ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΠΛΗΡΟΦΟΡΙΑΚΑ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

Αναγνώριση μη-έγκυρης πληροφορίας σχετικά με τη πανδημία του κορονοϊού Covid-19 σε πλατφόρμα κοινωνικής δικτύωσης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Λυγερού Γεώργιου
Α.Μ.3252019013

Επιβλέπων : Κωστούλας Θεόδωρος
Μέλη εξεταστικής επιτροπής: Σταματάτος Ευστάθιος, Συμεωνίδης Παναγιώτης

Σάμος, [Μάρτιος 2021]

Πρόλογος και ευχαριστίες

Θα ήθελα αρχικά να ευχαριστήσω τον καθηγητή κ. Κωστούλα Θεόδωρο που είχε την επίβλεψη της συγκεκριμένης Διπλωματικής Εργασίας για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον και σύγχρονο θέμα.

Επίσης θα ήθελα να ευχαριστήσω τους υιούς μου Κωνσταντίνο και Γιώργο για την κατανόηση που έδειξαν καθ' όλη την διάρκεια των σπουδών μου, καθώς και την σύζυγό μου Μαρίνα για την πίστη της σε εμένα χωρίς την οποία δεν θα ήταν δυνατή η ολοκλήρωση της παρούσας εργασίας.

Λυγερού Γεώργιου

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Πίνακας περιεχομένων

Λίστα Σχημάτων

Λίστα Πινάκων

Ακρωνύμια

PCA	Principal Component Analysis
SVM	Support Vector Machine
ML	Machine Learning
TF-IDF	Term Frequency- Inverse Document Frequency
NLP	Natural Language Processing
TTR	Type-to-Token Ratio
PoS	Part-of-Speech
DT	Decision Tree
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
API	Application Programming Interface

Περίληψη

Από την πρώτη στιγμή της εξάπλωσης της πανδημίας του κορονοϊού COVID-19, μία ακόμα πανδημία ξέσπασε ταυτόχρονα, αυτή της παραπληροφόρησης. Στο γεγονός αυτό καθοριστικό παράγοντα παίζουν τα μέσα κοινωνικής δικτύωσης, μέσω των οποίων οποιαδήποτε πληροφορία μεταδίδεται ραγδαία σε όλο τον κόσμο. Οι συνέπειες της μετάδοσης μη-έγκυρων πληροφοριών μπορεί να αποβούν χειρότερες από τις συνέπειες της ίδιας της πανδημίας. Θεωρίες συνωμοσίας και ψευδείς τρόποι θεραπείας είναι μόνο δύο από τις κατηγορίες που συναντώνται ευρέως και έχουν απρόβλεπτες συνέπειες στη δημόσια υγεία. Από την άλλη πλευρά, ο όγκος των πληροφοριών που διακινούνται καθημερινά καθιστά τον έλεγχο της αξιοπιστίας της πληροφορίας μία ιδιαίτερα απαιτητική πρόκληση.

Στην παρούσα Εργασία μελετάται η αυτόματη ανίχνευση μη έγκυρων ειδήσεων, που σχετίζονται με την εξελισσόμενη πανδημία του κορονοϊού στα κοινωνικά δίκτυα και συγκεκριμένα στο Twitter. Για το σκοπό αυτό αξιοποιούνται αλγόριθμοι Επεξεργασίας Φυσικής Γλώσσας (NLP) και Μηχανικής Μάθησης. Τα δεδομένα με τα οποία γίνεται η εκπαίδευση των αλγορίθμων προέρχονται από ένα δημόσια προσβάσιμο σύνολο δεδομένων το οποίο περιέχει tweets που σχετίζονται με την τρέχουσα πανδημία. Από το σύνολο των δεδομένων απομονώθηκε μόνο το περιεχόμενο που αφορά την ελληνική γλώσσα. Τα tweets αυτά διακρίθηκαν και χαρακτηρίστηκαν σε τρεις κατηγορίες, αληθή, άσχετα ή ψευδή. Αφού χαρακτηρίστηκε ένας επαρκής αριθμός από δεδομένα στη συνέχεια για κάθε κατηγορία οπτικοποιούνται οι πιο χαρακτηριστικές λέξεις σε νέφη λέξεων (Puds). Επιπλέον, εξάγεται από αυτά ένα σύνολο από γλωσσικά μορφολογικά χαρακτηριστικά, εφαρμόζοντας μεθόδους μετατροπής των κειμένων σε διανύσματα, καθώς και χαρακτηριστικά σχετικά με την υποκειμενικότητα των κειμένων. Επιπλέον χαρακτηριστικά υπολογίζονται με τη χρήση της μεθόδου TF-IDF τα οποία χρησιμοποιούνται σε συνδυασμό με τα μορφολογικά χαρακτηριστικά. Για τον υπολογισμό των χαρακτηριστικών αυτών αξιοποιούνται βιβλιοθήκες της Python όπως η NLTK, spaCy και Scikit-Learn. Πριν τα χαρακτηριστικά αυτά εισαχθούν στους αλγορίθμους μάθησης εφαρμόζεται PCA για τη μείωση των διαστάσεων. Εκπαιδεύονται τρεις αλγόριθμοι μάθησης, ο Random Forest, ο SVM και ο Multinomial Naive Bayes, εκ των οποίων ο Random Forest έχει τα πιο ενθαρρυντικά αποτελέσματα. Τα αποτελέσματά μας αποδεικνύουν ότι είναι εφικτή η αυτόματη ανίχνευση μη έγκυρης πληροφορίας σε δημοσιεύσεις στο Twitter παρά τις ιδιαιτερότητες που χαρακτηρίζουν την ελληνική γλώσσα.

Λέξεις Κλειδιά: *COVID-19, μέσα κοινωνικής δικτύωσης, παραπληροφόρηση, μη-έγκυρη πληροφορία, σύνολο δεδομένων, tweet, Μηχανική Μάθηση, NLP, ελληνική γλώσσα.*

Abstract

From the first moment of the spread of the COVID-19 coronavirus pandemic, another pandemic broke out at the same time, that of misinformation. Social media plays a key role in this, through which any information is rapidly transmitted around the world. The consequences of transmitting invalid information can be worse than the consequences of the pandemic itself. Conspiracy theories and false therapies are just two of the common categories that have unintended consequences for public health. On the other hand, the volume of information circulated on a daily basis makes checking the reliability of information a particularly demanding challenge.

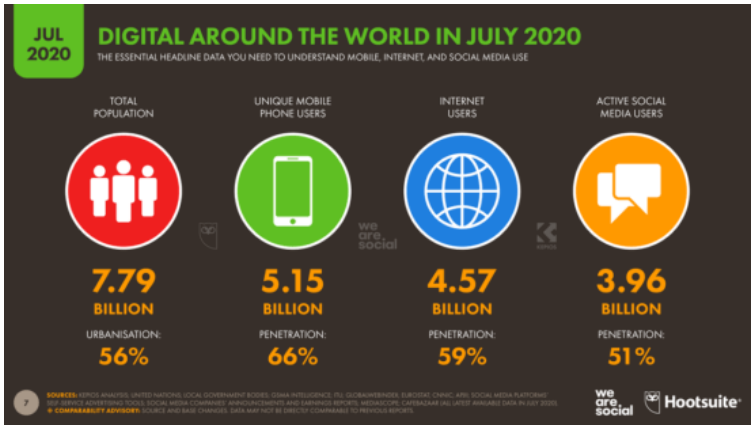
In this Thesis, the automatic detection of fake news related to the evolving coronavirus pandemic on social networks and specifically on Twitter is studied. For this purpose, algorithms of natural language processing (NLP) and Machine Learning are utilized. The data used to train the algorithms originates from a publicly accessible dataset that contains tweets related to the current pandemic. From the dataset, only the content concerning the Greek language was isolated. These tweets were classified and annotated in three categories, true, irrelevant, or false. Once a sufficient number of data has been annotated, the most common words are visualized through wordclouds for each category. In addition, a set of linguistic and morphological features are extracted from them by applying methods of converting texts into vectors, as well as features related to the subjectivity of the tweets' texts. Additional features are calculated using the TF-IDF method which are used in conjunction with the morphological features. Python libraries such as NLTK, spaCy and Scikit-Learn are used to calculate these features. Before these features are feeded into the learning algorithms, PCA is applied to reduce their dimensions. Three learning algorithms are trained, Random Forest, SVM and Multinomial Naive Bayes, of which Random Forest has the most encouraging results. Our results prove that it is possible to automatically detect invalid information in posts on Twitter despite the peculiarities that characterize the Greek language.

Keywords: *COVID-19, social media, misinformation, fake news, dataset, tweet, Machine Learning, NLP, Greek language.*

1

Εισαγωγή

Το διαδίκτυο και συγκεκριμένα τα κοινωνικά δίκτυα έχουν αδιαμφισβήτητα αλλάξει τον κόσμο. Από τα πρώτα μισά της δεκαετίας του 2000 που έκαναν την εμφάνισή τους οι πρώτες πλατφόρμες κοινωνικής δικτύωσης, η κοινωνία έχει μεταβληθεί σημαντικά καθώς η καθημερινότητα πλέον επηρεάζεται ιδιαίτερα από τη χρήση τους. Όπως φαίνεται στην Εικόνα 1, στα κοινωνικά δίκτυα συμμετέχει ενεργά το 59% του παγκόσμιου πληθυσμού, που αποτελεί ταυτόχρονα και το 87% του πληθυσμού που έχει πρόσβαση στο διαδίκτυο. Υπάρχει μεγάλο πλήθος από πλατφόρμες, κάθε μία προσανατολισμένη σε μία συγκεκριμένη διάσταση των αναγκών επικοινωνίας του σύγχρονου ανθρώπου. Επιπλέον, λόγω του ότι είναι εύκολο οποιοσδήποτε να αποκτήσει κάποιον λογαριασμό σε αυτά, όλο και περισσότεροι άνθρωποι τα επιλέγουν ως κύρια πηγή ενημέρωσης, αλλά και για να ενημερώσουν οι ίδιοι το κοινό με το οποίο αλληλοεπιδρούν.



Εικόνα 1. Τα ποσοστά χρήσης των μέσων κοινωνικής δικτύωσης σε παγκόσμια κλίμακα. Πηγή εικόνας: *[54]*

Καθώς το σύνολο των χρηστών αυτών αλληλοεπιδρούν μεταξύ τους, ένας τεράστιος όγκος πληροφορίας δημιουργείται και διακινείται ραγδαία. Η δραστηριότητα αυτή αποτελεί αντικείμενο μελέτης για διαφορετικούς επιστημονικούς κλάδους. Για παράδειγμα, η επιστήμη της πληροφορικής μελετά την αποδοτική διαχείριση, διακίνηση και αποθήκευση του τεράστιου όγκου της πληροφορίας αυτής. Η κοινωνιολογία μελετά τις κοινωνικές προεκτάσεις, τις τάσεις και τις συνδέσεις που δημιουργούνται στις διαδικτυακές κοινότητες. Πιο αναλυτικά, τα μέσα κοινωνικής δικτύωσης, μπορούν να μας βοηθήσουν να καταλάβουμε καλύτερα την διάδοση ιδεών και την επιρροή που έχουν οι χρήστες μεταξύ τους. Είναι σημαντικό να γνωρίζουμε πως διαχέεται η πληροφορία σε ένα δίκτυο καθώς έτσι μπορούμε να προβλέψουμε κοινωνικά φαινόμενα ή να αναγνωρίσουμε συμπεριφορές χρηστών που μας ενδιαφέρουν. Τα αποτελέσματα τέτοιων μελετών εφαρμόζονται με ποικίλους τρόπους όπως συστήματα συστάσεων, διαφήμιση κα.

1.1 Twitter

Μεταξύ των διαφορετικών πλατφορμών κοινωνικής δικτύωσης, ξεχωριστή θέση έχει το Twitter. Το Twitter είναι μία πλατφόρμα micro-blogging στην οποία μπορεί να κάνει λογαριασμό οποιοσδήποτε. Η πλατφόρμα αυτή ιδρύθηκε τον Μάρτιο του 2006 από τους Jack Dorsey, Noah Glass, Biz Stone και Evan Williams. Η αρχική ιδέα ήταν μία υπηρεσία μέσω της οποίας μικρές ομάδες ατόμων μπορούσαν να επικοινωνούν με γραπτά μηνύματα ονόματι 'twttr' το οποίο ως γνωστόν μεταβλήθηκε αργότερα σε "Twitter". Οι δημιουργοί επέλεξαν τον όρο 'tweet' για τα μηνύματα λόγο της φύσης τους ως μικρές εκρήξεις πληροφορίας και τα παρομοίασαν με τιτίβισμα πουλιών. Γνώρισε εκρηκτική ανάπτυξη και το 2008 έφτασε να δέχεται 100 εκατομμύρια δημοσιεύσεις ανά τρίμηνο. Η άνοδος της πλατφόρμας συνέχισε να προσελκύει ενεργούς χρήστες με ταχύτατους ρυθμούς, καταφέρνοντας να φτάσει τα 50 εκατομμύρια tweets ημερησίως τον Φεβρουάριο του 2010. Αυτή τη στιγμή αριθμεί 187 εκατομμύρια χρήστες παγκοσμίως [1] και αποτελεί ένα βασικό εργαλείο για την δημοσιογραφία καθώς τα νέα και οι τάσεις εξελίσσονται σε αυτό με ραγδαίους ρυθμούς. Οι παράγραφοι που ακολουθούν την πρώτη παράγραφο κάθε ενότητες έχουν μία εσοχή.

Το πρωταρχικό στοιχείο που το κάνει να ξεχωρίζει από τα υπόλοιπα κοινωνικά δίκτυα είναι ότι προσφέρει στους χρήστες του τη δυνατότητα να κάνουν μόνο δημοσιεύσεις μικρού μεγέθους στο προφίλ τους αλλά και το γεγονός ότι υπάρχει πρόσβαση πληροφορίας σε όλους. Κάθε χρήστης επιλέγει άλλους χρήστες που θέλει να «ακολουθεί», δηλαδή να παρακολουθεί τη δραστηριότητά τους στην πλατφόρμα. Ο καθένας έχει δικαίωμα να διαδώσει όσες προσωπικές πληροφορίες επιθυμεί στο εικονικό του προφίλ, ορισμένοι περισσότερο από άλλους, σε σημείο που μπορεί να μοιράζονται κάθε ενέργεια, που μπορεί να κάνουν στην καθημερινότητά τους, για τους ακολούθους τους. Αλλά ασχέτως με τις προσωπικές πληροφορίες, κάθε χρήστης με τη συμπεριφορά του, δηλαδή τις δημοσιεύσεις του, τις δημοσιεύσεις που του αρέσουν και τα σχόλιά του αντανακλά τη στάση του προς διάφορες τάσεις και συμβάντα.

Όπως αναφέρθηκε προηγουμένως το βασικό χαρακτηριστικό των δημοσιεύσεων στο Twitter είναι ότι περιορίζονται σε μήκος, το οποίο παλαιότερα περιοριζόταν σε 140 χαρακτήρες και πλέον σε 280. Επομένως, όταν ένας χρήστης θέλει να εκφραστεί, δεν μπορεί να πολυλογήσει, αλλά σε λίγες λέξεις πρέπει να συμπεριλάβει όση περισσότερη πληροφορία μπορεί. Για τη διευκόλυνσή τους οι χρήστες μπορούν να κάνουν χρήση ειδικών χαρακτήρων, όπως το 'at' (@) για την αναφορά

σε άλλους χρήστες, τη χρήση emoticons για την έκφραση συναισθήματος, καθώς και τη χρήση hashtag (#) με σκοπό να δηλώσουν ότι η δημοσίευσή τους σχετίζεται με ένα ευρύτερο θέμα. Ειδικά τα τελευταία είναι ιδιαίτερα σημαντικά καθώς μέσα από τη χρήση τους μπορεί κάποιος να παρακολουθήσει την εξέλιξη συμβάντων και τάσεων παρακολουθώντας τη δραστηριότητα των χρηστών οι οποίοι έχουν συμπεριλάβει στις δημοσιεύσεις τους το hashtag αυτό. Επομένως το Twitter αποτελεί μία τεράστια πηγή πληροφοριών που είναι διαθέσιμες σε πραγματικό χρόνο και που δημιουργούνται συνεχώς από χρήστες οποιουδήποτε κοινωνικού και μορφωτικού επιπέδου, για σχεδόν οποιοδήποτε θέμα. Για το λόγο αυτό, η συλλογή δεδομένων μπορεί να παρέχει σημαντική συμβολή σε σχεδόν όλους τους επιστημονικούς κλάδους και τους ερευνητικούς τομείς, καθώς συμβάλλουν στην εξαγωγή πολύτιμων πληροφοριών σχετικά με τις τελευταίες εξελίξεις σε πολλές ανθρώπινες δραστηριότητες και φυσικές διαδικασίες.

Για παράδειγμα, ένας από τους τομείς στους οποίους εφαρμόζονται συχνά προβλέψεις βάσει κοινωνικών δικτύων είναι οι επιδημίες. Οι ερευνητές στην εργασία [2], ήταν από τους πρώτους που χρησιμοποίησαν το Twitter για να προβλέψουν την εξέλιξη της πανδημίας της γρίπης των χοίρων. Συγκεκριμένα, ανέλυσαν πενήντα εκατομμύρια tweets χρησιμοποιώντας μεθόδους Τεχνητής Νοημοσύνης (AI). Τα αποτελέσματα ήταν ιδιαίτερα ενθαρρυντικά και επιβεβαιώθηκαν και από άλλους ερευνητές. Επιτυχημένοι αλγόριθμοι παρακολούθησης της πανδημίας της γρίπης έχουν επίσης τεκμηριωθεί και σε άλλα ερευνητικά έργα [3].

Σε οικονομικό επίπεδο, η ανάλυση κοινωνικών δικτύων προς υποστήριξη των προβλέψεων στις χρηματιστηριακές αγορές έχει επίσης προσελκύσει την προσοχή πολλών ερευνητών. Στην εργασία [4] έγινε ανάλυση των χρηματιστηριακών αγορών εφαρμόζοντας τεχνικές Twitter Sentiment Analysis. Οι συγγραφείς συνέλεξαν και επεξεργάστηκαν τέσσερα εκατομμύρια tweets, εντοπίζοντας μια ισχυρή συσχέτιση μεταξύ των τιμών των μετοχών και των αποτελεσμάτων της επεξεργασίας των δεδομένων του Twitter.

Όσον αφορά τις πολιτικές μελέτες, στην εργασία [5] περιγράφεται μια από τις πρώτες προσπάθειες χρήσης του Twitter για την εκτίμηση των εκλογικών αποτελεσμάτων των γερμανικών ομοσπονδιακών κρατιδίων. Η έρευνά έδειξε ότι τα tweets τα οποία έδειχναν θετική στάση για τα κύρια κόμματα ήταν παρόμοια με την αντίστοιχη κατανομή ψήφων στις εκλογές.

1.2 Διάδοση ψευδών ειδήσεων

Δυστυχώς, μαζί με τα ισχυρά πλεονεκτήματά του, η νέα γενιά του διαδικτύου έχει εξίσου ισχυρά μειονεκτήματα. Ένας απλός χρήστης, ακούσια ή εκούσια, μπορεί να δημιουργήσει μια λανθασμένη πληροφορία χωρίς κανένα έλεγχο και αυτή να διαδοθεί ραγδαία με απρόβλεπτες συνέπειες. Η παραπληροφόρηση μπορεί να αποτελεί ένα απλό ψέμα για την προσωπική ζωή κάποιου γνωστού προσώπου, μέχρι φήμες οι οποίες υποκρύπτουν συγκεκριμένη σκοπιμότητα. Πολλοί είναι αυτοί που εκμεταλλεύονται αυτή την αμεσότητα και ευχρηστία που προσφέρουν τα κοινωνικά δίκτυα στη διάδοση ειδήσεων, ώστε να διαδώσουν παραπλανητικές και ψευδείς ειδήσεις προ όφελός τους.

Οι ψευδείς ειδήσεις περιλαμβάνουν κάθε είδος κίτρινου τύπου, παραπληροφόρησης, συκοφαντίας, προκατάληψης, ρητορικής μίσους. Συνήθως τα κίνητρα για τη διάδοση ψευδών ειδήσεων διακρίνονται σε δύο κατηγορίες, τα οικονομικά και τα ιδεολογικά. Οικονομικά λόγω του ότι μια διαδεδομένη ψευδής είδηση θα φέρει στους παρόχους του περιεχομένου κέρδη μέσω

διαφημίσεων ή θα προωθήσει τους αναγνώστες για αγορά προϊόντων, και ιδεολογικά λόγω του ότι συχνά προωθούνται οντότητες και ιδεολογίες.

Οι άνθρωποι που διαδίδουν ψευδείς ειδήσεις εκμεταλλεύονται την απουσία ελέγχου του περιεχομένου που διαδίδεται στα κοινωνικά δίκτυα, διαδίδοντας τάχιστα ακόμα και μη έγκυρο περιεχόμενο. Λόγω του τεράστιου όγκου πληροφορίας που διακινείται, η δυνατότητα ελέγχου των δημοσιεύσεων από εξειδικευμένο και αξιόπιστο δημοσιογραφικό προσωπικό θα ήταν απίστευτα χρονοβόρα και ασύμφορη οικονομικά. Εκτός από την απουσία ελέγχου, το φαινόμενο διογκώνεται και από τους απλούς χρήστες της πλατφόρμας. Αρκετοί χρήστες όταν αναδημοσιεύουν μια είδηση (retweet) μπορεί να μην επιβεβαιώσουν τις πηγές της ή να μην διαβάσουν το άρθρο παρά μόνο τον τίτλο και έτσι άθελά τους να συμβάλλουν στη διάδοση των ψευδών αυτών δημοσιεύσεων, όπως συμβαίνει και στην πραγματική ζωή με τη διάδοση των φημών. Επίσης, μπορεί να υπάρχουν χρήστες που να μοιράζονται συγκεκριμένα ψευδείς ειδήσεις ή να είναι αυτοματοποιημένοι λογαριασμοί, δηλαδή bots, που είναι προγραμματισμένα στη διάδοση ειδήσεων από συγκεκριμένους λογαριασμούς. Έχουν γίνει πολλές προσπάθειες να εντοπιστούν τέτοιοι λογαριασμοί καθώς και υπάρχουν αρκετά εργαλεία που εξάγουν κάποια ένδειξη για το αν ένας χρήστης είναι bot ή όχι.

Εδώ πρέπει να σημειωθεί ότι η διασπορά ψευδών ειδήσεων είναι ένα διαχρονικό πρόβλημα με ποικίλες προεκτάσεις. Η μεγάλη διαφορά στη διάδοση των ψευδών ειδήσεων σήμερα σε σχέση με το παρελθόν είναι ότι πλέον μπορούμε να μελετήσουμε τη φύση των ψευδών ειδήσεων μέσα από τα μέσα κοινωνικής δικτύωσης. Ένα βασικό παράδειγμα που αξίζει να μελετηθεί προέκυψε από το ξέσπασμα της πανδημίας του ιού της νόσου του κορονοϊού 2019.

Η πανδημία της νόσου του κορονοϊού 2019 (COVID-19) είναι μια τρέχουσα πανδημία που προκλήθηκε από τον κορονοϊό SARS-CoV-2 και αναγνωρίστηκε για πρώτη φορά στην πόλη Ουχάν, πρωτεύουσα της επαρχίας Χουπέι της Κίνας, τον Δεκέμβριο του 2019. Ως και τις 22 Νοεμβρίου 2020 είχαν επιβεβαιωθεί πάνω από 58,5 εκατομμύρια κρούσματα σε 215 χώρες και περιοχές, είχαν σημειωθεί περισσότεροι από 1,38 εκατομμύρια θάνατοι που οφείλονται στη νόσο και είχαν ανακάμψει περισσότεροι από 40,2 εκατομμύρια άνθρωποι [6]. Εξαιτίας του υψηλού βαθμού μεταδοτικότητας του ιού πολλές κυβερνήσεις στον κόσμο έλαβαν μέτρα κοινωνικής αποστασιοποίησης με σκοπό τον περιορισμό της εξάπλωσής του ελαχιστοποιώντας την στενή επαφή μεταξύ των ανθρώπων. Τα μέτρα αυτά περιλαμβάνουν καραντίνα. ταξιδιωτικούς περιορισμούς · και το κλείσιμο σχολείων, χώρων εργασίας, σταδίων, θεάτρων, ή εμπορικών κέντρων και μπορεί να γίνουν πιο έντονα σε περιοχές που ο ιός βρίσκεται σε έξαρση. Η πανδημία όπως είναι λοιπόν αντιληπτό έχει επηρεάσει έντονα την καθημερινότητα των πολιτών σε όλο τον κόσμο, με εμφανείς συνέπειες σε πολλούς τομείς όπως στην πολιτική, στην οικονομία κα.

Οι συνέπειες αυτές αντανακλώνται πλήρως και στα κοινωνικά δίκτυα, και ιδιαίτερα στο Twitter, όπου το hashtag #covid19 αποτέλεσε το πιο χρησιμοποιούμενο hashtag για το 2020 [7]. Με την εμφάνιση ωστόσο της πανδημίας COVID-19, η πολιτική και ιατρική παραπληροφόρηση έχει αυξηθεί ραγδαία, δημιουργώντας συνέπειες που μπορούν στην πραγματικότητα να επιδεινώσουν τη διάδοση της ίδιας της επιδημίας. Θεωρίες συνωμοσίας, ψευδοεπιστημονικές θεραπείες και αγωγές είναι μόνο δύο ενδεικτικές κατηγορίες παραπληροφόρησης και ψευδών ειδήσεων που έχουν βρει πρόσφορο έδαφος λόγω της εξελισσόμενης κατάστασης. Η πανδημία αυτή παραπληροφόρησης δεν θα μπορούσε να αφήσει ανεπηρέαστη και τη χώρα μας.

Δεδομένων των κινδύνων που επιφυλάσσει η διάδοση ψευδών ειδήσεων, αποτελεί αναγκαιότητα η έγκαιρη αντιμετώπιση του φαινομένου. Έχοντας γνώση ωστόσο του πώς διαχέεται η πληροφορία σε ένα κοινωνικό δίκτυο, όπως είναι το Twitter, μπορούν να γίνουν αντιληπτά μοτίβα και ενδεχομένως κακόβουλες δραστηριότητες, οι οποίες αποσκοπούν στην παραπλάνηση των χρηστών. Η επιστήμη υπολογιστών και η Μηχανική Μάθηση, αποδεικνύεται ότι είναι κατάλληλες για τον σκοπό αυτό. Αξιοποιώντας μεθόδους Μηχανικής Μάθησης στην προσπάθεια ανίχνευσης ψεύτικων ειδήσεων, η διαδικασία μπορεί να αυτοματοποιηθεί, μειώνοντας τον χρόνο που απαιτείται στην προσπάθεια ελέγχου της πληροφορίας για τον εντοπισμό ψεύτικων ειδήσεων αλλά επιπλέον βοηθά και στην ανάσχεση της διάδοσής τους.

1.3 Στόχος της εργασίας

Ο σκοπός της εργασίας αυτής είναι να διερευνήσει την αυτοματοποιημένη κατηγοριοποίηση των Tweets που δημοσιεύονται στην πλατφόρμα του Twitter, τα οποία είναι γραμμένα στην ελληνική γλώσσα και έχουν ως θεματικό περιεχόμενο την πανδημία του COVID-19 και την εξέλιξη της. Η κατηγοριοποίηση στηρίζεται στον υπολογισμό ενός συνόλου μορφολογικών, σημασιολογικών (semantic), μερών του λόγου - PoS (Part of Speech) και στατιστικών χαρακτηριστικών, τα οποία προκύπτουν εφαρμόζοντας προχωρημένες τεχνικές Επεξεργασίας Φυσικής Γλώσσας - NLP (Natural Language Processing) στο κείμενο των tweets. Οι κατηγορίες στις οποίες είναι επιθυμητό για τα tweets να ταξινομηθούν προκύπτουν με βάση την εγκυρότητα του περιεχομένου τους δηλαδή αν περιέχουν αληθή, ψευδή ή μη-σχετική πληροφορία σε σχέση με την πανδημία. Η αυτόματη ταξινόμηση των tweets με βάση τα χαρακτηριστικά αυτά διερευνάται μέσω της εφαρμογής αλγορίθμων Μηχανικής Μάθησης.

1.4 Δομή της διπλωματικής

Στο Κεφάλαιο 2 παρουσιάζονται διάφορες εργασίες των οποίων το περιεχόμενο είναι συναφές με αυτό της Διπλωματικής Εργασίας αυτής. Στο Κεφάλαιο 3 αναφέρονται οι Προδιαγραφές και Απαιτήσεις του συστήματος που υλοποιείται, εξηγώντας συνήθεις τεχνικές Επεξεργασίας Φυσικής Γλώσσας (NLP) αλλά και τα εργαλεία και οι τεχνολογίες που χρησιμοποιήθηκαν. Στο Κεφάλαιο 4 αναλύεται η Αρχιτεκτονική του συστήματος και οι μεθοδολογίες που ακολουθήθηκαν ως προς την υλοποίησή του. Τα αποτελέσματα της έρευνας αλλά και τα συμπεράσματα που εξήχθησαν μέσω αυτής παρουσιάζονται στο Κεφάλαιο 5. Τέλος, στο τελευταίο κεφάλαιο που αποτελεί και τον επίλογο της συγκεκριμένης εργασίας γίνεται σύνοψη των όσων υλοποιήθηκαν και γίνεται αναφορά στις δυνατότητες μελλοντικής επέκτασης της έρευνας αυτής.

2

Σχετικές Εργασίες

2.1 NLP και ανίχνευση ψευδών ειδήσεων

Πολλές είναι έρευνες που έχουν πραγματοποιηθεί οι οποίες εστιάζουν στον τομέα της αυτοματοποίησης της ανίχνευσης ψευδών. Οι τεχνικές δυσκολίες που παρουσιάζονται στον τομέα αυτό μέσα από υλοποιήσεις με NLP, αλλά και συγκρίσεις μεταξύ των συνόλων δεδομένων (dataset) και των επιδόσεων των μοντέλων που χρησιμοποιούνται παρουσιάζονται στις εργασίες [8] και [9]. Στην εργασία [10] οι συγγραφείς προσεγγίζουν την ανίχνευση ψευδών ειδήσεων στην πλατφόρμα του Twitter με σκοπό να ορίσουν τα σημαντικότερα χαρακτηριστικά τα οποία ενδεχομένως να υποδεικνύουν μία ψευδή είδηση. Έχουν προταθεί αρκετές προηγούμενες προσπάθειες για την αξιολόγηση της αξιοπιστίας ενός συγκεκριμένου tweet [11] ή ενός χρήστη [12], ενώ άλλοι έχουν επικεντρωθεί περισσότερο στη πρόσκαιρη δυναμική διάδοσης φήμης [13].

Μία έρευνα η οποία είναι σχετική με το θέμα της συγκεκριμένης Διπλωματικής Εργασίας περιγράφεται ακολούθως. Η έρευνα [14] αφορά το πρόβλημα της αυτόματης διακριτοποίησης μεταξύ έγκυρων και μη-έγκυρων ειδήσεων οι οποίες πηγάζουν από δημοσιεύσεις που γίνονται στην πλατφόρμα του Twitter και αφορούν τις πρόσφατες διαδηλώσεις στην περιοχή του Hong Kong. Για την κατηγοριοποίηση των δημοσιεύσεων αυτών, χρησιμοποιήθηκαν διάφοροι αλγόριθμοι Μηχανικής Μάθησης (Machine Learning- ML). Στην συγκεκριμένη εργασία χρησιμοποιούνται Supervised Learning τεχνικές και αλγόριθμοι για την κατηγοριοποίηση των tweets. Τέτοιοι αλγόριθμοι απαιτούν η τροφοδότησή τους να είναι με δεδομένα τα οποία έχουν χαρακτηριστεί ότι ανήκουν σε κάποια κατηγορία και να εκπαιδευτούν πάνω σε αυτά, ώστε να αποκτήσουν την ικανότητα να κατηγοριοποιήσουν νέα tweets. Τα δεδομένα τα οποία περιείχαν ψευδείς ειδήσεις συγκεντρώθηκαν από διαπιστωμένα σύνολα δεδομένων τα οποία είχαν χαρακτηριστεί κατ' αυτόν τον τρόπο ενώ τα δεδομένα που περιείχαν αληθείς ειδήσεις δημιουργήθηκαν από τους συγγραφείς συγκεντρώνοντας tweets από φημισμένα πρακτορεία ειδήσεων και δημοσιογράφους.

.

2.2 Σύνολα δεδομένων σχετικά με τον COVID-19 και κατηγοριοποίηση των δεδομένων τους

Ένα μεγάλο ζήτημα σε έρευνες και υλοποιήσεις κατηγοριοποίησης κειμένου, όπως το ζήτημα που επεξεργάζεται στη συγκεκριμένη Διπλωματική Εργασία, είναι η σωστή ετικετοποίηση (labeling) των δεδομένων ώστε να τροφοδοτηθούν με ακρίβεια τα μοντέλα Μηχανικής Μάθησης με συνέπεια να επιτύχουν υψηλά ποσοστά επιτυχίας κατά την κατηγοριοποίησή νέων κειμένων.

Είναι ιδιαίτερα σπάνια η ύπαρξη συνόλων δεδομένων των οποίων τα δεδομένα είναι ήδη κατηγοριοποιημένα. Ειδικά για το θέμα της πρόσφατης πανδημίας, έχει δημοσιευτεί ένα πλήθος από σύνολα δεδομένων [15], [16], [17] των οποίων τα δεδομένα δεν είναι κατηγοριοποιημένα ή δεν είναι προσανατολισμένα στην ανίχνευση της παραπληροφόρησης. Αν και υπάρχουν σύνολα δεδομένων όπως το [18], τα οποία χαρακτηρίζουν τα δεδομένα ως δεδομένα παραπληροφόρησης, σύνολα δεδομένων που περιέχουν "αληθινές" ειδήσεις για τον COVID-19 είναι επίσης σπάνια.

Ιδιαίτερα σχετική με το ζήτημα αυτό είναι η έρευνα [15] είναι. Ο κύριος στόχος της [15] έχει να κάνει με τον χαρακτηρισμό των κοινοτήτων παραπληροφόρησης με θέμα τον Covid-19 μέσω δεδομένων που συγκεντρώνονται από την διαδικτυακή πλατφόρμα του Twitter. Για να επιτευχθεί όμως αυτό, πρέπει να αντληθούν tweets τα οποία ήδη έχουν χαρακτηριστεί και κατηγοριοποιηθεί. Μία τέτοια διαδικασία, δηλαδή μία διαδικασία δημιουργίας ενός συνόλου δεδομένων με κατηγοριοποιημένα ανάλογα με το περιεχόμενό τους tweets που αφορούν τον COVID-19 περιγράφεται αναλυτικά.

Οι συγγραφείς χρησιμοποιούν το Twitter API ώστε να αντλήσουν tweets που είναι σχετικά με τον COVID-19 χρησιμοποιώντας μία ομάδα από ποικίλες λέξεις κλειδιά και hashtags τα οποία έπειτα από πολλές συζητήσεις και αναθεωρήσεις εκ των συγγραφέων, διαχωρίστηκαν σε 17 κατηγορίες ανάλογα με το περιεχόμενο τους. Συνεπώς δημιουργήθηκε ένα σύνολο δεδομένων το οποίο, μπορεί να συμβάλλει σημαντικά στην δημιουργία μοντέλων με μεγαλύτερη επιτυχία στην κατηγοριοποίηση που πραγματοποιούν.

2.3 Συνεισφορά εργασίας

Η συνεισφορά της Διπλωματικής Εργασίας αυτής έγκειται στο γεγονός ότι για πρώτη φορά, από όσο γνωρίζουμε, πραγματοποιείται μια προσπάθεια ανίχνευσης ψευδών ειδήσεων και αυτόματης κατηγοριοποίησης ειδήσεων που πηγάζουν από tweets γραμμένα στα ελληνικά που αφορούν την πανδημία του COVID-19. Η γλώσσα μας καθιστά μία τέτοια ανάλυση ως ένα δύσκολο έργο, για αυτό και συναφείς εργασίες συναντιούνται αρκετά δύσκολα. Επίσης, για να επιταχυνθεί η έρευνα, κατασκευάστηκε ένα διαδικτυακό εργαλείο το οποίο δίνει την δυνατότητα μαζικής κατηγοριοποίησης των tweets. Το εργαλείο αυτό κάνει ευκολότερη και καλύτερα διαχειρίσιμη την διαδικασία της ετικετοποίησης των tweets για κάποιον εθελοντή και επίσης επιφέρει και την επιτάχυνση του ρυθμού πραγματοποίησης της διαδικασίας. Επιπλέον, στα χαρακτηριστικά μορφολογικού και σημασιολογικού χαρακτήρα που εξάγονται γίνεται πρόσθεση χαρακτηριστικών τα οποία προκύπτουν από την εφαρμογή της μεθόδου TF-IDF στο σύνολο δεδομένων. Η προσθήκη αυτή προσδίδει νέα πληροφορία που άλλες υλοποιήσεις δεν λαμβάνουν υπόψη.

3

Προδιαγραφές και Απαιτήσεις Συστήματος

3.1 Λήψη δεδομένων

Το πρώτο βήμα για τη χρήση δεδομένων από το Twitter είναι η δημιουργία ή λήψη ενός συνόλου δεδομένων. Η πλατφόρμα παρέχει τη δυνατότητα πρόσβασης σε περιεχόμενο που δημιουργείται σε πραγματικό χρόνο μέσα από το Twitter Streaming API [19]. Ωστόσο δεν είναι απαραίτητο να δημιουργήσει κάποιος ένα νέο σύνολο δεδομένων καθώς υπάρχουν ανοιχτά διαθέσιμα ήδη συλλεχθέντα σύνολα δεδομένων τα οποία εξυπηρετούν διαφορετικούς σκοπούς το καθένα. Πρέπει να σημειωθεί όμως ότι στα δεδομένα που περιέχονται στα σύνολα δεδομένων δεν υπάρχουν τα περιεχόμενα των tweet, γεωγραφικές τοποθεσίες, χρόνος, εικόνες και άλλες συνημμένες πληροφορίες που συνήθως συμπεριλαμβάνονται σε tweets. Το σύνολο των δεδομένων είναι ένα αρχείο απλού κειμένου που αποτελείται από μια λίστα μοναδικών αναγνωριστικών tweet. Ο βασικός λόγος που συμβαίνει αυτό είναι το μεγάλο μέγεθος των συσχετισμένων δεδομένων των tweets. Ένα αρχείο που περιέχει μόνο μια σειρά αριθμών (ID) είναι πολύ πιο εύκολα διαχειρίσιμο από, για παράδειγμα, ένα αρχείο CSV με εκατομμύρια tweets με τα μεταδεδομένα τους.

Επομένως, για την ανάκτηση του πλήρους περιεχομένου που αντιστοιχεί σε αυτά απαιτείται μία διαδικασία γνωστή ως “hydration” [20]. Η διαδικασία του hydration απαιτεί τη σύνδεση με το Twitter API μέσω του οποίου δίνεται η δυνατότητα αξιοποίησης των πλήρων δεδομένων. Για να γίνει ωστόσο κάτι τέτοιο απαιτείται η εξουσιοδότηση από την πλατφόρμα του Twitter, μέσω της απόκτησης Open Authorization Credentials (OAuth) [21].

Για την λήψη των συνόλων δεδομένων είτε απευθείας από το Twitter Streaming API ή από κάποιο αποθετήριο ανοιχτών δεδομένων, όπως το Zenodo [22] ή το Github [23], ιδιαίτερα χρήσιμη είναι η ανάπτυξη ενός εργαλείου Crawler. Το εργαλείο αυτό θα μπορεί να συλλέξει αρχικά μόνο τα δεδομένα που μας ενδιαφέρουν από το web. Στη συνέχεια, θα προχωρά στο hydration των συλλεχθέντων tweets και στην αποθήκευσή τους σε μία τοπική βάση με τρόπο που να τα καθιστά εύκολα προσβάσιμα και διαχειρίσιμα.

3.2 Διερευνητική ανάλυση

Η διερευνητική ανάλυση χρησιμοποιείται για να περιγράψει τα βασικά χακακτηριστικά των συλλεχθέντων δεδομένων σε μία μελέτη. Μέσα από αφαιρετικές αναπαραστάσεις μπορεί να οδηγήσει σε ενδιαφέροντα συμπεράσματα ακόμα και από απλά στατιστικά. Η εξαγωγή τέτοιων στατιστικών μπορεί να περιλαμβάνει την εύρεση των πιο συχνών hashtags, τις λέξεις που εμφανίζονται πιο συχνά ή το μήκος των tweets σε χαρακτήρες. Η οπτική απεικόνιση των χαρακτηριστικών αυτών σε κατάλληλα γραφήματα ανάλογα με την κατηγορία από την οποία προέρχονται ,μπορεί να παρέχει μία πρώτη εικόνα για τα κύρια χαρακτηριστικά κάθε κατηγορίας. Για την οπτικοποίηση ιδιαίτερα χρήσιμες είναι οι βιβλιοθήκες της Python, seaborn και matplotlib.

Μία τέτοια μέθοδος οπτικοποίησης είναι το νέφος λέξεων. Συγκεκριμένα, αποτελεί οπτική αναπαράσταση δεδομένων κειμένου, που χρησιμοποιείται συνήθως για την εμφάνιση μεταδεδομένων λέξεων-κλειδιών (ετικέτες) σε ιστότοπους ή για την απεικόνιση κειμένου ελεύθερης μορφής. Οι ετικέτες είναι μεμονωμένες λέξεις που απεικονίζονται στο νέφος λέξεων χρησιμοποιώντας διαφορετικά μεγέθη γραμματοσειρών και / ή χρώματα ανάλογα με τη σημαντικότητά τους στα δεδομένα από τα οποία προέρχονται. Η συγκεκριμένη απεικόνιση είναι χρήσιμη καθώς επιτρέπει τον γρήγορο προσδιορισμό των πιο σημαντικών όρων στα διαθέσιμα δεδομένα και ταυτόχρονα τη σύγκριση της σημαντικότητας διαφορετικών όρων μεταξύ τους. Επιπλέον, μπορούν να επισημανθούν σημαντικά στοιχεία κειμένου. Τα σύννεφα λέξεων χρησιμοποιούνται επίσης για την ανάλυση δεδομένων από κοινωνικά δίκτυα. Επιλέγεται η χρήση τους για την παρουσίαση των δεδομένων καθώς αφενός μεταφέρουν πληροφορίες με απλότητα και σαφήνεια και επιπλέον είναι οπτικά πιο ελκυστικά από ένα απλό πίνακα δεδομένων. Η ομώνυμη βιβλιοθήκη της Python WordCloud [24] επιτρέπει την εύκολη και γρήγορη εξαγωγή τέτοιων απεικονίσεων.

3.3 Μηχανική χαρακτηριστικών (feature engineering)

Η συλλογή απλών κειμένων όπως οι δημοσιεύσεις των χρηστών δεν έχει κάποιο νόημα αν αυτά τα δεδομένα δεν υποστούν πρώτα κάποια επεξεργασία. Τα δεδομένα που έχουν την μορφή κειμένου δεν αποτελούν πληροφορία επεξεργάσιμη από τα μαθηματικά μοντέλα που χρησιμοποιούν οι αλγόριθμοι της Μηχανικής Μάθησης. Ως αποτέλεσμα, υπάρχει η ανάγκη μετατροπής των κειμένων σε μορφή, με την οποία να είναι εφικτή η επεξεργασία τους από τους αλγόριθμους αυτούς, αλλά παράλληλα είναι επιθυμητό και πολλές φορές αναγκαίο, να διατηρείται η χρήσιμη πληροφορία που παρέχουν. Χρησιμοποιώντας την γλώσσα προγραμματισμού Python, η οποία έχει εξαιρετικές δυνατότητες επεξεργασίας συμβολοσειρών και μεγάλη ευκολία στη χρήση, είναι δυνατό να διευκολυνθεί η ανάπτυξη οποιασδήποτε διαδικασίας επεξεργασίας δεδομένων. Στη συνέχεια, περιγράφονται οι βασικές μέθοδοι επεξεργασίας που μπορούν να εφαρμοστούν στα δεδομένα καθώς και οι αντίστοιχες βιβλιοθήκες.

Αρχικά στα δεδομένα εφαρμόζονται μετασχηματισμοί οι οποίοι σχετίζονται με την αντιμετώπιση φράσεων, συμβόλων κ.λπ., προσεγγίσεις στατιστικής ανάλυσης για τα tweets και επιπλέον εφαρμόζονται διάφορες κανονικές εκφράσεις (Regular Expressions) στο αρχικό κείμενο με σκοπό την εύρεση ή κατάργηση συμβόλων και συμβολοσειρών. Το tokenisation [25] είναι

επίσης ένα βασικό βήμα της διαδικασίας κατά την οποία μια ροή κειμένου χωρίζεται σε μεμονωμένες οντότητες που ονομάζονται tokens. Στην απλούστερη μορφή τους, αυτές οι οντότητες είναι λέξεις, αλλά είναι επίσης εφικτή η δημιουργία tokens τα οποία αποτελούνται από φράσεις, σύμβολα κ.λπ.

Πολύ συχνά χρησιμοποιούμενες μέθοδοι προεπεξεργασίας, αποτελούν η stemming [26] και η lemmatization [26]. Η πρώτη είναι μέθοδος η οποία μέσω διαφόρων εξειδικευμένων συναρτήσεων οι καταλήξεις των λέξεων κόβονται σύμφωνα με κάποιους προκαθορισμένους κανόνες, ώστε να διατηρηθεί μόνο το θέμα (stem) κάθε λέξης. Για παράδειγμα αν ο κανόνας είναι "ες" τότε η λέξη "βάρυνες" μετατρέπεται σε "βαρυν". Το lemmatization από την άλλη μετατρέπει τη λέξη στο λήμμα από το οποίο προέρχεται (για παράδειγμα η λέξη "βάρυνες" μετατρέπεται σε "βαραίνω").

Μια άλλη χρήσιμη λειτουργία προεπεξεργασίας είναι η αφαίρεση λέξεων-κλειδιών (stopwords), δηλαδή λέξεων οι οποίες είναι πολύ κοινές σε μία γλώσσα και οποίες δεν συνεισφέρουν νοηματικά στο κείμενο όταν λαμβάνονται μεμονωμένα, όπως είναι για παράδειγμα τα άρθρα. Δεδομένης της παρουσίας και επανάληψης τέτοιων λέξεων σε όλα τα κείμενα, η αφαίρεση αυτών παρουσιάζει αρκετά πλεονεκτήματα στην εφαρμογή αλγορίθμων Μηχανικής Μάθησης. Κάποια από τα πλεονεκτήματα αφαίρεσης των λέξεων-κλειδιών, είναι η μείωση των διαστάσεων του συνόλου των δεδομένων εκμάθησης του αλγορίθμου, που έχει ως συνέπεια την μείωση της απαιτούμενης υπολογιστικής ισχύος που απαιτείται κατά την εκτέλεση του.

Μια σύνηθες επίσης πρακτική, είναι η αφαίρεση όλων των μη αλφαβητικών χαρακτήρων από το εκάστοτε κείμενο. Σε αυτό το στάδιο, αφαιρούνται τα σημεία στίξης και άλλοι χαρακτήρες που ενδεχομένως να υπάρχουν. Οι χαρακτήρες αυτοί, δεν αποτελούν λεξικολογικό χαρακτηριστικό στο κείμενο, και ενδέχεται να παρουσιάσουν σφάλματα στις συγκρίσεις μεταξύ των λέξεων.

Εξίσου σημαντική διαδικασία είναι η μετατροπή των κειμένων σε μη κεφαλαία γράμματα, ή το αντίθετο. Το βήμα αυτό, έχει ως σκοπό την ομοιομορφία του κειμένου, με αποτέλεσμα η ψηφιοποίηση και σύγκριση δύο όμοιων λέξεων να φέρει το επιθυμητό αποτέλεσμα. Πέραν της αποφυγής του σφάλματος σύγκρισης, το βήμα αυτό συμβάλλει και στην μείωση των διαστάσεων, αφού το συνολικό λεξιλόγιο της συλλογής δεδομένων, το οποίο αποτελείται από το σύνολο των διαφορετικών λέξεων που παρουσιάζεται στα δεδομένα, δεν θα εμφανίζει επανάληψη λέξεων.

Ακόμα χρησιμοποιούνται διάφορα διάφορες κανονικές εκφράσεις (regex) για την αναγνώριση emoticon, HTML tags, διευθύνσεων URL, αριθμών και άλλων χαρακτήρων, καθώς και στοχευμένων λέξεων. Η στατιστική ανάλυση αυτών των στοιχείων μπορεί να προσδώσει επιπλέον γνώση στο μοντέλο που θα χρησιμοποιηθεί για μάθηση.

Ένα κείμενο μπορεί να αποδώσει ένα μεγάλο αριθμό από χαρακτηριστικά. Αυτά που επιλέχθηκαν για να χαρακτηριστούν τα Tweets παρουσιάζονται στον Πίνακα 1.

	Χαρακτηριστικό	**Συντόμευση στο DataFrame**
1	Tweet length	tweet_length
2	Number of tokens	n_tokens
3	Type to Token Ratio (TTR)	TTR
4	Number of URLs	n_URLs
5	Number of hashtags	n_hashtags
6	Number of punctuation marks	n_punctuations
7	Number of “?!”	n_?!
8	Number of “?”	n_?
9	Number of “!”	n_!
10	Number of emojis	n_emojis
11	Negative emoji sentiment sum	emojis_sent_neg
12	Neutral emoji sentiment sum	emojis_sent_neu
13	Positive emoji sentiment sum	emojis_sent_pos
14	Number of periods	n_periods
15	Number of stopwords	n_stopwords
16	Number of words	n_words
17	Average number of chars per word	avg_chars_per_word
18	Average number of chars per sentence	avg_chars_per_sentene
19a	Overall text sentiment (polarity)	sentiment_polarity
19b	Overall text sentiment (subjectivity)	sentiment_subjectivity

20	Number of vowels	n_vowels
21	Number of consonants	n_consonants
22	Number of uppercase chars	n_uppercase
23	Number of lowercase chars	n_lowercase
24	Longest sequence of consecutive vowels	max_seq_vowels
25	Longest sequence of consecutive consonants	max_seq_consonants
26	Repetition of > 3 identical consecutive chars	n_identical_repetitions
27	Number of digits	n_digits
28	Number of letters	n_letters
29	Ratio of letters to digits	ratio_letters_digits
30	Number of pronouns	Pronouns
31	Number of determiners	Determiners
32	Number of nouns	Nouns
33	Number of adverbs	Adverbs
34	Number of verbs	Verb
35	Number of adjectives	n_Adjectives
36	Number of entities	n_entities
37	Tweet Entropy	text_entropy
38	TF-IDF	

Πίνακας 1. Χαρακτηριστικά που εξάγονται από το κείμενο των tweets.

3.3.1 Μορφολογικά χαρακτηριστικά

Μία πληθώρα από χαρακτηριστικά που επιλέχθηκαν έχουν να κάνουν με το μορφολογικό επίπεδο των κειμένων. Αναλυτικότερα, αυτά περιλαμβάνουν χαρακτηριστικά που έχουν να κάνουν με το μήκος (σε χαρακτήρες) του tweet καθώς επίσης τον αριθμό: των tokens, των URL, των hashtag, των διαφόρων σημείων στίξης (“?”, ”!”, ”?!”) αλλά και το συνολικό πλήθος τους, των emojis, των τελειών, των λέξεων-κλειδιών, των λέξεων, των φωνηέντων, των συμφώνων, των κεφαλαίων και των πεζών γραμμάτων, των ψηφίων και των γραμμάτων, τα οποία περιέχονται σε ένα tweet. Επίσης, χαρακτηριστικό αποτελεί η μέση τιμή του αριθμού των χαρακτήρων ανά λέξη αλλά και η μέση τιμή του αριθμού των λέξεων ανά τελεία και το εάν ένας χαρακτήρας επαναλαμβάνεται συνεχόμενα πάνω από τρεις φορές. Η αναλογία του αριθμού των γραμμάτων που περιέχονται σε ένα tweet ως προς τον αριθμό των ψηφίων αποτελεί ένα ακόμα χαρακτηριστικό.

Το TTR (Type-Token Ratio) είναι μία μετρική που δείχνει πόσο πλούσιο σε λεξιλόγιο είναι ένα κείμενο. Ουσιαστικά αυτή η μετρική υπολογίζεται από τον αριθμό των διαφορετικών μοναδικών λέξεων που υπάρχουν σε ένα κείμενο (για παράδειγμα εάν μία λέξη σε ένα κείμενο επαναλαμβάνεται θα μετρήσει μόνο μία φορά στο συνολικό πλήθος των λέξεων) προς τον αριθμό των tokens που αποτελούν το κείμενο αυτό. Τέλος το Tweet Entropy (εντροπία του tweet) είναι ένα χαρακτηριστικό του οποίου η τιμή υπολογίζεται από την εξίσωση $S = -\Sigma_i P_i \, log \, P_i$ όπου P_i είναι η πιθανότητα της λέξης i να εμφανίζεται στο σύνολο δεδομένων.

3.3.2 Χαρακτηριστικά από το Sentiment Analysis

Το Sentiment Analysis ενός κειμένου μελετά και προσδιορίζει συναισθηματικές καταστάσεις υποκειμενικής πληροφορίας. Γνωστή και ως «εξόρυξη γνώμης» έχει ως στόχο να καταλάβει από τα συμφραζόμενα ενός κειμένου την πολικότητά του, δηλαδή, αν είναι θετική, αρνητική ή ουδέτερη. Η μέθοδος αυτή συνήθως εφαρμόζεται από επιχειρήσεις που θέλουν να μάθουν τι αντίκτυπο έχει το προϊόν ή η υπηρεσία τους και πώς εκλαμβάνεται από την κοινωνία. Ο τομέας βρίσκεται σε ραγδαία ανάπτυξη και συνδυάζεται με μεθόδους μηχανικής μάθησης για τη βελτιστοποίηση των αποτελεσμάτων σε εφαρμογές μεγάλης κλίμακας.

Διάφορα σημασιολογικά (semantic) χαρακτηριστικά τα οποία προέκυψαν από το Sentiment Analysis μπορούν να υπολογιστούν μέσω της βιβλιοθήκης spaCy. Αυτά περιλαμβάνουν τα σύνολα των θετικών, αρνητικών και ουδέτερων emojis τα οποία δείχνουν αντίστοιχα το αν ένα περιεχόμενο ενός κειμένου είναι θετικό, αρνητικό ή ουδέτερο σημασιολογικά. Επίσης το συνολικό “συναίσθημα” (sentiment) ενός tweet αναδεικνύεται από δύο χαρακτηριστικά, το ένα βασισμένο στην πόλωση (θετικό, αρνητικό, ουδέτερο) και το δεύτερο στην αντικειμενικότητα του κειμένου.

3.3.3 Part-of-Speech (PoS) χαρακτηριστικά

Ένας αριθμός από PoS χαρακτηριστικά επίσης περιλαμβάνεται στο σύνολο των χαρακτηριστκών που μπορούν να χρησιμοποιηθούν. Αυτά τα χαρακτηριστικά αριθμούν τις εμφανίσεις των αντωνυμιών, προσδιορισμών (determiners), επιθέτων, ουσιαστικών, επιρρημάτων και των οντοτήτων (entities) μέσα ένα κείμενο. Τα χαρακτηριστικά αυτά υπολογίζονται μέσα από εξειδικευμένες συναρτήσεις της βιβλιοθήκης spaCy.

3.3.4 TF-IDF

Ένα πολύ σημαντικό χαρακτηριστικό αποτελεί το TF-IDF(Term Frequency [times] Inverse Document Frequency). Το TF-IDF είναι μία στατιστική μετρική που εκφράζει τη σημαντικότητα μιας λέξης σε ένα κείμενο και μπορεί να δώσει πολλές πληροφορίες που είναι ιδιαίτερα χρήσιμες για αλγόριθμους Μηχανικής Μάθησης πάνω στο NLP (αξιολογώντας για παράδειγμα τις λέξεις ενός κειμένου με ένα score). Ουσιαστικά, υπολογίζεται η σχετική συχνότητα μίας λέξης σε ένα κείμενο, σε σχέση με την εμφάνισή της στο σύνολο των κειμένων. Ο υπολογισμός αυτής της μετρικής γίνεται μέσω του πολλαπλασιασμού του σχετικού αριθμού των εμφανίσεων της λέξης μέσα σε ένα κείμενο (TF) και του Inverse Document Frequency (IDF) που ουσιαστικά εκφράζει το πόσο συχνή ή όχι είναι η εμφάνιση μίας λέξης σε όλη την συλλογή των κειμένων που αναλύονται. Για να εξαχθεί η μετρική αυτή πρέπει να έχουν ήδη εφαρμοστεί στα διαθέσιμα κείμενα οι διαδικασίες του καθαρισμού των λέξεων-κλειδιών και σημείων στίξης, stemming, lemmatization και vectorization. Το αποτέλεσμα της μετρικής αυτής είναι ένα σύνολο των πιο σημαντικών λέξεων που εμφανίζονται στο σύνολο των κειμένων του συνόλου δεδομένων. Στη συνέχεια για κάθε tweet υπολογίζεται ένα score για κάθε μία από αυτές τις λέξεις. Το score αυτό εξαρτάται από τον αριθμό των εμφανίσεων των λέξεων στο tweet αλλά και τη σημαντικότητα της λέξης στο σύνολο των δεδομένων.

3.4 Μετα-επεξεργασία

Μετά την εξαγωγή των χαρακτηριστικών από τις διαθέσιμες οντότητες, οι οποίες στην παρούσα Εργασία είναι τα συλλεχθέντα tweets, ακολουθούν δύο ακόμα διαδικασίες μετα-επεξεργασίας οι οποίες σκοπό έχουν να βελτιστοποιήσουν την απόδοση των αλγορίθμων Μάθησης.

3.4.1 Κλιμακοποίηση

Δεδομένου ότι οι περισσότεροι αλγόριθμοι Μηχανικής Μάθησης δεν αποδίδουν καλά όταν τα αριθμητικά χαρακτηριστικά που εισάγονται έχουν πολύ διαφορετικά εύρη τιμών, τα δεδομένα μετασχηματίζονται μέσω της αλγοριθμικής τεχνικής της κλιμακοποίησης χαρακτηριστικών (feature scaling). Συνήθης επιλογή κανονικοποίησης είναι τα δεδομένα να μετασχηματιστούν σε μία κλίμακα με μέση τιμή 0 και διασπορά 1 (standardization). Σε αλγόριθμους ωστόσο που κάτι τέτοιο δεν είναι εφικτό (όπως για παράδειγμα αλγόριθμοι που στηρίζονται στη χρήση πιθανοτήτων και επομένως δεν μπορούν να δεχθούν αρνητική τιμή στην είσοδο) τα δεδομένα μετασχηματίζονται σε κλίμακα στο διάστημα από 0 εώς 1 (min-max scaling).

3.4.2 Μείωση διαστατικότητας με PCA

Πολλά προβλήματα Μηχανικής Μάθησης περιλαμβάνουν χιλιάδες ή ακόμα και εκατομμύρια χαρακτηριστικά για κάθε οντότητα που χρησιμοποιείται στην εκπαίδευση (training) του

αλγορίθμου Μηχανικής Μάθησης. Αυτό οδηγεί σε μια εξαιρετικά αργή διαδικασία εκπαίδευσης και είναι δύσκολο να βρεθεί μια καλή λύση. Αυτό το πρόβλημα αναφέρεται συχνά ως "η κατάρα της διαστατικότητας" [27]. Ωστόσο, μειώνοντας σημαντικά τον αριθμό των χαρακτηριστικών, μετατρέπεται ένα δύσχρηστο πρόβλημα σε εύχρηστο [28].

Εκτός από την επιτάχυνση της εκπαίδευσης και τη μείωση των χαρακτηριστικών (διαστάσεων), μειώνοντας τον υψηλό αριθμό των διαστάσεων ενός συνόλου εκπαίδευσης, μόνο σε 2 ή 3, είναι δυνατή η οπτικοποίηση των οντοτήτων σε ένα γράφημα, δίνοντας έτσι τη δυνατότητα να εξαχθούν σημαντικές πληροφορίες μέσω της οπτικής ανίχνευσης μοτίβων, όπως συγκεκριμένες συστάδες. Το Principal Component Analysis (PCA) είναι μακράν ο πιο δημοφιλής αλγόριθμος μείωσης διαστάσεων. Αρχικά προσδιορίζει το υπερεπίπεδο που βρίσκεται πλησιέστερα στα δεδομένα και, στη συνέχεια, προβάλλει τα δεδομένα σε αυτό [29]. Μέσω της εφαρμογής του αλγορίθμου τα σημαντικότερα γραμμικά συσχετιζόμενα χαρακτηριστικά αποκαλύπτονται. Γίνεται έτσι η εξαγωγή των σημαντικότερων συσχετιζόμενων χαρακτηριστικών από αυτά που εξάγονται κατά την Μηχανική Χαρακτηριστικών (Feature Engineering) και η εξάλειψη των λιγότερο σημαντικών, ώστε να μειωθεί σημαντικά ο όγκος των δεδομένων και να αυξηθεί η απόδοση του αλγορίθμου Μηχανικής Μάθησης που θα χρησιμοποιηθεί.

3.5 Μηχανική Μάθηση

Ως Μηχανική Μάθηση (Machine Learning), ονομάζεται ο κλάδος της τεχνητής νοημοσύνης που ασχολείται με τη μελέτη ή την κατασκευή συστημάτων που έχουν την ικανότητα να «μαθαίνουν» από ένα πλήθος δεδομένων. Κύρια εφαρμογή της Μηχανικής Μάθησης, είναι η δυνατότητα ενός συστήματος να κατηγοριοποιεί δεδομένα, με σκοπό να αναγνωρίζει αντικείμενα και χαρακτηριστικά του φυσικού κόσμου. Περιλαμβάνει τη κατασκευή στατιστικών μοντέλων για την πρόβλεψη, ή εκτίμηση μιας εξόδου σε ένα σύστημα, βασισμένη σε μία ή περισσότερες εισόδους. Οι είσοδοι σε ένα σύστημα Μηχανικής Μάθησης αποτελούν τα χαρακτηριστικά του προβλήματος. Μετά την εφαρμογή των διάφορων μοντέλων πάνω στα χαρακτηριστικά αυτά εκτιμάται η κατηγορία ή ετικέτα η οποία χαρακτηρίζει τα αντίστοιχα δεδομένα εισόδου. Σε ένα πιο μαθηματικό ορισμό, ένα σύστημα Μηχανικής Μάθησης δέχεται στην είσοδο ένα σύνολο χαρακτηριστικά x_i και παράγει μία τιμή y. Η κύρια ταξινόμηση μεθόδων τεχνητής μάθησης προκύπτει ανάλογα με το είδος των δεδομένων και της πληροφορίας που είναι αρχικά διαθέσιμη στο σύστημα. Διακρίνεται σε δύο βασικές κατηγορίες, την Εποπτευόμενη (Supervised) και την Μη-εποπτευόμενη (Unsupervised) Μάθηση.

Στην Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning) το σύστημα "εκπαιδεύεται" χρησιμοποιώντας ήδη υπάρχουσα "γνώση". Αρχικά υποβάλλεται σε μία φάση εκπαίδευσης κατά την οποία σε αυτό εισάγεται ένα αντιπροσωπευτικό σύνολο από δεδομένα των οποίων η ετικέτα είναι ήδη γνωστή. Τα δεδομένα αυτά αποκαλούνται δεδομένα εκπαίδευσης (training data). Στηριζόμενο στις τιμές που αντιστοιχούν στα δεδομένα εκπαίδευσης αυτές, το σύστημα πρέπει να αντιστοιχίσει μελλοντικές εισόδους με όσο γίνεται μεγαλύτερη ακρίβεια στην καταλληλότερη από τις δυνατές εξόδους. Προβλήματα στα οποία χρησιμοποιούνται μέθοδοι Εποπτευόμενης μάθησης είναι η Κατηγοριοποίηση (Classification) και η Παλινδρόμηση (Regression). Το πρόβλημα της αναγνώρισης μη έγκυρων δημοσιεύσεων στο Twitter είναι ένα

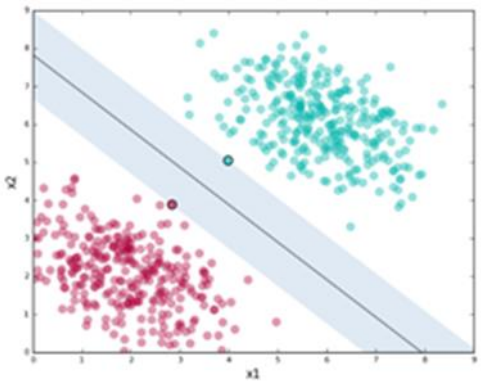
τέτοιο πρόβλημα ταξινόμησης, καθώς από κάθε tweet εξάγεται ένα σύνολο χαρακτηριστικών και με βάση αυτά, το επιλεγόμενο μοντέλο πρέπει να επιλέξει σε ποια από τις επιλεγμένες κλάσεις εντάσσεται.

3.5.1 Αλγόριθμος Multinomial Naive Bayes

Ο αλγόριθμος Multinomial Naive Bayes [30] εντάσσεται στους αλγορίθμους που επιλύουν το πρόβλημα της Κατηγοριοποίησης (Classification). Στηρίζεται στο γνωστό ομώνυμο θεώρημα της Θεωρίας Πιθανοτήτων, το οποίο υπολογίζει την πιθανότητα πραγματοποίησης ενός ενδεχομένου Α, δεδομένης μίας κατάστασης Β. Λαμβάνοντας στην είσοδο τις πιθανότητες πραγματοποίησης των ενδεχομένων του προβλήματος, υπολογίζει τις πιθανότητες για κάθε πιθανή απάντηση. Ως απόφαση θεωρείται η κατηγορία με τη μεγαλύτερη πιθανότητα. Ο αλγόριθμος αυτός αποκαλείται αφελής (naive), καθώς θεωρεί ότι όλα τα ενδεχόμενα τα οποία οδηγούν στην απόφαση είναι ανεξάρτητα μεταξύ τους, δεν επηρεάζει δηλαδή το ένα την πραγματοποίηση του άλλου. Παρά τις απλουστεύσεις που χαρακτηρίζουν την εφαρμογή του έχει εντυπωσιακά ακριβή αποτελέσματα.

3.5.2 Αλγόριθμος SVM

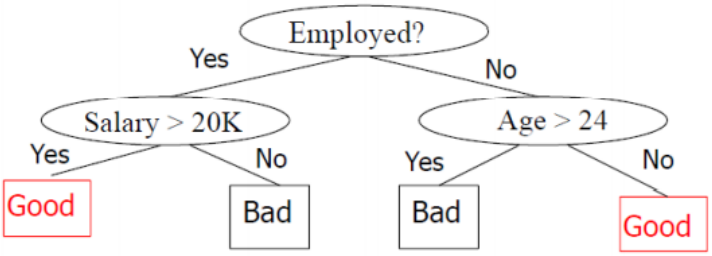
Τα SVM (Support Vector Machines) [31] επιλύουν επίσης το πρόβλημα της κατηγοριοποίησης. Η μέθοδος εντοπίζει μέσω κατάλληλων μαθηματικών μοντέλων το κατώφλι εκείνο το οποίο καθορίζει την συνάρτηση απόφασης για το σύνολο. Στη συνέχεια βρίσκει το βέλτιστο δυνατό εύρος για το οποίο το κατώφλι αυτό παραμένει αξιόπιστο. Η μέθοδος είναι ιδιαίτερα αξιόπιστη και χρησιμοποιείται σε πολλές εφαρμογές. Παρουσιάζει καλή συμπεριφορά ακόμα και αν τα δεδομένα εισόδου δεν έχουν υψηλό ποσοστό αξιοπιστίας (ύπαρξη θορύβου).



Εικόνα 2. Παράδειγμα Ταξινόμησης με χρήση SVM.

3.5.3 Αλγόριθμος Decision Trees (DT)

Τα δέντρα απόφασης [32] αποτελούν μία από τις πιο γνωστές μεθόδους κατηγοριοποίησης. Στηρίζονται στην πεποίθηση ότι η λήψη μιας απόφασης προϋποθέτει μία αλληλουχία προηγούμενων ενεργειών και γεγονότων. Αυτή η αλληλουχία είναι εύκολο να απεικονιστεί με ένα διάγραμμα δένδρου ή δένδρο αποφάσεων. Το διάγραμμα δένδρου είναι μια χρονολογική απεικόνιση όλων των πιθανών ακολουθιών ενεργειών και γεγονότων που οδηγούν στην τελική απόφαση. Κυριότερο πλεονέκτημά της μεθόδου είναι ότι αποτελεί τον καλύτερο τρόπο περιγραφής του προβλήματος γιατί παρουσιάζει κάθε ενέργεια (απόφαση), καθώς και τις αντίστοιχες δεδομένες εκβάσεις με σαφήνεια και απλότητα. Ωστόσο, αφού αποτελεί μία λογική δομή δεν είναι εύκολη η εφαρμογή μοντέλων και διαδικασιών από τα μαθηματικά και τη Θεωρία Γράφων που ενδεχομένως θα επιταχύνουν ή απλοποιούν την διαδικασία απόφασης.



Εικόνα 3. Παράδειγμα Ταξινόμησης με χρήση DT.

3.5.4 Αλγόριθμος Random Forest

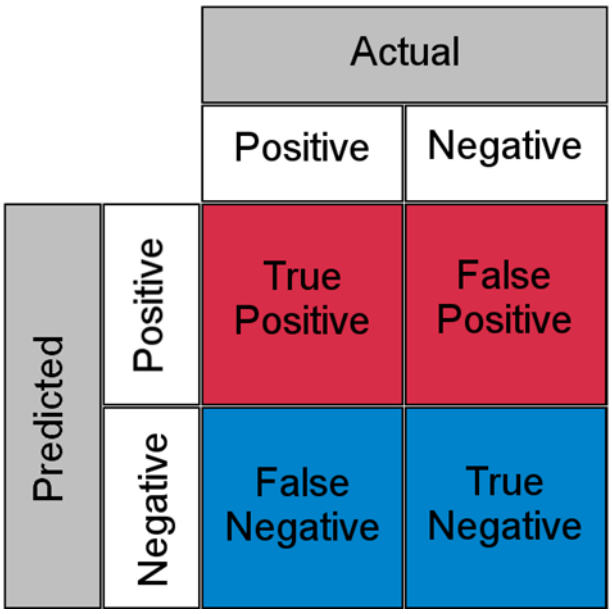
Ο αλγόριθμος Random Forest [33] αποτελεί μία γενίκευση των δέντρων απόφασης καθώς χρησιμοποιεί ένα σύνολο δέντρων απόφασης για να οδηγηθεί σε αποτέλεσμα. Το σύνολο αυτό προκύπτει με την χρήση κατάλληλων αλγορίθμων οι οποίοι εφαρμόζονται στα χαρακτηριστικά του προβλήματος. Στη συνέχεια κάθε δένδρο καταλήγει σε συμπέρασμα για ένα πρόβλημα ταξινόμησης. Η τελική και οριστική κατηγοριοποίηση γίνεται με το «τυχαίο δάσος» να διαλέγει την κατηγορία στην οποία τα περισσότερα δέντρα είχαν ταύτιση συμπεράσματος.

3.5.5 Εκπαίδευση και επιδόσεις Αλγορίθμων

Για την εκπαίδευση των αλγορίθμων απαιτείται οι οντότητες εισόδου (δηλαδή τα διανύσματα των εξαχθέντων χαρακτηριστικών για κάθε tweet), μαζί με τις αντίστοιχες ετικέτες τους, να διαχωριστούν σε δύο ομάδες, τις οντότητες που θα χρησιμοποιηθούν για εκπαίδευση των αλγορίθμων Μηχανικής Μάθησης (σύνολο εκπαίδευσης - train set) και τις οντότητες που θα χρησιμοποιηθούν για την επαλήθευση των αποτελεσμάτων μετά την εκπαίδευσή τους (σύνολο δοκιμών - test set). Για να εξαχθούν συμπεράσματα σχετικά με την αποτελεσματικότητα τους χρησιμοποιείται το test set. Εφόσον είναι ήδη γνωστή η ετικέτα που είχε κάθε οντότητα, αυτή

συγκρίνεται με την ετικέτα που έθεσε ο εκπαιδευμένος αλγόριθμος. Με αυτό τον τρόπο προκύπτουν οι διάφορες μετρικές αξιολόγησης. Χρησιμοποιώντας τις μετρικές αυτές, μπορεί κάθε αλγόριθμος να αξιολογηθεί ως προς την ακρίβεια αναγνώρισής του και αν αυτό επαρκεί ως προς τις απαιτήσεις που τίθενται. Παράλληλα, καθίσταται εφικτή η σύγκριση μεταξύ διαφορετικών αλγορίθμων.

Μία μέθοδος οπτικοποίησης των αποτελεσμάτων της κατηγοριοποίησης, αποτελούν τα confusion matrices (μητρώα σύγχυσης). Σε ένα πρόβλημα δυαδικής κατηγοριοποίησης, το confusion matrix αποτελεί ένα μητρώο 2x2, στο οποίο φαίνονται το πλήθος των True Positive (αληθώς θετικά - TP), True Negative (αληθώς αρνητικά - TN), False Positive (ψευδώς θετικά -FP) και False Negative (ψευδώς αρνητικά – FN). Η ερμηνεία των μεγεθών αυτών προκύπτει με βάση την Εικόνα 4.



Εικόνα 4. Παράδειγμα ενός confusion matrix.

Με βάση τα στοιχεία αυτά, μπορούμε να υπολογίσουμε κάποιες μετρικές, με τις οποίες δίνεται η δυνατότητα να καταλάβουμε την ακρίβεια της κατηγοριοποίησης. Δύο συνήθεις μετρικές που υπολογίζονται είναι οι μετρικές Precision (Ακρίβεια) και Recall (Ανάκληση). Το Precision, υπολογίζεται με τον ακόλουθο τύπο:

$$Precision=TP/(TP+FP)$$

Στον παραπάνω τύπο TP, FP είναι το πλήθος των στοιχείων True Positive και False Positive όπως αυτά προκύπτουν από το confusion matrix. Σε ένα πρόβλημα κατηγοριοποίησης, αν έχουμε Precision ίσο με 1, σημαίνει ότι κάθε στοιχείο που κατηγοριοποιήθηκε στην κατηγορία Υ όντως ανήκει σε αυτή την κατηγορία, χωρίς να λαμβάνονται υπόψη τα στοιχεία της κατηγορίας Υ που κατηγοριοποιήθηκαν λανθασμένα.

Το Recall υπολογίζεται με τον τύπο:

$$Recall=TP/(TP+FN)$$

Αν έχουμε Recall ίσο με 1, σημαίνει ότι κάθε στοιχείο της κατηγορίας Υ κατηγοριοποιήθηκε σωστά, χωρίς να λαμβάνονται υπόψη το πλήθος των στοιχείων της άλλης κατηγορίας που κατηγοριοποιήθηκαν εσφαλμένα στην κατηγορία Υ.

Λόγω του ότι οι δύο αυτές μετρικές αρκετές φορές έχουν αντίστροφη σχέση, συνήθως υπολογίζεται το F1-Score, το οποίο αποτελεί τον αρμονικό μέσο μεταξύ Precision και Recall, και υπολογίζεται με τον τύπο:

$$F1=2*(Precision*Recall)/(Precision+Recall)$$

3.6 Κίνδυνοι Υπερπροσαρμογής και Υποπροσαρμογής

Ανεξάρτητα από την επιλογή του αλγορίθμου και των παραμέτρων του, ένα βασικό χαρακτηριστικό που θα πρέπει να ελέγχεται κατά την διαδικασία εκπαίδευσης παρόμοιων μοντέλων είναι η ποιότητα των δεδομένων εκπαίδευσης και ελέγχου. Τα δεδομένα μπορούν να περιέχουν χρήσιμη πληροφορία που βοηθάει τον αλγόριθμο να μάθει, αλλά και θόρυβο, τον οποίο δεν μπορεί να ξεχωρίσει με αποτέλεσμα να μπερδεύει τις "γνώσεις" του. Η περίπτωση Υπερπροσαρμογής εμφανίζεται όταν ο αλγόριθμος μαθαίνει πάρα πολύ καλά να ξεχωρίζει κατηγορίες των δεδομένων εκπαίδευσής του αλλά δεν μπορεί να γενικεύσει τις γνώσεις του σε νέα δεδομένα. Αυτό συνήθως γίνεται όταν έχουμε μεγάλο αριθμό από χαρακτηριστικά για μικρό σύνολο παρατηρήσεων, όπου μπερδεύεται ο αλγόριθμος και αναλύει σε υπερβολικό βαθμό τα δεδομένα με αποτέλεσμα να εξετάζει το θόρυβο αντί για τη σχέση μεταξύ των μεταβλητών στα δεδομένα εκπαίδευσης. Αντίθετα, έχουμε Υποπροσαρμογή όταν τα δεδομένα εκπαίδευσης και τα χαρακτηριστικά τους δεν είναι αρκετά ώστε ο αλγόριθμος να μπορεί να αναγνωρίσει τάσεις στα δεδομένα εκπαίδευσης και έτσι να μην βρίσκεται σε θέση να γενικεύσει τις γνώσεις του σε νέα δεδομένα.

Επίσης, όταν τα δεδομένα περιέχουν θόρυβο τότε θα έχουν και τα χαρακτηριστικά, με αποτέλεσμα να μπαίνει μεγάλος όγκος θορύβου στην εκπαίδευση του αλγορίθμου, λόγο του ότι ο θόρυβος των αρχικών διαθέσιμων δεδομένων έχει περάσει από τη διεργασία εξαγωγής χαρακτηριστικών και έχει μεταφραστεί σε απλούς αριθμούς, όπου κάθε αριθμός έχει συμπυκνωμένη την πληροφορία και τον θόρυβο από όλα τα χαρακτηριστικά που έχουν εξαχθεί. Όπως προαναφέρθηκε, ένας μεγάλος αριθμός από χαρακτηριστικά αυξάνει την πιθανότητα Υπερπροσαρμογής. Για τον έλεγχο ύπαρξης προβλήματος αλλά και για την επίλυσή του, χρησιμοποιούνται μέθοδοι κανονικοποίησης στα χαρακτηριστικά των δεδομένων στο στάδιο της προεπεξεργασίας.

3.7 Web Annotator Tool

Οι Εποπτευόμενοι αλγόριθμοι Μηχανικής Μάθησης, όπως αυτοί που χρησιμοποιούνται στην συγκεκριμένη Διπλωματική Εργασία, καθιστούν απαραίτητη την εκ των προτέρων γνώση της κατηγορίας/ετικέτας (class/label) που ανήκουν τα δεδομένα που θα τροφοδοτηθούν σε αυτό, ούτως ώστε να εκπαιδευτούν αποτελεσματικά και να μπορέσουν να εξάγουν μοντέλα που θα κατηγοριοποιούν νέα δεδομένα με υψηλά ποσοστά επιτυχίας. Όσο πιο ακριβής είναι η ετικετοποίηση των δεδομένων τόσο πιο αποτελεσματικοί θα είναι και οι αλγόριθμοι. Για τους λόγους αυτούς υλοποιήθηκε ένα διαδικτυακό εργαλείο το οποίο επιτρέπει την άμεση και εύκολη ετικετοποίηση των tweets που έχουν συλλεχθεί. Η υλοποίηση του εργαλείου για το annotation των tweets πραγματοποιήθηκε με τη βοήθεια των τεχνολογιών Flask, HTML5 και CSS.

3.8 Τεχνολογίες που χρησιμοποιήθηκαν

3.8.1 Python

Η Python είναι μία από τις πιο δημοφιλείς γλώσσες προγραμματισμού υψηλού επιπέδου. Είναι συμβατή με τα περισσότερα λειτουργικά συστήματα της αγοράς. Στα βασικά πλεονεκτήματα της γλώσσας συγκαταλέγονται η εξαιρετική δυνατότητα επεξεργασίας συμβολοσειρών, η ευκολία στη συγγραφή καθώς το συντακτικό της προσομοιάζει στο συντακτικό της Αγγλικής γλώσσας,. καθώς και η επεκτασιμότητά της με ένα τεράστιο εύρος βιβλιοθηκών οι οποίες επιτρέπουν την ανάπτυξη λογισμικού εφαρμογών Web, διασύνδεση με άλλα συστήματα και βάσεις δεδομένων, χρήση εξειδικευμένων μαθηματικών μοντέλων κ.α.

Κάποιες από τις πιο βασικές βιβλιοθήκες που διαθέτει η Python και οι οποίες χρησιμοποιούνται στη συνέχεια είναι:

- Η RE: Η βιβλιοθήκη αυτή υποστηρίζει πράξεις χειρισμού αλφαριθμητικών μέσα από την χρήση κανονικών εκφράσεων (Regular Expressions) [34].
- Η collections: Με τη χρήση αυτής της βιβλιοθήκης καθίσταται ευκολότερος ο χειρισμός διαφορετικών τύπων δεδομένων και οντοτήτων οι οποίες οργανώνονται, είναι προσβάσιμες και μπορούν να προσπελαστούν με δομημένο τρόπο.
- Η numpy: Αποτελεί τη σημαντικότερη επέκταση της Python υποστηρίζοντας την εκτέλεση πολύπλοκων μαθηματικών υπολογισμών και πράξεων γραμμικής άλγεβρας με αποδοτικό τρόπο.

3.8.2 Tweepy

Το Tweepy [35] είναι ένα API του Twitter το οποίο επιτρέπει την λήψη δημόσια διαθέσιμων δεδομένων της πλατφόρμας και είναι από τις πιο διαδεδομένες μεθόδους συλλογής δεδομένων σε αυτήν. Δίνει πρόσβαση σε πολλές βιβλιοθήκες και συναρτήσεις που κάνουν την διαχείριση και λήψη δεδομένων από την πλατφόρμα απλή και γρήγορη. Απαιτεί ορισμένα διαπιστευτήρια, τα οποία αποκτώνται μετά από εγγραφή σε αντίστοιχη υπηρεσία του Twitter, χωρίς κάποιο κόστος για

τον χρήστη. Όταν, πλέον, είναι έτοιμο για χρήση, το Tweepy, σε λίγες γραμμές κώδικα μπορεί να δώσει πρόσβαση σε τεράστιο όγκο δεδομένων από tweets με βάση ένα σύνολο κριτηρίων που τίθενται από τον προγραμματιστή σε μορφή JSON(JavaScript Object Notation) [36]. Η μορφή αρχείων JSON είναι μία πρότυπη κωδικοποίηση η οποία επιτρέπει την αναπαράσταση δεδομένων με δομημένο τρόπο.

3.8.3 MongoDB

Για την αποθήκευση των δεδομένων των tweets που συλλέχθηκαν χρησιμοποιείται η τεχνολογία της MongoDB. Η MongoDB [37] είναι πρόγραμμα διαχείρισης εγγραφοστραφής βάσης δεδομένων και είναι το πιο δημοφιλές στην κατηγορία των NoSQL προγραμμάτων. Αναπτύχθηκε από την εταιρία MongoDB Inc., με πρώτη εμφάνιση τον Φεβρουάριο του 2009, σε C++. Χρησιμοποιούμε την έκδοση v4.4.4 και τα σημαντικότερα χαρακτηριστικά της είναι:

1. BSON: Αποθηκεύει την πληροφορία σε έγγραφα τα οποία είναι σε μορφή δυαδικού JSON γνωστό και ως BSON. Τα αρχεία BSON βασισμένα στα JSON είναι ιδανικότερα για δημιουργία, επεξεργασία, μεταφορά και διαγραφή μεγάλων αρχείων, κι έχουν λιγότερο ευανάγνωστη μορφή για τους ανθρώπους
2. Διαχείριση Αρχείων: Μία βάση δεδομένων MongoDB αποτελείται από τα λεγόμενα collections (συλλογές). Μία συλλογή αποτελείται από ομάδες εγγράφων (documents), τα οποία με τη σειρά τους έχουν πεδία με τιμές. Τα έγγραφα μέσα σε μια συλλογή μπορεί να έχουν διαφορετικά πεδία και τιμές αλλά έχουν παρόμοιο θέμα. Η γενική μορφή ενός εγγράφου είναι ζευγάρια κλειδιών- τιμών, ακριβώς σαν ένα αρχείο μορφής JSON.
3. Διαχείριση μεγάλου όγκου δεδομένων: Μέσω της MongoDB δίνεται η δυνατότητα διαχείρισης ακόμα και δεδομένων που είναι πολύ μεγάλου όγκου (big data). Επίσης οι πληροφορίες που χρειάζονται σε κάθε περίπτωση αναπαρίστανται άμεσα και στιγμιαία καθώς δεκάδες χιλιάδες μέχρι και εκατομμύρια MongoDB documents μπορούν να αναζητηθούν ταυτόχρονα. Στην περίπτωσή της παρούσας Εργασίας τα documents αυτά περιέχουν tweets και τα μεταδεδομένα τους.
4. Ερωτήματα (Queries): Η μέθοδος του indexing επιτρέπει την ταχύτατη αναζήτηση εγγράφων. Μπορούν επίσης, να κατασκευαστούν ερωτήματα (queries) τα οποία επιστρέφουν μόνο έγγραφα που πληρούν τις προϋποθέσεις που τίθενται από εκείνα, επιτρέποντας περαιτέρω βελτιστοποίηση του συστήματός με τη σωστή επιλογή αυτών.
5. Συμβασιμότητα με Pyhton: Τέλος, μέσω του MongoDB API PyMongo [38] υπάρχει εξαιρετική συμβατότητα με την Python. Η PyMongo είναι μία βιβλιοθήκη της Python η οποία παρέχει όλα εξειδικευμένες συναρτήσεις και εργαλεία για τη σύνδεση του συστήματος με την βάση δεδομένων, αλλά και για την εισαγωγή, την αναζήτηση και την επεξεργασία δεδομένων που υπάρχουν σε αυτή. Χρησιμοποιήθηκε η έκδοση 3.9.0 του API.

3.8.4 Pandas

Η διαχείριση και η επεξεργασία δεδομένων που υπάρχουν στην βάση δεδομένων γίνεται ακόμα ευκολότερη με τη χρήση της βιβλιοθήκης Pandas [39] . Το Pandas είναι μια βιβλιοθήκη λογισμικού της Python για ανάλυση και χειρισμό δεδομένων. Χρησιμοποιήσαμε την έκδοση 1.2.2. Η ονομασία προέρχεται από τον συνδυασμό των όρων PANel DAta, ο οποίος είναι οικονομετρικός όρος για σετ δεδομένων που περιλαμβάνουν τιμές στην πάροδο χρόνου. Το Pandas επιτρέπει την δόμηση δεδομένων σε αντικείμενα DataFrame τα οποία διευκολύνουν την διαχείριση τους και επιτρέπουν ταχεία πρόσβαση σε μεγάλο όγκο δεδομένων. Ένα DataFrame θυμίζει έναν πίνακα, όπου κάθε στήλη μπορεί να παίρνει διαφορετικού είδους τιμές αλλά και κάθε μία από τις στήλες να μπορεί να διαχειριστεί ξεχωριστά. Το Pandas αποτελεί ένα από τα πιο αξιόπιστα εργαλεία για την επεξεργασία και ανάλυση δεδομένων με αμέτρητα πολλές επιλογές και μεθόδους. Διαθέτει συναρτήσεις για να διαβάσει ή να γράψει σε άλλες δομές δεδομένων ασχέτως από το αν είναι διαφορετικού τύπου, εύκολη εκκαθάριση δεδομένων που είναι περιττά, καθώς και γρήγορο ανασχηματισμό των δεδομένων με βάση πολυάριθμων επιλογών για συνθήκες. Επιπλέον, ενώσεις διαφορετικών συνόλων δεδομένων, μετρικές για πλήθος τιμών που πληρούν κάποια συνθήκη, συχνότητα εμφάνισης και slicing, κάνουν το Pandas μια κατάλληλη βιβλιοθήκη για ανάλυση και διαχείριση μεγάλου όγκου δεδομένων. Τέλος, υπάρχει πολύ καλή συμβατότητα και με το PyMongo API της MongoDB. Η χρήση του Pandas ήταν καθοριστικός παράγοντας σε κάθε στάδιο της διαδικασίας διαχείρισης μεγάλου όγκου δεδομένων αποτελεσματικά και γρήγορα.

3.8.5 NLTK

Τα δεδομένα των tweets πρέπει να υποστούν μία επεξεργασία ώστε να είναι μετατραπούν σε κατάλληλη μορφή και να εξαχθούν τα χαρακτηριστικά αλλά και να αποκοπεί περιττή πληροφορία που φέρουν. Πολύ χρήσιμη για τέτοιους σκοπούς είναι η δημοφιλής πλατφόρμα Natural Language Toolkit (NLTK) [40] η οποία επιτρέπει την δημιουργία προγραμμάτων με Python που επεξεργάζονται δεδομένα φυσικής γλώσσας. Διαθέτει εργαλεία για επεξεργασία κειμένων, κατηγοριοποίηση, tokenization, lemmatization, stemming και πολλά άλλα. Είναι μία πλατφόρμα συνεχούς εξέλιξης προσθέτοντας νέες ενότητες και βελτιώνοντας τις υπάρχουσες, καθώς και συναναστρέφεται ενεργά με την κοινότητα της.

Στα πλαίσια της παρούσας Εργασίας γίνεται λήψη και αξιοποίηση του συνόλου των δεδομένων της NLTK το οποίο αφορά τη χρήση έτοιμων συναρτήσεων για την αναγνώριση και επεξεργασία γλωσσικών δομών στα Αγγλικά, Ελληνικά καθώς και των σημείων στίξης.

3.8.6 spaCy

Η spaCy [41] είναι μια βιβλιοθήκη λογισμικού ανοιχτού κώδικα για προηγμένη επεξεργασία φυσικής γλώσσας, γραμμένη στις γλώσσες προγραμματισμού Python και Cython. Σε αντίθεση με την NLTK, η οποία χρησιμοποιείται ευρέως για τη διδασκαλία και την έρευνα, η spaCy επικεντρώνεται στην παροχή κώδικα για χρήση σε επαγγελματική και εμπορική χρήση.

Μέσω της βιβλιοθήκης αυτής παρέχονται για την ελληνική γλώσσα και χρησιμοποιούνται ενημερωμένες συλλογές χαρακτηριστικών γλωσσικών ιδιωμάτων, οι οποίες συμπληρώνουν τις διαθέσιμες επιλογές από την NLTK, βελτιώνοντας με αυτό τον τρόπο την απόδοση των αλγορίθμων μάθησης.

3.8.7 Scikit-learn

Το Scikit-Learn [42] αποτελεί την πιο διαδεδομένη βιβλιοθήκη Μηχανικής Μάθησης και παρέχει έτοιμες υλοποιήσεις αλγορίθμων όπως ο Random Forest, ο SVM, και ο Naive Bayes. Επιπλέον συμπεριλαμβάνονται αλγόριθμοι για την εξαγωγή χαρακτηριστικών, την κλιμακοποίηση και κανονικοποίηση των χαρακτηριστικών, τη βελτιστοποίηση της απόδοσης των αλγορίθμων μάθησης και την εξαγωγή και οπτικοποίηση αποτελεσμάτων σχετικά με την απόδοση των αλγορίθμων. Οι κύριες συναρτήσεις της βιβλιοθήκης που αξιοποιούνται για τον υπολογισμό διαφορετικών μεθόδων και αλγορίθμων Μηχανικής Μάθησης στην παρούσα εργασία συνοψίζονται στον Πίνακας *2*.

Μέθοδος ή αλγόριθμος Μηχανικής Μάθησης	**Συνάρτηση της βιβλιοθήκης scikit-learn**
Tokenization και υπολογισμός TF-IDF calculation	`TfidfVectorizer`
Μείωση διαστατικότητας μητρώου χαρακτηριστικών	`PCA`
Κλιμακοποίηση του μητρώου χαρακτηριστικών	`StandardScaler, MinMaxScaler`
Αλγόριθμος Μηχανικής Μάθησης SVM	`svm.SVC`
Βελτιστοποίηση παραμέτρων αλγορίθμου SVM	`GridSearchCV`
Αλγόριθμος Μηχανικής Μάθησης RandomForest	`RandomForestClassifier`
Βελτιστοποίηση παραμέτρων αλγορίθμου Random Forest	`RandomizedSearchCV`
Αλγόριθμος Μηχανικής Μάθησης Multinomial Maive Bayes	`MultinomialNB`

Πίνακας 2. Κύριες μέθοδοι και αλγόριθμοι Μηχανικής Μάθησης που χρησιμοποιούνται και οι αντίστοιχες συναρτήσεις της βιβλιοθήκης scikit-learn.

3.8.8 Matplotlib και Seaborn

Η κύρια βιβλιοθήκη για την οπτικοποίηση δεδομένων είναι η Matplotlib [43], η οποία είναι μία αντικειμενοστραφής βιβλιοθήκη της Python για την δημιουργία γραφημάτων όλων των ειδών, με απλουστευμένη χρήση, καθώς επίσης διατίθενται πολυάριθμα παραδείγματα χρήσης και εφαρμογής στο διαδίκτυο. Η βιβλιοθήκη αυτή παρέχεται από τον μη-κερδοσκοπικό οργανισμό NumFOCUS, με πρώτη έκδοση το 2003. Η seaborn [44] είναι μία ακόμα βιβλιοθήκη οπτικοποίησης της python, η οποία είναι βασισμένη στην matplotlib. Η βιβλιοθήκη αυτή χαρακτηρίζεται από τον αφαιρετικό τρόπο με τον οποίο μπορεί να δημιουργήσει κάποιος, με χαρακτηριστική ευκολία, όμορφα και γεμάτα πληροφορίες γραφήματα.

3.8.9 Jupyter Notebook

Το Jupyter Notebook [45] είναι μια διαδικτυακή εφαρμογή, ελεύθερου λογισμικού και ανοιχτού κώδικα που επιτρέπει την δημιουργία και την διαμοίραση εγγράφων που περιέχουν κώδικα, εξισώσεις, γραφήματα, απεικονίσεις και επεξηγηματικό κείμενο. Το Jupyter notebook προσφέρει δυνατότητες παρουσίασης μορφοποιημένου κείμενου, μαθηματικών εξισώσεων και συμβόλων, και διαδραστικά στοιχεία με εκτελέσιμο κώδικα υπολογιστή, μέσα από ένα γνώριμο περιβάλλον ενός web browser. Ανάμεσα στις χρήσεις του Jupyter Notebook περιλαμβάνονται λειτουργίες όπως: ο καθαρισμός και ο μετασχηματισμός δεδομένων, η αριθμητική προσομοίωση, η στατιστική μοντελοποίηση, η μηχανική μάθηση και πολλά άλλα. Το Jupyter Notebook υποστηρίζει πάνω από 40 γλώσσες προγραμματισμού, συμπεριλαμβανομένων των δημοφιλών γλωσσών στην Επιστήμη των δεδομένων, όπως η Python, η R, η Julia και η Scala. Τέλος, προσφέρει διαδραστικά widgets που μπορούν να χρησιμοποιηθούν για τον χειρισμό και την οπτικοποίηση των δεδομένων σε πραγματικό χρόνο.

3.8.10 Flask

Το Flask [46] είναι ένα "microframework", δηλαδή ένα micro περιβάλλον για την ανάπτυξη web εφαρμογών σε Python. Είναι "microframework" καθώς έχει όλες τις βασικές λειτουργίες για μία απλή εφαρμογή web και η λειτουργικότητά του μπορεί να επεκταθεί με extensions. Επίσης το Flask δεν επιβάλλει τη χρήση κάποιων συγκεκριμένων εργαλείων, όπως ίσως συμβαίνει σε κάποια πιο σύνθετα frameworks. Αφήνει στο χρήστη την επιλογή για το ποια βάση δεδομένων θα χρησιμοποιήσει, με ποιο τρόπο θα συνδεθούν οι χρήστες στο λογαριασμό τους, κλπ.

3.8.11 HTML5

Η HTML5 είναι μία γλώσσα σήμανσης (markup language) η οποία χρησιμοποιείται για την δόμηση και την παρουσίαση του περιεχομένου στον Παγκόσμιο Ιστό. Ουσιαστικά αυτό που βλέπει κάποιος όταν επισκέπτεται μία ιστοσελίδα είναι ένα σύνολο από στοιχεία HTML. Η HTML5 ,όπως υποδεικνύει και το όνομα, είναι η πέμπτη και πολύ σημαντική έκδοση η οποία φέρει πολλές βελτιώσεις σε σχέση με προηγούμενες όσον αφορά τη σύνταξη της , τα στοιχεία που μπορεί να

παρουσιάσει αλλά και τη συνεργασία της με διάφορα API για τη δημιουργία πολύπλοκων εφαρμογών. Ένα παράδειγμα της εξέλιξης είναι ότι η HTML5 είναι αρκετή πλέον για την αναπαραγωγή βίντεο ενώ παλαιότερα χρειαζόταν ένα λογισμικό το οποίο αποκαλείται Flash.

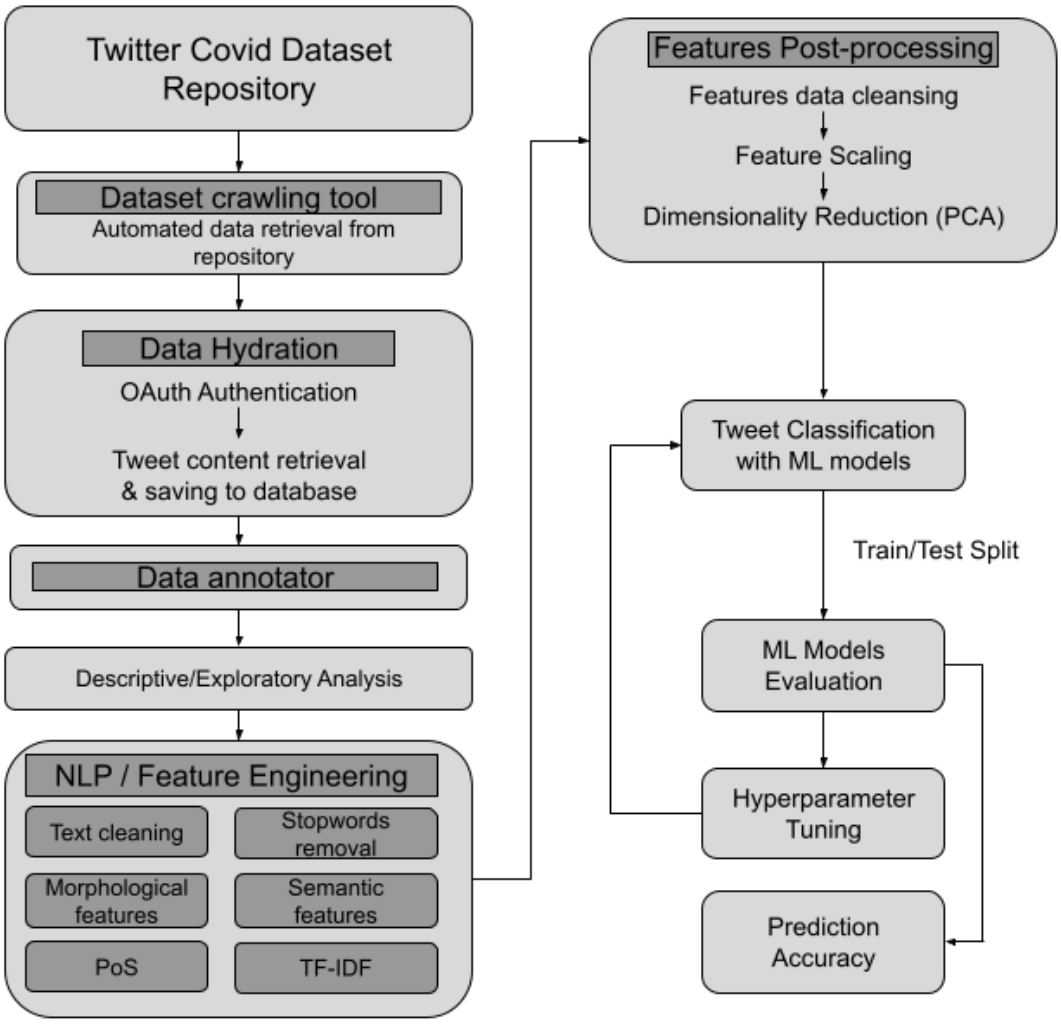
3.8.12 CSS (Cascading Style Sheets)

Η CSS (Cascading Style Sheets) είναι μία γλώσσα προγραμματισμού η οποία χρησιμοποιείται για τον έλεγχο της εμφάνισης ενός εγγράφου που γράφτηκε στις γλώσσες HTML και XHTML, δηλαδή για τον έλεγχο της εμφάνισης μιας ιστοσελίδας και γενικότερα ενός ιστοτόπου. Η CSS είναι μια γλώσσα υπολογιστή προορισμένη να αναπτύσσει στυλιστικά μια ιστοσελίδα δηλαδή να διαμορφώνει περισσότερα χαρακτηριστικά, χρώματα, στοίχιση και δίνει περισσότερες δυνατότητες σε σχέση με την HTML. Για μια όμορφη και καλοσχεδιασμένη ιστοσελίδα η χρήση της CSS κρίνεται ως απαραίτητη.

4

Αρχιτεκτονική και Υλοποίηση Συστήματος

Το σύστημα που υλοποιήθηκε στην Διπλωματική Εργασία αυτή μπορεί να αναλυθεί σύμφωνα με το διάγραμμα της Εικόνας 5.



Εικόνα 5. Γενική περιγραφή του συστήματος.

Αρχικά, ανακτώνται τα αναγνωριστικά των tweets που μας ενδιαφέρουν από το online αποθετήριο (repository) που φιλοξενεί το σύνολο δεδομένων. Για το σκοπό αυτό δημιουργήθηκε ειδικό εργαλείο crawler, το οποίο μπορεί να κάνει αναζήτηση των δεδομένων στο αποθετήριο με κριτήρια όπως η ημερομηνία και η γλώσσα των tweets. Στη συνέχεια, με βάση τα αναγνωριστικά των tweets, πραγματοποιείται σύνδεση στο API του Twitter για το hydration των tweets και η αποθήκευσή του στη βάση δεδομένων mongoDB. Ακολουθεί η διαδικασία ετικετοποίησης των δεδομένων σε τρεις κατηγορίες. Για την διευκόλυνση αυτής της χρονοβόρας διαδικασίας υλοποιήθηκε μία web εφαρμογή η οποία κάνει λήψη του περιεχομένου των αποθηκευμένων στη βάση tweets και μέσω κατάλληλων εργαλείων διεπαφής επιτρέπει την γρήγορη επιλογή της καταλληλότερης κατηγορίας στην οποία ανήκει το καθένα.

Αφού ολοκληρωθεί η διαδικασία ετικετοποίησης, γίνεται διερευνητική ανάλυση των δεδομένων για την βασική κατανόηση και εξαγωγή συμπερασμάτων σχετικά με τη δομή τους και την εύρεση χαρακτηριστικών μοτίβων. Ακολουθεί μία διαδικασία εφαρμογής μεθόδων επεξεργασίας φυσικής γλώσσας, μέσω της οποίας τα διαθέσιμα κείμενα από κάθε tweet καθαρίζονται από περιττά στοιχεία ή στοιχεία που δεν συνεισφέρουν νοηματικά στην πρόταση και εξάγεται ένα σύνολο χαρακτηριστικών από τη μορφολογία, το θέμα και τις λέξεις κάθε κειμένου. Επιπλέον εφαρμόζεται η μέθοδος TF-IDF μέσω της οποίας εντοπίζονται οι σημαντικότερες λέξεις στο σύνολο των διαθέσιμων δεδομένων.

Ακολουθεί η διαδικασία στην οποία το σύνολο των χαρακτηριστικών που εξήχθησαν, σε συνδυασμό με τις κατηγορίες των σημαντικών λέξεων που προέκυψαν από την εφαρμογή της TF-IDF προστίθενται σε ένα ενιαίο μητρώο χαρακτηριστικών (features matrix). Σε αυτό το μητρώο εφαρμόζονται τεχνικές μετεπεξεργασίας όπως κλιμακοποίηση, κανονικοποίηση και εξαγωγή των σημαντικότερων χαρακτηριστικών με τη μέθοδο PCA για τη μείωση των διαστάσεων του matrix. Μετά από αυτό το βήμα, το μητρώο χαρακτηριστικών χωρίζεται σε σύνολο εκπαίδευσης (train set) και σύνολο δοκιμών (test set) τα οποία στη συνέχεια χρησιμοποιούνται για την εκπαίδευση και αξιολόγηση των αλγορίθμων μηχανικής μάθησης που δοκιμάζονται. Τέλος, για κάθε αλγόριθμο γίνεται βελτιστοποίηση των βασικών του παραμέτρων για την εξαγωγή ακριβέστερων αποτελεσμάτων.

Στη συνέχεια του κεφαλαίου αυτού γίνεται περιγραφή των επιμέρους στοιχείων που απαρτίζουν το σύστημα που περιγράφηκε.

4.1 Συλλογή δεδομένων

Για την επίλυση του προβλήματος αναγνώρισης και κατηγοριοποίησης των tweets, χρησιμοποιήθηκε ένα τμήμα του COVID-19 Twitter chatter dataset [47]. Το σύνολο δεδομένων αυτό ξεκίνησε να δημιουργείται από τις 11 Μαρτίου 2020, αποδίδοντας πάνω από 4 εκατομμύρια tweets την ημέρα. Έχουν συμπεριληφθεί καθημερινά hashtags, αναφορές σε άλλους χρήστες, emojis και τις συχνότητές τους. Για την ευκολότερη χρήση του συνόλου δεδομένων, συμπεριλαμβάνονται εκτός των μοναδικών αναγνωριστικών των tweets (IDs) και η γλώσσα στην οποία είναι γραμμένα. Το πλήρες σύνολο δεδομένων αφορά όλες τις γλώσσες, ωστόσο οι επικρατέστερες είναι τα Αγγλικά, τα Ισπανικά και τα Γαλλικά και περιλαμβάνει 903.223.501 tweets και retweets. Επιπλέον παρέχεται και μια καθαρή έκδοση χωρίς retweets (226.582.903 μοναδικά

tweets). Για την εξυπηρέτηση εφαρμογών NLP παρέχονται επιπλέον οι κορυφαίοι 1000 πιο συχνοί όροι που συναντώνται, καθώς και τα 1000 κορυφαία bigrams και trigrams [48] τα οποία αποθηκεύονται, χωρισμένα ανά μέρα, καθημερινά σε ένα online αποθετήριο (repository) στην πλατφόρμα του github [49].

4.1.1 Περιγραφή και λήψη των δεδομένων

Οι πληροφορίες σχετικές με τα tweets εντός των συνόλων δεδομένων που είναι αποθηκευμένες στο github αποτελούνται μόνο από το μοναδικό χαρακτηριστικό των (tweet ID), την ημερομηνία και την ώρα της δημοσίευσης, την γλώσσα με την οποία είναι γραμμένα και την χώρα από την οποία προήλθαν.

Για τους σκοπούς της συγκεκριμένης Εργασίας χρησιμοποιήθηκαν tweets που δημοσιεύτηκαν στο διάστημα μεταξύ 1ης Νοεμβρίου του 2020 και 30 Δεκεμβρίου του 2020. Για να επιτευχθεί η συλλογή των συγκεκριμένων tweets κατασκευάστηκε με την γλώσσα Python ένα εργαλείο που επιτρέπει την αυτόματη αποθήκευση tweets από το github στον τοπικό χώρο αποθήκευσης του υπολογιστή σε μορφή CSV. Το εργαλείο αυτό, δίνει επίσης τη δυνατότητα της επιλογής του εύρους των ημερομηνιών για το οποίο είναι επιθυμητή η συλλογή των tweets αλλά και της γλώσσας στην οποία είναι γραμμένα. Στην συγκεκριμένη περίπτωση οι ημερομηνίες είναι το διάστημα που αναφέρθηκε προηγουμένως ενώ η γλώσσα είναι τα Ελληνικά της οποίας η συντομογραφία με την οποία είναι αποθηκευμένα τα tweets στο github είναι το “el”. Σημαντικό να ειπωθεί είναι το γεγονός ότι τα tweets συλλέγονται από τις καθαρές εκδόσεις των συνόλων δεδομένων έτσι δεν υπάρχει ανησυχία για διπλότυπα των tweets που θα συλλεχθούν. Συνολικά συλλέχθηκαν 61.147 tweet IDs.

4.1.2 Tweet hydrator

Προφανώς οι πληροφορίες που αποθηκεύτηκαν τοπικά δεν είναι επαρκής για τους σκοπούς αυτής της Διπλωματικής Εργασίας καθώς θα πρέπει να είναι διαθέσιμα τα πλήρη κείμενα των tweet ώστε να μπορέσει να εκτελεστεί η κατάλληλη ανάλυση. Για να επιτευχθεί αυτό κατασκευάστηκε ένα εργαλείο διασύνδεσης με το Twitter API. Μέσω του εργαλείου αυτού εντοπίζονται τα δεδομένα των tweets που συμπεριλαμβάνονται στα τοπικά δεδομένα, γίνεται λήψη του περιεχομένου τους τοπικά στον διακομιστή που εκτελείται το εργαλείο και στη συνέχεια αποθηκεύονται σε μία βάση δεδομένων mongoDB.

4.1.3 Data annotation

Οι αλγόριθμοι που χρησιμοποιούνται στην συγκεκριμένη Διπλωματική Εργασία απαιτούν την εκ των προτέρων γνώση της κατηγορίας που ανήκει ένα tweet ανάλογα με το πόσο έγκυρο είναι το θεματικό τους περιεχόμενο ώστε να εκπαιδευτούν σωστά (να γίνει ακριβής εκπαίδευση) και στη συνέχεια να παράξουν μοντέλα που θα κατηγοριοποιούν νέα tweets με ακρίβεια. Στα πλαίσια της παρούσας Εργασίας, επιλέχθηκε ο διαχωρισμός των tweets σε τρεις κατηγορίες, οι οποίες χαρακτηρίζουν ένα tweet σύμφωνα με το περιεχόμενο του . Πρέπει να σημειωθεί ότι στα πλαίσια

της παρούσας εργασίας δεν υπήρχε επαγγελματική δημοσιογραφική υποστήριξη στην διαδικασία της ετικετοποίησης, η οποία πραγματοποιήθηκε από ένα άτομο. Καθώς η επιλογή των κατηγοριών διαχωρισμού του συνόλου δεδομένων είναι μία πολύπλοκη εργασία η οποία απαιτεί καλή γνώση του προβλήματος, η ορθότητα των κατηγοριών και της μεθόδου χαρακτηρισμού των tweets μπορεί να αναθεωρηθεί στο μέλλον.

Οι κατηγορίες οι οποίες επιλέχθηκαν είναι:

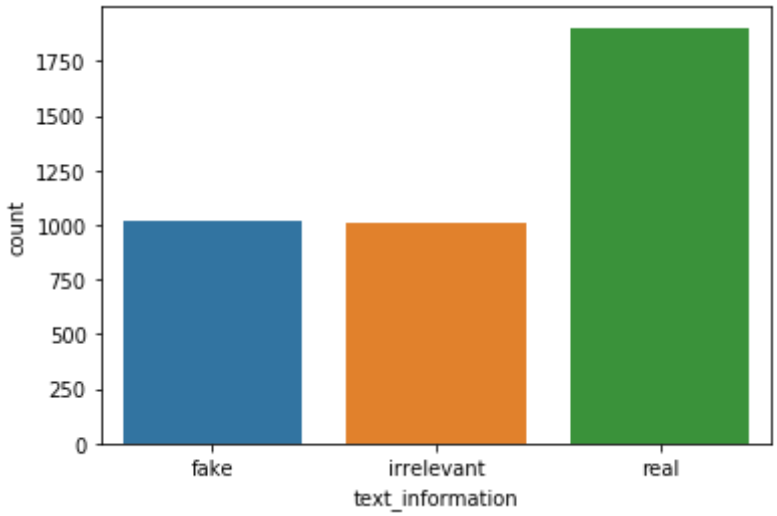
1) Τα tweets τα οποία είναι αντικειμενικά αληθή. Τέτοια tweets είναι για παράδειγμα δημοσιεύσεις ειδήσεων σχετικών με την πανδημία, νέα σχετικά με τον ιό Sars-Cov-2, δημοσιεύσεις σχετικές με τα μέτρα περιορισμού της πανδημίας κ.ο.κ.
2) Τα tweets τα οποία θεωρούνται ψευδή. Ως ψευδείς χαρακτηρίζονται δημοσιεύσεις οι οποίες σχετίζονται με τον ιό αλλά έχουν αμφιλεγόμενο ή και σατιρικό περιεχόμενο. Ο χαρακτηρισμός αυτός ωστόσο είναι δύσκολος καθώς απαιτεί γνώση των γεγονότων τα οποία αναφέρονται.
3) Τέλος, ως άσχετες χαρακτηρίζονται δημοσιεύσεις οι οποίες δεν σχετίζονται με την εξέλιξη των γεγονότων της πανδημίας. Για παράδειγμα η δημοσίευση "Με #COVID19 μόνο playstation και netflix" χαρακτηρίζεται εύκολα ως άσχετη.

Ωστόσο για την ετικετοποίηση υφίσταται ένα ακόμα πρόβλημα. Η επιλογή των tweets ένα προς ένα και η προσθήκη ετικέτας είναι μία εξαιρετικά χρονοβόρα διαδικασία. Για τον λόγο αυτό κατασκευάστηκε ένα online annotator tool (online εργαλείο ετικετοποίησης των tweets– Tweet Annotator) με τη γλώσσα Python και τη βοήθεια της βιβλιοθήκη της, Flask, ή οποία είναι προσανατολισμένη στην ανάπτυξη web εφαρμογών.

Μέσα από την εφαρμογή αυτή ένας χρήστης μπορεί να επιλέξει ένα εύρος ημερομηνιών για το οποίο επιθυμεί να ετικετοποιήσει tweets αλλά και και τον αριθμό των tweets που επιθυμεί να του εμφανιστούν στην οθόνη του. Τα tweets αυτά επιλέγονται τυχαία από την βάση δεδομένων, με κριτήριο να βρίσκονται στο εύρος που έχει επιλέξει ο χρήστης και να μην έχουν ήδη ετικετοποιηθεί στο παρελθόν, και εμφανίζονται στην ιστοσελίδα παρουσιάζοντας το πλήρες κείμενό τους. Στη συνέχεια, ο χρήστης επιλέγει την κατάλληλη ετικέτα για κάθε ένα από αυτά και εφόσον ολοκληρώσει τη διαδικασία, οι αντίστοιχες καταχωρήσεις στην βάση δεδομένων ενημερώνονται με την προσθήκη του πεδίου της ετικέτας τους.

Εικόνα 6. Το εργαλείο Tweet Annotator

Για τους σκοπούς της παρούσας Εργασίας συνολικά έγινε ετικετοποίηση σε 3931 tweets. Από τα tweets αυτά 1906 ανήκουν στην κατηγορία «αληθή» (real), 1017 στην κατηγορία «ψευδή» (fake) και 1018 στην κατηγορία «άσχετα» (irrelevant). Η κατανομή των χαρακτηρισμένων αυτών tweets φαίνεται στην Εικόνα 7.



Εικόνα 7. Κατανομή των tweets που έχουν χαρακτηριστεί ως fake, real και irrelevant.

4.2 Διερευνητική ανάλυση των δεδομένων

Για την κατανόηση των διαθέσιμων δεδομένων, τον έλεγχο της εγκυρότητας τους και την εξαγωγή πρωταρχικών συμπερασμάτων σχετικά με τις κατηγορίες που αναλύθηκαν προηγουμένως, γίνεται διερευνητική ανάλυση των διαθέσιμων δεδομένων. Πιο συγκεκριμένα, ελέγχεται αν υπάρχουν ελλιπή πεδία ή πεδία που λόγω σφάλματος έχουν απολεσθεί. Επιπλέον, εξάγονται περιγραφικά γραφήματα τα οποία απεικονίζουν διάφορα στατιστικά όπως για κάθε κατηγορία τον αριθμό των λέξεων που περιέχουν τα tweets καθώς και το μήκος σε χαρακτήρες.

4.3 Εξαγωγή των χαρακτηριστικών

Για την εξαγωγή των χαρακτηριστικών αρχικά ανακτώνται από την βάση δεδομένων mongoDB τα tweets εκείνα τα οποία έχουν κατηγοριοποιηθεί και έπειτα μετατρέπονται σε ένα Pandas DataFrame ώστε να γίνει πιο εύκολη η διαχείριση τους. Τα πεδία που επιλέχθηκαν να ανακτηθούν από τις πληροφορίες των tweet είναι το πλήρες κείμενό τους, η ημερομηνία δημοσίευσής τους και η κατηγορία που ανήκουν. Μέσα από μία δομή επανάληψης η οποία διατρέχει όλο το περιεχόμενο του DataFrame πραγματοποιούνται, μέσω κατάλληλων συναρτήσεων, διάφορες τεχνικές προεπεξεργασίας των κειμένων ώστε να εξαχθούν τα χαρακτηριστικά αποτελεσματικά. Η εξαγωγή όλων των χαρακτηριστικών γίνεται μέσω ειδικών συναρτήσεων που έχουν υλοποιηθεί για αυτό το σκοπό, σε συνδυασμό με συναρτήσεις από τις βιβλιοθήκες spaCy και NLTK σε όποιες περιπτώσεις αυτό είναι απαραίτητο.

4.3.1 Καθαρισμός κειμένων και εξαγωγή μορφολογικών χαρακτηριστικών

Υλοποιήθηκαν δύο βασικές συναρτήσεις με τις οποίες γίνεται ο καθαρισμός των κειμένων. Ο καθαρισμός των κειμένων ορίζεται ως η αποκοπή των URLs, των emojis, των οντοτήτων (mentions), και των hashtags από το περιεχόμενό τους. Η υλοποίηση αυτού έγινε με τη βοήθεια συνάρτησης που υλοποιήθηκε η οποία αποκόπτει τα παραπάνω σύμφωνα με συγκεκριμένα regex τα οποία καλύπτουν τα κριτήρια με τα οποία εμφανίζονται. Αυτό που επιστρέφει αυτή η συνάρτηση είναι το περιεχόμενο των κειμένων των tweet χωρίς τα τμήματα που έχουν αποκοπεί με την ονοματολογία clean_text, κάτι που είναι πολύ χρήσιμο για την εξαγωγή διάφορων χαρακτηριστικών καθώς σε μερικές περιπτώσεις είναι απαραίτητο τα κείμενα να είναι σε αυτή την μορφή.

Η επόμενη συνάρτηση που υλοποιήθηκε για τον καθαρισμό των κειμένων μετατρέπει τα κείμενα σε λίστες από tokens. Προτού γίνει αυτό, ακολουθούνται κάποια βήματα ώστε τα tokens να έχουν την επιθυμητή μορφή. Αυτά είναι τα ακόλουθα:

- Διαγραφή των emojis από το κείμενο το tweets
- Διαγραφή όλων των ψηφίων από το κείμενο και αντικατάστασή τους με το κενό
- Διαγραφή των σημείων στίξης
- Διαγραφή των λέξεων κλειδιών (μέσω συναρτήσεων της spaCy και της NLTK)
- Διαγραφή tokens μικρότερων των 3 χαρακτήρων

Οι λίστες των λέξεων που επιστρέφονται από την συνάρτηση αυτή περιέχουν το περιεχόμενο του κειμένου απαλλαγμένο από ότι διαγράφηκε κατά την εκτέλεσή της με την ονοματολογία words.

Το clean_text που επιστρέφεται από την παραπάνω συνάρτηση χρησιμοποιείται ως είσοδο στις συναρτήσεις οι οποίες εξάγουν τα περισσότερα από τα μορφολογικά χαρακτηριστικά. Για την ακρίβεια η εξαγωγή των χαρακτηριστικών που αφορούν , το μήκος του κειμένου, την καταμέτρηση των διαφόρων σημείων στίξης (“?!”, ”?”, ”!”, “.”, “;” κλπ.) αλλά και του συνολικού αριθμού τους και τέλος το tweet_entropy γίνεται μέσω συναρτήσεων υλοποιημένων με “απλή” Python σε συνδυασμό με τη βιβλιοθήκη *re* η οποία προσφέρει έναν εύκολο τρόπο για την αναγνώριση regex εντός του κειμένου. Το tweet_entropy εξάγεται από συνάρτηση που υλοποιήθηκε που κάνει τον μαθηματικό υπολογισμό για την εντροπία του κειμένου.

Υπάρχουν δύο συναρτήσεις ακόμη οι οποίες παίρνουν ως είσοδο το clean_text. Η πρώτη υπολογίζει τον αριθμό των συμφώνων αλλά και των φωνηέντων που υπάρχουν σε ένα tweet και συνεπώς μέσα από αυτή υπολογίζονται και τα αντίστοιχα χαρακτηριστικά. Αρχικά αφαιρούνται από το clean_text οποιαδήποτε κατάληξη και τόνος των λέξεων και έπειτα κάθε χαρακτήρας κατηγοριοποιείται ανάλογα με το αν είναι σύμφωνο ή φωνήεν. Ως αποτέλεσμα η καταμέτρηση τους είναι εύκολα υπολογίσιμη. Η δεύτερη συνάρτηση επιστρέφει με χρήση “απλής” Python τον αριθμό των κεφαλαίων, των πεζών, των ψηφίων, των γραμμάτων αλλά και την αναλογία γραμμάτων ως προς ψηφίων τα οποία αντιστοιχούν και στα ανάλογα χαρακτηριστικά.

Το χαρακτηριστικό που αντιστοιχίζεται με τον μέσο όρο λέξεων ανά πρόταση προέρχεται από μία συνάρτηση η οποία με όρισμα το clean_text αρχικά μετατρέπει σε tokens το περιεχόμενο του κειμένου με τη βοήθεια της βιβλιοθήκης NLTK η οποία υποστηρίζει την λειτουργία αυτή και στην ελληνική γλώσσα. Έπειτα αφαιρούνται τα σημεία στίξης και από εκεί και πέρα με έναν απλό υπολογισμό υπολογίζεται ο μέσος όρος αυτός.

Τα τελευταία χαρακτηριστικά που χρησιμοποιούν το clean_text για να παραχθούν έχουν να κάνουν με τον αριθμό των συνεχόμενων συμφώνων, τον αριθμό των συνεχόμενων φωνηέντων και

τον αριθμό από εμφανίσεις επαναλαμβανομένων ίδιων χαρακτήρων. Τα τρία αυτά χαρακτηριστικά υπολογίζονται μέσα από μία συνάρτηση η οποία έχει παρόμοια λογική με την συνάρτηση που υπολογίζει τον συνολικό αριθμό των συμφώνων και των φωνηέντων που περιγράφηκε παραπάνω. Ο υπολογισμός των δύο πρώτων χαρακτηριστικών υπολογίζεται άμεσα από λίστες χαρακτήρων (συμφώνων, φωνηέντων) που παράγονται κατά την εκτέλεση ενώ ο αριθμός συνεχόμενων εμφανίσεων ενός χαρακτήρα (>3) από λίστα που περιέχει όλους τους χαρακτήρες ενός κειμένου.

Η συνάρτηση που επιστρέφει τα words είναι χρήσιμη στην εξαγωγή δύο χαρακτηριστικών. Το πρώτο χαρακτηριστικό αφορά το μέσο μήκος των λέξεων το οποίο υπολογίζεται από μία συνάρτηση η οποία σαν όρισμα παίρνει το words. Το πρώτο βήμα που εκτελεί αυτή η συνάρτηση είναι να καλέσει την συνάρτηση που παράγει τα clean_text δίνοντας ως όρισμα την λίστα των words. Έπειτα ο υπολογισμός του μέσου όρου γίνεται πάλι μέσω ενός απλού μαθηματικού υπολογισμού. Το δεύτερο χαρακτηριστικό αφορά τον αριθμό των λέξεων υπάρχουν μέσα σε ένα κείμενο το οποίο είναι αρκετά εύκολο να υπολογιστεί μέσω του μήκους της λίστας words.

Τέλος, υπάρχουν χαρακτηριστικά τα οποία μπορούν να υπολογιστούν από το αρχικό κείμενο των tweets χωρίς να υποστεί κάποια επεξεργασία. Για παράδειγμα τα χαρακτηριστικά που αφορούν τον αριθμό των urls, mentions και hashtags παράγονται μέσω συναρτήσεων που ελέγχουν το περιεχόμενο για λέξεις που αρχίζουν με “http{.....}”, “@” και “#” αντίστοιχα, τα οποία και αριθμούν. Το χαρακτηριστικό που έχει να κάνει με τον αριθμό των οντοτήτων υπολογίζεται από το άθροισμα των παραπάνω. Σημαντικό να ειπωθεί είναι το γεγονός ότι πραγματοποιείται έλεγχος για το αν ένα url είναι λειτουργικό ή όχι, έτσι στην καταμέτρησή τους λαμβάνονται υπόψη μόνο τα λειτουργικά εξ αυτών. Οι λέξεις-κλειδιά είναι ένα από τα χαρακτηριστικά που υπολογίζονται μέσω συνάρτησης που στο όρισμά της έχει και πάλι το αρχικό κείμενο ενός tweet. Η συνάρτηση αυτή επιστρέφει των αριθμό των λέξεων-κλειδιών που περιέχονται σε ένα tweet με τη βοήθεια των συναρτήσεων που τις εξάγουν των βιβλιοθηκών spaCy και NLTK. Αν και αυτές οι συναρτήσεις αναγνωρίζουν τα περισσότερα από τις ελληνικές λέξεις-κλειδιά (επί, αν, από, θα), προστέθηκαν κάποια τα οποία δεν αναγνωρίζουν οι συγκεκριμένες συναρτήσεις (στις, στους κλπ.).

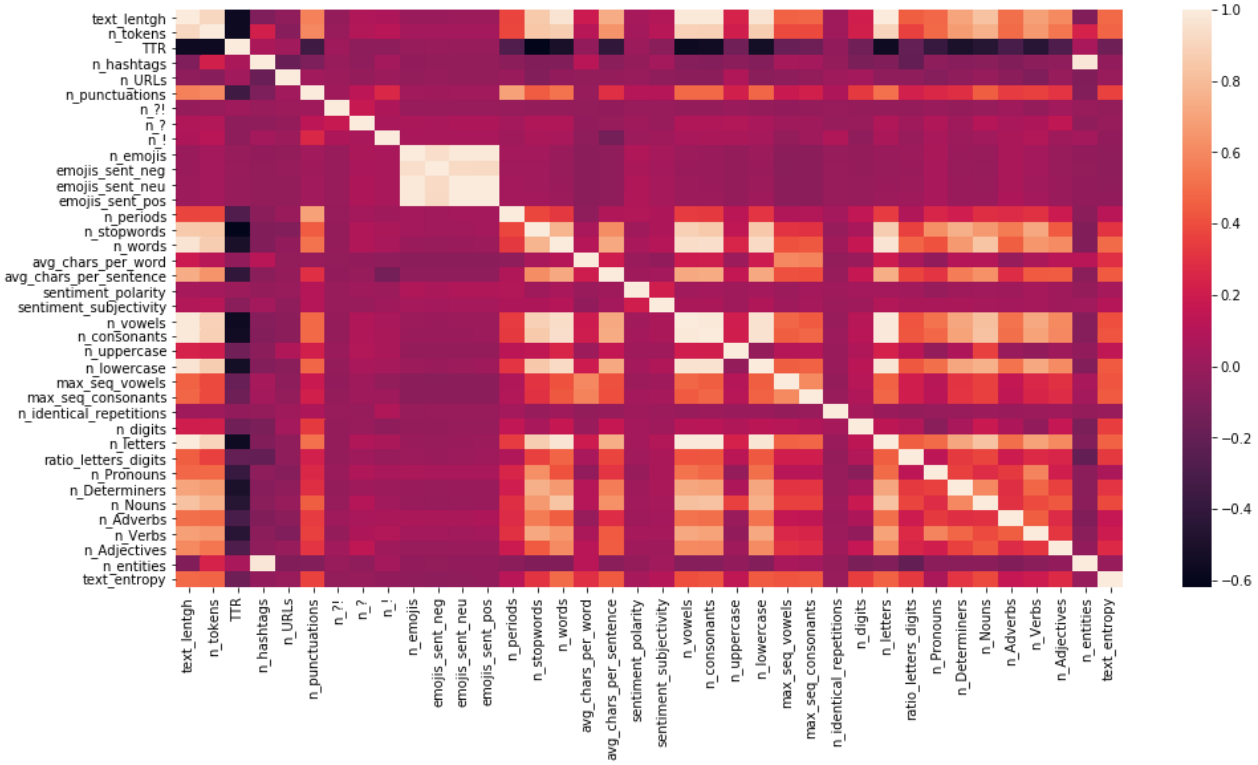
4.3.2 Εξαγωγή σημασιολογικών χαρακτηριστικών

Η εξαγωγή σημασιολογικών χαρακτηριστικών είναι ένα δύσκολο έργο για να υλοποιηθεί. Καθώς όμως ο τομέας συνεχώς εξελίσσεται, έχουν δημιουργηθεί βιβλιοθήκες οι οποίες μπορούν να εξάγουν τέτοια χαρακτηριστικά μερικώς αυτοματοποιημένα. Για τους σκοπούς της Διπλωματικής Εργασίας αυτής χρησιμοποιήθηκε η βιβλιοθήκη της spaCy μέσω της οποίας, με δύο απλές εντολές μπορεί να εξαχθεί η αντικειμενικότητα ενός κειμένου αλλά και η πόλωσή του, συνεπώς υπολογίζονται άμεσα και τα ανάλογα χαρακτηριστικά. Επίσης μία επιτυχής μετρική, είναι και η καταμέτρηση των θετικών, αρνητικών και ουδέτερων emoji που περιέχει ένα κείμενο. Ο υπολογισμός των χαρακτηριστικών αυτών γίνεται σε δύο στάδια. Στο πρώτο στάδιο αναζητούνται μέσω συνάρτησης όλα τα emoticons που περιέχονται σε ένα tweet και έπειτα σε δεύτερο στάδιο κατηγοριοποιούνται σε αρνητικά, θετικά και ουδέτερα, και καταμετρώνται. Τα προηγούμενα υλοποιούνται μέσω regex matching, έναν πολύ συνηθισμένο τρόπο για την εύρεση emoji καθώς επίσης η κατηγοριοποίηση γίνεται με βάση έναν κατάλογο από emojis που περιέχει όλα τα πιθανά emojis αλλά και την αξιολόγησή τους και ο οποίος παρέχεται από τη βιβλιοθήκη emosent [50].

4.3.3 Εξαγωγή PoS χαρακτηριστικών

Ο υπολογισμός των PoS χαρακτηριστικών γίνεται μέσω της συνάρτησης nlp που υλοποιήθηκε με την βοήθεια ειδικών της βιβλιοθήκης spaCy, η οποία περικλείει μία ακολουθία μεθόδων για τον υπολογισμό τέτοιων χαρακτηριστικών. Με τη χρήση ειδικού πακέτου της βιβλιοθήκης ([51]), η εξαγωγή χαρακτηριστικών είναι ειδικά προσαρμοσμένη για την ελληνική γλώσσα και γραμματική. Αρχικά χρησιμοποιείται το LexicalRichness() μέσω του οποίου υπολογίζεται μία μετρική η οποία αναπαριστά την "πλουσιότητα" του κειμένου όσο αναφορά την ποικιλία των λέξεων που χρησιμοποιούνται. Μέσω αυτής της μετρικής μπορεί να χρησιμοποιηθεί άλλη μία συνάρτηση του LexicalRichness η οποία υπολογίζει το TTR του κειμένου το οποίο αντιστοιχεί και στο ανάλογο χαρακτηριστικό. Έπειτα, δημιουργούνται τρεις λίστες που περιέχουν tokens στα οποία χωρίζεται το κείμενο, τα PoS κατηγοριοποιημένα που βρέθηκαν και τα lemmas (λήμματα) που παρήχθησαν αντίστοιχα. Από την λίστα με τα κατηγοριοποιημένα PoS που βρέθηκαν, υπολογίζονται τα χαρακτηριστικά που αφορούν τον αριθμό των εμφανίσεων των αντωνυμιών, προσδιορισμών (determiners), επιθέτων, ουσιαστικών, επιρρημάτων και των οντοτήτων (entities) που υπάρχουν σε ένα tweet. Από την ίδια διαδικασία εξάγονται η αντικειμενικότητα και η πόλωση του κειμένου που εξηγήθηκαν στο υποκεφάλαιο 4.3.2.

Το σύνολο των χαρακτηριστικών που εξήχθησαν καθώς και ο βαθμός της συσχέτισης μεταξύ τους απεικονίζεται στην Εικόνα 8.



Εικόνα 8. Μητρώο συσχέτισης (correlation matrix) των βασικών χαρακτηριστικών που εξάγονται από τα κείμενα των διαθέσιμων tweets.

4.3.4 Υπολογισμός των TF-IDF χαρακτηριστικών

Ακολούθως, γίνεται εφαρμογή της μεθόδου TF-IDF με χρήση της συνάρτησης TfidffVectorizer της scikit-learn βιβλιοθήκης. Για την δημιουργία των vectors των κειμένων που χρειάζονται για την εφαρμογή της TF-IDF χρησιμοποιείται μία συνάρτηση που ενώνει τα λήμματα που παρήχθησαν από την διαδικασία εξαγωγής των PoS χαρακτηριστικών. Αυτό γίνεται ώστε ο αριθμός των διαφορετικών λέξεων, και συνεπώς των διαστάσεων των χαρακτηριστικών, να μειωθεί αισθητά καθώς η συνάρτηση TfidffVectorizer παράγει χαρακτηριστικά για κάθε ξεχωριστή λέξη που εντοπίζει μέσα στα κείμενο. Εξάγονται έτσι ένα σύνολο από χαρακτηριστικά τα οποία αντιστοιχούν στις σημαντικότερες λέξεις του συνόλου δεδομένων. Τα χαρακτηριστικά αυτά συνδυάζονται με τα προηγούμενα μορφολογικά, PoS και σημασιολογικά χαρκατηριστικά, σε ένα ενιαίο μητρώο χαρακτηριστικών το οποίο έχει διαστάσεις (3931 x 10923), δηλαδή υπολογίζονται συνολικά 10923 χαρακτηριστικά.

4.4 Εκπαίδευση αλγορίθμων

Πριν την εκπαίδευση των αλγορίθμων, τα συλλεχθέντα χαρακτηριστικά μετασχηματίζονται ώστε να έχουν όλα το ίδιο εύρος τιμών. Για την κανονικοποίηση χρησιμοποιείται η κλάση StandardScaler της scikit-learn.

Αφού τα χαρακτηριστικά μετασχηματιστούν, στη συνέχεια εφαρμόζεται η μέθοδος PCA για τη μείωση της διαστατικότητας του μητρώου των χαρακτηριστικών, κάτι το οποίο κρίνεται απαραίτητο λόγω του πλήθους αυτών. Η μέθοδος παραμετροποιείται για να εξαγάγει τον ελάχιστο αριθμό διαστάσεων που απαιτούνται για τη διατήρηση του 95% της διακύμανσης του συνόλου δεδομένων που περιέχονται στο μητρώο χαρακτηριστικών. Τα αποτελέσματα δείχνουν ότι αντί για 10923, χρειάζονται μόνο 2745 βασικά χαρακτηριστικά. Προκύπτει επίσης ότι, τα δεδομένα αυτά αντιστοιχούν σχεδόν στο 25% του μεγέθους του αρχικού μητρώου. Κατά κάποιο τρόπο, αυτό είναι ένα είδος επιλογής χαρακτηριστικών (*Feature Selection*). Προφανώς, μετά τη μείωση των διαστάσεων, το μητρώο χαρακτηριστικών καταλαμβάνει πολύ λιγότερο χώρο και αυτό μπορεί να επιταχύνει σημαντικά έναν αλγόριθμο ταξινόμησης (όπως η ταξινόμηση που βασίζεται στον αλγόριθμο SVM).

Τέλος τα δεδομένα χωρίζονται σε σύνολο εκπαίδευσης και σύνολο δοκιμών. Τα δεδομένα εκπαίδευσης θα χρησιμοποιηθούν για την εκπαίδευση του αλγόριθμου μάθησης και τα δεδομένα δοκιμών για την επαλήθευση των αποτελεσμάτων. Για τον διαχωρισμό, 80% του συνόλου των διαθέσιμων δεδομένων χρησιμοποιούνται ως δεδομένα εκπαίδευσης και το 20% ως δεδομένα επαλήθευσης.

Ο αλγόριθμος SVM είναι ο πρώτος αλγόριθμος που δοκιμάζεται ως προς τη δυνατότητα ταξινόμησης των tweets με βάση τις κλάσεις που έχουν οριστεί. Αρχικά τίθενται οι παράμετροι του αλγόριθμου οι οποίες είναι οι εξής:

- Kernel: χρησιμοποιείται το “rbf” kernel.
- C: 1
- Gamma: Scale. Αυτό σημαίνει ότι το gamma προκύπτει από το ratio 1/ (<αριθμός Features> x <Διασπορά μητρώου εκπαίδευσης>)

Στη συνέχεια εφαρμόζεται *βελτιστοποίηση υπερπαραμέτρων (hyperparameter tuning)* με τη μέθοδο του Grid Search [52]. Στην περίπτωση αυτή δοκιμάζονται επιπλέον ως kernel το "linear", καθώς και διαφορετικοί συνδυασμοί των τιμών του C και του gamma.

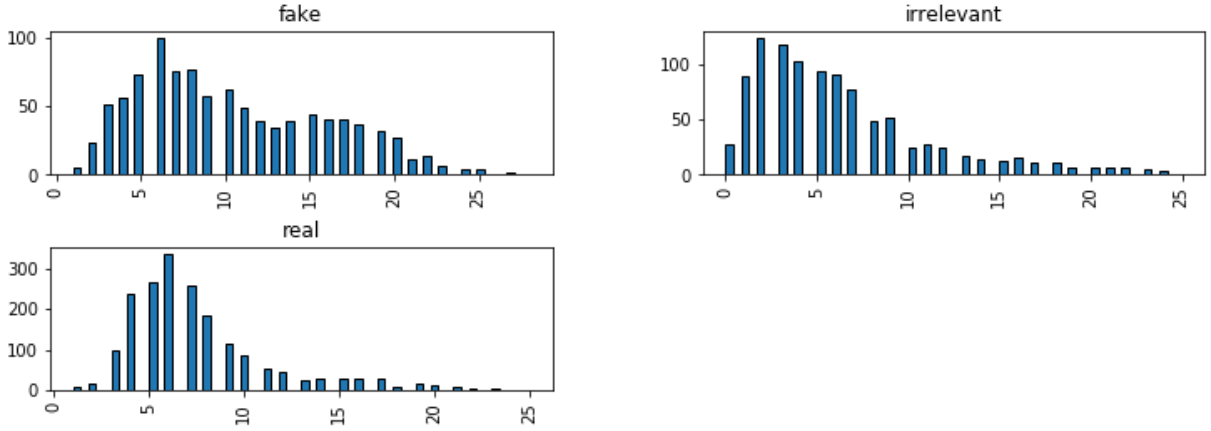
Επόμενος αλγόριθμος που δοκιμάζεται είναι ο Random Forest. Για τον Random Forest ακολουθείται παρόμοια διαδικασία. Αρχικά γίνεται εκπαίδευση με τις αρχικές παραμέτρους εκπαίδευσης. Στη συνέχεια με τη χρήση της συνάρτησης RandomizedSearchCV [53] δημιουργείται ένα σύνολο συνδυασμών των παραμέτρων, από τους οποίους ο αλγόριθμος επιλέγει με τυχαία ποιους θα εκπαιδευτεί.

Τέλος, γίνεται εκπαίδευση με χρήση του Multinomial Naive Bayes. Καθώς ο αλγόριθμος αυτός στηρίζεται στη χρήση πιθανοτήτων δεν μπορεί να δεχτεί αρνητικές τιμές. Επομένως στην περίπτωση αυτή γίνεται κανονικοποίηση των δεδομένων στο διάστημα 0 έως 1 με χρήση της κλάσης MinMaxScaler.

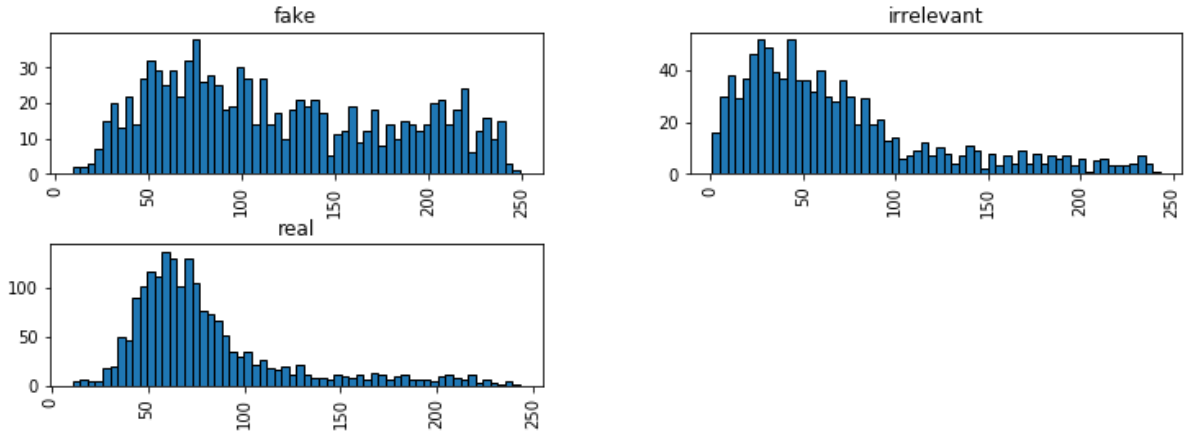
5

Αποτελέσματα και Συμπεράσματα

Στο παρόν κεφάλαιο γίνεται παράθεση των αποτελεσμάτων της διερευνητικής ανάλυσης καθώς και των αλγορίθμων Μηχανικής Μάθησης που εφαρμόστηκαν. Αρχικά, στα πλαίσια της διερευνητικής ανάλυσης γίνεται οπτικοποίηση κάποιων βασικών στατιστικών που χαρακτηρίζουν τα tweets.



Εικόνα 10. Πλήθος λέξεων σε ένα tweet ανάλογα με την κατηγορία που ανήκει.



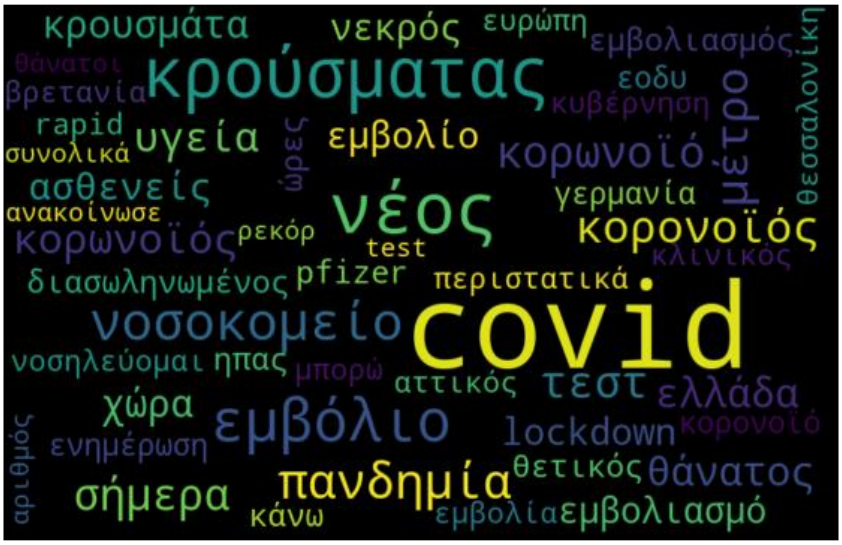
Εικόνα 9. Πλήθος χαρακτήρων σε ένα tweet ανάλογα με την κατηγορία που ανήκει.

Όπως προκύπτει από τις Εικόνες 9, 10 φαίνεται να υπάρχει κάποια συσχέτιση του μήκους του tweet σε λέξεις και αντίστοιχα του πλήθους χαρακτήρων του με την κατηγορία «ψευδή», δηλαδή συγκριτικά με τις άλλες δύο κατηγορίες, πολλά tweets έχουν μεγαλύτερο αριθμό λέξεων και χαρακτήρων.

Επιπλέον γίνεται οπτικοποίηση των πιο σημαντικών όρων σε νέφη λέξεων. Πιο συγκεκριμένα για κάθε κατηγορία εντοπίζονται και οπτικοποιούνται οι 50 πιο συχνές λέξεις που συναντώνται. Τα αποτελέσματα αυτά οπτικοποιούνται στις Εικόνες 12, 13, 14. Εκτός από τις πιο συχνές λέξεις ανά κατηγορία, γίνεται παρουσίαση και των πιο συχνών λέξεων στο σύνολο των δεδομένων, οι οποίες απεικονίζονται στην Εικόνα 11.

Εικόνα 11. Νέφος λέξεων για ολόκληρο το σύνολο δεδομένων.

Στο νέφος λέξεων που αφορά το σύνολο των συλλεχθέντων δεδομένων κυριαρχούν οι λέξεις οι οποίες είναι άμεσα συνδεδεμένες με την επικαιρότητα του διαστήματος από το οποίο επιλέχθηκαν οι δημοσιεύσεις. Όπως προκύπτει από τη μεγάλη συχνότητα την οποία έχουν οι λέξεις «εμβόλιο», «εμβολιασμός» και η συχνότητα εμφάνισης της εταιρίας που παρέχει το σχετικό εμβόλιο, το θέμα των εμβολιασμών κυριάρχησε στις διαδικτυακές συζητήσεις. Η λέξη που εμφανίζεται περισσότερο είναι «covid» κάτι που είναι αναμενόμενο καθώς αποτελεί και τη βασική λέξη-κλειδί με βάση την οποία δημιουργήθηκε το σύνολο δεδομένων αυτό.



Εικόνα 12. Νέφος λέξεων για την κατηγορία real.

Εικόνα 13. Νέφος λέξεων για την κατηγορία fake.

Εικόνα 14. Νέφος λέξεων για την κατηγορία irrelevant.

Ενδιαφέροντα συμπεράσματα προκύπτουν από την ανάλυση των νεφών λέξεων σε κάθε κατηγορία. Στην κατηγορία των αληθών δημοσιεύσεων κυριαρχούν λέξεις όπως «ενημέρωση», «κρούσματα», «περιστατικά» και άλλες οι οποίες συναντώνται συχνά σε επίσημα δελτία τύπου ενημέρωσης για την πορεία της πανδημίας.

Στην κατηγορία ψευδών δημοσιεύσεων, αν και πολλές λέξεις εμφανίζονται κοινές, φαίνεται να προκύπτουν με μεγαλύτερη συχνότητα, όπως η λέξη «εμβόλιο». Επιπλέον, με μεγάλη συχνότητα εμφανίζονται η λέξη «εκκλησία» και «πιστεύω», που δεν εμφανίζονται σε άλλο νέφος λέξεων.

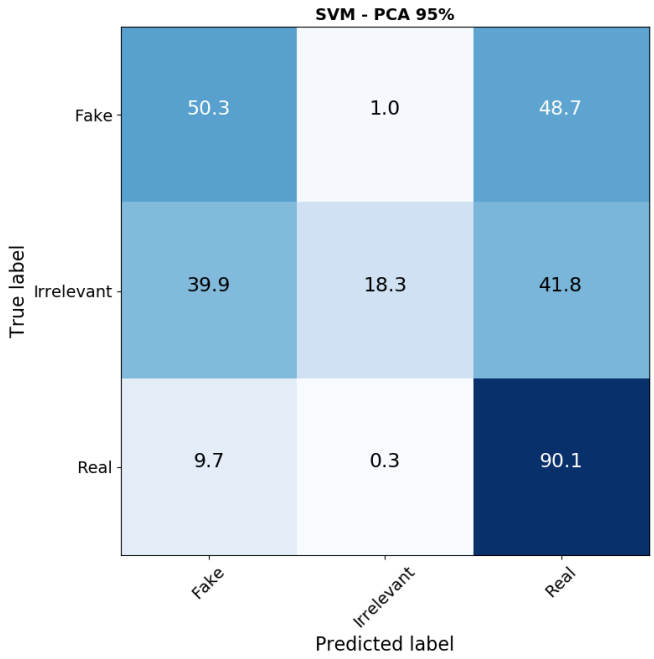
Τέλος, στην κατηγορία μη σχετικών δημοσιεύσεων, οι πιο συχνά εμφανιζόμενες λέξεις διαφοροποιούνται καθώς εμφανίζονται πιο συχνά λέξεις οι οποίες μπορούν να χρησιμοποιηθούν σε ένα πιο ευρύ φάσμα και να μην αφορούν αποκλειστικά την πανδημία, όπως η έντονη χρήση ρημάτων («δείτε», «διαβάστε», «μπορώ»). Σε κάθε περίπτωση, από την χρήση νεφών λέξεων μπορούν να εξαχθούν κάποια πρωταρχικά συμπεράσματα, ωστόσο χωρίς το νόημα των φράσεων, τη σειρά των λέξεων και το πώς αυτές συνδέονται μεταξύ τους, δεν μπορεί να προκύψει επαρκής πληροφορία για τη λήψη αποφάσεων.

Στη συνέχεια παρουσιάζονται και αναλύονται τα αποτελέσματα των αλγορίθμων που δοκιμάστηκαν για την κατηγοριοποίηση των tweets. Όπως περιγράφεται και στο προηγούμενο κεφάλαιο η εκπαίδευση του συνόλου των αλγορίθμων γίνεται με τη χρήση των χαρακτηριστικών εκείνων που αντιστοιχούν στο 95% του συνόλου δεδομένων. Στον Πίνακα 3 παρατίθενται οι μετρικές αξιολόγησης του αλγόριθμου SVM. Σύμφωνα με τα αποτελέσματα, ο αλγόριθμος είναι ικανός να διακρίνει με ικανοποιητική επιτυχία αν μία δημοσίευση μπορεί να χαρακτηριστεί ως αληθής, ωστόσο υπάρχει μεγάλη ασάφεια μεταξύ των δύο άλλων κατηγοριών, κάτι που φαίνεται και από το πολύ χαμηλό Recall.

	Precision	*Recall*	*F1-score*
fake	0.45	0.50	0.48
Irrelevant	0.93	0.18	0.31
real	0.65	0.90	0.76
Accuracy			0.61
Macro avg	0.68	0.53	0.51
Weighted avg	0.67	0.61	0.57

Πίνακας 3. Αποτελέσματα ταξινόμησης για τον αλγόριθμο SVM.

Στην Εικόνα 15 απεικονίζεται το ποσοστό της ταξινόμησης των tweets του συνόλου δοκιμών σε κάθε μία από τις κατηγορίες. Βάση των αποτελεσμάτων του, τα αληθή tweets αναγνωρίζονται με ικανοποιητικό ποσοστό, ωστόσο οι δύο άλλες κατηγορίες χαρακτηρίζονται σωστά σε μόλις 50.3 % και 18.3 % των περιπτώσεων για τις κατηγορίες των ψευδών και άσχετων δημοσιεύσεων αντίστοιχα. Στην περίπτωση της κατηγορίας «ψευδή», μεγάλο ποσοστό (48.7%) αναγνωρίζονται ως «αληθή». Η κατηγορία «άσχετα» ωστόσο δεν μπορεί να διακριθεί από τις άλλες δύο καθώς τα tweets αναγνωρίζονται είτε ως «ψευδή» ή ως «αληθή» και ένα μικρό ποσοστό (18.3%) ταξινομείται σωστά.



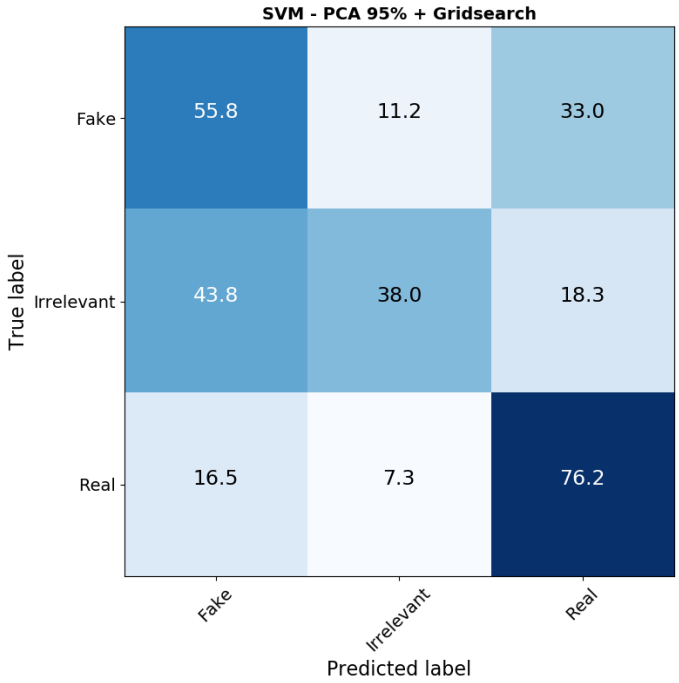
Εικόνα 15. Confusion matrix από την εκτέλεση του αλγόριθμου SVM στα χαρακτηριστικά που αντιπροσωπεύουν το 95% της διασποράς του συνόλου δεδομένων.

Στη συνέχεια εφαρμόζεται η μέθοδος Grid Search στις παραμέτρους του SVM με σκοπό την βελτίωση της απόδοσης. Η εφαρμογή του βελτιώνει το ποσοστό αναγνώρισης για τις κατηγορίες «ψευδή» και «άσχετα», ωστόσο αυτό παραμένει αρκετά χαμηλό. Παρατηρείται και πάλι ότι κείμενα της κατηγορίας «ψευδή» ταξινομούνται ως «αληθή». Υπήρξε όμως βελτίωση του ποσοστού ορθής ταξινόμησης της κατηγορίας «άσχετα», ωστόσο και πάλι αυτό παραμένει χαμηλό, καθώς tweets αυτής της κατηγορίας αναγνωρίζονται κυρίως ως «ψευδή». Οι παράμετροι του SVM οι οποίες τέθηκαν μέσω της μεθόδου Grid Search και με τις οποίες επιτυγχάνονται τα αποτελέσματα αυτά είναι:

- C: 10
- gamma: 0.0001
- kernel: rbf

	Precision	*Recall*	*F1-score*
fake	0.42	0.56	0.48
Irrelevant	0.61	0.38	0.47
Real	0.74	0.76	0.75
Accuracy			0.61
Macro avg	0.59	0.57	0.57
Weighted avg	0.62	0.61	0.61

Πίνακας 4. Αποτελέσματα ταξινόμησης για τον αλγόριθμο SVM και εφαρμογή της μεθόδου Grid Search.



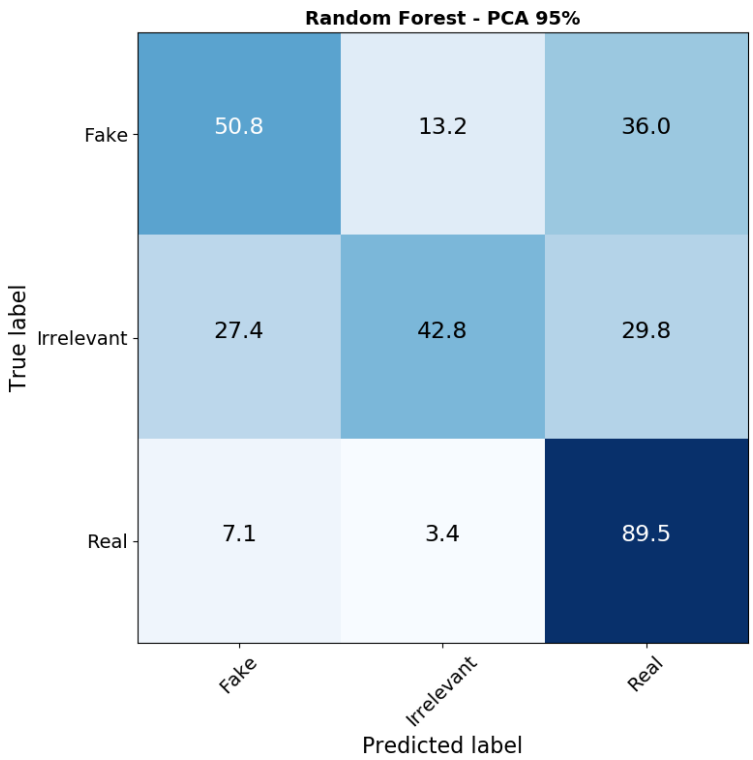
Εικόνα 16. Confusion Matrix από την εκτέλεση του αλγόριθμου SVM στα χαρακτηριστικά που αντιπροσωπεύουν το 95% και στη συνέχεια εφαρμογή της μεθόδου.

Ο αλγόριθμος που δοκιμάζεται στη συνέχεια είναι ο Random Forest. Παρόμοια με τον SVM πετυχαίνει υψηλό ποσοστό αναγνώρισης στην αναγνώριση των tweets της κατηγορίας «αληθή» ωστόσο δεν γίνεται σωστή διάκριση των δύο άλλων κατηγοριών. Και σε αυτή την περίπτωση

μεγάλο μέρος των ψευδών tweets χαρακτηρίζονται ως «αληθή», ενώ τα άσχετα tweets φαίνεται να μην μπορούν να διακριθούν εύκολα από τις άλλες δύο κατηγορίες. Τα αποτελέσματα του αλγορίθμου και το confusion matrix απεικονίζονται στον Πίνακα 5 και την Εικόνα 17 αντίστοιχα.

	Precision	*Recall*	*F1-score*
fake	0.54	0.51	0.52
Irrelevant	0.70	0.43	0.53
Real	0.72	0.90	0.85
Accuracy			0.67
Macro avg	0.65	0.61	0.62
Weighted avg	0.62	0.67	0.66

Πίνακας 5. Αποτελέσματα ταξινόμησης για τον αλγόριθμο Random Forest.



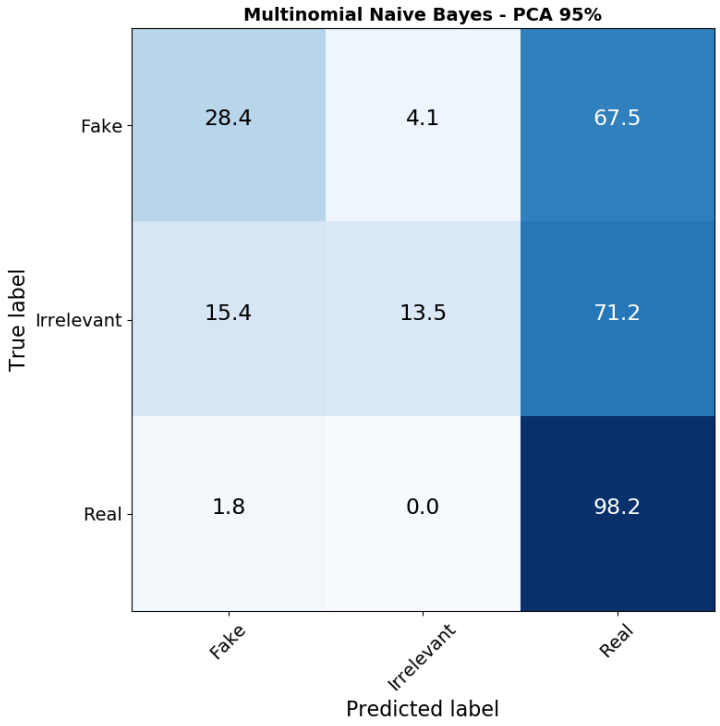
Εικόνα 17. Confusion matrix από την εκτέλεση του αλγόριθμου Random Forest στα χαρακτηριστικά που αντιπροσωπεύουν το 95% της διακύμανσης του συνόλου δεδομένων.

Στα πλαίσια βελτιστοποίησης των παραμέτρων του αλγορίθμου Random Forest έγινε προσπάθεια εφαρμογής της μεθόδου Randomized Search. Ωστόσο δεν υπήρξε κάποια εμφανής βελτίωση στα αποτελέσματα.

Τέλος δοκιμάζεται η εφαρμογή του αλγόριθμου Multinomial Naive Bayes. Ωστόσο, όπως προκύπτει από τον Πίνακα 6 και την Εικόνα 19, η απόδοση του αλγόριθμου κρίνεται μη ικανοποιητική καθώς αναγνωρίζει στις περισσότερες περιπτώσεις τα tweets ως αληθή ανεξάρτητα της κατηγορίας τους.

	Precision	*Recall*	*F1-score*
fake	0.54	0.51	0.52
Irrelevant	0.70	0.43	0.53
Real	0.72	0.90	0.85
Accuracy			0.67
Macro avg	0.65	0.61	0.62
Weighted avg	0.62	0.67	0.66

Πίνακας 6. Αποτελέσματα ταξινόμησης για τον αλγόριθμο Multinomial Naive Bayes.



Εικόνα 18. Confusion matrix από την εκτέλεση του αλγόριθμου Multinomial Naive Bayes στα χαρακτηριστικά που αντιπροσωπεύουν το 95% *της* διακύμανσης του συνόλου δεδομένων.

Στον Πίνακα 7 παρατίθενται τα συγκεντρωτικά αποτελέσματα για τους αλγόριθμους που δοκιμάστηκαν.

Αλγόριθμος	**Precision**	**Recall**	**F1-Score**
SVM	0.68	0.53	0.51
SVM + Grid Search	0.59	0.57	0.57
Random Forest	0.65	0.61	0.62
Multinomial Naive Bayes	0.65	0.47	0.45

Πίνακας 7. Συγκεντρωτικά αποτελέσματα των αλγορίθμων που δοκιμάστηκαν.

Όπως προκύπτει από τα αποτελέσματα των αλγορίθμων, αν και με χαμηλό ποσοστό διακρίνεται η δυνατότητα αναγνώρισης ενός tweet ως αληθές, ψευδές ή άσχετο βάση των χαρακτηριστικών που περιγράφονται στο προηγούμενο κεφάλαιο. Η κατηγορία «αληθή» είναι αυτή η οποία φαίνεται να ξεχωρίζει καλύτερα σε σχέση με τις υπόλοιπες. Η κατηγορία «ψευδή» συγχέεται πολλές φορές από τον αλγόριθμο με την κατηγορία «αληθή». Αυτό εξηγείται εύκολα καθώς ταξινόμηση ενός tweet σε ψευδές ή αληθές εξαρτάται άμεσα από τη γνώση των γεγονότων, καθώς μορφολογικά τα κείμενα μπορεί να είναι ίδια. Τέλος, η κατηγορία «άσχετα» είναι αυτή η

οποία συγχέεται κατά βάση με τις άλλες κατηγορίες καθώς στο 60% των περιπτώσεων ταξινομείται είτε ως «ψευδή» ή ως «αληθής». Τα ποσοστά αυτά δείχνουν επίσης ότι η διαδικασία ετικετοποίησης που ακολουθήθηκε μπορεί να μην είναι επαρκής για την εξαγωγή αποτελεσμάτων από τους αλγόριθμους ή να απαιτεί μεγαλύτερο πλήθος από χαρακτηρισμένα δεδομένα, ή οι κατηγορίες που επιλέχθηκαν για ταξινόμηση να μην είναι αντιπροσωπευτικές των δημοσιεύσεων που έχουν συλλεχθεί. Επομένως, για την βελτίωση των αποτελεσμάτων, απαραίτητος είναι ο έλεγχος του τρόπου με τον οποίο χαρακτηρίζεται κάθε tweet αλλά και η αναθεώρηση των κατηγοριών.

Από τους αλγόριθμους που δοκιμάστηκαν καλύτερη απόδοση πετυχαίνει ο αλγόριθμος Random Forest καθώς συγκριτικά με τους άλλους αλγόριθμους διακρίνει καλύτερα τα tweets κάθε κατηγορίας. Παρόμοια επίδοση πετυχαίνει και ο αλγόριθμος SVM με χρήση της μεθόδου Grid Search. Από τα αποτελέσματα αυτά προκύπτει επίσης και η σημασία της βελτιστοποίησης και ορθής παραμετροποίησης των αλγορίθμων. Χειρότερη επίδοση είχε ο αλγόριθμος Multinomial Naive Bayes.

6

Επίλογος

Στην παρούσα Διπλωματική Εργασία, ο στόχος ήταν να προσεγγίσει υπολογιστικά τις δημοσιεύσεις στα κοινωνικό δίκτυο του Twitter, υποβοηθώντας το έργο του εντοπισμού των ψευδών δημοσιεύσεων οι οποίες σχετίζονται με την τρέχουσα πανδημία COVID-19 και τη διάδοσή της στην Ελλάδα. Επικεντρωθήκαμε στο Twitter καθώς η δομή και τα κύρια χαρακτηριστικά του επιτρέπουν την ανίχνευση τάσεων οι οποίες σχετίζονται με τα τρέχοντα συμβάντα. Ο σκοπός των δραστηριοτήτων ήταν η εύρεση μορφολογικών χαρακτηριστικών των tweets, καθώς και χαρακτηριστικών τα οποία σχετίζονται με την υποκειμενικότητα των κειμένων και τα οποία επιτρέπουν την αυτόματη διακριτοποίηση μεταξύ των ψευδών και αληθών δηλώσεων ώστε να κατηγοριοποιηθούν ανάλογα με την αξιοπιστία τους.

Για τους σκοπούς της εργασίας μας χρησιμοποιήθηκε ένα έτοιμο σύνολο δεδομένων το οποίο σχετίζεται με tweets τα οποία αφορούν την πανδημία. Για λόγους ευκολίας η περίοδος της μελέτης μας αφορά μόνο το διάστημα μεταξύ 1ης Νοεμβρίου 2020 και 31 Δεκεμβρίου 2020. Υλοποιήθηκαν δύο εργαλεία, ένα για την αυτόματη λήψη από το σύνολο δεδομένων των αναγνωριστικών των tweets τα οποία δημοσιεύτηκαν κατά τη συγκεκριμένη περίοδο και είναι γραμμένα στα Ελληνικά. Επιπλέον υλοποιείται ένα ακόμα εργαλείο το οποίο για κάθε ένα από τα συλλεχθέντα αναγνωριστικά ανακτά από την πλατφόρμα του Twitter το πλήρες περιεχόμενό τους και το αποθηκεύει σε μία βάση MongoDB.

Καθώς οι αλγόριθμοι ταξινόμησης της Μηχανικής Μάθησης στηρίζονται στη χρήση δεδομένων εκπαίδευσης για τα οποία η κατηγορία ταξινόμησης είναι γνωστή εκ των προτέρων, ένα κομμάτι της εργασίας αφορούσε την καταγραφή της κατηγορίας για τα συλλεχθέντα tweets. Συνολικά ταξινομήθηκαν χειροκίνητα 3931 tweets σε τρεις κατηγορίες, αληθή (real), ψευδή (fake) και άσχετα (irrelevant). Για την επιτάχυνση της διαδικασίας αυτής δημιουργήθηκε νέο εργαλείο το οποίο επιτρέπει την γρήγορη προβολή των tweets και την καταγραφή της ετικέτας.

Από τα δεδομένα, εφαρμόζοντας μεθόδους μετασχηματισμού αυτά εξήχθησαν 38 χαρακτηριστικά τα οποία σχετίζονται με τη γλωσσική μορφολογία των tweets, την υποκειμενικότητά τους, την ανάλυση συναισθήματος και το είδος των λέξεων που χρησιμοποιούνται. Επιπλέον, μία καινοτομία της παρούσας Εργασίας είναι η χρήση χαρακτηριστικών τα οποία προκύπτουν μέσω της εφαρμογής της μεθόδου TF-IDF στα κείμενα των

συλλεχθέντων δημοσιεύσεων. Καθώς οι διαστάσεις ωστόσο των χαρακτηριστικών εκτοξεύτηκαν, απαραίτητη ήταν η χρήση της μείωσης της διαστατικότητας με χρήση της μεθόδου PCA.

Η εφαρμογή των χαρακτηριστικών αυτών σε αλγόριθμους Μηχανικής Μάθησης για την αυτόματη ταξινόμησή τους έδειξε μη ικανοποιητικά αλλά ενθαρρυντικά αποτελέσματα για την επίλυση του προβλήματος. Η αιτία της χαμηλής ακρίβειας των αλγορίθμων εντοπίζεται σε δύο παράγοντες. Αφενός στην δυσκολία διάκρισης ακόμα και για εξειδικευμένο προσωπικό μεταξύ των δημοσιεύσεων διαφορετικών κατηγοριών, καθώς η διάκριση μεταξύ ορθής και ψευδούς είδησης στηρίζεται στη γνώση γεγονότων. Κάτι τέτοιο δεν μπορεί να επιτευχθεί αξιοποιώντας μόνο γλωσσικά χαρακτηριστικά. Επίσης, το πλήθος των κατηγοριών στις οποίες διακρίθηκαν τα tweets αναμένεται να επηρεάζει σίγουρα την ακρίβεια, καθώς απαιτείται επίσης εξειδικευμένη γνώση για το πλήθος και τον χαρακτηρισμό των κατηγοριών που επιλέγονται.

Με εφαλτήριο τις δύο αυτές διαπιστώσεις, δίνεται το έναυσμα για περαιτέρω έρευνα στο πρόβλημα. Αρχικά μπορεί να μελετηθεί η εξόρυξη γνώσης σχετικά με την θεματολογία και η χρήση της πληροφορίας αυτής κατά την ταξινόμηση. Επίσης, σημαντική είναι και η συνεισφορά στην εγκυρότητα μίας δημοσίευσης, της αξιοπιστίας του χρήστη που πραγματοποίησε τη δημοσίευση. Η βελτίωση των αποτελεσμάτων και της ακρίβειας της κατηγοριοποίησης των αλγορίθμων SVM και Random Forest μπορεί να επιτευχθεί, όπως παρατηρήθηκε, από την εφαρμογή μεθόδων βελτιστοποίησης υπερπαραμέτρων όπως τη Grid Search και τη Randomized Search αντίστοιχα για τον κάθε αλγόριθμο. Συνεπώς, ένας από τους μελλοντικούς στόχους αποτελεί η διερεύνηση και η δοκιμή άλλων παραμέτρων οι οποίες πηγάζουν μέσω αυτών των μεθόδων, με σκοπό την περαιτέρω βελτίωση των αποτελεσμάτων. Τέλος, η διάκριση των tweets σε διαφορετικές κατηγορίες καθώς και ο χαρακτηρισμός μεγαλύτερου όγκου δεδομένων σίγουρα θα επιφέρει και νέους τρόπους για την επίτευξη του τελικού στόχου.

Βιβλιογραφία

[1] «Blog Hootsuite,» [Ηλεκτρονικό]. Available: https://blog.hootsuite.com/twitter-statistics/. [Πρόσβαση 2021].

[2] J. Ritterman, M. Osborne και E. Klein, «Using prediction markets and Twitter to predict a swine flu pandemic,» *1st International Workshop on Mining Social Media,* 11 2009.

[3] A. Signorini, A. M. Segre και P. M. Polgreen, «The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic,» *PLOS ONE,* τόμ. 6, pp. 1-10, 5 2011.

[4] T. Rao και S. Srivastava, «Analyzing Stock Market Movements Using Twitter Sentiment Analysis,» σε *ASONAM 2012*, 2012.

[5] A. Tumasjan, T. Sprenger, P. G. Sandner και I. Welpe, «Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment,» σε *ICWSM*, 2010.

[6] Wikipedia, «COVID-19,» [Ηλεκτρονικό]. Available: https://el.wikipedia.org/wiki/%CE%A0%CE%B1%CE%BD%CE%B4%CE%B7%CE%BC%CE%AF%CE%B1_COVID-19. [Πρόσβαση 2021].

[7] «twitter blog,» [Ηλεκτρονικό]. Available: https://blog.twitter.com/en_us/topics/insights/2020/spending-2020-together-on-twitter.html. [Πρόσβαση 2021].

[8] K. Shu, A. Sliva, S. Wang, J. Tang και H. Liu, «Fake News Detection on Social Media: A Data Mining Perspective,» *SIGKDD Explor. Newsl.*, τόμ. 19, p. 22–36, 9 2017.

[9] R. Oshikawa, J. Qian και W. Y. Wang, «A Survey on Natural Language Processing for Fake News Detection,» *CoRR,* τόμ. abs/1811.00770, 2018.

[10] C. Buntain και J. Golbeck, «I Want to Believe: Journalists and Crowdsourced Accuracy Assessments in Twitter,» *CoRR,* τόμ. abs/1705.01613, 2017.

[11] V. Qazvinian, E. Rosengren, D. Radev και Q. Mei, «Rumor has it: Identifying Misinformation in Microblogs,» 2011.

[12] B. Kang, J. O'Donovan και T. Höllerer, «Modeling topic specific credibility in Twitter,» 2012.

[13] S. Kwon, M. Cha, K. Jung, W. Chen και Y. Wang, «Prominent Features of Rumor Propagation in Online Social Media,» σε *2013 IEEE 13th International Conference on Data Mining*, 2013.

[14] A. Zervopoulos, A. G. Alvanou, K. Bezas, A. Papamichail, M. Maragoudakis και K. Kermanidis, «Hong Kong Protests: Using Natural Language Processing for Fake News Detection on Twitter,» σε *Artificial Intelligence Applications and Innovations*, Cham, 2020.

[15] S. A. Memon και K. M. Carley, *Characterizing COVID-19 Misinformation Communities Using a Novel Twitter Dataset*, 2020.

[16] L. Cui και D. Lee, *CoAID: COVID-19 Healthcare Misinformation Dataset,* 2020.

[17] U. Qazi, M. Imran και F. Ofli, *GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information*, 2020.

[18] J. Brennen, F. Simon, P. Howard και R. Nielsen, *Types, Sources, and Claims of COVID-19 Misinformation*, 2020.

[19] «Developer Twitter,» [Ηλεκτρονικό]. Available: https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time. [Πρόσβαση 2021].

[20] «Miami EDU,» [Ηλεκτρονικό]. Available: https://covid.dh.miami.edu/2020/06/11/hydrating-tweetsets/. [Πρόσβαση 2021].

[21] «oauth.net,» [Ηλεκτρονικό]. Available: https://oauth.net/2/. [Πρόσβαση 2021].

[22] «zenodo,» [Ηλεκτρονικό]. Available: https://zenodo.org/ . [Πρόσβαση 2021].

[23] «github,» [Ηλεκτρονικό]. Available: https://github.com/ . [Πρόσβαση 2021].

[24] «Wordcloud,» [Ηλεκτρονικό]. Available: http://amueller.github.io/word_cloud/. [Πρόσβαση 2021].

[25] «Tokenization,» [Ηλεκτρονικό]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html. [Πρόσβαση 2021].

[26] «Stemming and Lemmatization,» [Ηλεκτρονικό]. Available: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html. [Πρόσβαση 2021].

[27] «Curse of Dimensionality,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Curse_of_dimensionality. [Πρόσβαση 2021].

[28] L. Van Der Maaten, E. Postma και J. Van den Herik, «Dimensionality reduction: a comparative review,» *J Mach Learn Res*, τόμ. 10, pp. 66-71, 2009.

[29] «Dimensionality reduction: a comparative review,» *J Mach Learn Res,* τόμ. 10, pp. 66-71, 2009.

[30] S. M. Ross, «Introductory Statistics,» σε *Introductory Statistics (Fourth Edition)*, Fourth Edition επιμ., S. M. Ross, Επιμ., Oxford, Academic Press, 2017, pp. 797-800.

[31] «scikit-learn,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/svm.html. [Πρόσβαση 2021].

[32] «Decision Tree Classifier,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. [Πρόσβαση 2021].

[33] «scikit-learn,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.. [Πρόσβαση 2021].

[34] «Regular Expression,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/Regular_expression. [Πρόσβαση 2021].

[35] «Tweepy,» [Ηλεκτρονικό]. Available: https://www.tweepy.org/. [Πρόσβαση 2021].

[36] «JSON,» [Ηλεκτρονικό]. Available: https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Global_Objects/JSON. [Πρόσβαση 2021].

[37] «MongoDB,» [Ηλεκτρονικό]. Available: https://www.mongodb.com/. [Πρόσβαση 2021].

[38] «PyMongo,» [Ηλεκτρονικό]. Available: https://pymongo.readthedocs.io/en/stable/. [Πρόσβαση 2021].

[39] «Pandas,» [Ηλεκτρονικό]. Available: https://pandas.pydata.org. [Πρόσβαση 2021].

[40] «NLTK,» [Ηλεκτρονικό]. Available: https://www.nltk.org/. [Πρόσβαση 2021].

[41] «spaCy,» [Ηλεκτρονικό]. Available: https://spacy.io/. [Πρόσβαση 2021].

[42] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot και É. Duchesnay, «Scikit-Learn: Machine Learning in Python,» *J. Mach. Learn. Res.*, τόμ. 12, p. 2825–2830, 11 2011.

[43] «Matplotlib,» [Ηλεκτρονικό]. Available: https://matplotlib.org/stable/index.html. [Πρόσβαση 2021].

[44] «Seaborn,» [Ηλεκτρονικό]. Available: https://seaborn.pydata.org/. [Πρόσβαση 2021].

[45] «Jupyter,» [Ηλεκτρονικό]. Available: https://jupyter.org/. [Πρόσβαση 2021].

[46] «Flask,» [Ηλεκτρονικό]. Available: https://flask.palletsprojects.com/en/1.1.x/. [Πρόσβαση 2021].

[47] J. M. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, K. Artemova, E. Tutubalina και G. Chowell, *A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration*, 2020.

[48] «N-gram,» [Ηλεκτρονικό]. Available: https://en.wikipedia.org/wiki/N-gram. [Πρόσβαση 2021].

[49] «Covid19 Twitter,» [Ηλεκτρονικό]. Available: https://github.com/thepanacealab/covid19_twitter. [Πρόσβαση 2021].

[50] «emosent,» [Ηλεκτρονικό]. Available: https://pypi.org/project/emosent-py/. [Πρόσβαση 2021].

[51] Explosion.ai, «spaCy trained pipelines for Greek language,» spacy.io, 1 Feb 2021. [Ηλεκτρονικό]. Available: https://spacy.io/models/el. [Πρόσβαση 5 Mar 2021].

[52] «Grid-Search,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. [Πρόσβαση 2021].

[53] «Randomized Search,» [Ηλεκτρονικό]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html. [Πρόσβαση 2021].

[54] «Hootsuite,» 2021. [Ηλεκτρονικό]. Available: https://www.hootsuite.com/.