ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΑΙ ΕΞΟΡΥΞΗ ΓΝΩΜΗΣ ΤΩΝ ΧΡΗΣΤΩΝ ΤΟΥ TWITTER ΜΕ ΧΡΗΣΗ ΕΡΓΑΛΕΙΩΝ ΤΗΣ PYTHON. ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΡΩΣΟ-ΟΥΚΡΑΝΙΚΟΥ ΠΟΛΕΜΟΥ.

του

ΓΕΩΡΓΙΟΥ ΘΑΝΕΛΛΑ

Διπλωματική Εργασία

Υπεβλήθη για την εκπλήρωση μέρους των απαιτήσεων για

την απόκτηση του Διπλώματος Μηχανολόγου Μηχανικού

Βόλος, 2022

ΠΕΡΙΛΗΨΗ

Το Τουίτερ (Twitter) είναι μία από τις δημοφιλέστερες διαδικτυακές πλατφόρμες παγκοσμίως. Πληθώρα χρηστών το χρησιμοποιούν σε καθημερινή βάση, εκφράζοντας τις απόψεις τους πάνω σε διάφορα γεγονότα (της επικαιρότητας συνήθως), προϊόντα, υπηρεσίες κα.

Η ανάπτυξη αλγορίθμων, σε διάφορα προγραμματιστικά περιβάλλοντα (π.χ. Python, R κ.α.) ,που να μπορούν να επεξεργαστούν αυτά τα δεδομένα μορφής κειμένου (Semi Structured Data),και να παράγουν πληροφορία, είναι μια πολύ σημαντική διαδικασία.

Μια εκ των σημαντικότερων τέτοιων διαδικασιών είναι η Ανάλυση Συναισθήματος ή ΑΣ (Sentiment Analysis), η οποία αποτελεί μια από τις βασικότερες τεχνικές ΕΦΓ (Επεξεργασία Φυσικής Γλώσσας) ή NLP (Natural Language Processing). Οι ΕΦΓ ουσιαστικά ασχολούνται με την δημιουργία αλγορίθμων που να μπορούν να κατανοούν και να βγάζουν συμπεράσματα από την απλή καθημερινή γλώσσα των ανθρώπων, κάνοντας διάφορες διαδικασίες όπως μεταφράσεις, συνόψεις κειμένων (summarization), αναγνώριση λόγου (speech recognition) αλλά και το βασικό αντικείμενο αυτής της εργασίας και προαναφερθέν, την ΑΣ. Μέσω αυτού, ουσιαστικά ο αλγόριθμος διαβάζει ένα κείμενο ή μια πρόταση, και εξάγει ένα αποτέλεσμα που αφορά το συναίσθημα ή την γνώμη του συγγραφέα πάνω στο θέμα για το οποίο έγραψε το κείμενο η την πρόταση. Αυτό το αποτέλεσμα συνήθως είναι <<θετικό>>, <<ουδέτερο>> ή <<αρνητικό>> (συναίσθημα ή sentiment).

Στην συγκεκριμένη διατριβή λοιπόν, αντλήθηκε ένας αριθμός απο τουίτς μέσω του API του Twitter, σχετιζόμενα με τον πόλεμο Ρωσσίας – Ουκρανίας. Συγκεκριμένα πρόκειται για τουίτς που δημοσιεύθηκαν τις 4 πρώτες μέρες του πολέμου, και πάνω σε αυτά έγιναν διάφορες αναλύσεις. Η κυριότερη ήταν φυσικά η ΑΣ, όπου χρησιμοποιήθηκαν εργαλεία μηχανικής μάθησης καθώς και μέθοδοι χρήσης λεξικών (lexicon ή rule based methods).

Πέρα όμως από αυτήν, εφαρμόστηκαν και άλλες αναλύσεις στα δεδομένα. Η μία αφορά τις δημοφιλέστερες τοποθεσίες από όπου έγιναν τα περισσότερα τουίτς και η δεύτερη σχετίζεται με τις πιο πολυχρησιμοποιούμενες λέξεις σε αυτά.

Βέβαια, προτού γίνουν όλα τα παραπάνω χρειάστηκε να γίνει και η κατάλληλη προεπεξεργασία των δεδομένων, μιας και τα τουίτς στην αρχικη τους μορφή είναι τις περισσότερες φορές αρκετα δυσανάγνωστα για την μηχανή, και πρέπει επομένως να έρθουν σε μια μορφή πιο διαχειρίσιμη για περαιτέρω αναλύσεις.

Τέλος, σχολιάστηκαν τα αποτελέσματα της ερευνας και προτάθηκαν ιδέες για μελλοντικές εργασίες.

SUMMARY

Twitter is one of the most popular online platforms. Many users use it on a daily basis, expressing their views on various (current mostly) events, products, services, etc.

The development of algorithms, in various programming environments (e.g. Python, R etc.), that can process those textual data (Semi Structured Data), and produce information, is a very important process.

One of the most important such process is Sentiment Analysis, which is one of the basic techniques of NLP (Natural Language Processing). NLP essentially deals with the creation of algorithms that can understand and make conclusions from the simple everyday language of people,by doing various procedures such as translations, text summarization, speech recognition and the main object of this and the aforementioned, the Sentiment Analysis. Through this, the algorithm reads a text or a sentence, and extracts a result related to the author's feeling or opinion on the subject for which the text or sentence was written. This result is usually "positive", "neutral" or "negative" (feeling or sentiment).

In this dissertation, several tweets were extracted through the Twitter API, related to the Russia-Ukraine war. Specifically, these are tweets that were published in the first 4 days of the war, and various analyzes were made on them. The main one was of course the SA, where machine learning tools were used as well as lexicon (or rule) based methods.

But beyond that, other analyzes of the data were applied too. The first one, concerns the most popular sites from which most tweets were made and the second relates to the most frequently used words in them.

Of course, before all of the above can be done, the data must be properly pre-processed, since the tweets in their original form are often quite illegible for the machine, and must therefore come in a more manageable format for further analysis.

Finally, the results of the research were commented and ideas for future work were suggested.

ΠΕΡΙΕΧΟΜΕΝΑ

Λίστα Σχημάτων

1. ΕΙΣΑΓΩΓΗ

1.1 ΠΕΡΙΓΡΑΦΗ ΚΑΙ ΟΡΓΑΝΩΣΗ ΤΗΣ ΕΡΓΑΣΙΑΣ

Στην συγκεκριμένη διατριβή, στόχος είναι να εφαρμοστούν σε πρωτογενή,ακατέργαστα δεδομένα μορφής κειμένου,που είναι διατυπωμένα με φυσικό,καθημερινό λόγο,διάφορες τεχνικές ανάλυσης,μέσω των οποίων θα εξάγεται πολύτιμη πληροφορία. Αυτές οι τεχνικές ποικίλλουν, και η κάθε μία παρέχει την δική της πληροφορία, αλλά η βασικότερη εξ αυτών και αυτή που καταλαμβάνει το μεγαλύτερο μέρος της παρούσας εργασίας, είναι η τεχνική ανάλυσης συναισθήματος που αποτελεί μία από τις τεχνικές Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing) ή ΕΦΓ (NLP).

Η συγκεκριμένη τεχνική χρησιμοποιείται με σκοπό να αντληθεί από ένα κείμενο (text) το συναίσθημα (sentiment) του συγγραφέα,το οποίο μπορει να είναι θετικό,ουδέτερο η αρνητικό. Ουσιαστικά η πληροφορία είναι το τι νιώθει ο εκάστοτε συγγραφέας σχετικά με το θέμα το οποίο αναφέρει στο κείμενο του. Στην περίπτωση της συγκεκριμένης εργασίας, όπου τα δεδομένα αυτά μορφής κειμένου αντλούνται από την δημοφιλή πλατφόρμα κοινωνικής δικτύωσης Twitter, τα κείμενα είναι τα τουίτς, και ο συγγραφέας είναι χρήστης που δημοσίευσε το εκάστοτε τουίτ.

Όπως θα αναλυθεί και στο αντίστοιχο κεφάλαιο, η τεχνική αυτή βρίσκει ιδιαίτερη εφαρμογή στον επιχειρηματικό τομέα, μιας και πολλές εταιρείες την χρησιμοποιούν για να αντλούν την γνώμη των καταναλωτών σχετικά με το brand, τα προιόντα ή τις υπηρεσίες τους μέσω σχολίων που αφήνουν σε blogs, διάφορα sites ή -κυρίως- στα μέσα κοινωνικής δικτύωσης.

Η εργασία χωρίζεται σε 5 επιμέρους κεφάλαια. Στη συνέχεια του 1^{ου} Κεφαλαίου παρουσιάζεται η οργάνωση της εργασίας, το κίνητρο για την σύνταξη της καθώς και μία σύντομη βιβλιογραφική ανασκόπηση.

Στο κεφάλαιο 2 γίνεται μια αναφορά στις βασικές δομές που χρησιμοποιούνται,οι οποίες είναι η γλώσσα προγραμματισμού Python και η πλατφόρμα κοινωνικής δικτύωσης Twitter, καθώς και στο βασικό ζήτημα που απασχολεί την εργασία και που πάνω στο οποίο γίνεται η ανάλυση, που είναι ο πόλεμος Ρωσίας – Ουκρανίας.

Το Κεφάλαιο 3 είναι που παρουσιάζονται αναλυτικά οι τεχνικές Επεξεργασίας Φυσικής Γλώσσας και η χρησιμότητα τους, εμβαθύνοντας έπειτα στην τεχνική που αφορά την συγκεκριμένη έρευνα, που είναι η ανάλυση συναισθήματος.

Έπειτα, στο κεφάλαιο 4 παρουσιάζονται πιο συγκεκριμένα τα εργαλεία που χρησιμοποιούνται για τις αναλύσεις, τα οποία είναι βασικές βιβλιοθήκες της Python που χρησιμοποιούνται στην ανάλυση δεδομένων καθώς και σε εφαρμογές μηχανικής μάθησης.

Τέλος, το Κεφάλαιο 5 είναι αυτό όπου παρουσιάζονται βήμα-βήμα ολες οι διαδικασίες, με την σειρά με την οποία έγιναν και στον κώδικα (βλέπε παράρτημα). Στην αρχή και στην ενότητα 5.1 φαίνεται η αυθεντικοποίηση μέσω του API του Τουίτερ, ώστε να μπορούν να αντληθούν τα τουίτς, καθώς και η διαδικασία με την οποία αντλούνται. Έπειτα στην ενότητα 5.2 γίνεται η προεπεξεργασία των δεδομένων ώστε να έρθουν σε μια μορφή κατάλληλη για να αναλυθούν. Η ανάλυση ξεκινάει στην ενότητα 5.3, η οποία

είναι γενικού χαρακτήρα και έχει ως στόχο να εξορύξει μερικές ενδιαφέρουσες πληροφορίες από τα δεδομένα. Στις ενότητες 5.4 και 5.5 γίνεται η Ανάλυση Συναισθήματος, στην πρώτη χρησιμοποιώντας τις μεθόδους lexicon – based και στην δεύτερη τις μεθόδους μηχανικής μάθησης (οι δύο αυτές μέθοδοι αναλύονται και στην ενότητα 3.2.2).

1.2 ΥΠΟΒΑΘΡΟ – ΚΙΝΗΤΡΟ ΕΚΠΟΝΗΣΗΣ ΤΗΣ ΕΡΓΑΣΙΑΣ

Όπως θα αναλυθεί και στο κεφάλαιο 3, η ικανότητα άντλησης πληροφορίας από μια πλειάδα δεδομένων μορφής κειμένου (όπως είναι τα τουίτς στην περίπτωση μας), είναι εξέχουσας σημασίας και χρησιμοποιείται κατά κόρον στον επιχειρηματικό τομέα. Όταν μια εταιρία, έπειτα από σωστή εφαρμογή των τεχνικών Ανάλυσης Συναισθήματος, ξέρει πως νιώθουν οι καταναλωτές για το brand της ή για κάποιο καινούργιο προιόν της για παράδειγμα, τότε αποκτάει ένα πολύ σημαντικό πλεονέκτημα ώστε να φτάσει πιο γρήγορα και αποδοτικά στην ικανοποίηση τους.

Οι δυσκολίες στην ΑΣ έγκειται στο γεγονος πως η ανθρώπινη καθημερινή ομιλία περιέχει χαρακτηριστικά, όπως αναλύεται και στο κεφάλαιο 3, τα οποία πολύ δύσκολα αναγνωρίζονται από τους αλγορίθμους, όπως ιδιωματισμοί, έμμεσες αρνήσεις, ειρωνία κα. Ωστόσο μέχρι στιγμής έχουν αναπτυχθεί πολύ ακριβή μοντέλα και με την συνεχή έρευνα, αναμένεται να αναπτυχθούν και ακόμη αποδοτικότερα.

Η συγκεκριμένη εργασία, όντας στα πλαίσια προπτυχιακού επιπέδου, αρκείται στην ανάπτυξη βασικών μοντέλων με διόλου όμως κακή απόδοση. Ο σκοπός της είναι διπλός.Ο πρώτος είναι να βρεθεί το Συναίσθημα των τουίτς με χρήση τεχνικών βασισμένες σε λεξικό (lexicon based techniques),και ο δεύτερος,με βάση αυτό το Συναίσθημα, να εκπαιδευτούν μοντέλα μηχανικής μάθησης τα οποία θα βρίσκουν απευθείας το Συναίσθημα.

1.3 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Οι εργασίες που έχουν γίνει στην ΑΣ είναι πάμπολλες, γεγονός που δεν εκπλήσσει καθόλου δεδομένης της χρηστικότητας της. Όταν τα δεδομένα στο ίντερνετ άρχισαν να πληθαίνουν στα διάφορα blogs,forums, ιστοσελίδες κλπ, άρχισαν να υλοποιούνται και οι πρώτες σχετικές εργασίες. Η πρώτη δημοσιευμένη σχετική εργασία, είναι των Pang et. al (2002) και επικεντρώνεται σε ανάπτυξη αλγορίθμων μηχανικής μάθησης (Naive Bayes, maximum entropy, support vector machines) με σκοπό την κατηγοριοποίηση ενός πλήθους δεδομένων σχετιζόμενα με κριτικές ταινιών σε Θετικά ή Αρνητικά. Παρόμοιες εργασίες σε παρόμοιες βάσεις δεδομένων ταινιών που ακολούθησαν αμέσως ήταν οι Pang and Lee (2005) και Popescu & Etzioni (2005), ή προιόντων (Hu & Liu,2004), (Popescu and Etzioni,2005). Λίγο αργότερα ακολούθησαν και άλλες εργασίες οι οποίες επικεντρώθηκαν σε αναλύσεις πωλήσεων διαφόρων προιόντων (βιβλία, βιντεοπαιχνιδια κα) όπως των Chevalier and Mayzlin (2006), Liu et al. (2007), Zhu and Zhang (2010).

Αργότερα, και όταν πια οι πλατφόρμες κοινωνικης δικτύωσης όπως το Τουίτερ άρχισαν να γίνονται αρκετά δημοφιλείς και να παρέχουν μια πληθώρα διαθέσιμων για ανάλυση δεδομένων, πολλές μελέτες στράφηκαν προς τα εκεί (Jansen et al.,2009), (Asur and Huberman ,2010), (Arias et al., 2013). Οι H. Wang et al (2012) δημοσίευσαν μία εργασία όπου εφαρμόστηκαν τεχνικές ΑΣ σε τουίτς σχετικά με τις αμερικάνικες προεδρικές εκλογές το 2012, όπου φάνηκε η αποδοτικότητα αυτής της μεθόδου μέσω της

ταχείας εύρεσης της γνώμης του κοινού. Οι O. Almastrafi, S. Parack, B. Chavan et al (2010) πρότειναν και παρουσίασαν μία ανάλυση βασισμένη στις τοποθεσίες, εφαρμόζοντας τεχνικές ΑΣ με αλγορίθμους μηχανικής μάθησης όπως ο Naïve-Bayes σε 600 000 τουίτς σχετικά με τις ινδικές εκλογές. Τα αποτελέσματα παρουσιάστηκαν ανά περιοχή, δείχνοντας έτσι την δυναμικότητα του κάθε πολιτικού κόμματος στην κάθε μια.

Στον χώρο του χρηματιστηρίου έχουν δημοσιευθεί επίσης αρκετά ενδιαφέρουσες εργασίες,όπως των Lemmon & Portniaguina (2006) και Han (2008), που δείχνουν ότι υπάρχει σχέση ανάμεσα στο συναίσθημα (sentiment) και την εμπιστοσύνη (confidence) των επενδυτών με την διαμόρφωση των τιμών στο χρηματιστήριο. Επιπλέον,οι Gilbert & Karahalios (2010) δείχνουν ότι η πρόβλεψη των συναισθημάτων από διάφορα weblogs (διαδικτυακά μπλόγκ) μπορεί να παρέχει καίρια πληροφορία για τις μελλοντικές τιμές στο χρηματιστήριο.

Η ανάλυση συναισθήματος συνοδεύεται από πληθώρα εργασιών και σε ποικίλλους άλλους τομείς, οι οποίες σαφώς δεν γίνεται να παρουσιαστούν όλες εδώ.

2. ΒΑΣΙΚΕΣ ΔΟΜΕΣ – ΖΗΤΗΜΑΤΑ .

2.1 Το προγραμματιτικό περιβάλλον της Python

Η Python, σύμφωνα με την επίσιμη ιστοσελίδα της, «είναι μία αντικειμενοστραφής, υψηλού επιπέδου, γενικού σκοπού γλώσσα προγραμματισμού, που μπορεί να χρησιμοποιηθεί για μία μεγάλη ποικιλία προβλημάτων». «Έχει ενσωματωμένα δομοστοιχεία (modules), exceptions, dynamic typing, Πολυ υψηλού επιπέδου dynamic data types και classes» (δεν υπάρχουν αντιπροσωπευτικοί ελληνικοί όροι για αυτές τις έννοιες). «Η Python συνδυάζει αξιοσημείωτη δύναμη με πολύ σαφή σύνταξη».

Σύμφωνα πάλι με την επίσιμη ιστοσελίδα της, η Python «διαθέτει πληθώρα βιβλιοθηκών που καλύπτουν τομείς όπως επεξεργασία κειμένων, internet protocols, software engineering, και operating system interfaces».

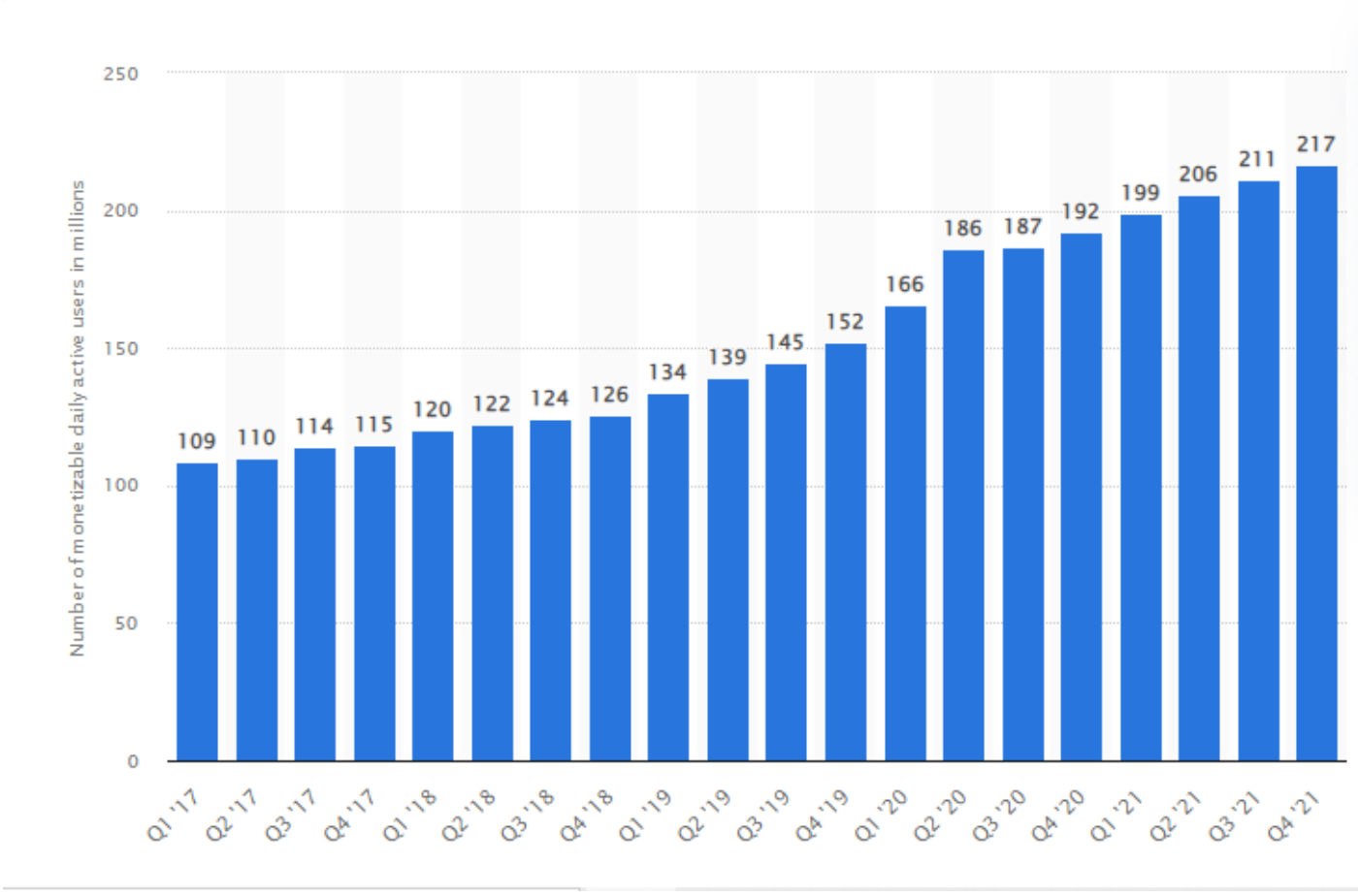
Όσον αφορά το θέμα της παρούσας εργασίας, που είναι η ανάλυση δεδομένων, και ειδικότερα δεδομένων μορφής κειμένου όπου εφαρμόζονται NLP τεχνικές, η Python αποτελεί όχι απλά μια αξιόπιστη, αλλά πολλές φορές την πρώτη επιλογή κάθε προγραμματιστή. Ακολουθούν τα περιβάλλοντα SQL, R κ.α. Ο λόγος είναι πως διαθέτει πληθώρα βιβλιοθηκών για κάθε στάδιο της ανάλυσης που θα προκύψει (π.χ Pandas, Numpy, MatplotLib, NLTK, scikit-learn κ.α.) , με πάρα πολύ αποδοτική και φιλική προς τον χρήστη δομή.

2.2 Η πλατφόρμα κοινωνικής δικτύωσης Twitter.

Το Twitter είναι μια πλατφόρμα κοινωνικής δικτύωσης («social networking or blogging platform») που ιδρύθηκε το 2006 από τους Jack Dorsey, Biz Stone, Noah Glass και Even Williams (Twitter, 2016) . Εκεί οι εγγεγραμένοι χρήστες μπορούν να κάνουν δημοσιεύσεις (tweets) που να αποτελούνται από κείμενα ή οπτικοακουστικό υλικό,να λάνουν like ή να αναδημοσιέυσουν υπάρχουσες δημοσιεύσεις. Οι μη εγγεγραμένοι χρήστες μπορούν μόνο να διαβάζουν δημοσιεύσεις άλλων.

Τα tweets έχουν την ιδιαιτερότητα ότι πρέπει να έχουν έκταση 140 χαρακτήρων, ενώ βασικά χαρακηριστικά τους είναι τα hashtags (#) και τα mentions(@). Τα hashtags είναι λέξεις οι οποίες έχουν το σύμβολο # στην αρχή τους, και αποτελούν ουσιαστικά τις λέξεις κλειδιά για το συγκεκριμένο τουίτ. Εάν για παράδειγμα πρόκειται για ένα τουίτ που περιέχει γνώμη για μια υποθετική κίνηση της κυβέρνησης σχετικά με την ακρίβεια, θα έχει ενδεχομένως τα hashtags «#κυβερνηση» , «#ακριβεια» ή άλλα παρόμοια. Οπότε όταν κάποιος χρήστης θέλει να διαβάσει σχόλια για κάποιο συγκεκριμένο θέμα,θα αναζητήσει τουίτς με τα σχετικά hashtags. Τα mentions, είναι ουσιαστικά η αναφορά άλλων χρηστών σε κάποιο τουίτ ενός χρήστη. Στο προηγούμενο υποθετικό παράδειγμα, ένα mention θα μπορούσε να είναι ο πρωθυπουργός ο ίδιος, και το οποίο θα ήταν «@kmitsotakis».

Οι καθημερινοί χρήστες του Twitter το τελευταίο τρίμηνο του 2021 ανέρχονταν στους 217 εκατομμύρια. Στην Εικόνα 1.1 φαίνονται οι ημερίσιοι χρήστες ανά τρίμηνο από το 2017 εώς και το 2021, και αντλήθηκε από τον ιστότοπο statista.com.



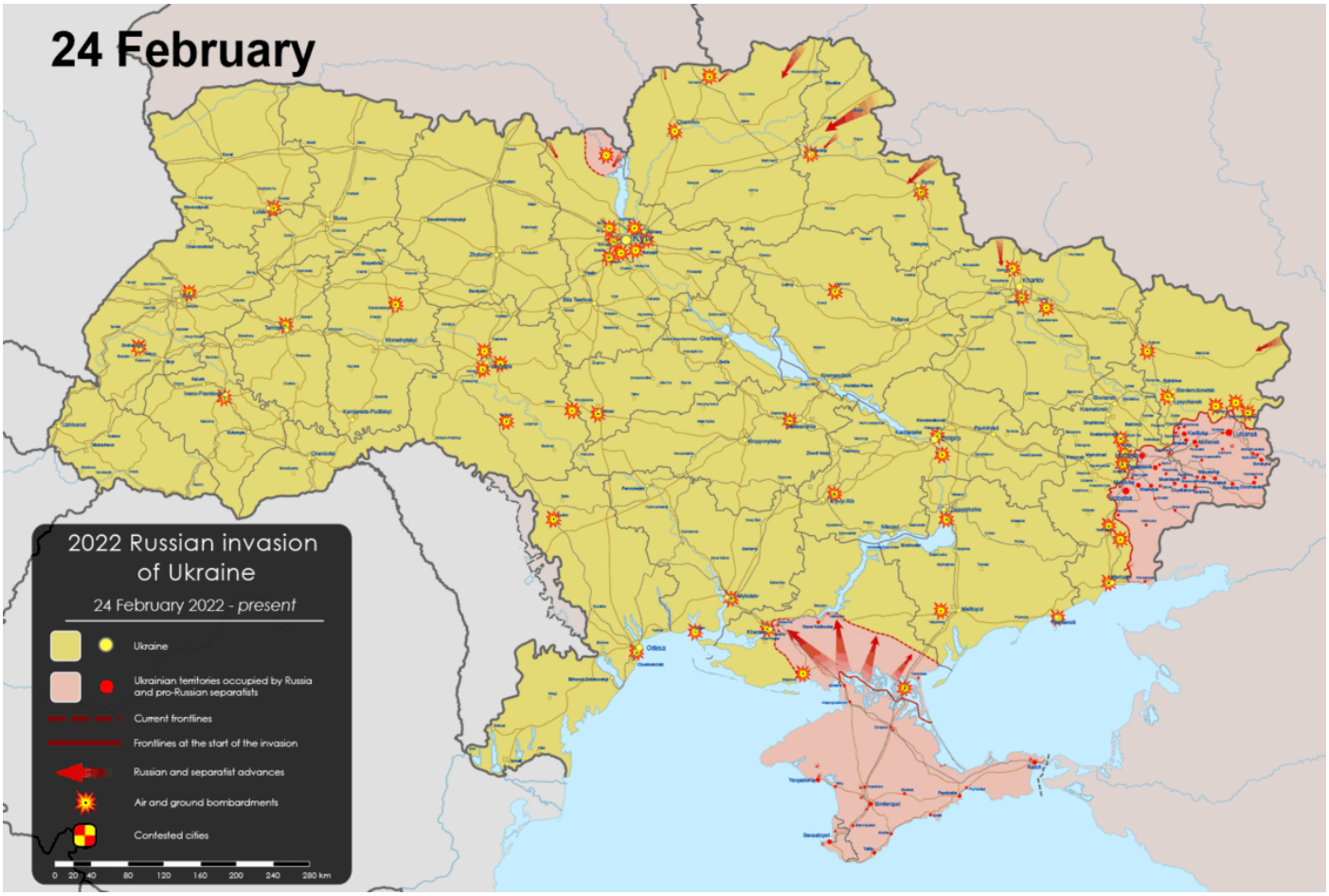
Εικόνα 2.1 . Καθημερινοί χρήστες του Twitter σε εκατομμύρια ανά τρίμηνο απότο 2017 έως το 2021 .

Σε αντίθεση με άλλες παρόμοιες πλατφόρμες (π.χ Facebook) το Twitter χρησιμοποιείται σχεδόν αποκλειστικά από τους χρήστες για σχολιασμό διαφόρων θεμάτων, είτε πρόκειται για διάφορες εξελίξεις στα τεκτενόμενα, είτε για έκφραση γνώμης πάνω σε προιόντα, υπηρεσίες , ανθρώπους κλπ . Γενικά, εάν κάποιος/α επιθυμεί να δει τι σκέφτεται ο κόσμος σχετικά με ένα (επίκαιρο, συνήθώς) θέμα, θα επισκεφθεί το Twitter. Αυτό ακριβώς το χαρακτηριστικό του είναι που το καθιστά κατάλληλο για εργασίες ανάλυσης συναισθήματος σαν την συγκεκριμένη.

2.3 Ο πόλεμος Ρωσσίας – Ουκρανίας.

Τα δεδομένα πάνω στα οποία θα γίνουν οι αναλύσεις της συγκεκριμένης διατριβής, επιλέχθηκε να είναι τουίτς που σχετίζονται με τον πόλεμο Ρωσσίας-Ουκρανίας.

Πρόκειται για ένα συμβάν που ξεκίνησε με την εισβολή της Ρωσσίας σε Ουκρανικό έδαφος στις 24/02/2022 με διάγγελμα του Βλαδιμίρ Πούτιν, ονομάζοντας την «ειδική στρατιωτική αποστολή» με σκοπό να «αποστρατικοποιήσει και να απαλλάξει από τους Ναζί» την Ουκρανία. Στην παρακάτω διαδραστική εικόνα φαίνεται το χρονικό της εισβολής από την πρώτη μέρα μέχρι και τις 21/04/2022.



Εικόνα 2.2 . Το χρονικό της εισβολής από την πρώτη μέρα ως τις 21/04/2022

Τα αίτια που οδήγησαν στο συγκεκριμένο συμβάν είναι πολύπλευρα, όπως και σε κάθε άλλο παρόμοιο συμβάν της ιστορίας άλλωστε. Αυτά, μαζί με τις λοιπές λεπτομέρειες του πολέμου δεν αποτελούν σαφώς αντικείμενο της συγκεκριμένης εργασίας, και ο αναγνώστης μπορεί να ανατρέξει στο διαδίκτυο για περισσότερες σχετικές λεπτομέρειες.

Εδώ απλά θα αναφερθεί πως έπειτα από αυτήν την επιχείρηση υπήρξε διχασμός σχετικά με τους υπαίτιους αυτής της τραγωδίας, η οποία προκάλεσε τον ξεριζωμό εκατοντάδων χιλιάδων ανθρώπων, θανάτους αμάχων και μεγάλες απώλειες περουσιών. Έχουν δωθεί ευκαιρίες στην διπλωματία, αλλά δυστυχώς δεν μπορεί να βρεθεί εύφορο έδαφος. Όλος ο πλανήτης παρακολουθεί αγωνιωδώς τις εξελίξεις, μιας και πρόκειται για μια κατάσταση που κάνεις δεν μπορεί να προβλέψει μέχρι που θα φτάσει και πόσο καταστροφικές συνέπειες θα έχει. Ήδη η παγκόσμια οικονομία έχει κλωνιστεί ριζικά, με τον κόσμο να υποφέρει με την ακρίβεια που έχει επέλθει σε πολλά βασικά αγαθά (καύσιμα, σιτηρά και πολλά άλλα).

3. ΤΕΧΝΙΚΕΣ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ (ΕΦΓ) ΚΑΙ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

3.1 ΟΙ ΤΕΧΝΙΚΕΣ ΕΦΓ

3.1.1 ΤΙ ΕΙΝΑΙ ΟΙ ΤΕΧΝΙΚΕΣ ΕΦΓ

Στο site της Βικιπαίδεια διαβάζουμε τον εξής ορισμό για την Επεξεργασία Φυσικής Γλώσσας ή ΕΦΓ (Natural Language Processing ή NLP). «Η επεξεργασία φυσικής γλώσσας (ΕΦΓ) είναι ένας διεπιστημονικός κλάδος της επιστήμης της πληροφορικής,της τεχνιτής νοημοσύνης και της υπολογιστικής γλωσσολογίας και ασχολείται με τις αλληλεπιδράσεις μεταξύ των υπολογιστών

και των ανθρώπινων (φυσικών) γλωσσών. Κατά συνέπεια, η ΕΦΓ συνδέεται στενά με την αλληλεπίδραση ανθρώπου-υπολογιστή. Προκλήσεις στην ΕΦΓ περιλαμβάνουν την κατανόηση φυσικής γλώσσας, δηλαδή την προσπάθεια να καταστούν ικανοί οι υπολογιστές να εξάγουν νοήματα από ανθρώπινα ή γλωσσικά δεδομένα, αλλά και την παραγωγή φυσικής γλώσσας.»

3.1.2 ΠΟΙΕΣ ΕΙΝΑΙ ΟΙ ΤΕΧΝΙΚΕΣ ΕΦΓ

Παρακάτω παρατίθενται κάποιες από τις βασικότερες από τις τεχνικές ΕΦΓ:

- Speech recognition (αυτόματη αναγνώριση ομιλίας). Πρόκειται για την δυνατότητα της μηχανής να μετατρέπει τον προφορικό ανθρώπινο λόγο σε γραπτό.
- Tokenization (splitting text into words (tokens) – or sentences-) (διαχωρισμός κειμένου σε λέξεις -ή προτάσεις-). Πρόκειται για μια πολύ συνήθης διαδικασία όταν προεπεξεργάζεται ένα κείμενο για να αναγνωστεί από μια μηχανή, καθώς το συνολικό νόημα μιας πρότασης συνήθως προκύπτει απο την ανάλυση της κάθε λέξης ξεχωριστά.
- Part-of-Speech tagging (επισήμανση μερών του λόγου). Πρόκειται για την διαδικασία κατά την οποία κάθε λέξη (token) της πρότασης κατατάσσεται με βάση το μέρος του λόγου στο οποίο ανήκει.
- Lemmatization and Stemming. Αποτελούν δύο τεχνικές απλοποίησης των λέξεων στις ρίζες τους (γίνεται εκτενέστερη αναφορά και στο κεφάλαιο 5.2). Η δεύτερη τεχνική το επιτυγχάνει αυτό αφαιρώντας κατάλληλα το πρόθυμα ή το επίθυμα απο μία λέξη, ενώ η πρώτη βρίσκει απευθείας μέσω ενσωματωμένου λεξικού που διαθέτει, την ρίζα της κάθε λέξης.
- Stopwords removal (αφαίρεση των stopwords). Πρόκειται για λέξεις όπως «το», «που», «εκεί» και άλλες παρόμοιες οι οποίες συνήθως δεν προσφέρουν κάποια ουσιαστική πληροφορία το κείμενο.Αναφορά στις «stopwords» γίνεται και στο κεφάλαιο 5.2
- Παραγωγή φυσικού λόγου. Είναι η δυνατότητα μιας μηχανής να παράγει φυσικό καθημερινό λόγο.
- Αυτόματη περίληψη ενός κειμένου.

- Ανάλυση Συναισθήματος. Πρόκειται για την τεχνική που χρησιμοποιείται στην παρούσα εργασία και αναλύεται στην επόμενη ενότητα

Σαφώς υπάρχει πλειάδα και άλλων σημαντικών ΕΦΓ τεχνικών, η παρουσίαση των οποίων σε αναλυτικότερο βαθμό θα ξέφευγε από τους σκοπούς της παρούσα διατριβής. Για περισσότερες πληροφορίες ο αναγνώστης μπορεί να επισκεφθεί την σχετική βιβλιογραφία.

3.2 Η ΤΕΧΝΙΚΗ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ (SENTIMENT ANALYSIS)

3.2.1 ΤΙ ΕΙΝΑΙ Η ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΑΙ ΓΙΑΤΙ ΕΙΝΑΙ ΧΡΗΣΙΜΗ

Σε αυτήν την ενότητα θα γίνει μια εμβάθυνση στην τεχνική της Ανάλυσης Συναισθήματος ή ΑΣ (Sentiment Analysis), μιας και ολόκληρη η διατριβή πλαισιώνεται γύρω από αυτήν. Σύμφωνα με το Oxford Languages, «η ανάλυση συναισθήματος είναι η διαδικασία της υπολογιστικής ταυτοποίησης και κατηγοριοποίησης της γνώμης που εκφράζεται σε ένα κομμάτι κειμένου,με στόχο κυρίως τον καθορισμό του κατά πόσο η στάση (γνώμη) του συγγραφέα απέναντι στο εκάστοτε θέμα, προιόν κλπ, είναι θετική, αρνητική ή ουδέτερη».

Πρόκειται για μια τεχνική που βρίσκει τεράστια εφαρμογή στον επιχειρηματικό τομέα, καθώς αποτελεί έναν άμεσο τρόπο για την εξόρυξη της γνώμης των καταναλωτών σχετικά με ένα καινούργιο προϊόν, υπηρεσία ή κάποιο brand. Μια τέτοιου είδους πληροφορία αποτελεί ένα πολύ σημαντικό εργαλείο για μια επιχείρηση, καθώς με σωστή χρήση αυτής της πληροφορίας μπορεί να έχει μια εικόνα του αντίκτυπου που έχει το νέο προϊόν για παράδειγμα στην αγορά, να βελτιώσει τις παροχές της και την σχέση της με τους πελάτες, καθώς και να κάνει προβλέψεις με βάση προηγούμενα δεδομένα για την γνώμη των πελατών σε μελλοντικές παροχές.

Παραδοσιακά, τέτοια δεδομένα μπορούσαν να εξαχθούν από διάφορα surveys (έρευνες) που διεξήγαγε η εκάστοτε εταιρία , σχόλια των πελατών στο site της ή σε κάποιο μπλόγκ, ή μέσω διαφόρων γκάλοπ σε συγκεκριμένες πληθυσμιακές ομάδες. Σήμερα, τα μέσα κοινωνικής δικτύωσης με την μεγάλη εξάπλωση τους αποτελούν μια ανεξάντλητη πηγή δεδομένων και έναν καινούργιο, πολύ υποσχόμενο χώρο άντλησης πληροφορίας. Επειδή όμως, λόγω ακριβώς αυτής της μεγάλης πληθώρας δεδομένων, η προσπέλαση τους απαιτεί την χρήση μηχανής,η ανάπτυξη εξελιγμένων αλγορίθμων εξόρυξης συναισθήματος βρίσκονται στο επίκεντρο πλέον της επιχειρηματικής ανάλυσης.

Εξόρυξη γνώμης μέσω των μέσων κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης, όπως το Twitter γα παράδειγμα, αποτελούν μια ανεξάντλητη πηγή δεδομένων που -μεταξύ άλλων- σχετίζονται με την γνώμη των καταναλωτών σχετικά με διάφορα brands, προιόντα κλπ, μιας και πληθώρα χρηστών το χρησιμοποιούν καθημερινά. Κάθε επιχείρηση θέλει να έχει την δυνατότητα να μπορεί μέσω αυτών των πολυάριθμων δεδομένων φυσικής γλώσσας, να μπορεί να εξάγει πληροφορία, η οποία πληροφορία στην περίπτωση μας να είναι το τι νιώθει ο καταναλωτής για το συγκεκριμένο brand, προιόν κλπ.

Μια πρώτη προσέγγιση, η οποία δεν χρειάζεται και κάποιο εξειδικευμένο εργαλείο,θα ήταν η παρακολούθηση των tweets (έστω η περίπτωση του Twitter) ποσοτικά. Ωστόσο αυτό δεν μπορεί να δώσει κάποια αξιόπιστη πληροφορία, μιας και τα πολλά tweets δεν σημαίνουν αναγκαστικά ότι το εκάστοτε προιόν για παράδειγμα είναι δημοφιλές, μιας και η πλειοψηφία αυτών των tweets μπορεί κάλλιστα να είναι αρνητικά. Φαίνεται λοιπόν πως η ανάπτυξη εργαλείων που να ανιχνεύουν συγκεκριμένα την γνώμη του κοινού είναι απαραίτητη.

Γιατί όμως; Γιατί να θέλει μια εταιρία να γνωρίζει τι γνώμη έχει ο κόσμος για το brand της; Η απάντηση είναι: για πολλούς λόγους.Και αυτοί ακριβώς οι λόγοι παρουσιάζονται αμέσως παρακάτω.

Ο πρώτος λόγος – που είναι και ο πιο προφανής- είναι πως η εκάστοτε εταιρία θέλει να μπορεί να παρακολουθεί την γνώμη του κόσμου για κάποιο προιόν της για παράδειγμα, ώστε να ξέρει αναλόγως αν πρέπει να αλλάξει κάτι στην στρατηγική της σε περίπτωση αρνητικού αντίκτυπου, ή να συνεχίσει να επενδύει στην αντίθετη, θετική περίπτωση.

Μέσω της ανάλυσης συναισθήματος, μια εταιρεία μπορεί να μάθει σε ποιούς τομείς είναι καλή και σε ποιούς πιο αδύναμη. Παράδειγμα αποτελεί η περίπτωση του αεροδρομίου του Heathrow (https://www.sciencedirect.com/science/article/abs/pii/S0969699719300079). Μέσω μιας έρευνας με εργαλεία ανάλυσης συναισθήματος βρέθηκε πως οι πελάτες ήταν ευχαριστημένοι με τις παροχές του αεροδρομίου σχετικά με το wifi, τα μπάνια, τα εστιατόρια και τα σαλόνια ενώ ήταν δυσαρεστημένοι σχετικά με τους χρόνους αναμονής, τα πάρκινγκ, το προσωπικό, τις διαδικασίες για τον ελεγχο του διαβατηρίου κ.α. Με αυτήν την γνώση λοιπόν, μπορεί η συγκεκριμένη εταιρία να στρέψει την καμπάνια της και να πλαισιώσει το brand της γύρω από εκείνους τους τομείς που την ξεχωρίζουν και ικανοποιούν τους πελάτες, ή μπορεί να επικεντρωθεί στα ζητήματα που δυσαρεστούν τους πελάτες και να τα βελτιώσει.

Ένα άλλο πεδίο που η ανάλυση συναισθήματος παίζει καθοριστικό ρόλο είναι αυτό της εξυπηρέτησης πελατών. Όταν μια εταιρία είναι σε θέση να γνωρίζει τι προβληματίζει την εκάστοτε χρονική στιγμή την πλειοψηφία του κοινού της, μπορεί να είναι σε θέση να παρέχει μια πολύ πιο στοχευμένη γραμμή εξυπηρέτησης. Έτσι, αυξάνονται οι πιθανότητες ο πελάτης εν τέλει να εξυπηρετηθεί με τον καλύτερο δυνατό τρόπο και να μείνει εν τέλει ικανοποιημένος.

Τέλος, η σωστή παρακολούθηση (monitoring) στα social media μπορεί πολλές φορές να δείξει κάποιες δυσαρέσκειες του κοινού σχετικά με μία κίνηση μιας εταιρίας, πριν αυτή η δυσαρέσκεια φτάσει να φανεί στις πωλήσεις της. Ένα παράδειγμα αποτελεί η Coca Cola, η οποία πριν κάποια χρόνια έβαλε μια διαφημιστική πινακίδα στο κέντρο του Amsterdam, με της οποίας το μήνυμα αποδείχθηκε ότι διαφωνούσε η πλειοψηφία του κόσμου. Η δυσαρέσκεια αυτή φάνηκε απευθείας στο Twitter, με αποτέλεσμα η Coca Cola να αποσύρει την πινακίδα πριν προκαλέσει γενικότερη δυσαρέσκεια των πελατών, και προτιμήσουν εν τέλει κάποιον ανταγωνιστή της. Εάν η Coca Cola περίμενε το αποτέλεσμα της λάθος αυτής κίνησης να αποτυπωθεί στις πωλήσεις, τότε θα ήταν ενδεχομένως αργά. Είναι γνωστό γενικά στον επιχειρηματικό τομέα ότι λίγα μόνο λάθη αρκούν για να χάσει μια εταιρία τους πελάτες της και να πάνε στους ανταγωνιστές της. Η άμεση ανίχνευση λοιπόν αυτών των λαθών είναι εξέχουσας σημασίας.

3.2.2 ΜΕΘΟΔΟΙ ΥΛΟΠΟΙΗΣΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΤΗΝ PYTHON.

Οι διαθέσιμοι τρόποι για την διεξαγωγή ανάλυσης συναισθήματος χωρίζονται σε δύο διακριτές κατηγορίες,ωτις βασισμένες σε λεξικό τεχνικές (lexicon based ή rule based), και τις τεχνικές μηχανικής μάθησης. Και στις δύο περιπτώσεις, το αρχικό κείμενο πρέπει να περάσει από ένα συγκεκριμένο στάδιο προεπεξεργασίας.Ο τρόπος που λειτουργεί η καθεμία παρουσιάζεται αναλυτικά αμέσως παρακάτω.

Βασισμένες σε λεξικό τεχνικές (lexicon based ή rule based techniques)

Πρόκειται για μια ομάδα τεχνικών που λειτουγούν με βάση ένα αποθηκευμένο λεξικό, όπου η κάθε λέξη φέρει και ένα σκορ που σχετίζεται με το συναίσθημα (polarity) που φέρει ή, αναλόγως το εργαλείο, και την υποκειμενικότητα (subjectivity) της. Έτσι, όταν η μέθοδος καλείται να βρει το συναίσθημα μιας πρότασης, ουσιαστικά συνδυάζει τα ήδη αποθηκευμένα polarities (ή subjectivities) της κάθε μίας, και εξάγει ένα συνολικό σκορ.

Προκειμένου να συμβούν τα παραπάνω, το εκάστοτε κείμενο πρέπει να λάβει μια συγκεκριμένη μορφή το οποίο θα είναι αποτέλεσμα μιας κατάλληλης προεπεξεργασίας. Στα πλαίσια της συγκεκριμένης τεχνικής, η προεπεξεργασία συνήθως περιλαμβάνει τα εξής στάδια:

- Απομάκρυνση σημείων στίξης,λέξεων χωρίς πληροφορία (stopwords - βλέπε σελίδα 7 -), και άλλων περιττών στοιχείων όπως τυχόν ύπαρξη URLs, emojis, αριθμών κλπ
- Tokenization
- Lemmatization ή Stemming

Συνήθως αυτές οι τεχνικές προεπεξεργασίας είναι αρκετές ώστε να μπορέσουν τα lexicon-based εργαλεία να δουλέψουν αποτελεσματικά.

Τέτοια εργαλεία υπάρχουν αρκετά με ποικίλους τρόπους λειτουργίας και επιδόσεις, παρά την κοινή τους βάση, και η επιλογή του καθενός εργαλείου στην εκάστοτε εργασία γίνεται με βάση τα δεδομένα που πρέπει να επεξεργαστούν καθώς και την κρίση και εμπειρία του κάθε αναλυτή. Στην παρούσα εργασία χρησιμοποιούνται δύο εκ των κοινότερων τέτοιων μεθόδων, η TextBlob και η Vader, των οποίων παρουσίαση γίνεται στο κεφάλαιο 4.

Τα μεγάλα θετικά αυτών των μεθόδων είναι δύο. Πρώτον, είναι η απλότητα και αμεσότητα τους όσον αφορά την ελάχιστη προεπεξεργασία που απαιτούν. Μπορούν να εφαρμοστούν άμεσα μιας και δεν χρειάζεται να προηγηθεί οποιαδήποτε εκπαίδευση όπως στα μοντέλα μηχανικής μάθησης. Το δεύτερο θετικό έγκειται στο γεγονός ότι για να βγάλουν αποτέλεσμα, δεν χρειάζονται τίποτα πέρα από τα δεδομένα τα οποία θα επεξεργαστούν. Αυτό έρχεται σε αντίθεση με τα μοντέλα μηχανικής μάθησης, τα οποία πρώτα πρέπει να εκπαιδευτούν σε ένα σετ δεδομένων το οποίο να είναι labeled (με ετικέτες) (να έχουν δηλαδή καταχωρηθεί ετικέτες σε κάθε πρόταση σχετικά με το sentiment τους με κάποιον τρόπο -με τις rule based μεθόδους ή με το χέρι-). Περισσότερες λεπτομέριες περί αυτού,ώστε να γίνει πιο κατανοητό,παρατίθενται αμέσως την συνέχεια, όπου θα αναλυθούν οι τεχνικές με μοντέλα μηχανικής μάθησης.

Το αρνητικό αυτών των μεθόδων είναι ότι εν γένει δεν φέρουν μεγάλες αποδόσεις. Πολλές φορές ωστόσο το ζητούμενο είναι η εξόρυξη μιας γενικής πρώτης εικόνας, και για τέτοιες περιπτώσεις αυτά τα

μοντέλα είνα ιδανικά. Παρ' όλα αυτά οι μέθοδοι αυτές δεν συνίσταται για περιπτώσεις όπου κυριαρχεί καθημερινός λόγος και απαιτείται μεγάλη ακρίβεια.Εκεί, προτιμούνται συνήθως τα καλά εκπαιδευμένα μοντέλα μηχανικής μάθησης. Η χαμηλή αυτή επίδοση είναι σχετικά αναμενόμενη μιας και οι μέθοδοι αυτές πολλές φορές δεν είναι αρκετά προχωρημένες ώστε να επεξεργαστούν σωστά τον καθημερινό λόγο που αποτελείται απο στοιχεία όπως ειρωνία,περίπλοκη άρνηση (complex negation), slang κλπ.

Τεχνικές με μοντέλα μηχανικής μάθησης

Οι τεχνικές αυτές χρησιμοποιούν κατάλληλα εκπαιδευμένα μοντέλα μηχανικής μάθησης, προς εξαγωγή του συναισθήματος του κάθε κειμένου (περισσότερες λεπτομέριες για την μηχανική μάθηση και όλες τις σχετικές διαδικασίες παρουσιάζονται αναλυτικά στο κεφάλαιο 5.5). Τα μοντέλα αυτά εκπαιδεύονται πάνω σε ένα σετ δεδομένων με κείμενα τα οποία είναι ήδη κατηγοριοποιημένα ως Θετικά, Ουδέτερα ή Αρνητικά (Positive,Neutral or Negative). Αυτήν η κατηγοριοποίηση έχει γίνει είτε μηχανικά «με το χέρι» από κάποιον αναλυτή, διαβάζοντας ένα ένα τα κείμενα και κατηγοριοποιώντας τα, ή από κάποια rule-based τεχνική. Αφού εκπαιδευτεί το μοντέλο, έπειτα είναι έτοιμο να χρησιμοποιηθεί σε καινούργια δεδομένα και να τα κατηγοριοποιήσει. Το πόσο καλά ή όχι θα λειτουργήσει, εξαρτάται κατά πολύ από το κατά πόσο ορθά ή όχι είναι κατηγοριοποιημένα τα δεδομένα στο αρχικό σετ εκπαίδευσης.

Όπως θα φανεί και στο κεφάλαιο 5.5 αναλυτικότερα, προκειμένου να μπορέσουν τα δεδομένα μορφής κειμένου να προσπελαθούν από ένα μοντέλο μηχανικής μάθησης, χρειάζονται μια συγκεκριμένη επεξεργασία. Αυτη αποτελείται από τα βήματα της προεπεξεργασίας της περίττωση των rule based μοντέλων, συν ένα ακόμα σημαντικό βήμα, το οποίο είναι η διανυσματοποίηση των δεδομένων σε έναν πίνακα TF – IDF (ή bag of words ή και άλλα). Αυτό γίνεται διότι τα μοντέλα αναγνωρίζουν μόνο αριθμούς, και αυτές οι τεχνικές αποσκοπούν ακριβώς σε αυτό, την μετατροπή δηλαδή των εκάστοτε κειμένων σε μια σειρά από αριθμούς.

Αφού λοιπόν το μοντέλο εκπαιδευτεί, είναι πλέον έτοιμο να κατηγοριοποιεί καινούργια δεδομένα, χωρίς να είναι απαραίτητο κάποιο λεξικό όπως στην προηγούμενη περίπτωση.

Το θετικό με αυτήν την τεχνική είναι πως ένα καλά εκπαιδευμένο μοντέλο που έχει εκπαιδευτεί με ορθώς κατηγοριοποιημένα δεδομένα, είναι πολύ πιο αποδοτικό από τις τεχνικές rule based (Pang et al., 2002), μιας και δεν αναλύει την εκάστοτε πρόταση λέξη-λέξη όπως αυτές, αλλά ως σύνολο. Εάν λοιπόν ένα μοντέλο έχει εκπαιδευτεί σε κείμενα που περιείχαν ειρωνία,περίπλοκη άρνηση και άλλα τέτοια στοιχεία του καθημερινού λόγου, τότε το μοντέλο θα μπορεί και να τα εντοπίζει με μεγάλη ακρίβεια, σε αντίθεση με τις rule based τεχνικές. Ένα άλλο θετικό είναι πως υπάρχει πληθώρα διαθέσιμων αλγορίθμων μηχανικής μάθησης από τους οποίους μπορεί κάθε φορά να επιλέγεται ο αποδοτικότερος.Υπάρχει έτσι μια πληθώρα επιλογών.

Εάν υπάρχει κάποιο αρνητικό με αυτά τα μοντέλα, είναι πως πρέπει αρχικά να εκπαιδευτούν. Χρειάζονται δηλαδή ένα σετ με ήδη κατηγοριοποημένα δεδομένα. Αυτό ωστόσο δεν αποτελεί μεγάλο ζήτημα μιας και τέτοια σετ υπάρχουν παντού στο διαδίκτυο, όπως για παράδειγμα στον ιστότοπο Github.

3.2.3 ΠΡΟΚΛΗΣΕΙΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Ανίχνευση ειρωνίας

Στον καθημερινό λόγο συναντάται πολύ συχνά η χρήση της ειρωνίας, κατά την οποία ουσιαστικά ο ομιλητής λέει κάτι θετικό εννοώντας κάτι αρνητικό ή και κάποιες φορές το αντίστροφο. Σε περιπτώσεις ειδικά όπως αυτή της συγκεκριμένης εργασίας όπου αντλούνται δεδομένα απο πλατφόρμα κοινωνικής δικτύωσης, τέτοια χρήση του λόγου είναι συχνή. Είναι προφανές πως αυτό το φαινόμενο θα κάνει τα μοντέλα να κατατάσσουν ως θετικά διάφορα σχόλια που θα ναι αρνητικά, ή και το αντίστροφο. Η ανάπτυξη τεχνικών ανίχνευσης αυτής τη δομής του λόγου βρίσκονται στην αιχμή του δόρατος, όπως και κάθε άλλη παρόμοια έρευνα που αποσκοπεί στην βελτίωση των μοντέλων ανάλυσης συναισθήματος.

Ανίχνευση περίπλοκης ή έμμεσης άρνησης

Ένα άλλο συχνό φαινόμενο στον καθημερινό λόγο το οποίο δυσκολεύει τα μοντέλα ανάλυσης συναισθήματος είναι η έμμεση άρνηση. Παραδείγματος χάριν, ένα πολύ απλό παράδειγμα είναι η χρήση της φράσης «καθόλου καλό» αντί της λέξης «κακό». Αυτές τις απλές περιπτώσεις, όπου ουσιαστικά είναι μια θετική ή αρνητική λέξη («καλό» στην περίπτωση μας), από την οποία προηγείται μια αρνητική λέξη («καθόλου») τα περισσότερα μοντέλα τις χειρίζονται πλέον σχετικά εύκολα χωρίς να χρειάζεται να είναι ιδιαίτερα εξειδικευμένα. Υπάρχουν όμως και περιπτώσεις τέτοιου είδους άρνησης που δεν είναι εύκολα διαχειρίσιμες, όπως για παράδειγμα η πρόταση «Λίγοι είναι αυτοί που θα χαρακτήριζαν το συγκεκριμένο εστιατόριο καλό». Εκει σαφώς απαιτούνται πολύ εξειδικευμένα μοντέλα και αποτελεί επίσης πεδίο έρευνας.

Ιδιωματισμοί

Πρόκειται για μια ακόμη δομή του λόγου όπου αυτό που διατυπώνεται είναι εντελώς διαφορετικό από αυτό που εννοείται, οπότε και εδώ χρειάζονται μοντέλα εξειδικευμένα όπου κατά την εκπαίδευση τους να έχουν εκτεθεί σε κείμενα που περιέχουν τέτοια στοιχεία ώστα να μπορούν έπειτα να τα αναγνωρίζουν. Σαφώς πρόκειται για μια δουλειά που δεν είναι τόσο εύκολη όσο ακούγεται, μιας και συχνα χρειάζονται περίπλοκες μέθοδοι.

Αυτές είναι κάποιες από τις πιο συνήθεις προκλήσεις που αντιμετωπίζονται κατά την ανάπτυξη αποδοτικών μοντέλων ανάλυσης συναισθήματος. Σαφώς υπάρχουν και άλλες, ανάλογα με το είδος των μοντέλων που θα αναπτυχθούν καθώς και το είδος των δεδομένων που θα κληθούν να αναλύσουν.

4. ΒΑΣΙΚΕΣ ΒΙΒΛΙΟΘΗΚΕΣ

Το περιβάλλον της Python υποστηρίζει την λειτουργία πληθώρας open source (ανοιχτού κώδικα) βιβλιοθηκών, οι οποίες έχουν η κάθε μία την δικιά της χρήση και το δικό της πεδίο εφαρμογών. Σε αυτό το κεφάλαιο θα παρουσιαστούν οι βασικές βιβλιοθήκες που χρησιμοποιήθηκαν στα διάφορα στάδια της συγκεκριμένης διατριβής (εισαγωγή,ανάλυση και παρουσιάση των δεδομένων,ανάπτυξη μοντέλων εξαγωγής συναισθήματος και μοντέλων μηχανικής μάθησης)

4.1 Η ΒΙΒΛΙΟΘΗΚΗ NUMPY

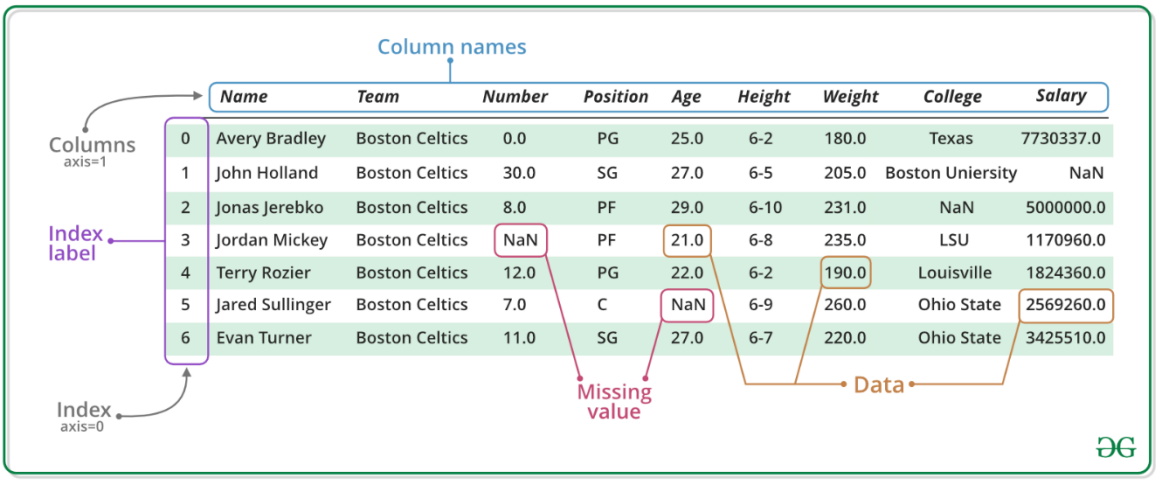
Η Numpy (Numerical Python) είναι μία από τις βασικότερες βιβλιοθήκες της Python. Χρησιμοποιείται σχεδόν σε κάθε επιστημονικό πεδίο προς διεξαγωγή διαφόρων υπολογισμών, καθώς αποτελεί το βασικό εργαλείο της Python που χρησιμοποιείται στον χειρισμό αριθμητικών δεδομένων.Το εύρος χρήσης της είναι τεράστιο, καθώς χρησιμοποιείται από αρχάριους για διεξαγωγή βασικών υπολογισμών,μέχρι από επιστήμονες διαφόρων τομέων προς διεξαγωγή καινοτόμων ερευνών.Το API της Numpy χρησιμοποιείται κατά κόρον στις βιβλιοθήκες Pandas,Matplotlib και scikit-learn,οι οποίες θα χρησιμοποιηθούν εκτενώς στην συγκεκριμένη εργασία - σε αντίθεση με την Numpy αυτήν κάθε αυτήν - ,αλλά και σε άλλες εξίσου σημαντικές βιβλιοθήκες που σχετίζονται με την ανάλυση δεδομένων,όπως η SciPy.

Η λειτουργία της βασίζεται στο βασικό πολυδιάστατο αντικείμενο πίνακα (multidimensional array object) που παρέχει, καθώς και σε μια σειρά γρήγορων υπολογισμών που εφαρμόζει πάνω σε αυτά τα αντικείμενα. Αυτοί οι υπολογισμοί μπορεί να είναι μαθηματικοί, λογικοί (logical) ,διαχείρισης πινάκων, ταξινόμησης, μετατροπές Fourier , βασικής Άλγεβρας ,στατιστικοί και πολλοί άλλοι.

Στη βάση της βιβλιοθήκης είναι το ndarray αντικείμενο. Αυτό περικλείει n – διάστατους πίνακες ομογενών τύπων δεδομένων στους οποίους γίνονται οι παραπάνω χειρισμοί σε μορφή μεταγλωττισμένου κώδικα για καλύτερη απόδοση.Οι πίνακες της NumPy διαφέρουν από άλλες παρόμοιες δομές της Python (όπως οι λίστες) για τους εξής κυρίως λόγους:

- Οι υπολογισμοί στους πίνακες της Numpy γίνονται πολύ πιο αποδοτικά και γρήγορα (ειδικά σε μεγάλες βάσεις δεδομένων) σε σχέση με τις υπόλοιπες παρόμοιες δομές.
- Οι πίνακες έχουν συγκεκριμένο μέγεθος που δεν αλλάζει κατά την διάρκεια των υπολογισμών.Εάν αλλάξει το μέγεθος ενός πίνακα, τότε ο αρχικός διαγράφεται και αντικαθίσταται από τον καινούργιο.
- Οι πίνακες χειρίζονται δεδομενα ίδιου τύπου.

4.2 Η ΒΙΒΛΙΟΘΗΚΗ PANDAS



Εικόνα 4.1 . Παράδειγμα δεδομένων σε μορφή Dataframe.

Η βιβλιοθήκη Pandas είναι το βασικότερο εργαλείο της Python για τον χειρισμό και την ανάλυση δεδομένων, και χρησιμοποιείται εκτενέστατα στην παρούσα έρευνα. Βασική δομή της βιβλιοθήκης είναι τα Dataframes, τα οποία αποτελούν την μορφή με την οποία η βιβλιοθήκη «διαβάζει» - και επεξεργάζεται έπειτα - τα διάφορα δεδομένα.Αυτά τα δεδομένα μπορούν να είναι αρχικά σε διάφορες μορφές αρχείων,όπως CSV (Comma Separated Values), Microsoft Excel, SQL database , JSON ή PARQUET. Κάποιες από τις πιο σύνηθεις και βασικότερες λειτουργίες της βιβλιοθήκης είναι οι εξής:

- Ανάγνωση αλλα και δημιουργία αρχείων διαφόρων μορφών (π.χ Excel)
- Σωστή κατάταξη και στοίχιση δεδομένων (data alignment) και χειρισμός μη καταχωρημένων τιμών (missing values)
- Αλλαγή μορφής (reshaping and pivoting) των σετ δεδομένων (data sets)
- Φιλτράρισμα δεδομένων (data filtering)
- Προσθήκη ή διαγραφή στηλών στα δεδομένα
- Δημιουργία υποσυνόλων ή τεμαχισμός του αρχικού σετ δεδομένων αφαιρώντας γραμμές με βάση τους δείκτες (index)
- Ομαδοποίηση των δεδομένων με βάση κάποιο χαρακτηριστικό για διάφορους σκοπούς (aggregation,transformation)
- Λειτουργίες με Time Series.

4.3 Η ΒΙΒΛΙΟΘΗΚΗ TWEEPY

Η συγκεκριμένη βιβλιοθήκη παρέχει το περιβάλλον ώστε να γίνει η «σύνδεση» της Python με το API του Twitter. Προκειμένου να γίνει αυτό και να μπορέσουν ακολούθως να χρησιμοποιηθούν οι δυνατότητες της βιβλιοθήκης, πρέπει πρώτα ο χρήστης να έχει δημιουργήσει έναν λογαριασμό προγραμματιστή στο Twitter, όπου από εκεί θα προμηθευτεί μια σειρά κωδικών (access keys and tokens),οι οποίοι θα χρησιμοποποιηθούν για την αυθεντικοποίηση του με αποτέλεσμα να μπορεί να χρησιμοποιήσει την βιβλιοθήκη. Περισσότερες πληροφορίες σχετικά με την αυθεντικοποίηση αλλά και το API του Twitter βρίσκονται στο κεφάλαιο 5.1.

Η Tweepy προσφέρει μια πληθώρα διαφόρων δυνατοτήτων,με κάποιες από τις βασικότερες να είναι οι εξής:

- Να εμφανίζει διάφορες πληροφορίες για κάποιον χρήστη του Twitter (όνομα, ψευδώνυμο, τοποθεσία, ακόλουθοι, δημοσιεύσεις κα).
- Να κάνει δημοσιεύσεις απευθείας από τον λογαριασμό του χρήστη (είτε κείμενο, είτε εικόνες και άλλα πολυμέσα)
- Να κάνει διάδραση (interact) με άλλους χρήστες (follow, unfollow κ.α.)
- Να κάνει like, follow, unfollow και retweets.
- Να αντλεί tweets κάποιου χρήστη, ή γενικά με βάση κάποιο hashtag ή λεξεις κλειδιά , τα οποία μετά μπορούν να χρησιμοποιηθούν για διάφορες αναλύσεις.

και πολλές άλλες.

4.4 ΟΙ ΒΙΒΛΙΟΘΗΚΕΣ MATPLOTLIB ΚΑΙ SEABORN

Η οπτικοποίηση των δεδομένων είναι ένα πολύ σημαντικό κομμάτι της εκάστοτε ανάλυσης, διότι κάνει πολλές φορές τα δεδομένα πιο κατανοητά και προσβάσιμα. Ειδικά σε μεγάλες βάσεις δεδομένων, μεγάλο μέρος της διαθέσιμης πληροφορίας μπορεί να φανεί μόνο μέσω κατάλληλων γραφημάτων. Επίσης, η οπτικοποίηση καθίσταται απαραίτητη σε εφαρμογές όπως εύρεση μοτίβων, διακυμάνσεων, προβλέψεων κλπ. Οι βιβλιοθήκες Matplotlib και Seaborn αποτελούν τα κύρια εργαλεία της Python για την οπτικοποίηση των δεδομένων.

Matplotlib

Η βιβλιοθήκη Matplotlib είναι η βασικότερη της Python για την δημιουργία γραφημάτων. Χρησιμοποιεί κατά βάσην την Numpy και την Pandas που παρουσιάστηκαν παραπάνω. Υποστηρίζει μεγάλη πληθώρα γραφημάτων (Bar graphs, Scatterplots, Histogramms, Lines, Pie plots) και χρησιμοποιείται συχνότερα για οπτικοποίηση δυσδιάστατων πινάκων. Η βιβλιοθήκη είναι κατασκευασμένη (από τον John D. Hunter το 2002) με τρόπο ώστε να μοιάζει με το λογισμικό MATLAB.

Seaborn

Η βιβλιοθήκη Seaborn είναι «χτισμένη» πάνω στην Matplotlib αλλά και στις Numpy και Pandas. Η Seaborn μπορεί εν γένει να κάνει ότι και η Matplotlib, είναι όμως πολυ πιο βολική και χρήσιμη όταν χρησιμοποιείται σε συνδυασμό με την Pandas προς οπτικοποίηση διαφόρων χαρακτηριστικων που σχετίζονται με τα Dataframes, καθώς παρέχει ευκολότερες εντολές και πιο ωραία θέματα (themes). Αυτός ακριβώς είναι και ο λόγος που χρησιμοποιείται σε όλη την έκταση της εργασίας.

4.5 TEXTBLOB

Η Textblob είναι μια βιβλιοθήκη της Python που χειρίζεται δεδομένα κειμένου. Πιο συγκεκριμένα, μπορεί και εφαρμόζει σε αυτά τα δεδομένα διάφορες NLP τεχνικές (βλέπε κεφάλιο 3.1) με τρόπο γρήγορο και βολικό προς τον χρήστη. Οι βασικότερες εκ των τεχνικών αυτών είναι οι εξής, κάποιες από τις οποίες παρουσιάστηκαν και στο κεφάλαιο 3.1:

- Noun phrase extraction (εύρεση υποκείμενου της πρότασης)
- Part-of-speech tagging (επισήμανση μέρους του λόγου)
- Sentiment analysis (ανάλυση συναισθήματος)
- Tokenization (splitting text into words and sentences) (διαχωρισμός κειμένου σε λέξεις η προτάσεις)
- Word and phrase frequencies (συχνότητα λέξεων ή φράσεων)
- Parsing (συντακτική ανάλυση)
- **n**-grams
- Word inflection (pluralization and singularization) and lemmatization (εύρεση ενικού ή πληθυντικού μιας λέξης και εύρεση της ρίζας μιας λέξης)
- Spelling correction (ορθογραφικός έλεγχος)
- Add new models or languages through extensions (προσθήκη καινούργιων μοντέλων ή γλωσσών μέσω επεκτάσεων)
- WordNet integration (δίκτυο λέξεων)

Στην περίπτωση της συγκεκριμένης εργασίας χρησιμοποιήθηκε μόνο το Sentiment Analysis (Ανάλυση Συναισθήματος) κομμάτι, κατά το οποίο η Textblob «διαβάζει» μία πρόταση ή ένα κομμάτι κειμένου, και επιστρέφει δύο τιμές (όπως θα φανεί και στο κεφάλαιο 5 επί του πρακτέου),το polarity (πολικότητα) και το subjectivity (υποκειμενικότητα) . Ο πρώτος αριθμός παίρνει τιμές από -1 έως 1 και ισχύει το εξής :

- Εάν polarity < 0 τότε το Sentiment είναι αρνητικό
- Εάν polarity = 0 τότε το sentiment είναι ουδέτερο
- Αλλιώς, το sentiment είναι θετικό

Ο δεύτερος αριθμός παίρνει τιμές από 0 έως 1, όπου το 0 υποδυκνείει πλήρη αντικειμενικότητα στο κείμενο, ενώ το 1 πλήρη υποκειμενικότητα.

Η TextBlob γενικά ανήκει στις lexicon – based (βασισμένες σε λεξικό) sentiment analysis τεχνικές, μιας και ο τρόπος που χρησιμοποιεί για την εξαγωγή των sentiments βασίζεται σε ένα αποθηκευμένο λεξικό (βάση δεδομένων) , όπως περιγράφηκε και στο κεφάλαιο 3.

4.6 Η ΒΙΒΛΙΟΘΗΚΗ NLTK

Η βιβλιοθήκη NLTK (Natural Language Toolkit) αποτελεί ένα ακόμα πακέτο της Python που χρησιμοποιείται για εφαρμογή NLP τεχνικών,αλλά που παρέχει επίσης και μια πληθώρα από διάφορα σετ δεδομένων. Πρόκειται για μια βιβλιοθήκη αρκετά παρόμοια με την TextBlob σχετικά με τις δυνατότητες που προσφέρει, το οποίο είναι αναμενόμενο μιας και η TextBlob έχει «χτιστεί» πάνω στην NLTK. Κάποιες από τις βασικότερες τεχνικές της βιβλιοθήκης είναι οι εξής:

- Filtering stopwords (φιλτράρισμα των «stopwords»)
- Sentiment analysis (ανάλυση συναισθήματος)
- Tokenization (splitting text into words and sentences) (διαχωρισμός κειμένου σε λέξεις η προτάσεις)
- Lemmatizing
- Part-of-speech tagging (εύρεση μέρους του λόγου)

- Word and phrase frequencies (συχνότητα λέξεων ή φράσεων)
- Spelling correction (ορθογραφικός έλεγχος)
- Stemming
- Chunking (εύρεση φράσεων)
- κ.α

Από τις παραπάνω τεχνικές, στην συγκεκριμένη εργασία χρησιμοποιήθηκαν οι 4 πρώτες. Συγκεκριμένα, οι τεχνικές stopwords filtering, tokenizing και lemmatizing χρησιμοποιήθηκαν στο κομμάτι του Data Preprocessing όπως φαίνεται και στο κεφάλαιο 5.2 , ενώ το sentiment analysis χρησιμοποιήθηκε στο ομώνυμο κεφάλαιο 5.4

Όσον αφορά το sentiment analysis, η NLTK διαθέτει την Vader, η οποία αποτελεί επίσης ένα lexicon – based sentiment analysis εργαλείο. Λειτουργεί με τρόπο παρόμοιο με την TextBlob , μόνο που χρησιμοποιεί ελαφρώς διαφορετικά μετρικά.Συγκεκριμένα, για κάθε κείμενο επιστρέφει 4 τιμές, τις Positive, Neutral, Negative και Compound ,οι οποιες σχετίζονται όλες με την πολικότητα (polarity) του κειμένου (δεν εξάγει αποτέλεσμα για το subjectivity). Η παράμετρος compound παίρνει τιμές από -1 έως 1 (-1 για εντελώς αρνητικό sentiment, έως 1 για εντελώς θετικό) και έχει παρόμοιο ρόλο με την παράμετρο polarity της TextBlob, με την διαφορά ότι τα όρια των τιμών τα οποία καθορίζουν αν το κείμενο έχει Positive, Neutral ή Negative sentiment δεν είναι σαφώς καθορισμένα, αλλά ορίζονται από τον χρήστη. Στο κεφάλαιο 5.3 φαίνεται επί του πρακτέου /αυτή η διαδικασία επιλογής ορίων (thresholds). Οι υπόλοιπες τρεις τιμές (0 έως 1 η κάθε μία) δείχουν το ποσοστό του κειμένου που έχει Positive, Neutral ή Negative sentiment αντίστοιχα. Αυτές οι τρεις τιμές αθροίζουν στην μονάδα.

4.7 Η ΒΙΒΛΙΟΘΗΚΗ SCIKIT – LEARN (SKLEARN)

Η scikit – learn είναι μία ανοιχτή βιβλιοθήκη της Python που χρησιμοποιείται στην επιβλεπόμενη και μη μηχανική μάθηση, η οποία παρουσιάζεται αναλυτικά στο κεφάλαιο 5.5 . Συγκεκριμένα, η βιβλιοθήκη προσφέρει :

- Τεχνικές προεπεξεργασίας. Τα δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση και την αξιολόγηση των μοντέλων, πολλές φορές δεν είναι σε μορφή κατάλληλη για να προσπελαθούν από τα μοντέλα μηχανικής μάθησης, όπως για παράδειγμα όταν έχουμε να κάνουμε με δεδομένα μορφής κειμένου. Η sklearn προσφέρει εργαλεία που μπορούν και τα μετατρέπουν σε μορφές κατάλληλες. Εφαρμογή αυτού γίνεται στο κεφάλαιο 5.5

- Μοντέλα μηχανικής μάθησης προς εκπαίδευση. Η βιβλιοθήκη προσφέρει πληθώρα μοντέλων μηχανικής μάθησης (Linear Regression, Support Vector Machines, K Nearest Neighbors κ.α.) τα οποία αρχικά εκπαιδεύονται με την μέθοδο .fit με χρήση δεδομένων εκπαίδευσης (train data) , και έπειτα κάνουν προβλέψεις με την μέθοδο .predict
- Τεχνικές αξιολόγησης των μοντέλων. Η βιβλιοθήκη προσφέρει τεχνικές διαχωρισμού σε δεδομένα εκπαίδευσης και αξιολόγησης (train – test split, cross – validation), καθώς και κάποια μετρικά που μετράνε ποσοτικά το πόσο αποτελεσματικά ή μη είναι τα αποτελέσματα. Περισσότερες λεπτομέριες, επισης στο κεφάλαιο 5.5 .
- Pipelines. Πρόκειται για ένα εργαλείο που ουσιαστικά ενώνει τα μοντέλα και τις τεχνικές προεπεξεργασίας σε ένα αντικείμενο, ώστε όταν καλείται να κάνει και τις δύο εργασίες ταυτόχρονα.
- Αυτόματη εύρεση βέλτιστων παραμέτρων. Κάθε μοντέλο έχει διάφορες παραμέτρους που καθορίζουν την λειτουργία του, οι οποίες ή ορίζονται από τον χρήστη ή παίρνουν ειδάλλως προκαθορισμένες τιμές. Πολλές φορές δεν είναι εύκολο για τον χρήστη να ξέρει ποιές τιμές είναι οι βέλτιστες για την κάθε παράμετρο. Αυτό το πρόβλημα το λύνει η sklearn, προσφέροντας εργαλεία (πχ GridSearchCV, RandomisedSearchCV) που με διάφορους τρόπους μπορούν και βρίσκουν τις βέλτιστες αυτές τιμές.

5. ΔΙΑΔΙΚΑΣΙΑ ΑΝΑΛΥΣΗΣ

5.1 ΑΥΘΕΝΤΙΚΟΠΟΙΗΣΗ –ΣΥΛΛΟΓΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ (ΤΟΥΙΤΣ)

Τα δεδομένα που χρησιμοποιήθηκαν σε αυτήν την εργασία συλλέχθηκαν απευθείας από το Twitter. Προκειμένου να καταστεί αυτό δυνατό, είναι απαραίτητη η δημιουργία ενός λογαριασμού προγραμματιστή (Twitter Developer Account), το οποίο απαιτεί την υποβολή αίτησης και την αποδοχή της από την πλατφόρμα.Έπειτα,μετά την αποδοχή, ο χρήστης προμηθεύεται με κάποια προσωπικά κλειδιά (access keys and tokens) ,τα οποία θα χρειαστούν προκειμένου να πραγματοποιηθεί η πρόσβαση στα tweets μέσω του προγραμματιστικού περιβάλλοντος που χρησιμοποιείται την εκάστοτε φορά, στην περίπτωση μας της Python.

Αυτή η συνεργασία Twitter – Python είναι δυνατή μέσω της χρήσης ενός API (Automatic Programming Interface). Τα APIs, γενικά, επιτρέπουν σε μία εφαρμογή να εξάγει πληροφορίες από ένα κομμάτι λογισμικού και να χρησιμοποιήσει αυτές τις πληροφορίες σε μια άλλη εφαρμογή για διάφορους σκοπούς, όπως για παράδειγμα (που ισχύει και στην περίπτωση μας) για ανάλυση δεδομένων. Έτσι, μιλώντας συγκεκριμένα για την περίπτωση που αφορά αυτήν την εργασία, το API του Twitter δίνει την πρόσβαση στην Python να συλλέγει tweets για ανάλυση ,αλλά και να κάνει δυνητικά και άλλες εργασίες όπως να δημοσιεύει tweets,να κάνει σχόλια σε δημοσιεύσεις τρίτων κ.α. Περισσότερες πληροφορίες σχετικά με το API του Twitter βρίσκονται στην αντίστοιχη βιβλιογραφία..

Η βιβλιοθήκη της Python που παρέχει το περιβάλλον για να γίνουν τα παραπάνω,είναι η Tweepy, για την οποία έχει γίνει λόγος στο Κεφάλαιο 4.3 .Αφού εισαχθεί η βιβλιοθήκη στον κώδικα, το πρώτο βήμα είναι να γραφτούν τα προαναφερθέντα κλειδιά πρόσβασης (access keys and tokens).

Εικόνα 5.1 . Κλειδιά πρόσβασης (access keys and tokens).

Έπειτα μένουν μόνο οι εντολές που φαίνονται στην εικόνα 5.2 , όπου χρησιμοποιούν τα παραπάνω κλειδιά για να κάνουν την αυθεντικοποίηση και την σύνδεση με το API. Τέλος, αρχικοποιείται το API όπως φαίνεται παρακάτω, και όλα είναι έτοιμα πλέον για να μπορέσει ο κώδικας να συλλέξει τα tweets.

```
# Twitter authentication and the connection to Twitter API
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

# Initializing Tweepy API
api = tweepy.API(auth, wait_on_rate_limit=True)
```

Εικόνα 5.2 . Αυθεντικοποίηση και σύνδεση με το API .

Η συλλογή των τουίτς γίνεται με τμήμα κώδικα στο οποίο πρέπει να δηλωθούν, με κατάλληλο τρόπο, διάφορες παράμετροι που αφορούν το περιεχόμενο και την δομή των tweets.

Η αρχική και κυριότερη παράμετρος που πρέπει να δηλωθεί είναι οι λέξεις-κλειδιά τις οποίες θα πρέπει να έχουν τα tweets που θα συλλεχθούν. Αυτές οι λέξεις-κλειδιά θα είναι ή απλές λέξεις που θα περιέχουν τα tweets (keywords) η θα είναι κάποια hashtags. Στην έρευνα αυτή επιλέχθηκε το keyword <<Ukraine>> . Οπότε τα tweets που θα αντληθούν θα περιέχουν οπωσδήποτε την λέξη <<Ukraine>> , είτε ως hashtag είτε μέσα στον κορμό τους.

Προφανώς,τα τουίτς που συλλέγονται δεν περιέχουν μόνο το κείμενο (text),αλλά και κάποια άλλα χαρακτηριστικά (attributes) τα οποία στην περίπτωση της συγκεκριμένης εργασίας είναι τα εξής :

- Η ημερομηνία δημιουργίας του tweet.
- Το όνομα του χρήστη (username) που το δημοσίευσε.
- Το σύνολο των hashtags μέσα στο tweet.
- Την τοποθεσία του χρήστη.
- Ο αριθμός των ακολούθων (followers) του χρήστη.
- Ο αριθμός των retweets του συγκεκριμένου tweetμέχρι και την στιγμή που συλλέχθηκε.
- Ο αριθμός των likes τουσυγκεκριμένου tweetμέχρι και την στιγμή που συλλέχθηκε.

Μέχρι στιγμής σε αυτήν την εργασία, αλλά και στην υπόλοιπη μέχρι το πέρας της, όταν αναφερόμαστε σε ένα τουίτ εννοούμε σαφώς το κείμενο (text).Το επόμενο λοιπόν μεγάλο βήμα είναι να δηλωθούν στον κώδικα αυτά τα χαρακτηριστικά (attributes) που θα συνοδεύουν το κάθε tweet.Έπειτα υπάρχει η δυνατότητα να δηλώσει ο προγραμματιστής τις ημερομηνίες από τις οποίες θα αντληθούν τα δεδομένα. Αυτήν η εργασία λοιπόν, άντλησε tweets από τις 24/02/2022 έως και τις 27/02/2022,που είναι φυσικά οι τέσσερις πρώτες μέρες του πολέμου Ρωσίας – Ουκρανίας.

Τέλος μένει να οριστεί ο αριθμός των tweets που θα συλλεχθούν από την κάθε μέρα.Ο αριθμός αυτός στην συγκεκριμένη περίπτωση ήταν 50.000 για την κάθε μία από τις τέσσερις ημέρες.Ένας αριθμός ιδανικός, που να συνδυάζει ταυτόχρονα μια πληθώρα δεδομένων για εξαγωγή ασφαλών συμπερασμάτων, με τον εύκολο σχετικά χειρισμό τους από το υπολογιστικό σύστημα που χρησιμοποιήθηκε (από άποψη διαθέσιμης μνήμης και επεξεργαστικής ισχύος). Τα δομικά στοιχεία του υπολογιστή παρουσιάζονται στο παράρτημα.

Αφού λοιπόν ο αλγόριθμος σύλλεξε τα δεδομένα (χρειάστηκαν προσεγγιστικά 20-25 ώρες), αυτά αποθηκεύτηκαν σε ένα αρχείο Microsoft Excel. Πλέον, όλα είναι έτοιμα για να εισαχθεί η βιβλιοθήκη Pandas (βλέπε Κεφάλαιο 4.3), η οποία θα ‘διαβάσει’ τα δεδομένα από τα προαναφερθέντα αρχεία, και θα τα διαθέσει σε μορφή DataFrame (βλέπε εικόνα 5.4). Θα έχουμε τέσσερα DataFrames , ένα για κάθε αρχείο,τα οποία θα είναι τα εξής :

- twt_27_02_pd για τα δεδομένα του αρχείου με τα 50.000 tweets από τις 27/02/2022
- twt_26_02_pd για τα δεδομένα του αρχείου με τα 50.000 tweets από τις 26/02/2022
- twt_25_02_pd για τα δεδομένα του αρχείου με τα 50.000 tweets από τις 25/02/2022
- twt_24_02_pd για τα δεδομένα του αρχείου με τα 50.000 tweets από τις 24/02/2022

Τα δεδομένα μας σε αυτό το σημείο είναι έτοιμα για το επόμενο πολύ σημαντικό βήμα, την προεπεξεργασία (Data Preprocessing).

	timestamp	tweet_text	username	all_hashtags	location	followers_count	retweet_count	favorite_count
0	2022-02-27 23:59:59	b"Rep. Lauren Boebert Says Canada And U.S. 'Ne...	b'kahornback1'	[]	New Jersey, USA	0	0	0
1	2022-02-27 23:59:59	b"@JennaEllisEsq This is so typical of republi...	b'BBullard007'	[]	Pittsburgh, PA	374	1	0
2	2022-02-27 23:59:59	b"I don't remember any war or any political is...	b'Katlysss'	[]	Deutschland	4383	4	147
3	2022-02-27 23:59:59	b'@terranova_billy @JonathanHoenig Is the "fin...	b'1stcitizen'	['OurManVlad', 'Ukraine', 'BidenHarris', 'IAsk...	Baltimore, MD	2326	0	1
4	2022-02-27 23:59:59	b'Ep. 102: Russia Invades Ukraine? Canada Off ...	b'Have2doxies'	[]	Southern midwest	1224	0	0

Εικόνα 5.3 . DataFrame με τα 5 πρώτα στοιχεία του αρχείου που περιέχει τα tweets από την ημερομηνία 27/02/2022.

5.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΤΩΝ ΔΕΔΟΜΕΝΩΝ.

Η διαδικασία που θα περιγραφτεί σε αυτήν την ενότητα είναι εξέχουσας σημασίας σε ότι έχει να κάνει με ανάλυση δεδομένων μορφής κειμένου, και πρόκειται για το στάδιο της προεπεξεργασίας (data preprocessing).Στην περίπτωση ειδικά των δεδομένων που αντλούνται από πλατφόρμες κοινωνικής δικτύωσης, όπου τα κείμενα περιέχουν καθημερινό λόγο, με την αργκό να κυριαρχεί, συνάμα με την ύπαρξη διαφόρων emojis, URLs αυτήν η διαδικάσια καθίσταται επιτακτική.

Ο στόχος είναι δηλαδή, το προς ανάλυση κείμενο να μεταβεί σε μια μορφή όσο το δυνατόν πιο απλή και λιτή, ώστε να μπορεί να 'διαβαστεί' από τον εκάστοτε αλγόριθμο ως προς την εξαγωγή πληροφορίας. Σαφώς, κατά την διάρκεια αυτής της επεξεργασίας ή 'απλοποίησης', το κείμενο θα πρέπει να χάνει καθόλου η έστω ελάχιστο από τον αρχικό όγκο πληροφορίας που φέρει.

Οι λόγοι για τους οποίους η ανάλυση πρέπει να γίνει στο επεξεργασμένο κείμενο και όχι απευθείας στο αρχικό,είναι κυρίως δύο. Ο πρώτος και σημαντικότερος λόγος είναι ότι έτσι αυξάνεται η αποδοτικότητα των μεθόδων που χρησιμοποιούνται για την ΑΣ (αλλά και των υπολοίπων NLP τεχνικών), μιας και ο αλγόριθμος της μεθόδου δεν 'ασχολείται' με περιττές λέξεις και σύμβολα, αυξάνοντας έτσι την ακρίβεια του (accuracy).Ο δεύτερος λόγος έχει να κάνει με την μνήμη και την ταχύτητα εκτέλεσης. Είναι προφανές ότι η εκάστοτε μέθοδος θα δουλέψει, εκτός από πιο αποδοτικά, και πιο γρήγορα πάνω σε ένα κείμενο σωστά επεξεργασμένο.

Για να γίνει σαφής η αναγκαιότητα της παραπάνω διαδικασίας επιλέγεται ένα τυχαίο tweet:

@Username_1 This frightening war is getting worse and worse everyday!!!!!!!!!! If you want to help you can donate here https://www.url.com #Ukraine#Together_for_Ukraine#War

Σε αυτό το τουίτ λοιπόν, υπάρχουν πολλά στοιχεία που δίνουν πληροφορία, αλλά και πολλά άλλα που δεν δίνουν. Σαφώς, όταν αναφερόμαστε στην λέξη πληροφορία, εννοούμε σχετικά με το αντικείμενο που μας ενδιαφέρει, το οποίο στην παρούσα περίπτωση είναι το συναίσθημα (sentiment) του χρήστη.

Ξεκινώντας από τα αχρείαστα στοιχεία λοιπόν, σε αυτά συγκαταλέγονται αμέσως τα mentions (@Username_1) και τα URLs (https://www.url.com).'Επειτα πρέπει να εξεταστεί η χρησιμότητα (ή μη) των hashtags. Κάποιες φορές τα hashtags ενδέχεται να περιέχουν κάποια πληροφορία (#Together_for_Ukraine) ,αλλά τις περισσότερες φορές περιέχουν απλά γενικότητες και θόρυβο,οπότε σχεδόν πάντα σε τέτοιες περιπτώσεις επιλέγεται η απομάκρυνση τους.

Ο κατάλογος των μη χρήσιμων στοιχείων όμως δεν τελειώνει στα προαναφερθέντα. Πολλές λέξεις είναι συχνά γραμμένες στον καθημερινό λόγο με τρόπο πιο πολύπλοκο από όσο χρειάζεται για μια μηχανή. Τι ορίζεται όμως ως πολύπλοκος; Πολυπλοκότητα λοιπόν προσφέρει ο οποιοσδήποτε άλλος τρόπος γραφής της λέξης πέρα από τον τελείως απλούστερο.

Για να γίνει κατανοητό αυτό, ας πάρουμε την λέξη worse από το tweet του παραδείγματος. Η λέξη worse προέρχεται από την λέξη bad. Αντίστοιχα,η λέξη frightening προέρχεται από το frighten. Όλες οι λέξεις λοιπόν πρέπει με παρόμοιο τρόπο να απλοποιηθούν στις βασικότερες δυνατές. Ο λόγος που καθιστά αυτήν την διαδικασία απαραίτητη θα φανεί αμέσως παρακάτω, και σχετίζεται με το TF-IDF vectorization κυρίως, όπως αναλύεται στην συνέχεια.

Η προαναφερθείσα τεχνική έχει κυρίαρχο ρόλο στην εκπαίδευση μοντέλων μηχανικής μάθησης –και όχι μόνο-, και ως εκ τούτου η λειτουργία της πρέπει να καταστεί απολύτως κατανοητή πριν γίνει οποιαδήποτε περαιτέρω ανάλυση.Ουσιαστικά,το TF-IDF vectorization μετατρέπει ένα σύνολο λέξεων σε αριθμούς, οι οποίοι σχετίζονται με την σημαντικότητα της κάθε λέξης.

Πως βρίσκει όμως ο εν λόγω αλγόριθμος την σημαντικότητα της κάθε λέξης; Αρχικά ας ξεκινήσουμε από τον όρο TF (Term Frequency) της κάθε λέξης. Ο όρος αυτός είναι ένα απλό κλάσμα, που ως αριθμητή έχει τον αριθμό των φορών που εμφανίζεται η κάθε λέξη (word) στο κάθε κείμενο (document –στην περίπτωση την συγκεκριμένη στο κάθε τουίτ),προς τον συνολικό αριθμό λέξεων του κειμένου αυτού.

$$TF(w,d) = \frac{occurences\ of\ w\ in\ document\ d}{total\ number\ of\ words\ in\ document\ d}$$

Ένα απλό παράδειγμα φαίνεται στις εικόνες X , X

Documents	Text	Total number of words in a document
A	Jupiter is the largest planet	5
B	Mars is the fourth planet from the sun	8

Εικόνα 5.4 . Παράδειγμα υπολογισμού του TF-IDF vector.

Words	TF (for A)	TF (for B)
Jupiter	1/5	0
Is	1/5	1/8
The	1/5	2/8
largest	1/5	0
Planet	1/5	1/8
Mars	0	1/8
Fourth	0	1/8
From	0	1/8
Sun	0	1/8

Εικόνα 5.5 . Πίνακας TF .

Έπειτα είναι ο όρος IDF της κάθε λέξης, ο οποίος είναι ο Νεπέριος λογάριθμος του κλάσματος με αριθμητή των συνολικό αριθμών των κειμένων και παρονομαστή τον αριθμό εκείνων των κειμένων που περιέχουν την συγκεκριμένη λέξη. Ο αριθμός αυτός αποτελεί ουσιαστικά ένα μέτρο της σπανιότητας, επομένως και της σημαντικότητας, της κάθε λέξης.

$$IDF(w, D) = \ln(\frac{Total\ number\ of\ documents\ (N)\ in\ corpus\ D}{number\ of\ documents\ containing\ w})$$

Στην εικόνα X φαίνεται ένα παράδειγμα υπολογισμού του IDF

Words	TF (for A)	TF (for B)	IDF
Jupiter	1/5	0	ln(2/1) = 0.69
Is	1/5	1/8	ln(2/2) = 0
The	1/5	2/8	ln(2/2) = 0
largest	1/5	0	ln(2/1) = 0.69
Planet	1/5	1/8	ln(2/2) = 0
Mars	0	1/8	ln(2/1) = 0.69
Fourth	0	1/8	ln(2/1) = 0.69
From	0	1/8	ln(2/1) = 0.69
Sun	0	1/8	ln(2/1) = 0.69

Εικόνα 5.6 . Παράδειγμα υπολογισμού του IDF.

Τέλος, οι δύο αυτοί αριθμοί πολλαπλασιάζονται για κάθε λέξη και για κάθε έγγραφο, δίνοντας τον τελικό πίνακα ή διάνυσμα TF-IDF,όπως φαίνεται και στην εικόνα X.

$$TFIDF\ (w, d, D) = TF(w, d) * IDF(w, D)$$

Words	TF (for A)	TF (for B)	IDF	TFIDF (A)	TFIDF (B)
Jupiter	1/5	0	ln(2/1) = 0.69	0.138	0
Is	1/5	1/8	ln(2/2) = 0	0	0
The	1/5	2/8	ln(2/2) = 0	0	0
largest	1/5	0	ln(2/1) = 0.69	0.138	0
Planet	1/5	1/8	ln(2/2) = 0	0.138	0
Mars	0	1/8	ln(2/1) = 0.69	0	0.086
Fourth	0	1/8	ln(2/1) = 0.69	0	0.086
From	0	1/8	ln(2/1) = 0.69	0	0.086
Sun	0	1/8	ln(2/1) = 0.69	0	0.086

Εικόνα 5.7 . Τελικός πίνακας (ή διάνυσμα) TF-IDF.

Επανερχόμαστε τώρα στο σημείο της απλούστευσης των λέξεων στην ρίζα τους (Lemmatizing). Αυτήν η διαδικασία έχει κάποια αρνητικά και κάποια θετικά στοιχεία.Το μεγάλο θετικό είναι πως μειώνει το πλήθος των λέξεων στον TF-IDF πίνακα, καθώς σε άλλη περίπτωση, η κάθε μορφή - πέρα της εντελώς βασικής της ρίζας- με την οποία θα εμφανιζόταν η κάθε λέξη στα tweets, θα καταλάμβανε και μια διαφορετική θέση. Οπότε γίνεται αντιληπτό πως όταν ο πίνακας πρέπει να έχει έναν μέγιστο αριθμό λέξεων για λόγους μνήμης, όπως και γίνεται σε όλους τους TF-IDF πίνακες αυτής της έρευνας, μας συμφέρει να μην υπάρχουν πολλές λέξεις με την ίδια ρίζα, ώστε να αποτυπωθεί όση περισσότερη πληροφορία είναι δυνατόν.

Ωστόσο υπάρχουν και κάποιοι λόγοι που καθιστούν ορισμένες φορές το Lemmatizing μη αναγκαίο. Αυτό σχετίζεται κυρίως με το γεγονός ότι οι lexicon based sentiment analysis τεχνικές που χρησιμοποιούνται, η TextBlob και η Vader, δίνουν διαφορετικό polarity και compound score αντίστοιχα, στις διαφορετικές παράγωγες λέξεις που έχουν ίδια ρίζα. Συγκεκριμένα δηλαδή, θα δοθεί διαφορετικό σκορ –αρνητικό μεν - στις λέξεις worse,worst και bad για παράδειγμα. Ωστόσο αυτές οι διαφορές είναι σχετικά μικρές, οπότε στην συντριπτική πλειοψηφία των περιπτώσεων, όταν κάποιο τουίτ πρόκειται να κατηγοριοποιηθεί ως αρνητικό, θα κατηγοριοποιηθεί ως αρνητικό και στις δύο περιπτώσεις,είτε γίνει Lemmatizing δηλαδή είτε όχι. Απλά ενδεχομένως να προκύψει ελαφρώς διαφορετικό polarity και compound score,το οποίο ακριβές σκορ δεν μας αφορά αυτό καθαυτό, παρά μόνο το τελικό label (positive,neutral,negative).

Όπως και να ‘χει, η χρήση ή μη του Lemmatization επαφίεται στην επιλογή του κάθε αναλυτή. Εδώ, όπως και στην πλειοψηφία παρόμοιων αναλύσεων, επιλέχθηκε να γίνει η χρήση του, διότι τα TF-IDF μοντέλα χρησιμοποιούνται εκτενώς, και απαιτείται η καλή ακρίβεια (accuracy) τους, ώστε να προκύψουν εν τέλει και ακριβή μοντέλα μηχανικής μάθησης.

Μέχρι στιγμής λοιπόν, έχει γίνει λόγος για τρόπους που μας επιτρέπουν να φέρουμε ένα κείμενο σε μια μορφή όσο το δυνατόν απλούστερη, ώστε να μπορεί η μηχανή (ο υπολογιστής ή ο εκάστοτε αλγόριθμος) να το επεξεργαστεί όσο το δυνατόν πιο αποδοτικά. Στη συνέχεια της ενότητας αυτής θα παρουσιαστεί ο τρόπος που εφαρμόστηκαν οι παραπάνω τεχνικές, καθώς και κάποιες ακόμα, ώστε να μετατραπούν τα τουίτς σε μια μορφή ιδανική,που να μπορούν να δεχθούν περαιτέρω ανάλυση.

Αρχικά λοιπόν, πριν αρχίσει η επεξεργασία των τουίτς αυτών καθαυτών, καλό είναι να απομακρυνθούν κάποιες στήλες από τα Dataframes που δημιουργήθηκαν στην ενότητα 5.1 . Πρόκειται για κάποιες στήλες

τις οποίες εξαρχής γνωρίζουμε πως δεν θα χρειαστούμε στην ανάλυση ("username", "all_hashtags","followers_count"), οπότε και διαγράφονται.

Επιπλέον, είναι σημαντικό να διαγραφούν οι τυχόν πανομοιότυπες γραμμές στο Dataframe (οι γραμμές που περιέχουν τα ίδια τουίτς δηλαδή), αν και αυτές αναμένεται να είναι ελάχιστες, μιας και κατά την συλλογή τους ρυθμίστηκε να μην συλλεχθούν και τα αντίστοιχα ριτουίτς (retweets) του καθενός, το οποίο θα ήταν και η κυριότερη πηγή διπλοτύπων γραμμών (duplicates).Έπειτα, φροντίζουμε να μορφοποιήσουμε την στήλη <<timestamp>>, με χρήση κατάλληλης εντολής από την βιβλιοθήκη Pandas, ώστε οι ώρες και οι ημερομηνίες των tweets να γραφούν σε μορφή κατάλληλη για να διαβαστούν αργότερα, εάν καταστεί αυτό αναγκαίο.

Τέλος, αφού παρατηρήσουμε ότι η στήλη της τοποθεσίας “location” έχει πολλά στοιχεία Nan, το οποίο στην Python σημαίνει ότι η συγκεκριμένη τιμή δεν υπάρχει (missing values),φροντίζεται να αντικατασταθούν με την λέξη unknown. Αυτό γίνεται διότι η λέξη unknown σε αντίθεση με την ένδειξη NaN είναι διαχειρίσιμη, το οποίο αργότερα θα χρειαστεί όταν θα γίνει σχετική ανάλυση με τις τοποθεσίες.

Πλέον όλα είναι έτοιμα για να λάβει τόπο το κυριότερο μέρος του preprocessing, αυτό που έχει να κάνει με τα tweets αυτά καθαυτά. Αρχικά, δημιουργείται μια δεύτερη και πανομοιότυπη στήλη με τα tweets (<<tweet.text>>) ,η οποία θα είναι και η στήλη που θα γίνει το preprocessing.Την ονομάζουμε "text_preprocessed_tokenized" .Ο λόγος που γίνεται αυτό είναι για να εξακολουθούν να υπάρχουν τα tweets και στην αρχική τους μορφή, κυρίως για λόγους σύγκρισης με τα επεξεργασμένα.

Μέσω της δημιουργίας καταλλήλου τμήματος κώδικα, όλα τα tweets και από τα 4 DataFrames (twt_24_02_pd, twt_25_02_pd, twt_26_02_pd, twt_27_02_pd) υπόκεινται στις εξής εργασίες:

- Αντικαθίσταται το hashtag #Ukraine με σκέτη τη λέξη Ukraine. Αυτό συμβαίνει διότι σε επόμενο βήμα θα διαγραφούν όλες οι λέξεις με hashtag (ο λόγος εξηγήθηκε παραπάνω) ,αλλά η συγκεκριμένη λέξη φέρει πληροφορία για την ανάλυση μας και επιλέγουμε να την κρατήσουμε.
- Αφαιρούνται όλα τα mentions και hashtags
- Αφαιρείται ο χαρακτήρας \n ο οποίος παρουσιάζεται στην πλειοψηφία των tweets και πρόκειται σαφώς για θόρυβο
- Αφαιρούνται τα σημεία στίξης
- Αφαιρούνται οι λέξεις <<stopwords>> . Stopwords είναι γενικά οι λέξεις που δεν φέρουν καμία σημασία σε αναλύσεις σαν αυτή της συγκεκριμένης έρευνας (text classification). Παραδείγματος χάριν, τέτοιες λέξεις είναι οι the, this, that, he, when και άλλες, οι οποίες σαφώς δεν παίζουν κάποιο ρόλο στην κατηγοριοποίηση του τουίτ ως θετικό, αρνητικό η ουδέτερο, οπότε και απομακρύνονται. Η βιβλιοθήκη NLTK προσφέρει έναν έτοιμο κατάλογο με τέτοιες λέξεις.
- Γίνεται το lemmatization,το οποίο αναλύθηκε εκτενώς παραπάνω. Η βιβλιοθήκη NLTK παρέχει την δυνατότητα να κάνει lemmatization.
- Γίνεται tokenization.Αυτή η τεχνική, η οποία επίσης παρέχεται από την βιβλιοθήκη NLTK, ουσιαστικά χωρίζει το κάθε tweet από ένα συνεχές κείμενο, σε μια λίστα λέξεων. Αυτό στην περίπτωση μας γίνεται διότι έτσι διευκολύνεται η διαδικασία απομάκρυνσης περιττών λέξεων,αλλά και η συνολική διαδικασία της προεπεξεργασίας.
- Τέλος, απομακρύνονται όλες οι λέξεις με αριθμό γραμμάτων μικρότερο ή ίσο του 2.Είναι σαφές πως καμία τέτοια λέξη δεν παρέχει κάποια πληροφορία.

Σε αυτό το σημείο, τα πρωτότυπα tweets που ήταν σε μορφή κειμένου (strings στην ορολογία της Python), έχουν μετατραπεί σε μία λίστα διαφορετικών λέξεων, όπου η καθεμία αποτελεί και ένα

string,το οποίο αποτελεί προιόν του tokenization.Το τελευταίο βήμα πριν την συνέχεια στο exploratory data analysis κομμάτι, είναι να μετατρέψουμε αντίστροφα την λίστα από strings, όπου το κάθε string είναι όπως αναφέραμε και μία λέξη, σε ένα ενιαίο string το οποίο θα είναι το επεξεργασμένο πλέον tweet. Αυτό γίνεται διότι η tokenized μορφή ήταν χρήσιμη μόνο στο στάδιο του preprocessing.Στην εικόνα 5.8 φαίνεται το καινούργιο DataFrame «twt_27_02_pd» (χρησιμοποιείται ενδεικτικά το ένα από τα τέσσερα, μιας και οι διαδικασίες –άρα και τα αποτελέσματα- σε όλα τα DataFrames ήταν πανομοιότυπα),έπειτα από όλες τις εργασίες που περιγράφηκαν στο παρόν κεφάλαιο.

	timestamp	tweet_text	location	retweet_count	favorite_count	text_preprocessed_tokenized	text_preprocessed_tokenized(str)
0	2022-02-27 23:59:59	b"Rep. Lauren Boebert Says Canada And U.S. 'Ne...	New Jersey, USA	0	0	[lauren, boebert, says, canada, liberated, lik...	lauren boebert says canada liberated like ukra...
1	2022-02-27 23:59:59	b"@JennaEllisEsq This is so typical of republi...	Pittsburgh, PA	1	0	[typical, republicans, country, pulling, toget...	typical republicans country pulling together c...
2	2022-02-27 23:59:59	b"I don't remember any war or any political is...	Deutschland	4	147	[remember, war, political, issue, entire, worl...	remember war political issue entire world unit...
3	2022-02-27 23:59:59	b'@terranova_billy @JonathanHoenig Is the "fin...	Baltimore, MD	0	1	[financial, calamity, sojourn, white, house]	financial calamity sojourn white house
4	2022-02-27 23:59:59	b'Ep. 102: Russia Invades Ukraine? Canada Off ...	Southern midwest	0	0	[russia, invades, ukraine, canada, rails, olym...	russia invades ukraine canada rails olympics a...

Εικόνα 5.8 . Το Dataframe twt_27_02_pd μετά το πέρας της προεπεξεργασίας.

5.3 ΔΙΕΡΕΥΝΗΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Το exploratory data analysis είναι ένα μείζον κομμάτι της ανάλυσης δεδομένων, και συνήθως είναι από τα πρώτα που γίνονται –μαζί με την προεπεξεργασία-, προτού ακολουθήσει οποιαδήποτε περαιτέρω εξειδικευμένη εργασία στα δεδομένα. Αυτό συμβαίνει διότι χωρίς την βαθιά κατανόηση του dataset (δομή δεδομένων) που επεξεργαζόμαστε, της δομής του, των βασικών στοιχείων του, αλλά και των βασικών πληροφοριών που μπορούν να αντληθούν αμέσως από αυτό, καμία περαιτέρω ανάλυση δεν είναι εφικτή.

Σε αυτό το κομμάτι της ανάλυσης, η βιβλιοθήκη Pandas –η οποία παρουσιάστηκε παραπάνω- , έχει τον κυρίαρχο ρόλο. Οι διάφορες δυνατότητες που προσφέρει την κάνουν το ιδανικό εργαλείο,όσον αφορά το περιβάλλον της Python, αλλά και από τα ιδανικότερα γενικά, όταν πρόκειται για ανάλυση και διαχείριση των δεδομένων.

Ένα από τα πρώτα πρώτα πράγματα που πρέπει να γίνουν, πριν καν το στάδιο της προεπεξεργασίας, είναι η κατανόηση της δομής των διαφόρων Dataframes που έχουν δημιουργηθεί. Με μια απλή εντολή η Pandas εμφανίζει όλες τις βασικές παραμέτρους ενός Dataframe, τις οποίες πρέπει να έχουμε υπόψιν. Στην εικόνα 5.9 παρακάτω φαίνεται ένα τέτοιο παράδειγμα, ξανά ενδεικτικά για το Dataframe <<twt_27_02_pd>>

```
twt_27_02_pd.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 8 columns):
timestamp          50000 non-null object
tweet_text         50000 non-null object
username           50000 non-null object
all_hashtags       50000 non-null object
location           31291 non-null object
followers_count    50000 non-null int64
retweet_count      50000 non-null int64
favorite_count     50000 non-null int64
dtypes: int64(3), object(5)
memory usage: 3.1+ MB
```

Εικόνα 5.9 . Οι βασικές πληροφορίες ενός Dataframe (εδώ, του <<twt_27_02_pd>>) .

Οι πληροφορίες που μας δίνει αυτή η εντολή, όπως φαίνεται στην εικόνα, είναι ο αριθμός των στηλών (columns) και των γραμμών (range index), τα ονόματα των στηλών ,ο τύπος δεδομένων που περιέχει η καθεμία καθώς επίσης και ο αριθμός των δεδομένων της κάθε στήλης που δεν είναι null (ή NaN, ή missing value,ή μη καταχωρημένη τιμή). Επίσης μας δίνεται και ο χώρος στην μνήμη που καταλαμβάνει το Dataframe.

Με βάση αυτές τις πληροφορίες, δύο είναι τα σημεία που συνήθως ενδιαφέρουν περισσότερο και χρήζουν περισσότερης προσοχής. Το πρώτο είναι το κατά πόσο ο τύπος των δεδομένων της κάθε στήλης είναι στην μορφή που πρέπει και επιθυμούμε να είναι. Για παράδειγμα μια στήλη που περιέχει αριθμούς, πρέπει να βεβαιωθούμε ότι οι αριθμοί είναι περασμένοι στο σωστό φορμάτ, int64 δηλαδή, όπως η στήλη <<followers_count>> για παράδειγμα. Αν τυχόν δεν είναι έτσι, στην συνέχεια θα προκύψουν προβλήματα με την επεξεργασία των δεδομένων της συγκεκριμένης στήλης.

Το δεύτερο σημείο που χρήζει ιδιαίτερη προσοχή , είναι ο εντοπισμός και κατάλληλος χειρισμός των μη καταχωρημένων τιμών (NaN) . Σε κάποιες περιπτώσεις επιλέγεται οι στήλες ή οι γραμμές που έχουν τέτοιες τιμές, να διαγράφονται. Σε κάποιες άλλες, το γεγονός ότι υπάρχει missing value μπορεί να αποτελεί ένα είδος πληροφορίας και καλό είναι να μένει ως έχει. Στην δικιά μας περίπτωση εφαρμόζεται, όπως φάνηκε και στο κεφάλαιο του preprocessing, η αντικατάσταση των τιμών NaN στην στήλη <<location>>με την λέξη unknown, όπως εξηγήθηκε και πιο πάνω.

Το κυρίως κομμάτι του exploratory data analysis στην παρούσα έρευνα, αφιερώνεται στην διερεύνηση δύο βασικών παραμέτρων. Η πρώτη παράμετρος η οποία αναλύεται προς διεξαγωγή πληροφορίας, είναι η τοποθεσία («location») ,και η δεύτερη έχει να κάνει με την εύρεση των πιο πολυχρησιμοποιημένων λέξεων στον κορμό των τεσσάρων αρχείων με τα tweets.

Ξεκινώντας από τις τοποθεσίες, κύριος στόχος είναι μέσω διαγραμμάτων να παρουσιαστούν οι πιο δημοφιλείς τοποθεσίες-χώρες, καθώς και να παρθούν και άλλες πληροφορίες σχετικά το πλήθος των χωρών, όπως θα φανεί και αμέσως στην συνέχεια.Πρόκειται για μια διαδικασία που παρέχει σημαντική πληροφορία, ειδικά αν σκεφτούμε το παράδειγμα μιας εταιρίας που έχει λανσάρει ένα καινούργιο προιόν.Εκεί ο αναλυτής θα θέλει σίγουρα να ξέρει σε ποιές περιοχές γίνεται η περισσότερη συζήτηση,ώστε έπειτα το τμήμα πωλήσεων να δώσει αναλόγως την αντίστοιχη βαρύτητα στο κάθε ένα. Ή γενικότερα στην περίπτωση ενός brand, είναι αντίστοιχα ιδιαίτερα ωφέλιμο για την εκάστοτε εταιρία να γνωρίζει τις περιοχές στις οποίες γίνεται η περισσότερη συζήτηση πάνω σε αυτό,ώστε ανάλογα να κινηθεί προς ανάλογες αποφάσεις, όπως

για παράδειγμα να επενδύσει και να ενισχύσει περαιτέρω την παρουσία της στις περιοχές αυτές. Αρχικά λοιπόν, ξεκινάμε βρίσκοντας τον συνολικό αριθμό των διαφορετικών τοποθεσιών σε κάθε τουίτ, και καταλήγουμε στα εξής:

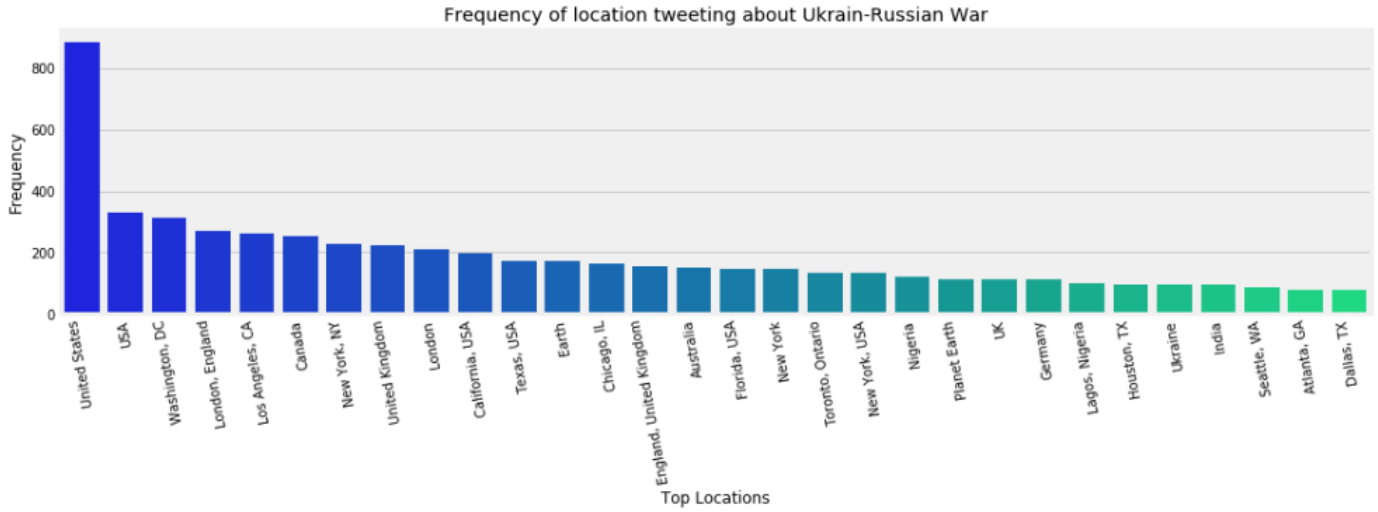
- 13.100 διαφορετικές τοποθεσίες στο Dataframe twt_24_02_pd.
- 11.791 διαφορετικές τοποθεσίες στο Dataframe twt_25_02_pd.
- 11.852 διαφορετικές τοποθεσίες στο Dataframe twt_26_02_pd.
- 11.454 διαφορετικές τοποθεσίες στο Dataframe twt_27_02_pd.

Έπειτα, θέλουμε να δείξουμε τις δημοφιλέστερες τοποθεσίες σε κάθε ένα Dataframe,το οποίο και γίνεται στην Εικόνα 5.10 . Εδώ πρέπει να παρατηρηθεί η συντριπτική υπεροχή των τιμών unknownσε κάθε Dataframe:

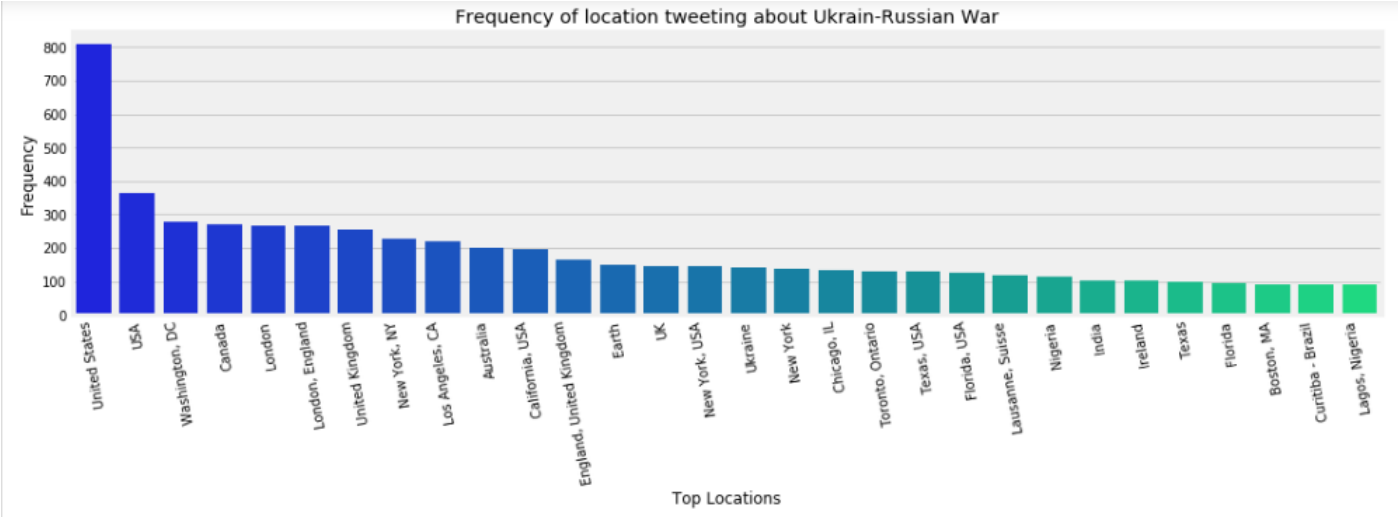
```
Value counts for 24/02/2022
unknown                 17570
United States             885
USA                       331
Washington, DC            315
London, England           272

Value counts for 25/02/2022
unknown                 18463
United States             809
USA                       366
Washington, DC            277
Canada                    272


Value counts for 26/02/2022
unknown                 18359
United States             888
USA                       427
London, England           290
Canada                    256


Value counts for 27/02/2022
unknown                 18710
United States             811
USA                       427
Canada                    300
Ukraine                   280
                          ...
```

Εικόνα 5.10 . Οι 5 δημοφιλέστερες τοποθεσίες σε κάθε Dataframe

Ένας καταλληλότερος τρόπος παρουσίασης των συχνότερων τοποθεσιών θα ήταν σίγουρα ένα σχετικό διάγραμμα. Πρέπει όμως σαφώς να απομακρυνθούν οι τιμές “unknown”, μιας και λόγω του μεγάλου ποσοστού που τους αντιστοιχεί, θα υπερκαλύπτουν όλες τις υπόλοιπες τιμές. Έτσι και έγινε, και τα αποτελέσματα φαίνονται στις εικόνες 5.11 – 5.14, όπου παρουσιάζονται οι 30 πιο δημοφιλείς τοποθεσίες για κάθε Dataframe.

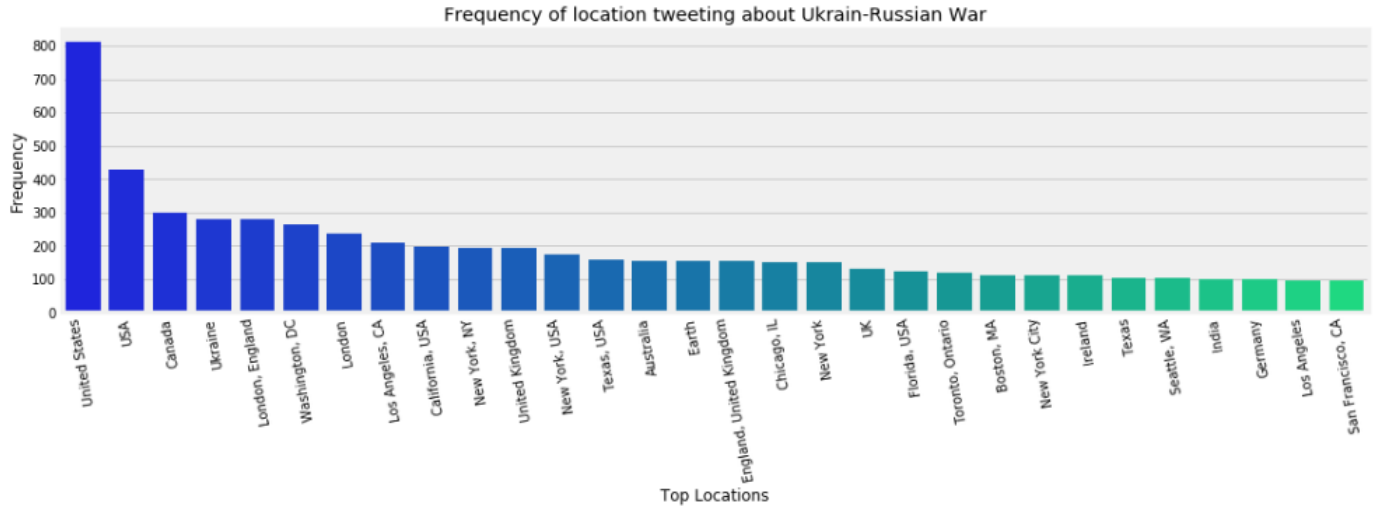


Εικόνα 5.11 . Οι 30 δημοφιλέστερες περιοχές στο Dataframe twt_24_02_pd



Εικόνα 5.12 . Οι 30 δημοφιλέστερες περιοχές στο Dataframe twt_25_02_pd

Εικόνα 5.13 . Οι 30 δημοφιλέστερες περιοχές στο Dataframe twt_26_02_pd



Εικόνα 5.14 . Οι 30 δημοφιλέστερες περιοχές στο Dataframe twt_27_02_pd

Σε επόμενο βήμα, θα ήταν χρήσιμο να ομαδοποιηθούν οι παραπάνω περιοχές ανά χώρα και να γίνει έπειτα η ανάλογη οπτικοποίηση. Η Python διαθέτει βιβλιοθήκη που αντιστοιχίζει μία κωδική ονομασία για κάθε χώρα. Για παράδειγμα “US” για το “United States”, “GB” για το «Great Britain” κοκ. Επιπλέον, μπορεί σε κάποιες περιπτώσεις να αναγνωρίζει και την χώρα στην οποία ανήκει κάποια περιοχή. Δηλαδή μπορεί να αντιστοιχίσει την ταμπέλα “US” ακόμα και στην περίπτωση που διαβάσει “Texas” ή “New York City” .Δυστυχώς τις περισσότερες περιοχές δεν τις αναγνωρίζει, οπότε δίνει την ταμπέλα “Unknown”.

Αυτό με λίγα λόγια σημαίνει ότι στην καινούργια στήλη που θα δημιουργηθεί με τις προαναφερθείσες ταμπέλες της κάθε περιοχής, τα “Unknown” στοιχεία αναμένεται να είναι πάρα πολλά. Και έτσι γίνεται. Όπως φαίνεται και στηις εικόνες 5.15-5.18 παρακάτω, οι τιμές “Unknown” αποτελούν πάνω από το 90% των τιμών σε όλα τα Dataframes.

```
Value counts for 24/02/2022
Number of unique values: 141
unknown    45750
US          1697
GB           391
CA           294
AU           164
            ...
TD             1
AM             1
GT             1
TN             1
MW             1
Name: country_code, Length: 140, dtype: int64
```

Εικόνα 5.15 . Αριθμός χωρών στο Dataframetwt_24_02_pd

```
Value counts for 25/02/2022
Number of unique values: 148
unknown    45421
US          1717
GB           431
CA           314
AU           233
            ...
NI             1
OM             1
AZ             1
GL             1
RW             1
Name: country_code, Length: 147, dtype: int64
```

Εικόνα 5.16 . Αριθμός χωρών στο Dataframe twt_25_02_pd

```
Value counts for 26/02/2022
Number of unique values: 136
unknown    45366
US          1716
GB           422
CA           293
UA           285
            ...
MZ             1
MS             1
RS             1
CD             1
RW             1
Name: country_code, Length: 136, dtype: int64
```

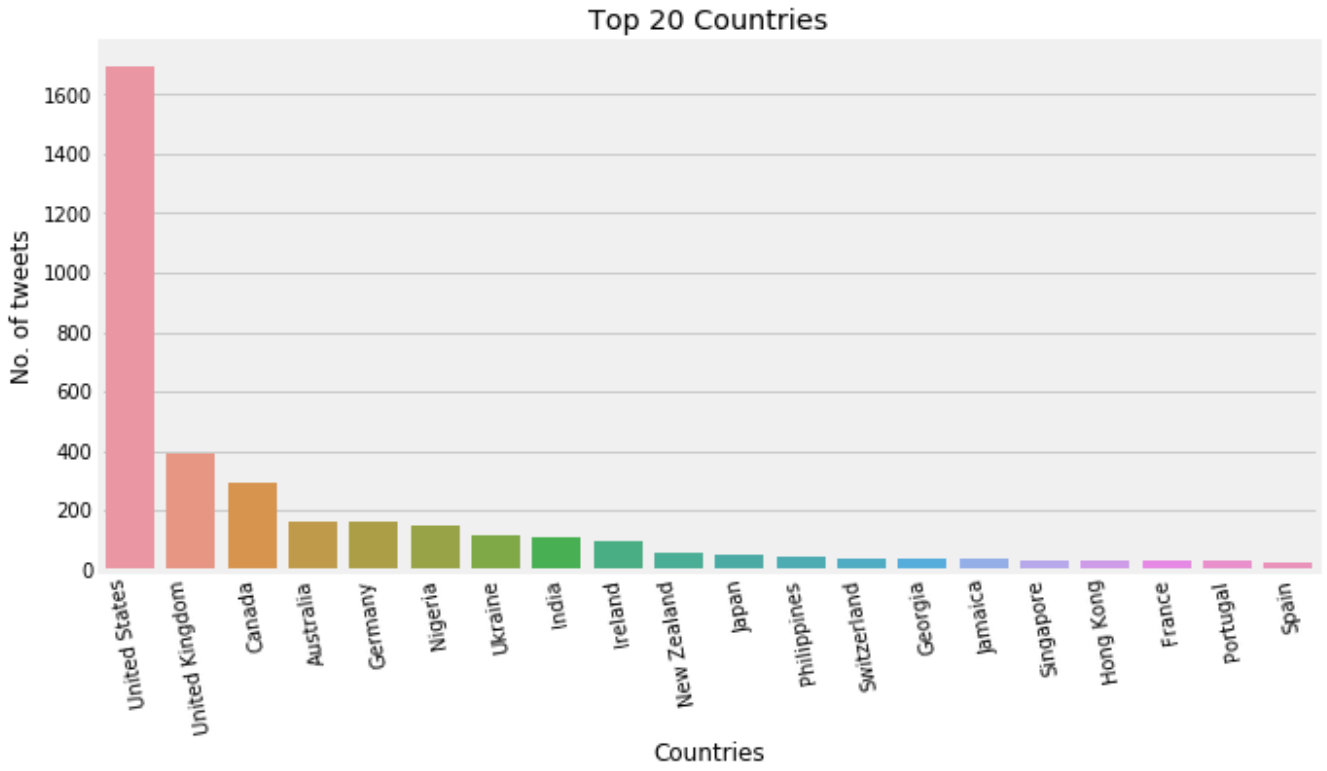
Εικόνα 5.17 . Αριθμός χωρών στο Dataframe twt_26_02_pd

```
Value counts for 27/02/2022
Number of unique values: 144
unknown    45416
US          1708
GB           484
CA           343
UA           280
            ...
DO             1
IM             1
PY             1
SL             1
MW             1
Name: country_code, Length: 144, dtype: int64
```

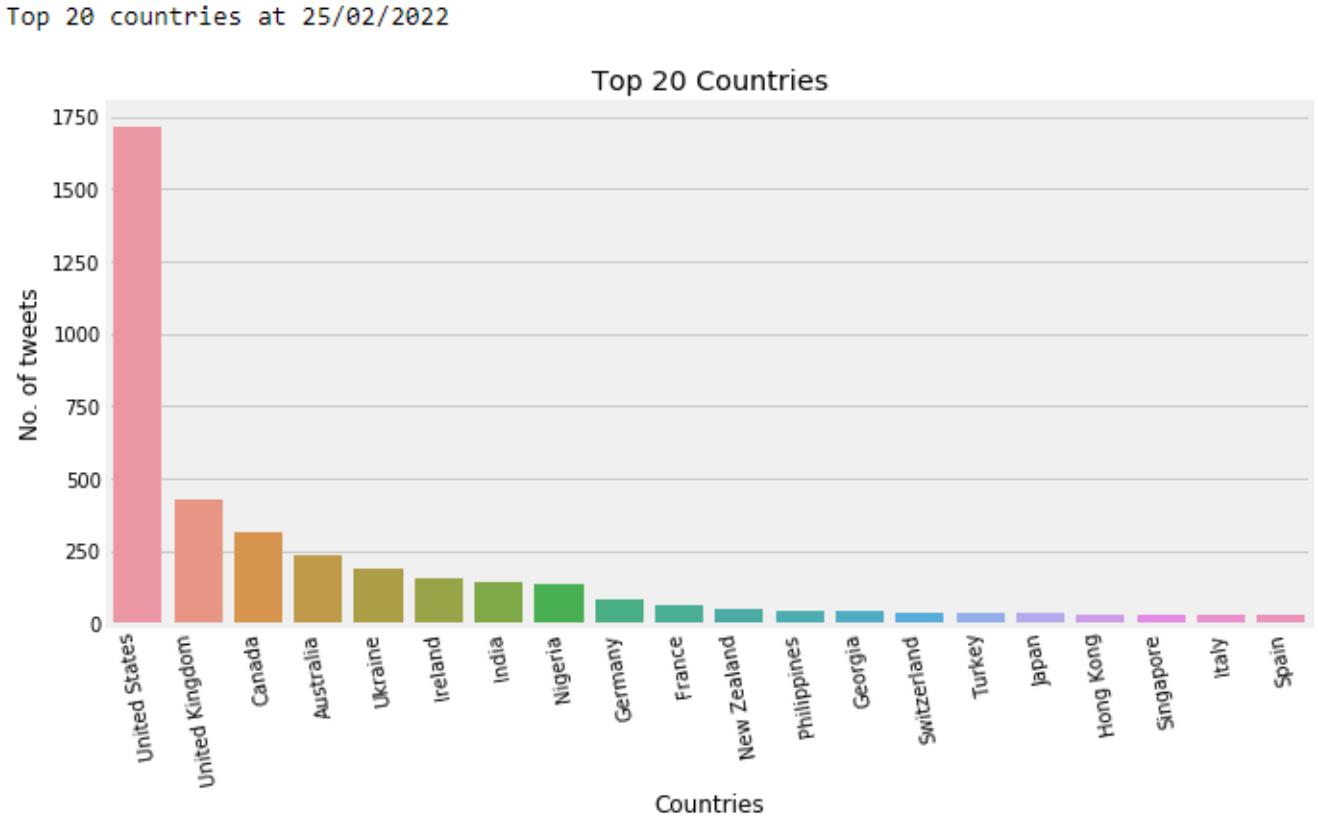
Εικόνα 5.18 . Αριθμός χωρών στο Dataframe twt_27_02_pd

Παρά την οφθαλμοφανή αδυναμία της συγκεκριμένη μεθόδου να κατατάξει αποδοτικά όλες τις περιοχές στις χώρες που αντιστοιχούν, τα αποτελέσματα εξακολουθούν να έχουν χρησιμότητα. Αυτή η χρησιμότητα πηγάζει ουσιαστικά από το γεγονός ότι δεν μας ενδιαφέρει τόσο πολύ ο απόλυτος αριθμός καταχωρήσεων της κάθε χώρας, παρά η σχετικότητα μεταξύ τους. Έτσι, άμα ο στόχος μας είναι να δούμε ποιες χώρες κυριαρχούν σε σχέση με τις υπόλοιπες στα tweets,αυτή η μέθοδος δεν είναι τόσο άσχημη. Στις εικόνες 5.19 – 5.22 παρουσιάζονται οι 20 πιο δημοφιλείς χώρες σε κάθε ένα από τα Dataframes. Όπως και προηγουμένως, οι τιμές “Unknown” δεν παρουσιάζονται στα διαγράμματα.

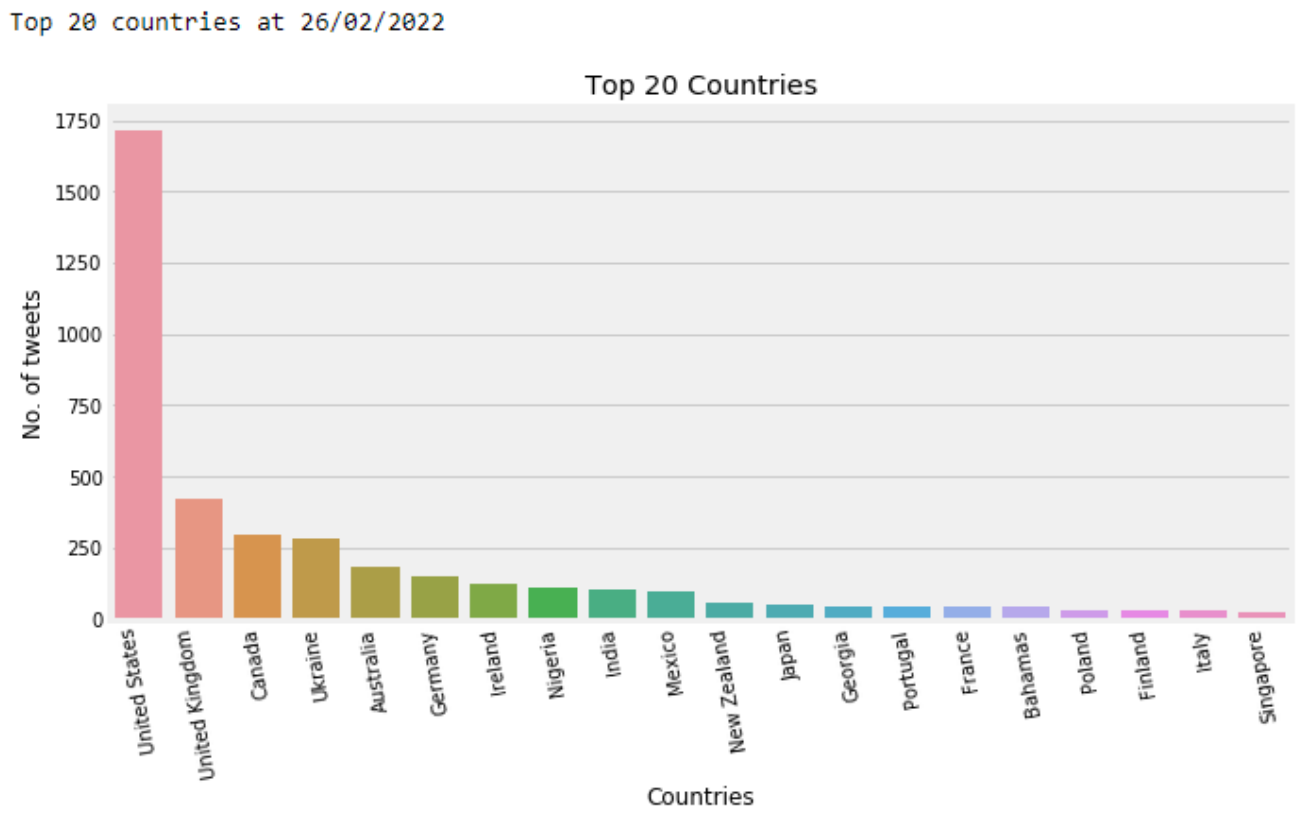
```
Top 20 countries at 24/02/2022
```



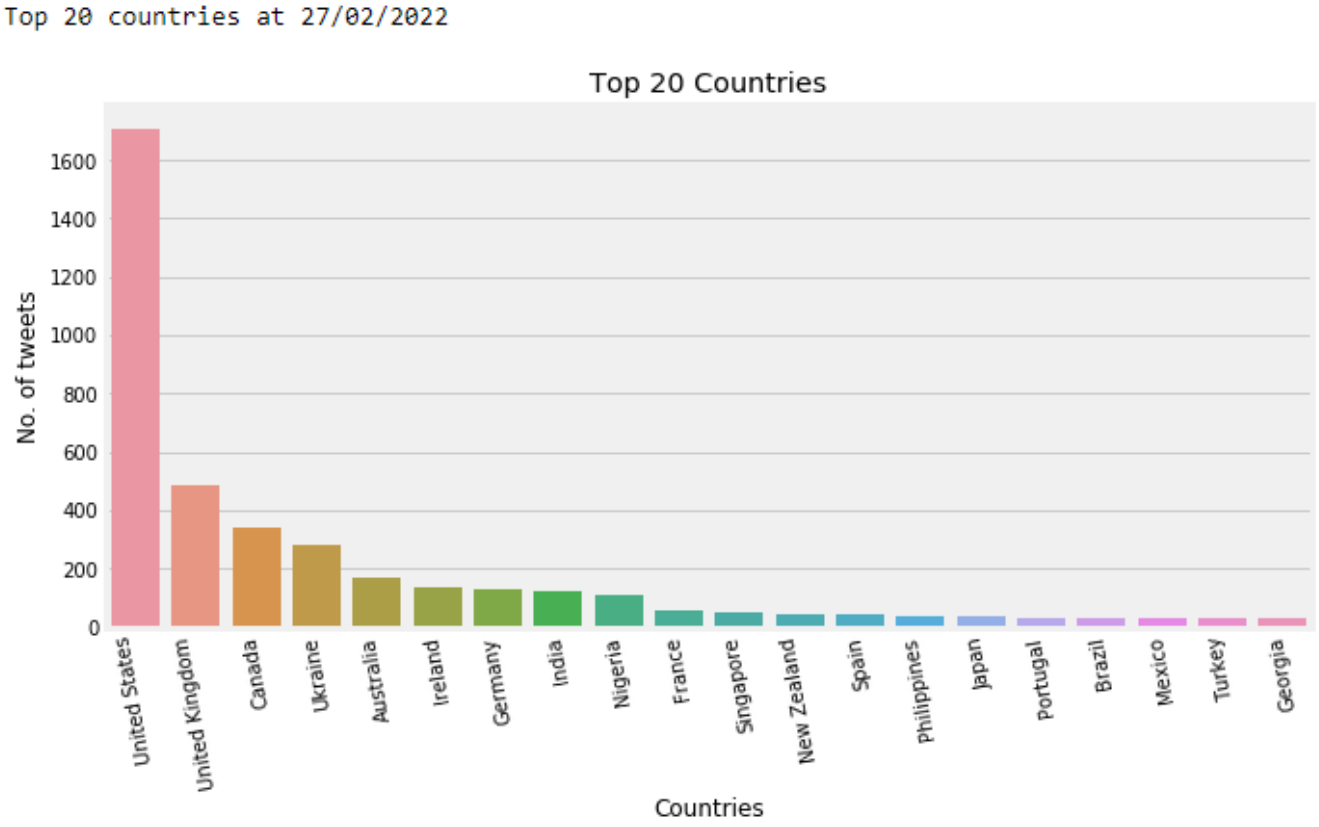
Εικόνα 5.19 . Οι 30 δημοφιλέστερες χώρες στο Dataframe twt_24_02_pd



Εικόνα 5.20 . Οι 30 δημοφιλέστερες χώρες στο Dataframe twt_25_02_pd



Εικόνα 5.21 . Οι 30 δημοφιλέστερες χώρες στο Dataframe twt_26_02_pd



Εικόνα 5.22 . Οι 30 δημοφιλέστερες χώρες στο Dataframe twt_27_02_pd

Η δεύτερη παράμετρος που εξετάζεται σε αυτήν την ενότητα είναι ,όπως αναφέρθηκε , η συχνότητα με την οποία παρουσιάζονται κάποιες λέξεις. Συγκεκριμένα, επιθυμούμε με την χρήση κατάλληλων εργαλείων να γίνει η εξόρυξη των πιο πολυχρησιμοποιούμενων λέξεων στον κορμό του κάθε ενός Dataframe.Αυτή η διαδικασία μπορεί να δώσει σημαντική πληροφορία σχετικά με την γνώμη του κόσμου σχετικά με το εκάστοτε αντικείμενο.

Έστω για παράδειγμα ότι μια εταιρεία έχει λανσάρει ένα νέο προιόν και θέλει να αντλήσει μέσω σχετικών tweets, την γνώμη του κόσμου για το εν λόγω προιόν της. Εκτός από τις τεχνικές sentiment analysis που θα εφαρμοστούν (οι οποίες φαίνονται και στην συνέχεια της συγκεκριμένης διατριβής) ,ένας πολύ άμεσος και γρήγορος τρόπος να φανεί η γνώμη του κόσμου, είναι να εξορύξει τις πιο πολυχρησιμοποιούμενες λέξεις. Έτσι, εάν λόγου χάριν φανεί ότι στα tweets κυριαρχούν λέξεις όπως ‘τέλειο’, ’συναρπαστικό’, ’πρωτοποριακό κλπ, αυτό σίγουρα θα αποτελεί μια ένδειξη ότι η γνώμη των καταναλωτών είναι σε γενικές γραμμές θετική.

Έτσι και στην περίπτωση μας λοιπόν, επιθυμούμε να ρίξουμε μια ματιά στις λέξεις γύρω από τις οποίες γίνεται περισσότερος λόγος. Το εργαλείο που χρησιμοποιείται εδώ είναι το προαναφερθέν TF-IDF vectorization.Η Python προσφέρει βιβλιοθήκη όπου μετατρέπει τα επεξεργασμένα tweets σε πίνακες TF-IDF. Από αυτούς τους πίνακες λοιπόν, μπορούμε εύκολα να βρούμε τις πιο συχνά χρησιμοποιούμενες λέξεις σε κάθε Dataframe.Τα αποτελέσματα φαίνονται στην παρακάτω εικόνα.

27/04/2022

	total frequency
putin	957.572860
people	878.621875
war	850.104242
support	600.881789
world	592.758899
nato	564.405535
help	555.317809
invasion	547.734617
like	502.543794
amp	485.223641
stand	408.090031
country	402.216635
president	378.197299
know	377.872087
trump	372.979150
think	371.411420
right	365.787494
good	363.251774
fight	359.702063
going	351.991876

26/04/2022

	total frequency
people	992.784463
putin	966.984670
war	849.900406
help	638.320527
support	586.425308
world	580.688054
nato	546.008633
invasion	508.984882
like	489.402555
amp	470.594225
stand	467.575898
trump	434.075032
right	412.879678
president	402.940348
country	385.113817
know	379.088458
think	371.250854
fight	369.833260
going	358.829631
need	357.099529

25/04/2022

	total frequency
putin	991.767446
people	989.495138
war	872.328773
nato	727.678717
invasion	639.193324
help	609.345858
world	574.551486
support	532.923133
amp	527.739463
stand	508.348344
like	501.224696
president	458.270311
country	429.241049
right	407.022505
stop	380.414847
going	370.359355
know	358.388876
trump	356.073398
think	355.980279
good	344.488994

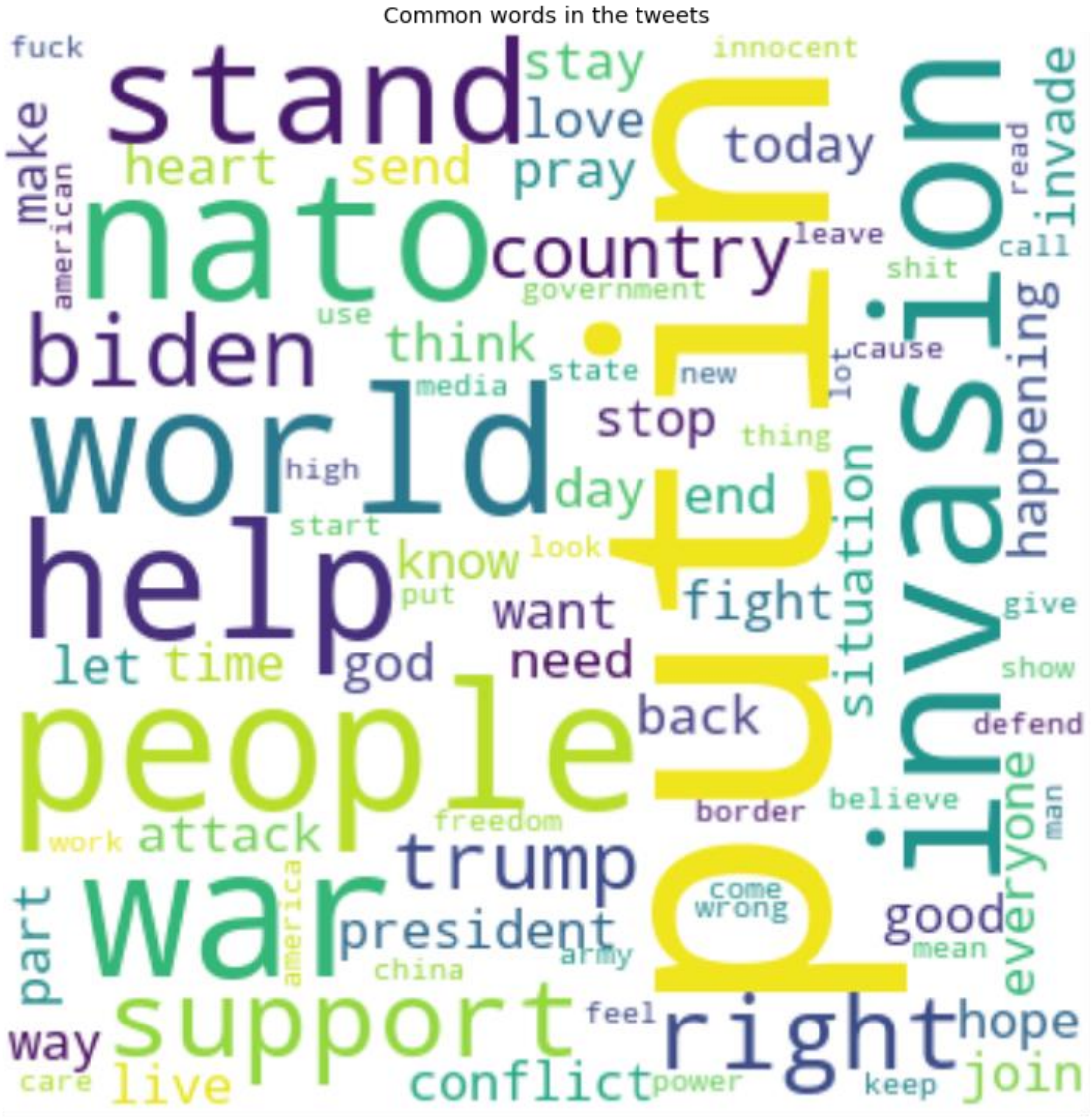
24/04/2022

	total frequency
putin	1110.084013
people	1051.082649
war	972.934010
invasion	741.543878
nato	711.788360
amp	531.837184
world	523.165329
like	500.543346
help	478.518635
stand	465.250924
support	460.741173
going	431.649672
right	431.292476
biden	427.732932
trump	420.853575
country	416.472472
president	397.666630
know	393.085505
think	388.108490
stop	374.076498

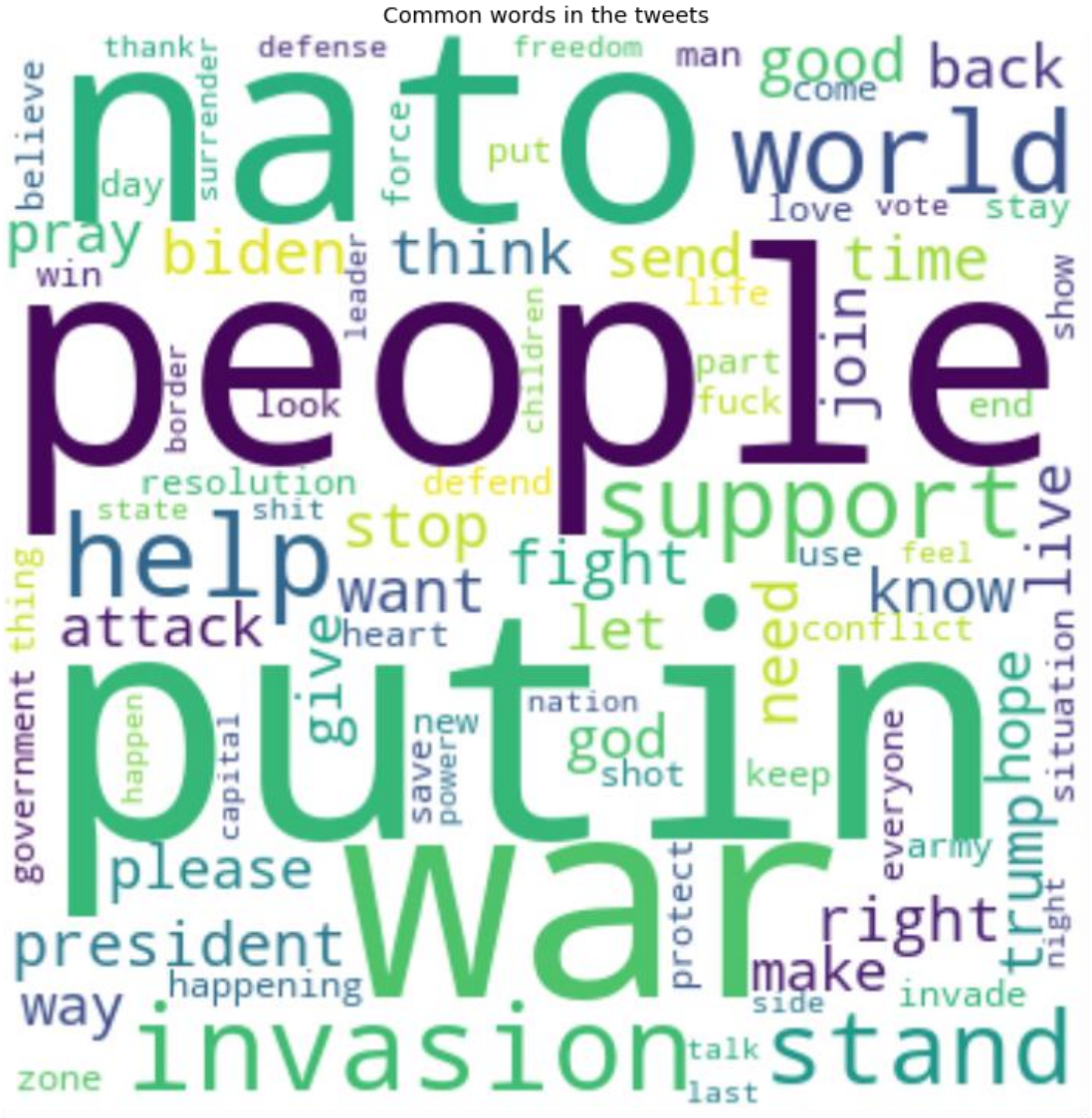
Εικόνα 5.23 . Οι πιο συχνά εμφανιζόμενες λέξεις στα tweets της κάθε ημέρας

Να σημειωθεί ότι στην καταμέτρηση των λέξεων για την εύρεση των συχνότερων, δεν προσμετρήθηκαν οι λέξεις που ανήκουν στις προαναφερθείσες “stopwords” ,καθώς επίσης και η λίστα λέξεων ["ukraine","russia","would","russian","ukrainian","get","one","via","see","says","take"] , οι οποίες εμφανίζονταν πολύ συχνά αλλά δεν πρόσδιδαν κάποια ιδιαίτερη πληροφορία και επιλέχθηκε να μην προσμετρηθούν.

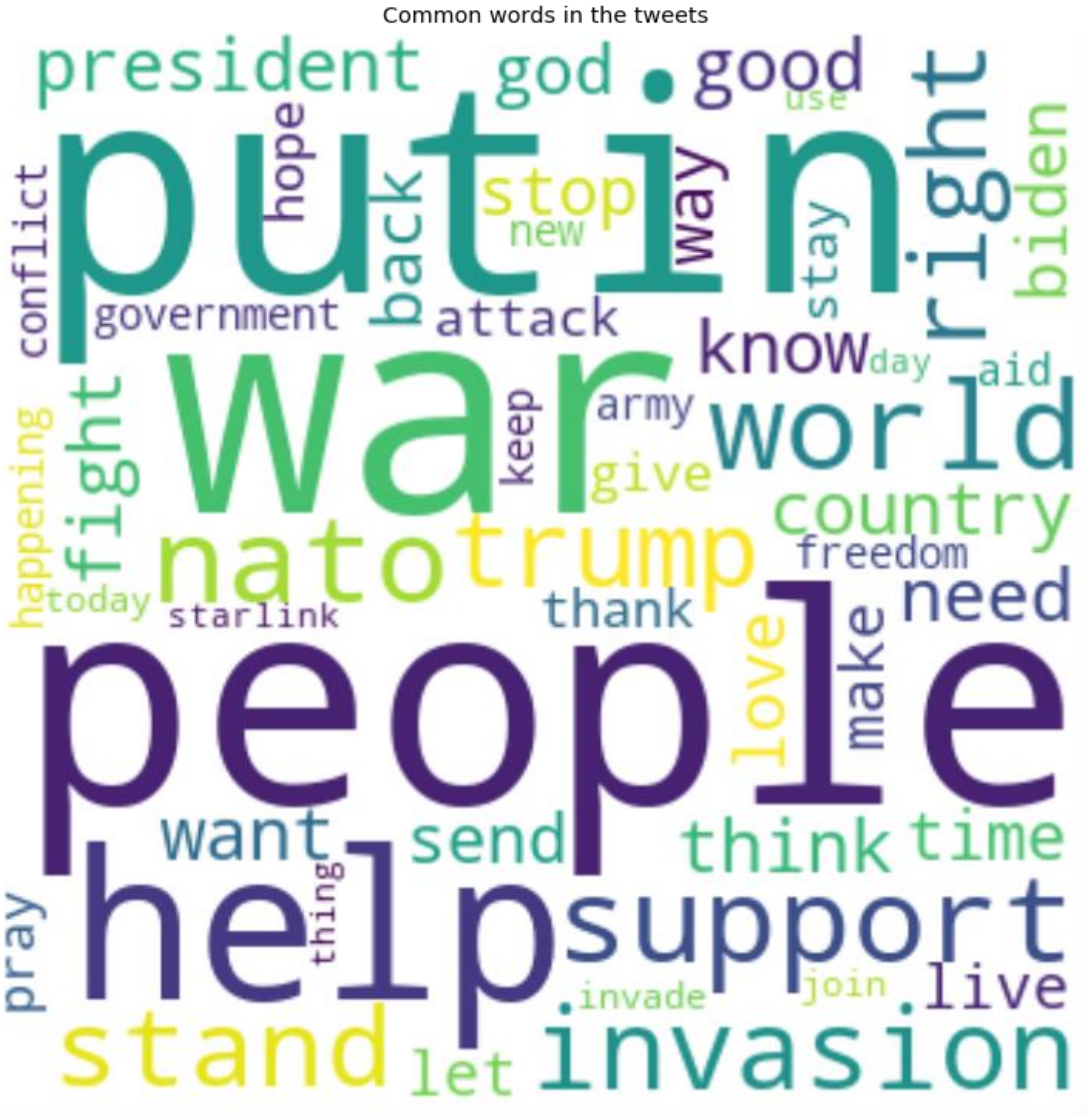
Ένας άλλος πιο οπτικός τρόπος παρουσίασης των παραπάνω δημοφιλέστερων λέξεων, είναι αυτός που στην Python εμφανίζεται με το όνομα WordCloud. Πρόκειται για μια βιβλιοθήκη της Pythonπου χρησιμοποιείται πολύ συχνά σε περιπτώσεις οπτικοποίησης συχνότητας εμφάνισης λέξεων, η οποία ουσιαστικά παρουσιάζει σε ένα πλαίσιο όλες τις δημοφιλείς λέξεις, με το μέγεθος της κάθε μίας να αντιστοιχεί και στο πόσο συχνά συναντιέται στα tweets.Τα αποτελέσματα φαίνονται στις παρακάτω εικόνες.



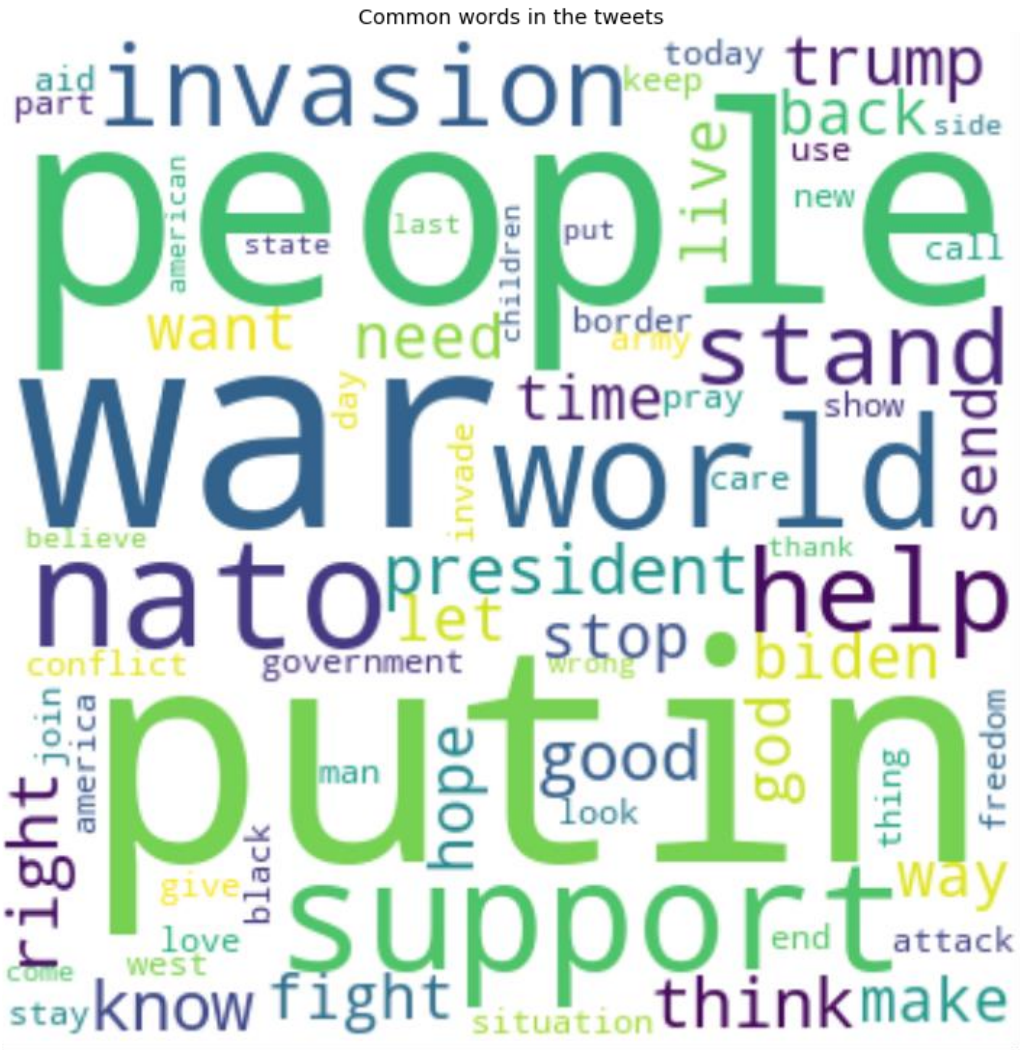
Εικόνα 5.24 . WordCloud για το Dataframe twt_24_02_pd.



Εικόνα 5.25 . WordCloud για το Dataframe twt_25_02_pd.



Εικόνα 5.26 . WordCloud για το Dataframe twt_26_02_pd.



Εικόνα 5.27 . WordCloud για το Dataframe twt_27_02_pd.

5.4 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ (SENTIMENT ANALYSIS)

Όπως εξηγήθηκε και παραπάνω, το βασικότερο αντικείμενο της συγκεκριμένης διατριβής είναι να εξαχθεί το "συναίσθημα" (sentiment) του κάθε tweet, με όσο το δυνατόν καλύτερη ακρίβεια –δεδομένων των εργαλείων που χρησιμοποιούνται-,ώστε έπειτα να εκπαιδευτούν πάνω στα κατηγοριοποιημένα (labeled) πλέον tweets,όσο το δυνατόν πιο ακριβή μοντέλα μηχανικής μάθησης.

Η κατηγοριοποίηση των tweets με βάση το συναίσθημα (sentiment) –θετικό,ουδέτερο και αρνητικό στην περίπτωση μας- ,γίνεται με δύο κυρίως τρόπους όπως αναφέρθηκε και στο κεφάλαιο 3.2.2.Ο πρώτος τρόπος είναι να χρησιμοποιηθεί μια lexicon-based τεχνική, όπως η TextBlob ή η Vader, οι οποίες κατηγοριοποιούν τα tweets με βάση το πόσο θετικό η αρνητικό πρόσημο φέρει κάθε μία από τις λέξεις τους. Ο δεύτερος τρόπος, είναι να χρησιμοποιήσουμε ένα μοντέλο μηχανικής μάθησης, το οποίο έχει εκπαιδευτεί πάνω σε ήδη κατηγοριοποιημένα δεδομένα, τα οποία συνήθως με την σειρά τους κατηγοριοποιήθηκαν από κάποιον αναλυτή ένα-ένα. Ο οποίος παρεμπιπτόντως αποτελεί και έναν τρίτο τρόπο να γίνει η κατηγοριοποίηση, να διαβαστούν δηλαδή από κάποιον αναγνώστη ο οποίος θα τα κατηγοριοποιεί ένα προς ένα. Προφανώς όμως στην συγκεκριμένη εργασία μας ενδιαφέρουν οι δύο πρώτοι, αυτοματοποιημένοι τρόποι.

Ας γυρίσουμε στις lexicon-based (ή rule_based) μεθόδους.Πρόκειται για τα πιο κοινά εργαλεία, ωστόσο και οι δύο παρουσιάζουν μειωμένη σχετικά απόδοση, με την Vader αρκετές φορές να υπερισχύει της TextBlob, μιας και η δεύτερη έχει την αδυναμία συχνά να μη κατατάσσει σωστά τα αρνητικά sentiments, όπως φαίνεται και στην εργασία του Mohit Kumar Barai ,«Sentiment Analysis with TextBlob and Vader» στην ιστοσελίδα Analytics Vidhya. Στην συγκεκριμένη εργασία φαίνεται επίσης πως όταν συγκρίθηκαν τα αποτελέσματα της TextBlob και της Vader, πάνω σε κείμενα που ήταν ήδη κατηγοριοποιημένα, η πρώτη είχε απόδοση 41.3% και η δεύτερη 61.3%.

Αυτό ωστόσο δεν σημαίνει ότι η Vader πάντα θα υπερισχύει της TextBlob. Έχει μεν ένα προβάδισμα, διότι όπως αναφέρθηκε κατατάσσει πιο σωστά εν γένει τα αρνητικά sentiments, ωστόσο πολλές φορές η συνολική απόδοση εξαρτάται από την βάση δεδομένων και τον τρόπο που χρησιμοποιούνται οι μέθοδοι.

Η τακτική που ακολουθήθηκε στην συγκεκριμένη εργασία για την κατηγοριοποίηση των tweets, ήταν ένας συνδυασμός των δύο παραπάνω. Συγκεκριμένα, βρέθηκαν τα sentiments σε κάθε tweet χρησιμοποιώντας την TextBlob αρχικά και έπειτα την Vader ρόπως θα φανεί και στην συνέχεια. Εκεί έγινε και μια σύντομη συζήτηση σχετικά με το πόσο όμοιες ή μη ήταν οι κατηγοριοποιήσεις του κάθε εργαλείου. Έπειτα, με τρόπο που θα φανεί αναλυτικά πιο κάτω, δημιουργήσαμε ένα νέο Dataframe για κάθε ένα από τα προηγούμενα τέσσερα, στο οποίο κρατήσαμε μόνο εκείνα τα τουίτς όπου το sentiment του ενός εργαλείου συμφωνούσε με το sentiment του άλλου. Αυτά θα είναι και τα Dataframes πάνω στα οποία θα εκπαιδευτούν αργότερα και τα μοντέλα μηχανικής μάθησης.

Ο λόγος που επιλέχθηκε αυτή η προσέγγιση είναι απλός. Είναι λογικό να συμπεράνουμε ότι σε εκείνα τα tweets που και οι δύο μέθοδοι έδωσαν το ίδιο label, είναι στατιστικά πιο πιθανό αυτό το label να είναι και το πραγματικό. Οπότε, είναι επίσης λογικό να επιλέξουμε να εκπαιδεύσουμε τα μοντέλα μας πάνω σε εκείνα τα tweets, και όχι με βάση την ετικέτα της μίας μεθόδου από τις δύο. Ας δούμε όμως συγκεκριμένα και με την σειρά, την διαδικασία που ακολουθήθηκε για να δοθούν οι ετικέτες (labels) σε κάθε ένα tweet,με τα εργαλεία που παρουσιάστηκαν παραπάνω.

Ανάλυση συναισθήματος με χρήση της Textblob

Ξεκινώντας από την TextBlob, υπενθυμίζεται ότι για κάθε κείμενο που επεξεργάζεται, βγάζει δύο αποτελέσματα. Το ένα είναι η υποκειμενικότητα του κειμένου, το οποίο στην συγκεκριμένη ανάλυση δε θα μας χρησιμεύσει, και το δεύτερο είναι η πολικότητα ή polarity,που παίρνει τιμές από -1 έως 1,και μας δείχνει ουσιαστικά το κατά πόσο το κείμενο έχει αρνητικό, ουδέτερο ή θετικό συναίσθημα.

Το συναίσθημα λοιπόν του κάθε τουίτ προκύπτει απευθείας από το Polarity,εφαρμόζοντας τον απλό κανόνα:

- Εάν polarity < 0 ,τότε Sentiment = ‘Negative’ (Αρνητικό)

- Εάν polarity = 0 ,τότε Sentiment = ‘Neutral’ (Ουδέτερο)
- Εάν polarity > 0 ,τότε Sentiment = ‘Positive’ (Θετικό)

Τα αποτελέσματα της κατάταξης φαίνονται στην εικόνα 5.28.

```
Sentiment Value Counts for 27/02/2022
Neutral     19664
Positive    18536
Negative    11798
Name: Sentiment, dtype: int64
-------------------------------------------
Sentiment Value Counts for 26/02/2022
Neutral     19990
Positive    18896
Negative    11111
Name: Sentiment, dtype: int64
-------------------------------------------
Sentiment Value Counts for 25/02/2022
Neutral     20293
Positive    18036
Negative    11670
Name: Sentiment, dtype: int64
-------------------------------------------
Sentiment Value Counts for 24/02/2022
Neutral     20136
Positive    17398
Negative    12465
Name: Sentiment, dtype: int64
```

Εικόνα 5.28 . Κατάταξη των tweets με βάση το sentimentμε χρήση της TextBlob .

Ανάλυση συναισθήματος με χρήση της Vader

Με αντίστοιχο τρόπο χρησιμοποιούμε και το δεύτερο εργαλείο μας,την Vader. Η Vader, όπως αναλύθηκε στο αντίστοιχο κεφάλαιο, δίνει από έναν αριθμό για τις τιμές Positive, Neutral και Negative, ο οποίος δε θα μας χρησιμεύσει εδώ, μιας και η Vader δίνει επίσης τον αριθμό Compound (Συνδυαστικός), μέσω του οποίου βρίσκουμε το sentiment με βάση τον κανόνα:

- Εάν compound < -0.05 , τότε Sentiment = ‘Negative’ (Αρνητικό)
- Εάν compound > 0.35 , τότε Sentiment = ‘Positive’ (Θετικό)
- Αλλιώς , Sentiment = ‘Neutral’ (Ουδέτερο)

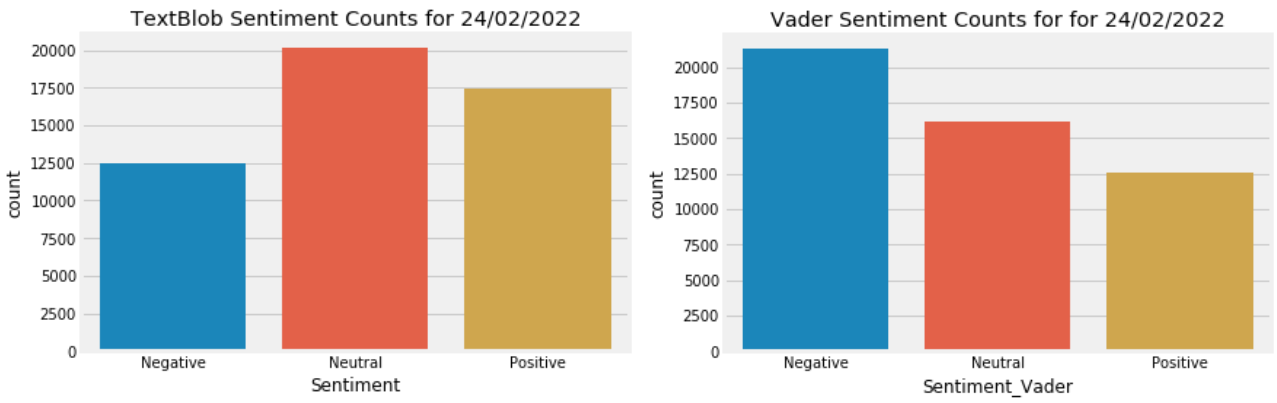
Σε αυτό το σημείο πρέπει να γίνει μια παρατήρηση. Όσον αφορά την TextBlob, τα όρια του polarity με βάση τα οποία τα tweets κατηγοριοποιούνται, είναι προκαθορισμένα και είναι αυτά που χρησιμοποιήθηκαν παραπάνω. Στην περίπτωση της Vader όμως, τα όρια αυτά δεν είναι προκαθορισμένα και ο εκάστοτε χρήστης έχει την ελευθερία να επιλέξει αυτά που θεωρεί καταλληλότερα. Στην εργασία του ο Pawan Bhandarkar στην σελίδα (https://www.kaggle.com/discussion/212043) συστήνει στα τουίτς να γίνει η χρήση των ορίων που έγινε και εδώ.

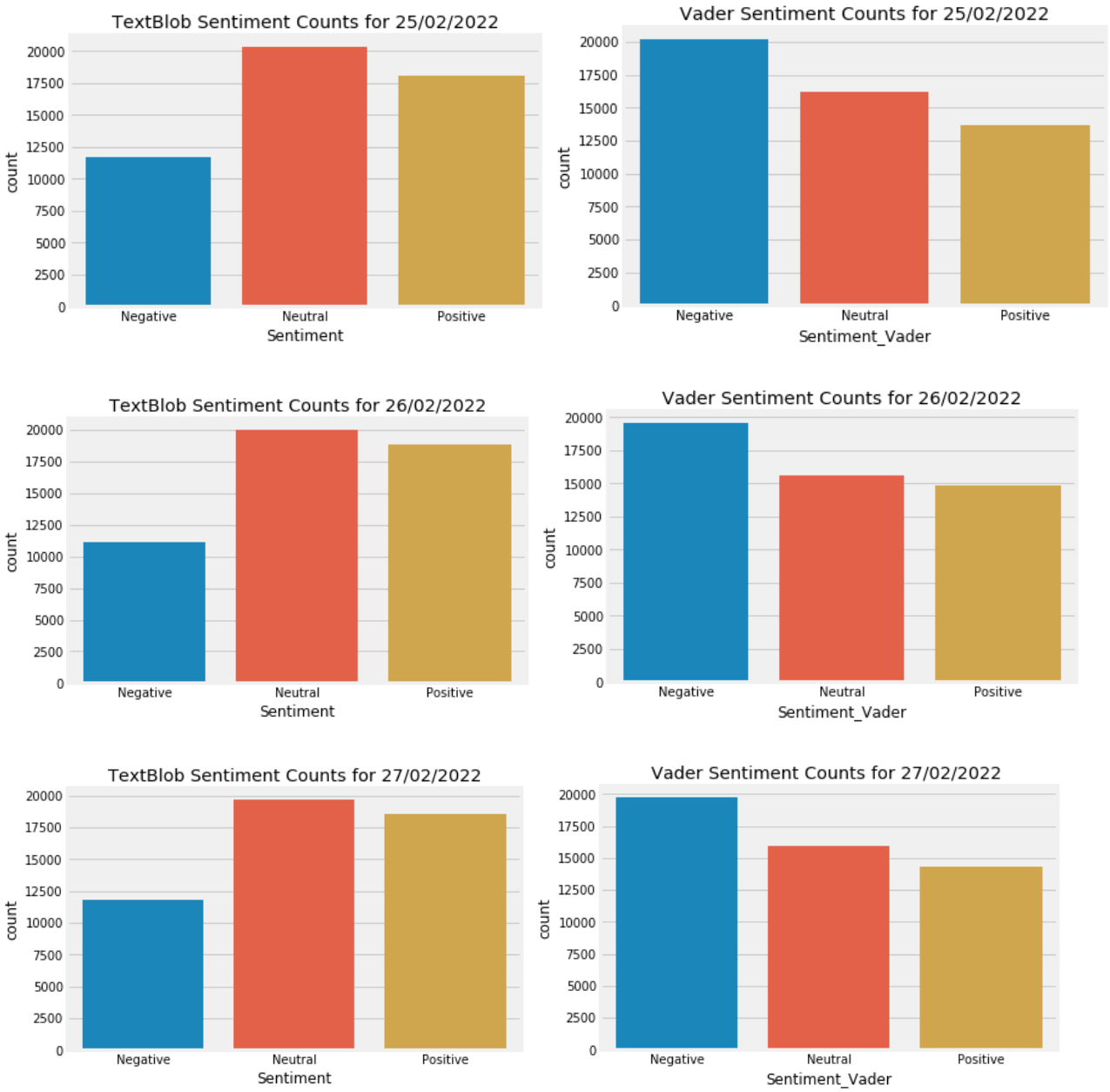
Τα αποτελέσματα που δίνει η Vader φαίνονται στην παρακάτω εικόνα.

```
Sentiment Value Counts for 27/02/2022
Negative    19726
Neutral     15934
Positive    14338
Name: Sentiment_Vader, dtype: int64
-----------------------------------------
Sentiment Value Counts for 26/02/2022
Negative    19541
Neutral     15593
Positive    14863
Name: Sentiment_Vader, dtype: int64
-----------------------------------------
Sentiment Value Counts for 25/02/2022
Negative    20188
Neutral     16177
Positive    13634
Name: Sentiment_Vader, dtype: int64
-----------------------------------------
Sentiment Value Counts for 24/02/2022
Negative    21280
Neutral     16163
Positive    12556
Name: Sentiment_Vader, dtype: int64
```

Εικόνα 5.29. Κατάταξη των tweets με βάση το sentiment με χρήση της Vader .

Είναι χρήσιμο να παρουσιαστούν και γραφικά οι κατηγοριοποιήσεις των δύο εργαλείων, ώστε να συγκρίνουμε -για κάθε μέρα- το κατά πόσο συμφωνούν τα sentiments που εξήγαγαν.





Εικόνα 5.30 . Γραφική σύγκριση των μεθόδων TextBlob και Vader .

Με μια ματιά φαίνεται αμέσως πως οι δύο μέθοδοι διαφέρουν αρκετά ως προς τα αποτελέσματα τους, κυρίως όσον αφορά τα αρνητικά tweets. Αυτό ήταν αναμενόμενο, μιας και όπως αναφέρθηκε πιο πάνω, η TextBlob παρουσιάζει μια αδυναμία στο να επεξεργαστεί σωστά τα αρνητικά κείμενα. Συν τοις άλλοις,θα έπρεπε έτσι κι αλλιώς να δημιουργεί προβληματισμό το γεγονός ότι τα αρνητικά tweets-που αναφέρονται σε πόλεμο-, είναι κατά τόσο πολύ λιγότερα από τα υπόλοιπα, όσο υποδεικνύει η TextBlob.

Πριν προχωρήσουμε, αποθηκεύουμε σε αρχεία τα τέσσερα επεξεργασμένα πλέον Dataframes (twt_24_02_pd_processed, twt_25_02_pd_processed, twt_26_02_pd_processed , twt_27_02_pd_processed) ,ώστε να μη χρειαστεί να γίνουν ξανά οι ίδιες διαδικασίες όταν κλείσει ο κώδικας και ξανατρέξει.

Σε αυτό το σημείο λοιπόν όλα είναι έτοιμα για το τελευταίο βήμα πριν την ανάπτυξη των μοντέλων μηχανικής μάθησης. Αυτό το βήμα είναι, όπως αναλύθηκε παραπάνω, να δημιουργήσουμε τέσσερα καινούργια Dataframes (twt_24_02_pd_sentiments, twt_25_02_pd_sentiments, twt_26_02_pd_sentiments , twt_27_02_pd_sentiments) για κάθε ένα από τα ήδη υπάρχοντα, όπου θα βρίσκονται μόνο τα τουίτς των οποίων το συναίσθημα που προέκυψε από τους δύο αλγορίθμους είναι ίδιο. Στην εικόνα 5.31 φαίνονται οι πρώτες γραμμές από το twt_27_02_pd_sentiments (μιας και τα υπόλοιπα είναι παρόμοια).

	text_preprocessed_tokenized(str)	Sentiment Textblob	Sentiment_Vader	Final_Sentiment
1	typical republicans country pulling together c...	Negative	Negative	Negative
2	remember war political issue entire world unit...	Positive	Positive	Positive
3	financial calamity sojourn white house	Neutral	Neutral	Neutral
4	russia invades ukraine canada rails olympics a...	Neutral	Neutral	Neutral
5	already provides visa free access	Positive	Positive	Positive

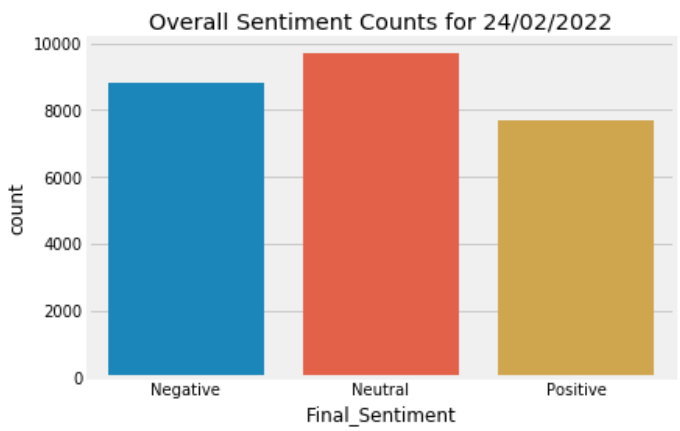
Εικόνα 5.31 .Οι 6 πρώτες γραμμές του Dataframe twt_27_02_sentiments

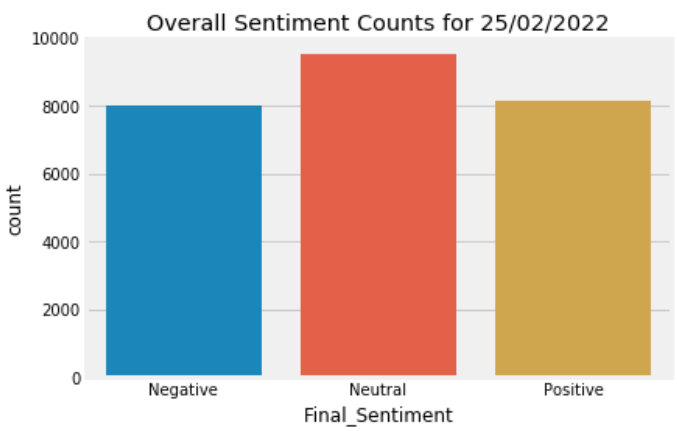
Το τελικό sentiment λοιπόν φαίνεται στην παρακάτω εικόνα (εικόνα 5.32). Παρατηρούμε ότι περίπου οι μισές τιμές από τα Dataframes που είχαμε έχουν διαγραφεί, το οποίο σημαίνει ότι είχαν κατηγοριοποιηθεί διαφορετικά από τα δύο εργαλεία. Πιο κάτω, στην εικόνα 5.33, φαίνονται επίσης και τα σχετικά γραφήματα.

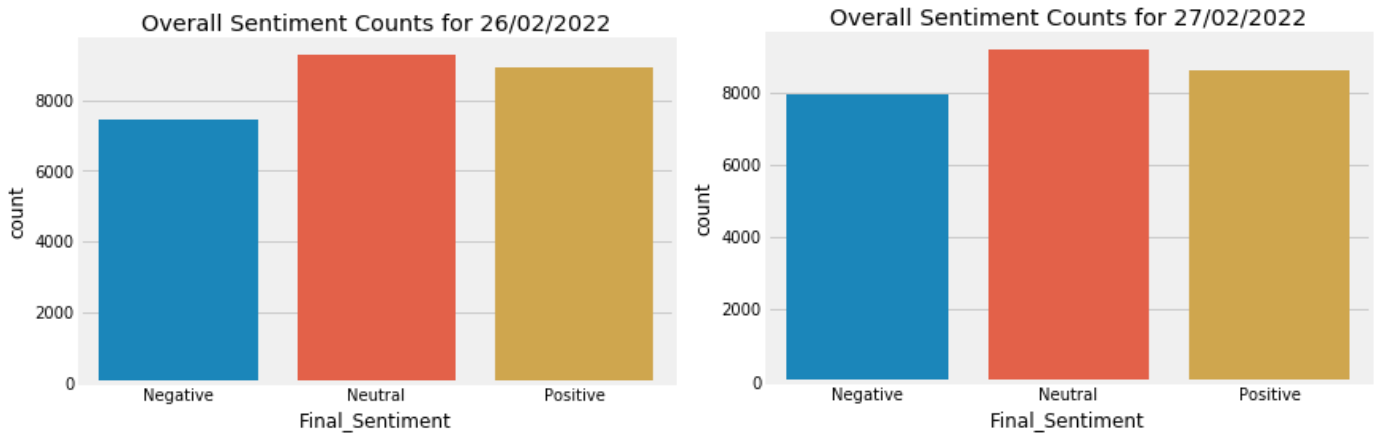
Τα καινούργια Dataframes αποθηκεύτηκαν και αυτά με την σειρά τους σε τέσσερα αρχεία, ώστε να μπορούν να διαβαστούν και να χρησιμοποιηθούν ανά πάσα στιγμή χωρίς να χρειαστεί να επαναληφθούν οι διαδικασίες. Τα αρχεία είναι τα : twt_24_02_sentiments , twt_25_02_sentiments , twt_26_02_sentiments , twt_27_02_sentiments

```
-------------------------------------------
Final Sentiment Value Counts for 27/02/2022
Neutral     9169
Positive    8601
Negative    7929
Name: Final_Sentiment, dtype: int64
Number of values: 25699
-------------------------------------------
Final Sentiment Value Counts for 26/02/2022
Neutral     9278
Positive    8940
Negative    7451
Name: Final_Sentiment, dtype: int64
Number of values: 25669
-------------------------------------------
Final Sentiment Value Counts for 25/02/2022
Neutral     9531
Positive    8151
Negative    8044
Name: Final_Sentiment, dtype: int64
Number of values: 25726
-------------------------------------------
Final Sentiment Value Counts for 24/02/2022
Neutral     9698
Negative    8841
Positive    7707
Name: Final_Sentiment, dtype: int64
Number of values: 26246
```

Εικόνα 5.32 . Το τελικό sentiment .







Εικόνα 5.33 . Γραφήματα με το τελικό sentiment.

5.5 ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Σε αυτήν την ενότητα θα παρουσιαστεί η διαδικασία η οποία ακολουθήθηκε ώστε να αναπτυχθούν αποδοτικά μοντέλα μηχανικής μάθησης καθώς και ο τρόπος με τον οποίο έγινε η σύγκριση μεταξύ τους για την εξαγωγή των καλυτέρων εξ αυτών.

Μηχανική μάθηση – Βασικές έννοιες

Πρώτα όμως είναι απαραίτητο να παρουσιαστούν οι βασικές αρχές γύρω από την μηχανική μάθηση. Καταρχήν, ας απαντηθεί το ερώτημα «τι είναι η μηχανική μάθηση»; Στην ιστοσελίδα Βικιπαίδεια διαβάζουμε «Μηχανική μάθηση είναι υποπεδίο της επιστήμης των υπολογιστών, που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην τεχνητή νοημοσύνη». Με πιο απλά λόγια, θα μπορούσε κάποιος να πει ότι μηχανική μάθηση είναι η εκπαίδευση ουσιαστικά ενός αλγορίθμου μέσω υπολογιστή, ώστε να μπορεί να αναγνωρίζει και να κατηγοριοποιεί δεδομένα. Ανάλογα με το είδος αυτής της κατηγοριοποίησης, τα μοντέλα διακρίνονται σε αυτά που κάνουν regression (παλινδρόμηση) και κατηγοριοποιούν τα δεδομένα σε συνεχείς τιμές, και σε αυτά που κάνουν classification (ταξινόμηση) που κατηγοριοποιούν τα δεδομένα σε διακριτές τιμές, όπως συμβαίνει και στην περίπτωση μας.

Υπάρχουν κυρίως δύο είδη μηχανικής μάθησης, η επιβλεπόμενη και η μη επιβλεπόμενη (supervised and unsupervised machine learning) .Κατά την επιβλεπόμενη, όπως παρουσιάζεται αμέσως πιο κάτω εκτενώς, ο αλγόριθμος εκπαιδεύεται πάνω σε ένα ήδη κατηγοριοποιημένο σετ δεδομένων, ενώ στην δεύτερη περίπτωση της μη επιβλεπόμενης, ο αλγόριθμος κατηγοριοποιεί από μόνος του τα δεδομένα σε ομάδες βάση ομοιοτήτων μεταξύ τους που από μόνος του βρίσκει. Στην παρούσα έρευνα χρησιμοποιούνται μόνο τα πρώτα, και όλες οι αναφορές σε μοντέλα μηχανικής μάθησης αναφέρονται μόνο στα επιβλεπόμενα, χωρίς να γίνεται ξανά αυτή η σημείωση.

Για να γίνει πιο κατανοητός ο τρόπος λειτουργίας ενός μοντέλου,ας σκεφτούμε το παράδειγμα όπου ένα μοντέλο μηχανικής μάθησης εκπαιδεύεται για να μπορεί να κατηγοριοποιεί φωτογραφίες με βάση το αν απεικονίζεται κάποιο δέντρο ή όχι. Κατά την εκπαίδευση του, το μοντέλο πρέπει να εκπαιδευτεί με όσο το δυνατόν περισσότερες φωτογραφίες με δέντρα και χωρίς. Σε αυτό το στάδιο θα γνωρίζει ποιες φωτογραφίες

περιέχουν δέντρα και ποιες όχι,μιας και αυτή η πληροφορία θα παρέχεται από τον χρήστη. Με βάση τα κατηγοριοποιημένα (labeled) δεδομένα αυτά, το μοντέλο θα πρέπει μετά την εκπαίδευση του να μπορεί να κατηγοριοποιεί σωστά καινούργιες φωτογραφίες που δεν θα έχει «ξαναδεί», με βάση το αν περιέχουν κάποιο δέντρο η όχι.

Γενικά λοιπόν, η διαδικασία που ακολουθείται για την δημιουργία οποιοδήποτε μοντέλου είναι η εξής συγκεκριμένη που περιεγράφηκε και πιο πάνω. Αρχικά, χρησιμοποιείται συνήθως ένα σετ κατηγοριοποιημένων δεδομένων. Αυτό αποτελείται από n δείγματα (samples) όπου το καθένα φέρει m χαρακτηριστικά (features) και τις ετικέτες (labels). Τα δύο πρώτα αποτελούν εκείνο το τμήμα των δεδομένων το οποίο το μοντέλο θα «διαβάσει» ώστε να κάνει την πρόβλεψη, ενώ τα labels είναι οι κατηγορίες ουσιαστικά,το κομμάτι δηλαδή των δεδομένων που προσπαθεί να προβλεφθεί. Στην περίπτωση του παραπάνω παραδείγματος, τα δείγματα είναι οι διάφορες φωτογραφίες με δέντρα ή χωρίς, και οι ετικέτες είναι οι δύο κατηγορίες : «Η φωτογραφία περιέχει δέντρο» και «Η φωτογραφία δεν περιέχει δέντρο».

Τα σετ αυτό των κατηγοριοποιημένων δεδομένων λοιπόν,χωρίζεται σε δύο τμήματα. Το ένα θα χρησιμοποιηθεί για να εκπαιδεύσει το μοντέλο, ενώ το άλλο θα χρησιμοποιηθεί για την αξιολόγηση του, συγκρίνοντας τις ετικέτες που προέβλεψε το μοντέλο, με τις πραγματικές και ήδη υπάρχουσες. Εάν η απόδοση του είναι αρκούντως καλή, τότε το μοντέλο μπορεί να χρησιμοποιηθεί πάνω σε καινούργια (ακατηγοριοποίητα) δεδομένα για την κατηγοριοποίηση τους.

Στην περίπτωση της συγκεκριμένης διατριβής, όπως ειπώθηκε, για την εκπαίδευση (training) και αξιολόγηση (evaluation) των μοντέλων, χρησιμοποιήθηκαν τα αρχεία που χρησιμοποιήθηκαν τελευταία (βλ. σελ. 46) ,μιας και οι ετικέτες που φέρουν είναι οι εγκυρότερες δυνατές σε αυτό το σημείο. Όπως εξηγήθηκε πιο πάνω, όσο πιο έγκυρη η εκάστοτε αρχική κατηγοριοποίηση των tweets , τόσο πιο έγκυρα θα είναι και τα μοντέλα μηχανικής μάθησης που θα εκπαιδευτούν πάνω σε αυτά.

Μία άλλη παράμετρος που καθορίζει την αποτελεσματικότητα ενός μοντέλου, είναι το πλήθος των δεδομένων που χρησιμοποιείται για την εκπαίδευση του. Είναι σαφές πως όσα περισσότερα δεδομένα συμβάλλουν στην δημιουργία του, τόσο πιο αποτελεσματικό θα είναι. Αυτό ισχύει διότι κάθε καινούργιο δεδομένο που χρησιμοποιείται από το μοντέλο, αποτελεί και μια πληροφορία, η οποία του επιτρέπει με αυτήν την επιπλέον γνώση, να κάνει καλύτερες προβλέψεις.

Όσον αφορά τα δεδομένα πάνω στα οποία γίνεται η αξιολόγηση, πρέπει και αυτά να είναι αρκούντως πολλά, ώστε η αξιολόγηση να είναι όσο πιο έμπιστη γίνεται.Όπως αναφέρθηκε και πιο πάνω, τα δεδομένα εκπαίδευσης και αξιολόγησης (train and test data) προέρχονται από την ίδια βάση δεδομένων, και επιλέγεται από τον χρήστη πόσα από αυτά θα διατεθούν για την μία δουλειά και πόσα για την άλλη.

Διαχωρισμός δεδομένων – train test split

Υπάρχουν παραπάνω από μία μέθοδοι που κάνουν αυτόν τον διαχωρισμό. Στην περίπτωση μας χρησιμοποιείται η δημοφιλέστερη (Train–Test Data Split), η οποία χωρίζει τα δεδομένα ακριβώς όπως περιγράφηκε. Με μια απλή αναζήτηση στο διαδίκτυο φαίνεται ότι συνήθως γίνεται ο διαχωρισμός 80-20 (80% των δεδομένων για το κομμάτι της εκπαίδευσης (train) και 20% για το κομμάτι της αξιολόγησης (test)), ενώ σε πολύ μεγάλες βάσεις δεδομένων μπορεί να χρησιμοποιηθεί πιο σπάνια και το 50-50.Όπως θα φανεί και θα εξηγηθεί πιο κάτω, στην περίπτωση μας θα χρησιμοποιηθεί ένα πρότυπο 95,5-0,5 .

Παράμετροι μέτρησης απόδοσης μοντέλων μηχανικής μάθησης

Αφού το εκάστοτε μοντέλο εκπαιδευτεί, δοκιμάζεται όπως αναφέρθηκε πάνω σε καινούργια δεδομένα προς εξαγωγή της απόδοσης του. Να σημειωθεί ότι στην περίπτωση της συγκεκριμένης εργασίας έχουμε κατηγοριοποίηση με βάση τρεις ετικέτες (Negative, Neutral, Positive) (multiclass classification) και όχι δυαδική (binary) ,οπότε όλες οι μέθοδοι μέτρησης (metrics) της απόδοσης των μοντέλω νπαρουσιάζονται με βάση την μορφή που εμφανίζονται σε αυτά τα συγκεκριμένα προβλήματα.

Μια ομάδα διαφόρων μετρικών που προσφέρει η βιβλιοθήκη sklearn της python, είναι το classification report (αναφορά κατάταξης).Αυτά είναι τα παρακάτω :

Accuracy (συνολική ορθότητα)

Η συγκεκριμένη παράμετρος, πέρα από το ότι συχνά είναι από τις πιο αντιπροσωπευτικές, είναι η μόνη που αποτελείται από έναν μόνο αριθμό αριθμό, με αποτέλεσμα να είναι συχνά και η πρώτη που κοιτάμε. Ορίζεται πολύ απλά ως ο αριθμός των κατηγοριοποιήσεων που έγιναν σωστά προς όλες τις κατηγοριοποιήσεις

$$accuracy = \frac{Αριθμός\ σωστών\ προβλέψεων}{Συνολικός\ αριθμός\ προβλέψεων}$$

Ενώ συχνά αποτελεί μια αντιπροσωπευτική μέτρηση, χρειάζεται πάντα να λαμβάνονται υπόψιν και οι υπόλοιπες παρακάτω. Ειδικά στην περίπτωση όπου οι κατηγορίες έχουν μεγάλες αποκλίσεις ως προς τον αριθμό δεδομένων τους, εκεί ένα μεγάλο accuracy δε σημαίνει αναγκαστικά ότι το μοντέλο δουλεύει καλά. Αυτό γίνεται αμέσως αντιληπτό αν σκεφτούμε την περίπτωση ενός μοντέλου που έχει εκπαιδευτεί να αναγνωρίζει σε φωτογραφίες ανθρώπινου δέρματος αν υπάρχει κάποιος καρκίνος, κατατάσσοντας τες αντίστοιχα σε δύο κατηγορίες, «φωτογραφία με καρκίνο» και «φωτογραφία χωρίς καρκίνο». Αν επεξεργαστεί 1000 φωτογραφίες από τις οποίες έχουν το καρκίνωμα οι 10,αλλά καταφέρει να εντοπίσει μόνο τις 3,το accuracyθα είναι με βάση τον παραπάνω ορισμό 993 / 1000 = 9.93, νούμερο άριστο παρά την προφανή μη ικανοποιητική λειτουργία του μοντέλου.

Precision (Ακρίβεια) – Recall (Ανάκληση) – F1 score

Αυτήν η ομάδα μετρικών αναφέρεται σε κάθε κατηγορία ξεχωριστά.Το precision, είναι το ποσοστό των στοιχείων που κατατάχθηκαν στην κατηγορία X από τον αλγόριθμο, τα οποία ανήκουν όντως στην κατηγορία X.

$$\text{Precision} = \frac{\text{κατατάχθηκε στην κατηγορία X και ανήκει όντως στην X}}{\text{κατατάχθηκε στην κατηγορία X}}$$

Στο παράδειγμα με τον καρκίνο, το precision της κατηγορίας «φωτογραφία με καρκίνο»θα ήταν 3 / 3 = 1

Το recall, είναι το ποσοστό των στοιχείων που ανήκουν στην κατηγορία X και κατατάχθηκαν όντως στην κατηγορία X.

$$\text{Recall} = \frac{\text{στοιχεία που κατατάχθηκαν στην X}}{\text{Σύνολο στοιχείων που ανήκουν στην X}}$$

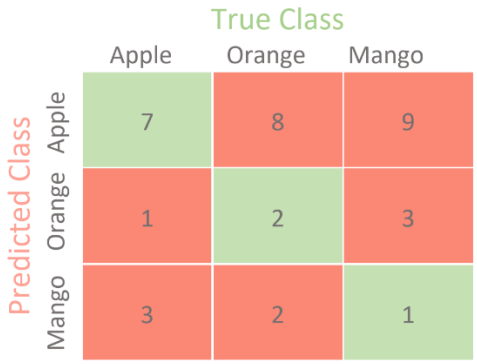
Στο ίδιο παράδειγμα,το recall της κατηγορίας «φωτογραφία με καρκίνο» θα ήταν 3 / 10 = 0.3. Οπότε αυτό θα ήταν το σημείο όπου θα φαινόταν η δυσλειτουργία του μοντέλου του παραδείγματος. Γι'αυτό πρέπει όλα τα μετρικά να λαμβάνονται υπόψιν.

Το F1 score είναι απλά ο μέσος όρος των precision και recall.

Macro avg – Weighted avg

Τέλος, στο classification report βρίσκονται και τα Macro και Weighted avg. Ουσιαστικά είναι ο μέσος όρος των F1 scores μιας κατηγορίας, χωρίς να δίνεται προσοχή στο ποσοστό των στοιχείων της συγκεκριμένης κλάσης σε σχέση με τις υπόλοιπες στην περίπτωση του Macro, και το αντίθετο στην περίπτωση του Weighted. Πρόκειται για παραμέτρους που δεν θα μας απασχολήσουν ιδιαίτερα.

Confusion Matrix



Εικόνα 5.34 . Παράδειγμα confusion matrix

Ο confusion matrix (πίνακας σύγχυσης, αλλά κρατείται ο αγγλικός όρος) αποτελεί μια οπτικοποίηση των σωστών προβλέψεων του μοντέλου. Γενικά, αν ονομάσουμε X1,X2,...,Xi τις γραμμές του, όπου θα βρίσκονται οι πραγματικές τιμές, και τις στήλες του αντίστοιχα όπου θα βρίσκονται οι προβλεπόμενες από το μοντέλο (όπου X1,X2,…,Xi είναι τα labels) ,τότε το κάθε στοιχείο - έστω (Xk, Xm)- θα μας δείχνει τον αριθμό των τιμών που είχαν label Xk και κατηγοριοποιήθηκαν από τον αλγόριθμο ως Xm. Προφανώς στα κελιά στην διαγώνιο του πίνακα k = m, οπότε εκεί φαίνονται πόσες τιμές κατατάχθηκαν σωστά από το μοντέλο στην πραγματική τους κλάση.

ROC – AUC score

Το ROC-AUC score είναι ένας αριθμός από το 0 μέχρι το 1, ο οποίος αναπαριστά τον χώρο κάτω από την καμπύλη ROC-AUC.Όσο μεγαλύτερος αυτός ο αριθμός, τόσο πιο αποδοτικό είναι το μοντέλο. Στην περίπτωση μας χρησιμοποιείται απλά ως ένα συμπληρωματικό μετρικό, καθώς όπως θα φανεί, τα παραπάνω μετρικά θα είναι αρκετά για να επιλεχθούν τα καλύτερα μοντέλα. Λόγω της δευτερεύουσας λοιπόν

σημασίας του σε αυτήν την εργασία, παραλείπεται η παράθεση περισσότερων λεπτομεριών σχετικά με την καμπύλη ROC-AUC και το γενικότερο πλαίσιο αυτού του αριθμού, μιας και προυποθέτει να οριστούν πρώτα διάφορες έννοιες που δεν συνάδουν με τα ενδιαφέροντα της συγκεκριμένης εργασίας.

Δημιουργία των δεδομένων εκπαίδευσης και αξιολόγησης

Αφού διατυπώθηκαν τα παραπάνω, θα αναλυθεί στην συνέχεια ο τρόπος που δημιουργήθηκαν τα δεδομένα για τα κομμάτια της εκπαίδευσης και της αξιολόγησης, πως ακριβώς αξιολογήθηκαν και πως προέκυψαν τα δύο καλύτερα.

Αρχικά λοιπόν εισάγονται τα δεδομένα από τα αρχεία twt_24_02_sentiments , twt_25_02_sentiments , twt_26_02_sentiments και twt_27_02_sentiments . Ένα από τα πρώτα πράγματα που πρέπει να αποφασιστούν , είναι το πώς θα χρησιμοποιηθούν αυτά τα αρχεία για την εκπαίδευση και την αξιολόγηση των μοντέλων. Επιλέχθηκε λοιπόν, τα 3 πρώτα αρχεία να ενωθούν σε ένα ενιαίο, στο οποίο θα εφαρμοστεί η τεχνική train-test split όπως παρουσιάστηκε παραπάνω. Έπειτα, τα μοντέλα που θα δημιουργηθούν θα δοκιμαστούν σε ακόμη μεγαλύτερη κλίμακα, βάζοντας τα να κάνουν προβλέψεις για όλα τα τουίτς του 4^{ου} αρχείου που δεν χρησιμοποιήθηκε προηγουμένως. Εκεί θα γίνει πάλι αξιολόγηση των μοντέλων, η οποία λόγω μεγαλύτερης κλίμακας θα είναι και πιο έμπιστη.

Λόγω αυτής της δεύτερης αξιολόγησης, αποφασίστηκε στο train-test split του ενιαίου αρχείου, το test κομμάτι να είναι τόσο μικρό (5% των συνολικών τιμών). Διότι, θέλουμε μεν να έχουμε μια αρχική εικόνα για το πόσο καλά λειτουργεί το εκάστοτε μοντέλο, πριν πάμε στην προαναφερθείσα δεύτερη και πιο ουσιαστική αξιολόγηση, αλλά περισσότερο μας αφορά σε αυτό το σημείο να έχει όσο το δυνατόν περισσότερα δεδομένα κατά την εκπαίδευση του. Οπότε αντί να γίνει ένα split 80-20 όπως ενδεχομένως να ήταν αναμενόμενο, όπου εκεί το 80 αντί για το 95% των δεδομένων θα αφιερωνόταν στην εκπαίδευση των μοντέλων, έγινε το 95 - 5 που περιεγράφηκε. Έχουμε λοιπόν τα :

- X_train: Το μέρος των samples που αφιερώνεται στην εκπαίδευση του μοντέλου (95%)
- X_test :Το μέρος των samples που αφιερώνεται στην αξιολόγηση του μοντέλου (5%)
- Y_train: Το μέρος των labels που αφιερώνεται στην εκπαίδευση του μοντέλου (95%)
- Y_test: Το μέρος των labels που αφιερώνεται στην αξιολόγηση του μοντέλου (5%)

Σε αυτό το σημείο είναι καθορισμένο ποια θα είναι τα samples (τα επεξεργασμένα tweets) και τα labels (το Final Sentiment).Ωστόσο όμως, δεν είναι ακόμα στην τελική τους μορφή, κι αυτό γιατί τα μοντέλα αναγνωρίζουν μόνο αριθμούς και όχι κείμενα. Οπότε πρέπει αμφότερα τα samples και τα labels να παραστούν με αριθμητικό τρόπο.

Στην περίπτωση των labels τα πράγματα είναι εύκολα. Μετατρέπονται οι λέξεις Negative,Neutral και Positive στους αριθμούς 0, 1, 2 κατ' αντιστοιχίαν ,χρησιμοποιώντας εργαλείο της βιβλιοθήκης sklearn (LabelEncoder).

Στην περίπτωση των samples όμως, προφανώς δεν μπορεί να γίνει κάτι αντίστοιχο. Ο τρόπος που χρησιμοποιείται για την αριθμητική παράσταση τους, είναι η προαναφερθείσατεχνική της διανυσματοποίησης TF-IDF (TF-IDF Vectorization).Εκπαιδεύεται λοιπόν ένα αντικείμενο (object) TF-IDFπάνω στα κείμενα (τουίτς) του X_train,και έπειτα χρησιμοποιείται το ίδιο εκπαιδευμένο object για την διανυσματοποίηση όλων των κειμένων πάνω στα οποία θα εφαρμοστεί το μοντέλο που θα εκπαιδευτεί για να κάνει τις προβλέψεις.

Δημιουργία και αξιολόγηση των μοντέλων μηχανικής μάθησης

Σε αυτό το κομμάτι, θα δημιουργηθούν τα 5 μοντέλα μηχανικής μάθησης. Για το κάθε ένα θα γίνει λόγος για τον τρόπο λειτουργίας του και την επίδοση του με βάση τα παραπάνω μετρικά.

Bernoulli Naive Bayes Model

Λίγα λόγια για τον αλγόριθμο

Σύμφωνα με το documentation (εγχειρίδιο) της βιβλιοθήκης scikit_learn, οι μέθοδοι Naïve Bayes είναι ένα σύνολο επιβλεπόμενων μοντέλων μηχανικής μάθησης που βασίζονται στο θεώρημα του Bayes (το οποίο παρουσιάζεται αναλυτικά στην ιστοσελίδα της scikit-learn), βασιζόμενα στην ισχυρή (“naive”) υπόθεση της υπό όρους ανεξαρτησίας των samples, δεδομένων των labels. Η Bernoulli Naïve Bayes μέθοδος είναι μία από αυτές, οι οποίες γενικά διαφέρουν η μία από την άλλη μόνο όσον αφορά την κατανομή με την οποία συσχετίζονται τα features με τα labels (Γκαουσσιανή, Πολυωνυμική, Μπερνούλλι και άλλες).

Παρά την απλή λειτουργία τους, οι συγκεκριμένοι αλγόριθμοι δουλεύουν αρκετά καλά σε αρκετά πραγματικά προβλήματα (κυρίως για κατάταξη κειμένων και ανίχνευση μηνυμάτων σπαμ), ενώ διακρίνονται για την ταχύτητα τους καθώς και για την μικρή απαίτηση τους σε training data.Με βάση αυτά, ο συγκεκριμένος αλγόριθμος αποτελεί μια καλή αρχή.

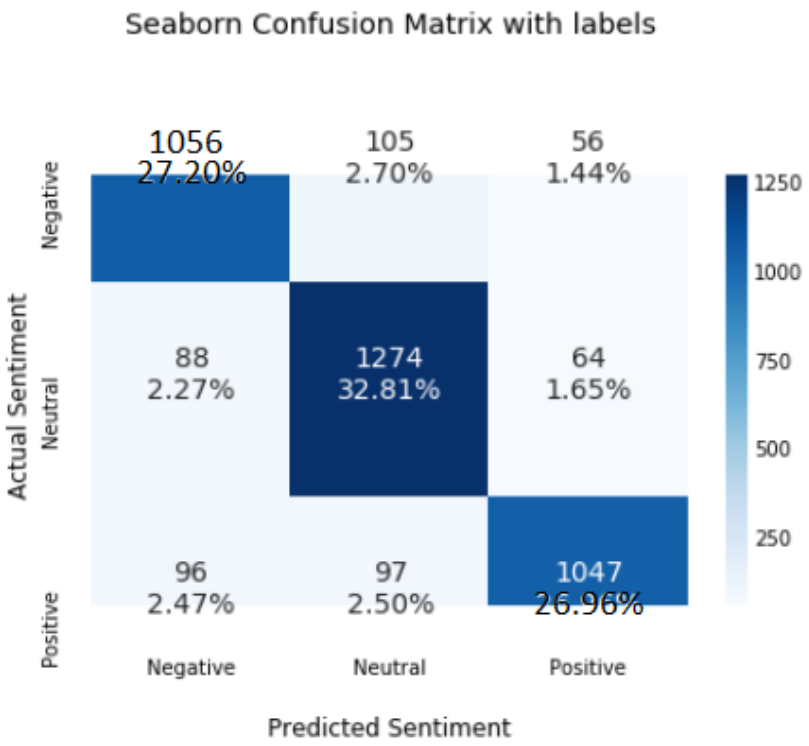
Όπως θα συμβεί και με κάποια από τα επόμενα μοντέλα, έτσι και εδώ το συγκεκριμένο μοντέλο είναι φτιαγμένο για προβλήματα δυαδικής κατάταξης (binary classification). Ωστόσο όμως υπάρχει η δυνατότητα να επεκταθεί και σε προβλήματα με τρεις ετικέτες, όπως και στην περίπτωση μας, εφαρμόζοντας την τεχνική «One Vs Rest».Αυτήν η τεχνική ουσιαστικά μετατρέπει τα προβλήματα με n ετικέτες σε πολλά δυαδικά, έτσι ώστε να μπορούν να εφαρμοστούν οι αλγόριθμοι.

Απόδοση

	precision	recall	f1-score	support
0	0.85	0.87	0.86	1217
1	0.86	0.89	0.88	1426
2	0.90	0.84	0.87	1240
accuracy			0.87	3883
macro avg	0.87	0.87	0.87	3883
weighted avg	0.87	0.87	0.87	3883

```
[[1056  105   56]
 [  88 1274   64]
 [  96   97 1047]]
------------------
ROC-AUC Score: 0.9699035896036937
```

Εικόνα 5.35 . Πίνακας με τα μετρικά για το μοντέλο Bernoulli Naïve Bayes .



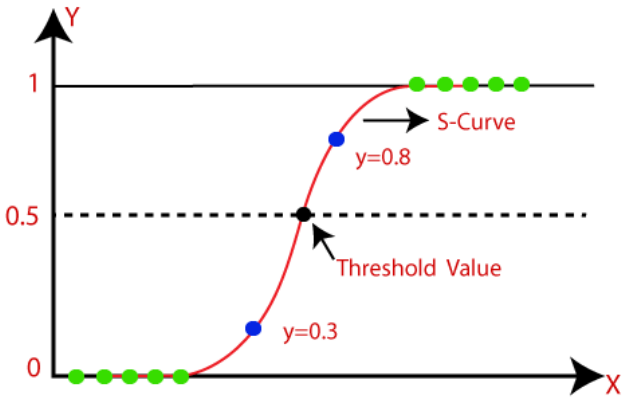
Εικόνα 5.36 . Confusion matrix για το μοντέλο Bernoulli Naïve Bayes .

Όλα τα μετρικά του συγκεκριμένου μοντέλου δείχνουν πολύ καλά, όλα τα νούμερα σχεδόν μαζί με το accuracy είναι πάνω από το 85%, το οποίο σημαίνει ότι το μοντέλο δείχνει να είναι παραπάνω από αρκούντως καλό.

Logistic Regression

Λίγα λόγια για τον αλγόριθμο

Πρόκειται για ένα από τα απλούστερα και πιο πολυχρησιμοποιημένα μοντέλα επιβλεπόμενης μηχανικής μάθησης, και χρησιμοποιείται για να κατατάσσει τα δεδομένα με βάση διακριτές τιμές (0 ή 1, True ή False κλπ). Με κατάλληλες τροποποιήσεις και με βάση την τεχνική «One vs Rest» που αναφέρθηκε και πιο πάνω, η χρήση του αλγορίθμου επεκτείνεται και σε προβλήματα μεγαλύτερης τάξης. Η πρόβλεψη που δίνει ο αλγόριθμος είναι πιθανολογικού χαρακτήρα και είναι μια τιμή από 0 έως 1.Αν αυτή η τιμή είναι μικρότερη από κάποια συγκεκριμένη τότε το feature κατατάσσεται στην κατηγορία 0,που μπορεί να είναι «η φωτογραφία δεν περιέχει δέντρο» με βάση το πιο πάνω παράδειγμα. Ομοίως αν είναι πάνω από αυτήν την τιμή κατατάσσεται στο 1. Το από ποιόν αριθμό και κάτω ή και πάνω (threshold value) θα καταταχθούν στο 0 ή το 1 αντίστοιχα, ορίζεται από τον χρήστη, με συχνή χρήση του αριθμού 0,5.



Εικόνα 5.37. Παράδειγμα Logistic Regression .

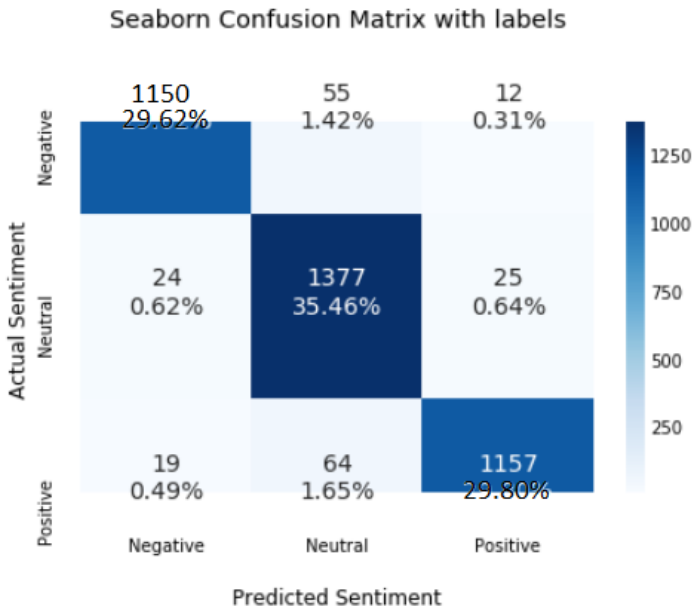
Η σιγμοειδής καμπύλη πάνω στην οποία γίνονται «fit» οι προβλέψεις προέρχεται από την logistic function (λογιστική συνάρτηση) και αναπαριστάται με την γνωστή ως S-Curve.

Απόδοση

	precision	recall	f1-score	support
0	0.96	0.94	0.95	1217
1	0.92	0.97	0.94	1426
2	0.97	0.93	0.95	1240
accuracy			0.95	3883
macro avg	0.95	0.95	0.95	3883
weighted avg	0.95	0.95	0.95	3883

```
[[1150   55   12]
 [  24 1377   25]
 [  19   64 1157]]
------------------
ROC-AUC Score: 0.9928178204363068
```

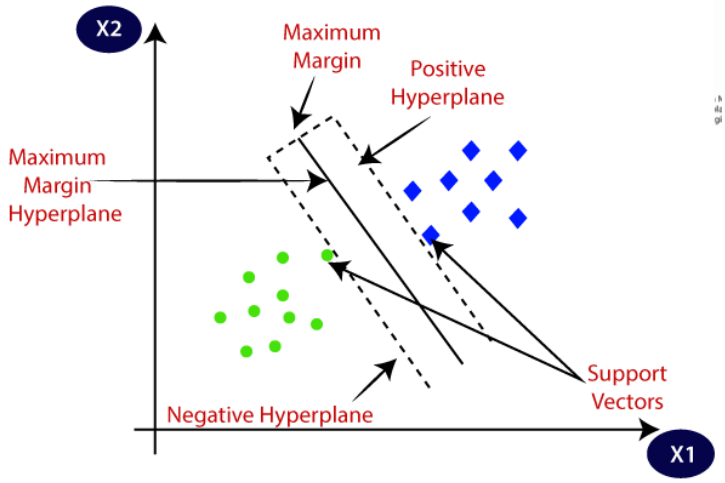
Εικόνα 5.38. Πίνακας με τα μετρικά για το μοντέλο Logistic Regression .



Εικόνα 5.39 . Confusion matrix για το μοντέλο Logistic Regression

Το συγκεκριμένο μοντέλο, όπως καταδεικνύουν και οι εικόνες παραπάνω, φαίνεται να λειτουργεί σχεδόν άψογα. Όλα τα μετρικά είναι πάνω από το 0.92, το οποίο σε συνδυασμό με ένα AUC-Score ουσιαστικά στην μονάδα, μας δείχνει ότι το μοντέλο είναι ιδανικό, και θα είναι όντως ένα από τα δύο καλύτερα. Αυτήν η συμπεριφορά ήταν σχετικά αναμενόμενη, καθώς αρκετές εργασίες έχουν δείξει ότι ο συγκεκριμένος αλγόριθμος δουλεύει πολύ καλά και σε δεδομένα μορφής κειμένου όπως στην περίπτωση μας.

Linear Support Vector Classifier



Λίγα λόγια για τον αλγόριθο

Ο αλγόριθμος linearSVC ανήκει στην γενικότερη οικογένεια των αλγορίθμων επιβλεπόμενης μηχανικής μάθησης Support Vector Machines, και χρησιμοποιείται και για classification και regression. Η λειτουργία τους είναι σχετικά απλή και σχετίζεται με την εύρεση του κατάλληλου συνόρου (hyperplane) που χωρίζει τα

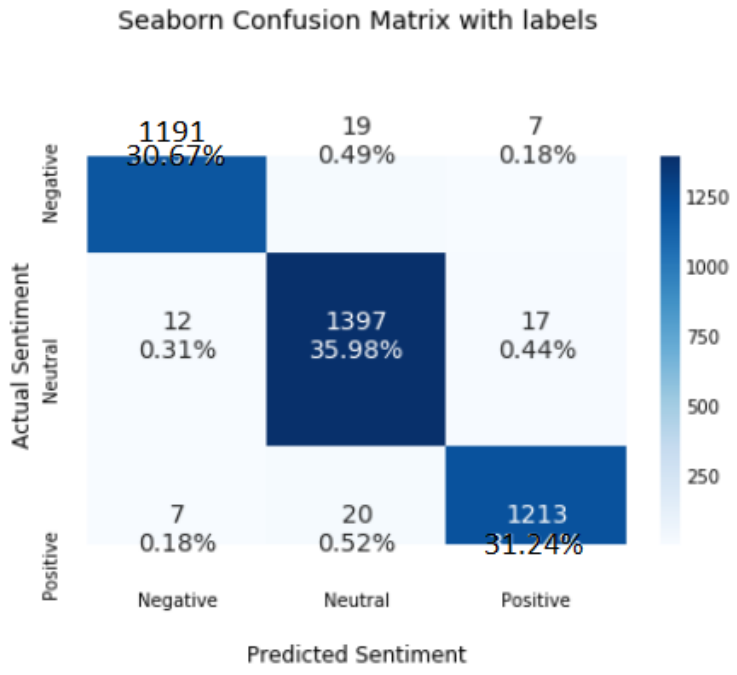
δεδομένα. Ως hyperplane ορίζεται το σύνορο που έχει διάσταση N-1 όταν ανήκει σε έναν χώρο διάστασης N. Με απλά λόγια,σε έναν χώρο 2 διαστάσεων –δηλαδή σε ένα επίπεδο- το hyperplane θα είναι μίας διάστασης –μια ευθεία δηλαδή-,όπως στην εικόνα παραπάνω. Αντίστοιχα εάν τα δεδομένα βρίσκονται στον τρισδιάστατο χώρο –όπως συμβαίνει στην περίπτωση μας- τα hyperplanes που αποτελούν τα σύνορα ανάμεσα στα Negative, Neutral και Positive δεδομένα θα είναι 2 διαστάσεων, δηλαδή επίπεδα. Το βέλτιστο σύνορο προκύπτει με βάση της αποστάσεις από τα στοιχεία της κάθε κλάσης που βρίσκονται πιο κοντά στο σύνορο (support vectors). Το ιδανικό σύνορο είναι αυτό που απέχει την μεγαλύτερη απόσταση (maximum margin) από τα support vectors,και με βάση αυτό κατηγοριοποιούνται και όλα τα υπόλοιπα στοιχεία που ο αλγόριθμος καλείται να κατηγοριοποιήσει. Πολλές εργασίες έχουν δείξει ότι ο αλγόριθμος δουλεύει πολύ καλά σε πολλές περιπτώσεις, όπου μία από αυτές είναι και η περίπτωση της συγκεκριμένης εργασίας, τα δεδομένα δηλαδή μορφής κειμένου.

Απόδοση

	precision	recall	f1-score	support
0	0.98	0.98	0.98	1217
1	0.97	0.98	0.98	1426
2	0.98	0.98	0.98	1240
accuracy			0.98	3883
macro avg	0.98	0.98	0.98	3883
weighted avg	0.98	0.98	0.98	3883

```
[[1191   19    7]
 [  12 1397   17]
 [   7   20 1213]]
```

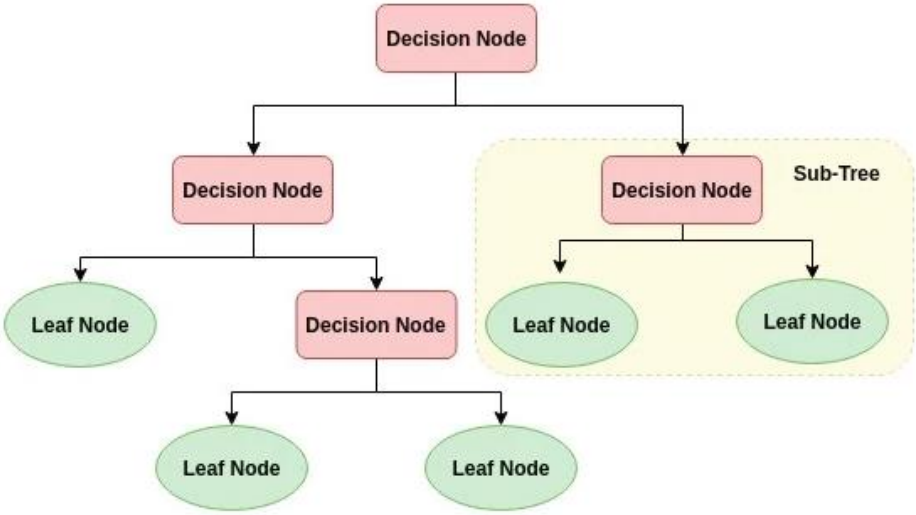
Εικόνα 5.40 . Πίνακας με τα μετρικά για το μοντέλο LinearSVC



Εικόνα 5.41 . Confusion matrix για το μοντέλο LinearSVC

Τα παραπάνω μετρικά δείχνουν πως και αυτό το μοντέλο δουλεύει άριστα. Όλα τα νούμερα είναι στο 0,98 (εκτός από ένα που είναι στο 0.97) , το οποίο είναι το καλύτερο σκορ ως τώρα. Να σημειωθεί πως δεν υπολογίστηκε ROC-AUC score στο συγκεκριμένο μοντέλο, μιας και για λόγους που δεν είναι του παρόντος, η λειτουργία αυτή δεν υποστηρίζεται για τον συγκεκριμένο αλγόριθμο.

.DECISION TREE CLASSIFIER



Λίγα λόγια για τον αλγόριθμο

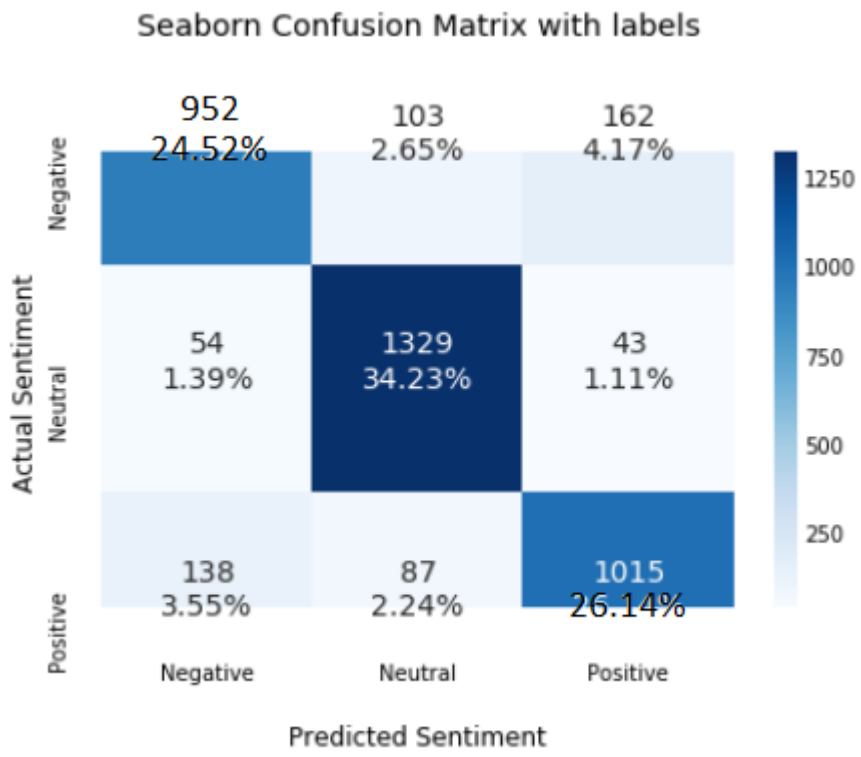
Οι Decision Trees Classifiers είναι μια οικογένεια αλγορίθμων επιβλεπόμενης μηχανικής μάθησης, οι οποίοι επίσης βρίσκουν χρήση και σε προβλήματα classification και regression.Ο τρόπος που λειτουργούν είναι διαφορετικός από τα παραπάνω μοντέλα, καθώς κάνουν τις κατηγοριοποιήσεις με βάση διαδοχικές αποφάσεις που παίρνουν σε κάθε βήμα (η κόμβο) του αλγορίθμου. Ο τρόπος με τον οποίο θα γίνονται αυτές οι αποφάσεις βασίζεται σαφώς στα δεδομένα που έλαβε ο αλγόριθμος κατά την εκπαίδευση του. Η διαδικασία των αποφάσεων ξεκινά από την ρίζα του δέντρου (root of the tree), όπου τα προς κατάταξη δεδομένα κινούνται από εκεί στους επόμενους κόμβους (nodes) με βάση το εάν φέρουν ή όχι διάφορα χαρακτηριστικά. Κάποια στιγμή θα καταλήξουν σε έναν κόμβο ο οποίος θα είναι τερματικός (leaf) ,ο οποίος εν τέλει θα καθορίζει και το label των δεδομένων. Πρόκειται για μοντέλα που φέρουν καλή απόδοση σε πολλές εφαρμογές, όπως κατάταξη εικόνων, ανάλυση αποφάσεων, ανάλυση στρατηγικών και άλλα. Για περισσότερες πληροφορίες ο αναγνώστης μπορεί να συμβουλευτεί την προτεινόμενη βιβλιογραφία.

Απόδοση

	precision	recall	f1-score	support
0	0.83	0.78	0.81	1217
1	0.87	0.93	0.90	1426
2	0.83	0.82	0.83	1240
accuracy			0.85	3883
macro avg	0.85	0.84	0.84	3883
weighted avg	0.85	0.85	0.85	3883

```
[[ 952  103  162]
 [  54 1329   43]
 [ 138   87 1015]]
ROC-AUC Score:
0.8843109879803936
```

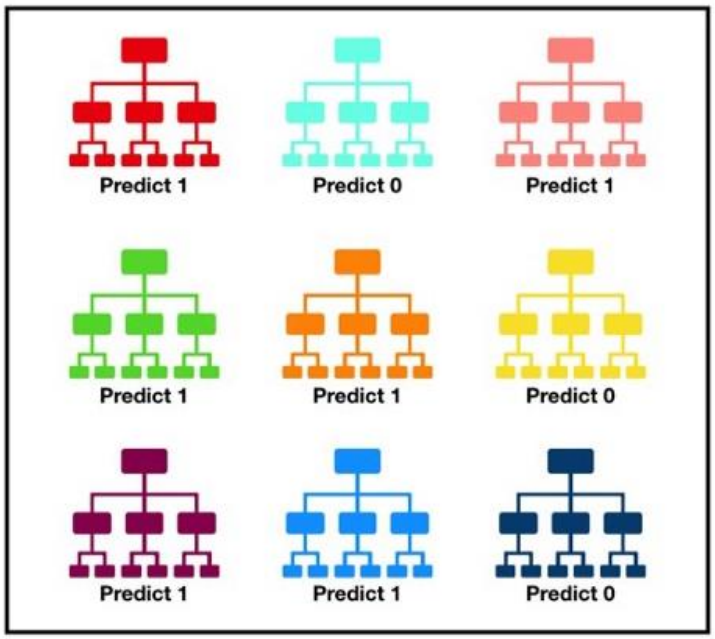
Εικόνα 5.42. Πίνακας με τα μετρικά για το μοντέλο Decision Tree Classifier .



Εικόνα 5.43. Confusion Matrix για το μοντέλο Decision Tree Classifier

Από τα παραπάνω μετρικά φαίνεται πως ο αλγόριθμος δεν λειτουργεί καθόλου άσχημα, αλλά σαφώς υστερεί μπροστά στα μοντέλα Logistic Regression και Linear Support Vectors Classifier.Επίσης, είναι και το αργότερο μέχρι στιγμής μοντέλο, το οποίο το αποτελεί ένα ακόμα μειονέκτημα σε σχέση με τα υπόλοιπα μοντέλα.

Random Forrest Classifier



Six 1s and Three 0s
Prediction: 1

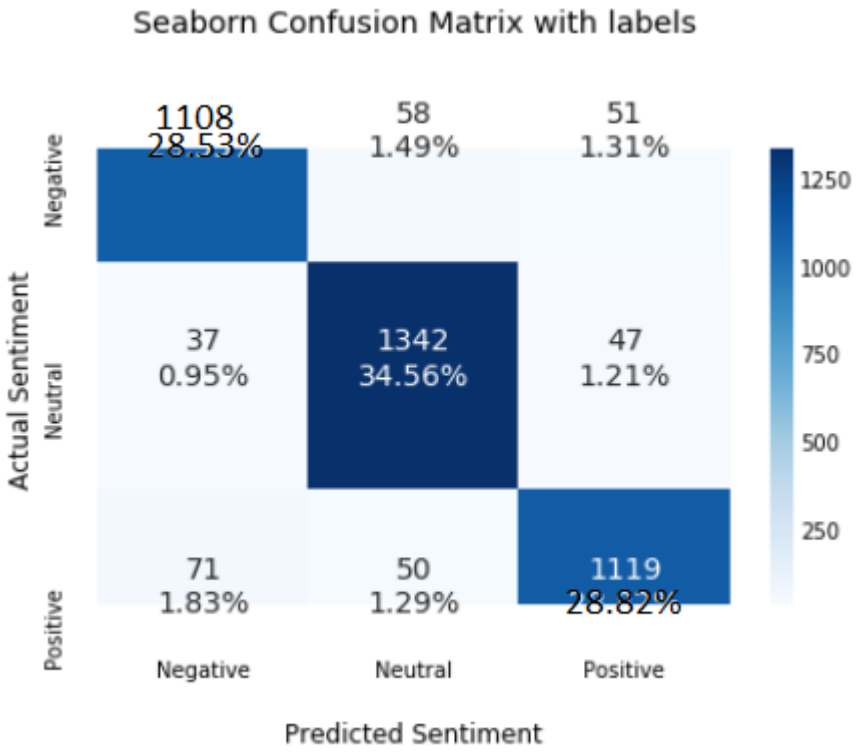
Το συγκεκριμένο μοντέλο σχετίζεται άμεσα με το προηγούμενο, καθώς η λειτουργία του αποτελείται ουσιαστικά από πολλά decision trees τα οποία εν τέλει συγχωνεύονται προς εξαγωγή ενός πιο έγκυρου, συνολικού αποτελέσματος. Αυτό βασίζεται στην λογική παραδοχή ότι ένας μεγάλος αριθμός σχετικά ανεξάρτητων μοντέλων που δουλεύουν για την εξαγωγή ενός κοινού αποτελέσματος, θα δουλέψουν στο σύνολο τους καλύτερα από το κάθε μεμονωμένο μοντέλο (decision tree). Γι αυτό συνήθως η απόδοση των Random Forrests είναι καλύτερη, όπως θα φανεί και στα μετρικά. Ένα από τα αρνητικά του αλγορίθμου είναι ο χρόνος που χρειάζεται για να τρέξει, μιας και ήταν με διαφορά ο μεγαλύτερος, από όλα τα προηγούμενα μοντέλα.

Απόδοση

	precision	recall	f1-score	support
0	0.91	0.91	0.91	1217
1	0.93	0.94	0.93	1426
2	0.92	0.90	0.91	1240
accuracy			0.92	3883
macro avg	0.92	0.92	0.92	3883
weighted avg	0.92	0.92	0.92	3883

```
[[1108   58   51]
 [  37 1342   47]
 [  71   50 1119]]
ROC-AUC Score:
0.9836849269357749
```

Εικόνα 5.44. Πίνακας με τα μετρικά για το Random Forrest Classifier .



Εικόνα 5.45. Confusion Matrix για το Random Forrest Classifier .

Από τα παραπάνω μετρικά φαίνεται μια σημαντική βελτίωση στην απόδοση σε σχέση με την χρήση ενός μοντέλου Decision Tree, όπως και αναμενόταν ως έναν βαθμό άλλωστε. Συγκριτικά με τα υπόλοιπα μοντέλα φέρει την τρίτη καλύτερη απόδοση, όντας ελαφρώς πιο πίσω από τον αλγόριθμο Logistic Regression. Παράλληλα όμως ο δεύτερος χρειάζεται λιγότερο χρόνο, δίνοντας του ακόμα περισσότερο προβάδισμα.

Όπως αναφέρθηκε, στο τέλος του σταδίου ανάπτυξης των 5 μοντέλων, θα γινόταν η επιλογή των δύο καλυτέρων. Με βάση την παραπάνω ανάλυση, και όπως ήδη αναφέρθηκε, είναι ξεκάθαρο πως σε αυτό το σημείο δείχνουν να είναι καλύτερα τα μοντέλα Linear Support Vector Classifier και Logistic Regression,καθώς συνδυάζουν πολύ υψηλές αποδόσεις με αρκετά χαμηλούς χρόνους εκτέλεσης.

Αυτά τα δύο μοντέλα λοιπόν είναι αυτά που θα χρησιμοποιηθούν στο δεύτερο «τεστ», το οποίο όπως περιγράφηκε προηγουμένως είναι να κάνουν προβλέψεις στα δεδομένα του dataframe twt_27_02_sentiments. Επιπλέον σε αυτό το σημείο αποθηκεύονται, ώστε να μη χρειαστεί να ξανά εκπαιδευτούν από την αρχή, με χρήση της βιβλιοθήκης pickle.

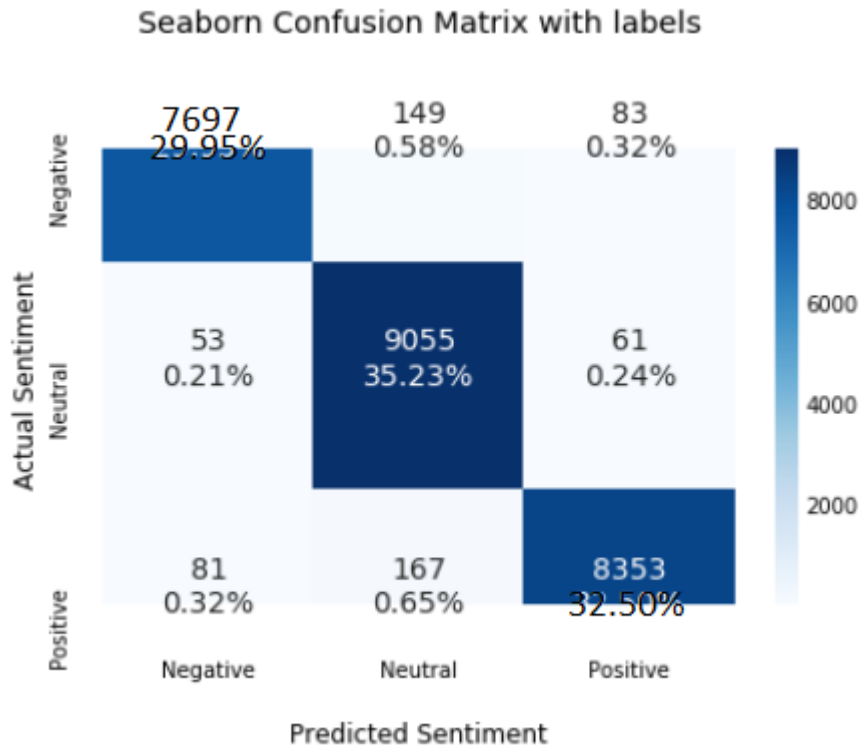
Εφαρμογή των μοντέλων στο dataframe twt_27_02_sentiments – Απόδοση

Αρχικά, όπως και προηγουμένως, πρέπει να ξεχωρίσουμε στο dataframe τα samples από τα labels, και να τα περάσουμε σε αντίστοιχες μεταβλητές. Τα samples σαφώς είναι τα προεπεξεργασμένα τουίτς, και τα labels είναι οι ετικέτες Negative, Neautral και Positive με τις οποίες θα συγκριθούν τα αντίστοιχα labels που θα προβλέψει ο αλγόριθμος. Υπενθυμίζεται ότι τα samples πρέπει να διανυσματοποιηθούν με τον TF-IDF

Vectorizer που εκπαιδεύτηκε πάνω στο training set στην αρχή της διαδικασίας ανάπτυξης των μοντέλων. Να σημειωθεί ξανά πως ο vectorizer που θα χρησιμοποιείται στα δεδομένα προς πρόβλεψη των labels τους , πρέπει να είναι ο ίδιος με αυτόν που εκπαιδεύτηκε πάνω στα δεδομένα τα οποία με την σειρά τους εκπαίδευσαν το εκάστοτε μοντέλο μηχανικής μάθησης (training data).

Απόδοση μοντέλου LinearSVC

	precision	recall	f1-score	support
0	0.98	0.97	0.98	7929
1	0.97	0.99	0.98	9169
2	0.98	0.97	0.98	8601
accuracy			0.98	25699
macro avg	0.98	0.98	0.98	25699
weighted avg	0.98	0.98	0.98	25699

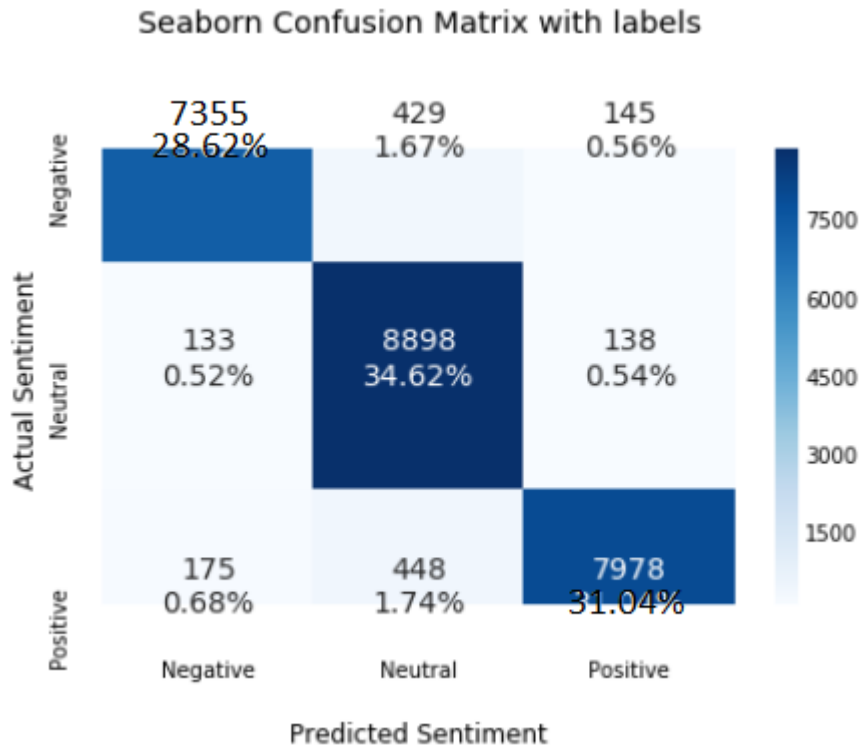


Εικόνα 5.46 . Απόδοση μοντέλου LinearSvc

Απόδοση μοντέλου Logistic Regression

	precision	recall	f1-score	support
0	0.96	0.93	0.94	7929
1	0.91	0.97	0.94	9169
2	0.97	0.93	0.95	8601
accuracy			0.94	25699
macro avg	0.95	0.94	0.94	25699
weighted avg	0.94	0.94	0.94	25699

```
[[7355  429  145]
 [ 133 8898  138]
 [ 175  448 7978]]
```



Εικόνα 5.47. Απόδοση μοντέλου Logistic Regression

Και τα δύο μοντέλα δουλεύουν εξίσου άριστα στο μεγάλο σέτ όπως και προηγουμένως, με το πρώτο όμως να υπερέχει. Οπότε εάν είναι να κρατήσουμε ένα μοντέλο από όλη την διαδικασία, αυτό θα είναι το LinearSVC.

Συμπέρασμα

Σε αυτό το σημείο καλό είναι να εξηγηθεί τι ακριβώς έχει επιτευχθεί στο κεφάλαιο 5.5. Όπως φαίνεται λοιπόν, έχουμε καταλήξει σε ένα μοντέλο με γενική απόδοση στο 98%. Τι σημαίνει όμως αυτό; Αυτό σημαίνει ουσιαστικά ότι το συγκεκριμένο μοντέλο κατέταξε τα τουίτς στην πλειοψηφία των φορών, στην ίδια κατηγορία που τα κατέταξαν η Textblob μάζι με την Vader. Το οποίο με την σειρά του σημάινει ότι σε ένα τυχαίο καινούργιο τουίτ, αντί να εφαρμόσουμε την Textblob και έπειτα την Vader, και να δούμε αν οι δύο κατηγοριοποιήσεις συμφωνούν ώστε να δωθεί το αντίστοιχο label στο τουίτ,μπορούμε απλά να

εφαρμόσουμε αυτό το μοντέλο,και στο 98% των περιπτώσεων θα έχουμε το ίδιο αποτέλεσμα.Και γενικότερα, πρόκειται για ένα μοντέλο που δεδομένου ότι έχει εκπαιδευτεί πάνω σε δεδομένα που έχουν label κοινό από Textblob και Vader, αναμένεται να δίνει πιο έγκυρη πρόβλεψη για το label ενός τουίτ από την αντίστοιχη πρόβλεψη που θα δώσουν η Textblob ή η Vader αν χρησιμοποιηθούν μεμονωμένα.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΕΣ ΕΡΓΑΣΙΕΣ

Συμπεράσματα

Στην παρούσα εργασία παρουσιάστηκαν διάφορες αναλύσεις που μπορούν να γίνουν σε τουίτς για εξαγωγή ποικιλίας πληροφορίας σχετιζόμενη γενικά με την γνώμη του κόσμου, με την μεγαλύτερη προσοχή να δίνεται στις τεχνικές ανάλυσης συναισθήματος.

Πιο συγκεκριμένα, αρχικά έγινε μια σύντομη αλλά ενδιαφέρουσα ανάλυση που σχετίζοταν με τις δημοφιλέστερες τοποθεσίες στα τουίτς, και ακολούθησε έπειτα μία δεύτερη σχετιζόμενη με τις δημοφιλέστερες λέξεις.Αυτές οι αναλύσεις, με κατάλληλη εφαρμογή βρίσκουν ιδιαίτερη χρησιμότητα εάν εφαρμοστούν από εταιρίες για να ενισχύσουν το brand τους ή να εξάγουν πληροφορία σχετικά με κάποιο καινούργιο προιόν η υπηρεσία,όπως αναλύθηκε και παραπάνω.

Έπειτα ακολούθησαν οι βασισμένες σε λεξικό τεχνικές για το πρώτο κομμάτι της ανάλυσης συναισθήματος, όπου φάνηκε μέσω αριθμών και γραφημάτων η γνώμη του κόσμου πάνω στο ζήτημα του πολέμου.Βέβαια, οι δύο τεχνικές που χρησιμοποιήθηκαν (TextBlob και Vader) φέρουν διαφορετικά αποτελέσματα,γιαυτό και χρησιμοποιήθηκε έπειτα και η τεχνική με τα μοντέλα μηχανικής μάθησης.

Τα μοντέλα αυτά λοιπόν,προκειμένου να είναι όσο το δυνατόν πιο αποδοτικά, εκπαιδεύτηκαν μόνο πάνω σε εκείνα τα τουίτς που το συναίσθημα των δύο προηγούμενων τεχνικών ήταν κοινό (άρα και πιο πιθανό να είναι έγκυρο),μιας και όπως αναφέρθηκε, όσο πιο σωστές είναι οι ετικέτες (labels) του σετ εκπαίδευσης, τόσο πιο αποδοτικό θα είναι και το μοντέλο πάνω στις οποίες θα εκπαιδευτεί.Από όλα τα μοντέλα που εκπαιδεύτηκαν, στο τέλος καταλήξαμε σε ένα, το LinearSVC.

Μελλοντικές εργασίες – Βελτιώσεις .

Οι βελτιώσεις που μπορούν να γίνουν αφορούν κυρίως τα μοντέλα μηχανικής μάθησης, μιας και εκεί είναι που οι διαδικασίες εκπαίδευσης ποικίλλουν.

Μια πρώτη βελτίωση θα αφορούσε το προαναφερθέν σετ εκπαίδευσης. Εδώ χρησιμοποιήθηκε σετ εκπαίδευσης που βασίστηκε στις τεχνικές lexicon based, μιας και ήταν στους στόχους της εργασίας να δείξει πώς οι δύο τεχνικές ανάλυσης συναισθήματος (βασισμένες σε λεξικό και μοντέλα μηχανικής μάθησης) μπορούν να συνδυαστούν, και πως γίνεται να εκπαιδευτεί ένα μοντέλο χωρίς να προυπάρχουν έτοιμα ,labeled (με ετικέτες), δεδομένα. Το μοντέλο λοιπόν θα ήταν εν γένει καλύτερο, εάν είχε χρησιμοποιηθεί ένα σετ εκπαίδευσης με 100% έγκυρες ετικέτες (υπάρχουν διάφορα στο διαδίκτυο). Ωστόσο σε αυτήν την εργασία δεν ακολουθήθηκε αυτή η τακτική για τον προαναφερθέν λόγο.

Δεδομένου ενός πιο έγκυρου σετ εκπαίδευσης λοιπόν, μια άλλη βελτίωση θα είχε να κάνει με το είδος των μοντέλων. Σε αυτήν την εργασία χρησιμοποιούνται μοντέλα μηχανικής μάθησης,τα οποία είναι κάπως ξεπερασμένα,δεδομένου ότι υπάρχουν μοντέλα βαθείας μάθησης (deep learning) και νευρωνικά δίκτυα. Οπότε ένα επόμενο βήμα για μια μελλοντική εργασία θα ήταν σίγουρα αυτό, η εκπαίδευση δηλαδη μοντέλων βαθείας μάθησης και νευρωνικών δικτύων

ΒΙΒΛΙΟΓΡΑΦΙΑ

Almatrafi, S. Parack, B. Chavan."*Application of location-based sentiment analysis using Twitter for identifying trends towards Indian general elections 2014*".Proc. The 9th National Conference on Ubiquitus Information Management and Communication. 2015

Arias, Marta, Argimiro Arratia, and Ramon Xuriguera. "*Forecasting with Twitter Data*." ACM Transactions on Intelligent Systems and Technology 5(1):1–24. 2013

Asur, Sitaram and Bernardo A. Huberman. "*Predicting the Future with Social Media.*" Pp. 492–99 in 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT '10. Washington, DC, USA: IEEE Computer Society. 2010

Bishop."*Pattern Recognition and Machine Learning*", Chapter 4.3.4. Springer .2006

Breiman, Friedman, Olshen, and C. Stone, *"Classification and Regression Trees"*, Wadsworth, Belmont, CA, 1984.

Chevalier, Judith A. and Dina Mayzlin. *"The Effect of Word of Mouth on Sales: Online Book Reviews."* Journal of Marketing Research 43(3):345–54.2006

Defazio,Bach,Lacoste-Julien *"SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives".* Montreal, Canada. 2014

Fan, Rong-En, et al., *"LIBLINEAR: A library for large linear classification."*, Journal of machine learning research 9.2008

Gilbert, Eric and Karrie Karahalios. *"Widespread Worry and the Stock Market."* Proceedings of the 4th International AAAI Conference on Weblogs and Social Media 58–65. 2010

Hastie, Tibshirani and Friedman. *"Elements of Statistical Learning"*, Springer, 2009.

Han, Bing. "*Investor Sentiment and Option Prices*." Review of Financial Studies 21(1):387–414. 2008

Hu, Minqing and Bing Liu. 2004b. *"Mining and Summarizing Customer Reviews."* P. 168 in Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04, KDD '04. New York, NY, USA: ACM. 2004

Jansen, Bernard J., Mimi Zhang, Kate Sobel, and Abdur Chowdury. *"Twitter Power: Tweets as Electronic Word of Mouth."* Journal of the American Society for Information Science and Technology 60(11):2169–88. 2009

Jurafsky and Martin. *"Speech and Language Processing",* 2nd edition. Pearson Prentice Hall. ISBN 978-0-13-187321-6. 2008

Lemmon, Michael and Evgenia Portniaguina. *"Consumer Confidence and Asset Prices: Some Empirical Evidence."* Review of Financial Studies 19(4):1499–1529. 2006

Liu, Yang, Xiangji Huang, Aijun An, and Xiaohui Yu. *"Arsa."* P. 607 in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07, SIGIR '07. New York, NY, USA: ACM. 2007

MacArthur, Amanda. *"The Real History of Twitter, In Brief - How the micro-messaging wars were won."* lifewire, 3 Oct. 2016, https://www.lifewire.com/history-ofTwitter-3288854.Ανακτήθηκε 08/05/2022.

Manning, P. Raghavan and H. Schütze , *"Introduction to Information Retrieval"*. Cambridge University Press, pp. 234-265. 2008

McCallum and K. Nigam *"A comparison of event models for Naive Bayes text classification"*. Proc. AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48.1998

Metsis, Androutsopoulos and Paliouras *"Spam filtering with Naive Bayes – Which Naive Bayes?"* 3rd Conf. on Email and Anti-Spam (CEAS).2006

Mohit Kumar Barai **,** [https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/ .2021. Ανακτήθηκε 30/03/2022]

Newberry Christina. 2020. [https://blog.hootsuite.com/social-media-sentiment-analysis-tools/. Ανακτήθηκε 20/03/2022]

Pang, Bo and Lillian Lee. *"Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales."* CoRR abs/cs/0506075.2005

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. *"Thumbs Up?"* Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - EMNLP '02 10:79–86.2002

Popescu, Ana-Maria and Oren Etzioni. *"Extracting Product Features and Opinion from Reviews."* Pp. 339–46 in Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, HLT '05. Stroudsburg, PA, USA: Association for Computational Linguistics.2005

Robinson Scott,Kentucky Farm Bureau,2021 ,[https://www.techtarget.com/searchcustomerexperience/tip/Sentiment-analysis-Why-its-necessary-and-how-it-improves-CX . Ανακτήθηκε 23/03/2022]

Roul Abhinandan. 2021 [https://medium.com/nerd-for-tech/sentiment-analysis-lexicon-models-vs-machine-learning-b6e3af8fe746. Ανακτήθηκε 20/03/2022]

Smola, Bernhard Schölkopf , "*A Tutorial on Support Vector Regression*", Statistics and Computing archive Volume 14 Issue 3, 2004, p. 199-222. 2004

Vaibhav Jayaswal. 2020. [https://towardsdatascience.com/text-vectorization-term-frequency-inverse-document-frequency-tfidf-5a3f9604da6d. Ανακτήθηκε 03/03/2022]

Van Rossum, Guido. "Python Programming Language." USENIX Annual Technical Conference. Vol. 41. 2007.

Wang,D. Can,F. Bar and S. Narayana. *"A system for real-time Twitter sentiment analysis of 2012 U.S presidential election cycle"*, Proc. ACL 2012 System Demonstration,pp. 115-120. 2012

Wu, Lin and Weng, *"Probability estimates for multi-class classification by pairwise coupling"*, JMLR 5:975-1005, 2004.

Zhu, Feng and Xiaoquan (Michael) Zhang. *"Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics."* Journal of Marketing 74(2):133–48. 2010

Links:

https://docs.python.org/3/faq/general.html. Ανακτήθηκε 04/04/2022

[https://getthematic.com/sentiment-analysis/. Ανακτήθηκε 01/05/2022]

[https://monkeylearn.com/natural-language-processing/. Ανακτήθηκε 28/04/2022]

[https://en.wikipedia.org/wiki/Natural_language_processing. Ανακτήθηκε 26/04/2022]

[https://scikit-learn.org/stable/getting_started.html. Ανακτήθηκε 22/04/2022]

[https://www.geeksforgeeks.org/difference-between-matplotlib-vs-seaborn/. 2021, Ανακτήθηκε 17/03/2022]

[https://developer.twitter.com/en/docs/twitter-api. Ανακτήθηκε 05/02/2022]

[https://numpy.org/doc/stable/user/absolute_beginners.html. Ανακτήθηκε 11/02/2022]

[https://en.wikipedia.org/wiki/Pandas_(software). Ανακτήθηκε 12/03/2022]

[https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Ανακτήθηκε 19/04/2022]

[https://en.wikipedia.org/wiki/Decision_tree_learning. Ανακτήθηκε 22/04/2022]

[https://textblob.readthedocs.io/en/dev/ .Ανακτήθηκε 27/03/2022]

ΠΑΡΑΡΤΗΜΑ

Ο κώδικας βρίσκεται στον ακόλουθο σύνδεσμο :

https://github.com/GiorgosThanellas/Russia-Ukraine-War-Tweet-Sentiment-Analysis