ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εύρεση Βέλτιστου Αλγόριθμου Μηχανικής Μάθησης και Μοντελοποίηση Θέματος για την Κατηγοριοποίηση Λογαριασμών Twitter σε Bot ή πραγματικούς χρήστες

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δέσποινα-Δάφνη Χ. Πετρή

Μαρία Χ. Χορτάτου

Επιβλέπων : Στέφανος Κόλλιας

Καθηγητής

Αθήνα, Νοέμβριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Εύρεση Βέλτιστου Αλγόριθμου Μηχανικής Μάθησης και Μοντελοποίηση Θέματος για την Κατηγοριοποίηση Λογαριασμών Twitter σε Bot ή πραγματικούς χρήστες

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Δέσποινα-Δάφνη Χ. Πετρή

Μαρία Χ. Χορτάτου

Επιβλέπων : Στέφανος Κόλλιας

Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30η Νοεμβρίου 2020.

.....	.....	.....
Στέφανος Κόλλιας Μέλος Δ.Ε.Π Καθηγητής	Ανδρέας-Γεώργιος Σταφυλοπάτης - Μέλος Δ.Ε.Π Καθηγητής	Γεώργιος Σιόλας ΕΔΙΠ

Αθήνα, Νοέμβριος 2020

.....

Δέσποινα-Δάφνη Χ. Πετρή

Πτυχιούχος του Τμήματος Πληροφορικής και Τηλεπικοινωνιών του Εθνικού Καποδιστριακού Πανεπιστημίου Αθηνών

Μαρία Χ. Χορτάτου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η παρούσα εργασία στοχεύει στην ανάλυση της αποτελεσματικότητας και της απόδοσης των διαφορετικών αλγορίθμων μηχανικής μάθησης στην ικανότητά τους να ταξινομούν τους λογαριασμούς Twitter είτε ως Bots είτε ως ανθρώπους. Για προφανείς λόγους, αυτό αντιμετωπίζεται ως πρόβλημα κατηγοριοποίησης. Αξιολογείται συνεπώς η απόδοση της γενίκευσης και πρόβλεψης των επιδόσεων όλων των μοντέλων σε σχέση με τα δεδομένα. Βάσει των αποτελεσμάτων, παρουσιάζεται ο πλέον κατάλληλος αλγόριθμος μηχανικής μάθησης για το δεδομένο χώρο υποθέσεων.

Είναι πλέον γεγονός ότι τα μέσα κοινωνικής δικτύωσης όχι μόνο έχουν κατακλύσει τις ζωές μας αλλά συμβάλλουν καθοριστικά στη διαμόρφωση της κοινής γνώμης γύρω από μείζονα ζητήματα πολιτικά και μη. Πολλά από όσα αναρτώνται στο **Twitter** δεν είναι γραμμένα από φυσικά πρόσωπα, αλλά από ψεύτικους και αυτοματοποιημένους λογαριασμούς τα λεγόμενα Bots. Η διαδικτυακή αγορά των Bots, εξαιρετικά δημοφιλής στο εξωτερικό, έχει επεκταθεί ως πρακτική και στην Ελλάδα.

Στην παρούσα εργασία προτείνεται ένας αποδοτικός τρόπος για την ανίχνευση ενός Bot λογαριασμού Twitter, χρησιμοποιώντας για την εκπαίδευση του μοντέλου χαρακτηριστικά χρηστών από το API του Twitter. Έμφαση δίνεται στη συναισθηματική ανάλυση πάνω στα κείμενα που δημοσιεύουν οι χρήστες. Ο συνδυασμός αυτών των στοιχείων δεν έχει χρησιμοποιηθεί ξανά σε προηγούμενες εργασίες, καθώς κατά τη διάρκεια της μελέτης πραγματοποιήθηκαν πειράματα με αλγορίθμους NaiveBayes, Μηχανές Διανυσμάτων Υποστήριξης και Νευρωνικά Δίκτυα. Η υλοποίηση τους έγινε στη γλώσσα Python με χρήση των βιβλιοθηκών scikit-learn. Επιπρόσθετα χρησιμοποιήθηκαν οι υπηρεσίες του IBM Watson Tone Analyzer για την ανάλυση του συναισθήματος πάνω στα Tweets.

Στο πρώτο μέρος της εργασίας, πραγματοποιείται εξόρυξη στα δεδομένα από το Kaggle και στη συνέχεια εισάγονται παραπάνω πληροφορίες για τους χρήστες χρησιμοποιώντας το API του Twitter. Εν συνεχεία, με την βοήθεια της υπηρεσίας Tone Analyzer, γίνεται εξαγωγή πρόσθετων χαρακτηριστικών από τα δεδομένα. Χρησιμοποιώντας διάφορα μοντέλα ταξινόμησης επιτυγχάνεται 92.2% ακρίβεια στις προβλέψεις.

Όλα τα πειράματα πραγματοποιήθηκαν σε δείγμα χρηστών αντιπροσωπευτικό του συνολικού πληθυσμού του Twitter για να γίνει βέβαιο ότι η προτεινόμενη μέθοδος μπορεί να γενικευτεί αξιόπιστα. Η εργασία αυτή καταδεικνύει ότι χρησιμοποιώντας μόνο έναν πολύ μικρό αριθμό χαρακτηριστικών από τα προφίλ χρηστών στο Twitter, είναι δυνατή η ανίχνευση εάν ένας λογαριασμός είναι φυσικό πρόσωπο ή Bot τους πετυχαίνοντας έναν πολύ καλό συνδυασμό δυνατότητας κλιμάκωσης (scalability) και ορθότητας (accuracy).

Λέξεις Κλειδιά

Κοινωνικά Δίκτυα, Twitter, Twitter Bots, Συλλογή δεδομένων, Ανάλυση δεδομένων, Ανάλυση Συναισθήματος, Οπτικοποίηση Δεδομένων, Μηχανική Μάθηση, Μετρικές Αξιολόγησης, Μοντελοποίηση Θέματος

Abstract

The initial purpose of this document has been to analyze the effectiveness and efficiency of different Machine Learning algorithms in their ability to classify Twitter accounts as either Bots or humans. For obvious reasons, this is treated as a categorization problem. The generalized performance and performance prediction of all models in relation to the data is therefore evaluated. Based on the results, the most appropriate machine learning algorithm for the given case space is presented.

It is now a fact that Social Media has not only flooded our lives but has been instrumental in shaping public opinion around major political and non-political issues. Much of what is posted on Twitter is not written by individuals, but by fake and automated accounts called Bots. The online market of Bots, extremely popular abroad, has expanded as a practice in Greece, as well.

This paper proposes an efficient way to detect a Twitter account Bot, using the Twitter API user features to train our model. Emphasis is placed on emotional analysis of texts published by users. The combination of these elements has not been used again in previous work, as during the study experiments were performed with Naïve Bayes algorithms, Support Vector Machines and Neural Networks. They were implemented in Python using the scikit-learn libraries. Additionally, the services of IBM Watson Tone Analyzer were used to analyze the emotion on the Tweets.

In the first part of the work performed, the data is extracted from Kaggle and afterwards more information about the users is extracted using the Twitter API. At that point, using Tone Analyzer service, additional features are extracted from the initial data. By using different classification models, 92.2% accuracy in the predictions is achieved.

All experiments have been performed on a sample of users that represent the total Twitter population to ensure that the proposed method can be reliably generalized. This work demonstrates that making use of only a very small number of features from Twitter user profiles, it is possible to detect whether an account is a natural person or a Bot, by achieving a very good combination of scalability and accuracy.

Key Words

Social Networks, Twitter, Twitter Bots, Data Mining, Data Analysis, Sentimental Analysis, Data Visualization, Machine Learning, Performance Metrics, Topic Modelling

Περιεχόμενα

1. Εισαγωγή

1.1 Μέσα Κοινωνικής Δικτύωσης

Τα μέσα κοινωνικής δικτύωσης είναι ισχυρά εργαλεία που συνδέουν εκατομμύρια ανθρώπους σε όλο τον κόσμο. Αυτές οι συνδέσεις σχηματίζουν το υπόστρωμα που υποστηρίζει τη διάδοση πληροφοριών, το οποίο τελικά επηρεάζει τις ιδέες, τις ειδήσεις και τις απόψεις στις οποίες εκτίθενται. Υπάρχουν οντότητες με ισχυρά κίνητρα και τεχνικά μέσα για κατάχρηση διαδικτυακών κοινωνικών δικτύων - από άτομα που στοχεύουν να αυξήσουν τεχνητά τη δημοτικότητά τους μέχρι να επηρεάσουν την κοινή γνώμη. Δεν είναι δύσκολο να στοχεύσει κάποιος αυτόματα συγκεκριμένες ομάδες χρηστών και να προωθήσει συγκεκριμένο περιεχόμενο ή απόψεις [12], [4]. Η εμπιστοσύνη στα κοινωνικά μέσα μπορεί επομένως να μας κάνει ευάλωτους σε χειραγώγηση. Τα Social Bots είναι λογαριασμοί που ελέγχονται από λογισμικό, δημιουργώντας αλγοριθμικά περιεχόμενο και δημιουργώντας αλληλεπιδράσεις. Πολλά κοινωνικά ρομπότ εκτελούν χρήσιμες λειτουργίες, όπως η διάδοση ειδήσεων και δημοσιεύσεων [25], [14] και ο συντονισμός εθελοντικών δραστηριοτήτων [29]. Ωστόσο, υπάρχει ένα αυξανόμενο ρεκόρ κακόβουλων εφαρμογών κοινωνικών Bots. Μερικοί μιμούνται την ανθρώπινη συμπεριφορά για να κατασκευάσουν ψεύτικη πολιτική υποστήριξη από το λαό [28], προώθηση τρομοκρατικής προπαγάνδας και στρατολόγησης [12], [3], [1], χειραγώγηση του χρηματιστηρίου [12] και διάδοση φημών και θεωριών συνωμοσίας [5]. Ένα αυξανόμενο σώμα έρευνας ασχολείται με τη δραστηριότητα των κοινωνικών Bot, τις επιπτώσεις της στο κοινωνικό δίκτυο και την ανίχνευση αυτών των λογαριασμών [12], [24], [8], [6], [28], [10]. Το μέγεθος του προβλήματος υπογραμμίστηκε από μια πρόκληση εντοπισμού Bot Twitter που διοργάνωσε πρόσφατα η DARPA (Defense Advanced Research Projects Agency) για να μελετήσει την διάδοση πληροφοριών από αυτοματοποιημένους λογαριασμούς, ώστε να ανιχνεύσει κακόβουλες δραστηριότητες που πραγματοποιήθηκαν μέσω αυτών των Bots [30].

1.2 Twitter Bot or not: Τι είναι Twitter Bot

Το Twitter είναι μια υπηρεσία κοινωνικής δικτύωσης και micro-blogging, που επιτρέπει στους εγγεγραμμένους χρήστες να διαβάζουν και να δημοσιεύουν σύντομα μηνύματα που ονομάζονται Tweets. Υπάρχουν περίπου 300 εκατομμύρια μηνιαίοι ενεργοί χρήστες που δημοσιεύουν 500 εκατομμύρια Tweets την ημέρα. Από όλους τους χρήστες περίπου 23 εκατομμύρια εκτιμάται ότι είναι Bots (αυτοματοποιημένα προγράμματα που δημοσιεύουν Tweets, κατεβάζουν δεδομένα, κλπ. χωρίς ανθρώπινη παρέμβαση). Στόχος είναι να δημιουργηθεί ένα αποτελεσματικό πρότυπο πρόβλεψης που ταξινομεί σωστά και ξεχωρίζει τα Bots από τον άνθρωπο. Ως επιπλέον πληροφορία για την πρόβλεψη θα χρησιμοποιηθεί το συναίσθημα που εξάγεται από τα Tweets.

Το «κλειδί» για την επιτυχή διάδοσή των ψευδών ειδήσεων είναι ίδιο σε όλο τον κόσμο και δεν είναι άλλο από την αυτοματοποίηση. Τα Social Media αποτελούν το πεδίο, όπου με την κατάλληλη τεχνική μπορεί με πέντε ανθρώπους να δημιουργηθεί η αίσθηση ότι πέντε εκατομμύρια άνθρωποι συζητούν για ένα θέμα.

Πιο συγκεκριμένα, ένα Twitter Bot είναι ένα είδος λογισμικού που ελέγχει έναν λογαριασμό Twitter μέσω του Twitter API. Το λογισμικό Bot μπορεί να εκτελεί αυτόνομα ενέργειες όπως Tweeting, re-Tweeting, like, following, unfollowingή άμεση ανταλλαγή μηνυμάτων με άλλους λογαριασμούς. Η αυτοματοποίηση των λογαριασμών Twitter διέπεται από ένα σύνολο κανόνων αυτοματοποίησης που περιγράφουν τις σωστές και ακατάλληλες χρήσεις της αυτοματοποίησης. Η σωστή χρήση περιλαμβάνει τη μετάδοση χρήσιμων πληροφοριών, την αυτόματη δημιουργία ενδιαφέροντος ή δημιουργικού περιεχομένου και την αυτόματη απάντηση στους χρήστες μέσω άμεσου μηνύματος. Η ακατάλληλη χρήση, από την άλλη, περιλαμβάνει την καταστρατήγηση των ορίων χρήσης του Twitter API, την παραβίαση της ιδιωτικής ζωής των χρηστών, και το spamming.

Στο Twitter, τα Bots είναι ουσιαστικά λογαριασμοί που αποστέλλουν αυτοματοποιημένα μηνύματα σε άλλους λογαριασμούς που έχει ορίσει ο διαχειριστής τους. Αρκετοί είναι αυτοί που πληρώνουν αδρά, προκειμένου να αποκτήσουν Bots, ώστε να φαίνεται, είτε ότι έχουν περισσότερους followers, είτε για να αποσυντονίζουν τη συζήτηση από αμφιλεγόμενα θέματα.

1.3 Κίνητρο Διπλωματικής

Με τον αυξανόμενο αριθμό των Bots στο Twitter, η αξιοπιστία της δημοτικότητας ενός χρήστη στο Twitter φαίνεται να μειώνεται και σταθερά τα spam Tweets έχουν πλέον κατακλύσει τις οθόνες μας. Επιπλέον, με την εμφάνιση των ψεύτικων ειδήσεων που διαδίδουν και επηρεάζουν τους χρήστες Twitter και όχι μόνο, είναι επιτακτική η ανάγκη να απομακρυνθούν αυτά τα Bots πριν προκαλέσουν περαιτέρω βλάβη. Υπάρχουν, φυσικά, καλής ποιότητας Bots που δημοσιεύουν χρήσιμες πληροφορίες. Ως εκ τούτου, είναι επιτακτική ανάγκη να διαπιστωθεί εάν ένας λογαριασμός είναι Bot αρχικά, και στη συνέχεια να καθοριστεί εάν είναι χρήσιμος ή όχι. Στην παρούσα εργασία, χρησιμοποιήθηκαν τεχνικές μάθησης μηχανών για να ταξινομηθούν οι λογαριασμοί χρηστών Twitter και να κατηγοριοποιηθούν σε Bot ή όχι.

Εμπειρικά, η αναγνώριση ενός ψεύτικου λογαριασμού μπορεί να γίνει παρατηρώντας τον αριθμό των ακολούθων του. Συνήθως ψεύτικοι λογαριασμοί ακολουθούν πολλά άτομα, αλλά τους ακολουθούν ελάχιστοι ή κανένας. Ένα άλλο χαρακτηριστικό είναι να μην έχει αλλάξει η φωτογραφία προφίλ. Επίσης, συχνά ψεύτικοι λογαριασμοί θα κάνουν Tweets με στόχο προώθηση κάποιο προϊόντος ή μιας ιδέας, ή και αντίστροφα με στόχο την συκοφάντηση του. Από το παραπάνω προέκυψε η ταξινόμηση των Tweets των χρηστών με βάση το συναίσθημα που εξάγουν ώστε να παρατηρηθεί εάν εν τέλει ένας έντονα δυσαρεστημένος ή ευχαριστημένος φαινομενικά χρήστης είναι πιο πιθανό να είναι ψεύτικος λογαριασμός.

1.4 Σχετικές Εργασίες

Το Twitter έχει γίνει πρόσφατα πολύ δημοφιλές, με αποτέλεσμα να έχει προσελκύσει spammers που επιθυμούν λόγω της δημοτικότητάς του να δημοσιεύσουν το περιεχόμενό τους. Η καταπολέμηση του spam Bot στο Twitter έχει ερευνηθεί σε πρόσφατα έργα. Οι χρήστες του Twitter κατηγοριοποιούνται βάσει των Tweets, των χαρακτηριστικών και της θέσης τους. Στην εργασία των H. Kwak et al. [16] μελετήθηκε το Twitter, καθώς αναλύοντας τον λόγο ακολούθων followers/following διαπιστώθηκε πως η δυσαναλογία με χαμηλό αριθμό followers, και η χαμηλή αμοιβαιότητα των Tweets, σηματοδοτούν την απόκλιση από τα γνωστά χαρακτηριστικά ενός ανθρώπινου λογαριασμού και κατ' επέκταση ενός ανθρώπινου κοινωνικού δικτύου. Στην εργασία Ο Chaijietal. [9] έχουν δείξει πώς να μεγιστοποιήσουν τη διάδοση περιεχομένου σε ένα δικό τους κοινωνικό δίκτυο. Αντίθετα, η παρούσα προσέγγιση στοχεύει στην επιλογή ενός σωστού συνόλου Bots στο Twitter για να αποτραπεί η διάδοση spam πληροφοριών.

Στόχος είναι να η αποτελεσματική ταξινόμηση των χρηστών του Twitter με βάση διάφορα χαρακτηριστικά ενός χρήστη. Το εργαλείο Sproutsocial[1] [26] παρέχει σημαντικά στοιχεία σχετικά με έναν λογαριασμό Twitter όπως εμφανίσεις Tweet, δραστηριότητα Tweet, και χρησιμοποιώντας αυτά τα δεδομένα μπορεί να διαφοροποιηθεί ένας λογαριασμό μεταξύ Bot και ανθρώπου. Στην εργασία [17] πραγματοποιήθηκε ανάλυση των λιστών Twitter ως πιθανή πηγή για την ανίχνευση των κρυφών χαρακτηριστικών και των ενδιαφερόντων των χρηστών. Πιο συγκεκριμένα, στην αρχική οθόνη ενός χρήστη στο Twitter εμφανίζονται οι ακόλουθοί του (followers), καθώς και τα Tweets των χρηστών που ο συγκεκριμένος χρήστης ακολουθεί (following). Η έρευνά τους έδειξε ότι οι λέξεις που εξάγονται από κάθε λίστα είναι αντιπροσωπευτικές όλων των μελών της λίστας, ακόμη και αν οι λέξεις δεν χρησιμοποιούνται από τα μέλη. Το συμπέρασμα αυτό είναι χρήσιμο για τη στόχευση χρηστών με συγκεκριμένα ενδιαφέροντα, πράγμα που χρησιμοποιούν κατά κόρον οι spammers για προσέγγιση νέων χρηστών. Σύμφωνα με τις παρατηρήσεις της συγκεκριμένης εργασίας, οι spammers στέλνουν περισσότερα μηνύματα από τους νόμιμους χρήστες και είναι επίσης πιο πιθανό να ακολουθήσουν άλλους spammers από τους νόμιμους χρήστες.

[1] http://sproutsocial.com/insights/twitter-data

2. Μηχανική Μάθηση

Μηχανική μάθηση είναι πεδίο της Επιστήμης των υπολογιστών που αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θεωρίας μάθησης στην Τεχνητή Νοημοσύνη. Η μηχανική μάθηση διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν να μαθαίνουν από τα δεδομένα και να κάνουν προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα.

Πιο συγκεκριμένα, η Μηχανική Μάθηση είναι μια περιοχή της τεχνητής νοημοσύνης, η οποία περιλαμβάνει την εκπαίδευση ενός υπολογιστικού συστήματος με την βοήθεια ενός συνόλου δεδομένων εκπαίδευσης (training dataset) και ενός αλγορίθμου, ώστε να εκτελούνται απαιτούμενες λειτουργίες ανά περίπτωση δίχως να απαιτείται ο προγραμματισμός του υπολογιστικού συστήματος σε κάθε ειδική περίπτωση. Στην συνέχεια, ακολουθεί η φάση της αξιολόγησης του συστήματος, η οποία γίνεται με τη χρήση ενός συνόλου δεδομένων δοκιμής (test dataset), τα οποία τροφοδοτούνται για πρώτη φορά στο σύστημα και εκείνο καλείται είτε να τα τοποθετήσει στις αντίστοιχες ομάδες είτε να παράγει τις κατάλληλες εξόδους ανάλογες πάντα με την φάση εκπαίδευσης. Ο διαχειριστής του συστήματος θα πρέπει να γνωρίζει εκ των προτέρων τα σωστά αποτελέσματα για το δεδομένο test dataset, ώστε να τα συγκρίνει με τα αποτελέσματα του συστήματος και να το αξιολογήσει. Τα παραπάνω σύνολα δεδομένων θα πρέπει να είναι αντιπροσωπευτικά δείγματα του τύπου δεδομένων όπου εφαρμόζεται το μοντέλο μηχανικής μάθησης, και συνήθως προκύπτουν από ένα αρχικό ενιαίο σύνολο δεδομένων. Η ικανότητα ενός μοντέλου Μηχανικής Μάθησης να προβλέπει αποτελέσματα για άγνωστες εισόδους με βάση τα δεδομένα εκπαίδευσης, ονομάζεται ικανότητα γενίκευσης.

2.1 Κατηγορίες Μηχανικής Μάθησης

Οι εργασίες μηχανικής μάθησης συνήθως ταξινομούνται σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση του εκπαιδευτικού «σήματος» ή την «ανατροφοδότηση» που είναι διαθέσιμα σε ένα σύστημα εκμάθησης. Αυτές είναι:

- **Επιτηρούμενη Μηχανική Μάθηση (αλλιώς επιβλεπόμενη μάθηση ή μάθηση με επίβλεψη – Supervised Learning):** Το υπολογιστικό πρόγραμμα δέχεται τις παραδειγματικές εισόδους καθώς και τα επιθυμητά αποτελέσματα από έναν «δάσκαλο», και ο στόχος είναι να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα.
- **Μη Επιτηρούμενη Μηχανική Μάθηση (αλλιώς μη επιβλεπόμενη μάθηση ή μάθηση χωρίς επίβλεψη -Unsupervised Learning):** Χωρίς να παρέχεται κάποια εμπειρία στον αλγόριθμο μάθησης, πρέπει να βρει την δομή των δεδομένων εισόδου. Η μη

επιτηρούμενη μάθηση μπορεί να είναι αυτοσκοπός (ανακαλύπτοντας κρυμμένα μοτίβα σε δεδομένα) ή μέσο για ταξινόμηση (χαρακτηριστικό της μάθησης).

- **Μηχανική Μάθηση με μερική επίβλεψη (Semi-Supervised Machine Learning):** Αποτελεί συνδυασμό των δυο προηγούμενων κατηγοριών μάθησης, καθώς σε αυτή το σύστημα τροφοδοτείται με λίγα ζεύγη εισόδου – εξόδου και αρκετές εισόδους χωρίς αντίστοιχες εξόδους.
- **Ενισχυτική Μηχανική Μάθηση (Reinforcement Learning):** Ένα πρόγραμμα υπολογιστή αλληλοεπιδρά με ένα δυναμικό περιβάλλον, και μέσω της αλληλεπίδρασης και της παρατήρησης των αποτελεσμάτων το σύστημα βελτιώνεται. Χαρακτηριστικά παραδείγματα αποτελούν η οδήγηση ενός οχήματος και η εκπαίδευση με ένα παιχνίδι εναντίον κάποιου αντιπάλου.

2.2 Αλγόριθμοι Επιβλεπόμενης Μάθησης – Το πρόβλημα της Ταξινόμησης

Η μάθηση με διακριτές τιμές ονομάζεται ταξινόμηση, ενώ η μάθηση συνέχων συναρτήσεων καλείται παλινδρόμηση. Το πρόβλημα της ταξινόμησης εντάσσεται στην κατηγορία της επιβλεπόμενης μάθησης. Κατά τη διαδικασία της ταξινόμησης, ένας αλγόριθμος επιβλεπόμενης μάθησης καλείται να αναγνωρίσει σε ποια κατηγορία ανήκει ένα δείγμα δεδομένων. Αρχικά ο αλγόριθμος τροφοδοτείται με ένα σύνολο δειγμάτων εκπαίδευσης όπου για κάθε δείγμα λαμβάνει τα σχετικά χαρακτηριστικά του και την κατηγορία στην οποία ανήκει. Όταν τελειώσει το στάδιο της εκπαίδευσης, ο αλγόριθμος είναι σε θέση να ταξινομήσει ανά κατηγορία καινούρια δείγματα δεδομένων.

2.3 Μοντέλα Μηχανικής Μάθησης

Στη συνέχεια αναφέρεται εν συντομία το θεωρητικό υπόβαθρο αλγορίθμων επιβλεπόμενης μάθησης που χρησιμοποιούνται συχνά σε παρόμοιες μελέτες ως ταξινομητές για την ανίχνευση αυτοματοποιημένων λογαριασμών του Twitter. Οι αλγόριθμοι αυτοί οι εξής:

2.3.1 Δέντρα Αποφάσεων

Στην εξόρυξη δεδομένων, ένα δέντρο απόφασης είναι ένα προγνωστικό μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για να αναπαραστήσει τους ταξινομητές καθώς και τα μοντέλα παλινδρόμησης.

- Κατάλληλα για διαχωρισμό γνώσης βάσει χαρακτηριστικών.
- Κατάλληλα για εξόρυξη γνώσης

Βασικές Προϋποθέσεις για τη δημιουργία τους είναι:

- Ύπαρξη παραδειγματικών περιπτώσεων για δημιουργία ενός **εκπαιδευτικού συνόλου** (training set),
- Ύπαρξη **αντίθετων χαρακτηριστικών** περιγραφών, αριθμητικών ή λογικών που θα αποτελέσουν τις ιδιότητες (προϋποθέσεις του επιδιωκόμενου προς εξαγωγή κανόνα),
- **Κλάσεις κατηγοριοποίησης** (συμπεράσματα του επιδιωκόμενου προς εξαγωγή κανόνα).

Πώς αναπαρίσταται ένα δέντρο απόφασης;

Κάθε εσωτερικός κόμβος ονοματίζεται με το όνομα ενός χαρακτηριστικού.

Κάθε κλαδί/σύνδεση ονοματίζεται με ένα κατηγόρημα που μπορεί να εφαρμοστεί στο χαρακτηριστικό που αποτελεί το όνομα του κόμβου-πατέρα.

Κάθε φύλλο ονοματίζεται με το όνομα μιας κλάσης.

Χρήση της τεχνικής διαίρει και βασίλευε για διαίρεση του χώρου αναζήτησης σε υποσύνολα.

2.3.2 Λογιστική Παλινδρόμηση

Η παλινδρόμηση μπορεί να χρησιμοποιηθεί και για την πρόβλεψη τιμών ονομαστικών πεδίων, μπορεί δηλαδή να εφαρμοστεί σε προβλήματα κατηγοριοποίησης. Ένας πολύ συνηθισμένος τύπος παλινδρόμησης, που χρησιμοποιείται για κατηγοριοποίηση, είναι Λογιστική Παλινδρόμηση.

Η Λογιστική Παλινδρόμηση δεν απαιτεί πολυπαραγοντικές κανονικές κατανομές, αλλά απαιτεί τυχαία ανεξάρτητη δειγματοληψία και γραμμικότητα μεταξύ του X και του logit. Το μοντέλο είναι πιθανό να είναι πιο ακριβές κοντά στη μέση των κατανομών και λιγότερο ακριβές προς τα άκρα.

Η λογιστική παλινδρόμηση, παρά το όνομά της, είναι ένα γραμμικό μοντέλο για ταξινόμηση και όχι για παλινδρόμηση. Η λογιστική παλινδρόμηση είναι επίσης γνωστή στη βιβλιογραφία ως παλινδρόμηση logit, ταξινόμηση μέγιστης εντροπίας (Max Ent) ή log-linear ταξινομητής. Σε αυτό το μοντέλο, οι πιθανότητες που περιγράφουν τα πιθανά αποτελέσματα μίας μόνο δοκιμής μοντελοποιούνται χρησιμοποιώντας μια λειτουργική λειτουργία .

Η εφαρμογή της λογιστικής παλινδρόμησης στη βιβλιοθήκη της Python scikit-learn μπορεί να προσεγγιστεί από την κλάση Logistic Regression. Αυτή η υλοποίηση μπορεί να κάνει fit

δυαδική, One-vsRest ή multinomial λογιστική παλινδρόμηση με προαιρετική κανονικοποίηση L2 ή L1.

Η Λογιστική Παλινδρόμηση συγκεντρώνει αρκετά πλεονεκτήματα:

● Είναι μια μέθοδος αρκετά απλή, δοκιμασμένη και ευρύτατα χρησιμοποιούμενη.

● Ο υπολογισμός των συντελεστών b1 … bn είναι ένα μέτρο της σημαντικότητας των ανεξάρτητων μεταβλητών. Υπό τη έννοια αυτή, η Λογιστική Παλινδρόμηση παρέχει μοντέλα ερμηνεύσιμα.

● Η Λογιστική Παλινδρόμηση επιτυγχάνει ικανοποιητικές επιδόσεις κατηγοριοποίησης.

Μειονεκτήματα της Λογιστικής Παλινδρόμησης είναι τα εξής:

● Το βασικό μειονέκτημα της Λογιστικής Παλινδρόμησης είναι η διατύπωση αυθαίρετων υποθέσεων, όπως η ύπαρξη γραμμικής σχέσης με τον λογάριθμο του κλάσματος των πιθανοτήτων.

● Σύμφωνα με τα πολλά ερευνητικά αποτελέσματα, άλλες μέθοδοι, όπως τα Νευρωνικά Δίκτυα ή οι Μηχανές διανυσμάτων Υποστήριξης, επιτυγχάνουν τουλάχιστον εφάμιλλες ή και καλύτερες επιδόσεις κατηγοριοποίησης.

2.3.3 Νευρωνικά Δίκτυα

Τα επικρατέστερα μοντέλα μηχανικής μάθησης, είναι τα λεγόμενα νευρωνικά δίκτυα. Τα μοντέλα αυτά, αποτελούν μια υπολογιστική απεικόνιση των νευρώνων του ανθρωπίνου εγκεφάλου. Ένα νευρωνικό δίκτυο αποτελείται από κόμβους (units), οι οποίοι συνδέονται μεταξύ τους με κατευθυνόμενους συνδέσμους (links). Κάθε κόμβος ***i***, έχει μια συνάρτηση ενεργοποίησης ***g***, η οποία λαμβάνει την είσοδο και παράγει έξοδο σε έναν άλλο κόμβο. Το νευρωνικό δίκτυο αλληλοεπιδρά με το περιβάλλον και μαθαίνει από αυτό, τα συναπτικά βάρη μεταβάλλονται συνεχώς, ενδυναμώνοντας ή αποδυναμώνοντας την ισχύ του κάθε δεσμού.

Όλη η εμπειρική γνώση που αποκτά επομένως το νευρωνικό δίκτυο από το περιβάλλον κωδικοποιείται στα συναπτικά βάρη. Αυτά αποτελούν το χαρακτηριστικό εκείνο που δίνει στο δίκτυο την ικανότητα για εξέλιξη και προσαρμογή στο περιβάλλον. Υπάρχουν δύο τρόποι να εκπαιδευτεί ένα δίκτυο. Κατά τον πρώτο τρόπο, η εκπαίδευση γίνεται με εποπτεία. Στην περίπτωση αυτή το δίκτυο τροφοδοτείται με ένα σύνολο γνωστών παραδειγμάτων, δηλαδή ένα σύνολο καταστάσεων στις οποίες μπορεί να περιέλθει το δίκτυο, μαζί με τα αποτελέσματα που τροφοδοτεί το δίκτυο για τις καταστάσεις αυτές. Για να μάθει το δίκτυο τα παραδείγματα αυτά, χρησιμοποιείται ένας αλγόριθμος εκπαίδευσης.

Ο αλγόριθμος εκπαίδευσης που θα χρησιμοποιηθεί εξαρτάται από το εκάστοτε πρόβλημα και από τη δομή του δικτύου που επιλέγεται. Κατά το δεύτερο τρόπο, η εκπαίδευση γίνεται χωρίς εποπτεία.

Στην παρούσα ανάλυση θα χρησιμοποιηθεί ο πρώτος τρόπος εκπαίδευσης του δικτύου καθώς αναλύεται ένα πρόβλημα επιβλεπόμενης μηχανικής μάθησης.

Το βασικό πλεονέκτημα των νευρωνικών δικτύων είναι ότι μπορούν να αποθηκεύσουν γνώση και εμπειρία από το περιβάλλον, την οποία μπορεί στη συνέχεια να ανακαλέσει. Επιπλέον, έχει τη δυνατότητα να γενικεύει, δηλαδή να εξάγει τα βασικά χαρακτηριστικά ενός συστήματος, ακόμα και όταν αυτά είναι κρυμμένα σε θορυβώδη δεδομένα.

2.3.3.1 Ο αισθητήρας (Perceptron)

Ο αισθητήρας (perceptron) είναι ένα δίκτυο με δύο επίπεδα. Το πρώτο επίπεδο απαρτίζεται από τις εισόδους του δικτύου, δεν έχει νευρώνες και επομένως δεν γίνεται καμία επεξεργασία πληροφορίας σε αυτό. Το δεύτερο επίπεδο αποτελείται από νευρώνες τύπου McCulloch-Pitts και είναι το επίπεδο εξόδου του δικτύου.

Ο στόχος του απλού αισθητήρα είναι να μάθει να λύνει προβλήματα ταξινόμησης, να αντιστοιχεί δηλαδή κάθε σετ εισόδων που δέχεται στη σωστή κλάση. Ο αισθητήρας μπορεί να λύσει πολλά τέτοια προβλήματα με επιτυχία. Ένα από τα πλεονεκτήματα του δικτύου αυτού είναι ότι υπάρχει ένας σαφής αλγόριθμος βάσει του οποίου μπορεί να εκπαιδευτεί, ώστε να δίνει σωστά αποτελέσματα.

2.3.3.2 Πολυεπίπεδοι Αισθητήρες (MultiLayer Perceptrons, MLPs)

Το τροποποιημένο αυτό μοντέλο του απλού αισθητήρα ονομάζεται πολυεπίπεδος αισθητήρας (multilayer perceptron ή multilayer feed forward networks). Σε ένα τέτοιο νευρωνικό δίκτυο, μεταξύ των επιπέδων εισόδου και εξόδου, μεσολαβούν και ένα ή περισσότερα επίπεδα ακόμα, τα λεγόμενα κρυφά επίπεδα (hidden layers).

Η ροή της πληροφορίας σε ένα τέτοιο δίκτυο γίνεται πάντα από τα αριστερά προς τα δεξιά, δεν υπάρχει κανένας βρόχος ανάδρασης. Θεωρείται επίσης ότι οι νευρώνες σε κάθε επίπεδο αλληλοεπιδρούν μόνο με εκείνους τους νευρώνες που ανήκουν στα άμεσα γειτονικά τους επίπεδα. Δηλαδή το πρώτο κρυφό επίπεδο δέχεται τις τιμές του επιπέδου εισόδου, τα αποτελέσματα του πρώτου κρυφού επιπέδου περνάνε στο δεύτερο κρυφό, του οποίου τα αποτελέσματα στη συνέχεια περνάνε στο επίπεδο εξόδου.

2.3.3.3 Τυχαία Δάση (Random Forests)

Τα τυχαία δάση αποτελούν μια ειδική κατηγορία των συνδυαστικών μεθόδων ταξινόμησης η οποία χρησιμοποιεί για ταξινομητές δέντρα απόφασης.

Για την κατασκευή ενός δέντρου απόφασης ανατίθεται αρχικά στη ρίζα του το σύνολο των δειγμάτων εκπαίδευσης. Κάθε ενδιάμεσος κόμβος περιέχει υποσύνολο των δειγμάτων το οποίο μέσω της εφαρμογής ενός κατάλληλου ελέγχου διαχωρίζεται σε δυο ή περισσότερα μικρότερα υποσύνολα (παιδιά) στο επόμενο επίπεδο. Ο έλεγχος συνήθως αφορά ένα υποσύνολο των χαρακτηριστικών των δειγμάτων εκπαίδευσης. Η επιλογή του καλύτερου διαχωρισμού γίνεται σύμφωνα με ένα κατάλληλο μέτρο όπως π.χ. Gini index, εντροπία, misclassification error. Τα δέντρα του δάσους αναπτύσσονται στο μέγιστο μέγεθος τους, χωρίς κλάδεμα.

Η διαδικασία ταξινόμησης «άγνωστων» παραδειγμάτων πραγματοποιείται μέσω της διάσχισης των δέντρων του δάσους ξεκινώντας από τη ρίζα και καταλήγοντας σε ένα από τα φύλλα του δέντρου και στη συνέχεια συνδυάζοντας τις προβλέψεις των ταξινομητών σύμφωνα με ένα πλειοψηφικό σύστημα ψηφοφορίας (majority voting scheme). Κάθε παράδειγμα ανατίθεται στην κατηγορία με τη μεγαλύτερη συχνότητα.

Παρακάτω αναφέρονται συνοπτικά οι ιδιότητες και τα κύρια πλεονεκτήματα των Random Forests:

- Μπορούν να εκπαιδευτούν σε σύνολα δεδομένων υψηλής διάστασης όπως είναι τα κείμενα και οι εικόνες, χωρίς να εμφανίσουν σημαντικό βαθμό υπερ-εκπαίδευσης.
- Εξαιτίας του μεγάλου πλήθους δέντρων στο δάσος, το σφάλμα γενίκευσης είναι περιορισμένο. Αυτό έχει ως αποτέλεσμα τη μη εμφάνιση φαινομένων υπερ-εκπαίδευσης.
- Μη επαναληπτική διαδικασία εκπαίδευσης, ο αλγόριθμος ολοκληρώνεται σε σταθερό αριθμό βημάτων.
- Η τυχαία επιλογή ενός υποσυνόλου των χαρακτηριστικών για τη διαμέριση των παραδειγμάτων κάθε ενδιάμεσου κόμβου ελαττώνει τη συσχέτιση ανάμεσα στα δέντρα και διατηρεί την πόλωση (bias) σε χαμηλά επίπεδα καθώς τα δέντρα αναπτύσσονται χωρίς κλάδεμα. Χρησιμοποιώντας ένα σύνολο δέντρων απόφασης μειώνεται και η διακύμανση (variance).
- Η διάσχιση ενός δέντρου από ένα παράδειγμα ξεκινώντας από τη ρίζα και καταλήγοντας σε έναν από τους τερματικούς κόμβους γίνεται σε λογαριθμικό ως προς το πλήθος των φύλλων του.

Τα τυχαία δάση παρουσιάζουν όμως και κάποια σημαντικά μειονεκτήματα ως προς την εφαρμογή τους τα οποία αναφέρονται συνοπτικά παρακάτω:

■Υψηλό υπολογιστικό κόστος.

■Υπάρχει σημαντικό πλήθος ελεύθερων παραμέτρων τις οποίες πρέπει να προσδιορίσει ο χρήστης π.χ. πλήθος δέντρων, βαθμός κόμβων, πλήθος παραδειγμάτων εκπαίδευσης, συνθήκη τερματισμού διαμέρισης των κόμβων.

■Για την επέκταση ενός μοντέλου με στόχο την εισαγωγή μιας ακόμα κατηγορίας απαιτείται η κατασκευή του μοντέλου από την αρχή.

■Κάθε νέο παράδειγμα πρέπει να διασχίσει όλα τα δέντρα του δάσους για την εκτίμηση της κατηγορίας του.

2.3.3.4 Gaussian Naive Bayes

Ο ταξινομητής Naive Bayes είναι ένα μοντέλο μηχανικής μάθησης που χρησιμοποιείται για εργασίες ταξινόμησης. Η ουσία του ταξινομητή βασίζεται στο θεώρημα Bayes. Ο απλός ταξινομητής Bayes (simple/naive Bayes classifier) είναι μια πρακτική μέθοδος μάθησης που στηρίζεται σε στατιστικά στοιχεία (κατανομές πιθανότητας).

Μια απλή τεχνική χρησιμοποιείται για την κατασκευή Naive Bayes ταξινομητών, όπου επικρατεί η αρχή ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής κλάσης.

Για ορισμένους τύπους μοντέλων πιθανότητας, οι αφελείς ταξινομητές Bayes μπορούν να εκπαιδευτούν πολύ αποτελεσματικά σε ένα εποπτευόμενο περιβάλλον μάθησης. Σε πολλές πρακτικές εφαρμογές, η εκτίμηση παραμέτρων για αφελείς μοντέλα Bayes χρησιμοποιεί τη μέθοδο της μέγιστης πιθανότητας. Με άλλα λόγια, μπορεί κανείς να συνεργαστεί με το αφελές μοντέλο Bayes χωρίς να αποδεχτεί την πιθανότητα Bayesian ή να χρησιμοποιήσει μεθόδους Bayesian.

2.3.3.5 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Ένα δυνατό μοντέλο, το οποίο περιέχει μια πληθώρα αλγορίθμων για επιβλεπόμενη μάθηση, είναι οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines). Η γενική χρήση των SVMs είναι η ταξινόμηση, ωστόσο ο αλγόριθμος έχει προσαρμοστεί για προβλήματα παλινδρόμησης. Για τα προβλήματα αυτά είναι πολύ χρήσιμα αυτά τα μοντέλα γιατί έχουν τρία μοναδικά χαρακτηριστικά:

1. Χρησιμοποιούν διαχώριση μεγίστου περιθωρίου (maximum margin separation), ο διαχωριστής δηλαδή που διαχωρίζει γραμμικά τα δεδομένα απέχει από τις διαφορετικές τάξεις, την μέγιστη απόσταση. Συνεπώς, τα στοιχεία που βρίσκονται πιο κοντά στον διαχωριστή είναι αυτά που έχουν μεγαλύτερη σημασία και τα υπόλοιπα μπορούν να αγνοηθούν. Τα σημεία αυτά ονομάζονται Διανύσματα Υποστήριξης (Support Vectors). Αυτό βοηθάει στην αποφυγή υπερπροσαρμογής (overfitting).

2. Παρόλο που δημιουργούν ένα γραμμικά διαχωρίσιμο μονοδιάστατο υπερεπίπεδο (hyperplane), έχουν την δυνατότητα να ενσωματώσουν τα δεδομένα, σε πολυδιάστατους χώρους χρησιμοποιώντας το κόλπο πυρήνα kernel-trick. Αυτό γιατί, δεδομένα που πολλές φορές δεν διαχωρίζονται στην μία διάσταση, διαχωρίζονται εύκολα σε περισσότερες.

3. Είναι μη παραμετρική μέθοδος, με την έννοια ότι κρατά στη μνήμη τα παραδείγματα της εκπαίδευσης. Όμως, σε αντίθεση με γενικές μη παραμετρικές μεθόδους μπορούν να αποκλείσουν ένα μεγάλο μέρος των παραδειγμάτων. Με αυτό τον τρόπο συνδυάζουν τα θετικά στοιχεία των παραμετρικών και μη παραμετρικών μεθόδων σε ένα μοντέλο. Δηλαδή μπορούν να αναπαριστούν περίπλοκες συναρτήσεις, αλλά και να κάνουν καλές γενικεύσεις.

2.3.4 Ensemble Methods

Ο στόχος των ensemble μεθόδων είναι να συνδυάσουν τις προβλέψεις πολλών εκτιμητών βάσης που κατασκευάστηκαν με έναν δεδομένο αλγόριθμο εκμάθησης, προκειμένου να βελτιωθεί η γενίκευση / ευρωστία σε έναν ενιαίο εκτιμητή.

2.3.4.1 Αλγόριθμος AdaBoost

Ο αλγόριθμος AdaBoost είναι ένας μετα-αλγόριθμος μηχανικής μάθησης που μπορεί να χρησιμοποιηθεί με άλλους αλγορίθμους μάθησης για να βελτιώσει την απόδοσή τους.

Ο πλέον γνωστός αλγόριθμος ώθησης είναι ο αλγόριθμος AdaBoost (Adaptive boosting), το όνομα του οποίου συχνά χρησιμοποιείται ως συνώνυμο της μεθοδολογίας της ώθησης. Η απλούστερη εκδοχή ώθησης θεωρεί προβλήματα κατηγοριοποίησης με δύο μόνον κατηγορίες στο σύνολο ***Y={0,1}***. Έστω ***X*** το σύνολο αναφοράς των δεδομένων που πρόκειται να κατηγοριοποιηθούν. Μετά την εκπαίδευσή του ο κατηγοριοποιητής μαθαίνει μια συνάρτηση ***h*** η οποία καλείται υπόθεση. Η τιμή ***h(x)*** ερμηνεύεται ως η πιθανότητα το ***x*** να ανήκει στην κατηγορία 1, συνεπώς ***1-h(x)*** είναι η πιθανότητα το x να ανήκει στην κατηγορία 0.

Οι περισσότεροι αλγόριθμοι ώθησης λειτουργούν επαναληπτικά, θεωρώντας μια συγκεκριμένη κατανομή πιθανότητας ορισμένη πάνω σε πεπερασμένο υποσύνολο εκπαίδευσης. Σε κάθε ασθενή κατηγοριοποιητή τυπικά αποδίδεται ένας συντελεστής βάρους ανάλογος με την ικανότητα του ασθενούς κατηγοριοποιητή για γενίκευση.

Μετά την ενδεχόμενη εισαγωγή ενός ασθενούς κατηγοριοποιητή υπολογίζεται εκ νέου η κατανομή πιθανότητας πάνω στο υποσύνολο εκπαίδευσης, έτσι ώστε για κάθε δεδομένο το οποίο κατηγοριοποιείται λανθασμένα να αυξάνεται η βαρύτητά του, ενώ για κάθε δεδομένο το οποίο κατηγοριοποιείται ορθά μειώνεται η βαρύτητα. Το αποτέλεσμα της

προαναφερθείσας τακτικής είναι ότι οι ασθενείς κατηγοριοποιητές επικεντρώνονται στο να μάθουν τα δεδομένα, τα οποία κατηγοριοποιούν λανθασμένα.

Στην παρούσα ανάλυση, χρησιμοποιείται Random Forest με Ada Boosting ώστε να είναι το μοντέλο ευαίσθητο σε noisy data και outliers.

2.3.4.2 Voting Classifier

Ψηφοφορία (Voting): κάθε ταξινομητής δίνει μια ψήφο για μία συγκεκριμένη κατηγορία. Κάθε παράδειγμα ανατίθεται στην κατηγορία η οποία έχει συγκεντρώσει τις περισσότερες ψήφους.

Η ιδέα πίσω από το Voting Classifier είναι να συνδυάσει εννοιολογικά διαφορετικούς ταξινομητές μηχανικής μάθησης και να χρησιμοποιήσει μια πλειοψηφία ή την μέση προβλεπόμενη πιθανότητα για την πρόβλεψη των ετικετών κλάσης. Ένας τέτοιος ταξινομητής μπορεί να είναι χρήσιμος για ένα σύνολο εξίσου καλά εκτελεστικών μοντέλων προκειμένου να εξισορροπηθούν οι επιμέρους αδυναμίες τους.

2.3.4.3 Gradient Boosting Classifier

Με την πάροδο των ετών, η ενίσχυση κλίσης έχει βρει εφαρμογές σε διάφορους τεχνικούς τομείς. Ο αλγόριθμος μπορεί να φαίνεται περίπλοκος στην αρχή, αλλά στις περισσότερες περιπτώσεις χρησιμοποιείται μόνο μία προκαθορισμένη διαμόρφωση για ταξινόμηση και μία για παλινδρόμηση, η οποία φυσικά μπορεί να τροποποιηθεί με βάση τις απαιτήσεις του προβλήματος.

Το Gradient Boosting έχει τρία κύρια συστατικά:

Loss Function (Συνάρτηση Απώλειας) - Ο ρόλος της συνάρτησης απώλειας είναι να εκτιμήσει πόσο καλό είναι το μοντέλο να κάνει προβλέψεις με τα δεδομένα. Αυτό μπορεί να διαφέρει ανάλογα με το πρόβλημα που αντιμετωπίζεται. Για παράδειγμα, εάν ο σκοπός είναι η πρόβλεψη του βάρους ενός ατόμου ανάλογα με ορισμένες μεταβλητές εισόδου (πρόβλημα παλινδρόμησης), τότε η λειτουργία απώλειας θα ήταν κάτι που βοηθά στην εύρεση της διαφοράς μεταξύ των προβλεπόμενων βαρών και των παρατηρούμενων βαρών. Από την άλλη πλευρά, εάν τίθεται πρόβλημα κατηγοριοποίησης εάν ένα άτομο θα ήθελε μια συγκεκριμένη ταινία με βάση την προσωπικότητά του, θα χρειαζόταν μια λειτουργία απώλειας που θα βοηθήσει στην κατανόηση πόσο ακριβές είναι το μοντέλο στην ταξινόμηση των ατόμων που το είδαν ή όχι ορισμένες ταινίες.

Weak Learner (Αδύναμος Μαθητής) - Ένας αδύναμος μαθητευόμενος είναι αυτός που ταξινομεί τα δεδομένα, αλλά το κάνει τόσο άσχημα, ίσως όχι καλύτερο από την τυχαία εικασία. Με άλλα λόγια, έχει υψηλό ποσοστό σφάλματος, και αυτά είναι συνήθως δέντρα αποφάσεων.

Additive Model (Πρόσθετο Μοντέλο) - Αυτή είναι η επαναληπτική και διαδοχική προσέγγιση της προσθήκης των δέντρων (Αδύναμοι Μαθητές) ένα βήμα τη φορά. Μετά από κάθε επανάληψη, πρέπει να βρίσκεται ο αλγόριθμος πιο κοντά στο τελικό μοντέλο. Με άλλα λόγια, κάθε επανάληψη θα πρέπει να μειώνει την αξία της λειτουργίας απώλειας.

2.3.4.4 XGboost

Το XGBoost είναι ένας αλγόριθμος μηχανικής εκμάθησης βασισμένος στα δέντρα αποφάσεων που χρησιμοποιεί ένα πλαίσιο ενίσχυσης κλίσης. Στην πρόβλεψη προβλημάτων που περιλαμβάνουν μη δομημένα δεδομένα (εικόνες, κείμενο κ.λπ.) τα τεχνητά νευρωνικά δίκτυα τείνουν να ξεπερνούν όλους τους άλλους αλγορίθμους. Ωστόσο, όταν πρόκειται για δομημένα δεδομένα μικρού έως μεσαίου μεγέθους, οι αλγόριθμοι βάσει δέντρων αποφάσεων θεωρούνται οι καλύτεροι στην κατηγορία αυτή τη στιγμή.

3. Εξόρυξη Γνώσης από Δεδομένα

Οι επιστημονικές πληροφορίες χρειάζονται διαχείριση, οργάνωση, τρόπους απεικόνισης των τεράστιων όγκων δεδομένων, και για το λόγο αυτό η εξόρυξη δεδομένων αποτελεί μια νέα σημαντική πρόκληση. Η εξόρυξη γνώσης συνεπώς, προέκυψε από την εκρηκτική ανάπτυξη και διαθεσιμότητα των δεδομένων, την ύπαρξη πολυδιάστατων και υψηλής πολυπλοκότητας δεδομένων από ανομοιογενείς πηγές, ενώ η ταχύτατη ανάπτυξη GPUs και ανάπτυξη αλγορίθμων κατέστησε αρκετά εύκολη την περαιτέρω πολυεπεξεργασία. Δυνητικές εφαρμογές της εξόρυξης γνώσης αποτελούν η ανάλυση δεδομένων και υποστήριξη αποφάσεων, η ανάλυση αγοράς και διαχείριση, η ανίχνευση ψευδών ειδήσεων ή ασυνήθιστων προτύπων (απόκλισης), η εξόρυξη κειμένου, γενικότερα η εξόρυξη πληροφοριών από το διαδίκτυο, η ανάλυση και διαχείριση ρίσκου καθώς και πολλές άλλες εφαρμογές.

Η εξόρυξη γνώσης από δεδομένα αποτελείται ουσιαστικά από την εξαγωγή ενδιαφερόντων (μη τετριμμένων, υπονοούμενων, αγνώστων ως την πρότερη γνώση, δυναμικά χρήσιμων) προτύπων ή γνώσης από μεγάλες δεξαμενές δεδομένων.

Για να εξαχθεί γνώση από τα δεδομένα πρέπει να αποθηκευτούν, να διαχειριστούν και να αναλυθούν. Δεδομένης της μεγάλης ποσότητας δεδομένων, χρήσιμη είναι η ανακάλυψη μοτίβων και μοντέλων τα οποία θα πρέπει να είναι :

- Έγκυρα: στηρίζονται σε νέα δεδομένα με κάποια βεβαιότητα,
- Χρήσιμα: θα πρέπει να έχουν πεδίο εφαρμογής,
- Μη αναμενόμενα: μη εμφανή στο σύστημα,
- Κατανοητά: οι άνθρωποι θα πρέπει να είναι σε θέση να ερμηνεύσουν το μοτίβο/μοντέλο.

Η διαδικασία ανίχνευσης γνώσης (Knowledge Discovery Database - KDD) αποτελείται από τα παρακάτω βήματα:

1. Αφαιρούνται ασυνεπή δεδομένα και δεδομένα θορύβου. Συνδυάζονται πολλαπλές πηγές δεδομένων.
2. Δημιουργία στοχευμένου συνόλου δεδομένων.
3. Μείωση και μετασχηματισμός δεδομένων. Εύρεση χρήσιμων χαρακτηριστικών. Μείωση διαστάσεων / μεταβλητών. Μετατροπή δεδομένων στη κατάλληλη φόρμα για επεξεργασία.
4. Σύνοψη, ταξινόμηση, παλινδρόμηση, συσχέτιση, ομαδοποίηση. Επιλογή του αλγόριθμου εξόρυξης. Εξόρυξη δεδομένων: αναζήτηση μορφών ενδιαφέροντος → πρότυπα.
5. Αξιολογούνται τα πρότυπα δεδομένων.

6. Μετασχηματισμός, απομάκρυνση πλεονασμάτων. Χρήση γνώσης που ανακαλύφθηκε.

Η εξόρυξη δεδομένων εφαρμόζεται σε απλά ή σύνθετα σύνολα δεδομένων, και μπορεί ουσιαστικά να αποκαλύψει ανθρώπινα ερμηνεύσιμα μοτίβα που περιγράφουν τα δεδομένα, ή να χρησιμοποιήσει μερικές μεταβλητές για να προβλέψει άγνωστες ή μελλοντικές τιμές άλλων μεταβλητών, καθώς η Μηχανική Μάθηση αποτελεί μια τεχνολογία που χρησιμοποιείται ευρέως για εξόρυξη δεδομένων. Αποτελεί τελικά η εξόρυξη δεδομένων μια φυσική εξέλιξη της επιστήμης και της τεχνολογίας των πληροφοριών, με μεγάλη ζήτηση και με ευρείες εφαρμογές.

Η εξόρυξη γνώσης από τα δεδομένα περιλαμβάνει:

- τη γνώση των δεδομένων και
- την εύρεση σχέσης μεταξύ των δεδομένων.

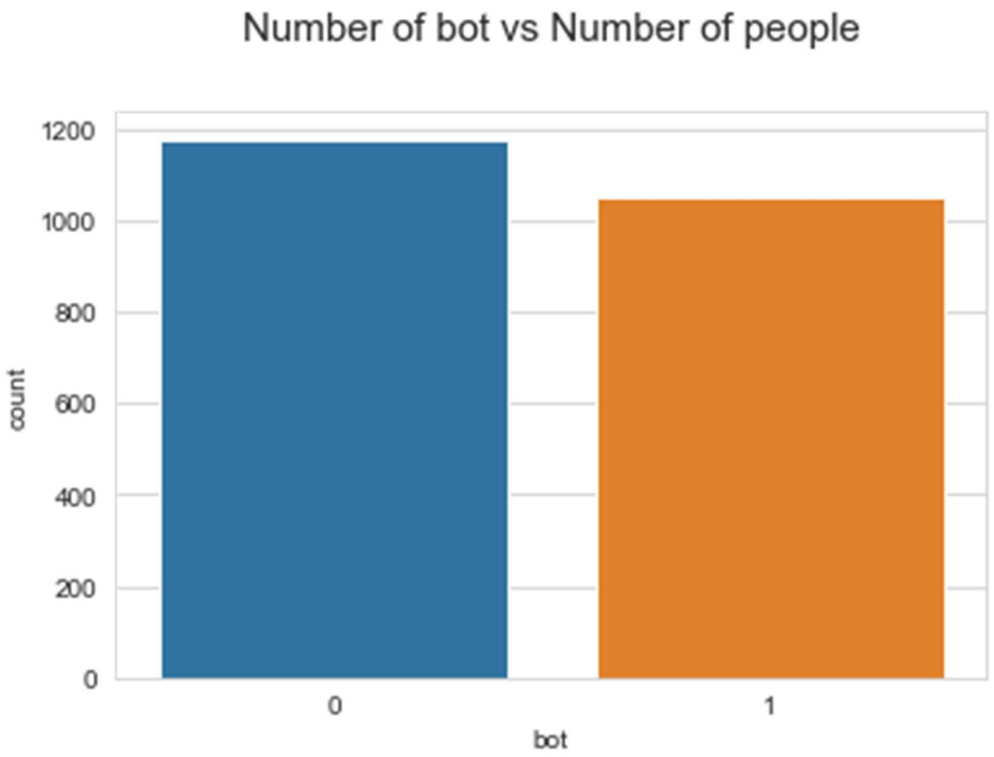
Για να ανακαλυφθούν ωστόσο τα δεδομένα, είναι απαραίτητη η επεξεργασία των συνόλων δεδομένων, τα αντικείμενα δεδομένων, τα χαρακτηριστικά γνωρίσματα δεδομένων και των τύπων των χαρακτηριστικών δεδομένων.

3.1 Το Σύνολο Δεδομένων

Το σύνολο δεδομένων (dataset) περιλαμβάνει τα δεδομένα που θα χρησιμοποιηθούν κατά την υλοποίηση της Μηχανικής Μάθησης. Πρέπει να αποτελείται από ένα ισορροπημένο δείγμα δεδομένων, αντιπροσωπευτικό ως προς την επιθυμητή κλάση στην οποία θα γίνει η ταξινόμηση. Όπως έχει αναφερθεί, τα σύνολα δεδομένων που χρησιμοποιούνται στις διάφορες φάσεις της Μηχανικής Μάθησης προκύπτουν από τη διαίρεση του αρχικά ενιαίου συνόλου δεδομένων σε σύνολο εκπαίδευσης (training dataset) και σύνολο ελέγχου (test dataset), και η τεχνική αυτή ονομάζεται Hold Out. Ο συνήθης διαχωρισμός του αρχικού συνόλου δεδομένων γίνεται με κάποια λογική αναλογία (για παράδειγμα 70% και 30%). Σε περιπτώσεις που τα δεδομένα είναι περιορισμένα, χρησιμοποιείται η μέθοδος n-fold validation, κατά την οποία τα δεδομένα χωρίζονται σε nίσα υποσύνολα και στην συνέχεια, με βάση αυτά, δημιουργούνται ημοντέλα γνώσης. Σε καθένα από τα ημοντέλα γνώσης εξαιρείται από το σύνολο εκπαίδευσης κάποιο από τα n υποσύνολα, το οποίο χρησιμοποιείται ως σύνολο ελέγχου.

Τα δεδομένα που χρησιμοποιήθηκαν, και αναλύθηκαν περαιτέρω, περιλαμβάνουν ισορροπημένο ποσοστό Bots και κανονικών χρηστών, αποτελούνται από επαληθευμένους λογαριασμούς ανθρώπων ή Bot, και αποκτήθηκαν από dataset από το Kaggle[2].

Από την παρακάτω εικόνα μπορεί εύκολα να επιβεβαιωθεί ότι χρησιμοποιήθηκε ένα ισορροπημένο δείγμα δεδομένων.



Εικόνα 1: Δεδομένα Bot VS human

3.2 Εξαγωγή Χαρακτηριστικών από τα Δεδομένα

Τα δεδομένα που έχουν συλλεχθεί με την βοήθεια του Twitter API, έχουν κατηγοριοποιηθεί σε δυο κλάσεις, όπως αναλύεται στην συνέχεια.

[2] https://www.kaggle.com/charvijain27/detecting-Twitter-Bot-data

Το Twitter API επιτρέπει την πρόσβαση μέσω προγραμματισμού στο Twitter με μοναδικούς και προηγμένους τρόπους. Χρησιμοποιείται κυρίως για την ανάλυσης των Tweet, των δεδομένων των χρηστών, καθώς και άλλα βασικά χαρακτηριστικά του Twitter.

3.2.1 Χαρακτηριστικά βασισμένα στο χρήστη

Χαρακτηριστικά που εξάγονται από τα metadata του κάθε χρήστη και χρησιμοποιήθηκαν για την ταξινόμηση των χρηστών σε Bot ή Human παρατίθενται στην συνέχεια. Το API Twitter επιτρέπει την λήψη των μεταδεδομένων του χρήστη, από τα οποία εξάγονται χαρακτηριστικά όπως ο αριθμός των ακολούθων (followers), ο αριθμός των φίλων, ο συνολικός αριθμός από Tweets που έχουν παραχθεί από τον χρήστη, το όνομα που χρησιμοποιεί, την περιγραφή, τη θέση, τις γλώσσες που χρησιμοποιεί και κάποιες επιπλέον ρυθμίσεις, όπως εξηγούνται λεπτομερώς παρακάτω.

- **Screenname:** Παρέχει το ψευδώνυμο που έχει επιλέξει ο χρήστης για τον Twitter λογαριασμό του.
- **Location:** Παρέχει την τοποθεσία του χρήστη.
- **Description:** Σύντομες πληροφορίες για τον χρήστη. Μερικά Bots παρέχουν την πληροφορία ότι πρόκειται για Βοτστην περιγραφή.
- **FollowersCount:** Ο αριθμός των ακόλουθων που έχει αυτήν τη στιγμή ο λογαριασμός.
- **FriendsCount:** Ο αριθμός των χρηστών που ακολουθεί αυτός ο λογαριασμός.
- **ListedCount:** Ο αριθμός των δημόσιων λιστών στις οποίες είναι μέλος ο χρήστης.
- **Createdat:** Ημερομηνία δημιουργίας λογαριασμού από τον χρήστη.
- **Verified:**Αναφέρει εάν πρόκειται για επαληθευμένο χρήστη (εμφανίζεται ως μπλε σημάδι tick δίπλα στο όνομα).
- **Statuscount:** Αριθμός Tweet που έχουν πραγματοποιηθεί από τον χρήστη.
- **Languages:** Ο κωδικός για τη γλώσσα διεπαφής χρήστη που έχει δηλωθεί από τον χρήστη.
- **DefaultProfilebackground:** Είναι η δυαδική τιμή που παρέχει πληροφορίες εάν ο χρήστης έχει την προεπιλεγμένη φωτογραφία για το φόντο, ή όχι.
- **DefaultProfile Picture:** Επίσης, μια τιμή boolean που παρέχει την πληροφορία εάν ο χρήστης έχει προεπιλεγμένη φωτογραφία προφίλ ή όχι.
- **Name:** Το όνομα του χρήστη, όπως το έχει προσδιορίσει. Δεν είναι απαραίτητα το πραγματικό όνομα του χρήστη.
- **FavouriteCount:** Αριθμός τωνTweets που έχει χαρακτηρίσει ο χρήστης ως αγαπημένα.
- **Diversity:** Χαρακτηριστικό που αφορά στο μήκος του ονόματος.
- **CreatedHour:** Ώρα κατά την οποία ο λογαριασμός δημιουργήθηκε, ανεξάρτητα από την ημερομηνία.

- **URL:** Μια διεύθυνση URL που παρέχεται από τον χρήστη σε συνδυασμό με το προφίλ του. Μπορεί να είναι κενό.
- **Accountage:** Ο αριθμός των ημερών που ο λογαριασμός είναι ενεργός. Υπολογίζεται από την ημερομηνία δημιουργίας λογαριασμού και την ημερομηνία του τελευταίου Tweet.
- **AverageTweets per day:** Όπως υποδηλώνει το όνομα είναι ο μέσος αριθμός Tweets που έγιναν από τον χρήστη. Είναι ο λόγος μεταξύ του συνολικού αριθμού των Tweets και της ηλικίας του λογαριασμού.
- **FollowerFriendsRatio:** Είναι ο λόγος μεταξύ των οπαδών και των φίλων.
- **ScreennameLength:** Μήκος ονόματος που παρέχεται από τον χρήστη.
- **Descriptionlength:** Μήκος περιγραφής που παρέχεται από τον χρήστη.
- **Lexicaldiversity:** Ποικιλία λέξεων που χρησιμοποιούνται.
- **NullUrl:** Boolean τιμή που υποδηλώνει εάν υπάρχει URL ή όχι.
- **Nameratio:** Αναλογία μεταξύ του μήκους του ονόματος και του αριθμού των λέξεων στο όνομα.
- **NameBot:** Αληθές εάν το Bot είναι παρόν στο όνομα.
- **DescriptionBot:** Αληθές αν το Bot είναι παρόν στην περιγραφή του λογαριασμού.
- **Number of words in name:** Αριθμός λέξεων στο όνομα του χρήστη.
- **NameLength:** Μήκος ονόματος που παρέχεται από τον χρήστη.

3.2.2 Χαρακτηριστικά βασισμένα σε Tweet

Χαρακτηριστικά που έχουν εξαχθεί από το τελευταίο Tweet του χρήστη έχουν επίσης χρησιμοποιηθεί. Με την βοήθεια του Twitter API, μπορεί να εξαχθεί το τελευταίο Tweet για έναν συγκεκριμένο χρήστη. Από το τελευταίο Tweet μπορούν να εξαχθούν διάφορα χαρακτηριστικά όπως το κείμενο του Tweet, ή ώρα που έγινε το Tweet, εάν είναι re-Tweet ή όχι, εάν το Tweet είναι απάντηση σε κάποιον άλλο χρήστη, #hashtags ή σύνδεσμοι (link) που περιλαμβάνονται στο κείμενο, και διάφορα επιπλέον χαρακτηριστικά όπως εξηγούνται παρακάτω.

- **Truncated:** Εάν το κείμενο του Tweet είναι σύντομο ώστε να χωρέσει μόνο 140 χαρακτήρες (μήκος Tweet)
- **Text:** Παρέχει το κείμενο Tweet που δημιουργήθηκε από τον χρήστη.
- **InreplytoTweetid:** Αναγνωριστικό id του κύριου Tweet, αν αφορά απάντηση σε Tweet.
- **Id:** Ένα μοναδικό αναγνωριστικό που δίνεται σε κάθε Tweet.
- **FavouriteCount:** Ο αριθμός των χρηστών που έχουν χαρακτηρίσει ως αγαπημένο αυτό το συγκεκριμένο Tweet.
- **Coordinates:** Παρέχει την τοποθεσία από την οποία γίνεται Tweet.
- **UserMentions:** Χαρακτηριστικό που αναφέρει τους χρήστες αναφέρονται στο Tweet.

- **Hashtags:**Hashtags που υπάρχουν στο Tweet.
- **ReTweetcount:** Ο αριθμός των re-Tweet στο συγκεκριμένο Tweet.
- **Createdat:** Ώρα που δημιουργήθηκε τοTweet.
- **ReTweeted:** Παρέχει τις πληροφορίες εάν το Tweet έχει γίνει re-Tweet από κάποιον άλλο χρήστη.

Συμπληρωματικά με τα παραπάνω χαρακτηριστικά, συλλέχθηκαν τα τελευταία 100 Tweets από κάθε χρήστη με τις αντίστοιχες παραμέτρους. Τα χαρακτηριστικά που εξάγονται από τα δεδομένα εξηγούνται περαιτέρω στην παρακάτω ενότητα. Όλες οι τιμές χαρακτηριστικών που αναφέρονται αφορούν τα τελευταία 100 Tweets.

- **In-replycount:** Αριθμός των Tweet ως απάντηση σε κάποιον άλλο χρήστη στα τελευταία 100 Tweets.
- **ReTweetcount:** Αριθμός των Tweet που έγιναν re-Tweet από κάποιον άλλον χρήστη.
- **Favouritecount:** Αριθμός Tweets που αρέσουν σε άλλους χρήστες.
- **Usermentions:** Αριθμός των χρηστών που αναφέρονται στα Tweets.
- **Createddates:** Ημερομηνίες δημιουργίας για τα τελευταία 100 Tweets
- **Texts:** Λίστα που περιέχει το κείμενο για τα τελευταία 100 Tweet.
- **Average in-replycount:** Αναλογία μεταξύ του αριθμού των απαντήσεων και των 100 Tweets.
- **AveragereTweetcount:** Αναλογία μεταξύ του αριθμού retweet και του αριθμού των Tweets.
- **Averagefavoritecount:** Αναλογία μεταξύ του Favoritecount και του αριθμού των Tweets.
- **Averageusermentions:** Αναλογία μεταξύ του Usermentions και του αριθμού των Tweets.
- **TweetDays:** Dictionary που παρέχει την κατανομή των τελευταίων 100 Tweets σε σχέση με την ημέρα της εβδομάδας (Δευτέρα, Τρίτη, κ.λπ.)
- **Tweethours:** Dictionary που παρέχει την κατανομή των τελευταίων 100 Tweets σε σχέση με την ώρα της ημέρας.
- **TweetDaysktest:** Εφαρμόστηκε Kolmogorov-Smirnov στην κατανομή ημερών Tweet.
- **Tweethoursktest:** Εφαρμόστηκε Kolmogorov-Smirnovστην κατανομή ωρών και ημερών των Tweets.

3.3 Προ-επεξεργασία Δεδομένων

Η επιτυχία της εφαρμογής της Μηχανικής Μάθησης σε ένα υπολογιστικό σύστημα, εξαρτάται σε πολύ μεγάλο βαθμό από την ποιότητα των δεδομένων στα οποία βασίζεται.

Εφόσον τα αρχικά δεδομένα είναι ποιοτικά, εξίσου ποιοτικά θα είναι και τα αποτελέσματα της εξόρυξης δεδομένων.

Τα δεδομένα πρέπει να χαρακτηρίζονται από:

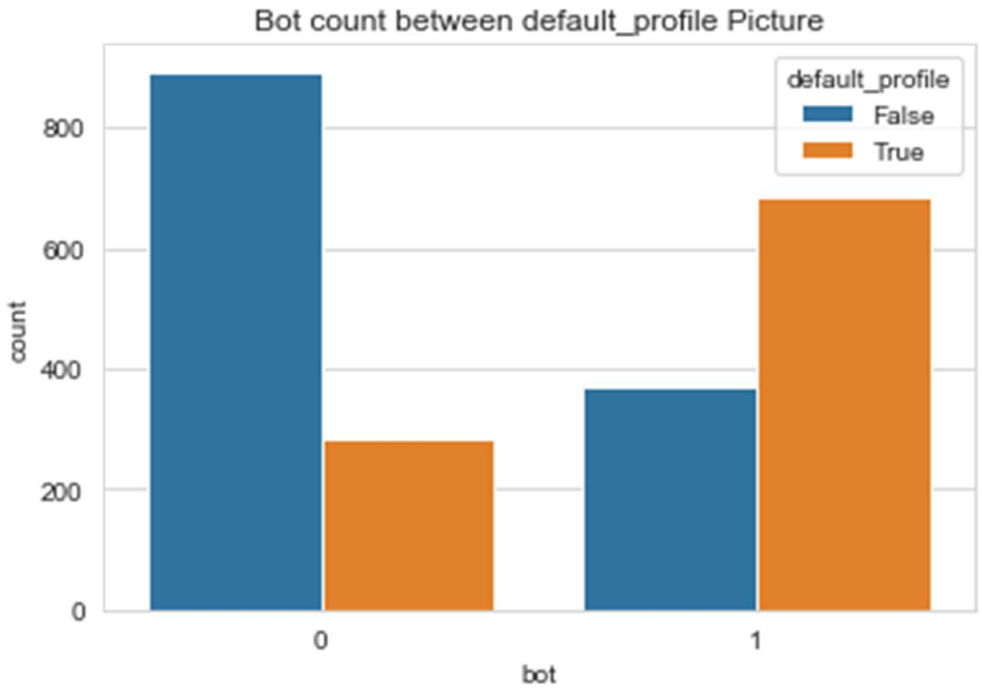
- Ακρίβεια (Accuracy)
- Πληρότητα (Completeness)
- Συνάφεια (Consistency)
- Επικαιρότητα (Timeliness)
- Πιστότητα (Believability)
- Προστιθέμενη αξία (Value added)
- Διερμηνεία (Interpretability)
- Προσβασιμότητα (Accessibility)

Εάν υπάρχει πλεονάζουσα πληροφορία ή τα δεδομένα δεν είναι αξιόπιστα, τότε η εκπαίδευση του μοντέλου και κατ' επέκταση η ανακάλυψη νέας γνώσης καθίσταται εξαιρετικά δύσκολη [22]. Είναι συνεπώς απαραίτητο να προηγηθεί της Μηχανικής Μάθησης η προ-επεξεργασία δεδομένων (data preprocessing), ώστε να βελτιωθεί η απόδοση του μοντέλου και να μειωθεί ο απαιτούμενος χρόνος εκπαίδευσης [13]. Τα τελικά δεδομένα που προκύπτουν από την προ-επεξεργασία ονομάζονται χαρακτηριστικά (features).

3.3.1 Ανάλυση Δεδομένων

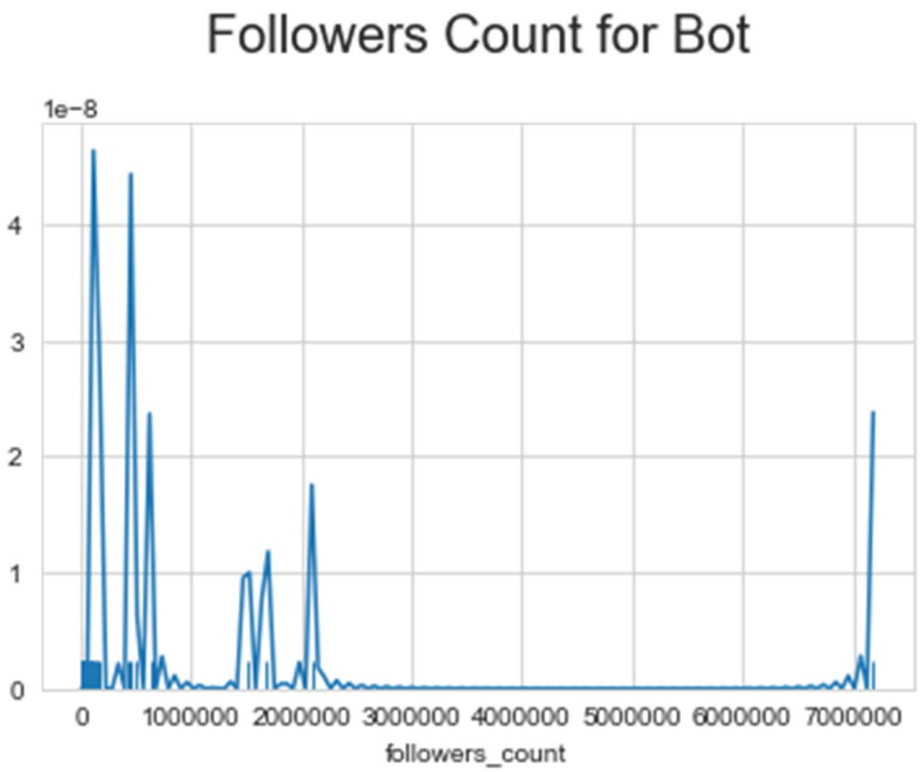
Μια πρώτη εκτίμηση των χαρακτηριστικών που έχουν εξαχθεί για κάθε χρήστη γίνεται παρακάτω, στην προσπάθεια κατανόησης της σημασίας του κάθε χαρακτηριστικού και κατά πόσο εκείνο βοηθά στην κατηγοριοποίηση ενός λογαριασμού σε Bot ή όχι.

Στην Εικόνα 2 φαίνεται εύκολα ότι η πιθανότητα του χρήστη να είναι Bot είναι υψηλή εάν ο χρήστης χρησιμοποιεί την προεπιλεγμένη φωτογραφία προφίλ, καθώς αντίστοιχα συμπεράσματα μπορούν να εξαχθούν και για το προεπιλεγμένο φόντο προφίλ του χρήστη. Το συμπέρασμα είναι εύκολα κατανοητό, καθώς οι πραγματικοί χρήστες τείνουν να προσαρμόζουν το προφίλ τους πέρα από τα προεπιλεγμένα.

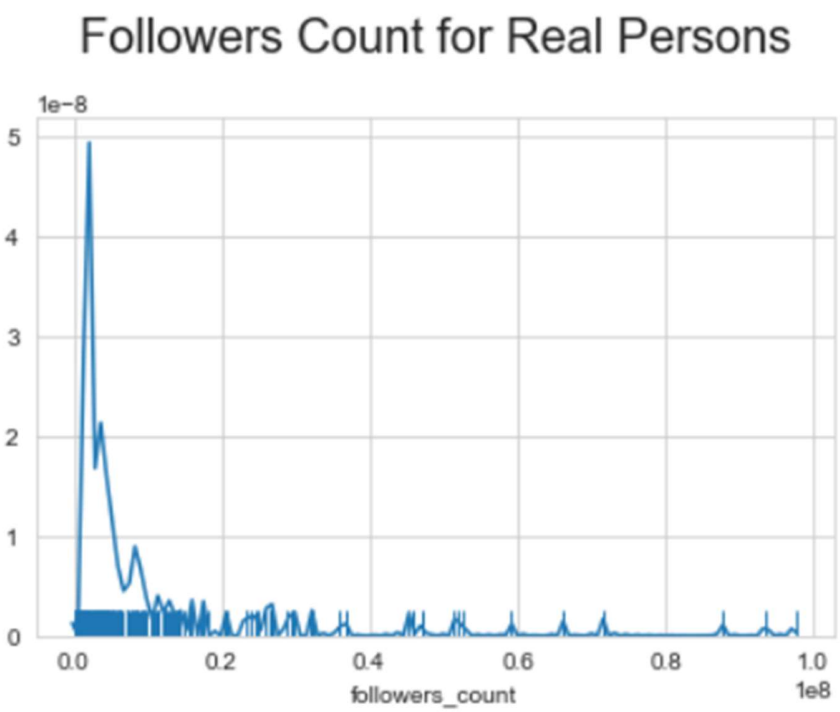


Εικόνα 2: Προεπιλεγμένη φωτογραφία προφίλ σε Bot ή Human

Στην Εικόνα 3, που αποτελεί ένα γράφημα συχνότητας του αριθμού των ακόλουθων που έχουν οι Bot λογαριασμοί, μπορεί να φανεί ότι ο αριθμός των ακολούθων για ένα Bot λογαριασμό είναι γενικά υψηλός. Μπορεί να σημειωθεί ότι η πιθανότητα του λογαριασμού να είναι Bot και ο αριθμός των ακόλουθων συσχετίζονται θετικά. Αντίστοιχα συμπεράσματα μπορούν να εξαχθούν για τους πραγματικούς λογαριασμούς στην Εικόνα 4, όπου ο αριθμός των ακολούθων και η πιθανότητα του λογαριασμού να είναι άνθρωπος συσχετίζονται αρνητικά.

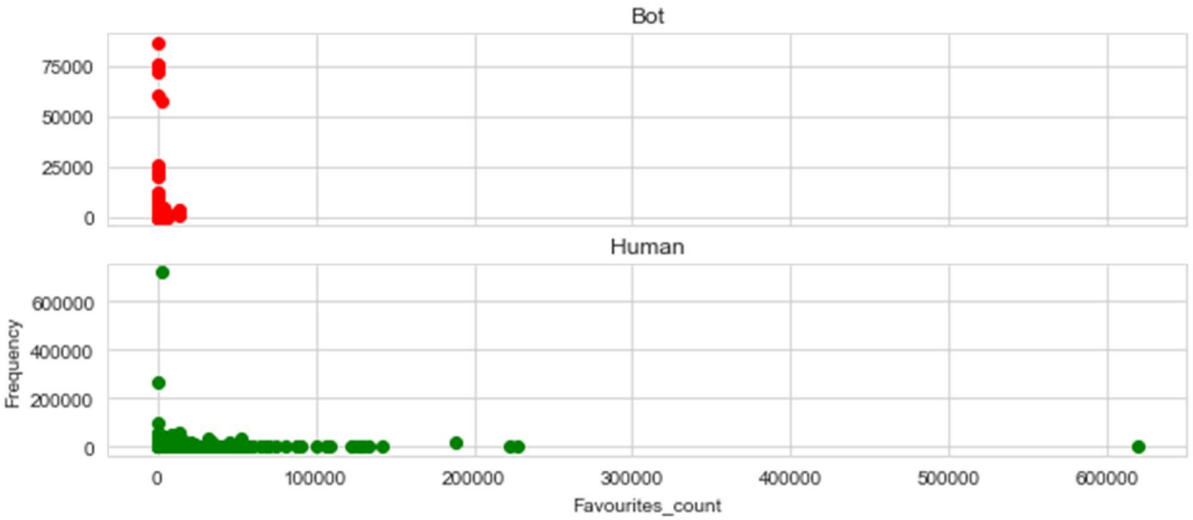


Εικόνα 3: Αριθμός Ακολούθων σε Bot λογαριασμό



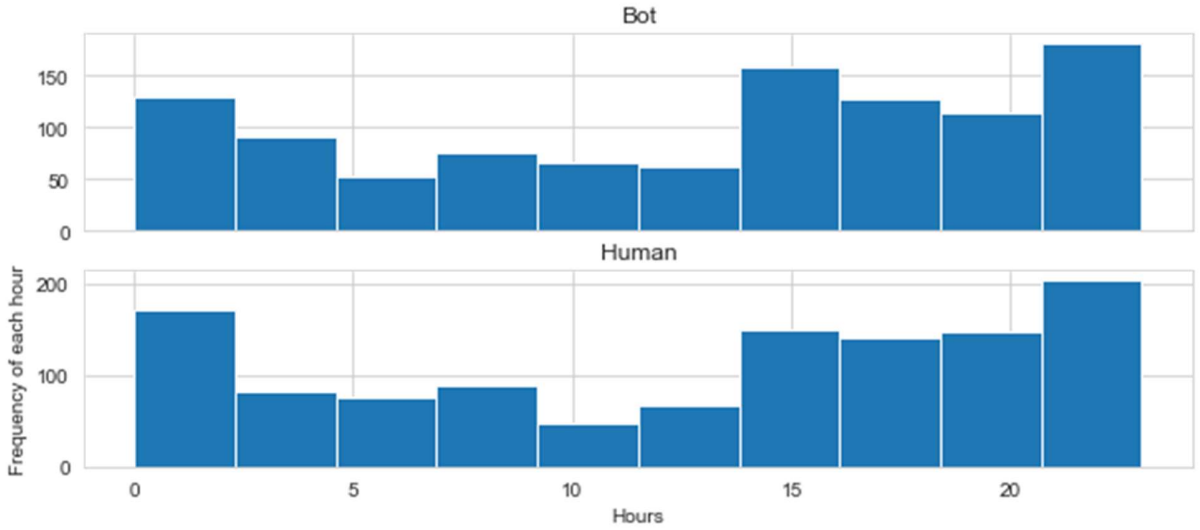
Εικόνα 4: Αριθμός Ακολούθων σε πραγματικό λογαριασμό

Η Εικόνα 5 απεικονίζει ότι ένας λογαριασμός που δεν χαρακτηρίζει συχνά ένα Tweet ως αγαπημένο, τότε ο λογαριασμός αυτός είναι πολύ πιθανό να είναι Bot.

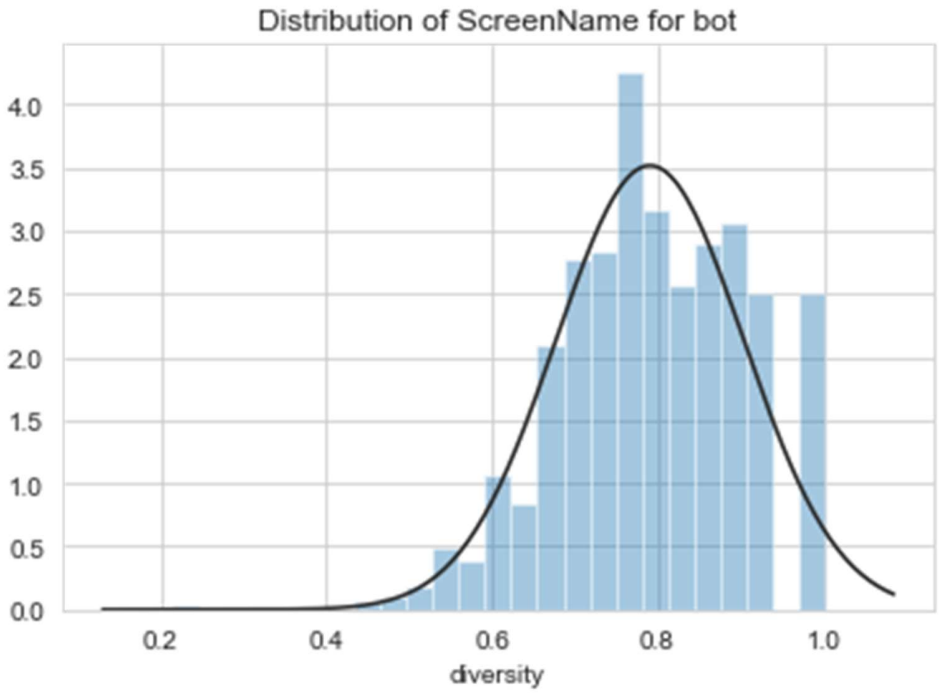


Εικόνα 5: Αριθμός Αγαπημένων Tweet

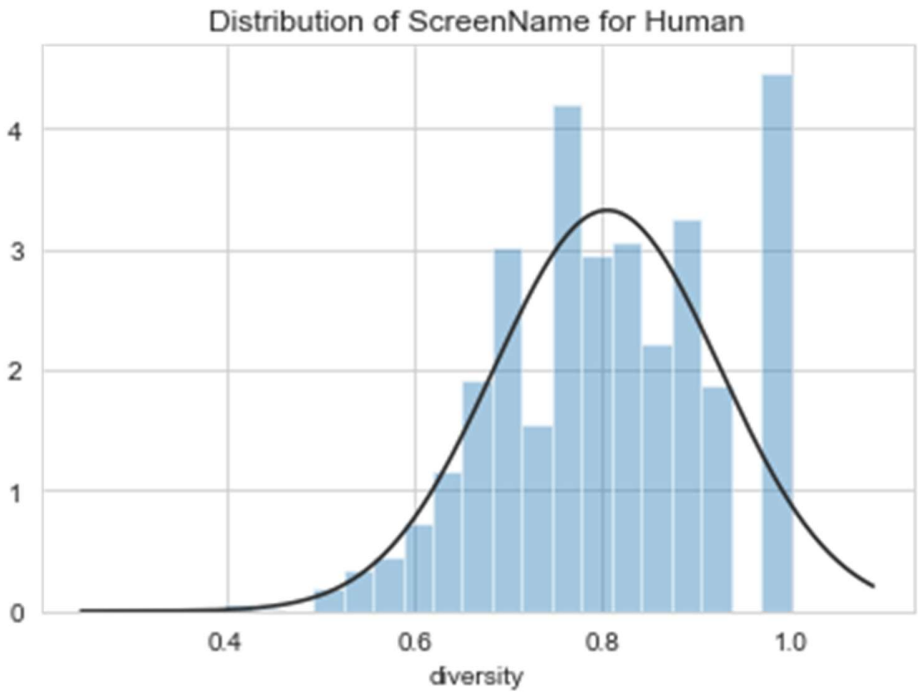
Στις Εικόνες 6,7 και 8 φαίνεται ότι οι ώρες που πραγματοποιούνται Tweets, και το μήκος του ονόματος του χρήστη, έχουν αντίστοιχη συμπεριφορά τόσο για Bot όσο και για ανθρώπινο λογαριασμό. Συνεπώς, τα χαρακτηριστικά «Createdate» και «Diversity» δεν μπορούν να θεωρηθούν σημαντικά στον χαρακτηρισμό ενός λογαριασμού σε Bot ή human.



Εικόνα 6: Συχνότητα Tweet ανά ώρα

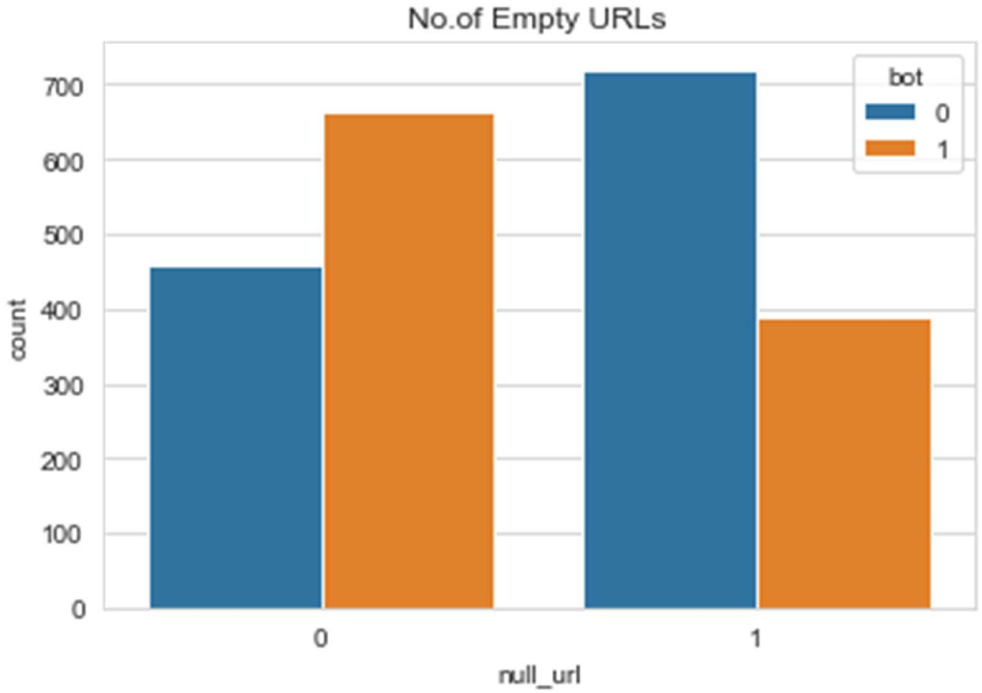


Εικόνα 7: Κατανομή μήκους ονόματος σε Bot



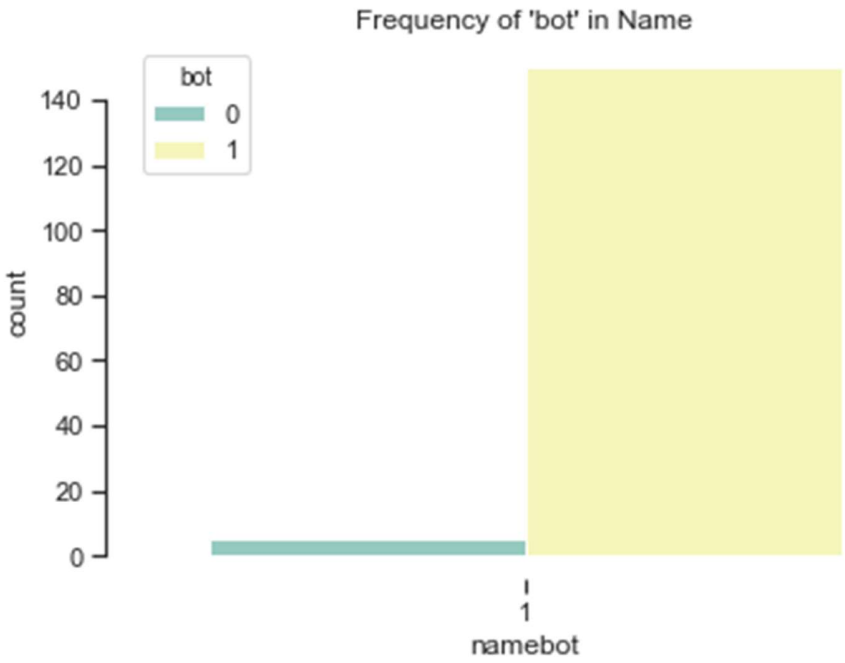
Εικόνα 8: Κατανομή μήκους ονόματος σε human

Στην Εικόνα 9, φαίνεται να είναι πιο πιθανό ένας Bot λογαριασμός να έχει συμπληρωμένο και έγκυρο URL, που φαίνεται λογικό καθότι οι Bot λογαριασμοί δημιουργούνται κυρίως για διαφημιστικούς σκοπούς. Αντίστοιχα, ένας πραγματικός λογαριασμός είναι πιο πιθανό να έχει αφήσει κενό το πεδίο για το URL.



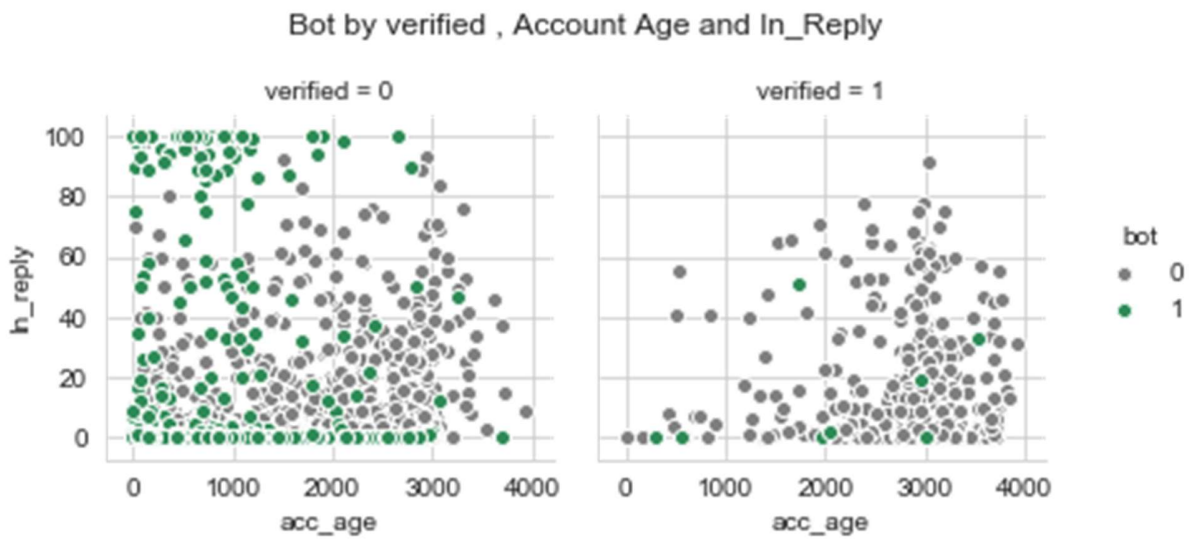
Εικόνα 9: Αριθμός Bot ή human με κενό URL

Επίσης, η εμφάνιση της λέξης Bot φαίνεται να είναι σημαντικά συχνότερη σε έναν Bot λογαριασμό, όπως φαίνεται στην Εικόνα 10.



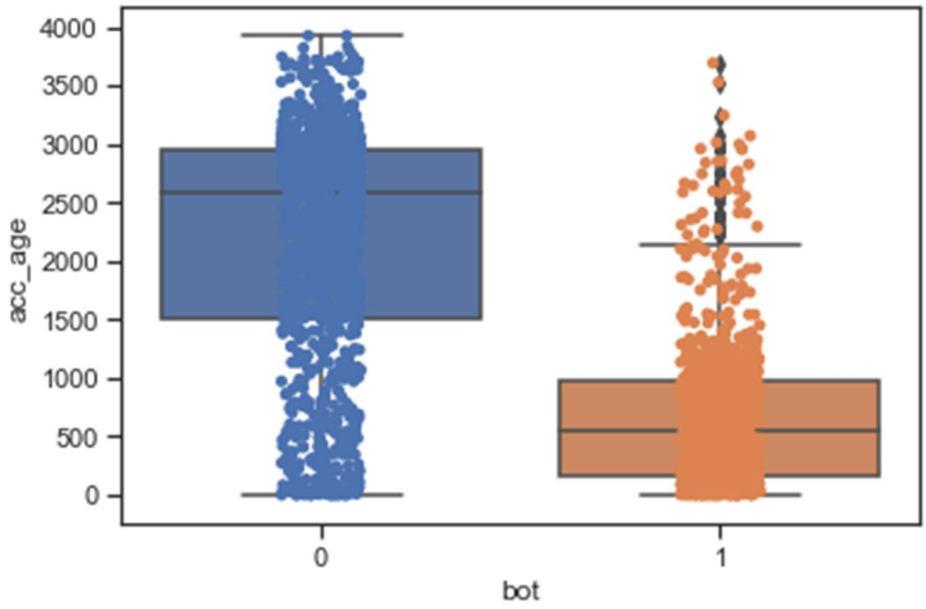
Εικόνα 10: Εμφάνιση της λέξης Bot στο όνομα

Η Εικόνα 11 περιγράφει τη σχέση μεταξύ επαληθευμένου λογαριασμού, της ηλικίας λογαριασμού και του αριθμού απαντήσεων (δεδομένα που συλλέγονται από το API του Twitter). Μπορεί να φανεί ότι ο λογαριασμός δημιουργήθηκε πρόσφατα και τα δεδομένα απάντησης είναι υψηλά, τότε ο λογαριασμός είναι Bot. Αντίστοιχα, εάν η ηλικία του λογαριασμού είναι μεγάλη και ο αριθμός των απαντήσεων δεν είναι χαμηλός (μεσαίο έως υψηλό), ο λογαριασμός είναι πιο πιθανό να είναι άνθρωπος. Επίσης, φαίνεται πως ένας επιβεβαιωμένος λογαριασμός είναι πολύ πιο πιθανό να είναι πραγματικός.



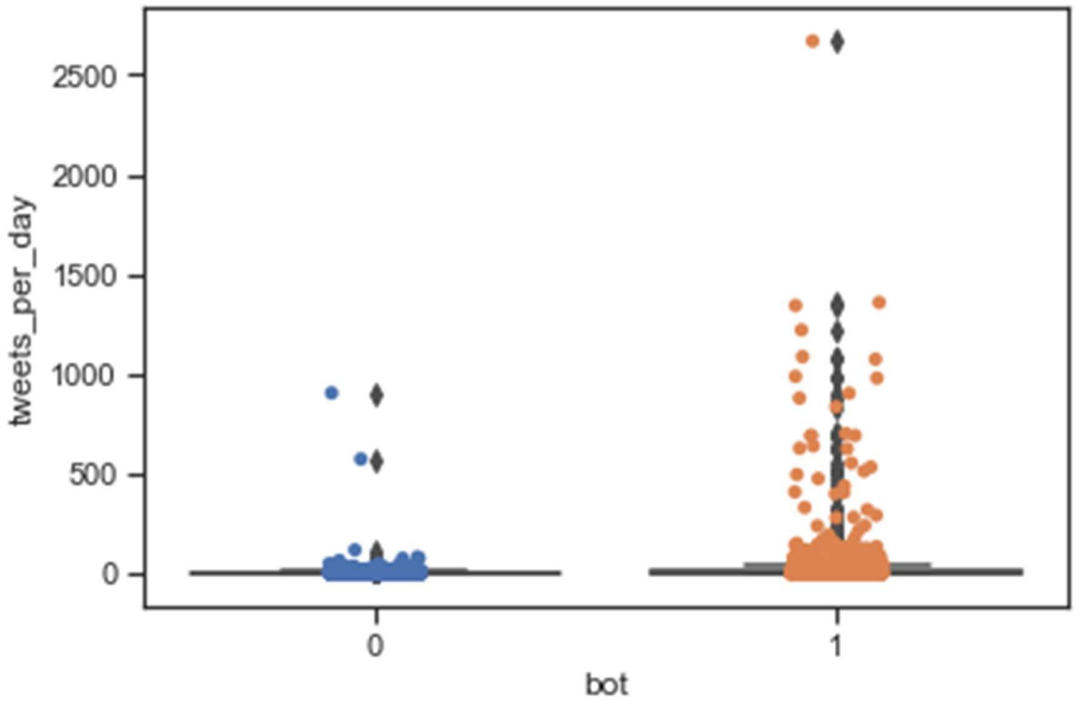
Εικόνα 11: Ηλικία Λογαριασμού σε σχέση με αριθμό απαντήσεων σε επιβεβαιωμένους και μη λογαριασμούς

Στην Εικόνα 12 απεικονίζεται τη σχέση μεταξύ ηλικίας λογαριασμού και συμπεριφοράς λογαριασμού. Μπορεί να φανεί, η συμπεριφορά ενός λογαριασμού που δημιουργήθηκε πρόσφατα είναι πιο πιθανό να είναι Bot. Από το γράφημα υπονοείται ότι το 75% των δεδομένων Bot, αποτελούνται από λογαριασμούς που δημιουργήθηκαν πρόσφατα, καθώς τα υπόλοιπα 25% των δεδομένων έχουν ηλικία λογαριασμού μικρότερη από τη μέση ηλικία ενός ανθρώπινου λογαριασμού.

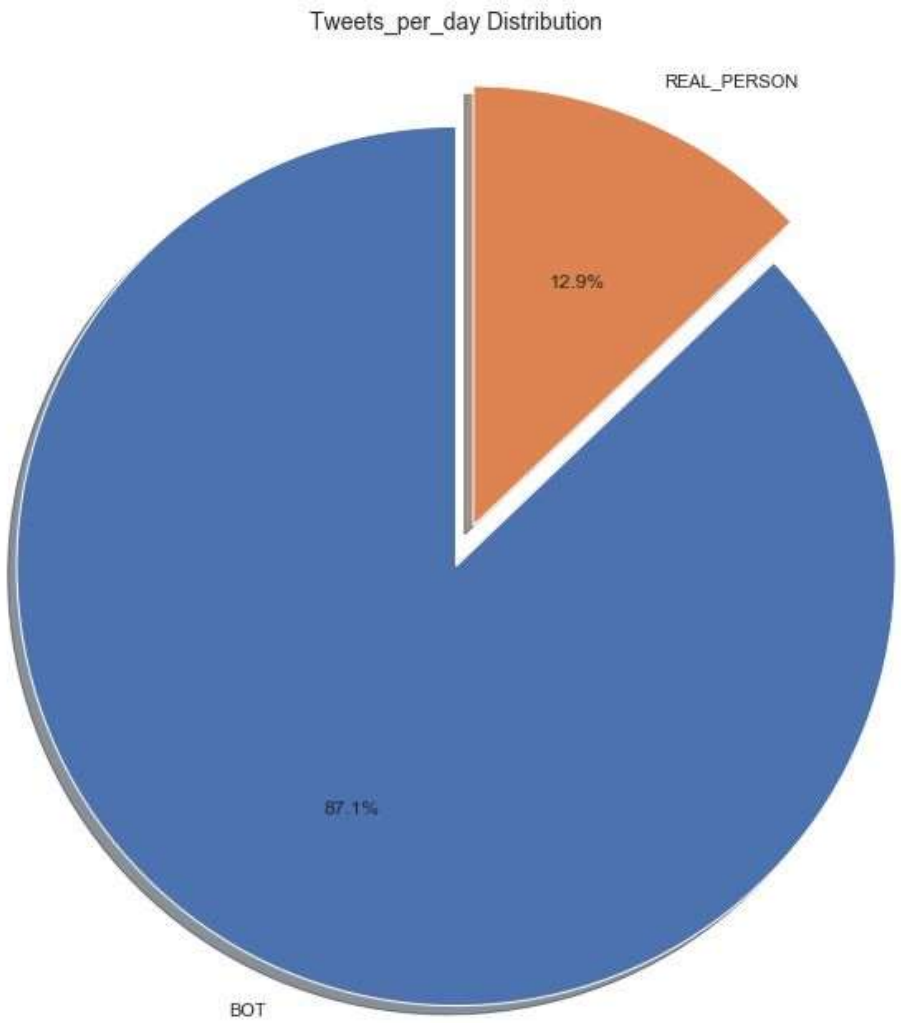


Εικόνα 12: Ηλικία Λογαριασμού Bot ή Human

Στις Εικόνες 13 και 14 φαίνεται η κατανομή μεταξύ Tweets ανά ημέρα και συμπεριφοράς λογαριασμού. Μπορεί να συναχθεί ότι εάν τα Tweets ανά ημέρα είναι υψηλά, τότε η συμπεριφορά του λογαριασμού δεν είναι ανθρώπινη.

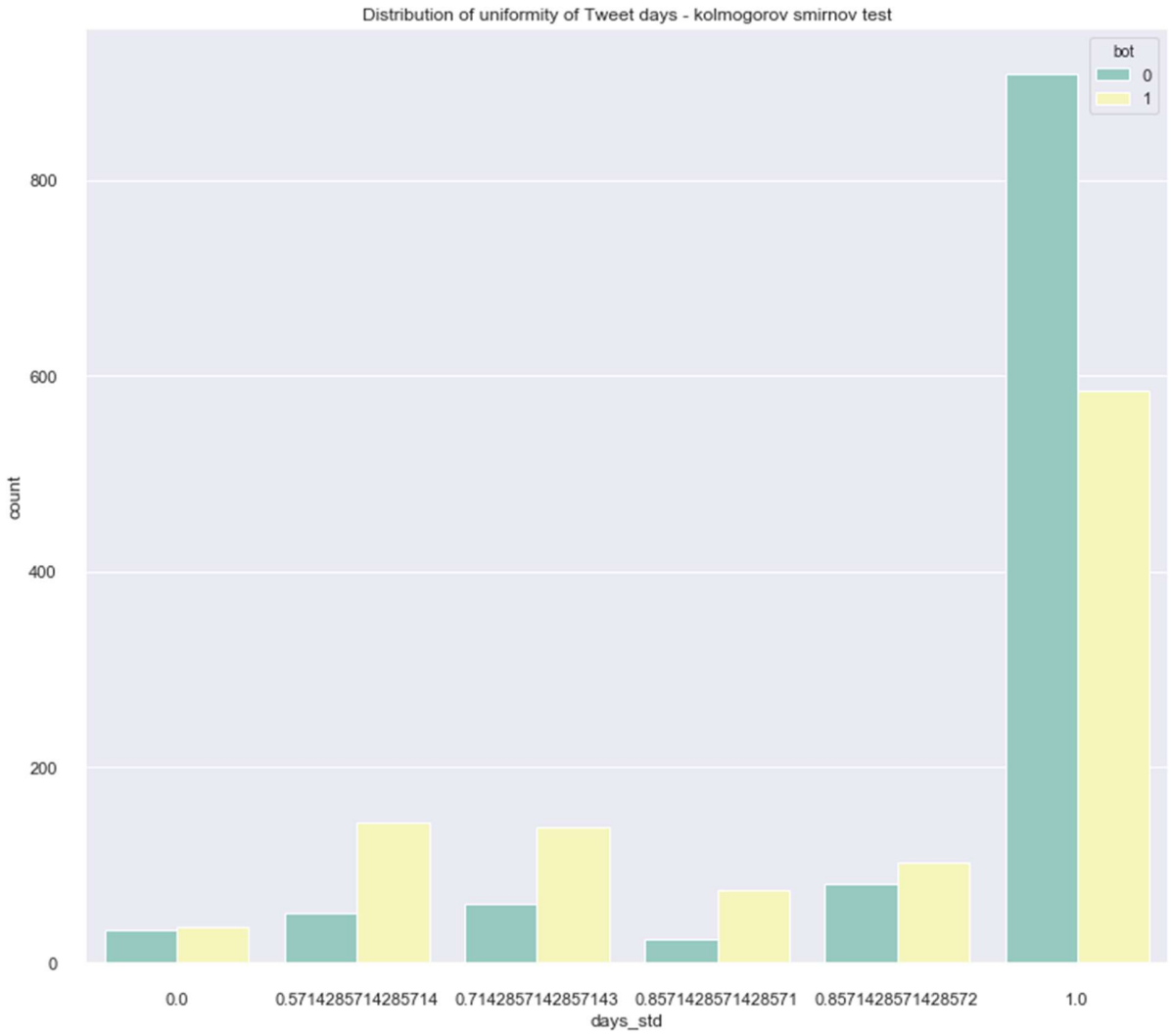


Εικόνα 13: Κατανομή Tweets ανά ημέρα



Εικόνα 14: Κατανομή Tweets ανά ημέρα

Η Εικόνα 15 περιγράφει την ομοιομορφία των Tweets που δημοσιεύονται στις μέρες της εβδομάδας. Από το γράφημα φαίνεται ότι αν η ομοιομορφία είναι 1 τότε ο λογαριασμός είναι πιθανότατα Bot.



Εικόνα 15: Κατανομή Ημερών Tweet

3.3.2 Ανάλυση Συναισθήματος

Αφενός η Ανάλυση Συναισθήματος αποτελεί πολύ ενδιαφέρον αντικείμενο έρευνας και μελέτης, αφετέρου μπορεί να βρει πολλές και χρήσιμες εφαρμογές. Η καινούρια πληροφορία που προστίθεται στις ήδη υπάρχουσες που βρίσκονται στο διαδίκτυο, είναι ότι μπορεί να παραχθεί ουσιαστικά μια περίληψη συναισθήματος στα δεδομένα που αναλύονται, κάνοντας πολύ γρήγορο και απλό στον εκάστοτε ενδιαφερόμενο να διαμορφώσει μια εικόνα δίχως να χρειαστεί να μελετήσει σε βάθος την κάθε άποψη που εκφράζεται.

Τα τελευταία χρόνια, επιπρόσθετα, πολλές εταιρείες έχουν αναλάβει το έργο της Ανάλυσης Συναισθήματος, καθώς αρκετά είναι και τα εργαλεία αυτόματης Ανάλυσης Συναισθήματος που έχουν αναπτυχθεί και είναι στη διάθεση οποιουδήποτε απλού χρήστη.

Η Ανάλυση Συναισθήματος σε κείμενο, χρησιμοποιείται από εταιρίες που επιθυμούν να ερευνήσουν την αποδοχή των προϊόντων τους από το ευρύ κοινό, να σχεδιάσουν τις μελλοντικές τους κινήσεις και να βελτιώσουν ενδεχόμενες ατέλειές τους. Ακόμη, δημοφιλή πρόσωπα αξιοποιούν την Ανάλυση Συναισθήματος διαδικτυακής πληροφορίας για να μάθουν την γνώμη που έχει το κοινό προς το πρόσωπό τους, ενώ απλοί χρήστες μπορούν να διαμορφώσουν μια γενική εικόνα για κάποιο προϊόν, έναν προορισμό ή μια ταινία που τους ενδιαφέρει και τυχαίνει να έχει αξιολογηθεί από άλλους χρήστες. Επίσης, με την χρήση της Ανάλυσης Συναισθήματος έχουν γίνει προσπάθειες και για πρόβλεψη αποτελεσμάτων πολιτικών εκλογών.

Στον τομέα της Ανάλυσης Συναισθήματος, η Μηχανική Μάθηση χρησιμοποιείται ώστε να είναι σε θέση το μοντέλο μάθησης να εκτιμήσει αυτόματα το συναισθηματικό περιεχόμενο ενός οποιουδήποτε κειμένου του δοθεί προς ανάλυση, εφόσον εκπαιδευτή με μια μέθοδο Μηχανικής Μάθησης.

Με σκοπό την περαιτέρω επεξεργασία και ανάλυση των δεδομένων, χρησιμοποιήθηκε το εργαλείο WatsonToneAnalyzer της IBM, για την εισαγωγή του συναισθήματος στα δεδομένα.

Η υπηρεσία IBM Watson ™ ToneAnalyzer χρησιμοποιεί γλωσσική ανάλυση για τον εντοπισμό συναισθηματικών και γλωσσικών τόνων σε γραπτό κείμενο. Η υπηρεσία μπορεί να αναλύσει τον τόνο τόσο σε επίπεδο εγγράφου όσο και σε επίπεδο προτάσεων.

Παρακάτω περιγράφονται οι τόνοι γενικής χρήσης που μπορεί να επιστρέψει η υπηρεσία. Ένας τόνος του οποίου η βαθμολογία είναι μικρότερη από 0,5 παραλείπεται, υποδεικνύοντας ότι το συναίσθημα είναι απίθανο να γίνει αντιληπτό στο περιεχόμενο. Μια βαθμολογία μεγαλύτερη από 0,75 δείχνει μια μεγάλη πιθανότητα να γίνει αντιληπτός ο τόνος.

Τόνος	Περιγραφή
Θυμός	Ο θυμός προκαλείται λόγω αδικίας, σύγκρουσης, ταπείνωσης, αμέλειας ή προδοσίας. Εάν ο θυμός είναι ενεργός, το άτομο επιτίθεται στον στόχο, προφορικά ή σωματικά. Εάν ο θυμός είναι παθητικός, το άτομο σιωπά σιγά και αισθάνεται ένταση και εχθρότητα. (Ένας συναισθηματικός τόνος.)
Φόβος	Ο φόβος είναι μια απάντηση στον επικείμενο κίνδυνο. Είναι ένας μηχανισμός επιβίωσης που ενεργοποιείται ως αντίδραση σε κάποια αρνητική διέγερση. Ο φόβος μπορεί να είναι ήπια προσοχή ή ακραία φοβία. (Ένας συναισθηματικός τόνος.)
Χαρά	Η χαρά (ή ευτυχία) έχει αποχρώσεις απόλαυσης, ικανοποίησης και ευχαρίστησης. Η χαρά φέρνει μια αίσθηση ευεξίας, εσωτερικής ειρήνης, αγάπης, ασφάλειας και ικανοποίησης. (Ένας συναισθηματικός τόνος.)
Θλίψη	Η θλίψη δείχνει ένα αίσθημα απώλειας και μειονεξίας. Όταν ένα άτομο είναι

	ήσυχο, λιγότερο ενεργητικό και αποσυρθεί, μπορεί να συναχθεί ότι αισθάνεται θλίψη. (Ένας συναισθηματικός τόνος.)
Αναλυτικός	Ένας αναλυτικός τόνος δείχνει τη συλλογιστική ενός ατόμου και την αναλυτική στάση για τα πράγματα. Ένα αναλυτικό άτομο μπορεί να γίνει αντιληπτό ως διανοητικό, λογικό, συστηματικό, χωρίς συναισθήματα ή απρόσωπο. (Ένας τόνος γλώσσας.)
Βέβαιος	Ένας τόνος με αυτοπεποίθηση δείχνει το βαθμό βεβαιότητας ενός ατόμου. Ένα άτομο με αυτοπεποίθηση μπορεί να γίνει αντιληπτό ως σίγουρο, συλλεγόμενο, ελπιδοφόρο ή εγωιστικό. (Ένας τόνος γλώσσας.)
Διστακτικός	Ένας διστακτικός τόνος δείχνει τον βαθμό αναστολής ενός ατόμου. Ένα διστακτικό άτομο μπορεί να θεωρηθεί αμφισβητήσιμο, αμφίβολο ή συζητήσιμο. (Ένας τόνος γλώσσας.)

Πίνακας 1: Συναισθήματα Tone Analyzer

Στην προκειμένη περίπτωση, το συνολικό κείμενο των τελευταίων 100 Tweet από κάθε χρήστη τροφοδοτήθηκε στο ToneAnalyzer ώστε να αποκτηθεί μια αίσθηση του συνολικού τόνου για τον χρήστη. Το αποτέλεσμα που προέκυψε από την ανάλυση είναι το συναίσθημα ή τα συναισθήματα του εκάστοτε κειμένου με την αντίστοιχη βαθμολογία.

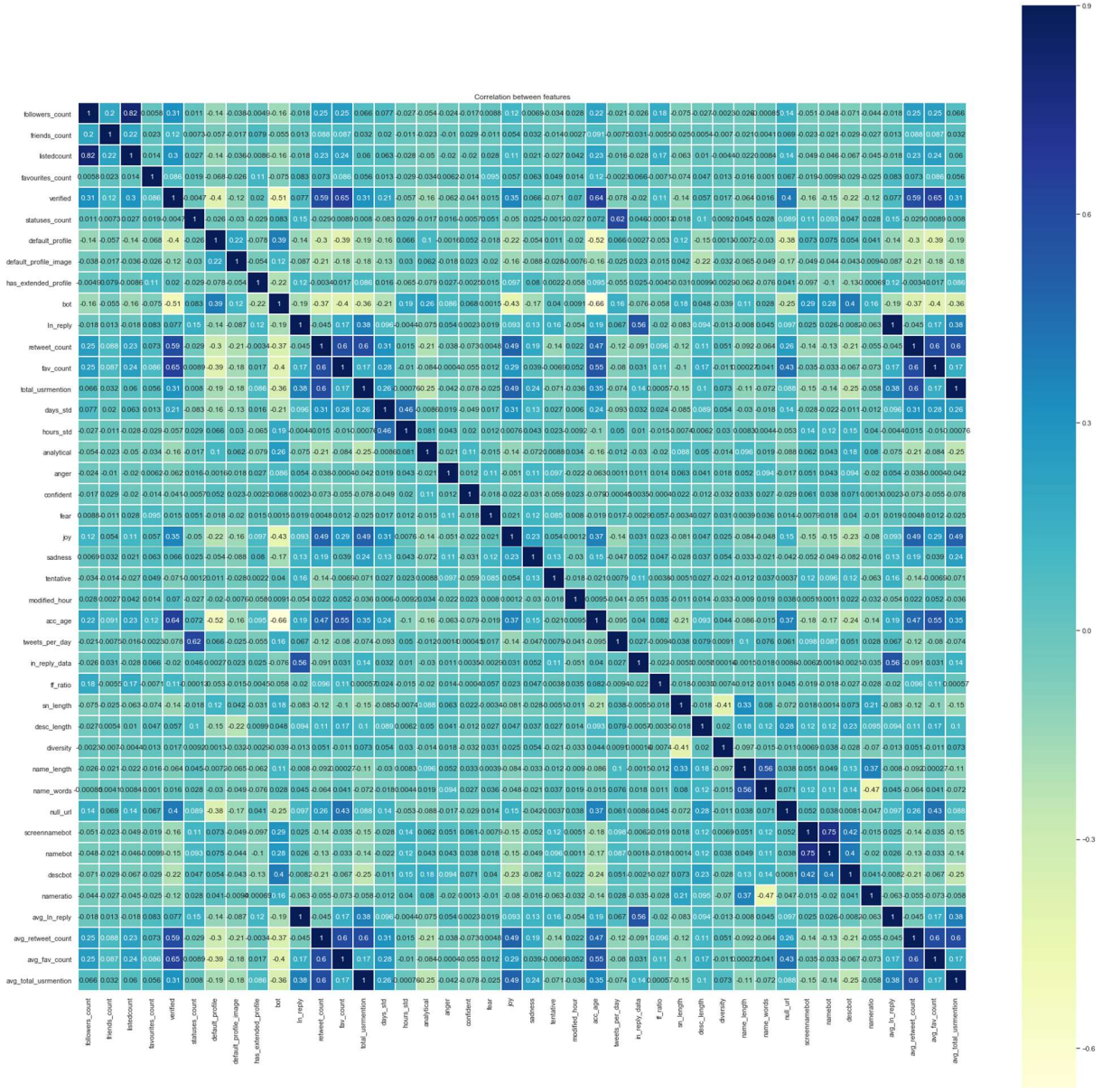
3.3.3 Μήτρα Συσχέτισης Χαρακτηριστικών

Στην Εικόνα 16 απεικονίζεται η Μήτρα Συσχέτισης των συνολικών δεδομένων, χρησιμοποιώντας την μέθοδο Pearson. Μια συσχέτιση Pearson είναι ένας αριθμός μεταξύ -1 και 1 που δείχνει τον βαθμό στον οποίο δύο μεταβλητές σχετίζονται γραμμικά. Θα πρέπει να σημειωθεί ότι για τα χαρακτηριστικά που εμφανίζουν υψηλή συσχέτιση μεταξύ τους, θα πρέπει να αφαιρεθεί ένα εκ των δυο στην τελική επιλογή χαρακτηριστικών. Τα χαρακτηριστικά με υψηλή συσχέτιση όπως προέκυψαν είναι τα εξής:

- Tweets_per_day – statuses_count
- followers_count – listedcount
- nameBot – screennameBot

Επίσης, προέκυψαν και τα παρακάτω με απόλυτη συσχέτιση 1 μεταξύ τους, το οποίο φυσικά ήταν αναμενόμενο.

- In_reply – avg_In_reply
- reTweet_count – avg_reTweet_count
- fav_count – avg_fav_count
- total_usrmention – avg_total_usrmention



Εικόνα 16: Μήτρα Συσχέτισης Χαρακτηριστικών

3.3.4 Αναδρομική Εξάλειψη Χαρακτηριστικών – Recursive Feature Elimination

Η εύρεση των βέλτιστων χαρακτηριστικών, που πρέπει να χρησιμοποιηθούν για την εκπαίδευση μοντέλου Machine Learning, δεν αποτελεί δύσκολη διαδικασία, ωστόσο υπάρχουν πολλές μέθοδοι που μπορούν να χρησιμοποιηθούν. Μέθοδοι όπως η Ανάλυση Κύριων Συστατικών (Principal Component Analysis) είναι αρκετά καλές, χωρίς όμως να επιστρέφουν

ως αποτέλεσμα ποια χαρακτηριστικά είναι τα πιο σημαντικά - επιστρέφουν τα κύρια στοιχεία που είναι στην πραγματικότητα συνδυασμοί χαρακτηριστικών.

Προκειμένου να αντιμετωπιστεί αυτό το ζήτημα και να εντοπιστούν σημαντικά χαρακτηριστικά που επηρεάζουν την πρόβλεψη, η τεχνική Recursive Feature Elimination χρησιμοποιείται. Και πιο συγκεκριμένα η Αναδρομική Εξάλειψη Χαρακτηριστικών με Διασταυρούμενη Επικύρωση (RFECV), επειδή χρησιμοποιείται συχνότερα από την επιλογή χωρίς διασταυρούμενη επικύρωση.

Η διασταυρούμενη επικύρωση, γνωστή ως cross-validations, αποτελεί μια τεχνική για την αξιολόγηση μοντέλων ML εκπαιδεύοντας διάφορα μοντέλα ML σε υποσύνολα των διαθέσιμων δεδομένων εισόδου, με τελική αξιολόγησή τους στο συμπληρωματικό υποσύνολο των δεδομένων.

Η αναδρομική εξάλειψη χαρακτηριστικών (RFE) αποτελεί βασικά μια επιλογή των μεταβλητών με έλεγχο προς τα πίσω. Αυτή η τεχνική ξεκινά δημιουργώντας ένα μοντέλο σε ολόκληρο το σύνολο των μεταβλητών και υπολογίζοντας μια βαθμολογία σπουδαιότητας για κάθε μεταβλητή. Στη συνέχεια αφαιρούνται οι λιγότερο σημαντικές μεταβλητές, το μοντέλο ξαναχτίζεται και οι βαθμολογίες σπουδαιότητας υπολογίζονται ξανά. Στην πράξη, ο αναλυτής είναι εκείνος που καθορίζει τον αριθμό των υποσυνόλων των μεταβλητών που θα χρησιμοποιηθούν για την πρόβλεψη και αξιολόγηση, καθώς και το μέγεθος κάθε υποσυνόλου. Επομένως, το μέγεθος υποσυνόλου είναι μια παράμετρος συντονισμού για την RFE. Το μέγεθος υποσυνόλου που βελτιστοποιεί τα κριτήρια απόδοσης χρησιμοποιείται για την επιλογή των προβλέψεων με βάση τις βαθμολογίες σπουδαιότητας. Στη συνέχεια χρησιμοποιείται το βέλτιστο υποσύνολο που υπολογίστηκε για την εκπαίδευση του τελικού μοντέλου.

Η επιλογή προς τα πίσω χρησιμοποιείται συχνά με RandomForest μοντέλα για δύο λόγους. Πρώτον, το τυχαίο δάσος τείνει να μην αποκλείει μεταβλητές από την εξίσωση πρόβλεψης. Ο λόγος σχετίζεται με τη φύση των ensemble μεθόδων. Η αυξημένη απόδοση στα σύνολα σχετίζεται με την ποικιλία των μοντέλων που απαρτίζουν μια ensemble μέθοδο.

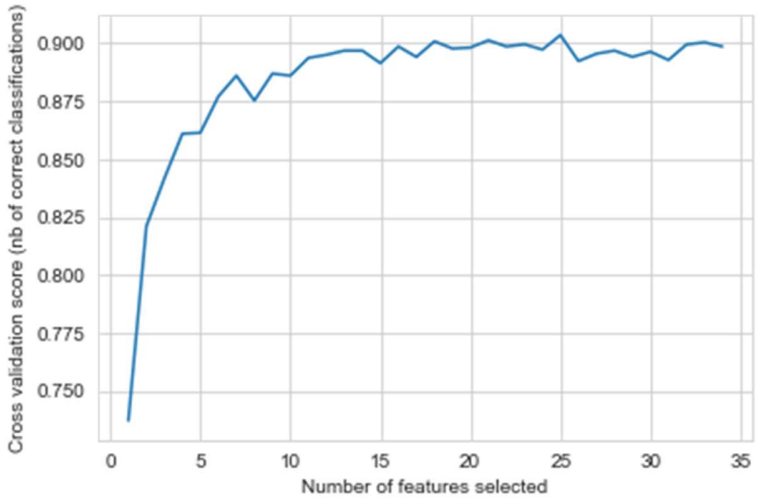
Ο δεύτερος λόγος για τον οποίο χρησιμοποιείται τυχαίο δάσος με RFE είναι επειδή αυτό το μοντέλο έχει μια γνωστή εσωτερική μέθοδο για τη μέτρηση της σημασίας χαρακτηριστικών, καθώς επίσης ολόκληρο το σετ προβλέψεων χρησιμοποιείται για τον υπολογισμό της κατάταξης χαρακτηριστικών.

Ένα αξιοσημείωτο ζήτημα με τη μέτρηση της σημασίας στα δέντρα σχετίζεται με την πολυσυγγραμμικότητα. Εάν υπάρχουν πολύ συσχετισμένοι προγνωστικοί παράγοντες σε ένα σύνολο εκπαίδευσης, που είναι χρήσιμοι για την πρόβλεψη του αποτελέσματος, τότε ποιος προγνωστικός δείκτης επιλέγεται για την κατανομή των δειγμάτων είναι ουσιαστικά μια τυχαία επιλογή. Στο σενάριο όπου ένα σύνολο συσχετισμένων μεταβλητών χρησιμοποιείται, η προγνωστική απόδοση του μοντέλου φαίνεται να μην επηρεάζεται από εξαιρετικά

συσχετισμένα και χρήσιμα χαρακτηριστικά. Ωστόσο, ο πλεονασμός των χαρακτηριστικών χαμηλώνει τις βαθμολογίες σπουδαιότητας των μεταβλητών.

Για τον λόγο αυτό, συστήνεται όταν χρησιμοποιείται RFE με τυχαίο δάσος ή άλλα δέντρα, να μην χρησιμοποιούνται οι μεταβλητές με υψηλή συσχέτιση κατά την εκπαίδευση του μοντέλου. Αφαιρούνται συνεπώς, οι μεταβλητές που φαίνονται να έχουν υψηλή συσχέτιση μεταξύ τους, όπως προέκυψε από την Μήτρα Συσχέτισης παραπάνω.

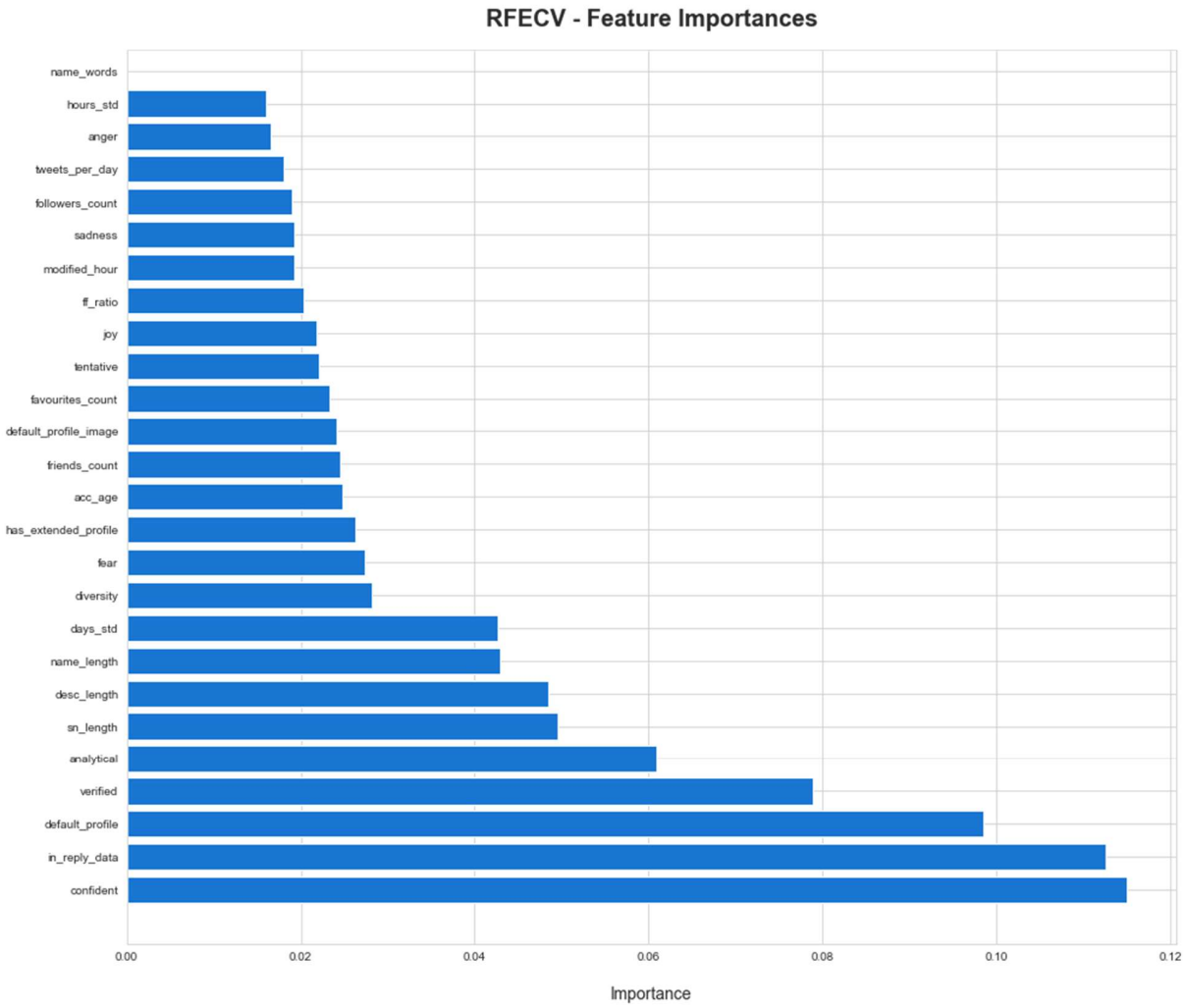
Από την Εικόνα 17, προκύπτει ότι ο βέλτιστος αριθμός των χαρακτηριστικών που πρέπει να χρησιμοποιηθούν για την εκπαίδευση των δεδομένων είναι 25. Αυτά τα χαρακτηριστικά αναμένεται να μειώσουν το overfitting και να αυξήσουν την ακρίβεια του μοντέλου.



Εικόνα 17: Βέλτιστος αριθμός χαρακτηριστικών

Τα χαρακτηριστικά που προέκυψαν από τον RFE έλεγχο είναι τα παρακάτω, αντικατοπτρίζονται και στην Εικόνα 18 ανάλογα την σημαντικότητά τους, με πιο σημαντικό τον τόνο “confident”όπως προέκυψε από την ανάλυση συναισθήματος.

```
['anger', 'followers_count', 'friends_count', 'favourites_count',
'verified', 'default_profile', 'has_extended_profile', 'days_std',
'hours_std', 'fear', 'diversity', 'analytical', 'joy',
'modified_hour', 'acc_age', 'Tweets_per_day', 'ff_ratio', 'sn_length',
'desc_length', 'name_length','sadness', 'tentative',
'default_profile','in_reply_data', 'confident']
```



Εικόνα 18: Βέλτιστος αριθμός χαρακτηριστικών αναλόγως σημαντικότητας

4. Προετοιμασία για εκτέλεση αλγορίθμων

Πριν την διαδικασία εκπαίδευσης του μοντέλου, πραγματοποιείται προ επεξεργασία στα δεδομένα. Αρχικά, παράγονται polynomial features και μετά κανονικοποιούνται (standardizing).

Πολλοί αλγόριθμοι μηχανικής μάθησης λειτουργούν καλύτερα όταν τα χαρακτηριστικά βρίσκονται σε σχετικά παρόμοια κλίμακα και πλησιάζουν κανονικά. Ο StandardScaler που χρησιμοποιείται είναι μέθοδος του scikit-learning για την προεπεξεργασία δεδομένων για μηχανική εκμάθηση.

Scale γενικά σημαίνει αλλαγή του εύρους των τιμών. Το σχήμα της κατανομής δεν αλλάζει. Η κανονικοποίηση γενικά σημαίνει αλλαγή των τιμών έτσι ώστε η τυπική απόκλιση κατανομής από τον μέσο όρο να ισούται με ένα. Εξάγει κάτι πολύ κοντά σε μια κανονική κατανομή. Πολλοί αλγόριθμοι μηχανικής μάθησης αποδίδουν καλύτερα ή συγκλίνουν γρηγορότερα όταν τα χαρακτηριστικά είναι σε σχετικά ίδια κλίμακα η σε κανονική κατανομή. Παραδείγματα τέτοιων αλγορίθμων είναι:

- linear and logistic regression
- nearest neighbors
- neural networks
- support vector machines with radial bias kernel functions
- principal components analysis
- linear discriminant analysis

Ο StandardScaler τυποποιεί ένα χαρακτηριστικό αφαιρώντας το μέσο όρο και μετά κλιμακώνοντας τη διακύμανση μονάδας. Διακύμανση μονάδας σημαίνει διαίρεση όλων των τιμών με την τυπική απόκλιση. Το StandardScaler οδηγεί σε κατανομή με τυπική απόκλιση ίση με 1. Η διακύμανση ισούται με 1 επίσης, επειδή η διακύμανση = τετράγωνη τυπική απόκλιση. Και 1 εις το τετράγωνο ισούται 1.

Όταν κανονικοποιούνται τα δεδομένα, λαμβάνονται οι μεταβλητές, από τις οποίες αφαιρείται η μέση τιμή και εκφράζονται με τυπική απόκλιση (επίσης γνωστό ως z-score).

Το Polynomial Features δημιουργεί έναν νέο πίνακα με όλους τους συνδυασμούς χαρακτηριστικών πολυωνύμων με τον δοθέντα βαθμό.

Για να αποφευχθεί το under-fitting, χρειάζεται αύξηση της πολυπλοκότητας του μοντέλου. Για παράδειγμα ένα γραμμικό μοντέλο μπορεί να μετατραπεί εύκολα σε πολυωνυμικό. Για να μετατραπούν τα αρχικά χαρακτηριστικά σε υψηλότερους όρους χρησιμοποιείται η Polynomial Features κλάση από scikit-learn.

4.1 Αναζήτηση Πλέγματος – Τυχαία αναζήτηση

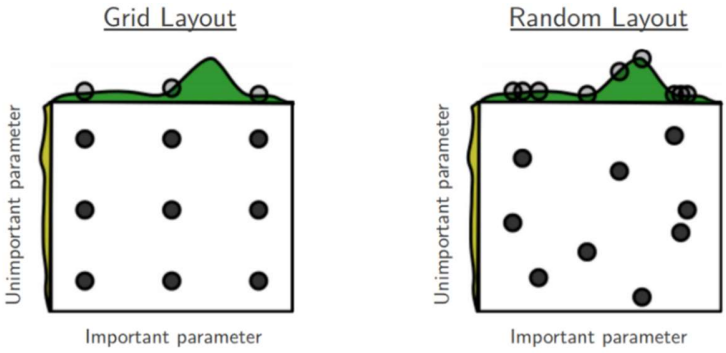
Κατά την διεξαγωγή πειραμάτων με έναν αλγόριθμο επιβλεπόμενης μάθησης, χρειάζεται πολύ συχνά η εύρεση των υπερπαραμέτρων του αλγορίθμου που οδηγούν στην βέλτιστη απόδοση του. Οι υπερπαράμετροι ορίζονται πριν τη διαδικασία μάθησης και καθορίζουν ιδιότητες του μοντέλου όπως η πολυπλοκότητα του, η μορφή της επιφάνειας απόφασης κ.α. Μια δημοφιλής τεχνική που χρησιμοποιείται για αυτό το σκοπό είναι η αναζήτηση πλέγματος (gridsearch). Κατά την τεχνική αυτή πραγματοποιείται εξαντλητικός έλεγχος σε συνδυασμούς υπερπαραμέτρων με προκαθορισμένο εύρος τιμών προκειμένου να βρεθεί ο συνδυασμός που έχει την καλύτερη απόδοση.

Χρησιμοποιήθηκε την βιβλιοθήκη Gridsearchcv στους ταξινομητές Decision Trees και Gradient Boosting. Ενώ στην αναζήτηση των καλύτερων παραμέτρων για τον ταξινομητή Random Forest χρησιμοποιείται RandomizedSearchCV.

Η αναζήτηση πλέγματος μπορεί να θεωρηθεί ως εξαντλητική αναζήτηση για την επιλογή ενός μοντέλου. Στην Αναζήτηση πλέγματος, δημιουργείται ένα πλέγμα τιμών υπερπαραμέτρων και για κάθε συνδυασμό, εκπαιδεύεται ένα μοντέλο και βαθμολογεί τα δεδομένα δοκιμών. Σε αυτήν την προσέγγιση, δοκιμάζεται κάθε συνδυασμός τιμών υπερπαραμέτρων που μπορεί να είναι πολύ αναποτελεσματικός. Για παράδειγμα, η αναζήτηση 20 διαφορετικών τιμών παραμέτρων για καθεμία από τις 4 παραμέτρους θα απαιτήσει 160.000 δοκιμές εγκυρότητας. Αυτό ισοδυναμεί με 1.600.000 προσαρμογές μοντέλου και 1.600.000 προβλέψεις εάν χρησιμοποιείται διασταυρούμενη επικύρωση 10 φορές. Ενώ η Scikit Learn προσφέρει τη λειτουργία GridSearchCV για απλοποίηση της διαδικασίας, θα ήταν μια εξαιρετικά δαπανηρή εκτέλεση τόσο σε υπολογιστική ισχύ όσο και σε χρόνο.

Αντίθετα, η τυχαία αναζήτηση δημιουργεί ένα πλέγμα τιμών υπερπαραμέτρων και επιλέγει τυχαίους συνδυασμούς για να εκπαιδεύσει το μοντέλο και να βαθμολογήσει. Αυτό επιτρέπει τον ελέγχο τον αριθμό των συνδυασμών παραμέτρων που επιχειρούνται. Ο αριθμός των επαναλήψεων αναζήτησης ορίζεται με βάση το χρόνο ή τους πόρους. Το Scikit Learn προσφέρει τη συνάρτηση RandomizedSearchCV για αυτήν τη διαδικασία.

Παρόλο που είναι πιθανό το RandomizedSearchCV να μην βρει τόσο ακριβές αποτέλεσμα όσο το GridSearchCV, επιλέγει απροσδόκητα το καλύτερο αποτέλεσμα πιο συχνά από ό, τι όχι και σε ένα κλάσμα του χρόνου που χρειάζεται το GridSearchCV. Δεδομένων των ίδιων πόρων, η τυχαία αναζήτηση μπορεί ακόμη και να ξεπεράσει την αναζήτηση πλέγματος. Αυτό μπορεί να απεικονιστεί στο παρακάτω γράφημα όταν χρησιμοποιούνται συνεχείς παράμετροι.



Εικόνα 19: Grid vs Random Layout

5. Αξιολόγηση αλγορίθμων Επιβλεπόμενης Μάθησης

Τα μοντέλα επιβλεπόμενης μάθησης παρουσιάζουν μεγάλες διαφορές στην απόδοση τους ανάλογα με τον αλγόριθμο υλοποίησης τους, τα χαρακτηριστικά των δεδομένων εισόδου, τις επιθυμητές εξόδους και τη φύση του προβλήματος που καλούνται να επιλύσουν. Για να προσδιοριστεί η απόδοση ενός μοντέλου χρησιμοποιούνται διάφορες μέθοδοι και μετρικές. Μια μετρική σχετίζεται με τον δείκτη απόδοσης που λαμβάνεται υπόψη για την ποσοτικοποίηση της επιτυχίας ενός μοντέλου, ενώ οι μέθοδοι καθορίζουν τη διαδικασία με την οποία εξάγεται αυτή η μετρική. Παρακάτω αναφέρονται οι μετρικές και οι μέθοδοι που χρησιμοποιούνται συχνά για την αξιολόγηση των δυαδικών ταξινομητών και εξηγείται ποιες επιλέχθηκαν στην παρούσα εργασία.

5.1 Overfitting

Για την πραγματοποίηση πειραμάτων με έναν ταξινομητή επιβλεπόμενης μάθησης υπάρχουνε διαφορετικές μέθοδοι σχετικά με τον τρόπο που γίνεται η εκπαίδευση και η αξιολόγηση του συστήματος. Μερικές από τις πιο διαδεδομένες μεθόδους είναι οι Holdout, k-fold cross-validation και leave-one-outcross-validation.

Κατά τη διαδικασία του k-fold cross-validation τα δεδομένα χωρίζονται σε k υποσύνολα. Σε κάθε έναν από τους k γύρους της μεθόδου αυτής, τα k-1 υποσύνολα χρησιμοποιούνται ως training set για τον αλγόριθμο και το υποσύνολο που απομένει χρησιμοποιείται ως test set με βάση το οποίο εξάγονται οι επιθυμητές μετρικές. Στο τέλος όλης της διαδικασίας υπολογίζεται ο μέσος όρος για κάθε μετρική και οι προκύπτοντες αριθμοί αποτελούν τον δείκτη απόδοσης του εκάστοτε αλγορίθμου μηχανικής μάθησης. Τα πλεονεκτήματα της μεθόδου k-fold cross-validation είναι ότι υπάρχει μικρή μεροληψία (bias) στη διαδικασία μάθησης του αλγορίθμου καθώς ένα μεγάλο ποσοστό των δεδομένων χρησιμοποιούνται για την εκπαίδευση του αλγορίθμου και επιπλέον μειώνεται η διακύμανση (variance) καθώς όλα τα δεδομένα περνούν από το test set. Με άλλα λόγια, το k-fold cross-validation προσφέρει αντικειμενικά αποτελέσματα για την επίδοση ενός μοντέλου αποφεύγοντας το overfitting του αλγορίθμου στα δεδομένα μέσω της χρήσης των διαφορετικών folds, και προσφέρει έναν τρόπο εξαγωγής συμπερασμάτων για το πόσο καλά γενικεύει αυτό το μοντέλο και πόσο μπορεί δυνητικά να αποδώσει καλά σε καινούρια δεδομένα.

Στην παρούσα διπλωματική εργασία επιλέχθηκε η μέθοδος 10-fold cross-validation για να υπολογιστεί το accuracy των αλγορίθμων επιβλεπόμενης μάθησης που δοκιμάστηκαν. Ο λόγος επιλογής αυτής της μεθόδου είναι ότι προσφέρει ένα καλό συνδυασμό κατανάλωσης υπολογιστικών πόρων και αξιόπιστων αποτελεσμάτων.

5.2 Μετρικές Αξιολόγησης

Οι μετρικές που χρησιμοποιούνται συνήθως για την αξιολόγηση της απόδοσης ενός δυαδικού ταξινομητή αναφέρονται ακολούθως μαζί με τους αντίστοιχους τύπους:

$$accuracy = \frac{\#correct\ predictions}{\#all\ predictions}$$

Στην παρούσα μελέτη χρησιμοποιήθηκε ως μετρική των μοντέλων επιβλεπόμενης μάθησης το accuracy. Το accuracy επιλέχθηκε για δυο λόγους. Αρχικά, είναι μια μετρική που χρησιμοποιείται ευρέως σε παρόμοιες έρευνες και αποτελεί την μετρική που τονίζεται κυρίως για τη σύγκριση των αποτελεσμάτων μεταξύ των ερευνών αυτών. Δεύτερον, μέσω του accuracy λαμβάνεται αρκετά αξιόπιστη πληροφορία για την απόδοση των δυαδικών ταξινομητών.

Confusion matrix:

Ένας πίνακας σύγχυσης είναι μια σύνοψη των αποτελεσμάτων πρόβλεψης για ένα πρόβλημα ταξινόμησης. Ο αριθμός των σωστών και λανθασμένων προβλέψεων συνοψίζεται με τιμές μέτρησης και κατανέμεται ανά κατηγορία. Αυτό είναι το κλειδί για τον πίνακα σύγχυσης. Ο πίνακας σύγχυσης δείχνει τους τρόπους με τους οποίους το μοντέλο ταξινόμησης μπερδεύεται όταν κάνει προβλέψεις. Μας δίνει μια εικόνα όχι μόνο για τα λάθη που γίνονται από έναν ταξινομητή, αλλά το πιο σημαντικό είναι τα είδη σφαλμάτων που γίνονται.

	Class 1 Predicted	*Class 2 Predicte*
Class 1 Actual	**TP**	**FN**
Class 2 Actual	**FP**	**TN**

- Class 1 : Positive
- Class 2 : Negative

Στην περίπτωση των δεδομένων που χρησιμοποιούνται για το test οι Bot λογαριασμοί είναι στο σύνολο 280 , ενώ οι πραγματικοί – φυσικοί λογαριασμοί είναι 278.

Ως θετική κλάση θεωρείται να είναι οι φυσικοί λογαριασμοί ενώ ως αρνητική κλάση οι Bots λογαριασμοί.

Άλλη μετρική που λαμβάνεται από το classification report είναι το **Recall**. Το Recall υπολογίζεται ως εξής:

$$Recall = \frac{TP}{(TP + FN)}$$

όπου TP: true positive και FN: False negative.

Το recall υπολογίζει από όλες τις θετικές κλάσεις πόσες έχουν προβλεφθεί σωστά από τον αλγόριθμό, και ιδανικά θα πρέπει η τιμή του να είναι υψηλή. Στην περίπτωση που εξετάζεται, από όλους τους λογαριασμούς που είναι όντως φυσικοί, υπολογίζεται πόσοι ποσοτικά όντως ταξινομήθηκαν ως πραγματικοί χρήστες και όχι Bots.

Η επόμενη μετρική, η οποία εξάγεται πάλι από το classification report είναι η Precision, και υπολογίζεται ως:

$$Precision = \frac{TP}{(TP + FP)}$$

όπου FP : false positive.

Η μετρική precision υπολογίζει από όλες τις θετικές κλάσεις που έχουν προβλεφθεί, πόσες είναι όντως θετικές, δηλαδή στην παρούσα περίπτωση, η ακρίβεια του συστήματος κατά πόσο έχει προβλέψει λογαριασμούς ως πραγματικούς χρήστες, ενώ είναι στην πραγματικότητα Bots. Αποτελεί δηλαδή μια μετρική που υπολογίζει την ποιότητα των εξαγόμενων συμπερασμάτων.

Ακόμη μία μετρική που εξάγεται από το classification report είναι το F1 score.

Το **F1 score** είναι μια καλή μετρική για να βρίσκει την ισορροπία ανάμεσα σε precision και recall. Μπορεί να ερμηνευθεί ως σταθμισμένος μέσος όρος της ακρίβειας και ανάκλησης, όπου η βαθμολογία F1 φτάνει την καλύτερη τιμή της στο 1 και τη χειρότερη βαθμολογία στο 0.

$$F1 = \frac{2 * (precision * recall)}{(precision + recall)}$$

Η σύγκριση της απόδοσης των αλγορίθμων μπορεί επίσης να πραγματοποιηθεί με την καμπύλη Receiver Operator Characteristic (**ROC**), μια μέτρηση αξιολόγησης για προβλήματα δυαδικής ταξινόμησης. Η καμπύλη ROC αποτελεί μια καμπύλη πιθανότητας που σχεδιάζει το TPR έναντι FPR σε διάφορες τιμές κατωφλίου και ουσιαστικά διαχωρίζει το «σήμα» από το «θόρυβο». Η περιοχή κάτω από την καμπύλη (**AUC**) είναι το μέτρο της ικανότητας ενός ταξινομητή να διακρίνει μεταξύ τάξεων και χρησιμοποιείται ως σύνοψη της καμπύλης ROC.

Σε μια καμπύλη ROC, μια υψηλότερη τιμή άξονα X υποδηλώνει μεγαλύτερο αριθμό ψευδών θετικών από ό, τι τα αληθινά αρνητικά. Ενώ η υψηλότερη τιμή του άξονα Y δείχνει υψηλότερο

αριθμό θετικών από τα ψευδώς αρνητικά. Έτσι, η επιλογή του κατωφλίου εξαρτάται από την ικανότητα εξισορρόπησης μεταξύ Ψευδών θετικών και Ψευδών αρνητικών.

6. Αξιολόγηση Αποτελεσμάτων

Οι παρακάτω αλγόριθμοι δοκιμάστηκαν αρχικά στα original χαρακτηριστικά του dataset, αφαιρώντας μόνο τα features με υψηλή συσχέτιση, όπου παρατηρήθηκε overfitting. Μια βασική πρόκληση με το overfitting, και με τη μηχανική μάθηση γενικά, είναι ότι δεν μπορεί να είναι γνωστό εκ των προτέρων πόσο καλά θα αποδώσει το μοντέλο σε νέα δεδομένα.

Για να αντιμετωπιστεί αυτό, μπορεί να διαιρεθεί το αρχικό σύνολο δεδομένων σε ξεχωριστά υποσύνολα εκπαίδευσης και δοκιμής. Αυτή η μέθοδος μπορεί να προσεγγίσει πόσο καλά θα αποδώσει το μοντέλο σε νέα δεδομένα.

Στη περίπτωση που μελετήθηκε, το μοντέλο απέδιδε με ένα ποσοστό 3-4% καλύτερα στο training set από ότι στο testing set.

Με την χρήση των τελικών χαρακτηριστικών:

'anger', 'followers_count', 'friends_count', 'favourites_count', 'verified', 'default_profile', 'has_extended_profile', 'days_std', 'hours_std', 'fear', 'diversity', 'analytical', 'joy', 'modified_hour', 'acc_age', 'Tweets_per_day', 'ff_ratio', 'sn_length', 'desc_length', 'name_length', 'sadness', 'tentative', 'default_profile', 'in_reply_data', 'confident'

Η ακρίβεια στο testing set αυξήθηκε σημαντικά ενώ παράλληλα μειώθηκε το overfitting, παρατηρώντας χαμηλότερο AUC Score στο training set.

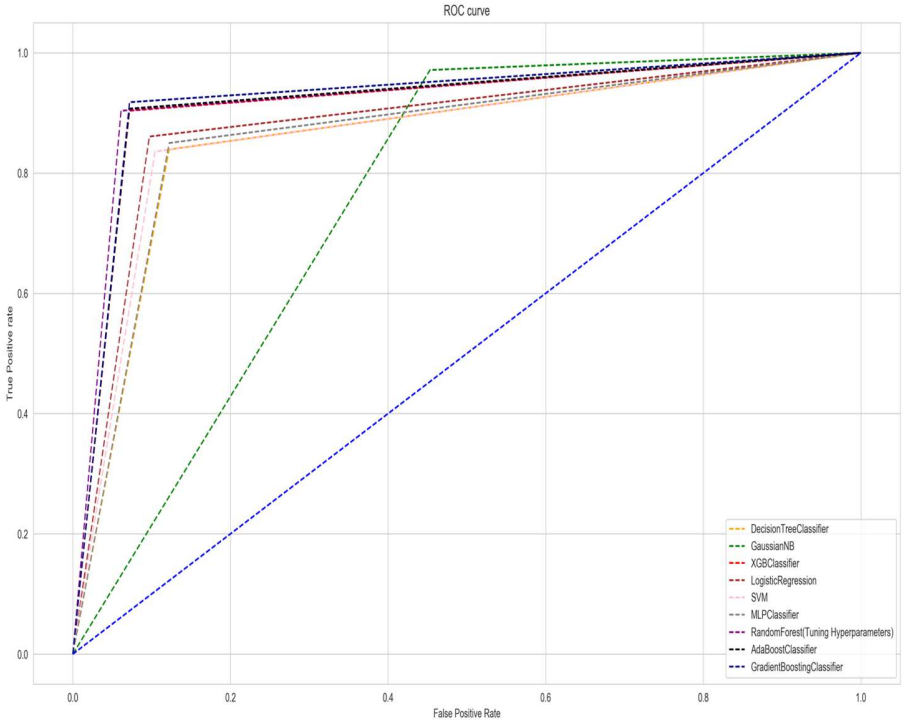
Συγκεντρωτικά, σε αύξουσα σειρά, φαίνονται παρακάτω όλες οι μετρικές που υπολογιστήκαν στο test set στα 'τελικά' δεδομένα, δηλαδή στα δεδομένα που έχουν πλέον υποστεί επεξεργασία, καθαρισμό, προσθήκη νέων χαρακτηριστικών (όπως sentiment) καθώς και αφαίρεση χαρακτηριστικών.

ΑΛΓΟΡΙΘΜΟΣ	**ACCURACY**	**PRECISION** (avg / total)	**RECALL** (avg / total)	**F1-SCORE** (avg / total)
Gaussian Naive Bayes	0.759	0.82	0.76	0.75
Logistic Regression	0.881	0.88	0.88	0.88
Decision Trees (Classification tree)	0.858	0.86	0.86	0.86
Neural Networks - MLPClassifier	0.863	0.86	0.86	0.86
Support vector machine	0.865	0.87	0.87	0.87

Random forest	0.892	0.90	0.90	0.90
XGBoost	0.915	0.92	0.92	0.92
AdaBoost	0.917	0.92	0.92	0.92
Random Forest with Tuning Hyperparameters	0.921	0.92	0.92	0.92
GradientBoostingClassifier	0.922	0.92	0.92	0.92

Παρατηρείται ότι οι πιο δυνατοί αλγόριθμοι , με ποσοστό 0.92% σε όλες τις μετρικές, είναι οι: XGBoost, AdaBoost, Random Forest with Tuning Hyperparameters, GradientBoostingClassifier

Επίσης, παρακάτω φαίνεται ένα συγκεντρωτικό διάγραμμα AUC-ROC curve.



Εικόνα 20: AUC-ROC Curve Διάγραμμα

Παρατηρείται και από το παραπάνω να υπερισχύουν ο Random Forest (με υπερπαραμέτρους) και ο Gradiend Boosting ταξινομητής.

Με βάση τα αποτελέσματα που ελήφθησαν χρησιμοποιώντας τις μετρήσεις που καθορίστηκαν στην ενότητα αποτελεσμάτων και το αποτέλεσμα ακρίβειας των δεδομένων δοκιμής που ελήφθησαν από το Kaggle, ο αλγόριθμος τυχαίων δασών που χρησιμοποιεί την ενίσχυση ADA έχει αποδώσει καλύτερα από τους υπόλοιπους αλγόριθμους. Στη συνέχεια, συνδυάστηκε η λογιστική παλινδρόμηση με αναζήτηση πλέγματος για επιλογή βέλτιστων δυνατοτήτων, και η ακρίβεια ήταν περίπου 90 τοις εκατό. Χρησιμοποιήθηκαν συνεπώς τεχνικές συνόλων όπως το τυχαίο δάσος για καλύτερη απόδοση. Κατά τη δημιουργία του μοντέλου, βελτιώθηκε περαιτέρω η ακρίβεια χρησιμοποιώντας Tuning Hyperparameters σε τυχαίο δασικό εκτιμητή όπου η βαθμολογία ακρίβειας του μοντέλου ήταν 92,6 τοις εκατό.

Ως γενικότερο συμπέρασμα, μελετήθηκε η συμπεριφορά του εκάστοτε λογαριασμού Twitter χρησιμοποιώντας χαρακτηριστικά που παρέχονται από το Twitter API και με περαιτέρω τροποποίηση των εν λόγω δεδομένων. Για την καλύτερη κατανόηση της συμπεριφοράς του λογαριασμού Twitter, έχουν χρησιμοποιηθεί διάφορες τεχνικές μηχανικής εκμάθησης με διαφορετικές τεχνικές προ-επεξεργασίας. Ως διαφορετικοί τύποι εκτιμητών και λήψη δεδομένων γίνεται μια προσέγγιση της πλήρους κατανόησης της φύσης του λογαριασμού.

7. Topic Modeling

Η μοντελοποίηση θέματος είναι ένας κλάδος επεξεργασίας φυσικής γλώσσας χωρίς επίβλεψη που χρησιμοποιείται για την αναπαράσταση ενός κειμένου με τη βοήθεια διαφόρων θεμάτων, που μπορούν να εξηγήσουν καλύτερα τις υποκείμενες πληροφορίες σε ένα συγκεκριμένο έγγραφο. Αυτό μπορεί να εξεταστεί από άποψη ομαδοποίησης, αλλά με μια διαφορά. Αντί για αριθμητικά χαρακτηριστικά, υπάρχει μια συλλογή λέξεων που θα μπορούσε να ομαδοποιηθεί με τέτοιο τρόπο ώστε κάθε ομάδα να αντιπροσωπεύει ένα θέμα σε ένα έγγραφο[3]. Το λεγόμενο Topic Modeling μπορεί να υλοποιηθεί με πολλούς τρόπους, ένας από αυτούς είναι η χρήση του Latent Dirichlet Allocation (LDA) αλγορίθμου, ο οποίος χρησιμοποιείται και αναλύεται παρακάτω.

7.1 Latent Dirichlet Allocation Analysis (LDA)

Ένα στατιστικό μοντέλο για την ανίχνευση θεμάτων σε σύνολα κειμένων είναι το μοντέλο Λανθάνουσας Κατανομής Dirichlet. Αυτό το μοντέλο αναπτύχθηκε από τους David Blei et al [7]. Όσον αφορά την πρακτική εφαρμογή του, αποτελεί έναν από τους σημαντικότερους αλγόριθμους που έχουν αναπτυχθεί και αναλυθεί για την εξαγωγή θεμάτων από μεγάλα αρχεία κειμένων. Συνεπώς, το πεδίο εφαρμογής του είναι αρκετά ευρύ, αφού δοθέντος ενός κατάλληλου λεξιλογίου, μπορεί να οδηγήσει σε ακριβή αποτελέσματα.

Η ιδέα πίσω από αυτό το μοντέλο είναι απλή: το κάθε κείμενο αποτελείται από ένα σύνολο θεμάτων, τα οποία προσδιορίζονται από συγκεκριμένες λέξεις συνδεδεμένες με κάποια πιθανότητα. Μια συλλογή κειμένων μοιράζεται κοινά θέματα περιεχομένου, αλλά το κάθε κείμενο εκφράζει τα θέματα σε διαφορετικές αναλογίες. Ο τρόπος που κατανέμονται τα θέματα σε μια συλλογή κειμένων, σε συνδυασμό με την αναλογία των θεμάτων και την κατανομή των λέξεων στο κάθε κείμενο αποτελούν την “κρυμμένη δομή” της συλλογής. Για να αναγνωριστεί η θεματολογία των κειμένων, η οποία αποτελεί την άγνωστη μεταβλητή του μοντέλου, το υπολογιστικό πρόβλημα ανάγεται στην αντιστροφή της διαδικασίας παραγωγής τους: αν υπολογιστεί η δομή με την οποία δημιουργήθηκαν αρχικά τα κείμενα, έχει επιτευχθεί η ανίχνευση των θεμάτων που την αποτελούν

Πιο συγκεκριμένα, το LDA είναι ένα γενετικό μοντέλο πιθανοτήτων που υποθέτει ότι κάθε θέμα είναι ένα μείγμα πάνω από ένα υποκείμενο σύνολο λέξεων και κάθε έγγραφο είναι ένα μείγμα πάνω από ένα σύνολο πιθανοτήτων θέματος.

[3] https://medium.com/analytics-vidhya/topic-modeling-using-lda-and-gibbs-sampling-explained-49d49b3d1045

Το LDA λειτουργεί αρχικά επιλέγοντας ένα σύνολο θεμάτων και έπειτα για κάθε θέμα επιλέγοντας ένα σύνολο λέξεων. Για να βρει τα θέματα αντιστρέφει τους μηχανισμούς αυτής της διαδικασίας. Για να το κάνει αυτό κάνει τα εξής για κάθε έγγραφο m:

1. Υποθέτει ότι υπάρχουν k θέματα σε όλα τα έγγραφα

2. Διανέμει αυτά τα θέματα k στο έγγραφο m (αυτή η κατανομή είναι γνωστή ως α και μπορεί να είναι συμμετρική ή ασύμμετρη), εκχωρώντας σε κάθε λέξη ένα θέμα.

3. Για κάθε λέξη w στο έγγραφο m, ας υποθέτει ότι το θέμα της είναι λάθος, αλλά σε κάθε άλλη λέξη αντιστοιχεί το σωστό θέμα.

4. Πιθανοτικά εκχωρεί τη λέξη με ένα θέμα που βασίζεται σε δύο πράγματα:

   - Ποια θέματα βρίσκονται στο έγγραφο m.
   - Πόσες φορές η λέξη w έχει εκχωρηθεί ένα συγκεκριμένο θέμα σε όλα τα έγγραφα[4].

7.2 Χρήση Αλγορίθμου σε Δεδομένα Εκλογών US 2018

Με στόχο τον έλεγχο της εκπαίδευσης και ακρίβειας των αλγορίθμων σε διαφορετικά dataset, χρησιμοποιήθηκαν δεδομένα των Εκλογών της Αμερικής το 2018 από το site Bot Repository. Η επιλογή αυτών τον δεδομένων έγινε λόγω υψηλού ενδιαφέροντος στο πως αυτοματοποιημένοι λογαριασμοί του Twitter επηρεάζουν τις εκλογές.

Στις εκλογές του 2016, ανάμεσα στα δύο προεδρικά debates, παρατηρήθηκε ότι ένα μεγάλο μέρος των προεκλογικών Tweets του Trump και της Clinton προήλθαν από Bot λογαριασμούς[5]. Σε ανάλυση που έχει πραγματοποιηθεί στα συγκεκριμένα δεδομένα φαίνεται έντονα το πόσο δημοφιλής και ενεργός στα Social Media ήταν ο Trump σε αντίθεση με την πολιτική αντίπαλό του, Clinton. Με την χρήση Bot λογαριασμών είναι εύκολο να δημιουργηθεί μια ψευδαίσθηση δημοσιότητας πάνω σε επίκαιρα θέματα. Παράλληλα, μπορούν να κάνουν χρήστες πιο δημοφιλείς και άλλους λιγότερο με το να αναδημοσιεύουν απόψεις του υποψήφιου που υποστηρίζουν ή επικρίνοντας επανειλημμένα τον αντίπαλό του. Η πραγματική προπαγανδιστική δύναμη των Bots όμως έγκειται στο ότι δύσκολα ανιχνεύονται και πολλοί δεν γνωρίζουν την ύπαρξή τους. Οι απόψεις που δημοσιεύονται γίνονται πιο δυνατές όταν περισσότεροι άνθρωποι φαίνεται να πιστεύουν σε αυτές και τα Bots δημιουργούν αυτήν την ψευδαίσθηση.

[4] https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd
[5] https://www.theatlantic.com/technology/archive/2016/11/election-Bots/506072/

Το παραπάνω μοτίβο πάνω στα Tweets χρηστών την περίοδο των εκλογών οδήγησε, επίσης, σε εκτενέστερη επεξεργασία κειμένου και στην μοντελοποίηση θέματος, η οποία αναλύεται παρακάτω.

Συνοπτικά, χρησιμοποιείται το νέο dataset για τον έλεγχο ότι το Topic Modeling λειτουργεί σωστά καθώς θα περίμενε κάποιος λέξεις που αφορούν τις εκλογές και την ψηφοφορία. Έπειτα από τον αρχικό έλεγχο στο σύνολο των εκλογών, το Topic Modeling εφαρμόζεται επίσης στα αρχικά δεδομένα.

7.2.1 Δεδομένα και Εξόρυξη Δεδομένων

Το dataset που χρησιμοποιήθηκε περιείχε 7500 λογαριασμούς χαρακτηρισμένους ως φυσικούς λογαριασμούς ή Bot. Από τα συγκεκριμένα δεδομένα έλειπαν τα Tweets και σχετικές πληροφορίες πάνω σε αυτά, και με σκοπό την άντληση των απαραίτητων πληροφοριών χρησιμοποιήθηκε το API του Twitter. Αρχικά, τέθηκε χρονικός περιορισμός στην περίοδο των εκλογών, ώστε να εξαχθούν όλα τα 'screen_name' και userId από το Twitter API. Τα Tweets που αντλήθηκαν αφορούν 50 φυσικούς λογαριασμούς, καθώς δεν ήταν εφικτή η εξόρυξη των Tweets που είχαν δημιουργηθεί από τους υπόλοιπους λογαριασμούς. Τα νέα εξαχθέντα δεδομένα συνενώθηκαν με τα υπάρχοντα δεδομένα στην έως τώρα ανάλυση, ωστόσο λόγω του μικρού μεγέθους του νέου dataset εκλογών καθώς και των νέων δεδομένων που περιλαμβάνουν μόνο πραγματικούς χρήστες, δεν πραγματοποιήθηκε περαιτέρω εκπαίδευση και πρόβλεψη στο νέο σύνολο δεδομένων. Ωστόσο, παρακάτω περιγράφεται η διαδικασία που ακολουθήθηκε σχετικά με την εξαγωγή νέων χαρακτηριστικών κειμένου, όπως το Topic Modeling.

7.2.2 Εξαγωγή Χαρακτηριστικών από Κείμενο

Για την ανάλυση των δεδομένων πραγματοποιήθηκε επεξεργασία και ανάλυση κειμένου πάνω στο περιεχόμενο του description των χρηστών του καινούριου dataset.

Τα χαρακτηριστικά που εξάχθηκαν είναι τα παρακάτω:

Πλήθος hashtags ανά περιγραφή: Τα hashtag έχουν ιδιαίτερη σημασία για την εξαγωγή αποτελεσμάτων πάνω στο κείμενο διότι δίνουν μια γενική ιδέα του κειμένου.

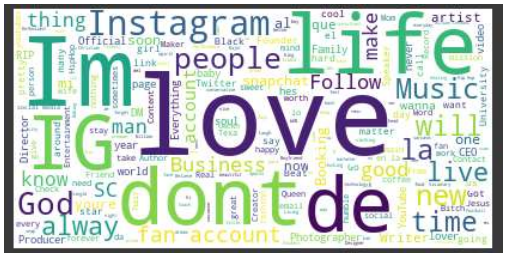
Πλήθος tag (αναφορά στο κείμενο σε χρήστη): Όταν υπάρχει αναφορά σε χρήστες στα Tweets ή στην παρούσα εργασία στην περιγραφή του λογαριασμού ενός χρήστη, σημαίνει ότι το κείμενο αποτελεί συζήτηση ανάμεσα σε δύο μέλη.

Πλήθος λέξεων ανά κείμενο: παρατηρείται ότι πολλά κείμενα χρειάζονται περισσότερες λέξεις για να περιγράψουν μια κατάσταση. Στην περίπτωση της παρούσας μελέτης προκύπτει ότι θετικά-χαρούμενα κείμενα έχουν περισσότερες λέξεις από τα κείμενα με πιο μέτριο ή θυμωμένο ύφος, πράγμα που διαπιστώνεται και στη συναισθηματική ανάλυση κειμένου που πραγματοποιήθηκε.

Μέσο μήκος λέξης: Το συγκεκριμένο χαρακτηριστικό αποκτά σημασία με την οπτική ότι παρατηρείται σε επιστημονικά παραδείγματος χάρη κείμενα το μέσο μήκος λέξης να είναι μεγαλύτερο από το μέσο μήκος λέξης σε κείμενα με καθημερινό ύφος συζήτησης.

Sentiment analysis με χρήση της βιβλιοθήκη Text Blob εφαρμόστηκε στην συνέχεια στα νέα κείμενα που προέκυψαν, καθώς επίσης και τεχνικές για καθαρισμό κειμένου, όπως η αφαίρεση ειδικών χαρακτήρων # και @, url links και σημείων στίξης. Επιπλέον, όλο το κείμενο μετατράπηκε σε χαρακτήρες lowercase.

Προκειμένου να επαληθευτεί η προ επεξεργασία, δημιουργήθηκε ένα σύννεφο λέξεων χρησιμοποιώντας το πακέτο wordcloud με σκοπό την οπτική αναπαράσταση των πιο κοινών λέξεων. Είναι βασική η κατανόηση των δεδομένων και η διασφάλιση ότι ακολουθείται ορθή στρατηγική, ειδάλλως απαιτείται περισσότερη προ επεξεργασία πριν από την εκπαίδευση του μοντέλου. Το αποτέλεσμα της επεξεργασίας φαίνεται παρακάτω:



Εικόνα 21: Σύννεφο λέξεων πριν την προετοιμασία δεδομένων

Παρατηρείται ότι δεν έχει γίνει σωστός καθαρισμός στο κείμενο. Οι λέξεις που προκύπτουν έχουν νόημα, ωστόσο περαιτέρω καθαρισμός απαιτείται, όπως περιγράφεται στην συνέχεια.

7.2.3 Καθαρισμός Δεδομένων Κειμένου

Η πιο σημαντική ενέργεια που πρέπει να προηγηθεί του αλγόριθμου μοντελοποίησης θεμάτων είναι η προ-επεξεργασία και ο καθαρισμός των Tweets. Τα Tweets ενδέχεται να περιέχουν μη τυπικά αγγλικά, κεφαλαία γράμματα, συνδέσμους, hashtag, @users, σημεία στίξης και emoticon. Ο σκοπός της ανάλυσης είναι η κατανόηση των θεμάτων για τα οποία μιλούν οι άνθρωποι, και όχι για ποιον μιλάνε ή τους συνδέσμους Ιστού που μοιράζονται. Ο στόχος του

καθαρισμού δεδομένων είναι η τελική μετατροπή του κειμένου με τέτοιο τρόπο, ώστε να μπορεί να χρησιμοποιηθεί ως είσοδος σε έναν αλγόριθμο.

Ο **Καθαρισμός** πραγματοποιείται με σκοπό να αφαιρεθούν τα λιγότερο χρήσιμα μέρη του κειμένου με αφαίρεση των stopwards, αφαίρεση σπανίων λέξεων, μέσα από την κατανόηση και μεταχείριση των κεφαλαίων χαρακτήρων και διάφορους άλλους τρόπους όπως περιγράφονται πιο κάτω.

Το κείμενο έχει συχνά μια ποικιλία κεφαλαίων που είθισται να αντικατοπτρίζει την αρχή των προτάσεων, δίνοντας έμφαση στα ουσιαστικά. Η πιο συνηθισμένη προσέγγιση είναι να μετατραπούν όλοι οι χαρακτήρες σε πεζούς, ωστόσο είναι σημαντικό να ληφθούν υπόψη λέξεις όπως "US" που αλλάζουν νόημα με αντίστοιχη μετατροπή.

Επιπρόσθετα, πραγματοποιήθηκε η αφαίρεση των stopwords από τα Tweets. Η πλειονότητα των λέξεων σε ένα δεδομένο κείμενο συνδέει τμήματα μιας πρότασης αντί να δείχνει θέματα, αντικείμενα ή πρόθεση. Λέξεις όπως "the" ή "and" θα πρέπει να αφαιρούνται συγκρίνοντας το κείμενο με μια λίστα από stopwards.

Ενδεικτικά παρατίθενται μερικά Tweets χωρίς αφαίρεση των stopwords :

```
0    washington ave memorial park donphan f iv 53 u...
1    rt syeddoha ty my friend willkommen dohalaptop...
2    matt lieber is a better and everlasting spouse...
3    single cell rna sequencing to dissect the mole...
4    rt rightwingangel dear trumptrain vote may 2nd...
```

Και μετά από την αφαίρεση:

```
0       washington ave memorial park donphan f iv 53 1...
1       syeddoha ty friend willkommen dohalaptops deut...
2       matt lieber better everlasting spouse matt lie...
3       single cell rna sequencing dissect molecular h...
4       rightwingangel dear trumptrain vote may 2nd pr...
```

Μπορεί να παρατηρηθεί εύκολα πως λέξεις όπως ‘my’, ‘is’, ‘a’, ‘and’ έχουν αφαιρεθεί.

Στην συνέχεια, αφαιρέθηκαν οι πολύ σπάνιες λέξεις όπως είναι ονόματα, επωνυμίες, ονόματα προϊόντων και κάποιοι χαρακτήρες θορύβου, λόγου χάριν html.

Παράδειγμα τέτοιων λέξεων είναι το παρακάτω:

```
<bound method Series.sort_value
omg                 1
joerogan            1
votexexxdnnlooks    1
waltz               1
record              1
natural             1
pet                 1
xexdxaxefxbxf       1
hey                 1
donated             1
westside            1
youxexxve           1
puglife             1
atlanta             1
quite               1
grou                1
resource            1
susanbanthony       1
donxexxt            1
dtype: int64>
```

Εν συνεχεία, με χρήση της βιβλιοθήκης Textblob, πραγματοποιήθηκε διόρθωση ορθογραφίας σε όλες τις λέξεις των Tweets με την μέθοδο correct(). Η διόρθωση ορθογραφίας διενεργήθηκε με σκοπό των περαιτέρω καθαρισμό του κειμένου, και την εξάλειψη παραλλαγών ορθογραφίας που ενδέχεται να εμφανίζονται στα κείμενα. Η μερική εξομάλυνση της γλωσσικής ετερογένειας, θα συμβάλει στον να παραχθούν ουσιαστικά αποτελέσματα.

Η **Κανονικοποίηση** αποτελείται από την χαρτογράφηση των ορολογιών που χρησιμοποιούνται ή από αφαιρέσεις λέξεων μέσω των διαδικασιών Stemming, Lemmazation και άλλων μορφών κανονικοποίησης. Το Stemming είναι μια διαδικασία όπου οι λέξεις μειώνονται στην ρίζα τους, αφαιρώντας ανεπιθύμητους χαρακτήρες και συνήθως την κατάληξη. Υπάρχουν διάφορα μοντέλα Stemming, όπως ενδεικτικά το Porter και το Snowball. Τα αποτελέσματα μπορούν να χρησιμοποιηθούν για τον εντοπισμό σχέσεων και ομοιότητας σε μεγάλα σύνολα δεδομένων.

Lemmazation είναι μια εναλλακτική μέθοδος παρόμοια με την Stemming, ωστόσο και οι δύο πραγματοποιούν κανονικοποίηση κειμένου (ή αλλιώς κανονικοποίηση λέξης) στον τομέα της Επεξεργασίας Φυσικής Γλώσσας (NLP – Natural Language Processing). Επίσης, χρησιμοποιούνται στην προετοιμασία κειμένου, λέξεων και γενικότερων εγγράφων για περαιτέρω επεξεργασία.

Οι γλώσσες που χρησιμοποιούνται αποτελούνται από λέξεις που συχνά παράγονται η μία από την άλλη μέσα από διάφορους γραμματικούς κανόνες χρησιμοποιώντας συνήθως κάποιο πρόθεμα ή κατάληξη. Οι μέθοδοι Stemming και Lemmazation βοηθούν συνεπώς στην εξαγωγή της λέξης βάσης ή ρίζας από τις παραγόμενες λέξεις. Η μέθοδος Stemming είναι διαφορετική από τη Lemmatization στην προσέγγιση που χρησιμοποιεί για να παράγει ριζικές μορφές λέξεων και τη λέξη που παράγεται.

Και οι δύο μέθοδοι χρησιμοποιούνται ευρέως σε συστήματα προσθήκης ετικετών, συστήματα κατάταξης, βελτιστοποίηση μηχανών αναζήτησης, αποτελέσματα αναζήτησης στο Web και ανάκτηση πληροφοριών.

Για την αγγλική γλώσσα, θα μπορούσε κάποιος να επιλέξει μεταξύ Porter Stemmer ή Lancaster Stemmer, με το Porter Stemmer να είναι το παλαιότερο που αρχικά αναπτύχθηκε το 1979. Το Lancaster Stemmer αναπτύχθηκε το 1990 και χρησιμοποιεί μια πιο επιθετική προσέγγιση από τον αλγόριθμο Porter Stemming.

Το Porter Stemmer χρησιμοποιεί αφαίρεση καταλήξεων για την παραγωγή στελεχών. Παρατηρήστε πώς το Porter Stemmer δίνει τη ρίζα (στέλεχος) της λέξης "cats" απλώς αφαιρώντας το "s" μετά τη λέξη cat. Πρόκειται για μια κατάληξη που προστίθεται στη λέξη cat για να την κάνει πληθυντικό ως cats. Αλλά αν παρατηρήσει κάποιος τις λέξεισ «trouble», «troubling» και «troubled» γίνονται «trouble» επειδή ο αλγόριθμος Porter Stemmer δεν ακολουθεί τη γλωσσολογία αλλά ένα σύνολο κανόνων για διαφορετικές περιπτώσεις που εφαρμόζονται σε φάσεις (βήμα προς βήμα) για τη δημιουργία των στελεχών. Αυτός είναι ο λόγος για τον οποίο το Porter Stemmer δεν δημιουργεί συχνά στελέχη που είναι πραγματικές αγγλικές λέξεις. Δεν διατηρεί έναν πίνακα αναζήτησης για πραγματικά στελέχη της λέξης, αλλά εφαρμόζει αλγοριθμικούς κανόνες για τη δημιουργία στελεχών. Χρησιμοποιεί τους κανόνες για να αποφασίσει εάν είναι συνετό να αφαιρεθεί μια κατάληξη.

Το Porter Stemmer είναι γνωστό τελικά για την απλότητα και την ταχύτητά του. Είναι συνήθως χρήσιμο σε περιβάλλοντα ανάκτησης πληροφοριών γνωστά ως περιβάλλοντα υπερύθρων (IR) για γρήγορη ανάκληση και ανάκτηση ερωτημάτων αναζήτησης. Σε ένα τυπικό IR, τα περιβαλλοντικά έγγραφα παρουσιάζονται ως διανύσματα λέξεων ή όρων. Οι λέξεις που έχουν το ίδιο στέλεχος θα έχουν παρόμοια σημασία. Για παράδειγμα:

CONNECT
CONNECTIONS------> CONNECT
CONNECTED------> CONNECT
CONNECTING------> CONNECT
CONNECTION------> CONNECT

Το Lancaster Stemmer (Paice-Husk stemmer) είναι ένας επαναληπτικός αλγόριθμος με κανόνες που αποθηκεύονται εξωτερικά. Ένας πίνακας που περιέχει περίπου 120 κανόνες χρησιμοποιώντας για κατάταξη το τελευταίο γράμμα μιας κατάληξης. Σε κάθε επανάληψη, προσπαθεί να βρει έναν εφαρμόσιμο κανόνα από τον τελευταίο χαρακτήρα της λέξης. Κάθε κανόνας καθορίζει είτε τη διαγραφή είτε την αντικατάσταση μιας κατάληξης. Εάν δεν υπάρχει τέτοιος κανόνας, τότε ο αλγόριθμος τερματίζει. Τερματίζει επίσης εάν μια λέξη ξεκινά με ένα φωνήεν και απομένουν μόνο δύο γράμματα ή εάν μια λέξη ξεκινά με ένα σύμφωνο και απομένουν μόνο τρεις χαρακτήρες. Διαφορετικά, εφαρμόζεται ο κανόνας και η διαδικασία επαναλαμβάνεται.

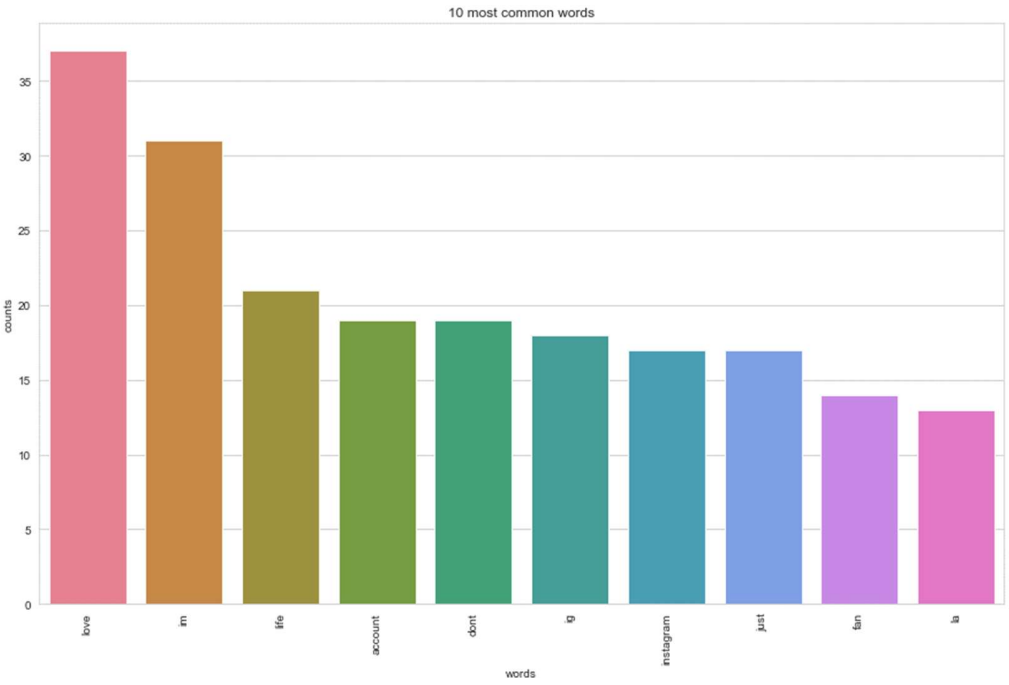
Το Lancaster Stemmer αν και απλό, μπορεί να προκαλέσει heavy stemming λόγω πολλαπλών επαναλήψεων ή over-stemming. Το over-stemming ενδέχεται να έχει ως αποτέλεσμα οι λέξεις να υπάρχουν ή να μην έχουν γλωσσικό νόημα.

Ο Porter Stemmer χρησιμοποιήθηκε τελικά, καθώς ο Lancaster φαίνεται να είναι πιο ‘επιθετικός’ stemming αλγόριθμος με αποτέλεσμα πολλές μικρές λέξεις να γίνονται εντελώς ασαφείς.

7.3 Προετοιμασία δεδομένων για εκτέλεση LDA

Στη συνέχεια, τα δεδομένα κειμένου μετατράπηκαν σε μορφή κατάλληλη ως είσοδος για την εκπαίδευση του μοντέλου LDA, ξεκινώντας μετατρέποντας τα έγγραφα σε μια απλή διανυσματική αναπαράσταση (Bag of Words BOW), και τελικώς, μετατρέποντας μια λίστα τίτλων σε λίστες διανυσμάτων, όλα με μήκος ίσο με το λεξιλόγιο.

Παρακάτω σχεδιάστηκαν οι δέκα πιο συχνές λέξεις με βάση το αποτέλεσμα αυτής της λειτουργίας (λίστα διανυσμάτων εγγράφων). Ως έλεγχος, αυτές οι λέξεις πρέπει επίσης να εμφανίζονται στο σύννεφο λέξεων.



Εικόνα 22: Λίστα Διανυσμάτων Εγγράφων

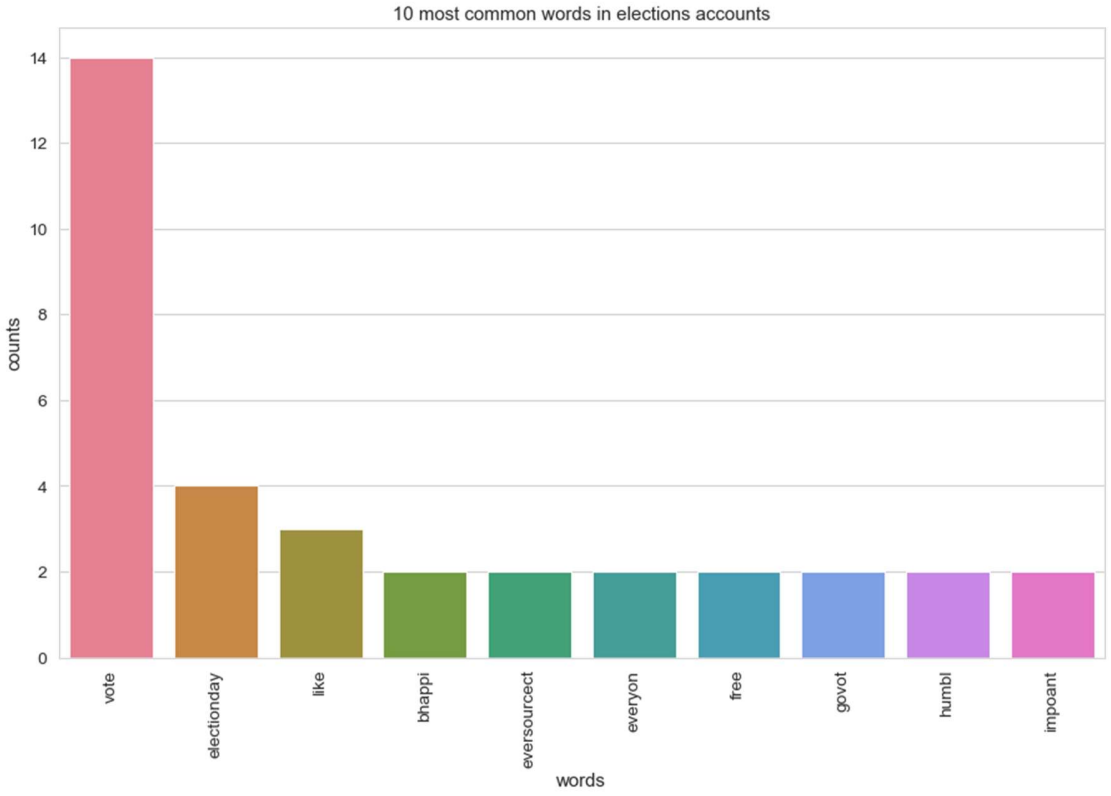
Επιβεβαιώνεται ότι οι λέξεις που παρουσιάζονται είναι οι ίδιες με εκείνες στο σύννεφο, καθώς αυτός είναι συχνά ένας τρόπος για την επαλήθευση του σωστού καθαρισμού και επεξεργασίας των δεδομένων του dataset.

7.3.1 LDA σε Tweets των Δεδομένων των Εκλογών

Τα ίδια βήματα και διαδικασία ακολουθήθηκε στα Tweets των 50 λογαριασμών που προέκυψαν από τα δεδομένα των εκλογών με την χρήση του Twitter API. Το σύννεφο λέξεων που προέκυψε δείχνει έντονα το θέμα των εκλογών να κυριαρχεί στα Tweets των χρηστών.

Εικόνα 23: Σύννεφο λέξεων στα Δεδομένα Εκλογών

Παρακάτω φαίνονται και οι 10 πιο σύνηθες λέξεις, οι οποίες παρατηρείται ότι σωστά ανήκουν και στο παραπάνω εξαγόμενο σύννεφο.



Εικόνα 24: Λίστα Διανυσμάτων Εγγράφων στα Δεδομένα Εκλογών

Στην περίπτωση του dataset των εκλογών επιλέχθηκαν μόνο 3 θέματα, διότι έχοντας λίγα Tweets διαθέσιμα παρατηρείται πως με περισσότερα θέματα εισάγεται θόρυβος στα δεδομένα.

Το αποτέλεσμα για το dataset των εκλογών είναι:

	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights
0	theworkplace_ct	2.1	everyon	2.1	vote	11.4
1	eversourcect	2.1	bhappi	2.0	electionday	3.7
2	workshop	2.1	poll	1.5	last	2.0

Εικόνα 25: Θέματα Δεδομένων Εκλογών

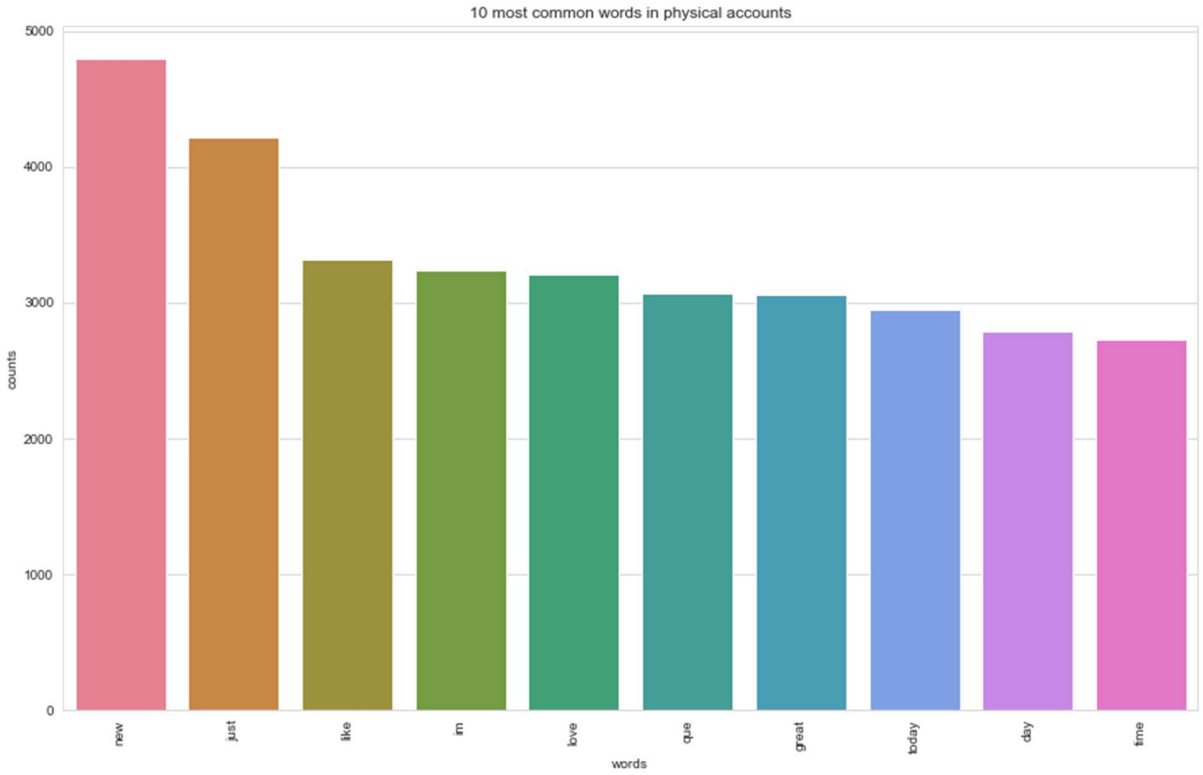
Διαπιστώνεται ότι τα topics σχεδόν όλα είχαν λέξεις σχετικά με τις εκλογές στην λίστα τους. Αυτό θα μπορούσε να υποδηλώσει ότι πρέπει να προστεθούν αυτές τις λέξεις στις λέξεις-κλειδιά (stopwords), καθώς δεν αναφέρουν κάτι που δεν είναι ήδη γνωστό, μιας και το σύνολο δεδομένων ήταν Tweets για την περίοδο των εκλογών.

Το συγκεκριμένο αποτέλεσμα μπορεί επίσης να προήλθε από το γεγονός ότι τα Tweets είναι πολύ σύντομα, καθώς αυτή η LDA (η οποία λειτουργεί πολύ καλά για μεγαλύτερα έγγραφα κειμένου), έχει έναν περιορισμό σε συντομότερα έγγραφα κειμένου όπως τα Tweets.

7.3.2 LDA σε Tweets του αρχικού Συνόλου Δεδομένων

Αρχικά, δημιουργήθηκε ένα σύννεφο λέξεων μόνο για τους πραγματικούς λογαριασμούς και σχετικά με τα Tweets που έχουν αναρτήσει:

Εικόνα 26: Σύννεφο λέξεων από Tweet πραγματικών λογαριασμών



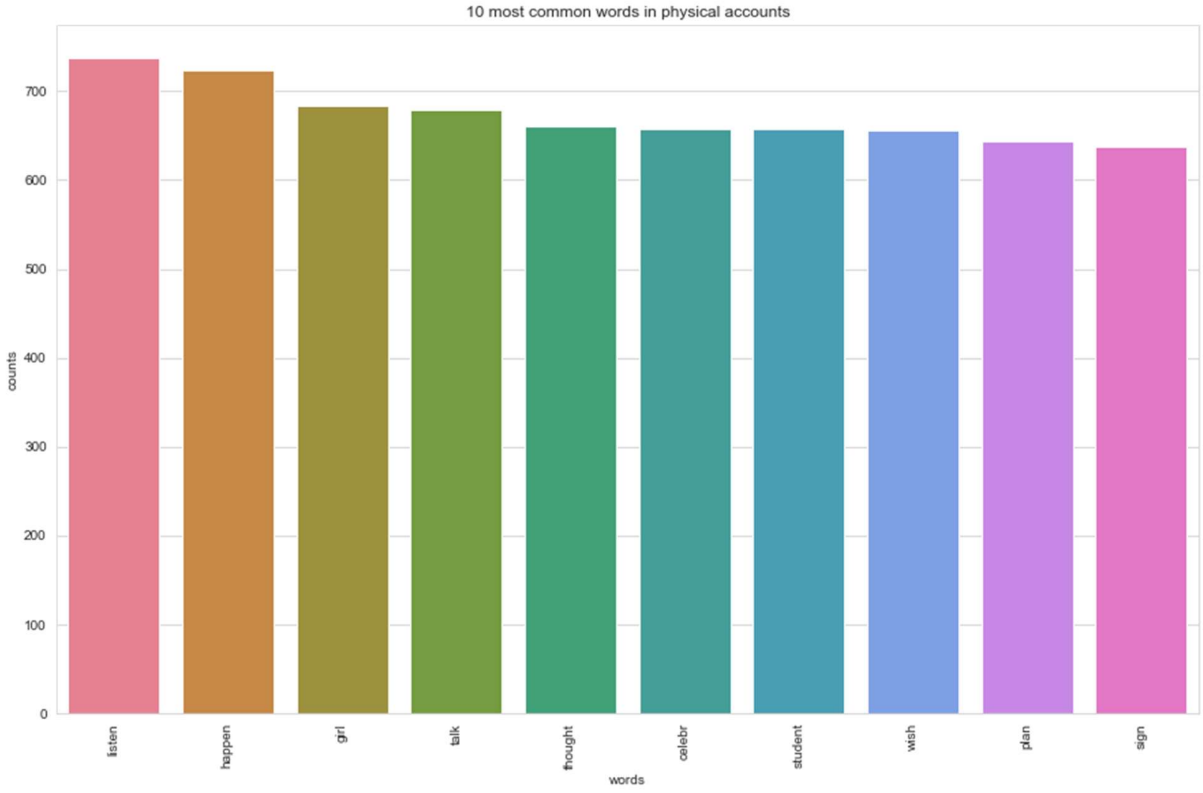
Εικόνα 27: Λίστα Διανυσμάτων Εγγράφων από Tweet πραγματικών λογαριασμών

Παρατηρήθηκε πως υπάρχουν πολλές λέξεις που δεν ανήκουν στο σύνολο των stopwords και δεν δίνουν κάποια επιπλέον σημασιολογική αξία στο γενικό θέμα ενός Tweet όπως η λέξεις: ‘thank’, ‘now’, ‘one’ .

Επεκτάθηκε λοιπόν το σύνολο των stopwords με τέτοιες κοινές εκφράσεις, καθώς επίσης, πραγματοποιήθηκε περαιτέρω αναζήτηση πολύ συχνών λέξεων. Αυτό επιτεύχθηκε με το να χωριστεί το σύνολο των Tweets data σε λέξεις για τις οποίες υπολογίστηκε η συχνότητα τους. Αυτές με την μεγαλύτερη συχνότητα αφαιρέθηκαν.

Το νέο σύννεφο λέξεων που προκύπτει φαίνεται παρακάτω.

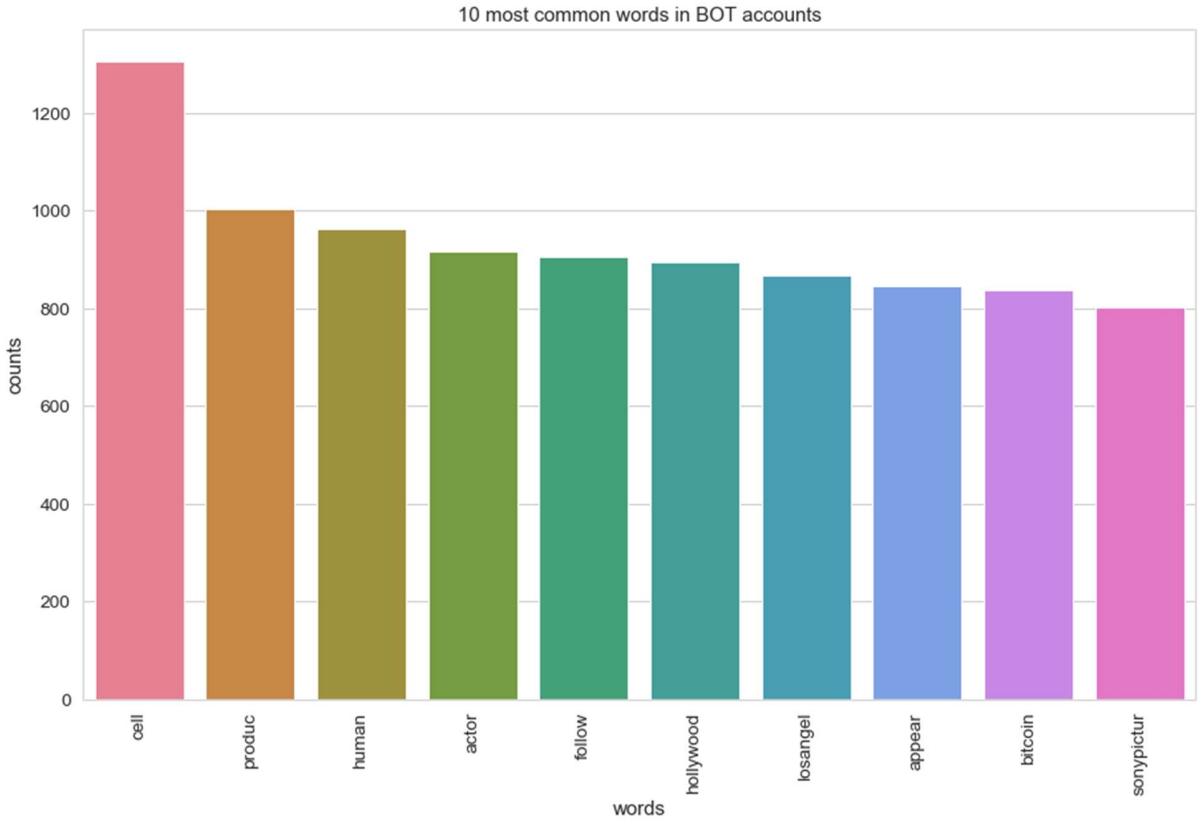
Εικόνα 28: Σύννεφο λέξεων από Tweet πραγματικών λογαριασμών έπειτα από επιπλέον καθαρισμό



Εικόνα 29: Λίστα Διανυσμάτων Εγγράφων από Tweet πραγματικών λογαριασμών έπειτα από επιπλέον καθαρισμό

Αντίστοιχα και μόνο για τους Bot λογαριασμούς:

Εικόνα 30: Σύννεφο λέξεων από Tweet Bot λογαριασμών έπειτα από επιπλέον καθαρισμό



Εικόνα 31: Λίστα Διανυσμάτων Εγγράφων από Tweet Bot λογαριασμών έπειτα από επιπλέον καθαρισμό

Από το σύννεφο λέξεων και την λίστα διανυσμάτων εγγράφων παρατηρούνται λέξεις που σχετίζονται με την παραγωγή ταινίας, πράγμα λογικό εφόσον είθισται οι λογαριασμοί Bot να χρησιμοποιούνται επί το πλείστον για λόγους διαφήμισης και προβολής.

7.3.3 Μοντέλο Εκπαίδευσης

Το μοντέλο εκπαιδεύτηκε με τις παρακάτω παραμέτρους για τον Αλγόριθμο LDA[6]:

1. **Αριθμός θεμάτων:** Με την παράμετρο αυτή προσδιορίζεται ο αριθμός των θεμάτων που επιθυμεί ο ερευνητής να εξάγει από τα δεδομένα κείμενα.
2. **Παράμετρος Άλφα:** Η τιμή της παραμέτρου αντιπροσωπεύει την πυκνότητα θέματος εγγράφου. Όσο υψηλότερη είναι η τιμή της παραμέτρου Άλφα, τότε τα κείμενα αποτελούνται από περισσότερα topics και οδηγούν σε πιο συγκεκριμένη κατανομή θεμάτων ανά έγγραφο.
3. **Beta:** Η τιμή της παραμέτρου αντιπροσωπεύει την πυκνότητα θέματος λέξης, που σημαίνει ότι όταν η τιμή betaείναι υψηλή, τα topics αποτελούνται από περισσότερες λέξεις και οδηγούν σε πιο συγκεκριμένη κατανομή λέξεων ανά θέμα)

Στην προσπάθεια εκτέλεσης του αλγορίθμου με διαφορετικό αριθμό θεμάτων προκύπτουν τα εξής αποτελέσματα για το σύνολο τον **Bots** λογαριασμών:

	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights
0	cell	1325.1	bitcoin	815.7	produc	955.6
1	appear	896.0	potu	486.0	actor	895.1
2	analysi	823.3	davidkwilliam	475.3	hollywood	876.3
3	gene	773.4	burst	460.6	losangel	852.2
4	human	707.0	mididittybot	352.6	am	851.4
5	wild	679.3	guerbuez	348.4	sonypictur	785.4
6	xcxb	672.2	woman	314.1	butterflycaught	784.5
7	pubm	641.2	thcircuit	311.9	follow	761.8
8	genom	620.3	xfxfxcxb	310.3	bong	615.7
9	cancer	595.9	review	306.0	fuck	615.6

Εικόνα 32: Θέματα Bot Λογαριασμών

Παρατηρείται ότι μόνο για την τελευταία στήλη θα μπορούσε να εξαχθεί η ιδέα του θέματος, το οποίο φαίνεται να είναι σχετικό με την παραγωγή ταινιών. Για την πρώτη κατηγορία επίσης

[6] https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

θα μπορούσε το θέμα να θεωρηθεί σχετικό με τον ‘άνθρωπο’ αφού κυριαρχούν λέξεις όπως «κύτταρο», «γένος» και συναφείς λέξεις.

Αντίστοιχα, για την εκτέλεση του αλγορίθμου μόνο για **φυσικούς** λογαριασμούς τα αποτελέσματα είναι τα εξής:

	Topic 0 words	Topic 0 weights	Topic 1 words	Topic 1 weights	Topic 2 words	Topic 2 weights
0	bihday	591.9	talk	1103.1	para	732.8
1	music	579.9	learn	1049.4	vocxcxaa	487.6
2	ticket	490.8	woman	1000.2	como	399.9
3	song	487.0	make	925.3	minha	348.2
4	photo	445.4	feel	890.1	sxcxb	332.1
5	wait	413.4	chang	879.1	todo	323.5
6	friend	404.5	vote	873.0	pessoa	299.2
7	tomorrow	403.8	friend	841.3	muito	280.9
8	perform	394.5	give	781.8	hoje	246.3
9	album	379.7	work	751.7	aqui	236.7

Εικόνα 33: Θέματα Φυσικών Λογαριασμών

Εδώ παρατηρείται ότι τα θέματα αποτελούνται από λέξεις με μεγαλύτερη συνάφεια, ωστόσο η εξαγωγή κάποιου θέματος είναι δύσκολη.

Όπως αναφέρθηκε και παραπάνω, αυτό το αποτέλεσμα μπορεί να προήλθε από το γεγονός ότι τα Tweets είναι πολύ σύντομα και αυτή η συγκεκριμένη μέθοδος, η LDA, δεν λειτουργεί καλά σε συντομότερα έγγραφα κειμένου όπως Tweets.

Τα μοντέλα θέματος παράγουν συχνά ανεξήγητα θέματα που γεμίζουν με θορυβώδεις λέξεις. Ο λόγος είναι ότι οι λέξεις στη μοντελοποίηση θεμάτων έχουν ίσα βάρη. Οι λέξεις υψηλής συχνότητας κυριαρχούν στις λίστες λέξεων του κορυφαίου θέματος, αλλά οι περισσότερες από αυτές είναι λέξεις χωρίς νόημα.

Το LDA από την φύση του επιτρέπει πιο ασαφείς δημιουργίες ομάδων. Αυτό παρέχει έναν τρόπο να εντοπίζονται παρόμοια αντικείμενα και διπλότυπα ή να ανακαλύπτονται συσχετίσεις που εξαπλώνονται σε μεγάλο μέρος κειμένων[7].

[7] https://towardsdatascience.com/lda-topic-modeling-an-explanation-e184c90aadcd

8. Συμπεράσματα

Έπειτα από την μελέτη και ανάλυση του μέσου Κοινωνικής Δικτύωσης Twitter, και με την βοήθεια της Μηχανικής Μάθησης και της Εξόρυξης Γνώσης από Δεδομένα, είναι δυνατόν να αξιολογηθεί και με πολύ μεγάλη ακρίβεια εάν ένας λογαριασμός Twitter αποτελεί Bot ή πραγματικό χρήστη. Η αξιολόγηση αυτή μπορεί να πραγματοποιηθεί έπειτα από την ανάλυση των δεδομένων του χρήστη, δηλαδή χαρακτηριστικά όπως ο αριθμός των ακολούθων (followers), ο αριθμός των φίλων, ο συνολικός αριθμός από Tweets που έχουν παραχθεί από τον χρήστη, ή το όνομα που χρησιμοποιεί, η περιγραφή και η θέση, αλλά και από την ανάλυση χαρακτηριστικών που προκύπτουν από τα κείμενα Tweet του κάθε χρήστη.

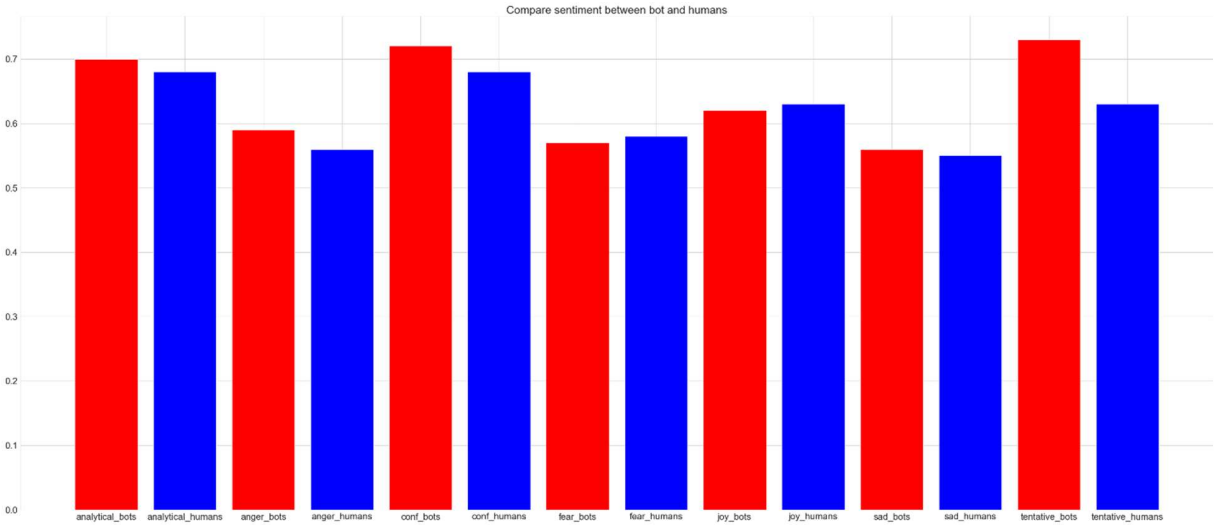
Στην συνέχεια, πραγματοποιήθηκε ανάλυση συναισθήματος στα κείμενα που εξάχθηκαν από το Twitter για τον εντοπισμό συναισθηματικών και γλωσσικών τόνων σε κάθε γραπτό κείμενο, καθώς φαίνεται από τα αποτελέσματα ο γλωσσικός τόνος που χρησιμοποιείται να καθορίζει σε μεγάλο βαθμό εάν ένας λογαριασμός είναι Bot ή όχι.

Επιπρόσθετα, με την χρήση της μοντελοποίησης θέματος ανιχνεύθηκαν τα θέματα που χρησιμοποιούνται από τους πραγματικούς ή μη χρήστες, πράγμα που επιβεβαιώνει την χρήση Bot λογαριασμών για λόγους διαφήμισης και προβολής. Συνοπτικά, η θεματολογία των πραγματικών χρηστών αποτελείται από λέξεις της καθημερινότητας όπως ευχαριστώ, αγάπη, ημέρα, ενώ η θεματολογία των Bot λογαριασμών κυρίως πραγματεύεται ταινίες, ηθοποιούς ή και διάφορα πρόσωπα της δημοσιότητας.

Εξαιρετική ιδέα για μελλοντική έρευνα θα μπορούσε να είναι η δημιουργία μιας εφαρμογής, η οποία θα μπορεί σε πραγματικό χρόνο να συλλέγει δεδομένα από το Twitter, να εφαρμόζει ανάλυση συναισθήματος και μοντελοποίηση θέματος, και τελικά να μπορεί να προβλέπει με βάση τον βέλτιστο αλγόριθμο της παρούσας εργασίας εάν ένας λογαριασμός είναι **Bot or Not**.

Η εφαρμογή θα μπορούσε επίσης να εμφανίζει σε πραγματικό χρόνο ποια είναι η τάση και τα θέματα που πραγματεύονται οι ψευδείς λογαριασμοί και οι πραγματικοί, ώστε να μπορεί εύκολα να εξαχθεί ένα συμπέρασμα. Εάν για παράδειγμα το θέμα που πραγματεύεται ένας Bot λογαριασμός αφορά είτε τις Αμερικανικές Εκλογές ή τον Κορωνοϊό, και το κυρίαρχο συναίσθημα είναι συγκεκριμένο, θα μπορούσε να εξαχθεί ένα συμπέρασμα για την περίπτωση που το Bot χρησιμοποιείται για την διαμόρφωση της κοινής γνώμης.

Παρατηρείται και στο παρακάτω διάγραμμα ότι σε όλα τα συναισθήματα που εξάγονται από τα Tweets των χρηστών, οι Bots λογαριασμοί σημειώνουν υψηλότερο σκορ, με εμφανή διαφορά στα συναισθήματα: ‘confident’, ‘anger’, ‘tentative’, ενώ κατά ένα πολύ μικρό ποσοστό έχουν χαμηλότερο σκορ στο συναίσθημα ‘joy’ .



Εικόνα 34: Σύγκριση Συναισθημάτων Μεταξύ Bot και Human

Αυτό μπορεί να επιβεβαιωθεί καθώς φαίνεται τα Bots να προσπαθούν να ακολουθήσουν το ανθρώπινο συναίσθημα στις απαντήσεις που δίνουν σε Tweets χρηστών δίνοντας έμφαση σε μια πιο έντονη εκδήλωση συναισθήματος.[8]

Το παραπάνω αποτέλεσμα εξηγείται, επίσης, από το γεγονός ότι ένα Bot επιδιώκει να εξαπατήσει τους ανθρώπινους χρήστες αποκλίνοντας το συναίσθημα που εκφράζεται στα Tweets τους από το βασικό συναίσθημα του αντίστοιχου γεγονότος. Πιο συγκεκριμένα, τα Bots τείνουν να είναι πιο θετικά σε γεγονότα που προκαλούν πώληση (όπως πολιτικά γεγονότα) και πιο αρνητικά κατά τη διάρκεια θετικών γεγονότων.[9]

[8] https://psyarxiv.com/cbv5j/

[9] https://nm.wu.ac.at/nm/strembeck/publications/complexis18-Bots.pdf

9. Μελλοντικές Επεκτάσεις - Χρήση εξαγόμενων Topics ως Features για Supervised Learning

Μια σειρά από άρθρα έχουν παρουσιαστεί τα τελευταία χρόνια σχετικά με εφαρμογή μεθόδων μηχανικής μάθησης και ειδικότερα βαθιάς μάθησης (deep learning) στην αναγνώριση συναισθήματος [18], [19], [20] και σε άλλες εφαρμογές [21], όπως και τεχνικών προσοχής (attention) στην εξαγωγή πληροφορίας για ταυτοποίηση χρηστών σε οπτικοακουστικά δεδομένα [2], [27]. Αυτές οι τεχνικές μπορούν να αποτελέσουν τη βάση για επέκταση των μεθόδων που χρησιμοποιήθηκαν στην παρούσα διπλωματική για αναγνώριση πραγματικών χρηστών ή bots σε κοινωνικά δίκτυα.

Μια υλοποίηση που θα μπορούσε να πραγματοποιηθεί ως επέκταση της παρούσας εργασίας είναι η χρήση των topics ως νέα χαρακτηριστικά των αρχικών δεδομένων, και έγκειται στην μετατροπή των topics σε feature vector όπως αναλύεται παρακάτω[10].

Θεωρείται ως δεδομένο ότι η μέθοδος LDA δηλώνει πως κάθε έγγραφο σε μία ομάδα είναι ένας συνδυασμός ενός καθορισμένου αριθμού θεμάτων. Ένα θέμα έχει την πιθανότητα να δημιουργήσει διάφορες λέξεις, όπου οι λέξεις είναι όλες οι λέξεις που παρατηρούνται στην ομάδα. Αυτά τα «κρυμμένα» θέματα εμφανίζονται στη συνέχεια με βάση την πιθανότητα συνύπαρξης λέξεων[11].

Έστω ένα παράδειγμα, όπου έχει εκπαιδευτεί ένα μοντέλο LDA με μεταβλητή επιλογής θεμάτων ίση με 3. Μετά την εκπαίδευση, λαμβάνονται όλα τα Tweets και να υπολογίζεται η κατανομή των θεμάτων για κάθε Tweet. Με άλλα λόγια, ορισμένα Tweets ενδέχεται να είναι 100% Θέμα 1, άλλα μπορεί να είναι 33% / 33% / 33% του Θέματος 1/2/3 και ούτω καθεξής. Αυτή η έξοδος είναι απλώς ένα διάνυσμα για κάθε Tweet και δείχνει την κατανομή των θεμάτων στο κείμενο. Η ιδέα εδώ είναι να ελεγχθεί εάν η κατανομή ανά Tweet κρυφών σημασιολογικών πληροφοριών θα μπορούσε να προβλέψει αν ο λογαριασμός είναι **Bot or not**.

Ο στόχος της συγκεκριμένης ιδέας είναι δηλαδή ο εξής:

[10] https://monkeylearn.com/blog/introduction-to-topic-modeling/
[11] https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28

Για την υλοποίηση χρησιμοποιείται η βιβλιοθήκη Gensims με την οποία τα Tweets μετατρέπονται σε unigrams και bigrams και εντάσσονται σε μια λίστα από την οποία θα εξαχθεί ο αριθμός συχνότητας λέξεων, για κάθε λέξη και για κάθε Tweet.

Με την χρήση του LDA αλγορίθμου θα εξαχθεί η κατανομή των λόγου χάριν 20 topics για κάθε Tweet. Αυτός ο 20-vector θα είναι ο feature vector για την εποπτευόμενη ταξινόμηση, με στόχο να αποφασίζει για τον χρήστη εάν είναι **Bot or not**.

Βιβλιογραφία

[1] [Abokhodair, Yoo, and McDonald 2015] Abokhodair, N.; Yoo, D.; and McDonald, D. W. 2015. Dissecting a social Botnet: Growth, content and influence in Twitter. In Proc. of the 18th ACM Conf. on Computer Supported Cooperative Work & Social Computing, 839–851. ACM.

[2] [Avrithis, Yannis and Tsapatsoulis, Nicolas and Kollias, Stefanos 2000] Avrithis, Yannis and Tsapatsoulis, Nicolas and Kollias, Stefanos; Broadcast news parsing using visual cues: A robust face detection approach, 2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast-Changing World of Multimedia (Cat. No. 00TH8532), v.3, 1469-1472.

[3] [Berger and Morgan 2015] Berger, J., and Morgan, J. 2015. The isis Twitter census: Defining and describing the population of isis supporters on Twitter. The Brookings Project on US Relations with the Islamic World 3:20.

[4] [Bessi and Ferrara 2016] Bessi, A., and Ferrara, E. 2016. Social Bots distort the 2016 us presidential election online discussion. First Monday 21(11).

[5] [Bessi et al. 2015] Bessi, A.; Coletto, M.; Davidescu, G. A.; Scala, A.; Caldarelli, G.; and Quattrociocchi, W. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. PLoS ONE 10(2): e0118093.

[6] [Beutel et al. 2013] Beutel, A.; Xu, W.; Guruswami, V.; Palow, C.; and Faloutsos, C. 2013. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In Prov. 22nd Intl. ACM Conf. World Wide Web (WWW), 119–130.

[7] [Blei David M., Ng Andrew, Jordan Michael. (2003)] "Latent Dirichlet Allocation"

[8] [Boshmaf et al. 2011] Boshmaf, Y.; Muslukhov, I.; Beznosov, K.; and Ripeanu, M. 2011. The socialBot network: when Bots socialize for fame and money. In Proc. 27th Annual Computer Security Applications Conf.

[9] [Chaoji, V., Ranu, S., Rastogi, R., and Bhatt, R] Recommendations to boost content spread in social networks., In WWW, 2012.

[10] [Chavoshi, Hamooni, and Mueen 2016] Chavoshi, N.; Hamooni, H.; and Mueen, A. 2016. Identifying correlated Bots in Twitter. In Social Informatics: 8th Intl. Conf., 14–21.

[11] [Emilio Ferrara, OnurVarol, Clayton Davis, Filippo Menczer and Alessandro Flammini] The Rise of Social Bots. X, X, Article XX (201X), 11 pages.

[12] [Ferrara et al. 2016a] Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016a. The rise of social Bots. Comm. ACM 59(7):96–104.

[13] [Haddi, E., Liu, X., Shi, Y. (2013)] The Role of Text Pre-Processing Sentiment Analysis, Information Technology and Quantitative Management, Procedia Computer Science 26 -32, Elsevier 2013.

[14] [Haustein et al. 2016] Haustein, S.; Bowman, T. D.; Holmberg, K.; Tsou, A.; Sugimoto, C. R.; and Larivi`ere, V. 2016. Tweets as impact indicators: Examining the implications of automated “Bot” accounts on Twitter. Journal of the Association for Information Science and Technology 67(1):232–238.

[15] [H. Kopka and P. W. Daly] A Guide to L ATEX, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[16] [H.Kwak,C.Lee,H.Park,andS.Moon] What Is Twitter,a Social Network or a News Media? Proc. 19th Intl Conf. World Wide Web, pp. 591-600, 2010

[17] [I.-C.M. Dongwoo Kim, Y. Jo, and A. Oh] Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users, Proc. CHI Workshop Microblogging: What and How Can We Learn From It?, 2010.

[18] [Kollias, Dimitrios and Zafeiriou, Stefanos P 2018] Kollias, Dimitrios and Zafeiriou, Stefanos P; Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm, IEEE 2018 International Joint Conference on Neural Networks (IJCNN), 1-8.

[19] [Kollias, Dimitrios and Zafeiriou, Stefanos P 2020] Kollias, Dimitrios and Zafeiriou, Stefanos P; Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset, IEEE Transactions on Affective Computing, 2020.

[20] [Kollias, Dimitris and Marandianos, George and Raouzaiou, Amaryllis and Stafylopatis, Andreas-Georgios 2015] Kollias, Dimitris and Marandianos, George and Raouzaiou, Amaryllis and Stafylopatis, Andreas-Georgios; Interweaving deep learning and semantic techniques for emotion analysis in human-machine interaction, IEEE 2015 10th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), 1-6.

[21] [Kollias, Dimitrios and Tagaris, Athanasios and Stafylopatis, Andreas and Kollias, Stefanos and Tagaris, Georgios 2018] Kollias, Dimitrios and Tagaris, Athanasios and Stafylopatis, Andreas and Kollias, Stefanos and Tagaris, Georgios; Deep neural architectures for prediction in healthcare, Complex \& Intelligent Systems, v.4, n.2, 119-131.

[22] [Kotsiantis, S., Kanellopoulos, D., &Pintelas, P. (2006)] Data preprocessing for supervised learning. International Journal of Computer Science, 1, 111–117

[23] [Kyumin Lee, Jalal Mahmud, Jilin Chen, Michelle Zhou and Jeffrey Nichols] Who Will ReTweet This? Automatically Identifying and Engaging Strangers on Twitter to Spread Information. Harlow, England: AddisonWesley, 1999.

[24] [Lee, Eoff, and Caverlee 2011] Lee, K.; Eoff, B. D.; and Caverlee, J. 2011. Seven months with the devils: A long-term study of content polluters on Twitter. In Proc. 5th AAAI Intl. Conf. on Web and Social Media.

[25] [Lokot and Diakopoulos 2016] Lokot, T., and Diakopoulos, N. 2016. News Bots: Automating news and information dissemination on Twitter. Digital Journalism 4(6):682–699.

[26] [OnurVarol, Emilio Ferrara, Clayton A. Davis, Filippo Menczer, Alessandro Flammini] Human-Bot Interactions: Detection, Estimation, and Characterization

[27][Rapantzikos, Konstantinos and Tsapatsoulis, Nicolas and Avrithis, Yannis and Kollias, Stefanos 2007] Rapantzikos, Konstantinos and Tsapatsoulis, Nicolas and Avrithis, Yannis and Kollias, Stefanos; Bottom-up spatiotemporal visual attention model for video analysis, IET Image Processing, v.1, n.2, 237-248.

[28] [Ratkiewicz et al. 2011] Ratkiewicz, J.; Conover, M.; Meiss, M.; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Detecting and tracking political abuse in Social Media. In 5th Int Conf on Weblogs & Soc Med, 297–304.

[29] [Savage, Monroy-Hernandez, and Höllerer 2016] Savage, S.; Monroy-Hernandez, A.; and Höllerer, T. 2016. Botivist: Calling volunteers to action using online Bots. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, 813–822. ACM.

[30] [Subrahmanian et al. 2016] Subrahmanian,V.; Azaria,A.; Durst,S.; Kagan,V.; Galstyan,A.; Lerman,K.; Zhu,L.; Ferrara,E.; Flammini, A.; Menczer, F.; et al. 2016. The DARPA Twitter BotChallenge. IEEE Computer 6(49):38–46.

[31] [Yang et al. 2014] Yang, Z.; Wilson, C.; Wang, X.; Gao, T.; Zhao, B. Y.; and Dai, Y. 2014. Uncovering social network sybils in the wild. ACM Trans. Knowledge Discovery from Data 8(1):2.