ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΣΤΗΝ ΑΝΑΛΥΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ ΚΑΙ ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

ΜΕΛΕΤΗ ΤΩΝ ΑΠΟΨΕΩΝ ΤΩΝ ΧΡΗΣΤΩΝ ΤΟΥ TWITTER ΓΙΑ ΤΗΝ ΚΛΙΜΑΤΙΚΗ ΑΛΛΑΓΗ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΕΡΓΑΛΕΙΑ ΤΗΣ PYTHON (RESEARCH FOR CLIMATE CHANGE USING PYTHON)

Διπλωματική Εργασία

της

Μαρίας Παπαευθυμίου

Θεσσαλονίκη, Φεβρουαριος 2023

ΜΕΛΕΤΗ ΤΩΝ ΑΠΟΨΕΩΝ ΤΩΝ ΧΡΗΣΤΩΝ ΤΟΥ TWITTER ΓΙΑ ΤΗΝ ΚΛΙΜΑΤΙΚΗ ΑΛΛΑΓΗ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΕΡΓΑΛΕΙΑ ΤΗΣ PYTHON (RESEARCH FOR CLIMATE CHANGE USING PYTHON)

Μαρία Παπαευθυμίου

Πτυχίο Οικονομικών, ΑΠΘ, 2017

Διπλωματική Εργασία

υποβαλλόμενη για τη μερική εκπλήρωση των απαιτήσεων του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΤΙΤΛΟΥ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ ΚΑΙ ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Επιβλέπων Καθηγητής
Παπαθανασίου Ιάσων

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Παπαθανασίου Ιάσων Ταραμπάνης Κωνσταντίνος Καλαμπόκης Ευάγγελος

Μαρία Παπαευθυμίου

Ευχαριστίες

Πριν από την έναρξη της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω αρχικα την οικογένεια και τους φίλους μου για την συνεχή υποστήριξη και ενθάρρυνση τους καθ'όλη τη διάρκεια εκπόνησής της καθώς και τον καθηγητή μου κ. Παπαθανασίου Ιάσων για την διαρκή του καθοδήγηση και την άμεση ανταπόκριση σε τυχόν δυσκολίες που μπορεί να προέκυψαν από την αρχή του εργου εως και την υλοποίηση του.

Περίληψη

Η έλευση του Παγκόσμιου Ιστού 2.0 (Web 2.0) και μια μεγάλη γκάμα πλατφορμών ξεκίνησαν να προσφέρουν νέα διαδικτυακά εργαλεία, τα οποία προώθησαν νέες τακτικές επικοινωνίας, προσδίδοντας μια πιο κοινωνική διάσταση στο διαδίκτυο. Η ραγδαία ανάπτυξη των κοινωνικών δικτύων και η ενσωμάτωση τους στην καθημερινή ζωή, την τελευταία δεκαετία, έχει αλλάξει ολοκληρωτικά τις ανθρώπινες σχέσεις, τον τρόπο επικοινωνίας και διάδρασης σε ολόκληρο τον πλανήτη. Η ατέρμονη χρήση των μέσων κοινωνικής δικτύωσης,με 3,8 εκατομμύρια ενεργούς χρήστες ανά τον κόσμο, έχει δημιουργήσει έναν τεράστιο όγκος δεδομένων που παράγεται ανελλιπώς. Καθως αυξάνεται ο όγκος των δεδομένων αυξάνεται και το ενδιαφέρον για την συλλογή, την ανάλυση και την αξιοποίηση τους. Τα αποτελέσματα που προκύπτουν από τις ενέργειες αυτές συνεισφέρουν στη διαδικασία λήψης αποφάσεων σε διάφορους τομείς, οικονομικούς, κοινωνικούς και πολιτικούς.

Οι παραδοσιακοί μέθοδοι διαχείρισης αδυνατούν πλέον να ανταπεξέλθουν στο τεράστιο μέγεθος δεδομένων που παράγεται. Καθώς λοιπόν τα δεδομένα αυξάνονται και αναβαθμίζονται σε Μεγάλα Δεδομένα (Big Data) προκύπτουν νέες καινοτόμες τεχνικές αποθήκευσης, επεξεργασίας και ανάλυσης. Η διαδικασία επεξεργασίας της φυσικής γλώσσας (NLP) και η ανάλυση συναισθήματος αποτελεί ένα από τα σημαντικότερα εργαλεία διαχείρισης των Big Data.

Η κλιματική αλλαγή αποτελεί επίσης ένα από τα πιο δημοφιλή θέματα την τελευταία 20ετία. Ιδιαίτερα τα τελευταία χρόνια, η αλλαγή αυτή γίνεται πιο αισθητή από τον γενικό πληθυσμό και όχι μόνο από τους επιστήμονες. Οι άνθρωποι πια έχουν την ευκαιρία να δημοσιεύουν τις απόψεις τους στα μέσα κοινωνικής δικτύωσης για διάφορα θέματα, όπως η κλιματική αλλαγή, μέσω σχολίων θετικών, αρνητικών ή ουδέτερων. Η ανάλυση συναισθήματος στα κοινωνικά δίκτυα , έχει σημειώσει μεγάλη ανάπτυξη με τη χρήση της επεξεργασίας φυσικής γλώσσας (NLP) και εφαρμόζεται πλέον ευρέως για διάφορους εμπορικούς αλλά και μη εμπορικούς τομείς.

Η παρούσα διπλωματική εργασία πραγματεύεται την ανάλυση συναισθήματος σε περιεχόμενο του Twitter για την κλιματική αλλαγή. Η εξόρυξη των δεδομένων γίνεται μέσω τεχνικής Web scraping. Χρησιμοποιήθηκαν 5 διαφορετικά μοντέλα αναπαράστασης λέξεων NLP και 6 ταξινομητές Machine Learning. Έπειτα έγινε η σύγκριση των αποδόσεών τους, από την οποία τα μοντέλα αναπαράστασης One Hot Unigram, One hot Bigram και TF-IDF με τον ταξινομητή SVM δίνουν την υψηλότερη ακρίβεια, 0,91.

Λέξεις Κλειδιά: Κλιματική Αλλαγή, Υπερθέρμανση του πλανήτη, Κοινωνικά Δίκτυα, Μεγάλα Δεδομένα, Ανάλυση Συναισθήματος, Εξόρυξη Δεδομένων, Twitter, web scraper, climate, climatechange, globalwarming, agw.

Abstract

The advent of the World Wide Web 2.0 and the wide range of platforms started to offer new online tools, which rekindled new tactics of communication, lending a more social dimension to the Internet. The rapid development of social networks and their influence in our lives in the last decade has completely changed human relations and their way of communication and interaction across the planet. The ceaseless use of social media networks, and their 3.8 million active users worldwide, has created a huge amount of data that is continuously generated every day.This is also the reason why there is interest in the collection, analysis and utilization of data from social media.The outcomes of these actions contribute to the decision-making process in various fields such as economic, social and political.

Traditional data management methods are no longer able to cope with the enormous amount of data produced. As data multiplies and evolves into Big Data, new innovative techniques for storing, processing and analyzing data have emerged. Natural language processing and sentiment analysis are some of the emerging fields concerning Big Data.

Climate change is also one of the most popular topics in the last 20 years. Recently this change has become more noticeable not only by scientists, but the general population as well. People now have the opportunity to post their views on various topics such as climate change on social media, through positive, negative or neutral comments. The emotion analysis in social networks, has grown greatly with the use of natural language processing (NLP). And it is now widely applied in various commercial as well as non-commercial areas.

This thesis deals with sentiment analysis in Twitter content on climate change. Data mining is done through Web scraping technique. In addition to that, 5 different NLP word representation models and 6 Machine Learning classifiers were used. From the comparison of their performances, the representation models of One Hot

Unigram, One hot Bigram and TF-IDF with SVM classifiers gave the best results, with an accuracy of 0.91.

Περιεχόμενα

Κατάλογος Σχημάτων x

Κατάλογος Πινάκων xiii

1 Εισαγωγή 1
1.1 Σκοπός της Διπλωματικής . . . 1
1.2 Οργάνωση της Διπλωματικής . . . 3

2 Ανασκόπηση Βιβλιογραφίας 4
2.1 Μέσα Κοινωνικής Δικτύωσης (Social Media) . . . 4
2.1.1 Twitter . . . 6
2.1.2 Κλιματική Αλλαγή στο Twitter . . . 8
2.2 Εξόρυξη Δεδομένων - Μεγάλα Δεδομένα . . . 10
2.2.1 Μεγάλα Δεδομένα (Big Data) . . . 12
2.2.2 Εξόρυξη Δεδομενων (Data Mining) . . . 15
2.2.3 Εξόρυξη Δεδομενων απο το Twitter (Twitter Data Mining) . . . 17
2.3 Sentiment Analysis . . . 21
2.4 Τεχνητή Νοημοσύνη (Artificial Intelligence) . . . 23

3 Μεθοδολογία 31
3.1 Μοντέλα αναπαράστασης λέξεων στην NLP . . . 31
3.1.1 Κλασικά Μοντέλα . . . 32
3.1.2 Word Embeddings - Ενσωματώσεις Λέξεων . . . 37
3.2 Machine Learning Models . . . 44
3.2.1 Logistic Regression . . . 46
3.2.2 Random Forest . . . 47
3.2.3 Support Vector Machine - SVM . . . 48
3.2.4 Naive Bayes . . . 50
3.2.5 K-Nearest Neighbours . . . 51
3.2.6 Extreme Gradient Boosting - XGBoost . . . 53

4 Ανάπτυξη του Θέματος 55
4.1 Εξόρυξη δεδομένων . . . 55
4.1.1 Εξόρυξη δεδομένων απο το Twitter . . . 55
4.1.2 Μεθοδος Web Scraping vs Twitter API . . . 55
4.1.3 Λογαριασμοί και hashtags . . . 56
4.1.4 Αποτελέσματα από το snscrape . . . 59
4.2 Περιγραφική Στατιστική των Δεδομενων . . . 61
4.2.1 Ισορροπία Κατηγοριών . . . 61
4.2.2 Αριθμοί των Likes . . . 62
4.2.3 Αριθμός Αναπαραγωγής του Tweet (Retweet) . . . 63
4.2.4 Αριθμός Απαντήσεων . . . 64
4.2.5 Αριθμός Αναφορών σε άλλο Tweet . . . 65
4.2.6 Αριθμός Λέξεων . . . 66

4.2.7 Αριθμός Ιστοσελίδων . . . 67
4.2.8 Αριθμός Αναφορών σε άλλους Λογαριασμούς . . . 68
4.2.9 Αριθμός Σημείων Στίξεων, όπως !, ? . . . 69
4.2.10 Αριθμός Ετικετών (Hashtags) . . . 70
4.2.11 Αριθμός Emojis . . . 71
4.2.12 Αριθμός Κεφαλαίων Λέξεων . . . 72
4.3 Εκκαθάριση Δεδομένων . . . 73
4.3.1 Αναφορές σε άλλους χρήστες . . . 73
4.3.2 Παραθέση ιστοσελίδων . . . 74
4.3.3 Emojis . . . 74
4.3.4 Σημεία στίξης . . . 75
4.3.5 Συντομογραφίες . . . 75
4.3.6 Νούμερα . . . 75
4.3.7 Μετατροπή σε πεζά . . . 76
4.3.8 Αφαίρεση stopwords, ή αλλιώς πολύ συχνών χαρακτήρων . . . 76
4.3.9 Αποκοπή Καταλήξεων . . . 77
4.3.10 Τελική Βάση Δεδομένων . . . 78
4.4 Μοντέλα αναπαράστασης λέξεων στην NLP . . . 79
4.5 One Hot Encoding - Bag Of Words (BoW) . . . 80
4.5.1 Uni-grams . . . 80
4.5.2 Bi-grams . . . 81
4.5.3 Term Frequency – Inverse Document Frequency (TF-IDF) . . . 82
4.5.4 Word2Vec . . . 82
4.5.5 FastText . . . 83

5 Ανάλυση Δεδομένων και Ερμηνεία των Ευρημάτων 84
5.1 One Hot Encoding - Unigrams . . . 85
5.2 One Hot Encoding - Bigrams . . . 88
5.3 TF-IDF . . . 91
5.4 Word2Vec . . . 94
5.5 FastText . . . 97

6 Συμπεράσματα και Προτάσεις 99

Βιβλιογραφία/Αρθρογραφία 100

Α΄ Κώδικας viii
Α΄.1 Εξόρυξη Δεδομένων . . . viii
Α΄.1.1 Λογαριασμοί και Hashtags . . . viii
Α΄.2 Μοντέλα αναπαράστασης λέξεων NLP . . . xvi
Α΄.2.1 One Hot Encoding - Unigrams . . . xvi
Α΄.2.2 One Hot Encoding - Bigrams . . . xvi
Α΄.2.3 TF-IDF . . . xvii
Α΄.2.4 Word2vec . . . xvii
Α΄.2.5 FastText . . . xviii

Α΄.3 Μοντέλα Μηχανικής Μάθησης - Training xviii
Α΄.3.1 Logistic Regression xix
Α΄.3.2 Random Forests xix
Α΄.3.3 Support Vector Machine xix
Α΄.3.4 Gaussian Naive Bayes xix
Α΄.3.5 Multinomial Naive Bayes xx
Α΄.3.6 K Nearest Neighbors xx
Α΄.3.7 XGBoost xx
Α΄.4 Μοντέλα Μηχανικής Μάθησης - Testing xx

Κατάλογος Σχημάτων

1 Κλασικά μοντέλα Αναπαράστασης Λέξεων 42
2 Ισορροπία Κατηγοριών 61
3 Αριθμός Like 62
4 Αριθμός Retweet 63
5 Αριθμός Replies 64
6 Αριθμός Quotes 65
7 Αριθμός Λέξεων 66
8 Αριθμός Ιστοσελίδων 67
9 Αριθμός Αναφορών σε άλλους Λογαριασμούς 68
10 Αριθμός Σημείων Στίξεων 69
11 Αριθμός Hashtags 70
12 Αριθμός Emojis 71
13 Αριθμός Κεφαλαίων Λέξεων 72
14 Τα πιο συχνά θέματα λέξεων 78
15 Καμπύλη ROC για τα μοντέλα One Hot Encoding - Unigrams 87
16 Καμπύλη ROC για τα μοντέλα One Hot Encoding - Bigrmas 90
17 Καμπύλη ROC για τα μοντέλα TF-IDF 93
18 Καμπύλη ROC για τα μοντέλα Word2Vec 96
19 Καμπύλη ROC για τα μοντέλα FastText 98
20 Hashtags viii
21 Υποστηρικτές viii
22 Αρνητές viii
23 Συλλογή αρνητικών tweet viii
24 Συλλογή θετικών tweet ix
25 Τελικό dataset ix
26 Ισσοροπία dataset ix
27 Dataset με περαιτέρω πληροφορίες ix
28 Replies x

29 Retweets x
30 Likes x
31 Quotes x
32 Words x
33 Mentions xi
34 Hashtags xi
35 Capital words xi
36 Marks words xi
37 Sites words xi
38 Emojis words xii
39 Αφαίρεση αναφορών xii
40 Αφαίρεση ιστοσελίδων xii
41 Μετατροπή των emojis xii
42 Αφαίρεση σημείων στίξεως xii
43 Συνάρτηση για μετατροπή της αργκό xiii
44 Μετατροπή των λέξεων αργκό xiii
45 Αφαίρεση αριθμών xiii
46 Μετατροπή όλων σε μικρά xiii
47 Αφαίρεση των stopwords xiv
48 Εφαρμογή stemmer xiv
49 Επιπλέον αφαιρέσεις xiv
50 Συχνά εμφανιζόμενες λέξεις xiv
51 Αποθήκευση "καθαρισμένου" Dataset xv
52 Δημιουργία των train και test data xvi
53 Διανυσματοποιήση One Hot Encoding - Unigrams xvi
54 Διανυσματοποίηση One Hot Encoding - Bigrams xvi
55 Διανυσματοποίηση TF-IDF xvii
56 Διανυσματοποίηση Word2Vec xvii
57 Διανυσματοποίηση FastText xviii
58 Λίστα των διάφορων training set για κάθε αναπαράσταση xviii

59 Logistic Regression xix
60 Random Forests xix
61 Support Vector Machine xix
62 Gaussian Naive Bayes xix
63 Multinomial Naive Bayes xx
64 K Nearest Neighbors xx
65 XGBoost xx
66 Λίστα των διάφορων testing set για κάθε αναπαράσταση xx
67 Συνάρτηση summary για τη δημιουργία διαγραμμάτων xxi
68 Τελικά αποτελέσματα - Εφαρμογή συνάρτησης summary xxi

Κατάλογος Πινάκων

1 Διαφορές μεταξύ Twitter API και Web Scraper 20
2 Hashtags που χρησιμοποιήθηκαν στην αναζήτηση tweets 56
3 Λογαριασμοί απο τους οποίους εξορύχθηκαν tweets 58
4 Στήλες που επιστρέφονται από την βιβλιοθήκη snscrape 60
5 Περιγραφική Στατιστική για τον Αριθμό των Likes 62
6 Περιγραφικη Στατιστική για τον Αριθμό Αναπαραγωγης του Tweet (Retweet) 63
7 Περιγραφική Στατιστική για τον Αριθμό Απαντήσεων 64
8 Περιγραφική Στατιστική για τον Αριθμό Αναφορών σε άλλο Tweet 65
9 Περιγραφική Στατιστική για τον Αριθμό Λέξεων 66
10 Περιγραφική Στατιστική για τον Αριθμό Ιστοσελίδων 67
11 Περιγραφική Στατιστική για τον Αριθμό Αναφορών σε άλλους Λογαριασμούς 68
12 Περιγραφική Στατιστική για τον Αριθμό Σημείων Στίξεων 69
13 Περιγραφική Στατιστική για τον Αριθμό Hashtags 70
14 Περιγραφική Στατιστική για τον Αριθμό Emojis 71
15 Περιγραφική Στατιστική για τον Αριθμό Κεφαλαίων Λέξεων 72
16 Πίνακας των Unigrams στα Bag of Words 80
17 Πίνακας των Bigrams στα Bag of Words 81
18 Οι τιμές των λέξεων του tweet με την αναπαράσταση της TF-IDF 82
19 Ακρίβεια για τα μοντέλα του One Hot Encoding - Unigrams 87
20 Ακρίβεια για τα μοντέλα του One Hot Encoding - Bigrams 90
21 Ακρίβεια για τα μοντέλα του TF-IDF 93
22 Ακρίβεια για τα μοντέλα του Word2Vec 96
23 Ακρίβεια για τα μοντέλα του FastText 98

1 Εισαγωγή

1.1 Σκοπός της Διπλωματικής

Η κλιματική αλλαγή και η υπερθέρμανση του πλανήτη, που έχουν προκύψει από το φαινόμενο του θερμοκηπίου αποτελούν ένα από τα μείζονα θέματα συζήτησης τόσο στον επιστημονικό χώρο όσο και στον γενικό πληθυσμό. Το 2022 βάση των διεθνών στατιστικών αποτελεί το δεύτερο θερμότερο έτος που έχει καταγραφεί ποτέ στην Ευρώπη και πέμπτο θερμότερο έτος παγκοσμίως. Οι περισσότεροι επιστήμονες πιστεύουν ότι η υπερθέρμανση του πλανήτη είναι αποτέλεσμα της ανθρώπινης δραστηριότητας, ενώ λίγοι είναι εκείνοι που πιστεύουν ότι είναι μία κοροϊδία και αποτελεί μια φυσική εξέλιξη. Ο γενικός πληθυσμός έχει χωριστεί σε αυτούς που πιστεύουν στην κλιματική αλλαγή έναντι αυτών που παραμένουν πιστοί στη γνώμη τους, όσο χρόνια και αν περάσουν.

Από τη βιομηχανική επανάσταση έως και σήμερα οι ανθρωπογενείς εκπομπές έχουν προκαλέσει την άνοδο του διοξειδίου του άνθρακα (CO2) στις συγκεντρώσεις των θερμοκηπιακών αερίων. Το 2022 η συγκέντρωση του στην ατμόσφαιρα ήταν 48% υψηλότερη από την προβιομηχανική εποχή. Καθώς η παραγωγή του διοξειδίου του άνθρακα (CO2) και άλλων ρυπογόνων αερίων με την πάροδο του χρόνου αυξάνονται και απελευθερώνονται στην ατμόσφαιρα της γης, παγιδεύουν την ηλιακή ακτινοβολία στη γη και εμποδίζουν τη διάχυση της στο διάστημα με αποτέλεσμα την υπερθέρμανση του πλανήτη. Η πιο χαρακτηριστική ένδειξη της παγκόσμιας κλιματικής αλλαγής είναι η αύξηση της μέσης παγκόσμιας θερμοκρασίας. Η αύξηση της στάθμης της θάλασσας, η μεταβολή της χιονοκάλυψης, της κάλυψης από πάγο καθώς και η αύξηση των υδρατμών κάνουν την υπερθέρμανση του πλανήτη να γίνεται επίσης αντιληπτή. Τα κλιματικά συστήματα της γης και η γεωγραφία της, αλλάζουν με ανεπανόρθωτο τρόπο. Η περίοδος 2011-2020 ήταν η θερμότερη δεκαετία που έχει καταγραφεί ποτέ, καθώς η παγκόσμια μέση θερμοκρασία ξεπέρασε τα προβιομηχανικά επίπεδα κατά 1,1 °C το 2019.¨ [1]. Η θερμοκρασία της γης θα αυξηθεί κατά 4°C από την προβιομηχανική εποχή μέχρι το 2100 σύμφωνα με επιστημονικές μελέτες [2].

Η χρήση των μέσων κοινωνικής δικτύωσης αποτελεί το βασικό μέσω κοινωνικής διάδρασης και αναφαίρετο κομμάτι της παγκόσμιας κοινωνίας. Τα μέσα κοινωνικής δικτύωσης έχουν προσφέρει τη δυνατότητα στους χρήστες τους, από παθητικοί δέκτες πληροφορίας, να γίνουν ενεργητικοί δημιουργοί πληροφορίας καθώς μέσω αυτών μπορούν όχι μόνο να επικοινωνούν μεταξύ τους αλλά και να θέτουν ερωτήματα, να ανταλλάσουν απόψεις, να εκφράζουν τη γνώμη τους σε διάφορα θέματα επικαιρότητας ή να αξιολογούν προϊόντα και υπηρεσίες.

Το Twitter είναι ένα μέσο κοινωνικής δικτύωσης και μια ευρέως χρησιμοποιούμενη πλατφόρμα microblogging που δημιουργεί έναν τεράστιο όγκο αδόμητων πληροφοριών καθώς διαθέτει 237,8 εκατομμύρια ενεργούς χρήστες μηνιαία, 500 εκατομμύρια δημοσιευμένα tweets την ημέρα, εκ των οποίων το καθένα από αυτά τα μπορεί να περιέχει μέχρι και 140 χαρακτήρες [3].

Από το 2006 που δημιουργήθηκε το Twitter αποτελεί την καλύτερη πλατφόρμα συλλογής απόψεων για οτιδήποτε συμβαίνει στον κόσμο. Πιο συγκεκριμένα, σε αυτήν την πλατφόρμα γίνεται καλύτερα αντιληπτή, η διχοτόμηση των γνώμεων στο θέμα της κλιματικής αλλαγής. Δεδομένου ότι οι χρήστες της πλατφόρμας μπορούν ελεύθερα να σχολιάσουν το tweet κάποιου άλλου χρήστη και να διαφωνήσουν στα λεγόμενα του, γεγονός που διευκολύνει την αναγνώριση των διφορούμενων απόψεων.

Οι εξελιγμένες μέθοδοι εξόρυξης δεδομένων (data mining) και ανάλυσης συναισθήματος (sentiment analysis) μπορούν, πλέον, να κατηγοριοποιήσουν τα δεδομένα κειμένου (όπως tweet) και να δημιουργήσουν έτσι μια εικόνα για τις απόψεις των χρηστών που αποτελεί ενα μεγάλο μέρος της κοινωνίας ολόκληρης. Στην συγκεκριμένη περίπτωση, της κλιματικής αλλαγής, στην έρευνα που θα διεξαχθεί, θα διαπιστωθεί ποιοι χρήστες πιστεύουν ότι η κλιματική αλλαγή είναι ανθρωπογενές φαινόμενο (υπερασπιστές) και ποιοι πιστεύουν ότι είναι απλά μέρος του κύκλου ζωής του πλανήτη (αρνητές).

Η συγκεκριμένη εργασία έχει ως σκοπό να αναγνωρίσει, αποδοτικά, πότε κάποιος χρήστης είναι της μίας πλευράς, αρνητής, και πότε της άλλης, υποστηρικτής, χωρίς φυσικά να λάβει υπόψη τους χρήστες που δεν φέρουν γνώμη ή η γνώμη τους δεν είναι κατασταλαγμένη σε μια από τις δυο παρατάξεις. Για αυτό τον λόγο τα δεδομένα που θα εξορυχθούν, από το Twitter που αφορούν την κλιματική αλλαγή και την υπερθέρμανση του πλανήτη, προσκομίζονται ακριβώς έτσι όπως αποτυπώνονται από τους δογματικούς χρήστες του. Στη συνέχεια θα χρησιμοποιηθούν μέθοδοι αναπαράστασης δεδομένων κειμένου (word embeddings) και θα εισαχθούν σε διάφορες μεθόδους μηχανικής μάθησης (machine learning) για να δημιουργηθεί ο αλγόριθμος. Τέλος συνδυάζοντας τις διάφορες μεθόδους (παλιές και σύγχρονες) θα συγκριθούν τα αποτελέσματα με τις ήδη διεξαγόμενες μεθόδους.

1.2 Οργάνωση της Διπλωματικής

Στο Κεφάλαιο 2 θα αναπτυχθούν οι τεχνικές και οι μεθόδοι που συνάδελφοι και λοιποί ερευνητές έχουν αναπτύξει. Στο Κεφάλαιο 3, θα γίνει μια περιγραφή των τεχνικών και των μεθόδων που θα παρουσιαστούν στη συγκεκριμένη εργασία. Στη συνέχεια, στο Κεφάλαιο 4, θα αναλυθεί λεπτομερώς ο τρόπος με τον οποίο η εργασία δομήθηκε και αναπτύχθηκε και κατόπιν στο Κεφάλαιο 5 θα παρουσιαστούν τα σημαντικότερα αποτελέσματα αυτής της εργασία. Τέλος στο Κεφάλαιο 6 θα συγκριθούν τα αποτελέσματα της εργασίας και η καινοτομία της, με τις ήδη υπάρχουσες έρευνες καθώς και οι ενέργειες που μπορούν να ακολουθηθούν για την βελτίωση αυτής της εργασίας.

2 Ανασκόπηση Βιβλιογραφίας

Σε αυτήν την ενότητα θα αναπτυχθούν οι υπάρχουσες γνώσεις περί των μέσων κοινωνικής δικτύωσης, της εξόρυξης δεδομένων, της ανάλυσης συναισθήματος και τέλος της τεχνιτής νοημοσύνης. Όλα τα παραπάνω, θα βοηθήσουν στην ανάπτυξη του θέματος καθώς και στην σύγκριση των αποτελεσμάτων με τυχόν ήδη δημοσιευμένες τεχνικές.

2.1 Μέσα Κοινωνικής Δικτύωσης (Social Media)

Η σημερινή κοινωνία χαρακτηρίζεται από ένα σύνολο συνεχών εξελίξεων και σπουδαίων τεχνολογικών καινοτομιών, οι οποίες έχουν ως σκοπό να κάνουν πιο εύκολη την καθημερινότητα των ανθρώπων. Τα Μέσα Κοινωνικής Δικτύωσης (ΜΚΔ), ή παγκοσμίως γνωστά ως social media, αποτελούν ένα τέτοιο σπουδαίο τεχνολογικό επίτευγμα, καθώς επίσης κι ένα αναπόσπαστο κομμάτι της καθημερινότητας αυτής. Τα social media, ως ένα σύνολο εικονικού χώρου, όπου άτομα όλων των ηλικιών κατέχουν την δυνατότητα να αλληλεπιδράσουν με άλλους ανθρώπους, να ανταλλάξουν πληροφορίες κι ιδέες και να δημιουργήσουν μεταξύ τους κοινότητες, τείνουν να καταρρίψουν τα φυσικά όρια του χώρου και του χρόνου.

Το Facebook, το Twitter, το Instagram, το YouTube και το LinkedIn είναι μερικοί από τους πιο ευρέως χρησιμοποιούμενους ιστότοπους κοινωνικής δικτύωσης. Ο μεγαλύτερος ιστότοπος κοινωνικής δικτύωσης είναι το Facebook, το οποίο δημιουργήθηκε το 2004 και έχει περισσότερους από 2,8 δισεκατομμύρια ενεργούς χρήστες ανά μήνα. Αποτελεί μια πλατφόρμα σύνδεσης για φίλους και οικογένεια, επιπλέον δίνει τη δυνατότητα συμμετοχής σε ομάδες ενδιαφέροντος καθώς και δυνατότητα παρακολούθησης εταιρειών και δημοσίων προσώπων. Το Twitter με περιορισμό 280 χαρακτήρων, αποτελεί μια πλατφόρμα microblogging που ιδρύθηκε το 2006 και επιτρέπει στους χρήστες να κάνουν σύντομες παρατηρήσεις και ενημερώσεις ειδήσεων σε πραγματικό χρόνο. Το Instagram, που πλέον διαθέτει πάνω από 1 δισεκατομμύριο ενεργούς χρήστες μηνιαίως, δημιουργήθηκε το 2010. Είναι ένας ιστότοπος κοινής χρήσης φωτογραφιών και βίντεο που δίνει μεγάλη έμφαση στην οπτική αφήγηση. Τα τελευταία χρόνια έχει εξελιχθεί

σε ένα ισχυρό εργαλείο για τις επιχειρήσεις που αλληλεπιδρούν με το κοινό τους. Η μεγαλύτερη διαδικτυακή πλατφόρμα βίντεο στον κόσμο είναι το YouTube και δημιουργήθηκε το 2005. Η πλατφόρμα αυτή επιτρέπει στους χρήστες της να δημοσιεύουν, να παρακολουθούν και να μοιράζονται βίντεο, με περισσότερους από 740 εκατομμύρια χρήστες σε περισσότερες από 200 χώρες, το LinkedIn, που ξεκίνησε το 2002, είναι ένα επαγγελματικό μέσο δικτύωσης που απευθύνεται σε επαγγελματίες και άτομα που αναζητούν εργασία.

Υπάρχουν πολλά πλεονεκτήματα στην ανάπτυξη των μέσων κοινωνικής δικτύωσης, όπως η καλύτερη συνδεσιμότητα και η γρήγορη ανταλλαγή πληροφοριών. Έχει αποδειχθεί η αύξηση των δυνατοτήτων, του κοινωνικού κεφαλαίου και των πόρων που είναι διαθέσιμοι στους ανθρώπους μέσω των κοινωνικών δικτύων τους [4]. Έχει επίσης αποδειχθεί ότι είναι ένα μέσο για κοινωνική υπεράσπιση και ομαδική δράση [5]. Τα μέσα κοινωνικής δικτύωσης έχουν εξελιχθεί σε ένα σημαντικό εργαλείο για τις εταιρείες, στην προσπάθειά τους να αλληλεπιδρούν με την αγορά-στόχο τους, δίνοντάς τους την ευκαιρία να έρχονται σε άμεση επαφή με τους πελάτες και να αυξήσουν την αναγνωρισιμότητα της επωνυμίας τους [6].

Ένα από τα μεγαλύτερα πλεονεκτήματα των μέσων κοινωνικής δικτύωσης είναι το γεγονός ότι άλλαξαν τον τρόπο που διαδίδονται τα νέα. Στον σημερινό κόσμο, θα έλεγε κανείς ότι χρειάζεται μόνο ένα tweet για ένα περιστατικό και η είδηση αυτή θα φτάσει στις ηλεκτρονικές συσκευές όλων όσων είναι συνδεδεμένοι στο διαδίκτυο. Ένα άτομο που είναι αυτόπτης μάρτυρας ενός περιστατικού έχει τη δυνατότητα να μοιραστεί άμεσα πληροφορίες ή εικόνες για το γεγονός αυτό, όπως για παράδειγμα ο μαζικός πυροβολισμός στην Aurora του Κολοράντο το 2012 ή την βομβιστική επίθεση στον μαραθώνιο της Βοστώνης το 2013 [7]. Αυτές οι ειδήσεις έγιναν viral (διάσημες, τις είδαν πολλοί άνθρωποι) πριν προβληθούν από τα τηλεοπτικά κανάλια, στις ειδήσεις. Το γεγονός αυτό, αποδεικνύει πόσο σημαντικά και πολύτιμα μπορούν να είναι τα μέσα κοινωνικής δικτύωσης για μια κοινωνία. Τέλος, από την οπτική ενός data scientist, τα μέσα κοινωνικής δικτύωσης αποτελούν ένα ¨χρυσωρυχείο¨, μια πολύτιμη πηγή, ανάκτησης

πληροφοριών για διάφορους σκοπούς, από τις επερχόμενες εκλογές έως και την κλιματική αλλαγή.

Παρόλα αυτά υπάρχουν και σημαντικά μειονεκτήματα από την εξέλιξη των social media. Ένα από τα κύρια ζητήματα είναι η διεξαγωγή και η διοχέτευση ψευδών πληροφοριών και ειδήσεων. Πολλές μελέτες για παράδειγμα, έχουν δείξει ότι οι ψευδείς ειδήσεις μεταδίδονται πιο γρήγορα και σε μεγαλύτερο φάσμα από ότι οι αληθείς [8]. Επίσης ολοένα και περισσότεροι επιστήμονες τείνουν να υποστηρίζουν ότι τα μέσα κοινωνικής δικτύωσης έχουν συσχετιστεί με συμπτώματα άγχους και κατάθλιψης, ειδικά μεταξύ των νεαρών σε ηλικία χρηστών [9]. Επιδεινώνονται επίσης οι ανησυχίες σχετικά με το απόρρητο και τη χρήση των προσωπικών δεδομένων για τις διαφημίσεις. Για παράδειγμα, έχει αποδειχθεί ότι οι επιχειρήσεις κοινωνικών μέσων στοχεύουν σε χρήστες με διαφημίσεις τους, χρησιμοποιώντας αλγόριθμους οι οποίοι μπορούν να οδηγήσουν σε χειραγώγηση των αντιλήψεων και των ενεργειών των χρηστών αυτών [10].

Εν κατακλείδι, η ανάπτυξη των μέσων κοινωνικής δικτύωσης έχει φέρει μια νέα εποχή στον τρόπο επικοινωνίας και ανταλλαγής πληροφοριών, δεν πρέπει όμως να παραγκωνίζονται και τα σημαντικά μειονεκτήματα που έχει επιφέρει. Η ορθή και υπεύθυνη χρήση τους σε συνδυασμό με την ενημέρωση των χρηστών για τις επιπτώσεις της αυξανόμενης χρήσης αποτελούν τον ακρογωνιαίο λίθο για μια ισορροπημένη διαδικτυακή και ψηφιακή διάδραση.

2.1.1 Twitter

Το Twitter, όπως αναφέρθηκε προηγουμένως, είναι ένα μέσω κοινωνικής δικτύωσης όπου οι χρήστες μπορούν να στέλνουν σύντομα μηνύματα, που ονομάζονται tweets μέγιστου αριθμού 280 χαρακτήρων και διαθέτει περισσότερους από 330 εκατομμύρια ενεργούς χρήστες μηνιαίως. Δημιουργήθηκε το 2006 και έκτοτε έχει εξελιχθεί σε μία από τις πιο δημοφιλείς πλατφόρμες κοινωνικής δικτύωσης παγκοσμίως. Θεωρείται επίσης η πιο μελετημένη πλατφόρμα κοινωνικών μέσων [11]. Οι χρήστες του Twitter μπορούν να ανταλλάσσουν γρήγορα ειδήσεις, σκέψεις και ενημερώσεις με τους υπόλοι-

πους χρήστες, δίνοντας στην πλατφόρμα τη φήμη ότι είναι ο πιο άμεσος ανταποκριτής σε πραγματικό χρόνο [12].

Έχουν διεξαχθεί εκτενείς μελέτες από πολλούς ακαδημαϊκούς και ερευνητές για διάφορα ζητήματα που αφορούν το Twitter, από την κρίση της διαπροσωπικής επικοινωνίας, την επίδρασή του στα παραδοσιακά μέσα επικοινωνίας έως την επιρροή που μπορεί να ασκεί στην πολιτική επικοινωνία. Ενώ το Twitter έχει λάβει επαίνους για την παροχή άμμεσης επικοινωνίας μεταξύ των πολιτικών εκπροσώπων και των πολιτών, αλλά και για την τόνωση της πολιτικής δέσμευσης, έχει επίσης δεχθεί και κριτική για την ενίσχυση αμφιλεγόμενων απόψεων, τη διάδοση ψεύτικων πληροφοριών και τη δυνατότητα χειραγώγησης της κοινής γνώμης [13].

Η κατηγοριοποίηση hashtag που ξεκίνησε από το Twitter και επεκτάθηκε σε όλα τα social media, επιτρέπει στους χρήστες να κατατάσσουν τα tweets τους με βάση το περιεχόμενό τους, να θέτουν δηλαδή μια θεματολογία ή αλλιώς μια ετικέτα. Έτσι για παράδειγμα κάθε tweet με hashtag #climatechange θα απευθύνεται στην κλιματική αλλαγή. Με αυτόν τον τρόπο πολλές κοινωνικές παρατάξεις χρησιμοποιούν το Twitter ως ¨φωνή λαού¨ για κοινωνικοπολιτική αφύπνιση και κινητοποίηση [12].

Οι επιχειρήσεις και οι οργανισμοί χρησιμοποιούν ολοένα και περισσότερο το Twitter ως έναν τρόπο αλληλεπίδρασης με τους πελάτες, με στόχο την προώθησής τους. Το Twitter προσφέρει μια ανεπανάληπτη ευκαιρία στις επιχειρήσεις να προσεγγίσουν μεγαλύτερο κοινό λόγω της παγκόσμιας εμβέλειας και της ικανότητάς του να συνδέει άτομα με παρόμοια ενδιαφέροντα. Για παράδειγμα, το Pew Research Center κατέληξε στο συμπέρασμα ότι το 22% των Αμερικανών χρησιμοπεί το Twitter και ότι σχεδόν τα δύο τρίτα των ενηλίκων στην Αμερική χρησιμοποιούν ενεργά τα μέσα κοινωνικής δικτύωσης. Οι επιχειρήσεις έχουν τώρα μια τεράστια ευκαιρία να συνδέονται με δυνητικούς πελάτες και να τους προωθούν διαδραστικά μηνύματα.

Βέβαια το Twitter παρουσιάζει ορισμένες δυσκολίες και αντιφάσεις. Ο ιστότοπος έχει δεχθεί έναν καταιγισμό αρνητικών κριτικών που αναφέρονται σε πολλαπλασιασμό

ψευδών ειδήσεων, διαδικτυακό εκφοβισμό και παραβίαση δικαιωμάτων απορρήτου. Ως αποτέλεσμα οι διαχειριστές της πλατφόρμας έχουν θέσει μια σειρά από περιορισμούς για να μπορέσουν να οριοθετήσουν τα προβλήματα αυτά. Παραδείγματα τέτοιου είδους περιορισμών είναι η προσθήκη ετικετών ελέγχου δεδομένων και η δυνατότητα επισήμανσης tweets που δεν συνάδουν με τους κανόνες της πλατφόρμας [14].

Ωστόσο,ο λόγος για τον οποίον επικεντρώνεται η εργασία στη συγκεκριμένη πλατφόρμα είναι η χρήση της για ένα ευρύ φάσμα επιχειρημάτων, τις περισσότερες φορές τα επιχειρήματα αυτά είναι διφορούμενα. Συγκεκριμένα, θα χρησιμοποιηθεί το Twitter ως πηγή πληροφοριών για τους ακόλουθους λόγους [15]:

- Οι πλατφόρμες microblogging χρησιμοποιούνται από διαφορετικούς ανθρώπους για να εκφράσουν τη γνώμη τους για διαφορετικά θέματα, επομένως είναι μια πολύτιμη πηγή απόψεων ενός μεγάλου μέρους του πληθυσμού.
- Το Twitter περιέχει έναν τεράστιο αριθμό αναρτήσεων κειμένου που αυξάνεται καθημερινά. Το ποσό των κειμένων που συλλέγονται μπορεί να είναι αναμφίβολα μεγάλο.
- Οι χρήστες του Twitter ποικίλλουν από διασημότητες, εκπροσώπους εταιρειών, πολιτικούς, έως και προέδρους χωρών. Επομένως, είναι δυνατή η συλλογή αναρτήσεων κειμένου χρηστών από διαφορετικές κοινωνικές ομάδες και ομάδες συμφερόντων.
- Οι χρήστες του Twitter μπορούν να είναι από οποιαδήπτοτε χώρα, άρα είναι ένα πολυπολιτισμικό μέσο.

2.1.2 Κλιματική Αλλαγή στο Twitter

Ένας άλλος λόγος για τον οποίο επιλέγεται το Twitter για την έρευνα αυτή είναι επειδή περιέχει πολλές πληροφορίες σχετικά με το θέμα της κλιματικής αλλαγής. Όπως αναφέρθηκε παραπάνω, το Twitter μπορεί να θεωρηθεί ως ειδησεογραφικό μέσο, προβάλλοντας νέες πληροφορίες σχετικά με ένα θέμα κάθε στιγμή. Αυτό το γεγονός παρατηρήθηκε ιδιαίτερα για την κλιματική αλλαγή.

Σύμφωνα με ορισμένες ακαδημαϊκές έρευνες, το Twitter μπορεί να θεωρηθεί ως ένας ενεργός χώρος αμφισβήτησης για την κλιματική αλλαγή [16]. Μια τέτοια μελέτη, η "Climate Change Discourse on Twitter: A Study of Public Involvement and Polarization", αποκάλυψε ότι, άτομα άκρως ενεργά στην πλατφόρμα και ομάδες με ακραίες απόψεις ελέγχουν συχνά τη συζήτηση για την κλιματική αλλαγή στο Twitter. Σύμφωνα με μια άλλη μελέτη, "Twitter and Climate Change: A Study of Public Perceptions and Attitudes", ο ιστότοπος αυτός μπορεί να χρησιμοποιηθεί ως μέσο για τη συλλογή των tweets, των μελών, πάνω σε ειδήσεις και πολιτικές επιλογές που σχετίζονται με την κλιματική αλλαγή, καθώς και για το πώς οι άνθρωποι βλέπουν και αν κατανοούν το πρόβλημα.

Το Twitter έχει επίσης χρησιμοποιηθεί σε διεθνείς συζητήσεις για το κλίμα, όπως για παράδειγμα η Συμφωνία του Παρισιού, που χρησιμοποιήθηκε ως διπλωματικό όπλο στον αγώνα κατά της κλιματικής αλλαγής. Επίσης χρησιμοποιείται από τη Σύμβαση-Πλαίσιο των Ηνωμένων Εθνών για την Κλιματική Αλλαγή (UNFCCC) ως εργαλείο για να αλληλεπιδρά με το κοινό και για να συλλέγει πληροφορίες σχετικά με το έργο της για την κλιματική αλλαγή.

Το συγκεκριμένο μέσο φαίνεται να είχε τεράστιο αντίκτυπο στις συζητήσεις γύρω από την κλιματική αλλαγή, προσφέροντας μια ειδική πλατφόρμα στους ανθρώπους, τις ομάδες και τις κυβερνήσεις όπου μπορούν να μοιραστούν τη γνώση, να αλληλεπιδράσουν και να εμπνεύσουν το κοινό για δράση. Ακόμα κι αν η πλατφόρμα έχει αντιμετωπίσει δυσκολίες και διαμάχες, εξακολουθεί να είναι ένα χρήσιμο εργαλείο για την προώθηση της ευαισθητοποίησης και του δημόσιου διαλόγου σε ένα από τα πιο κρίσιμα ζητήματα της εποχής, όπως η κλιματική αλλαγή.

Θα μπορούσε όμως να αποτελέσει ένα αμερόληπτο μέσο διαδυκτιακής αλληλεπίδρασης ; Ορισμένοι χρήστες έχουν χρησιμοποιήσει το Twitter για να διαδώσουν θεωρίες συνωμοσίας και εσφαλμένες πληροφορίες σχετικά με την κατάσταση, προκαλώντας κριτική για τον τρόπο με τον οποίο χρησιμοποιείται το μέσο. Ως αποτέλεσμα, τέθηκαν ορισμένα ερωτήματα σχετικά με το αν το Twitter έχει την ικανότητα να προωθεί

συζητήσεις με επικαιροποιημένες πληροφορίες που βοηθούν στη λήψη αποφάσεων.

Συμπερασματικά, το Twitter είναι ένα χρήσιμο εργαλείο για να αντιληφθεί κανείς πώς αισθάνεται το ευρύ κοινό για την κλιματική αλλαγή. Οι χρήστες το χρησιμοποιούν για να εκφράσουν τη γνώμη τους σχετικά με το αν η κλιματική αλλαγή είναι πραγματική ή όχι. Στην ανάλυση που ακολουθεί, γίνεται λόγος για το αντίκτυπο αυτών των tweets και την αναγνώριση του συναισθήματος του χρήστη.

2.2 Εξόρυξη Δεδομένων - Μεγάλα Δεδομένα

Στην σημερινή εποχή, καθώς οι πληροφορίες ρέουν όλο και πιο εύκολα και σε μεγάλες ποσότητες, δημιουργείται η ευκαιρία συλλογής τους. Η δυσκολία έγκειται στον τρόπο με τον οποίο μπορεί κανείς χειριστεί μια τόσο μεγάλη εισροή πληροφοριών. Πριν από μια δεκαετία, οι επιστήμονες αντλούσαν από το διαδίκτυο ένα διαχειρίσιμο αριθμό πληροφοριών, μικρής χωρητικότητας, και τα δεδομένα που υποβάλλονταν σε επεξεργασία ήταν σχεδόν ομοιογενή. Αυτό ονομαζόταν τότε εξόρυξη δεδομένων (Data mining). Ωστόσο, τα τελευταία χρόνια έχει δημιουργηθεί μια αναδυόμενη νέα πρόκληση που ονομάζεται μεγάλα δεδομένα (Big Data). Αυτό ενσωματώνει τη μαζική παραγωγή και ανταλλαγή πληροφοριών στο διαδίκτυο και τις προκλήσεις για τη διαχείριση και την απόκτηση γνώσεων από αυτές τις πληροφορίες [17].

Η εξόρυξη δεδομένων και τα μεγάλα δεδομένα είναι δύο θέματα σχετικά, αλλά ταυτόχρονα ξεχωριστά, που γίνονται τα τελευταία χρόνια συνεχώς πιο σημαντικά, καθώς η τεχνολογία των πληροφοριών εξελίσσεται και τα δεδομένα είναι διαθέσιμα πιο εύκολα. Ως εξόρυξη δεδομένων θεωρείται η τεχνική εξαγωγής προτύπων και γνώσης από τεράστιους όγκους δεδομένων. Για την αναζήτηση συσχετίσεων, μοτίβων και σχέσεων στα δεδομένα και για να εξαχθούν ενδιαφέρουσες πληροφορίες, χρησιμοποιούνται αλγόριθμοι, στατιστικά μοντέλα και προσεγγίσεις μηχανικής μάθησης. Από την άλλη πλευρά οι ογκώδεις ποσότητες οργανωμένων και μη δομημένων δεδομένων που δημιουργούνται από ποικίλες πηγές, συνυπολογίζοντας τα μέσα κοινωνικής δικτύωσης,

τις φορητές συσκευές και το InternetofThings, αναφέρονται ως ¨μεγάλα δεδομένα¨ [18].

Στον αντίποδα η πληθώρα των δεδομένων έχει δημιουργήσει πλέον νέα προβλήματα όπως η αποθήκευση, η επεξεργασία και η ανάλυση τους. Για την ορθή εξόρυξη δεδομένων και την ανάλυση μεγάλων δεδομένων είναι αναγακαία η χρήση προηγμένων υπολογιστικών και στατιστικών ικανότητων, καθώς και εξειδικευμένο λογισμικό και υλικό.

Η ανάλυση μεγάλων δεδομένων μπορεί επίσης να προσφέρει πολλές και χρήσιμες πληροφορίες για μια ποικιλία φαινομένων, όπως αυτή της συμπεριφοράς των καταναλωτών, τις οικονομικές τάσεις και τα κοινωνικά δίκτυα [19]. Για παράδειγμα, κάποιοι οργανισμοί που ενδιαφέρονται για τη συμπεριφορά των καταναλωτών μπορούν να μελετήσουν καλύτερα και να αξιολογήσουν τα δεδομένα συναλλαγών από διάφορες πηγές. Οι ακαδημαϊκοί μπορούν να μελετήσουν τη στάση και τη γνώμη του κοινού χρησιμοποιώντας τεράστιες ποσότητες δεδομένων μέσων κοινωνικής δικτύωσης, όπως είναι το Twitter, που χρησιμοποιείται και στην παρούσα εργασία.

Τα πλεονεκτήματα της εξόρυξης δεδομένων και των μεγάλων δεδομένων είναι αρκετά, όμως δημιουργούν κάποιες ανησυχίες σχετικά με τις επιπτώσεις στο απόρρητο, την ασφάλεια και την ηθική [20, 21]. Τα δεδομένα πρέπει να λαμβάνονται και να χρησιμοποιούνται σύμφωνα με τους κανόνες απορρήτου που ισχύουν για το διαδίκτυο, σε συνδυασμό πάντα με τη λήψη σωστών αποφάσεων και όχι για να εκθέσουν αυτές τις πληροφορίες. Τα μεγάλα δεδομένα ενδέχεται να αυξήσουν τις απειλές ασφαλείας αυξάνοντας την πιθανότητα παραβιάσεων δεδομένων.

Στην επόμενη ενότητα θα παρουσιαστεί η έννοια και οι προκλήσεις που υπάρχουν πίσω από τον όρο Big Data. Θα παρατεθούν επίσης τα οφέλη της εξαγωγής πληροφοριών από το Διαδίκτυο και τις ηλεκτρονικές υπηρεσίες. Αυτό θα ρίξει λίγο φως στον όρο Εξόρυξη Δεδομένων (Data Mining) και θα αναδείξει πόσο σημαντικός είναι ο όρος αυτός στην τρέχουσα έρευνα. Τέλος, θα γίνει σύνδεση όλων των παραπάνω, για να αποδειχθεί ο λόγος για τον οποίο είναι απαραίτητες οι τεχνικές εξόρυξης δεδομένων για το έργο αυτό

και θα οριστούν οι τεχνικές που χρησιμοποιούνται για την εξόρυξη δεδομένων Twitter.

2.2.1 Μεγάλα Δεδομένα (Big Data)

Στη σημερινή κοινωνία, οι έξυπνες συσκευές όπως τα smartphones, τα smartwatches κ.λπ. αποθηκεύουν σχεδόν ολόκληρη τη δραστηριότητά τους με τη μορφή δεδομένων. Για παράδειγμα, το smartphone ενός ατόμου μπορεί να μεταδίδει πληροφορίες σε διάφορους παρόχους σχετικά με τις κλήσεις, τα μηνύματα, τη μουσική, τις φωτογραφίες, τα βίντεο και την τοποθεσία του για κάθε λεπτό που είναι ενεργό [22]. Επιπλέον, υπάρχουν και οι υπηρεσίες Διαδικτύου που αυξάνουν κατά πολύ αυτόν τον όγκο δεδομένων [23]. Για παράδειγμα, αναζητήσεις στο Διαδίκτυο, μηνύματα και αναρτήσεις στο facebook, ακόλουθοι, likes, tweets και retweets στο Twitter, likes, μηνύματα, ακόλουθοι στο Instagram κ.λπ. Με βάση τα παραπάνω, ο όγκος δεδομένων που παράγει ένα άτομο με μία μόνο ¨έξυπνη¨ συσκευή σε ένα λεπτό είναι τεράστιος. Σε μια προσπάθεια υπολογισμού του όγκου των δεδομένων για ολόκληρο τον πληθυσμό και τις διάφορες συσκευές που διαθέτει ο καθένας, δύσκολα θα μπορούσε να αντιληφθεί κανείς πόσο μεγάλα είναι τα δεδομένα που δημιουργούνται [24].

Θα μπορούσε να πει κανείς, ότι αυτά τα δεδομένα είναι σαν ¨ορυχείο¨ για τις κυβερνήσεις και τις διαδικτυακές υπηρεσίες, ο τρόπος εξαγωγής, διαχείρισης και μετατροπής τους σε πολύτιμες πληροφορίες είναι μια αναδυόμενη ανάγκη. Δυστυχώς, αυτά τα δεδομένα αποτελούν πραγματική πρόκληση για τις παραπάνω οντότητες, καθώς η φύση τους δεν είναι συμβατή με τις τεχνικές που έχουν παραχθεί μέχρι τώρα. Συγκεκριμένα, έχουν υπάρξει δύο σημαντικές ερμηνείες για το πώς είναι και τι είναι τα Big Data. Το θεώρημα HACE. Τα Μεγάλα Δεδομένα ειναι μεγάλου όγκου δεδομένα από ετερογενείς (Heterogeneous) και αυτόνομες (Autonomous) πηγές με κατανεμημένο και αποκεντρωμένο έλεγχο και επιδιώκουν να εξερευνήσουν σύνθετες (Complex) και εξελισσόμενες (Evolving) σχέσεις μεταξύ των δεδομένων. Αυτά τα χαρακτηριστικά το καθιστούν μια εξαιρετική πρόκληση για την ανακάλυψη χρήσιμης γνώσης [25].

Για να αποδοθεί ο όρος Big Data, θα χρησιμοποιηθεί, επίσης, ο ορισμός των 3Vs, Volume, Velocity, Variety [26]. Ωστόσο, καθώς ο κόσμος εξελίσσεται, το ίδιο συμβαίνει και με τη φύση των Big Data. Σήμερα τα Big Data μπορούν να περιγραφούν από τα 5Vs [27]. Συγκεκριμένα :

- **Volume (Ογκος)**

  Ο όγκος των δεδομένων αποτελεί μία από τις μεγαλύτερες προκλήσεις όσον αφορά τα Big Data. Οι εταιρείες, η κυβέρνηση και άλλοι πάροχοι υπηρεσιών λαμβάνουν πλήθος πληροφοριών σε μορφή ακατέργαστων δεδομένων. Ωστόσο, η μη αποδοτική αποθήκευση ή χειρισμός αυτών των δεδομένων μπορεί να οδηγήσει σε απώλεια πληροφοριών, ενώ επίσης κατακλύζει τη χωρητικότητα αποθήκευσης των παρόχων.

- **Velocity (Ταχύτητα)**

  Ο όρος ταχύτητα αναφέρεται στην αυξανόμενη ταχύτητα με την οποία δημιουργούνται αυτά τα δεδομένα. Αυτό συνεπάγεται και την δημιουργία, την αποθήκευση και την ανανέωση των βάσεων δεδομένων. Επιπλέον, εξαιτίας της ταχύτητας παραγωγής δεδομένων, η ζήτηση για ταχύτερες τεχνικές απόδοσης πληροφοριών αυξάνεται. Συγκεκριμένα τον Οκτώβριο του 2012 η πρώτη προεδρική αντιπαράθεση μεταξύ του προέδρου Μπαράκ Ομπάμα και του κυβερνήτη Μιτ Ρόμνεϊ πυροδότησε περισσότερα από 10 εκατομμύρια tweets μέσα σε 2 ώρες [28].

- **Variety (Ποικιλία)**

  Το επόμενο χαρακτηριστικό των μεγάλων δεδομένων είναι η ποικιλία τους. Τα δεδομένα αυτά, επειδή στην ουσία είναι η ταχεία εισροή αφιλτράριστων πληροφοριών από το ίντερνετ, μπορούν να πάρουν διάφορες μορφές δόμησης. Απο δομημένα δεδομένα (structured data), όπως για παράδειγμα ένα αρχείο excel , σε ημιδομημένα δεδομένα (semistructured data), όπως για παράδειγμα ένα email, και τέλος σε αδόμητα δεδομένα (unstructured data) όπως για παράδειγμα μια ακτινογραφία. Η ενασχόληση με τα δομημένα και μη δομημένα δεδομένα αυξάνει σημαντικά την πολυπλοκότητα τόσο της αύξησης του μεγέθους όσο και της ανάλυσης των Big Data. Το 90% των δεδομένων που παράγονται είναι δεδομένα

σε μη δομημένη μορφή.

- **Veracity (Φιλαλήθεια)**

  Για να χρησιμοποιηθούν τα δεδομένα τα οποία συλλέγονται θα πρέπει να πιστοποιηθεί ότι είτε προέρχεται από αξιόπιστες πηγές, ή ότι οι χρήστες οι οποίοι παρήγαγαν αυτά τα δεδομένα είναι άτομα της κοινωνίας και όχι μηχανές που δημιουργήθηκαν για να προκαλέσουν ρωγμές στην εξακρίβωση της κοινής γνώμης.

- **Value (Αξία)**

  Η αξία αποτελεί την πιο σημαντική πτυχή στα μεγάλα δεδομένα. Ο λόγος είναι ότι, από αυτήν την μεγάλη εισροή δεδομένων μόνο μερικά από αυτά μπορούν να φανούν χρήσιμα σε περαιτέρω τεχνολογίες και μπορούν να επιφέρουν πρόοδο σε τεχνικές. Η διαπίστωση της αξίας των δεδομένων και η ενδεχόμενη ανάπτυξη, που μπορεί να επιφέρουν, είναι ένας από τους σημαντικότερους σκοπούς των μεγάλων δεδομένων.

Όπως είναι φανερό τα μεγάλα δεδομένα είναι δύσκολο να επεξεργαστούν με παλιές καθιερωμένες τεχνικές διαχείρισης δεδομένων, λόγω της υψηλής ταχύτητας με την οποία δημιουργούνται, της ποικιλίας και του όγκου τους. Ο τρόπος συγκέντρωσης, αποθήκευσης και αξιολόγησης των μεγάλων δεδομένων από τις εταιρείες, οργανισμούς, τις κυβερνήσεις έχει αλλάξει ολοκληρωτικά. Η τεράστια αύξηση δεδομένων έχει επιφέρει τη δημιουργία υπερεξελιγμένης τεχνολογίας, τη δημιουργία βάσεων δεδομένων, όπως οι: Hadoop, Spark και NoSQL. Τα ογκώδη πλέον δεδομένα μπορούν τώρα να αποθηκευτούν και να επεξεργαστούν αποτελεσματικά και οικονομικά μέσω αυτών των βάσεων. Κάποιες τάσεις, μοτίβα, λίστες, πίνακες που πριν ήταν αδύνατον να εξερευνηθούν ή να επεξεργαστούν είναι πλέον εφικτές με τη χρήση αυτών των βάσεων δεδομένων.

2.2.2 Εξόρυξη Δεδομενων (Data Mining)

Η τεχνική εξαγωγής προτύπων και γνώσης από τεράστιους όγκους δεδομένων είναι γνωστή ως εξόρυξη δεδομένων. Για την εξαγωγή πληροφοριών από δεδομένα και τη μετατροπή τους σε μια κατανοητή δομή για περαιτέρω χρήση, συνεπάγεται η χρήση αλγορίθμων και στατιστικών μοντέλων [18, 29]. Η εξόρυξη δεδομένων είναι η λήψη πληροφοριών από τα δεδομένα και η οργάνωση τους με τρόπο που να έχει νόημα για μελλοντική χρήση [30].

Όπως αναφέρθηκε παραπάνω, υπάρχει άμεση ανάγκη για τεχνικές χειρισμού μεγάλων δεδομένων και παροχής πολύτιμων πληροφοριών. Αυτόν τον ρόλο έρχεται να διαδραματίσει η Εξόρυξη Δεδομένων [19]. Όπως ένας άνθρωπος κάνει εξορύξεις για να βρει χρυσό και όταν το βρίσκει γίνεται πλούσιος, το ίδιο συμβαίνει και με τα δεδομένα. Στην αναζήτηση πληροφοριών, οι δεξιότητες που χρησιμοποιούνται είναι απαραίτητες για την εύρεση των πληροφορίων που απαιτούνται σε πολύ σύντομο χρονικό διάστημα [31]. Επομένως εξελιγμένες μέθοδοι αναζήτησης και εύρεσης πληροφοριών, σημαίνει πρόοδος στο προϊόν ή την υπηρεσία που διατίθεται. Έτσι αναπτύσσονται οι εταιρείες και οι διαδικτυακές υπηρεσίες αυτές τις μέρες, με την Εξόρυξη Δεδομένων.

Συνολικά, η εξόρυξη δεδομένων είναι ένας τομέας που αναπτύσσεται συνεχώς και εξακολουθεί να είναι μείζονος σημασίας για την ανάλυση και την ερμηνεία τεράστιων ποσοτήτων δεδομένων. Η εξόρυξη δεδομένων αναμένεται να συνεχίσει να κερδίζει εξέχουσα θέση τα επόμενα χρόνια, δεδομένου των πολλών εφαρμογών της και του αυξανόμενου όγκου δεδομένων που συλλέγονται.

Ο όρος εξόρυξη δεδομένων μπορεί εύκολα να ερμηνευτεί στο παράδειγμα της καθημερινής ζωής, ωστόσο αυτό που αποτελεί πρόκληση είναι οι τεχνικές που κρύβονται πίσω από αυτή τη ¨μεγάλη¨ λέξη. Ειδικότερα, η εξαγωγή πολύτιμων πληροφοριών ακολουθεί την παρακάτω πορεία [30]:

1. **Καθαρισμός δεδομένων (Data Cleaning)**

   Σε αυτό το πρώτο βήμα γίνεται ένα φιλτράρισμα των δεδομένων έτσι ώστε τα προ-

βληματικά δεδομένα ή δεδομένα με μηδενικό όφελος να αφαιρεθούν.

2. **Ενοποίηση δεδομένων (Data Integration)**

   Σε αυτό το βήμα συνδυάζονται διάφορες πηγές δεδομένων που αφορούν τον τελικό σκοπό της κάθε έρευνας. Παρόλο που τα δεδομένα αυτά μπορεί να διαφέρουν ανάλογα με την πηγή τους, η εκμετάλλευση τους θα είναι πολύ πιο εύκολη, έπειτα την επεξεργασία τους στα επόμενα βήματα.

3. **Επιλογή δεδομένων (Data Selection)**

   Αφού έχει δημιουργηθεί μια βάση δεδομένων, σε αυτό το βήμα θα χρησιμοποιηθεί η (υπάρχουσα) γνώση ως αντικείμενο της έρευνας, έτσι ώστε να επιλεχθούν δεδομένα που είναι χρήσιμα για τον τελικό σκοπό του έργου και θα αποδειχθεί αποδοτικά η επιτυχία τους. Στη συνέχεια θα γίνει χρήση όλων των δεδομένων, ανεξαρτήτως της σημασίας τους, για τον τελικό σκοπό. Η μέθοδος αυτή μπορεί να αποδειχθεί επιζήμια και ως προς την αποθήκευση και μεταφορά του μεγέθους των δεδομένων, αλλά και ως προς την εισαγωγή άσχετων δεδομένων με αποτέλεσμα να μπερδέψει ή να αποπροσανατολίσει τις τεχνικές που έχουν χρησιμοποιηθεί.

4. **Μετασχηματισμός δεδομένων (Data Transformation)**

   Ένα σημαντικό βήμα πριν την εισαγωγή των δεδομένων στις τεχνικές εξόρυξης, είναι η ομοιογένεια και η συμβατότητα των δεδομένων που προέρχονται από διαφορετικές πηγές. Επιπλέον τα δεδομενα θα πρέπει να μετασχηματιστούν σε μορφές που ο εξορυκτικός μηχανισμός μπορεί να αντιληφθεί.

5. **Εξόρυξη δεδομένων (Data Mining)**

   Σε αυτό το σημαντικό βήμα, τα δεδομένα που έχουν συλλεχθεί θα εισαχθούν στις τεχνικές εξόρυξης πληροφορίας. Με αυτό τον τρόπο λαμβάνονται πολύτιμες πληροφορίες που μπορούν να βοηθήσουν στην ανάπτυξη του στόχου δηλαδή του προϊόντος ή της υπηρεσίας.

6. **Αξιολόγηση προτύπων (Pattern Evaluation)**

   Σε αυτό το βήμα ελέγχεται αν οι πληροφορίες που εξορύχθηκαν προσδίδουν στην πράξη πρόοδο, όσον αφορά το προϊόν ή την υπηρεσία.

7. **Παρουσίαση γνώσης (Knowledge Presentation)**

   Τέλος θα χρησιμοποιούνται τεχνικές απεικόνισης και αναπαράστασης γνώσης για να παρουσιαστεί η εξέλιξη και η πρόοδος που επέφερε η εξόρυξη πληροφοριών.

Όπως γίνεται αντιληπτό σε αυτήν την ανάλυση των βημάτων, η εξόρυξη δεδομένων είναι μόνο ένα βήμα στην όλη διαδικασία. Ωστόσο, τα αποτελέσματα αυτού του βήματος μπορούν να καθορίσουν την επιτυχία ή την αποτυχία της όλης διαδικασίας.

2.2.3 Εξόρυξη Δεδομενων απο το Twitter (Twitter Data Mining)

Η διαδικασία συλλογής σημαντικών μοτίβων και πληροφοριών από τον τεράστιο όγκο των tweets που παράγονται από εκατομμύρια χρήστες στην πλατφόρμα κοινωνικών μέσων είναι γνωστή ως εξόρυξη δεδομένων από το Twitter. Η εξόρυξη δεδομένων έχει γίνει ένα βασικό εργαλείο τόσο για τον ακαδημαϊκό κόσμο όσο και για τις επιχειρήσεις λόγω της αυξανόμενης δημοτικότητας του Twitter και του σημαντικού όγκου δεδομένων που δημιουργούνται από τους χρήστες που κατέχει [32].

Η αναγνώριση συμβάντων, η εξόρυξη γνώμης, η ανάλυση συναισθήματος είναι μερικές μόνο από τις πολλές χρήσεις για την εξόρυξη δεδομένων Twitter. Η ανάλυση συναισθήματος είναι η τεχνική για τον προσδιορισμό του συναισθηματικού τόνου, θετικού ή αρνητικού, πίσω από τα tweets που συλλέγονται [33]. Η εξόρυξη γνώμης είναι η διαδικασία προσδιορισμού απόψεων και στάσεων σε tweets που μπορούν να χρησιμοποιηθούν για πολιτικές εκστρατείες, ανάπτυξη προϊόντων και έρευνα αγοράς [34]. Ο εντοπισμός σημαντικών γεγονότων και τάσεων σε πραγματικό χρόνο από tweets, όπως φυσικές καταστροφές, έκτακτες ειδήσεις και κοινωνικά κινήματα, είναι γνωστός ως ανίχνευση συμβάντων.Η ανάλυση τάσεων εξετάζει τη συχνότητα και τη δημοτικότητα συγκεκριμένων θεμάτων μέσω των hashtags και των λέξεων-κλειδιών στα tweets, έτσι δημιουργείται μια άποψη για τα κοινά ενδιαφέροντα και τις συμπεριφορές των χρηστών του Twitter.

Η αντιμετώπιση του τεράστιου όγκου δεδομένων και του αστραπιαίου ρυθμού με τον οποίο παράγονται τα tweets είναι μία από τις βασικές προκλήσεις της εξόρυξης

δεδομένων από το Twitter. Ερευνητές και επαγγελματίες χρησιμοποιούν συχνά τεχνικές επεξεργασίας δεδομένων, όπως είναι η μηχανική εκμάθηση και η επεξεργασία φυσικής γλώσσας, για να εξάγουν σχετικές και χρήσιμες πληροφορίες από τα δεδομένα.

Ένα άλλο ζήτημα που προκύπτει είναι αυτό της ποιότητας των δεδομένων του Twitter. Τα tweets μπορεί να περιέχουν τυπογραφικά λάθη, αργκό και συντομογραφίες που κάνουν το κείμενο ακόμα πιο περίπλοκο για να κατανοηθεί και να αναλυθεί. Επιπλέον, καθώς τα tweets είναι συχνά σύντομα και ανακριβή, οι ερευνητές πρέπει να εφαρμόζουν μεθόδους, σύνοψης κειμένου και διαλογής δεδομένων για να κάνουν τα δεδομένα χρήσιμα και αξιόπιστα για ανάλυση.

Παρά τα εμπόδια, η εξόρυξη δεδομένων Twitter έχει παράγει έναν πλούτο ευρημάτων που την καθιστούν συναρπαστική για μελέτες ακαδημαϊκού τύπου και όχι μόνο. Πολλές μελέτες έχουν δείξει ότι η ανάλυση συναισθήματος από το Twitter μπορεί να προβλέψει σωστά τα πρότυπα των χρηματιστηρίων και τα αποτελέσματα των πολιτικών εκλογών. Επιπλέον, η εξόρυξη δεδομένων από το Twitter έχει εφαρμοστεί στην έρευνα ποικίλων θεμάτων, συμπεριλαμβανομένων των κοινωνικών κινημάτων, της δημόσιας υγείας και των φυσικών καταστροφών.

Πιο συγκεκριμένα επιστρέφοντας στην παρούσα μελέτη, βασικός στόχος είναι η συλλογή δεδομένων από την πλατφόρμα Twitter. Για να γίνει αυτό, υπάρχουν πολλοί τρόποι. Οι πιο δημοφιλείς είναι η απόξεση ιστού (web scraping), το API και τέλος η χειροκίνητη εξαγωγή. Η απόξεση ιστού (web scraping) και το API είναι αυτοματοποιημένες τεχνικές και είναι οι πιο πρακτικοί τρόποι συλλογής δεδομένων [35].

- **Απόξεση ιστού (Web scraping)**

  Το web scraping είναι η συλλογή δεδομένων από έναν ιστότοπο. Αυτή η διαδικασία γίνεται με συστηματικό τρόπο, δηλαδή ένας "πράκτορας" λογισμικού, ένα ρομπότ ή αρχείο κώδικα, μιμείται την αλληλεπίδραση περιήγησης μεταξύ των διακομιστών Ιστού και του ανθρώπου. Δεδομένου ότι το σενάριο συνήθως πρέπει να αλλάξει για να ταιριάζει στην αρχιτεκτονική κάθε ιστότοπου, αυτή η μέθοδος

είναι προτιμότερη όταν για την ανακτήση μεγάλου όγκου πληροφοριών από έναν συγκεκριμένο ιστότοπο [36]. Για να λειτουργήσει, πρέπει να τροφοδοτηθεί η διεύθυνση URL του ιστότοπου στο scraper. Στη συνέχεια, το scraper θα φορτώσει τον ιστότοπο και θα εξαγάγει είτε όλα τα δεδομένα που είναι διαθέσιμα στον ιστότοπο είτε συγκεκριμένα δεδομένα που ζητήθηκαν από τον χρήστη. Τέλος, όταν ολοκληρωθεί η διαδικασία, ο κώδικας θα επιστρέψει ένα εύκολα αναγνώσιμο αρχείο, όπως ένα csv αρχείο, που περιέχει όλες τις πληροφορίες που ζητήθηκαν.

- **Διασύνδεση Προγραμματισμού Εφαρμογών Twitter (Twitter API)**
  Ένα API (Application Programming Interface) αποτελεί βασικά έναν μηχανισμό που απαντά σε συγκεκριμένα αιτήματα - ερωτήματα. Υπάρχουν ιστοσελίδες και πάροχοι που δεν παρέχουν την υπηρεσία αυτή. Παρόλα αυτά το Twitter την πρόσφερει και μάλιστα δωρεάν. Το Twitter API λαμβάνει ένα ερώτημα για τα δεδομένα που μπορεί να είναι χρήσιμα για την εκάστοτε έρευνα, επεξεργάζεται το αίτημα και επιστρέφει το αρχείο με τα ζητούμενα δεδομένα. Δεδομένου ότι το API χρησιμοποιείται μέσα από τον ιστότοπο ή την εταιρεία, είναι πολύ πιο προσιτό και εύκολο στη χρήση του από το web scraping.

Συμπερασματικά, ανάλογα με τους στόχους της κάθε μελέτης και τον τύπο των δεδομένων που θα εξορυχθούν, μπορεί κανείς να επιλέξει ποια μέθοδο θα χρησιμοποιήσει, το Twitter API ή το web scraping, για την εξόρυξη των δεδομένων του. Πριν από την επιλογή της μεθόδου για την παρούσα εργασία, θα αναφερφούν στον παρακάτω Πίνακα 1 κάποιες κύριες διαφορές μεταξύ των δύο μεθόδων [37]. Βάση του Πίνακα αυτού, θα χρησιμοποιηθεί η τεχνική απόξεσης ιστού για την επερχόμενη ανάλυση.

Πίνακας 1: Διαφορές μεταξύ Twitter API και Web Scraper

API	**Web Scraping**
Το API είναι μια εφαρμογή δημιουργημένη από την ίδια την ιστοσελίδα του Twitter, εξαιτίας αυτού δεν υπάρχει περίπτωση να μπλοκάρει κάποια αίτηση από χρήστη ως κακόβουλη, διότι λειτουργεί ήδη σύμφωνα με τους κανόνες της ιστοσελίδας.	Στο web scraper μπορεί να αποκλειστούν από την ιστοσελίδα οι χρήστες αφού προκαλούν ¨σύγχυση¨ στον αριθμό των πραγματικών επισκέψεων της ιστοσελίδας. Ωστόσο οι σύγχρονοι κώδικες του web scraping αλλάζουν την προέλευση του αιτήματος εισόδου. Με αυτόν τον τρόπο η ιστοσελίδα δεν λαμβάνει τις επισκέψεις, ως επισκέψεις από ένα άτομο και επιτρέπει στους χρήστες να συνεχίσουν την εξαγωγή.
Ένα μειονέκτημα που έχει το API είναι ότι ενδέχεται να μην μπορούν να ληφθούν όλες οι πληροφορίες που ζητήθηκαν ή που είναι διαθέσιμες στην ιστοσελίδα. Τα περισσότερα API επιτρέπουν την εξαγωγή μόνο βασικών πληροφοριών, συνοδευμένη με την αιτιολόγηση χρήσης αυτών των δεδομένων.	Το web scraper μπορεί να εξαγάγει ότι πληροφορία υπάρχει στην ιστοσελίδα, δηλαδή ότι είναι ορατό σε μια δημόσια ιστοσελίδα. Αυτό γίνεται σύμφωνα με τους περιορισμούς που έχει καθορίσει η ιστοσελίδα στους όρους και τις προϋποθέσεις της.
Για τη χρήση του Twitter API χρειάζεται να δημιουργηθεί ένας λογαριασμός και μετέπειτα ένας εξατομικευμένος κωδικός που θα περιλαμβάνει τα κλειδιά πρόσβασης και τον καθορισμό των δεδομένων που χρειάζονται.	Η κατασκευή ενός web scraper από την αρχή είναι αρκετά περίπλοκη υπόθεση. Παρόλα αυτά, για εφαρμογές όπως εξαγωγή πληροφοριών απο το Twitter υπάρχουν πολλοί έτοιμοι κώδικες, εντελώς δωρεάν (open-source) που βοηθούν στην διεκπεραίωση της διαδικασίας.
Τα αποτελέσματα που επιστρέφει το API, μπορεί να είναι δύσκολα στην διαχείριση και στο φιλτράρισμα τους.	Το web scraping παρέχει συνήθως ολόκληρο το περιεχόμενο από μια ιστοσελίδα. Ωστόσο, από στους κώδικες web scraping που υπάρχουν ήδη για το Twitter μπορούν να εξακριβωθούν τα κομμάτια πουν ενδιαφέρουν τον ερευνητή.
Αφού το API παρέχεται από την ιστοσελίδα είναι εντελώς νόμιμο.	Το web scraping είναι νόμιμο, εφόσον ο κώδικας ακολουθεί τους όρους και τις προϋποθέσεις του ιστότοπου που καθορίζονται στο αρχείο robot.txt.

2.3 Sentiment Analysis

Η άνοδος των μέσων κοινωνικής δικτύωσης και της ψηφιακής επικοινωνίας, έχει συμπαρασύρει και την αύξηση της ανάλυση συναισθήματος. Αυτό συμβαίνει διότι μέσω των μέσων κοινωνικής δικτύωσης μπορούμε να λάβουμε πραγματικά πολύτιμες πληροφορίες για τις απόψεις και τις αντιλήψεις των χρηστών. Από τις κριτικές που αφήνουν οι χρήστες μιας ιστοσελίδας έως και τις αναφορές σε πολιτικά debate, βοηθούν να διεξαχθούν συμπεράσματα για την κοινή γνώμη [38].

Η ανάλυση συναισθήματος, ή όπως αλλιώς αναφέρεται ως εξόρυξη γνώμης, είναι ένας κλάδος έρευνας που εξετάζει τις στάσεις, τις απόψεις και τα συναισθήματα των ανθρώπων σχετικά με διάφορα θέματα, αγαθά και υπηρεσίες. Στην πραγματικότητα η κατανόηση του τι μπορεί να αισθάνεται ένα άτομο σε μια συνομιλία είναι ζωτικής σημασίας για το πώς θα εξελιχθεί η συνομιλία και πώς θα ενεργήσει ο ακροατής. Μερικές φορές είναι δύσκολο ακόμα και για τους ανθρώπους να κατανοήσουν τι νιώθουν οι άλλοι, πόσο μάλλον για μια μηχανή. Η ανάλυση συναισθήματος (sentiment analysis) αναλαμβάνει αυτήν ακριβώς την πρόκληση.

Συγκεκριμένα, η ανάλυση συναισθήματος είναι η μέθοδος διδασκαλίας μιας μηχανής έτσι ώστε να κατανοεί τι αισθάνεται ένα άτομο/χρήστης, μέσα από ένα κείμενο, έναν ήχο, μια εικόνα ή ακόμα και από ένα βίντεο. Αυτή η μέθοδος χρησιμοποιείται για την κατανόηση της στάσης ενός ατόμου σε ορισμένα σημαντικά θέματα, τις απόψεις και τα γενικά συναισθήματά του, στη δεδομένη περίπτωση για την κλιματική αλλαγή [39]. Στην ανάλυση αυτή θα χρησιμοποιηθούν κείμενα και πιο συγκεκριμένα tweets σχετικά με την κλιματική αλλαγή, και θα αποδοθεί η στάση του χρήστη ως προς το θέμα αυτό. Τα επιλεγμένα tweets είναι συναισθηματικά φορτισμένα για το συγκεκριμένο ζήτημα, ώστε να μάθει η μηχανή να ταξινομεί τα συναισθήματα [40].

Στο πλαίσιο της εξόρυξης συναισθημάτων, τα κείμενα αναλύονται για να συλλάβουν τις απόψεις των ανθρώπων, που συνήθως εμπίπτουν σε δυο κατηγορίες, θετικές ή αρνητικές, με περιστασιακή εξέταση της ουδέτερης άποψης [41].Για να ταξινομηθεί

συναισθηματικά το κείμενο, η ανάλυση συναισθήματος πρέπει να βασιστεί σε προκαθορισμένους γλωσσικούς κανόνες και λεξικά. Από την άλλη πλευρά, η ανάλυση συναισθήματος μπορεί να βασιστεί στη μηχανική μάθηση, για την διενέργεια προβλέψεων με βάση τα μοτίβα που έχει ανακαλύψει στα δεδομένα κειμένου [42]. Αφού ταξινομηθούν τα κείμενα ως προς την πολικότητα τους, υπάρχουν και άλλες δυσκολίες οι οποίες πρέπει να ληφθούν υπόψη, όπως η διαχείριση του κειμένου.

Οι τεχνικές ανάλυσης συναισθήματος έχουν βοηθήσει στο να αναλύονται κείμενα με γραμματικά και συντακτικά λάθη, κείμενα που περιέχουν emojis καθώς και μικρά κείμενα με λίγους χαρακτήρες, όπως τα tweets. Θα πρέπει ακόμα να αναφερθεί ότι, πέρα από τις σύγχρονες μεθόδους για την αντιμετώπιση όλων αυτών των δυσκολιών, υπάρχουν και οι περιπτώσεις που αφορούν την ανάλυση συναισθήματος σε αρχεία κειμένου, όπως είναι η ανάλυση ειρωνικών κειμένων και μη [43].

Η υποκειμενική πτυχή του συναισθήματος, η οποία μπορεί να επηρεαστεί από τις ατομικές εμπειρίες, το πολιτισμικό υπόβαθρο και το πλαίσιο στο οποίο δημιουργήθηκε το κείμενο, αποτελεί ένα ακόμα εμπόδιο για στην ανάλυση συναισθημάτων. Εξαιτίας αυτού, η ανάλυση συναισθήματος μπορεί να είναι εξαιρετικά ευαίσθητη στο πλαίσιο και μια μεμονωμένη λέξη ή φράση μπορεί να φέρει διαφορετικές σημασίες ανάλογα με την κατάσταση.

Στις μελέτες που έχουν διεξαχθεί για την συναισθηματική ανάλυση, μεγάλη έμφαση δίνεται στη βελτίωση της ακρίβειας, στη δημιουργία καλύτερων αλγορίθμων και στην κατανόηση των δυσκολιών της ανάλυσης συναισθήματος. Ο Pang et al. (2002), πρότεινε τη χρήση ταινιών ως πηγή για έρευνα ανάλυσης συναισθημάτων, και οι Kim και Hovy (2004), που ερεύνησαν την επίδραση των λέξεων άρνησης στην ανάλυση συναισθημάτων, είναι δύο από τις πιο ενδεικτικές μελέτες στο θέμα.

Στην επίλυση αυτών των ζητημάτων που προαναφέρθηκαν, βοηθούν οι ραγδαίες εξελίξεις στην τεχνητή νοημοσύνη και τη μηχανική μάθηση, ενισχύοντας τη χρήση της ανάλυσης συναισθημάτων. Έτσι το πεδίο της ανάλυσης συναισθήματος επεκτείνεται

ολοένα και περισσότερο προσφέροντας πιο ακριβή και αξιόπιστα αποτελέσματα.

2.4 Τεχνητή Νοημοσύνη (Artificial Intelligence)

Η τεχνητή νοημοσύνη (Artificial Intelligence) αποτελεί τον κλάδο της τεχνολογίας και της επιστήμης των υπολογιστών, που ασχολείται με την δυνατότητα των έξυπνων μηχανών να μιμούνται τη συμπεριφορά των ανθρώπων, όπως η αναγνώριση ομιλίας, η μετάφραση γλώσσας και η οπτική αντίληψη, δηλαδή ενέργειες που απαιτούν ανθρώπινη νοημοσύνη [29]. Αρχικά, ο όρος «τεχνητή νοημοσύνη» πρωτοεμφανίστηκε σε ένα συνέδριο στο Κολέγιο του Ντάρτμουθ, το 1956. Την ίδια χρονιά, εντάχθηκε σε ακαδημαϊκό επίπεδο.

Η κύρια ενασχόλησή της αφορά την εφαρμογή των υπολογιστών σε εργασίες που απαιτούν γνώση, αντίληψη, συλλογισμό, κατανόηση και γνωστικές ικανότητες [44]. Ο βασικός σκοπός της τεχνητής νοημοσύνης είναι να μιμηθεί τη νοητική λειτουργία του ανθρώπου και να πραγματοποιήσει δραστηριότητες που κανονικά θα πραγματοποιούσε ο άνθρωπος. Τα εργαλεία της τεχνητής νοημοσύνης μπορούν να διαμορφωθούν με τέτοιο τρόπο, ώστε να αξιοποιήσουν ατομικές συμπεριφορές, προτιμήσεις, πεποιθήσεις και ενδιαφέροντα για να εξατομικεύσουν εμπειρίες. Με άλλα λόγια εκπαιδεύουν τις μηχανές να συμπεριφέρονται όπως οι άνθρωποι, δηλαδή να βλέπουν, να ακούν, να μιλούν, να γράφουν [45]. Χωρίς να διδαχθούν από τους ανθρώπους, οι μηχανές χρησιμοποιούν τη δική τους εμπειρία, ώστε να επιλύσουν ένα πρόβλημα.

Η επεξεργασία φυσικής γλώσσας (Natural language processing -NLP), η οποία ασχολείται με την ερμηνεία της φυσικής γλώσσας των ανθρώπων από τους υπολογιστές, είναι μια από τις σημαντικότερες εφαρμογές της τεχνητής νοημοσύνης. Η αναγνώριση ομιλίας, η ανάλυση συναισθημάτων και η αυτόματη μετάφραση είναι μερικές μόνο από τις πολλές χρήσεις του NLP [19]. Αυτοί οι αλγόριθμοι χρησιμοποιούνται στην ανάλυση συναισθήματος για τον προσδιορισμό της πολικότητας ενός κειμένου όπως αναφέρθηκε παραπάνω.

Η τεχνητή νοημοσύνη έχει τη δυνατότητα να αλλάξει ριζικά τον τρόπο ζωής και εργασίας. Για παράδειγμα, η τεχνολογία μπορεί να αυτοματοποιήσει επαναλαμβανόμενες λειτουργίες όπως η εισαγωγή δεδομένων και η υποστήριξη πελατών, απελευθερώνοντας τους ανθρώπους να επικεντρωθούν σε πιο δύσκολες και ευφάνταστες εργασίες. Η τεχνητή νοημοσύνη μπορεί επίσης να βοηθήσει στην εξεύρεση λύσεων σε μερικά από τα πιο επείγοντα ζητήματα που αντιμετωπίζει η ανθρωπότητα, όπως η φτώχεια, η ασθένεια και η κλιματική αλλαγή. Δεν πρέπει επίσης να παραλείπονται ούτε να παραμελούνται οι εξελίξεις που έχουν επέλθει στην υγεία ή την ενέργεια ή πολλούς άλλους τομείς, χάρη στην τεχνητή νοημοσύνη. Αυτές οι εξελίξεις, παρόλο που έγιναν από ανθρώπους, δεν θα είχαν ποτέ τέτοιου είδους αποτελέσματα στο ίδιο χρονικό πλαίσιο, χωρίς τη βοήθεια των μηχανών. Από αυτό και μόνο γίνεται λοιπόν αντιληπτό πόσο ισχυρό εργαλείο είναι η τεχνητή νοημοσύνη.

Παρόλα αυτά η εισχώρηση της τεχνητής νοημοσύνης, στην καθημερινότητα έχει δημιουργήσει μια νέα πραγματικότητα που μπορεί να αποκτήσει έναν απειλητικό ρόλο, λαμβάνοντας υπόψη τον απόλυτο έλεγχο που έχει στη σημερινή ζωή όλων. Ανησυχίες επίσης υπάρχουν για το πώς η τεχνητή νοημοσύνη θα επηρεάσει την κοινωνία, συμπεριλαμβανομένης της απώλειας θέσεων εργασίας και της πιθανότητας κακόβουλης χρήσης των συστημάτων τεχνητής νοημοσύνης.

Συμπερασματικά, η τεχνητή νοημοσύνη είναι ένα ταχέως αναπτυσσόμενο πεδίο που έχει τη δυνατότητα να αλλάξει σημαντικά πολλές πτυχές της ζωής ενός ανθρώπου αρκεί κανείς να συγκρίνει την ζωή του σήμερα με τη ζωή του πριν 10 χρόνια. Η ταχύτητα εξέλιξης της τεχνικής νοημοσύνης έχει επιφέρει αυτή την τόσο μεγάλη αλλαγή και η ανεξέλεγκτη πορεία της μπορεί να προκαλέσει, πέρα από ευεργετικές διευκολύνσεις, ανεπανόρθωτα προβλήματα [46].

Μηχανική μάθηση (Machine Learning)

Αυτή η υποκατηγορία τεχνητής νοημοσύνης περιέχει μια μεγάλη ποικιλία αλγορίθμων και στατιστικών μοντέλων που χρησιμοποιούνται για τη δημιουργία προγνωστικών μοντέλων και για την επίλυση προβλημάτων λήψης αποφάσεων [44]. Η μηχανική μάθηση βασίζεται στην ανάλυση δεδομένων όπως επίσης και στην τροποποίηση αυτών, για να δώσει πολύτιμες πληροφορίες σχετικά με αυτά τα δεδομένα και τις πιθανές διαδικασίες λήψης αποφάσεων. Εφαρμόζεται σε ένα ευρύ φάσμα βιομηχανιών και εφαρμογών, όπως τα συστήματα προτάσεων και η προγνωστική ανάλυση.

Παρόλο που τα αυτοματοποιημένα συστήματα είναι πολύ διαδεδομένα πλέον στην κοινωνία, μερικές φορές είναι δύσκολο να αναγνωρίσει κανείς πόσο σημαντικές και πόσο ευρείες μπορεί να είναι οι υπηρεσίες μηχανικής εκμάθησης [47]. Από τη δυνατότητα επισήμανσης ενός φίλου σε μια φωτογραφία (tagging) έως την πρόταση μιας νέα ταινία (recommendation systems), όλα αυτά αποτελούν εκροές της μηχανικής μάθησης. Είναι προφανές ότι η μηχανική μάθηση μπορεί να ωφελήσει σε διάφορες εργασίες, που ούτως ή άλλως θα μπορούσαν να διεκπεραιωθούν, αλλά πλέον γίνονται πολύ πιο γρήγορα και αποτελεσματικά. Με το μέλλον της μηχανικής μάθησης να προσανατολίζεται σε εργασίες που δεν είναι ανθρωπίνως δυνατές [48].

Παρακάτω παρατίθεται ο τρόπος με τον οποίο λειτουργεί συνήθως η μηχανική εκμάθηση. Αρχικά χρησιμοποιούνται δεδομένα για να εκπαιδεύσουν τις μηχανές να απαντούν σε μια συγκεκριμένη ερώτηση ή εργασία. Η τροφοδοσία δεδομένων στον αλγόριθμο ονομάζεται εκπαίδευση (training). Υπάρχουν πολλοί τύποι μηχανικής μάθησης, αλλά σε αυτή την φάση θα αναφερθούν οι δύο πιο εξέχουσες, η εποπτευόμενη (supervised) και η μη εποπτευόμενη(unsupervised) μάθηση. Και οι δύο τύποι μηχανικής μάθησης γίνονται από μοντέλα εκπαίδευσης σε βάσεις δεδομένων.

Στην εποπτευόμενη εκμάθηση (supervised learning) δίνονται στον αλγόριθμο οδηγίες για το ποια θα πρέπει να είναι η σωστή απάντηση. Για παράδειγμα, για την

αναγνώριση εικόνας, ο αλγόριθμος έχει ως εισαγωγικά δεδομένα πολλές εικόνες. Ωστόσο, για να μάθει ο αλγόριθμος να αναγνωρίζει τους ανθρώπους στις εικόνες, πρέπει να του δοθεί μια ένδειξη για το που απεικονίζονται άνθρωποι σε ποια εικόνα και στη συνέχεια να ελεγχθεί πόσο καλά προβλέπει τους ανθρώπους σε μια άλλη, άγνωστη εικόνα. Αυτό που μόλις περιγράφθηκε είναι η εκπαίδευση (training) και η επικύρωση (validation) μιας εποπτευόμενης μαθησιακής εργασίας.

Από την άλλη πλευρά, η μη εποπτευόμενη μάθηση (unsupervised learning) για μια εργασία εικόνας θα μοιάζει ως εξής: Έχοντας πολλές εικόνες φρούτων (μήλα, μπανάνες και πορτοκάλια) και χωρίς να δοθεί καμία άλλη πληροφορία στη μηχανή, θα πρέπει να παράγει 3 διακριτές ομάδες φρούτων, πιο συγκεκριμένα μια ομάδα εικόνων που να μοιάζουν με μήλα, μια ομάδα εικόνων που μοιάζουν με μπανάνες, και μια ομάδα εικόνων που μοιάζουν με πορτοκάλια.

Ένα από τα βασικά χαρακτηριστικά των αλγορίθμων αυτών είναι η δυνατότητα αυτόνομης βελτίωσης της απόδοσής τους με την πάροδο του χρόνου μέσω της διαδικασίας της εκπαίδευσης. Κατά τη διάρκεια της εκπαίδευσης, ο αλγόριθμος παρουσιάζεται με ένα σύνολο δεδομένων εισόδου και την αντίστοιχη αναμενόμενη έξοδο. Στη συνέχεια, ο αλγόριθμος προσαρμόζει τις παραμέτρους του προκειμένου να παράγει μια πιο ακριβή έξοδο με βάση το αρχικό ερώτημα. Αυτή η διαδικασία επαναλαμβάνεται συνεχώς μέχρι ο αλγόριθμος να φτάσει σε ένα ικανοποιητικό επίπεδο ακρίβειας [49].

Το ποσοστό επιτυχίας των αλγορίθμων μηχανικής μάθησης εξαρτάται κατά κύριο λόγο από την ποιότητα των δεδομένων που χρησιμοποιούνται στην εκπαίδευση. Τα ανακριβή ή μεροληπτικά δεδομένα μπορεί να οδηγήσουν σε ανακριβή αποτελέσματα, γι' αυτό είναι σημαντικό να γίνεται σωστή προεπεξεργασία και να καθαρίζονται τα δεδομένα πριν από την εκπαίδευση. Η επιλογή χαρακτηριστικών και η μείωση διαστάσεων είναι επίσης σημαντικά βήματα στο στάδιο της προεπεξεργασίας, καθώς μπορούν να βοηθήσουν στη μείωση της πολυπλοκότητας των δεδομένων και στη βελτίωση της απόδοσης του αλγορίθμου.

Ένας αριθμός δημοφιλών αλγορίθμων μηχανικής μάθησης έχει αναπτυχθεί τα τελευταία χρόνια, συμπεριλαμβανομένων των Logistic regressions, Support vector machines, Random forest και K-nearest neighbours που θα αναλυθούν παρακάτω και πολλοί άλλοι. Η επιλογή του αλγόριθμου που θα χρησιμοποιηθεί εξαρτάται από το είδος του προβλήματος και το σύνολο δεδομένων που είναι διαθέσιμα για την εκπαίδευση.

Παρά την εκτενή χρήση των αλγορίθμων μηχανικής μάθησης, εξακολουθούν να υπάρχουν ορισμένοι περιορισμοί και προκλήσεις. Το overfitting (η υπερπροσαρμογή) είναι μια από αυτές τις προκλήσεις που καλούνται να αντιμετωπίσουν οι data scientists στη μηχανική μάθηση. Το overfitting συμβαίνει όταν ένα μοντέλο ταιριάζει πολύ καλά στα δεδομένα εκπαίδευσης, αλλά έχει κακή απόδοση στα δεδομένα δοκιμής. Αυτό μπορεί να προκύψει λόγω των πολλών παραμέτρων του μοντέλου ή στο γεγονός ότι είναι πολύ περίπλοκο για τα δεδομένα. Μια άλλη πρόκληση είναι η έλλειψη ερμηνείας των μοντέλων. Πολλοί αλγόριθμοι μηχανικής μάθησης θεωρούνται ¨μαύρα κουτιά¨ επειδή είναι δύσκολο να κατανοηθεί πώς το μοντέλο κατέληξε σε μια συγκεκριμένη απόφαση. Αυτό μπορεί να είναι ένα σημαντικό πρόβλημα σε εφαρμογές όπου πρέπει να γίνεται φανερός ο τρόπος διαχείρισης.

Παρά τις προκλήσεις και τις δυσκολίες που μπορεί να προκύπτουν, η μηχανική μάθηση συνεχώς αναπτύσσεται και να εξελίσσεται, με νέες τεχνικές και αλγόριθμους, επιτρέποντας στους υπολογιστές να βελτιώνουν αυτόματα την απόδοσή τους μέσω της εμπειρίας. Η μηχανική εκμάθηση έχει τη δύναμη να αντιμετωπίζει σύνθετες εργασίες και να παρέχει νέες πληροφορίες και λύσεις. Είναι ένας ταχέως αναπτυσσόμενος τομέας που έχει τη δυνατότητα να φέρει την επανάσταση στον τρόπο με τον οποίο λαμβάνονται αποφάσεις και επιλύονται προβλήματα.

Βαθιά Μάθηση (Deep Learning)

Η βαθιά μάθηση είναι μια υποκατηγορία της μηχανικής μάθησης που βασίζεται κυρίως στην αρχιτεκτονική των τεχνητών νευρωνικών δικτύων [50, 44]. Δηλαδή, αλγόριθμους που μιμούνται τον τρόπο με τον οποίο λειτουργούν οι νευρώνες στον ανθρώπινο εγκέφαλο. Όπως οι νευρώνες μεταφέρουν πληροφορίες στον ανθρώπινο εγκέφαλο, έτσι και οι νευρώνες σε έναν αλγόριθμο τεχνητών νευρωνικών δικτύων μεταφέρουν πληροφορίες και προσπαθούν να αποκτήσουν μια αξιόλογη ερμηνεία των δεδομένων κάνοντας αυτές τις συνδέσεις.

Η βαθιά εκμάθηση είναι πιο προηγμένη από την απλή μηχανική μάθηση, καθώς μπορεί να χειριστεί πολύ πιο περίπλοκα και αφηρημένα δεδομένα, έχει πολύ πιο περίπλοκη αρχιτεκτονική, που αποτελείται από αρκετούς νευρώνες και πολλά κρυφά επίπεδα.Τα τεχνητά νευρωνικά δίκτυα που χρησιμοποιεί προϋπήρχαν, αλλά αναζωπυρωθήκαν μόνο όταν οι μεγάλες ποσότητες σχολιασμένων δεδομένων κατάφεραν να επιτύχουν αποτελέσματα σε ένα ευρύ φάσμα εργασιών. Ο όρος "deep" στη βαθιά μάθηση αναφέρεται στον μεγάλο αριθμό επιπέδων στο νευρωνικό δίκτυο, σε αντίθεση με τα ρηχά δίκτυα του παρελθόντος. Τα βαθύτερα δίκτυα επιτρέπουν στο μοντέλο να μάθει πιο σύνθετες αναπαραστάσεις και να κάνει πιο ακριβείς προβλέψεις.

Οι αλγόριθμοι βαθιάς μάθησης μπορούν να χωριστούν σε νευρωνικά δίκτυα τροφοδοσίας και σε επαναλαμβανόμενα δίκτυα [51]. Τα δίκτυα τροφοδοσίας επεξεργάζονται τα δεδομένα εισόδου με διαδοχικό τρόπο, περνώντας τα δεδομένα μέσω πολλαπλών νευρωνικών επιπέδων μέχρι να γίνει μια τελική πρόβλεψη. Από την άλλη τα επαναλαμβανόμενα δίκτυα έχουν συνδέσεις ανατροφοδότησης που επιτρέπουν στο δίκτυο να διατηρεί πληροφορίες από προηγούμενες εισόδους, καθιστώντας τα κατάλληλα για την επεξεργασία διαδοχικών δεδομένων όπως ομιλία ή κείμενο.

Ένας ευρέως διαδεδομένος τύπος βαθιάς μάθησης είναι τα συνελικτικά νευρωνικά δίκτυα (CNN). Αυτό ο τύπος δικτύου χρησιμοποιείται κατά βάση για εργασίες ανα-

γνώρισης εικόνας. Τα CNN επεξεργάζονται δεδομένα εικόνας μέσα από φίλτρα συλλογή δεδομένων σε περιοχές εμφάνισης της εικόνας για να δημιουργήσει αναπαραστάσεις υψηλότερου επιπέδου. Η διαδικασία αυτή δίνει τη δυνατότητα στα CNN να μαθαίνουν αυτόματα τα σημαντικά χαρακτηριστικά μιας εικόνας και να βελτιώνουν την ακρίβεια των προβλέψεων.

Ένας άλλος τύπος αλγορίθμου βαθιάς μάθησης είναι το γενετικό αντιφατικό δίκτυο (GAN), το οποίο αποτελείται από δύο δίκτυα: τον δημιουργό και τον διαχωριστή. Ο δημιουργός επινοεί πλασματικά δείγματα, παρόμοια με τα πραγματικά δεδομένα, ενώ ο διαχωριστής διεξάγει έρευνα για να διακρίνει τα πραγματικά από τα πλασματικά δείγματα. Κατά τη διάρκεια της εκπαίδευσης, ο δημιουργός και ο διαχωριστής εκπαιδεύονται ταυτόχρονα, με τον δημιουργό να προσπαθεί να δημιουργήσει δείγματα που δεν διαφέρουν από τα πραγματικά δεδομένα και τον διαχωριστή να προσπαθεί να αναγνωρίσει τα πλασματικά δείγματα. Ως αποτέλεσμα τα δίκτυα GAN μπορούν να παράγουν εικόνες υψηλής ποιότητας ή άλλα δείγματα δεδομένων που είναι παρόμοια με τα πραγματικά δεδομένα.

Όπως γίνεται φανερό ακόμα και με περιορισμένα δεδομένα εκπαίδευσης οι αλγόριθμοι βαθιάς μάθησης έχουν την ικανότητα να δημιουργούν πολύπλοκες αναπαραστάσεις και να κάνουν ακριβείς προβλέψεις. Αυτό συμβαίνει διότι επιτρέπουν στο μοντέλο να μάθει τα χαρακτηριστικά και τις αναπαραστάσεις απευθείας από τα δεδομένα, αντί να τα προσδιορίζει εκ των προτέρων. Επιπλέον, οι αλγόριθμοι βαθιάς μάθησης μπορούν να αναπτυχθούν καλύτερα σε μεγάλες ποσότητες δεδομένων και μπορούν να βελτιώσουν τις προβλέψεις τους με την πάροδο του χρόνου καθώς προστίθενται περισσότερα δεδομένα.

Υπάρχουν βέβαια ορισμένοι περιορισμοί και σε αυτού του είδους τους αλγόριθμους. Για την επιτυχημένη λειτουργία τους χρήζουν μεγάλες ποσότητες σχολιασμένων δεδομένων, η συλλογή των οποίων απαιτεί χρόνο και πολύ μνήμη που σημαίνει ότι μπορεί να είναι κοστοβόρα [52]. Επιπλέον, μπορεί να είναι δύσκολο να ερμηνευθούν και μερικές φορές μπορεί να λάβουν αποφάσεις που είναι δύσκολο να κατανοηθούν.

Παρότι υπάρχουν αυτοί οι περιορισμοί, η βαθιά εκμάθηση έχει σημειώσει σημαντική πρόοδο σε ένα ευρύ φάσμα εφαρμογών. Για παράδειγμα, αλγόριθμοι βαθιάς μάθησης έχουν χρησιμοποιηθεί για να βελτιώσουν την ακρίβεια των συστημάτων αναγνώρισης ομιλίας και να τα καταστήσουν πιο ανθεκτικά [53, 54]. Στην επεξεργασία φυσικής γλώσσας, αλγόριθμοι βαθιάς μάθησης έχουν χρησιμοποιηθεί για την ανάπτυξη μοντέλων που μπορούν να μεταφράσουν γλώσσες, να συνοψίσουν κείμενα ακόμα και να απαντήσουν σε ερωτήσεις. Εργασίες όπως η αναγνώριση ομιλίας, η κατάτμηση εικόνας και η επεξεργασία φυσικής γλώσσας αντιμετωπίζονται συνήθως με τεχνικές βαθιάς μάθησης [55, 56].

3 Μεθοδολογία

3.1 Μοντέλα αναπαράστασης λέξεων στην NLP

Τα μέσα κοινωνικής δικτύωσης και τα διαδικτυακά forum έχουν σχεδόν αποκτήσει το μονοπώλειο των ανθρώπων όσον αφορά την έκφραση των απόψεων τους και των συναισθημάτων τους. Οι χρήστες αυτών των μέσων γράφουν με διαφορετικούς τρόπους, χρησιμοποιούν συντομογραφίες, σημεία στίξης, emoticons, hastags, αργκό και συχνά χρησιμοποιούν και URL. Με την πρωταρχική τους μορφή τα κείμενα αυτά είναι δύσκολο να κατανοηθούν από μια μηχανή.

Τα δεδομένα λοιπόν που υπάρχουν στα κείμενα αυτά αυξάνονται με ραγδαίους ρυθμούς αλλά αποτελούνται από μη δομημένα δεδομένα χαμηλής ποιότητας τα οποία είναι δύσκολο να επεξεργαστούν και ταξινομηθούν ως έχουν. Πάραυτα αποτελούν μία πηγή πλούσια σε πληροφορίες τα οποία δίνουν περισσότερες ευκαιρίες στην κατανόηση ωφέλιμης γνώσης που δεν μπορεί να επιτευχθεί μέσω των ποσοτικών δεδομένων. Πολλές έρευνες και μεθοδολογίες έχουν επικεντρώσει την προσοχή τους στην επίλυση αυτών των προβλημάτων και στη διαδικασία μετατροπής ακατέργαστων δεδομένων σε αριθμητικά διανύσματα που μπορούν να κατανοήσουν οι μηχανές.

Η αναπαράσταση λέξεων, συμπεριλαμβανομένης της σημασίας τους, ήταν ανέκαθεν μια ιδιαίτερα σημαντική ευρευνητική περιοχή για το πεδίο της NLP (Natural language Processing). Οι διανυσματικές αναπαραστάσεις λέξεων θεωρούνται μια από τις πιο σημαντικές ανακαλύψεις της βαθιάς μάθησης για την επίλυση προβλημάτων επεξεργασίας φυσικής γλώσσας.Το ίδιο σημαντική είναι και η κατανόηση της σημασιολογίας και των συντακτικών πληροφοριών για την ανάλυση κειμένου. Στο συγκεκριμένο σημείο θα αναφερθούν διάφορα μοντέλα αναπαράστασης λέξεων και πως έχουν εξελιχθεί στο χρόνο από τα κλασικά (προ εποχής της βαθιάς μάθησης) μέχρι τα πιο σύγχρονα μοντέλα.

3.1.1 Κλασικά Μοντέλα

Σε αυτή την ενότητα παρουσιάζονται μερικά από τα κλασικά μοντέλα που χρησιμοποιούνται για την ταξινόμηση του κειμένου. Χωρίζονται σε κατηγορικές και σταθμισμένες αναπαραστάσεις λέξεων και μετατρέπουν το κείμενο σε μια διανυσματική μορφή.

One Hot Encoding

Είναι η απλούστερη μέθοδος αναπαράστασης των λέξεων με μια απλή κωδικοποίηση. Με την κωδικοποίηση one-hot, μια κατηγορική μεταβλητή, όπως μια λέξη ή ένας χαρακτήρας, μετατρέπεται σε ένα αριθμητικό διάνυσμα που μπορεί να χρησιμοποιηθεί ως είσοδος για αλγόριθμους μηχανικής μάθησης και στην ανάλυση δεδομένων.

Η τεχνική χρησιμοποιείται για τη μετατροπή κατηγορικών δεδομένων σε μορφή αριθμητικού διανύσματος. Το μήκος αυτού του διανύσματος θα είναι ίσο με την ποσότητα των όρων που υπάρχουν στο λεξιλόγιο που έχει δοθεί. Κάθε όρος του λεξιλογίου μετατρέπεται σε μια δυαδική μεταβλητή με μηδέν (0) ή ένα (1), δηλαδή κάθε λέξη αναπαριστάται από ένα δυαδικό διάνυσμα που έχει σε όλες τις θέσεις του μηδενικά (0) εκτός από την θέση που αντιστοιχεί στην συγκεκριμένη λέξη, η οποία ισούται με ένα(1). Στην ουσία η κεντρική ιδέα της One Hot Encording μεθοδολογία είναι να δημιουργήσει ένα πίνακα με μηδενικά (0) και ένα (1). Οι γραμμές του πίνακα αντιστοιχούν στις προτάσεις που έχουν δοθεί, όπως για παράδειγμα tweets, και οι στήλες του πίνακα αντιστοιχούν στις μοναδικές λέξεις που υπάρχουν σε αυτά τα tweets. Όταν μια λέξη υπάρχει στο εκάστοτε tweet τότε παίρνει την τιμή ένα (1) στη θέση που αντιστοιχεί ειδάλλως εάν δεν υπάρχει παίρνει την τιμή μηδέν(0).

Η σχέση μεταξύ των λέξεων σε αυτή την αναπαράσταση δεν λαμβάνεται υπόψη. Αυτή την παράλειψη προσπαθεί να καλύψει η κωδικοποίηση One Hot Enonding - Bigrams. Η διαφορά είναι ότι δημιουργούνται αρχικά δυάδες λέξεων (Bigrams) οι οποίες βρίσκονται η μία δίπλα από την άλλη στο tweet και οι οποίες αντιμετωπίζονται πλέον ως ένας όρος (μία οντότητα). Στη συνέχεια γίνεται η παράθεση των νέων όρων

στον πίνακα σε συνδυασμό με τα tweets για να δημιουργηθεί ο πίνακας και να γίνει η τοποθέτηση της μονάδας (1) στην παρουσία ή του μηδενικού (0) στην απουσία.

Ένα πλεονέκτημα της one-hot κωδικοποίησης είναι ότι είναι απλή στην εφαρμογή της, δεν απαιτεί μεγάλη υπολογιστική ισχύ και μπορεί να χρησιμοποιηθεί σε ένα ευρύ φάσμα μεθόδων μηχανικής μάθησης, όπως Neural networks, Support vector machines, Decision trees και Random forests, εξαιτίας της ευελιξίας της. Ένα ακόμα πλεονέκτημά της είναι η ικανότητά της να χειρίζεται αραιά δεδομένα. Τα αραιά δεδομένα είναι δεδομένα στην πλειονότητά τους απαρτίζονται από ελλιπείς τιμές ή μηδενικά. Η κωδικοποίηση αυτή έχει τη δυνατότητα μετατροπής τέτοιων δεδομένων σε συμπαγή μορφή, καθιστώντας ευκολότερη την επεξεργασία τους. Αυτό το πλεονέκτημα είναι ιδιαίτερα χρήσιμο για μεγάλα σύνολα δεδομένων όπου ο χρόνος επεξεργασίας και η χρήση μνήμης μπορεί να γίνουν ανασταλτικός παράγοντας. Επίσης η One hot encoding είναι ότι αντιμετωπίζει χωριστά την κάθε λέξη χωρίς να λαμβάνει αποτυπώνει τις σχέσεις μεταξύ τους, σημασιολογικές και συντακτικές. Αυτό μπορεί να οδηγήσει σε απώλεια πληροφοριών που θα μπορούσαν να είναι πολύτιμες για τους αλγόριθμους μηχανικής εκμάθησης.

Υπάρχουν όμως και αρκετά μειονεκτήματα, το σημαντικότερο εξ αυτών είναι γνωστό ως η κατάρα της διάστασης και στην ουσία είναι η τάση που έχει να αυξάνει τη διάσταση του συνόλου δεδομένων. Κάθε κατηγορικό χαρακτηριστικό στο σύνολο των δεδομένων μετατρέπεται σε ένα δυαδικό διάνυσμα, το οποίο μπορεί να οδηγήσει σε πολύ μεγάλο αριθμό χαρακτηριστικών. Σημαντικό είναι να αναφερθεί ότι η απόδοση των αλγορίθμων μηχανικής μάθησης, επηρεάζεται εξαιρετικά από την αυξημένη πολυπλοκότητα του συνόλου δεδομένων. Επίσης η One hot encoding είναι ότι αντιμετωπίζει χωριστά την κάθε λέξη χωρίς να αποτυπώνει τις σχέσεις μεταξύ τους, σημασιολογικές και συντακτικές. Αυτό μπορεί να οδηγήσει σε απώλεια πληροφοριών που θα μπορούσαν να είναι πολύτιμες για τους αλγόριθμους μηχανικής εκμάθησης.

Συμπερασματικά, παρά τα μειονεκτήματα που εμφανίζει, παραμένει ένα πολύτιμο εργαλείο για την ανάλυση δεδομένων και τη μηχανική μάθηση, καθώς μπορεί να

συνδυαστεί με άλλες τεχνικές, όπως η μείωση διαστάσεων, για να μετριάσει τους περιορισμούς του και έπειτα εισάγοντας τα δυαδικά διανύσματα που δημιουργεί σε αλγορίθμους μηχανικής μάθησης να τους επιτρέψει να ταξινομήσουν το συναίσθημα του κειμένου ή να προβλέψουν το θέμα του εγγράφου.

Bag of Words

Το Bag of Words (BoW) είναι μια τεχνική όπου πρακτικά κάθε έγγραφο αναπαρίσταται ως ιστόγραμμα των συχνοτήτων των λέξεων. Αυτό το ιστόγραμμα προκύπτει διαιρώντας πρώτα το έγγραφο σε λέξεις (token) και στη συνέχεια μετρώντας τον αριθμό των εμφανίσεων κάθε λέξης [57] . Χρησιμοποιείται συνήθως στην ανάλυση κειμένου και στην επεξεργασία φυσικής γλώσσας. Αντιπροσωπεύει έγγραφα κειμένου ως αριθμητικά διανύσματα, τα οποία στη συνέχεια μπορούν να χρησιμοποιηθούν ως είσοδος για διάφορους αλγόριθμους μηχανικής εκμάθησης και έχει αποδειχθεί μια αποτελεσματική προσέγγιση για ταξινόμηση κειμένου.

Αυτή η μέθοδος είναι μια επέκταση της κωδικοποίησης One Hot. Στην ουσία προσθέτει τις One Hot αναπαραστάσεις των λέξεων μέσα σε μια πρόταση, ένα άρθρο ή ένα tweet [58, 59]. Είναι μια αναπαράσταση κειμένου που απαριθμεί την εμφάνιση των λέξεων μέσα στο κείμενο (στη συγκεκριμένη περίπτωση σε ένα tweet) χωρίς να λαμβάνει υπόψη την σειρά που εμφανίζονται οι λέξεις στο κείμενο και πληροφορίες γραμματικής και σύνταξης, εξού και η ονομασία Bag of words (¨τσάντα¨ λέξεων). Παρακάτω παρατίθενται τα βήματα δημιουργίας μια Bag of words αναπαράστασης.

Αρχικά γίνεται tokenization, δηλαδή ο διαχωρισμός του κειμένου σε token. Αυτό συνήθως επιτυγχάνεται με διαίρεση του κειμένου χρησιμοποιώντας κενά και σημεία στίξης. Έπειτα γίνεται το φιλτράρισμα, οι λέξεις διακοπής (όπως ¨το¨, ¨και¨ και ¨του¨) και άλλα διακριτικά που θεωρούνται άσχετα, χωρίς νόημα και αφαιρούνται από το κείμενο σε αυτό το στάδιο. Το φιλτράρισμα ακολουθείται από την κωδικοποίηση των tokens σε μια αριθμητική αναπαράσταση όπως στην one-hot κωδικοποίηση. Και τέλος γίνεται η καταμέτρηση της συχνότητας εμφάνισης κάθε token. Αυτό επιτυγχάνεται με

τον υπολογισμό του αριθμού των φορών που κάθε λέξη ή τοκεν εμφανίζεται στο κείμενο. Ως αποτέλεσμα, δημιουργείται ένας αραιός πίνακας, οι περισσότερες καταχωρήσεις του οποίου είναι μηδενικές και μόνο μερικές είναι μη μηδενικές, αυτές που αντιστοιχούν στις λέξεις που υπάρχουν στο έγγραφο. Κάθε σειρά του πίνακα αυτού υποδηλώνει ένα έγγραφο ενώ κάθε στήλη ένα token.

Πλεονεκτήματα που παρουσιάζει το BOW σε σχέση με τις άλλες τεχνικές αναπαράστασης κειμένου, είναι η απλότητα της υλοποίησης και η ικανότητα χειρισμού μεγάλων ποσοτήτων δεδομένων κειμένου. Επιπλέον, παρέχει μια τυπική αναπαράσταση της δομής του κειμένου, η οποία μπορεί να χρησιμοποιηθεί ως είσοδος για πιο εξελιγμένες τεχνικές ανάλυσης κειμένου.

Ωστόσο, το BoW έχει αρκετούς περιορισμούς ένας από αυτούς είναι η απώλεια νοήματος που προκύπτει από τη μείωση των λέξεων που χρησιμοποιεί. Για παράδειγμα, λέξεις με πολλές σημασίες ή παρόμοιες λέξεις με διαφορετικές έννοιες, αντιμετωπίζονται ως ίδιες, οδηγώντας σε απώλεια σημασιολογικών πληροφοριών. Επιπλέον δεν καταγράφει τη σειρά ή τις σχέσεις μεταξύ των λέξεων στο έγγραφο, κάτι που είναι σημαντικό για πολλές εργασίες ανάλυσης κειμένου.Παρά τους περιορισμούς αυτούς, το BoW παραμένει μια δημοφιλής και ευρέως χρησιμοποιούμενη τεχνική στον τομέα της ανάλυσης κειμένου. Έχει χρησιμοποιηθεί σε ποικίλες εφαρμογές, όπως ταξινόμηση κειμένου, ανάλυση συναισθημάτων, μοντελοποίηση θεμάτων και ανάκτηση πληροφοριών.

Term Frequency/–Inverse Document Frequency (TF-IDF)

Είναι μια κατηγορία σταθμισμένης αριθμητικής αναπαράσταση δεδομένων κειμένου που αντικατοπτρίζει τη σημασία μιας συγκεκριμένης λέξης σε ένα έγγραφο. Ο στόχος της TF-IDF αναπαράστασης είναι να παρέχει μια σταθμισμένη αναπαράσταση των πληροφοριών κειμένου σε ένα έγγραφο και να αποτυπώνει τις πιο κατατοπιστικές λέξεις [60].

Ο όρος Term Frequency (TF-συχνότητα όρου) υπολογίζεται διαιρώντας τον αριθμό συχνότητας εμφάνισης μιας λέξης με τον συνολικό αριθμό τον λέξεων του κειμένου. Παρόλα αυτά η συχνότητα εμφάνισης των όρων δεν αποτελεί τον καλύτερο τρόπο αναπαράστασης ενός κειμένου, καθώς σε ένα κείμενο λέξεις όπως άρθρα, προθέσεις κλπ. εμφανίζουν σχεδόν πάντα την μεγαλύτερη συχνότητα. Αυτό το μειονέκτημα έρχεται να καλύψει ο όρος Inverse Document Frequency (IDF- αντίστροφη συχνότητα εγγράφου) ο οποίος θέτει μία κατανομή βάρους στον TF, μειώνοντας έτσι την επίδραση των κοινών λέξεων [61]. Η αντίστροφη συχνότητα εγγράφου (IDF) υπολογίζεται ως ο λογάριθμος του συνολικού αριθμού εγγράφων διαιρεμένος με τον αριθμό των εγγράφων που περιέχουν τη συγκεκριμένη λέξη . Αυτός ο συνδυασμός μεθόδου TF και IDF είναι γνωστός ως TF-IDF και προκύπτει πολλαπλασιάζοντας τον όρο συχνότητα με την αντίστροφη συχνότητα εγγράφου, όπως φαίνεται μαθηματικά από την παρακάτω εξίσωση.

$$TF - IDF(t, d, D) = TF(t, d)(x)log(\frac{D}{dft})$$

Όπου t οι όροι, d κάθε έγγραφο, D η συλλογή των εγγράφων, dft το άθροισμα των εγγράφων με t όρους. Το μοντέλο TF-IDF είναι βασισμένο στην κωδικοποίηση BOW.

Η TF-IDF αναπαράσταση, εμφανίστηκε πρώτη φορά από τον K. Spärck Jones το 1972 και χρησιμοποιείται σε διάφορες εφαρμογές NLP όπως η ταξινόμηση, η ομαδοποίηση και η ανάκτηση κειμένου. Ένα από τα πλεονεκτήματα της χρήσης του TF-IDF είναι ότι εξαλείφει τη σημασία των λέξεων που εμφανίζονται συχνά, όπως για παράδειγμα οι τερματικές λέξεις, οι οποίες δεν περιέχουν χρήσιμη πληροφορία.

Στην ταξινόμηση κειμένου, το TF-IDF χρησιμοποιείται συχνά ως η αναπαράσταση χαρακτηριστικών των δεδομένων κειμένου. Έχει αποδειχθεί ότι το TF-IDF είναι μια αποτελεσματική αναπαράσταση για ταξινόμηση κειμένου, ξεπερνώντας άλλες αναπαραστάσεις όπως το One Hot Encoding, το Bag of Words και το Word2Vec σε αρκετές μελέτες [57, 62].Στην ομαδοποίηση κειμένου, χρησιμοποιείται για την ομαδοποίηση παρόμοιων εγγράφων με βάση το περιεχόμενό τους. Η ομοιότητα μεταξύ των εγγράφων

υπολογίζεται με βάση την αναπαράσταση των εγγράφων TF-IDF . Στην ανάκτηση κειμένου, το TF-IDF χρησιμοποιείται για την ταξινόμηση των εγγράφων με βάση τη συνάφειά τους με ένα δεδομένο ερώτημα. Η συνάφεια ενός εγγράφου υπολογίζεται με βάση την αναπαράσταση TF-IDF του εγγράφου και του ερωτήματος [63].

3.1.2 Word Embeddings - Ενσωματώσεις Λέξεων

Η έλλειψη εξαγωγής χαρακτηριστικών των κλασικών μοντέλων καθώς και η δημιουργία αραιών και μεγάλης διάστασης πινάκων θέτουν περιορισμούς στην δημιουργία ενός κατάλληλου μοντέλου Machine Learning.

Στον τομέα του NLP, τα Word Embeddings (Ενσωματώσεις λέξεων) έχουν πλέον αντικαταστήσει τις κατηγορικές μεθόδους αναπαράστασης. Στην ουσία πρόκειται για μια μέθοδο εκμάθησης χαρακτηριστικών όπου μια λέξη από το λεξιλόγιο αντιστοιχίζεται σε ένα διάνυσμα διαστάσεων. Οι διανυσματικές αυτές αναπαραστάσεις κωδικοποιούν το νόημα των λέξεων, με τέτοιο τρόπο έτσι ώστε οι λέξεις που βρίσκονται πιο κοντά στο διανυσματικό χώρο να μοιάζουν σημασιολογικά. Οι μέθοδοι αυτοί βασίζονται σε νευρωνικά δίκτυα και αποτελούν κατανεμημένες μεθόδους αναπαράστασης [64].

Ο Bengio ήταν ο πρώτος που εισήγαγε το 2003 στα NLP τις τεχνικές νευρωνικών δικτύων για τη δημιουργία διανυσματικών αναπαραστάσεων. Το NNLM Neural Network Language Models, που παρουσιάστηκε από τους συγγραφείς ήταν ένα από τα πρώτα μοντέλα ενσωμάτωσης. Το μοντέλο πρακτικά, χρησιμοποιούσε τη προηγούμενη λέξη n1 για να μαντέψει την επερχόμενη λέξη n. Τέτοιου είδες μοντέλα καθιστούν τα uni-gram χρήσιμα και κατανοητά στους αλγόριθμους machine learning και συνήθως χρησιμοποιούνται στο πρώτο επίπεδο εκπαίδευσης μοντέλων που βασίζονται στα βαθιά νευρωνικά δίκτυα.

Οι ενσωματώσεις λέξεων χρησιμοποιούνται συχνά ως στοιχεία εισόδου για τα μοντέλα μηχανικής εκμάθησης, όπως Neural networks, Decision trees και Support vector

machines. Το πλεονέκτημα της χρήσης ενσωματώσεων λέξεων σε σχέση με τις παραδοσιακές αναπαραστάσεις Bag of words ή του One hot encoding είναι ότι αποτυπώνουν τη σημασιολογική σημασία των λέξεων και μπορούν να χειριστούν αποτελεσματικά τη σημασιολογική σχέση μεταξύ των λέξεων. Τέτοιου είδους ενσωματώσεις έχουν αποδειχτεί πολύ χρήσιμες σε διάφορες εργασίες NLP . Παρακάτω παρατίθενται ορισμένες από αυτές τις ενσωματώσεις, που χρησιμοποιήθηκαν κατά βάση και στην παρούσα εργασία.

Word2Vec

Το Word2Vec είναι ένα μοντέλο βαθιάς μάθησης που χρησιμοποιείται για την παραγωγή πυκνών διανυσμάτων (ενσωματώσεις λέξεων) που αποτυπώνουν το σημασιολογικό νόημα των λέξεων. Αναπτύχθηκε το 2013 από τον Tomas Mikolov και είναι μια κατηγορία πολύ αποτελεσματικών μοντέλων για ενσωματώσεις λέξεων, που έχουν διαμορφώσει μια ισχυρή τάση στις εργασίες ομοιότητας λέξεων NLP [65].

Αποτελεί μια μη επιβλεπόμενη τεχνική, που πρακτικά λαμβάνει μία είσοδο (ένα ολόκληρο κείμενο ή μια λέξη) και ως έξοδο αποτυπώνει word embeddings. Η κατηγορία αυτή αποτελείται από δυο διαφορετικά μοντέλα τεχνικής μάθησης το Combined Bag of Words (CBOW) και το Skip-gram. Οι δύο αυτοί μέθοδοι χρησιμοποιούν νευρωνικά δίκτυα λίγων επιπέδων (τα λεγόμενα ρηχά νευρωνικά δίκτυα) για να προσδιορίσουν τα διανύσματα των word embeddings.

Η εκπαίδευση της κάθε λέξης γίνεται σε ένα συγκεκριμένο πλαίσιο γειτονικών λέξεων (το λεγόμενο παράθυρο), στο corpus των κειμένων που χρησιμοποιούνται και η αναπαράσταση γίνεται σε έναν διανυσματικό χώρο N. Κάθε λέξη αναπαριστάται ως ένα διάνυσμα διάστασης N (συνήθως N=300). Με αυτόν τον τρόπο παρέχετε ένα ισχυρό εργαλείο για την ανάπτυξη σχέσεων στο corpus και την ομοιότητα μεταξύ των λέξεων.

Το CBOW λαμβάνει ως είσοδο ένα συγκείμενο πλαίσιο και προσπαθεί να προβλέψει την πιθανότητα μιας λέξης, ενώ το Skip-gram, εργάζεται για την αντίστροφη διαδικασία, παίρνει ως είσοδο μία λέξη και προσπαθεί να προβλέψει το συγκείμενο πλαίσιο. Τέτοιου

είδους αναπαραστάσεις διαθέτουν συντακτικές και σημασιολογικές πληροφορίες των λέξεων του παραθύρου.Το skip- gram είναι πιο αποτελεσματικό σε λιγότερα δεδομένα, με τις μη συνηθισμένες λέξεις να παρουσιάζονται καλύτερα, ενώ το CBOW είναι πιο γρήγορο και αποδίδει καλύτερα στις επαναλαμβανόμενες λέξεις.

Τα διανύσματα που δημιουργεί το Word2Vec μπορούν να χρησιμοποιηθούν στην ανάλυση συναισθημάτων, στην ταξινόμηση κειμένου αλλά κα στημετάφραση γλώσσας. Το μοντέλο Word2vec μετατρέπει τις λέξεις σε διανύσματα που αποτυπώνουν το νόημά τους. Αυτά τα διανύσματα μπορούν να χρησιμοποιηθούν για την εκπαίδευση μοντέλων ανάλυσης συναισθήματος, την εκπαίδευση μοντέλων ταξινόμησης κειμένου και την εκπαίδευση μοντέλων μετάφρασης γλώσσας αντίστοιχα.

Glove

Το Word2vec μπορεί να αποτυπώνει καλύτερα τη έννοια των λέξεων αλλά περιορίζεται στο παράθυρο των γειτονικών λέξεων χάνοντας τις πληροφορίες παγκόσμιας συχνότητας των λέξεων σε ένα μεγάλο σώμα. Αυτό το κενό έρχεται να καλύψει το Glove (Global Vectors for Word Representation) το οποίο προτάθηκε το 2014 στο πανεπιστήμιο Stanford, από τους Jeffrey Pennington, Richard Socher και Christopher D. Manning στην εργασία "GloVe: Global Vectors for Word Representation" [66]. Στο άρθρο αυτό, γίνεται σύγκριση της απόδοσης του GloVe με άλλες δημοφιλείς μεθόδους αναπαράστασης λέξεων, μία από αυτές ήταν και το Word2vec. Τα αποτελέσματα αποτυπώνουν την άποψη ότι το GloVe ξεπερνά τις άλλες μεθόδους κυρίως σε εργασίες σημασιολογικής συνάφειας.

Η εκπαίδευση του αλγορίθμου GLOVE εργάζεται πάνω σε συγκεντρωτικά παγκόσμια στατιστικά στοιχεία συν-εμφάνισης λέξεων από ένα corpus με καθολική παραγοντοποίηση πίνακα, και όχι σε συγκεκριμένο corpus κειμένων όπως το Word2vec. Βάση ενός πίνακα συν-εμφάνισης λέξεων σε ένα corpus, μαθαίνει τα διανύσματα λέξεων παραγοντοποιώντας αυτόν τον πίνακα σε δύο πίνακες χαμηλής κατάταξης, που αντιπροσωπεύουν αντίστοιχα τα διανύσματα λέξεων και τα διανύσματα περιβάλλοντος. Η

αντικειμενική λειτουργία του, είναι να ελαχιστοποιήσει τη διαφορά μεταξύ των διανυσμάτων λέξης και περιβάλλοντος και το εσωτερικό γινόμενό τους θα πρέπει να ισούται με τον λογάριθμο πιθανότητας συν-εμφάνισης των λέξεων. Με αυτή του την λειτουργία επιτυγχάνει να συλλάβει πληροφορίες συνεμφάνισης λέξεων και τις πληροφορίες συχνότητας λέξεων ταυτόχρονα. Η αντικειμενική συνάρτηση του είναι η εξής:

$$J = \sum_{k,j=1}^{v} f(X_{kj})(w_k^T w_j^{'} + b_k + b_j) - logX_{kj}$$

Όπου, V είναι το μέγεθος του λεξιλογίου, X είναι η μήτρα συν-εμφάνισης, X_{kj} η συχνότητα συνύαπαρξης της λέξης k με την λέξη j, X_k ο συνολικός αριθμός εμφάνισης της λέξης k στο corpus, P_{jk} η πιθανότητα εμφάνισης της λέξης j στο πλαίσιο της λέξης k, w είναι μία ενσωμάτωση της λέξης σε d διάσταση, w' είναι το πλαίσιο ενσωμάτωσης της λέξης σε d διάσταση.

Ένα από τα βασικά πλεονεκτήματα του GloVe είναι η ικανότητά του να διαχειρίζεται αποτελεσματικά τις λέξεις εκτός λεξιλογίου (OOV). Η ιδέα πίσω από το GloVe είναι να μάθουμε τα διανύσματα λέξεων λαμβάνοντας υπόψη τόσο τις τοπικές όσο και τις καθολικές πληροφορίες των λέξεων, κάτι που του επιτρέπει να καταγράφει τις σημασιολογικές σχέσεις μεταξύ των λέξεων. Η χρήση του GloVe και άλλων μεθόδων ενσωμάτωσης λέξεων έχει προχωρήσει πολύ το πεδίο του NLP και έχει επιτρέψει στα μοντέλα μηχανικής μάθησης να κάνουν καλύτερες προβλέψεις.

FastText

Το FastText αποτελεί μία ανοιχτού κώδικα βιβλιοθήκη που δίνει το δικαίωμα στους χρήστες να εκτελούν αναπαραστάσεις κειμένου και εργασίες ταξινόμησης κειμένου. Το 2016 ο Bojanowski πρότεινε το FastText το οποίο βασίζεται στη CBOW τεχνική και πρακτικά χρησιμοποιεί υπολέξεις για να διαχειριστεί περιπτώσεις λέξεων εκτός του λεξιλογίου [64]. Πρωταρχικός του στόχος είναι αποτελεσματική εκμάθηση και ταξινόμηση δεδομένων κειμένου, συμπεριλαμβανομένων κειμένων που αποτελούνται από πολλές

λέξεις ή φράσεις.

Η λειτουργία του. Αρχικά αντί να χρησιμοποιήσει ολόκληρη τη λέξη τη διασπά σε n-grams και έπειτα την παρέχει σε ένα νευρωνικό δίκτυο, το οποίο αποσπά το νόημα μεμονωμένων λέξεων, καθώς και τις σχέσεις μεταξύ λέξεων σε μια πρόταση. Το FastText μειώνει τον χρόνο εκπαίδευσης, διατηρεί την απόδοση και δίνει καλύτερα αποτελέσματα έχοντας καλύτερη αναπαράσταση κυρίως των σπάνιων λέξεων [67].

Το FastText λαμβάνει υπόψη τόσο τις λέξεις όσο και τις δευτερεύουσες λέξεις μέσα σε ένα κείμενο. Αυτό επιτρέπει στο FastText να χειρίζεται πιο αποτελεσματικά λέξεις εκτός λεξιλογίου με ορθογραφικά λάθη, τα οποία βρίσκονται συχνά σε κείμενα των μέσων κοινωνικής δικτύωσης. Ένα μεγάλο πλεονέκτημα της ενσωμάτωσης αυτής είναι ότι έχει τη δυνατότητα να εφαρμοστεί σε 294 διαφορετικές γλώσσες όπως και σε μια ποικιλία εργασιών ταξινόμησης κειμένου, συμπεριλαμβανομένης της ανάλυσης συναισθήματος, της ταξινόμησης θεμάτων και του φιλτραρίσματος ανεπιθύμητων μηνυμάτων. Σε μια σύγκριση του FastText με τις παραδοσιακές ενσωματώσεις λέξεων, στη μελέτη που διεξήγαγαν ο Joulin et al. αποδείχθηκε ότι το FastText αποδίδει καλύτερα σε εργασίες ταξινόμησης κειμένου ιδιαίτερα για μεγάλα σύνολα δεδομένων, με ακρίβεια παρόμοια με αυτή των βαθιών νευρωνικών δικτύων.

ELMO

Το ELMO (embeddings from language models) παρουσιάστηκε στην εργασία "Deep Contextualized Word Representations" από τον Peters et al. το 2018 και αποτελεί μια πιο ουσιώδη μέθοδο αναπαράστασης λέξεων με βάση τα συμφραζόμενα [68]. Είναι ένα μοντέλο προεκπαιδευμένο σε ένα μεγάλου μεγέθους corpus και τα διανύσματα ενσωμάτωσης λέξεων που δημιουργεί, εξαρτάται από το γλωσσικό πλαίσιο που τις απαρτίζει.

Τα τελικά διανύσματα των λέξεων μαθαίνονται από ένα μοντέλο γλώσσας αμφίδρομης κατεύθυνσης LM, με αυτόν τον τρόπο καταφέρνει να μεταβιβάσει στην τρέχουσα

λέξη το περιεχόμενο της προηγούμενης και της επόμενης λέξης από αυτήν. Το ELMO παρέχει διαφορετικές αναπαραστάσεις λέξεων για την ίδια λέξη, όταν αυτή βρίσκεται σε διαφορετικές προτάσεις. Οι αναπαραστάσεις λέξεων αυτές μαθαίνονται από το δίγλωσσο μοντέλο BiLM. Η τελική αναπαράσταση είναι το σταθμισμένο άθροισμα από τα τρέχοντα διανύσματα λέξης με τα διανύσματα της προηγούμενης και της επόμενης λέξης από αυτήν. Πιο συνοπτικά το ELMO εξάγει, σε μια downstream εργασία, τις αναπαραστάσεις που μαθαίνει από το BiLM και εκτελεί έναν γραμμικό συνδυασμό για κάθε λέξη.

Τέτοιου είδους ενσωματώσεις καταφέρνουν να αποτυπώσουν τόσο τη σύνταξη όσο και τη σημασιολογία των λέξεων ενός πλαισίου, γεγονός που τους επιτρέπει να παρέχουν πιο λεπτομερή αναπαράσταση λέξεων σε σχέση με τις παραδοσιακές ενσωματώσεις λέξεων όπως το Word2Vec και το GloVe. Επίσης είναι ευαίσθητες στο πλαίσιο, που σημαίνει ότι η αναπαράσταση μιας λέξης αλλάζει με βάση το πλαίσιο στο οποίο εμφανίζεται. Παρακάτω παρουσιάζεται ένας συγκεντρωτικός πίνακας για τα κλασικά μοντέλα.

ΓΛΩΣΣΙΚΑ ΜΟΝΤΕΛΑ	ΣΗΜΑΣΙΟΛΟΓΙΑ	ΣΥΝΤΑΚΤΙΚΟ	ΠΛΑΙΣΙΟ	ΕΚΤΟΣ ΛΕΞΙΛΟΓΙΟΥ
One Hot encoding	X	X	X	X
Bag of Words	X	X	X	X
TF - IDF	X	X	X	X
Word2vec	✓	✓	X	X
GLOVE	✓	✓	X	X
FastText	✓	✓	X	✓
ELMO	✓	✓	✓	✓

Σχήμα 1: Κλασικά μοντέλα Αναπαράστασης Λέξεων

BERT

BERT σημαίνει αναπαραστάσεις αμφίδρομου κωδικοποιητή από μετασχηματιστές. Όπως και η ELMO και αυτή είναι μια μέθοδος αναπαράστασης λέξεων LM που βασίζεται στα συμφραζόμενα, με τη διαφορά ότι εδώ υπάρχει ο μετασχηματιστής NN ο οποίος χρησιμοποιεί παράλληλα επίπεδα προσοχής αντί να κάνει διαδοχικές επαναλήψεις. Πρόκειται για την εκπαίδευση ενός τεράστιου LM σε ένα ελεύθερο κείμενο, που τελειοποιεί μεμονωμένα τις εργασίες του.

Η καινοτομία αυτού του μοντέλου είναι ότι εκπαιδεύεται σε δύο μη-εποπτευόμενες εργασίες, για να δημιουργήσει πρόβλεψη διπλής κατεύθυνσης, που σημαίνει ότι λαμβάνει υπόψη τόσο το αριστερό όσο και το δεξί πλαίσιο κάθε λέξης σε μια πρόταση, για να συλλέξει το νόημα των προτάσεων. Η πρώτη εργασία που χρησιμοποιεί είναι ένα μοντέλο καλυμμένης γλώσσας MLM το οποίο αντικαθιστά αυθαίρετα το 15% των λέξεων με τη μεταβλητή "MASK" και το μοντέλο εκπαιδεύεται για να προβλέψει τις λέξεις που θα βρίσκονταν σε κάθε σημείο "MASK". Στη δεύτερη εργασία χρησιμοποιεί το μοντέλο NSP που επιδιώκει την πρόβλεψη της επόμενης πρότασης. Παρέχεται έτσι ένα ζεύγος προτάσεων και το μοντέλο εκπαιδεύεται για να αναγνωρίσει πότε η δεύτερη πρόταση ακολουθεί την πρώτη. Με τη δεύτερη εργασία συλλέγονται ρεαλιστικές πληροφορίες.

Το BERT εκπαιδεύεται εκ των προτέρων σε ένα μεγάλο σύνολο δεδομένων κειμένου και στη συνέχεια προσδιορίζονται συγκεκριμένα οι εργασίες που πρέπει να εκτελέσει. Αυτή η προεκπαίδευση επιτρέπει στον BERT να συλλέξει ένα μεγάλο όγκο γλωσσικών γνώσεων, καθιστώντας τον χρήσιμο για πολλές και διαφορετικές εργασίες NLP, όπως η ανάλυση συναισθήματος, η ταξινόμηση κειμένου, η αναγνώριση ονομαστικών οντοτήτων και η απάντηση ερωτήσεων. Αυτός είναι και ο λόγος που πολλά από τα μεταγενέστερα μοντέλα που αναπτύχθηκαν αποτελούν επέκταση του [69, 70, 71].

3.2 Machine Learning Models

Ο όρος αλγόριθμος μηχανική μάθηση εκφράζει έναν κώδικα προγράμματος που έχει ως βασική του εργασία της διεξαγωγή μελέτης, ανάλυσης και κατανόησης μεγάλων σύνθετων συνόλων δεδομένων. Καθένας από αυτούς υπόκειται σε προκαθορισμένη ροή ενεργειών για να πραγματοποιήσει προβλέψεις, να κατηγοριοποιήσει πληροφορίες μέσω της εκμάθησης και να ανακαλύψει μοτίβα που είναι ενσωματωμένα στα δεδομένα.

Αυτοί οι αλγόριθμοι αναλύουν και προσομοιώνουν δεδομένα για να προβλέψουν το αποτέλεσμα εντός ενός προκαθορισμένου εύρους. Επίσης είναι αρμόδιοι να καθορίσουν τους κανόνες και τις διαδικασίες που πρέπει να ακολουθήσει ένα σύστημα για την αντιμετώπιση του κάθε προβλήματος που μπορεί να δημιουργηθεί. Τέλος τείνουν να γίνονται ¨ευφυέστεροι¨ κάθε φορά που νέα δεδομένα τροφοδοτούνται σε αυτούς τους αλγορίθμους καθώς αυτοί μαθαίνουν, βελτιστοποιούνται και βελτιώνονται με βάση την ανατροφοδότηση σχετικά με τις προηγούμενες επιδόσεις στην πρόβλεψη των αποτελεσμάτων.

Στην ουσία, για τη διεξαγωγή αυτής της διαδικασίας πρέπει να υπάρχουν δυο τύποι δεδομένων, τα δεδομένα εκπαίδευση (training data) και τα δεδομένα δοκιμής (testing data). Αρχικά χρησιμοποιούνται από ένα αυτόματο ταξινομητή (classifier) τα δεδομένα εκπαίδευση, με σκοπό να αναγνωρίσει τα διαφορετικά χαρακτηριστικά των δεδομένων και να δημιουργήσει τα μοντέλα δεδομένων. Η διαδικασία αυτή ονομάζεται εκπαίδευση (training). Στη συνέχεια χρησιμοποιούνται τα δεδομένα δοκιμής τα οποία είναι δεδομένα που δεν έχουν συμμετάσχει στη διαδικασία εκπαίδευσης, για να αξιολογηθεί η ακρίβεια και η αποδοτικότητά του συστήματος.

Υπάρχουν τέσσερεις τύποι μηχανικής μάθησης. Ο διαχωρισμός τους γίνεται βάση της τεχνικής τους: μάθηση με επίβλεψη, μάθηση με ημιεπίβλεψη, μάθηση χωρίς επίβλεψη και μάθηση με ενίσχυση. Οι αλγόριθμοι ταξινόμησης και παλινδρόμησης είναι οι πιο συνήθεις επιλογές για την πρόβλεψη τιμών, τον εντοπισμό ομοιοτήτων και την εύρεση μοτίβων που μπορεί να παρουσιάζουν τα δεδομένα.

Στη μάθηση με επίβλεψη (Supervised learning) για να γίνουν προβλέψεις χρησιμοποιούνται σύνολα δεδομένων με ετικέτες. Αυτή η τεχνική είναι πιο χρήσιμη όταν είναι γνωστός σκοπός ή το είδος του αποτελέσματος που θέλουμε να πετύχουμε.

Στη μάθηση χωρίς επίβλεψη (Unsupervised learning) χρησιμοποιούνται δεδομένα χωρίς ετικέτες. Αυτή η τεχνική μάθησης εντοπίζει τα μη επισημασμένα δεδομένα και τα κατηγοριοποιεί με βάση τον τύπο, τη μορφή ή τη δομή τους. Αυτή η τεχνική είναι ωφέλιμη όταν ο τύπος του αποτελέσματος είναι άγνωστος.

Στη μάθηση με ημιεπίβλεψη (Semisupervised learning (SSL)) συνδυάζονται οι δύο παραπάνω τύποι. Χρησιμοποιούν δεδομένα με ή χωρίς ετικέτες. Στόχος τους είναι η κατηγοριοποίηση των μη επισημασμένων δεδομένων με βάση τις πληροφορίες που αντλούν από τα επισημασμένα δεδομένα.

Στην ενισχυτική μάθηση (Reinforcement learning) χρησιμοποιούνται τα αποτελέσματα ως σημείο αφετηρίας για να προγραμματιστεί το επόμενο βήμα επεξεργασίας. Με άλλα λόγια, αυτοί οι αλγόριθμοι μαθαίνουν από προηγούμενα αποτελέσματα, λαμβάνουν ανατροφοδότηση μετά από κάθε βήμα και στη συνέχεια αποφασίζουν αν θα προχωρήσουν στο επόμενο βήμα ή όχι. Αυτός ο τύπος μάθησης είναι σχεδιασμένος να λαμβάνει αποφάσεις με ελάχιστη ανθρώπινη παρέμβαση και γι αυτόν ακριβώς τον λόγο μπορεί εύκολα να χρησιμοποιηθεί από αυτοματοποιημένα συστήματα.

Κάθε ένας αλγόριθμος μηχανικής μάθησης διαχειρίζεται ένα μεμονωμένο πρόβλημα. Στην ανάλυση συναισθήματος ως επί το πλείστον χρησιμοποιείται η μάθηση με επίβλεψη, αλλά σε ορισμένες περιπτώσεις, οι data analysts επιλέγουν έναν συνδυασμό αυτών των αλγορίθμων, για να λύσουν ένα πρόβλημα το οποίο δεν μπορεί ένας συγκεκριμένος τύπος αλγόριθμου να επιλύσει. Στην συνέχεια παρουσιάζονται οι αλγόριθμοι ταξινόμησης κειμένου που χρησιμοποιήθηκαν στην παρούσα εργασία καθώς και τα πλεονεκτήματα και τα μειονεκτήματά τους.

3.2.1 Logistic Regression

Είναι μια από τις παλιότερες τεχνικές ταξινόμησης που χρησιμοποιήθηκαν. Αυτός ο τύπος ανάλυσης περιγράφει τη μοντελοποίηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Τα δεδομένα προβλέπουν μια πιθανότητα ύπαρξης ενός συμβάντος με μια συνάρτηση logit. Η μαθηματική αναπαράσταση της είναι η εξής :

$$y = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}$$

x = τιμή εισόδου, y = προβλεπόμενη έξοδος, b_0 = όρος μεροληψίας ή διακοπής, b_1 = συντελεστής για την είσοδο x

Είναι ένα στατιστικό μοντέλο και χρησιμοποιείται για προβλήματα δυαδικής ταξινόμησης, με στόχο να προβλεφθεί η πιθανότητα να συμβεί ένα γεγονός, δεδομένου ενός συνόλου προγνωστικών. Λειτουργεί καλύτερα στην περίπτωση των κατηγορικών αποτελεσμάτων, αλλά αποτυγχάνει στην περίπτωση των συνεχών αποτελεσμάτων. Επίσης, απαιτεί από κάθε δεδομένο να είναι ανεξάρτητο από όλα τα άλλα και αν τα ευρήματα είναι αλληλένδετα μεταξύ τους, τότε ο ταξινομητής δεν θα διεξάγει καλές προβλέψεις [72, 73] .

Το πλεονέκτημά της είναι ότι είναι απλή, την καθιστά εύκολη στην κατανόηση και στην εφαρμογή αλλά αποτελεί και μία εύκολα προσαρμόσιμη μέθοδο. Μπορεί να χειριστεί διάφορους τύπους δεδομένων, όπως συνεχών, δυαδικών και κατηγορικών και είναι κατάλληλη για μεγάλα σύνολα δεδομένων [74] . Επίσης τα νέα δεδομένα μπορούν να ενσωματωθούν στο μοντέλο χωρίς να χρειαστεί να επανεκπαιδευτεί ολόκληρο το μοντέλο από την αρχή. Υπάρχουν όμως και ορισμένοι περιορισμοί. Ένας από αυτούς είναι ότι η σχέση μεταξύ των προβλέψεων και της εξαρτημένης μεταβλητής πρέπει να είναι γραμμική. Στην πράξη, αυτή η υπόθεση συχνά δεν τηρείται και το μοντέλο μπορεί να παρέχει ψευδή αναπαράσταση της σχέσης μεταξύ των προγνωστικών παραγόντων και της εξαρτημένης μεταβλητής [75] .

3.2.2 Random Forest

Πρόκειται για έναν αλγόριθμο μηχανικής μάθησης με επίβλεψη, όπου διαφορετικά δέντρα αποφάσεων κατασκευάζονται σε διαφορετικά δείγματα κατά την εκπαίδευση, για να χειριστούν προβλήματα ταξινόμησης και παλινδρόμησης [76] . Ο αλγόριθμος Random Forest (RF) επικεντρώνεται σε μεθόδους σύγκρισης των αποτελεσμάτων πολλών εκπαιδευμένων μοντέλων στη σειρά για να δώσει τον βέλτιστο ταξινομητή και την καλύτερη απόδοση από ένα μεμονωμένο μοντέλο. Η Ho το 1998 πρότεινε τον ταξινομητή RF, που αποτελεί μια επέκταση του αλγορίθμου του Decision trees, στον οποίο κατασκευάζονται και συνδυάζονται πολλαπλά δέντρα για να σχηματίσουν ένα ενιαίο μοντέλο [77] . Ο ταξινομητής RF αποτελείται από τον αριθμό των ταξινομητών DT, όπου κάθε δέντρο εκπαιδεύεται από ένα υποσύνολο του κειμένου. Στην ουσία πρόκειται για ένα αυθαίρετο υποσύνολο των χαρακτηριστικών, που επιλέγεται σε κάθε κόμβο απόφασης και το μοντέλο θα εξετάσει μόνο ένα μέρος αυτών των χαρακτηριστικών. Στη συνέχεια επαναλαμβάνει τη διαδικασία πολλές φορές με διαφορετικά υποσύνολα χαρακτηριστικών, για να δημιουργήσει ένα δάσος δέντρων. Οι αλγόριθμοι Random Forest ακολουθούν τα εξής βήματα:

Αρχικά επιλέγονται αμερόληπτα δείγματα δεδομένων από το σύνολο των δεδομένων που δίνεται. Στο επόμενο βήμα δημιουργείται ένα δέντρο απόφασης για κάθε δείγμα δεδομένων και εκτυπώνονται τα αποτέλεσμα πρόβλεψης για κάθε ένα δέντρο απόφασης. Η διαδικασία αυτή επαναλαμβάνεται για να δημιουργηθεί ένα δάσος δέντρων. Έπειτα διεξάγεται ψηφοφορία για κάθε αναμενόμενο αποτέλεσμα και τέλος επιλέγεται το τελικό αποτέλεσμα πρόβλεψης με βάση το υψηλότερο αποτέλεσμα πρόβλεψης που έχει ψηφιστεί.

Βασική αρχή του αλγόριθμου αυτού είναι ότι έχει μαζική διακύμανση, με αποτέλεσμα η διάταξη των δεδομένων εκπαίδευσης και των χαρακτηριστικών να μπορεί να επηρεάσει τα τελικά αποτελέσματά του. Αυτοί οι αλγόριθμοι βοηθούν στην εκτίμηση των δεδομένων που λείπουν. Οι Random Forest έχουν πολλά πλεονεκτήματα σε σχέση με τα παραδοσιακά Decision Trees, όπως βελτιωμένη ακρίβεια, μειωμένη υπερπροσαρ-

μογή και αυξημένη σταθερότητα. Έχει καλές επιδόσεις τόσο με κατηγορικές όσο και με συνεχείς μεταβλητές, καθιστώντας το ως ένα ευέλικτο εργαλείο για πολλούς τύπους εργασιών ανάλυσης δεδομένων. Επίσης μπορεί να χειριστεί αυτόματα τις τιμές που λείπουν, είναι ανθεκτικός στις ακραίες τιμές και επηρεάζεται λιγότερο από το θόρυβο. Όσον αφορά την ακρίβεια, τα Random Forest έχει αποδειχθεί ότι ξεπερνούν τα Decision Trees σε πολλές εφαρμογές. Ένα από τα βασικά του πλεονεκτήματα είναι ότι ελαχιστοποιεί τον κίνδυνο υπερπροσαρμογής, ένα πρόβλημα πολύ κοινό στους αλγόριθμους Decision Trees. Αυτοί οι ταξινομητές εκπαιδεύονται γρήγορα για δεδομένα κειμένου, αλλά είναι αργοί στο να δίνουν προβλέψεις μετά την εκπαίδευση [78] .

Επιπλέον η εκπαίδευση ενός τεράστιου αριθμού δέντρων μπορεί να είναι υπολογιστικά δαπανηρή, να απαιτεί περισσότερο χρόνο εκπαίδευσης και χρήση μεγάλου μέρους της μνήμης. Ένα ακόμα μειονέκτημά τους είναι ότι είναι ακατάλληλα για δεδομένα υψηλών διαστάσεων, καθώς ο αριθμός των χαρακτηριστικών μπορεί να γίνει τεράστιος και να επιφέρει κακή απόδοση.

3.2.3 Support Vector Machine - SVM

Αυτός ο τύπος αλγορίθμου μηχανικής μάθησης έχει την δεξιότητα να χειρίζεται πολύπλοκα και υψηλών διαστάσεων δεδομένα, χρησιμοποιείται κυρίως για ταξινόμηση και ανάλυση παλινδρόμησης. Προτάθηκε το 1963 από τους Vapnik και Chervonenkis και αποτελεί ένα γραμμικό μοντέλο με δυνατότητα επίλυσης γραμμικών αλλά και μη γραμμικών προβλημάτων [79] .

Βασική αρχή του είναι να χωρίσει τα δεδομένα σε δύο κλάσεις μεταξύ του υπερεπίπεδου. Ο SVM είναι ένας αλγόριθμος που λαμβάνει ως είσοδο τα δεδομένα και εκτυπώνει μια γραμμή που χωρίζει τα δεδομένα σε δύο κλάσεις. Το υπερεπίπεδο βασίζεται στην αρχή του μέγιστου περιθωρίου [80] . Η αρχή αυτή έχει τη δυνατότητα να μεγιστοποιεί την απόσταση μεταξύ των πλησιέστερων σημείων κάθε κατηγορίας με το υπερεπίπεδο. Αυτά τα σημεία ονομάζονται διανύσματα υποστήριξης. Ως περιθώριο ο-

ρίζεται ο χώρος μεταξύ του υπερεπίπεδου και των διανυσμάτων υποστήριξης. Βασικός στόχος επίτευξης του SVM είναι η μεγιστοποίηση του περιθωρίου, για την εύρεση του βέλτιστου υπερεπίπεδου.

Στην πραγματικότητα για την εύρεση του βέλτιστου υπερεπίπεδου ανάμεσα σε εκατομμύρια σύνολα δεδομένων απαιτεί πολύ χρόνο. Για την επίλυση αυτού του ζητήματος έχουν οριστεί κάποιοι παράμετροι κανονικοποίησης και το Gamma, που ρυθμίζουν τον ταξινομητή SVM. Η ρύθμιση των παραμέτρων αυτών συμβάλει στην διεξαγωγή εύρεση μιας μη γραμμικής γραμμής υπερεπιπέδου με μεγαλύτερη ακρίβεια και σε σύντομο χρονικό διάστημα [81] . Οι παράμετροι αυτοί είναι οι εξής:

Kernel (Πυρήνας)

Η παράμετρος αυτή μετασχηματίζει το πρόβλημα χρησιμοποιώντας γραμμική άλγεβρα για την εύρεση του υπερεπιπέδου. Πρακτικά η παράμετρος χαρτογραφεί τα δεδομένα εκπαίδευσης σε έναν χώρο υψηλότερων διαστάσεων ώστε να μπορέσει να βρει ένα γραμμικό όριο. Αυτό ονομάζεται Kernel trick.

Regularization (Τακτοποίηση)

Η παράμετρος αυτή, η οποία συμβολίζεται με C στη βιβλιοθήκη sklearn της Python, προσδιορίζει το ποσοστό αποφυγής σφάλματος για την εκπαίδευση κάθε ταξινόμησης. Ένα C με μεγάλη τιμή θα αποδώσει ένα υπερεπίπεδο μικρότερου περιθωρίου, με προϋπόθεση ότι το υπερεπίπεδο αυτό θα ταξινομεί σωστά όλα τα σημεία εκπαίδευσης. Αντίθετα, μια πολύ μικρή τιμή του C θα οδηγήσει τον βελτιστοποιητή σε εύρεση υπερεπιπέδου με μεγαλύτερο περιθώριο, ακόμη και αν αυτό το υπερεπίπεδο ταξινομεί εσφαλμένα τα περισσότερα σημεία.

Gamma

Η παράμετρος αυτή προσδιορίζει την απόσταση των σημείων κάθε μεμονωμένου παραδείγματος εκπαίδευσης από το υπερεπίπεδο. Για ένα χαμηλό gamma, τα σημεία βρίσκονται πολύ μακριά από την γραμμή διαχωρισμού. Ενώ υψηλό gamma σημαίνει ότι τα σημεία βρίσκονται κοντά στο πιθανό υπερεπίπεδο. Το gamma λαμβάνεται υπόψη στον υπολογισμό του υπερεπίπεδου.

Ο αλγόριθμος SVM είναι χρήσιμος σε διάφορους τομείς όπως η ταξινόμηση κειμένου, η ταξινόμηση εικόνων και η βιοπληροφορική. Ένα από τα δυνατά σημεία του SVM είναι η ικανότητά του να χειρίζεται γραμμικά και μη γραμμικά προβλήματα, καθώς και μεγάλα σύνολα δεδομένων με υψηλή πολυπλοκότητα. Ο μεγαλύτερος περιορισμός του παρόλα αυτά είναι η ευαισθησία του στην επιλογή του πυρήνα και στην επιλογή των παραμέτρων του.

3.2.4 Naive Bayes

Οι Naive Bayes ταξινομητές είναι ένα σύνολο αλγορίθμων μηχανικής μάθησης και χρησιμοποιούνται για την αντιμετώπιση προβλημάτων ταξινόμησης κειμένου και την ανάκτηση πληροφοριών [74] . Η θεμελιώδης υπόθεση των αλγορίθμων αυτών είναι ότι τα χαρακτηριστικά που εξετάζονται είναι ανεξάρτητα μεταξύ τους και ότι μια αλλαγή στην τιμή του ενός δεν επηρεάζει την τιμή του άλλου. Εάν για παράδειγμα μια λέξη αλλάξει θέση μέσα σε μια πρόταση τότε μπορεί να αλλάξει η έννοιά της ή και το συναίσθημα και το οποίο είναι φορτωμένη [29] . Παρακάτω παρουσιάζεται η μαθηματική αναπαράσταση του αλγορίθμου που βασίζεται στο θεώρημα πιθανοτήτων του Bayes και υπολογίζει την πιθανότητα ενός γεγονότος με βάση την προηγούμενη γνώση.

$$P(X|Y) = (P(Y|X) * P(X)) \ /P(Y)$$

αν X, Y είναι πιθανοτικά γεγονότα, $P(X)$ η πιθανότητα να είναι αληθές το X, $P(X|Y)$ η υπό συνθήκη πιθανότητα να είναι αληθές το X σε περίπτωση που είναι αληθές το Y [82] .

Στόχος του αλγόριθμου αυτού είναι να προβλέψει την πιθανότητα ένα δείγμα δεδομένων να ανήκει σε μια συγκεκριμένη κατηγορία με βάση τα στατιστικά χαρακτηριστικά του δείγματος. Αρχικά υπολογίζει την πιθανότητα ταξινόμησης κάθε μιας λέξης από τα δεδομένα εκπαίδευσης και έπειτα ταξινομεί τα νέα δεδομένα με βάση τις πιθανότητες που υπολόγισε από τα δεδομένα εκπαίδευσης [83] .

Αυτού του είδους αλγόριθμοι αποδεικνύονται πολύ χρήσιμοι για την πραγματοποίηση προβλέψεων σε πραγματικό χρόνο. Επίσης είναι ιδιαίτερα χρήσιμοι σε περιπτώσεις όπου τα δεδομένα είναι μεγάλα και οι δυνατότητες είναι πολλές, αλλά λειτουργεί καλύτερα όταν τα training data είναι λιγότερα. Τέλος έχει χαμηλότερο υπολογιστικό κόστος από άλλους.

Παρά την απλότητα και την αποτελεσματικότητά του, ο αλγόριθμος Naive Bayes έχει ορισμένους περιορισμούς. Ένα βασικό του μειονέκτημα είναι και η θεμελιώδης αρχή ανεξαρτησίας, που υποθέτει για τα δεδομένα. Σε ένα πραγματικό σύνολο δεδομένων σπάνια συναντάται το χαρακτηριστικό αυτό. Αυτή η υπόθεση συχνά οδηγεί σε υποεκτίμηση της αλληλεξάρτησης μεταξύ των χαρακτηριστικών και οδηγεί σε μείωση της ακρίβειας του αλγόριθμου. Επιπλέον, το Naive Bayes μπορεί να έχει κακή απόδοση όταν ο αριθμός των χαρακτηριστικών είναι πολύ μεγαλύτερος από τον αριθμό των δειγμάτων, καθώς ο αλγόριθμος μπορεί να υπεργενικεύσει τα δεδομένα.

3.2.5 K-Nearest Neighbours

Ο KNN αλγόριθμος είναι αλγόριθμος εποπτευόμενης μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί για την επίλυση προβλημάτων τόσο ταξινόμησης όσο και παλινδρόμησης [74] . Υποθέτει ότι παρόμοια δεδομένα υπάρχουν σε κοντινή απόσταση

μεταξύ τους, εξού και η ονομασία του αλγορίθμου, και βασίζεται στο ότι αυτή η υπόθεση είναι αρκετά αληθής ώστε ο αλγόριθμος να είναι χρήσιμος. Στην ουσία το KNN αποτυπώνει την ιδέα της ομοιότητας με τον υπολογισμό της απόστασης μεταξύ σημείων των δεδομένων.

Το αρχικό "K" αναφέρεται στον αριθμό των γειτονικών δεδομένων για ένα συγκεκριμένο σημείο των δεδομένων, χρησιμοποιώντας τις ετικέτες κλάσεων των γειτόνων αυτών προσπαθεί να κάνει προγνώσεις. Η λειτουργία του, αρχικά αποθηκεύει όλες τις γνωστές περιπτώσεις απόστασης από τα training data και με βάση τις αποστάσεις αυτές ταξινομεί τις νέες περιπτώσεις απόστασης με αύξουσα σειρά, διαχωρίζοντάς τες σε διαφορετικές κλάσεις [48] . Η ταξινόμηση αυτή επιτυγχάνεται με βάση της ομοιότητας των πρόσφατων περιπτώσεων παρατήρησης με τις ήδη υπάρχουσες. Ο αλγόριθμος εκτυπώνει το μέσο όρο των K εάν η εργασία που πρόκειται να εκτελέσει είναι εργασία παλινδρόμησης, ενώ αν πρόκειται για εργασία ταξινόμησης, επιστρέφει τον τρόπο λειτουργίας των ετικετών K.

Ο KNN είναι ένας απλός και εύκολος στην εφαρμογή αλγόριθμος και χρήζει ελάχιστων δεδομένων κατά την εκπαίδευση. Επίσης είναι αρκετά ευέλικτος σε σχέση με άλλους αλγόριθμους καθώς δύναται να χειριστεί μη γραμμικές σχέσεις και αποδέχεται την υποκειμενική κατανομή των δεδομένων, χωρίς περιορισμούς και υποθέσεις [84] .

Ένα από τα σημαντικά μειονεκτήματά του είναι ότι ελλατώνεται η ταχύτητά του αισθητά καθώς αυξάνεται το μέγεθος των δεδομένων που χρησιμοποιούνται. Επίσης καθώς αυξάνονται τα δεδομένα αυξάνεται υπολογιστικά και το κόστος, που αποτελεί απόρροια του υπολογισμού της απόστασης μεταξύ των δεδομένων και της εύρεσης των K πλησιέστερων γειτόνων. Τέλος ο αλγόριθμος παρουσιάζει μεγάλη ευαισθησία ως προς την επιλογή του K, αν το K είναι πολύ μεγάλο τότε μπορεί να υποκύψει στο θόρυβο, ενώ αν είναι πολύ μικρό, μπορεί να ταιριάζει υπερβολικά στα δεδομένα.

3.2.6 Extreme Gradient Boosting - XGBoost

Ο ταξινομητής XGBoost είναι η δημοφιλέστερη μέθοδος μηχανικής εκμάθησης λόγω της απόδοσης του σε μια μεγάλη γκάμα εφαρμογών. Το XGBoost, που σημαίνει Extreme Gradient Boosting, βασίζεται στα δέντρα αποφάσεων και χρησιμοποιεί ένα πλαίσιο ενίσχυσης κλίσης. Είναι ένας εξαιρετικά επεκτάσιμος και αποτελεσματικός ταξινομητής που μπορεί να χειριστεί τεράστια σύνολα δεδομένων και ιδιαίτερα σύνθετα προβλήματα.

Παρουσιάστηκε το 2016 στο Πανεπιστήμιο της Ουάσιγκτον από τον Tianqi Chen και τον Carlos Guestrin και αποτελεί έναν συνδυασμό τεχνικών βελτιστοποίησης λογισμικού για την απόδοση καλύτερων αποτελεσμάτων με τη χρήση λιγότερων πόρων στο συντομότερο χρονικό διάστημα [85] . Το XGBoost πρακτικά βελτιώνει το βασικό πλαίσιο ενίσχυσης κλίσης ενσωματώνοντας βελτιστοποίησης συστημάτων και αλγοριθμικών βελτιώσεων. Η βασική του καινοτομία είναι η χρήση ενός αλγορίθμου προσέγγισης κλίσης δεύτερης τάξης, που επιτρέπει την ακριβέστερη μοντελοποίηση πολύπλοκων σχέσεων στα δεδομένα [86] .

Ο XGBoost εφαρμόζεται με επιτυχία σε ταξινόμηση εικόνων, αναγνώριση ομιλίας και επεξεργασία της φυσικής γλώσσας. Άλλα και στην επίλυση περίπλοκων επιχειρηματικών προβλημάτων, όπως ο εντοπισμός απάτης και η βελτιστοποίηση μάρκετινγκ [87] .Επίσης φαίνεται να είναι αποτελεσματικότερος σε σύγκριση με άλλους αλγόριθμους ταξινόμησης όπως ο Random Forest και ο SVM. Βασικό συστατικό της επιτυχίας του είναι η ικανότητά του να χειρίζεται μεγάλο αριθμό λειτουργιών, ελλείπουσες τιμές και διαφορετικούς τύπους δεδομένων, όπως κατηγορικά και αριθμητικά, χωρίς να τα προεπεξεργαστεί [88] . Επιπλέον,παρέχει έναν αριθμό παραμέτρων που μπορούν να βελτιστοποιηθούν για την επίτευξη βέλτιστης απόδοσης, καθιστώντας τον έναν ιδιαίτερα προσαρμόσιμο και ευέλικτο αλγόριθμο [89] .

Παρά την πληθώρα πλεονεκτημάτων, παρουσιάζει και ορισμένους περιορισμούς. Όταν ο αριθμός των δέντρων είναι ιδιαίτερα μεγάλος μπορεί να οδηγήσει σε ένα πολύ

περίπλοκο μοντέλο, με καλή απόδοση στα δεδομένα εκπαίδευσης αλλά κακή απόδοση στα δεδομένα δοκιμής. Για ζυγοστάθμιση του προβλήματος αυτού, είναι σημαντικό να χρησιμοποιηθεί η διασταυρούμενη επικύρωση και η έγκαιρη διακοπή του αλγόριθμου για να βρεθεί ο βέλτιστος αριθμός δέντρων.

4 Ανάπτυξη του Θέματος

4.1 Εξόρυξη δεδομένων

4.1.1 Εξόρυξη δεδομένων απο το Twitter

Όπως αναφέρθηκε παραπάνω, τα δεδομένα που πρόκειται να χρησιμοποιηθούν είναι δεδομένα που συλλέγονται από το Twitter. Δεδομένου ότι το Twitter είναι μια ευρέως χρησιμοποιούμενη πλατφόρμα όπου απλοί άνθρωποι και δημόσια πρόσωπα εκφράζουν τη γνώμη και τις απόψεις τους, είναι ένα εξαιρετικό μέσο για να ληφθούν πολωτικά tweets για συγκεκριμένα θέματα. Στην συγκεκριμένη περίπτωση, υπήρξε μια μεγάλη συζήτηση μεταξύ των χρηστών που είτε ισχυρίστηκαν ότι η κλιματική αλλαγή είναι ανθρωπογενές αποτέλεσμα και πρέπει να αναλάβουν δράση γι' αυτήν (Υποστηρικτές οι οποίοι θα εκφράζονται ως sentiment 1) ή ισχυρίστηκαν ότι η κλιματική αλλαγή δεν είναι ανθρωπογενές αποτέλεσμα (Αρνητές ως sentiment 0). Όποια και αν είναι η αλήθεια, υπάρχουν πολλά δεδομένα σχετικά με την κλιματική αλλαγή σε αντίθετους ¨πόλους¨.

4.1.2 Μεθοδος Web Scraping vs Twitter API

Από τις αρχές του 2022, το Twitter νομιμοποίησε την ελεύθερη χρήση του scrapper από όλα τα άτομα-προγραμματιστές. Αυτή η εξέλιξη ανοίγει το δρόμο σε περισσότερες μεθόδους-βιβλιοθήκες για την εξαγωγή των δεδομένων. Ωστόσο, υπάρχει μεγάλη διαφορά μεταξύ αυτών των μεθόδων. Όπως φαίνεται από τον ίδιο τον ιστότοπο, του Twitter API , χρειάζεται αρχικά έναν λογαριασμό διαπιστευτηρίων, χρειάζεται να είναι κάποιος μέλος της κοινότητας Twitter για να το χρησιμοποιήσει και επιπλέον υπάρχουν και ορισμένοι περιορισμοί χρονικού εύρους. Το web scraping από την άλλη πλευρά παρέχει έναν πιο αποτελεσματικό τρόπο συλλογής πολλών και χρονικά μεταβαλλόμενων δεδομένων. Ωστόσο η χρήση του web scraping θα πρέπει να γίνεται υπό παρακολούθηση επειδή ο ιστότοπος μπορεί ενδεχομένως να αποκλείσει τον προγραμματιστή. Στη συγκεκριμένη διπλωματική εργασία θα χρησιμοποιηθεί η μέθοδος Web scraping και συγκεκριμένα η βιβλιοθήκη snscrape που είναι διαθέσιμη για scraping

ιστούς μέσω υπηρεσιών κοινωνικής δικτύωσης (SNS), καθώς θα πρέπει να αντληθούν πολλά δεδομένα από διαφορετικά χρονικά πλαίσια.

4.1.3 Λογαριασμοί και hashtags

Η αναζήτησή των δεδομένων θα πρέπει να εστιαστεί στην κλιματική αλλαγή, για αυτό τον λόγο θα συλλεγχθούν tweets με λέξεις ή φράσεις που αφορούν την κλιματική αλλαγή (Πίνακας 2). Ιδανικά θα έπρεπε να ελεγχθούν όλα τα tweets που συλλέγονται ένα προς ένα (δύο ή τρία άτομα θα ήταν πολύ καλύτερα) για να τα ταξινομηθούν σωστά σε θετικά ή αρνητικά. Ωστόσο, το χρονικό περιθώριο της διατριβής καθώς επίσης και η έλλειψη προσωπικού δεν επιτρέπουν την πραγματοποίηση αυτού του είδους ελέγχου. Για αυτόν τον λόγο θα γίνει ένας περιορισμός στους λογαριασμούς Twitter. Θα επιλεχθούν λογαριασμοί, οι οποίοι μοιράζονται δημόσια τη γνώμη τους για την κλιματική αλλαγή, δηλαδή την υποστήριξη ή την αποδοκιμασία τους.

Πίνακας 2: Hashtags που χρησιμοποιήθηκαν στην αναζήτηση tweets

Hashtags			
climate	climate action	climate change	climate crisis
global warming	climate emergency	climate hysteria	climate hoax
climate alarmism	climate change alarmism	climate change hoax	climate change hysteria
global warming hysteria	climate change fraud	climate emergency hoax	climate change scam
global warming hoax	global warming alarmism	global warming cult	global warming fraud
global warming scam	global warming nonsense	climate fraud	climate scam
climate cult	climate nonsense	climate change nonsense	climate change cult
COP24	Paris Agreement	climate realist	agw

Έγινε αρχικά μια περιήγηση στο Twitter για τις λέξεις-κλειδιά (Πίνακας 2) και περιορίστηκε η αναζήτησή είτε σε έναν επαληθευμένο λογαριασμό (π.χ. realDonaldTrump) είτε σε έναν λογαριασμό με πολλούς ακόλουθους. Η αναζήτηση η οποία έγινε απέδωσε την παρακάτω λίστα (Πίνακας 3). Όπως είναι φανερό, το μεγαλύτερο μέρος των Αρνητών αποτελείται από συντηρητικούς πολιτικούς ή παρουσιαστές ειδήσεων και ορισμένους επιστήμονες. Από την άλλη πλευρά, οι Υποστηρικτές αποτελούνται κυρίως από οργανώσεις και ακτιβιστές.

Πίνακας 3: Λογαριασμοί απο τους οποίους εξορύχθηκαν tweets

Αρνητές		**Υποστηρικτές**	
HeartlandInst	Δυναμικό Εμπειρογνωμόνων για τη Δημόσια Πολιτική	climateWWF	ΜΚΟ
KurtSchlichter	Δικηγόρος	UNFCCC	ΜΚΟ
BjornLomborg	Συγγραφέας	WWF	ΜΚΟ
JunkScience	Δικηγόρος	GreenpeaceUK	ΜΚΟ
TomFitton	Ακτιβιστής	MichaelEMann	Επιστήμονας
brithume	Δημοσιογράφος	ayanaeliza	Επιστήμονας
CatoInstitute	ΜΚΟ	FrannyArmstrong	Σκηνοθέτρια
DineshDSouza	Συνωμοσιολόγος	pooley_eric	Ακτιβιστής
GovMikeHuckabee	Πολιτικός	jrockstrom	Επιστήμονας
JimInhofe inhofeforsenate InhofePress	Πολιτικός	elonmusk	Επιχειρηματίας
LtGovTimGriffin	Πολιτικός - Δικηγόρος	bruneski	Ακτιβιστής
LanceForman	Πολιτικός	xiyebastida	Ακτιβιστής
Truth_Freedom23	Σελίδα για την πλύση εγκεφάλου και την α-φύπνιση	vict_barrett	Ακτιβιστής
LisaMarieBoothe	Ρεπόρτερ	johnpauljos	Ακτιβιστής
mattwridley	Συγγραφέας Δημοσιογράφος	israhirsi	Ακτιβιστής
MaximeBernier	Πολιτικός	luisamneubauer	Ακτιβιστής
TeamBachmann MicheleBachmann	Πολιτικός	severnsuzuki	Ακτιβιστής
Origsmartassam	Πρώην ποδοσφαιριστής	iraghislain	Επιχειρηματίας
PrisonPlanet	Youtuber	alexandriav2005	Ακτιβιστής
realDonaldTrump POTUS45	Πολιτικός	kehkashanbasu	Ακτιβιστής
joebarto	Πολιτικός	felixfinkbeiner	Ακτιβιστής
RonPaul	Πολιτικός	jeromefosterii	Ακτιβιστής
SeibtNaomi	Πολιτικός	laurajoparker	Συγγραφεας
MelissaAFrancis	Τηλεπαρουσιάστρια	jamie_margolin	Καλλιτέχνης
AGToddRokita ToddRokitaIN ToddRokita	Δικηγόρος	fisher_danar	Ακτιβιστής

4.1.4 Αποτελέσματα από το snscrape

Αφού εγκατασταθεί η βιβλιοθήκη snscrape στον κώδικα, θα εκτελεστεί η ακόλουθη εντολή:

snscrape - jsonl -max-results 1000000 twitter-search ' phrase-key from: account'> tweets.json

Η εντολή αυτή υποδηλώνει ότι θα χρησιμοποιηθεί η βιβλιοθήκη snscrape, και ο τύπος του αρχείου που θα ληφθεί θα είναι ένα αρχείο json, που στη δεδομένη περίπτωση ονομάζεται «tweets.json». Αναζητήθηκαν έως και 1.000.000 tweets που να περιέχουν τη συγκεκριμένη φράση-κλειδί (phrase-key) και να έχουν γραφτεί από τον συγκεκριμένο λογαριασμό (account). Επιπλέον έγινε συνδυασμός κάθε hashtag με κάθε λογαριασμό και αναλύθηκαν μέσω αυτής της γραμμής εντολών. Αυτή η ενέργεια είχε ως αποτέλεσμα συνολικά 108.311 tweets από τα οποία, τα 60.207 είναι θετικά και τα 48.104 είναι αρνητικά. Πριν τον καθαρισμό και την οργάνωση των δεδομένων, θα παρουσιαστούν τα δεδομένα που ανασύρθηκαν από το snscrape, εκτός από το κείμενο του tweet. Το όνομα των στηλών που επιστράφηκαν και το αντίστοιχο περιεχόμενό τους φαίνεται στον παρακάτω πίνακα (Πίνακας 4):

Πίνακας 4: Στήλες που επιστρέφονται από την βιβλιοθήκη snscrape

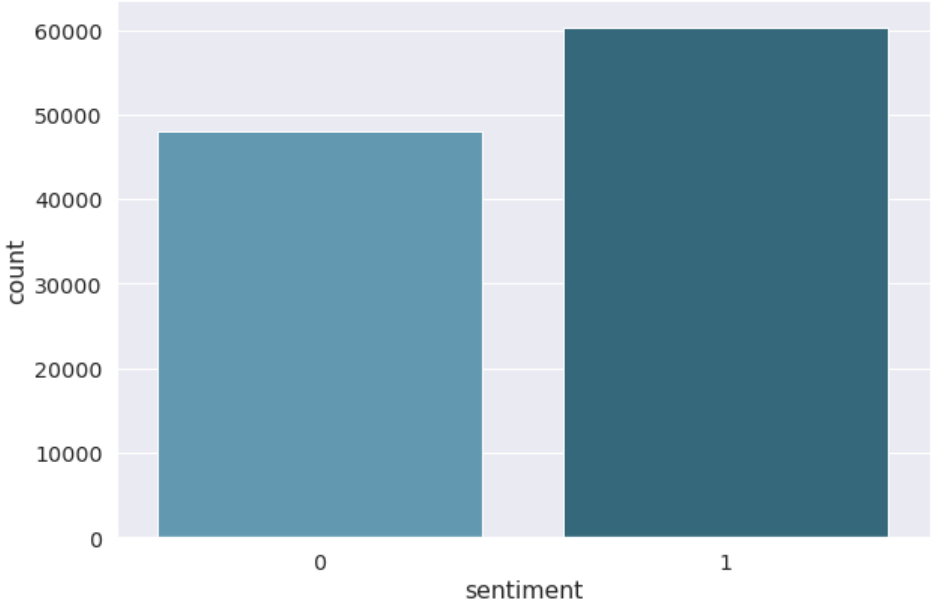
Χαρακτηριστικό	Λεπτομέρειες	Μετέπειτα Χρήση
url	Σύνδεσμος που δείχνει την τοποθεσία του tweet	Όχι
date	Ημερομηνίας δημιουργίας του tweet	Όχι
content	Το περιεχόμενο κειμένου του tweet	Ναί
renderedContent	Φαίνεται ότι είναι επίσης περιεχόμενο κειμένου του tweet	Όχι
id	Ταυτότητα του tweet	Όχι
user	Αντικείμενο του χρήστη που περιέχει τα ακόλουθα δεδομένα: username, displayname, id, description, descriptionUrls, verified, created, followersCount, friendsCount, statusCount, favouritesCount, listedCount, mediaCount, location, protected, linkUrl, profileimageUrl, profileBannerUrl	Όχι
outlinks		Όχι
toooutlist		Όχι
replyCount	Αριθμός απαντήσεων	Ναί
retweetCount	Αριθμός retweets	Ναί
likeCount	Αριθμός likes	Ναί
quoteCount	Αριθμός των χρηστών που ανέφεραν το tweet και απάντησαν	Ναί
coversationId	Φαίνεται να είναι ίδιο με την ταυτότητα του tweet	Όχι
lang	Απο γεννητρια η οποία υποθέτει την γλώσσα στην οποία το tweet είναι γραμμένο	Όχι
source	Από πού αναρτήθηκε το tweet, π.χ: iPhone, Android κλπ.	Όχι
media	Πολυμέσων, που περιέχουν previewUrl, fullUrl, και type	Όχι
retweetedTweet	Αν είναι retweet, η ταυτότητα του αρχικού tweet	Όχι
quotedTweet	Εάν αναφέρεται σε άλλο tweet, την ταυτότητα του αρχικού tweet	Όχι
mentionedUsers	Χρήστες οι οποίοι αναφέρονται στο tweet	Όχι

4.2 Περιγραφική Στατιστική των Δεδομενων

Σε αυτήν την ενότητα θα εξερευνηθούν οι τάσεις που παρουσιάζουν τα δεδομένα της έρευνας. Πιο συγκεκριμένα θα παρουσιαστεί ο τρόπος με τον οποίο κατανέμονται τα δεδομένα στις δύο διαφορετικές κατηγορίες (Αρνητές και Υποστηρικτές). Από τα αποτελέσματα του snscrape, θα εξεταστούν τα εξής : likes, retweets, replies, quotes. Από τη μορφολογία των tweets θα εξερευνηθεί επίσης ο αριθμός των λέξεων, ο αριθμός παράθεσης ιστοσελίδων (π.χ. https://t.co/Gb7At6AvlB)), ο αριθμός αναφοράς σε άλλους λογαριασμούς (π.χ. @UN), ο αριθμός κεφαλαίων λέξεων (π.χ. NEED), ο αριθμός emojis, ο αριθμός hashtags (π.χ. #TimeForAction) και ο αριθμός των σημείων στίξεων, πιο συγκεκριμένα ο αριθμός των θαυμαστικών μαζί με τον αριθμό των ερωτηματικών .

4.2.1 Ισορροπία Κατηγοριών

Από το Σχήμα 2 φανερώνεται ότι τα δεδομένα ανα κατηγορία δεν είναι ισορροπημένα. Συγκεκριμένα, υπάρχουν 48.104 δεδομένα Αρνητών και 60.207 δεδομένα Υποστηρικτών. Αν και η διαφορά είναι σημαντική, προς το παρόν δεν θα υπάρξει περαιτέρω επεξεργασία. Η διαφορά αυτή θα επιβεβαιωθεί και από τα αποτελέσματα που θα επαληθευτεί και από τα ML μοντέλα που θα χρησιμοποιηθούν αργότερα.



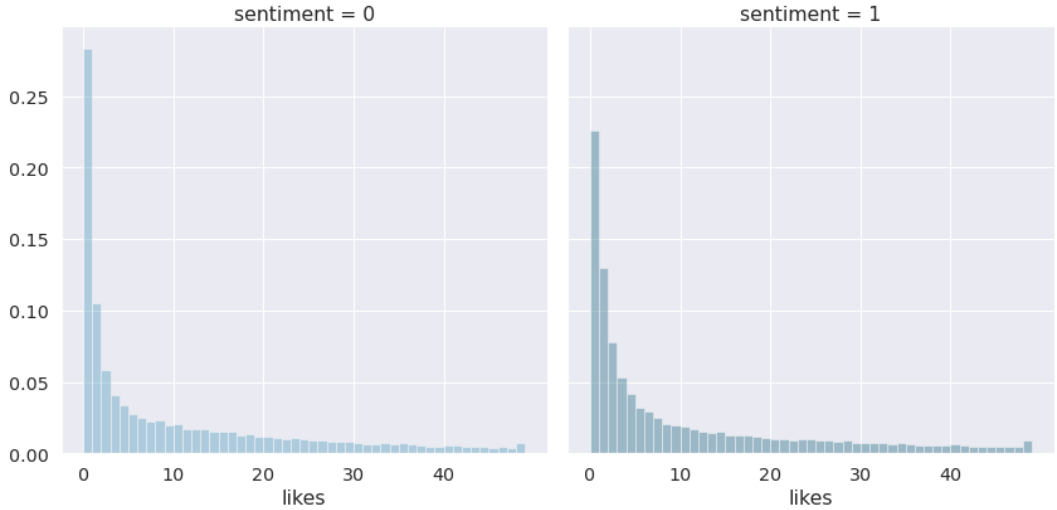
Σχήμα 2: Ισορροπία Κατηγοριών

4.2.2 Αριθμοί των Likes

Από τον Πίνακα 5 φαίνεται ότι η Υποστηρικτική στάση είναι πιο αρεστή και αποδεκτή από τους υπόλοιπους χρήστες του Twitter. Συγκεκριμένα ο μέγιστος αριθμός likes που έχει πάρει κάποιο tweet αρνητικης στάσης είναι 65.434, ενώ το tweet με τα περισσότερα likes της Υποστηρικτικής στάσης φτάνει μέχρι και τα 699.864 likes. Επιπλέον από το Σχήμα 3 φαίνεται ότι τα αρνητικά tweets τείνουν να συλλέγονται κοντά στο 0.

Πίνακας 5: Περιγραφική Στατιστική για τον Αριθμό των Likes

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	219,67	1.241,35	0	1	11	52	65.434
1	60.207	300,72	6.944,38	0	1	10	47	699.864



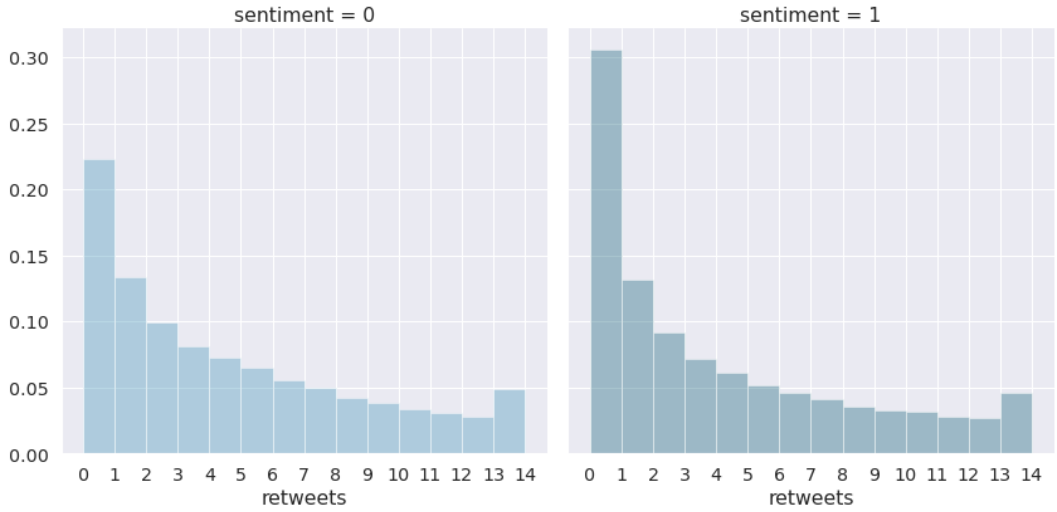
Σχήμα 3: Αριθμός Like

4.2.3 Αριθμός Αναπαραγωγής του Tweet (Retweet)

Για τον αριθμό των retweets, απο τον Πίνακα 6 γίνεται αντιληπτό ότι δεν υπάρχουν μεγάλες διαφορές πέρα από το ότι τα retweets της Υποστηρικτικής στάσης έχουν μεγαλύτερη διασπορά και πολύ μεγαλύτερη μέγιστη τιμή. Παρόλα αυτά, από το Σχήμα 4 διακρίνεται ότι τα tweets των Υποστηρικτών έχουν μεγαλύτερη τάση να μην αναπαράγονται, από τα tweet των Αρνητών.

Πίνακας 6: Περιγραφικη Στατιστική για τον Αριθμό Αναπαραγωγης του Tweet (Retweet)

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	72,51	371,07	0	2	7	24	20.350
1	60.207	65,46	1.148,93	0	1	8	32	188.794



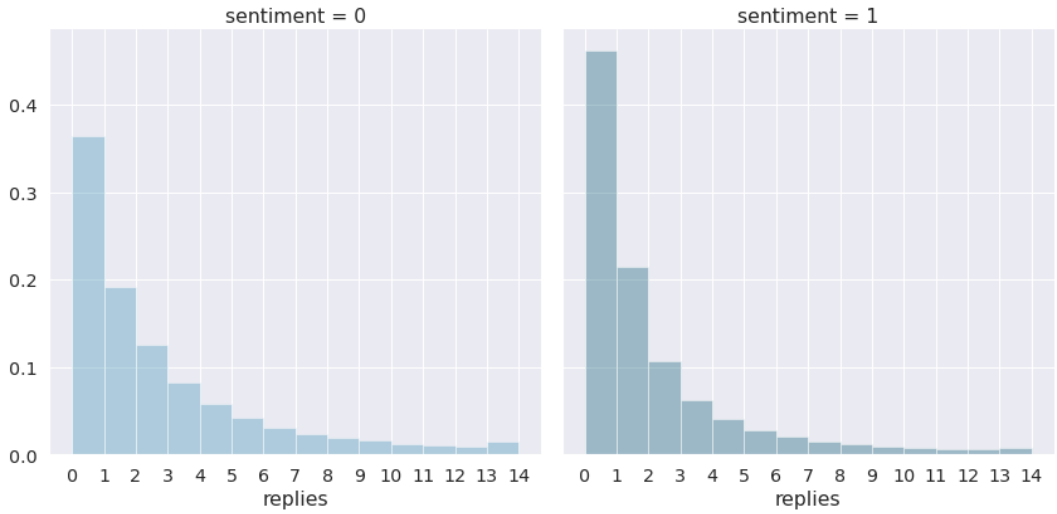
Σχήμα 4: Αριθμός Retweet

4.2.4 Αριθμός Απαντήσεων

Ο αριθμός των απαντήσεων ανά tweet διαθέτει παρόμοιες τάσεις με τον αριθμό των retweets. Όπως φαίνεται απο τον Πίνακα 7 τα tweet των Αρνητών κατά μέσο όρο έχουν περισσότερες απαντήσεις από τα tweet των Υποστηρικτών. Εν αντιθέσει, οι περισσότερες απαντήσεις τις εμφανίζονται σε tweet των Υποστηρικτων με τον μέγιστο αριθμό απαντήσεων να φτάνει τις 143.513 απαντήσεις. Το Σχήμα 5 παρουσιάζει τον αριθμό retweet ανα κατηγορία. Όπως φαίνεται είναι μεγαλύτερη η ποσότητα των tweets των Υποστηρικτών που συλλέγουν 0 απαντήσεις από αυτές των Αρνητών.

Πίνακας 7: Περιγραφική Στατιστική για τον Αριθμό Απαντήσεων

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	20,54	159,86	0	0	2	6	17.438
1	60.207	17,48	722,65	0	0	1	3	143.513



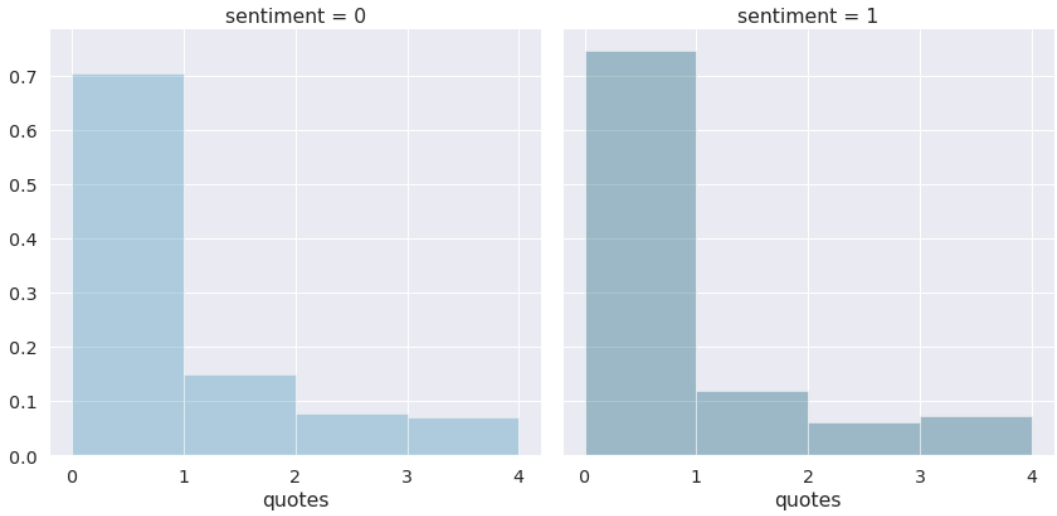
Σχήμα 5: Αριθμός Replies

4.2.5 Αριθμός Αναφορών σε άλλο Tweet

Στο Σχήμα 6, το οποίο σχετίζεται με τον αριθμό των αναφορών σε άλλο tweet, δεν φαίνονται μεγάλες διαφορές μεταξύ Αρνητών και Υποστηρικτών. Στον Πίνακα 8, όμως, γίνεται εμφανές ότι ακόμα και αν η μέση τιμή των δύο κατηγοριών είναι κοντά αριθμητικά, ο μέγιστος αριθμός αναφορών των Υποστηρικτικών tweets μπορεί να φτάσει μέχρι και τα 40.281 tweets ενώ στα Αρνητικά tweets μπορεί να φτάσει μέχρι και τις 6.356 αναφορές.

Πίνακας 8: Περιγραφική Στατιστική για τον Αριθμό Αναφορών σε άλλο Tweet

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	7,27	60,54	0	0	0	2	6.356
1	60.207	6,44	215,15	0	0	0	1	40.281



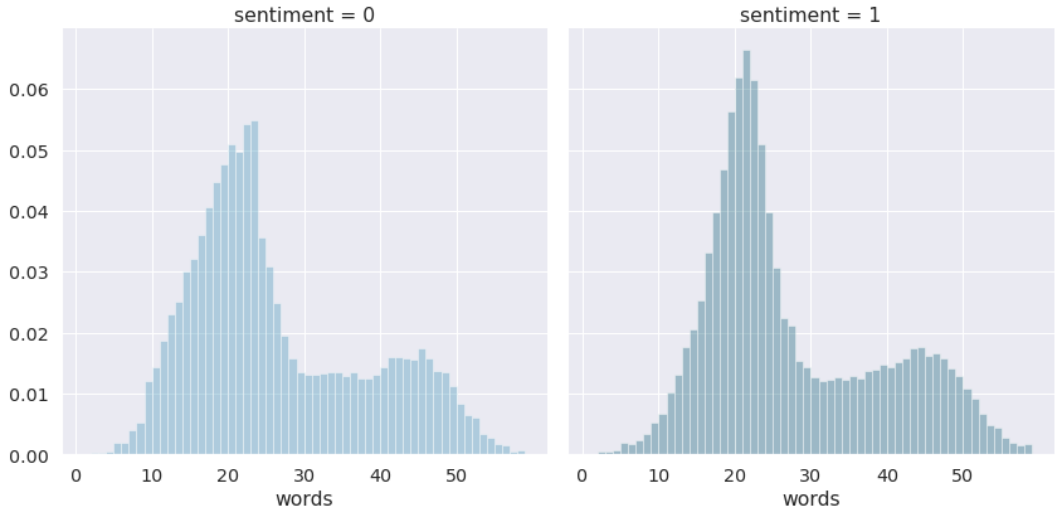
Σχήμα 6: Αριθμός Quotes

4.2.6 Αριθμός Λέξεων

Όσον αφορά τον αριθμό λέξεων, κοιτώντας τον Πίνακα 9, δεν φαίνεται να υπάρχουν σημαντικές διαφορές μεταξύ των κατηγοριών. Παρόλα αυτά, στο Σχήμα 7 παρατηρούνται κάποιες διαφορές όσων αφορά την κατανομή του μεγέθους (αριθμό λέξεων) των tweets. Φαίνεται ότι το μέγεθος των tweets παρουσιάζει δύο σημαντικές τάσεις. Συγκεκριμένα, τα tweets των 15-25 λέξεων είναι πολύ πιο συχνά και στις δυο κατηγορίες και αμέσως επόμενα έρχονται τα tweets των 40-50 λέξεων. Επίσης είναι φανερό ότι οι τάσεις αυτές είναι πολύ πιο έντονες στα tweets των Υποστηρικτών από τα tweets των Αρνητών.

Πίνακας 9: Περιγραφική Στατιστική για τον Αριθμό Λέξεων

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	25,97	11,58	1	18	23	34	94
1	60.207	27,12	11,67	1	19	23	36	94



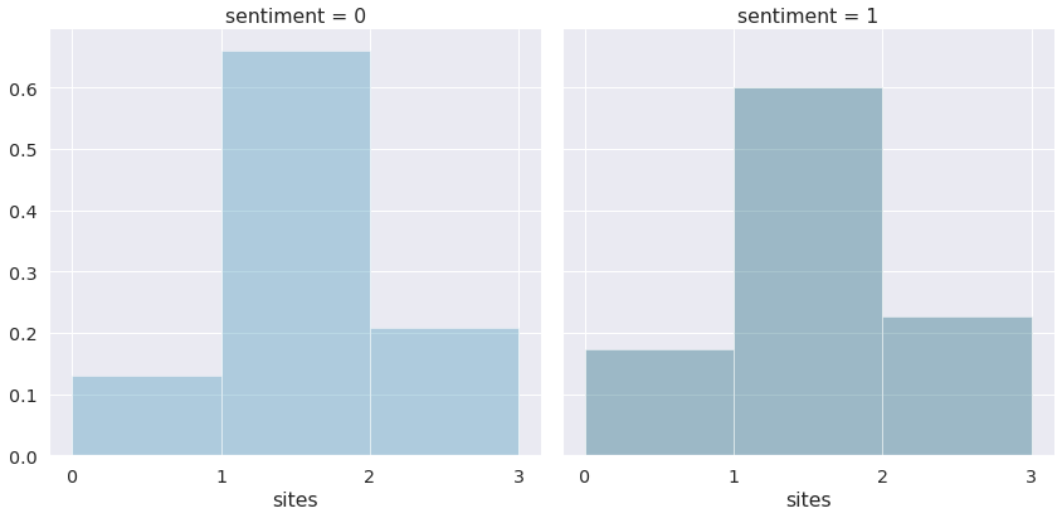
Σχήμα 7: Αριθμός Λέξεων

4.2.7 Αριθμός Ιστοσελίδων

Για τον αριθμό παράθεσης ιστοσελίδων δεν γίνεται αντιληπτή κάποια σημαντική διαφορά μεταξύ των κατηγοριών, ούτε στα περιγραφικά δεδομένα του Πίνακα 10, αλλά ουτε οπτικά στο Σχήμα 8.

Πίνακας 10: Περιγραφική Στατιστική για τον Αριθμό Ιστοσελίδων

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	1,09	0,6	0	1	1	1	5
1	60.207	1,06	0,65	0	1	1	1	6



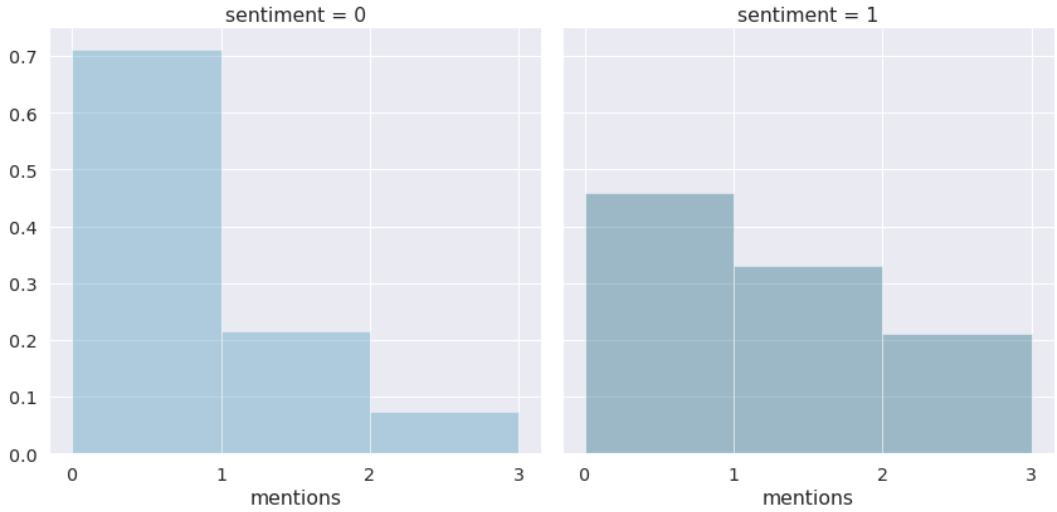
Σχήμα 8: Αριθμός Ιστοσελίδων

4.2.8 Αριθμός Αναφορών σε άλλους Λογαριασμούς

Στον Πίνακα 11 φαίνεται ότι τα Υποστηρικτικά tweets τείνουν να έχουν περισσότερες αναφορές σε άλλους λογαριασμούς, κατά μέσο όρο. Το ίδιο φαίνεται επίσης και από το Σχήμα 9.

Πίνακας 11: Περιγραφική Στατιστική για τον Αριθμό Αναφορών σε άλλους Λογαριασμούς

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	0,4	0,82	0	0	0	1	50
1	60.207	1,08	1,68	0	0	1	2	50



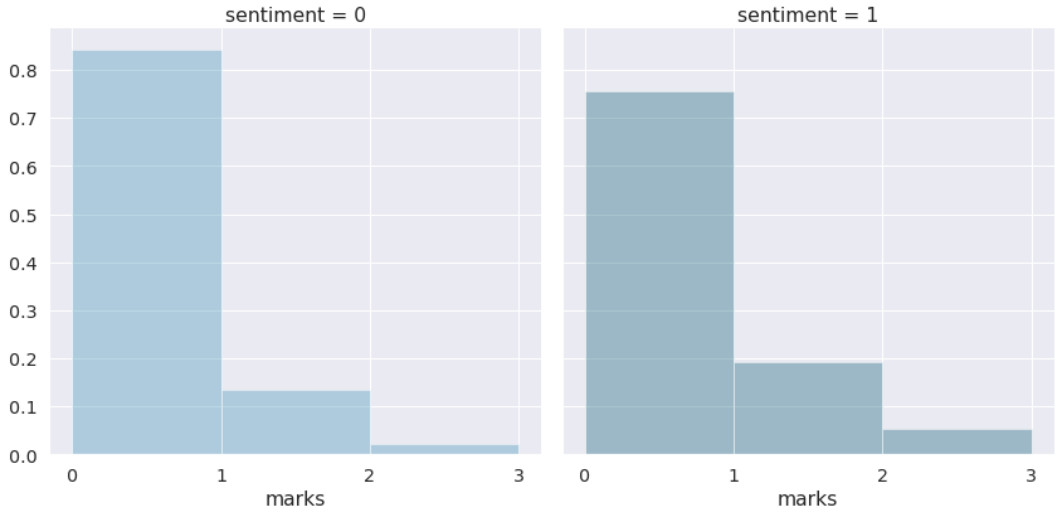
Σχήμα 9: Αριθμός Αναφορών σε άλλους Λογαριασμούς

4.2.9 Αριθμός Σημείων Στίξεων, όπως !, ?

Για τα σημεία στίξεων δεν υπάρχει κάποια ιδιαίτερη τάση, όπως φαίνεται από το Σχήμα 10. Θα πρέπει, παρόλα αυτά, να αναφερθεί ότι το tweet με τα περισσότερα σημεία στίξης ανήκει σε tweet Αρνητών, με συνολικό αριθμό ερωτηματικών και θαυμαστικών, 48 (Πίνακας 12).

Πίνακας 12: Περιγραφική Στατιστική για τον Αριθμό Σημείων Στίξεων

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	0,2	0,59	0	0	0	0	48
1	60.207	0,32	0,65	0	0	0	0	13



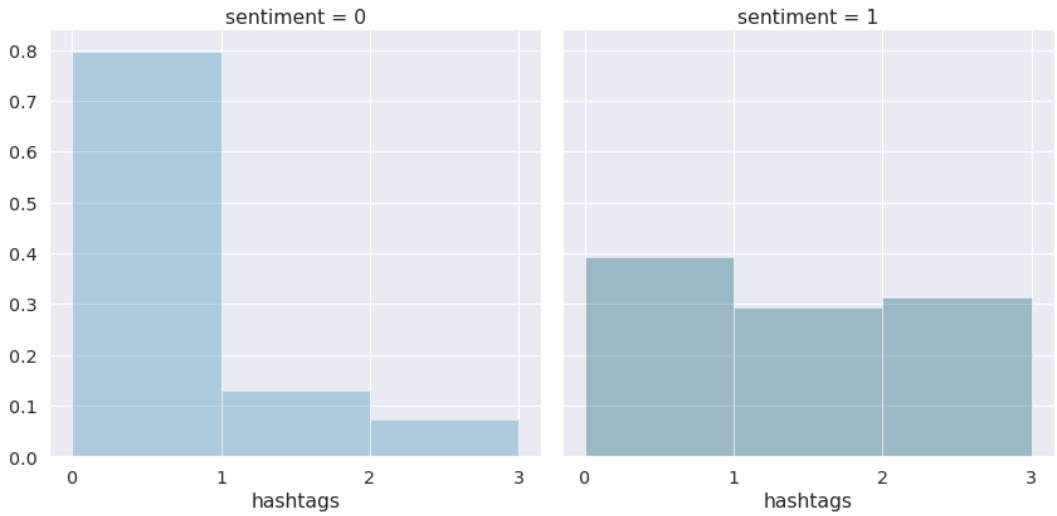
Σχήμα 10: Αριθμός Σημείων Στίξεων

4.2.10 Αριθμός Ετικετών (Hashtags)

Στον Πίνακα 13 παρατηρείται ότι η χρήση των hashtags σε tweets Αρνητών είναι σχεδόν μηδαμινή. Συγκεκριμένα παραπάνω από το 75% της Αρνητικής Στάσης tweet είναι 0. Για την Υποστηρικτική Στάση η κατανομή της χρήσης hashtag είναι πιο ισορροπημένη. Γύρω στο 40% δεν χρησιμοποιεί hashtags , γύρω στο 30% χρησιμοποιεί ένα hashtag , ενώ γύρω στο 30% χρησιμοποιούν τουλάχιστον 2 hashtags με τον μεγαλύτερο αριθμό να φτάνει τα 18 hashtags (Σχήμα 11).

Πίνακας 13: Περιγραφική Στατιστική για τον Αριθμό Hashtags

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	0,39	0,92	0	0	0	0	12
1	60.207	1,28	1,37	0	0	1	2	18



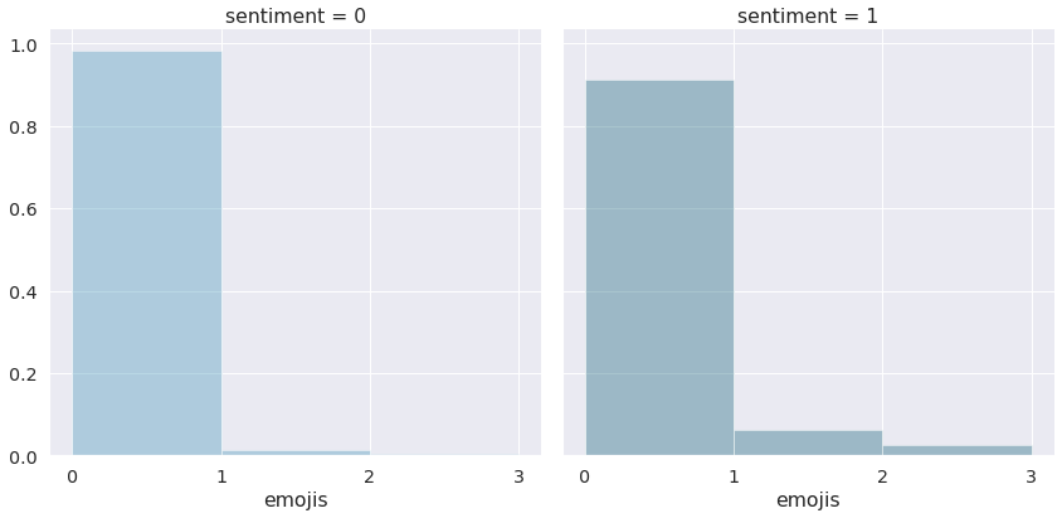
Σχήμα 11: Αριθμός Hashtags

4.2.11 Αριθμός Emojis

Η κατανομή της χρήσης των emojis δεν εμφανίζει κάποια ξεκάθαρη τάση (Σχήμα 12).Καμία από τις δύο παρατάξεις δεν φαίνεται χρησιμοποιεί συχνά τα emojis, μπορεί όμως κανείς να φτάσει στο συμπέρασμα ότι είναι πιο πιθανό για έναν Υποστηρικτή να περιέχει emoji στο tweet του, παρά για έναν Αρνητή. Σύμφωνα με τον Πίνακα 14, έχουν χρησιμοποιηθεί σε ένα μόνο tweet μεχρι 40 και 20 emojis, από Υποστηρικτή και Αρνητή χρήστη, αντίστοιχα.

Πίνακας 14: Περιγραφική Στατιστική για τον Αριθμό Emojis

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	0,03	0,26	0	0	0	0	20
1	60.207	0,15	0,64	0	0	0	0	40



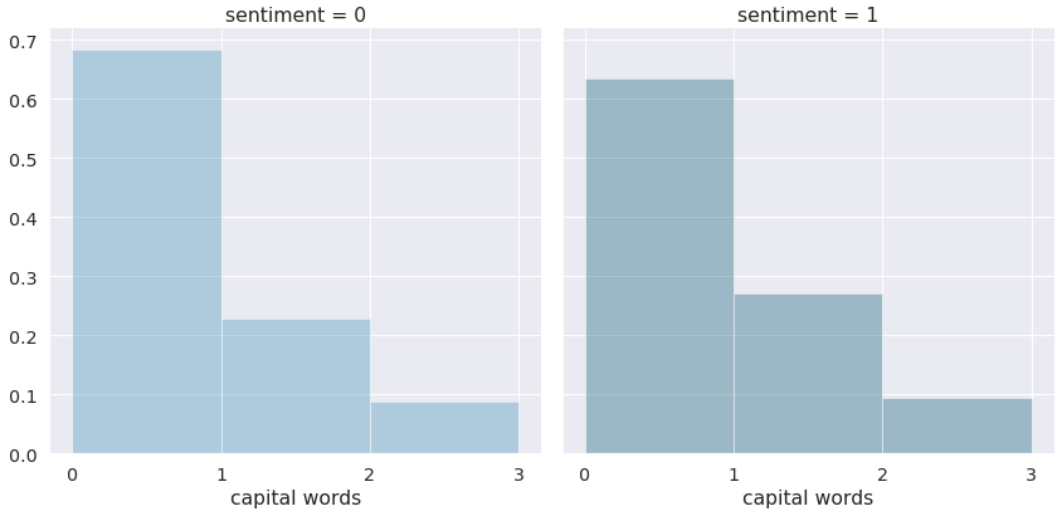
Σχήμα 12: Αριθμός Emojis

4.2.12 Αριθμός Κεφαλαίων Λέξεων

Η χρήση κεφαλαίων λέξεων στα social media, συνήθως, υποδεικνύει ένα μήνυμα εκνευρισμού ή απελπισίας, αλλά θα μπορούσε κανείς να πει ότι αντιστοιχεί και με μια ¨κραυγή¨ στον προφορικό λόγο. Ενώ η χρήση των κεφαλαίων λέξεων μπορεί να βοηθήσει τον κάθε λογαριασμό, Αρνητή ή Υποστηρικτή, να περάσει το μήνυμά του, από τον Πίνακα 15 αλλά και από το Σχήμα 13, σαν τεχνική φαίνεται να χρησιμοποιείται με παρόμοιο τρόπο και από τις δύο παρατάξεις.

Πίνακας 15: Περιγραφική Στατιστική για τον Αριθμό Κεφαλαίων Λέξεων

sentiment	count	mean	std	min	25%	50%	75%	max
0	48.104	0,53	1,07	0	0	0	1	34
1	60.207	0,55	1,04	0	0	0	1	45



Σχήμα 13: Αριθμός Κεφαλαίων Λέξεων

4.3 Εκκαθάριση Δεδομένων

Στην συνέχεια της ανάλυσης αυτής θα γίνει η μετατροπή των tweets σε διανυσματικές αναπαραστάσεις των λέξεων, θα πρέπει πρώτα να απαλειφθούν χαρακτηριστικά τα οποία είτε μπορεί να δημιουργήσουν σφάλματα ή δεν είναι απαραίτητα. Οι επικείμενες αλλαγές των tweets θα παρουσιαστούν χρησιμοποιώντας το επόμενο tweet ως παράδειγμα :

OMG today was that day went I went on LIVE @cbsnews and told all the adults in America that they are threatened by us! emoji emoji
#ClimateStrike
#SchoolStrike4Climate
#WhateverItTakes
@GretaThunberg @MichaelEMann @EricHolthaus
https://t.co/8aPGNCqogI

4.3.1 Αναφορές σε άλλους χρήστες

Αρχικά θα αφαιρεθεί κάθε αναφορά σε άλλο χρήστη. Αφού λοιπόν κάθε αναφορά σε άλλο χρήστη συνοδεύεται από τον ειδικό χαρακτήρα @,θα αφαιρεθεί ολόκληρος ο χαρακτήρας, μαζί με τον όνομα του χρήστη. Το tweet θα διαμορφωθεί ως εξής :

OMG today was that day went I went on LIVE and told all the adults in America that they are threatened by us! emoji emoji
#ClimateStrike
#SchoolStrike4Climate
#WhateverItTakes
https://t.co/8aPGNCqogI

4.3.2 Παραθέση ιστοσελίδων

Στη συνέχεια θα αφαιρεθούν όλες οι παραθέσεις ιστοσελίδων, της μορφής ***http://***. Μετά από την αφαίρεση των ιστοσελίδων το tweet θα έχει τη μορφή :

OMG today was that day went I went on LIVE and told all the adults in America that they are threatened by us! emoji emoji
#ClimateStrike
#SchoolStrike4Climate
#WhateverItTakes

4.3.3 Emojis

Σε πολλά tweets οι χρήστες χρησιμοποιούν emojis είτε για να εκφράσουν ένα συναίσθημα, είτε για να κάνουν το tweet τους αισθητικά πιο ωραίο. Πλέον η επιλογή των emojis μπορεί να ερμηνευθεί και να αλλάξει εντελώς το μήνυμα ενός κειμένου. Για το λόγο αυτό θα χρησιμοποιηθεί ένας «διερμηνέας» emoji. Ο «διερμηνέας» αυτός έχει την εξής αρμοδιότητα, να βρίσκει τα emojis στο κείμενο και να τα αντικαταστεί με την ερμηνεία τους. Στο συγκεκριμένο παράδειγμα όπου εντοπίσει το emoji , ο διερμηνέας θα γράψει την ερμηνεία του δηλαδή face_with_tears_of_joy. Το tweet θα μετατραπεί ως εξής :

OMG today was that day went I went on LIVE and told all the adults in America that they are threatened by us! face_with_tears_of_joy face_with_tears_of_joy
#ClimateStrike
#SchoolStrike4Climate
#WhateverItTakes

4.3.4 Σημεία στίξης

Σε αυτό το κομμάτι της εκκαθάρισης κειμένου, θα αφαιρεθούν τα σημεία στίξης, όπως :, _, # κ.λπ. Ως αποτέλεσμα, το κείμενο θα μετατραπεί σε:

OMG today was that day went I went on LIVE and told all the adults in America that they are threatened by us face with tears of joy face with tears of joy ClimateStrike SchoolStrike4Climate WhateverItTakes

4.3.5 Συντομογραφίες

Ακόμα ένα γνώρισμα της εποχής είναι οι συντομογραφίες. Ειδικά στα social media οι εκφράσεις όπως YOLO (You only live once), IDK (I don't know), ILU (I love you), OMG (Oh my god) είναι αρκετά συχνές. Για τον λόγο αυτό θα χρησιμοποιηθούν οι ήδη υπάρχουσες γνώσεις για αυτές τις συντομογραφίες, για να γίνει το κείμενο λίγο πιο κατανοητό. Η νέα του μορφή είναι η εξής:

oh my god today was that day went I went on LIVE and told all the adults in America that they are threatened by us face with tears of joy face with tears of joy ClimateStrike SchoolStrike4Climate WhateverItTakes

4.3.6 Νούμερα

Το επόμενο βήμα εκκαθάρισης του κειμένου θα είναι να αφαιρεθούν τα τυχόν νούμερα. Τα νούμερα δεν φέρουν κάποια σημασία ακόμα και αν διατηρηθούν μέσα στο tweet. Οπότε αφαιρώντας τα δημιουργείται μια πιο ευκρινής εικόνα του κειμένου.

oh my god today was that day went I went on LIVE and told all the adults in America that they are threatened by us face with tears of joy face with tears of joy ClimateStrike SchoolStrikeClimate WhateverItTakes

4.3.7 Μετατροπή σε πεζά

Στο τελευταίο βήμα πριν την παρουσίαση της διανυσματικής αναπαράστασης των λέξεων θα μετατραπούν όλα τα γράμματα από όλες λέξεις σε πεζά έτσι ώστε να είναι ομογενείς όλες οι λέξεις μεταξύ τους.

oh my god today was that day went i went on live and told all the adults in america that they are threatened by us face with tears of joy face with tears of joy climatestrike schoolstrikeclimate whateverittakes

4.3.8 Αφαίρεση stopwords, ή αλλιώς πολύ συχνών χαρακτήρων

Όπως προαναφέρθηκε σε αυτή την εργασία θα χρησιμοποιηθούν τεχνικές sentiment analysis για να προσδιοριστεί αν κάποιος είναι θετικά ή αρνητικά κατασταλαγμένος σε σχέση με την κλιματική αλλαγή. Σε αυτές τις τεχνικές παίζει πολύ μεγάλο ρόλο η βαρύτητά της κάθε λέξης δηλαδή το πόσο συχνά χρησιμοποιείται. Στο γραπτό και στον προφορικό λόγο χρησιμοποιούνται πολύ συχνά συγκεκριμένες λέξεις οι οποίες δεν φέρουν σημαντική βαρύτητα όπως είναι οι συνδετικές λέξεις για παράδειγμα το: 'and', 'I', 'you', 'is', 'are'. Οι λέξεις αυτές δεν φέρουν καμία σημαντική βαρύτητα όσον αφορά το θέμα της εργασίας, την κλιματική αλλαγή και επειδή θα εμφανίζονται πολύ πιο συχνά από τις άλλες σημαντικές λέξεις πρέπει να αφαιρεθούν.

oh god today day went went live told adults america threatened us face tears joy face tears joy climatestrike schoolstrikeclimate whateverittakes

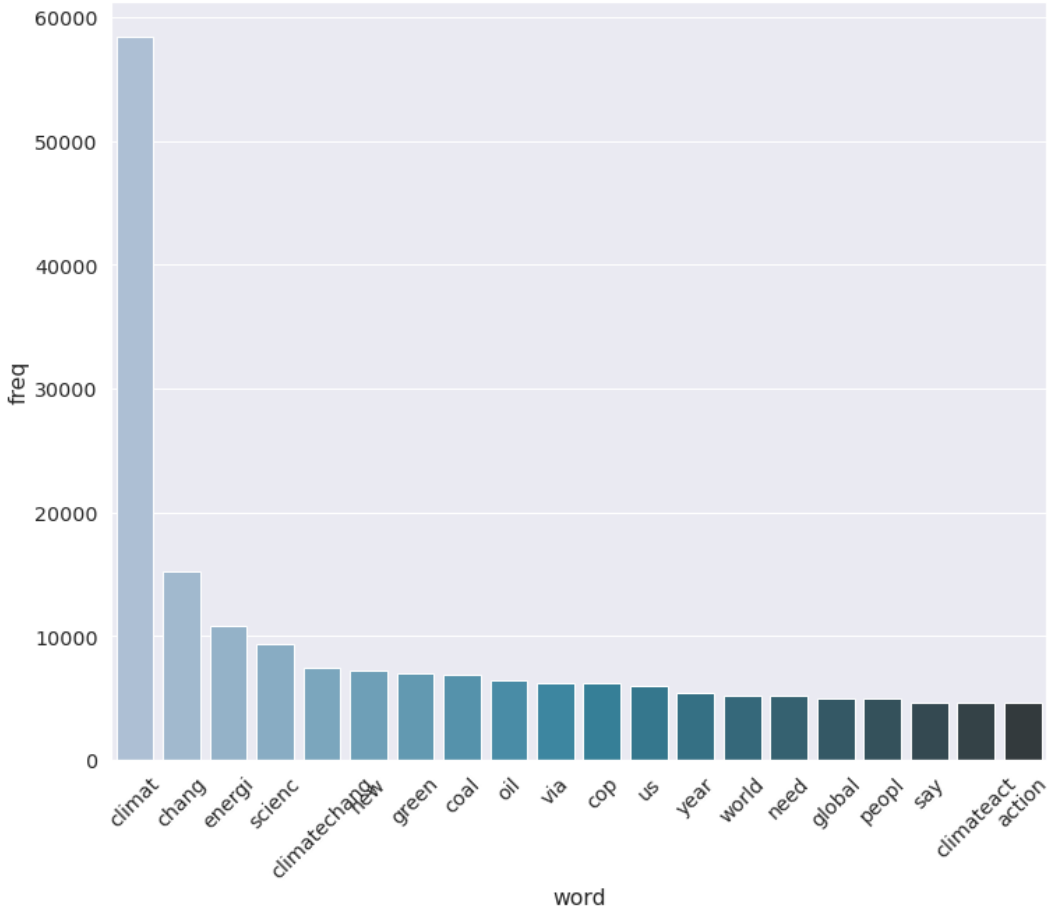
4.3.9 Αποκοπή Καταλήξεων

Η αποκοπή καταλήξεων, όπως το λέει και η φράση, είναι μία μέθοδος στην οποία αφαιρείται η κατάληξη μιας λέξης και μένει μόνο η ρίζα - το θέμα της. Αυτή η τεχνική χρησιμοποιείται πολύ συχνά στην natural language processing κοινότητα. Ο λόγος είναι διότι με αυτό τον τρόπο μειώνεται αισθητά ο αριθμός απαιτούμενων υπολογισμών. Για παράδειγμα αντί να χρειαστεί να αναλυθούν οι επόμενες τέσσερις διαφορετικές λέξεις talk, talked, talking, talkative, θα αναλυθεί μόνο το θέμα των προηγούμενων λέξεων που είναι το talk. Έτσι η ανάλυση των δεδομένων γίνεται πιο γρήγορα και αποδοτικά χωρίς να εισαχθούν προκαταλήψεις ως προς τις λέξεις.

oh god today day went went live told adult america threaten us face tear joy face tear joy climatestrik schoolstrikeclim whateverittak

4.3.10 Τελική Βάση Δεδομένων

Στο Σχήμα 14 που ακολουθεί παρουσιάζονται τα πιο συχνά θέματα λέξεων από τα κείμενα που συλλεχθεί. Είναι προφανές ότι η πλειοψηφία των λέξεων είναι λέξεις συνδεδεμένες με το θέμα της εργασίας, την κλιματική αλλαγή. Επίσης είναι φανερό, ότι υπάρχουν και λέξεις οι οποίες δεν έχουν ξεκάθαρη σχέση με την κλιματική αλλαγή, όπως via, us, say. Τέλος, φαίνεται ότι η λέξη/το θέμα cop χρησιμοποιείται αρκετά και αυτό γιατί όπως προαναφέρθηκε και στην Ενότητα των hashtags που χρησιμοποιήθηκαν, η λέξη COP σημαίνει Conference of Parties. Το COP είναι ένα συνέδριο που διοργανώνεται με πρωτοβουλία των Ηνωμένων Εθνών για να λύσει το θέμα της κλιματικής αλλαγής.



Σχήμα 14: Τα πιο συχνά θέματα λέξεων

4.4 Μοντέλα αναπαράστασης λέξεων στην NLP

Σε αυτή την ενότητα θα εφαρμοστούν μερικές από τις μεθόδους ενσωμάτωσης λέξεων που αναφέρονται στο Κεφάλαιο 3. Θα εφαρμοστεί το Unigram Bag of Words, το Bigram Bag of Words, το TF-IDF, το Word2Vec και οι αναπαραστάσεις λέξεων FastText. Θα χρησιμοποιηθεί το ίδιο tweet που χρησιμοποιήθηκε και για την διαδικασία εκκαθάρισης δεδομένων. Παρακάτω βρίσκεται το tweet πριν από τη διαδικασία εκκαθάρισης:

OMG today was that day went I went on LIVE @cbsnews and told all the adults in America that they are threatened by us! emoji emoji
#ClimateStrike
#SchoolStrike4Climate
#WhateverItTakes
@GretaThunberg @MichaelEMann @EricHolthaus
https://t.co/8aPGNCqogI

Με το πέρας της διαδικασίας καθαρισμού το ίδιο tweet θα έχει την εξής μορφή:

oh god today day went went live told adult america threaten us face tear joy face tear joy climatestrik schoolstrikeclim whateverittak

4.5 One Hot Encoding - Bag Of Words (BoW)

4.5.1 Uni-grams

Με την εφαρμογή αυτής της μεθόδου θα γίνει προσπάθεια διανυσματικής αναπαράστασης της κάθε λέξης του συνόλου στο συγκεκριμένο tweet. Για το σύνολο των λέξεων στο συγκεκριμένο tweet, ο υπολογισμός είχε ως αποτέλεσμα 6.679 λέξεις/δείκτες. Κάθε λέξη στο σύνολο καταλαμβάνει μια συγκεκριμένη θέση σε αυτόν τον πίνακα, 6.679 περιπτώσεις. Οι λέξεις του προαναφερθέντος tweet αντιστοιχούν στους δείκτες που εμφανίζονται στον Πίνακα 16.

Πίνακας 16: Πίνακας των Unigrams στα Bag of Words

word	oh	god	today	day	went	live
index	4132	2526	6010	1451	6483	3488
word	told	adult	america	threaten	face	
index	6016	81	196	5968	2123	
word	tear	joy	climatestrik	schoolstrikeclim	whateverittak	
index	5890	3223	1059	5208	6497	

Τα αποτελέσματα της μεθόδου αυτής δείχνουν ποιες από τις λέξεις του συνόλου που έχει δημιουργηθεί εμφανίζονται στο συγκεκριμένο tweet. Η παρουσία της λέξης στο tweet συμβολίζεται με 1 και η απουσία με 0. Αυτό σημαίνει ότι για το συγκεκριμένο tweet υπάρχει 1 στους δείκτες 81, 196, 1059, 1451, 2123, 2526, 3223, 3488, 4132, 5208, 5890, 5968, 6010, 6497 ενώ 0 σε όλους τους υπόλοιπους.

4.5.2 Bi-grams

Σε αντίθεση με την προηγούμενη μέθοδο εδώ γίνεται αναζήτηση της εμφάνιση κάθε δύο λέξεων. Όπως και πριν, αυτός ο υπολογισμός λαμβάνει υπόψη ολόκληρο το σύνολο. Ο υπολογισμός κατέληξε σε 15.045 συνδυασμούς των δύο λέξεων. Αυτό σημαίνει ότι κάθε συνδυασμός δύο λέξεων καταλαμβάνει μια συγκεκριμένη θέση σε έναν πίνακα 15.045 περιπτώσεων. Χρησιμοποιώντας το Bi-gram για το προαναφερθέν tweet αποτυπώνονται οι δείκτες που εμφανίζονται στον Πίνακα 17.

Πίνακας 17: Πίνακας των Bigrams στα Bag of Words

word	oh	oh god	god	today
index	9457	9458	5934	13542
word	today day	day	went	live
index	13555	3606	14535	8045
word	told	adult	america	threaten
index	13590	225	450	13443
word	face	face tear	tear	tear joy
index	5064	5084	13251	13252
word	joy	climatestrik	schoolstrikeclim	whateverittak
index	7454	2707	11855	14557
word	climatestrik schoolstrikeclim			
index	2712			

Μετά τον υπολογισμό των Uni-gram και Bi-gram, η αναπαράσταση αυτού του tweet θα είναι ένας πίνακας μήκους 15.045. Σε αυτόν τον πίνακα θα υπάρχει 1 στους δείκτες 225, 450, 2707, 2712, 3606, 5064, 5084, 5934, 7454, 8045, 9457, 9458, 11855, 13251, 13252, 13443, 13542, 13555, 13590, 14535, 14557 ενώ ο υπόλοιπος πίνακας θα γεμίσει με 0.

4.5.3 Term Frequency – Inverse Document Frequency (TF-IDF)

Για αυτόν τον υπολογισμό, όπως αναφέρεται στο Κεφάλαιο 3, λαμβάνεται υπόψη η συχνότητα της λέξης στο tweet καθώς και η συχνότητα της λέξης σε ολόκληρο το σώμα. Αυτή η μέθοδος θα δώσει μια τιμή στην κάθε λέξη. Αφού εφαρμοστεί αυτός ο υπολογισμός σε ολόκληρο το σώμα, δημιουργείται ένας πίνακας 6.679 περιπτώσεων για το συγκεκριμένο tweet. Σε αυτές τις περιπτώσεις, οι λέξεις στους αντίστοιχους δείκτες τους θα έχουν τις τιμές που φαίνονται στον Πίνακα 18. Όλοι οι άλλοι δείκτες του πίνακα θα συμπληρωθούν με 0.

Table 18: Οι τιμές των λέξεων του tweet με την αναπαράσταση της TF-IDF

word	oh	god	today	day	went	live
index	0.185	0.204	0.118	0.134	0.396	0.129
word	told	adult	america	threaten	face	
index	0.187	0.205	0.147	0.169	0.267	
word	tear	joy	climatestrik	schoolstrikeclim	whateverittak	
index	0.423	0.436	0.158	0.233	0.269	

4.5.4 Word2Vec

Το Word2vec είναι ένα μοντέλο πρόβλεψης που βασίζεται σε βαθιά μάθηση για τον υπολογισμό και τη δημιουργία υψηλής ποιότητας διανυσμάτων αναπαράστασης λέξεων, οι οποίες καταγράφουν ομοιότητες με βάση τα συμφραζόμενα και τη σημασιολογική ομοιότητα. Το μοντέλο αυτό συνήθως αγνοεί τη μορφολογική δομή κάθε λέξης και θεωρεί την λέξη ως μια ενιαία οντότητα. Παρακάτω παρουσιάζεται η διανυσματική αναπαράσταση του συγκεκριμένου tweet για κάθε μια λέξη χωριστά.

$$[0.08731418, 0.03902944, ..., 0.06603411, -0.04472182]$$

Η μεγαλύτερη αδυναμία του μοντέλου αυτού είναι ότι δεν μπορεί να αναπαραστήσει λέξεις που δεν εμφανίζονται στο σύνολο δεδομένων εκπαίδευσης.

4.5.5 FastText

Το FastText είναι μια επέκταση του Word2Vec. Στην ουσία διασπά τις λέξεις σε n-grams δηλαδή σε υπολέξεις. Μετά την εκπαίδευση του Νευρωνικού Δικτύου, αποτυπώνεται η ενσωμάτωση των λέξεων για όλα τα n-grams βάση του συνόλου δεδομένων εκπαίδευσης. Παρακάτω παρουσιάζεται το διάνυσμα ενσωμάτωσης λέξεων για το συγκεκριμένο tweet.

$$[1.65509935 * 10^{-2}, -1.97551344 * 10^{-2}, ..., 1.99307827 * 10^{-3}, 3.04156281 * 10^{-2}]$$

Οι σπάνιες λέξεις με αυτή τη μέθοδο μπορούν να αναπαρασταθούν σωστά, καθώς είναι πολύ πιθανό κάποιο από τα n-gram τους να εμφανίζονται και σε άλλες λέξεις.

5 Ανάλυση Δεδομένων και Ερμηνεία των Ευρημάτων

Στην ενότητα αυτή, θα παρουσιαστούν τα αποελέσματα ολόκληρης της έρευνας. Συνολικά δημιουργήθηκαν 7 μοντέλα για κάθε διανυσματική αναπαράσταση. Εδώ θα πρέπει να σημειωθεί ότι για τις διανυσματικές αναπαραστάσεις των Word2Vec και FastText δεν χρησιμοποιήθηκε το μοντελο Multinomial Naive Bayes αφού το συγκεκριμένο μοντέλο δεν μπορεί να ανταπεξέλθει στις αρνητικές τιμές που μπορεί να έχουν τα παραπάνω διανύσματα.

Στην ανάλυση συναισθήματος, χρησιμοποιούνται διαφορετικοί αλγόριθμοι μηχανικής μάθησης για την πρόβλεψη του συναισθήματος ενός κειμένου, είτε θετικό είτε αρνητικό. Τα αποτελέσματα που λήφθηκαν δείχνουν την ακρίβεια και τις βαθμολογίες ROC (Receiver Operating Characteristic) AUC (Area Under the Curve) για τα διάφορα μοντέλα ανάλυσης συναισθήματος: Logistic Regression, Random Forests, Support Vector Machine, Gaussian Naive Bayes, Multinomial Naive Bayes, 5- Nearest Neighbors και XGBoost Classifier.

Η ακρίβεια είναι ένα μέτρο της ικανότητας του μοντέλου να προβλέπει σωστά το συναίσθημα, εκφρασμένο ως ποσοστό. Η βαθμολογία ROC AUC, από την άλλη πλευρά, μετρά την ικανότητα του μοντέλου να διακρίνει μεταξύ θετικών και αρνητικών συναισθημάτων. Μια βαθμολογία ROC AUC 1 υποδεικνύει ότι το μοντέλο διαχωρίζει τέλεια τα θετικά από τα αρνητικά συναισθήματα, ενώ η βαθμολογία 0,5 υποδηλώνει μια τυχαία απόδοση.

Η ανάλυση θα χωριστεί συμφωνα με την διανυσματική αναπαράσταση των tweets, ξεκινώντας με το One Hot Encoding - Unigram.

5.1 One Hot Encoding - Unigrams

Τα αποτελέσματα (όπως φαίνονται στον Πίνακα 19 και στο Διάγραμμα 15) δείχνουν ότι τα μοντέλα Logistic Regression και Random Forests είχαν παρόμοια απόδοση, με ακρίβεια 0,89 και βαθμολογία ROC AUC 0,97 και 0,96 αντίστοιχα. Αυτά τα μοντέλα είναι σχετικά απλά και αποτελεσματικά στην ανάλυση συναισθήματος, καθιστώντας τα δημοφιλείς επιλογές για ανάλυση συναισθήματος. Βασίζονται στις αρχές της στατιστικής μάθησης και των γραμμικών μοντέλων και είναι αποτελεσματικά στο χειρισμό μεγάλων συνόλων δεδομένων.

Το μοντέλο Support Vector Machine πέτυχε υψηλότερη ακρίβεια 0,91 και βαθμολογία ROC AUC 0,97, υποδεικνύοντας υψηλότερη ικανότητα πρόβλεψης του συναισθήματος σε σύγκριση με τα μοντέλα Logistic Regression και Random Forest. Αυτό οφείλεται στην ικανότητα του SVM να εκτελεί σύνθετους διαχωρισμούς σε χώρους υψηλών διαστάσεων, καθιστώντας το ένα πιο ισχυρό μοντέλο για ανάλυση συναισθήματος.

Το μοντέλο Gaussian Naive Bayes είχε χαμηλότερη βαθμολογία ακρίβειας και ROC AUC, με ακρίβεια 0,73 και βαθμολογία ROC AUC 0,82 , υποδηλώνοντας κακή απόδοση. Τα μοντέλα Gaussian Naive Bayes βασίζονται στο θεώρημα πιθανοτήτων του Bayes και χρησιμοποιούνται συχνά σε εργασίες ταξινόμησης κειμένων, ωστόσο, έχουν περιορισμούς στον χειρισμό σύνθετων συνόλων δεδομένων και μπορεί να δυσκολεύονται σε εργασίες ανάλυσης συναισθήματος.

Το μοντέλο Multinomial Naive Bayes είχε σχετικά καλή απόδοση, με ακρίβεια 0,88 και βαθμολογία ROC AUC 0,97, ωστόσο, εξακολουθεί να έχει καλύτερη απόδοση από τα μοντέλα Logistic Regression, Random Forests και Support Vector Machine.

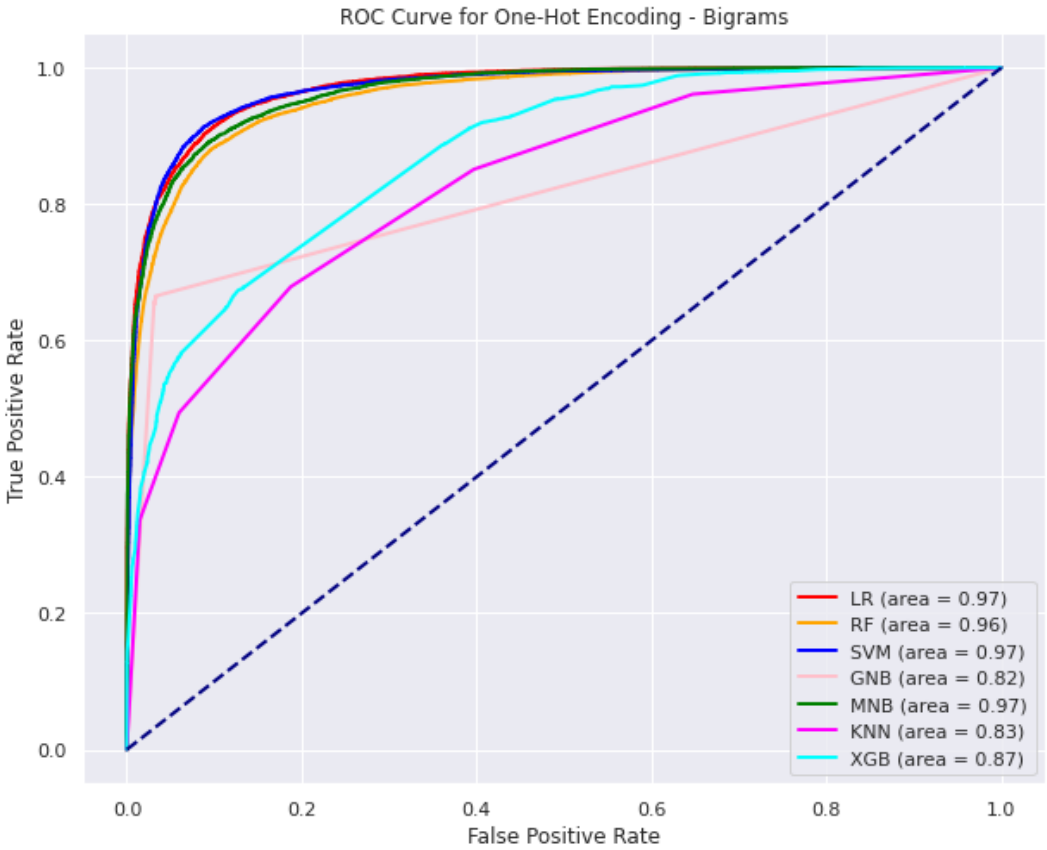
Το μοντέλο 5-Nearest Neighbors είχε παρόμοια απόδοση με το μοντέλο Gaussian Naive Bayes, με ακρίβεια 0,73 και βαθμολογία ROC AUC 0,83 , υποδεικνύοντας περιορισμένη ικανότητα πρόβλεψης συναισθημάτων. Το μοντέλο 5-Nearest Neighbors βασίζεται σε έναν απλό αλγόριθμο που βρίσκει τους πλησιέστερους γείτονες σε ένα δεδομένο κείμενο και το ταξινομεί με βάση την κλάση των πλησιέστερων γειτόνων.

Το μοντέλο XGBoost Classifier είχε τη χαμηλότερη βαθμολογία ακρίβειας και ROC AUC, με ακρίβεια 0,78 και βαθμολογία ROC AUC 0,87 , υποδεικνύοντας κακή απόδοση σε εργασίες ανάλυσης συναισθήματος. Το μοντέλο XGBoost Classifier είναι ένα μοντέλο συνόλου που χρησιμοποιείται συχνά για μεγάλα σύνολα δεδομένων, ωστόσο, η απόδοσή του μπορεί να επηρεαστεί από την πολυπλοκότητα της εργασίας ανάλυσης συναισθήματος.

Συμπερασματικά, τα αποτελέσματα των μοντέλων ανάλυσης συναισθήματος για το One Hot Encoding - Unigrams δείχνουν ότι τα μοντέλα Logistic Regression, Random Forests και Support Vector Machine έχουν την καλύτερη απόδοση, με υψηλή ακρίβεια και βαθμολογίες ROC AUC. Αυτά τα μοντέλα είναι κατάλληλα για εργασίες ανάλυσης συναισθήματος και παρέχουν αξιόπιστα αποτελέσματα. Τα μοντέλα Gaussian Naive Bayes, Multinomial Naive Bayes, 5-Nearest Neighbors και XGBoost Classifier, από την άλλη πλευρά, έχουν περιορισμένη απόδοση στην ανάλυση συναισθήματος και ενδέχεται να μην είναι κατάλληλα για πιο σύνθετες εργασίες.

Πίνακας 19: Ακρίβεια για τα μοντέλα του One Hot Encoding - Unigrams

	LR	RF	SVM	GNB	MNB	KNN	XGB
Ακρίβεια	0.89	0.89	0.91	0.73	0.88	0.73	0.78



Σχήμα 15: Καμπύλη ROC για τα μοντέλα One Hot Encoding - Unigrams

5.2 One Hot Encoding - Bigrams

Τα αποτελέσματα (όπως φαίνονται στον Πίνακα ;; και στο Διάγραμμα ;;) των μοντέλων ανάλυσης συναισθήματος που εκτελούνται με την αναπαράσταση One Hot Encoding - Bigrams υποδεικνύουν ότι ορισμένα μοντέλα είχαν καλύτερη απόδοση από άλλα. Το μοντέλο Logistic Regression είχε την υψηλότερη ακρίβεια με 0,91 και βαθμολογία ROC AUC 0,97. Το μοντέλο Random Forests είχε επίσης υψηλή ακρίβεια 0,89, αλλά η βαθμολογία ROC AUC ήταν χαμηλότερη στο 0,96. Το μοντέλο Support Vector Machine είχε την ίδια ακρίβεια με την Logistic Regression αλλά με ελαφρώς χαμηλότερη βαθμολογία ROC AUC 0,97. Το μοντέλο Gaussian Naive Bayes είχε ακρίβεια 0,79 και βαθμολογία ROC AUC 0,82. Το μοντέλο Multinomial Naive Bayes είχε καλή απόδοση με ακρίβεια 0,90 και βαθμολογία ROC AUC 0,97. Το μοντέλο 5-Nearest Neighbors είχε ακρίβεια 0,74 και βαθμολογία ROC AUC 0,83. Τέλος, το μοντέλο XGBoost Classifier είχε ακρίβεια 0,78 και βαθμολογία ROC AUC 0,87.

Τα αποτελέσματα των μοντέλων που εκτελούνται με την Αναπαράσταση One Hot Encoding - Bigram έδειξαν κάποιες βελτιώσεις σε σύγκριση με τα μοντέλα που εκτελούνται με την Αναπαράσταση One Hot Encoding - Unigram. Τα μοντέλα Logistic Regression, Support Vector Machine και Multinomial Naive Bayes παρουσίασαν μικρή βελτίωση στην ακρίβειά τους, σημειώνοντας 0,91, 0,91 και 0,90 αντίστοιχα. Η βαθμολογία ROC AUC για αυτά τα μοντέλα παρέμεινε η ίδια στο 0,97. Το Random Forests δεν παρουσίασε βελτίωση στην ακρίβεια, σημειώνοντας 0,89, αλλά είχε ελαφρώς χαμηλότερη βαθμολογία ROC AUC 0,96.

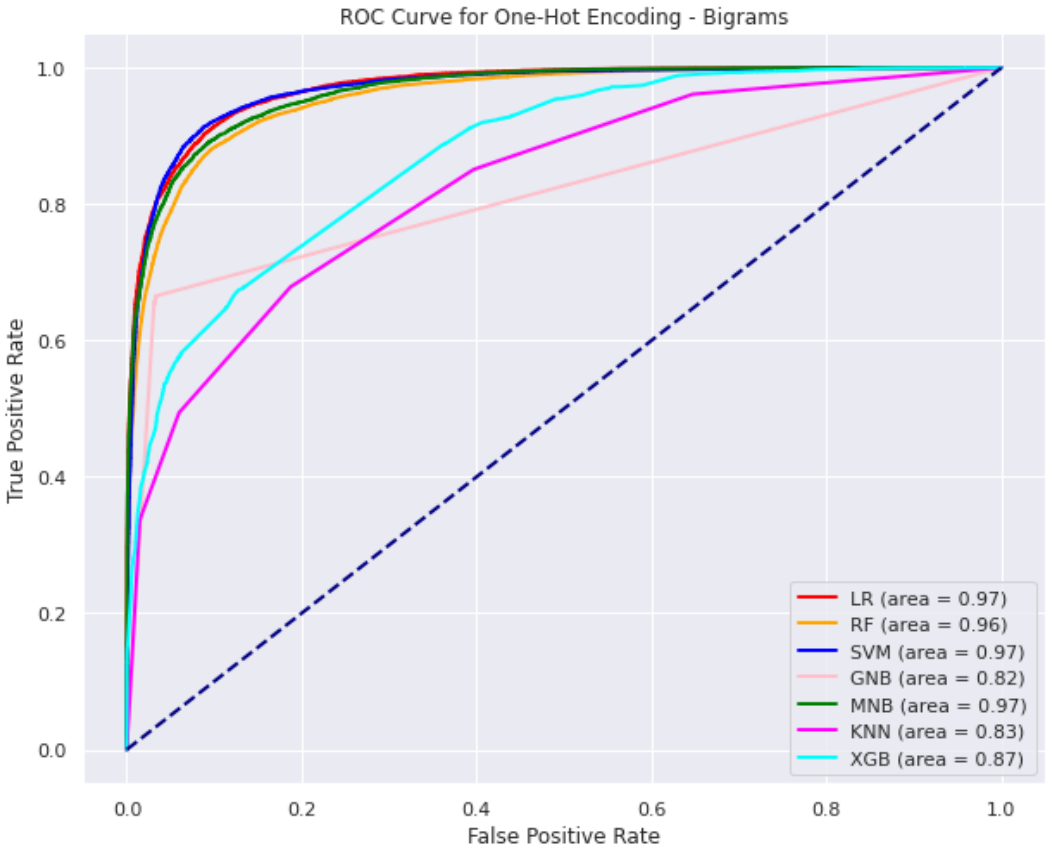
Από την άλλη πλευρά, τα μοντέλα Gaussian Naive Bayes και 5-Nearest Neighbors παρουσίασαν αισθητή βελτίωση στην ακρίβεια, σημειώνοντας 0,79 και 0,74 αντίστοιχα. Ωστόσο, οι βαθμολογίες ROC AUC παρέμειναν σχετικά χαμηλές στο 0,82 και 0,83 αντίστοιχα. Το μοντελο XGBoost Classifier δεν έδειξε βελτίωση στην ακρίβεια ή στη βαθμολογία ROC AUC, σημειώνοντας 0,78 και 0,87 αντίστοιχα.

Η βελτίωση της ακρίβειας για ορισμένα μοντέλα και η αμετάβλητη ακρίβεια για άλλα μπορεί να αποδοθεί στην αναπαράσταση που χρησιμοποιείται. Η αναπαράσταση One Hot Encoding - Bigram λαμβάνει υπόψη τον συνδυασμό δύο γειτονικών λέξεων σε ένα κείμενο και όχι μόνο μία λέξη τη φορά. Αυτή η αναπαράσταση αποτυπώνει περισσότερο πλαίσιο και νόημα στο κείμενο, οδηγώντας σε βελτιωμένη απόδοση για ορισμένα μοντέλα.

Με βάση τα αποτελέσματα, τα μοντέλα που είχαν την καλύτερη απόδοση με την αναπαράσταση Bigram One Hot Encoding είναι τα μοντέλα Logistic Regression, Support Vector Machine και Multinomial Naive Bayes. Αυτά τα μοντέλα πέτυχαν υψηλή ακρίβεια και βαθμολογίες ROC AUC και μπορούν να θεωρηθούν καλές επιλογές για εργασίες ανάλυσης συναισθήματος.

Πίνακας 20: Ακρίβεια για τα μοντέλα του One Hot Encoding - Bigrams

	LR	RF	SVM	GNB	MNB	KNN	XGB
Ακρίβεια	0.91	0.89	0.91	0.79	0.90	0.74	0.78



Σχήμα 16: Καμπύλη ROC για τα μοντέλα One Hot Encoding - Bigrmas

5.3 TF-IDF

Τα αποτελέσματα (όπως φαίνονται στον Πίνακα 21 και στο Διάγραμμα 17) των μοντέλων ανάλυσης συναισθήματος που χρησιμοποιούν την αναπαράσταση TF-IDF δείχνουν μια σειρά επιδόσεων μεταξύ των διαφορετικών μοντέλων. Το μοντέλο Support Vector Machine (SVM) είχε την καλύτερη απόδοση όσον αφορά την ακρίβεια, με βαθμολογία 0,91. Επιπλέον, το μοντέλο SVM είχε επίσης την υψηλότερη βαθμολογία ROC AUC 0,97. Αυτό δείχνει ότι το μοντέλο είναι σε θέση να διακρίνει σωστά μεταξύ θετικών και αρνητικών συναισθημάτων με υψηλή ακρίβεια.

Από την άλλη πλευρά, το μοντέλο Gaussian Naive Bayes (GNB) είχε σχετικά κακή απόδοση σε σύγκριση με τα άλλα μοντέλα, σημειώνοντας μόνο 0,73 ως προς την ακρίβεια και 0,79 ως προς το ROC AUC. Αυτό είναι πιθανό επειδή το GNB είναι ένα σχετικά απλό μοντέλο που προϋποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών, κάτι που μπορεί να μην ισχύει για την ανάλυση συναισθήματος.

Τα μοντέλα Random Forests και Logistic Regression είχαν παρόμοια απόδοση με βαθμολογίες ακρίβειας 0,88 και 0,90 αντίστοιχα και ROC AUC 0,95 και 0,96 αντίστοιχα. Αυτά τα μοντέλα είναι καλά για την εκτέλεση εργασιών δυαδικής ταξινόμησης, όπως η ανάλυση συναισθήματος και συνήθως λειτουργούν καλά με χώρους χαρακτηριστικών υψηλών διαστάσεων.

Τα μοντέλα Multinomial Naive Bayes (MNB) και XGBoost Classifier είχαν σχετικά καλή απόδοση με βαθμολογίες ακρίβειας 0,88 και 0,77 και βαθμολογία ROC AUC 0,95 και 0,88 αντίστοιχα. Το MNB είναι κατάλληλο για προβλήματα ταξινόμησης κειμένου, όπως η ανάλυση συναισθήματος, καθώς λαμβάνει υπόψη τη συχνότητα εμφάνισης των λέξεων, η οποία είναι σημαντική για τον προσδιορισμό του συναισθήματος ενός κειμένου.

Το μοντέλο 5-Narest Neighbors (KNN) είχε τη χαμηλότερη ακρίβεια και βαθμολογία AUC ROC μεταξύ όλων των μοντέλων, με βαθμολογία 0,66 ως προς την ακρίβεια και 0,73 ως προς την AUC ROC. Αυτό μπορεί να οφείλεται στην κατάρα της διάστασης, όπου

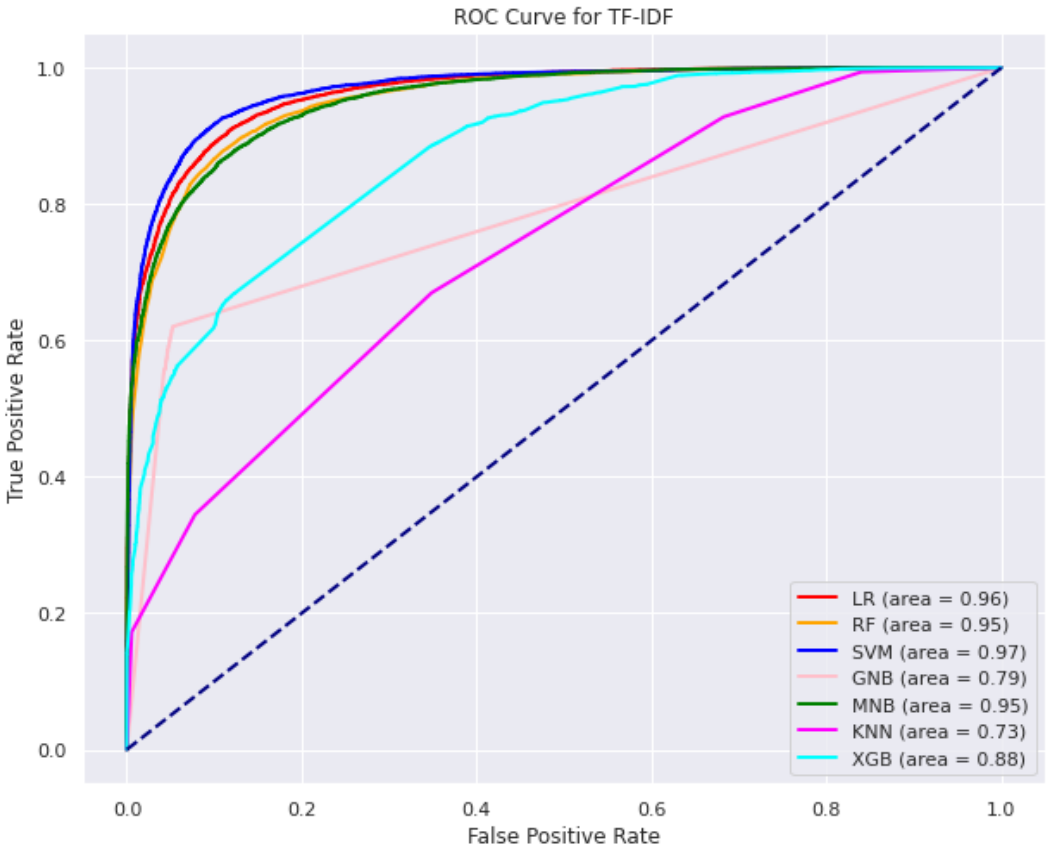
ο αριθμός των χαρακτηριστικών στο χώρο χαρακτηριστικών είναι μεγάλος και το μοντέλο δεν είναι σε θέση να αποτυπώσει αποτελεσματικά τη σχέση μεταξύ των χαρακτηριστικών και της μεταβλητής στόχου.

Κατά τη σύγκριση των αποτελεσμάτων με τις προηγούμενες εκτελέσεις χρησιμοποιώντας την Αναπαράσταση Unigram One Hot Encoding και την One Hot Encoding Bigram Representation, φαίνεται ότι η αναπαράσταση TF-IDF είχε ως αποτέλεσμα σχετικά χαμηλότερη απόδοση μεταξύ ορισμένων μοντέλων. Ωστόσο, αυτό μπορεί να μην συμβαίνει πάντα και η επιλογή της αναπαράστασης θα εξαρτηθεί από το συγκεκριμένο σύνολο δεδομένων και το πρόβλημα που θα λυθεί.

Συμπερασματικά, τα αποτελέσματα των μοντέλων ανάλυσης συναισθήματος που χρησιμοποιούν την αναπαράσταση TF-IDF δείχνουν ότι το μοντέλο SVM είχε την καλύτερη απόδοση όσον αφορά την ακρίβεια και τη βαθμολογία ROC AUC. Άλλα μοντέλα όπως η Logistic Regression και τα Random Forests είχαν επίσης καλή απόδοση, ενώ τα μοντέλα GNB και KNN είχαν σχετικά κακή απόδοση. Η επιλογή της αναπαράστασης θα εξαρτηθεί από τις ιδιαιτερότητες του προβλήματος και του συνόλου δεδομένων που χρησιμοποιείται, αλλά τα αποτελέσματα από αυτό το πείραμα δείχνουν ότι το μοντέλο SVM είναι μια καλή επιλογή για εργασίες ανάλυσης συναισθήματος.

Πίνακας 21: Ακρίβεια για τα μοντέλα του TF-IDF

	LR	RF	SVM	GNB	MNB	KNN	XGB
Ακρίβεια	0.90	0.88	0.91	0.73	0.88	0.66	0.77



Σχήμα 17: Καμπύλη ROC για τα μοντέλα TF-IDF

5.4 Word2Vec

Τα αποτελέσματα (όπως φαίνονται στον Πίνακα ;; και στο Διάγραμμα 18) των μοντέλων ανάλυσης συναισθήματος με την αναπαράσταση Word2Vec δείχνουν ότι αρκετοί αλγόριθμοι μηχανικής μάθησης είχαν καλή απόδοση. Τα μοντέλα με τις καλύτερες επιδόσεις ήταν τα Support Vector Machine, Random Forests και 5-Nearest Neighbours, με βαθμολογίες ακρίβειας 0,84, 0,79 και 0,80, αντίστοιχα. Αυτά τα μοντέλα είχαν επίσης σχετικά υψηλές βαθμολογίες ROC AUC, υποδεικνύοντας ότι είναι σε θέση να διακρίνουν αποτελεσματικά μεταξύ θετικών και αρνητικών συναισθημάτων.

Η απόδοση των Logistic Regression και Gaussian Naive Bayes ήταν σχετικά χαμηλότερη σε σύγκριση με τα άλλα μοντέλα, με βαθμολογίες ακρίβειας 0,78 και 0,67, αντίστοιχα. Το XGBoost Classifier είχε αξιοπρεπή απόδοση, με βαθμολογία ακρίβειας 0,76 και βαθμολογία ROC AUC 0,84.

Ένας λόγος για τη διαφορά στην απόδοση μεταξύ των μοντέλων θα μπορούσε να είναι η διαφορετική φύση των αναπαραστάσεων που χρησιμοποιούνται. Η αναπαράσταση Word2Vec περιλαμβάνει την εκπαίδευση ενός νευρωνικού δικτύου για να μάθει τις σχέσεις μεταξύ των λέξεων, με αποτέλεσμα μια πυκνή διανυσματική αναπαράσταση κάθε λέξης. Αυτή η αναπαράσταση αποτυπώνει τη σημασιολογική σημασία των λέξεων και είναι πιο κατάλληλη για χρήση σε εργασίες NLP. Ωστόσο, αυτή η αναπαράσταση μπορεί να μην είναι βέλτιστη για όλους τους αλγόριθμους, οδηγώντας σε χαμηλότερη απόδοση για ορισμένα μοντέλα.

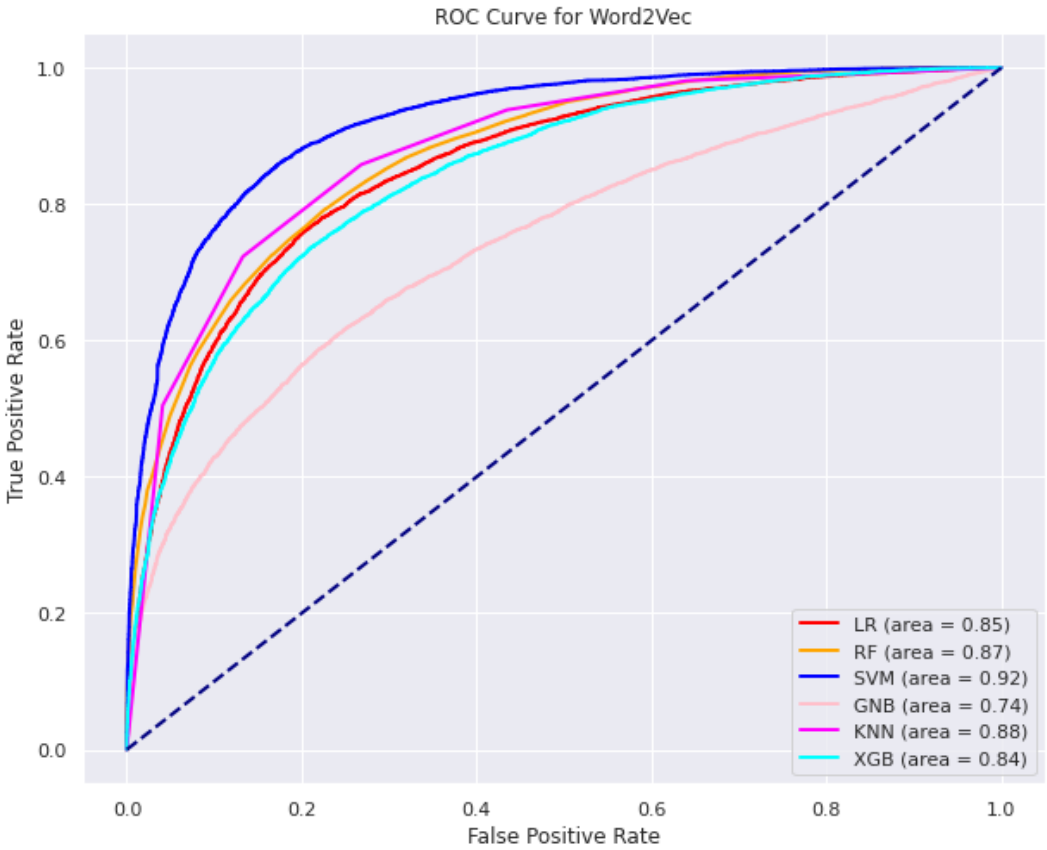
Ένας άλλος παράγοντας που θα μπορούσε να επηρεάσει την απόδοση των μοντέλων είναι η επιλογή των υπερπαραμέτρων. Κάθε αλγόριθμος μηχανικής μάθησης έχει ένα σύνολο υπερπαραμέτρων που πρέπει να ρυθμιστούν προσεκτικά για να επιτευχθεί η βέλτιστη απόδοση. Τα μοντέλα που είχαν καλή απόδοση μπορεί να είχαν καλύτερες ρυθμίσεις υπερπαραμέτρων, ενώ τα μοντέλα που είχαν κακή απόδοση μπορεί να είχαν μη βέλτιστες ρυθμίσεις υπερπαραμέτρων.

Σε σύγκριση με τα προηγούμενα αποτελέσματα με το One Hot Encoding Unigram Representation, One Hot Encoding Bigram Representation και την αναπαράσταση TF-IDF, τα μοντέλα έδειξαν βελτιωμένη απόδοση με την αναπαράσταση Word2Vec. Αυτό δείχνει ότι η πυκνή διανυσματική αναπαράσταση λέξεων που παρέχεται από το μοντέλο Word2Vec ήταν σε θέση να συλλάβει τη σημασιολογική σημασία των λέξεων, οδηγώντας σε καλύτερη απόδοση στην ανάλυση συναισθημάτων.

Συμπερασματικά, τα αποτελέσματα των μοντέλων ανάλυσης συναισθήματος με την αναπαράσταση Word2Vec δείχνουν ότι αρκετοί αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για αυτήν την εργασία και ότι η επιλογή της αναπαράστασης μπορεί να έχει σημαντικό αντίκτυπο στην απόδοση των μοντέλων. Τα μοντέλα με τις καλύτερες επιδόσεις ήταν το Support Vector Machine, τα Random Forests και τα 5-Nearest Neighbours, τα οποία είχαν σχετικά υψηλές βαθμολογίες ακρίβειας και βαθμολογίες ROC AUC.

Πίνακας 22: Ακρίβεια για τα μοντέλα του Word2Vec

	LR	RF	SVM	GNB	KNN	XGB
Ακρίβεια	0.78	0.79	0.84	0.67	0.80	0.76



Σχήμα 18: Καμπύλη ROC για τα μοντέλα Word2Vec

5.5 FastText

Τα αποτελέσματα (όπως φαίνονται στον Πίνακα 23 και στο Διάγραμμα 19) των μοντέλων ανάλυσης συναισθήματος δείχνουν ότι τα μοντέλα με την καλύτερη απόδοση ήταν το SVM, με ακρίβεια 0,85 και βαθμολογία ROC AUC 0,93, ακολουθούμενα από Logistic Regression, με ακρίβεια 0,78 και βαθμολογία ROC AUC 0,86. Τα άλλα μοντέλα, Random Forests, GNB, KNN και XGBoost Classifier, είχαν χαμηλότερη βαθμολογία ακρίβειας και ROC AUC, με το μοντέλο με τη χειρότερη απόδοση να είναι το GNB, με ακρίβεια 0,67 και βαθμολογία ROC AUC 0,74.

Η απόδοση των μοντέλων μπορεί να εξηγηθεί από τη φύση της αναπαράστασης FastText. Το FastText είναι μια τεχνική ενσωμάτωσης λέξεων που αναπαριστά τις λέξεις ως διανύσματα με βάση το περιβάλλον τους. Αυτή η αναπαράσταση αποτυπώνει τη σημασιολογική σημασία των λέξεων και τις σχέσεις τους με άλλες λέξεις, καθιστώντας την χρήσιμη για εργασίες ανάλυσης συναισθήματος. Ωστόσο, ενδέχεται να μην καταγράφει όλες τις αποχρώσεις και την πολυπλοκότητα του συναισθήματος, γεγονός που μπορεί να έχει ως αποτέλεσμα χαμηλότερη απόδοση σε σύγκριση με άλλες αναπαραστάσεις όπως το One Hot Encoding Unigram, το One Hot Encoding Bigram, το TF-IDF και το Word2Vec.

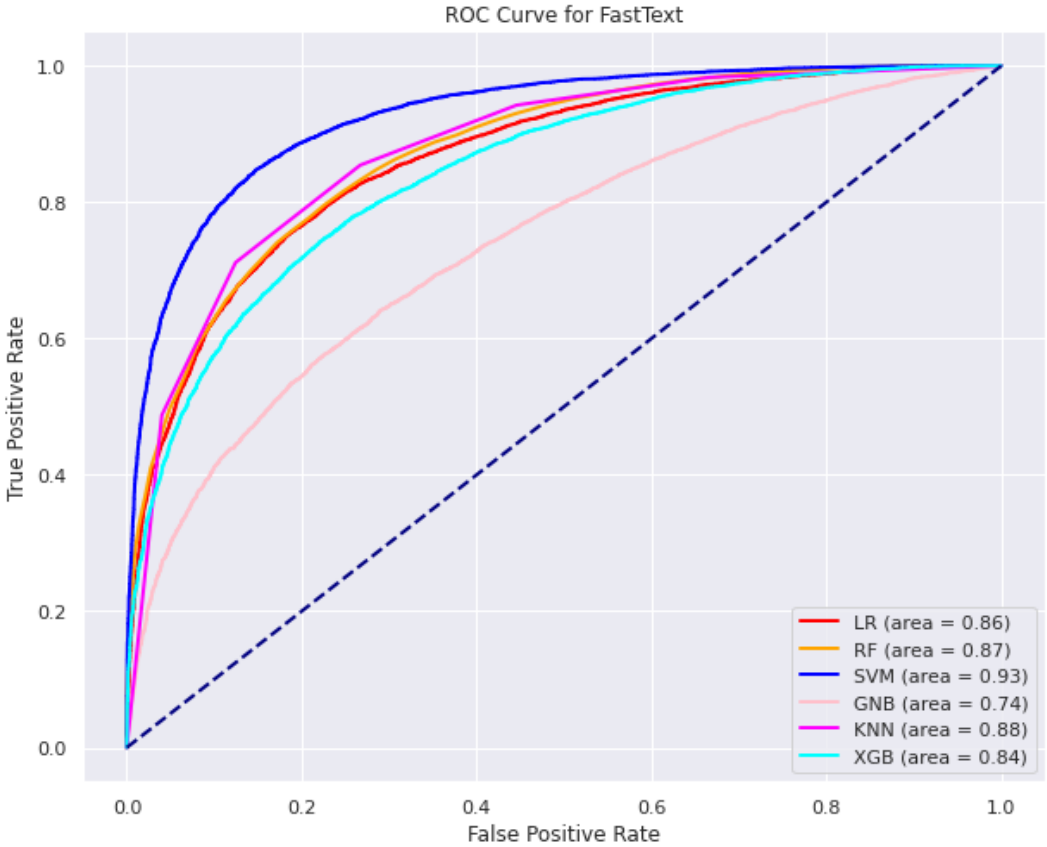
Τα αποτελέσματα των μοντέλων ανάλυσης συναισθήματος με αναπαράσταση FastText είναι συγκρίσιμα με τα αποτελέσματα των μοντέλων με τις άλλες αναπαραστάσεις. Η αναπαράσταση One Hot Encoding Unigram έδειξε παρόμοια αποτελέσματα με την αναπαράσταση FastText, με ελαφρώς χαμηλότερη ακρίβεια και βαθμολογίες ROC AUC. Η αναπαράσταση One Hot Encoding Bigram είχε υψηλότερη ακρίβεια και βαθμολογίες ROC AUC σε σύγκριση με το FastText, ενώ οι αναπαραστάσεις TF-IDF και Word2Vec έδειξαν τα καλύτερα αποτελέσματα, με την υψηλότερη ακρίβεια και βαθμολογίες ROC AUC.

Συμπερασματικά, τα αποτελέσματα των μοντέλων ανάλυσης συναισθήματος με αναπαράσταση FastText δείχνουν ότι αυτή η αναπαράσταση μπορεί να χρησιμοποιηθεί

αποτελεσματικά για εργασίες ανάλυσης συναισθήματος, αν και μπορεί να μην παρέχει την υψηλότερη απόδοση σε σύγκριση με άλλες αναπαραστάσεις. Τα μοντέλα με την καλύτερη απόδοση ήταν το SVM και η Logistic Regression και τα αποτελέσματά τους είναι συγκρίσιμα με τα αποτελέσματα των μοντέλων που εκπαιδεύτηκαν με τις άλλες αναπαραστάσεις. Η επιλογή της αναπαράστασης θα εξαρτηθεί τελικά από τις συγκεκριμένες απαιτήσεις της εργασίας ανάλυσης συναισθήματος και είναι σημαντικό να ληφθούν υπόψη πολλαπλές αναπαραστάσεις προκειμένου να καθοριστεί η καλύτερη προσέγγιση για μια δεδομένη εργασία.

Πίνακας 23: Ακρίβεια για τα μοντέλα του FastText

	LR	RF	SVM	GNB	KNN	XGB
Ακρίβεια	0.78	0.79	0.85	0.67	0.80	0.76



Σχήμα 19: Καμπύλη ROC για τα μοντέλα FastText

6 Συμπεράσματα και Προτάσεις

Στην παρούσα διπλωματική, αναζητήθηκε ο καλύτερος συνδοιασμός διανυσματικής αναπαράστασης και μοντέλου μηχανικής μάθησης για την ανάλυση συναισθήματος με υλικό από το Twitter. Με βάση την παραπάνω έρευνα, τα καλύτερα αποτελέσματα επιτεύχθηκαν χρησιμοποιώντας One Hot Unigrams, One Hot Bigrams και TF-IDF ως ενσωματώσεις λέξεων και Support Vector Machine ως μέθοδο Machine Learning.

Σε περαιτέρω συζήτηση, θα μπορούσαν να διερευνηθούν τυχόν περιορισμοί του μοντέλου, όπως η ανάγκη για υψηλούς υπολογιστικούς πόρους ή η πρόκληση αντιμετώπισης του σαρκασμού και της ειρωνείας στα δεδομένα Twitter.

Τέλος, για την μετεπειτα έρευνα θα μπορούσαν να διευρευνηθούν τα παρακάτω :

- **Πιο προηγμένες ενσωματώσεις λέξεων**

  Ενώ έχουχρησιμοποιήθεί πολλές δημοφιλείς ενσωματώσεις λέξεων, σκόπιμο θα ήταν να χρησιμοποιηθούν και πιο προηγμένες τεχνικές όπως BERT ή ELMo, οι οποίες έχει αποδειχθεί ότι παράγουν καλύτερα αποτελέσματα για ορισμένες εργασίες επεξεργασίας φυσικής γλώσσας.

- **Εκμάθηση συνόλου**

  Θα μπορούσε να χρησιμοποιηθεί ένα σύνολο πολλών μεθόδων μηχανικής εκμάθησης, όπως ο ταξινομητής ψηφοφορίας ή στοίβαξης, για να βελτιωθεί η απόδοση του μοντέλου.

- **Συντονισμός υπερπαραμέτρων**

  Ο συντονισμός των υπερπαραμέτρων των αλγορίθμων μηχανικής εκμάθησης μπορεί να βοηθήσει στο να αξιοποιηθεί στο έπακρο το μοντέλο. Η χρήση τεχνικών όπως η αναζήτηση πλέγματος ή η τυχαία αναζήτηση για να βρεθούν οι βέλτιστες υπερπαράμετροι για τον συγκεκριμένο αλγόριθμο.

Βιβλιογραφία/Αρθρογραφία

[1] Ευρωπαϊκό Συμβούλιο Συμβούλιο της Ευρωπαϊκής Ένωσης (2023) Κλιματική αλλαγή: Η δράση της ΕΕ. [Online] Available from: https://www.consilium.europa.eu/el/policies/climate-change/. [Accessed: 07 Φεβρουαρίου 2023].

[2] NRDC (2022) Are the Effects of Global Warming really that Bad. [Online] Available from: https://www.nrdc.org/stories/are-effects-global-warming-really-ba [Accessed: 23 May 2022].

[3] Saad, S.E. and Yang, J. (2019) Twitter Sentiment Analysis Based on Ordinal Regression. IEEE Access, 7, p.p. 163677-163685.
Doi: 10.1109/ACCESS.2019.2952127

[4] Lin, N. (2001) Social capital : A Theory of Social Structure and Action, 19th ed., New York: Cambridge University Press.

[5] Karpf, D. (2012) Collective Action in Organizations: Interaction and Engagement in an Era of Technological Change, by Bruce Bimber, Andrew J. Flanagin, and Cynthia Stohl, 4, p.p. 463-466.

[6] Kaplan, A. M. and Haenlein, M. (2010) Users of the World, Unite! The Challenges and Opportunities of Social Media, Business Horizons, 53(1), p.p. 59-68.
Doi: 10.1016/j.bushor.2009.09.003

[7] Amedie, J. (2015) The Impact of Social Media on Society. SantaClaraUniversity. 2. [Online] Available from: https://scholarcommons.scu.edu/engl_176/2 [Accessed: 09 March 2015]

[8] Aral, S., Roy, D. and Vosoughi, S. (2018) The spread of true and false news online, 359(6380), p.p. 1146-1151.
Doi: 10.1126/science.aap9559

[9] Twenge, J. M. and Campbell, W. K (2009) The narcissism epidemic: Living in the age of entitlement, 1st ed., New York : Free Press.

[10] Carpini, M. X. D. and Keeter, S. (1996). What Americans know about politics and why it matters, 2th ed., New Haven: Yale University Press

[11] Rogers, R. (2013) Debanalizing Twitter: The transformation of an object of study, Association for Computing Machinery, 1, p.p. 356-365.

[12] Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V. and García Villalba, L. J. (2019). Using Twitter Data to Monitor Natural Disaster Social Dynamics: A Recurrent Neural Network Approach with Word Embeddings and Kernel Density Estimation,Multidisciplinary Digital Publishing Institute, 19(7), p.p. 1-13.
Doi: 10.3390/s19071746

[13] Gerguri, D. (2016). Political power of social media in Kosovo. Romanian Review of Political Sciences and International Relations, 13(1), p.p. 95-111.

[14] Weller, K., Bruns, A., Burgess, J., Mahrt, M. and Puschmann, C. (Eds.) (2013) Twitter and society, New York: Peter Lang, p.p. 4

[15] Kurka, D. B., Godoy, A. and Von Zuben, F. J. (2015) Online social network analysis: A survey of research applications in computer science, arXiv preprint arXiv:1504.05655.

[16] Fownes, J. R., Yu, C. and Margolin, D. B. (2018) Twitter and climate change, Sociology Compass, 12(6), e12587. https://doi.org/10.1111/soc4.12587

[17] Han, J., Pei, J. and Tong, H. (2022) Data mining: concepts and techniques, 4th ed., Waltham: Morgan Kaufmann.

[18] Aggarwal, C.C. and Reddy, C.K. (2014) Data Clustering: Algorithms and Applications, 1st ed., New York: Chapman and Hall/CRC.

[19] Witten, I. H. and Frank, E., (2005) Data Mining:Practical Machine Learning Tools and Techniques,2th ed., San Francisco: Morgan Kaufmann.

[20] Wong, K. C. (2016) Big data analytics in genomics, 1th ed., Springer.
Doi: 10.1007/978-3-319-41279-5

[21] Zikopoulos, P., Eaton, C., deRoos, D., Deutsch, T. and Lapis, G. (2011) Understanding big data: Analytics for enterprise class hadoop and streaming data, 1th ed., USA: McGraw-Hill Osborne Media.

[22] Bughin, J., Chui, M. and Manyika, J. (2010). Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. McKinsey quarterly, 56(1), p.p. 75-86.

[23] Shmueli, G., Bruce, P. C. and Patel, N. R. (2016) Data mining for business analytics: Concepts, techniques, and applications with XLMiner, 3th ed., John Wiley and Sons.

[24] Gandomi, A. and Haider, M. (2015) Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), p.p. 137-144.

[25] Wu, X., Zhu, X., Wu, G. Q. and Ding, W. (2013) Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), p.p. 97-107.

[26] Sagiroglu, S. and Sinanc, D. (2013) Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS), San Diego : IEEE, p.p. 42-47.
Doi: 10.1109/CTS.2013.6567202.

[27] Anuradha, J. (2015) A brief introduction on Big Data 5Vs characteristics and Hadoop technology, Procedia computer science, 48, p.p. 319-324.

[28] Twitter Blog (2012) Dispatch from the Denver Debate. [Online] Available from: https://blog.twitter.com/en_us/a/2012/dispatch-from-the-denver-debate. [Accessed: 04 October 2012].

[29] Chollet, F. (2018) Deep learning with Python, 1th ed., New York: Manning.

[30] Han, J., Kamber, M. and Pei, J. (2012). Data mining concepts and techniques, 3th ed., Waltham: MorganKaufmann.

[31] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) From data mining to knowledge discovery in databases, AI magazine, 17(3), p.p.37-40.

[32] Asur, S. and Huberman, B. A. (2010) Predicting the future with social media, In 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, 1, p.p. 492-499.

[33] Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M. (2011) Election forecasts with Twitter: How 140 characters reflect the political landscape, Social

science computer review, 29(4), p.p. 402-418.

[34] Bollen, J., Mao, H. and Zeng, X. (2011) Twitter mood predicts the stock market, Journal of computational science, 2(1), p.p. 1-8.

[35] Slamet, C., Andrian, R., Maylawati, D. S. A., Darmalaksana, W. and Ramdhani, M. A. (2018) Web scraping and Naïve Bayes classification for job search engine, In IOP conference series: Materials science and engineering, 288(1), p. 012038, IOP Publishing.

[36] Glez-Peña, D., Lourenço, A., López-Fernández, H., Reboiro-Jato, M. and Fdez-Riverola, F. (2014) Web scraping technologies in an API world, Briefings in bioinformatics, 15(5), p.p. 788-797.

[37] AIMultiple (2023) Top 7 Differences of Web Scraping vs API in 2023. [Online] Available from:
https://research.aimultiple.com/web-scraping-vs-api/.
[Accessed: 28 January 2023].

[38] Pang, B., Lee, L., and Vaithyanathan, S. (2002) Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint cs/0205070.

[39] Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M. and Lee, S. (2017) Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression, In 2017 international conference on information and communication technology convergence (ICTC 2017), IEEE, p.p. 138-140.

[40] Joyce, B. and Deng, J. (2017) Sentiment analysis of tweets for the 2016 US presidential election, In 2017 ieee mit undergraduate research technology conference (urtc), IEEE, p.p. 1-4.

[41] Chatzakou, D. and Vakali, A. (2015) Harvesting opinions and emotions from social media textual resources, IEEE Internet Computing, 19(4), p.p. 46-50.

[42] Kim, S. M. and Hovy, E. (2004) Determining the sentiment of opinions, In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, p.p. 1367-1373.

[43] Maynard, D. G. and Greenwood, M. A. (2014) Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis., In Lrec 2014 proceedings, ELRA.

[44] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep learning, MIT press.

[45] Sadiku, M., Ashaolu, T. J., Ajayi-Majebi, A. and Musa, S. M. (2021) Artificial Intelligence in Social Media, International Journal of Scientific Advances, 2(1), p.p. 15-20.

[46] Kelleher, J. D., Mac Namee, B., and Darcy, A. (2020) Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies, MIT press.

[47] Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009) The elements of statistical learning: data mining, inference, and prediction , 2th ed., New York: Springer, p.p. 1-758.

[48] Bishop, C. M. and Nasrabadi, N. M. (2006) Pattern recognition and machine learning, 4(4). New York: Springer, p.p. 738.

[49] Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspec-

tives, and prospects. Science, 349(6245), p.p. 255-260.

[50] LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning, Nature, 521(7553), p.p. 436-444.

[51] Kaelbling, L. P., Littman, M. L. and Moore, A. W. (1996) Reinforcement learning: A survey, Journal of artificial intelligence research, 4, p.p. 237-285.

[52] Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory, Neural computation, 9(8), p.p. 1735-1780.

[53] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research, 15(1), p.p. 1929-1958.

[54] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. and Polosukhin, I. (2017) Attention is all you need, Advances in neural information processing systems, p.p. 30.

[55] Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017) Imagenet classification with deep convolutional neural networks, Communications of the ACM, 60(6), p.p. 84-90.

[56] He, K., Zhang, X., Ren, S. and Sun, J. (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, In Proceedings of the IEEE international conference on computer vision, p.p. 1026-1034.

[57] Joachims, T. (2005) Text categorization with support vector machines: Learning with many relevant features, In Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceed-

ings , Berlin, Heidelberg: Springer Berlin Heidelberg, p.p. 137-142.

[58] Ramage, D., Hall, D., Nallapati, R. and Manning, C. D. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, In Proceedings of the 2009 conference on empirical methods in natural language processing, p.p. 248-256.

[59] Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) Latent dirichlet allocation, Journal of machine Learning research, 3(1), p.p. 993-1022.

[60] Aggarwal, C. C. and Zhai, C. (2012) A survey of text classification algorithms, Mining text data, p.p. 163-222.

[61] Aliwy, A. H. and Ameer, E. A. (2017) Comparative study of five text classification algorithms with their improvements, International Journal of Applied Engineering Research, 12(14), p.p. 4309-4319.

[62] Chen, Y. (2015) Convolutional neural network for sentence classification (Master's thesis, University of Waterloo).

[63] Salton, G. (1983) Introduction to modern information retrieval, McGraw-Hill.

[64] Bojanowski, P., Grave, E., Joulin, A. and Mikolov, T. (2017) Enriching word vectors with subword information, Transactions of the association for computational linguistics, 5,p.p. 135-146.

[65] Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) Efficient estimation of word representations in vector space, arXiv:1301.3781.

[66] Pennington, J., Socher, R. and Manning, C. D. (2014, October) Glove: Global

vectors for word representation, In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), p.p. 1532-1543.

[67] Joulin, A., Grave, E., Bojanowski, P. and Mikolov, T. (2016) Bag of tricks for efficient text classification, arXiv:1607.01759.

[68] Ilić, S., Marrese-Taylor, E., Balazs, J. A., and Matsuo, Y. (2018) Deep contextualized word representations for detecting sarcasm and irony, arXiv:1809.09795.

[69] Devlin, J., Chang, M. W., Lee, K. and Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv:1810.04805.

[70] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D. and Stoyanov, V. (2019) Roberta: A robustly optimized bert pretraining approach, arXiv:1907.11692.

[71] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. and Soricut, R. (2019) Albert: A lite bert for self-supervised learning of language representations, arXiv:1909.11942.

[72] Hagenaars, J. A. and McCutcheon, A. L. (2002) Applied latent class analysis, Cambridge University Press, p.p. 163-287.

[73] Agresti, A. (2012) Categorical data analysis, John Wiley and Sons, 792(1), p.p. 165- 635.

[74] Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009) The elements of statistical learning: data mining, inference, and prediction , 2th ed., New York: Springer, p.p. 1-758.

[75] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An introduction

to statistical learning , 2th ed., New York: Springer.

[76] Breiman, L. (2001) Random forests. Machine learning, Springer, 45 (1), p.p. 5-32.

[77] Ho, T. K. (1998) The random subspace method for constructing decision forests, IEEE transactions on pattern analysis and machine intelligence, 20(8), p.p. 832-844.

[78] Bansal, H., Shrivastava, G., Nguyen, G. N. and Stanciu, L. M. (Eds.). (2018) Social network analytics for contemporary business organizations,1th ed., Pennsylvania: IGI Global.

[79] Vapnik, V. (1999) The nature of statistical learning theory, 1th ed., New York: Springer.

[80] Burges, C. J. (1998) A tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery, 2(2), p.p. 121-167.

[81] Schölkopf, B., Smola, A. J. and Bach, F. (2002), Learning with kernels: support vector machines, regularization, optimization, and beyond, MIT press, 2(1), p.p. 187-421.

[82] Rish, I. (2001) An empirical study of the naive Bayes classifier, In IJCAI 2001 workshop on empirical methods in artificial intelligence, 3(22), p.p. 41-46.

[83] Zhang, H. (2004) The optimality of naive Bayes, Aa, 1(2), p.p. 1-6.

[84] Alpaydin, E. (2020) Introduction to machine learning. 4th ed., MIT press.

[85] Chen, T. and Guestrin, C. (2016) Xgboost: A scalable tree boosting system, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794.

[86] Jurczuk, K., Czajkowski, M. and Kretowski, M. (2021) Multi-GPU approach to global induction of classification trees for large-scale data mining, Applied Intelligence, 51(8), p.p. 5683-5700.

[87] Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net, Journal of the royal statistical society: series B (statistical methodology), 67(2), p.p. 301-320.

[88] Domingos, P. (2015) The master algorithm: How the quest for the ultimate learning machine will remake our world, 1th ed., New York : Basic Books.

[89] Friedman, J. H. (2001) Greedy function approximation: a gradient boosting machine, Annals of statistics, p.p. 1189-1232.

Α΄ Κώδικας

Α΄.1 Εξόρυξη Δεδομένων

Α΄.1.1 Λογαριασμοί και Hashtags

```
Hashtags = ['climate', 'climate action', 'climate change', 'climate crisis', 'global warming', 'climate emergency', 'climate hysteria', 'climate hoax',
            'climate alarmism', 'climate change alarmism', 'climate change hoax', 'climate change hysteria',
                    'global warming hysteria', 'climate change fraud', 'climate emergency hoax', 'climate change scam', 'global warming hoax', 'global warming alarmism',
                    'global warming cult', 'global warming fraud', 'global warming scam', 'global warming nonsense',
                    'climate fraud', 'climate scam', 'climate cult', 'climate nonsense', 'climate change nonsense',
                    'climate change cult',  'COP24', 'ParisAgreement', 'climate realist', 'agw']
```

Σχήμα 20: Hashtags

```
Believers = ['climateWWF', 'UNFCCC', 'WWF', 'GreenpeaceUK', 'MichaelEMann', 'ayanaeliza', 'FrannyArmstrong', 'pooley_eric', 'jrockstrom', 'elonmusk', 'bruneski',
             'xiyebastida', 'vict_barrett', 'johnpauljos', 'israhirsi', 'luisamneubauer', 'severnsuzuki', 'iraghislain', 'alexandriav2005', 'kehkashanbasu',
             'felixfinkbeiner', 'jeromefosterii', 'laurajoparker', 'jamie_margolin', 'fisher_danar']
```

Σχήμα 21: Υποστηρικτές

```
Deniers = [ 'HeartlandInst', 'KurtSchlichter', 'BjornLomborg', 'JunkScience', 'TomFitton', 'brithume', 'Carbongate', 'CatoInstitute', 'DineshDSouza',
           'GovMikeHuckabee', 'GrrrGraphics', 'JimInhofe', 'inhofeforsenate','InhofePress',  'LtGovTimGriffin', 'LanceForman', 'Truth_Freedom23',
           'LisaMarieBoothe', 'mattwridley', 'MaximeBernier', 'TeamBachmann', 'MicheleBachmann', 'Origsmartassam', 'PrisonPlanet', 'realDonaldTrump',
           'POTUS45', 'joebarto', 'RonPaul', 'SeibtNaomi', 'MelissaAFrancis', 'AGToddRokita', 'ToddRokitaIN', 'ToddRokita']

```

Σχήμα 22: Αρνητές

```
list_of_df = []
# Δημιουργούμε τους δυνατους συνδυασμους των Deniers και Hashtags
for element in tqdm(list(itertools.product(*[Deniers, Hashtags]))):
  # Εφαρμόζουμε την εντολή εξόρυξης των tweets που περιέχει ένα συγκεκριμένο hashtag
  # και το οποίο προέρχεται από το συγκεκριμένο Denier χρήστη
  os.system("snscrape --jsonl --max-results {} twitter-search ' {} from:{}'> tweets.json".format(1000000, element[1], element[0]))
  # Δημιουργούμε ένα dataframe με τα tweets που μόλις εξορύξαμε
  tweets_df1 = pd.read_json('tweets.json', lines=True)
  # Αν δεν έχουμε βρεί tweets, συνεχίζουμε την αναζήτηση
  if len(tweets_df1) == 0 :
    print(element)
    continue
  # Συλλέγουμε κάθε dataframe του κάθε συνδιασμου των Deniers και Hashtags
  list_of_df.append(tweets_df1)

# Αφου τελειωσει η διαδικασία εξόρυξης των tweets όλων των συνδυασμών, δημιουργούμε
# ένα τελικό dataframe στο οποίο κρατάμε μόνο το περιεχόμενο του κάθε tweet.
# Επίσης προσθέτουμε σε αυτό το dataframe την κατηγοριοποίηση αυτών των tweets
# που είναι αρνητική (0)

denier_tweets = pd.concat(list_of_df)
denier_tweets['sentiment'] = 0
```

Σχήμα 23: Συλλογή αρνητικών tweet

```python
list_of_df = []
# Δημιουργούμε τους δυνατους συνδυασμους των Believers και Hashtags
for element in tqdm(itertools.product(*[Believers, Hashtags])):
  # Εφαρμόζουμε την εντολή εξόρυξης των tweets που περιέχει ένα συγκεκριμένο hashtag
  # και το οποίο προέρχεται από το συγκεκριμένο Believer χρήστη
  os.system("snscrape --jsonl --max-results {} twitter-search ' {} from:{}'> tweets.json".format(10000, element[1], element[0]))
  # Δημιουργούμε ένα dataframe με τα tweets που μόλις εξορυξαμε
  tweets_df1 = pd.read_json('tweets.json', lines=True)
  # Αν δεν έχουμε βρεί tweets, συνεχίζουμε την αναζήτηση
  if len(tweets_df1) == 0 :
    continue
  # Συλλέγουμε κάθε dataframe του κάθε συνδιασμου των Believers και Hashtags
  list_of_df.append(tweets_df1)

# Αφου τελειωσει η διαδικασία εξόρυξης των tweets όλων των συνδυασμών, δημιουργούμε
# ένα τελικό dataframe στο οποίο κρατάμε μόνο το περιεχόμενο του κάθε tweet.
# Επίσης προσθέτουμε σε αυτό το dataframe την κατηγοριοποίηση αυτών των tweets
# που είναι θετική (1)

believer_tweets = pd.concat(list_of_df)
believer_tweets = believer_tweets[['content']]
believer_tweets['sentiment'] = 1
```

Σχήμα 24: Συλλογή θετικών tweet

```python
final_dataset = pd.concat([believer_tweets, denier_tweets])
```

Σχήμα 25: Τελικό dataset

```python
sns.factorplot(x="sentiment", data=df, kind="count", size=6, aspect=1.5, palette="PuBuGn_d")
plt.show()

# Το dataset μας δεν φαίνεται να είναι ισορροπημένο αλλά η διαφορά μεταξύ κατηγοριων
# δεν φαίνεται να είναι μεγάλη
```

Σχήμα 26: Ισοροπία dataset

```python
'''
Θα δημιουργησουμε ενα dataframe στο οποίο θα έχουμε το τα εξής :
- Πόσες απαντήσεις
- Πόσα retweets
- Πόσα likes
- Πόσες αναφορές σε άλλα tweets
- Πόσες λέξεις υπάρχουν στο tweet
- Πόσες αναφορές σε άλλους λογαριασμούς υπαρχουν στο tweet
- Πόσα hashtags υπάρχουν στο tweet
- Πόσες λέξεις είναι γραμμένες με κεφαλαία γράμματα στο tweet(Συνήθως εκφράζουν μια οργή/ουρλιαχτό στον γραπτό λόγο)
- Πόσα σημεία στίξης με σημασία, όπως το θαυμαστικό ή το ερωτηματικό
- Πόσες sites/ιστοσελίδες είναι παραθετηένεο στο tweet
- Πόσα emojis είναι στο tweet
'''

dataframe = []
for i in range(len(df)):
  final_row = []
  final_row.append(df.replyCount[i])
  final_row.append(df.retweetCount[i])
  final_row.append(df.likeCount[i])
  final_row.append(df.quoteCount[i])
  final_row.append(len(re.findall(r'\w+', df['content'][i]))) # λέξεις
  final_row.append(len(re.findall(r'@\w+', df['content'][i]))) # αναφορές
  final_row.append(len(re.findall(r'#\w+', df['content'][i]))) # hashtags
  final_row.append(len(re.findall(r'\b[A-Z]{2,}\b', df['content'][i]))) # κεφαλαία
  final_row.append(len(re.findall(r'!|\?', df['content'][i]))) # σημεία στίξης
  final_row.append(len(re.findall(r'http.?://[^\s]+[\s]?', df['content'][i]))) # ιστοσελίδες
  final_row.append(len(re.findall(r':[a-z_&]+:', emoji.demojize(df['content'][i])))) # emojis
  dataframe.append(final_row)

df_eda = pd.DataFrame(columns=['replies', 'retweets', 'likes', 'quotes', 'words', 'mentions', 'hashtags', 'capital words', 'marks', 'sites', 'emojis'], data = dataframe)
df_eda['sentiment'] = df.sentiment
df_eda.dropna(inplace = True)
```

Σχήμα 27: Dataset με περαιτέρω πληροφορίες

```
table = df_eda.groupby('sentiment')["replies"].describe()
print(table)
bins = np.arange(df_eda["replies"].min(), 15)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "replies", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 28: Replies

```
table = df_eda.groupby('sentiment')["retweets"].describe()
print(table)
bins = np.arange(df_eda["retweets"].min(), 15)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "retweets", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 29: Retweets

```
table = df_eda.groupby('sentiment')["likes"].describe()
print(table)
bins = np.arange(df_eda["likes"].min(), 50)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "likes", kde=False, norm_hist=True, bins=bins)
g.set(xticks=np.arange(df_eda["likes"].min(), 50, 10))
plt.show()
```

Σχήμα 30: Likes

```
table = df_eda.groupby('sentiment')["quotes"].describe()
print(table)
bins = np.arange(df_eda["quotes"].min(), 5)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "quotes", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 31: Quotes

```
table = df_eda.groupby('sentiment')["words"].describe()
print(table)
bins = np.arange(df_eda["words"].min(), 60)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "words", kde=False, norm_hist=True, bins=bins)
g.set(xticks=np.arange(0, 60, 10))
plt.show()
```

Σχήμα 32: Words

```
table = df_eda.groupby('sentiment')["mentions"].describe()
print(table)
bins = np.arange(df_eda["mentions"].min(), 4)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "mentions", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 33: Mentions

```
table = df_eda.groupby('sentiment')["hashtags"].describe()
print(table)
bins = np.arange(df_eda["hashtags"].min(), 4)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "hashtags", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 34: Hashtags

```
table = df_eda.groupby('sentiment')["capital words"].describe()
print(table)
bins = np.arange(df_eda["capital words"].min(), 4)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "capital words", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 35: Capital words

```
table = df_eda.groupby('sentiment')["marks"].describe()
print(table)
bins = np.arange(df_eda["marks"].min(), 4)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "marks", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 36: Marks words

```
table = df_eda.groupby('sentiment')["sites"].describe()
print(table)
bins = np.arange(df_eda["sites"].min(), 4)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "sites", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 37: Sites words

```
table = df_eda.groupby('sentiment')["emojis"].describe()
print(table)
bins = np.arange(df_eda["emojis"].min(), 4)
g = sns.FacetGrid(df_eda, col='sentiment', size=6, hue='sentiment', palette="PuBuGn_d")
g = g.map(sns.distplot, "emojis", kde=False, norm_hist=True, bins=bins)
g.set(xticks=bins)
plt.show()
```

Σχήμα 38: Emojis words

```
cleaned_text = []
for i in tqdm(range(len(df))):
  cleaned_text.append(re.sub(r'@\w+', '',df.content[i]))
df_clean['cleaned_text'] = cleaned_text
```

Σχήμα 39: Αφαίρεση αναφορών

```
for i in tqdm(range(len(df))):
  df_clean['cleaned_text'][i] = re.sub(r'http.?://[^\s]+[\s]?', '',df_clean['cleaned_text'][i])
```

Σχήμα 40: Αφαίρεση ιστοσελίδων

```
for i in tqdm(range(len(df))):
  df_clean['cleaned_text'][i] = emoji.demojize(df_clean['cleaned_text'][i]).replace(':', ' ')
```

Σχήμα 41: Μετατροπή των emojis

```
punct = string.punctuation
punct = punct + '’'
trantab = str.maketrans(punct, len(punct)*' ')
for i in tqdm(range(len(df))):
  df_clean['cleaned_text'][i] = df_clean['cleaned_text'][i].translate(trantab)
```

Σχήμα 42: Αφαίρεση σημείων στίξεως

```python
def remove_slang(input_text):
  '''
  Αύτη η συνάρτηση παίρνει ώς input το αρχικό κείμενο και μας γυρίζει
  το κείμενο αλλά χωρίς slang (αργκώ)
  '''
  #Χωρίζουμε κάθε λέξη του κειμένου
  input_text = input_text.split(" ")
  j = 0
  # Θα πάρουμε κάθε λέξη του κειμένου και θα ελέγξουμε αν είναι αργκώ ή οχι.
  for word in input_text:
    # έχουμε ήδη ένα αρχείο txt με τις πιθανές slang λέξεις
    with open(r"/content/drive/MyDrive/Mary/slang.txt", 'r') as myCSVfile:
      slang = csv.reader(myCSVfile, delimiter="=")
      # με αυτόν τον τρόπο (delimiter="=") το πρώτο κομμάτι είναι η αργκώ και το δεύτερο η σημασία της

      # Για να μπορέσουμε να ελέγξουμε την καθε λέξη θα πρέπει να αφαιρέσουμε
      # χαρακτήρες οι οποίοι δεν χρειάζονται
      word = re.sub('[^a-zA-Z0-9-_.]', '', word)
      for element in slang:
        # Επειδη στο αρχείο slang οι λέξεις είναι στα κεφαλαία θα μετατρέψουμε
        # κάθε λέξη στα κεφαλαία, για να δούμε αν υπάρχει στο αρχείο slang
        if word.upper() == element[0]:

          # Αν υπάρξει τέτοια λέξη θα την αλλάξουμε με τη σημασία της
          input_text[j] = element[1]
      myCSVfile.close()
      j = j + 1
  return (' '.join(input_text))

```

Σχήμα 43: Συνάρτηση για μετατροπή της αργκό

```python
for i in tqdm(range(len(df))):
  df_clean['cleaned_text'][i] = remove_slang(df_clean['cleaned_text'][i])
```

Σχήμα 44: Μετατροπή των λέξεων αργκό

```python
for i in tqdm(range(len(df))):
  df_clean['cleaned_text'][i] = re.sub('\d+', '',df_clean['cleaned_text'][i])
```

Σχήμα 45: Αφαίρεση αριθμών

```python
for i in tqdm(range(len(df))):
  df_clean['cleaned_text'][i] = df_clean['cleaned_text'][i].lower()
```

Σχήμα 46: Μετατροπή όλων σε μικρά

```
stopwords_list = stopwords.words('english')
stopwords_list.append("n't")
stopwords_list.append('not')
stopwords_list.append('amp')
stopwords_list.append("no")
words = df_clean['cleaned_text'][48065].split()
clean_words = [word for word in words if (word not in stopwords_list) and len(word) > 1]
for i in tqdm(range(len(df))):
  words = df_clean['cleaned_text'][i].split()
  clean_words = [word for word in words if (word not in stopwords_list) and len(word) > 1]
  df_clean['cleaned_text'][i] = " ".join(clean_words)
```

Σχήμα 47: Αφαίρεση των stopwords

```
porter = PorterStemmer()
words = df_clean['cleaned_text'][48065].split()
stemmed_words = [porter.stem(word) for word in words]
for i in tqdm(range(len(df))):
  words = df_clean['cleaned_text'][i].split()
  stemmed_words = [porter.stem(word) for word in words]
  df_clean['cleaned_text'][i] = " ".join(stemmed_words)
```

Σχήμα 48: Εφαρμογή stemmer

```
# αφαίρεση κάποιων επιπλέον χαρακτήρων
for i in tqdm(range(len(df))):
  df_clean['cleaned_text'][i] = df_clean['cleaned_text'][i].replace("\r", " ")
  df_clean['cleaned_text'][i] = df_clean['cleaned_text'][i].replace("\n", " ")
  df_clean['cleaned_text'][i] = df_clean['cleaned_text'][i].replace("    ", " ")
  df_clean['cleaned_text'][i] = df_clean['cleaned_text'][i].replace('"', '')
```

Σχήμα 49: Επιπλέον αφαιρέσεις

```
# Χρησιμοποιώντας το CountVectorizer θα έχουμε μια λίστα με λέξεις(αποκομένες)
cv = CountVectorizer()
# Στο Bag of Words θα συλλέξουμε όλες τις λέξεις/tokens
bow = cv.fit_transform(df_clean['cleaned_text'])
# Θα μετρησουμε όσο συνά εμφανίζζεται το κάθε token
word_freq = dict(zip(cv.get_feature_names(), np.asarray(bow.sum(axis=0)).ravel()))
word_counter = collections.Counter(word_freq)
word_counter_df = pd.DataFrame(word_counter.most_common(20), columns = ['word', 'freq'])
fig, ax = plt.subplots(figsize=(12, 10))
yo = sns.barplot(x="word", y="freq", data=word_counter_df, palette="PuBuGn_d", ax=ax)
yo.set_xticklabels(word_counter_df['word'],rotation=45)
plt.show()
```

Σχήμα 50: Συχνά εμφανιζόμενες λέξεις

```
# Τέλος θα αποθηκεύσουμε την καθαρισμένη βάση δεδομένων
df.to_csv("/content/drive/MyDrive/Mary/CleanDataset.csv")
```

Σχήμα 51: Αποθήκευση ¨καθαρισμένου¨ Dataset

Α΄.2 Μοντέλα αναπαράστασης λέξεων NLP

```
X_train, X_test, Y_train, Y_test = train_test_split(np.array(df.cleaned_text.tolist()), df.sentiment,
                                                    test_size=0.2, random_state=17,
                                                    shuffle=True, stratify = df.sentiment)
```

Σχήμα 52: Δημιουργία των train και test data

Α΄.2.1 One Hot Encoding - Unigrams

```
OHU_vectorizer = CountVectorizer(stop_words = 'english',
                                 binary=True,
                                 min_df = 10)
OHU_vectorizer.fit(X_train)
OHU_train = OHU_vectorizer.transform(X_train)
OHU_test = OHU_vectorizer.transform(X_test)
```

Σχήμα 53: Διανυσματοποιήση One Hot Encoding - Unigrams

Α΄.2.2 One Hot Encoding - Bigrams

```
OHB_vectorizer = CountVectorizer(stop_words = 'english',
                                 binary=True,
                                 min_df = 10,
OHB_vectorizer.fit(X_train)

OHB_train = OHB_vectorizer.transform(X_train)
OHB_test = OHB_vectorizer.transform(X_test)
```

Σχήμα 54: Διανυσματοποίηση One Hot Encoding - Bigrams

Α΄.2.3 TF-IDF

```
tfidf_vectorizer = TfidfVectorizer(stop_words='english',
                                   min_df=10) # used for now for consistency
tfidf_vectorizer.fit(X_train)
tfidf_train = tfidf_vectorizer.transform(X_train)
tfidf_test = tfidf_vectorizer.transform(X_test)

# Construct Scaled Datasets

scaler_tfidf = MaxAbsScaler()
tfidf_train_scaled = scaler_tfidf.fit_transform(tfidf_train)
tfidf_test_scaled = scaler_tfidf.transform(tfidf_test)
```

Σχήμα 55: Διανυσματοποίηση TF-IDF

Α΄.2.4 Word2vec

```
import gensim.downloader as api
wv = api.load('word2vec-google-news-300')

df.cleaned_text = df.cleaned_text.apply(func=lambda tweet: ' '.join([word for word in tweet.split() \
                                                                      if word in wv.vocab]))
print(f"Count of text_clean_known entries which are null: {sum(df.cleaned_text.isnull())}")
print(f"Count of text_clean_known entries which empty: \
{sum(df.cleaned_text.apply(func=lambda x: x.strip() == ''))}")

# Remove both rows with either null or empty `text_clean_known` entries

df = df.loc[(~df.cleaned_text.isnull()), :]
df = df.loc[~df.cleaned_text.apply(func=lambda x: x.strip() == ''), ]

print("After cleaning:", end="\n")
print(f"Count of text_clean_known entries which are null: {sum(df.cleaned_text.isnull())}")
print(f"Count of text_clean_known entries which empty: {sum(df.cleaned_text.apply(func=lambda x: x.strip() == ''))}")

def Word2VectorizeTweet(tweet):
    tweet_vector = np.zeros(shape=(300,), dtype="float32")
    n_vectorizable = 0
    for word in tweet.split():
        try:
#             print(f"Adding {word} to word representation")
            tweet_vector = np.add(tweet_vector, wv.get_vector(word))
            n_vectorizable += 1
        except KeyError:
            print(f"Could not vectorize {word}")
    return (tweet_vector/n_vectorizable)

df.w2v = df.cleaned_text.apply(func=Word2VectorizeTweet)
```

Σχήμα 56: Διανυσματοποίηση Word2Vec

Α΄.2.5 FastText

```
1 !git clone https://github.com/facebookresearch/fastText.git
2 !cd fastText
3 !pip install fastText
4 import fasttext.util
5 fasttext.util.download_model('en', if_exists='ignore')  # English
6 model = fasttext.load_model('cc.en.300.bin')
7
8 def FastTextVectorizeTweet(tweet):
9     tweet_vector = np.zeros(shape=(300,), dtype="float32")
10    n_vectorizable = 0
11    for word in tweet.split():
12        try:
13 #             print(f"Adding {word} to word representation")
14            tweet_vector = np.add(tweet_vector, model.get_word_vector(word))
15            n_vectorizable += 1
16        except KeyError:
17            print(f"Could not vectorize {word}")
18    return (tweet_vector/n_vectorizable)
19
20 df['FT'] = df.cleaned_text.apply(func=FastTextVectorizeTweet)
```

Σχήμα 57: Διανυσματοποίηση FastText

Α΄.3 Μοντέλα Μηχανικής Μάθησης - Training

```
1 X_train_embedded = [OHU_train, OHB_train, tfidf_train_scaled, w2v_train, FT_train]
2 word_embeddings = ['OHU', 'OHB', 'TFIDF', 'W2V', 'FT']
```

Σχήμα 58: Λίστα των διάφορων training set για κάθε αναπαράσταση

Α΄.3.1 Logistic Regression

```
# Logistic Regression - Maximum Entropy
clf_LG = LogisticRegression()
start = time.time()
clf_LG.fit(X_train_embedded[i], Y_train)
pickle.dump(clf_LG, open( dir + '/B_LR_'.format(word_embeddings[i]), 'wb')) # Baseline Logistic Regression
clear_output()
```

Σχήμα 59: Logistic Regression

Α΄.3.2 Random Forests

```
# Random Forests
clf_RF = RandomForestClassifier()
start = time.time()
clf_RF.fit(X_train_embedded[i], Y_train)
pickle.dump(clf_RF, open( dir + '/B_RF_'.format(word_embeddings[i]), 'wb')) # Baseline Random Forests
clear_output()
```

Σχήμα 60: Random Forests

Α΄.3.3 Support Vector Machine

```
# Support Vector Classifier
clf_SVM = SVC(probability=True)
start = time.time()
clf_SVM.fit(X_train_embedded[i], Y_train)
pickle.dump(clf_SVM, open( dir + '/B_SVM_'.format(word_embeddings[i]), 'wb')) # Baseline Support Vector Machine
clear_output()
```

Σχήμα 61: Support Vector Machine

Α΄.3.4 Gaussian Naive Bayes

```
# Gaussian Naive Bayes
clf_GNB =  GaussianNB()
start = time.time()
clf_GNB.fit(X_train_embedded[i], Y_train)
pickle.dump(clf_GNB, open( dir + '/B_GNB_'.format(word_embeddings[i]), 'wb')) # Baseline Gaussian Naive Bayes
clear_output()
```

Σχήμα 62: Gaussian Naive Bayes

Α΄.3.5 Multinomial Naive Bayes

```
# Multinomial Naive Bayes
clf_MNB =  MultinomialNB()
start = time.time()
clf_MNB.fit(X_train_embedded[i], Y_train)
pickle.dump(clf_MNB, open( dir + '/B_MNB_'.format(word_embeddings[i]), 'wb')) # Baseline Multinomial Naive Bayes
clear_output()
```

Σχήμα 63: Multinomial Naive Bayes

Α΄.3.6 K Nearest Neighbors

```
clf_KNN = KNeighborsClassifier(n_neighbors=5)
start = time.time()
clf_KNN.fit(X_train_embedded[i], Y_train)
pickle.dump(clf_KNN, open( dir + '/B_KNN_'.format(word_embeddings[i]), 'wb')) # Baseline K-Nearest Neighbors
clear_output()
```

Σχήμα 64: K Nearest Neighbors

Α΄.3.7 XGBoost

```
clf_XGB = XGBClassifier()
start = time.time()
clf_XGB.fit(X_train_embedded[i], Y_train)
pickle.dump(clf_XGB, open( dir + '/B_XGB_'.format(word_embeddings[i]), 'wb')) # Baseline XGBoost
clear_output()
```

Σχήμα 65: XGBoost

Α΄.4 Μοντέλα Μηχανικής Μάθησης - Testing

```
X_test_embedded = [OHU_test, OHB_test, tfidf_test_scaled, w2v_test, FT_test]
word_embeddings = ['OHU', 'OHB', 'TFIDF', 'W2V', 'FT']
```

Σχήμα 66: Λίστα των διάφορων testing set για κάθε αναπαράσταση

```python
def summary(list_best_grid,X_test,Y_test, word_representation):
  acc = []
  idx = ["Logistic Regression", "Random Forest", "Support Vector Machine",
         "Gaussian Naive Bayes", 'Multinomial Naive Bayes', 'K-Nearest Neighbors',
         'XGBoost']
  color = ["red", "orange", "blue", "pink", "green", 'magenta', 'cyan']
  i = 0
  for clf in list_best_grid:
    y_score = clf.predict_proba(X_test.toarray())[:, 1]
    acc.append(round(accuracy_score(clf.predict(X_test.toarray()), Y_test), 2))
    fpr, tpr, _ = roc_curve(Y_test, y_score)
    roc_auc = auc(fpr, tpr)
    plt.plot(fpr, tpr, color=color[i],lw=2, label=idx[i] + '(area = %0.2f)' % roc_auc)
    i +=1
  df = pd.DataFrame(acc, index = idx, columns = [word_representation])
  plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
  plt.xlabel('False Positive Rate')
  plt.ylabel('True Positive Rate')
  plt.title('ROC Curve for ' + word_representation )
  plt.legend(loc="lower right")
  plt.show()
  return df
```

Σχήμα 67: Συνάρτηση summary για τη δημιουργία διαγραμμάτων

```python
BG_EST = [clf_LG,
          clf_RF,
          clf_SVM,
          clf_GNB,
          clf_MNB,
          clf_KNN,
          clf_XGB]
summary(BG_EST, X_test_embedded[i],Y_test, word_embeddings[i])
```

Σχήμα 68: Τελικά αποτελέσματα - Εφαρμογή συνάρτησης summary