Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Σύγκριση Μεθόδων Ανίχνευσης Bots στο Twitter με χρήση Θεωρίας Δικτύων και Μεγάλων Γλωσσικών Μοντέλων

Διπλωματικη Εργασια

της

ΜΑΣΑΛΗ ΑΘΗΝΑΣ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

Σύγκριση Μεθόδων Ανίχνευσης Bots στο Twitter με χρήση Θεωρίας Δικτύων και Μεγάλων Γλωσσικών Μοντέλων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΜΑΣΑΛΗ ΑΘΗΝΑΣ

Επιβλέπων: Συμεών Παπαβασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 17η Φεβρουαρίου 2025.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Συμεών Παπαβασιλείου	Ιωάννα Ρουσσάκη	Ελένη Στάη
Καθηγητής Ε.Μ.Π.	Αναπληρώτρια Καθηγήτρια Ε.Μ.Π.	Επίκουρη Καθηγήτρια Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2025

Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Επικοινωνιων, Ηλεκτρονικης και Συστηματων Πληροφορικης

ΔΗΛΩΣΗ ΜΗ ΛΟΓΟΚΛΟΠΗΣ ΚΑΙ ΑΝΑΛΗΨΗΣ ΠΡΟΣΩΠΙΚΗΣ ΕΥΘΥΝΗΣ

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ενυπογράφως ότι είμαι αποκλειστικός συγγραφέας της παρούσας Πτυχιακής Εργασίας, για την ολοκλήρωση της οποίας κάθε βοήθεια είναι πλήρως αναγνωρισμένη και αναφέρεται λεπτομερώς στην εργασία αυτή. Έχω αναφέρει πλήρως και με σαφείς αναφορές, όλες τις πηγές χρήσης δεδομένων, απόψεων, θέσεων και προτάσεων, ιδεών και λεκτικών αναφορών, είτε κατά κυριολεξία είτε βάσει επιστημονικής παράφρασης. Αναλαμβάνω την προσωπική και ατομική ευθύνη ότι σε περίπτωση αποτυχίας στην υλοποίηση των ανωτέρω δηλωθέντων στοιχείων, είμαι υπόλογος έναντι λογοκλοπής, γεγονός που σημαίνει αποτυχία στην Πτυχιακή μου Εργασία και κατά συνέπεια αποτυχία απόκτησης του Τίτλου Σπουδών, πέραν των λοιπών συνεπειών του νόμου περί πνευματικών δικαιωμάτων. Δηλώνω, συνεπώς, ότι αυτή η Πτυχιακή Εργασία προετοιμάστηκε και ολοκληρώθηκε από εμένα προσωπικά και αποκλειστικά και ότι, αναλαμβάνω πλήρως όλες τις συνέπειες του νόμου στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής άλλης πνευματικής ιδιοκτησίας.

(Υπογραφή)

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

Περίληψη

Η ανίχνευση αυτοματοποιημένων λογαριασμών (bots) στο κοινωνικό δίκτυο του Twitter αποτελεί ένα κρίσιμο πρόβλημα, καθώς τα bots χρησιμοποιούνται συχνά για παραπληροφόρηση, προπαγάνδα και κακόβουλες δραστηριότητες. Η παρούσα διπλωματική εργασία εξετάζει το πρόβλημα της ανίχνευσής τους μέσω διαφόρων τεχνικών και μεθόδων Μηχανικής Μάθησης. Συγκεκριμένα, για τον σκοπό αυτό επιστρατεύονται κλασικά μοντέλα μηχανικής μάθησης όπως το Logistic Regression (Λογιστική Παλινδρόμηση) και το Random Forest (Τυχαίο Δάσος), πιο σύγχρονα νευρωνικά δίκτυα όπως τα Graph Convolutional Networks (Συνελικτικό Δίκτυο Γράφων) και τα Graph Attention Networks (Δίκτυο Γράφων Προσοχής), και τα, επίκαιρα, Μεγάλα Γλωσσικά Μοντέλα. Γίνεται ανάπτυξη των παραπάνω μοντέλων ώστε να εφαρμοστούν στο πρόβλημα της αναγνώρισης των κακόβουλων χρηστών, και ακολούθως, γίνεται η σύγκρισή τους με βάση την απόδοσή τους στο υπό εξέταση πρόβλημα. Αρχικά δίνεται το θεωρητικό υπόβαθρο που θεωρείται απαραίτητο για την κατανόηση των εννοιών και μεθόδων που θα παρουσιαστούν παρακάτω. Έπειτα, γίνεται μια αναλυτική επισκόπηση της σχετικής με το θέμα ερευνητικής βιβλιογραφίας, και παρουσιάζεται ενδελεχώς το σύνολο δεδομένων που αξιοποιήθηκε, καθώς και η προεπεξεργασία που υπέστη πριν χρησιμοποιηθεί για την εκπαίδευση και αξιολόγηση των μοντέλων. Στη συνέχεια, περιγράφεται η μεθοδολογία που ακολουθήθηκε, και ακολούθως αναλύεται διεξοδικά η υλοποίηση των μοντέλων και το πειραματικό σκέλος, όπου γίνεται παρουσίαση και ερμηνεία των αποτελεσμάτων της απόδοσης κάθε μοντέλου, και σύγκριση μεταξύ τους. Τέλος, διατυπώνονται τα συνολικά συμπεράσματα και προτείνονται πιθανές μελλοντικές επεκτάσεις της έρευνας. Η αξιολόγηση έγινε με τις μετρικές accuracy, precision, recall, f1 score και καλύτερη απόδοση στην ανίχνευση bots σημείωσε το μοντέλο Graph Convolutional Network, ενώ την, απροσδόκητα, χαμηλότερη σημείωσαν τα Μεγάλα Γλωσσικά Μοντέλα.

Λέξεις Κλειδιά

Μηχανική Μάθηση, Τεχνητή Νοημοσύνη, Γραφικά Συνελικτικά Δίκτυα, Twitter, Μέσα Κοινωνικής Δικτύωσης, Αυτοματοποιημένοι Λογαριασμοί, Ανίχνευση Bots, Νευρωνικά Δίκτυα Γράφων, Μεγάλα Γλωσσικά Μοντέλα

Abstract

The subject of this diploma thesis is the detection of automated accounts, known as bots, on the social media platform of Twitter using a variety of methods. More specifically, for the classification of Twitter accounts as bots, the following methods are utilized; classical machine models like Logistic Regression and Random Forest, the state-of-the-art graph-based models of Graph Convolutional Networks and Graph Attention Networks, and the, lately, much-discussed Large Language Models. The above models are developed in order to be applied to the problem of malicious bot user identification, and then, they are compared based on their performance in the problem under consideration. First, the theoretical background is given which is considered necessary for understanding the concepts and methods presented below. Then, a detailed review of the relevant research literature is given, and the dataset utilized is thoroughly presented, along with the preprocessing it underwent before being used for model training and evaluation. The methodology of the process is then described, followed by a detailed discussion of the implementation of the models and the experimental part, where the results of the performance of each model are presented and interpreted, and compared with each other. Finally, overall conclusions are drawn and possible future improvements are suggested. The evaluation was performed with the metrics accuracy, precision, recall, f1 score and the best performance in detecting bots was achieved by the Graph Convolutional Network model, while the, unexpectedly, lowest was achieved by the Large Language Models. .

Keywords

Machine Learning, Artificial Intelligence, Graph Convolutional Networks, Twitter, Social Media, Bots, Bot Detection, Graph Neural Networks, Large Language Models

Ευχαριστίες

Η παρούσα διπλωματική εργασία σηματοδοτεί την ολοκλήρωση της φοίτησής μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσοβίου Πολυτεχνείου, και ταυτόχρονα το τέλος ενός ιδιαίτερα σημαντικού κεφαλαίου της ζωής μου. Με την δυνατότητα που μου δίνεται εδώ, θα ήθελα να εκφράσω τις ευχαριστίες και την ευγνωμοσύνη μου στους ανθρώπους που στήριξαν, ο καθένας με τον δικό του τρόπο, την επίτευξη αυτού μου του στόχου.

Πρωτίστως, θα ήθελα να ευχαριστήσω θερμά τον κύριο Συμεών Παπαβασιλείου, ο οποίος υπήρξε ο επιβλέπων καθηγητής μου σε αυτή την διπλωματική εργασία και μου έδωσε την ευκαιρία να ασχοληθώ με το συγκεκριμένα επιστημονικό πεδίο και ένα τόσο ενδιαφέρον θέμα. Εν συνεχεία, θα ήθελα να ευχαριστήσω τον ερευνητή Δρ. Κωνσταντίνο Τσιτσεκλή για την διαρκή καθοδήγηση και τις συμβουλές του, που υπήρξαν καθοριστικά καθ' όλη τη διάρκεια εκπόνησης αυτής της διπλωματικής.

Ύστερα, θα ήθελα να ευχαριστήσω τις φίλες και τους φίλους μου, για όλες τις όμορφες αναμνήσεις που μου πρόσφεραν απλόχερα αυτά τα χρόνια και για την στήριξή τους σε κάθε μου δυσκολία. Ιδιαίτερα, θα ήθελα να ευχαριστήσω θερμά τα άτομα με τα οποία διασταυρώθηκαν οι δρόμοι μας στα πλαίσια αυτής της σχολής και που με αφορμή την κοινή μας ακαδημαϊκή πορεία, αναπτύξαμε σημαντικές για μένα φιλίες. Η βοήθεια και η συμπαράστασή τους υπήρξε ανεκτίμητη.

Τέλος, θέλω να ευχαριστήσω από τα βάθη της καρδιάς μου τους γονείς μου και την αδερφή μου, αλλά και τα άτομα που αν και δεν μας ενώνουν δεσμοί αίματος, θεωρώ οικογένειά μου. Η αγάπη τους, η υποστήριξη τους και η άνευ όρων παρουσία τους στις δύσκολες στιγμές μου, συνετέλεσαν καταλυτικά στην διαμόρφωσή μου στο άτομο που είμαι σήμερα, και για αυτό η ευγνωμοσύνη μου είναι ανυπολόγιστη.

Αθήνα, Νοέμβριος 2024

Αθηνά Μασαλή

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Μέσα Κοινωνικής Δικτύωσης

Τα μέσα κοινωνικής δικτύωσης (social media) αποτελούν αναπόσπαστο κομμάτι της σύγχρονης ψηφιακής εποχής, προσφέροντας πρωτόγνωρες δυνατότητες επικοινωνίας, αλληλεπίδρασης και δημιουργίας περιεχομένου. Μέσω αυτών, οι χρήστες μπορούν να μοιράζονται ιδέες, ενδιαφέροντα, σκέψεις, βιώματα και οπτικοακουστικό υλικό, δημιουργώντας εικονικές κοινότητες χωρίς γεωγραφικούς περιορισμούς. Η διάδοσή τους συνδέεται με την ανάπτυξη των τεχνολογιών του διαδικτύου που επέτρεψαν τη μετάβαση από την απλή κατανάλωση πληροφοριών στη διαδραστική συμμετοχή.

Οι πλατφόρμες κοινωνικής δικτύωσης δεν περιορίζονται μόνο στην επικοινωνία και την κοινωνική δικτύωση, αλλά έχουν εξελιχθεί σε εργαλεία με πολλαπλές λειτουργίες. Το Facebook, για παράδειγμα, συνδυάζει την κοινωνική αλληλεπίδραση με λειτουργίες όπως η δημιουργία εκδηλώσεων, οι διαδικτυακές κοινότητες και η ψυχαγωγία. Το Instagram και το TikTok επικεντρώνονται κυρίως στην οπτική ψυχαγωγία και τη δημιουργικότητα, με ιδιαίτερη έμφαση στη γρήγορη κατανάλωση περιεχομένου. Το YouTube είναι σημείο αναφοράς για την αναζήτηση και δημοσίευση οπτικοακουστικού υλικού, ενώ το LinkedIn αποτελεί βασικό εργαλείο επαγγελματικής δικτύωσης, βοηθώντας τους χρήστες να προβάλλουν την επαγγελματική και ακαδημαϊκή τους δραστηριότητα. Τέλος, το Twitter, που πλέον έχει μετονομαστεί σε X, ενθαρρύνει την ελεύθερη έκφραση, τη δημόσια συζήτηση και την άμεση ενημέρωση.

Στη σημερινή εποχή, η δυναμική των μέσων κοινωνικής δικτύωσης εκτείνεται πέρα από την απλή επικοινωνία και επηρεάζει την καθημερινότητα, διαδραματίζοντας σημαντικό ρόλο σε διάφορες πτυχές της ανθρώπινης δραστηριότητας, όπως την ενημέρωση, την ψυχαγωγία, την προώθηση προϊόντων και υπηρεσιών, τις επιχειρηματικές δραστηριότητες, ακόμα και τις πολιτικές εξελίξεις. Παράλληλα, η ανάπτυξή τους συνοδεύεται από προκλήσεις και κινδύνους, όπως την διαχείριση των προσωπικών δεδομένων και την προστασία της ιδιωτικότητας, την εξάπλωση της παραπληροφόρησης και τις αρνητικές επιπτώσεις (ψυχικές, κοινωνικές κλπ.) της υπερβολικής χρήσης. Ωστόσο, η δυνατότητα ανάλυσης του περιεχομένου που παράγεται στις πλατφόρμες αυτές έχει ανοίξει νέους δρόμους για την επιστημονική έρευνα, καθιστώντας τα social media εργαλεία μελέτης της ανθρώπινης συμπεριφοράς, της ψυχικής υγείας και των κοινωνικών τάσεων. [1] [2]

Με τις τεχνολογίες των μέσων κοινωνικής δικτύωσης να εξελίσσονται διαρκώς, η κατανόηση της λειτουργίας, των επιπτώσεων και των δυνατοτήτων τους αποτελεί ζήτημα κρίσιμης

σημασίας για την ανάλυση των σύγχρονων κοινωνικών δομών και σχέσεων.

Σχήμα 1.1: *Μέσα Κοινωνικής Δικτύωσης (AI Generated Image)* [3]

1.2 Twitter (X)

Το Twitter, το οποίο πρόσφατα μετονομάστηκε σε "X", είναι μία από τις πιο δημοφιλείς πλατφόρμες κοινωνικής δικτύωσης παγκοσμίως, με εκατοντάδες εκατομμύρια χρήστες. Ιδρύθηκε το 2006 από τους Jack Dorsey, Noah Glass, Biz Stone και Evan Williams, με έδρα το Σαν Φρανσίσκο, Καλιφόρνια. Η πλατφόρμα σχεδιάστηκε αρχικά για να παρέχει στους χρήστες έναν γρήγορο και εύκολο τρόπο να μοιράζονται σύντομες ενημερώσεις για τις δραστηριότητές τους, κάτι που αντικατοπτρίζεται στο όριο χαρακτήρων των δημοσιεύσεων, γνωστές και ως "tweets", το οποίο αρχικά ήταν 140 και αυξήθηκε σε 280 το 2017.

Το Twitter ξεχωρίζει για τη δυνατότητα που προσφέρει στους χρήστες να δημοσιεύουν tweets τα οποία μπορεί να περιλαμβάνουν κείμενο, εικόνες, βίντεο ή συνδέσμους. Η αναδημοσίευση περιεχομένου (retweet), οι απαντήσεις (replies) και οι αναφορές σε άλλους χρήστες (mentions) μέσω του χαρακτήρα "@" ενισχύουν την αλληλεπίδραση μεταξύ των χρηστών, ενώ τα hashtags "#" λειτουργούν ως εργαλεία κατηγοριοποίησης και εντοπισμού περιεχομένου. Επιπλέον, οι χρήστες μπορούν να διαχειρίζονται την εμπειρία τους μέσω λιστών (lists) οργανώνοντας τους λογαριασμούς που ακολουθούν σε κατηγορίες, να ανταλλάσσουν άμεσα μηνύματα (Direct Messages) και να ενημερώνονται για τις παγκόσμιες τάσεις (trends), δηλαδή τα θέματα που είναι δημοφιλή την εκάστοτε στιγμή. Κάθε χρήστης ακολουθεί (Following/Friends) τους χρήστες για την δραστηριότητα των οποίων θέλει να παραμένει ενήμερος, ενώ αντίστοιχα ακολουθείται (Followers) από τα άτομα που θέλουν να ενημερώνονται για τη δική του δραστηριότητα.

Το Twitter εξελίχθηκε σε σημείο αναφοράς του δημόσιου διαλόγου, της πληροφόρησης και της ψυχαγωγίας. Είναι ιδανικό για τη διάδοση ειδήσεων σε πραγματικό χρόνο, την ανταλλαγή απόψεων και τη συμμετοχή σε συζητήσεις χωρίς γεωγραφικούς περιορισμούς. Οι δυνατότητες αναζήτησης και παρακολούθησης τάσεων έχουν αναδείξει την πλατφόρμα σε σημαντικό εργαλείο για δημοσιογράφους, ερευνητές και επαγγελματίες του χώρου της επικοινωνίας [4].

Το 2022, η εξαγορά του Twitter από τον Elon Musk σηματοδότησε μια νέα εποχή για την πλατφόρμα, με αλλαγές στις πολιτικές της και τη στρατηγική της. Παρά τις αντιδράσεις που προκλήθηκαν, το Twitter διατηρεί τη θέση του ως μια πλατφόρμα που ενσωματώνει τις φωνές εκατομμυρίων χρηστών παγκοσμίως, συνδυάζοντας την κοινωνική δικτύωση με την καινοτομία στην επικοινωνία.

1.3 Ορισμός Προβλήματος

Σε συνέχεια των παραπάνω, η ευρεία χρήση του Twitter ως μέσου κοινωνικής δικτύωσης και ενημέρωσης, και συνεπώς ως τόπο διάδοσης πληροφοριών, έχει φέρει στην επιφάνεια μια κρίσιμη πρόκληση: την αυξανόμενη παρουσία αυτοματοποιημένων λογαριασμών, γνωστούς ως bots, στην πλατφόρμα. Τα bots λειτουργούν ανεξάρτητα και έχουν τη δυνατότητα να μιμούνται ανθρώπινες συμπεριφορές, δημοσιεύοντας και αναπαράγοντας περιεχόμενο με σκοπό να επηρεάσουν την κοινή γνώμη, να διαδώσουν ψευδείς ειδήσεις και να χειραγωγήσουν συζητήσεις.

Τα τελευταία χρόνια έχει παρατηρηθεί ραγδαία αύξηση της δράσης τους και υπολογίζεται πως, πλέον, ένα σημαντικό ποσοστό των συνολικών λογαριασμών του Twitter είναι bots [5]. Αυτοί οι λογαριασμοί έχουν συσχετιστεί με την εξάπλωση της παραπληροφόρησης, του εχθρικού λόγου και της προπαγάνδας, προκαλώντας έντονες ανησυχίες σχετικά με την αξιοπιστία της πλατφόρμας. Τα bots δεν περιορίζονται μόνο στη διάδοση ψευδών πληροφοριών, αλλά χρησιμοποιούνται και για τη διαμόρφωση συγκεκριμένου επενδυτικού και επιχειρηματικού κλίματος, την προώθηση προϊόντων και την εξυπηρέτηση πολιτικών σκοπών.

Η ύπαρξη και δράση των bots ενισχύει την πολυπλοκότητα της διάκρισης των αληθινών και των ψευδών πληροφοριών στα κοινωνικά δίκτυα. Καθώς οι αυτοματοποιημένοι αυτοί λογαριασμοί εξελίσσονται συνεχώς, αποκτούν τη δυνατότητα να μιμούνται ακόμη πιο πειστικά τη συμπεριφορά πραγματικών χρηστών, καθιστώντας την ανίχνευσή τους ιδιαίτερα δύσκολη. Αυτή η εξέλιξη υπονομεύει την εμπιστοσύνη των χρηστών στην πλατφόρμα και ενισχύει την ανάγκη για αποτελεσματικά εργαλεία εντοπισμού και διαχείρισης των bots.

Σχήμα 1.2: *Social Media Bots (AI Generated Image)* [3]

Η κατανόηση του προβλήματος των bots είναι ζωτικής σημασίας για τη διασφάλιση της ακεραιότητας του περιεχομένου που δημοσιεύεται στο Twitter. Στο πλαίσιο αυτό, η ανίχνευση και η εξάλειψη αυτών των ψευδών λογαριασμών αποτελεί προτεραιότητα, όχι μόνο για το Twitter αλλά και για τα υπόλοιπα μέσα κοινωνικής δικτύωσης, ώστε να διασφαλιστεί ένα περιβάλλον με μεγαλύτερη διαφάνεια και αξιοπιστία.

1.4 Αντικείμενο και Συνεισφορά της Διπλωματικής Εργασίας

Όπως έχει ήδη αναφερθεί στην προηγούμενη ενότητα, είναι επιτακτική η ανάγκη περιορισμού της παρουσίας των bots στα μέσα κοινωνικής δικτύωσης, οπότε η έρευνα που επικεντρώνεται στην αναγνώρισή τους είναι το πρώτο βήμα προς αυτή την κατεύθυνση. Σε αυτό το μήκος κύματος κινείται και η παρούσα διπλωματική εργασία που εστιάζει στο μείζον ζήτημα της ανίχνευσης των αυτοματοποιημένων λογαριασμών στην πλατφόρμα του Twitter. Αντικείμενο της έρευνας αποτελεί η ανάπτυξη και διερεύνηση διαφόρων εργαλείων και τεχνικών για την ταυτοποίηση αυτοματοποιημένων λογαριασμών, με σκοπό την κατανόηση της δράσης τους και τη συμβολή στην αποτελεσματική αντιμετώπισή τους.

Η εργασία στοχεύει στη συστηματική διερεύνηση των χαρακτηριστικών και των μοτίβων συμπεριφοράς των bots, καθώς και στην ανάπτυξη μοντέλων που χρησιμοποιούν σύγχρονες τεχνικές ανάλυσης δεδομένων και μηχανικής μάθησης. Συγκεκριμένα, το ζήτημα της αναγνώρισης bots προσεγγίζεται με τρεις μεθόδους: με κλασικά μοντέλα μηχανικής μάθησης, με μοντέλα που βασίζονται στην γραφηματική (ή δικτυακή) απεικόνιση των χαρακτηριστικών, και με Μεγάλα Γλωσσικά Μοντέλα.

Στα πλαίσια της εργασίας προτείνεται η εξαγωγή και αξιοποίηση πολύπλευρων χαρακτηριστικών των χρηστών, από αριθμητικά στατιστικά των προφίλ και κειμενικά στοιχεία τους, όπως η περιγραφή και το περιεχόμενο των δημοσιεύσεων, μέχρι σύνθετα γραφικά χαρακτηριστικά που προκύπτουν από την μορφή του δικτύου του Twitter που εξετάζεται. Στην διαδικασία επεξεργασίας των δεδομένων δίνεται έμφαση, ώστε η τελική τους μορφή να συνεισφέρει στην βέλτιστη απόδοση των μοντέλων που αναπτύσσονται και αξιολογούνται.

Η συνεισφορά της διπλωματικής εργασίας εντοπίζεται σε δύο κύριους άξονες. Πρώτον, εξετάζει εκτενώς την υπάρχουσα σχετική βιβλιογραφία και σε συμμόρφωση με αυτή, προτείνει καινοτόμες προσεγγίσεις για την ανίχνευση και την κατηγοριοποίηση χρηστών σε humans (πραγματικούς χρήστες) και bots (αυτοματοποιημένους χρήστες), συνδυάζοντας υπάρχουσες μεθόδους με νέες τεχνικές που αξιοποιούν τη δυναμική της σύγχρονης τεχνολογίας. Από τα κλασικά μοντέλα Logistic Regression, Random Forest και τα πιο σύγχρονα πολύπλοκα συνελικτικά μοντέλα Graph Convolutional Network, Graph Attention Network, μέχρι τα πολυσυζητημένα Μεγάλα Γλωσσικά Μοντέλα, εξετάζεται η ευελιξία και η απόδοση τους στο πρόβλημα εντοπισμού αυτοματοποιημένων λογαριασμών. Δεύτερον, παρέχει μια αναλυτική σύγκριση της αποτελεσματικότητας των προτεινόμενων προσεγγίσεων, συμβάλλοντας στην συζήτηση που εγείρεται γύρω από την αποδοτικότητα και την καταλληλότητα των σχετικών συστημάτων στην αντιμετώπιση του υπό εξέταση ζητήματος.

1.5 Δομή Διπλωματικής

Η Διπλωματική Εργασία οργανώνεται σε επτά επί μέρους κεφάλαια. Στο Κεφάλαιο 1 δίνεται μια σύντομη εισαγωγή στο θέμα και περιγράφεται το πρόβλημα που πραγματεύεται η εν λόγω διπλωματική. Στο Κεφάλαιο 2 παρέχεται το απαραίτητο, για την κατανόηση του κειμένου, θεωρητικό υπόβαθρο, τόσο γενικών εννοιών της Θεωρίας Γράφων, της Μηχανικής Μάθησης και των Μεγάλων Γλωσσικών Μοντέλων, όσο και συγκεκριμένων τεχνολογιών και μεθόδων, τα οποία χρησιμοποιούνται κατά την εκπόνηση της παρούσας εργασίας. Στο Κεφάλαιο 3 γίνεται μια ανασκόπηση της πρόσφατης βιβλιογραφίας και αναφέρονται οι τεχνικές και τα εργαλεία που έχουν αναπτυχθεί από την διεθνή επιστημονική κοινότητα προκειμένου να αντιμετωπιστεί το πρόβλημα της ανίχνευσης των αυτοματοποιημένων λογαριασμών στα κοινωνικά δίκτυα. Στο Κεφάλαιο 4 παρουσιάζεται αναλυτικά το σύνολο δεδομένων που αξιοποιήθηκε κατά την ερευνητική διαδικασία, καθώς και η προεπεξεργασία στην οποία υποβλήθηκαν τα υποσύνολα δεδομένων του που επιλέχθηκαν για την εκπαίδευση και την αξιολόγηση των μοντέλων. Στο Κεφάλαιο 5 περιγράφεται η μεθοδολογία που ακολουθήθηκε, δίνοντας έμφαση στην κατασκευή του δικτύου από το σύνολο δεδομένων, την εξαγωγή και επιλογή χαρακτηριστικών από αυτό, και, τέλος, τον καθορισμό των μοντέλων με τα οποία έγινε η υλοποίηση. Στο Κεφάλαιο 6 παρουσιάζεται η υλοποίηση των μοντέλων που αναπτύχθηκαν για την ανίχνευση αυτοματοποιημένων λογαριασμών, και καταγράφονται τα αποτελέσματά τους, ενώ στη συνέχεια δίνεται η ερμηνεία τους και η μεταξύ τους σύγκριση. Τέλος, το Κεφάλαιο 7 αποτελεί τον επίλογο της Διπλωματικής Εργασίας, όπου διατυπώνονται τα βασικά συμπεράσματα που προέκυψαν, και προτείνονται πιθανές μελλοντικές επεκτάσεις της έρευνας που πραγματοποιήθηκε στα πλαίσια αυτής της εργασίας.

Μέρος I

Θεωρητικό Μέρος

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

Σε αυτό το κεφάλαιο παρέχεται το θεωρητικό υπόβαθρο που κρίνεται απαραίτητο για την κατανόηση των εννοιών και των μεθόδων που χρησιμοποιούνται κατά την εκπόνηση της παρούσας διπλωματικής εργασίας. Συγκεκριμένα, δίνεται μια εισαγωγή στη Θεωρία Γράφων και στα Σύνθετα Δίκτυα, επεξηγούνται βασικές έννοιες της Μηχανικής Μάθησης και παρουσιάζονται τα μοντέλα που θα αξιοποιηθούν στα πλαίσια της ερευνητικής διαδικασίας, και, τέλος, γίνεται εκτεταμένη αναφορά στα Μεγάλα Γλωσσικά Μοντέλα και στην Ανίχνευση Αυτοματοποιημένων Λογαριασμών.

2.1 Θεωρία Γράφων και Σύνθετα Δίκτυα

Σε αυτή την ενότητα γίνεται μια σύντομη εισαγωγή στην Θεωρία Γράφων και τα Σύνθετα Δίκτυα παρουσιάζοντας κάποιες βασικές έννοιες που θα μας απασχολήσουν κατά την ανάλυση αυτής της εργασίας. [6] [7]

2.1.1 Βασικοί Ορισμοί

Γράφος

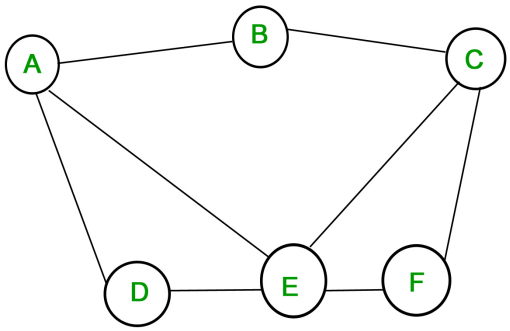
Ένας **γράφος (ή γράφημα)** είναι μια αφηρημένη αναπαράσταση ενός συνόλου στοιχείων, όπου μερικά ζεύγη στοιχείων συνδέονται μεταξύ τους με δεσμούς.

Πιο αυστηρά, ένας γράφος G ορίζεται ως ένα διατεταγμένο ζεύγος:

$$G = \{V, E\}$$

που περιέχει ένα σύνολο από κορυφές (ή κόμβους) V και ένα σύνολο ακμών E. Μια ακμή $(u, v) \in E$ αναπαριστά μια διασύνδεση μεταξύ των κόμβων $u, v \in V$.

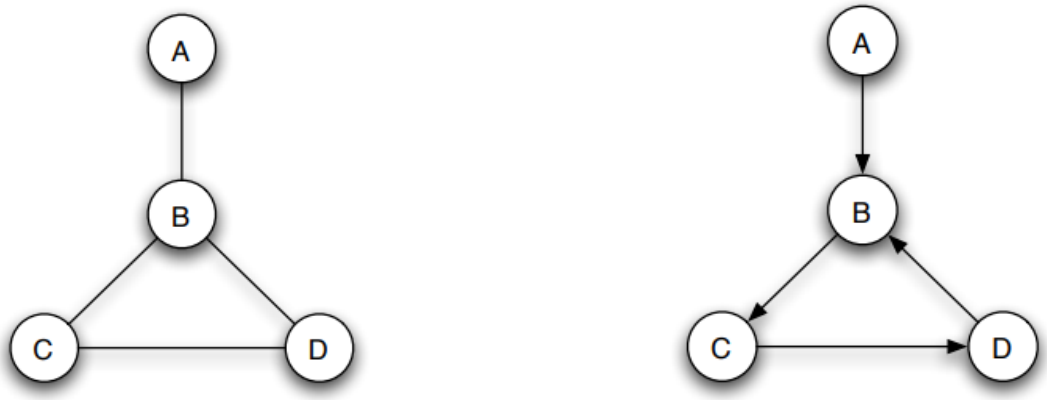
Οι γράφοι μπορούν εναλλακτικά να αναφερθούν και σαν δίκτυα. Χρησιμοποιούνται για τη μοντελοποίηση ποικίλων σύγχρονων σημαντικών προβλημάτων. Συγκεκριμένα, βρίσκουν εφαρμογή σε τηλεπικοινωνιακά, οδικά, ηλεκτρικά και κοινωνικά δίκτυα, διαδρομές, δρομολόγηση, ανάθεση πόρων κλπ [6].



Σχήμα 2.1: *Παράδειγμα ενός γράφου με 6 κόμβους και 8 ακμές*
[8]

Κατευθυνόμενος και μη-κατευθυνόμενος Γράφος

Ένα γράφημα μπορεί να είναι **κατευθυνόμενο** ή **μη-κατευθυνόμενο**. Στα κατευθυνόμενα γραφήματα κάθε ακμή έχει έναν προσανατολισμό, δηλαδή έχει έναν κόμβο-πηγή και έναν κόμβο-προορισμό. Στα μη-κατευθυνόμενα γραφήματα οι ακμές δεν έχουν προσανατολισμό.



(α') *Μη κατευθυνόμενος γράφος με 4 κόμβους.* (β') *Κατευθυνόμενος γράφος με 4 κόμβους.*

Σχήμα 2.2: *Δύο γράφοι: (α') ένας μη κατευθυνόμενος γράφος, και (β') ένας κατευθυνόμενος γράφος.*
[7]

Γειτονιά και γειτονικοί κόμβοι

Σε ένα μη κατευθυνόμενο γράφημα G, δύο κόμβοι u και v λέγονται γειτονικοί αν συνδέονται μεταξύ τους με μια ακμή $(u, v) \in E$. Η γειτονιά N_u ενός κόμβου u αποτελείται από το σύνολο των γειτονικών του κόμβων. Δηλαδή, ισχύει:

$$N_u = \{v \in V : (u, v) \in E\}$$

Από την άλλη, σε ένα κατευθυνόμενο γράφημα G, κάθε κόμβος συνδέεται με δύο ειδών ακμές, αυτές που ξεκινούν από αυτόν και εκείνες που τερματίζουν σε αυτόν. Συνεπώς,

ορίζεται η έσω γειτονιά N_u^- ενός κόμβου u, ως το σύνολο των κόμβων με ακμή προς τον u, δηλαδή:

$$N_u^- = \{v \in V : (v, u) \in E\}$$

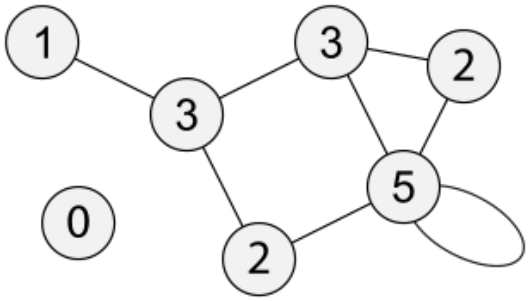
Αντίστοιχα, ορίζεται η έξω γειτονιά N_u^+ ενός κόμβου u, ως το σύνολο των κόμβων με ακμή που ξεκινά από τον u, δηλαδή:

$$N_u^+ = \{v \in V : (u, v) \in E\}$$

Συνεκτικότητα και Συνιστώσες

Μια συνιστώσα ενός απλού μη κατευθυνόμενου γραφήματος είναι ένα υποσύνολο κορυφών V', όπου $V' \subseteq V$, για το οποίο ισχύει ότι για κάθε δύο κορυφές του V' υπάρχει μονοπάτι που τις συνδέει.

Ένα γράφημα λέγεται συνεκτικό αν για κάθε δύο κορυφές του υπάρχει μονοπάτι που να τις συνδέει. Δηλαδή, αν αποτελείται από μία και μοναδική συνιστώσα, η οποία είναι το ίδιο το γράφημα. [9]



Σχήμα 2.3: *Ένας μη κατευθυνόμενος γράφος που έχει δύο συνιστώσες, και άρα δεν είναι συνεκτικός.*

[9]

2.1.2 Σύνθετα Δίκτυα

Εισαγωγή στα Σύνθετα Δίκτυα

Στα πλαίσια της θεωρίας δικτύων, ένα **σύνθετο** δίκτυο είναι ένα γράφημα με μη τετριμμένα τοπολογικά χαρακτηριστικά, με μοτίβα σύνδεσης μεταξύ των στοιχείων του που δεν είναι απολύτως τακτικά (regular), αλλά και ούτε απολύτως τυχαία (random). Αυτά τα χαρακτηριστικά δεν εμφανίζονται σε απλά δίκτυα, αλλά συχνά παρατηρούνται σε δίκτυα που αναπαριστούν πραγματικά συστήματα.

Συνεπώς, τα σύνθετα δίκτυα αποτελούν ένα ισχυρό εργαλείο για τη μελέτη διαφόρων κοινωνικών, βιολογικών και τεχνολογικών συστημάτων. Μεταξύ αυτών, σημαντική θέση κατέχουν οι κοινωνικοί γράφοι, δηλαδή τα γραφήματα που αποτυπώνουν τις σχέσεις και τις συνδέσεις μεταξύ ατόμων σε κοινωνικά δίκτυα, όπως το Twitter, το Facebook, το LinkedIn

κλπ [10]. Σε ένα κοινωνικό γράφημα, τα άτομα απεικονίζονται ως κόμβοι και οι σχέσεις μεταξύ τους αναπαρίστανται ως γραμμές που συνδέουν αυτούς τους κόμβους.

Στην επιστημονική βιβλιογραφία, οι όροι γράφημα και δίκτυο χρησιμοποιούνται συχνά εναλλακτικά. Ωστόσο, υπάρχει μια λεπτή διάκριση: το δίκτυο συνδέεται περισσότερο με τα πραγματικά συστήματα και τις σχέσεις που περιγράφουν, ενώ το γράφημα αναφέρεται στη μαθηματική αναπαράσταση αυτών των συστημάτων. Οι κοινωνικοί γράφοι ενσωματώνουν δεδομένα για τις οντότητες που απαρτίζουν τους κόμβους, προσδίδοντάς τους πολυδιάστατη φύση και καθιστώντας τους κατάλληλους για σύνθετες αναλύσεις [6].

Η μελέτη των σύνθετων δικτύων, και ιδιαίτερα των κοινωνικών γραφημάτων, προσφέρει μοναδικές ευκαιρίες για την κατανόηση της δομής και της δυναμικής των κοινωνικών αλληλεπιδράσεων, ενισχύοντας τη γνώση μας σε τομείς όπως η κοινωνιολογία, η πληροφορική και η ανάλυση δεδομένων [11].

Αναπαράσταση και οπτικοποίηση γράφων

Η **αναπαράσταση** και **οπτικοποίηση** των δεδομένων ενός γραφήματος αποτελούν κρίσιμα βήματα για την κατανόηση και ανάλυση των δικτύων. Υπάρχουν δύο κύριες προσεγγίσεις για την αναπαράσταση γραφημάτων: ο πίνακας γειτνίασης και η λίστα γειτνίασης.

Ο πίνακας γειτνίασης είναι ένας τετραγωνικός πίνακας διαστάσεων n x n (όπου n ο αριθμός των κορυφών), που καταγράφει με τιμή 1 την ύπαρξη ακμής μεταξύ δύο κορυφών και με 0 την απουσία της. Αυτή η αναπαράσταση είναι αποδοτική για τον άμεσο έλεγχο της ύπαρξης ακμών, αλλά μπορεί να καταναλώνει πολύ χώρο μνήμης, ειδικά σε αραιά γραφήματα. Αντίθετα, η λίστα γειτνίασης συνδέει κάθε κορυφή με τις γειτονικές της μέσω συνδεδεμένων λιστών, παρέχοντας έναν πιο οικονομικό τρόπο αποθήκευσης, ιδανικό για αραιά γραφήματα.

Η οπτικοποίηση των γραφημάτων γίνεται συνήθως με διαγράμματα όπου οι κορυφές αναπαρίστανται ως σημεία και οι ακμές ως γραμμές που συνδέουν αυτά τα σημεία. Αυτό διευκολύνει την ανάλυση των σχέσεων και τη χαρτογράφηση πολύπλοκων δικτύων, όπως κοινωνικά, βιολογικά ή τεχνολογικά συστήματα [6].

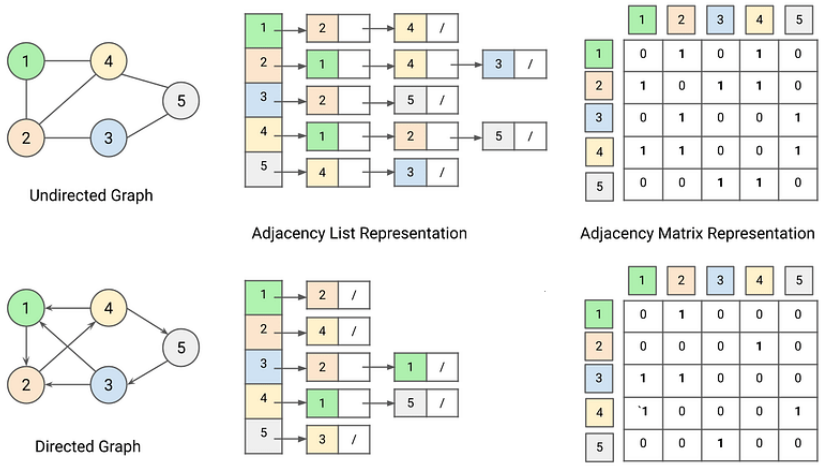
2.1.3 Μετρικές Σύνθετων Δικτύων

Σε αυτή την υποενότητα θα παρουσιαστούν κάποιες μετρικές που χρησιμοποιούνται συχνά για τη μελέτη της δομής και της συμπεριφοράς των σύνθετων δικτύων.

Κατανομή Βαθμών

Ο βαθμός ενός κόμβου (node degree) σε ένα δίκτυο είναι ο αριθμός των άμεσων συνδέσεων (ακμών) που έχει με άλλους κόμβους. Δηλαδή υπολογίζεται ως ο αριθμός των γειτονικών του κόμβων.

Στα κατευθυνόμενα γραφήματα υπάρχει ο εισερχόμενος βαθμός κόμβου (in-degree), που είναι ο αριθμός των εισερχόμενων ακμών από γειτονικούς του κόμβους, και ο εξερχόμενος βαθμός κόμβου (out-degree), που ορίζεται ως ο αριθμός των εξερχόμενων ακμών προς άλλους γειτονικούς του κόμβους.



Σχήμα 2.4: *Υπολογισμός λίστας και πίνακα γειτνίασης σε μη κατευθυνόμενο και κατευθυνόμενο γράφο.*

[12]

Η κατανομή βαθμών $P(k)$ ενός δικτύου (degree distribution), συνεπώς, ορίζεται σαν την κατανομή πιθανοτήτων που περιγράφει την πιθανότητα ενός κόμβου του δικτύου να έχει έναν συγκεκριμένο βαθμό. Έτσι, αν υπάρχουν n κόμβοι συνολικά σε ένα δίκτυο και n_k από αυτούς έχουν βαθμό k, τότε η κατανομή βαθμών ορίζεται ως το κλάσμα:

$$P(k) = \frac{n_k}{n}$$

Η κατανομή βαθμών σαν μετρική μπορεί να δώσει σημαντικές πληροφορίες για τη δομή ενός γράφου.

Συντελεστής Συσταδοποίησης

Στα σύνθετα δίκτυα οι κόμβοι έχουν την τάση να σχηματίζουν συμπλέγματα (clusters) με έντονη εσωτερική διασύνδεση μεταξύ των κόμβων τους. Η μετρική που χρησιμοποιείται για την καταγραφή αυτής της συμπεριφοράς ονομάζεται συντελεστής συσταδοποίησης (clustering coefficient), και μπορεί να υπολογιστεί είτε σε όλη την έκταση του δικτύου είτε μεμονωμένα για έναν κόμβο.

Καθολικός Συντελεστής Συσταδοποίησης

Ο καθολικός συντελεστής συσταδοποίησης (global clustering coefficient) C είναι ένας τρόπος να ποσοτικοποιείται η συσταδοποίηση που παρουσιάζει ένα δίκτυο στο σύνολό του. Ο υπολογισμός του βασίζεται στις τριπλέτες κόμβων. Μια τριπλέτα ορίζεται σαν ένα σύνολο τριών κόμβων που συνδέονται μεταξύ τους με δύο (ανοιχτή τριπλέτα) ή τρεις (κλειστή τριπλέτα) ακμές.

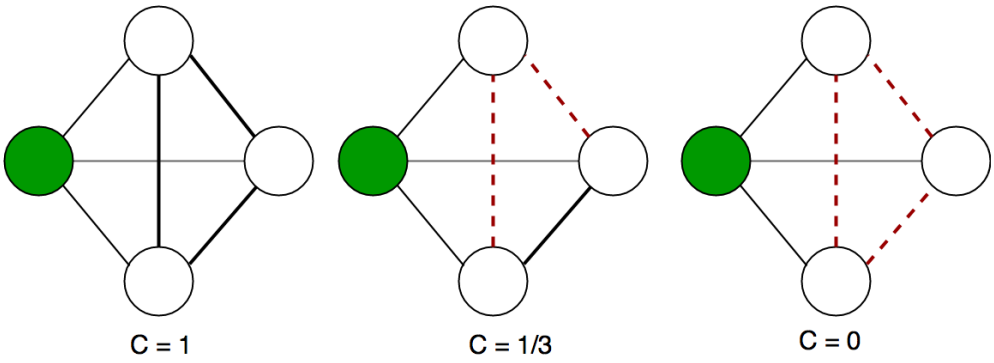
$$C = \frac{\text{αριθμός κλειστών τριπλετών}}{\text{αριθμός τριπλετών (ανοιχτών και κλειστών)}}$$

Τοπικός Συντελεστής Συσταδοποίησης

Ο τοπικός συντελεστής συσταδοποίησης (local clustering coefficient) αφορά έναν μεμονωμένο κόμβο και μετράει πόσο στενά συνδεδεμένοι είναι οι γείτονές του. Υπολογίζεται από τη διαίρεση του αριθμού των συνδέσεων (ακμών) στη γειτονιά του κόμβου, προς το συνολικό αριθμό συνδέσεων που θα μπορούσαν να έχουν. Πιο απλά, αυτός ο συντελεστής δείχνει πόσο κοντά είναι οι γείτονες ενός κόμβου στο να σχηματίσουν μια κλίκα, δηλαδή ένα υποσύνολο πλήρους γράφου.

Σε έναν κατευθυνόμενο γράφο, η ακμή e_{ij} είναι διαφορετική από την ακμή e_{ji}, και συνεπώς για κάθε γειτονιά N_i υπάρχουν $k_i(k_i - 1)$ πιθανές συνδέσεις - ακμές μεταξύ των κόμβων της (k_i είναι ο αριθμός των γειτόνων ενός κόμβου). Άρα, ο τοπικός συντελεστής συσταδοποίησης σε έναν κατευθυνόμενο γράφο ορίζεται ως:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$



Σχήμα 2.5: *Παράδειγμα του υπολογισμού του τοπικού συντελεστή συσταδοποίησης του πράσινου κόμβου σε τρία διαφορετικά σενάρια.*

[13]

Κεντρικότητες

Οι κεντρικότητες κόμβων είναι μετρικές που αναδεικνύουν την σημαντικότητα κάθε κόμβου μέσα σε ένα δίκτυο. Σε αυτή την υποενότητα θα παρουσιαστούν κάποιες από τις πιο δημοφιλής κεντρικότητες που χρησιμοποιούνται και στο πειραματικό κομμάτι αυτής της διπλωματικής. [6]

Κεντρικότητα Βαθμού

Η απλούστερη κεντρικότητα είναι η κεντρικότητα βαθμού (degree centrality), που ισούται με τον βαθμό του εξεταζόμενου κόμβου. Οι κόμβοι με πολλούς γείτονες αξιολογούνται ως σημαντικότεροι, διότι έχουν περισσότερες διασυνδέσεις και, άρα, μεγαλύτερη επιρροή στη ροή της πληροφορίας μέσα στο δίκτυο.

Έστω A ο πίνακας γειτνίασης ενός δικτύου $G = (V, E)$ με n κόμβους. Τότε η κεντρικότητα βαθμού ενός κόμβου $i \in V$ ορίζεται ως:

$$C_D(i) = \sum_{j=1}^{n} a_{ij}$$

Κεντρικότητα Εγγύτητας

Η κεντρικότητα εγγύτητας (closeness centrality) εκτιμά πόσο κοντά βρίσκεται ένας κόμβος σε όλους τους άλλους κόμβους. Είναι ένας τρόπος να γίνει πρόβλεψη της ταχύτητας με την οποία ταξιδεύει η πληροφορία από έναν κόμβο προς το υπόλοιπο δίκτυο.

Υπολογίζεται διαιρώντας το συνολικό αριθμό των κόμβων μείον τον αρχικό κόμβο-πηγή με την απόσταση του αρχικού κόμβου-πηγή προς όλους τους άλλους κόμβους. Για αυτόν τον λόγο, η κεντρικότητα εγγύτητας ορίζεται μόνο εάν το δίκτυο είναι συνδεδεμένο, δηλαδή μόνο αν υπάρχει μονοπάτι μεταξύ οποιουδήποτε ζεύγους κόμβων.

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

Κεντρικότητα Διαμεσολάβησης

Η κεντρικότητα διαμεσολάβησης (betweenness centrality) είναι μια μετρική που αξιολογεί τη σημασία ενός κόμβου, λαμβάνοντας υπόψη τον αριθμό των συντομότερων διαδρομών ανάμεσα σε όλους τους κόμβους του δικτύου που περνούν μέσα από αυτόν.

Ένας κόμβος με υψηλή κεντρικότητα διαμεσολάβησης ελέγχει μεγάλο μέρος της πληροφορίας που ρέει στο δίκτυο, καθώς πολλές διαδρομές διέρχονται από αυτόν. Αντίθετα, ένας κόμβος με χαμηλή κεντρικότητα διαμεσολάβησης μπορεί εύκολα να παρακαμφθεί μέσω άλλων κόμβων.

Για έναν κόμβο $u \in V$, η κεντρικότητα διαμεσολάβησης $C_B(u)$ ορίζεται εξετάζοντας όλα τα ζευγάρια κόμβων $i,j \in V$ και διαιρώντας τον αριθμό των συντομότερων μονοπατιών $\sigma_{ij}(u)$ μεταξύ των i και j που περνούν από τον u, με τον συνολικό αριθμό των συντομότερων μονοπατιών σ_{ij} μεταξύ των i και j.

$$C_B(u) = \sum_{u \in V, i \neq j} \frac{\sigma_{ij}(u)}{\sigma_{ij}}$$

Κεντρικότητα Ιδιοδιανύσματος

Η κεντρικότητα ιδιοδιανύσματος (eigenvector centrality) είναι μια μετρική που εκτιμά την σημαντικότητα ενός κόμβου σε ένα δίκτυο βασιζόμενη τόσο στον αριθμό των συνδέσεών του, όσο και στη σημασία των κόμβων με τους οποίους συνδέεται. Συγκεκριμένα, ένας κόμβος έχει υψηλή κεντρικότητα ιδιοδιανύσματος αν συνδέεται με άλλους κόμβους που είναι επίσης σημαντικοί. Υπολογίζεται ως το κύριο ιδιοδιάνυσμα του πίνακα γειτνίασης του γράφου, όπου η τιμή κάθε κόμβου στο ιδιοδιάνυσμα αντιπροσωπεύει την κεντρικότητά του.

Μαθηματικά, αν A είναι ο πίνακας γειτνίασης ενός γράφου με n κόμβους και x είναι το διάνυσμα κεντρικότητας, τότε η κεντρικότητα ιδιοδιανύσματος ορίζεται ως η λύση της εξίσωσης:

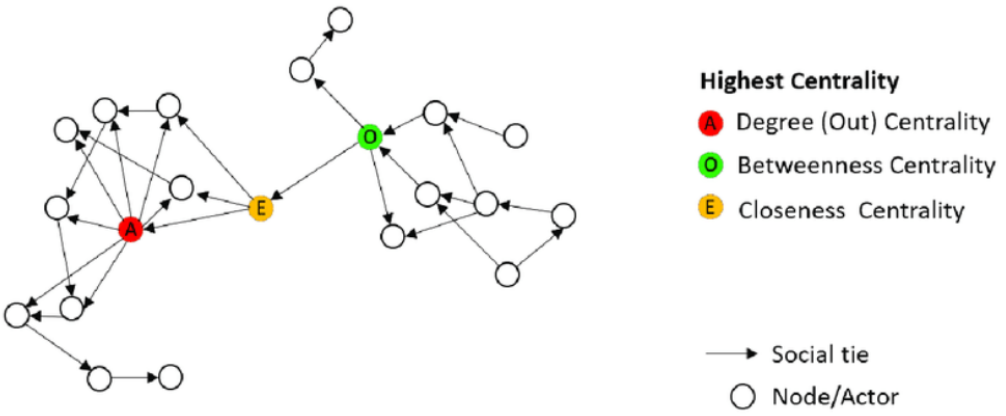
$$A\mathbf{x} = \lambda\mathbf{x}$$

όπου λ είναι η μέγιστη ιδιοτιμή του πίνακα A.

Η τιμή της κεντρικότητας για κάθε κόμβο i δίνεται από την εξής σχέση:

$$x_i = \frac{1}{\lambda} \sum_j A_{ij} x_j$$

δηλαδή, η κεντρικότητα ενός κόμβου i είναι ανάλογη με το άθροισμα των κεντρικοτήτων των γειτονικών του κόμβων. Το διάνυσμα $\mathbf{x}$ που προκύπτει από αυτήν την ιδιοτιμή είναι το κύριο ιδιοδιάνυσμα του πίνακα A και αντιπροσωπεύει τη σχετική σπουδαιότητα κάθε κόμβου στο δίκτυο.



Σχήμα 2.6: *Απεικόνιση των κόμβων με την υψηλότερη κεντρικότητα βαθμού, εγγύτητας και διαμεσολάβησης σε έναν κατευθυνόμενο γράφο.*
[14]

Αλγόριθμος PageRank

Ο αλγόριθμος PageRank είναι μια μέθοδος κατάταξης κόμβων σε έναν γράφο, που αρχικά αναπτύχθηκε από τους ιδρυτές της Google για την ταξινόμηση ιστοσελίδων με βάση τη σημαντικότητά τους. Ο αλγόριθμος βασίζεται στην ιδέα ότι ένας κόμβος είναι σημαντικός εάν συνδέεται με άλλους σημαντικούς κόμβους, και υπολογίζει μια επαναληπτική κατανομή πιθανοτήτων που εκφράζει τη σχετική σημασία κάθε κόμβου στο δίκτυο. [15]

2.2 Μηχανική Μάθηση

Αυτή η διπλωματική εργασία εξετάζει πώς η Τεχνητή Νοημοσύνη, και ειδικότερα οι μέθοδοι μηχανικής μάθησης και ανάλυσης δικτύων, μπορούν να εφαρμοστούν στην ανίχνευση αυτοματοποιημένων λογαριασμών σε κοινωνικά δίκτυα όπως το Twitter. Συνεπώς, σε αυτή την ενότητα παρέχεται στον αναγνώστη το απαραίτητο θεωρητικό υπόβαθρο γενικών αρχών της Μηχανικής Μάθησης, καθώς και συγκεκριμένων τεχνολογιών και μεθόδων ανάλυσης γράφων οι οποίες χρησιμοποιούνται στο ερευνητικό μέρος.

2.2.1 Εισαγωγή στην Τεχνητή Νοημοσύνη

Η **Τεχνητή Νοημοσύνη (Artificial Intelligence - AI)** αποτελεί ένα πεδίο της επιστήμης των υπολογιστών που στοχεύει στην ανάπτυξη συστημάτων ικανών να μιμούνται ανθρώπινες γνωστικές λειτουργίες, όπως η λογική, η μάθηση, η λήψη αποφάσεων και η επίλυση προβλημάτων. Στον πυρήνα της, η Τεχνητή Νοημοσύνη επιδιώκει να αυτοματοποιήσει τη νοημοσύνη, επιτρέποντας στα μηχανήματα να αναλύουν πολύπλοκα προβλήματα, να προσαρμόζονται σε νέα δεδομένα και να λαμβάνουν αποφάσεις χωρίς την ανάγκη για ρητές οδηγίες από τον άνθρωπο. Αυτή η ικανότητα βασίζεται σε μαθηματικά μοντέλα, στατιστικές τεχνικές και υπολογιστικούς αλγορίθμους που επιτρέπουν στα συστήματα να αντλούν γνώση από δεδομένα [16].

Μέσα στην πάροδο λίγων δεκαετιών, η Τεχνητή Νοημοσύνη έχει εξελιχθεί ραγδαία από συστήματα βασισμένα σε κανόνες έως τη σύγχρονη εποχή της Μηχανικής Μάθησης (Machine Learning - ML) που βασίζεται στα δεδομένα [16]. Αυτή η μετατόπιση οφείλεται στις εξελίξεις στην υπολογιστική ισχύ, στη διαθεσιμότητα μεγάλων συνόλων δεδομένων και στις αλγοριθμικές καινοτομίες. Η μηχανική μάθηση, ως βασικό υποσύνολο της Τεχνητής Νοημοσύνης, δίνει έμφαση στη δημιουργία μοντέλων που μαθαίνουν από δεδομένα αντί να βασίζονται σε κανόνες. Αυτή η προσέγγιση επιτρέπει στα συστήματα να γενικεύουν από παραδείγματα εκπαίδευσης και να κάνουν προβλέψεις ή ταξινομήσεις σε νέα δεδομένα. Ένα από τα πιο χρήσιμα και ενδιαφέροντα χαρακτηριστικά της Τεχνητής Νοημοσύνης είναι η ικανότητά της να αναλύει και να μοντελοποιεί σύνθετα συστήματα, όπως τα κοινωνικά δίκτυα, που παρουσιάζουν μεγάλης κλίμακας, δυναμική και διασυνδεδεμένη συμπεριφορά. Τα κοινωνικά δίκτυα παρουσιάζουν μοναδικές προκλήσεις ανάλυσης λόγω του μεγέθους τους, της ποικιλομορφίας της συμπεριφοράς των χρηστών και της δυναμικής τους εξέλιξης [7]. Αυτά τα δίκτυα συχνά περιγράφονται καλύτερα μέσω αναπαραστάσεων με τη μορφή γράφων, όπου οι κόμβοι αναπαριστούν οντότητες (π.χ. χρήστες ή λογαριασμούς) και οι ακμές αναπαριστούν αλληλεπιδράσεις (π.χ. likes, follows ή retweets).

Στο πλαίσιο των κοινωνικών δικτύων, η ανίχνευση κακόβουλων δραστηριοτήτων, όπως οι λογαριασμοί bot, αποτελεί ένα επιτακτικό ζήτημα [5]. Τα bots είναι αυτοματοποιημένες οντότητες προγραμματισμένες να μιμούνται ανθρώπινη συμπεριφορά και συχνά επηρεάζουν τη δημόσια συζήτηση, ενισχύουν τη διάδοση παραπληροφόρησης ή χειραγωγούν τις κοινωνικές δυναμικές για πολιτικούς ή εμπορικούς σκοπούς. Η ανίχνευση τέτοιων λογαριασμών είναι ένα πρόβλημα δυαδικής ταξινόμησης, όπου τα συστήματα Τεχνητής Νοημοσύνης πρέπει να αποφασίσουν εάν ένας λογαριασμός διαχειρίζεται από άνθρωπο ή από αυτοματοποιημένο σύστημα.

Η Τεχνητή Νοημοσύνη προσφέρει μια ευρεία γκάμα προσεγγίσεων για την αντιμετώπιση αυτού του προβλήματος. Παραδοσιακές μέθοδοι μηχανικής μάθησης, όπως η λογιστική παλινδρόμηση (Logistic Regression) και τα τυχαία δάση (Random Forest), παρέχουν αποδοτικούς τρόπους ταξινόμησης λογαριασμών βασισμένους σε προκαθορισμένα χαρακτηριστικά, όπως η συχνότητα δημοσιεύσεων ή τα στατιστικά αλληλεπίδρασης. Ωστόσο, τα κοινωνικά δίκτυα είναι εγγενώς γραφηματικά, γεγονός που καθιστά τις μεθόδους μάθησης βασισμένες σε γραφήματα, όπως τα Συνελικτικά Δίκτυα Γράφων (Graph Convolutional Networks - GCNs) και τα Δίκτυα Γράφων Προσοχής (Graph Attention Networks - GATs), ι-

διαίτερα αποτελεσματικές. Αυτές οι μέθοδοι αξιοποιούν τη δομική και σχεσιακή πληροφορία που κωδικοποιείται στα γραφήματα δικτύου, ανιχνεύοντας μοτίβα που μπορεί να διαφεύγουν από τα παραδοσιακά μοντέλα μηχανικής μάθησης.

2.2.2 Εισαγωγή στη Μηχανική Μάθηση

Η **Μηχανική Μάθηση (Machine Learning - ML)** είναι ένας κλάδος της Τεχνητής Νοημοσύνης που εστιάζει στη δημιουργία μεθόδων και αλγορίθμων που επιτρέπουν στους υπολογιστές να «μαθαίνουν» από δεδομένα, χωρίς να απαιτείται ο προγραμματισμός αυστηρών κανόνων. Αντί να ακολουθούν προσχεδιασμένες οδηγίες, τα συστήματα μηχανικής μάθησης αναλύουν μεγάλα σύνολα δεδομένων, αναγνωρίζουν μοτίβα (pattern recognition) και χρησιμοποιούν αυτή τη γνώση για να λαμβάνουν αποφάσεις ή να κάνουν προβλέψεις. Αυτή η προσέγγιση καθιστά τη μηχανική μάθηση εξαιρετικά ευέλικτη και ισχυρή για την αντιμετώπιση πολύπλοκων προβλημάτων [17].

Η θεμελιώδης αρχή της μηχανικής μάθησης έγκειται στη δυνατότητα των αλγορίθμων να προσαρμόζονται συνεχώς, βελτιώνοντας την απόδοσή τους καθώς εκτίθενται σε νέα δεδομένα. Οι μέθοδοι ML αντλούν τη φιλοσοφία τους από την αναγνώριση προτύπων και τη θεωρία της υπολογιστικής μάθησης, συνδυάζοντας την στατιστική και την επιστήμη δεδομένων για να επιτύχουν την μάθηση. Στη βάση τους βρίσκονται μαθηματικά και στατιστικά μοντέλα, τα οποία επιτρέπουν στους αλγορίθμους να γενικεύουν από δεδομένα εκπαίδευσης και να λειτουργούν με ακρίβεια ακόμη και σε πρωτόγνωρα σύνολα δεδομένων.

Οι αλγόριθμοι μηχανικής μάθησης έχουν ενσωματωθεί σε ένα ευρύ φάσμα εφαρμογών που αγγίζουν την καθημερινότητα. Από τις μηχανές αναζήτησης, όπως το Google, και τα συστήματα συστάσεων στις πλατφόρμες streaming, έως την ανάλυση ιατρικών δεδομένων και τη διάγνωση ασθενειών, η μηχανική μάθηση τα τελευταία χρόνια πρωταγωνιστεί με τις καινοτομίες που έχει φέρει σε διάφορους κλάδους. Ειδικότερα, η ικανότητα των αλγορίθμων να αυτοβελτιώνονται με την πάροδο του χρόνου, τους καθιστά ιδιαίτερα χρήσιμους για την ανάλυση δεδομένων μεγάλης κλίμακας, όπου τα παραδοσιακά προγραμματιστικά μέσα αποτυγχάνουν.

Η μηχανική μάθηση κατηγοριοποιείται σε τέσσερις βασικές προσεγγίσεις, ανάλογα με τη φύση των δεδομένων εκπαίδευσης και τη μέθοδο μάθησης που χρησιμοποιείται:

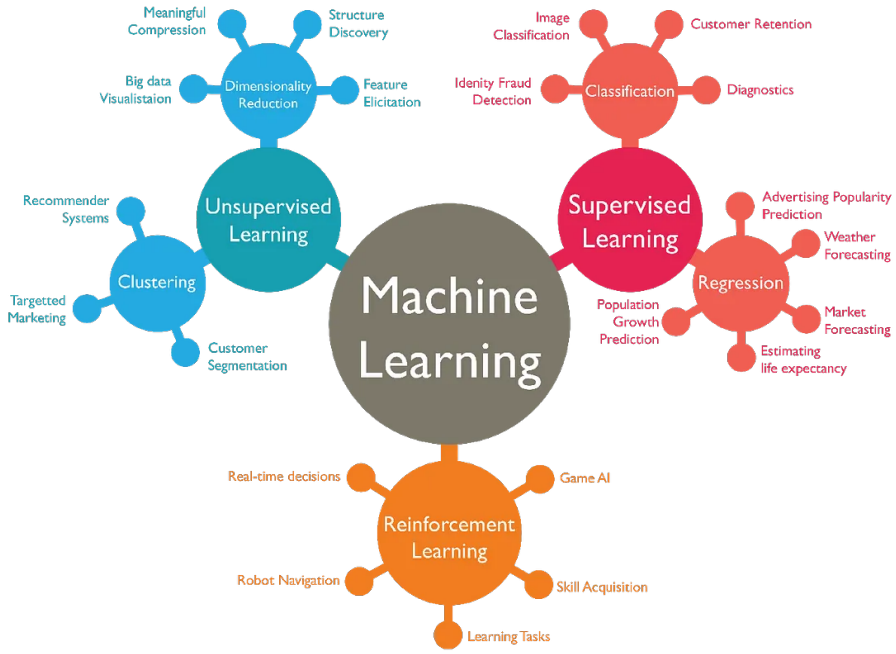
- **Επιβλεπόμενη Μάθηση (Supervised Learning)**: Η επιβλεπόμενη μάθηση είναι μια μέθοδος Μηχανικής Μάθησης όπου το μοντέλο εκπαιδεύεται με δεδομένα εισόδου που συνοδεύονται από γνωστές εξόδους (labelled data). Σκοπός της είναι η δημιουργία ενός μοντέλου που συσχετίζει τα χαρακτηριστικά εισόδου (features) με την επιθυμητή έξοδο (target value) και μπορεί να κάνει προβλέψεις για άγνωστα δεδομένα. Η διαδικασία εκπαίδευσης περιλαμβάνει τη χρήση ενός συνόλου δεδομένων εκπαίδευσης, όπου κάθε παράδειγμα αποτελείται από ένα ζεύγος εισόδου - εξόδου. Το μοντέλο μαθαίνει από αυτά τα δεδομένα, προσαρμόζοντας τις παραμέτρους του για να ελαχιστοποιήσει το σφάλμα μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Μετά την εκπαίδευση, αξιολογείται η απόδοση του μοντέλου σε ένα ξεχωριστό σύνολο δεδομένων, γνωστό ως σύνολο δοκιμών (test set), για να διασφαλιστεί η ικανότητά του να γενικεύει σωστά. Οι αλγόριθμοι επιβλεπόμενης μάθησης χωρίζονται σε ταξινόμηση (classification),

όπου η έξοδος είναι διακριτές κατηγορίες (π.χ. ανίχνευση κακόβουλων χρηστών), και παλινδρόμηση (regression), όπου η έξοδος είναι συνεχείς τιμές (π.χ. πρόβλεψη τιμής ακινήτου). Χρησιμοποιείται σε πλήθος εφαρμογών, όπως οι μηχανές συστάσεων και η ανάλυση δεδομένων κυκλοφορίας, αξιοποιώντας την ανίχνευση μοτίβων στα δεδομένα για τη συνεχή βελτίωση της ακρίβειας των προβλέψεων. [17]

- **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)**: Η μη επιβλεπόμενη μάθηση είναι μια προσέγγιση της Μηχανικής Μάθησης όπου τα δεδομένα εισόδου δεν συνοδεύονται από ετικέτες ή γνωστές τιμές εξόδου (unlabelled data). Ο στόχος των αλγορίθμων της είναι να ανακαλύψουν κρυφά μοτίβα, πρότυπα ή δομές μέσα στα δεδομένα, χωρίς να γνωρίζουν εκ των προτέρων τις κατηγορίες ή τις σχέσεις που υπάρχουν. Βασικές υποκατηγορίες προβλημάτων της μη επιβλεπόμενης μάθησης περιλαμβάνουν τη συσταδοποίηση (clustering), όπου τα δεδομένα ομαδοποιούνται βάσει ομοιοτήτων, τη συσχέτιση (association), την ανίχνευση ανωμαλιών (anomaly detection) και τη χρήση αυτόματων κωδικοποιητών (autoencoders) για τη συμπίεση και αναπαράσταση δεδομένων. Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη λόγω της αφθονίας μη επισημασμένων δεδομένων και βρίσκει εφαρμογές όπως η αναγνώριση προσώπου, η ανάλυση γονιδιακών αλληλουχιών και η μείωση διαστάσεων δεδομένων, παρέχοντας σημαντική γνώση σε ποικίλα πεδία χωρίς την ανάγκη εξαντλητικής επισήμανσης. [17]

- **Ενισχυτική Μάθηση (Reinforcement Learning)**: Η ενισχυτική μάθηση είναι μια προσέγγιση της Μηχανικής Μάθησης στην οποία ένα μοντέλο, γνωστό ως agent (πράκτορας), εκπαιδεύεται μέσω αλληλεπίδρασης με ένα δυναμικό περιβάλλον, εφαρμόζοντας τη διαδικασία «δοκιμή - σφάλμα» (trial and error). Ο agent λαμβάνει ανατροφοδότηση για τις ενέργειές του με τη μορφή ανταμοιβών (rewards) ή ποινών (penalties), χωρίς να έχει εκ των προτέρων πληροφορίες για τη σωστή λύση του προβλήματος. Στόχος είναι η ανάπτυξη μιας στρατηγικής (policy) που μεγιστοποιεί τη συνολική ανταμοιβή, και συνεπώς βελτιστοποιεί την συνολική απόδοση. Το περιβάλλον προσδιορίζει τις επιτρεπόμενες δράσεις, τους κανόνες και τις πιθανές καταστάσεις, ενώ η πολιτική ανταμοιβών και ποινών καθορίζεται από τον σχεδιαστή του συστήματος. Εφαρμογές ενισχυτικής μάθησης περιλαμβάνουν τη ρομποτική, την εκπαίδευση συστημάτων οδήγησης και την ανάπτυξη στρατηγικών παιχνιδιών, όπου το μοντέλο βελτιώνεται μέσω συνεχούς εμπειρίας και αυτοβελτιούμενης συμπεριφοράς. [17]

- **Ημι-Επιβλεπόμενη Μάθηση (Semi-Supervised Learning)**: Η ημι-επιβλεπόμενη μάθηση είναι ένα επιπλέον είδος Μηχανικής Μάθησης που εμφανίζεται συχνά και τοποθετείται μεταξύ της Επιβλεπόμενης και της Μη Επιβλεπόμενης Μάθησης. Χρησιμοποιείται όταν το μοντέλο καλείται να διαχειριστεί τόσο δεδομένα που διαθέτουν ετικέτα κατηγοριοποίησης, όσο και δεδομένα που δεν διαθέτουν ετικέτα. Ένα παράδειγμα χρήσης της Ημι-Επιβλεπόμενης Μάθησης είναι η ανίχνευση ανωμαλιών (anomaly detection), που στοχεύει στην ανίχνευση μοτίβων και σημείων που διαφέρουν από την νόρμα. Βρίσκει εφαρμογή σε τομείς όπως η τραπεζική δραστηριότητα και το χρηματιστήριο. [18]



Σχήμα 2.7: *Ποικιλία εφαρμογών της Μηχανικής Μάθησης* [19]

Μηχανική Μάθηση σε Προβλήματα Ταξινόμησης

Η μηχανική μάθηση αποτελεί ένα ισχυρό εργαλείο για την επίλυση προβλημάτων ταξινόμησης, όπου ο στόχος είναι η ανάθεση ετικετών (labels) ή κατηγοριών σε δεδομένα βάσει των χαρακτηριστικών τους (features). Σε αυτό το πλαίσιο, οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν εκπαιδευτικά σύνολα δεδομένων για να εντοπίσουν πρότυπα και συσχετίσεις, μαθαίνοντας να προβλέπουν την κατηγορία μιας νέας εισόδου. Οι αλγόριθμοι ταξινόμησης διακρίνονται κυρίως σε επιβλεπόμενους, όπως οι Logistic Regression, Random Forest, και Support Vector Machines (SVM), οι οποίοι εκπαιδεύονται με δεδομένα που περιλαμβάνουν γνωστές ετικέτες, καθώς και μη επιβλεπόμενους, που εφαρμόζουν τεχνικές όπως clustering όταν οι ετικέτες είναι άγνωστες. Τα προβλήματα ταξινόμησης διακρίνονται σε δυαδικά (binary classification), όπου υπάρχουν δύο κατηγορίες (π.χ. ναι/όχι, άνθρωπος/bot), και πολυκατηγορικά (multi-class classification), με περισσότερες από δύο κατηγορίες. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε τομείς όπως η ανίχνευση απάτης, η ανάλυση συναισθήματος, και η αναγνώριση αντικειμένων σε εικόνες, όπου η ακρίβεια και η αυτοματοποίηση είναι ζωτικής σημασίας. [20]

Παρακάτω δίνονται σύντομοι ορισμοί για έννοιες της μηχανικής μάθησης που θα αναφερθούν στην συνέχεια αυτής της ενότητας.

- **Νευρωνικά Δίκτυα: Τα Νευρωνικά Δίκτυα (Neural Networks)** είναι μοντέλα μηχανικής μάθησης εμπνευσμένα από τη λειτουργία του ανθρώπινου εγκεφάλου. Αποτελούνται από έναν αριθμό διασυνδεδεμένων τεχνητών νευρώνων, οργανωμένων σε στρώσεις. Κάθε νευρώνας δέχεται είσοδο, την επεξεργάζεται μέσω μιας συνάρτησης ενεργοποίησης και προωθεί την έξοδό του στους επόμενους νευρώνες. Μέσω της διαδικασίας εκπαίδευσης, τα βάρη των συνδέσεων ανάμεσα στους νευρώνες προσαρμόζονται

ώστε να ελαχιστοποιηθεί το σφάλμα πρόβλεψης. Τα Νευρωνικά Δίκτυα χρησιμοποιούνται ευρέως σε προβλήματα όπως η ταξινόμηση, η αναγνώριση εικόνων και η επεξεργασία φυσικής γλώσσας. [21]

- **Graph Neural Networks (GNNs)**: Τα Graph Neural Networks (GNNs) είναι μια κατηγορία αλγορίθμων βαθιάς μάθησης που έχουν σχεδιαστεί για να επεξεργάζονται δεδομένα που αναπαρίστανται σε μορφή γράφων. Αντί να λειτουργούν σε δομημένα δεδομένα όπως πίνακες ή εικόνες, τα GNNs αξιοποιούν τη δομή των γράφων, ενσωματώνοντας πληροφορίες από τους κόμβους και τις ακμές τους. Μέσω επαναληπτικών διαδικασιών, οι αναπαραστάσεις των κόμβων ενημερώνονται με βάση τα χαρακτηριστικά των γειτόνων τους. Αυτή η μέθοδος επιτρέπει την επίλυση πολύπλοκων προβλημάτων, όπως η πρόβλεψη συνδέσεων, η κατηγοριοποίηση κόμβων και η ανάλυση κοινωνικών δικτύων. [22]

- **Βαθιά Μάθηση**: Η Βαθιά Μάθηση (Deep Learning) είναι ένας τομέας της μηχανικής μάθησης που βασίζεται σε πολυεπίπεδα νευρωνικά δίκτυα. Τα μοντέλα βαθιάς μάθησης περιλαμβάνουν πολλαπλές στρώσεις που επιτρέπουν την ιεραρχική εξαγωγή χαρακτηριστικών από τα δεδομένα, μεταβαίνοντας από απλά μοτίβα σε πιο σύνθετες δομές. Αυτή η προσέγγιση έχει φέρει επανάσταση σε τομείς όπως η αναγνώριση φωνής και η αυτόνομη οδήγηση, χάρη στη δυνατότητά της να επεξεργάζεται μεγάλα σύνολα δεδομένων και να βρίσκει πολύπλοκες σχέσεις. [23]

2.2.3 Logistic Regression

Η λογιστική παλινδρόμηση (Logistic Regression) είναι μια στατιστική μέθοδος και αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται κυρίως για δυαδικά προβλήματα ταξινόμησης, δηλαδή όταν η έξοδος ανήκει σε δύο διακριτές κατηγορίες (0 ή 1). Παρά την ονομασία του, το Logistic Regression δεν είναι ένας αλγόριθμος παλινδρόμησης, αλλά βασίζεται στην εκτίμηση της πιθανότητας εμφάνισης μιας κατηγορίας, χρησιμοποιώντας μια γραμμική συνάρτηση εισόδων και τη λογιστική συνάρτηση (sigmoid) για να παράγει τιμές μεταξύ 0 και 1. [24] [25]

Βασική Ιδέα και Αρχές Λειτουργίας

Η βασική ιδέα του Logistic Regression είναι η χρήση μιας γραμμικής συνάρτησης z για την πρόβλεψη, που υπολογίζεται ως εξής:

$$z = \mathbf{w}^T\mathbf{x} + b$$

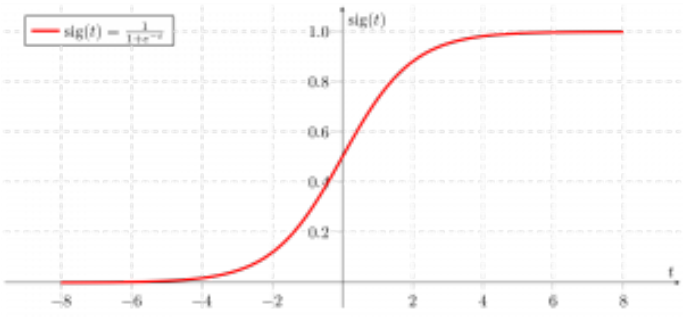
όπου $\mathbf{w}$ είναι το διάνυσμα των βαρών (weights), $\mathbf{x}$ είναι το διάνυσμα χαρακτηριστικών (features) και b είναι ο όρος απόκλισης (bias).

Για να μετατραπεί το αποτέλεσμα της γραμμικής συνάρτησης σε πιθανότητα, εφαρμόζεται η λογιστική συνάρτηση:

$$h_\theta(\mathbf{x}) = \sigma(z) = \frac{1}{1 + e^{-z}}$$

Η λογιστική συνάρτηση (sigmoid function) είναι μια μαθηματική συνάρτηση που χρησιμοποιείται για να μετασχηματίσει τις προβλεπόμενες τιμές σε πιθανότητες. Μετατρέπει οποιαδήποτε πραγματική τιμή σε μια άλλη τιμή εντός του διαστήματος 0 έως 1. Η τιμή $h_\theta(\mathbf{x})$ αντιπροσωπεύει, επομένως, την πιθανότητα το δείγμα $\mathbf{x}$ να ανήκει στην κατηγορία «1».

Η τιμή της λογιστικής παλινδρόμησης πρέπει να βρίσκεται ανάμεσα στο 0 και το 1, χωρίς να μπορεί να ξεπεράσει αυτά τα όρια, σχηματίζοντας έτσι μια καμπύλη με τη μορφή "S". Αυτή η καμπύλη με τη μορφή "S" ονομάζεται λογιστική συνάρτηση (sigmoid function) ή λογιστική καμπύλη (logistic function).



Σχήμα 2.8: *Η λογιστική συνάρτηση (sigmoid function)*
[24]

Στη λογιστική παλινδρόμηση, χρησιμοποιείται η έννοια της τιμής κατωφλίου (threshold value), η οποία ορίζει την πιθανότητα για την κατηγορία 0 ή 1. Δηλαδή, τιμές που βρίσκονται πάνω από την τιμή κατωφλίου τείνουν προς το 1, ενώ τιμές που βρίσκονται κάτω από την τιμή κατωφλίου τείνουν προς το 0. Συνήθως ως κατώφλι ορίζεται η τιμή 0.5.

Εκπαίδευση του Μοντέλου

Η εκπαίδευση του Logistic Regression πραγματοποιείται μέσω βελτιστοποίησης των παραμέτρων w και b, ώστε να ελαχιστοποιηθεί μια συνάρτηση απώλειας, συνήθως η λογιστική απώλεια (log-loss):

$$L(w, b) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log h_\theta(\mathbf{x}_i) + (1 - y_i)\log(1 - h_\theta(\mathbf{x}_i))\right]$$

όπου N είναι ο αριθμός των δειγμάτων, y_i είναι η πραγματική ετικέτα (label) του δείγματος i και $h_\theta(\mathbf{x}_i)$ είναι η προβλεπόμενη πιθανότητα για το δείγμα i.

Η βελτιστοποίηση πραγματοποιείται συχνά μέσω αλγορίθμων όπως η Στοχαστική Καταβατική Κλίση (Stochastic Gradient Descent - SGD).

Ιδιότητες και Εφαρμογές

Το Logistic Regression θεωρείται εύκολο στην εφαρμογή και αποδοτικό μοντέλο για μικρά και μεσαίου μεγέθους σύνολα δεδομένων. Είναι γραμμικό μοντέλο, που σημαίνει ότι υποθέτει πως τα δεδομένα είναι γραμμικά διαχωρίσιμα στον χώρο των χαρακτηριστικών. Ωστόσο, μπορεί να επεκταθεί σε πιο πολύπλοκες καταστάσεις μέσω της χρήσης πολυωνυμικών

χαρακτηριστικών (polynomial features) ή αλληλεπιδράσεων μεταξύ χαρακτηριστικών. Αυτό το μοντέλο χρησιμοποιείται εκτενώς σε προβλήματα όπως η ανίχνευση απάτης, η διάγνωση ασθενειών, και η ανάλυση συναισθημάτων, χάρη στην απλότητα και την ικανότητά του να παρέχει προβλέψεις που ερμηνεύονται χωρίς ιδιαίτερη δυσκολία.

2.2.4 Random Forest

Τα τυχαία δάση (Random Forest) είναι μια ισχυρή και ευέλικτη μέθοδος μηχανικής μάθησης, που ανήκει στους αλγόριθμους συνόλων (ensemble learning). Βασίζεται στη δημιουργία πολλών τυχαίων δέντρων απόφασης (decision trees) και στον συνδυασμό των αποτελεσμάτων τους, ώστε να βελτιωθεί η ακρίβεια και η γενίκευση του μοντέλου. Η μέθοδος χρησιμοποιείται τόσο για προβλήματα ταξινόμησης (classification) όσο και παλινδρόμησης (regression). [26] [27] [28]

Βασική Ιδέα και Αρχές Λειτουργίας

Το Random Forest δημιουργεί πολλά δέντρα απόφασης μέσω μιας διαδικασίας που ονομάζεται ενσάκιση, ή bootstrap aggregation (bagging). Στο βήμα αυτό, δημιουργούνται τυχαία υποσύνολα του αρχικού συνόλου δεδομένων χρησιμοποιώντας δειγματοληψία με επανατοποθέτηση (sampling with replacement). Κάθε υποσύνολο χρησιμοποιείται για την εκπαίδευση ενός ξεχωριστού δέντρου απόφασης.

Κατά τη δημιουργία κάθε δέντρου, για τον διαχωρισμό (split) των κόμβων δεν λαμβάνονται υπόψη όλα τα χαρακτηριστικά (features), αλλά μόνο ένα τυχαίο υποσύνολό τους. Αυτός ο τυχαίος περιορισμός μειώνει την συσχέτιση (correlation) μεταξύ των δέντρων, βελτιώνοντας έτσι την ακρίβεια του τελικού μοντέλου.

Για προβλήματα ταξινόμησης, τα δέντρα δίνουν τις προβλέψεις τους για την κατηγορία, και το τελικό αποτέλεσμα προκύπτει μέσω πλειοψηφίας ψήφων (majority voting).

Για προβλήματα παλινδρόμησης, η πρόβλεψη προκύπτει από τον μέσο όρο (mean) των αποτελεσμάτων όλων των δέντρων.

Υπολογισμοί και Εξισώσεις

Διαχωρισμός Κόμβων (Node Splitting)

Η επιλογή του χαρακτηριστικού για τον διαχωρισμό γίνεται με στόχο την ελαχιστοποίηση του μέτρου ανωμαλίας (impurity measure). Ένα δημοφιλές μέτρο είναι η εντροπία (entropy) ή το Gini index.

Η εντροπία δίνεται από τη σχέση:

$$H(S) = -\sum_{i=1}^{C} p_i \log_2 p_i$$

όπου p_i είναι η πιθανότητα μιας παρατήρησης να ανήκει στην κατηγορία i και C το σύνολο των κατηγοριών.

Το Gini index δίνεται από τη σχέση:

$$G(S) = 1 - \sum_{i=1}^{C} p_i^2$$

Τελική Πρόβλεψη

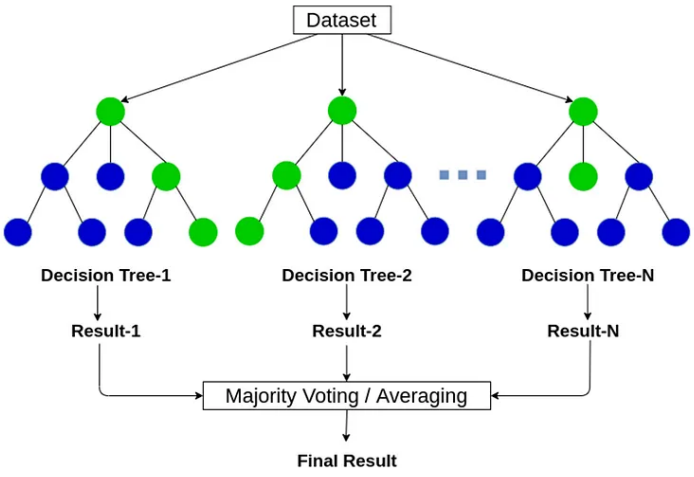
Για ταξινόμηση:

$$y = \text{mode}(y_1, y_2, \ldots, y_T)$$

όπου y_i είναι η πρόβλεψη του i-οστού δέντρου και T ο συνολικός αριθμός δέντρων.

Για παλινδρόμηση:

$$y = \frac{1}{T} \sum_{i=1}^{T} y_i$$



Σχήμα 2.9: *Απεικόνιση της ροής του αλγορίθμου Random Forest μέχρι την ταξινόμηση.* [29]

Ιδιότητες και Εφαρμογές

Το Random Forest είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης που συνδυάζει πολλά δέντρα απόφασης, μειώνοντας έτσι τον κίνδυνο υπερπροσαρμογής και βελτιώνοντας τη γενίκευση των προβλέψεων. Διαχειρίζεται αποτελεσματικά μη γραμμικά δεδομένα, καθώς τα δέντρα του μπορούν να αποτυπώσουν πολύπλοκες σχέσεις μεταξύ των χαρακτηριστικών. Παράλληλα, η τυχαία επιλογή υποσυνόλου χαρακτηριστικών σε κάθε διακλάδωση μειώνει τη συσχέτιση μεταξύ των δέντρων και βελτιστοποιεί την απόδοσή του. Είναι ανθεκτικό σε ασαφή ή ελλιπή δεδομένα, ενώ επιτρέπει την εκτίμηση της σημαντικότητας των χαρακτηριστικών, καθιστώντας το ιδιαίτερα χρήσιμο σε εφαρμογές όπως η ιατρική διάγνωση, η ανίχνευση απάτης και η ανάλυση χρηματοοικονομικών δεδομένων.

2.2.5 Graph Convolutional Networks

Τα Graph Convolutional Networks (GCNs) είναι μια κατηγορία νευρωνικών δικτύων σχεδιασμένων για τη μοντελοποίηση δεδομένων που αναπαρίστανται σε μορφή γράφων. Οι γράφοι αποτελούν γενικευμένες δομές δεδομένων, στις οποίες κόμβοι συνδέονται μέσω ακμών, αποτυπώνοντας πολύπλοκες σχέσεις και αλληλεπιδράσεις. Τα GCNs επεκτείνουν τις ιδέες των παραδοσιακών συνελικτικών νευρωνικών δικτύων (CNNs) από δομημένα δεδομένα, όπως εικόνες, σε μη δομημένα ή ημι-δομημένα δεδομένα που χαρακτηρίζουν τους γράφους. [30] [31] [32]

Βασική Ιδέα και Αρχές Λειτουργίας

Η βασική λειτουργία των GCNs είναι η συσσώρευση πληροφορίας (message passing) από τους γειτονικούς κόμβους, ώστε να προκύψουν αναπαραστάσεις (embeddings) που περιλαμβάνουν πληροφορίες από τη δομή του γράφου και τα χαρακτηριστικά των κόμβων. Η λειτουργία τους μπορεί να περιγραφεί μέσω επαναλαμβανόμενων συνελικτικών βημάτων (convolutional steps), κατά τα οποία κάθε κόμβος ενημερώνει τα χαρακτηριστικά του χρησιμοποιώντας πληροφορίες από τους γείτονές του.

Υπολογισμοί και Εξισώσεις

Ένας γράφος $G = (V, E)$ αποτελείται από: V: το σύνολο των κόμβων ($|V| = N$), και E: το σύνολο των ακμών.

Τα δεδομένα του γράφου αναπαρίστανται μέσω:

- Πίνακα γειτνίασης A:

$$A_{ij} = \begin{cases} 1, & \text{αν υπάρχει ακμή μεταξύ των κόμβων } i \text{ και } j \\ 0, & \text{διαφορετικά} \end{cases}$$

- Πίνακα χαρακτηριστικών X:

$$X \in \mathbb{R}^{N \times F}$$

όπου F είναι ο αριθμός των χαρακτηριστικών ανά κόμβο.

Βασική Ενημέρωση Χαρακτηριστικών

Η διαδικασία ενημέρωσης για έναν κόμβο v_i σε ένα GCN περιλαμβάνει δύο στάδια:

- Συσσώρευση πληροφορίας (aggregation): Ο κόμβος συλλέγει πληροφορίες από τους γείτονές του.

- Ενημέρωση (update): Ο κόμβος ενημερώνει τα χαρακτηριστικά του χρησιμοποιώντας τη συσσωρευμένη πληροφορία.

Η λειτουργία ενός GCN περιγράφεται γενικά από τη σχέση:

$$H^{(l+1)} = \sigma\left(D'^{-\frac{1}{2}} A' D'^{-\frac{1}{2}} H^{(l)} W^{(l)}\right)$$

όπου:

- $H^{(l)}$: τα χαρακτηριστικά στη στρώση l, με $H^{(0)} = X$.

- $A' = A + I$: ο κανονικοποιημένος πίνακας γειτνίασης με προσθήκη αυτοσυνδέσεων (self-loops).

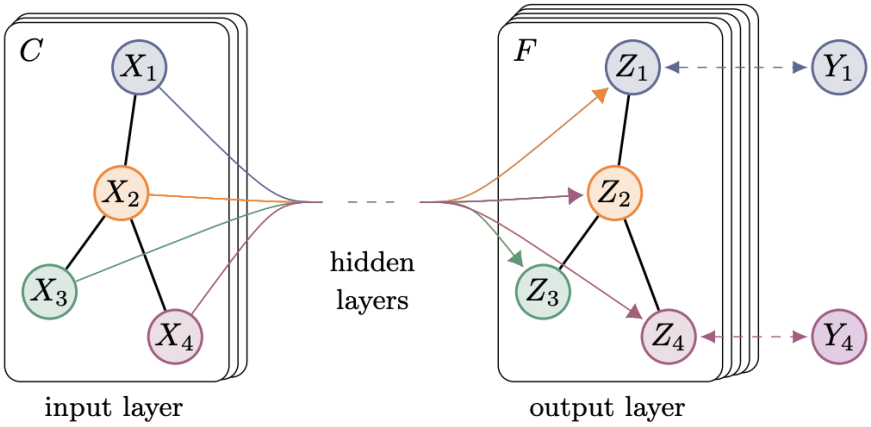
- D': ο διαγώνιος πίνακας βαθμού, με $D'_{ii} = \sum_j A'_{ij}$.

- $W^{(l)}$: οι παράμετροι της στρώσης l που εκπαιδεύονται.

- σ: μια μη γραμμική συνάρτηση ενεργοποίησης, όπως ReLU.

Κανονικοποίηση Πληροφορίας

Η χρήση του $D'^{-\frac{1}{2}}A'D'^{-\frac{1}{2}}$ εξασφαλίζει ότι η πληροφόρηση από τους γείτονες ζυγίζεται ανάλογα με τον βαθμό (degree) τους, ώστε να αποφευχθούν διακυμάνσεις λόγω υψηλής σύνδεσης (over-smoothing).



Σχήμα 2.10: *Graph Convolutional Network* [33]

Ιδιότητες και Εφαρμογές

Τα Graph Convolutional Networks (GCNs) αξιοποιούν τη δομή των γράφων για να επεξεργαστούν δεδομένα με περίπλοκες σχέσεις μεταξύ οντοτήτων, επιτρέποντας την αποδοτική μάθηση σε μη ευκλείδειους χώρους. Μέσω αυτοματοποιημένης μάθησης αναπαραστάσεων, δημιουργούν embeddings που κωδικοποιούν τόσο τις συνδέσεις όσο και τα χαρακτηριστικά των κόμβων, βελτιώνοντας την ανάλυση δικτύων. Η ικανότητά τους να γενικεύουν τα καθιστά κατάλληλα για διάφορες εφαρμογές, όπως ταξινόμηση κόμβων, πρόβλεψη ακμών και συστήματα συστάσεων. Αποτελούν βασικό εργαλείο στην ανάλυση δεδομένων γράφων, συνδυάζοντας τη μαθηματική μοντελοποίηση με τη δύναμη των νευρωνικών δικτύων.

2.2.6 Graph Attention Networks

Τα Graph Attention Networks (GATs) αποτελούν μια εξελιγμένη αρχιτεκτονική βαθιάς μάθησης για την επεξεργασία δεδομένων που αναπαρίστανται σε γράφους. Τα GATs ενσωματώνουν τη δύναμη της μηχανικής μάθησης με προσοχή (attention mechanism) για να αντιμετωπίσουν ορισμένους περιορισμούς των Graph Convolutional Networks (GCNs), όπως η αδυναμία δυναμικής στάθμισης των γειτονικών κόμβων. [34] [35]

Βασική Ιδέα και Αρχές Λειτουργίας

Η βασική ιδέα πίσω από τα GATs είναι να επιτρέπουν στο μοντέλο να αναθέτει διαφορετική σημασία (attention weights) στους γείτονες ενός κόμβου, ανάλογα με τη σχετικότητά τους, κατά τη διαδικασία ενημέρωσης των χαρακτηριστικών. Αυτή η ευελιξία καθιστά τα GATs ιδιαίτερα κατάλληλα για δεδομένα γράφων με ανομοιογενείς σχέσεις.

Το κάθε GAT αποτελείται από στρώσεις προσοχής (attention layers) που λειτουργούν μέσω των παρακάτω βημάτων:

1. Υπολογισμός της Προσοχής: Οι γειτονικοί κόμβοι ενός κόμβου i λαμβάνουν διαφορετικά βάρη ανάλογα με τη σημασία τους.

2. Συγκέντρωση (Aggregation): Τα χαρακτηριστικά των γειτόνων ενσωματώνονται χρησιμοποιώντας τα βάρη αυτά.

3. Μη γραμμικός Μετασχηματισμός: Εφαρμόζεται μια συνάρτηση ενεργοποίησης για την τελική αναπαράσταση.

Υπολογισμοί και Εξισώσεις

Έστω ένας γράφος $G = (V, E)$, με N κόμβους και F χαρακτηριστικά ανά κόμβο. Τα αρχικά χαρακτηριστικά των κόμβων δίνονται από τον πίνακα $X \in \mathbb{R}^{N \times F}$.

Υπολογισμός Συντελεστή Προσοχής

Για κάθε ζεύγος γειτονικών κόμβων i, j:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^T \left[W\mathbf{h}_i^{(l)} \parallel W\mathbf{h}_j^{(l)}\right]\right)$$

όπου:

- $\mathbf{h}_i^{(l)}$: τα χαρακτηριστικά του κόμβου i στη στρώση l,
- W: ο πίνακας βαρών που εφαρμόζεται στα χαρακτηριστικά,
- $\mathbf{a}$: το διάνυσμα βαρών προσοχής,
- $\parallel$: η συνένωση (concatenation) δύο διανυσμάτων.

Κανονικοποίηση Συντελεστών Προσοχής

Οι συντελεστές προσοχής κανονικοποιούνται χρησιμοποιώντας τη συνάρτηση ενεργοποίησης softmax, ώστε το άθροισμά τους να είναι 1:

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})}$$

όπου $\mathcal{N}(i)$ είναι το σύνολο των γειτόνων του κόμβου i.

Ενημέρωση Χαρακτηριστικών Κόμβων

Οι τελικές αναπαραστάσεις υπολογίζονται ως στάθμιση των χαρακτηριστικών των γειτόνων:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} a_{ij} W \mathbf{h}_j^{(l)}\right)$$

όπου σ μια μη γραμμική συνάρτηση ενεργοποίησης, όπως το ReLU.

ReLU

Το Rectified Linear Unit (ReLU) είναι μια συνάρτηση ενεργοποίησης που ορίζεται ως:

$$f(x) = \max(0, x)$$

Διατηρεί τις θετικές τιμές αναλλοίωτες και μετατρέπει τις αρνητικές σε μηδέν, επιταχύνοντας τη σύγκλιση των νευρωνικών δικτύων και μειώνοντας το πρόβλημα της εξαφάνισης του gradient.

Πολυκεφαλική Προσοχή (Multi-Head Attention)

Για να ενισχυθεί η ισχύς του μοντέλου, τα GATs χρησιμοποιούν πολλαπλές κεφαλές προσοχής (multi-head attention). Αυτό σημαίνει ότι οι υπολογισμοί της προσοχής επαναλαμβάνονται με ανεξάρτητα σύνολα βαρών, και τα αποτελέσματα:

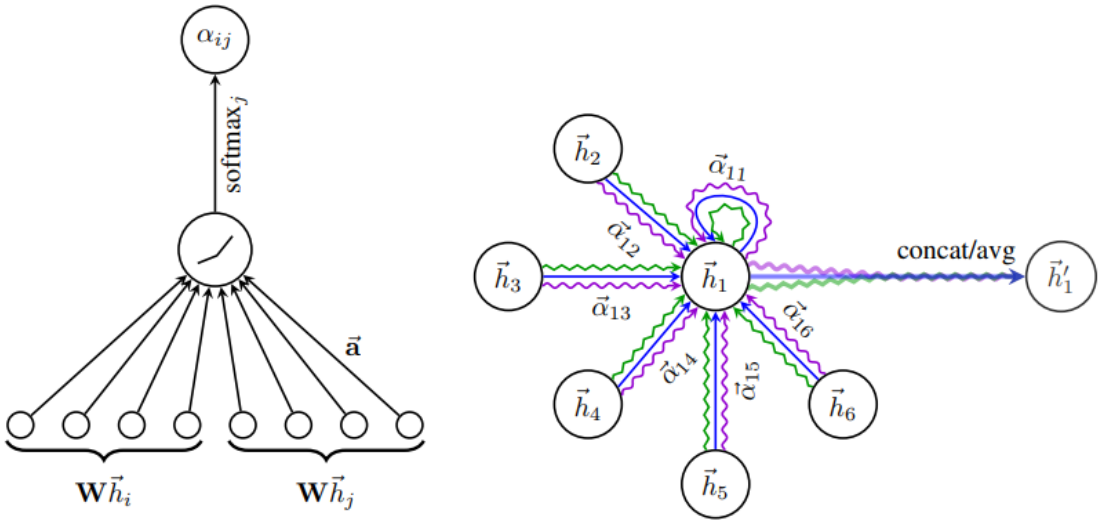
- Συγχωνεύονται (concatenation):

$$\mathbf{h}_i^{(l+1)} = \Big\|_{m=1}^{K} \sigma\left(\sum_{j \in \mathcal{N}(i)} a_{ij}^{(m)} W^{(m)} \mathbf{h}_j^{(l)}\right)$$

- Μέσος όρος (averaging): Συνδυάζονται μέσω μέσης τιμής για σταθεροποίηση.

Ιδιότητες και Εφαρμογές

Τα Graph Attention Networks (GATs) εισάγουν έναν μηχανισμό προσαρμοστικής προσοχής που επιτρέπει τη δυναμική στάθμιση της σημασίας κάθε γείτονα, βελτιώνοντας τη μάθηση σημασιολογίας στα γραφήματα. Σε αντίθεση με τα GCNs, δεν απαιτούν κανονικοποιημένο πίνακα γειτνίασης, προσφέροντας μεγαλύτερη ευελιξία στη μοντελοποίηση σύνθετων σχέσεων. Επιπλέον, η πολυκεφαλική προσοχή ενισχύει την ακρίβεια και τη γενίκευση, καθιστώντας τα GATs ιδανικά για εφαρμογές σε κοινωνικά δίκτυα, συστήματα συστάσεων και ανάλυση βιολογικών δεδομένων. Αυτή η καινοτόμος προσέγγιση συνδυάζει τη δύναμη των γράφων με την προσαρμοστική προσοχή, επεκτείνοντας τις δυνατότητες ανάλυσης πολύπλοκων δικτύων.



Σχήμα 2.11: *Αριστερά: Ο μηχανισμός προσοχής που χρησιμοποιείται σε ένα GAT μουτέλο, παραμετροποιημένος από ένα διάνυσμα βαρών, με εφαρμογή της ενεργοποίησης LeakyReLU. Δεξιά: Μια απεικόνιση της πολυκεφαλικής προσοχής (με 3 κεφαλές) από τον κόμβο 1 προς τους γείτονές του. Οι διαφορετικοί τύποι και χρώματα βελών αντιπροσωπεύουν ανεξάρτητους υπολογισμούς προσοχής.*

[35]

2.3 Μεγάλα Γλωσσικά Μοντέλα

Σε αυτή την ενότητα δίνεται μια εισαγωγή στα Μεγάλα Γλωσσικά Μοντέλα, τις χρήσεις και τον τρόπο λειτουργίας τους. Έπειτα, διατυπώνονται ορισμένα πλεονεκτήματα που προκύπτουν από τη χρήση τους, αλλά και ορισμένες προκλήσεις που παρουσιάζονται [36] [37] [38]. Τέλος, δίνεται μια συνοπτική παρουσίαση της ταξινόμησης κειμένου με την αξιοποίηση Μεγάλων Γλωσσικών Μοντέλων. [39]

2.3.1 Εισαγωγή στα Μεγάλα Γλωσσικά Μοντέλα

Μεγάλα Γλωσσικά Μοντέλα: Ορισμός και Χαρακτηριστικά

Ένα Μεγάλο Γλωσσικό Μοντέλο (Large Language Model - LLM) αποτελεί έναν τύπο αλγόριθμου τεχνητής νοημοσύνης (AI) που έχει σχεδιαστεί με σκοπό να επεξεργάζεται και να παράγει κείμενο με τρόπο που να μοιάζει με την ανθρώπινη γραφή. Αυτού του είδους τα μοντέλα βασίζονται σε προηγμένες τεχνικές μηχανικής μάθησης, ιδίως στη βαθιά μάθηση (deep learning), και εκπαιδεύονται σε τεράστιους όγκους δεδομένων, που συχνά αντλούνται από πηγές στο διαδίκτυο. Η εκτεταμένη εκπαίδευσή τους, τους επιτρέπει να κατανοούν και να ερμηνεύουν τα μοτίβα της γλώσσας, δίνοντάς τους τη δυνατότητα να εκτελούν μια πληθώρα εργασιών επεξεργασίας φυσικής γλώσσας (NLP).

Τα LLMs χαρακτηρίζονται από την τεράστια κλίμακά τους, η οποία συχνά περιλαμβάνει εκατομμύρια έως και τρισεκατομμύρια παραμέτρους. Οι παράμετροι αυτές είναι μαθηματικές τιμές που βοηθούν το μοντέλο να κατανοεί τις σχέσεις και τα μοτίβα μεταξύ λέξεων και φράσεων. Για παράδειγμα, το GPT-4, ένα από τα πιο γνωστά LLMs, εκτιμάται ότι διαθέτει

έως και ένα τρισεκατομμύριο παραμέτρους [40]. Αυτή η κλίμακα επιτρέπει στα LLMs να εκτελούν εργασίες όπως η δημιουργία κειμένου, η ανάλυση συναισθήματος, η μετάφραση γλωσσών και πολλά άλλα, με υψηλό βαθμό ακρίβειας και συνοχής.

Ουσιαστικά, τα LLMs είναι ισχυρά εργαλεία που γεφυρώνουν το χάσμα μεταξύ της μηχανικής επεξεργασίας και της ανθρώπινης επικοινωνίας, καθιστώντας τα απαραίτητα για εφαρμογές όπως οι αυτοματοποιημένοι συνομιλητές (chatbots), η δημιουργία περιεχομένου και η αυτοματοποιημένη ανάλυση. Αν και βασίζονται σε σύνθετους αλγόριθμους και τεράστια σύνολα δεδομένων, ο τελικός τους στόχος είναι να παράγουν απαντήσεις και αποτελέσματα που προσομοιάζουν τη σκέψη και την έκφραση των ανθρώπων.

Χρήσεις των Μεγάλων Γλωσσικών Μοντέλων

Μία από τις βασικές εφαρμογές των LLMs είναι η δημιουργία και αναδιατύπωση κειμένου, καθώς συμβάλλουν στη συγγραφή και βελτίωση περιεχομένου, όπως δοκίμια, άρθρα και παραφρασμένα κείμενα. Επιπλέον, αξιοποιούνται στη μετάφραση γλώσσας, καθώς μοντέλα εκπαιδευμένα σε γλωσσικά δεδομένα χρησιμοποιούνται ευρέως για τη μετάφραση κειμένου μεταξύ διαφορετικών γλωσσών, ενσωματώνοντας εργαλεία όπως το Google Translate για την παροχή μεταφράσεων με ακρίβεια.

Μια άλλη σημαντική χρήση τους είναι η ανάλυση συναισθήματος, η οποία επιτρέπει την κατανόηση του τόνου ενός κειμένου. Αυτό είναι ιδιαίτερα χρήσιμο για επιχειρήσεις που θέλουν να διαχειριστούν τα σχόλια των πελατών και να βελτιώσουν τις στρατηγικές μάρκετινγκ. Παράλληλα, τα LLMs τροφοδοτούν συστήματα συνομιλίας (chatbots) με τεχνητή νοημοσύνη, όπως το ChatGPT, καθιστώντας δυνατές τις φυσικές και συμφραζόμενες (context-aware) συνομιλίες με τους χρήστες.

Τα γλωσσικά μοντέλα διαθέτουν επίσης ικανότητες περίληψης και κατηγοριοποίησης, επιτρέποντας τη συμπύκνωση πληροφοριών και την αποτελεσματική οργάνωση περιεχομένου. Στον τομέα του προγραμματισμού, παρέχουν βοήθεια στη συγγραφή κώδικα και στον εντοπισμό σφαλμάτων, με εργαλεία όπως το GitHub Copilot να αποτελούν πολύτιμους βοηθούς.

Ένα ακόμη πεδίο εφαρμογής τους είναι η μετατροπή ομιλίας σε κείμενο (speech-to-text), επιτρέποντας την αυτόματη μετατροπή της προφορικής γλώσσας σε γραπτό κείμενο, διευκολύνοντας την προσβασιμότητα. Επιπλέον, βελτιώνουν και διευκολύνουν την εμπειρία της διαδικτυακής αναζήτησης, καθιστώντας την πιο ακριβή και αποδοτική.

Συνολικά, τα Μεγάλα Γλωσσικά Μοντέλα αποτελούν ευέλικτα εργαλεία που βελτιώνουν την επικοινωνία, αυτοματοποιούν εργασίες και ενισχύουν την καινοτομία σε διάφορους τομείς.

Τρόπος Λειτουργίας των Μεγάλων Γλωσσικών Μοντέλων

Τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) λειτουργούν μέσω μιας πολυεπίπεδης διαδικασίας που συνδυάζει την εκτενή εκπαίδευση πάνω σε δεδομένα, τεχνικές βαθιάς μάθησης και προηγμένες αρχιτεκτονικές νευρωνικών δικτύων, όπως οι μετασχηματιστές. Η πρώτη φάση της διαδικασίας είναι η συλλογή και προεπεξεργασία δεδομένων, κατά την οποία τα LLMs εκπαιδεύονται σε τεράστια σύνολα δεδομένων προερχόμενα από βιβλία, άρθρα και το διαδίκτυο. Τα κείμενα αυτά μετατρέπονται σε αριθμητικές αναπαραστάσεις, ώστε να μπορούν

να αναλυθούν από τους αλγόριθμους μηχανικής μάθησης.

Η εκπαίδευση ξεκινά με μη εποπτευόμενη μάθηση (unsupervised learning), όπου το μοντέλο εντοπίζει μοτίβα και σχέσεις σε μη δομημένα και μη επισημασμένα δεδομένα. Στη συνέχεια, εφαρμόζεται αυτο-εποπτευόμενη μάθηση (self-supervised learning), που χρησιμοποιεί μερικώς επισημασμένα δεδομένα για να βελτιώσει την ικανότητα του μοντέλου να αναγνωρίζει έννοιες και συσχετισμούς. Η βαθιά μάθηση διαδραματίζει κρίσιμο ρόλο, καθώς τα LLMs χρησιμοποιούν πιθανολογικές μεθόδους για να αναλύσουν μοτίβα και να προβλέψουν πιθανές ακολουθίες λέξεων. Αυτή η επαναληπτική διαδικασία βοηθά το μοντέλο να αποκτήσει δεξιότητες κατανόησης κειμένου και δημιουργίας του με τρόπο που μοιάζει με αυτόν του ανθρώπου.

Στην καρδιά των LLMs βρίσκονται τα νευρωνικά δίκτυα, τα οποία μιμούνται τον τρόπο που ο ανθρώπινος εγκέφαλος επεξεργάζεται πληροφορίες. Αυτά τα δίκτυα αποτελούνται από διασυνδεδεμένα στρώματα που αναλύουν και μεταδίδουν δεδομένα. Τα LLMs βασίζονται ειδικά σε μοντέλα μετασχηματιστών, μια αρχιτεκτονική νευρωνικών δικτύων που χρησιμοποιεί μηχανισμό αυτο-προσοχής (self-attention). Αυτός ο μηχανισμός αξιολογεί τη σημασία κάθε λέξης σε μια πρόταση και καθορίζει τις σχέσεις της με άλλες, συλλαμβάνοντας αποτελεσματικά το πλαίσιο και το νόημα. Οι μετασχηματιστές είναι εξαιρετικοί στη διαχείριση εξαρτήσεων μεταξύ προτάσεων και παραγράφων, επιτρέποντας στα LLMs να κατανοούν πολύπλοκες γλωσσικές δομές.

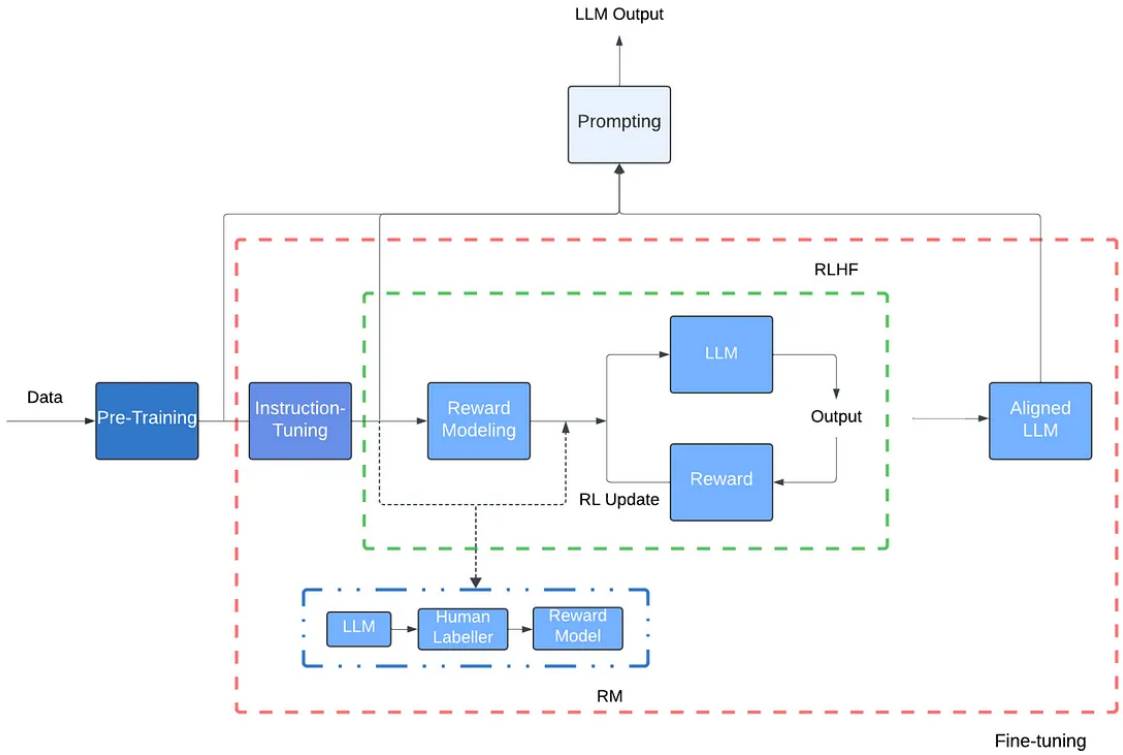
Αφού ολοκληρωθεί η εκπαίδευση, τα LLMs μπορούν να τελειοποιηθούν για συγκεκριμένες εργασίες ή τομείς χρησιμοποιώντας μικρότερα, επιμελημένα σύνολα δεδομένων. Αυτή η εξειδίκευση βελτιώνει την απόδοσή τους σε τομείς όπως η μετάφραση, η ανάλυση συναισθήματος και η σύνταξη κώδικα σε διάφορες γλώσσες προγραμματισμού.

Μετά την ολοκλήρωση της εκπαίδευσής τους, τα LLMs μπορούν να ανταποκρίνονται σε ερωτήσεις χρηστών, δημιουργώντας κείμενο, συνοψίζοντας πληροφορίες, απαντώντας σε ερωτήματα ή πραγματοποιώντας ανάλυση συναισθήματος. Η ικανότητά τους να κατανοούν και να προσαρμόζονται στο πλαίσιο που τους δίνεται, τα καθιστά ικανά να χειρίζονται αποτελεσματικά μια ευρεία γκάμα εργασιών επεξεργασίας φυσικής γλώσσας. Αυτή η πολυεπίπεδη προσέγγιση - από τη μαζική εκπαίδευση δεδομένων έως την εξειδικευμένη τελειοποίηση - καθιστά τα LLMs ισχυρά εργαλεία για την ερμηνεία και τη δημιουργία ανθρώπινης γλώσσας.

Τύποι Μεγάλων Γλωσσικών Μοντέλων

Τα Μεγάλα Γλωσσικά Μοντέλα μπορούν να κατηγοριοποιηθούν σε διάφορους τύπους, ανάλογα με την εκπαίδευσή τους και τις λειτουργίες τους [42] [17]. Οι κύριοι τύποι είναι οι εξής:

- Μοντέλα Zero-shot: Αυτά τα μοντέλα εκπαιδεύονται σε ένα ευρύ και γενικευμένο σύνολο δεδομένων, γεγονός που τους επιτρέπει να αποδίδουν καλά σε ποικίλες εργασίες χωρίς επιπλέον εκπαίδευση. Δίνουν αρκετά καλές και ακριβείς απαντήσεις σε εφαρμογές και ερωτήσεις γενικής χρήσης. Ένα παράδειγμα είναι το GPT-3, το οποίο μπορεί να διαχειριστεί ένα ευρύ φάσμα ερωτήσεων και εργασιών χωρίς προσαρμογή και περαιτέρω εκπαίδευση.



Σχήμα 2.12: *Ένα βασικό διάγραμμα ροής που απεικονίζει τα διάφορα στάδια των Μεγάλων Γλωσσικών Μοντέλων (LLMs), από την προεκπαίδευση έως την προτροπή/χρήση. Η προτροπή (prompting) των LLMs για τη δημιουργία απαντήσεων είναι δυνατή σε διαφορετικά στάδια εκπαίδευσης, όπως η προεκπαίδευση, η εκπαίδευση με οδηγίες (instruction-tuning) ή η ευθυγράμμιση (alignment tuning). Ο όρος "RL" αναφέρεται στη μάθηση μέσω ενίσχυσης (reinforcement learning), το "RM" στην μοντελοποίηση ανταμοιβής (reward-modeling) και το "RLHF" στη μάθηση μέσω ενίσχυσης με ανθρώπινη ανατροφοδότηση (reinforcement learning with human feedback).*

[41]

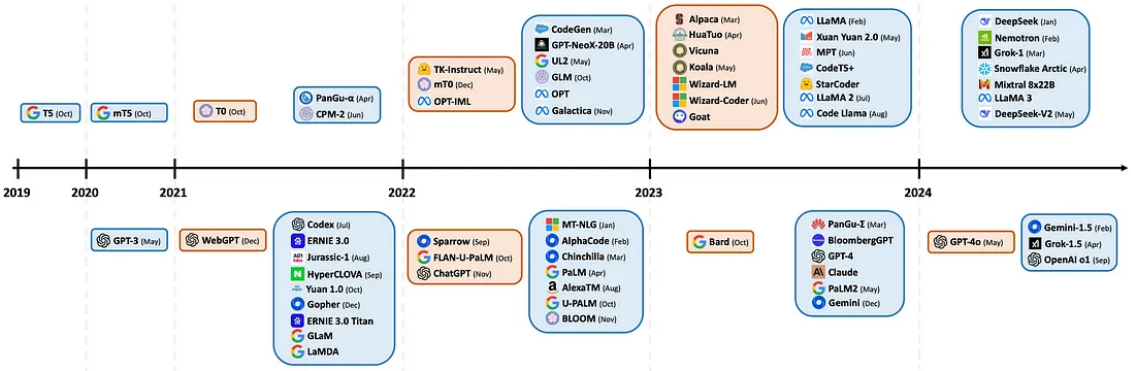
- Προσαρμοσμένα (Fine-tuned) ή Μοντέλα Ειδικού Τομέα (Domain-specific): Αυτά είναι μοντέλα zero-shot που έχουν υποβληθεί σε πρόσθετη εκπαίδευση με δεδομένα συγκεκριμένου τομέα, ώστε να εξειδικευτούν σε συγκεκριμένους κλάδους. Για παράδειγμα, το OpenAI Codex, βασισμένο στο GPT-3, είναι προσαρμοσμένο για προγραμματιστικές εργασίες, επιτρέποντάς του να δημιουργεί και να κατανοεί κώδικα με αποτελεσματικότητα.

- Μοντέλα Αναπαράστασης Γλώσσας: Αυτά τα μοντέλα επικεντρώνονται στην κατανόηση και αναπαράσταση δεδομένων κειμένου, καθιστώντας τα ιδανικά για εργασίες όπως η επεξεργασία φυσικής γλώσσας (NLP). Ένα γνωστό παράδειγμα είναι το BERT της Google, το οποίο χρησιμοποιεί βαθιά μάθηση για να εκτελεί αποτελεσματικά εργασίες κατανόησης και ανάλυσης κειμένου.

- Πολυτροπικά Μοντέλα (Multimodal Models): Σε αντίθεση με τα παραδοσιακά LLMs που επικεντρώνονται αποκλειστικά στο κείμενο, τα πολυτροπικά μοντέλα έχουν σχεδιαστεί για να επεξεργάζονται πολλαπλές μορφές δεδομένων, όπως κείμενο και εικόνες. Ένα παράδειγμα είναι το GPT-4, το οποίο μπορεί να ερμηνεύει και να δημιουργε-

ί απαντήσεις βασισμένες τόσο σε κείμενο όσο και σε οπτικά δεδομένα, διευρύνοντας σημαντικά τη χρησιμότητά του.

Αυτές οι κατηγορίες αναδεικνύουν την ευελιξία και την εξειδίκευση των LLMs, καλύπτοντας ένα ευρύ φάσμα εφαρμογών σε διάφορους τομείς.



Σχήμα 2.13: *Χρονολογική απεικόνιση της κυκλοφορίας Μεγάλων Γλωσσικών Μοντέλων (LLMs): Οι μπλε κάρτες αντιπροσωπεύουν τα ¨προεκπαιδευμένα¨ μοντέλα, ενώ οι πορτοκαλί κάρτες αντιστοιχούν σε μοντέλα που έχουν υποστεί εκπαίδευση με οδηγίες. Τα μοντέλα στο άνω μισό του διαγράμματος είναι ανοιχτού κώδικα, ενώ αυτά στο κάτω μισό είναι κλειστού κώδικα. Το διάγραμμα καταδεικνύει την αυξανόμενη τάση προς τα μοντέλα εκπαίδευσης με οδηγίες και ανοιχτού κώδικα, αναδεικνύοντας το εξελισσόμενο τοπίο και τις τάσεις στην έρευνα της επεξεργασίας φυσικής γλώσσας.*

[41]

Πλεονεκτήματα της χρήσης των Μεγάλων Γλωσσικών Μοντέλων

Η χρήση των Μεγάλων Γλωσσικών Μοντέλων προσφέρει πολλά πλεονεκτήματα, καθιστώντας τα ισχυρά εργαλεία για διάφορες εφαρμογές. Ένα από τα βασικά τους πλεονεκτήματα είναι η επεκτασιμότητα και η προσαρμοστικότητα, καθώς μπορούν να διαμορφωθούν ώστε να εξυπηρετούν συγκεκριμένες ανάγκες οργανισμών και να εξειδικευτούν για συγκεκριμένες εργασίες. Παράλληλα, διαθέτουν μεγάλη ευελιξία, αφού ένα μόνο μοντέλο μπορεί να εκτελεί διαφορετικές εργασίες, όπως σύνοψη κειμένου, μετάφραση και απαντήσεις σε ερωτήματα, γεγονός που τα καθιστά ιδιαίτερα χρήσιμα σε πολλούς τομείς.

Ένα ακόμα σημαντικό πλεονέκτημα των LLMs είναι η υψηλή απόδοση και ακρίβεια που προσφέρουν. Τα σύγχρονα μοντέλα, με δισεκατομμύρια παραμέτρους, είναι σε θέση να παρέχουν γρήγορες και ακριβείς απαντήσεις, βελτιώνοντας συνεχώς την απόδοσή τους όσο αυξάνεται ο όγκος των δεδομένων πάνω στα οποία εκπαιδεύονται. Επιπλέον, διευκολύνουν τη διαδικασία εκπαίδευσης, καθώς βασίζονται κυρίως σε μη επισημασμένα δεδομένα, μειώνοντας έτσι τον χρόνο και την προσπάθεια που απαιτούνται για την αρχική τους διαμόρφωση, ενώ ταυτόχρονα αξιοποιούν τεράστιες ποσότητες πληροφοριών.

Τέλος, τα LLMs συμβάλλουν σημαντικά στην αποδοτικότητα, καθώς αυτοματοποιούν επαναλαμβανόμενες εργασίες, εξοικονομώντας χρόνο και αυξάνοντας την παραγωγικότητα των χρηστών τους. Με αυτόν τον τρόπο, επιτρέπουν στους ανθρώπους να επικεντρωθούν σε πιο

σημαντικές και δημιουργικές δραστηριότητες, ενισχύοντας έτσι τη συνολική αποδοτικότητα και αποτελεσματικότητα σε διάφορους τομείς.

Προκλήσεις που προκύπτουν από την χρήση των Μεγάλων Γλωσσικών Μοντέλων

Παρά τα σημαντικά πλεονεκτήματα που προσφέρουν, η χρήση των Μεγάλων Γλωσσικών Μοντέλων συνοδεύεται από προκλήσεις που πρέπει να ληφθούν σοβαρά υπόψη. Μία από τις κύριες προκλήσεις είναι το υψηλό κόστος, καθώς η ανάπτυξη και η λειτουργία τους απαιτούν τεράστιους υπολογιστικούς πόρους και σημαντική οικονομική επένδυση. Αυτό περιορίζει τη δυνατότητα μικρότερων επιχειρήσεων και ερευνητικών ομάδων να τα χρησιμοποιήσουν αποτελεσματικά.

Επιπλέον, προκύπτουν σοβαρές ηθικές ανησυχίες και κίνδυνοι ασφαλείας. Τα LLMs που εκπαιδεύονται σε μη φιλτραρισμένα δεδομένα μπορεί να αναπαράγουν κοινωνικές προκαταλήψεις ή να παράγουν επιβλαβές περιεχόμενο, ενώ παράλληλα εγείρονται ζητήματα ιδιωτικότητας και προστασίας δεδομένων, καθώς οι χρήστες συχνά εισάγουν ευαίσθητες πληροφορίες. Η έλλειψη αξιοπιστίας αποτελεί επίσης σημαντικό πρόβλημα, καθώς τα LLMs μπορεί να επινοούν πληροφορίες ή να παρέχουν ανακριβείς απαντήσεις, γεγονός που υπονομεύει την εμπιστοσύνη και τη χρησιμότητά τους, ειδικά σε κρίσιμες εφαρμογές.

Μια ακόμη πρόκληση είναι η επεξηγησιμότητα και η πολυπλοκότητα αυτών των μοντέλων. Λόγω της σύνθετης αρχιτεκτονικής τους, είναι δύσκολο να κατανοηθεί πώς καταλήγουν σε συγκεκριμένα αποτελέσματα, καθιστώντας την ενσωμάτωσή τους σε υπάρχοντα συστήματα και την αντιμετώπιση πιθανών σφαλμάτων ιδιαίτερα απαιτητική. Τέλος, η συντήρηση, η αναβάθμιση και η κλιμάκωση των LLMs αποτελούν λειτουργικές προκλήσεις, καθώς το τεράστιο μέγεθός τους και η ανάγκη για συνεχή εκπαίδευση απαιτούν εξειδικευμένους πόρους και τεχνική υποδομή.

2.3.2 Εισαγωγή στην ταξινόμηση κειμένου με Μεγάλα Γλωσσικά Μοντέλα

Η ταξινόμηση κειμένου (text-based classification) αποτελεί μια από τις σημαντικότερες δυνατότητες που παρέχουν τα εργαλεία επεξεργασίας φυσικής γλώσσας (NLP). Χαρακτηριστικές εφαρμογές της είναι η ανίχνευση ανεπιθύμητων μηνυμάτων, η ανάλυση συναισθήματος και η κατηγοριοποίηση θεμάτων. Παραδοσιακά, η ταξινόμηση κειμένου βασίζεται σε μοντέλα που έχουν εκπαιδευτεί ειδικά για αυτόν τον σκοπό, απαιτώντας συχνά εκτεταμένα σύνολα δεδομένων με ετικέτες (labeled data). Ωστόσο, η έλευση των LLMs έχει διευρύνει το τοπίο των εργασιών NLP.

Τα LLMs, όπως τα GPT-4 και BERT, δεν είναι εγγενώς σχεδιασμένα για την ταξινόμηση κειμένου, αλλά διαπρέπουν στην επεξεργασία και κατανόηση σύνθετων γλωσσικών προτύπων χάρη στην εκτεταμένη εκπαίδευσής τους σε μεγάλης κλίμακας και ποικιλόμορφα δεδομένα. Αυτή η προσαρμοστικότητά τους, τους επιτρέπει να εκτελούν εργασίες ταξινόμησης με ελάχιστη προσαρμογή ή ακόμη και σε zero-shot σενάρια, όπου δεν απαιτείται περαιτέρω εξειδικευμένη εκπαίδευση. Με την αξιοποίηση της βαθιάς κατανόησης των συμφραζομένων από τα LLMs, οι χρήστες μπορούν να ταξινομήσουν αποτελεσματικά κείμενο με βάση γλωσσικά χαρακτηριστικά. Έτσι, ανοίγονται νέες δυνατότητες για τις εφαρμογές που απαιτούν την ταξινόμηση ενός όγκου δεδομένων βάσει χαρακτηριστικών κειμένου.

Συνεπώς, η διπλή εφαρμοσιμότητά τους στις προσεγγίσεις εποπτευόμενης (supervised) και μη εποπτευόμενης (unsupervised) μάθησης καθιστά τα LLMs ευέλικτα εργαλεία, ικανά να παρέχουν αποτελέσματα υψηλής ποιότητας ενώ μειώνουν την ανάγκη για εκτεταμένα δεδομένα με ετικέτες. [39]

2.4 Ανίχνευση Αυτοματοποιημένων Λογαριασμών

Σε αυτή την ενότητα δίνεται μια εισαγωγή στην λειτουργία των αυτοματοποιημένων λογαριασμών στα μέσα κοινωνικής δικτύωσης, και στη χρήση τους. Ακολούθως, γίνεται αναφορά στην έντονη παρουσία τους και τους τρόπους διαχωρισμού τους από γνήσιους χρήστες, δηλαδή ανθρώπους [43] [44] [45]. Οι αυτοματοποιημένοι λογαριασμοί ονομάζονται "Bots", και έτσι θα αναφέρονται και στη συνέχεια του κειμένου.

2.4.1 Εισαγωγή στους Αυτοματοποιημένους Λογαριασμούς

Bots στα Μέσα Κοινωνικής Δικτύωσης: Ορισμός και Χαρακτηριστικά

Τα bots στα μέσα κοινωνικής δικτύωσης (social media bots) είναι αυτοματοποιημένα προγράμματα λογισμικού σχεδιασμένα να αλληλεπιδρούν με χρήστες (users) και περιεχόμενο (content) σε πλατφόρμες κοινωνικών μέσων, όπως για παράδειγμα το Twitter. Είναι προγραμματισμένα ώστε να μιμούνται πιστά την ανθρώπινη συμπεριφορά και είναι ικανά να εκτελούν διάφορες εργασίες (tasks) όπως να ανεβάζουν δημοσιεύσεις, να κάνουν "likes" και κοινοποιήσεις, να αφήνουν σχόλια, να ακολουθούν άλλους λογαριασμούς χρηστών και να στέλνουν μηνύματα.

Ορισμένα bots εξυπηρετούν νόμιμους σκοπούς, όπως ο προγραμματισμός αναρτήσεων ή η υποστήριξη πελατών, όμως τα περισσότερα έχουν σχεδιαστεί ώστε να συμπεριφέρονται σαν ανθρώπινοι χρήστες και να παραπλανούν, έτσι, τους υπόλοιπους χρήστες του δικτύου. Τα bots λειτουργούν αυτόνομα ή ημιαυτόνομα. Επίσης, δρουν μεμονωμένα ή ως μέρος ενός μεγαλύτερου δικτύου, γνωστού ως botnet, το οποίο ελέγχεται από έναν χειριστή που αποκαλείται botmaster.

Χρήσεις των Bots στα Μέσα Κοινωνικής Δικτύωσης

Τα Bots στα μέσα κοινωνικής δικτύωσης είναι σύνθετα εργαλεία που χρησιμοποιούνται για καλούς αλλά και κακούς σκοπούς, ανάλογα με την πρόθεση του ανθρώπου-χειριστή που τα σχεδίασε. Αφενός, μπορούν να προσφέρουν χρήσιμες υπηρεσίες όπως το να στέλνουν ενημερώσεις για τον καιρό ή για αποτελέσματα αθλητικών παιχνιδιών και να παρέχουν αυτοματοποιημένη βοήθεια πελατών. Όταν τα bots χρησιμοποιούνται για καλό σκοπό, η ταυτότητα τους ως ρομπότ είναι φανερή στους χρήστες και δεν προσπαθεί να τους παραπλανήσει.

Ωστόσο, η χρήση των bots δεν περιορίζεται μόνο σε νόμιμες και διαφανείς πρακτικές, καθώς πολλά από αυτά αξιοποιούνται για κακόβουλους σκοπούς. Ένας από τους κυριότερους τρόπους με τους οποίους χρησιμοποιούνται είναι η τεχνητή αύξηση της δημοτικότητας. Τα bots μπορούν να δημιουργούν ψεύτικους ακολούθους, ψευδείς αντιδράσεις και σχόλια,

προσδίδοντας σε ορισμένους λογαριασμούς μια πλασματική επιρροή. Με αυτόν τον τρόπο, άτομα ή οργανισμοί εμφανίζονται πιο σημαντικοί και επιδραστικοί από ό,τι είναι στην πραγματικότητα, επηρεάζοντας την αντίληψη του κοινού [46].

Ένας ακόμη ανησυχητικός τρόπος αξιοποίησης των bots είναι η διάδοση παραπληροφόρησης. Μέσω της μαζικής και ταχύτατης διασποράς ψευδών ή παραπλανητικών ειδήσεων, τα bots μπορούν να επηρεάσουν την κοινή γνώμη, δημιουργώντας σύγχυση ή προωθώντας συγκεκριμένες πολιτικές ατζέντες. Αυτό το φαινόμενο έχει παρατηρηθεί ιδιαίτερα σε πολιτικές εκστρατείες και κρίσιμες κοινωνικές συζητήσεις, όπου η διαστρέβλωση των γεγονότων μπορεί να οδηγήσει σε παραπλανητικές αποφάσεις από τους πολίτες [47].

Επιπλέον, τα bots χρησιμοποιούνται για τη χειραγώγηση της κοινής γνώμης. Μέσω της δημιουργίας μεγάλου όγκου αναρτήσεων, σχολίων και επισημάνσεων ¨μου αρέσει¨ (likes), καλλιεργούν την εντύπωση ευρείας αποδοχής ή απόρριψης συγκεκριμένων ιδεών, επηρεάζοντας τις αντιλήψεις των χρηστών. Παρόμοια, συμβάλλουν στην πολιτική πόλωση, καθώς συχνά προωθούν ακραίες απόψεις και διευρύνουν τα χάσματα μεταξύ διαφορετικών ιδεολογικών ομάδων, ενισχύοντας τις κοινωνικές εντάσεις [48].

Ένας ακόμη τομέας στον οποίο τα bots διαδραματίζουν σημαντικό ρόλο είναι η χειραγώγηση του χρηματιστηρίου. Μέσω της διάδοσης ψευδών ειδήσεων και φημών για εταιρείες ή οργανισμούς, μπορούν να επηρεάσουν τις τιμές των μετοχών, δημιουργώντας κέρδη για όσους βρίσκονται πίσω από αυτά. Αυτή η πρακτική αποτελεί σοβαρή απειλή για τη σταθερότητα των χρηματοπιστωτικών αγορών και έχει προσελκύσει την προσοχή ρυθμιστικών αρχών [49].

Παράλληλα, τα bots μπορούν να χρησιμοποιηθούν για την καταστολή της ελευθερίας του λόγου. Σε πολλές περιπτώσεις, κυβερνήσεις και κρατικοί φορείς έχουν αξιοποιήσει bots για να πλημμυρίσουν τις διαδικτυακές πλατφόρμες με άσχετο ή παραπλανητικό περιεχόμενο, καταπνίγοντας έτσι τις φωνές διαμαρτυρίας και αποτρέποντας την ελεύθερη ανταλλαγή απόψεων [50].

Τέλος, τα bots εμπλέκονται σε ποικίλες μορφές απάτης. Μπορούν να διευκολύνουν επιθέσεις phishing, να διανέμουν spam ή να συμμετέχουν σε click fraud, εξαπατώντας χρήστες και εταιρείες για οικονομικό όφελος. Οι αυτοματοποιημένες αυτές δραστηριότητες υπονομεύουν την ασφάλεια του διαδικτύου και καθιστούν απαραίτητη την ανάπτυξη πιο αποτελεσματικών μεθόδων ανίχνευσης και αντιμετώπισης των κακόβουλων bots [51].

Μηχανισμός Λειτουργίας των Bots

Τα κοινωνικά bots λειτουργούν αυτοματοποιώντας εργασίες και αλληλεπιδράσεις στις πλατφόρμες κοινωνικών δικτύων με τη χρήση εξειδικευμένου λογισμικού. Αυτά τα προγράμματα επιτρέπουν στα bots να εκτελούν δραστηριότητες όπως δημοσίευση περιεχομένου, likes, κοινοποιήσεις, σχόλια, ακολούθηση χρηστών και αποστολή άμεσων μηνυμάτων. Συλλέγουν δεδομένα από πηγές όπως τα δημοφιλή θέματα της επικαιρότητας (trending topics), τα hashtags και οι αλληλεπιδράσεις χρηστών για να προσαρμόζουν τη συμπεριφορά και τις ενέργειές τους.

Για να μιμούνται την ανθρώπινη συμπεριφορά και να αποφεύγουν επιτυχώς τον εντοπισμό από τους αλγόριθμους ανίχνευσης bots των πλατφορμών, τα bots συχνά χρησιμοποιούν

τεχνικές όπως τυχαιοποίηση (randomizing) των χρόνων δημοσίευσης, διαφοροποίηση του περιεχομένου που ανεβάζουν και μίμηση των φυσικών προτύπων αλληλεπίδρασης. Μπορεί, ακόμα, μέσω προσαρμοστικών αλγόριθμους να ρυθμίζουν τη συμπεριφορά τους με βάση την ανατροφοδότηση των χρηστών ή τα δεδομένα από το περιβάλλον τους, διατηρώντας έτσι την πειστικότητα των ενεργειών τους.

Τα κοινωνικά bots συνήθως ενσωματώνονται με τα APIs (Διεπαφές Προγραμματισμού Εφαρμογών) των πλατφορμών, επιτρέποντάς τους να αλληλεπιδρούν προγραμματιστικά με λειτουργίες της πλατφόρμας, όπως η δημοσίευση περιεχομένου ή η λήψη δεδομένων. Αυτή η ενσωμάτωση ενισχύει την αποδοτικότητα και την κλιμακωσιμότητά τους, επιτρέποντάς τους να λειτουργούν ταυτόχρονα σε πολλούς λογαριασμούς ή πλατφόρμες.

Συνδυάζοντας αυτές τις στρατηγικές, τα κοινωνικά bots μπορούν να εκτελούν ένα ευρύ φάσμα δραστηριοτήτων, από την προώθηση περιεχομένου και την αύξηση των αλληλεπιδράσεων, μέχρι και πιο επιβλαβείς συμπεριφορές. [43] [44] [45]

2.4.2 Αυτοματοποιημένοι Λογαριασμοί και Πραγματικοί Χρήστες

Παρουσία Λογαριασμών Bots στα Κοινωνικά Δίκτυα

Ο ακριβής αριθμός λογαριασμών στα μέσα κοινωνικής δικτύωσης που είναι bots είναι δύσκολο να προσδιοριστεί, καθώς πολλά bots είναι σχεδιασμένα να μιμούνται πειστικά την ανθρώπινη συμπεριφορά. Έτσι, είναι δύσκολο να εντοπιστούν ακόμα και από ειδικούς. Παρ' όλα αυτά, εκτιμήσεις υποδεικνύουν ότι τα bots αποτελούν ένα σημαντικό ποσοστό χρηστών στις μεγάλες πλατφόρμες, ιδιαίτερα στο Twitter.

Για παράδειγμα, οι εκπρόσωποι του Twitter έχουν δηλώσει ότι μόνο περίπου 5% των λογαριασμών είναι bots, αλλά ορισμένες ανεξάρτητες έρευνες δίνουν μια διαφορετική εικόνα. Αυτές οι μελέτες έχουν εκτιμήσει ότι μεταξύ 15% και 68% των λογαριασμών στο Twitter μπορεί να είναι bots, ανάλογα με τα κριτήρια και τις μεθόδους που χρησιμοποιούνται. Επιπλέον, ορισμένες έρευνες δείχνουν ότι, ενώ τα bots μπορεί να αποτελούν μειοψηφία στους λογαριασμούς, συνεισφέρουν δυσανάλογα στην πλατφόρμα. Για παράδειγμα, εκτιμάται ότι τα bots είναι υπεύθυνα για 20.8% έως 29% όλων των tweets.

Αυτές οι αποκλίσεις υπογραμμίζουν τις προκλήσεις στον ακριβή εντοπισμό των bots και τονίζουν την πιθανή επιρροή που μπορούν να ασκήσουν, ανεξαρτήτως του πραγματικού αριθμού τους. Αν και τα ποσοστά μπορεί να διαφέρουν μεταξύ των πλατφορμών, το ζήτημα των bots είναι ευρέως διαδεδομένο και εγείρει σημαντικές ανησυχίες για την αυθεντικότητα των αλληλεπιδράσεων στα μέσα κοινωνικής δικτύωσης. [43] [44]

Τρόποι Διάκρισης Λογαριασμών Bots από Πραγματικούς Χρήστες

Η διάκριση ενός bot από έναν πραγματικό χρήστη στα μέσα κοινωνικής δικτύωσης μπορεί να είναι δύσκολη, ειδικά όταν πρόκειται για εξελιγμένα bots. Παρ' όλα αυτά, υπάρχουν διάφορες στρατηγικές και ενδείξεις που μπορούν να βοηθήσουν στον εντοπισμό λογαριασμών που ανήκουν σε bot.

Μία από τις βασικές ενδείξεις είναι οι πληροφορίες προφίλ. Τα bots συχνά έχουν ελλιπή ή γενικά προφίλ, με περιορισμένες πληροφορίες, απουσία φωτογραφιών ή γενικόλογες πε-

Σχήμα 2.14: *Social Media Bot (AI Generated Image)* [3]

ριγραφές. Επιπλέον, μπορεί να χρησιμοποιούν επαναλαμβανόμενες φράσεις ή να φαίνονται λιγότερο προσωπικά και συγκεκριμένα σε σχέση με τους ανθρώπινους λογαριασμούς.

Τα πρότυπα δραστηριότητας αποτελούν επίσης κρίσιμο παράγοντα αναγνώρισης. Τα bots συνήθως δημοσιεύουν περιεχόμενο σε ασυνήθιστες ώρες, χωρίς να ακολουθούν το φυσιολογικό πρόγραμμα ενός ανθρώπου. Επιπλέον, παρουσιάζουν εξαιρετικά υψηλά επίπεδα δραστηριότητας, με συχνές δημοσιεύσεις, likes ή κοινοποιήσεις μέσα σε σύντομο χρονικό διάστημα, κάτι που δεν είναι χαρακτηριστικό των πραγματικών χρηστών.

Η ποιότητα του περιεχομένου και η αλληλεπίδραση μπορούν επίσης να προσφέρουν πολύτιμες ενδείξεις. Τα bots συχνά παράγουν χαμηλής ποιότητας, μη λογικό ή τυποποιημένο περιεχόμενο που στερείται προσωπικών στοιχείων. Επίσης, σπάνια συμμετέχουν σε ουσιαστικές συνομιλίες, με τις απαντήσεις τους να φαίνονται τυποποιημένες ή να μην έχουν σχέση με το θέμα συζήτησης.

Τα συμπεριφορικά μοτίβα ενός λογαριασμού είναι ένας ακόμα δείκτης. Πολλά bots εστιάζουν στην ενίσχυση συγκεκριμένου περιεχομένου, όπως πολιτικά μηνύματα ή εμπορικά προϊόντα, με περιορισμένη ποικιλία ενδιαφερόντων. Επιπλέον, ορισμένα bots λειτουργούν μέσω χακαρισμένων λογαριασμών, οι οποίοι μπορεί να φαίνονται πιο πειστικοί λόγω της υπάρχουσας ιστορίας αναρτήσεων και συνδέσεων.

Τέλος, η χρήση εργαλείων εντοπισμού και τεχνικών τεχνητής νοημοσύνης μπορεί να ενισχύσει την αποτελεσματικότητα της αναγνώρισης bots. Οι μη αυτόματες μέθοδοι, όπως η αντίστροφη αναζήτηση εικόνων και η ανάλυση προτύπων αναρτήσεων, μπορούν να εντοπίσουν ύποπτους λογαριασμούς. Παράλληλα, εξειδικευμένες υπηρεσίες, όπως το Botcheck.me, χρησιμοποιούν αλγορίθμους μηχανικής μάθησης για την ταυτοποίηση λογαριασμών που εμφανίζουν τα χαρακτηριστικά ενός bot. Επιπλέον, προηγμένα συστήματα τεχνητής νοημοσύνης αναλύουν μοτίβα αλληλεπίδρασης, κινήσεις ποντικιού και τη διάρκεια των συνεδριών

σύνδεσης, ώστε να εντοπίσουν αποκλίσεις που υποδηλώνουν αυτοματοποιημένη δραστηριότητα.

Συνδυάζοντας όλες αυτές τις προσεγγίσεις, τόσο οι χρήστες όσο και οι διαδικτυακές πλατφόρμες μπορούν να βελτιώσουν την ικανότητά τους να ξεχωρίζουν τα bots από τους πραγματικούς λογαριασμούς, διασφαλίζοντας πιο αυθεντικές αλληλεπιδράσεις και ασφαλέστερη επικοινωνία στα μέσα κοινωνικής δικτύωσης.

Κεφάλαιο 3

Συναφής Έρευνα και Βιβλιογραφία

Σε αυτό το κεφάλαιο παρουσιάζεται μια βιβλιογραφική ανασκόπηση των ήδη υπαρχόντων ερευνών που έχουν γίνει στον τομέα της ανίχνευσης και αναγνώρισης αυτοματοποιημένων λογαριασμών bot στα κοινωνικά δίκτυα, και ειδικά στο Twitter. Γίνεται αναφορά στις μεθόδους, τις τεχνικές, τις μετρικές αξιολόγησης και τα σύνολα δεδομένων που αξιοποιήθηκαν. Οι σχετικές έρευνες κατηγοριοποιούνται ανάλογα με το είδος της προσέγγισης που υιοθέτησαν και των μοντέλων που επιστράτευσαν. Συγκεκριμένα, το κεφάλαιο χωρίζεται σε τέσσερις ενότητες που περιλαμβάνουν τις έρευνες που αξιοποιούν κλασικούς αλγορίθμους μηχανικής μάθησης, αυτές που χρησιμοποιούν συνελικτικά δίκτυα ή δίκτυα με προσοχή, καθώς και εκείνες που έχουν υβριδική προσέγγιση ή που επιστρατεύουν γλωσσικά μοντέλα.

3.1 Ανίχνευση Bots με Κλασικούς Αλγορίθμους Μηχανικής Μάθησης

Οι Ramalingaiah et al. [52] χρησιμοποίησαν το σύνολο δεδομένων "Detecting Twitter Bot Data" [53] από την πλατφόρμα Kaggle, το οποίο περιλαμβάνει χαρακτηριστικά λογαριασμών χρηστών, για την ανίχνευση bots. Εφάρμοσαν διάφορες μεθόδους εξαγωγής χαρακτηριστικών, όπου αξιοποιήθηκε το μοντέλο Bag of Bots' Words. Στη συνέχεια, δοκίμασαν ταξινομητές όπως Decision Trees, Logistic Regression, K Nearest Neighbors (KNN), και Naïve Bayes, καθώς και έναν νέο ταξινομητή που προτάθηκε στην εργασία. Για την αξιολόγηση, χρησιμοποιήθηκαν οι μετρικές Accuracy, ROC, και AUC, με τα Decision Trees να επιτυγχάνουν την καλύτερη ακρίβεια. Τα μοντέλα Logistic Regression και Naïve Bayes είχαν χαμηλότερη απόδοση λόγω υπερεκπαίδευσης και λανθασμένων υποθέσεων ανεξαρτησίας μεταξύ χαρακτηριστικών.

Οι Fagni et al. [54] αξιοποιώντας μόνο τα tweets και τα χαρακτηριστικά που προκύπτουν από το περιεχόμενό τους, εξερεύνησαν την αποτελεσματικότητα διαφόρων τεχνικών Μηχανικής Μάθησης και Βαθιάς Μάθησης μέσω τεσσάρων προσεγγίσεων. Στο πλαίσιο της πρώτης προσέγγισης, χρησιμοποίησαν τις μεθόδους Bag of Words (BoW) και Term Frequency-Inverse Document Frequency (TF-IDF), σε συνδυασμό με ταξινομητές όπως Logistic Regression, Random Forest, και Support Vector Machine (SVM). Οι μετρικές αξιολόγησης περιλάμβαναν τα Precision, Recall, F1-Score, και Accuracy. Για τα πειράματα, χρησιμοποιήθηκε το TweepFake, ένα σύνολο δεδομένων που συγκεντρώθηκε από τους ίδιους τους

συγγραφείς και περιλάμβανε 25.572 tweets (50% από bots και 50% από ανθρώπους). Η εργασία αυτή όχι μόνο εντόπισε προκλήσεις που συναντά κανείς κατά την ανίχνευση deepfake tweets, αλλά επίσης αξιολόγησε 13 μεθόδους ανίχνευσης, δημιουργώντας μια ισχυρή βάση για μελλοντική έρευνα στην αναγνώριση deepfake περιεχομένου στα κοινωνικά δίκτυα.

Οι Heidari et al. [55] παρουσίασαν μια καινοτόμο προσέγγιση για την ανίχνευση bots στα κοινωνικά δίκτυα, επικεντρώνοντας την έρευνά τους στα χαρακτηριστικά συναισθήματος των tweets. Η μέθοδος περιλαμβάνει την εξαγωγή νέων χαρακτηριστικών που βασίζονται σε ψυχολογικές και κοινωνικές πτυχές του κειμένου. Αυτά τα χαρακτηριστικά περιλαμβάνουν το πλήθος θετικών, αρνητικών και ουδέτερων tweets ενός χρήστη, το άθροισμα και τον μέσο όρο της πολικότητας (θετικής και αρνητικής) των tweets, και αξιολογήθηκαν χρησιμοποιώντας μοντέλα όπως Random Forest, Support Vector Machine (SVM), Logistic Regression, και ένα Feedforward Neural Network. Η ανάλυση έγινε τόσο σε λογαριασμό χρήστη (account level) όσο και σε ομαδικό επίπεδο (group level). Τα πειράματα έδειξαν ότι η ενσωμάτωση αυτών των χαρακτηριστικών βελτίωσε σημαντικά την ακρίβεια των μοντέλων.

Οι Dukić et al. [56] πρότειναν μια μέθοδο ανίχνευσης bots στο Twitter, εστιάζοντας στην επεξεργασία των tweets μέσω διαφορετικών προσεγγίσεων εξαγωγής χαρακτηριστικών. Ανέπτυξαν δύο μοντέλα: ένα μοντέλο Logistic Regression και ένα Feedforward Deep Neural Network, χρησιμοποιώντας word2vec embeddings για την αναπαράσταση των δεδομένων. Το σύνολο δεδομένων PAN-19, το οποίο περιέχει tweets στα αγγλικά, αποτέλεσε τη βάση της ανάλυσης, ενώ ως μετρική αξιολόγησης χρησιμοποιήθηκε το Weighted F1-Score. Η ανάλυση των αποτελεσμάτων υποδεικνύει ότι τα επιφανειακά μοντέλα, όπως η Λογιστική Παλινδρόμηση, αποδίδουν καλύτερα από τα βαθιά νευρωνικά δίκτυα σε αυτό το πρόβλημα, καθιστώντας τα πιο αποτελεσματικά και λιγότερο περίπλοκα. Στα μελλοντικά σχέδια των ερευνητών περιλαμβάνεται η περαιτέρω βελτίωση της απόδοσης μέσω προσαρμογής των παραμέτρων του BERT και η ενσωμάτωση συναισθηματικής ανάλυσης (sentiment analysis) ως επιπλέον χαρακτηριστικό.

Οι Ilias και Roussaki [57] ανέπτυξαν δύο μεθόδους για την ανίχνευση bots στο Twitter, με κύριο άξονα την Επεξεργασία Φυσικής Γλώσσας (NLP). Στην πρώτη μέθοδο, εστίασαν στην εξαγωγή 71 χαρακτηριστικών από τους λογαριασμούς χρηστών. Στη συνέχεια, χρησιμοποιώντας τεχνικές επιλογής χαρακτηριστικών, όπως οι Mutual Information, ANOVA F-Value και Chi-squared test, επέλεξαν το πιο σχετικό υποσύνολο χαρακτηριστικών. Για την αντιμετώπιση ανισορροπιών στα δεδομένα, εφάρμοσαν κατάλληλες τεχνικές εξισορρόπησης και στη συνέχεια εκπαίδευσαν μοντέλα μηχανικής μάθησης όπως Support Vector Machine, Decision Trees, Random Forests, AdaBoost, Logistic Regression και KNN. Τα αποτελέσματα αξιολογήθηκαν με μετρικές όπως Precision, Recall, F1-Score, AUROC και Accuracy, χρησιμοποιώντας τα σύνολα δεδομένων Cresci-2017 και Social Honeypot Dataset. Η μελέτη τους υπογραμμίζει τη σημασία της προσεκτικής επιλογής χαρακτηριστικών για την ανίχνευση bots και εισάγει μια δεύτερη μέθοδο που περιλαμβάνει αρχιτεκτονική βαθιάς μάθησης με μηχανισμό προσοχής. Αυτή η προσέγγιση, η οποία εφαρμόστηκε για πρώτη φορά για τον εντοπισμό bots, προσφέρει πλεονεκτήματα συγκριτικά με υπάρχουσες μεθόδους, απο-

δεικνύοντας την αποτελεσματικότητά της μέσω πειραμάτων σε δύο μεγάλα datasets.

Οι Fonseca Abreu et al. [58] διερεύνησαν την ανίχνευση bots στο Twitter χρησιμοποιώντας δεδομένα από τους Cresci et al. Εφαρμόζοντας τεχνικές επιλογής χαρακτηριστικών, επικεντρώθηκαν στα πιο σημαντικά στοιχεία των λογαριασμών και αξιολόγησαν την απόδοση τεσσάρων αλγορίθμων μηχανικής μάθησης: Random Forest, Support Vector Machine (SVM), Naive Bayes και One-Class SVM (ocSVM). Οι μετρικές αξιολόγησης περιλάμβαναν Accuracy, Recall, AUC και F1-Score. Η ανάλυση έδειξε ότι όλοι οι κλασικοί ταξινομητές πέτυχαν υψηλές επιδόσεις, με AUC μεγαλύτερο από 0.9, υποδεικνύοντας τη χρησιμότητά τους για την ανίχνευση bots. Η έρευνα καταλήγει ότι μια περιορισμένη σειρά χαρακτηριστικών μπορεί να είναι εξαιρετικά αποτελεσματική για τον εντοπισμό bots. Αυτό το αποτέλεσμα ενισχύει την προοπτική χρήσης απλών χαρακτηριστικών για την ανάπτυξη αποδοτικών μοντέλων μηχανικής μάθησης.

Οι Rodrigues et al. [59] παρουσίασαν μια μέθοδο για τον εντοπισμό των spam tweets και την ανάλυση συναισθημάτων τους, βασισμένη σε διαφορετικά σύνολα δεδομένων και τεχνικές επεξεργασίας φυσικής γλώσσας. Για την ανίχνευση spam tweets, χρησιμοποίησαν δεδομένα SMS που περιλαμβάνουν 4.825 "human" και 747 "spam" tweets, ενώ για την ανάλυση συναισθημάτων χρησιμοποίησαν ένα μεγαλύτερο σύνολο δεδομένων από την πλατφόρμα Kaggle, με 31.015 tweets διαχωρισμένα σε ουδέτερα, θετικά και αρνητικά συναισθήματα. Η επεξεργασία περιλάμβανε στάδια όπως φιλτράρισμα, λεκτική ανάλυση, αφαίρεση stop words, αποκατάληξη και λημματοποίηση, ενώ τα χαρακτηριστικά εξήχθησαν μέσω τεχνικών TF-IDF και Bag of Words. Οι ταξινομητές όπως Decision Tree, Logistic Regression και Random Forest χρησιμοποιήθηκαν για την ανίχνευση των spam tweets, ενώ μέθοδοι όπως Support Vector Machine (SVM) και Naive Bayes εφαρμόστηκαν για την ανάλυση συναισθημάτων. Επιπλέον, αξιολογήθηκαν μοντέλα βαθιάς μάθησης, όπως LSTM και CNN, με το LSTM να πετυχαίνει ακρίβεια 98.74% για spam detection και 73.81% για sentiment analysis. Οι μελλοντικές κατευθύνσεις περιλαμβάνουν τη μελέτη της σχέσης μεταξύ των λογαριασμών και της συχνότητας δημοσίευσης spam tweets, καθώς και την ανάλυση της αναλογίας followers προς followings για περαιτέρω ενδείξεις περί spam δραστηριότητας.

Ο Knauth [60] εστίασε στον εντοπισμό bots στο Twitter μέσω της ανάλυσης ενός συνόλου δεδομένων με 8.385 λογαριασμούς, περιλαμβάνοντας ανθρώπινους και αυτοματοποιημένους λογαριασμούς. Τα χαρακτηριστικά που χρησιμοποιήθηκαν ανήκαν σε τρεις βασικές κατηγορίες: χαρακτηριστικά λογαριασμού, χαρακτηριστικά tweets και συμπεριφορικά χαρακτηριστικά. Η έμφαση δόθηκε σε γλωσσικά ανεξάρτητα χαρακτηριστικά, ώστε να διευκολυνθεί η μελλοντική επαναχρησιμοποίηση του συστήματος. Για την αντιμετώπιση ανισορροπιών στα δεδομένα χρησιμοποιήθηκε η μέθοδος SMOTE-ENN. Στη συνέχεια, τα χαρακτηριστικά αυτά τροφοδοτήθηκαν σε διάφορα μοντέλα μηχανικής μάθησης, όπως Logistic Regression, Support Vector Machine, Random Forest, Multi-Layer Perceptron (MLP), με το AdaBoost να πετυχαίνει την καλύτερη απόδοση, σημειώνοντας ακρίβεια 0.988 και AUC-ROC 0.995. Οι καμπύλες εκμάθησης που αναλύθηκαν έδειξαν ότι ακόμα και μικρότερα σύνολα δεδομένων μπορούν να οδηγήσουν σε αποδοτικά αποτελέσματα, κάνοντας το σύστημα ιδιαίτερα

πρακτικό για εφαρμογές με περιορισμένα δεδομένα.

3.2 Ανίχνευση Bots με Graph Convolutional Networks και Graph Attention Networks

Οι Liu et al. [61] υλοποίησαν το SEGCN (Subgraph Encoding Graph Convolutional Network), που αποτελεί μια καινοτόμο προσέγγιση για την ανίχνευση κοινωνικών bots, βελτιώνοντας την εκφραστική ισχύ των γραφικών συνελικτικών δικτύων (GCNs) μέσω της χρήσης κωδικοποίησης υπογραφημάτων αντί της παραδοσιακής χρήσης άμεσων γειτόνων. Το μοντέλο συνδυάζει ποικίλα χαρακτηριστικά, όπως πληροφορίες από το περιεχόμενο, τη συμπεριφορά και τις σχέσεις των χρηστών, και δοκιμάστηκε στα δημόσια σύνολα δεδομένων Twibot-20 και Twibot-22, όπου παρουσίασε βελτίωση της ακρίβειας κατά 2.4% και 3.1%, αντίστοιχα, συγκρινόμενο με τις μεθόδους αιχμής. Παρά τους περιορισμούς που θέτει η περιορισμένη πρόσβαση σε δεδομένα σε πλατφόρμες όπως το Twitter (X), το SEGCN προτείνει ένα γενικευμένο και αξιόπιστο πλαίσιο ανίχνευσης bots, ανοίγοντας τον δρόμο για μελλοντική έρευνα σε ετερογενή γραφήματα που περιλαμβάνουν πολλαπλούς τύπους σχέσεων χρηστών.

Οι Zhau et al. [62] προτείνουν ένα ημι-επιβλεπόμενο μοντέλο βασισμένο σε δίκτυο γραφημάτων με προσοχή (Graph Attention Network - GAT) για την ανίχνευση spam bots σε κοινωνικά δίκτυα. Το μοντέλο ενσωματώνει χαρακτηριστικά χρηστών και γειτονικές σχέσεις σε ένα κατευθυνόμενο γράφημα, χρησιμοποιώντας τόσο μεθόδους βασισμένες σε χαρακτηριστικά όσο και μεθόδους διάδοσης πληροφορίας. Συγκεκριμένα, συνδυάζονται δύο διαφορετικοί τύποι σχέσεων μεταξύ χρηστών, οι σχέσεις παρακολούθησης (following) και αναδημοσίευσης (retweeting), για τη βελτίωση της ακρίβειας. Η χρήση ενός βελτιωμένου μηχανισμού προσοχής επιτρέπει στο μοντέλο να αναλύει τις διαφορετικές αλληλεπιδράσεις μεταξύ κόμβων και να εξάγει πλούσιες πληροφορίες για τον εντοπισμό ανωμαλιών. Τα πειράματα με πραγματικά δεδομένα από το Twitter επιβεβαιώνουν την αποδοτικότητα και τη σταθερότητα του προτεινόμενου μοντέλου στην ανίχνευση spam bots. Στο μέλλον, προτείνεται η διερεύνηση πλουσιότερων αναπαραστάσεων χαρακτηριστικών και η αξιοποίηση περισσότερων δομικών και σημασιολογικών δεδομένων για περαιτέρω βελτίωση της απόδοσης.

Οι Feng et al. [63] προτείνουν το BotRGCN, ένα πλαίσιο ανίχνευσης bots στο Twitter βασισμένο σε Σχεσιακά Γραφικά Συνελικτικά Δίκτυα (Relational Graph Convolutional Networks - RGCNs), σχεδιασμένο για να αντιμετωπίζει δύο βασικές προκλήσεις: τη συλλογική δράση των bots (bot communities) και την ικανότητά τους να φαίνονται σαν πραγματικοί χρήστες. Το μοντέλο κατασκευάζει ένα ετερογενές γράφημα από τις σχέσεις παρακολούθησης (follow relationships) στο Twitter, ενσωματώνοντας πολυτροπικές πληροφορίες χρηστών (π.χ. χαρακτηριστικά λογαριασμού και σημασιολογικά δεδομένα) για να αποφύγει την παραδοσιακή εξόρυξη χαρακτηριστικών (feature engineering) και να ενισχύσει την ικανότητά του να αναγνωρίζει διαφορετικούς τύπους bots. Τα πειράματα σε ένα ολοκληρωμένο σύνολο δεδομένων, το TwiBot-20, αποδεικνύουν ότι το BotRGCN υπερέχει των ανταγωνιστικών μεθόδων, καταδεικνύοντας την αποτελεσματικότητα της στρατηγικής κωδικοποίησης πλη-

ροφοριών χρηστών και της μεθόδου γραφημάτων του μοντέλου. Το BotRGCN αποτελεί ένα ολοκληρωμένο πλαίσιο για την ανίχνευση bots, με έμφαση στην προσαρμογή σε πραγματικές προκλήσεις των κοινωνικών δικτύων.

Οι Shi et al. [64] παρουσιάζουν το MSGS (Multi-scale Graph Neural Network with Signed-Attention), ένα καινοτόμο μοντέλο ανίχνευσης κοινωνικών bots που αξιοποιεί πληροφορίες από χαμηλές και υψηλές συχνότητες στα κοινωνικά δίκτυα, βελτιώνοντας την αναπαραστατική ικανότητα των γραφημάτων. Το μοντέλο ενσωματώνει μια πολυκλιμακωτή δομή για τη δημιουργία αναπαραστάσεων σε διαφορετικές κλίμακες και χρησιμοποιεί έναν μηχανισμό υπογεγραμμένης προσοχής (signed-attention) για τη σύνθεση αυτών των αναπαραστάσεων. Επιπλέον, μέσω πολυεπίπεδων νευρωνικών δικτύων (MLP), παράγεται το τελικό αποτέλεσμα. Η θεωρητική ανάλυση από την οπτική της συχνότητας δείχνει ότι το MSGS επεκτείνει το φάσμα ρύθμισης της συχνότητας σε σύγκριση με τα υπάρχοντα φίλτρα γραφημάτων και αντιμετωπίζει αποτελεσματικά το πρόβλημα της υπερ-ομαλοποίησης που παρατηρείται σε βαθιές δομές GNN. Πειράματα σε πραγματικά δεδομένα επιβεβαιώνουν ότι το MSGS υπερέχει σταθερά έναντι κορυφαίων μεθόδων ανίχνευσης bots, προσφέροντας ένα ευέλικτο και προσαρμοστικό πλαίσιο για την ανάλυση κοινωνικών γραφημάτων.

Οι Kalam et al. [65] εστιάζουν στην ανίχνευση bots στο Twitter, αναδεικνύοντας τη σημασία της αντιμετώπισης της παραπληροφόρησης και των κυβερνοαπειλών που προκαλούνται από αυτά. Χρησιμοποιώντας γραφήματα και βαθιά μάθηση μέσω Graph Neural Networks (GNNs), δημιούργησαν ένα δυναμικό γραφικό μοντέλο του Twitter, στο οποίο οι αλληλεπιδράσεις χρηστών αναπαρίστανται ως ακμές. Το γραφικό μοντέλο εμπλουτίστηκε με δεδομένα χρηστών, όπως η διάρκεια λειτουργίας των λογαριασμών και η δραστηριότητά τους, για να παρέχει μια ολοκληρωμένη εικόνα των σχέσεων και συμπεριφορών. Η προσαρμοσμένη αρχιτεκτονική GNN που αναπτύχθηκε βασίζεται σε υπάρχοντα μοντέλα, όπως το BotRGCN και το BIC, για την εκμάθηση ενσωμάτωσης χρηστών και τον εντοπισμό μοτίβων που χαρακτηρίζουν τα bots. Η μέθοδος αξιοποιεί την ανάλυση δικτύων, τη μηχανική μάθηση και τις αναλύσεις κοινωνικών μέσων, συμβάλλοντας σημαντικά στη διασφάλιση της ακεραιότητας της πλατφόρμας και στη μείωση των επιπτώσεων από bots που διαδίδουν παραπληροφόρηση.

3.3 Ανίχνευση Bots με Υβριδικές Μεθόδους

Οι Hayawi et al. [66] προτείνουν το DeeProBot, ένα υβριδικό μοντέλο βαθιάς μάθησης για την ανίχνευση bots στο Twitter, βασισμένο σε μεταδεδομένα από τα προφίλ χρηστών. Χρησιμοποιεί χαρακτηριστικά όπως η περιγραφή του προφίλ, ο αριθμός των ακολούθων και ο αριθμός των tweets, για να διακρίνει τους ανθρώπινους λογαριασμούς από τα bots. Ιδιαίτερη καινοτομία του DeeProBot είναι η αξιοποίηση του κειμένου από το πεδίο περιγραφής του προφίλ, το οποίο ενσωματώνεται μέσω του GLoVe (Global Vectors for Word Representation). Το μοντέλο επεξεργάζεται συνδυασμένα χαρακτηριστικά (αριθμητικά, δυαδικά και κειμενικά) με υβριδική αρχιτεκτονική, χρησιμοποιώντας μονάδες LSTM για την ανάλυση του κειμένου και πυκνές στρώσεις για τα υπόλοιπα χαρακτηριστικά. Το DeeProBot επιτυγχάνει υψηλή απόδοση με AUC 0,97 στις δοκιμές, δείχνοντας την αποτελεσματικότητά του στην ανίχνευση

bots, ενώ διασφαλίζει την γενικευσιμότητά του μέσω της επιλογής χαρακτηριστικών και της ενσωμάτωσης τακτικών για την αποφυγή υπερβολικής προσαρμογής (overfitting). Στο μέλλον, το μοντέλο προγραμματίζεται να δοκιμαστεί με δεδομένα σε πραγματικό χρόνο και να εξελιχθεί για να αναγνωρίζει νέες συμπεριφορές bots.

Οι Ellaky et al. [67] προτείνουν μια νέα υβριδική αρχιτεκτονική για την ανίχνευση bots στα κοινωνικά δίκτυα, βασισμένη στην αναπαράσταση λέξεων μέσω GloVe και σε επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs). Η προσέγγιση συνδυάζει τα Bidirectional Gated Recurrent Units (BiGRU) και τα Long Short-Term Memory (LSTM) δίκτυα για την ταξινόμηση κειμένων που παράγονται από τα tweets. Η αρχιτεκτονική εκπαιδεύτηκε με τα σύνολα δεδομένων Cresci-2017 και Twibot-20 και αξιολογήθηκε με βάση τέσσερις μετρικές: Ακρίβεια, Ανάκληση, F1-Score, και Precision. Η μέθοδος πέτυχε εξαιρετικά αποτελέσματα στην ανίχνευση bots μόνο από το κείμενο των tweets, με Precision 100%, Accuracy 99,73%, Recall 99,56%, και F1-Score 99,63% χρησιμοποιώντας το σύνολο δεδομένων Twibot-20. Η προτεινόμενη αρχιτεκτονική υπερέβη τα αποτελέσματα άλλων προηγμένων μεθόδων και απέδειξε την ανθεκτικότητά της στο overfitting, ενώ είναι αποτελεσματική στην ανίχνευση bots σε νέα και αόρατα δεδομένα. Αυτή η μελέτη αναδεικνύει τη σημασία των μεθόδων βαθιάς μάθησης και της αναπαράστασης λέξεων για την αποτελεσματική ανίχνευση και διαχείριση των bots στα κοινωνικά δίκτυα.

Οι Vitkova et al. [68] παρουσιάζουν μια υβριδική προσέγγιση για την ανάλυση κοινωνικών δικτύων με σκοπό την ανίχνευση ύποπτων προφίλ χρηστών. Η προσέγγιση συνδυάζει στατιστικές τεχνικές, εξόρυξη δεδομένων και οπτική ανάλυση, χρησιμοποιώντας περιορισμένα δεδομένα κοινωνικών δικτύων, όπως "likes" σε ομάδες και συνδέσεις μεταξύ χρηστών, που συνήθως είναι διαθέσιμα σε ανοιχτή πρόσβαση. Το πλεονέκτημα αυτής της μεθόδου είναι η δυνατότητά της να επεξεργάζεται δεδομένα μεγάλης κλίμακας χωρίς να απαιτεί εκτεταμένα δεδομένα από τα προφίλ των χρηστών. Στα πειράματα που πραγματοποιήθηκαν, επιβεβαιώθηκε η αποτελεσματικότητα της προτεινόμενης προσέγγισης στην ανίχνευση ύποπτων χρηστών. Η υβριδική ανάλυση επιτρέπει τη συλλογή πιο ετερογενών πληροφοριών για τα προφίλ χρηστών και ενδυναμώνει τη δυνατότητα αναγνώρισης αναρτήσεων ή χρηστών που διαδίδουν ακατάλληλο περιεχόμενο. Αν και οι παραδοσιακές μέθοδοι ανάλυσης δεν επαρκούν για να ανιχνεύσουν τέτοιους χρήστες λόγω του όγκου των δεδομένων, η προτεινόμενη προσέγγιση προσφέρει νέα χαρακτηριστικά για τη λήψη μέτρων προστασίας των χρηστών. Στο μέλλον, η μεθοδολογία θα επεκταθεί για να αναλύσει τις ροές πληροφορίας, με προοπτική συνδυασμού αυτής της προσέγγισης με άλλες που επικεντρώνονται στην ανάλυση ροών πληροφορίας.

3.4 Ανίχνευση Bots με Γλωσσικά Μοντέλα

Οι Heidari και Jones [69] ανέπτυξαν ένα μοντέλο ανίχνευσης bots που χρησιμοποιεί τον αλγόριθμο BERT για την ταξινόμηση συναισθημάτων (θετικά, αρνητικά, ουδέτερα) από tweets. Το μοντέλο διακρίνεται για τη χρήση topic-independent χαρακτηριστικών, επιτρέποντας ανίχνευση χωρίς εξαρτήσεις από δικτυακά δεδομένα. Επίσης, δημιούργησαν νέο

σύνολο δεδομένων με 4 εκατομμύρια tweets, βασισμένο στην ταξινόμηση συναισθημάτων με BERT. Το μοντέλο πέτυχε ακρίβεια 94% στην ανίχνευση bots στο σύνολο δεδομένων Cresci, ξεπερνώντας προηγούμενες μεθόδους. Εφαρμόζει μεταφορά γνώσης από άλλες πλατφόρμες μέσω labeled δεδομένων για βελτίωση των επιδόσεων.

Οι Garcia-Silva, Berrio και Gomez-Perez [70] μελέτησαν τη χρήση προεκπαιδευμένων μοντέλων γλώσσας όπως το GPT-2 και το BERT για την ανίχνευση bots στο Twitter, εστιάζοντας στο πώς επηρεάζονται οι εσωτερικές αναπαραστάσεις τους μέσω fine-tuning. Ανακάλυψαν ότι το GPT-2 έχει καλύτερη ακρίβεια από το BERT στην αναγνώριση περιεχομένου bots, διατηρώντας περισσότερες γραμματικές πληροφορίες (στοιχεία όπως η σύνταξη και η μορφολογία, π.χ. η δομή μιας πρότασης ή η χρήση της γλώσσας) και αποδίδοντας καλύτερα σε οντότητες όπως hashtags και URLs. Δηλαδή κατάφερε να αναγνωρίσει καλύτερα τα μοτίβα χρήσης αυτών των οντοτήτων (π.χ. πώς τα bots χρησιμοποιούν hashtags διαφορετικά από ανθρώπους). Το GPT-2 παρουσιάζει μεγαλύτερη σταθερότητα και καλύτερη διάκριση των αναπαραστάσεων μετά το fine-tuning, κάνοντας το πιο κατάλληλο για δεδομένα Twitter.

Οι Tourille, Sow και Popescu [71] μελέτησαν την αυτόματη ανίχνευση bot-generated tweets, επικεντρωμένοι στην πρόκληση των τελευταίων γλωσσικών μοντέλων και της μικρής έκτασης των tweets. Προσδιόρισαν το πρόβλημα χωρίς να κάνουν υποθέσεις για τον λογαριασμό ή το δίκτυο του bot, χρησιμοποιώντας δύο προσεγγίσεις βασισμένες σε προεκπαιδευμένα γλωσσικά μοντέλα. Χρησιμοποίησαν το μοντέλο RoBERTa για την ανίχνευση bot-generated tweets, επιτυγχάνοντας συνολική ακρίβεια 91.3% με F1-scores 0.912 για τους ανθρώπους και 0.914 για τα bots. Η απόδοση βελτιώθηκε στη διάκριση των tweets που δημιουργήθηκαν με GPT-2 (0.826), αν και παρέμεινε χαμηλότερη από άλλες μεθόδους όπως τα RNNs. Δοκίμασαν δύο προσεγγίσεις αυτόματης ενίσχυσης δεδομένων, οι οποίες βελτίωσαν τις επιδόσεις για ορισμένα σύνολα δεδομένων. Η γενίκευση παραμένει πρόκληση, καθώς η ακρίβεια κυμαίνεται από 61.5% έως 93.9% ανάλογα με την τυχαία κατανομή του συνόλου δεδομένων.

Οι Sallah, Alaoui και Agoujil [72] παρουσιάζουν μια μέθοδο για την ανίχνευση bots στο Twitter, χρησιμοποιώντας προεκπαιδευμένα μοντέλα γλώσσας (PLMs) και τεχνικές εξήγησης. Η προσέγγιση περιλαμβάνει τη χρήση μοντέλων BERT, GPT-3 και RoBERTa για την απόκτηση πλούσιων αναπαραστάσεων των tweets, οι οποίες τροφοδοτούνται σε ένα βαθύ νευρωνικό δίκτυο για τη διάκριση ανθρωπίνων χρηστών από bots. Τα αποτελέσματα δείχνουν σημαντική βελτίωση στην ανίχνευση bots με F1-score 93%, υπερβαίνοντας παραδοσιακές μεθόδους όπως τα Word2Vec και GloVe. Η μελέτη επίσης εξετάζει τη χρήση τεχνικών τεχνητής νοημοσύνης για την ενίσχυση της διαφάνειας των μοντέλων.

Κεφάλαιο 4

Σύνολο Δεδομένων

Στο Κεφάλαιο αυτό παρουσιάζεται αναλυτικά το Σύνολο Δεδομένων (Dataset) το οποίο χρησιμοποιήθηκε για την εκπόνηση αυτής της διπλωματικής, καθώς και η προεπεξεργασία στην οποία υποβλήθηκαν τα υποσύνολα δεδομένων που το ίδιο εμπεριέχει, για να προκύψουν εν τέλει τα τελικά σύνολα δεδομένων που επιστρατεύτηκαν για την εκπαίδευση και αξιολόγηση των προτεινόμενων μοντέλων, τα οποία θα παρουσιαστούν στα επόμενα κεφάλαια.

4.1 Περιγραφή Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στα πλαίσια της παρούσας διπλωματικής ονομάζεται Twibot-20 και παρουσιάστηκε στο επιστημονικό άρθρο με τίτλο "TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark" [73]. Το Twibot-20 είναι ένα από τα μεγαλύτερα και πιο ολοκληρωμένα σύνολα δεδομένων που έχουν αναπτυχθεί για την ανίχνευση bots στο Twitter. Δημιουργήθηκε για να καλύψει ελλείψεις σε υπάρχοντα datasets, όπως η χαμηλή ποικιλομορφία χρηστών, η περιορισμένη πληροφορία ανά χρήστη και η έλλειψη δεδομένων σε μεγάλη κλίμακα. Το Twibot-20 περιλαμβάνει 229.573 χρήστες, 33.488.192 tweets, 8.723.736 στοιχεία χρηστών (properties) και 455.958 σχέσεις ακολουθίας (follow relationships). Στόχος του είναι να παρέχει μια ποικιλόμορφη αναπαράσταση της τρέχουσας κατάστασης της σφαίρας του Twitter (Twittersphere). Πρόκειται, λοιπόν, για ένα dataset που προσφέρει δεδομένα υψηλής ποιότητας, σχεδιασμένα να υποστηρίζουν τόσο την ταξινόμηση μεμονωμένων χρηστών όσο και προσεγγίσεις ανάλυσης κοινοτήτων.

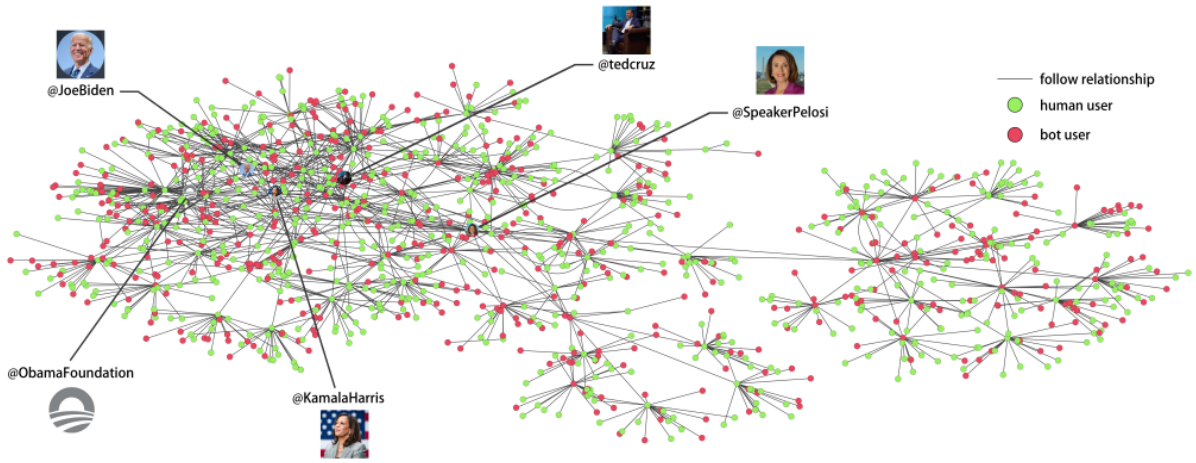
Συλλογή και Οργάνωση Δεδομένων

Το Twibot-20 περιλαμβάνει τέσσερα κύρια αρχεία: train.json, dev.json, test.json, και support.json, τα οποία περιέχουν δεδομένα χρηστών σε μορφή JSON. Αυτή η μορφοποίηση διευκολύνει την επεξεργασία των αρχείων με σύγχρονα εργαλεία ανάλυσης δεδομένων. Το train.json χρησιμοποιείται για την εκπαίδευση μοντέλων, το dev.json για την αξιολόγηση της απόδοσης των μοντέλων κατά τη διάρκεια της εκπαίδευσης και το test.json για την τελική αξιολόγηση. Το support.json περιέχει χρήστες που δεν συμπεριλαμβάνονται άμεσα στα υπόλοιπα αρχεία, αλλά προσφέρουν χρήσιμες πληροφορίες για semi-supervised μάθηση. Τέλος, υπάρχει και το αρχείο TwiBot-20_Seed_Users.txt , το οποίο απαριθμεί τους seed users (αρχικούς χρήστες) που χρησιμοποιήθηκαν για την έναρξη της συλλογής δεδομένων.

Βασική Μορφή Εγγραφών στο Twibot-20

Μια τυπική εγγραφή στο dataset περιλαμβάνει:

- ID: Το μοναδικό αναγνωριστικό χρήστη στο Twitter.
- Profile: Πληροφορίες προφίλ που εξάγονται μέσω του Twitter API, π.χ. το όνομα χρήστη, η περιγραφή του προφίλ, ο αριθμός ακολούθων και ακολουθούμενων, καθώς και μεταδεδομένα όπως η ημερομηνία δημιουργίας του λογαριασμού. Συγκεκριμένα συλλέγονται 38 ιδιότητες για κάθε χρήστη από το Twitter API.
- Tweet: Περιλαμβάνονται τα 200 πιο πρόσφατα tweets κάθε χρήστη. Διατηρείται η αρχική μορφή των tweets, συμπεριλαμβανομένων emojis, URLs, και retweets, επιτρέποντας τη λεπτομερή ανάλυση.
- Neighbor: Λίστες με έως 10 ακολούθους (followers) και 10 ακολουθούμενους (followings) για κάθε χρήστη, επιλεγμένους με τη μέθοδο BFS. Με την συμπερίληψη σχέσεων ακολούθησης (followers και followings), δημιουργείται ένα κοινωνικό γράφημα για κάθε χρήστη, προσέγγιση η οποία δύναται να προσθέσει δυναμική διάσταση στην μελέτη του συνόλου δεδομένων.
- Domain: Οι τομείς ενδιαφέροντος του χρήστη, π.χ. πολιτική, επιχειρήσεις, ψυχαγωγία ή αθλητισμός.
- Label: Ένδειξη αν ο χρήστης είναι bot (1) ή άνθρωπος (0).



Σχήμα 4.1: *Απεικόνιση μιας συστάδας χρηστών στο TwiBot-20 με την @SpeakerPelosi ως τον αρχικό χρήστη (seed user). Οι πράσινοι κόμβοι αντιπροσωπεύουν ανθρώπινους χρήστες στη συστάδα χρηστών, οι κόκκινοι κόμβοι αντιπροσωπεύουν bot χρήστες, και οι ακμές στο γράφημα υποδεικνύουν τις σχέσεις ακολούθησης μεταξύ των χρηστών.*

[73]

Αρχικοί Χρήστες (Seed Users)

Το αρχείο TwiBot-20_Seed_Users.txt περιλαμβάνει μια λίστα με αρχικούς χρήστες (seed users) που χρησιμοποιήθηκαν για τη συλλογή του dataset. Αυτοί χωρίζονται σε τέσσερις κατηγορίες:

- Politics: π.χ. @realDonaldTrump, @SpeakerPelosi.
- Business: π.χ. @amazon, @elonmusk.
- Entertainment: π.χ. @iamcardib, @samsmith.
- Sports: π.χ. @StephenCurry30, @Cristiano.

Αυτή η επιλογή διασφαλίζει ότι οι χρήστες καλύπτουν διαφορετικές κοινότητες και θέματα, βοηθώντας στη δημιουργία ενός αντιπροσωπευτικού δείγματος του Twittersphere.

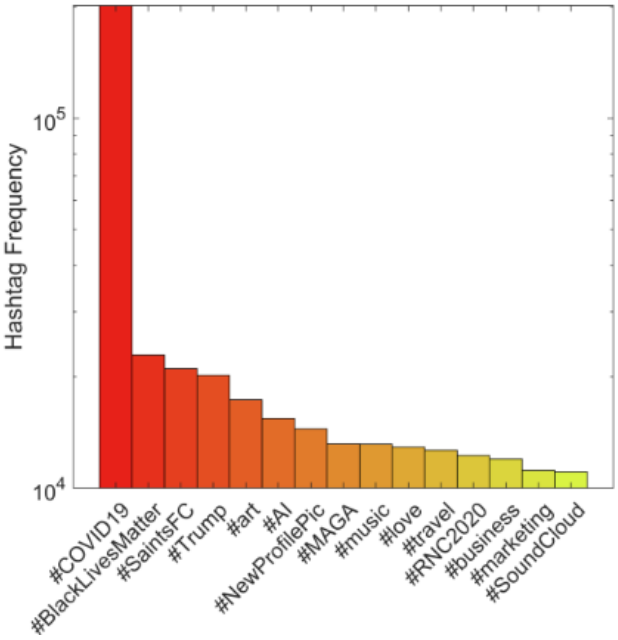
Ποικιλία Χρηστών

Το Twibot-20 σχεδιάστηκε για να αντικατοπτρίζει την ποικιλία των χρηστών του Twitter:

- Γεωγραφική Ποικιλία: Οι χρήστες, τα προφίλ των οποίων έχουν συγκεντρωθεί στο Twibot-20 παρουσιάζουν σημαντική γεωγραφική ποικιλία. Οι χρήστες προέρχονται από όλο τον κόσμο, αλλά οι δύο χώρες με την συχνότερη εμφάνιση είναι η Ινδία και οι Ηνωμένες Πολιτείες. Παράλληλα έχουμε σημαντική συμμετοχή χρηστών από την Ευρώπη και την Αφρική.
- Ενδιαφέροντα Χρηστών: Καλύπτονται τέσσερις βασικοί τομείς: πολιτική, επιχειρήσεις, ψυχαγωγία και αθλητισμός. Συχνά hashtags, όπως το #COVID19, #Trump, #business, #marketing, #love και #travel, δείχνουν την πολυπλοκότητα και τη διαφοροποίηση των ενδιαφερόντων των χρηστών.

Πληροφορίες Συλλογής Δεδομένων

Τα δεδομένα συλλέχθηκαν με τη βοήθεια ενός ελεγχόμενου αλγορίθμου αναζήτησης πλάτους (BFS). Αυτή η μέθοδος επιλέγει χρήστες μέσω των σχέσεών τους στο δίκτυο, εξασφαλίζοντας μια πυκνή και συνεκτική δομή γραφήματος. Η δομή αυτή επιτρέπει την ανάλυση τόσο σε επίπεδο μεμονωμένων χρηστών όσο και σε επίπεδο κοινοτήτων. Πιο συγκεκριμένα, η αναζήτηση ξεκινά από καθορισμένους seed users, 40 στο σύνολο, οι οποίοι επιλέχθηκαν για να αντιπροσωπεύουν διαφορετικούς τομείς ενδιαφέροντος (politics, business, entertainment, sports). Στη συνέχεια με την BFS αναζήτηση πλάτους, εντοπίζονται χρήστες που συνδέονται με τους seed users μέσω σχέσεων ακολούθησης (follow relationships), είτε άμεσα είτε έμμεσα μέσω άλλων χρηστών.



Σχήμα 4.2: *Τα 15 πιο συχνά hashtags που χρησιμοποιούν οι εγγεγραμμένοι χρήστες του Twibot-20. Παρόμοια hashtags όπως #COVID19 και #coronavirus έχουν συγχωνευτεί ώστε να δίνεται καλύτερη εικόνα για τα θέματα που απασχολούν συχνότερα τους χρήστες.*
[73]

Στρατηγική Επισήμανσης Χρηστών ως Bot ή Human

Η επισήμανση των χρηστών ως bots (αυτόματοι λογαριασμοί) ή humans (ανθρώπινοι λογαριασμοί) αποτελεί μια κρίσιμη διαδικασία στη δημιουργία του Twibot-20, καθώς επηρεάζει την ποιότητα και την αξιοπιστία του συνόλου δεδομένων. Για να γίνει αυτό με ακρίβεια, ακολουθήθηκε μια πολυεπίπεδη στρατηγική με σαφή βήματα.

1. Καθορισμός Κριτηρίων Επισήμανσης
   Για να καθοριστεί αν ένας λογαριασμός είναι bot ή human, χρησιμοποιήθηκαν κριτήρια βασισμένα σε προηγούμενες επιστημονικές έρευνες. Αυτά τα κριτήρια επικεντρώνονται σε χαρακτηριστικά που συνήθως εμφανίζουν οι λογαριασμοί bot:

   - Αυτοματοποιημένη δραστηριότητα: Τα bots συχνά εκτελούν μαζικές ή πολύ γρήγορες ενέργειες, όπως tweets ή retweets, χρησιμοποιώντας αυτοματοποιημένα εργαλεία ή APIs.
   - Επαναλαμβανόμενο περιεχόμενο: Αν ένας λογαριασμός δημοσιεύει συνεχώς τα ίδια tweets ή retweets, αυτό μπορεί να είναι ένδειξη ότι είναι bot.
   - Εξωτερικοί σύνδεσμοι που προωθούν phishing ή διαφημίσεις: Τα bots συχνά περιλαμβάνουν URLs που οδηγούν σε επικίνδυνες ή εμπορικές ιστοσελίδες.
   - Έλλειψη πρωτοτυπίας και σχετικότητας: Τα tweets ενός bot μπορεί να φαίνονται άσχετα ή να είναι «γενικά» χωρίς καμία μοναδικότητα ή προσωπικότητα.

2. Crowdsourcing για Αρχική Επισήμανση
   Η διαδικασία επισήμανσης περιλάμβανε τη συμμετοχή ανθρώπων αξιολογητών μέσω μιας καμπάνιας crowdsourcing. Συγκεκριμένα:

- Ομάδα αξιολογητών: Για κάθε λογαριασμό ανατέθηκαν πέντε διαφορετικοί αξιολογητές.
- Εκπαίδευση αξιολογητών: Οι αξιολογητές ήταν ενεργοί χρήστες του Twitter και εκπαιδεύτηκαν με τη βοήθεια ενός οδηγού που περιείχε τα κριτήρια ανίχνευσης bot, μαζί με παραδείγματα για καλύτερη κατανόηση.
- Αξιολόγηση: Οι αξιολογητές έκριναν αν ένας λογαριασμός είναι bot, human, ή ασαφής. Παράλληλα, έπρεπε να απαντήσουν σε συγκεκριμένες πρότυπες ερωτήσεις, που σχεδιάστηκαν για να αξιολογήσουν την αξιοπιστία τους.

Για κάθε λογαριασμό, η πλειοψηφική απόφαση (majority voting), δηλαδή, αν τέσσερις από τους πέντε αξιολογητές συμφωνούσαν, χρησιμοποιήθηκε για την αρχική επισήμανση.

3. Τελική Επιβεβαίωση Επισήμανσης
   Οι περιπτώσεις όπου δεν υπήρχε σαφής συμφωνία ή κρίθηκαν «αμφίβολες» υποβλήθηκαν σε περαιτέρω έλεγχο:

   - Αλληλεπίδραση με τον λογαριασμό: Οι ερευνητές έστειλαν μηνύματα μέσω του Twitter σε χρήστες που δεν μπορούσαν να χαρακτηριστούν με σαφήνεια, ζητώντας απλές πληροφορίες για να καθορίσουν αν πρόκειται για bot ή human.
   - Χειροκίνητος έλεγχος: Η ερευνητική ομάδα εξέτασε διεξοδικά αυτούς τους λογαριασμούς. Σε περιπτώσεις όπου υπήρχε ακόμα διαφωνία, οι λογαριασμοί αποκλείστηκαν από το dataset.

Η παραπάνω στρατηγική εξασφαλίζει ότι η επισήμανση των χρηστών είναι αξιόπιστη, ακριβής και αντιπροσωπευτική, μειώνοντας τον κίνδυνο σφαλμάτων.

Συνοψίζοντας, το Twibot-20 επιλέχθηκε διότι παρέχοντας πλήρεις πληροφορίες από τρεις διαστάσεις (σημασιολογικές, ιδιότητες, και γειτονικές πληροφορίες), υπερέχει σε σχέση με άλλα σύνολα δεδομένων που χρησιμοποιούνται για αναγνώριση bot. Διασφαλίζει την αξιοπιστία της επισήμανσης μέσω ενός συνδυασμού crowdsourcing, αυτοματοποιημένων μεθόδων και χειροκίνητης επιβεβαίωσης, ενώ η ποικιλομορφία χρηστών και θεμάτων προσφέρει μια ρεαλιστική απεικόνιση του Twittersphere. Αυτά τα χαρακτηριστικά καθιστούν το Twibot-20 μια πρωτοποριακή πηγή για την ανίχνευση bots και την ανάπτυξη γενικευμένων και αποτελεσματικών αλγορίθμων.

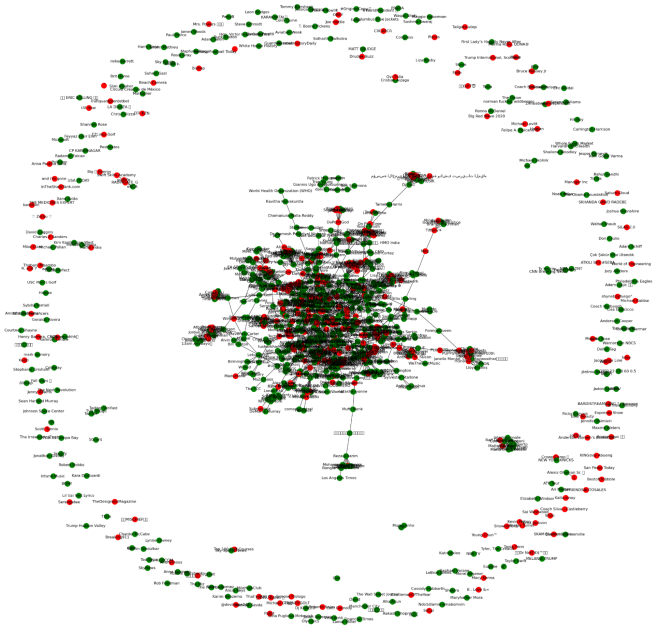
4.2 Προεπεξεργασία Συνόλου Δεδομένων

Φυσικά, στα μοντέλα μηχανικής μάθησης και στα Μεγάλα Γλωσσικά Μοντέλα που εφαρμόστηκαν στην παρούσα διπλωματική, το σύνολο δεδομένων Twibot-20 χρειάστηκε προεπεξεργασία προκειμένου να χρησιμοποιηθεί. Σε αυτή την ενότητα, λοιπόν, θα αναφερθούν τα βήματα προεπεξεργασίας που απαιτήθηκαν ώστε να προκύψουν τα υποσύνολα του Twibot-20 που αξιοποιήθηκαν.

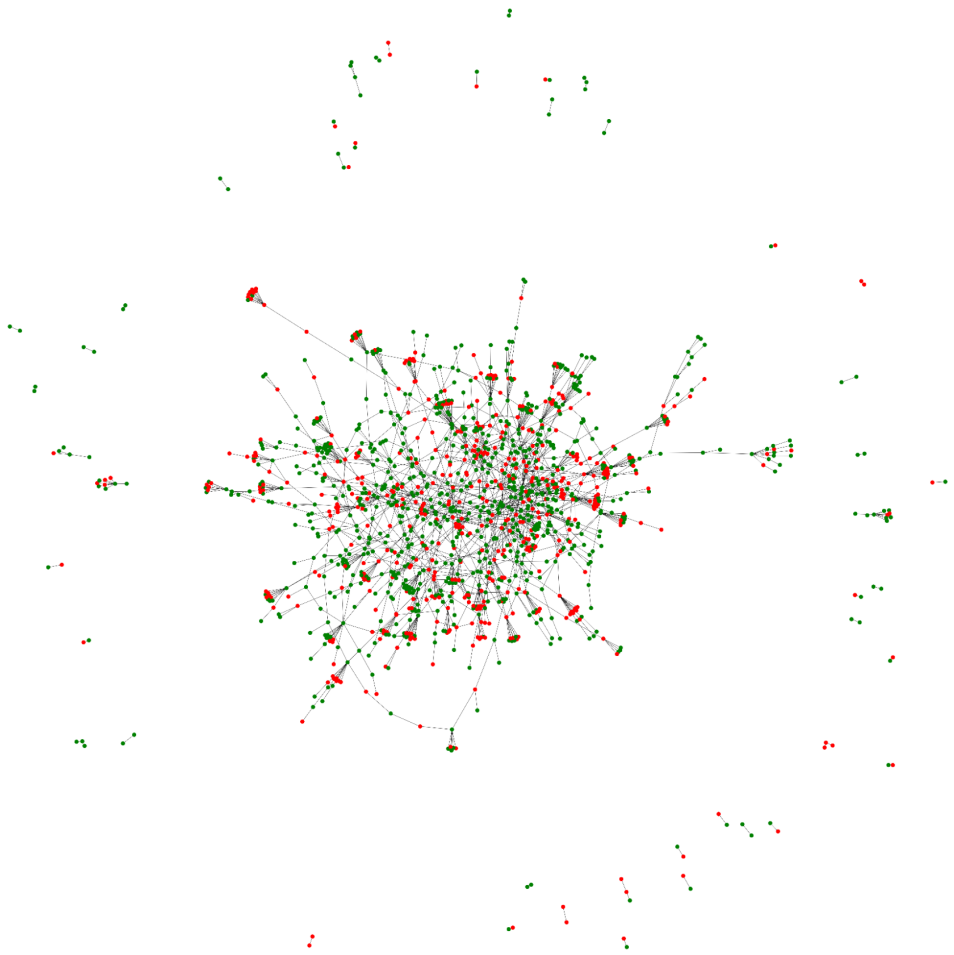
Σε πρώτη φάση, αυτό που παρατηρήθηκε είναι πως τα αρχεία του Twibot-20 είναι εξαιρετικά μεγάλα σε μέγεθος, γεγονός που προκαλεί δυσκολία στο να ανοιχτούν και να υποστούν επεξεργασία τοπικά σε έναν υπολογιστή, και ειδικά στα πλαίσια του Jupyter Notebook, όπου και έγινε η ανάπτυξη των μοντέλων. Το αρχείο που επιλέχθηκε είναι το train.json με μέγεθος 239 MB. Αφού ανοίχτηκε πετυχημένα, μέσω κώδικα στην python, απομονώθηκε ένα υποσύνολό του, με μέγεθος 43 MB, που ονομάστηκε "trainDataFromGraph.json".

Όπως υποδηλώνει και το όνομα του "trainDataFromGraph.json" αρχείου, προήλθε από έναν μεγάλο γράφο. Συγκεκριμένα, δημιουργήθηκε ένα μεγάλο δίκτυο όπου οι κόμβοι-χρήστες που το αποτελούν πρέπει να έχουν following ή follower σχέση με τουλάχιστον άλλον ένα κόμβο-χρήστη του δικτύου. Αυτό έγινε προκειμένου σε αυτό το υποσύνολο του Twibot-20 να υπάρχουν χρήστες με γειτονικές σχέσεις μεταξύ τους, ώστε να εφαρμοστούν, μεταξύ άλλων, τεχνικές μηχανικής μάθησης που βασίζονται στην μορφή του γραφήματος. Αν το τελικό αρχείο περιείχε χρήστες αποκομμένους από το υπόλοιπο δίκτυο, τα μοντέλα θα ήταν δυσκολότερο να εφαρμοστούν. Επιπλέον, στο δίκτυο που χρησιμοποιήθηκε απεικονίζονται με κόκκινο χρώμα οι κόμβοι - χρήστες bot και με πράσινο οι κόμβοι - χρήστες άνθρωποι.

Στα δύο παρακάτω σχήματα φαίνεται το δίκτυο σε δύο φάσεις, πριν και μετά την αφαίρεση των αποκομμένων χρηστών (και την αφαίρεση του ονόματος χρήστη για πιο ευπαρουσίαστο δίκτυο). Σημαντικό να σημειωθεί πως στα παρακάτω δίκτυα για καλύτερη απεικόνιση έχουν χρησιμοποιηθεί λιγότεροι κόμβοι - χρήστες από ότι υπάρχουν στο τελικό αρχείο "trainDataFromGraph.json".



Σχήμα 4.3: *Απεικόνιση ενός υποσυνόλου του δικτύου που χρησιμοποιήθηκε κατά το πειραματικό σκέλος της εργασίας. Στο συγκεκριμένο δίκτυο περιλαμβάνονται οι μη συνδεδεμένοι κόμβοι - χρήστες μαζί με τις ετικέτες του ονόματός τους, που στη συνέχεια αφαιρέθηκαν ώστε να προκύψει πιο συνεκτικό δίκτυο. Οι πράσινοι κόμβοι αντιπροσωπεύουν ανθρώπινους χρήστες στη συστάδα χρηστών, οι κόκκινοι κόμβοι αντιπροσωπεύουν bot χρήστες, και οι ακμές στο γράφημα υποδεικνύουν τις σχέσεις ακολούθησης μεταξύ των χρηστών.*



Σχήμα 4.4: *Απεικόνιση υποσυνόλου του δικτύου που αξιοποιήθηκε στα πειράματα, μετά την αφαίρεση των κόμβων που δεν είχαν τουλάχιστον έναν γείτονα. Οι πράσινοι κόμβοι αντιπροσωπεύουν ανθρώπινους χρήστες στη συστάδα χρηστών, οι κόκκινοι κόμβοι αντιπροσωπεύουν bot χρήστες, και οι ακμές στο γράφημα υποδεικνύουν τις σχέσεις ακολούθησης μεταξύ των χρηστών.*

Προκειμένου το δίκτυο που επιλέχθηκε σαν υποσύνολο του Twibot-20 να μπορεί να χρησιμοποιηθεί εύκολα στα μοντέλα, δηλαδή να φορτώνεται με μεγάλη ταχύτητα, ακολούθησε άλλος ένας γύρος προεπεξεργασίας ώστε να μειωθεί το μέγεθός του. Συγκεκριμένα για τα μοντέλα μηχανικής μάθησης που χρησιμοποιούν μόνο τα χαρακτηριστικά του προφίλ και όχι το περιεχόμενο των tweets κάθε λογαριασμού, αφαιρέθηκαν τα tweets, που καταλαμβάνουν την πλειονότητα του αρχείου. Δηλαδή, τελικά προέκυψε ένα json αρχείο με τους χρήστες ίδιο με το αρχικό, αλλά χωρίς την λίστα "tweet": [..] που περιέχει τα τελευταία 200 tweets του εκάστοτε χρήστη.

Από την άλλη, για τις τεχνικές κατηγοριοποίησης με τη βοήθεια Μεγάλων Γλωσσικών Μοντέλων (LLMs) ακολουθήθηκε η αντίστροφη διαδικασία. Εφόσον σε αυτή την προσέγγιση, στόχος ήταν να γίνει η πρόβλεψη από το LLM με βάση τα χαρακτηριστικά κειμένου (text-

based), δηλαδή το περιεχόμενο των tweets κάθε χρήστη και της περιγραφής του προφίλ του, απομονώθηκαν μόνο αυτά τα πεδία του json αρχείου. Φυσικά, όμως, υπήρχαν και το ID κάθε λογαριασμού για mapping και το label (0 ή 1, bot ή άνθρωπος) για σύγκριση των προβλέψεων του LLM με το groundtruth του συνόλου δεδομένων.

Κατά τη διάρκεια της ανάπτυξης των μοντέλων που χρησιμοποιήθηκαν στην έρευνα αυτής της διπλωματικής, χρησιμοποιήθηκαν διάφορα υποσύνολα του "trainDataFromGraph.json". αρχείου. Εξετάστηκε η απόδοση των μοντέλων σε μικρότερα και μεγαλύτερα dataset, εξασφαλίζοντας πάντα μια σχετική εξισορρόπηση μεταξύ των labels, ώστε να υπάρχει πάντα ικανοποιητική ποσότητα δειγμάτων bot και ανθρώπων χρηστών. Πιο λεπτομερώς, αυτά τα υποσύνολα του Twibot-20, το μέγεθός τους, το πλήθος χρηστών της κάθε κατηγορίας και τα χαρακτηριστικά (features) που επιλέχθηκαν για την εκπαίδευση των μοντέλων θα εξεταστούν στα επόμενα κεφάλαια με τη μεθοδολογία και τα πειράματα.

Μέρος II

Πρακτικό Μέρος

Κεφάλαιο 5

Μεθοδολογία

5.1 Επισκόπηση Μεθοδολογίας

Σε αυτό το κεφάλαιο παρουσιάζεται η μεθοδολογία της ερευνητικής διαδικασίας που ακολουθήθηκε στην παρούσα διπλωματική εργασία για την ανίχνευση αυτοματοποιημένων λογαριασμών bot στο κοινωνικό δίκτυο Twitter. Η μεθοδολογία περιλαμβάνει τρία βασικά στάδια: την κατασκευή του δικτύου από το σύνολο δεδομένων, την εξαγωγή και επιλογή χαρακτηριστικών και, τέλος, τον καθορισμό των μοντέλων με τα οποία θα γίνει η υλοποίηση.

Κατασκευή Δικτύου: Το γράφημα δημιουργήθηκε από τα δεδομένα του Twibot-20, όπου οι χρήστες του Twitter αντιπροσωπεύονται από κόμβους και οι σχέσεις ακολούθησης από ακμές. Το γράφημα αυτό παρέχει μια δομή που προσφέρεται για κατανόηση των σχέσεων και της αλληλεπίδρασης μεταξύ των χρηστών.

Εξαγωγή και Επιλογή Χαρακτηριστικών: Για την μελέτη των χρηστών και την προσπάθεια κατηγοριοποίησής τους ως λογαριασμοί που ανήκουν σε άνθρωπο χρήστη ή bot, έγινε εξαγωγή χαρακτηριστικών από το σύνολο δεδομένων και ακολούθως από το κατασκευασμένο δίκτυο. Και στη συνέχεια, πραγματοποιήθηκε η επιλογή των καταλληλότερων χαρακτηριστικών για την εκπαίδευση των μοντέλων πάνω σε αυτά. Ενδεικτικά χρησιμοποιήθηκαν χαρακτηριστικά που μπορούν να κατανεμηθούν στις εξής κατηγορίες:

- Τοπολογικά χαρακτηριστικά (π.χ., αριθμός γειτόνων, κεντρικότητα).
- Αριθμητικά χαρακτηριστικά (π.χ., αριθμός tweets, followers, και ακόλουθων).
- Χαρακτηριστικά κειμένου (π.χ. περιγραφή προφίλ χρήστη, περιεχόμενο των δημοσιευμένων tweets)

Μοντέλα Ανίχνευσης: Επιλέχθηκαν διάφορες μέθοδοι για την ανίχνευση bots, όπως κλασικά μοντέλα μηχανικής μάθησης (Logistic Regression, Random Forest), γραφο-νευρωνικά δίκτυα (GCN, GAT), και μεγάλα γλωσσικά μοντέλα (LLMs).

Τα παραπάνω στάδια, όσον αφορά τις λεπτομέρειες της μεθοδολογίας, αναλύονται διεξοδικά στις επόμενες ενότητες αυτού του κεφαλαίου.

5.2 Κατασκευή Δικτύου

Αυτή η ενότητα εστιάζει στην κατασκευή των δικτύων που χρησιμοποιήθηκαν για να εκπαιδεύτούν και να αξιολογηθούν τα μοντέλα. Στο προηγούμενο κεφάλαιο, στην ενότητα 4.2, αναλύθηκε η διαδικασία προεπεξεργασίας που υπέστει το σύνολο δεδομένων Twibot-20 προκειμένου να προκύψει το τελικό δίκτυο - υποσύνολό του, που αποθηκεύτηκε στο αρχείο "trainDataFromGraph.json". Στα διάφορα πειράματα που έγιναν χρησιμοποιήθηκε ολόκληρο αυτό το δίκτυο ή μικρότερα του υποδίκτυα, ώστε να εξεταστεί την απόδοση των μοντέλων σε διαφορετικού μεγέθους δίκτυα. Στη συνέχεια θα αναφερθούν όλα τα υπο-δίκτυα που χρησιμοποιήθηκαν και τις τεχνικές λεπτομέρειες του καθενός.

Πηγή Δεδομένων

Η κατασκευή του δικτύου πάνω στο οποίο εφαρμόστηκαν τα μοντέλα για την ανίχνευση αυτοματοποιημένων λογαριασμών bot βασίστηκε, όπως έχει ήδη αναφερθεί, στο σύνολο δεδομένων Twibot-20. Τα δεδομένα του περιλαμβάνουν πληροφορίες για χρήστες, σχέσεις ακολούθησης (follow relationships), και χαρακτηριστικά χρηστών, όπως περιγραφές προφίλ, αριθμό ακολούθων κλπ. Συνεπώς, έχουν μια δομή που διευκολύνει την κατασκευή δικτύων και μέσα σε αυτά ευνοεί την μελέτη των σχέσεων των χρηστών και της ροής της πληροφορίας ανάμεσά τους. Παράλληλα, καθιστά απλούστερη την παρατήρηση της δράσης κάθε μεμονωμένου χρήστη μέσα στο δίκτυο.

Μετατροπή Δεδομένων σε Γράφο

Προκειμένου να αναπαρασταθούν τα δεδομένα σε ένα γράφημα έπρεπε να ακολουθηθεί η εξής διαδικασία μετατροπής:

- Κόμβοι: Κάθε κόμβος του δικτύου αντιστοιχεί σε έναν χρήστη του Twitter. Ανάλογα με την τιμή (0 ή 1) που έχει στο πεδίο "label" του συνόλου δεδομένων ο κάθε χρήστης, ερμηνεύεται σαν χρήστης άνθρωπος ή bot.

- Ακμές: Οι ακμές αναπαριστούν τις σχέσεις ακολούθησης μεταξύ των χρηστών. Δηλαδή, αν ένας χρήστης Α ακολουθεί έναν χρήστη Β, τότε αυτή η σχέση ακολούθησης αναπαρίσταται με μια κατευθυνόμενη ακμή από τον χρήστη Α στον Β, και αντιστρόφως.

Συνεπώς, το δίκτυο που προκύπτει είναι ένα κατευθυνόμενο γράφημα. Το οποίο είναι και φυσιολογικό μιας και οι σχέσεις ακολούθησης στο Twitter έχουν κατεύθυνση και δεν είναι απαραίτητα αμφίδρομες, όπως σε άλλα μέσα κοινωνικής δικτύωσης.

Είναι σημαντικό να αναφερθεί πως κάθε κόμβος του δικτύου ενώνεται με τουλάχιστον μια (¨εισερχόμενη¨ ή ¨εξερχόμενη¨) ακμή με κάποιον άλλο κόμβο. Δηλαδή, κάθε χρήστης ακολουθεί τουλάχιστον έναν άλλο χρήστη ή/και ακολουθείται από τουλάχιστον άλλον έναν χρήστη. Αυτό έχει ως αποτέλεσμα να μην υπάρχει κανένας εντελώς αποκομμένος κόμβος στο δίκτυο, και έτσι να μελετηθούν πιο αποτελεσματικά τα γραφικά χαρακτηριστικά του.

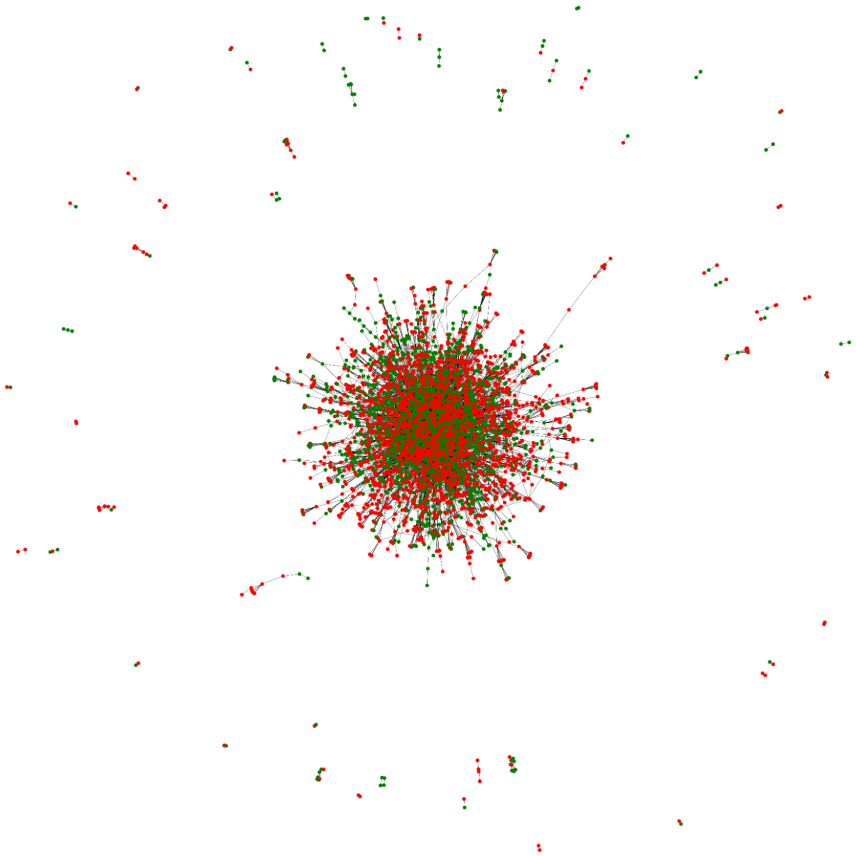
Διαθέσιμα δίκτυα που αξιοποιήθηκαν

Προκειμένου να γίνει η αξιολόγηση της απόδοσης των μοντέλων αξιοποιήθηκαν δίκτυα διαφορετικών μεγεθών, τα οποία προήλθαν από το αρχικό δίκτυο που προέκυψε από την προεπεξεργασία του Twibot-20. Συγκεκριμένα έγινε χρήση τριών δικτύων για την ανίχνευση bot με τα μοντέλα Logistic Regression, Random Forest, Graph Convolutional Network και Graph Attention Network και ένα ¨δίκτυο¨ για την ανίχνευση με Μεγάλα Γλωσσικά Μοντέλα (Large Language Models - LLMs).

Τα τρία δίκτυα - σύνολα δεδομένων που χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση των μοντέλων μηχανικής μάθησης επιγραμματικά είναι τα εξής:

- smallMLdataset
- midMLdataset
- bigMLdataset

Τα στατιστικά τους αναφέρονται παρακάτω.



Σχήμα 5.1: *Οπτικοποίηση του Γράφου που αναπαριστούν τα δεδομένα του "bigMLdataset". Με κόκκινο τα bots και πράσινο οι άνθρωποι χρήστες.*

Από την άλλη, όσον αφορά την κατηγοριοποίηση λογαριασμών χρηστών ως bots ή άνθρωποι με προεκπαιδευμένα Μεγάλα Γλωσσικά Μοντέλα έγινε χρήση ενός μόνο δικτύου μικρής κλίμακας. Ονομάστηκε "LLMdataset" και αποθηκεύτηκε σε ομώνυμο json αρχείο. Συνολικά αποτελείται από 100 χρήστες, 50 εκ των οποίων είναι bots και 50 άνθρωποι. Όπως αναφέρεται και παρακάτω, στην ανάλυση με LLMs έγινε χρήση μόνο των χαρακτηριστικών περιγραφή προφίλ και περιεχόμενο tweets (description, tweet) κάθε χρήστη. Συνεπώς, η γραφική αναπαράσταση του δικτύου δεν ήταν αναγκαία, μιας και τα γραφικά χαρακτηριστικά του εδώ ήταν αδιάφορα. Για αυτό το λόγο σε αυτό το μικρό σύνολο δεδομένων, δεν επιλέχθηκαν αποκλειστικά λογαριασμοί που έχουν σχέση ακολούθησης με τουλάχιστον άλλον έναν λογαριασμό του δικτύου. Επίσης, το μέγεθός του είναι τόσο μικρό λόγω των χρονικών και υπολογιστικών περιορισμών της εκτέλεσης API calls σε LLMs. Περισσότερα θα αναφερθούν στο κεφάλαιο 6 με την υλοποίηση των πειραμάτων.

Στατιστικά των Δικτύων

Για την ανάλυση και την καλύτερη κατανόηση της δομής κάθε δικτύου, υπολογίστηκαν μερικά βασικά στατιστικά μεγέθη (Πίνακας 4.1):

- Αριθμός Κόμβων και Ακμών: Το εκάστοτε δίκτυο περιλαμβάνει X κόμβους και Y ακμές, όπου X είναι ο συνολικός αριθμός χρηστών και Y ο συνολικός αριθμός σχέσεων ακολούθησης.

- Αριθμός Χρηστών Bot και Χρηστών Ανθρώπων: Ο αριθμός των κόμβων - χρηστών κάθε δικτύου ισούται με τον αριθμό bot και ανθρώπων χρηστών που το αποτελούν. Όπως φαίνεται και στον πίνακα με τα στατιστικά που ακολουθεί, σε κάθε δίκτυο υπάρχει μια εξισορρόπηση δειγμάτων από κάθε κατηγορία ώστε να έχουμε επαρκή δεδομένα για την εκπαίδευση και αξιολόγηση των μοντέλων ανίχνευσης.

- Συνδεδεμένες Συνιστώσες (Σ.Σ.): Υπολογίστηκε ο αριθμός των συνδεδεμένων συνιστωσών για την κατανόηση της συνοχής του δικτύου. Το γράφημα περιλαμβάνει Z συνδεδεμένες συνιστώσες, υποδεικνύοντας ότι υπάρχουν περιοχές απομονωμένων ή λιγότερο συνδεδεμένων χρηστών.

- Βαθμός Δικτύωσης (Β.Δ.): Αναλύθηκε η κατανομή του βαθμού (degree distribution), που περιγράφει τον αριθμό εισερχόμενων και εξερχόμενων συνδέσεων για κάθε κόμβο. Παρατηρήθηκε ότι η κατανομή ακολουθεί χαρακτηριστικά ενός scale-free δικτύου, όπου λίγοι κόμβοι έχουν πολλούς συνδέσμους (hubs), ενώ οι περισσότεροι κόμβοι έχουν λίγους.

Οπτικοποίηση των Δικτύων με Γραφήματα

Για την κατανόηση της δομής των δικτύων, πραγματοποιήθηκε οπτικοποίηση του γραφήματος. Χρησιμοποιήθηκαν εργαλεία όπως το NetworkX για τη δημιουργία γραφικών παραστάσεων. Συγκεκριμένα, και για τα τρία δίκτυα που χρησιμοποιήθηκαν στα μοντέλα μηχανικής μάθησης έγιναν οι εξής οπτικοποιήσεις:

Δίκτυο	Κόμβοι	Ακμές	Bots	Humans	Σ.Σ.	Β. Δ.
smallMLdataset	1448	1661	524	924	47	0.0016
midMLdataset	3756	4607	1801	1955	62	0.0007
bigMLdataset	6316	8062	3524	2792	58	0.0004
LLMdataset	100	-	50	50	-	-

Πίνακας 5.1: *Στατιστικά των Δικτύων που χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση των μεθόδων ανίχνευσης Bots*

- Αναπαράσταση του πλήρους γραφήματος.
- Αποτύπωση μιας περιοχής με υψηλή συνδεσιμότητα, για να αναδειχθούν τα hubs.
- Η κατανομή του βαθμού (degree distribution) σε μορφή γραφήματος, που αποτυπώνει τη συχνότητα των εισερχόμενων και εξερχόμενων συνδέσεων.

Παρουσιάζοναται κατά σειρά αναφοράς οι παραπάνω οπτικοποιήσεις για το δίκτυο smallMLdataset (Σχήμα 4.2).

5.3 Εξαγωγή και Επιλογή Χαρακτηριστικών

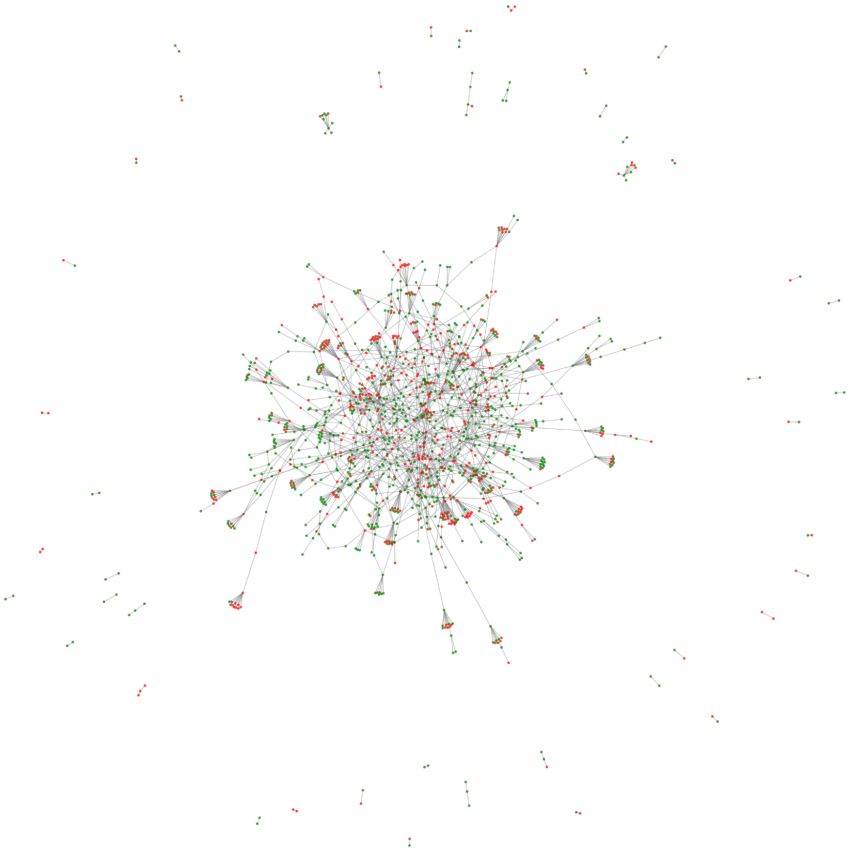
Η εξαγωγή χαρακτηριστικών (features) από το σύνολο δεδομένων Twibot-20 και, στη συνέχεια, η επιλογή των καταλληλότερων για την εκπαίδευση και την αξιολόγηση των μοντέλων με αυτά, αποτέλεσε κρίσιμο βήμα κατά την διεκπεραίωση της παρούσας διπλωματικής εργασίας. Η διαδικασία αυτή περιείχε την επιλογή χαρακτηριστικών που ενσωματώνουν σημαντικές πληροφορίες τόσο για τη δικτυακή δομή του γράφου όσο και για τις αριθμητικές και κειμενικές (textual) πληροφορίες των χρηστών.

Τύποι Εξαγόμενων Χαρακτηριστικών

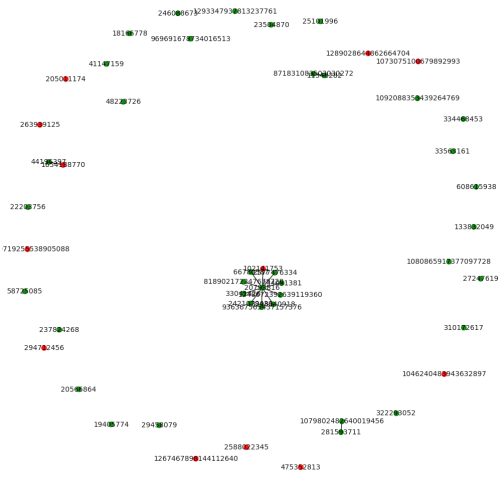
Τα χαρακτηριστικά που εξήχθησαν από το σύνολο δεδομένων και επιλέχθηκαν για την εκπαίδευση των μοντέλων κατηγοριοποιούνται σε τρεις βασικές κατηγορίες: τοπολογικά χαρακτηριστικά, αριθμητικά και κειμενικά. Κάποια από αυτά προήλθαν άμεσα από τις εγγραφές χρηστών του Twibot-20, ενώ κάποια άλλα προέκυψαν από τη δομή του γράφου κατά την εκπαίδευση των μοντέλων μηχανικής μάθησης. Παρακάτω παρουσιάζονται τα χαρακτηριστικά που αξιοποιήθηκαν. Συγκεκριμένα μαζί με το όνομά τους, δίνεται και μια περιγραφή του περιεχομένου τους.

Τοπολογικά Χαρακτηριστικά (Topological Features)

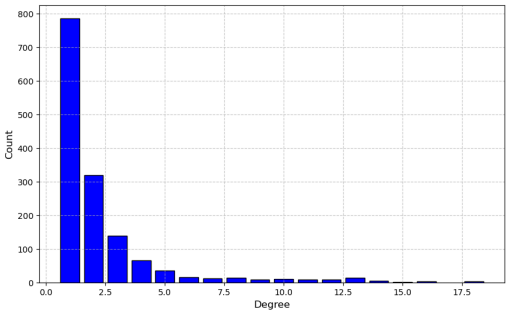
Τα τοπολογικά χαρακτηριστικά βασίζονται στη δομή του γράφου και παρέχουν πληροφορίες για τη θέση κάθε κόμβου (χρήστη) στο δίκτυο. Προκειμένου να κατασκευαστεί το δίκτυο με τους κόμβους - χρήστες και τις κατευθυνόμενες ακμές που αναπαριστούν τις σχέσεις ακολούθησης, χρειάστηκε να εξαχθούν από το σύνολο δεδομένων τα χαρακτηριστικά "following" και "follower" που ανήκουν στο πεδίο "neighbor" του Twibot-20. Για αυτά τα χαρακτηριστικά ισχύουν τα εξής:



(α') *Αναπαράσταση του πλήρους γραφήματος*



(β') *Αποτύπωση μιας περιοχής με υψηλή συνδεσιμότητα*



(γ') *Κατανομή βαθμού*

Σχήμα 5.2: *Οπτικοποίηση του δικτύου smallMLdataset.*

- Following: Περιλαμβάνει μια λίστα με το πολύ 10 χρήστες που ακολουθεί ο εξεταζόμενος χρήστης. Αποθηκεύονται τα IDs των χρηστών αυτών.

- Follower: Περιλαμβάνει μια λίστα με το πολύ 10 χρήστες που ακολουθούν τον εξεταζόμενο χρήστη. Αποθηκεύονται τα IDs των χρηστών αυτών.

Εδώ, αξίζει να σημειωθεί ότι τα IDs (μοναδικό αναγνωριστικό κάθε λογαριασμού) των χρηστών, επίσης εξήχθησαν από το σύνολο δεδομένων. Είναι απαραίτητο χαρακτηριστικό κάθε λογαριασμού ώστε να γίνεται η συσχέτισή του με άλλους χρήστες. Επειδή, λοιπόν, είναι διαθέσιμα αυτά τα IDs, είναι εύκολο να εντοπιστούν οι σχέσεις ακολούθησης, και μετέπειτα να αναπαρασταθούν σε μορφή γράφου για την οπτικοποίηση του δικτύου και την περαιτέρω μελέτη του με αλγορίθμους μηχανικής μάθησης.

Από την δομή γράφου που έχουν τα δεδομένα, μπορούν να υπολογιστούν και άλλα, πιο σύνθετα, τοπολογικά χαρακτηριστικά. Συγκεκριμένα, επιλέχθηκαν οι εξής κεντρικότητες και μετρικές:

- Κεντρικότητα Βαθμού (Degree Centrality): Αναπαριστά τον βαθμό του εξεταζόμενου κόμβου.

- Κεντρικότητα Εγγύτητας (Closeness Centrality): Υπολογίζει πόσο κοντά βρίσκεται ο εξεταζόμενος κόμβος από όλους τους άλλους.

- Κεντρικότητα Διαμεσολάβησης (Betweenness Centrality): Δίνει μια τιμή που εκτιμά πόσο συχνά ένας κόμβος βρίσκεται στις συντομότερες διαδρομές μεταξύ άλλων κόμβων.

- Κεντρικότητα Ιδιοδιανύσματος (Eigenvector Centrality): Είναι ένα μέτρο της επιρροής του εξεταζόμενου κόμβου στο δίκτυο.

- Συντελεστής Συσταδοποίησης (Clustering Coefficient): Υπολογίζει τον βαθμό στον οποίο οι γειτονικοί κόμβοι ενός κόμβου - χρήστη είναι επίσης συνδεδεμένοι μεταξύ τους.

Ακόμα, άλλο ένα χαρακτηριστικό που αξιοποιήθηκε είναι ο αλγόριθμος του PageRank. Στο πλαίσιο της ανίχνευσης bots στο Twitter, ο PageRank μπορεί να εφαρμοστεί σε κοινωνικά δίκτυα χρηστών, όπως αυτά που προκύπτουν από το Twibot-20, ώστε να εντοπιστούν οι πιο επιδραστικοί ή αξιόπιστοι λογαριασμοί. Οι κόμβοι του γράφου αντιπροσωπεύουν τους χρήστες και οι ακμές τις αλληλεπιδράσεις τους (όπως οι σχέσεις ακολούθησης). Η ταξινόμηση των χρηστών με βάση το PageRank επιτρέπει την εξαγωγή χαρακτηριστικών που διακρίνουν τα αυθεντικά προφίλ από τα bots, καθώς αυτά συχνά εμφανίζουν χαμηλή διασύνδεση ή ασυνήθιστα μοτίβα συνδέσεων.

Επιπλέον, στα τοπολογικά χαρακτηριστικά μπορεί να αναφερθεί και η ¨Φήμη Λογαριασμού¨ (Account Reputation) που χρησιμοποιήθηκε. Είναι ένα χαρακτηριστικό που αναφέρεται στο "Twibot-20: A comprehensive twitter bot detection benchmark" [73] και μπορεί να εξαχθεί εύκολα από τα χαρακτηριστικά που ήδη παρέχονται στο σύνολο δεδομένων. Μέσω της μετρικής Account Reputation διερευνάται το reputation score ενός χρήστη bot και ενός

αυθεντικού ανθρώπινου χρήστη. Πιο συγκεκριμένα, η φήμη (reputation) είναι ένας συντελεστής που συνυπολογίζοντας τον αριθμό των ακολούθων (followers_count) ενός χρήστη και των λογαριασμών που ακολουθεί ο ίδιος (friends_count), καταλήγει στο αν αυτός έχει υψηλό ή χαμηλό reputation score.

$$Reputation(u) = \frac{|N^t(u)|}{|N^t(u)| + |N^f(u)|}$$

Το $|N^t(u)|$ αναπαριστά το σύνολο των ακολούθων (followers) του χρήστη u και το $|N^f(u)|$ αναπαριστά το σύνολο των χρηστών που ακολουθεί ο χρήστης u.

Αριθμητικά Χαρακτηριστικά (Numerical Features)

Τα αριθμητικά χαρακτηριστικά σχετίζονται με πληροφορίες που προκύπτουν από τα δεδομένα των χρηστών [74]. Τα εξαγόμενα χαρακτηριστικά περιλαμβάνουν:

- Αριθμός Ακολούθων (Followers Count): Ο συνολικός αριθμός χρηστών που ακολουθούν έναν λογαριασμό.
- Αριθμός Φίλων (Friends Count) : Ο συνολικός αριθμός χρηστών που ακολουθεί ένας λογαριασμός.
- Statuses Count: Ο συνολικός αριθμός των Tweets, συμπεριλαμβανομένων και των retweet (αναδημοσιεύσεις tweet άλλων χρηστών) που έχει δημοσιεύσει ένας λογαριασμός.
- Favourites Count: Ο συνολικός αριθμός των Tweets στα οποία ο χρήστης έχει πατήσει "like" ("Μου αρέσει").
- Listed Count: Ο συνολικός αριθμός δημόσιων λιστών στις οποίες ο χρήστης είναι μέλος.
- Verified: Ένα boolean πεδίο, που όταν είναι αληθές δηλώνει πως ο λογαριασμός είναι επιβεβαιωμένος από το Twitter.
- Geo_Enabled: Είναι ένα boolean πεδίο, που όταν είναι αληθές, σημαίνει πως ο χρήστης έχει συναινέσει στο να δημοσιεύονται τα δεδομένα της τοποθεσίας του όταν δημοσιεύει κάποιο Tweet.
- Default_Profile: Ένα boolean πεδίο, που όταν είναι αληθές δηλώνει πως ο χρήστης δεν έχει αλλάξει το θέμα (theme) ή το background του προφίλ του.
- Default_Profile_Image: Ένα boolean πεδίο, που όταν είναι αληθές δηλώνει πως ο χρήστης δεν έχει ανεβάσει την δική του εικόνα προφίλ, αλλά εξακολουθεί να χρησιμοποιεί την default εικόνα προφίλ που παρέχεται από το Twitter.
- Location_Present: Αυτό το χαρακτηριστικό έχει οριστεί στα πλαίσια αυτής της διπλωματικής, και δηλώνει την ύπαρξη ή μη ορισμένης τοποθεσίας για έναν λογαριασμό. Όταν είναι αληθές, σημαίνει πως έχει οριστεί τοποθεσία.

- Profile_Location_Present: Αυτό το χαρακτηριστικό έχει οριστεί στα πλαίσια αυτής της διπλωματικής, και δηλώνει την ύπαρξη ή μη ορισμένης εκτεταμένης τοποθεσίας για έναν λογαριασμό. Όταν είναι αληθές, σημαίνει πως έχει οριστεί εκτεταμένη τοποθεσία.

- Url_Present: Αυτό το χαρακτηριστικό έχει οριστεί στα πλαίσια αυτής της διπλωματικής, και δηλώνει την ύπαρξη ή μη url σε έναν λογαριασμό. Όταν είναι αληθές, σημαίνει πως έχει οριστεί ένα url από τον χρήστη για μεταπήδηση σε άλλη σελίδα που έχει ορίσει ο ίδιος (πχ. σελίδα της επιχείρησής του).

- Description_Present: Αυτό το χαρακτηριστικό έχει οριστεί στα πλαίσια αυτής της διπλωματικής, και δηλώνει την ύπαρξη ή μη περιγραφής προφίλ ενός λογαριασμού. Είναι ένα boolean πεδίο που είναι αληθές όταν η περιγραφή προφίλ δεν είναι κενή.

Κειμενικά Χαρακτηριστικά (Textual Features)

Τα κειμενικά χαρακτηριστικά προκύπτουν άμεσα από τα δεδομένα των χρηστών. Στην παρούσα έρευνα χρησιμοποιούνται μόνο στην ανίχνευση αυτοματοποιημένων λογαριασμών bot με προεκπαιδευμένα Μεγάλα Γλωσσικά Μοντέλα, και όχι στην ανίχνευση με μοντέλα μηχανικής μάθησης. Τα εξαγόμενα κειμενικά χαρακτηριστικά περιλαμβάνουν:

- Description: Πρόκειται για το περιεχόμενο που έχει αναρτήσει ο χρήστης στην περιγραφή του προφίλ του.

- Tweet: Είναι μια λίστα με τα Tweets που έχει αναρτήσει ή αναδημοσιεύσει ένας χρήστης στο προφίλ του.

Τέλος, από τα χαρακτηριστικά του συνόλου δεδομένων, έγινε σαφώς και χρήση του χαρακτηριστικού "label" που συνοδεύει κάθε χρήστη. Πρόκειται για μια boolean μεταβλητή που λειτουργεί ως το ground truth για τα πειράματα, δηλαδή είναι τιμή του συνόλου δεδομένων που δηλώνει αν ένας χρήστης είναι bot ή άνθρωπος. Είναι εξαρχής γνωστή, ώστε να εκπαιδευτούν τα μοντέλα με τα υπόλοιπα δεδομένα και με αυτή την τιμή να αξιολογηθεί η απόδοσή τους αργότερα. Έτσι προετοιμάζονται τα μοντέλα μηχανικής μάθησης, πριν εφαρμοστούν σε ξένα δεδομένα και πραγματικές συνθήκες για πρόβλεψη και ανίχνευση.

Εφαρμογή των Χαρακτηριστικών

Τα εξαγόμενα χαρακτηριστικά αξιοποιήθηκαν σε διάφορα μοντέλα μηχανικής μάθησης, από κλασικά μοντέλα, όπως το Random Forest και το Logistic Regression, μέχρι προηγμένα γραφικά μοντέλα, όπως το Graph Attention Network και το Graph Convolutional Network. Η ποικιλία των χαρακτηριστικών που επιλέχθηκαν διασφάλισε ότι τα μοντέλα μπορούσαν να ενσωματώσουν πληροφορίες τόσο από τη δομή του δικτύου όσο και από τις αριθμητικές μεταβλητές, ενισχύοντας την ακρίβεια της κατηγοριοποίησης.

Συνολική Παρουσίαση Χαρακτηριστικών

Τα εξαγόμενα χαρακτηριστικά συνοψίζονται στον παρακάτω πίνακα:

Χαρακτηριστικό	**Τύπος Δεδομένων**
ID	Int64
Following	List of IDs
Follower	List of IDs
Degree Centrality	Dictionary
Closeness Centrality	Dictionary
Betweenness Centrality	Dictionary
Eigenvector Centrality	Dictionary
Clustering Coefficient	Float
Account Reputation	Float
PageRank	Dictionary
Followers Count	Int
Friends Count	Int
Statuses Count	Int
Favourites Count	Int
Listed Count	Int
Verified	Boolean
Geo_Enabled	Boolean
Default_Profile	Boolean
Default_Profile_Image	Boolean
Location_Present	Boolean
Profile_Location_Present	Boolean
Url_Present	Boolean
Description_Present	Boolean
Description	String
Tweet	List of Strings
Label	Boolean

Πίνακας 5.2: *Παρουσίαση των χαρακτηριστικών που εξήχθησαν από το σύνολο δεδομένων*

Η σημασία των παραπάνω χαρακτηριστικών για τα αποτελέσματα θα αναλυθεί περαιτέρω στο επόμενο κεφάλαιο που περιγράφει τα πειράματα και την αξιολόγηση.

5.4 Παρουσίαση Μοντέλων

Σε αυτή την ενότητα θα περιγραφούν συνοπτικά τα μοντέλα που χρησιμοποιήθηκαν για την ανίχνευση αυτοματοποιημένων λογαριασμών bots στην παρούσα διπλωματική και θα αναλυθεί η καταλληλότητά τους για εφαρμογή στο πρόβλημα.

5.4.1 Logistic Regression

Ο αλγόριθμος μηχανικής μάθησης Logistic Regression χρησιμοποιήθηκε ως το βασικό μοντέλο αναφοράς για την ανίχνευση bots στο δίκτυο χρηστών που προέκυψε από το σύνολο δεδομένων Twibot-20. Αξιοποιεί τόσο τοπολογικές όσο και αριθμητικές πληροφορίες προφίλ των χρηστών για τη δυαδική κατηγοριοποίηση (bot ή άνθρωπος).

Καταλληλότητα Μοντέλου για Εφαρμογή στο Πρόβλημα

Το Logistic Regression είναι κατάλληλο για το πρόβλημα της κατηγοριοποίησης λογαριασμών του Twitter σε bots ή ανθρώπους για αρκετούς λόγους. Πρώτον, η απλότητά του καθιστά εύκολη την ερμηνεία των αποτελεσμάτων και την κατανόηση της επίδρασης κάθε χαρακτηριστικού. Επιπλέον, η αποδοτικότητά του, λόγω της ταχύτητας εκπαίδευσης και πρόβλεψης, το καθιστά ιδανικό για μεγάλα σύνολα δεδομένων, όπως το Twibot-20. Τέλος, λειτουργεί ως μια αρχική προσέγγιση, παρέχοντας ένα σημείο αναφοράς για τη σύγκριση πιο πολύπλοκων μοντέλων, όπως τα Graph Convolutional Networks και τα Graph Attention Networks.

5.4.2 Random Forest

Ο αλγόριθμος μηχανικής μάθησης Random Forest χρησιμοποιήθηκε ως ένα ισχυρό μη γραμμικό μοντέλο για την ανίχνευση bots. Εφαρμόστηκε σε ένα σύνολο χαρακτηριστικών που εξήχθησαν από τη δομή του γράφου και από πληροφορίες από τα προφίλ των χρηστών. Η υλοποίηση αυτή βασίστηκε σε έναν αλγόριθμο που συνδυάζει πολλά δέντρα απόφασης για να βελτιώσει την ακρίβεια και τη γενίκευση του μοντέλου.

Καταλληλότητα Μοντέλου για Εφαρμογή στο Πρόβλημα

Το Random Forest κρίθηκε κατάλληλο για την ανίχνευση bots λόγω της ανθεκτικότητάς του στα θορυβώδη δεδομένα. Η ενσωμάτωση πολλών δέντρων μειώνει την επίδραση των θορυβωδών δεδομένων ή των ανωμαλιών. Επιπλέον, το μοντέλο έχει την ικανότητα να χειρίζεται σύνθετες σχέσεις μεταξύ χαρακτηριστικών, κάτι που είναι κρίσιμο για δεδομένα όπως αυτά που εξάγονται από γράφους. Παρέχει επίσης τη δυνατότητα αξιολόγησης της σχετικής σημασίας κάθε χαρακτηριστικού μέσω του feature importance, ενώ η αρχιτεκτονική του επιτρέπει την εύκολη κλιμάκωση σε μεγαλύτερα σύνολα δεδομένων αυξάνοντας τα δέντρα απόφασης.

5.4.3 Graph Convolutional Network

Το μοντέλο Graph Convolutional Network (GCN) εφαρμόστηκε ως ένα σύγχρονο, state-of-the-art εργαλείο για την ανίχνευση bots, αξιοποιώντας τα δεδομένα του γράφου και τη δομή των σχέσεων μεταξύ χρηστών. Με τη δυνατότητά του να ενσωματώνει πληροφορίες από τη γειτονιά κάθε κόμβου, το GCN είναι ένα εξαιρετικά χρήσιμο εργαλείο για προβλήματα όπου τα δεδομένα έχουν μορφή γραφήματος.

Καταλληλότητα Μοντέλου για Εφαρμογή στο Πρόβλημα

Το GCN είναι ιδανικό για την ανίχνευση bots χάρη στην ικανότητά του να εκμεταλλεύεται τη δομή του κοινωνικού γράφου (σχέσεις follow/following). Ενσωματώνει πληροφορίες από τους γείτονες κάθε κόμβου, κάτι που είναι κρίσιμο για τη διάκριση μεταξύ bots και ανθρώπων. Επίσης, το GCN διαχειρίζεται αποτελεσματικά υψηλής διάστασης δεδομένα, ενσωματώνοντας χαρακτηριστικά όπως followers, friends, verified flags, και άλλα, στο ίδιο πλαίσιο με τη δομή του γράφου.

5.4.4 Graph Attention Network

Το μοντέλο Graph Attention Network (GAT) αποτελεί μία εξελιγμένη μορφή νευρωνικού δικτύου για γράφους που εισάγει την έννοια της προσοχής (attention) στη μάθηση από γειτονικά δεδομένα. Είναι κατάλληλο για προβλήματα που περιλαμβάνουν δεδομένα γραφηματικής μορφής, όπου οι σχέσεις μεταξύ των κόμβων διαδραματίζουν κεντρικό ρόλο.

Καταλληλότητα Μοντέλου για Εφαρμογή στο Πρόβλημα

Το GAT είναι ιδανικό για την ανίχνευση bots λόγω της ικανότητάς του να εκμεταλλεύεται τη δομή του κοινωνικού γράφου, λαμβάνοντας υπόψη τη σημασία των σχέσεων ακολούθησης (follow/following). Η δυναμική εκμάθηση βαρών προσοχής επιτρέπει στο μοντέλο να δίνει μεγαλύτερη έμφαση σε σημαντικούς γείτονες, ενισχύοντας την αποτελεσματικότητα της διάκρισης μεταξύ ανθρώπων και bots. Επιπλέον, το GAT μπορεί να συνδυάσει τοπικά χαρακτηριστικά κάθε κόμβου με πληροφορίες από τη γειτονιά του, ενισχύοντας την ακρίβεια και την προσαρμοστικότητα στις ιδιαιτερότητες του προβλήματος.

5.4.5 Μεγάλα Γλωσσικά Μοντέλα

Τα Μεγάλα Γλωσσικά Μοντέλα, όπως αυτά που χρησιμοποιούνται σε αυτή τη διπλωματική, προσφέρουν προηγμένες δυνατότητες κατανόησης και επεξεργασίας κειμένου, καθιστώντας τα ιδανικά εργαλεία για την ανάλυση περιγραφών και tweets χρηστών, με στόχο τον διαχωρισμό των bots από τους πραγματικούς χρήστες. Η εφαρμογή ενός LLM σε αυτό το πλαίσιο επιτρέπει την αποτελεσματική επεξεργασία δεδομένων φυσικής γλώσσας, εξασφαλίζοντας υψηλή ακρίβεια και προσαρμοστικότητα.

Περιγραφή Λειτουργίας Μοντέλου

Κατά την πραγματοποίηση των πειραμάτων, χρησιμοποιήθηκαν διάφορα γλωσσικά μοντέλα για την ταξινόμηση χρηστών. Κάθε προφίλ αναλύεται με βάση την περιγραφή του (Description) και τα πρώτα 20 tweets του χρήστη. Τα tweets χωρίζονται σε μικρότερα πακέτα (batches) ή αποστέλλονται όλα μαζί για επεξεργασία. Το μοντέλο λειτουργεί με την εξής λογική: πρώτα, τα tweets διαχωρίζονται σε πακέτα των 3-4 tweets ή στέλνονται όλα μαζί για τον χρήστη που εξετάζεται. Στη συνέχεια, χρησιμοποιείται ειδικά σχεδιασμένο prompt, ώστε το LLM να κάνει μια απλή ταξινόμηση (Bot/Human). Όταν τα tweets αποστέλλονται σε batches, οι προβλέψεις για κάθε πακέτο συγκεντρώνονται και εφαρμόζεται κανόνας πλειοψηφίας (majority voting) για την τελική πρόβλεψη.

Καταλληλότητα Μοντέλου για Εφαρμογή στο Πρόβλημα

Τα LLMs θεωρούνται αρκετά ικανά για την ανίχνευση bots. Η ικανότητά τους να κατανοούν και να αναλύουν κείμενα, όπως περιγραφές και tweets, επιτρέπει τη διάκριση μεταξύ ανθρώπινων και μη ανθρώπινων προφίλ. Επιπλέον, η ευελιξία τους στην επεξεργασία δεδομένων τους δίνει τη δυνατότητα να προσαρμόζονται σε διαφορετικές εισόδους και μοτίβα γλώσσας, βελτιώνοντας την ακρίβεια των αποτελεσμάτων.

Κεφάλαιο 6

Πειράματα και Αποτελέσματα

Σε αυτό το κεφάλαιο παρουσιάζονται τα πειράματα που διεξήχθησαν για την αξιολόγηση των μοντέλων που αναπτύχθηκαν, καθώς και τα αποτελέσματα που προέκυψαν. Αρχικά, γίνεται μια εισαγωγή στις μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση της απόδοσης, με έμφαση στην ερμηνεία τους και την καταλληλότητά τους για το πρόβλημα. Στη συνέχεια, περιγράφεται λεπτομερώς η υλοποίηση κάθε μοντέλου και τα μέσα που απαιτήθηκαν για αυτή, ενώ γίνεται και αναλυτική παρουσίαση των αποτελεσμάτων που επιτεύχθηκαν. Ακολουθεί η συγκριτική ανάλυση μεταξύ των διαφορετικών μοντέλων, και κυρίως η σύγκριση της αποτελεσματικότητας των μοντέλων μηχανικής μάθησης και των μεγάλων γλωσσικών μοντέλων (LLMs) στην ανίχνευση αυτοματοποιημένων λογαριασμών. Τέλος, ερμηνεύονται τα αποτελέσματα, επισημαίνοντας τα πλεονεκτήματα και τα μειονεκτήματα κάθε προσέγγισης, καθώς και τη σημασία των ευρημάτων για την ανάλυση του προβλήματος.

6.1 Επιλογή Μετρικών Αξιολόγησης

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν τέσσερις βασικές μετρικές: accuracy, precision, recall και F1 score. Κάθε μία από αυτές τις μετρικές επιλέχθηκε λόγω της δυνατότητάς της να προσφέρει συγκεκριμένες πληροφορίες σχετικά με την απόδοση των μοντέλων στο πρόβλημα της διάκρισης των λογαριασμών που ανήκουν σε bots και ανθρώπους χρήστες σε δίκτυα του Twitter. Παρακάτω περιγράφεται ο ρόλος κάθε μετρικής και η σημασία της. [78] [79]

Accuracy

Η μετρική accuracy είναι ο λόγος των σωστών προβλέψεων προς το συνολικό αριθμό των δειγμάτων.

$$\text{Accuracy} = \frac{\text{Σωστές προβλέψεις}}{\text{Συνολικά Δείγματα}}$$

- Υψηλή τιμή: Υποδεικνύει ότι το μοντέλο προβλέπει σωστά για το μεγαλύτερο μέρος των δεδομένων.
- Χαμηλή τιμή: Μπορεί να υποδηλώνει ότι το μοντέλο δυσκολεύεται να γενικεύσει ή ότι τα δεδομένα είναι μη ισορροπημένα.

Η μετρική accuracy επιλέχθηκε διότι παρέχει μια γενική εικόνα της απόδοσης, αλλά μπορεί να είναι παραπλανητική σε περιπτώσεις με ανισορροπία στις κλάσεις στο σύνολο δεδομένων.

Precision

Η μετρική precision είναι το ποσοστό των σωστών θετικών προβλέψεων επί των συνολικών θετικών προβλέψεων.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

- Υψηλή τιμή: Υποδεικνύει ότι το μοντέλο έχει λίγα false positives, δηλαδή δεν χαρακτηρίζει λανθασμένα ανθρώπους ως bots.
- Χαμηλή τιμή: Σημαίνει ότι το μοντέλο κάνει πολλές λανθασμένες θετικές προβλέψεις.

Η μετρική precision είναι κρίσιμη σε εφαρμογές όπου η λανθασμένη κατηγοριοποίηση ενός ατόμου ως bot μπορεί να έχει σημαντικές συνέπειες.

Recall

Η μετρική recall είναι το ποσοστό των σωστών θετικών προβλέψεων επί των συνολικών πραγματικών θετικών δειγμάτων.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

- Υψηλή τιμή: Υποδεικνύει ότι το μοντέλο αναγνωρίζει το μεγαλύτερο μέρος των πραγματικών bots.
- Χαμηλή τιμή: Σημαίνει ότι το μοντέλο αποτυγχάνει να ανιχνεύσει πολλά πραγματικά θετικά δείγματα.

Η μετρική recall είναι σημαντική σε περιπτώσεις όπου το να μην εντοπιστεί ένα bot μπορεί να οδηγήσει σε σοβαρές επιπτώσεις.

F1 Score

Το F1 score είναι ο αρμονικός μέσος της μετρικής precision και της recall.

$$F1\text{-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision + Recall}}$$

- Υψηλή τιμή: Υποδεικνύει καλή ισορροπία μεταξύ precision και recall.
- Χαμηλή τιμή: Σημαίνει ότι υπάρχει ανισορροπία ή χαμηλές τιμές και στις δύο μετρικές.

Η μετρική F1 score επιλέχθηκε γιατί είναι κατάλληλη για περιπτώσεις με ανισορροπία στις κλάσεις και δίνει μια συνολική εικόνα της απόδοσης.

Επιλογή Μετρικών

Οι παραπάνω μετρικές επιλέχθηκαν επειδή παρέχουν μια ολοκληρωμένη εικόνα της απόδοσης των μοντέλων. Ενώ η μετρική accuracy προσφέρει μια γενική εικόνα, οι μετρικές precision και recall είναι πιο κατάλληλες για τη μέτρηση της απόδοσης σε μη ισορροπημένα δεδομένα (όπως, εν μέρει, μπορούν να θεωρηθούν και τα δεδομένα της παρούσας έρευνας). Τέλος, το F1 score εξισορροπεί τις δύο τελευταίες μετρικές, παρέχοντας μια πιο αξιόπιστη αξιολόγηση σε προβλήματα όπου τόσο τα false positives όσο και τα false negatives είναι σημαντικά.

6.2 Υλοποίηση Μοντέλων και Αποτελέσματα

Σε αυτή την ενότητα γίνεται παρουσίαση της υλοποίησης των μοντέλων που αναπτύχθηκαν με σκοπό την ανίχνευση αυτοματοποιημένων λογαριασμών bot. Επίσης, δίνονται τα αποτελέσματα που επιτεύχθηκαν με κάθε μοντέλο για τα επιλεγμένα σύνολα δεδομένων.

Σε αυτό το σημείο πρέπει να αναφερθεί πως όλα τα μοντέλα αναπτύχθηκαν με την γλώσσα προγραμματισμού Python (έκδοση 3.12.5) και η ανάπτυξη και αξιολόγησή τους έγινε στο περιβάλλον του Jupyter Notebook, είτε τοπικά, είτε στο Google Colab που προσφέρει την υπηρεσία του Jupyter Notebook με παροχή ελεύθερης πρόσβασης σε υπολογιστικούς πόρους online. Χρησιμοποιήθηκε ένας υπολογιστής με επεξεργαστή Intel i5 11ης γενιάς, 8GB μνήμη RAM και λειτουργικό σύστημα Microsoft Windows 11.

6.2.1 Logistic Regression

Σε αυτή την ενότητα θα δοθούν όλες οι λεπτομέρειες υλοποίησης του μοντέλου Logistic Regression.

Χαρακτηριστικά που αξιοποιήθηκαν

Τα χαρακτηριστικά (features) που αξιοποιήθηκαν για την εκπαίδευση του μοντέλου Logistic Regression περιλάμβαναν τόσο τοπολογικά χαρακτηριστικά όσο και αριθμητικά χαρακτηριστικά, που προέρχονταν απευθείας από τα υποσύνολα του Twibot-20 (πχ. Followers Count) ή και από υπολογισμούς πριν την έναρξη της διαδικασίας εκπαίδευσης (πχ. κεντρικότητες) στο μοντέλο. Παρουσιάζονται αναλυτικά στον παρακάτω πίνακα (Πίνακας 6.1).

Πρέπει να αναφερθεί το γεγονός πως, αρχικά, είχαν επιλεχθεί περισσότερα χαρακτηριστικά (πχ. Closeness Centrality). Ωστόσο, εφαρμόστηκε μελέτη απόρριψης χαρακτηριστικών, η οποία υπέδειξε πως ορισμένα χαρακτηριστικά δεν συνεισφέρουν θετικά στην απόδοση του μοντέλου. Η μελέτη απόρριψης χαρακτηριστικών (Feature Ablation Study) είναι μια τεχνική στη μηχανική μάθηση που χρησιμοποιείται για την αξιολόγηση της σημασίας των χαρακτηριστικών ενός μοντέλου. Αφαιρώντας διαδοχικά ή ομαδικά συγκεκριμένα χαρακτηριστικά από τα δεδομένα εισόδου, εξετάζεται η επίδραση αυτών στην απόδοση του μοντέλου. Με αυτόν τον τρόπο, εντοπίζονται τα χαρακτηριστικά που δεν συνεισφέρουν θετικά ή είναι περιττά για τη βελτίωση της ακρίβειας ή της γενίκευσης του μοντέλου. Συνεπώς, στο συγκεκριμένο

Χαρακτηριστικό	**Είδος Χαρακτηριστικού**	**Τύπος Δεδομένων**
ID	Μοναδικό αναγνωριστικό χρήστη	Int64
Following	Τοπολογικό	List of IDs
Follower	Τοπολογικό	List of IDs
Degree Centrality	Μετρική Δικτύου	Dictionary
Eigenvector Centrality	Μετρική Δικτύου	Dictionary
Clustering Coefficient	Μετρική Δικτύου	Float
Account Reputation	Τοπολογικό	Float
PageRank	Μετρική Δικτύου	Dictionary
Followers Count	Αριθμητικό	Int
Friends Count	Αριθμητικό	Int
Statuses Count	Αριθμητικό	Int
Favourites Count	Αριθμητικό	Int
Verified	Αριθμητικό	Boolean
Geo_Enabled	Αριθμητικό	Boolean
Default_Profile	Αριθμητικό	Boolean
Default_Profile_Image	Αριθμητικό	Boolean
Location_Present	Αριθμητικό	Boolean
Profile_Location_Present	Αριθμητικό	Boolean
Url_Present	Αριθμητικό	Boolean
Description_Present	Αριθμητικό	Boolean
Label	Ground Truth	Boolean

Πίνακας 6.1: *Χαρακτηριστικά που αξιοποίησε το Logistic Regression*

μοντέλο, αυτά τα χαρακτηριστικά εντοπίστηκαν και αφαιρέθηκαν. Και, έτσι, προέκυψαν τα παραπάνω χαρακτηριστικά που παρουσιάστηκαν.

Στον παρακάτω πίνακα (Πίνακας 6.2) παρουσιάζεται η απόδοση του μοντέλου κατά τη μελέτη απόρριψης χαρακτηριστικών σε ένα υποσύνολο δεδομένων του Twibot-20. Είναι σημαντικό να τονιστεί πως η παρακάτω αξιολόγηση γίνεται σε σύνολο δεδομένων διαφορετικό από αυτά στα οποία γίνεται η συνολική αξιολόγηση παρακάτω.

Υλοποίηση

Η υλοποίηση του μοντέλου Logistic Regression ξεκίνησε με την επεξεργασία των δεδομένων που περιλαμβάνονται σε ένα αρχείο JSON που δίνεται ως είσοδος και αποτελεί το τρέχον σύνολο δεδομένων για εκπαίδευση και αξιολόγηση. Τα δεδομένα φορτώθηκαν και μετατράπηκαν σε κατευθυνόμενο γράφο, χρησιμοποιώντας τη βιβλιοθήκη NetworkX. Σε αυτόν τον γράφο, κάθε κόμβος αντιπροσωπεύει έναν χρήστη, με τις ακμές να υποδεικνύουν σχέσεις ακολούθων ή ακολουθούμενων χρηστών. Οι ετικέτες των κόμβων, οι οποίες χρησιμοποιήθηκαν ως εξαρτημένες μεταβλητές (labels), αναπαριστούσαν τις κατηγορίες προς πρόβλεψη.

Αφού δημιουργήθηκε ο γράφος, υπολογίστηκαν τοπολογικά χαρακτηριστικά γραφη-

Χαρακτηριστικό	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Baseline	0.6914	0.6073	0.9251	0.7280
Degree Centrality	0.6373	0.5613	0.8114	0.6752
Closeness Centrality	0.6971	0.6142	0.9281	0.7362
Betweenness Centrality	0.6942	0.6109	0.9336	0.7339
Eigenvector Centrality	0.6832	0.5989	0.9144	0.7153
Clustering Coefficient	0.6903	0.6053	0.9182	0.7130
Account Reputation	0.6907	0.6100	0.8783	0.7126
PageRank	0.6820	0.6027	0.9173	0.7198
Followers Count	0.6856	0.5890	0.8716	0.6949
Friends Count	0.6723	0.5764	0.8665	0.6893
Favourites Count	0.6801	0.5998	0.8882	0.7102
Statuses Count	0.6838	0.6021	0.8971	0.7182
Listed Count	0.6999	0.6141	0.9273	0.7355
Verified	0.6909	0.5841	0.9012	0.6957
Geo_Enabled	0.6901	0.6007	0.9211	0.7221
Default_Profile	0.6895	0.5964	0.9123	0.7195
Default_Profile_Image	0.6876	0.5903	0.9183	0.7134
Location_Present	0.6891	0.5927	0.9192	0.7104
Profile_Location_Present	0.6863	0.5879	0.9232	0.7277
Url_Present	0.6912	0.5997	0.9185	0.7264
Description_Present	0.6900	0.6044	0.9184	0.7238

Πίνακας 6.2: *Απόδοση του Logistic Regression κατά τη μελέτη απόρριψης χαρακτηριστικών*

μάτων όπως η κεντρικότητα βαθμού (degree centrality), ο συντελεστής συσσωμάτωσης (clustering coefficient), το PageRank, και η κεντρικότητα ιδιοδιανύσματος (eigenvector centrality). Αυτά τα χαρακτηριστικά αποτέλεσαν το θεμέλιο της ανάλυσης, καθώς συλλαμβάνουν βασικές ιδιότητες του γράφου που σχετίζονται με τη δομή του κοινωνικού δικτύου. Επιπλέον, εξήχθησαν χαρακτηριστικά από τα προφίλ των χρηστών, όπως ο αριθμός ακολούθων, ο αριθμός tweets, το αν ο λογαριασμός ήταν επαληθευμένος, και άλλα δημογραφικά δεδομένα. Όλα τα παραπάνω συνδυάστηκαν σε έναν ενιαίο πίνακα χαρακτηριστικών.

Η επόμενη φάση περιλάμβανε την κανονικοποίηση των δεδομένων. Δεδομένου ότι τα χαρακτηριστικά παρουσιάζουν διαφορετικές κλίμακες τιμών (π.χ. αριθμός ακολούθων σε σύγκριση με δυαδικές μεταβλητές), εφαρμόστηκε η μέθοδος StandardScaler της βιβλιοθήκης scikit-learn. Η κανονικοποίηση διασφάλισε ότι όλα τα χαρακτηριστικά θα συνέβαλαν ισοδύναμα στην απόδοση του μοντέλου, μειώνοντας τον κίνδυνο υπερβολικής επιρροής χαρακτηριστικών με μεγάλες τιμές.

Για την αντιμετώπιση του προβλήματος της ανισορροπίας των κλάσεων υπολογίστηκαν βάρη για κάθε κλάση μέσω της μεθόδου compute_class_weight. Αυτά τα βάρη εισήχθησαν ως υπερπαράμετρος στο Logistic Regression, ώστε να ενισχυθεί η συμβολή των λιγότερο αντιπροσωπευτικών κλάσεων κατά την εκπαίδευση.

Ο διαχωρισμός των δεδομένων πραγματοποιήθηκε σε σύνολα εκπαίδευσης (80%) και

δοκιμής (20%). Το σύνολο εκπαίδευσης χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου και τη ρύθμιση των παραμέτρων του, ενώ το σύνολο δοκιμής παρέμεινε ανεξάρτητο για την τελική αξιολόγηση της απόδοσης.

Το Logistic Regression μοντέλο εκπαιδεύτηκε με χρήση του αλγορίθμου μέγιστης πιθανοφάνειας, με τον αριθμό επαναλήψεων (max_iter) να έχει οριστεί στις 1000. Αυτό διασφάλισε τη σύγκλιση του μοντέλου, ακόμη και με τα εκτεταμένα χαρακτηριστικά που περιλαμβάνονταν στα δεδομένα. Η επιλογή αυτής της υπερπαραμέτρου βασίστηκε σε πειραματική ανάλυση, κατά την οποία μικρότεροι αριθμοί επαναλήψεων δεν παρείχαν ικανοποιητικά αποτελέσματα.

Το εκπαιδευμένο μοντέλο αξιολογήθηκε μέσω διασταυρούμενης επικύρωσης (cross-validation), χρησιμοποιώντας πέντε πτυχές (5-fold cross-validation). Αυτή η διαδικασία εξασφάλισε ότι τα αποτελέσματα του μοντέλου ήταν σταθερά και δεν επηρεάζονταν υπερβολικά από συγκεκριμένα υποσύνολα δεδομένων. Οι μετρικές που χρησιμοποιήθηκαν για την αξιολόγηση περιλάμβαναν τις accuracy, precision, recall και την F1 score.

Η διαδικασία υλοποίησης και επιλογής υπερπαραμέτρων σχεδιάστηκε προσεκτικά για να διασφαλιστεί η βέλτιστη απόδοση του Logistic Regression. Η εστίαση στη αποτελεσματική διαχείριση των χαρακτηριστικών, της κανονικοποίησης και της ανισορροπίας των κλάσεων αποτέλεσε κύριο στόχο της προσέγγισης που ακολουθήθηκε.

Αποτελέσματα

Το μοντέλο που αναπτύχθηκε για το Logistic Regression τεσταρίστηκε σε τρία δίκτυα - σύνολα δεδομένων, διαφορετικού μεγέθους ώστε να παρατηρηθεί η εξέλιξη της απόδοσης των επιλεγμένων μετρικών αξιολόγησης ανάλογα με το πλήθος των δειγμάτων που λαμβάνει το μοντέλο κατά την διαδικασία εκπαίδευσης και αξιολόγησης.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα στα τρία διαφορετικά δίκτυα:

Network	**Nodes**	**Bots**	**Humans**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
smallMLdataset	1448	524	924	0.7103	0.5902	0.9231	0.7200
midMLdataset	3756	1801	1955	0.6902	0.6142	0.9239	0.7379
bigMLdataset	6316	3524	2792	0.7035	0.6719	0.9475	0.7862

Πίνακας 6.3: *Αποτελέσματα του μοντέλου Logistic Regression*

Η ερμηνεία των παραπάνω αποτελεσμάτων παρουσιάζεται στην επόμενη ενότητα αυτού του κεφαλαίου.

6.2.2 Random Forest

Σε αυτή την ενότητα θα δοθούν όλες οι λεπτομέρειες υλοποίησης του μοντέλου Random Forest που αναπτύχθηκε.

Χαρακτηριστικά που αξιοποιήθηκαν

Τα χαρακτηριστικά (features) που αξιοποιήθηκαν για την εκπαίδευση του μοντέλου Random Forest περιλάμβαναν τόσο τοπολογικά χαρακτηριστικά όσο και αριθμητικά χαρακτηριστικά, που προέρχονταν απευθείας από τα υποσύνολα του Twibot-20 (πχ. Statuses Count) ή και από υπολογισμούς πριν την έναρξη της διαδικασίας εκπαίδευσης (πχ. κεντρικότητες) στο μοντέλο. Παρουσιάζονται αναλυτικά στον παρακάτω πίνακα (Πίνακας 6.4).

Χαρακτηριστικό	**Είδος Χαρακτηριστικού**	**Τύπος Δεδομένων**
ID	Μοναδικό αναγνωριστικό χρήστη	Int64
Following	Τοπολογικό	List of IDs
Follower	Τοπολογικό	List of IDs
Degree Centrality	Μετρική Δικτύου	Dictionary
Eigenvector Centrality	Μετρική Δικτύου	Dictionary
Clustering Coefficient	Μετρική Δικτύου	Float
Account Reputation	Τοπολογικό	Float
PageRank	Μετρική Δικτύου	Dictionary
Followers Count	Αριθμητικό	Int
Friends Count	Αριθμητικό	Int
Statuses Count	Αριθμητικό	Int
Favourites Count	Αριθμητικό	Int
Verified	Αριθμητικό	Boolean
Geo_Enabled	Αριθμητικό	Boolean
Default_Profile	Αριθμητικό	Boolean
Default_Profile_Image	Αριθμητικό	Boolean
Location_Present	Αριθμητικό	Boolean
Profile_Location_Present	Αριθμητικό	Boolean
Url_Present	Αριθμητικό	Boolean
Description_Present	Αριθμητικό	Boolean
Label	Ground Truth	Boolean

Πίνακας 6.4: *Χαρακτηριστικά που αξιοποίησε το Random Forest*

Τα χαρακτηριστικά είναι τα ίδια που εφαρμόστηκαν και για την εκπαίδευση του μοντέλου Logistic Regression. Όπως και σε εκείνο, έτσι και εδώ, αρχικά είχαν επιλεχθεί περισσότερα χαρακτηριστικά (πχ. Betweenness Centrality). Μετά, όμως, από την πραγματοποίηση μελέτης απόρριψης χαρακτηριστικών (Feature Ablation Study), εντοπίστηκαν τα χαρακτηριστικά που δεν επιδρούν θετικά στην απόδοση του μοντέλου και αφαιρέθηκαν, καταλήγοντας, έτσι, στα παραπάνω.

Υλοποίηση

Η διαδικασία υλοποίησης του μοντέλου Random Forest ξεκίνησε με τη φόρτωση των δεδομένων από ένα αρχείο JSON (υποσύνολο του δικτύου Twibot-20), το οποίο περιείχε πληροφορίες για τους χρήστες και τις σχέσεις μεταξύ τους. Αυτά τα δεδομένα μετατράπηκαν

σε έναν κατευθυνόμενο γράφο (directed graph) χρησιμοποιώντας τη βιβλιοθήκη NetworkX. Σε αυτό το δίκτυο, κάθε κόμβος αντιστοιχούσε σε έναν χρήστη, ενώ οι κατευθυνόμενες ακμές υποδήλωναν σχέσεις ακολούθησης (follower και following). Επιπλέον, κάθε κόμβος είχε μια ετικέτα (label) που αντιπροσωπεύει την κατηγορία του (bot ή άνθρωπος) προς πρόβλεψη.

Για την εκπαίδευση του μοντέλου εξήχθησαν χαρακτηριστικά τόσο τοπολογικά, δηλαδή προερχόμενα από τη δομή του γράφου, όσο και αριθμητικά, από τα δεδομένα που αφορούν το προφίλ των χρηστών. Από τον γράφο εξήχθησαν μετρικές όπως η κεντρικότητα βαθμού, ο συντελεστής συσσωμάτωσης (clustering coefficient), το PageRank, και η κεντρικότητα ιδιοδιανύσματος (eigenvector centrality). Τα δεδομένα των χρηστών περιλάμβαναν τον αριθμό των ακολούθων, τον αριθμό των φίλων, τον αριθμό των tweets, και χαρακτηριστικά όπως το αν ο λογαριασμός είναι επαληθευμένος ή αν υπάρχει περιγραφή στο προφίλ κλπ. Όλα τα παραπάνω συνδυάστηκαν σε έναν ενιαίο πίνακα χαρακτηριστικών. Όπως αναφέρθηκε και προηγουμένως, για τη βελτίωση της απόδοσης του μοντέλου, εφαρμόστηκε η μελέτη απόρριψης χαρακτηριστικών (Feature Ablation Study), όπου χαρακτηριστικά που δεν συνέβαλαν θετικά στα αποτελέσματα αφαιρέθηκαν.

Η επόμενη φάση περιλάμβανε τον διαχωρισμό των δεδομένων σε τρία σύνολα: εκπαίδευσης, επικύρωσης, και δοκιμής. Ο διαχωρισμός αυτός διασφαλίζει την αντικειμενικότητα στην αξιολόγηση του μοντέλου. Πιο συγκεκριμένα, το σύνολο εκπαίδευσης χρησιμοποιήθηκε για τη δημιουργία του μοντέλου, ενώ το σύνολο επικύρωσης αξιοποιήθηκε για τη βελτιστοποίηση των υπερπαραμέτρων. Τέλος, το σύνολο δοκιμής παρέμεινε ανεξάρτητο για την τελική αξιολόγηση. Πριν την εισαγωγή των δεδομένων στο μοντέλο, τα χαρακτηριστικά κανονικοποιήθηκαν με τη χρήση του StandardScaler της scikit-learn, ώστε να διασφαλιστεί ότι όλα τα χαρακτηριστικά συνέβαλαν ισότιμα στη διαδικασία εκπαίδευσης.

Για την εκπαίδευση του μοντέλου, χρησιμοποιήθηκε ο αλγόριθμος Random Forest μέσω της κλάσης RandomForestClassifier της scikit-learn. Ο αριθμός των δέντρων ορίστηκε σε 100 (n_estimators=100), ενώ χρησιμοποιήθηκε συγκεκριμένος σπόρος τυχαιοποίησης (random_state=42) για να διασφαλιστεί η αναπαραγωγιμότητα των αποτελεσμάτων. Επιπλέον, εφαρμόστηκε διασταυρούμενη επικύρωση (cross-validation) με χρήση της μεθόδου StratifiedKFold, όπου τα δεδομένα χωρίστηκαν σε πέντε υποσύνολα (n_splits=5) και οι ετικέτες διατηρήθηκαν ισορροπημένες σε κάθε υποσύνολο.

Κατά τη διάρκεια της εκπαίδευσης, το μοντέλο εκπαιδεύτηκε επανειλημμένα με διαφορετικά υποσύνολα δεδομένων, και τα αποτελέσματα αξιολογήθηκαν μέσω των μετρικών accuracy, precision, recall, και F1 score. Οι υπερπαράμετροι του μοντέλου ρυθμίστηκαν με τρόπο τέτοιο ώστε να επιτευχθεί η βέλτιστη απόδοση, ενώ η διασταυρούμενη επικύρωση συνέβαλε στη μείωση της πιθανότητας υπερεκπαίδευσης (overfitting).

Αποτελέσματα

Το μοντέλο που αναπτύχθηκε για το Random Forest τεσταρίστηκε σε τρία δίκτυα - σύνολα δεδομένων, διαφορετικού μεγέθους ώστε να παρατηρηθεί η εξέλιξη της απόδοσης των επιλεγμένων μετρικών αξιολόγησης ανάλογα με το πλήθος των δειγμάτων που λαμβάνει το μοντέλο κατά την διαδικασία εκπαίδευσης και αξιολόγησης.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα στα τρία διαφορετικά δίκτυα:

Network	Nodes	Bots	Humans	Accuracy	Precision	Recall	F1 Score
smallMLdataset	1448	524	924	0.7103	0.5902	0.9231	0.7200
midMLdataset	3756	1801	1955	0.6902	0.6142	0.9239	0.7379
bigMLdataset	6316	3524	2792	0.7035	0.6719	0.9475	0.7862

Πίνακας 6.5: *Αποτελέσματα του μοντέλου Random Forest*

Η ερμηνεία των παραπάνω αποτελεσμάτων παρουσιάζεται στην επόμενη ενότητα αυτού του κεφαλαίου.

6.2.3 Graph Convolutional Network

Σε αυτή την ενότητα θα δοθούν όλες οι λεπτομέρειες της υλοποίησης του μοντέλου Graph Convolutional Network που αναπτύχθηκε για την ανίχνευση αυτοματοποιημένων λογαριασμών σε κοινωνικά δίκτυα.

Χαρακτηριστικά που Αξιοποιήθηκαν

Τα χαρακτηριστικά (features) που αξιοποιήθηκαν για την εκπαίδευση του συνελικτικού μοντέλου Graph Convolutional Network (Πίνακας 6.6) περιλάμβαναν τοπολογικά χαρακτηριστικά από το δίκτυο, αλλά και αριθμητικά χαρακτηριστικά από τα προφίλ των χρηστών. Πριν εισαχθούν στο GCN για την εκπαίδευση του μοντέλου, κανονικοποιήθηκαν.

Χαρακτηριστικό	Είδος Χαρακτηριστικού	Τύπος Δεδομένων
ID	Μοναδικό αναγνωριστικό χρήστη	Int64
Following	Τοπολογικό	List of IDs
Follower	Τοπολογικό	List of IDs
Degree Centrality	Μετρική Δικτύου	Dictionary
Closeness Centrality	Μετρική Δικτύου	Dictionary
Betweenness Centrality	Μετρική Δικτύου	Dictionary
Clustering Coefficient	Μετρική Δικτύου	Float
Followers Count	Αριθμητικό	Int
Friends Count	Αριθμητικό	Int
Favourites Count	Αριθμητικό	Int
Verified	Αριθμητικό	Boolean
Geo_Enabled	Αριθμητικό	Boolean
Default_Profile	Αριθμητικό	Boolean
Default_Profile_Image	Αριθμητικό	Boolean
Location_Present	Αριθμητικό	Boolean
Profile_Location_Present	Αριθμητικό	Boolean
Url_Present	Αριθμητικό	Boolean
Description_Present	Αριθμητικό	Boolean
Label	Ground Truth	Boolean

Πίνακας 6.6: *Χαρακτηριστικά που αξιοποίησε το μοντέλο Graph Convolutional Network*

Όπως και στα δύο προηγούμενα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν, υλοποιήθηκε μελέτη απόρριψης χαρακτηριστικών (Feature Ablation Study). Εδώ θα δοθούν, ενδεικτικά, περισσότερες λεπτομέρειες για αυτή την ανάλυση.

Η επαναληπτική διαδικασία που ακολουθήθηκε για την απόρριψη χαρακτηριστικών είναι η εξής: Από το σύνολο χαρακτηριστικών (feature set), αφαιρείται ένα χαρακτηριστικό και επανεκπαιδεύεται το μοντέλο GCN με αυτό το ¨ελαττωμένο¨ σύνολο χαρακτηριστικών. Έπειτα, μετά την εκπαίδευσή του, η απόδοση του μοντέλου αξιολογείται στο test set και καταγράφονται τα αποτελέσματα που δίνει για τις μετρικές αξιολόγησης χωρίς αυτό το χαρακτηριστικό. Αυτά τα αποτελέσματα συγκρίνονται με την baseline απόδοση, δηλαδή όταν συμμετέχουν όλα τα χαρακτηριστικά, και αποφασίζεται αν η έλλειψη του συγκεκριμένου χαρακτηριστικού επιδρά θετικά η αρνητικά στην απόδοση του μοντέλου. Αυτή η διαδικασία, γίνεται επαναληπτικά για όλα τα χαρακτηριστικά και στο τέλος δίνονται τα χαρακτηριστικά, η έλλειψη των οποίων οδηγεί σε υψηλότερες επιδόσεις του μοντέλου, ώστε να αναγνωριστούν και να αφαιρεθούν από μελλοντικά πειράματα.

Ακολουθώντας την παραπάνω ανάλυση, για το μοντέλο GCN, παρατηρήθηκε πως τα χαρακτηριστικά Statuses Count, Listed Count, Account Reputation και Eigenvector Centrality, δεν επιδρούν θετικά κατά τη διαδικασία εκπαίδευσης και, συνεπώς, αφαιρέθηκαν από αυτή.

Στον πίνακα 6.7 παρουσιάζεται η απόδοση του μοντέλου σε ένα υποσύνολο δεδομένων του Twibot-20. Είναι σημαντικό να τονιστεί πως η συγκεκριμένη αξιολόγηση γίνεται σε σύνολο δεδομένων διαφορετικό από αυτά στα οποία γίνεται η συνολική αξιολόγηση παρακάτω. Επίσης, πρόκειται για το μοντέλο GCN πριν υποβληθεί σε διασταυρούμενη επικύρωση.

Υλοποίηση

Η υλοποίηση του μοντέλου Graph Convolutional Network (GCN) ξεκίνησε με την προετοιμασία των δεδομένων. Αρχικά, τα δεδομένα φορτώθηκαν από ένα αρχείο JSON, που αναπαριστά ένα δίκτυο χρηστών του Twitter, και επεξεργάστηκαν ώστε να μετατραπούν σε μορφή συμβατή με το PyTorch Geometric. Κάθε χρήστης αντιστοιχούσε σε έναν κόμβο του γράφου, και οι σχέσεις ακολούθησης (follower και following) μεταξύ χρηστών καθόρισαν τις ακμές του γράφου. Από κάθε προφίλ χρήστη εξήχθησαν χαρακτηριστικά, όπως, μεταξύ άλλων, ο αριθμός των followers, friends, favourites και άλλες πληροφορίες όπως το αν έχει γίνει η επαλήθευση του λογαριασμού (verified). Παράλληλα, χαρακτηριστικά όπως η χρήση προεπιλεγμένου προφίλ ή εικόνας, καθώς και η παρουσία τοποθεσίας, περιγραφής ή URL στο προφίλ, συμπεριλήφθηκαν για την καλύτερη μοντελοποίηση της συμπεριφοράς του χρήστη.

Για την ενίσχυση των χαρακτηριστικών, χρησιμοποιήθηκαν μετρικές κεντρικότητας που υπολογίστηκαν με τη βοήθεια του networkx. Τα χαρακτηριστικά αυτά προστέθηκαν στα αρχικά χαρακτηριστικά των κόμβων, παρέχοντας ένα εμπλουτισμένο σύνολο δεδομένων. Τα τελικά δεδομένα μετατράπηκαν σε τανυστές PyTorch: ο τανυστής x περιλάμβανε τα χαρακτηριστικά των κόμβων, ο y τις ετικέτες (bot/human), και ο edge_index περιέγραφε τις ακμές του γράφου.

Η αρχιτεκτονική του GCN σχεδιάστηκε ώστε να αποτελείται από διαδοχικά επίπεδα

Χαρακτηριστικό	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Baseline	0.770	0.6619	0.8392	0.7401
Degree Centrality	0.6284	0.5714	0.0714	0.1270
Closeness Centrality	0.6351	0.5833	0.1250	0.2059
Betweenness Centrality	0.7605	0.6866	0.8114	0.7380
Eigenvector Centrality	0.7856	0.6790	0.8267	0.7564
Clustering Coefficient	0.7703	0.6579	0.8294	0.7376
Account Reputation	0.7905	0.6812	0.8393	0.7520
PageRank	0.7798	0.6726	0.8402	0.7675
Followers Count	0.7432	0.6154	0.8571	0.7164
Friends Count	0.6587	0.5851	0.0967	0.1376
Favourites Count	0.6216	0.5987	0.3453	0.4562
Statuses Count	0.7832	0.6851	0.8374	0.7538
Listed Count	0.7784	0.6712	0.8305	0.7602
Verified	0.6254	0.5838	0.0842	0.1386
Geo_Enabled	0.7560	0.6456	0.8275	0.7309
Default_Profile	0.7689	0.6590	0.8298	0.7309
Default_Profile_Image	0.7578	0.6489	0.8376	0.7284
Location_Present	0.7703	0.6602	0.8187	0.7294
Profile_Location_Present	0.7309	0.6530	0.8245	0.7413
Url_Present	0.7683	0.6478	0.8294	0.7206
Description_Present	0.7498	0.6587	0.8201	0.7235

Πίνακας 6.7: *Απόδοση του Graph Convolutional Network κατά τη μελέτη απόρριψης χαρακτηριστικών*

Graph Convolutional Network (GCNConv). Το πρώτο εισαγωγικό επίπεδο (input layer) λαμβάνει ως είσοδο τα χαρακτηριστικά των κόμβων και τα μετασχηματίζει σε έναν κρυφό χώρο χαρακτηριστικών. Ενδιάμεσα επίπεδα (hidden layers) χρησιμοποιήθηκαν για την περαιτέρω επεξεργασία των χαρακτηριστικών, με την εφαρμογή ενεργοποιητή ReLU μετά από κάθε συνελικτικό επίπεδο, για την εισαγωγή μη γραμμικότητας. Επιπλέον, ενσωματώθηκε batch normalization για τη σταθεροποίηση της εκπαίδευσης, καθώς και dropout με ποσοστό 0.3, για την αποφυγή υπερεκπαίδευσης. Το τελικό επίπεδο του GCN παρήγαγε προβλέψεις κατηγοριοποίησης για κάθε κόμβο σε μία από τις δύο κλάσεις (bot ή human).

Η εκπαίδευση του μοντέλου πραγματοποιήθηκε μέσω 10-fold διασταυρούμενης επικύρωσης (10-fold cross-validation) με τη χρήση της StratifiedKFold. Για κάθε fold, το σύνολο δεδομένων χωρίστηκε σε σύνολα εκπαίδευσης και δοκιμών με διατήρηση της αναλογίας μεταξύ των κατηγοριών. Το μοντέλο εκπαιδεύτηκε για 100 εποχές σε κάθε fold, με τη συνάρτηση απώλειας cross-entropy loss να βελτιστοποιείται μέσω του αλγορίθμου AdamW. Ο ρυθμός μάθησης ορίστηκε στο 0.01, ενώ η αποσύνθεση βαρών (weight decay) ρυθμίστηκε στο 0.01 για την αποφυγή υπερπροσαρμογής.

Κατά τη διαδικασία εκπαίδευσης, για κάθε εποχή, το μοντέλο εκτέλεσε διαδοχικά περάσματα στα δεδομένα του training set, υπολογίζοντας τις προβλέψεις και ενημερώνοντας τα

βάρη με βάση τη συνάρτηση κόστους. Μετά την ολοκλήρωση της εκπαίδευσης, το μοντέλο αξιολογήθηκε στο test set του κάθε fold. Οι προβλέψεις συγκρίθηκαν με τις πραγματικές ετικέτες, και υπολογίστηκαν οι μετρικές απόδοσης.

Η διαδικασία διασταυρούμενης επικύρωσης έδωσε ένα στατιστικά αξιόπιστο μέτρο απόδοσης του GCN. Οι μετρικές από κάθε fold καταγράφηκαν και υπολογίστηκε ο μέσος όρος τους, αποκαλύπτοντας την ικανότητα του μοντέλου να διακρίνει αποτελεσματικά σε ποια κατηγορία (bot ή human) ανήκει ο κάθε χρήστης.

Πειραματισμοί με τις υπερπαραμέτρους

Κατά τη διάρκεια ανάπτυξης του μοντέλου GCN, μέχρι να φτάσει στην τελική του μορφή, έγιναν διάφοροι πειραματισμοί με τις υπερπαραμέτρους. Δοκιμάστηκαν διάφορες παράμετροι ώστε να επιτευχθεί η βέλτιστη απόδοση. Αρχικά, εξετάστηκε ο αριθμός των επιπέδων του GCN, όπου διαπιστώθηκε ότι λίγα επίπεδα (1-2) είναι επαρκή για τη σύλληψη γενικών σχέσεων μεταξύ κόμβων. Αντίθετα, περισσότερα επίπεδα (3) κατέστησαν το μοντέλο ικανό να αποδώσει πιο λεπτομερείς συσχετίσεις, αλλά αύξησαν τον κίνδυνο υπερπροσαρμογής του στα δεδομένα.

Παράλληλα, εξετάστηκε το μέγεθος της κρυφής διάστασης (hidden dimension), δοκιμάζοντας τιμές όπως 16, 32, 64, και 128. Τα αποτελέσματα έδειξαν ότι τιμές γύρω στο 16-32 προσέφεραν ικανοποιητική ισορροπία μεταξύ της ικανότητας του μοντέλου να μαθαίνει και της αποφυγής υπερπροσαρμογής. Τελικά, επιλέχθηκε η τιμή 32. Στον τομέα της βελτιστοποίησης δοκιμάστηκαν ο Adam και ο AdamW, αλλά ο δεύτερος προσέφερε καλύτερη σταθερότητα και αποτελεσματική κανονικοποίηση με τη χρήση weight decay.

Στα πλαίσια της ρύθμισης υπερπαραμέτρων, δοκιμάστηκαν διαφορετικοί ρυθμοί εκμάθησης (learning rates) από 0.001 έως 0.01. Ένας ρυθμός εκμάθησης 0.01 επέτρεψε ταχύτερη εκπαίδευση, αλλά η σταθερότητα της εκπαίδευσης βελτιώθηκε με χαμηλότερους ρυθμούς. Αυτός είναι που τελικά επιλέχθηκε βάση σύγκρισης των αποδόσεων. Επιπλέον, η κανονικοποίηση μέσω weight decay αποδείχθηκε κρίσιμη για την αποφυγή υπερπροσαρμογής, με βέλτιστες τιμές γύρω στο 0.01. Σημαντικό ρόλο έπαιξε επίσης η χρήση dropout για τον έλεγχο της υπερπροσαρμογής, με διάφορα ποσοστά (0.3, 0.5, 0.7) να δοκιμάζονται, και το 0.3 να παρέχει την καλύτερη ισορροπία.

Τέλος, η ενσωμάτωση batch normalization ανάμεσα στα επίπεδα βοήθησε στη σταθεροποίηση της εκπαίδευσης και στη βελτίωση της γενίκευσης. Οι παραπάνω πειραματισμοί κατέδειξαν τη σημασία της προσεκτικής ρύθμισης των υπερπαραμέτρων για την επίτευξη βέλτιστων αποτελεσμάτων.

Αποτελέσματα

Το μοντέλο Graph Convolutional Network που αναπτύχθηκε αξιολογήθηκε σε τρία δίκτυα - σύνολα δεδομένων. Τα δίκτυα αυτά ήταν διαφορετικού μεγέθους ώστε να παρατηρηθεί η εξέλιξη της απόδοσης των μετρικών αξιολόγησης στο μοντέλο ανάλογα με την ποσότητα χρηστών που λαμβάνει ως είσοδο κατά τη διάρκεια εκπαίδευσης.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα στα τρία διαφορετικά δίκτυα:

Δίκτυο	Κόμβοι	Bots	Humans	Accuracy	Precision	Recall	F1 Score
smallMLdataset	1448	524	924	0.8108	0.7000	0.8750	0.7778
midMLdataset	3756	1801	1955	0.8270	0.7236	0.8934	0.7924
bigMLdataset	6316	3524	2792	0.8426	0.7598	0.9104	0.8090

Πίνακας 6.8: *Αποτελέσματα του μοντέλου Graph Convolutional Network*

6.2.4 Graph Attention Network

Σε αυτή την ενότητα θα παρουσιαστεί λεπτομερώς η υλοποίηση του μοντέλου Graph Attention Network που κατασκευάστηκε ώστε να εφαρμοστεί στο πρόβλημα του εντοπισμού αυτοματοποιημένων λογαριασμών bots μεταξύ των χρηστών στα κοινωνικά δίκτυα.

Χαρακτηριστικά που Αξιοποιήθηκαν

Τα χαρακτηριστικά (features) του συνόλου δεδομένων Twibot-20 που αξιοποιήθηκαν για την εκπαίδευση του μοντέλου Graph Attention Network (Πίνακας 6.9) περιλάμβαναν τοπολογικά χαρακτηριστικά που προέκυψαν από το δίκτυο και τη δομή του, καθώς και χαρακτηριστικά από τα προφίλ των χρηστών.

Χαρακτηριστικό	Είδος Χαρακτηριστικού	Τύπος Δεδομένων
ID	Μοναδικό αναγνωριστικό χρήστη	Int64
Following	Τοπολογικό	List of IDs
Follower	Τοπολογικό	List of IDs
Degree Centrality	Μετρική Δικτύου	Dictionary
Closeness Centrality	Μετρική Δικτύου	Dictionary
Eigenvector Centrality	Μετρική Δικτύου	Dictionary
Clustering Coefficient	Μετρική Δικτύου	Float
Followers Count	Αριθμητικό	Int
Friends Count	Αριθμητικό	Int
Favourites Count	Αριθμητικό	Int
Listed Count	Αριθμητικό	Int
Verified	Αριθμητικό	Boolean
Geo_Enabled	Αριθμητικό	Boolean
Default_Profile	Αριθμητικό	Boolean
Default_Profile_Image	Αριθμητικό	Boolean
Location_Present	Αριθμητικό	Boolean
Profile_Location_Present	Αριθμητικό	Boolean
Url_Present	Αριθμητικό	Boolean
Description_Present	Αριθμητικό	Boolean
Label	Γρουνδ Τρυτη	Boolean

Πίνακας 6.9: *Χαρακτηριστικά που αξιοποίησε το μοντέλο Graph Attention Network*

Όπως και στα προαναφερθέντα μοντέλα, έτσι και στο Graph Attention Network, πραγμα-

τοποιήθηκε μελέτη απόρριψης χαρακτηριστικών (Feature Ablation Study) προκειμένου να συμπεριληφθούν στην διαδικασία εκπαίδευσης μόνο τα χαρακτηριστικά που συνεισφέρουν θετικά σε αυτή. Όσα έχουν αρνητική επίδραση (πχ. Betweenness Centrality, Statuses Count κλπ.) εντοπίστηκαν και αφαιρέθηκαν.

Υλοποίηση

Η υλοποίηση του μοντέλου Graph Attention Network (GAT) ξεκίνησε με την προετοιμασία των δεδομένων και τη μετατροπή τους σε γράφο. Αρχικά, τα δεδομένα φορτώθηκαν από ένα αρχείο JSON, που περιείχε τις πληροφορίες χρηστών ενός δικτύου του Twitter, και επεξεργάστηκαν ώστε να μετατραπούν σε μορφή συμβατή με το PyTorch Geometric. Κάθε χρήστης αυτού του δικτύου αναπαριστούσε έναν κόμβο του γραφήματος, ενώ οι ακμές καθορίστηκαν από τις σχέσεις ακολούθησης (follower και following) μεταξύ χρηστών. Ακολούθησε η εξαγωγή χαρακτηριστικών από τα προφίλ των χρηστών. Πιο συγκεκριμένα, εξήχθησαν χαρακτηριστικά όπως αριθμητικά δεδομένα (π.χ. αριθμός friends και followers) και λογικές τιμές (π.χ. αν ο χρήστης είναι verified). Ταυτόχρονα, για την καλύτερη μοντελοποίηση της συμπεριφοράς κάθε χρήστη, συμπεριλήφθηκαν και δευτερεύοντα χαρακτηριστικά όπως όπως η χρήση προεπιλεγμένου προφίλ ή εικόνας, καθώς και η παρουσία τοποθεσίας, περιγραφής ή URL στο προφίλ. Στη συνέχεια αυτά τα χαρακτηριστικά μετατράπηκαν σε tensors PyTorch, δηλαδή σε μορφή κατάλληλη για επεξεργασία από το μοντέλο GAT.

Για τη δημιουργία του γράφου και μετέπειτα τον εμπλουτισμό των χαρακτηριστικών του, φτιάχτηκαν συναρτήσεις που προετοιμάζουν το γράφο, υπολογίζουν την φήμη ενός λογαριασμού (account reputation), εξάγουν χαρακτηριστικά προφίλ, και υπολογίζουν τοπολογικές μετρικές, δηλαδή κεντρικότητες, όπως την degree και closeness centrality. Τα χαρακτηριστικά αυτά συνδυάστηκαν σε ένα ενιαίο διάνυσμα χαρακτηριστικών για κάθε κόμβο, ενώ δημιουργήθηκε ένας κατευθυνόμενος γράφος (directed graph) που βασίστηκε, όπως αναφέρθηκε και παραπάνω, στις σχέσεις ακολούθησης. Οι ακμές του γράφου μετατράπηκαν σε μορφή edge_index, διασφαλίζοντας ότι τα δεδομένα θα είναι συμβατά με τα μοντέλα γραφημάτων του PyTorch Geometric.

Στη συνέχεια, τα δεδομένα οργανώθηκαν σε ένα αντικείμενο Data της βιβλιοθήκης PyTorch Geometric, το οποίο περιλαμβάνει τη μήτρα χαρακτηριστικών, το edge_index για τις σχέσεις των κόμβων, και τις ετικέτες. Ακολούθως, πραγματοποιήθηκε διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, προετοιμάζοντας το γράφημα για την εκπαίδευση και αξιολόγηση του GAT μοντέλου.

Το μοντέλο GAT αναπτύχθηκε με τη χρήση της βιβλιοθήκη PyTorch Geometric και βασίστηκε στην αρχιτεκτονική του GATConv. Το πρώτο επίπεδο λαμβάνει ως είσοδο τα χαρακτηριστικά των κόμβων και τα μετασχηματίζει σε ένα χώρο χαρακτηριστικών. Το μοντέλο, ακολούθως, υποστηρίζει πολλαπλές κρυφές στρώσεις, ρυθμιζόμενες διαστάσεις κρυφών χαρακτηριστικών (hidden channels), αριθμό κεφαλών προσοχής (heads), και ποσοστά αποκοπής (dropout). Η ενσωμάτωση Layer Normalization μετά από κάθε κρυφή στρώση βελτιώνει τη σταθερότητα της εκπαίδευσης, ενώ η χρήση ενεργοποίησης ELU και dropout μειώνει το overfitting. Στην τελική στρώση, το μοντέλο εξάγει προβλέψεις μέσω της συνάρτησης ενεργοποίησης softmax για την ταξινόμηση των κόμβων.

Η εκπαίδευση του μοντέλου πραγματοποιήθηκε με τη χρήση του αλγορίθμου βελτιστοποίησης AdamW, ο οποίος περιλαμβάνει L2 κανονικοποίηση (weight decay), και ενός προγραμματιστή ρυθμού εκμάθησης (ReduceLROnPlateau) που προσαρμόζει δυναμικά τον ρυθμό εκμάθησης. Κατά τη διάρκεια της εκπαίδευσης, εφαρμόστηκαν εμπρόσθια και οπίσθια διάδοση για την ελαχιστοποίηση της αρνητικής λογαριθμικής πιθανότητας μεταξύ προβλέψεων και πραγματικών ετικετών. Το μοντέλο εκπαιδεύτηκε για 200 εποχές, με ενδιάμεσες αναφορές κάθε 10 εποχές, και η αξιολόγησή του έγινε με τον υπολογισμό των μετρικών accuracy, precision, F1 score και recall.

Πειραματισμοί με τις υπερπαραμέτρους

Κατά την ανάπτυξη του Graph Attention Network (GAT) μοντέλου, μέχρι να φτάσει στην τελική του μορφή και να επιτευχθεί η βέλτιστη απόδοση, έγιναν διάφοροι πειραματισμοί, όπως εκτενής ρύθμιση παραμέτρων και τροποποιήσεις στο αρχιτεκτονικό του σχεδιασμό. Αρχικά, πραγματοποιήθηκαν δοκιμές σχετικά με την επέκταση της αρχιτεκτονικής του μοντέλου. Αυξήθηκε ο αριθμός των attention heads, με περισσότερα heads στο πρώτο στρώμα για να καταγραφούν ποικίλα μοτίβα προσοχής, ενώ στα τελικά στρώματα χρησιμοποιήθηκαν λιγότερα heads για τη συμπύκνωση των χαρακτηριστικών. Επιπλέον, προστέθηκαν κρυφά στρώματα για την ανάλυση πιο σύνθετων μοτίβων.

Παράλληλα, εξετάστηκε ο αριθμός των hidden channels. Παρατηρήθηκε πως τιμές που κυμαίνονται μεταξύ 8 και 32 αποδίδουν καλύτερα, καθώς αυτή η προσέγγιση εξισορροπεί την υπολογιστική ισχύ και αποτρέπει το overfitting. Στο μοντέλο καλύτερα απέδωσαν τα 32 hidden channels, οπότε αυτή η τιμή επιλέχθηκε τελικά.

Στα πλαίσια της ρύθμισης υπερπαραμέτρων, δοκιμάστηκαν διάφορες τιμές για τον ρυθμό εκμάθησης (learning rate). Συγκεκριμένα, ελέγχθηκαν τιμές μεταξύ 0.001 και 0.1, πριν επιλεχθεί τελικά το 0.1 καθώς εκεί σταθεροποιήθηκε το learning rate. Επιπλέον, το weight decay (L2 Regularization) προσαρμόστηκε σε διάφορες τιμές, ώστε να μειωθεί η πιθανότητα υπερπροσαρμογής, ιδιαίτερα για datasets με περιορισμένα δεδομένα εκπαίδευσης. Επίσης, ο ρυθμός dropout βελτιστοποιήθηκε ξεχωριστά για κάθε στρώμα, ώστε να επιτευχθεί επαρκής κανονικοποίηση.

Για τη βελτίωση της εκπαίδευσης, χρησιμοποιήθηκαν διαφορετικοί βελτιστοποιητές όπως AdamW αντί του παραδοσιακού Adam, ενώ προστέθηκε learning rate scheduler (ReduceLROnPlateau) που επέτρεψε τη δυναμική ρύθμιση του ρυθμού εκμάθησης κατά τη διάρκεια της εκπαίδευσης. Τέλος, ενσωματώθηκε Layer Normalization μετά από κάθε κρυφό GAT στρώμα για τη σταθεροποίηση της διανομής των χαρακτηριστικών, βοηθώντας στην αποδοτικότερη εκμάθηση από τα γραφήματα.

Είναι σημαντικό να αναφερθεί πως δοκιμάστηκε και η τεχνική της διασταυρούμενης επικύρωσης, ωστόσο απορρίφθηκε, διότι είχε θετική επίδραση στην απόδοση του μοντέλου.

Αποτελέσματα

Το μοντέλο Graph Attention Network που αναπτύχθηκε αξιολογήθηκε σε τρία δίκτυα - σύνολα δεδομένων. Τα δίκτυα αυτά ήταν διαφορετικού μεγέθους ώστε να παρατηρηθεί

η εξέλιξη της απόδοσης των μετρικών αξιολόγησης στο μοντέλο ανάλογα με την ποσότητα χρηστών που λαμβάνει ως είσοδο κατά τη διάρκεια εκπαίδευσης.

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα στα τρία διαφορετικά δίκτυα:

Δίκτυο	**Κόμβοι**	**Bots**	**Humans**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
smallMLdataset	1448	524	924	0.7967	0.6953	0.8301	0.7575
midMLdataset	3756	1801	1955	0.8124	0.7121	0.8645	0.7892
bigMLdataset	6316	3524	2792	0.8389	0.7490	0.9623	0.8003

Πίνακας 6.10: *Αποτελέσματα του μοντέλου Graph Attention Network*

6.2.5 Μεγάλα Γλωσσικά Μοντέλα

Σε αυτή την ενότητα θα παρουσιαστεί λεπτομερώς η υλοποίηση της προσέγγισης με Μεγάλα Γλωσσικά Μοντέλα που ακολουθήθηκε για να εφαρμοστεί στο πρόβλημα της αναγνώρισης αυτοματοποιημένων λογαριασμών bot μεταξύ των χρηστών σε δίκτυα του Twitter.

Περιβάλλον Υλοποίησης

Η υλοποίηση της λύσης με τη χρήση μεγάλων γλωσσικών μοντέλων (LLMs) πραγματοποιήθηκε σε περιβάλλον Jupyter Notebook, ενώ αξιολογήθηκαν πέντε διαφορετικά μεγάλα γλωσσικά μοντέλα μέσω API κλήσεων και τοπικής εκτέλεσης. Συγκεκριμένα, εγκαταστάθηκε το LM Studio τοπικά, και από αυτό χρησιμοποιήθηκαν τα μοντέλα Llama 3.2 1b και Llama 3.2 3b. Η διαφορά των δύο μοντέλων έγκειται στο ότι το 3b έχει εκπαιδευτεί με περισσότερους παραμέτρους, άρα θεωρητικά είναι καλύτερο και δύναται να προσφέρει καλύτερες απαντήσεις. Επίσης, εγκαταστάθηκε τοπικά και το Ollama, από το οποίο δοκιμάστηκε το μοντέλο Phi 4. Τέλος, από τα διαδικτυακά LLMs αξιοποιήθηκαν το ChatGPT-4 και το Cohere.

Για την προεπεξεργασία των δεδομένων, την πραγματοποίηση των κλήσεων στο API των μοντέλων και την, μετέπειτα, διαχείριση των απαντήσεών τους, υλοποιήθηκε κώδικας σε Python.

Τα API endpoints που χρησιμοποιήθηκαν με τη βιβλιοθήκη requests για κάθε γλωσσικό μοντέλο, ώστε να πραγματοποιηθούν τα ερωτήματα, είναι τα εξής:

- Llama 3.2 (μέσω LM Studio): http://localhost:11434/api/generate
- Ollama (για τοπική εκτέλεση): http://localhost:11434/api/generate
- ChatGPT-4 (OpenAI API): https://api.openai.com/v1/chat/completions
- Cohere API: https://api.cohere.ai/v1/generate

Παρατηρείται πως το LM Studio και το Ollama που εκτελούνται τοπικά, έχουν το ίδιο endpoint.

Χαρακτηριστικά που Αξιοποιήθηκαν

Τα χαρακτηριστικά (features) του συνόλου δεδομένων Twibot-20 που αξιοποιήθηκαν για τον εντοπισμό των αυτοματοποιημένων λογαριασμών στην προσέγγιση με Μεγάλα Γλωσσικά Μοντέλα είναι μόνο τα κειμενικά χαρακτηριστικά. Ο σκοπός ήταν να εξεταστεί αν τα LLMs μπορούν να ξεχωρίσουν τα bots από τους πραγματικούς χρήστες έχοντας μόνο την περιγραφή του προφίλ και το περιεχόμενο των tweets που δημοσιεύει. Συνεπώς, ήταν απαραίτητα μόνα τα χαρακτηριστικά Tweet και Description.

Εννοείται πως για τη σύγκριση των απαντήσεων του εκάστοτε γλωσσικού μοντέλου με το ground truth του συνόλου δεδομένων, και την αξιολόγηση της απόδοσής του, ήταν απαραίτητη η χρήση του ID κάθε χρήστη για ταυτοποίηση και του Label (0 ή 1, ανάλογα με το αν είναι άνθρωπος ή bot).

Χαρακτηριστικό	**Είδος Χαρακτηριστικού**	**Τύπος Δεδομένων**
ID	Μοναδικό αναγνωριστικό χρήστη	Int64
Description	Κειμενικό	String
Tweet	Κειμενικό	List of Strings
Label	Ground Truth	Boolean

Πίνακας 6.11: *Χαρακτηριστικά που αξιοποιήθηκαν από τα Μεγάλα Γλωσσικά Μοντέλα*

Υλοποίηση

Η υλοποίηση της διαδικασίας ανίχνευσης αυτοματοποιημένων λογαριασμών bot με τη βοήθεια Μεγάλων Γλωσσικών Μοντέλων ξεκίνησε με την προετοιμασία των δεδομένων. Αρχικά, έγινε η φόρτωση ενός dataset από ένα αρχείο JSON, το οποίο περιείχε πληροφορίες για χρήστες που χαρακτηρίζονται ως bots ή πραγματικοί χρήστες, με ετικέτες (labels) ¨1¨ και ¨0¨ αντίστοιχα. Τα δεδομένα χωρίστηκαν σε δύο λίστες, μία για bots και μία για πραγματικούς χρήστες, και από κάθε κατηγορία επιλέχθηκαν τυχαία 50 προφίλ. Οι δύο λίστες ενώθηκαν, και τα προφίλ ανακατεύτηκαν τυχαία ώστε να δημιουργηθεί η τελική λίστα με τα 100 προφίλ που θα χρησιμοποιούνταν στα επόμενα βήματα της διαδικασίας.

Σε αυτό το σημείο, πρέπει να αναφερθεί πως χρησιμοποιήθηκαν ενδεικτικά 100 προφίλ για την αξιολόγηση της απόδοσης του κάθε LLM, καθώς οι κλήσεις στο API των μοντέλων είναι κοστοβόρα υπολογιστικά και χρονικά. Για παράδειγμα, στο Ollama, οι 100 κλήσεις στο API του έγιναν σε διάρκεια 7 ωρών.

Στη συνέχεια, κατά την επεξεργασία των προφίλ, έγινε εξαγωγή του αναγνωριστικού χρήστη (ID), της περιγραφής του προφίλ του (description), και μιας λίστας με τα πρώτα 20 tweets του. Η περιγραφή του προφίλ και τα tweets πέρασαν, έπειτα, από τη συνάρτηση remove_emojis. Η συνάρτηση αυτή αφαιρεί emojis από το κείμενο χρησιμοποιώντας ένα πρότυπο regex που καλύπτει διάφορες κατηγορίες emojis, όπως εικονίδια προσώπων, σύμβολα, σημαίες και άλλα ειδικά σύμβολα. Εφαρμόστηκε στα tweets και τις περιγραφές (descriptions) των προφίλ, ώστε να ¨καθαριστούν¨ πριν δοθούν σαν είσοδος στα LLMs. Έγιναν δοκιμές και χωρίς τον καθαρισμό του κειμένου, ωστόσο τα LLMs, κυρίως τα μοντέλα του LM Studio (Llama 3.2 1b και Llama 3.2 3b), σε πολλές περιπτώσεις αδυνατούσαν να

επεξεργαστούν τα emojis και έδιναν άσχετες απαντήσεις. Για αυτό και αποφασίστηκε να αξιοποιηθεί το ¨καθαρισμένο¨ σύνολο δεδομένων για τις κλήσεις στα APIs. Αν κάποιο προφίλ δεν περιείχε tweets, ορίστηκε ως κενή λίστα.

Τα δεδομένα κάθε προφίλ, που περιλάμβαναν το αναγνωριστικό, την ¨καθαρισμένη¨ περιγραφή, και τα ¨καθαρισμένα¨ tweets, αποθηκεύτηκαν σε μια λίστα (test_profiles). Παράλληλα, οι ετικέτες των προφίλ μετατράπηκαν σε ακέραιες τιμές και αποθηκεύτηκαν στη λίστα ground_truth. Έτσι, προέκυψε το σύνολο δεδομένων των 100 προφίλ που θα χρησιμοποιηθούν για την αξιολόγηση των δυνατοτήτων κάθε LLM στην αναγνώριση bots.

Η διαδικασία συνεχίστηκε με την υλοποίηση της ταξινόμησης χρηστών σε κατηγορίες ("Bot" ή "Human") χρησιμοποιώντας Μεγάλα Γλωσσικά Μοντέλα (LLM) μέσω επαναλαμβανόμενων αιτημάτων σε API. Αρχικά, για κάθε χρήστη στη λίστα (από τους 100 συνολικά), τα tweets χωρίστηκαν σε παρτίδες (batches) των 4 tweets, διασφαλίζοντας έτσι ότι κάθε αίτημα προς το μοντέλο θα περιλαμβάνει ένα διαχειρίσιμο μέγεθος δεδομένων. Αυτό σημαίνει πως για κάθε προφίλ χρήστη θα γίνουν το πολύ 5 κλήσεις στο API του LLM, καθώς, όπως αναφέρθηκε πριν, έχουν επιλεγεί 20 tweets για κάθε προφίλ. Δοκιμάστηκε και η εκδοχή του να σταλούν όλα τα tweets και η περιγραφή ενός προφίλ σε μία παρτίδα, αλλά αυτό δημιουργούσε ένα τεράστιο prompt, που ήταν δύσκολα διαχειρίσιμο από τα LLMs και συχνά οδηγούσε σε απροσδόκητες απαντήσεις αντί για κατηγοριοποίηση ενός προφίλ ως bot ή άνθρωπο. Για αυτό και προτιμήθηκε η μέθοδος με τις παρτίδες (batches).

Κατόπιν, δημιουργήθηκε μια προτροπή (prompt) που περιγράφει το στόχο του ερωτήματος στο LLM, δηλαδή την κατηγοριοποίηση ενός προφίλ ως bot ή άνθρωπο, περιλαμβάνοντας τις πληροφορίες της περιγραφής του προφίλ και της εκάστοτε παρτίδας tweets. Το prompt απαιτεί από το μοντέλο να απαντήσει μόνο με τη λέξη "Bot" ή "Human", χωρίς άλλες εξηγήσεις. Δοκιμάστηκαν διάφοροι τύποι prompts, zero-shot και few-shot, ώστε να βρεθεί το πιο αποδοτικό για κάθε μοντέλο. Παρακάτω, μετά την περιγραφή της υλοποίησης της μεθόδου, παρατίθενται κάποια από τα εναλλακτικά prompts.

Για κάθε παρτίδα, στελνόταν ένα αίτημα POST στο API του LLM με το εκάστοτε επιλεγμένο prompt. Αν η απάντηση περιείχε έγκυρη πρόβλεψη, αυτή καταγραφόταν ως "Bot" ή "Human". Αν η απάντηση δεν ήταν έγκυρη ή προέκυπτε κάποιο σφάλμα κατά τη διαδικασία, καταγραφόταν ως "Unknown". Αφού ολοκληρώνονταν τα αιτήματα για όλες τις παρτίδες ενός χρήστη, οι προβλέψεις συνδυάζονταν και εφαρμοζόταν η μέθοδος της ψήφου πλειοψηφίας (majority voting), ώστε να προκύψει η τελική πρόβλεψη για κάθε προφίλ.

Η τελική απόφαση για κάθε χρήστη καταγραφόταν σε μια λίστα προβλέψεων. Αυτές οι προβλέψεις μετατρέπονταν σε δυαδικές ετικέτες, με την ετικέτα 1 να αντιστοιχεί στο "Bot" και την ετικέτα 0 στο "Human". Χρησιμοποιώντας αυτές τις ετικέτες, υπολογίζονταν οι επιλεγμένες μετρικές αξιολόγησης ώστε να μετρηθεί η απόδοση του LLM στην ταξινόμηση των χρηστών.

Prompts

Όπως αναφέρθηκε και παραπάνω, χρησιμοποιήθηκαν διάφοροι τύποι prompts μέχρι να καταλήξουμε στο αποδοτικότερο για κάθε γλωσσικό μοντέλο.

Χρησιμοποιήθηκε η μέθοδος των zero-shot prompts. Ως zero-shot ορίζεται μια προτρο-

πή (prompt) όπου το γλωσσικό μοντέλο καλείται να επιλύσει ένα πρόβλημα ή να απαντήσει σε μια ερώτηση χωρίς να του έχουν δοθεί προηγούμενα παραδείγματα του τι αναμένεται. Το μοντέλο βασίζεται αποκλειστικά στη γενική γνώση που έχει αποκτήσει κατά την εκπαίδευσή του. Κάποια παραδείγματα από τα zero-shot prompts που αξιοποιήθηκαν είναι τα εξής:

```
Task: Classify a Twitter profile as either 'Bot' or 'Human'. Instructions: Based on the
provided profile description and tweets, respond with one word only: either 'Bot' or 'Human'.
Do not include explanations, additional text, or alter the provided text in your response.
<Description> profile['description'] </Description> <Tweets> ' '.join(batch) </Tweets>
```

```
Task: Classify a Twitter profile as either 'Bot' or 'Human'. Instructions: Based on the
provided profile description and tweets, respond with one word only: 'Bot' or 'Human'. <Description>
profile['description'] </Description> <Tweets> ' '.join(batch) </Tweets>
```

```
Task: Below is a Twitter profile that I want you to classify as 'Bot' or 'Human'. <Description>
profile['description'] </Description> <Tweets> ' '.join(batch) </Tweets>
```

Επίσης, δοκιμάστηκαν και few-shot prompts. Ως few-shot ορίζεται μια προτροπή όπου το γλωσσικό μοντέλο καλείται να επιλύσει ένα πρόβλημα ή να απαντήσει σε μια ερώτηση, αλλά του παρέχονται λίγα παραδείγματα (συνήθως 1 έως 5) που υποδεικνύουν το επιθυμητό μοτίβο ή το πλαίσιο της απάντησης. Τα παραδείγματα βοηθούν το μοντέλο να κατανοήσει καλύτερα τη δομή που αναμένεται. Κάποια παραδείγματα από τα few-shot prompts που αξιοποιήθηκαν είναι τα εξής:

```
Task: Classify a Twitter profile as either 'Bot' or 'Human'. Instructions: Based on the
provided profile description and tweets, respond with one word only: 'Bot' or 'Human'.
Examples: <Description> Tech news and updates from around the world. </Description> <Tweets>
Breaking: New iPhone launched today. Specs look amazing! Stay tuned for updates. </Tweets>
Response: Human
<Description> Affordable travel deals and tips. </Description> <Tweets> Flights to Bali
now 299 roundtrip! Plan your dream vacation today. Link in bio. </Tweets> Response: Bot

<Description> Fitness enthusiast sharing workout routines. </Description> <Tweets> Great
leg day today! Tried new squats variation. Feeling pumped. </Tweets> Response: Human
<Description> Daily horoscope and zodiac advice. </Description> <Tweets> Capricorns: Today
is your day to shine! Embrace new opportunities. </Tweets> Response: Bot
<Description> Just another coffee lover sharing thoughts. </Description> <Tweets> Iced coffee
on a sunny day is unbeatable.  CaffeineAddict </Tweets> Response: Human
```

```
Task: Classify a Twitter profile as either 'Bot' or 'Human'. Instructions: Based on the
provided profile description and tweets, respond with one word only: 'Bot' or 'Human'. Do
not include explanations, additional text, or alter the provided text in your response.
```

```
Examples: <Description> Celebrity gossip and entertainment news. </Description> <Tweets>
The latest scoop on the red carpet tonight! Who wore it best? </Tweets> Response: Human

<Description> Automated tweets about marketing strategies. </Description> <Tweets> Learn
how to boost your business with social media marketing. </Tweets> Response: Bot
<Description> Personal trainer sharing fitness tips and motivation. </Description> <Tweets>
Just finished an intense HIIT session. Ready to crush tomorrow's workout! </Tweets> Response:
Human
<Description> Daily motivational quotes and productivity hacks. </Description> <Tweets>
"Success is not final, failure is not fatal: It is the courage to continue that counts."
</Tweets> Response: Human
<Description> AI-driven promotional content for eCommerce platforms. </Description> <Tweets>
Check out our latest offers: Buy 1 Get 1 Free on all items! </Tweets> Response: Bot
```

Αποτελέσματα

Η μέθοδος της αναγνώρισης αυτοματοποιημένων λογαριασμών με Μεγάλα Γλωσσικά Μοντέλα αξιολογήθηκε σε ένα σύνολο δεδομένων, αυτό των 100 προφίλ που προέκυψε με τη διαδικασία που αναλύθηκε προηγουμένως. Χρησιμοποιήθηκε αυτό το, θεωρητικά μικρό, σύνολο δεδομένων διότι οι κλήσεις στο API των μοντέλων (ειδικά όσων έχουν εγκατασταθεί και τρέχουν τοπικά) είναι απαιτητική υπολογιστικά και χρονικά. Δοκιμάστηκαν διάφορα prompts, η αποστολή των tweets ενός προφίλ σε παρτίδες ή όλα μαζί, με ή χωρίς τα emojis τους. Στον παρακάτω πίνακα επιλέχθηκαν τα καλύτερα αποτελέσματα που προέκυψαν για κάθε μοντέλο συνδυάζοντας τις προαναφερθείσες εναλλακτικές προσεγγίσεις.

Γλωσσικό Μοντέλο	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Llama 3.2 1b	0.54	0.52	0.70	0.69
Llama 3.2 3b	0.53	0.59	0.28	0.34
Ollama - Phi 4	0.52	0.60	0.12	0.20
ChatGPT 4	0.54	0.59	0.26	0.36
Cohere	0.51	0.67	0.04	0.08

Πίνακας 6.12: *Αποτελέσματα ταξινόμησης αυτοματοποιημένων λογαριασμών από διάφορα Μεγάλα Γλωσσικά Μοντέλα*

6.3 Σύγκριση και Ερμηνεία Αποτελεσμάτων

Σε αυτή την ενότητα ακολουθεί η ερμηνεία και η συγκριτική ανάλυση των αποτελεσμάτων που δίνουν οι προαναφερθείσες υλοποιήσεις των μοντέλων Logistic Regression, Random Forest, Graph Convolutional Network (GCN), Graph Attention Network (GAT), και Μεγάλων Γλωσσικών Μοντέλων (LLMs), όταν εφαρμόζονται στο πρόβλημα ανίχνευσης αυτοματοποιημένων λογαριασμών bot.

Η αξιολόγηση βασίζεται στις μετρικές accuracy, precision, recall, και F1 score και επικεντρώνεται στη συμπεριφορά τους στα διαφορετικού μεγέθους datasets.

Αποτελέσματα των Κλασικών Μοντέλων Μηχανικής Μάθησης

Logistic Regression

Το Logistic Regression σημείωσε μέτρια απόδοση, με τις 3 από τις 4 μετρικές αξιολόγησής του να παρουσιάζουν σταδιακή βελτίωση όσο αυξανόταν το μέγεθος του dataset. Χρήζουν προσοχής τα εξής σημεία:

Accuracy: Ξεκινώντας από 0.7103 στο μικρό dataset, παρατηρείται μικρή μείωση στα δύο επόμενα datasets, φτάνοντας μόνο το 0.7035 στο μεγάλο dataset, υποδεικνύοντας έναν πιθανό κορεσμό της απόδοσης. Αυτό μπορεί να αποδοθεί στη γραμμική φύση του Logistic Regression, το οποίο δυσκολεύεται να μοντελοποιήσει περίπλοκες σχέσεις στα δεδομένα.

Recall: Η εξαιρετική τιμή του recall (0.9475 στο μεγάλο dataset) δείχνει ότι το μοντέλο καταφέρνει να εντοπίσει την πλειονότητα των bots, αλλά αυτό συνοδεύεται από χαμηλή τιμή precision.

Precision: Η χαμηλή τιμή του precision (0.5902 στο μικρό dataset και 0.6719 στο μεγάλο) υποδηλώνει αυξημένο αριθμό false positives. Το μοντέλο μπορεί να θεωρεί ανθρώπινους χρήστες ως bots, κάτι που μπορεί να είναι προβληματικό σε πραγματικές εφαρμογές.

Random Forest

Το Random Forest παρουσίασε σαφώς καλύτερη απόδοση σε σύγκριση με το Logistic Regression, αποδεικνύοντας πως ήταν πιο ισχυρό στη μοντελοποίηση μη γραμμικών σχέσεων. Αξίζει να γίνουν οι εξής παρατηρήσεις:

Accuracy: Παρόλο που η ακρίβειά του μειώνεται καθώς αυξάνεται το μέγεθος των datasets (0.7475 στο μεγάλο), η συνολική απόδοσή του παραμένει σταθερά υψηλότερη σε σχέση με το Logistic Regression.

Precision και Recall: Το Random Forest πετυχαίνει καλύτερη ισορροπία μεταξύ precision και recall. Στο μεγάλο dataset, σημειώνει precision 0.7325 και recall 0.8668, που οδηγεί σε υψηλό F1 score (0.7940). Αυτό δείχνει ότι το μοντέλο είναι ικανό να εντοπίζει bots χωρίς να θυσιάζει την ακρίβεια της πρόβλεψης.

Από τη σύγκριση των δύο κλασσικών μοντέλων μηχανικής μάθησης, προκύπτει πως το Random Forest είναι μια αποτελεσματικότερη μέθοδος, ιδιαίτερα για datasets με μεγαλύτερη ποικιλομορφία.

Αποτελέσματα των Γραφικών Μοντέλων

Graph Convolutional Network

Το GCN παρουσίασε την καλύτερη συνολική απόδοση σε όλες τις μετρικές, ειδικά στα μεγαλύτερα datasets, όπου η δομή των δεδομένων γραφημάτων μπορεί να αξιοποιηθεί στο έπακρο.

Accuracy: Η ακρίβεια αυξήθηκε από 0.8108 στο μικρό dataset σε 0.8426 στο μεγάλο, υποδεικνύοντας ότι το μοντέλο επωφελείται από τη μεγαλύτερη ποσότητα δεδομένων.

Precision: Με τιμές από 0.7000 έως 0.7598, το GCN δείχνει ότι μπορεί να διατηρήσει ικανοποιητική ακρίβεια ακόμα και σε μεγαλύτερα σύνολα δεδομένων.

Recall: Το recall του GCN παραμένει σταθερά υψηλό, φτάνοντας το 0.9104 στο μεγάλο dataset, δείχνοντας την ικανότητα του μοντέλου να ανιχνεύει σχεδόν όλα τα bots.

F1 Score: Το υψηλό F1 score (0.8090 στο μεγάλο dataset) αποδεικνύει τη συνολική αποτελεσματικότητα του GCN.

Graph Attention Network

Το GAT παρουσίασε παρόμοια, ελαφρώς υποδεέστερη, απόδοση με το GCN, αλλά με κάποιες διαφορές που αξίζουν προσοχής:

Recall: Το GAT σημείωσε την υψηλότερη τιμή recall (0.9623 στο μεγάλο dataset), υποδεικνύοντας ότι έχει μεγαλύτερη ευαισθησία στον εντοπισμό bots.

Precision και F1 Score: Παρόλο που το precision του GAT είναι ελαφρώς χαμηλότερο από αυτό του GCN (0.7490 στο μεγάλο dataset), η συνολική του απόδοση είναι εξαιρετική, με F1 score 0.8003.

Από τη σύγκριση των δύο graph-based μοντέλων, φαίνεται πως το GAT υπερτερεί σε εφαρμογές όπου το recall είναι η κύρια προτεραιότητα. Ωστόσο, συνολικά πιο αποτελεσματικό είναι το GCN.

Αποτελέσματα των Μεγάλων Γλωσσικών Μοντέλων (LLMs)

Τα LLMs παρουσίασαν σημαντικά χαμηλότερες επιδόσεις σε σχέση με τις υπόλοιπες μεθόδους. Πιο συγκεκριμένα:

Accuracy: Η ακρίβεια κυμαίνεται μεταξύ 0.51 και 0.54, υποδηλώνοντας περιορισμένη ικανότητα γενίκευσης.

Precision και Recall: Ενώ κάποια LLMs (π.χ. Cohere) παρουσίασαν σχετικά ικανοποιητικό precision (0.67), οι τιμές recall ήταν εξαιρετικά χαμηλές (0.04 για το Cohere), οδηγώντας σε πολύ χαμηλό F1 score.

Από τα αποτελέσματα, προκύπτει πως τα LLMs (Llama 3.2, ChatGPT 4, Ollama - Phi 4, Cohere) δυσκολεύονται στην αναγνώριση των bots. Συνολικά έχουν σαφώς χαμηλότερη απόδοση σε σχέση με τα παραδοσιακά και graph-based μοντέλα, ειδικά όταν μελετάται η απόδοσή τους σε μικρά datasets, καθώς επηρεάζεται, ενδεχομένως, από την έλλειψη επαρκούς πληροφορίας. Ωστόσο, η αξιολόγηση με περισσότερα δεδομένα καθίσταται εξαιρετικά απαιτητική λόγω της μεγάλης υπολογιστικής πολυπλοκότητας. Συνεπώς, παρά τις δυνατότητές τους στη γλωσσική κατανόηση, οι χαμηλές τιμές accuracy (μέγιστο 0.54), precision, και recall υποδεικνύουν ότι η χρήση τους σε αυτό το πρόβλημα είναι μάλλον άστοχη. Η καλύτερη επίδοση παρατηρήθηκε στο μοντέλο Llama 3.2 1b με F1 score 0.69, το οποίο, όμως, ήταν υποδεέστερο των υπόλοιπων μεθόδων.

Συνολική Παρουσίαση Αποτελεσμάτων

Παρακάτω παρουσιάζονται τα συνολικά αποτελέσματα των μοντέλων ανά σύνολο δεδομένων (Πίνακας 6.13 - Πίνακας 6.16).

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7103	0.5902	0.9231	0.7200
Random Forest	0.7859	0.6876	0.7518	0.7177
GCN	0.8108	0.7000	0.8750	0.7778
GAT	0.7967	0.6953	0.8301	0.7575

Πίνακας 6.13: *Συνολικός Πίνακας Αποτελεσμάτων για το smallMLdataset*

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6902	0.6142	0.9239	0.7379
Random Forest	0.7535	0.7095	0.8234	0.7621
GCN	0.8270	0.7236	0.8934	0.7924
GAT	0.8124	0.7121	0.8645	0.7892

Πίνακας 6.14: *Συνολικός Πίνακας Αποτελεσμάτων για το midMLdataset*

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7035	0.6719	0.9475	0.7862
Random Forest	0.7475	0.7325	0.8668	0.7940
GCN	0.8426	0.7598	0.9104	0.8090
GAT	0.8389	0.7490	0.9623	0.8003

Πίνακας 6.15: *Συνολικός Πίνακας Αποτελεσμάτων για το bigMLdataset*

Model	Accuracy	Precision	Recall	F1 Score
Llama 3.2 1b	0.54	0.52	0.70	0.69
Llama 3.2 3b	0.53	0.59	0.28	0.34
Ollama - Phi 4	0.52	0.60	0.12	0.20
ChatGPT 4	0.54	0.59	0.26	0.36
Cohere	0.51	0.67	0.04	0.08

Πίνακας 6.16: *Συνολικός Πίνακας Αποτελεσμάτων για το LLMdataset*

Γενική Ερμηνεία Αποτελεσμάτων

Παραδοσιακά Μοντέλα: Το Random Forest ξεπέρασε το Logistic Regression λόγω της ικανότητάς του να μοντελοποιεί μη γραμμικές σχέσεις. Ωστόσο, υστερεί σε σχέση με τα graph-based μοντέλα, τα οποία εκμεταλλεύονται καλύτερα τη δομή των δεδομένων.

Γραφικά Μοντέλα: Το GCN παρουσίασε την καλύτερη συνολική απόδοση, ειδικά σε μεγάλα datasets, αποδεικνύοντας την αποτελεσματικότητά του σε δεδομένα με σχέσεις γραφημάτων. Το GAT, αν και αρκετά κοντά σε επίδοση υπερτερώντας σε recall, υστερεί στις υπόλοιπες μετρικές.

Μεγάλα Γλωσσικά Μοντέλα: Παρά την υποσχόμενη θεωρητική τους βάση, τα LLMs απέτυχαν σημαντικά να ανταγωνιστούν τις άλλες μεθόδους.

Συμπέρασμα

Τα αποτελέσματα δείχνουν ότι τα graph-based μοντέλα (GCN, GAT) είναι οι καλύτερες επιλογές για το συγκεκριμένο πρόβλημα, ειδικά όταν τα σύνολα δεδομένων περιέχουν πλούσια πληροφορία σχέσεων μεταξύ κόμβων. Οι παραδοσιακές μέθοδοι μπορούν να αποτελέσουν εναλλακτική λύση, ενώ τα LLMs, αν και υποσχόμενα αρχικά, απαιτούν βελτιώσεις και προσαρμογές, ακόμα και στον τρόπο εκπαίδευσής τους, για να γίνουν ανταγωνιστικά.

Μέρος III

Επίλογος

Κεφάλαιο 7

Επίλογος

Σε αυτό το, τελευταίο, κεφάλαιο γίνεται μια σύνοψη των αποτελεσμάτων και των συμπερασμάτων που προέκυψαν από την παρούσα διπλωματική εργασία, ενώ δίνονται και προτάσεις για πιθανές μελλοντικές επεκτάσεις της έρευνας που παρουσιάστηκε.

7.1 Σύνοψη και Συμπεράσματα

Η παρούσα διπλωματική εργασία επικεντρώθηκε στην ανίχνευση αυτοματοποιημένων λογαριασμών bots στην πλατφόρμα του Twitter, προσφέροντας μια αναλυτική προσέγγιση για την αντιμετώπιση αυτού του ιδιαίτερα επίκαιρου προβλήματος των κοινωνικών δικτύων. Μέσα από την μελέτη, ανάπτυξη και σύγκριση διαφόρων τεχνικών και μεθόδων, διερευνήθηκαν οι δυνατότητες αναγνώρισης και ταξινόμησης λογαριασμών ως bots, παρέχοντας σημαντικά αποτελέσματα και συμπεράσματα για την έρευνα που διεξάγεται επί του συγκεκριμένου θέματος.

Στην έρευνα αξιοποιήθηκε ένα ευρύ και αντιπροσωπευτικό του δικτύου του Twitter σύνολο δεδομένων, το Twibot-20. Περιελάμβανε γνήσιους λογαριασμούς ανθρώπων, αλλά και λογαριασμούς bots. Το σύνολο δεδομένων υποδιαιρέθηκε σε συγκεκριμένα υποσύνολα που χρησιμοποιήθηκαν για την εκπαίδευση και αξιολόγηση των μοντέλων. Από αυτά αντλήθηκαν άμεσα κάποια χαρακτηριστικά (αριθμητικά, κειμενικά), ενώ κάποια άλλα, κυρίως γραφικά, χαρακτηριστικά εξήχθησαν ύστερα από επεξεργασία των δεδομένων στα μοντέλα. Η χρήση ποικιλόμορφων χαρακτηριστικών επέτρεψε την καλύτερη δυνατή μοντελοποίηση της συμπεριφοράς και των χαρακτηριστικών των bots.

Τα αποτελέσματα των πειραμάτων που διεξήχθησαν κατέδειξαν σημαντικές διαφορές στην απόδοση των μεθόδων που εξετάστηκαν. Ιδιαίτερα εντυπωσιακή και απροσδόκητη ήταν η κακή επίδοση που σημείωσαν τα Μεγάλα Γλωσσικά Μοντέλα. Όλα όσα εξετάστηκαν σημείωσαν απογοητευτικά αποτελέσματα στην ανίχνευση των bots βάσει κειμενικών χαρακτηριστικών (περιγραφή προφίλ, περιεχόμενο tweets), παρά τις εντυπωσιακές τους δυνατότητες σε άλλες εφαρμογές. Πιθανός λόγος για την αδυναμία τους σε αυτού του είδους τις προβλέψεις μπορεί να είναι η έλλειψη εξειδικευμένης εκπαίδευσης των LLMs στα δεδομένα που σχετίζονται με τη συγκεκριμένη περίπτωση χρήσης. Παράλληλα, η ολοένα και καλύτερη μοντελοποίηση των bots κατά την δημιουργία τους, ώστε να προσομοιάζουν τους ανθρώπινους χρήστες και την συμπεριφορά τους στα μέσα κοινωνικής δικτύωσης, φαίνεται να δουλεύει όταν είναι τα LLMs αυτά που καλούνται να κάνουν την κατηγοριοποίηση. Όσο προχωρά η έρευνα που στοχεύει

στην ανίχνευση των bots, τόσο προοδεύουν και οι τεχνολογίες ανάπτυξής τους, καταφέρνοντας να ενσωματώνονται πετυχημένα στις διαδικτυακές κοινότητες των γνήσιων χρηστών, χωρίς να κινούν υποψίες με την δράση τους. Ειδικά όσον αφορά το περιεχόμενο που δημοσιεύουν, τα καταφέρνουν πλέον πολύ καλά, έτσι ώστε να ¨ξεγελούν¨ τα Μεγάλα Γλωσσικά Μοντέλα, που είναι εκπαιδευμένα στην κατανόηση και παραγωγή γλωσσικού περιεχομένου. Συνεπώς, τουλάχιστον σε αυτό τον τομέα, τα LLMs έχουν ακόμα πολλά περιθώρια βελτίωσης στην προεκπαίδευση που λαμβάνουν.

Αντίθετα, η καλύτερη απόδοση σημειώθηκε από τα συνελικτικά μοντέλα γράφων, Graph Convolutional Network και Graph Attention Network, με το πρώτο να σημειώνει ελαφρώς καλύτερα αποτελέσματα. Τα μοντέλα αυτά συνδύασαν αριθμητικά χαρακτηριστικά του προφίλ αλλά και γραφικά που προέκυψαν από το εξεταζόμενο δίκτυο και τις σχέσεις μεταξύ των χρηστών που το αποτελούσαν. Κατάφεραν μέσω της πολύπλοκης δομής τους να μοντελοποιήσουν τα χαρακτηριστικά και την συμπεριφορά των bots κατά την διαδικασία της εκπαίδευσής τους, ώστε να επιτύχουν την ανίχνευσή τους με αρκετά καλά ποσοστά επιτυχίας. Σε επιδόσεις ακολούθησαν τα κλασικά μοντέλα μηχανικής μάθησης Logistic Regression και Random Forest.

Συνοψίζοντας, η εργασία αυτή συνέβαλε στην έρευνα που διεξάγεται γύρω από το μείζον πρόβλημα της ανίχνευσης αυτοματοποιημένων λογαριασμών bots στο Twitter, αναδεικνύοντας τόσο τις προκλήσεις όσο και τις δυνατότητες που προκύπτουν από τη χρήση διαφόρων σύγχρονων τεχνολογιών για αυτό το ζήτημα. Τα ευρήματα της μελέτης συνεισφέρουν στην κατανόηση της φύσης του προβλήματος και ανοίγουν τον δρόμο για περαιτέρω έρευνα, που θα μπορούσε να επικεντρωθεί στη βελτίωση της απόδοσης των LLMs και στη δημιουργία πιο εξειδικευμένων μοντέλων που συνδυάζουν γλωσσικά και μη γλωσσικά δεδομένα. Με αυτόν τον τρόπο, η εργασία φιλοδοξεί να αποτελέσει ένα χρήσιμο εργαλείο για την επιστημονική κοινότητα και την περαιτέρω ανάπτυξη του πεδίου.

7.2 Μελλοντικές Επεκτάσεις

Παρότι τα αποτελέσματα που προέκυψαν από την παρούσα εργασία ήταν αρκετά ικανοποιητικά και ενθαρρυντικά, λόγω των χρονικών και υπολογιστικών περιορισμών που υπήρχαν κατά την εκπόνησή της, αφέθηκαν περιθώρια για βελτιώσεις και επεκτάσεις. Στη συνέχεια, λοιπόν, παρατίθενται τόσο μερικές πιθανές βελτιώσεις που θα μπορούσαν να υλοποιηθούν ώστε να αυξηθεί η πληρότητα της προτεινόμενης προσέγγισης, όσο και κάποιες γενικότερες προτάσεις για την επέκταση της έρευνας στην ανίχνευση των bots στα μέσα κοινωνικής δικτύωσης.

Ένα πρώτο βήμα για την επέκταση της εργασίας θα ήταν η εφαρμογή των πειραμάτων σε περισσότερα δείγματα δικτύων και λογαριασμών. Αφενός, στο σύνολο δεδομένων που αξιοποιήθηκε, Twibot-20, μπορούν να δοκιμαστούν τα μοντέλα σε μεγαλύτερο δείγμα χρηστών, κάτι το οποίο ήταν δύσκολο λόγω υπολογιστικών περιορισμών κατά την διεξαγωγή των πειραμάτων της παρούσας εργασίας. Και αφετέρου, μπορεί να γίνει και η επέκταση σε άλλα σύγχρονα σύνολα δεδομένων.

Μια ακόμα κατεύθυνση για μελλοντική ανάπτυξη θα μπορούσε να είναι η επιστράτευση και άλλων Μεγάλων Γλωσσικών Μοντέλων, καθώς και η αξιολόγησή τους με σαφώς μεγαλύτε-

ρη ποσότητα δειγμάτων. Όπως, έχει ήδη αναφερθεί, η απόδοση των LLMs ήταν απογοητευτική στα πειράματα που πραγματοποιήθηκαν, ακόμα και αν αξιολογήθηκαν με λίγους μόνο χρήστες Twitter έκαστο, λόγω υπολογιστικών περιορισμών. Πέρα από την δοκιμή και άλλων LLMs που ίσως είναι πιο ικανά στην αναγνώριση των bots από τα κειμενικά τους χαρακτηριστικά, μπορούν αυτά, αλλά και τα ήδη υλοποιημένα μοντέλα, να εξεταστούν σε μεγαλύτερα σύνολα δεδομένων ώστε να αποκτηθεί μια πιο ολοκληρωμένη εικόνα για τις δυνατότητές τους στην ανίχνευση bots.

Παράλληλα, θα ήταν ενδιαφέρον να εξεταστεί η επίδοση των LLMs, όταν τους δίνονται σαν είσοδος και τα υπόλοιπα χαρακτηριστικά του συνόλου δεδομένων (πχ. followers count, statuses count).

Στο ίδιο μήκος κύματος, αλλά αντίστροφα, στα μοντέλα μηχανικής μάθησης και γραφικών νευρωνικών δικτύων, θα μπορούσε να γίνει χρήση και των κειμενικών χαρακτηριστικών (περιγραφή προφίλ, περιεχόμενο tweets) κατά την εκπαίδευση των μοντέλων. Δηλαδή, θα μπορούσαν να ενσωματωθούν και τεχνικές επεξεργασίας φυσικής γλώσσας σε αυτές τις μεθόδους ανίχνευσης.

Επιπλέον, θα ήταν χρήσιμη η επέκταση της έρευνας ώστε να περιλαμβάνει την υλοποίηση και αξιολόγηση περισσότερων μοντέλων, από κλασικά μοντέλα μηχανικής μάθησης όπως τα Support Vector Machines (SVM), μέχρι και διάφορες, πιο σύγχρονες, τεχνικές Βαθιάς Μάθησης (Deep Learning).

Μια άλλη πιθανή κατεύθυνση για βελτίωση της προτεινόμενης υλοποίησης είναι η εξαγωγή και χρήση, κατά την εκπαίδευση, χαρακτηριστικών που δεν έχουν αποτελέσει αντικείμενο έρευνας ως τώρα. Για παράδειγμα, μπορεί να μελετηθεί η ώρα και ημερομηνία δημιουργίας ενός προφίλ (created_at) και των λογαριασμών που ακολουθεί και το ακολουθούν. Αν είναι ύποπτα κοντινές χρονικά, ενδεχομένως είναι ένδειξη μαζικής δημιουργίας bots.

Ακόμα, θα μπορούσε να δοθεί ακόμα πιο πολλή βαρύτητα στην ρύθμιση υπερπαραμέτρων (hyperparameter tuning) και να δοκιμαστεί μεγαλύτερο εύρος πιθανών συνδυασμών μέσω της υλοποίησης ενός κώδικα βελτιστοποίησης, ώστε να επιτευχθούν τα καλύτερα δυνατά και πιο αποδοτικά μοντέλα.

Τέλος, σημαντικός μελλοντικός στόχος είναι και η εξέλιξη της υλοποιημένης έρευνας για την ανίχνευση των αυτοματοποιημένων λογαριασμών bot ώστε να εφαρμοστεί και σε άλλα μέσα κοινωνικής δικτύωσης εκτός του Twitter.

Βιβλιογραφία

[1] Haenlein M. Kaplan, A. M. *Users of the world, unite! The challenges and opportunities of Social Media*. *Business Horizons*, 1(53):59–68, 2010.

[2] Ellison N. B. Boyd, D. *Social Network Sites: Definition, History, and Scholarship*. *Journal of Computer-Mediated Communication*, 1(13):210–230, 2007.

[3] *Free AI Image Generator: Online Text to Image App | Canva*. Online, 2024. Available: https://www.canva.com/ai-image-generator/.

[4] Burgess J. E. Bruns, A. *Researching News Discussion on Twitter: New Methodologies*. *Journalism Studies*, 5(13):801–814, 2012.

[5] S. C. Woolley και P. N. Howard. *Automation, Algorithms, and Politics: Political Communication in the Era of Social Media Bots*. *International Journal of Communication*, 10:4882–4890, 2016.

[6] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[7] D. Easley και J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[8] GeeksforGeeks. *Mathematics | Graph Theory Basics - Set 2*, 2024. Available: https://www.geeksforgeeks.org/mathematics-graph-theory-basics/.

[9] Wikipedia. *Graph (discrete mathematics)*, 2024. Available: https://en.wikipedia.org/wiki/Graph_(discrete_mathematics).

[10] M. E. J. Newman. *The Structure and Function of Complex Networks*. *SIAM Review*, 45(2):167–256, 2003.

[11] A. L. Barabási. *Network Science*. Cambridge University Press, 2016.

[12] EnjoyAlgorithms. *Representation of Graph in Data Structures*, 2024. Available: https://www.enjoyalgorithms.com/blog/graph-representation-in-data-structures.

[13] GeeksforGeeks. *Clustering Coefficient in Graph Theory*, 2024. Available: https://www.geeksforgeeks.org/clustering-coefficient-graph-theory/.

[14] Subhajyoti Samaddar, Sudip Roy, Fatima Akter και Hirokazu Tatano. *Diffusion of Disaster-Preparedness Information by Hearing from Early Adopters to Late Adopters in Coastal Bangladesh*. 2022.

[15] Wikipedia. *PageRank*, 2024. Available: https://en.wikipedia.org/wiki/PageRank.

[16] S. J. Russell και P. Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2021.

[17] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, 2020.

[18] GeeksforGeeks. *Semi-Supervised Learning in ML*. Online. Available: https://www.geeksforgeeks.org/ml-semi-supervised-learning/.

[19] Vaani Rawatt. *Introduction to Machine Learning*, 2024. Available: https://medium.com/@vaani.rawatt/introduction-to-machine-learning-a5ddc31ce404.

[20] DataCamp. *Classification in Machine Learning: A Guide for Beginners*. Online. Available: https://www.datacamp.com/blog/classification-machine-learning.

[21] GeeksforGeeks. *What is a Neural Network?* Online. Available: https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/.

[22] Simi Job, X. Tao, T. Cai, Lin Li, H. Xie και J. Yong. *Towards Causal Classification: A Comprehensive Study on Graph Neural Networks*. *arXiv*, 2401.154441, 2024.

[23] Wikipedia. *Deep Learning*. Online. Available: https://en.wikipedia.org/wiki/Deep_learning.

[24] GeeksforGeeks. *Logistic Regression in Machine Learning*. Online. Available: https://www.geeksforgeeks.org/understanding-logistic-regression/.

[25] Daniel Jurafsky και James H. Martin. *Speech and Language Processing*. Stanford, 2024.

[26] GeeksforGeeks. *Random Forest Algorithm in Machine Learning*. Online. Available: https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/.

[27] Built In. *Random Forest: A Complete Guide for Machine Learning*. Online. Available: https://builtin.com/data-science/random-forest-algorithm.

[28] DataCamp. *Random Forest Classification with Scikit-Learn*. Online. Available: https://www.datacamp.com/tutorial/random-forests-classifier-python.

[29] Abhishek Jain. *Everything about Random Forest*, 2024. Available: https://medium.com/@abhishekjainindore24/everything-about-random-forest-90c106d63989.

[30] Towards Data Science. *Graph Convolutional Networks: Introduction to GNNs*, 2024. Available: https://towardsdatascience.com/graph-convolutional-networks-introduction-to-gnns-24b3f60d6c95.

[31] GeeksforGeeks. *Graph Convolutional Networks (GCNs): Architectural Insights and Applications*, 2024. Available: https://www.geeksforgeeks.org/graph-convolutional-networks-gcns-architectural-insights-and-applications/.

[32] Felix Wu, Tianyi Zhang, Amauri Holandade Souza Jr., Christopher Fifty, Tao Yu και Kilian Q. Weinberger. *Simplifying Graph Convolutional Networks*. *arXiv*, 2019.

[33] Thomas N. Kipf και Max Welling. *Semi-Supervised Classification with Graph Convolutional Networks*. *International Conference on Learning Representations (ICLR)*, 2017.

[34] Farzad Karami. *Understanding Graph Attention Networks: A Practical Exploration*, 2024. Available: https://medium.com/@farzad.karami/understanding-graph-attention-networks-a-practical-exploration-cf033a8f3d9d.

[35] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò και Yoshua Bengio. *Graph Attention Networks*. *International Conference on Learning Representations (ICLR)*, 2018.

[36] Cloudflare. *What is an LLM (large language model)?* Online.

[37] Big Blue Academy. *Large Language Models*. Online. Available: https://bigblue.academy/gr/large-language-models.

[38] TechTarget. *What are Large Language Models (LLMs)? | Definition from TechTarget*. Online. Available: https://www.techtarget.com/whatis/definition/large-language-model-LLM.

[39] Striveworks. *Text Classification With LLMs: A Roundup of the Best Methods*. Online. Available: https://www.striveworks.com/blog/text-classification-with-llms-a-roundup-of-the-best-methods.

[40] Exploding Topics. *Number of Parameters in GPT-4 (Latest Data)*. Online. Available: https://explodingtopics.com/blog/gpt-parameters.

[41] H. Naveed, A. Ullah Khan, Shi Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes και Ajmal Mian. *A Comprehensive Overview of Large Language Models*. 2024.

[42] Label Your Data. *Types of LLMs: Classification Guide in 2025*. Online. Available: https://labelyourdata.com/articles/types-of-llms.

[43] *Do Social Bots Shape Public Opinion? | CHEQ*. Online, 2024. Available: https://cheq.ai/blog/social-bots-how-do-they-shape-public-opinion/.

[44] Cloudflare. *What is a social media bot? | Social media bot definition*. Available: https://www.cloudflare.com/learning/bots/what-is-a-social-media-bot/.

[45] Indusface. *Social Media Bots - Definition, Purpose, Preventive Measures*. Available: https://www.indusface.com/learning/social-media-bots/.

[46] S. Cresci. *A Decade of Social Bot Detection*. *Communications of the ACM*, 63(10):72–83, 2020.

[47] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini και F. Menczer. *The Spread of Low-credibility Content by Social Bots*. *Nature Communications*, 9:4787, 2018.

[48] Alessandro Bessi και Emilio Ferrara. *Social bots distort the 2016 U.S. Presidential election online discussion*. *First Monday*, 21(11), 2016.

[49] Sergey Kogan, Tobias J. Moskowitz και Monika Niessner. *Fake news in financial markets*. *Journal of Finance*, 74(6):2911–2953, 2019.

[50] Samantha Bradshaw και Philip N. Howard. *Troops, trolls and troublemakers: A global inventory of organized social media manipulation*. 2017. Available: `https://comprop.oii.ox.ac.uk/research/troops-trolls-and-troublemakers/`.

[51] J. Zhang, Z. Zhang, J. Chen και Y. Xiang. *Bot Detection in Social Networks: A Data Mining Perspective*. *IEEE Transactions on Knowledge and Data Engineering*, 33(5):2075–2089, 2021.

[52] A. Ramalingaiah, S. Hussaini και S. Chaudhari. *Twitter bot detection using supervised machine learning*. *Journal of Physics: Conference Series*, 1950(1):012006, 2021.

[53] *Kaggle - Detecting Twitter Bot Data*. Online, 2021. Available: `https://www.kaggle.com/datasets/charvijain27/detecting-twitter-bot-data`.

[54] T. Fagni, F. Falchi, M. Gambini, A. Martella και M. Tesconi. *TweepFake: About detecting deepfake tweets*. *PLoS ONE*, 16(5):ε0251415, 2021.

[55] M. Heidari, J. H. J. Jones και O. Uzuner. *An Empirical Study of Machine Learning Algorithms for Social Media Bot Detection*. *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, σελίδες 1–5, 2021.

[56] D. Dukić, D. Keča και D. Stipić. *Are You Human? Detecting Bots on Twitter Using BERT*. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, σελίδες 631–636, 2020.

[57] L. Ilias και I. Roussaki. *Detecting malicious activity in Twitter using deep learning techniques*. *Applied Soft Computing*, 107:107360, 2021.

[58] J. V. Fonseca Abreu, C. Ghedini Ralha και J. J. Costa Gondim. *Twitter Bot Detection with Reduced Feature Set*. *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, σελίδες 1–6, 2020.

[59] A. P. Rodrigues, R. Fernandes, A. Shetty, K. Lakshmanna και R. M. Shafi. *Real-time Twitter Spam Detection and Sentiment Analysis Using Machine Learning and Deep Learning Techniques*. *Computational Intelligence and Neuroscience*, 2022.

[60] J. Knauth. *Language-agnostic Twitter-bot Detection*. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, σελίδες 550–558, 2019.

[61] Feng Liu, Zhenyu Li, Chunfang Yang, Daofu Gong, Haoyu Lu και Fenlin Liu. *SE-GCN: A Subgraph Encoding Based Graph Convolutional Network Model for Social Bot Detection. Scientific Reports*, 14, 2024.

[62] C. Zhau, Y. Xin, X. Li, H. Zhu, Y. Yang και Y. Chen. *An Attention-Based Graph Neural Network for Spam Bot Detection in Social Networks. Applied Sciences*, 10(22), 2020.

[63] S. Feng, H. Wan, N. Wang και M. Luo. *BotRGCN: Twitter Bot Detection with Relational Graph Convolutional Networks. Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, σελίδες 236–239, 2021.

[64] S. Shi, K. Qiao, Z. Wang, J. Yang, B. Song, J. Chen και B. Yan. *Multi-scale Graph Neural Network with Signed-attention for Social Bot Detection: A Frequency Perspective. Journal of LaTeX Files*, 14(8), 2023.

[65] Md. Abul Kalam, K. Akash, Md. Khadeer και R. Vamshi. *A Survey on Twitter Bot Detection Using Graph Based Deep Learning (Graph Neural Networks). International Research Journal of Modernization in Engineering Technology and Science*, 6(3), 2024.

[66] K. Hayawi, S. Mathew, N. Venugopal, M. M. Masud και P. H. Ho. *Deeprobot: A Hybrid Deep Neural Network Model for Social Bot Detection Based on User Profile Data. Social Network Analysis and Mining*, 12(1):43, 2022.

[67] Z. Ellaky, F. Benabbou, Y. Matrane και S. Qaqa. *A Hybrid Deep Learning Architecture for Social Media Bots Detection Based on BiGRU-LSTM and GloVe Word Embedding. IEEE Access*, (99), 2024.

[68] I. Kotenko, M. Kolomeets, A. Chechulin και L. Vitkova. *Hybrid Approach for Bots Detection in Social Networks Based on Topological, Textual and Statistical Features. Advances in Intelligent Systems and Computing*, 2020.

[69] M. Heidari και J. H. Jones. *BERT Model for Social Media Bot Detection*, 2022.

[70] A. Garcia-Silva, C. Berrio και J. M. Gomez-Perez. *Understanding Transformers for Bot Detection in Twitter. arXiv preprint arXiv:2104.06182*, 2021.

[71] J. Tourille, B. Sow και A. Popescu. *Automatic Detection of Bot-generated Tweets. Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, σελίδες 44–51, 2022.

[72] A. Sallah, A. Alaoui και S. Agoujil. *Fine-Tuned Understanding: Enhancing Social Bot Detection With Transformer-Based Classification. IEEE Access*, 2024.

[73] S. Feng, H. Wan, N. Wang, J. Li και M. Luo. *Twibot-20: A Comprehensive Twitter Bot Detection Benchmark. Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, σελίδες 4485–4494, 2021.

[74] Twitter Developer Platform. *User object | Docs*, 2024. Available: https://developer.x.com/en/docs/x-api/v1/data-dictionary/object-model/user.