοι] (confidence: 0.229)

[εμβολιασμοι] --> [εμβολιασμοι\_εμβολιο] (confidence: 0.242)

[εμβολιασμενοι] --> [ανεμβολιαστοι] (confidence: 0.256)

[ανεμβολιαστοι] --> [αντιεμβολιαστες] (confidence: 0.264)

[κρουσματα] --> [εμβολιασμοι] (confidence: 0.265)

[ανεμβολιαστοι] --> [εμβολιασμενοι] (confidence: 0.276)

[εμβολιασμοι] --> [εμβολια] (confidence: 0.280)

[εμβολιασμοι] --> [εμβολιασμος\_εμβολιασμοι] (confidence: 0.285)

[κανω] --> [κανω\_εμβολιο] (confidence: 0.473)

[κανουν] --> [κανουν\_εμβολιο] (confidence: 0.477)

[κανει] --> [κανει\_εμβολιο] (confidence: 0.556)

[εκανε] --> [εκανε\_εμβολιο] (confidence: 0.583)

[συνταγμα] --> [αντιεμβολιαστες] (confidence: 0.619)

[κανει\_εμβολιο] --> [κανει] (confidence: 1.000)

[αντιεμβολιαστες\_negative] --> [αντιεμβολιαστες] (confidence: 1.000)

[εμβολια\_εμβολιασμος] --> [εμβολια] (confidence: 1.000)

[εμβολιασμος\_εμβολιασμοι] --> [εμβολιασμοι] (confidence: 1.000)

[εμβολιασμοι\_εμβολιο] --> [εμβολιασμοι] (confidence: 1.000)

[εμβολιο\_εμβολιασμοι] --> [εμβολιασμοι] (confidence: 1.000)

[εμβολιασμοι\_negative] --> [εμβολιασμοι] (confidence: 1.000)

[κανουν\_εμβολιο] --> [κανουν] (confidence: 1.000)

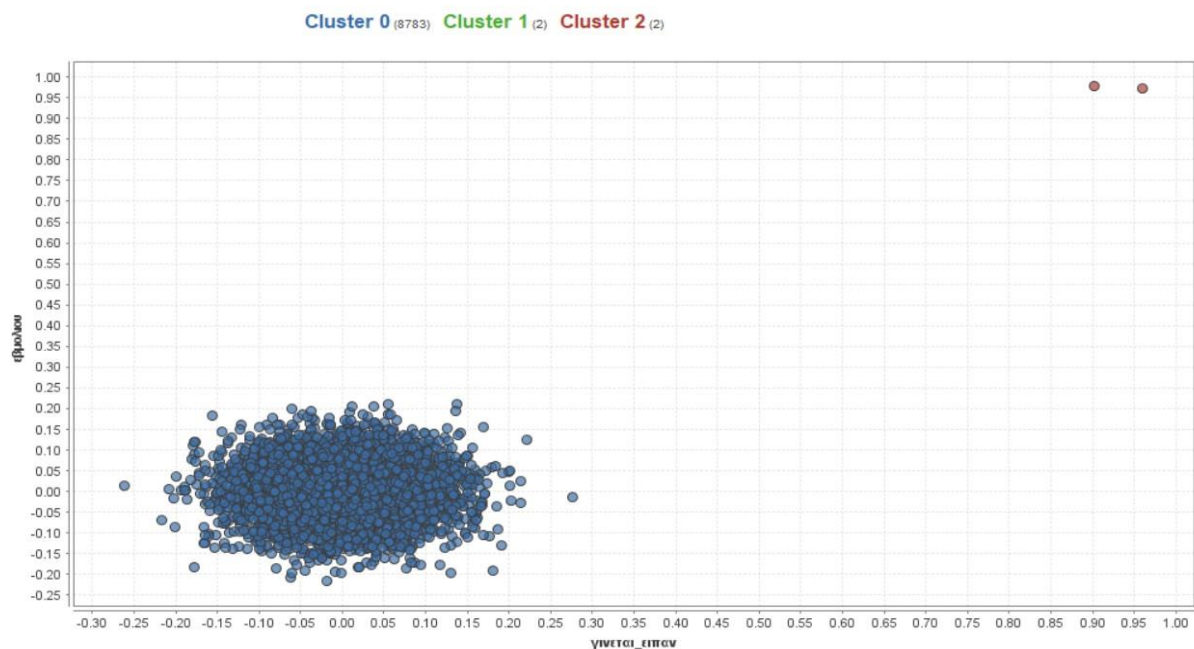
[κανω\_εμβολιο] --> [κανω] (confidence: 1.000)

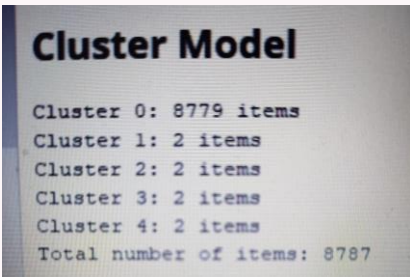
[εκανε\_εμβολιο] --> [εκανε] (confidence: 1.000)

Εικόνα 14. Κανόνες συσχέτισης αρνητικού όφους

## ΟΜΑΔΟΠΟΙΗΣΗ – ΣΥΣΤΑΔΟΠΟΙΗΣΗ (CLUSTERING)

Στην προσπάθεια να ερευνήσουμε περισσότερο τα αρνητικά tweets εκτελέσαμε ομαδοποίηση αυτών, θεωρώντας ότι, μαζί με τους κανόνες συσχέτισης θα μας βοηθούσε να εξάγουμε πιο χρήσιμη πληροφορία. Επειδή δε γνωρίζαμε από πριν τον κατάλληλο αριθμό των κλάσεων προχωρήσαμε την ομαδοποίηση με τη χρήση του αλγορίθμου x-means ούτως ώστε να υπολογίσει μόνος του το μοντέλο τον ιδανικό αριθμό των κλάσεων. Το μοντέλο δημιούργησε αυτόματα τρεις (3) συνολικά κλάσεις όπου η πρώτη περιέχει 8783 παραδείγματα, η δεύτερη και η τρίτη δυο (2) όπως φαίνεται και στην Εικόνα 15 που ακολουθεί συνεπώς δε μας βοηθά περαιτέρω. Ομοίως με τη χρήση του αλγορίθμου k-means και τον ορισμό κλάσεων από το χρήστη δεν είχαμε καλύτερα αποτελέσματα. Επειδή δεν υπάρχει κάποιος αυτοματοποιημένος τρόπος επιλογής του “k” σε τυχαίες δοκιμές, επηρεασμένοι και από το x-means που δημιούργησε 3 συστάδες, εκτελέσαμε ομαδοποίηση με 3, 5, 7, 18 και 80 συστάδες αντίστοιχα. Όλες τις φορές είχαμε ως αποτέλεσμα μια συστάδα που συγκέντρωνε τη συντριπτική πλειοψηφία των χαρακτηριστικών χωρίς να δίνει καμία χρήσιμη πληροφορία και τις υπόλοιπες συστάδες με ελάχιστα παραδείγματα χωρίς αξία. Ακόμη και η επιλογή των 80 συστάδων όπου αναμέναμε ισομοιρασμό των 8.000 και πλέον παραδειγμάτων είχαμε ως αποτέλεσμα παρόμοια συμπεριφορά.

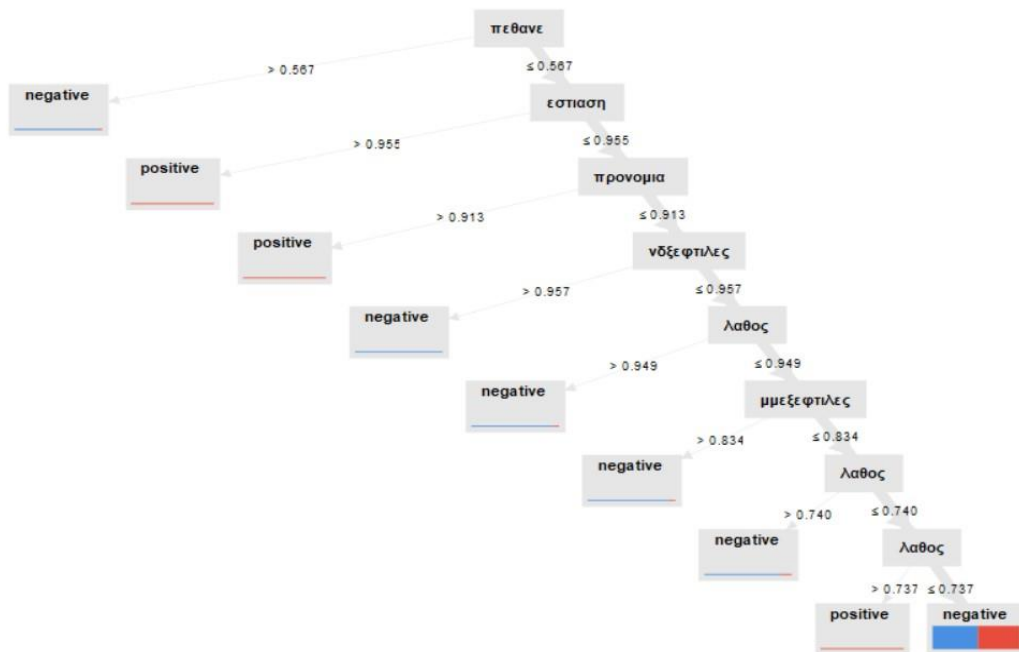




Εικόνα 15. Ομαδοποίηση (clustering) 3 συστάδων επάνω και 5 κάτω εικόνα

## ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ

Μια ενδιαφέρουσα προσέγγιση είναι αυτή που προέρχεται από τα δέντρα απόφασης. Με τη χρήση του αλγορίθμου “decision tree” στην εφαρμογή RapidMiner Studio παίρνω τα εξής αποτελέσματα που φαίνονται και στην εικόνα που ακολουθεί.



## Tree

```
πεθανε > 0.567: negative {negative=179, positive=8}
πεθανε ≤ 0.567
|   εστίαση > 0.955: positive {negative=0, positive=6}
|   εστίαση ≤ 0.955
|   |   προνομία > 0.913: positive {negative=0, positive=3}
|   |   προνομία ≤ 0.913
|   |   |   νδξεφτίλες > 0.957: negative {negative=2, positive=0}
|   |   |   νδξεφτίλες ≤ 0.957
|   |   |   |   λαθος > 0.949: negative {negative=13, positive=1}
|   |   |   |   λαθος ≤ 0.949
|   |   |   |   |   μμεξεφτίλες > 0.834: negative {negative=13, positive=1}
|   |   |   |   |   μμεξεφτίλες ≤ 0.834
|   |   |   |   |   |   λαθος > 0.740: negative {negative=34, positive=5}
|   |   |   |   |   |   λαθος ≤ 0.740
|   |   |   |   |   |   |   λαθος > 0.737: positive {negative=0, positive=2}
|   |   |   |   |   |   |   λαθος ≤ 0.737: negative {negative=8582, positive=8024}
```

### Εικόνα 16. Δέντρο απόφασης (decision tree)

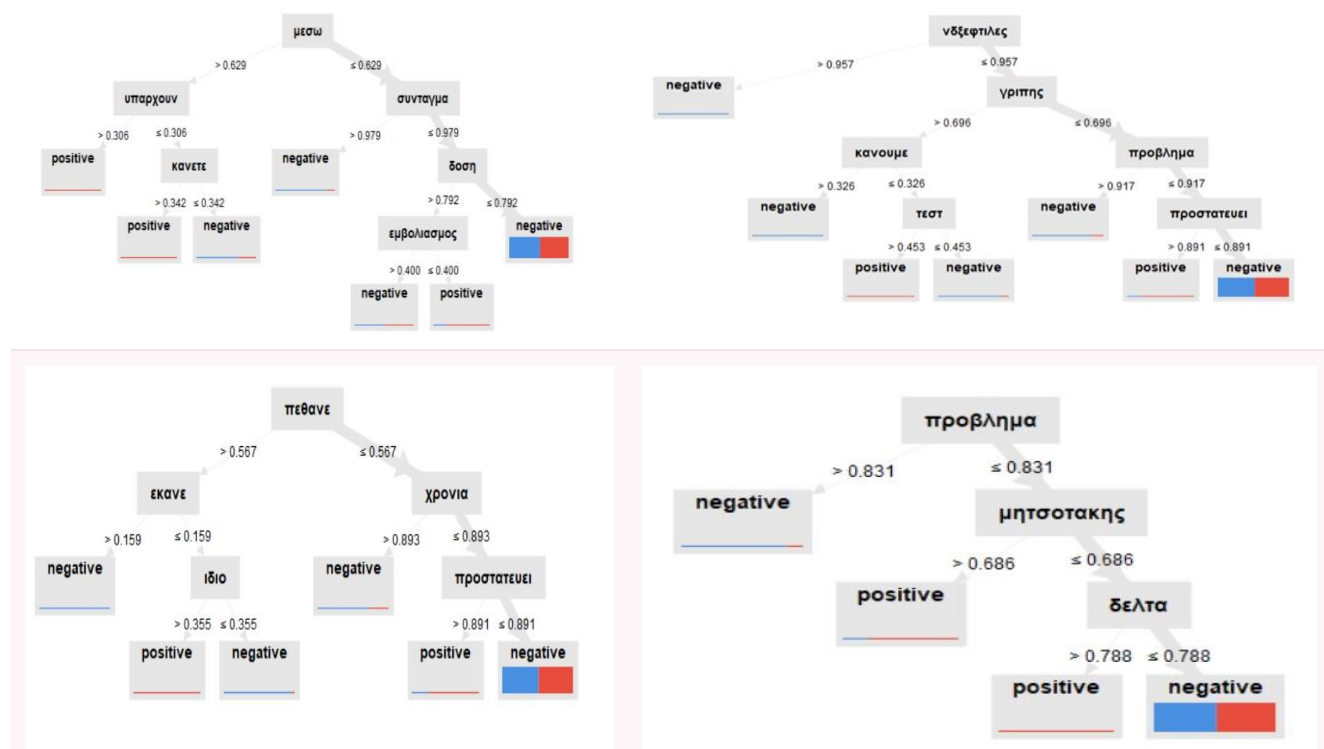
Για τη δημιουργία του δέντρου απόφασης επιλέχθηκαν τα κείμενα με αρνητικό ύφος. Στη συνέχεια έγινε διανυσματοποίηση τους με την ίδια διαδικασία που αναφέρθηκε παραπάνω στους κανόνες συσχέτισης με μόνη διαφορά ότι αυτή τη φορά επιλέχθηκε κριτήριο TF-IDF με μέθοδο «ψαλιδίσματος» (prune method) μικρότερο από 1% και μεγαλύτερο από 66%. Αυτό οδήγησε σε ένα πιο διαχειρίσιμο αριθμό χαρακτηριστικών (attributes) περί τα 2100 αυτή τη φορά. Ο λόγος που επιλέχθηκε η μέθοδος Terms Frequency – Inverse Document Frequency είναι διότι μπορεί να ποσοτικοποιήσει τη συνάφεια προσθέτοντας «βαρύτητα» σε εκείνα τα attributes σύμφωνα με τη συχνότητα εμφάνισης τους στο κείμενο σε σχέση με την αντίστροφη εμφάνισή τους σε ολόκληρο το σετ δεδομένων. Με λίγα λόγια αποδίδει μεγαλύτερη βαρύτητα σε λέξεις που εμφανίζονται σπανιότερα σε σχέση με τις κοινότερες λέξεις που κυριαρχούν στο σετ δεδομένων και δεν έχουν καμία επιπρόσθετη αξία.

Από το δέντρο φαίνεται καθαρά πως καθοριστικής σημασίας κριτήριο για να γραφτεί ένα σχόλιο με αρνητικό ύφος έχει να κάνει με τη λέξη κλειδί «πέθανε». Ακολουθεί η «εστίαση» όπου στις περισσότερες φορές που απαντάται στα δεδομένα μας έχει να κάνει με θετικά σχόλια εκτός αν γίνεται αναφορά στα «προνόμια». Τις περισσότερες φορές που υπάρχει η λέξη κλειδί «προνόμια» στα δεδομένα έχει θετικό ύφος εκτός αν ακολουθείται από πολιτικό σχόλιο όπως πχ με τη λέξη κλειδί «νδξεφτίλες» όπου έχει να κάνει με αρνητικού ύφος σχόλιο. Τέλος αν μετά τα ανωτέρω εξακολουθεί να αναφέρεται και η λέξη κλειδί «λάθος» τότε τα σχόλια είναι αρνητικού ύφους.

## RANDOM FOREST

Αρκετά χρήσιμα στην περίπτωσή μας αποτελέσματα μπορούμε να εξάγουμε με τη χρήση του αλγόριθμου Random Forest επιλέγοντας τη δημιουργία 30 (τυχαίων) διαφορετικών δέντρων απόφασης. Με τη «σοφία του πλήθους» που τον χαρακτηρίζει και τη χρήση μερικών εκ των

δέντρων που δημιουργήθηκαν μπορούμε να εξάγουμε χρήσιμα συμπεράσματα. Εστιάζοντας σε ορισμένα -από τα τριάντα (30) τυχαία- δέντρα παρατηρούμε σε ένα εξ αυτών ότι όταν αναφέρεται η λέξη «Σύνταγμα» στο σετ κυριαρχεί το αρνητικό ύφος. Σε αντίθετη περίπτωση γίνεται αναφορά στον «εμβολιασμό» και το ύφος ισομοιράζεται.

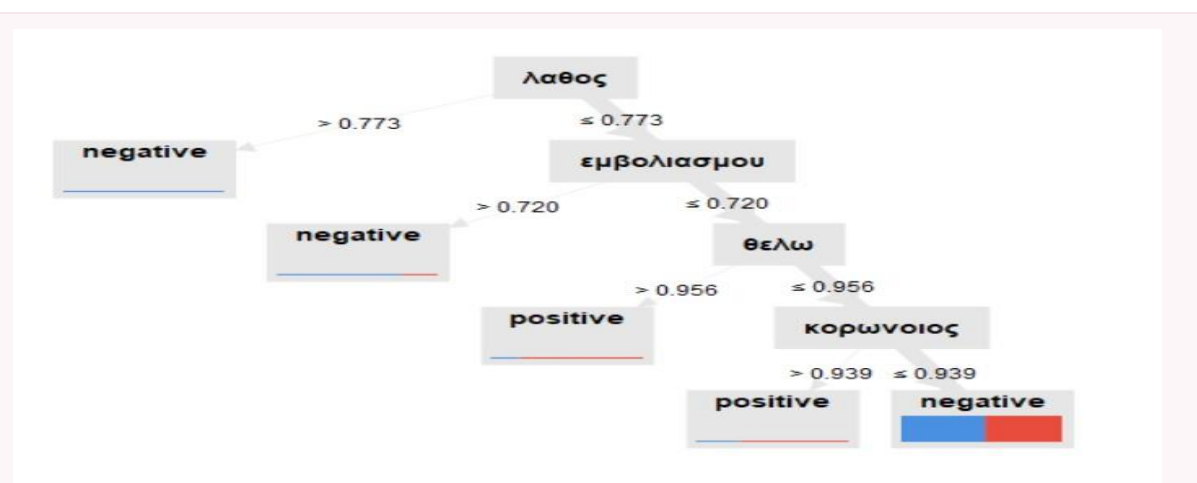
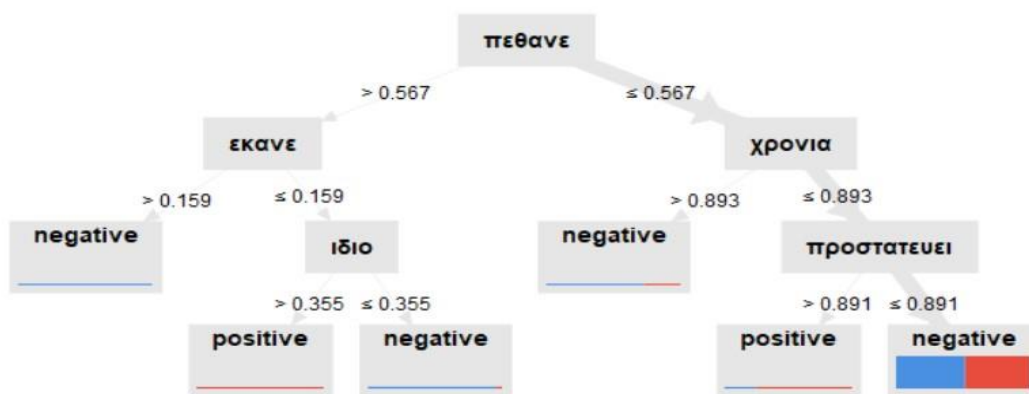


Εικόνα 17. Random Forest -Παρατίθενται 4 από τα 30 δέντρα

Σε ένα άλλο δέντρο παρατηρούμε ότι ένα αρχικό κριτήριο έχει να κάνει με την κυβέρνηση και τους κυβερνητικούς χειρισμούς που επί το πλείστον κυριαρχούν τα αρνητικά ύφους σχόλια. Όταν όμως ακολουθεί πιθανή αναφορά σε εμβόλιο πχ της γρίπης και ότι πρέπει να γίνει ή γίνεται αναφορά σε κάποιο πρόβλημα τότε πάλι κυριαρχεί το αρνητικό ύφος εκτός αν γίνονται αναφορές στην προστασία που αυτό προσφέρει.

Ένα άλλο δέντρο που ξεχωρίσαμε και κρίναμε άξιο αναφοράς έχει να κάνει με τη μετάλλαξη δέλτα του κορονοϊού, την εστίαση και την υποχρεωτικότητα. Παρατηρούμε ότι όπου γίνονται αναφορές σε αυτά οδηγεί σε θετικό ύφος.

Επίσης για τους σκοπούς της εργασίας μας ξεχωρίσαμε και τα ακόλουθα.



Εικόνα 18. Random Forest – επιπλέον 2 από 30 δέντρα

Από τα δέντρα στην Εικόνα 18 φαίνεται το αρνητικό ύφος που έχουν τα tweets με τις λέξεις «πέθανε», «έκανε» και «λάθος», «εμβολιασμού» που αναφέρονται προφανώς σε εμβόλια και εμβολιασμό ενώ όταν γίνεται αναφορά σε «χρόνια» και «προστατεύει» ή «θέλω» (πιθανότητα αναφέρεται σε εμβολιασμό) και «κορωνοϊός» τότε συνήθως οδηγούμαστε σε θετικό ύφος στην πλειοψηφία των περιπτώσεων χωρίς βέβαια να εκλείπει το αρνητικό ύφος που απαντάται σε μικρότερη συχνότητα.

Πριν ολοκληρώσουμε τη μελέτη του αρνητικού ύφους, με τη χρήση της επαγωγής θα προσπαθήσουμε να προσεγγίσουμε ένα νέο μοντέλο ταξινόμησης των παραδειγμάτων μας.

## ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ – ΤΑΞΙΝΟΜΗΣΗ

Με τη βοήθεια του εργαλείου RapidMiner Studio εκτελέσαμε ταξινόμηση (classification). Για τις ανάγκες της εργασίας δε συμπεριλήφθηκαν τα παραδείγματα με ουδέτερο ύφος στην ταξινόμηση διότι πέρα από το γεγονός ότι δε μας απασχολούν, θα περιόριζαν την απόδοση των μοντέλων και του RapidMiner Studio γενικότερα. Επίσης στην προσπάθεια αποφυγής

προβλημάτων επεξεργασίας έγινε περιορισμός των 45.000+ attributes με τη χρήση της μεθόδου prune σε ποσοστιαία μείωση 1% & 33% (prune below - prune above) ούτως ώστε το σετ να είναι επεξεργάσιμο<sup>2</sup> και να οδηγεί σε αποτέλεσμα. Οπότε με την επιλογή μόνο θετικών και αρνητικών παραδειγμάτων και τη χρήση επί αυτών έξι (6) διαφορετικών μοντέλων παρατίθενται τα αποτελέσματα που ακολουθούν.

| Αλγόριθμος               | Accuracy |
|--------------------------|----------|
| Naïve Bayes              | 49.3%    |
| Generalized Linear Model | 64.5%    |
| Fast Large Margin        | 60%      |
| Decision Tree            | 48.7%    |
| Random Forest            | 48.7%    |
| Gradient Boosted Trees   | 64.9%    |

Πίνακας 4. Μετρική Accuracy διαφόρων μοντέλων.

Με τη μετρική Accuracy που μας δείχνει το ποσοστό των σωστά ταξινομημένων παραδειγμάτων στο σύνολο των παραδειγμάτων φαίνεται ένα σαφέστατο προβάδισμα στο μοντέλο με τη χρήση του αλγορίθμου Gradient Boosted Trees. Για το κάθε ένα από τα ανωτέρω παρατίθενται και ο ανάλογος πίνακας σύγχυσης ή αλλιώς confusion matrix.

| Gradient Boosted Trees | TN     | TP     | precision |
|------------------------|--------|--------|-----------|
| Pred. Negative         | 2261   | 1432   | 61.22%    |
| Pred. Positive         | 259    | 868    | 77.02%    |
| recall                 | 89.72% | 37.74% |           |

Πίνακας 5. Confusion Matrix Gradient Boosted Trees

| Random Forest  | TN    | TP     | precision |
|----------------|-------|--------|-----------|
| Pred. Negative | 49    | 2      | 96.08%    |
| Pred. Positive | 2472  | 2298   | 48.18%    |
| recall         | 1.94% | 99.91% |           |

Πίνακας 6. Confusion Matrix Random Forest

<sup>2</sup> Χρήση όλων των attributes οδήγησε επανειλημμένως σε κατάρρευση το RapidMiner Studio.

| <b>Decision Tree</b>  | <b>TN</b> | <b>TP</b> | <b>precision</b> |
|-----------------------|-----------|-----------|------------------|
| <b>Pred. Negative</b> | 50        | 5         | 90.91%           |
| <b>Pred. Positive</b> | 2470      | 2295      | 48.16%           |
| <b>recall</b>         | 1.98%     | 99.78%    |                  |

*Πίνακας 7. Confusion Matrix Decision Tree*

| <b>Fast Large Margin</b> | <b>TN</b> | <b>TP</b> | <b>precision</b> |
|--------------------------|-----------|-----------|------------------|
| <b>Pred. Negative</b>    | 773       | 178       | 81.28%           |
| <b>Pred. Positive</b>    | 1748      | 2122      | 54.83%           |
| <b>recall</b>            | 30.66%    | 92.26%    |                  |

*Πίνακας 8. Confusion Matrix Fast Large Margin*

| <b>Generalized Linear Model</b> | <b>TN</b> | <b>TP</b> | <b>precision</b> |
|---------------------------------|-----------|-----------|------------------|
| <b>Pred. Negative</b>           | 1092      | 282       | 79.48%           |
| <b>Pred. Positive</b>           | 1428      | 2018      | 58.56%           |
| <b>recall</b>                   | 43.33%    | 87.74%    |                  |

*Πίνακας 9. Confusion Matrix Generalized Linear Model*

| <b>Naïve Bayes</b>    | <b>TN</b> | <b>TP</b> | <b>precision</b> |
|-----------------------|-----------|-----------|------------------|
| <b>Pred. Negative</b> | 89        | 13        | 87.25%           |
| <b>Pred. Positive</b> | 2432      | 2287      | 48.46%           |
| <b>recall</b>         | 3.53%     | 99.43%    |                  |

*Πίνακας 10. Confusion Matrix Naive Bayes*

Με τη βοήθεια των μετρικών από τους παραπάνω πίνακες σύγκρισης βλέπουμε ότι γενικά καλύτερες αποδόσεις έχουν και οι:

- Fast Large Margin,
- Generalized Linear Model

Αυτό φαίνεται ξεκάθαρα από τις τιμές των μετρικών precision και recall. Η πρώτη μετρική αναφέρεται στο ποσοστό που είναι θετικά σε σχέση με όλα τα θετικά που έχουν προβλεφθεί και η δεύτερη το ποσοστό που προβλέφθηκε θετικό σε σχέση με το σύνολο των θετικών.

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN})$$

Εν αντιθέσει με τους Naive Bayes, Decision Tree και Random Forest που έχουν απογοητευτικά χαμηλά σκορ στη μετρική Recall στην πρόβλεψη των αρνητικών, διότι βλέπουμε ότι τα

κατατάσσουν όλα τα παραδείγματα σε θετικά με αντίστοιχα θεαματικά υψηλό σκορ της μετρικής στα θετικά.

Οι Gradient Boosted Trees, Decision Tree και Random Forest έχουν αρκετές ομοιότητες μεταξύ τους. Μπορούμε κάλλιστα να πούμε ότι ανήκουν στην ίδια οικογένεια, την οικογένεια των δέντρων απόφασης όπου είναι ιδιαίτερα χρήσιμοι και στην εξαγωγή κανόνων ταξινόμησης μέσω ενός ιεραρχικού μοντέλου αποφάσεων. Προσπαθώντας να απλοποιήσουμε τον τρόπο όπου λειτουργούν οι συγκεκριμένοι αλγόριθμοι θα λέγαμε ότι αρχικά στόχος είναι να βρίσκουν το χαρακτηριστικό με το μέγιστο κέρδος σε πληροφορία και να το τοποθετούν στην κορυφή – τη ρίζα- του δέντρου. Το αμέσως επόμενο στο επόμενο κλαδί δημιουργώντας κόμβους κοκ. Όπως ήταν αναμενόμενο η δημιουργία μοντέλου με τη χρήση του Gradient Boosted Trees επέφερε υψηλότερο σκορ συγκριτικά με τους Decision Tree και Random Forest. Αυτό οφείλεται στο χαρακτηριστικό του (boosting) ότι δηλαδή κάθε νέο δέντρο που δημιουργεί «φροντίζει» να είναι καλύτερο από το προηγούμενο διορθώνοντας τα λάθη του προηγούμενου δέντρου. Όποτε από την μεγάλη «οικογένεια των δέντρων» καταγράφει τα καλύτερα αποτελέσματα.

Ο Naive Bayes βασίζεται περισσότερο στα μαθηματικά και συγκεκριμένα στις πιθανότητες υπολογίζοντας την πιθανότητα ενός προτύπου να ανήκει στη συγκεκριμένη κλάση. Στα μαθηματικά επίσης βασίζονται οι Fast Large Margin και Generalized Linear Model. Ο πρώτος εκπαιδεύει γραμμικούς κατηγοριοποιητές ενώ ο δεύτερος είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης που επιτρέπει μεταβλητές απόκρισης που έχουν μοντέλα κατανομής σφάλματος διαφορετικά από μια απλή κατανομή (Hayes, 2019).

Ένας από τους κύριους λόγους που πέτυχαν αρκετά χαμηλά σκορ οι αλγόριθμοι που χρησιμοποιήθηκαν έχει να κάνει με το γεγονός ότι στο συγκεκριμένο σετ δεδομένων με τη χρήση της μεθόδου prune έγινε μια γενναία περικοπή των χαρακτηριστικών attributes συγκριτικά με το εξαιρετικά βραδυκίνητο προς επεξεργασία αρχικό τους αριθμό. Αυτό επηρέασε κατά πολύ την εκπαίδευση των μοντέλων όπου χρειαζόνταν περισσότερα και πιο ποιοτικά attributes για τη δημιουργία πιο αποδοτικών μοντέλων. Επιπλέον θα ήταν μείζονος σημασίας η πραγματοποίηση feature selection για την επιλογή των πιο κατάλληλων χαρακτηριστικών για ταξινόμηση<sup>3</sup>.

Σε αυτό ίσως μπορεί να αποδοθεί και ο μεγάλος όγκος σε False Positive ταξινομημένα παραδείγματα μερικών μοντέλων. Μια επιπλέον μείωση των True Negative σε αριθμό καθώς είναι περισσότερα των True Positive με ταυτόχρονη χρήση feature selection για επιλογή πιο κατάλληλων χαρακτηριστικών ίσως βοηθήσει τα συγκεκριμένα μοντέλα να σκοράρουν καλύτερα αποτελέσματα.

---

<sup>3</sup> Προέκυψαν περιορισμοί στην εκτέλεση μέσω RapidMiner Studio καθώς παρατηρήθηκαν υψηλές απαιτήσεις σε πόρους συστήματος σε συνδυασμό με υπερβολικά μεγάλα χρονικά διαστήματα για την κατασκευή ενός μοντέλου.

Επίσης ένας εξίσου σημαντικός λόγος τον οποίο έχουμε αναφέρει και παραπάνω έχει να κάνει με την ποιότητα των δεδομένων. Παρόλο που στη μηχανική μάθηση, γίνεται χρήση εξελιγμένων αλγορίθμων για την εξαγωγή χρήσιμης γνώσης από αδόμητα δεδομένα, ισχύει περίτρανα ο όρος “GiGo” δηλ. garbage in garbage out. Δηλαδή, δε θα γίνει καμία χρήσιμη εξαγωγή πληροφορίας με αυτόματο τρόπο από δεδομένα που περιέχουν σφάλματα, ή γενικά είναι κακής ποιότητας αλλά, αυτό θα οδηγήσει σε κακής ποιότητας, μη χρήσιμα ή και ακόμη σε μη κατανοητά αποτελέσματα. Άρα, επί της ουσίας, δε θα πραγματοποιηθεί μηχανική μάθηση! Έτσι και στο παράδειγμά μας η μη ύπαρξη βιβλιοθηκών ελληνικής γλώσσας στο RapidMiner Studio οδήγησε σε καταναγκαστικές «εκπτώσεις» τόσο στο φιλτράρισμα μέσω των stopwords όσο και στη διαδικασία stemming. Άμεση συνέπεια αυτού η υποβάθμιση κατά πολύ της ποιότητας των αποτελεσμάτων δημιουργώντας ένα εξαιρετικά δύσχρηστο προς επεξεργασία σετ διανυσματοποιημένων δεδομένων, άνω των 45.000 χαρακτηριστικών που στη πλειοψηφία τους ήταν μη ποιοτικά.

## ΣΥΜΠΕΡΑΣΜΑΤΑ

Με τη βοήθεια τεχνικών μηχανικής μάθησης έγινε μελέτη και ανάλυση δεδομένων σχετικά με τον εμβολιασμό κατά τη διάρκεια της πανδημίας προερχόμενα από την πλατφόρμα κοινωνικής δικτύωσης του Twitter. Το Twitter ως ένα μέσο κοινωνικής δικτύωσης προσέφερε απλόχερα δεδομένα για την πανδημία και τον εμβολιασμό καθώς, όπως φαίνεται, προτιμάται για **άμεση** ενημέρωση τόσο από κυβερνητικούς όσο και από ιατρικούς και ερευνητικούς οργανισμούς. Ίσως και γι’ αυτό προσέελκυσε το έντονο ενδιαφέρον του επιχειρηματία Elon Musk για την εξαγορά του τον Απρίλιο του 2022. Ομοίως εξακολουθεί και ο υπόλοιπος πλανήτης να το χρησιμοποιεί για να ενημερώνεται και να εκφράζει ταυτόχρονα και τις προσωπικές του απόψεις.

Μέσω του εργαλείου RapidMiner Studio προσπαθήσαμε να εξάγουμε χρήσιμη γνώση και πληροφορίες, από ένα τεράστιο σετ δεδομένων σχετικό με τον εμβολιασμό, για τους λόγους που η ελληνική κοινωνία αντιμετώπισε με δισταγμό το εμβόλιο κατά του κορονοϊού. Τα αρνητικά συναισθήματα και κατ’ επέκταση το αρνητικό ύφος που εκφράστηκε κατά κύριο λόγο φάνηκε να είναι αποτέλεσμα κατακριτέων κυβερνητικών χειρισμών σε ένα ευαίσθητο όπως αυτό θέμα. Ο «εξαναγκασμός» που χρησιμοποιήθηκε ως τακτική από την κυβέρνηση και η υποχρεωτικότητα του εμβολιασμού μέσω αποκοπής – αποκλεισμού των μη εμβολιασμένων από πολλές κοινωνικές δραστηριότητες όπως ήταν λογικό επέφερε κύμα αντιδράσεων και σχολιάστηκε καυστικά στα μέσα κοινωνικής δικτύωσης.

Ήταν πλέον ολοφάνερο πως η κυβέρνηση δεν προσπάθησε ποτέ να κατανοήσει την κοινή γνώμη και την κοινωνία που αντιπροσώπευε, ούτως ώστε να λάβει πιο «προσωποποιημένα» για την κοινωνία μέτρα. Αντιθέτως υιοθέτησε και επέβαλε αυστηρές τακτικές που δεν ήταν έτοιμη η κοινωνία να δεχτεί. Αυτό φαίνεται από τις αναφορές αρνητικού ύφους που γίνονται στο σύνταγμα και κατ’ επέκταση στις συνταγματικές ελευθερίες καθώς και γενικά στις αρνητικές αναφορές που συγκεντρώνει η κυβέρνηση εκείνο το χρονικό διάστημα.

Στην αντίπερα όχθη αντιπαρατάσσεται το παράδειγμα κυβέρνησης που «άκουσε» τη φωνή του λαού και της κοινής γνώμης, αυτό της Σουηδίας. Παρά το γεγονός ότι επικρίθηκε από πολλούς ως ακραίο και επικίνδυνο για τη δημόσια υγεία παράδειγμα, τόλμησε και δεν έλαβε μέτρα περιορισμού των ελευθεριών, όπως οι υπόλοιπες χώρες. Επίσης δεν ακολούθησε την τακτική εξαναγκασμού στους πολίτες της, χωρίς αυτό να σημαίνει πως όποτε βρέθηκε αντιμέτωπη με εξάρσεις ότι δεν έλαβε ουσιαστικά και δραστικά μέτρα, υιοθετώντας γενικότερα μέτρα κοινωνικώς αποδεχτά και χωρίς αντιδράσεις. Πέρα από της παγκόσμιας κλίμακας αρνητικές επιπτώσεις που επέφερε ο κορονοϊός, εκ των υστέρων το πείραμα της Σουηδίας στέφθηκε με απόλυτη επιτυχία, αφού εν έτη 2022 παραμένει στην πρώτη δεκάδα των χωρών με τους πιο ευτυχισμένους κατοίκους στον πλανήτη (Report, 2022).

Στην προσπάθεια να εμβαθύνουμε περισσότερο στους λόγους του αρνητικού ύφους δεν είχαμε την αντίστοιχη επιτυχία καθώς το εργαλείο RapidMiner Studio δεν υποστηρίζει βιβλιοθήκες για την ελληνική γλώσσα στην ανάλυση κειμένου. Άμεσο αποτέλεσμα αυτού να μην μπορούμε να προσεγγίσουμε, στο βάθος τουλάχιστον που θα επιθυμούσαμε, την εξαγωγή γνώσης του αρνητικού ύφους που κυριαρχούσε κατά των εμβολίων στο συγκεκριμένο σετ δεδομένων και στην κοινωνία γενικότερα.

Επίσης πέρα των περιορισμών του εργαλείου, ένα άλλο στοιχείο καταπέλτης που απαντάται συχνά στην ανάλυση κειμένων, έχει να κάνει με το ειρωνικό ύφος των χρηστών που είναι αδύνατον να εντοπιστεί και επηρεάζει αρνητικά την εξαγωγή γνώσης από τα κείμενα αυτά.

Σε αυτό το σημείο, κλείνοντας, να υπενθυμίσουμε πως οποιοδήποτε γενικευμένο για τον πληθυσμό συμπέρασμα πέρα από τους ερευνητικούς σκοπούς της συγκεκριμένης εργασίας θα ήταν λανθασμένο διότι το δείγμα μας δε θεωρείται αντιπροσωπευτικό.

## ΒΙΒΛΙΟΓΡΑΦΙΑ

About-Twitter. (χ.χ.). Ανάκτηση από [about.twitter.com](https://about.twitter.com/en/who-we-are/our-company): <https://about.twitter.com/en/who-we-are/our-company>

Eibensteiner F, R. V.-K.-P. (2021). *People's Willingness to Vaccinate Against COVID-19 Despite Their Safety Concerns: Twitter Poll Analysis*. J Med Internet Res.

Hayes, G. (2019, Sep 22). Ανάκτηση από [https://towardsdatascience.com](https://towardsdatascience.com/https://towardsdatascience.com/beyond-linear-regression-an-introduction-to-glms-7ae64a8fad9c): <https://towardsdatascience.com/beyond-linear-regression-an-introduction-to-glms-7ae64a8fad9c>

Hegarty, S. (2020, February). Ανάκτηση από [www.bbc.com](http://www.bbc.com): <https://www.bbc.com/news/world-asia-china-51364382>

IBM. (2021). *Data mining*. Ανάκτηση από <https://www.ibm.com/cloud/learn/data-mining>

iefimerida.gr. (2021, July 17). Ανάκτηση από <https://www.iefimerida.gr/>

- Lazarus JV, R. S. (2021). *A global survey of potential acceptance of a COVID-19 vaccine*. Nature Medicine.
- Report, W. H. (2022, March 03). *www.fortunegreece.com*. Ανάκτηση από <https://www.fortunegreece.com/photo-gallery/i-deka-pio-eftichismenes-chores-tou-kosmou-gia-to-2022-ke-i-thesi-tis-elladas/#3>
- Sohaib R Rufai, C. B. (2020). World leaders' usage of Twitter in response to the COVID-19 pandemic: a content analysis. *Journal of Public Health*, 510-516.
- Twitter-Help-Center. (χ.χ.). *help.twitter.com*. Ανάκτηση από <https://help.twitter.com/en/resources/new-user-faq>
- Twitter-Inc. (2020). *Fiscal year 2020 Annual report*. San Franscisco, California: twitter inc.
- WHO-Director. (2020, March). Ανάκτηση από [www.who.int](https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020): <https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020>
- Wikipedia. (χ.χ.). Ανάκτηση από [en.wikipedia.org](https://en.wikipedia.org/wiki/Twitter#:~:text=Twitter%20was%20created%20by%20Jack,billion%20search%20queries%20per%20day): <https://en.wikipedia.org/wiki/Twitter#:~:text=Twitter%20was%20created%20by%20Jack,billion%20search%20queries%20per%20day>.
- Επιστήμη Δεδομένων*. (χ.χ.). Ανάκτηση από <https://datascience.aueb.gr/>