Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΥΠΟΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Ανίχνευση Bots στο Twitter με Τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κωνσταντίνου Τσιγγέλη

Επιβλέπων: Ασκούνης Δημήτρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Ηλεκτρικών Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων
Εργαστήριο Συστημάτων Υποστήριξης Αποφάσεων και Διοίκησης

Ανίχνευση Bots στο Twitter με Τεχνικές Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κωνσταντίνου Τσιγγέλη

Επιβλέπων: Ασκούνης Δημήτρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 19η Οκτωβρίου, 2023.

.....	.....	.....
Ασκούνης Δημήτρης	Ιωάννης Ψαρράς	Ευάγγελος Μαρινάκης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2023

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

Περίληψη

Η διπλωματική αυτή εργασία ασχολείται με την ανάπτυξη ενός μοντέλου ανίχνευσης bot λογαριασμών στο Twitter με τεχνικές μηχανικής μάθησης. Το Twitter είναι ένα μέσο κοινωνικής δικτύωσης το οποίο επιτρέπει την αλληλεπίδραση των χρηστών μέσω σύντομων μηνυμάτων τα οποία ονομάζονται τουιτς (tweets). Έχει εδραιωθεί στην σημερινή κοινωνία ως ένα από τα κυρίαρχα μέσα κοινωνικής δικτύωσης ενώ συχνά παρομοιάζεται ως μία 'παγκόσμια εφημερίδα' η οποία επηρεάζει το κοινωνικοπολιτικό γίγνεσθαι και διαμορφώνει την κοινή γνώμη. Η άνθιση αυτή βέβαια του Twitter αντιμετωπίσθηκε σαν ευκαιρία από πολλούς χρήστες οι οποίοι προσπάθησαν να εκμεταλλευτούν την δύναμή του για να υλοποιήσουν δικούς τους, συχνά κακοπροαίρετους σκοπούς. Ως επακόλουθο, εμφανίστηκαν τα bots δηλαδή χρήστες οι οποίοι προκύπτουν από αυτοματοποιημένα λογισμικά και αποσκοπούν στην υπονόμευση της αξιοπιστίας και ανεξαρτησίας του Twitter.

Βασικά χαρακτηριστικά τους, το οποία καθιστούν την ανίχνευσή τους πολύ απαιτητική, είναι η ποικιλομορφία και η εξελισσιμότητά τους. Διαφορετικά είδη Bot αξιοποιούν διαφορετικά χαρακτηριστικά και μέσα του Twitter προκειμένου να προωθήσουν την ατζέντα τους. Κάποιοι τύποι επιθέσεων είναι οι ακόλουθοι: διανέμουν κακόβουλους συνδέσμους, προσποιούνται τους κοινωνικούς φίλους σε χρήστες για να αποσπάσουν επικίνδυνες και ζημιογόνες πληροφορίες, αναδημοσιεύουν ειδήσεις με μεροληπτικό περιεχόμενο προκειμένου να επηρεάσουν την κοινή γνώμη κτλ. Παράλληλα τα bot εξελίσσονται συνέχεια ώστε να ξεπερνούν τα είδη υπάρχοντα μέτρα ανίχνευσης αλλά και για να αναβαθμίσουν την αληθοφάνεια τους και ως επακόλουθο να αυξήσουν την επιρροή τους. Με τον καιρό λοιπόν γίνονται όλο και πιο ευφυή, προσομοιώνοντας την συμπεριφορά ρεαλιστικών χρηστών.

Η εργασία ανίχνευσης bot λογαριασμών είναι πολύ ουσιώδης και απαιτητική. Οι ήδη υπάρχουσες μέθοδοι γενικά μπορούν να διαιρεθούν σε δύο κατηγορίες: μέθοδοι που βασίζονται στην μηχανικη εξαγωγής χαρακτηριστικών και μέθοδοι που χρησιμοποιούν δίκτυα βαθιάς μάθησης. Οι πρώτες εξάγουν τα χαρακτηριστικα των χρηστών από τα tweets και από την πληροφορία του λογαριασμού τους και τα τροφοδοτούν σε κλασικούς ταξινομητές μηχανικής μάθησης ενώ οι μετέπειτα στηρίζονται σε αρχιτεκτονικές βαθιών νευρωνικών δικτύων. Παρά το αρχικό θετικό αποτέλεσμα, η αναζήτηση ενός μοντέλου που θα αντιμετωπίζει αποδοτικά τις απαιτήσεις του ζητήματος και θα γενικεύει στην πραγματική σφαίρα του Twitter παραμένει ανοιχτή. Στο μοντέλο που προτείνουμε χρησιμοποιούμε παράλληλα πολυτροπικές πληροφορίες για κάθε χρήστη χωρίς μηχανική χαρακτηριστικών. Συνδυάζουμε τεχνικές επιβλεπόμενης και μη επιβλεπόμενης μηχανικής μάθησης ούτως ώστε να κατηγοριοποιήσουμε τους χρήστες σε Bot ή γνήσιους χρήστες. Ιδιαίτερα εφαρμόζουμε μοντέλα Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing), ώστε να επιτευχθεί η εξαγωγή πληροφορίας από αδόμητες πηγές δεδομένων (tweets) και νευρωνικά δίκτυα για να βρούμε την αναπαράσταση των χαρακτηριστικών κάθε χρήστη, επιλέγοντας αυτά που βελτιστοποιουν το μοντέλο μας. Ακολούθως, κατασκευάζουμε έναν ετερογενή γράφο που καλύπτει τις σχέσεις ακολουθησης που αναπτύσσονται στο Twitter (follower και following) και εφαρμόζουμε δομές νευρωνικών δικτύων σε γράφους (Graph Neural Networks) ούτως ώστε να συμπεριλάβουμε στην πρόβλεψή μας και την κοινωνική δραστηριότητα των χρηστών. Τέλος, βασισμένοι στο ολοκληρωμένο σύνολο δεδομένων Twibot-20 που αποτελεί σημείο αναφοράς εκτελούμε πειράματα που αναδεικνύουν την αποδοτικότητα του μοντέλου μας και την ανταγωνιστική του επίδοση σε σχέση με τις υπάρχουσες υλοποιήσεις.

Λέξεις Κλειδιά — Twitter , Bot , Μέσα Κοινωνικής Δικτύωσης , Νευρωνικά Δίκτυα σε Γράφους, Επεξεργασία Φυσικής Γλώσσας , tweets , Μηχανική Μάθηση

Abstract

This thesis aims to the development of a bot account detection model on Twitter using machine learning techniques. Twitter is a social media platform that allows users to interact through short messages called tweets. It has become firmly established in today's society as one of the dominant social media platforms and is often referred to as a "global newspaper" that influences social and political events and shapes public opinion. The rise of Twitter, however, was seen as an opportunity by many users who tried to exploit its power for their own, often malicious, purposes. As a result, bots emerged, which are users generated by automated software aiming to undermine Twitter's credibility and independence.

Key characteristics that make bot detection challenging are their diversity and adaptability. Different types of bots leverage different features and means within Twitter to promote their agendas. Some types of attacks include distributing malicious links, impersonating social friends to extract dangerous and harmful information, reposting biased news to influence public opinion, and more. Bots constantly evolve to surpass existing detection measures and enhance their authenticity to increase their influence. Over time, they become increasingly intelligent, simulating the behavior of real users.

The task of bot account detection is crucial and demanding. Existing methods can generally be categorized into two groups: methods based on feature extraction and methods using deep learning networks. The former extracts user features from tweets and account information and feeds them into traditional machine learning classifiers, while the latter relies on deep neural network architectures. Despite initial positive results, finding a model that efficiently addresses the challenges of the issue and generalizes to the real Twitter sphere remains an open question. In the proposed model we utilize multi-modal information for each user without relying solely on feature engineering.We combine supervised and unsupervised machine learning techniques to categorize users into bots or genuine users. Specifically, we apply Natural Language Processing (NLP) models to extract information from unstructured data (tweets) and neural networks to find representations of user features, selecting those that optimize our model. Subsequently, we construct a heterogeneous graph that covers the following relationships that develop on Twitter: follower and following. We apply Graph Neural Networks (GNNs) to include social activity in our predictions. Finally, based on the integrated Twibot-20 dataset that serves as a reference point, we conduct experiments that highlight the efficiency of our model and its competitive performance compared to existing implementations.

Keywords — Twitter , Bot , Social Media Platforms , Graph Neural Networks , Natural Language Processing , tweets , Machine Learning

Ευχαριστίες

Η διπλωματική αυτή σηματοδοτεί τη λήξη ενός κεφαλαίου στη ζωή μου, αυτού των φοιτητικών μου χρόνων στη σχολή ΗΜΜΥ του ΕΜΠ. Μέσα σε διάστημα πέντε ετών, βίωσα πρωτόγνωρες εμπειρίες - άγχη, αγωνίες, αλλά και στιγμές ενθουσιασμού και χαράς - γνώρισα εκπληκτικούς ανθρώπους και φίλους, αποκόμισα γνώσεις και αναμνήσεις που διαμόρφωσαν το άτομο που είμαι σήμερα, γράφοντας αυτό το κείμενο.

Αρχικά, θα ήθελα να ευχαριστήσω τον κύριο Ασκούνη Δημήτρη , καθηγητή της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσοβίου Πολυτεχνείου, ο οποίος υπήρξε επιβλέπων της διπλωματικής αυτής εργασίας και μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα.

Έπειτα, οφείλω ένα ευχαριστώ στον Ηλία Λουκά, ερευνητής του εργαστηρίου Συστημάτων Υποστήριξης Αποφάσεων και Διοίκησης, ο οποίος μου παρείχε διαρκή καθοδήγηση και στήριξη κατά την εκπόνηση της εργασίας μέσα σε ένα πνεύμα εξαιρετικής συνεργασίας.

Στη συνέχεια, νιώθω την ανάγκη να ευχαριστήσω τους φίλους που έκανα στη σχολή αλλά και την οικογένειά μου που βρίσκονταν πάντα δίπλα μου για να με στηρίζουν.

Κωνσταντίνος Τσιγγέλης

Οκτώβριος 2023

Περιεχόμενα

Λίστα Σχημάτων

Λίστα Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Κοινωνικά Δίκτυα και Μέσα Κοινωνικής Δικτύωσης

Τα κοινωνικά δίκτυα (online social networks) είναι κοινωνικές δομές οι οποίες απαρτίζονται από άτομα τα οποία συνδέονται ή δεσμεύονται μεταξύ τους με έναν ή περισσότερους τύπους αλληλεξάρτησης όπως η φιλία, η συγγένεια, τα κοινά ενδιαφέροντα ή τις οικονομικές συναλλαγές. Ο όρος ακούστηκε για πρώτη φορά από τους Durkheim και Tonnies [1] οι οποίοι προωθούσαν την ιδέα των κοινωνικών δικτύων στις θεωρίες τους και την έρευνα των κοινωνικών ομάδων. Στα χρόνια που ακολούθησαν ο τομέας έλαβε μεγάλη αναγνώριση εφόσον συνδέθηκε άμεσα με την διαμόρφωση της ανθρώπινης πρακτικής και της κοινωνικής ταυτότητας , φτάνοντας έτσι στους Walker et al. [2], οι οποίοι όρισαν ως κοινωνικό δίκτυο το άθροισμα των προσωπικών επαφών μέσω των οποίων το άτομο διατηρεί την κοινωνική του ταυτότητα, λαμβάνει συναισθηματική υποστήριξη, υλική ενίσχυση και συμμετοχή στις υπηρεσίες, έχει πρόσβαση στις πληροφορίες και δημιουργεί νέες κοινωνικές επαφές.

Η κοινωνική δικτύωση βρήκε πρόσφορο έδαφος και εδραιώθηκε στην πραγματικότητα εκατομμυρίων χρηστών όταν συνδέθηκε με το διαδίκτυο. Στο διαδίκτυο χρησιμοποιούνται ιστοσελίδες γνωστές ως ιστότοποι ή πλατφόρμες κοινωνικής δικτύωσης οι οποίες λειτουργούν ως εικονικές κοινότητες χρηστών που επιτρέπουν στα άτομα να παρουσιάσουν τους εαυτούς τους, να αναπτύξουν την κοινωνική τους δικτύωση, καθώς και να δημιουργήσουν ή να διατηρήσουν συνδέσεις με άλλους χρήστες με τους οποίους μοιράζονται κοινά ενδιαφέροντα, χόμπι ,πολιτικές πεποιθήσεις κτλ. Έκαναν την εμφάνισή τους το 2002 με το Friendster ενώ στα επόμενα χρόνια γνώρισαν τεράστια ανάπτυξη βελτιώνοντας συνέχεια την εμπειρία και τις δυνατότητες των χρηστών. Οι πιο γνωστές από αυτές τις ιστοσελίδες είναι το Facebook, το Twitter, το Instagram, το LinkedIn και το Pinterest. Μάλιστα, η καθιέρωσή τους στην διάπλαση διαπροσωπικών σχέσεων αλλά και το γεγονός ότι αποτελούν αναπόσπαστο κομμάτι της ανθρώπινης έκφρασης καθιστούν τα μέσα κοινωνικής δικτύωσης πηγές πληροφορίας για ανάλυση της ψυχοσύνθεσης των ατόμων [3],[4].

1.2 Twitter

Το Twitter , το οποίο αποτελεί και το μέσο κοινωνικής δικτύωσης που θα με απασχολήσει στην παρούσα διπλωματική εργασία, είναι ένα μέσο κοινωνικής δικτύωσης που επιτρέπει στους χρήστες του να στέλνουν και να διαβάζουν σύντομα μηνύματα τα οποία ονομάζονται τουίτς (tweets). Δημιουργήθηκε την 21 Μαρτίου του 2006 από τον Τζακ Ντόρσεϊ και σήμερα αριθμεί πάνω από 200 δισεκατομμύρια χρήστες παγκοσμίως και περίπου 140 εκατομμύρια tweets ημερησίως. Η βασική ιδέα γύρω από τον συγκεκριμένο ιστότοπο είναι αυτή των ακολούθων (followers). Όταν επιλέγει ο χρήστης να συνδεθεί με έναν άλλον , τότε τα tweets του δεύτερου εμφανίζονται με αντίστροφη χρονολογική σειρά στην κεντρική σελίδα του πρώτου. Το περιεχόμενο των tweets μπορεί να εκφράζει την προσωπική άποψη του χρήστη για κάποιο επίκαιρο ψυχαγωγικό, κοινωνικό ή πολιτικό γεγονός, την αναπαραγωγή κάποιας πρόσφατης είδησης ή την ενημέρωση για κάποιο ζήτημα.

Συχνά οι χρήστες του Twitter χρησιμοποιούν hashtags στα tweets τους δηλαδή λέξεις ή φράσεις που ξεκινούν με το σύμβολο #, και χρησιμοποιούνται για την ομαδοποίηση των tweets με βάση το θέμα τους και mentions δηλαδή λέξεις που περιέχουν το σύμβολο @ ακολουθούμενο από ένα όνομα χρήστη σε κάποιο tweet ώστε να στέλνεται το συγκεκριμένο tweet κατευθείαν σε αυτόν το χρήστη. Επίσης υπάρχει η δυνατότητα ένα tweet να περιλαμβάνει φωτογραφίες, βίντεο και συνδέσμους. Ο κάθε χρήστης ακολουθεί (Following) τα άτομα για τα οποία θέλει να ενημερώνεται όταν δημοσιεύουν ένα tweet και αντιστοίχως ακολουθείται και αυτός από άλλους χρήστες (Followers). Οι χρήστες έχουν τη δυνατότητα να απαντήσουν στα tweets άλλων χρηστών (reply) ή να τα αναδημοσιεύσουν (retweet) ενώ υπάρχει η δυνατότητα για ανταλλαγή απευθείας ιδιωτικών μηνυμάτων μεταξύ των χρηστών (direct messages).

Η διαφοροποίηση του Twitter απο τα υπόλοιπα μέσα κοινωνικής δικτύωσης και η πρωτοτυπία που παρουσιάζει στον τρόπο ανάδρασης και ενημέρωσης των χρηστών έχει εδραιώσει την θέση του στην κοινωνία ως μία ‘παγκόσμια εφημερίδα’ η οποία επηρεάζει το κοινωνικοπολιτικό γίγνεσθαι. Πληθώρα αναδημοσιεύσεων ειδήσεων αλλά και πολυποίκιλες αναλύσεις συμβάλλουν στην πολύπλευρη ενημέρωση των χρηστών και στην διαμόρφωση της κοινής γνώμης.

1.3 Bots(software robots) σε Κοινωνικά Δίκτυα

Με τον όρο Bot περιγράφουμε μία ποικιλία αυτοματοποιημένων λογισμικών γενικού ή ειδικού σκοπού όπως τα αυτοματοποιημένα συστήματα σχεδιασμένα να συνομιλούν με ανθρώπους (ChatBots) [5]. Παρόλα αυτά στα πλαίσια της συγκεκριμένης διπλωματικής μας ενδιαφέρουν τα Bot που λειτουργούν στα κοινωνικά δίκτυα (Social Bots). Για αυτό τον λόγο απαιτούνται πιο

συγκεκριμένοι όροι για να προσδιορίσουμε τόσο την ιδιαιτερότητά τους αλλά και την ποικιλομορφία τους. Οι Morstatter et al. [6] περιέγραψαν τα κοινωνικά Bot σαν αυτοματοποιημένους λογαριασμούς οι οποίοι δρουν σε κάποιο διαδικτυακό μέσο κοινωνικής δικτύωσης. Με παρόμοιο τρόπο περιέγραψαν οι Forelle et al. [7] τα κοινωνικά Bot στο Twitter σαν υπολογιστικά προγράμματα τα οποίο συνομιλούν, ποστάρουν και ανεβάζουν tweets κατά την δική τους βούληση. Μετέπειτα, οι Grimme et al. [8] πρότειναν έναν καινούριο ορισμό για τα Social Bots. Τα παρομοίασαν με ανεξάρτητα προγράμματα σχεδιασμένα να εκπληρώσουν έναν συγκεκριμένο στόχο μέσω της διάδρασής τους με τους χρήστες στα κοινωνικά δίκτυα. Τέλος, οι Ferrara et al. [9] κατάφεραν να τα περιγράψουν ολοκληρωμένα και συμπεριληπτικά υπογραμμίζοντας ότι είναι λογαριασμοί που προκύπτουν απο υπολογιστικούς αλγόριθμους και παράγουν περιεχόμενο και αλληλεπιδρούν με τους υπόλοιπους χρήστες προσπαθώντας να μιμηθούν την ανθρώπινη συμπεριφορά.

Καθώς λοιπόν τα μέσα κοινωνικής δικτύωσης αποκτούν όλο μεγαλύτερη δύναμη και κοινωνική επιρροή, μεμονωμένα άτομα ή ολόκληροι οργανισμοί προσπαθούν να εκμεταλλευτούν τις δυνατότητες αυτές για να υλοποιήσουν δικούς τους, συχνά κακοπροαίρετους, σκοπούς. Για αυτό δημιουργούνται τα Social Bot. Ιδιαίτερα κάποια από τα είδη των κακόβουλων κοινωνικών bot που έχουν ανιχνευθεί είναι τα ακόλουθα. Οι Wang et al. [10] αναφέρθηκε στα bot που διανέμουν κακόβουλους συνδέσμους (links) και ανεπιθύμητα μηνύματα. Οι Elyasar et al. [11] περιέγραψαν μία κατηγορία bot που προσποιούνται τους κοινωνικούς φίλους σε χρήστες ούτως ώστε να αποσπάσουν επικίνδυνες και ζημιογόνες πληροφορίες. Οι Abokhodair et al. [12] αναφέρθηκε στα bot που προσπαθούν να επηρεάσουν την κοινή γνώμη πάνω σε πολιτικά ζητήματα και διαμάχες αναδημοσιεύοντας ειδήσεις με μεροληπτικό περιεχόμενο ή με αλλοιωμένες πληροφορίες για να αποπροσανατολίσουν τους χρήστες. Οι Ratkiewicz et al. [13] περιέγραψαν τα Social Bots που προσπαθούν να επηρεάσουν τα εκλογικά αποτελέσματα αναδημοσιεύοντας ψευδείς ειδήσεις ή κάνοντας συνεχόμενα tweet δείχνοντας την στήριξή τους προς κάποιο υποψήφιο. Τέλος, ο Cresci [14] ανίχνευσε συντονισμένες ομάδες από Bot τα οποία προωθούν μετοχές μικρής αξίας ούτως ώστε να επηρεάσουν την χρηματιστηριακή αγορά.

Μάλιστα με τον καιρό τα Social Bot εξελίσσονται συνεχόμενα και παράλληλα γίνονται πιο ευφυη προσομοιώνοντας την συμπεριφορά γνήσιων χρηστών σε επίπεδο που η διάκρισή τους τόσο από τους απλούς χρήστες όσο και από τους αλγορίθμους ανίχνευσης είναι αδύνατη. Εφόσον λοιπόν ευδοκιμούν στο περιβάλλον των διαδικτυακών κοινωνικών δικτύων συχνά οι επιπτώσεις που προκύπτουν από την δράση τους επηρεάζουν την κοινή γνώμη και ως επακόλουθο την πορεία της κοινωνίας. Ιδιαίτερα σύμφωνα με τους Cresci et al. [15] κατά την διάρκεια των δημοτικών εκλογών στην Ρώμη το 2014 ανιχνεύθηκαν πάνω από 1000 Bot που χρησιμοποιήθηκαν από έναν υποψήφιο για να δημοσιοποιήσει την πολιτική του μέσω του Twitter. Στο διάστημα μεταξύ 2014 και 2017 σύμφωνα με τους Broniatowski et al. [16] κοινωνικά Bot συμμετείχαν σε διαλόγους που αφορούσαν το εμβολιο κατα του ιού Covid-19 προωθώντας αντιεμβολιαστικές πεποιθήσεις. Το 2019 πανω απο 5000 Bot ανιχνεύθηκαν στο Twitter τα οποία ήταν υπέρ του Trump και παρότρειναν τον κόσμο να διαμαρτυρηθεί απέναντι στην απάτη της ‘‘Russiagate’’. Επιπλέον, στις προεδρικές εκλογές των ΗΠΑ το 2016 ,σύμφωνα με τους Bessi και Ferrara [15], χιλιάδες Bot στο Twitter συνέβαλαν στην εκλογή του Trump είτε επειδή έδειχναν την στήριξή τους στον Trump είτε με το να επιτίθενται στους πολιτικούς του αντιπάλους.

1.4 Αντικείμενο της Διπλωματικής

Η παρούσα διπλωματική εργασία επικεντρώνεται στην ανίχνευση των Bot λογαριασμών που δρουν στο Twitter με χρήση γράφων. Προσεγγιστικά, τα Bot αντιπροσωπεύουν το 8.5% των συνολικών χρηστών του Twitter όπως αποκάλυψε ο ίδιος ο ιστότοπος [18] ενώ μία έρευνα στα Social bots ανέδειξε ότι από όλους τους Αγγλόφωνους χρήστες του Twitter , το 9% με 15% παρουσιάζει συμπεριφορά παρόμοια με αυτή των Bot [19]. Βασισμένοι σε ένα εκτενές σύνολο δεδομένων που περιλαμβάνει τις πληροφορίες χιλιάδων χρηστών όπως αυτές προκύπτουν απευθείας απο την πλατφόρμα του Twitter(Twitter Application Programming Interface) εφαρμόζουμε τεχνικές επιβλεπόμενης και μη επιβλεπόμενης μηχανικής μάθησης ούτως ώστε να κατηγοριοποιήσουμε τους χρήστες σε Bot ή γνήσιους χρήστες. Ιδιαίτερα, για τον σκοπό αυτό, χρησιμοποιούνται μοντέλα Επεξεργασίας Φυσικής Γλώσσας σε συνδυασμό με αλγορίθμους συσταδοποίησης (clustering algorithms), ώστε να επιτευχθεί η εξαγωγή πληροφορίας από αδόμητες πηγές δεδομένων (tweets) και νευρωνικά δίκτυα για να βρούμε την αναπαράσταση των χαρακτηριστικών κάθε χρήστη , επιλέγοντας αυτά που βελτιστοποιουν το μοντέλο μας. Τέλος , εφαρμόζουμε δομές νευρωνικών δικτύων σε γράφους ούτως ώστε να συμπεριλάβουμε στην πρόβλεψή μας και την κοινωνική δραστηριότητα των χρηστών και συγκρίνουμε τα αποτελέσματά μας με σύγχρονες και ανταγωνιστικές υλοποιήσεις. Οι κύριες συνεισφορές μας συνοψίζονται ως εξής:

- Προτείνουμε ένα πρωτοπόρο μοντέλο για την πραγματοποίηση γενικευμένης ανίχνευσης bot στο Twitter. Η αρχιτεκτονική μας είναι ένα πλαίσιο end-to-end που χρησιμοποιεί από κοινού σημασιολογικές , ιδιωματικές και πληροφορίες γειτονιάς των χρηστών χωρίς την ανάγκη για μηχανική χαρακτηριστικών.

- Βασιζόμενοι στην αρχιτεκτονική του μοντέλου BotRGCN [20] εξελίσσουμε την επίδοσή του και την δράση του εφαρμόζοντας αλγορίθμους clustering για την ομαδοποίηση των tweets σε θεματικές ομάδες. Ιδιαίτερα, η έρευνα αυτή εισάγει το πρώτο μοντέλο το οποίο συνδυάζει αλγορίθμους συσταδοποίησης (KMeans clustering) με νευρωνικά δίκτυα σε γράφους (GNN) προκειμένου να ενσωματώσει με μεγαλύτερη εκφραστικότητα την κοινωνική δραστηριότητα των χρηστών και να αναβαθμίσει την επίδοση της ανίχνευσης bot.

- Διενεργούμε εκτενείς πειραματικές μελέτες σε ένα περιεκτικό σύνολο δεδομένων Twibot-20. Τα αποτελέσματα δείχνουν ότι το μοντέλο μας εμφανίζει ανταγωνιστικές επιδόσεις υπερτερώντας των περισσότερων προτεινόμενων μεθόδων. Επιπλέον, περαιτέρω ανάλυση επιβεβαιώνει την αποτελεσματικότητα της συνδυαστικής αρχιτεκτονικής.

1.5 Οργάνωση Κειμένου

Η εργασία διαρθρώνεται σε έξι διακριτά κεφάλαια. Το Κεφάλαιο 2 παρέχει στον αναγνώστη το απαραίτητο θεωρητικό υπόβαθρο αναφορικά με τις βασικές αρχές της μηχανικής μάθησης και τα δομικά συστατικά των νευρωνικών δικτύων , περιγράφει μοντέλα Επεξεργασίας Φυσικής Γλώσσας από την περιοχή της Βαθιάς Μάθησης και αναλύει την λειτουργία τεχνητών νευρωνικών δικτύων σε γράφους . Στο Κεφάλαιο 3 παρουσιάζονται και αναλύονται ήδη υπάρχουσες μέθοδοι και μοντέλα που αποσκοπούν στην ανίχνευση bot στο Twitter . Στο Κεφάλαιο 4 περιγράφεται η μεθοδολογία που ακολουθήσαμε για την ανάπτυξη της αρχιτεκτονικής δικτύου του μοντέλου μας. Στο Κεφάλαιο 5 παρουσιάζονται τα πειράματα και τα αποτελέσματά μας , συνοδευόμενα από τον κατάλληλο σχολιασμό . Τέλος, το Κεφάλαιο 6 παρέχει μια τελική σύνοψη της εργασίας, καθώς και μελλοντικές επεκτάσεις της από την επιστημονική κοινότητα.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Μηχανική Μάθηση

Η Μηχανική Μάθηση είναι ένας κλάδος της Τεχνητής Νοημοσύνης (AI: Artificial Intelligence) και της Επιστήμης των Υπολογιστών που εστιάζει στη χρήση δεδομένων και αλγορίθμων για τη μίμηση του τρόπου με τον οποίο μαθαίνουν οι άνθρωποι με σταδιακά βελτιωνόμενη ακρίβεια, δίχως ρητό προγραμματισμό [21]. Στόχος της Μηχανικής Μάθησης είναι η εκπαίδευση συστημάτων τα οποία μπορούν να κατανοήσουν και να χειριστούν αποδοτικά τεράστιους όγκους αδόμητων δεδομένων με απώτερο σκοπό την εκτέλεση κάποιας πρόβλεψης είτε την λήψη αποφάσεων. Γενική αρχή πάνω στην οποία θεμελιώνονται τα συστήματα αυτά είναι η εκμάθηση μέσω της ‘’εμπειρίας’’. Με άλλα λόγια, η επίδοση τους βελτιώνεται όσο αποκτούν μεγαλύτερη εμπειρία στην εκτέλεση μίας διεργασίας, χωρίς να απαιτείται ανθρώπινη υποβοήθηση ή κάποια προγραμματιστική ανάδραση. Μέσω της "επαφής" τους με ένα πλήθος διαφορετικών εισόδων παράγουν αποτελέσματα και με βάση αυτά, ανεξαρτήτως της ορθότητας ή μη των αποτελεσμάτων τους, αυτοβελτιώνονται, έως ότου να παράγουν όσο το δυνατόν καλύτερο αποτέλεσμα πάνω στην εργασία για την οποία δουλεύουν.

Η Μηχανική Μάθηση και οι επιστημονικές καινοτομίες που εισάγει αποτελούν κυρίαρχο κομμάτι που απασχολεί την σύγχρονη επιστημονική κοινότητα. Η τεράστια αναγνώριση και εξέλιξη που έχει λάβει οφείλεται τόσο στις προοπτικές της όσο και στις τεράστιες βάσεις δεδομένων που έχουμε σήμερα στην διάθεσή μας. Πληθώρα συνόλων δεδομένων τα οποία σχετίζονται με εικόνα, κείμενο ή ήχο τροφοδοτούνται σε αλγορίθμους Μηχανικής Μάθησης με σκοπό την εύρεση μοτίβων τα οποία θα οδηγήσουν το σύστημα στην λήψη βέλτιστων αποφάσεων. Μάλιστα, όπως τονίζουν οι ειδικοί στην επιστήμη δεδομένων (data scientists), η ιδιομορφια που συνοδεύει την Μηχανική Μάθηση είναι ότι δεν υπάρχει κάποιος αλγόριθμος που να λύνει βέλτιστα όλα τα προβλήματα. Το είδος του αλγορίθμου που χρησιμοποιείται εξαρτάται από το είδος του προβλήματος, τον αριθμό των μεταβλητών, το είδος του μοντέλου που θα του ταίριαζε καλύτερα και ούτω καθεξής. Η ποικιλομορφία λοιπόν που συνοδεύει την Μηχανική Μάθηση έχει οδηγήσει

στην δημιουργία πολλαπλών μοντέλων και αρχιτεκτονικών μάθησης, των οποίων οι αρχές λειτουργίας θα αναλυθούν ακολούθως.

Στον τομέα της Μηχανικής Μάθησης, τα προβλήματα ταξινομούνται σε ευρείες κατηγορίες με βάση το πώς μαθαίνει ένα υπολογιστικό σύστημα και το είδος της ανάδρασης που λαμβάνει. Παρακάτω αναλύονται οι πιο διαδεδομενες κατηγορίες μάθησης του κλάδου.

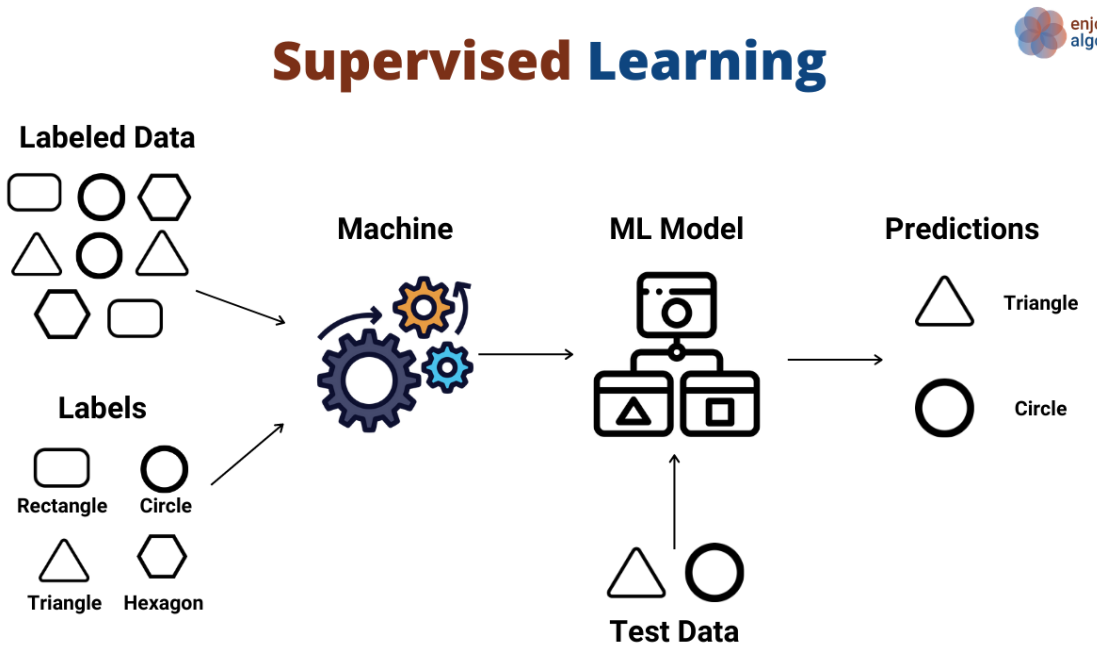
(I) Επιβλεπόμενη Μάθηση (Supervised Learning)

Στην επιβλεπόμενη μάθηση, ένα σύστημα τροφοδοτείται με ένα σύνολο δεδομένων τα παραδείγματα του οποίου έχουν ετικέτες (labels). Κάθε δείγμα λοιπόν αντιστοιχίζεται σε μία ετικέτα η οποία ανάλογα με την διεργασία μπορεί να αποτελεί την κατηγορία, την κλάση ή οποιαδήποτε άλλη ιδιότητα του δείγματος. Σκοπός λοιπόν στην επιβλεπόμενη μάθηση είναι η δημιουργία ενός αλγορίθμου ο οποίος μαθαίνει μία συνάρτηση απεικόνισης απο την είσοδο στην έξοδο.

$$Y = f(X) \qquad (2.1.1)$$

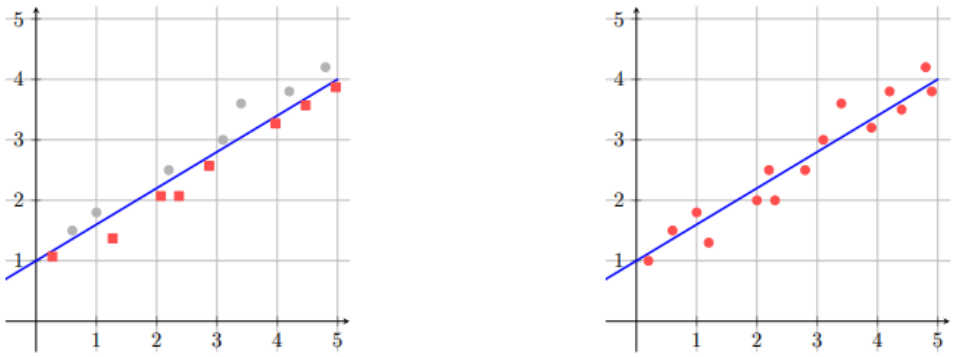
X :μεταβλητές εισόδου, Y:μεταβλητές εξόδου, f:συνάρτηση απεικόνισης

Επομένως, η επίδοση του αλγορίθμου εξαρτάται άμεσα από το πόσο καλά θα προσεγγίσουμε την συνάρτηση απεικόνισης f ώστε για κάθε δείγμα εισόδου να μπορούμε να προβλέψουμε την ορθή μεταβλητή εξόδου. Η διαδικασία της μάθησης ακολουθεί την εξής ροή: Αρχικά χωρίζουμε το σύνολο δεδομένων σε ένα σύνολο εκπαίδευσης (training set) και σε ένα σύνολο αξιολόγησης (test set). Κατά την εκπαίδευση τροφοδοτούμε το training set στον αλγόριθμο μηχανικής μάθησης ο οποίος λαμβάνει επαναληπτικά προβλέψεις της εξόδου και προβαίνει σε διορθωτικές κινήσεις με στόχο τη μείωση του σφάλματος. Η μάθηση τερματίζει μόλις ο αλγόριθμος φτάσει σε ένα ικανοποιητικό επίπεδο απόδοσης και στην συνέχεια ελέγχεται η επίδοσή του κάνοντας προβλέψεις πάνω στο test set.



Σχήμα 2.1.1: Ροή εργασιών στην επιβλεπόμενη μάθηση [22].

Τα προβλήματα επιβλεπόμενης μάθησης διακρίνονται σε δύο κατηγορίες: τα προβλήματα ταξινόμησης (classification problems) και τα προβλήματα παλινδρόμησης(regression problems). Τα πρώτα αφορούν στην πρόβλεψη των κατηγοριών ή κλάσεων που ανήκει μια άγνωστη παρατήρηση ενώ τα δεύτερα αφορούν στην απεικόνιση ενός δείγματος εισόδου σε μία συνεχής τιμή εξόδου, όπως έναν ακέραιο ή μια τιμή κινητής υποδιαστολής.



(α) Ταξινόμηση *(β) Παλινδρόμηση*

Σχήμα 2.1.2: Προβλήματα επιβλεπόμενης Μηχανικής Μάθησης.

(II) Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning)

Στην μη επιβλεπόμενη μάθηση το σύνολο δεδομένων δεν περιέχει τις ετικέτες των δειγμάτων. Με άλλα λόγια, ο χρήστης έχει στην διάθεσή του μόνο το σύνολο των διανυσμάτων εκπαίδευσης X χωρίς την αντιστοίχιση σε κάποια τιμή στόχο. Η μη επιβλεπόμενη μάθηση μελετά τον τρόπο με τον οποίο τα συστήματα τεχνητής νοημοσύνης μπορούν να μάθουν να αναπαριστούν μοτίβα τα οποία αντανακλούν την στατιστική δομή των δεδομένων εισόδου χωρίς την ύπαρξη κάποιας ανάδρασης αξιολόγησης [23].

Καθώς τα μη επισημασμένα δεδομένα είναι σε μεγαλύτερη αφθονία από τα επισημασμένα, η μέθοδοι μηχανικής μάθησης που επιστρατεύουν τεχνικές μη επιβλεπόμενης μάθησης είναι ιδιαίτερα ωφέλιμες. Κυρίαρχη εφαρμογή βρίσκουν στην εκμάθηση και παραγωγή προτύπων και χαρακτηριστικών (feature learning and engineering) από το σύνολο δεδομένων, τα οποία στην συνέχεια επιτρέπουν στην υπολογιστική μηχανή να βρει τις αναπαραστάσεις που απαιτούνται για την ταξινόμησή τους. Η μεγαλύτερη υποκατηγορία προβλημάτων της μη επιβλεπόμενης μάθησης είναι η συσταδοποίηση (Clustering). Στην συσταδοποίηση διαιρούμε το σύνολο των μη επισημασμένων δεδομένων σε συστάδες (clusters) έτσι ώστε παρόμοια σημεία δεδομένων να ανήκουν στην ίδια συστάδα. Με απλά λόγια, ο στόχος της διαδικασίας αυτής είναι να διαχωρίσει ομάδες με παρόμοια χαρακτηριστικά και να τις αντιστοιχίσει σε συστάδες.

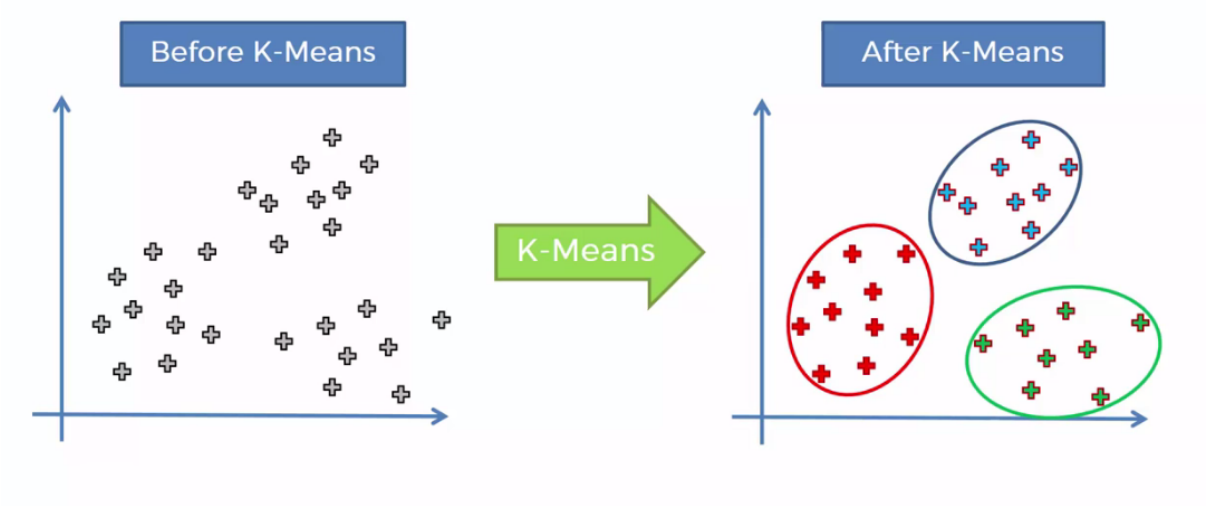
Ο πιο διάσημος αλγόριθμος συσταδοποίησης είναι ο K-Means. Η διαδικασία ακολουθεί έναν εύκολο και απλό τρόπο για να ταξινομήσει το σύνολο δεδομένων σε έναν συγκεκριμένο αριθμό συστάδων. Αφού λοιπόν καθορίσουμε τον αριθμό συστάδων στη συνέχεια κάθε δείγμα του συνόλου δεδομένων αντιστοιχίζεται τυχαία σε κάποια συστάδα. Η κύρια ιδέα του αλγορίθμου είναι να βρούμε τα κέντρα των συστάδων που ομαδοποιούν βέλτιστα τα δεδομένα μας. Ως κέντρο ορίζουμε την μέση τιμή των δειγμάτων που έχουμε αναθέσει σε κάποιο cluster. Για τον σκοπό αυτό, ο αλγόριθμος επαναληπτικά αναθέτει κάθε δείγμα του συνόλου δεδομένων στην συστάδα με το πλησιέστερο κέντρο και επαναυπολογίζει την τιμή των κέντρων μέχρι αυτά να σταθεροποιηθούν ή το ύψος της μεταβολής τους να είναι μικρότερο από μία τιμή κατωφλίου (threshold).

Algorithm 1 K-means Algorithm

Input: Data Matrix $A \in \mathcal{R}^{M \times N}$, number of clusters (K) and number of iterations (I).
Output: K clusters with its centroids $(cntd)$.

1: **Initialization** Randomly selects K number of elements from the matrix A. Initially set them as $cntd^k = A_k$ for $k = 1$ to K. Initially $cltr^k$ is a null vector for $k = 1$ to K.
2: **for** $i \leftarrow 1$ to I **do**
3: **for** $j \leftarrow 1$ to N **do**
4: **for** $k \leftarrow 1$ to K **do**
5: $\lambda = \text{argmin} \, |A_j - cntd^k|_2^2$
6: **end for**
7: $cltr^\lambda = [cltr^\lambda A_j]$
8: **end for**
9: $d \in \mathcal{R}^{M \times 1} = 0$
10: **for** $k \leftarrow 1$ to K **do**
11: **for** $j \leftarrow 1$ to $size(cltr^k)$ **do**
12: $d = d + cltr_j^k$
13: **end for**
14: $cntd^k = d/size(cltr^k)$
15: **end for**
16: **end for**

Σχήμα 2.1.3: Υλοποίηση του αλγορίθμου K-means [24].

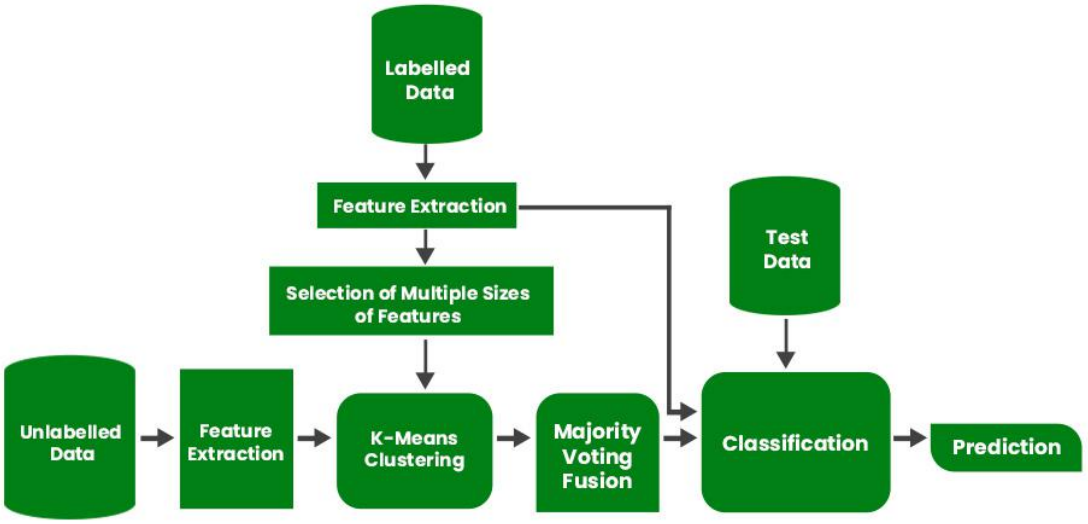


Σχήμα 2.1.4: Οπτικοποίηση της συσταδοποίησης μέσω K-Means [24].

(III) Ημι-επιβλεπόμενη Μάθηση (Semi-supervised Learning)

Η Ημι-επιβλεπόμενη μηχανική μάθηση συνδυάζει τις τεχνικές της επιβλεπόμενης και της μη-επιβλεπόμενης μάθησης προκειμένου να επιλύσει τις θεμελιώδεις προκλήσεις τους. Η βασική ιδέα είναι ότι εκπαιδεύουμε ένα αρχικό μοντέλο σε μερικά δείγματα με ετικέτα και, στη συνέχεια, το εφαρμόζουμε επαναληπτικά στον μεγαλύτερο αριθμό δεδομένων χωρίς ετικέτα.

Μάλιστα στην περίπτωση που έχουμε ένα σύνολο δεδομένων το οποίο είναι εν μέρη επισημασμένο τα πλεονεκτήματα που αποκτούμε από την εφαρμογή τεχνικών ημι επιβλεπόμενης μάθησης είναι συχνά πολυ καρποφόρα. Σε αντίθεση με την μάθηση χωρίς επίβλεψη, η ημι-επιβλεπόμενη μάθηση λειτουργεί για μια ποικιλία προβλημάτων από την ταξινόμηση και την παλινδρόμηση έως την ομαδοποίηση και τη συσχέτιση ενώ ταυτόχρονα παρουσιάζει υψηλά επίπεδα γενίκευσης (generalization) και προσαρμογής (adaptation). Επίσης, σε αντίθεση με την εποπτευόμενη μάθηση, η μέθοδος χρησιμοποιεί μικρές ποσότητες δεδομένων με ετικέτα και επίσης μεγάλες ποσότητες δεδομένων χωρίς ετικέτα, γεγονός που μας απαλλάσσει από την επίπονη διαδικασία της ανάθεσης ετικετών και μειώνει τον χρόνο προετοιμασίας δεδομένων.

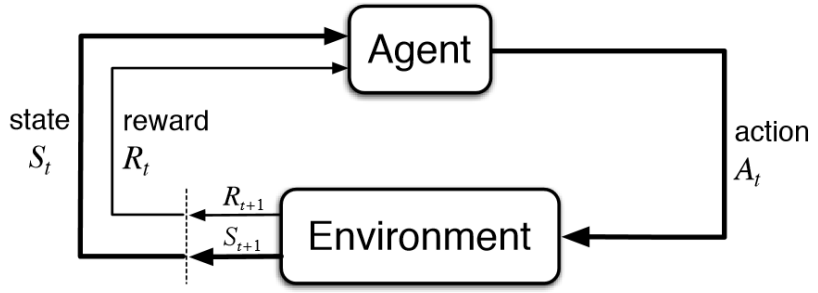


Σχήμα 2.1.5: Ροή εργασιών στην ημι-επιβλεπόμενη μάθηση [25].

(IV) Ενισχυτική Μάθηση (Reinforcement Learning)

Η Ενισχυτική Μάθηση είναι ένας τομέας της Μηχανικής Μάθησης ο οποίος ασχολείται με το πως οι πράκτορες λογισμικού (software agents) πρέπει να αναλαμβάνουν ενέργειες μέσα σε ένα περιβάλλον ώστε να μεγιστοποιηθεί κάποια έννοια ανταμοιβής. Στην επιστήμη των υπολογιστών, ένας πράκτορας λογισμικού είναι ένα πρόγραμμα υπολογιστή που ενεργεί για έναν χρήστη ή άλλο πρόγραμμα με μια σχέση αντιπροσωπείας. Η βασική δομή των μοντέλων αυτών περιλαμβάνει έναν τέτοιο πράκτορα λογισμικού ο οποίος συνδέεται με το περιβάλλον του μέσω δράσης ή αντίληψης. Σε κάθε βήμα αλληλεπίδρασης ο πράκτορας λαμβάνει σαν είσοδο κάποια ένδειξη της τρέχουσας κατάστασης του περιβάλλοντος και στη συνέχεια επιλέγει μια ενέργεια. Η ενέργεια αυτή αλλάζει την κατάσταση του περιβάλλοντας και η μεταβολή αυτή ανατροφοδοτείται στον πράκτορα μέσω ενός σήματος ενίσχυσης (reinforcement signal). Στόχος λοιπόν είναι να επιλέγονται ενέργειες οι οποίες τείνουν να μεγιστοποιήσουν το μακροχρόνιο άθροισμα αυτών των

σημάτων ενίσχυσης. Το μοντέλο εκπαιδεύεται στην ολοκλήρωση του παραπάνω στόχου μέσω της συστηματικής δοκιμής και του σφάλματος. Ο σχεδιαστής του προβλήματος ορίζει την πολιτική ανταμοιβής ενώ δεν δίνει στο μοντέλο υποδείξεις ή προτάσεις για το πώς να λύσει το πρόβλημα. Με το τέλος της διαδικασίας το μοντέλο πρέπει να καταλάβει πώς να εκτελέσει την εργασία για να μεγιστοποιήσει την ανταμοιβή και τελικά να αφομοιώσει εξελιγμένες τακτικές.



Σχήμα 2.1.6: Διάγραμμα Ενισχυτικής Μάθησης.

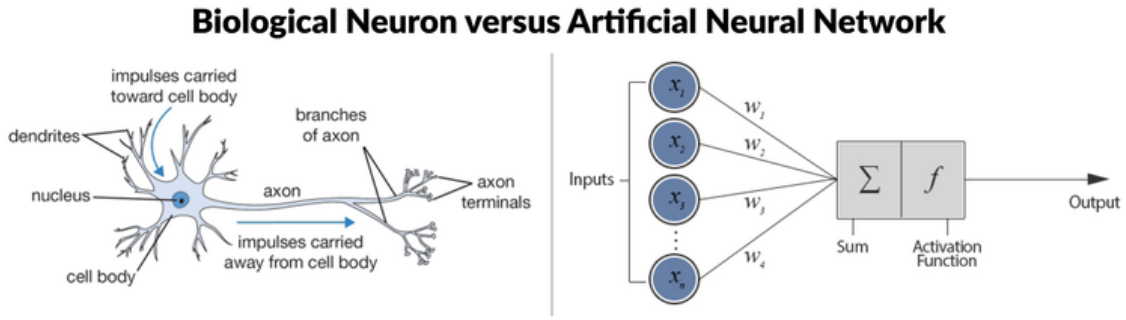
2.2 Βαθιά Μάθηση

Η Βαθιά Μάθηση συνιστά κλάδο της Μηχανικής Μάθησης ο οποίος αξιοποιεί πολλά επίπεδα μη γραμμικής επεξεργασίας πληροφοριών για επιβλεπόμενη ή μη επιβλεπόμενη εξαγωγή χαρακτηριστικων, για αναγνώριση προτύπων αλλά και για ταξινόμηση. Βασίζεται σε αλγορίθμους οι οποίοι μαθαίνουν πολυεπίπεδες αναπαραστάσεις προκειμένου να μοντελοποιήσουν σύνθετες σχέσεις που αναπτύσσονται μεταξύ των δεδομένων. Οι δύο κυρίαρχες αιτίες που οδήγησαν στην ραγδαία ανάπτυξη αρχιτεκτονικών Βαθιάς Μάθησης είναι αρχικά ο τεράστιος όγκος των συνόλων δεδομένων που οι κλασικές τεχνικές μηχανικής μάθησης αδυνατούσαν να χειριστούν αποδοτικά αλλα και τα χαρακτηριστικά υψηλού επιπέδου που παράγει. Εφόσον τα μοντέλα αυτά περιλαμβάνουν πολλαπλά επίπεδα μη γραμμικής επεξεργασίας του συνόλου δεδομένων έχουν την δυνατότητα να παράγουν πιο αντιπροσωπευτικές αναπαραστάσεις χαρακτηριστικών και να αναλύουν σε μεγαλύτερο βάθος τις σχέσεις που αναπτύσσονται μεταξύ των δεδομένων, παρουσιάζοντας έτσι σημαντική βελτίωση από τα μοντέλα παραδοσιακής μηχανικής μάθησης τα οποία χρησιμοποιούν χαρακτηριστικά χαμηλού επιπέδου. Η άνθιση της βαθιάς μάθησης συνέβαλε καθοριστικά στην πρόοδο τομέων, όπως είναι η ΄Οραση Υπολογιστών, η Επεξεργασία Φυσικής Γλώσσας και η Αναγνώριση Φωνής.

2.2.1 Νευρωνικά Δίκτυα

Ο θεμελιώδης λίθος της Βαθιάς Μάθησης είναι τα Τεχνητά Νευρωνικά Δίκτυα, περίπλοκα υπολογιστικά μοντέλα σχεδιασμένα ώστε να προσομοιώνουν τη λειτουργία των ανθρώπινων

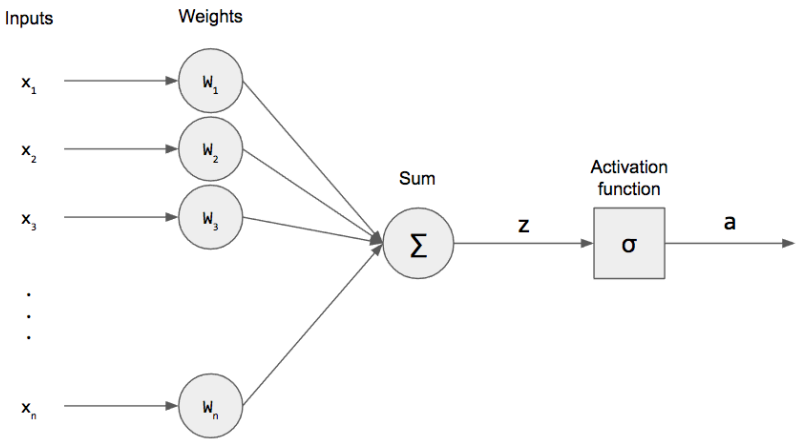
νευρώνων. Η έμπνευση λοιπόν για τα νευρωνικά δίκτυα προέρχεται αρχικά από μελέτες των μηχανισμών επεξεργασίας πληροφοριών στα βιολογικά νευρικά συστήματα και ιδιαίτερα στον ανθρώπινο εγκέφαλο ενώ πράγματι μεγάλο μέρος της τρέχουσας έρευνας στους αλγορίθμους νευρωνικών δικτύων επικεντρώνονται στην απόκτηση βαθύτερης κατανόησης της επεξεργασίας πληροφοριών σε βιολογικά συστήματα. Ένα προωθητικό νευρωνικό δίκτυο (feedforward neural network) μπορεί να θεωρηθεί ως μια μη γραμμική μαθηματική συνάρτηση που μετασχηματίζει ένα σύνολο μεταβλητών εισόδου σε ένα σύνολο μεταβλητών εξόδου. Η ακριβής μορφή του μετασχηματισμού διέπεται από ένα σύνολο παραμέτρων που ονομάζονται βάρη (weights) οι τιμές των οποίων προσδιορίζονται κατά το στάδιο της εκμάθησης (learning phase). Η διαδικασία προσδιορισμού αυτών των τιμών των παραμέτρων συχνά είναι υπολογιστικά ακριβή παρόλα αυτά μόλις καθοριστούν τα νέα δεδομένα επεξεργάζονται από το δίκτυο πολύ γρήγορα.



Σχήμα 2.2.1: Απεικόνιση των βιολογικών νευρώνων (αριστερά) και της μαθηματικής προτυποποίησής τους (δεξιά) [26].

Κάθε Νευρωνικό Δίκτυο απαρτίζεται από τρία μέρη: το επίπεδο εισόδου στο οποίο τροφοδοτούνται τα δεδομένα εισόδου x_i, ένα ή περισσότερα κρυφά επίπεδα h_i, και ένα επίπεδο εξόδου y_i. Τα δεδομένα εισόδου συνδυάζονται με διαφορετικά βάρη τα οποία στην ουσία υποδηλώνουν την επίδραση του κάθε νευρώνα. Κατά τη διάδοση των τιμών μέσα στο Νευρωνικό Δίκτυο σε κάθε επίπεδο ανάλογα με την είσοδο ενεργοποιείται κάθε φορά ένας νευρώνας από κάθε επίπεδο μέσω μιας Συνάρτησης Ενεργοποίησης (Activation Function) *f*. Μερικές από τις πιο συνήθεις συναρτήσεις ενεργοποίησης θα παρουσιαστούν στην επόμενη ενότητα.

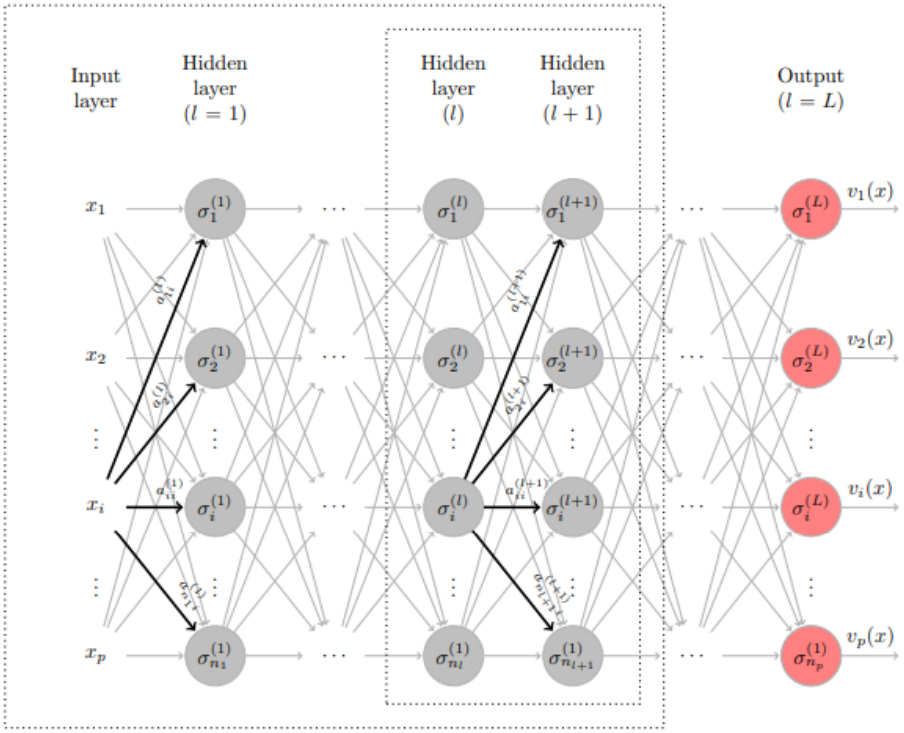
Η θεμελιώδη μονάδα πάνω στην οποία βασίζονται όλες οι προχωρημένες και περίπλοκες αρχιτεκτονικές Νευρωνικών Δικτύων είναι το perceptron. Το perceptron εμπνεύστηκε από τον Rosenblatt [27] και αποτελεί την πιο απλή μορφή νευρωνικού δικτύου με μόνο ένα κρυφό επίπεδο. Οι δυνατότητές του ήταν περιορισμένες εφόσον λειτουργούσε αποδοτικά μόνο σε γραμμικώς διαχωρίσιμα σύνολα δεδομένων παρόλα αυτά η καινοτόμα μαθηματική του θεμελίωση έδωσε το έναυσμα για την μετέπειτα ραγδαία εξέλιξη του τομέα.



Σχήμα 2.2.2: Η αρχιτεκτονική του perceptron [27].

Τα δεδομένα εισόδου λοιπόν x_i πολλαπλασιάζονται με τα βάρη τους w_i. Στην συνέχεια αθροίζονται υπολογίζοντας έτσι το σταθμισμένο άθροισμα (weighted sum) το οποίο ακολούθως διέρχεται μέσα από μία συνάρτηση ενεργοποίησης λαμβάνοντας τελικά την έξοδο.

Προκειμένου να μοντελοποιηθούν περίπλοκες δομές δεδομένων και να δημιουργηθούν μοντέλα ικανά να μάθουν μη γραμμικές συναρτήσεις και να αποδώσουν καρπούς σε προβλήματα μηχανικής μάθησης, σχεδιάστηκαν αρχιτεκτονικές που συνδυάζουν τεχνητούς νευρώνες. Η απλούστερη περίπτωση τέτοιων αρχιτεκτονικών είναι το πολυεπίπεδο perceptron (Multi-layer Perceptron - MLP), το οποίο στην ουσία αποτελείται από πολλά επίπεδα perceptron τοποθετημένα το ένα κάτω από το άλλο. Κάθε κόμβος τόσο του κρυφού επιπέδου όσο και του επιπέδου εξόδου χρησιμοποιεί μία μη γραμμική συνάρτηση ενεργοποίησης ή οποία αποτελεί το κλειδί στην αποδοτική επεξεργασία μη γραμμικών διαχωρίσιμων συνόλων δεδομένων. ΄Ενα MLP το οποίο αποτελείται από περισσότερα από ένα κρυφά επίπεδα, συνιστά ένα Βαθύ Νευρωνικό Δίκτυο (Deep Neural Network- DNN). Η προσθήκη πολλαπλών κρυφών επιπέδων μαζί με τις μη γραμμικές συναρτήσεις ενεργοποίησης που συμπεριλαμβάνουν μας επιτρέπουν να εξάγουμε από τα δεδομένα αναπαραστάσεις χαρακτηριστικών υψηλότερου επιπέδου και έτσι να αναλύσουμε σε μεγαλύτερο βάθος τις σχέσεις και τα μοτίβα που αναπτύσσονται μεταξύ των δεδομένων.



Σχήμα 2.2.3: Αρχιτεκτονική Πολυ-Επίπεδου perceptron(MLP) [28].

2.2.2 Συναρτήσεις Ενεργοποίησης(Activation Function)

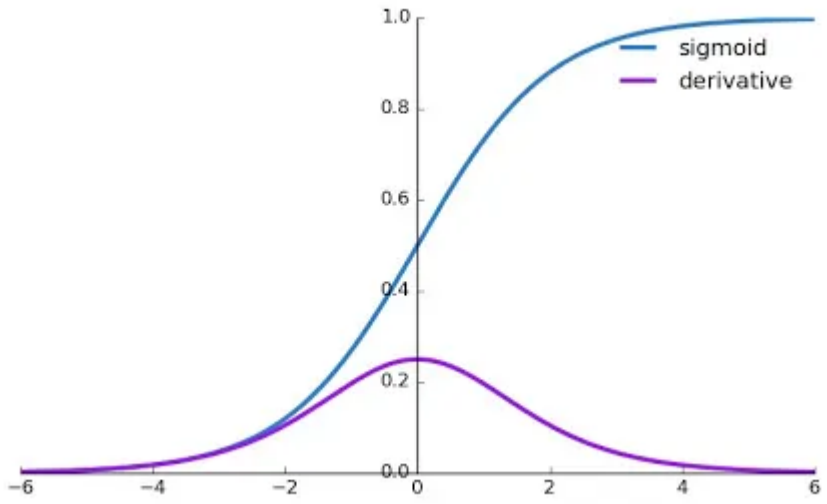
Στην συγκεκριμένη ενότητα αναλύονται οι πιο δημοφιλείς μη γραμμικές συναρτήσεις ενεργοποίησης και παρουσιάζονται κάποια από τα πλεονεκτήματα και τα μειονεκτήματά τους. Γενικά δεν υπάρχει κάποια συγκεκριμένη συνάρτηση ενεργοποίησης οι οποία να εγγυάται βέλτιστα αποτελέσματα και η αποδοτικότητά της εξαρτάται τόσο από το σύνολο δεδομένων όσο και από την διεργασία που θέλουμε να εκτελέσουμε. Ταυτόχρονα, η συμβολή τους στην μοντελοποίηση μη γραμμικών φαινομένων και συστημάτων είναι καθοριστική. Χωρίς την παρουσία τους το σήμα εξόδου θα εκφυλιζόταν σε μία απλή γραμμική συνάρτηση.

(I) Σιγμοειδής Συνάρτηση (Sigmoid Function)

Η σιγμοειδής συνάρτηση είναι μία μαθηματική συνάρτηση που δίνεται από τον ακόλουθο τύπο:

$$S(x) = \frac{1}{1+e^{-x}} \qquad (2.2.1)$$

Η συνάρτηση αυτή δέχεται σαν είσοδο οποιαδήποτε πραγματική τιμή και την συμπιέζει στο διάστημα [0,1.0]. Όσο μεγαλύτερη είναι η τιμή εισόδου τόσο πιο κοντα στην μοναδα θα είναι η τιμή εξόδου ενώ αντίθετα όσο πιο μικρή είναι η τιμή εισόδου τοσο πιο κοντα στο μηδεν θα ειναι η τιμη εξόδου. Χρησιμοποιείται συνήθως για μοντέλα όπου πρέπει να προβλέψουμε την πιθανότητα ως έξοδο. Δεδομένου ότι η πιθανότητα ανήκει στο εύρος 0 και 1 η σιγμοειδής συνάρτηση είναι η σωστή επιλογή λόγω του εύρους της. Ωστόσο, σε κάθε ουρά στο 0 ή στο 1, οι τιμές της παραγώγου της είναι πολύ μικρές, συγκλίνοντας στο 0. Ως εκ τούτου, τα διανύσματα κλίσης "εξαφανίζονται" (φαινόμενο vanishing gradient), περιορίζοντας τις δυνατότητες μάθησης του μοντέλου.



Σχήμα 2.2.4: Η γραφική της σιγμοειδής συνάρτησης και της παραγώγου της [29].

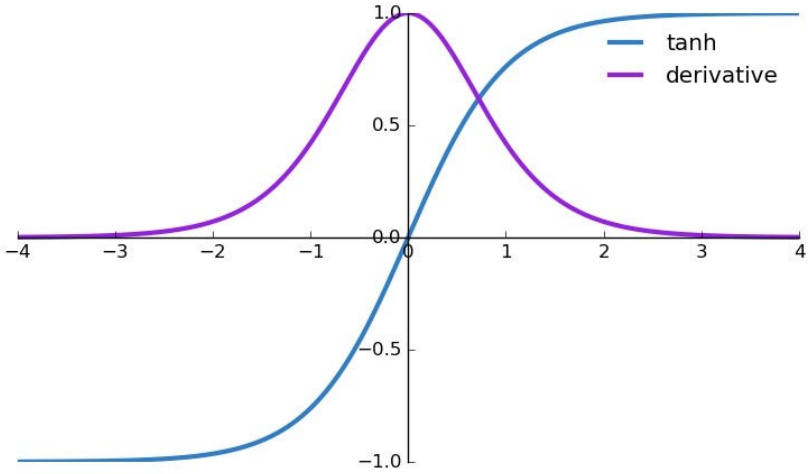
(II) Συνάρτηση Υπερβολικής Εφαπτομένης (Hyperbolic Tangent)

Η συνάρτηση υπερβολικής εφαπτομένης δίνεται από τον ακόλουθο μαθηματικό τύπο:

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad (2.2.2)$$

Είναι παρόμοια με την σιγμοειδή συνάρτηση με την μόνη διαφορά ότι η έξοδος συμπιέζεται στο διάστημα [-1.0 , 1.0]. Επομένως, όσο μεγαλύτερη είναι η τιμή εισόδου τόσο πιο κοντα στην μοναδα θα είναι η τιμή εξόδου ενώ αντίθετα όσο πιο μικρή είναι η τιμή εισόδου τοσο πιο κοντα

στην αρνητική μονάδα θα ειναι η τιμη εξόδου. Βασικό πλεονέκτημα της tanh είναι ότι η έξοδός της είναι κεντραρισμένη στο μηδέν. Συνήθως χρησιμοποιείται ως συνάρτηση ενεργοποίησης αποκλειστικά στα κρυφά επίπεδα των νευρωνικών δικτύων καθώς βοηθάει στο κεντράρισμα των δεδομένων (centering data) και ως επακόλουθο διευκολύνει την διαδικασία της εκμάθησης των επόμενων επιπέδων. Παρόλα αυτά, ομοίως με την σιγμοειδή συνάρτηση, επιβαρύνεται από τα προβλήματα του μηδενισμού της παραγώγου στα άκρα της (vanishing gradient) και επιπλέον η παράγωγός της φθίνει πιο απότομα.



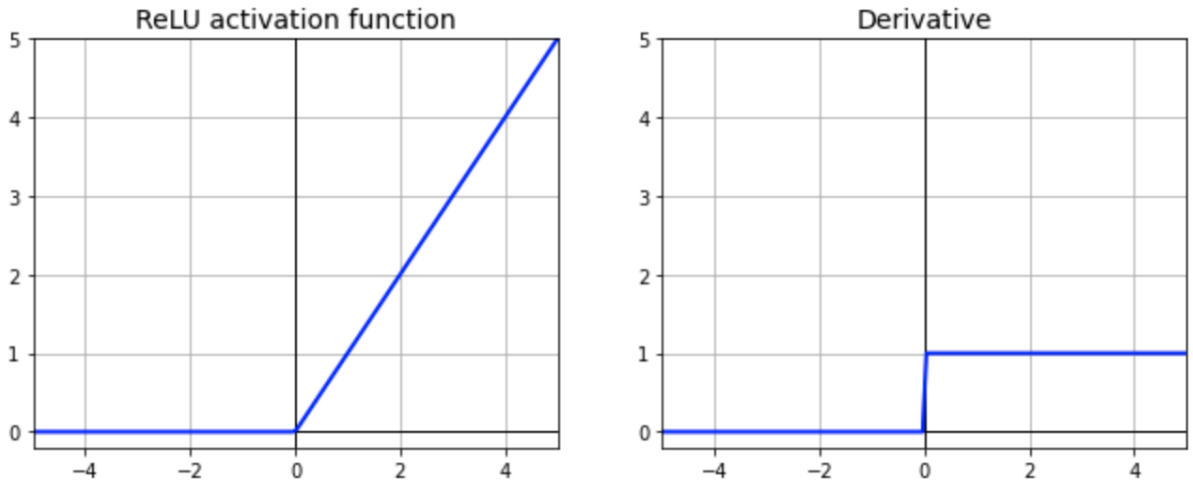
Σχήμα 2.2.5: Γραφική της υπερβολικής εφαπτομένης (tanh) και της παραγώγου της [29].

(III) Συνάρτηση ReLU (Rectified Linear Unit)

Η συνάρτηση ReLU δίνεται από την ακόλουθη σχέση:

$$f(x) = max(0, x) \qquad (2.2.3)$$

Ουσιαστικά αποτελεί ένα κατώφλι της εισόδου στο μηδεν. Παρόλο που μοιάζει με μία γραμμική συνάρτηση η παράγωγός της επιτρέπει την οπίσθια διάδοση (back propagation) και παράλληλα είναι υπολογιστικά αποδοτική. Η κυρίαρχη ιδέα της ReLU είναι ότι δεν ενεργοποιεί ταυτόχρονα όλους τους νευρώνες εφόσον ενεργοποιούνται μόνο αυτοί που έχουν θετικό σταθμισμένο άθροισμα. Έτσι προσφέρει γρήγορη σύγκλιση και παράλληλα καταναλώνει ελάχιστους υπολογιστικούς πόρους. Το βασικό της μειονέκτημα όμως είναι ότι για όλες τις αρνητικές εισόδους η παράγωγος μηδενίζεται γεγονός που δυσχεραίνει την δυνατότητα του συστήματος στην μοντελοποίηση των δεδομένων κατά την διαδικασία της εκμάθησης.



Σχήμα 2.2.6: Γραφική της συνάρτησης ReLU (αριστερά) και της παραγώγου της (δεξιά) [29].

2.2.2 Αλγόριθμος Οπίσθιας Διάδοσης (Backpropagation)

Ο αλγόριθμος Backpropagation χρησιμοποιείται ευρέως στην διαδικασία εκπαίδευσης των νευρωνικών δικτύων (training process). Η εκπαίδευση είναι μια επαναληπτική διαδικασία με απώτερο σκοπό την εύρεση των βέλτιστων παραμέτρων (weights) οι οποίες ελαχιστοποιούν κάποια συνάρτηση κόστους L. Κάθε επανάληψη περιλαμβάνει δύο φάσεις: Αρχικά εκτελείται η πρόσθια διάδοση (forward pass) η οποία επιτρέπει την ροή της πληροφορίας από την είσοδο στην έξοδο. Στην συνέχεια εκτελείται η οπίσθια διάδοση στην οποία η ροή της πληροφορίας έχει αντίθετη κατεύθυνση δηλαδή από την έξοδο προς την είσοδο ενώ παράλληλα ανανεώνονται οι τιμές των βαρών. Για να είναι εφικτή όμως αυτή η ανανέωση των τιμών απαιτείται ο υπολογισμός παραγώγων από περίπλοκες και σύνθετες μαθηματικες εκφράσεις. Παρόλα αυτά ο αλγόριθμος Backpropagation [30] καταφέρνει να τις υπολογίσει μεθοδικά και αποδοτικά βασιζόμενος στον κανόνα της αλυσίδας. Ο κανόνας της αλυσίδας αναδεικνύει τον τρόπο με τον οποίο υπολογίζονται αναδρομικά οι παράγωγοι από μία σύνθετη συνάρτηση. Για την ανανέωση των βαρών του δικτύου, μετά από κάθε υπολογισμό της συνάρτησης κόστους L, υπολογίζονται οι μερικοί παράμετροι της συνάρτησης κόστους ως προς τα βάρη $\frac{\partial L}{\partial w}$ [31]. Ο τρόπος με τον οποίο ανανεώνονται οι τιμές των βαρών προκύπτει από τους αλγορίθμους βελτιστοποίησης οι οποίοι θα αναλυθούν στην επόμενη ενότητα. Με αυτόν τον τρόπο το μοντέλο εκπαιδεύεται βελτιώνοντας συνεχώς την επίδοσή του.

2.2.3 Βελτιστοποίηση (Optimization)

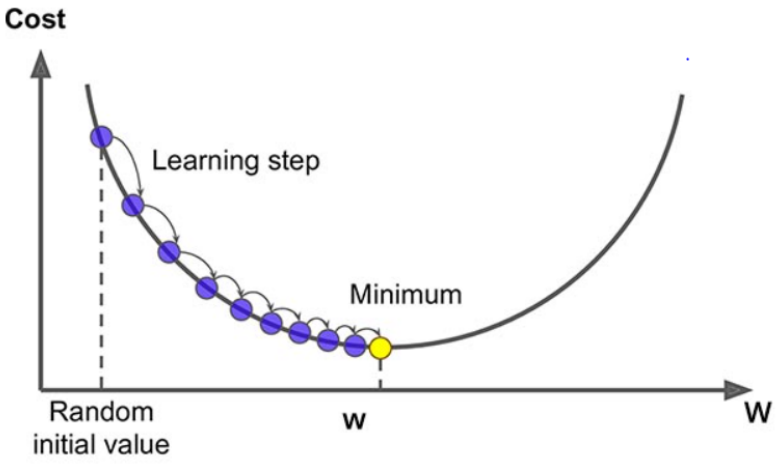
Βελτιστοποίηση είναι η διαδικασία κατά την οποία εκπαιδεύουμε ένα μοντέλο επαναληπτικά ούτως ωστε να οδηγηθεί στην μεγιστοποίηση κάποιας συνάρτησης αξιολόγησης. Όπως αναφέραμε και προηγουμένως στα νευρωνικά δίκτυα, κατά την εκπαίδευση ενός μοντέλου, οι παράμετροι ανανεώνονται επαναληπτικά έως ότου να ελαχιστοποιηθεί κάποια συνάρτηση κόστους. Οι αλγόριθμοι βελτιστοποίησης καθορίζουν τον τρόπο με τον οποίο προσαρμόζονται τα βαρη του δικτύου. Η πιο δημοφιλείς κατηγορίες τέτοιων αλγορίθμων στηρίζονται σε υπολογισμούς κλίσεων (gradients) της συνάρτησης κόστους. Διαισθητικά ανανεώνουμε τα βάρη προς την αντίθετη κατεύθυνση από αυτή του διανύσματος κλίσης προσεγγίζοντας κάποιο ελάχιστο της συνάρτησης κόστους. Οι κυριότεροι αλγόριθμοι βελτιστοποίησης αναλύονται ακολούθως.

(I) Κάθοδος Κλίσης (Gradient Descent)

Ο αλγόριθμος Κάθοδος Κλίσης υπολογίζει την κλίση της συνάρτησης κόστους για ολόκληρο το σύνολο δεδομένων σε σχέση με τις παραμέτρους θ του μοντέλου σε κάθε επανάληψη [32]. Η μαθηματική σχέση ανανέωσης των βαρών είναι η ακόλουθη:

$$\theta^{'} = \theta - \eta \cdot \nabla_{\theta} J(\theta) \qquad (2.2.4)$$

Η μεταβλητή *η* συμβολίζει τον ρυθμό μάθησης ,δηλαδή το πόσο μεγάλα βήματα κατάβασης θα εκτελούνται σε κάθε ανανέωση βαρών. Η συνάρτηση $J(\theta)$ συμβολίζει κάποια συνάρτηση κόστους. Ο αλγόριθμος Gradient Descent δεν εγγυάται σύγκλιση σε ολικό ελάχιστο αλλά μπορεί να εγκλωβιστεί σε κάποιο τοπικό ελάχιστο. Επιπλέον παρά την εύκολη υλοποίησή του δεν είναι υπολογιστικά αποδοτικός εφόσον για μεγάλα σύνολα δεδομένων ο υπολογισμός του σφάλματος για κάθε δείγμα είναι χρονοβόρος. Ιδιαίτερα ο αλγόριθμος εμφανίζει χρονική πολυπλοκότητα $O(kn^2)$ όπου k το πλήθος των επαναλήψεων και n το πλήθος των δειγμάτων στο σύνολο δεδομένων.



Σχήμα 2.2.7 : Αλγόριθμος βελτιστοποίησης Gradient Descent [32].

(II) Στοχαστική Κάθοδος Κλίσης (Stochastic Gradient Descent-SGD)

Ο αλγόριθμος Stochastic Gradient Descent έρχεται να δώσει λύση στην υψηλή χρονική πολυπλοκότητα που εμφανίζει ο Gradient Descent. Εμφανίζουν παρόμοια λειτουργία με την διαφορά ότι ο SGD χρησιμοποιεί ένα υποσύνολο των δειγμάτων για να ελαχιστοποιήσει την συνάρτηση κόστους. Η στοχαστικότητά του στην πραγματικότητα βασίζεται στον πιθανοθεωρητικό τρόπο με τον οποίο επιλέγει τυχαία τα δείγματα που θα χρησιμοποιήσει. Μαθηματικά η συνάρτηση ανανέωσης των παραμέτρων είναι η ακόλουθη:

$$\theta' = \theta - \eta \cdot \nabla_\theta J(\theta;\, x;\, y) \qquad (2.2.5)$$

όπου x είναι τα επιλεγμένα δείγματα και y οι αντίστοιχες ετικέτες τους. Βέβαια η μέθοδος αυτή πάλι δεν εγγυάται συγκλιση σε κάποιο ολικό ελάχιστο. Επιταχύνει όμως την σύγκλιση γεγονός που ευνοεί την χρήση μεγάλων συνόλων δεδομένων.

Algorithm 8.1 Stochastic gradient descent (SGD) update at training iteration k

Require: Learning rate ϵ_k.
Require: Initial parameter $\boldsymbol{\theta}$
while stopping criterion not met **do**
Sample a minibatch of m examples from the training set $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ with corresponding targets $\boldsymbol{y}^{(i)}$.
Compute gradient estimate: $\hat{\boldsymbol{g}} \leftarrow +\frac{1}{m} \nabla_{\boldsymbol{\theta}} \sum_i L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), \boldsymbol{y}^{(i)})$
Apply update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \hat{\boldsymbol{g}}$
end while

Σχήμα 2.2.8: Ο αλγόριθμος Stochastic Gradient Descent [32].

(III) Adam (Adaptive moment estimation)

Ο αλγόριθμος βελτιστοποίησης Adam αποτελεί επέκταση του Stochastic Gradient Descent και τελευταία έχει υιοθετηθεί σε εφαρμογές βαθιάς μάθησης κυρίως στους τομείς της Φυσικής Επεξεργασίας Γλώσσας και στην Οραση Υπολογιστών. Η κύρια διαφορά του είναι ότι δεν χρησιμοποιεί ένα σταθερό ρυθμό εκμάθησης, κοινό για την ανανέωση όλων των παραμέτρων. Αντίθετα, ένας ρυθμός εκμάθησης διατηρείται για κάθε βάρος του δικτύου και προσαρμόζεται ξεχωριστά καθώς εξελίσσεται η μάθηση. Οι βελτιστοποιητές Adam ,Adagrad, Adadelta και BertAdam είναι ευρέως χρησιμοποιούμενοι λόγω της αποδοτικότητάς τους αλλά και λόγω των χαμηλών απαιτήσεων σε μνήμη.

2.2.4 Συναρτήσεις Κόστους (Loss Functions)

Οι συναρτήσεις κόστους (Loss Functions) είναι ένα από τα πιο σημαντικά κομμάτια των νευρωνικών δικτύων, εφόσον είναι υπεύθυνες για την εκπαίδευση των μοντέλων και την σωστή αφομοίωση του συνόλου δεδομένων. Η γενική αρχή των συναρτησεων αυτών είναι να ποσοτικοποιήσουν την επίδοση του δικτύου στην μοντελοποίηση των δεδομένων εκπαίδευσης, συγκρίνοντας την τιμή στόχο y με την προβλεπόμενη τιμή $\widehat{y}$. Η επιλογή της συνάρτησης κόστους εξαρτάται από τον τύπο λειτουργίας του νευρωνικού δικτύου και χωρίζονται σε δύο μεγάλες κατηγορίες: Οι συναρτήσεις κόστους που είναι κατάλληλες για προβλήματα παλινδρόμησης (regression problems), όπου η έξοδος είναι μία συνεχής τιμή και οι συναρτήσεις κόστους που είναι κατάλληλες για προβλήματα ταξινόμησης (classification problems), όπου η έξοδος είναι απλά μία ετικέτα. Οι πιο δημοφιλείς συναρτήσεις κόστους αναλύονται στην συνέχεια.

(I) Μέσο Τετραγωνικό Σφάλμα (Mean Square Error - MSE)

Η συνάρτηση Μέσου Τετραγωνικού Σφάλματος είναι ίσως η πιο διάσημη συνάρτηση κόστους η οποία υπολογίζει την μέση τετραγωνική διαφορά μεταξύ της τιμής στόχου y και της πρόβλεψης $\widehat{y}$. Η μαθηματική της έκφραση είναι η ακόλουθη :

$$MSE = \frac{1}{n}\sum_{i=0}^{n} (y^{(i)} - \widehat{y}^{(i)})^2 \qquad (2.2.6)$$

όπου n συμβολίζει το πλήθος των δειγμάτων του συνόλου εκπαίδευσης. Η διαφορά είναι υψωμένη στο τετράγωνο, πράγμα που σημαίνει ότι δεν έχει σημασία αν η προβλεπόμενη τιμή είναι πάνω ή κάτω από την τιμή στόχο και το σφάλμα θα είναι πάντα μία θετική τιμή. Επιπλέον, οι τιμές με μεγάλο σφάλμα τιμωρούνται περισσότερο λόγω του τετραγωνισμού της διαφοράς τους. Ένα σημαντικό πλεονέκτημα είναι ότι η συνάρτηση αυτή είναι κυρτή με ένα ξεκάθαρα διακεκριμένο ολικό ελάχιστο. Άρα μπορούμε εύκολα να αξιοποιήσουμε τον αλγόριθμο Stochastic Gradient Descent για να ανανέωση των παραμέτρων του δικτύου. Παρόλα αυτά ένα σημαντικό μειονέκτημά της είναι ότι το σφάλμα αυξάνεται ραγδαία για τις ακραίες τιμές του συνόλου δεδομένων. Επομένως, συχνά αδυνατεί να συγκλίνει σε κάποιο ελάχιστο και επιπλέον τα βάρη

του δικτύου αποκτάνε πολύ υψηλές τιμές, γεγονός που δυσχεραίνει την δυνατότητα του δικτύου στην μοντελοποίηση του συνόλου δεδομένων καταλήγοντας έτσι σε υπερεκπαίδευση (overfitting). Υπο αυτές τις συνθήκες είναι αναγκαία η προσθήκη κάποιου όρου κανονικοποίησης ούτως ώστε οι τιμές των βαρών να παραμένουν χαμηλές.

(II) Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE)

Η συνάρτηση αυτή υπολογίζει την μέση τιμή της απόλυτης διαφοράς μεταξύ της τιμής στόχου y και της προβλεπόμενη τιμής $\widehat{y}$. Η μαθηματική της έκφραση είναι η ακόλουθη:

$$MAE = \frac{1}{n}\sum_{i=0}^{n} |y^{(i)} - \widehat{y}^{(i)}| \qquad (2.2.7)$$

Η συνάρτηση αυτή έχει παρόμοια λειτουργία με την προηγούμενη με την μόνη διαφορά ότι δεν είναι τόσο ευαίσθητη στις ακραίες τιμές. Λόγω της απουσίας του τετραγωνισμού τα μεγάλα σφάλματα δεν τιμωρούνται και έτσι χαρακτηρίζεται κατάλληλη για σύνολα δεδομένων τα οποία περιλαμβάνουν πολλές ακραίες τιμές (outliers). Παρόλα αυτά, ένα σημαντικό μειονέκτημά της είναι ότι η παράγωγός της δεν ορίζεται κοντά στο 0.

(III)Σφάλμα Huber (Huber Loss)

Η συνάρτηση κόστους Huber συνδυάζει τα θετικά των συναρτήσεων *MSE* και *MAE*. Ιδιαίτερα είναι μία δίκλαδη συνάρτηση όπου ορίζουμε ένα κατώφλι σφάλματος δ κάτω από το οποίο εφαρμόζεται η συνάρτηση *MSE* ενώ για τιμές σφάλματος μεγαλύτερες του δ εφαρμόζεται η συνάρτηση *MAE*. Η μαθηματική της έκφραση είναι η ακόλουθη:

$$Huber\ loss = \begin{cases} \frac{1}{n}\sum_{i=1}^{N}\left(y^{(i)} - \widehat{y}^{(i)}\right)^2 & for\ \left|y^{(i)} - \widehat{y}^{(i)}\right| \le \delta \\ \frac{1}{n}\sum_{i=1}^{N}\left|y^{(i)} - \widehat{y}^{(i)}\right| & otheriwise \end{cases} \qquad (2.2.8)$$

(III)Απόκλιση Kullback Leibler (KL Divergence)

Η απόκλιση Kullback Leibler είναι μία στατιστική μέτρηση εμπνευσμένη από την θεωρία της πληροφορίας η οποία ποσοτικοποιεί την απόκλιση μεταξύ δύο στατιστικών κατανομών . Ιδιαίτερα μετράει την σχετική εντροπία μεταξύ των κατανομών ακολουθώντας την παρακάτω μαθηματική σχέση:

$$D_{KL}(p(x)|\,|q(x)) = \sum_{x \in X} p(x) ln \frac{p(x)}{q(x)} \quad (2.2.9)$$

Κύρια εφαρμογή βρίσκει στην παρακολούθηση μοντέλου (model monitoring). Η παρακολούθηση μοντέλου είναι ένα λειτουργικό στάδιο στον κύκλο ζωής της μηχανικής μάθησης που έρχεται μετά την ανάπτυξη του μοντέλου. Συνεπάγεται την παρακολούθηση των μοντέλων ML για αλλαγές όπως η υποβάθμιση του μοντέλου, η μετατόπιση δεδομένων, η μετατόπιση εννοιών και η διασφάλιση ότι το μοντέλο μας διατηρεί ένα αποδεκτό επίπεδο απόδοσης.

(III)Σφάλμα Διασταυρούμενης Εντροπίας (Cross Entropy Loss)

Η συγκεκριμένη συνάρτηση κόστους εφαρμόζεται μόνο σε προβλήματα ταξινόμησης και αποτελεί εξέλιξη της στατιστικής μέτρησης KL Divergence. Με τον όρο εντροπία αναφερόμαστε στο πλήθος των bits που απαιτούνται για την μεταφορά ενός τυχαίου γεγονότος από μία στατιστική κατανομή. Το Cross Entropy Loss λοιπόν βασίζεται στην ιδέα της εντροπίας από τη θεωρία πληροφοριών και υπολογίζει τον αριθμό των bit που απαιτούνται για την αναπαράσταση ή τη μετάδοση ενός γεγονότος από μια κατανομή σε σύγκριση με μια άλλη. Η μαθηματική σχέση από την οποία διέπεται είναι η ακόλουθη:

$$CE\ Loss = -\frac{1}{n}\sum_{j=1}^{M}\sum_{i=1}^{N} y_{ij}\, log(p_{ij}) \quad (2.2.10)$$

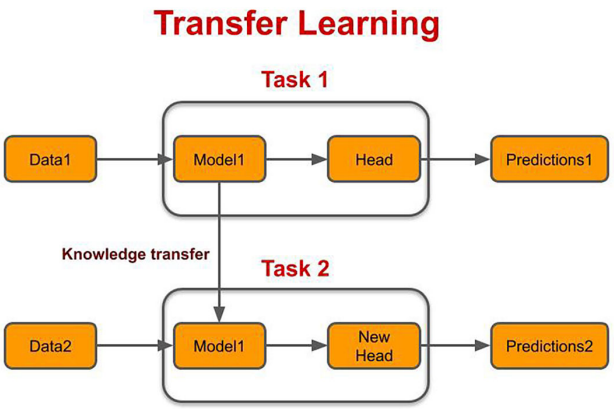
όπου το y_{ij} αναφέρεται στην κλάση που ανήκει το δείγμα i ενώ το p_{ij} την πιθανότητα του δείγματος i να ανήκει στην κλάση j. Απαραίτητα προϋπόθεση είναι οι ετικέτες μας να κωδικοποιηθούν δυαδικά σε 0 και 1 (one hot encoding). Το Cross Entropy Loss έχει κυριαρχήσει στις μέρες σε προβλήματα ταξινόμησης λόγω της αποδοτικότητας του.

2.2.5 Μεταφορά Μάθησης (Transfer Learning)

Όπως αναφέραμε και προηγουμένως , οι αλγόριθμοι μηχανικής μάθησης είναι πολύ ευαίσθητοι στα δεδομένα εισόδου. Ιδιαίτερα απαιτούνται τεράστια σύνολα δεδομένων ούτως ώστε να εκπαιδεύσουμε ένα σύστημα το οποίο θα παρουσιάζει υψηλή απόδοση. Παρόλα αυτά συχνά δεν έχουμε την δυνατότητα πρόσβασης σε τέτοια σύνολα δεδομένων ούτε τους υπολογιστικούς πόρους που απαιτούνται για την επεξεργασία τους. Στις περιπτώσεις αυτές επιστρατεύεται η τεχνική της Μεταφοράς Μάθησης (Transfer Learning).

Η μεταφορά μάθησης λοιπόν αποτελεί σημαντικό εργαλείο της μηχανικής μάθησης που συνεισφέρει στην επίλυση του θεμελιώδους προβλήματος της έλλειψης ικανοποιητικών συνόλων εκπαίδευσης. Προσπαθεί να μεταφέρει την γνώση από ένα προεκπαιδευμένο μοντέλο και να την

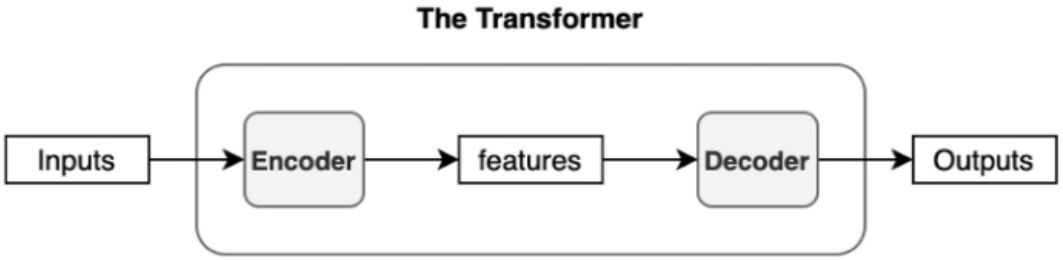
αξιοποιήσει στο πρόβλημα στόχο, χαλαρώνοντας έτσι την βασική υπόθεση της μηχανικής μάθησης κατά την οποία τα δείγματα του συνόλου εκπαίδευσης και του συνόλου εξολόγησης αντλούνται ανεξάρτητα από την ίδια τυχαία κατανομή. Με άλλα λόγια χρησιμοποιούμε την γνώση που συγκέντρωσε ένα μοντέλο κατά την εκπαίδευσή του σε ένα πρόβλημα -πηγή και την ενσωματώνουμε (fine tuning) στο πρόβλημα-στόχο. Έτσι εκπαιδεύουμε μοντέλα με πρότερη γνώση τα οποία παρουσιάζουν υψηλή γενίκευση και σημαντικά βελτιωμένη απόδοση.



Σχήμα 2.2.9 : Ροή εργασιών στην Μεταφορά Μάθησης [33].

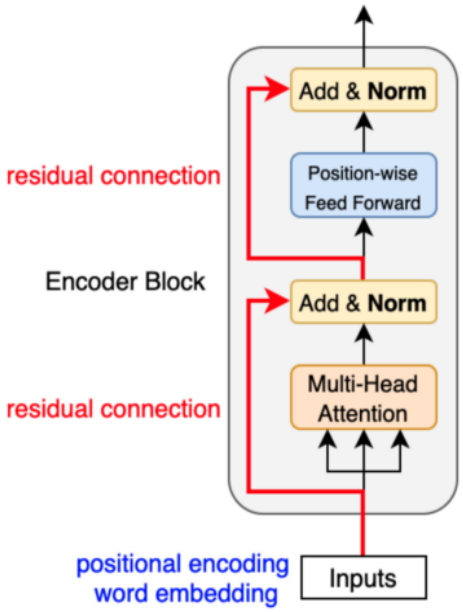
2.2.6 Η αρχιτεκτονική των Transformers

Η αρχιτεκτονική του δικτύου Transformer προτάθηκε από τους Vaswani et al. [34] επιτυγχάνοντας ανώτερες επιδόσεις σε σχέση με πρότερες αρχιτεκτονικές αλλά και μικρότερες χρονικές απαιτήσεις όσον αφορά την εκπαίδευση του μοντέλου. Στηρίζεται στην αρχιτεκτονική Encoder-Decoder δηλαδή σε μία δομή κωδικοποιητή και αποκωδικοποιητή. Αναλυτικότερα ο κωδικοποιητής εξάγει χαρακτηριστικά, δηλαδή συνεχείς αναπαραστάσεις, από μια ακολουθία εισόδου και ο αποκωδικοποιητής χρησιμοποιεί τα χαρακτηριστικά αυτά για να παράγει μία ακολουθία εξόδου.



Σχήμα 2.2.10 : Η αρχιτεκτονική του μοντέλου Transformer.

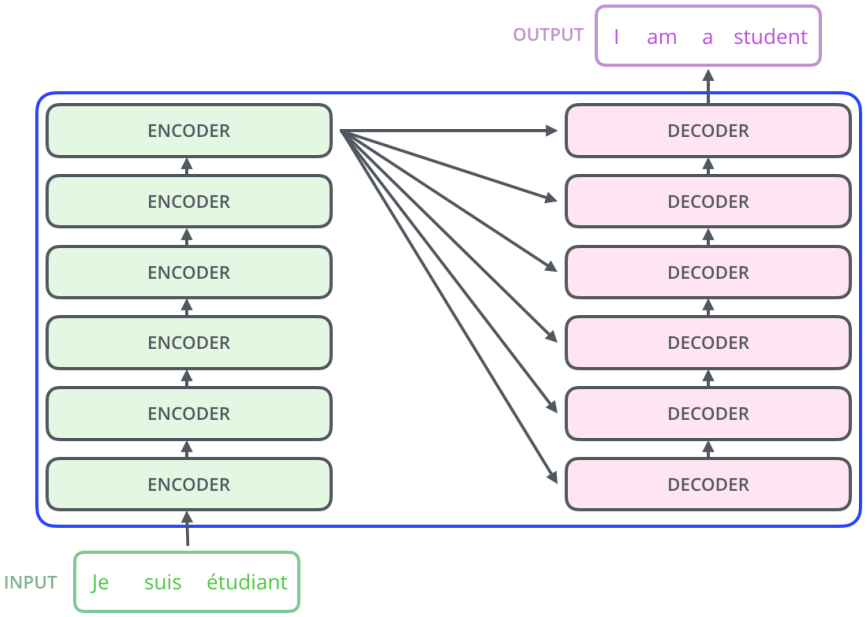
Η στοίβα Κωδικοποιητή (Encoder Stack) αποτελείται από πολλαπλά επίπεδα κωδικοποίησης τα οποία είναι πανομοιότυπα μεταξύ τους. Κάθε επίπεδο περιλαμβάνει ένα μηχανισμό multi-head αυτοπροσοχής και έναν μηχανισμό ο οποίος κωδικοποιεί την θέση κάθε στοιχείου εισόδου (positional encoding mechanism). Ανάλογα με τα γύρω στοιχεία/λέξεις, κάθε στοιχείο/λέξη μπορεί να έχει περισσότερες από μία σημασιολογικές λειτουργίες. Ως εκ τούτου, ο μηχανισμός αυτοπροσοχής χρησιμοποιεί πολλαπλές κεφαλές (οκτώ παράλληλους υπολογισμούς προσοχής) έτσι ώστε το μοντέλο να μπορεί να αξιοποιήσει διαφορετικούς υποχώρους των εμφυτευμάτων. Όσον αφορά το δεύτερο υπόστρωμα είναι ένα απλό position-wise fully connected feed forward δίκτυο. Η τεχνική της υπολειπόμενης σύνδεσης (Residual Connection) [35] επιστρατεύεται επίσης γύρω από κάθε υποστρωμα μεταφέροντας στην ουσία τα προηγούμενα λεκτικά εμφυτεύματα στο ακριβώς επόμενο επίπεδο. Τέλος μετά από κάθε υπολειπόμενη συνδεση εφαρμόζεται ένα επίπεδο κανονικοποίησης που σκοπεύει στην αντιμετώπιση της αρνητικής επίδρασης που επιφέρει κάποια αλλαγή στην κατανομή των δεδομένων εισόδου (Covariate shift). Επομένως, η έξοδος κάθε υποστρώματος είναι LayerNorm(x + Sublayer(x)), όπου Sublayer(x) είναι η συνάρτηση που εφαρμόζει το ίδιο το υπόστρωμα.



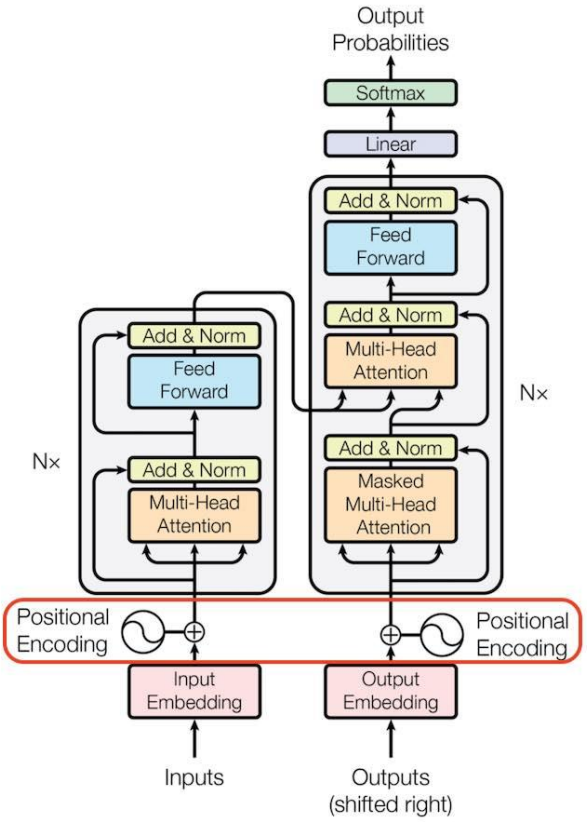
Σχήμα 2.2.11: Οπτικοποίηση της εσωτερικής αρχιτεκτονικής της στοίβας του κωδικοποιητή [36].

Η στοίβα Αποκωδικοποιητή (Decoder Stack) αποτελείται ομοίως από πολλαπλά επίπεδα αποκωδικοποίησης τα οποία είναι πανομοιότυπα μεταξύ τους. Κάθε επίπεδο εμφανίζει όμοια αρχιτεκτονική με τα αντίστοιχα επίπεδα του Encoder με την διαφορά ότι προστίθεται και ένα τρίτο υπόστρωμα προσοχής πηγής-στόχου (source-target attention). Η source-target προσοχή είναι

στην ουσία άλλο ένα επίπεδο multihead προσοχής το οποίο εφαρμόζεται μεταξύ των χαρακτηριστικών που εξάγει ο Encoder και της εξόδου κάθε επιπέδου του Decoder. Τέλος, εφαρμόζεται ένα είδος μάσκας στον μηχανισμό αυτοπροσοχής, ώστε η πρόβλεψη της θέσης να μην εξαρτάται από τις κρυφές καταστάσεις Hidden States).



Σχήμα 2.2.12: Οπτικοποίηση Της λειτουργίας του Transformer . Ο Encoder και ο Decoder περιλαμβάνουν έξι επίπεδα ο καθένας [36].



Σχήμα 2.2.13: Η αναλυτική αρχιτεκτονική του μοντέλου Transformer [37].

2.3 Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing) είναι ένας τομέας έρευνας και εφαρμογής που διερευνά τον τρόπο με τον οποίο οι υπολογιστές μπορούν να χρησιμοποιηθούν για την κατανόηση και τον χειρισμό κειμένων ή της ομιλίας φυσικής γλώσσας. Συμπεριλαμβάνει διάφορες υπολογιστικές τεχνικές για την ανάλυση και την αναπαράσταση φυσικών κειμένων σε ένα ή περισσότερα επίπεδα γλωσσικής ανάλυσης με απώτερο σκοπό την επεξεργασία της ανθρώπινης γλώσσας. Το πεδίο του NLP, λοιπόν, εστιάζει σε δύο ξεχωριστές προσεγγίσεις: την επεξεργασία και την παραγωγή της φυσικής γλώσσας. Η πρώτη περίπτωση αναφέρεται στην ανάλυση της φυσικής γλώσσας με σκοπό την παραγωγή μιας αντιπροσωπευτικής αναπαράστασης, ενώ το τελευταίο αναφέρεται στην παραγωγή της γλώσσας από μια αναπαράσταση.

Η λειτουργική δομή των συστημάτων NLP μπορεί να χαρακτηριστεί από κάποια υποπροβλήματα τα οποία αντιπροσωπεύουν επίπεδα ανάλυσης και καλούνται ‘επίπεδα γλώσσας’. Η κεντρική ιδέα

βασίζεται στο γεγονός ότι κάθε επίπεδο συνεισφέρει ακολουθιακά στην διαδικασία της κατανόησης και εφόσον οι άνθρωποι χρησιμοποιούν όλα τα γλωσσικά επίπεδα, ένα σύστημα NLP είναι πιο αποδοτικό όσο περισσότερα γλωσσικά επίπεδα αξιοποιεί. Παρακάτω αναλύονται τα γλωσσικά επίπεδα του κλάδου:

- ***Phonology***: το επίπεδο αυτό ασχολείται με την ερμηνεία του φυσικού λόγου μεταξύ των λέξεων . Εφαρμόζεται μόνο όταν η πηγή ενός κειμένου είναι ο προφορικός λόγος.

- ***Morphology***: ασχολείται με την σύνθετη φύση των λέξεων οι οποίες αποτελούνται από μορφήματα. Το σύστημα του NLP αναλύει την λέξη στα συνθετικά της και έτσι αποκτά ανώτερη κατανόηση και παράγει αναπαραστάσεις υψηλότερου επιπέδου.

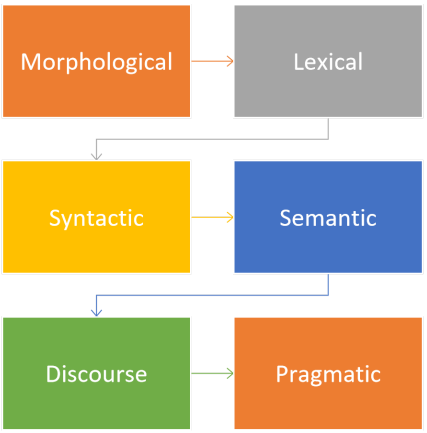
- ***Lexical***: το τρίτο επίπεδο ασχολείται με την κατανόηση διακριτών λέξεων με βάση τη θέση τους στον λόγο, τη σημασία και τη σχέση τους με άλλες λέξεις. Σε κάθε λέξη λοιπόν ανατίθεται μια ετικέτα αναφορικά με την χρήση της και το μέρος του λόγου. Σπάνια κάποια λέξη είναι μονοσήμαντη και έτσι το σύστημα αποκτά μεγαλύτερη ευελιξία.

- ***Syntactic***: Αυτό το επίπεδο εστιάζει στην ανάλυση των λέξεων σε μια πρόταση, ώστε να αποκαλύψει τη γραμματική δομή της πρότασης.

- ***Semantic***: Η σημασιολογική επεξεργασία καθορίζει τις πιθανές έννοιες μιας πρότασης εστιάζοντας στις αλληλεπιδράσεις μεταξύ των σημασιών των λέξεων στην πρόταση.

- ***Discourse***:Ενώ η σύνταξη και η σημασιολογία λειτουργούν με μονάδες μήκους προτάσεων, το επίπεδο λόγου αυτό λειτουργεί με μονάδες κειμένου μεγαλύτερες από μια πρόταση. Δηλαδή, δεν ερμηνεύει τα κείμενα πολλαπλών προτάσεων ως απλώς συνδυασμένες προτάσεις, καθεμία από τις οποίες μπορεί να ερμηνευτεί μεμονωμένα. Αντίθετα εστιάζει στις ιδιότητες του κειμένου ως συνόλου που μεταφέρουν νόημα και δημιουργούν συνδέσεις .

- ***Pragmatic***:Αυτό το τελευταίο επίπεδο ασχολείται με τη σκόπιμη χρήση της γλώσσας σε καταστάσεις και χρησιμοποιεί πλαίσιο πέρα από τα περιεχόμενα του κειμένου. Ο στόχος είναι η εξήγηση του πώς το επιπλέον νόημα διαβάζεται στα κείμενα χωρίς να κωδικοποιείται πραγματικά σε αυτά.



Σχήμα 2.3.1 : Τα φωνολογικά επίπεδα γλώσσας ενός NLP [38].

2.3.1 Γλωσσικές Αναπαραστάσεις (Word Representations)

Με τον όρο γλωσσικές αναπαραστάσεις αναφερόμαστε στη διαδικασία αντιστοίχισης κάθε λέξης σε κάποια μαθηματική αναπαράσταση ούτως ώστε να χρησιμοποιηθεί ακολούθως ώς είσοδο σε μοντέλα μάθησης για την επίλυση NLP προβλημάτων. Σύγχρονες προσεγγίσεις του χώρου, αντιμετωπίζουν κάθε λέξη ενός λεξιλογίου ως διανυσματικές αναπαραστάσεις, οι οποίες ονομάζονται εμφυτεύματα λέξεων (word embeddings). Ο κύριος στόχος είναι να εξάγουμε αντιπροσωπευτικά διανύσματα τα οποία καταφέρνουν να χαρακτηρίσουν το νοηματικό περιεχόμενο των λέξεων. Ιδέα κλειδί στις προσεγγίσεις αυτές είναι οι λέξεις με κοντινό περιεχόμενο να έχουν κοντινά εμφυτεύματα στο διανυσματικό χώρο των αναπαραστάσεων, εφαρμόζοντας κάποιο κριτήριο ομοιότητας.

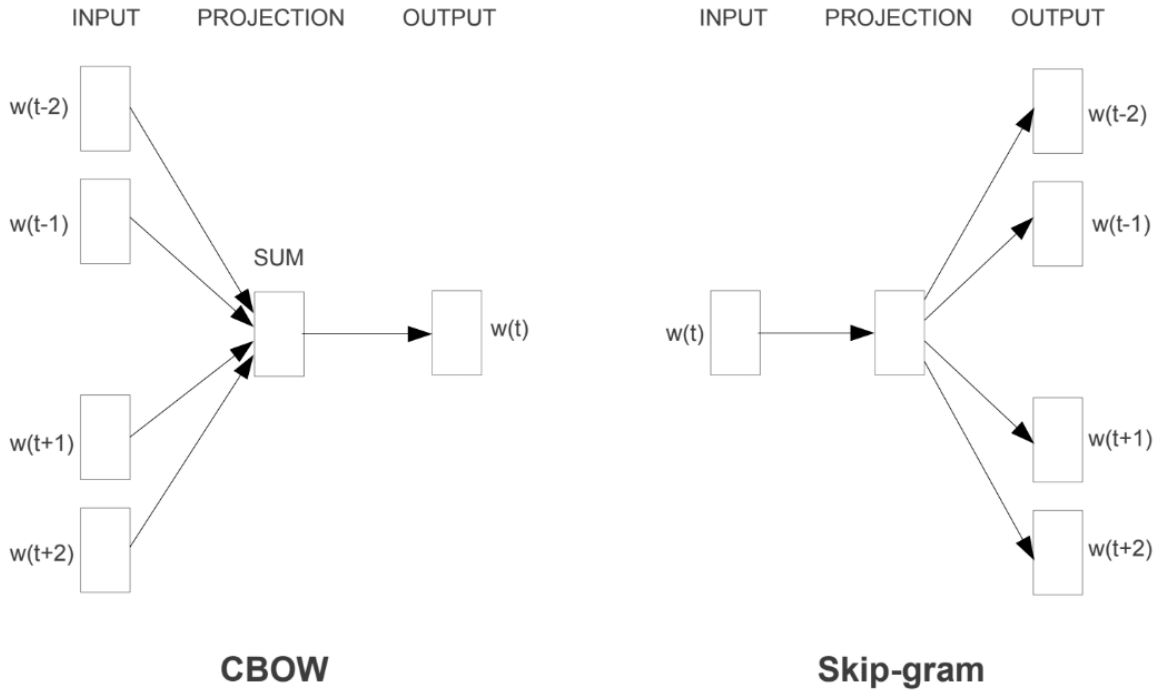
Υπάρχουν δύο κύριες προσεγγίσεις για την κατασκευή των γλωσσικών αναπαραστάσεων. Η πρώτη, γνωστή ως δηλωτική σημασιολογία (denotational semantics) αντιμετωπίζει κάθε λέξη σαν ξεχωριστό σύμβολο, δημιουργώντας έτσι αραιές αναπαραστάσεις χωρίς να ενσωματώνει κάποια έννοια ομοιότητας. Αντίθετα στην σημασιολογία κατανομής (distributional semantics) η λεκτική αναπαράσταση μαθαίνετε με βάση την χρήση της λέξης. Αυτό επιτρέπει σε λέξεις που χρησιμοποιούνται με παρόμοιους τρόπους να έχουν παρόμοιες αναπαραστάσεις, συλλαμβάνοντας φυσικά το νόημά τους. Στην συνέχεια αναλύονται κάποιες σύγχρονες μέθοδοι εύρεσης των λεκτικών εμφυτευμάτων.

(I)Word2Vec

Οι πρώτες προσπάθειες κατασκευής λεκτικών εμφυτευμάτων στηρίχτηκαν σε μια σειρά συχνοτικών μεθόδων με στόχο την στατιστική περιγραφή των λέξεων όπως αυτή προκύπτει από την χρήση τους μέσα στο κείμενο (One-Hot Vectorization, Count Vectorization, TF-IDF Vectorization, Window based co-occurrence,vectorization κλπ.). Παρόλα αυτά οι διαδικασίες αυτές ήταν υπολογιστικά ακριβές και κατασκεύαζαν αρκετά αραιές αναπαραστάσεις οι οποίες δεν ήταν αποδοτικές. Για αυτό τον λόγο στην συνέχεια υιοθετήθηκαν οι επαναληπτικές μεθόδοι όπως η Word2Vec. Βασική ιδέα των προσεγγίσεων αυτών είναι ότι λέξεις με παρόμοιο σημασιολογικό περιεχόμενο έχουν κοινά συμφραζόμενα και άρα πρέπει να έχουν γεωμετρικά κοντινές διανυσματικές αναπαραστάσεις. Με άλλα λόγια πρέπει να αναπαριστώνται κοντά η μία στην άλλη στον χώρο των εμφυτευμάτων (embedding space).

Η Word2Vec είναι μια στατιστική μέθοδος για την αποτελεσματική εκμάθηση μιας αυτόνομης λέξης που ενσωματώνεται από ένα σώμα κειμένου. Αναπτύχθηκε από τον Mikolov [39] το 2013 και για αρκετά χρόνια είχε κυριαρχήσει στον τομέα του NLP για την παραγωγή εμφυτευμάτων. Στην μέθοδο αυτή χρησιμοποιούνται τρία επίπεδα νευρωνικών δικτύων: το επίπεδο εισόδου στο οποίο τροφοδοτείται ένα μεγάλο σώμα λέξεων (corpus text), το επίπεδο προβολής (projection) στο οποίο το μοντέλο εκπαιδεύεται με βάση τα γλωσσικά συμφραζόμενα των λέξεων και το επίπεδο εξόδου. Η μέθοδος διαπερνά το σώμα λέξεων επαναληπτικά ούτως ώστε να μάθει τη συσχέτιση μεταξύ των λέξεων βασιζόμενη στην υπόθεση ότι κοντινα σημασιολογικές λέξεις έχουν γεωμετρικά κοντινές διανυσματικές αναπαραστάσεις. Η μέτρηση της ομοιότητας γίνεται με χρήση της μετρικής ομοιότητας συνημιτόνου (cosine similarity). Στην πραγματικότητα απλά μετράμε το συνημίτονο της γωνίας που σχηματίζουν τα διανύσματα των λέξεων. Για κοντινές λέξεις

στοχεύουμε σε γωνία που προσεγγίζει το μηδέν και ως επακόλουθο σε συνημίτονο που προσεγγίζει την μονάδα. Η μέθοδος Word2Vec συνδυάζει δύο προβλεπτικές μεθόδους: την CBOW (Continuous Bag of Words) και την Skip-Gram. ΄Οπως φαίνεται και στο ακόλουθο σχήμα ο μηχανισμός CBOW επιχειρεί να προβλέψει την "κεντρική" λέξη δεδομένων των συμφραζομένων, ενώ ο Skip-Gram μηχανισμός προσπαθεί να προβλέψει τα συμφραζόμενα δοθέντος την "κεντρική" λέξη.

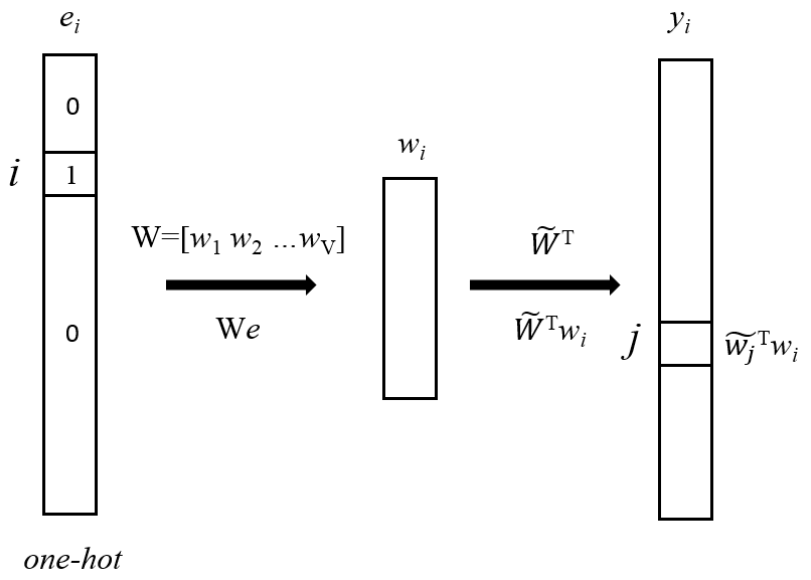


Σχήμα 2.3.2 : Οι μηχανισμοί CBOW και Skip-Gram στη μέθοδο Word2Vec [39].

(II) GloVe

Ο αλγόριθμος GloVe [40] (Global Vectors) αποτελεί επέκταση της μεθόδου Word2Vec για πιο αποδοτική εκμάθηση των γλωσσικών αναπαραστάσεων. Η κύρια διαφορά του είναι ότι δεν αξιοποιεί αποκλειστικά τα τοπικά στατιστικά ,δηλαδή τα τοπικά συμφραζόμενα λέξεων, αλλά και τα ολικά (Global) στατιστικά για την εύρεση των λεκτικών εμφυτευμάτων. Η αξιοποίηση του συνόλου του κειμένου λοιπόν γίνεται με την κατασκευή ενός πίνακα συνήπαρξης (coocurrence matrix) κάθε στοιχείο του οποίου A_{ij} υποδηλώνει πόσες φορές η λέξη i έχει εμφανιστεί μαζί με τη λέξη j . Έτσι, αφού αρχικοποιηθούν τα διανύσματα αναπαράστασης ξεκινάει η επαναληπτική διαδικασία εκμάθησης με στόχο την ελαχιστοποίηση κάποιας συνάρτησης κόστους η οποία συνήθως είναι το μέσο τετραγωνικό σφάλμα. Με αυτόν τον τρόπο, το μοντέλο αξιοποιεί τα κύρια οφέλη των ολικών

στατιστικών, μετρώντας συνεμφανίσεις λέξεων, ενώ συγχρόνως χρησιμοποιεί και αποδοτικές γραμμικές υποδομές, όπως το Word2Vec.



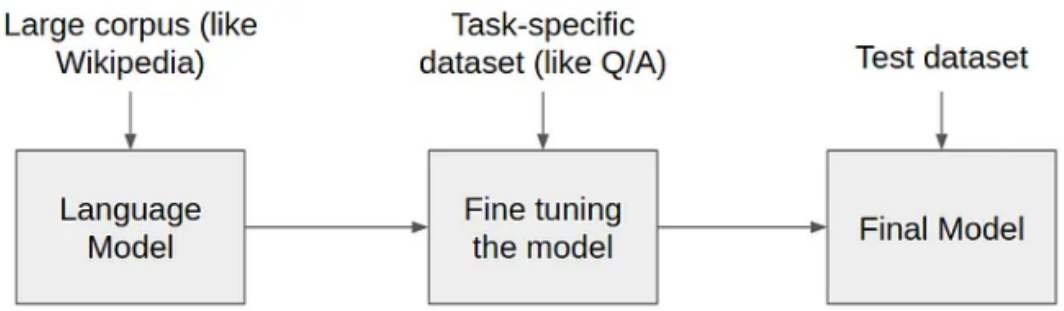
Σχήμα 2.3.3 : Η αρχιτεκτονική του μοντέλου Glove [40].

2.3.2 Προεκπαιδευμένα Μοντέλα για Επεξεργασία Φυσικής Γλώσσας

Τα προεκπαιδευμένα μοντέλα (pretrained models) για επεξεργασία φυσικής γλώσσας (NLP) είναι μοντέλα βαθιάς μάθησης τα οποία έχουν εκπαιδευτεί σε τεράστια σύνολα δεδομένων προκειμένου να εκτελούν μία συγκεκριμένη εργασία. Έχουν την δυνατότητα να μαθαίνουν καθολικές γλωσσικές αναπαραστάσεις, αφου εκπαιδεύονται σε τεράστια σώματα λέξεων στα οποία ο χρήστης είτε δεν έχει πρόσβαση είτε δεν διαθέτει συνήθως τους απαραίτητους υπολογιστικούς πόρους, γεγονός που τα καθιστά πολύ χρήσιμα σε διάφορες εφαρμογές NLP όπως η σύνοψη κειμένου, η αναγνώριση ονομαστικής οντότητας, η ανάλυση συναισθήματος, η επισήμανση μέρους του λόγου, η μετάφραση γλώσσας, η δημιουργία κειμένου κτλ.. Αυτό εξαλείφει την ανάγκη να εκπαιδεύετε ένα νέο μοντέλο από την αρχή κάθε φορά. Με άλλα λόγια, τα προεκπαιδευμένα μοντέλα μπορούν να θεωρηθούν ως επαναχρησιμοποιήσιμα μοντέλα που μπορούν να αξιοποιήσουν οι προγραμματιστές για να δημιουργήσουν γρήγορα και πιο αποδοτικά εφαρμογές NLP.

Η ροή των εργασιών είναι σειριακή και παράλληλα χαρακτηρίζεται από μία ιεραρχία εξειδίκευσης. Αναλυτικότερα, το γλωσσικό μοντέλο τροφοδοτείται πρώτα με ένα μεγάλο αριθμό μη

επισημασμένων δεδομένων (για παράδειγμα, η πλήρης βιβλιοθήκη του Wikipedia). Αυτό επιτρέπει στο μοντέλο να μάθει τη χρήση διαφόρων λέξεων και γενικά να αφομοιώσει την δομή της γλώσσας εξάγοντας πολύ αποδοτικές και αντιπροσωπευτικές διανυσματικές αναπαραστάσεις. Στην συνέχεια, το μοντέλο τροφοδοτείται με ένα άλλο μικρότερο σύνολο δεδομένων ούτως ώστε να εξειδικευτεί σε μια συγκεκριμένη εργασία NLP. Στο βήμα αυτό εκτελείται μία ακριβής προσαρμογή (fine tuning) των παραμέτρων του ήδη εκπαιδευμένου μοντέλου με σκοπό την τελειοποίησή του πάνω στην προαναφερθείσα εργασία. Με τον τρόπο αυτό προσπαθούμε να συγκεκριμενοποιήσουμε τις γενικές και καθολικές αναπαραστάσεις λέξεων που έχουν προκύψει καταλήγοντας έτσι σε μοντέλα υψηλής απόδοσης. Στην συνέχεια αναλύονται τα σημαντικότερα προεκπαιδευμένα μοντέλα .



Σχήμα 2.3.4 : Ροή εργασιών στην χρήση προεκπαιδευμένων μοντέλων Γλώσσας

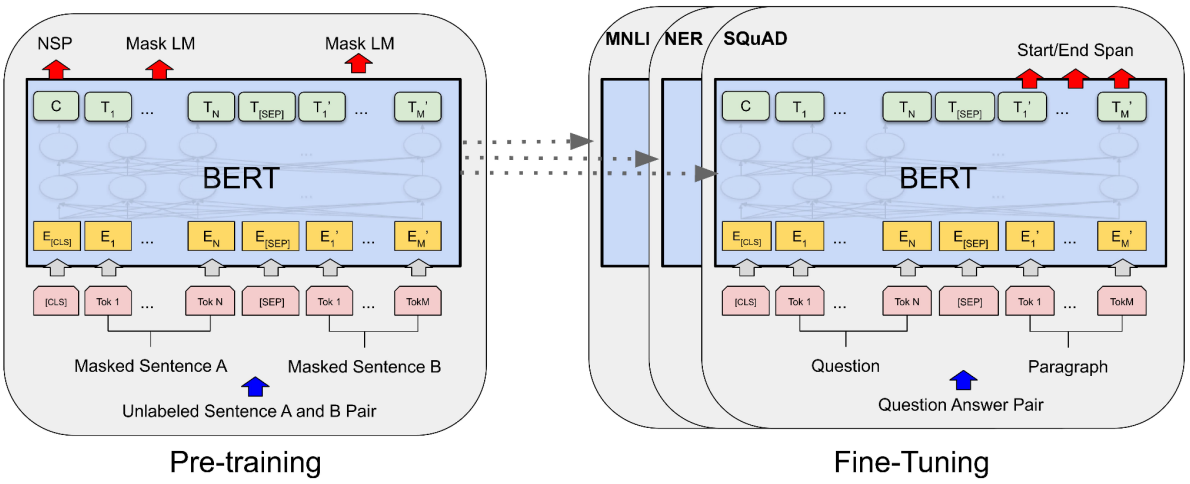
(I) Το μοντέλο BERT(Bidirectional Encoder Representations from Transformers)

Το μοντέλο BERT αναπτύχθηκε από μία ομάδα ερευνητών της Google AI το 2018 και αποτέλεσε κομβικό σημείο που συνέφερε καθοριστικά στην εξέλιξη του τομέα της επεξεργασίας της φυσικής γλώσσας. Στην πραγματικότητα, η κυκλοφορία του μοντέλου BERT σηματοδότησε την έναρξη μιας νέας εποχής στον χώρο του NLP λόγω της αποδοτικότητάς του αλλά και της ταχύτητας εκπαίδευσης που παρουσιάζει. Η βασική τεχνική καινοτομία του είναι η εφαρμογή της αμφίδρομης εκπαίδευσης στην αρχιτεκτονική του Transformer επιτυγχάνοντας έτσι μία βαθύτερη αίσθηση του γλωσσικού πλαισίου και ανώτερη κατανόηση των μοτίβων που διέπουν την γλώσσα. Και όλα αυτά σε συνδυασμό με το γεγονός ότι εκπαιδεύτηκε σε ένα τεράστιο σύνολο δεδομένων που περιελάμβανε πάνω από 3.3 δισεκατομμύρια λέξεις(Wikipedia ~2.5B λέξεις και Google’s BooksCorpus ~ 800M λέξεις).

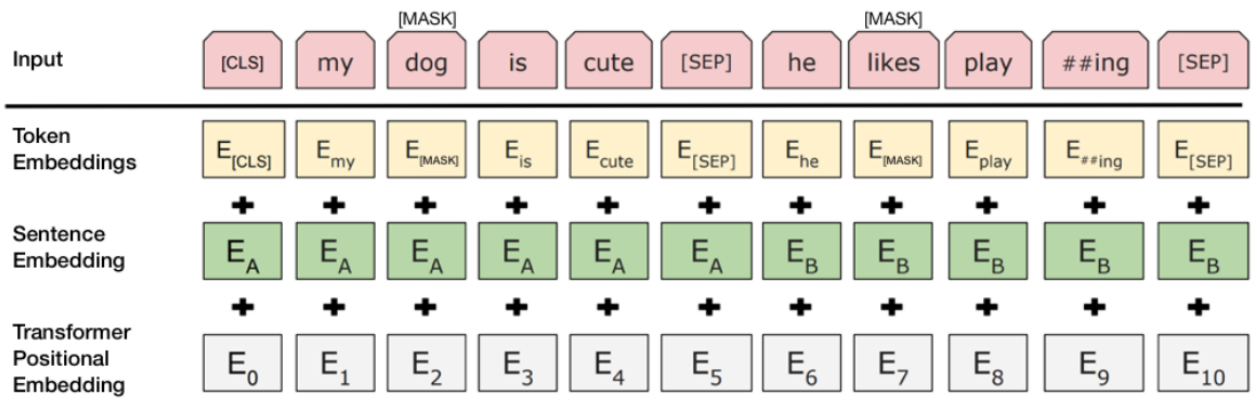
Η πρώτη στρατηγική με την οποία εκπαιδεύτηκε το μοντέλο ονομάζεται "**Masked Language Modeling**" (MLM). Στο εν λόγω πρόβλημα, το 15% όλων των WordPiece ενδείξεων (tokens) κάθε πρότασης "κρύβεται" με τυχαίο τρόπο (χρησιμοποιώντας το token [MASK]) και το μοντέλο επιχειρεί να προβλέψει την αρχική τιμή των καλυμμένων λέξεων, με βάση το πλαίσιο που

παρέχεται από τις άλλες, μη καλυμμένες, λέξεις της πρότασης. Τεχνικά αυτό γίνεται με την προσθήκη ενός επιπέδου ταξινόμησης στην έξοδο του Encoder.

Η δεύτερη στρατηγική με την οποία εκπαιδεύτηκε το μοντέλο ονομάζεται **"Next Sentence Prediction"** (NSP). Το μοντέλο λοιπόν λαμβάνει ζεύγη προτάσεων ως είσοδο και μαθαίνει να προβλέπει εάν η δεύτερη πρόταση στο ζεύγος είναι η επόμενη πρόταση στο αρχικό έγγραφο. Κατά τη διάρκεια της εκπαίδευσης, το 50% των εισροών είναι ένα ζευγάρι στο οποίο η δεύτερη πρόταση είναι η επόμενη πρόταση στο αρχικό έγγραφο (επισημασμένη ως IsNext), ενώ στο άλλο 50% μια τυχαία πρόταση από το σώμα (επισημασμένη ως Not Next). Η υπόθεση είναι ότι η τυχαία πρόταση θα αποσυνδεθεί από την πρώτη πρόταση. Κατά την διάρκεια εκπαίδευσης του μοντέλου οι δύο αυτές στρατηγικές υλοποιούνται παράλληλα και ο σκοπός είναι η ελαχιστοποίηση του από κοινού σφάλματος.



Σχήμα 2.3.5 : Οι διαδικασίες προεκπαίδευσης και fine-tuning του μοντέλου BERT [41].



Σχήμα 2.3.6 : Αναπαράσταση εισόδου του BERT.Τα εμφυτεύματα εισόδου είναι το άθροισμα των εμφυτευμάτων των tokens, των τμημάτων πρότασης και της θέσης [37].

Στο αρχικό άρθρο παρουσιάστηκαν δύο εκδοχές του μοντέλου BERT. Το BERTBASE που διαθέτει 12 στρώματα στη στοίβα του κωδικοποιητή και το BERTLARGE με 24 στρώματα. Οι αρχιτεκτονικές BERT (BASE και LARGE) έχουν επίσης μεγαλύτερα δίκτυα πρόσθιας-τροφοδότησης (με 768 και 1024 κρυφές μονάδες αντίστοιχα) και περισσότερα κεφαλές προσοχής (12 και 16 αντίστοιχα) από την πρωτότυπη αρχιτεκτονική των Transformers. Πάνω στην αρχιτεκτονική του BERT έχουν θεμελιωθεί μία σειρά από προεκπαιδευμένα μοντέλα τα σημαντικότερα εκ των οποίων αναφέρονται ακολούθως: CodeBERT, OpenNMT, RoBERTa, ELMo, XLNet, ULMFit.

2.4 Νευρωνικά Δίκτυα σε Γράφους

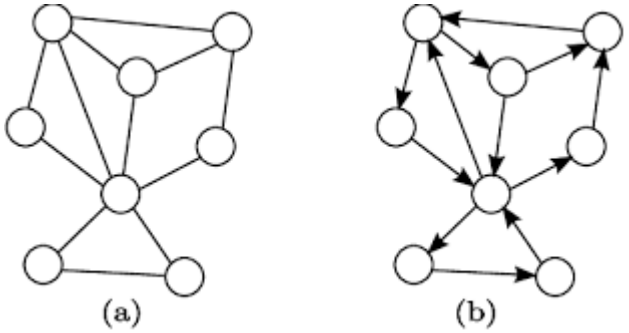
Οι γράφοι είναι ένα είδος δομής δεδομένων που μοντελοποιεί ένα σύνολο αντικειμένων, που ονομάζονται κόμβοι, και τις σχέσεις και αλληλεπιδράσεις που αναπτύσσονται μεταξύ τους (ακμές). Τα τελευταία χρόνια, έρευνες πάνω στην ανάλυση γραφημάτων με μηχανική μάθηση λαμβάνουν όλο και μεγαλύτερη προσοχή λόγω της μεγάλης εκφραστικής τους δύναμης αλλά και των σχέσεων που μπορούν να ενσωματώσουν. Έχουν εφαρμοστεί σε διάφορους τομείς συμπεριλαμβανομένων των κοινωνικών επιστημών [42], τις φυσικές επιστήμες [43], τα δίκτυα αλληλεπίδρασης πρωτεΐνης-πρωτεΐνης [44], γραφήματα γνώσης [45] και πολλές άλλες ερευνητικές περιοχές [46]. Εστιάζει κυρίως σε εργασίες όπως η ταξινόμηση κόμβων, η πρόβλεψη συνδέσμων και η ομαδοποίηση. Η κύρια πρωτοτυπία που εισάγει είναι ότι στην κλασική Μηχανική Μάθηση επεξεργαζόμαστε Ευκλείδεια δεδομένα τα οποία θεωρούνται ανεξάρτητα μεταξύ τους. Παρόλα αυτά οι γράφοι αποτελούν ακανόνιστες δομές χωρις κάποια συγκεκριμένη γεωμετρία και με κόμβους οι οποίοι πλέον αλληλεπιδρούν μεταξύ τους με διάφορους τύπους συνδέσμων. Στην συνέχεια παρουσιάζουμε την λειτουργία βαθιών νευρωνικών δικτύων σε γράφους.

2.4.1 Βασικά Στοιχεία Θεωρίας Γραφημάτων

Γράφος ή γράφημα (graph) είναι μία δομή που αποτελείται από ένα σύνολο κορυφών (vertices) ή κόμβων (nodes) που συνδέονται μεταξύ τους με ένα σύνολο ακμών (edges) ή γραμμών (lines). Κάθε γράφος ορίζεται ως ένα ζεύγος $G=(V,E)$ όπου V είναι το πλήθος των κόμβων και E το πλήθος των ακμών. Το πλήθος των κόμβων ενός γράφου συμβολίζεται με n= |V| και ονομάζεται τάξη (order) του γράφου. Το πλήθος των ακμών συμβολίζεται με m =|E| και ονομάζεται μέγεθος (size) του γράφου. Η **πυκνότητα** (density) ενός γράφου, ορίζεται από το σύνολο των ακμών του και των κόμβων του. Εάν δύο γράφοι έχουν την ίδια τάξη, αλλά διαφορετικό μέγεθος, τότε έχουν διαφορετική πυκνότητα (ο γράφος με το μεγαλύτερο μέγεθος λέμε ότι είναι πιο πυκνός από αυτόν με το μικρότερο μέγεθος).

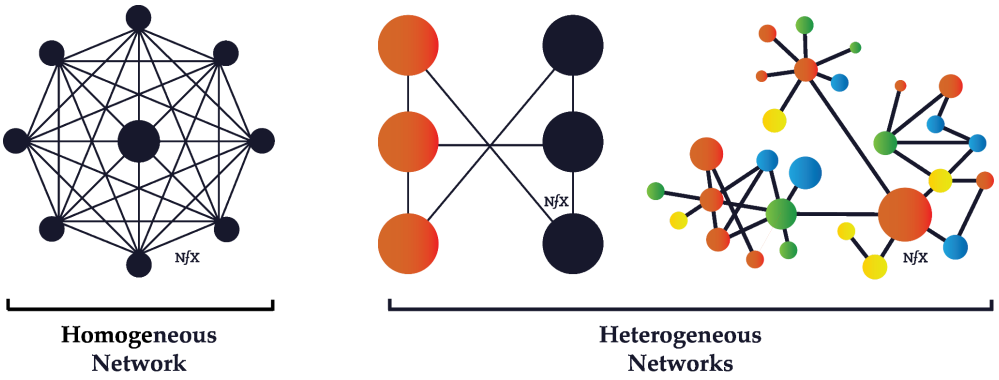
Δεδομένης μιας ακμής (u, v), ο κόμβος u ονομάζεται **γείτονας** του κόμβου v. Η **γειτονιά** (neighborhood) ενός κόμβου u συμβολίζεται με N(u) και είναι το σύνολο των κόμβων που ορίζεται από την σχέση : $N(u) = [u \in V(G) | (u, v) \in E(G)]$. Το πλήθος των γειτόνων-ακμών που προσπίπτουν στο κόμβο u, ονομάζεται **βαθμός** (degree) του κόμβου u και συμβολίζεται με d(u). Αν για κάποιον κόμβο ισχύει d(u) = 0, ο κόμβος ονομάζεται **απομονωμένος** (isolated) και αντίστοιχα, αν ισχύει d(u) = 1, ο κόμβος ονομάζεται **εκκρεμή**ς (pendant). Ένας γράφος, για τον οποίο κάθε κόμβος του έχει βαθμό d ονομάζεται **d-κανονικός**. Ενας γράφος G λέγεται **συνδεδεμένος ή συνεκτικός**, εάν υπάρχει μονοπάτι που να συνδέει όλα τα ζεύγη κόμβων μεταξύ τους.

Οι γράφοι ανάλογα με την κατεύθυνση των ακμών διακρίνονται σε δύο κατηγορίες : τους κατευθυνόμενους γράφους (directed) και τους μη κατευθυνόμενους γράφους (undirected).Ένας γράφος ονομάζεται κατευθυνόμενος, εάν περιλαμβάνει αποκλειστικά κατευθυνόμενες ακμές. Οι ακμές ενός γράφου ονομάζονται κατευθυνόμενες, όταν τα ζεύγη των ακμών (u, v) και (v, u) λαμβάνονται ως διατεταγμένα, δηλαδή (u, v) ≠ (v, u). Στην περίπτωση που οι ακμές δεν θεωρούνται διατεταγμένες, τότε ονομάζονται μη διατεταγμένες ή απλές. Αντίστοιχα, στον μη κατευθυνόμενο γράφο εμφανίζονται μόνο απλές ακμές.



Σχήμα 2.4.1 : Ένας μη κατευθυνόμενο γράφος (αριστερά) και ένας κατευθυνόμενος (δεξιά).

Επιπλέον οι γράφοι διακρίνονται με βάση τον τύπο των κόμβων και των ακμών τους σε ομογενείς (Homogeneous graph) και ετερογενείς (Heterogeneous graph). Ένας ομογενής γράφος είναι ένας γράφος στον οποίο όλοι οι κόμβοι και όλες οι ακμές είναι του ίδιου τύπου. Για παράδειγμα, σε έναν ομογενή γράφο κοινωνικής δικτύωσης, όλοι οι κόμβοι μπορεί να είναι άτομα και οι σχέσεις που αναπτύσσονται μεταξύ τους να είναι σχέσεις φιλίας. Αντίθετα σε έναν ετερογενή γράφο μπορεί να έχουμε κόμβους ή συνδέσεις διαφόρων τύπων. Για παράδειγμα, σε έναν ετερογενή γράφο κοινωνικής δικτύωσης, οι κόμβοι μπορεί να αντιπροσωπεύουν όχι μόνο άτομα, αλλά και οργανισμούς και ομάδες ενώ οι σχέσεις που τους συνδέουν μπορεί να είναι φιλίας, κοινά ενδιαφέροντα και άλλα.

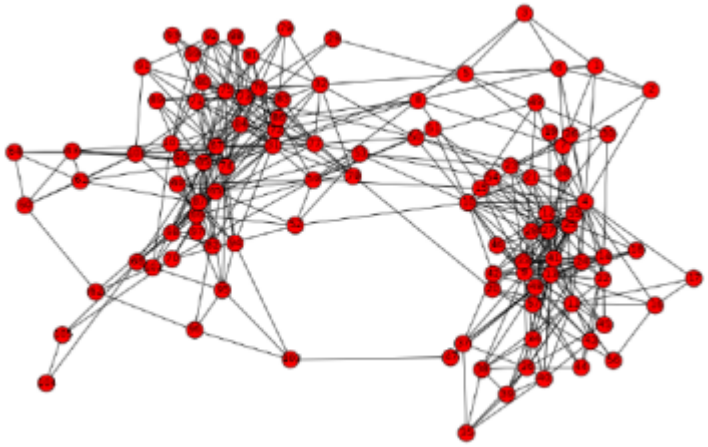


Σχήμα 2.4.2 : Ομογενής γράφος (αριστερά) και ένας ετερογενής γράφος (δεξιά).

Τέλος, η τοπολογική πληροφορία του γράφου *G=(V,E)* περιγράφεται από τον **Πίνακα Γειτνίασης** (Adjacency Matrix) ο οποίος είναι ένας $|V|x|V|$ πίνακας που ορίζεται από την ακόλουθη σχέση:

$$A = (a_{ij})_{nxn} = \{ 1,\ if\ (i,\ j) \in E,\ \forall i,\ j \in 1,\ \ldots,\ n\ \ and\ 0\ ,otherwise\ \} \qquad (2.4.1)$$

Ο Adjacency Matrix για μη κατευθυνόμενους γράφους είναι συμμετρικός ενώ τα στοιχεία της διαγωνίου είναι πάντα μηδενικά. Σε περίπτωση που οι ακμές είναι σταθμισμένες (weighted) ορίζεται ο αντίστοιχος Πίνακας Βαρών (Weighted Matrix) όπου κάθε στοιχείο w_{ij} απεικονίζει το βάρος της ακμής που συνδέει του κόμβους i και j.



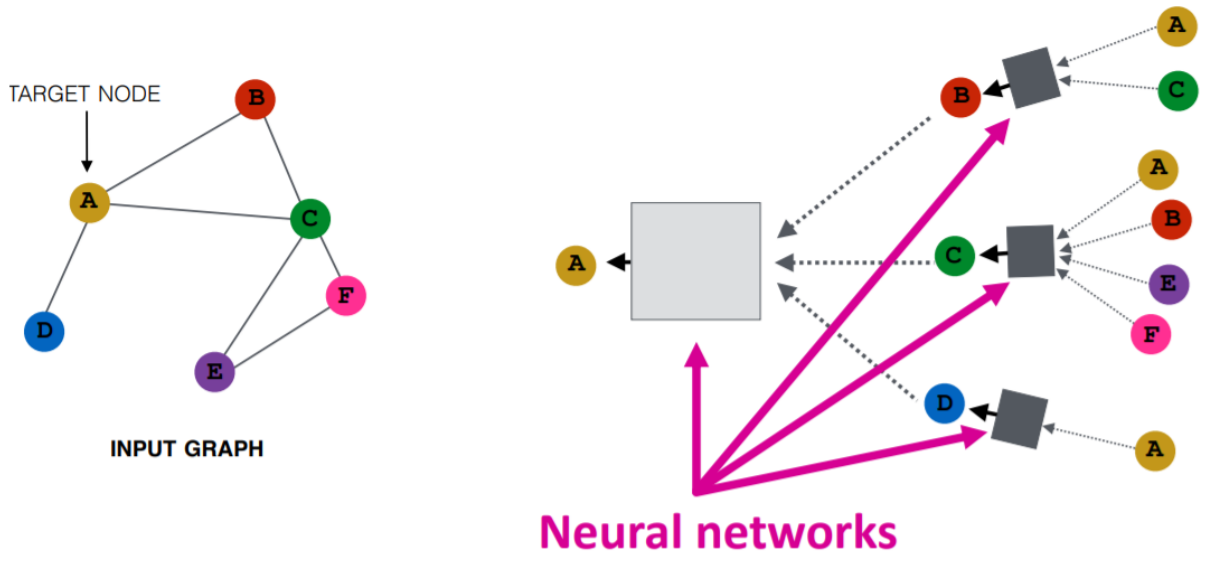
Σχήμα 2.4.3: Παράδειγμα γράφου του δικτύου the polbooks network [47].

2.4.2 Νευρωνικά Δίκτυα σε Γράφους (Graph Neural Network's - GNN)

Τα GNN (Graph Neural Networks) ανήκουν στην ευρύτερη κλάση των νευρωνικών δικτύων και είναι σχεδιασμένα για να επεξεργάζονται και να αναλύουν δεδομένα τα οποία είναι δομημένα σε γράφους. Σκοπός τους είναι να μάθουν εκφραστικές αναπαραστάσεις (representations) για τους κόμβους ενσωματώνοντας την πληροφορία που προκύπτει από τους γειτονικούς κόμβους. Διακρίνονται σε δύο μεγάλες κατηγορίες:

- Η **χωρική προσέγγιση γραφήματος (spatial based graph)** είναι απλή και λειτουργεί στην τοπολογία γραφήματος συγκεντρώνοντας πληροφορίες από την γειτονιά του κόμβου. Οι τεχνικές που βασίζονται σε χωρικά γραφήματα προτιμώνται καθώς είναι πολύ αποδοτικές και ταυτόχρονα εμφανίζουν μικρή υπολογιστική πολυπλοκότητα.
- Το **φιλτράρισμα γραφήματος με βάση το φάσμα (spectral based graph filtering)** βασίζεται σε μια αποσύνθεση Eigen του πίνακα Laplacian του γραφήματος που επιτρέπει τον μετασχηματισμό Fourier του γραφήματος. Προτιμάται κυρίως στην ανάλυση σημάτων όπου η μεταφορά στο πεδίο της συχνότητας λειτουργεί πιο αποτελεσματικά.

Η βασική ιδέα στην οποία στηρίζονται είναι η **μεταβίβαση μηνυμάτων (message passing)** η οποία τους επιτρέπει να αξιοποιήσουν τις σχέσεις και τις συνδέσεις που αναπτύσσονται στους γράφους. Αναλυτικότερα , ορίζουμε έναν γράφο G=(V,E) όπου περιλαμβάνει ένα σύνολο κόμβων $S \subseteq V$ οι οποίοι περιλαμβάνουν ετικέτες δηλαδή την κλάση στην οποία ανήκουν. Αρχικά τα διανύσματα των κόμβων αρχικοποιούνται με ένα διάνυσμα χαρακτηριστικών το οποίο μπορεί να συμπεριλαμβάνει κάποιες ιδιότητες των οντοτήτων που αντικατοπτρίζουν οι κόμβοι όπως αριθμητικά ή κατηγορηματικά χαρακτηριστικά. Στην συνέχεια για κάθε κόμβο υπολογίζεται ο **υπολογιστικός γράφος (computational graph)**. Οι υπολογιστικοί γράφοι είναι ένας τύπος κατευθυνομενων γραφων που αξιοποιείται για την αναπαράσταση μαθηματικών εκφράσεων αλλά και για να οπτικοποιήσει την ροή της πληροφορίας στο message passing.

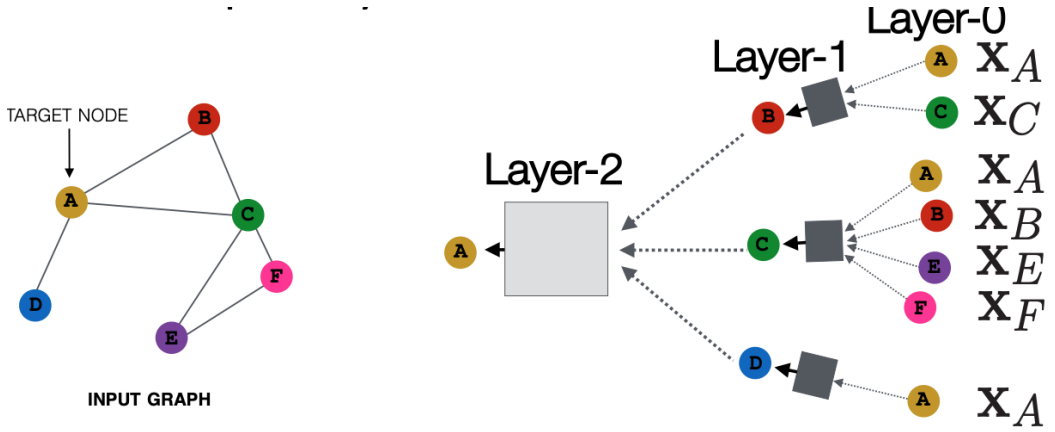


Σχήμα 2.4.4 : Ο υπολογιστικός γράφος του κόμβου A σε ένα GNN 2-επιπέδων [48].

Στην συνέχεια παρουσιάζουμε την ροή που ακολουθείται κατά την πρόσθια διάδοση (forward propagation) που καθορίζει τον τρόπο με τον οποίο η πληροφορία μεταφέρεται από την είσοδο στην έξοδο του νευρωνικού δικτύου. Σε κάθε λοιπόν επανάληψη το νευρωνικό δίκτυο για την μεταβίβαση μηνυμάτων συνδυάζει (aggregates) την πληροφορία που προκύπτει από τους γειτονικούς κόμβους και μέσω μίας συνάρτησης ανανέωσης παράγει τα τελικά embeddings. Αναλυτικότερα χρησιμοποιούμε μία συνάρτηση(aggregation function) για να συνδυάσουμε την πληροφορία που προκύπτει από τους γειτονικούς κόμβους. Η συνάρτηση αυτή μπορεί να είναι η μέση τιμή, το συνολικό άθροισμα ή κάποια σταθμισμένη μέση τιμή. Στην συνέχεια το αποτέλεσμα διέρχεται από ένα fully connected layer (MLP) για να παράγουμε την τελική αναπαράσταση για τον κόμβο. Η μαθηματική σχέση που ακολουθείται για το l-layer του GNN είναι η εξής:

$$h_v^l = \sigma\left(\frac{1}{|N(v)|}\sum_{u\in N(v)} W^l \cdot h_u^{l-1} + \mathrm{B}^l \cdot h_v^{l-1}\right) \qquad (2.4.2)$$

όπου σ είναι η μη γραμμική συνάρτηση ενεργοποίησης, N(v) το σύνολο των κόμβων που ανήκουν στην γειτονιά του κόμβου v και W^l , B^l είναι το προς εκμάθηση βάρη. Ο όρος $\mathrm{B}^l \cdot h_v^{l-1}$ είναι απλά ένα self-loop, δηλαδή η αναπαράσταση του κόμβου v του προηγούμενου επιπέδου πολλαπλασιασμένο με έναν πίνακα βαρών. Να σημειωθεί ότι η συνάρτηση συνδυασμού που αξιοποιήθηκε είναι η μέση τιμή για αυτό και ο όρος κανονικοποίησης $\frac{1}{|N(v)|}$.

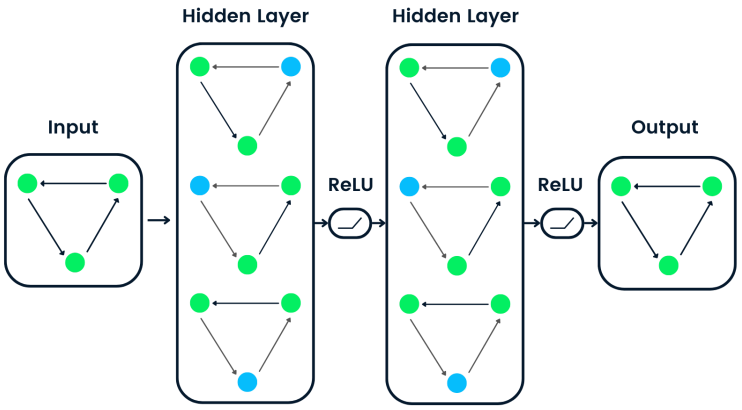


Σχήμα 2.4.5 : Οι υπολογιστικοί γράφοι για κάθε κόμβο του δικτύου [48].

Στην συνέχεια απλά συνδυάζουμε σειριακά πολλαπλά επίπεδα GNN. Κάθε επίπεδο που προσθέτουμε αυξάνει το βάθος του computational graph και το εύρος των γειτόνων που αξιοποιεί κάθε κόμβος για να εξάγει την τελική του αναπαράσταση. Αναλυτικότερα κάθε layer αντιστοιχεί στο βήμα (step) που αξιοποιείται δηλαδή για 1 layer αξιοποιούμε την πληροφορία που προκύπτει

από τους γείτονες του κόμβου (step 1) ενώ για 2 layer αξιοποιουμε επιπλέον την πληροφορία από τους γείτονες των γειτόνων (step 2). Η επιλογή του πλήθους των layers πρέπει να γίνεται προσεκτικά καθώς για υψηλό αριθμό επιπέδων υπάρχει ο κίνδυνος να αλλοιωθεί η πληροφορία και να επικρατήσει μία ομοιομορφία στις τελικές αναπαραστάσεις των κόμβων.

Η εκπαίδευση των GNN ακολουθεί τα κλασικά πρότυπα της μηχανικής μάθησης, ορίζοντας κάποια συνάρτηση κόστους την οποία προσπαθούμε να ελαχιστοποιήσουμε μέσω μίας συνάρτησης βελτιστοποίησης (optimization function).



Σχήμα 2.4.6 : Η αρχιτεκτονική ενός GCN 2-επιπέδων [49].

Διάφορα είδη GNN έχουν προταθεί τα οποία διαφοροποιούνται κυρίως στον τρόπο με τον οποίο εκτελείται το aggregation Τα πιο δημοφιλή είναι : τα Graph Convolutional Networks (GCN) [49], GraphSAGE [50] και τα Graph Attention Networks(GAT) [51] τα οποία προσθέτουν και έναν μηχανισμό προσοχής.

Κεφάλαιο 3

Πρότερη Δουλειά

Στην συγκεκριμένη ενότητα παρέχουμε μια ολοκληρωμένη επισκόπηση της υπάρχουσας βιβλιογραφίας, μελετών και προηγμένων εξελίξεων οι οποίες σχετίζονται με το θέμα που ερευνάται δηλαδή την ανίχνευση bot λογαριασμών. Περιλαμβάνει μια εμπεριστατωμένη αναθεώρηση και σύνθεση των προηγούμενων ερευνών γεγονός που συμβάλλει στην τοποθέτηση της συγκεκριμένης εργασίας αλλά και στην ανάδειξη της συμβολής της. Θα ακολουθήσουμε μια χρονολογική σειρά ώστε να επισημάνουμε την λογική εξέλιξη των αρχιτεκτονικών που αξιοποιήθηκαν αλλά και τις καινοτομίες που εισήγαγαν. Γενικά το σύνολο των μεθόδων μπορεί να διαιρεθεί σε δύο κατηγορίες: (i) μέθοδοι που στηρίζονται στην μηχανική εξαγωγής χαρακτηριστικών (feature based methods) και (ii) μέθοδοι η οποίες αξιοποιούν αρχιτεκτονικές βαθιάς μάθησης (Deep Learning Methods).

3.1 Feature Based Methods

Το σύνολο των προτεινόμενων μεθοδολογιών που στηρίζονται στην μηχανική εξαγωγής χαρακτηριστικών αξιοποιούν τεχνικές επιβλεπόμενης και μη επιβλεπόμενης μάθησης, με την συντριπτική πλειοψηφία να ανήκει στην πρώτη κατηγορία.

3.1.1 Μέθοδοι που βασίζονται σε Τεχνικές Επιβλεπόμενης Μάθησης

Οι Lee et al. [52] πρότειναν μία από τις πρώτες μεθόδους ανίχνευσης των bot στο Twitter. Αναλυτικότερα, ανέπτυξαν έναν ταξινομητή τον οποίο ονόμασαν Decorate ο οποίο συνδύαζε πληροφορίες περιεχομένου , λογαριασμού και στατιστικά χαρακτηριστικά που περιγράφουν την χρήση και δραστηριότητα του λογαριασμού πάνω σε ένα σύνολο δεδομένων που παρήγαγαν οι ίδιοι. Προκειμένου να αντλήσουν αντιπροσωπευτικά προφίλ bot έθεσαν διάφορες κοινωνικές παγίδες στα μέσα κοινωνικής δικτύωσης οι οποίες θα προσέλκυαν την λειτουργία τους και στην συνέχεια εκτελώντας μία στατιστική ανάλυση των χαρακτηριστικών τους ανέπτυξαν έναν ταξινομητή ο οποίος θα είχε απτά αποτελέσματα στον πραγματικό κόσμο του Twitter. Σε παρόμοια πλαίσια κινήθηκαν και οι Wang et al. [11] οι οποίοι όμως λαμβάνουν τα χαρακτηριστικά μόνο από τα τελευταία 20 tweets των χρηστών και αξιοποιούν ένα ταξινομητη Naive-Bayes.

Στα χρόνια που ακολούθησαν η ερευνητική κοινότητα έδωσε έμφαση στην αξιοποίηση και εξαγωγή εκτενεστερων διανυσμάτων χαρακτηριστικών με παράλληλη μεγαλύτερη εκφραστικότητα προκειμένου να αναβαθμίσουν το επίπεδο επίδοσης των μοντέλων τους. Οι Chu et al. [53] παρατήρησαν τη διαφορά μεταξύ ανθρώπων και bot όσον αφορά τη συμπεριφορά στις αναρτήσεις, το περιεχόμενο των αναρτήσεων και τις ιδιότητες του λογαριασμού και προτείνουν ένα σύστημα ταξινόμησης που περιλαμβάνει τα παρακάτω μέρη: ένα στοιχείο βασισμένο στην εντροπία, ένα στοιχείο ανίχνευσης ανεπιθύμητων αναρτήσεων, ένα στοιχείο ιδιοτήτων του λογαριασμού και έναν τελικό ταξινομητή RandomForest. Έναν χρόνο αργότερα οι Yang et al. [54] βελτίωσαν ακόμα περισσότερο την ακρίβεια του μοντέλου χρησιμοποιώντας τον ίδιο ταξινομητή αλλά χαρακτηριστικά μεγαλύτερης εκφραστικότητας όπως ο τρόπος εισόδου του χρήστη στην πλατφόρμα Twitter ή την τακτικότητα της ανάρτησης tweet από τον χρήστη. Παράλληλα ανέλυσαν την ανθεκτικότητα των χαρακτηριστικών αλλά και την διακριτότητα που συνεισφέρουν.

Οι Cresci et al [55] μοντελοποίησαν την συμπεριφορά των χρηστών μέσω μιας ψηφιακής αλυσίδας DNA (Digital DNA Modelling). Στην ουσία δημιουργεί μία βάση DNA οι οποία περιέχει ένα αλφάβητο που αντιστοιχεί σε δράσεις του χρήστη (user_activities) και έτσι μοντελοποιεί την δραστηριότητα κάθε χρήστη μέσω ενός string που περιέχει τα ακολουθιακά του δεδομένα. Η αλυσίδα αυτή λοιπόν χρησιμοποιείται για την κατηγοριοποίηση των χρηστών βασιζόμενοι στην ιδέα ότι τα spam bots, τα οποία παράγονται από αυτοματοποιημένα λογισμικά, θα εμφανίζουν μεγάλη ομοιομορφία στο ψηφιακό DNA τους. Έτσι εφάρμοσε την μετρική της Μεγαλύτερης Κοινής Υποακολουθίας (Longest Common Subsequence - LCS) προκειμένου να ανιχνεύσει πρότυπα και ομοιότητες. Με άλλα λόγια ψάχνει για ομάδες χρηστών οι οποίες εμφανίζουν μεγάλο LCS. Η ιδέα του Cresci εμφάνισε αρκετά υψηλά επίπεδα απόδοσης. Παρόλα αυτά σε σύνολα δεδομένων που εμπεριέχουν διαφόρων ειδών Bot’s αδυνατεί να ανιχνεύσει ομοιότητες αποδοτικά.

Ο Davis το 2016 [56] παρουσίασε το BotOrNot μία αρχιτεκτονική η οποία παράγει πάνω από 1000 χαρακτηριστικά για κάθε χρήστη και στην συνέχεια αναγεται σε ένα κλασικό πρόβλημα ταξινόμησης (classification problem) όπου η έξοδός του είναι η πιθανότητα ένας χρήστης να είναι bot (bot likelihood). Είναι η πρώτη μέθοδος η οποία εισήγαγε τοπολογικά χαρακτηριστικά τα οποία ενσωματώνουν τις συνδέσεις και αλληλεπιδράσεις που αναπτύσσονται μεταξύ των χρηστών. Στην συνέχεια τα διανύσματα χαρακτηριστικών των χρηστών προωθούνται σε έναν αλγόριθμο μηχανικής μάθησης και συγκεκριμένα στον αλγόριθμο RandomForest αξιοποιώντας ξεχωριστό ταξινομητή για κάθε τύπο χαρακτηριστικών. Ένα μεγάλο πρόβλημα της συγκεκριμένης μεθόδου είναι ότι αυτή η χειροκίνητη εξαγωγή χαρακτηριστικών (handcrafted features) εμφανίζει μεγάλη υπολογιστική πολυπλοκότητα. Ιδιαίτερα στους γράφους πρέπει να ξαναυπολογίζονται κάθε φορά που προστίθεται ένας καινούριος χρήστης ή γίνεται ένα καινούριο follow εφόσον αλλάζει η δομή του γράφου κάτι που προφανώς δεν είναι αποδοτικό.

Οι Gilani et al. [57] υιοθέτησαν μία διαφορετική μεθοδολογία η οποία βασίστηκε στα πολυμεσικά στοιχεία που περιέχουν τα tweets. Προκειμένου να βελτιώσουν την επίδοση του μοντέλου τους διαίρεσαν το σύνολο δεδομένων σε τέσσερις κατηγορίες δημοτικότητας ανάλογα με το πλήθος των ακολούθων των χρηστών και στην συνέχεια εφάρμοσαν έναν ταξινομητή RandomForest. Στις πειραματικές τους δοκιμές, βρέθηκε ότι χρησιμοποιώντας την παραπάνω αναπαράστασή για τους χρήστες είναι πιο δύσκολο να ανιχνευθεί ένα bot μεταξύ λογαριασμών που έχουν λιγότερους από χίλιους ακόλουθους.

Οι Hayes et al. [58] κωδικοποιούν πάλι το ιστορικό δημοσίευσης ενός λογαριασμού στο Twitter ως μία ακολουθία χαρακτήρων μέσω της ψηφιακής αλυσίδας DNA. Η βασική τους ιδέα στηρίχθηκε στην υπόθεση ότι η μακροπρόθεσμη συμπεριφορά ενός bot λογαριασμού είναι λιγότερο τυχαία από την συμπεριφορά ενός γνήσιου χρήστη. Η μέτρηση αυτής της τυχαιότητας και απροβλεψιμότητας έγινε μέσω της αξιοποίησης ενός αλγορίθμου συμπίεσης (compression algorithm) μέσω του οποίου εξάγει κάποια στατιστικά στοιχεία όπως το μέγεθος της μη συμπιεσμένης ακολουθίας DNA, το μέγεθος της συμπιεσμένης ακολουθίας DNA και την τιμή του ποσοστού συμπίεσης. Αυτά τα χαρακτηριστικά χρησιμοποίησε στην ταξινόμηση των χρηστών σε γνήσιους και bot.

Οι Bacciu et al. [59] στα πλαίσια της ερευνητικής εργασίας του PAN 2019 πάνω στην σκιαγράφηση συγγραφικού προφίλ στο Twitter παρουσίασαν την δικιά τους μέθοδο πάνω στην ταξινόμηση των χρηστών σε γνήσιους και bot. Βασισμένοι στην μηχανική εξαγωγής χαρακτηριστικών υλοποιούν μία εκτενή προεπεξεργασία των δεδομένων προκειμένου να καταλήξουν στο τελικό διάνυσμα χαρακτηριστικών. Αναλυτικότερα εξάγουν χαρακτηριστικά από τα tweets μέσω του μέτρου ομοιότητας συνημιτόνου, της ανάλυσης συναισθημάτων (sentiment analysis) και της παραμόρφωσης κειμένου. Στην συνέχεια αξιοποιούν ένα ταξινομητή δύο επιπέδων προς επίλυση του ζητήματος. Στο πρώτο layer χρησιμοποιούν έναν SVM ταξινομητή με πυρήνα την συνάρτηση ακτινικής βάσης και έναν Adaboost 30 ταξινομητών. Στο επόμενο layer απλά συνδυάζουν τις προβλέψεις των επιμέρους ταξινομητών μέσω ενός ταξινομητή με χαλαρή ψηφοφορία (soft voting classifier) για να αντλήσουν το τελικό αποτέλεσμα.

Η προσέγγιση των Loyola-González [60] αποσκοπεί στην ανάπτυξη ενός κατανοητού και επεξηγήσιμου μοντέλου ταξινόμησης για την ανίχνευση των bot στο Twitter. Το μοντέλο που εισάγουν είναι βασισμένο σε μοτίβα αντιδιαστολής (contrast pattern_based model). Αναλυτικότερα τα μοντέλα αυτά χρησιμοποιούν μία συλλογή από μοτίβα αντιδιαστολής για να αναπτύξουν έναν ταξινομητή ο οποίος θα αντιστοιχεί κάποιο ερώτημα σε μία προκαθορισμένη κατηγορία. Παράλληλα για κάθε χρήστη συνδυάζει πληροφορίες οι οποίες εξάγονται από το περιεχόμενο των tweets, τον λογαριασμό τους και από την ανάλυση συναισθήματος (sentiment analysis) του περιεχομένου τους. Η παραπάνω μέθοδος εμφανίζει ανταγωνιστική απόδοση. Επιπλέον, η κυρίαρχη συνεισφορά της είναι η επεξηγηματικότητα που συνοδεύει κάθε απόφαση του μοντέλου και η συμβολή της στον προσδιορισμό του τι θεωρείται ως μη ανθρώπινη δραστηριότητα σε ένα online δίκτυο.

Οι Rodríguez-Ruiz et al. [61] αναγνώρισαν ότι βασική προϋπόθεση και παράλληλα περιορισμός που συνοδεύει την χρήση ταξινομητών πολλών κλάσεων (multi-class classifiers) είναι η ύπαρξη αντιπροσωπευτικού συνόλου εκπαίδευσης. Μάλιστα στην περίπτωση των bot η εργασία γίνεται ακόμα πιο απαιτητική εφόσον η ανίχνευσή τους είναι αρκετά δύσκολη και απαιτεί εκτεταμένες επικυρώσεις. Έτσι, προτείνουν μία μεθοδολογία η οποία στηρίζεται σε έναν ταξινομητή μίας κλάσης (one-class classifier) αντιμετωπίζοντας την πρόκληση της ανίχνευσης των bot σαν ένα πρόβλημα ανίχνευσης ανωμαλιών (anomaly detection problem). Πειραματίζονται λοιπόν με διάφορους ταξινομητές οι οποίοι μοντελοποιούν την συμπεριφορά γνήσιων χρηστών και στην συνέχεια ανιχνευουν τα bot σαν χρήστες οι οποίοι αποκλίνουν από την συμπεριφορά αυτή. Ένα βασικό πλεονέκτημα της προσέγγισης αυτής είναι ότι βρίσκεται σε θέση να ανιχνεύσει νέους τύπους bot ενώ οι δυαδικοί ταξινομητές θα είναι σε θέση να διακρίνουν μόνο bot ίδιου τύπου με αυτά που χρησιμοποιήθηκαν κατά την διαδικασία εκπαίδευσης του ταξινομητή. Ο ταξινομητής που λειτούργησε βέλτιστα στην υπόθεση αυτή είναι ο Bayes με πολύ υψηλά επίπεδα επίδοσης.

Οι Shevtsov et al. [62] παρουσιάζουν μια νέα μεθοδολογία βασισμένη σε ένα πλαίσιο επιβλεπόμενης Μηχανικής Μάθησης για την αναγνώριση bot έναντι γνήσιων χρηστών στο Twitter. Συγκεκριμένα το προτεινόμενο σύστημα συμπεριλαμβάνει την εξαγωγή μιας πληθώρας χαρακτηριστικών τα οποία καλύπτουν από χαρακτηριστικά προφίλ και περιβάλλοντος έως χρονικά και χαρακτηριστικά αλληλεπίδρασης. Πειραματικά αποδεικνύουν ότι το βέλτιστο μοντέλο σε συνδυασμό με το σετ των χαρακτηριστικών είναι το XGBoost η αρχιτεκτονική του οποίου στηρίζεται στην ενίσχυση κλίσης (gradient boosting). Στην συνέχεια ελέγχουν την δυνατοτητα γενικευσης του στο σύνολο δεδομένων US Elections 2020 στο οποίο παρουσιάζει υψηλή επίδοση και αναλύουν την επεξηγηματικότητα του μοντέλου, αποκαλύπτωντας σημαντικές εισηγήσεις από την άποψη ανάλυσης δεδομένων του Twitter.

3.1.2 Μέθοδοι που βασίζονται σε Τεχνικές Μη-Επιβλεπόμενης Μάθησης

Η πρώτη μέθοδος μη επιβλεπόμενης μάθησης προτάθηκε το 2012 από τους Ahmed και Abulaish [63]. Πρότειναν τη χρήση του αλγορίθμου Markov clustering για τον εντοπισμό ομάδων ύποπτων λογαριασμών που προωθούν ανεπιθύμητο περιεχόμενο (spam). Αρχικά δημιούργησαν ένα μη κατευθυνόμενο, πλήρως συνδεδεμένο γράφο όπου κάθε κόμβος αντιπροσωπεύει κάποιο χρήστη. Αυτός ο γράφος περιλαμβάνει το βάρος των συνδέσμων, που είναι η ομοιότητα μεταξύ των λογαριασμών χρησιμοποιώντας χαρακτηριστικά περιεχομένου των tweet. Στη συνέχεια, εφαρμόζεται ο αλγόριθμος συσταδοποίησης των χρηστών αντιμετωπίζοντας τα bot σαν ακραίες τιμές των ομάδων (outliers of the clusters).

Το 2014 παρουσιάστηκε από τον Miller [64] μία αρχιτεκτονική μηχανικής μάθησης η οποία αποσκοπούσε κυρίως στην ανίχνευση μηνυμάτων με κακόβουλο περιεχόμενο (spams). Ο Miller αντιμετώπισε την πρόκληση της ανίχνευσης των spammers σαν ένα πρόβλημα ανίχνευσης ανωμαλιών (anomaly detection problem). Αναλυτικότερα εφάρμοσε αλγορίθμους ομαδοποίησης (Clustering) στους αληθινούς χρήστες και θεώρησε τους spammers σαν ακραίες τιμές (outliers) δηλαδή χρήστες οι οποίοι εμφανίζουν απόκλιση από την προσδοκώμενη συμπεριφορά. Για τον σκοπό αυτό συνδύασε δύο αλγορίθμους clustering τον DENstream και τον StreamKM++ πάνω στα ακολουθιακά δεδομένα. Η μηχανική χαρακτηριστικών περιελάμβανε κάποια απλά αριθμητικά χαρακτηριστικά που προέκτυπαν από το API του Twitter ενώ τα υπόλοιπα 95 προέκυπταν από την επεξεργασία των tweets μέσω του κώδικα ASCII.

Ο Chavoshi [65] παρατήρησε ότι είναι πολύ απίθανο να αναδημοσιεύσουν πολλοί χρήστες ένα μήνυμα σε λιγότερο από 20 δευτερόλεπτα μετά τη δημοσίευσή του, ή να δημοσιεύσουν ένα μήνυμα με σχετικό τρεντινγκ θέμα το ίδιο ακριβώς δευτερόλεπτο. Για να το αποδείξει αυτό, αξιοποίησε μία τεχνική κατά την οποία λαμβάνεται υπόψη η καθυστέρηση (lag-sensitive hashing technique) και μια μέτρηση αντιστοίχισης ανεξάρτητη από τον χρόνο (warping-invariant correlation measure) προκειμένου να οργανώσουν τους χρήστες σε ομάδες ανορθοδόξα συσχετισμένων λογαριασμών. Η κύρια υπόθεση βασίζεται στο ότι χρήστες οι οποίοι συσχετίζονται ανορθόδοξα είναι πολύ πιθανό να χειρίζονται απο αυτοματοποιημένα προγράμματα. Τα αποτελέσματα έδειξαν ότι, όταν οι λογαριασμοί χρησιμοποιούνται για να δημοσιεύσουν σχετικά με το ίδιο τρεντινγκ θέμα σε μικρό χρονικό πλαίσιο, εντοπίζονται ομάδες bot για το Twitter με υψηλή ακρίβεια.

3.2 Deep Learning Methods

Σήμερα στην εποχή των Big Data [66] έχουμε στην διάθεσή μας έναν τεράστιο όγκο δεδομένων τα οποία χαρακτηρίζονται από μεγάλη ένταση και ποικιλομορφία. Ιδιαίτερα όσον αφορά την ανίχνευση των bot στο Twitter υπάρχουν τεράστια και πλήρη σύνολα δεδομένων τα οποία επιτρέπουν την επεξεργασία τους με τεχνικές βαθιάς μάθησης [67]. Βασικό χαρακτηριστικό των τεχνικών αυτών είναι ότι δεν απαιτούν καμία μηχανική εξαγωγής χαρακτηριστικών. Οι αλγόριθμοι αυτοί μπορούν να απορροφούν και να επεξεργάζονται μη δομημένα δεδομένα, όπως κείμενο και εικόνες, αυτοματοποιώντας την διαδικασία εξαγωγής μοτίβων. Έχουν προταθεί αρκετές μεθοδολογίες οι οποίες προσπαθούν να αντιμετωπίσουν αποδοτικά το ζήτημα της ανίχνευσης bot στο Twitter οι αρχιτεκτονικές των οποίων μπορούν να διαιρεθούν στις ακόλουθες κατηγορίες: μέθοδοι που χρησιμοποιούν επαναλαμβανόμενα νευρωνικά δίκτυα (RNN), μέθοδοι που χρησιμοποιούν νευρωνικά συνελικτικά δίκτυα σε γράφους (GCN) και μέθοδοι οι οποίοι χρησιμοποιούν transformers..

3.2.1 Μέθοδοι που χρησιμοποιούν RNN

Οι Kudugunta et al. [68] χρησιμοποίησαν ένα βαθύ αναδρομικό νευρωνικό δίκτυο Long Short Term Memory (LSTM) ούτως ώστε να αξιοποιήσουν την πληροφορία που εμπεριέχεται στα tweets με μεγαλύτερη εκφραστικότητα. Αναλυτικότερα με την χρήση του LSTM βρίσκει ακολουθιακές δομές στο εσωτερικό των tweets καταφέρνοντας έτσι να σκιαγραφήσει το μοτίβο συγγραφής του κάθε χρήστη. Μία πολύ σημαντική συνεισφορά του είναι το γεγονός ότι απέδειξε πως μπορούμε να αναπτύξουμε μοντέλα με υψηλή απόδοση αξιοποιώντας ένα ελάχιστο πλήθος χαρακτηριστικών. Επιπλέον είναι η πρώτη τεχνική που προτάθηκε η οποία κάνει πρόβλεψη σε επίπεδο των tweets. Επειδή τα dataset που είχε στην διάθεσή του αποτελούνταν από χιλιάδες χρήστες αλλά από εκατομμύρια tweets αποφάσισε να αντιστοιχίσει τις ετικέτες στα tweets. Η αρχιτεκτονική που προτείνεται καλείται ContextualLSTM. Αναλυτικότερα κωδικοποιεί τα tweets μέσω του προεκπαιδευμένου μοντέλου GLoVE και στην συνέχεια τα προωθεί σε ένα LSTM δύο επιπέδων. Το τελευταίο βήμα είναι αυτό που χαρακτηρίζει την αρχιτεκτονική ως Contextual καθώς ενσωματώνει στην τελική ταξινόμηση και την πληροφορία του λογαριασμού των χρηστών.

Οι Wei et al. [69] πρότειναν μία αρχιτεκτονική η οποία αξιοποιεί ένα αμφίδρομο LSTM (Bidirectional LSTM-BiLSTM). Αμφίδρομο σημαίνει ότι λειτουργούν παράλληλα δύο κρυφά επίπεδα για την ίδια έξοδο ούτως ώστε να αξιοποιείται τόσο η παρελθοντική όσο και η μελλοντική πληροφορία. Για την εύρεση των embeddings των tweets χρησιμοποιείται το προεκπαιδευμένο μοντέλο GLoVe και στην συνέχεια τροφοδοτούνται στο BiLSTM τριών επιπέδων. Η αρχιτεκτονική αυτή είναι η πρώτη η οποία δεν ενσωματώνει τα χαρακτηριστικά των χρηστών από το API του Twitter στην τελική ταξινόμηση.

Οι Feng et al. [70] πρότειναν το SATAR που είναι η πρώτη μέθοδος αυτο-επιβλεπόμενης μάθησης. Δεν απαιτεί κανέναν μηχανισμό εξαγωγής χαρακτηριστικών αλλά κωδικοποιεί την

σημασιολογική και τοπολογική πληροφορία μέσω προεκπαιδευμένων μοντέλων. Η συνολική αρχιτεκτονική απαρτίζεται από τέσσερα κομμάτια: Στο πρώτο κομμάτι εφαρμόζουν LSTM με μηχανισμούς προσοχής για να αναπαραστήσουν τόσο την σημασιολογική πληροφορία μεταξύ των tweet των χρηστών (tweet level) αλλά και σε επίπεδο λέξεων για να αναπαραστήσουν την λεξιλογική του συμπεριφορά (word level). Στο επόμενο επίπεδο αξιοποιεί τα χαρακτηριστικά των χρηστών που προκύπτουν από το API του Twitter ενώ στο τρίτο μέρος δημιουργεί ένα δίκτυο (Γράφος) με συνδέσεις ακολούθησης και εξάγει την τοπολογική πληροφορία μέσω κάποιον μετασχηματισμών, χωρίς την χρήση νευρωνικών δικτύων. Τέλος τα πολυτροπικά χαρακτηριστικά συνδυάζονται μέσω του Co-influence Aggregator το οποίο υπολογίζει την μεταξύ τους συσχέτιση μέσω παραμέτρων προσοχής. Η συνολική αρχιτεκτονική εκπαιδεύεται με αυτο-επιβλεπόμενη μάθηση. Δηλαδή το συνολικό μοντέλο εκπαιδεύεται πρώτα πάνω στο σήμα του #πλήθους_Followers και στην συνέχεια βελτιστοποιεί τις παραμέτρους πάνω στο ζητούμενο της ανίχνευσης των Bot. Το μεγάλο πλεονέκτημα της συγκεκριμένης μεθόδου είναι ότι εμφανίζει ανταγωνιστικά επίπεδα απόδοσης ενώ παράλληλα γενικεύει (generalization) και προσαρμόζεται (adaptation) πολύ καλά στον πραγματικό κόσμο του Twitter.

3.2.2 Μέθοδοι που χρησιμοποιούν GCN

Οι Alhosseini et al. [71] πρότειναν την πρώτη αρχιτεκτονική η οποία αξιοποιεί νευρωνικά δίκτυα σε γράφους για να συμπεριλάβει την τοπολογική πληροφορία του γράφου. Αναλυτικότερα εφάρμοσε συνελικτικά νευρωνικά δίκτυα σε γράφους (Graph Convolutional Networks - GCN) τα οποία εκμεταλλεύονται τόσο τα χαρακτηριστικά του χρήστη όπως αυτά προκύπτουν απο το API του Twitter αλλά και την πληροφορία από την δομή του γράφου.Ο γράφος έχει σαν συνδέσεις μόνο τις σχέσεις ακολούθου ενώ για το αρχικό διάνυσμα του κάθε χρήστη χρησιμοποιούνται 7 low-level χαρακτηριστικά. Το GCN που χρησιμοποιείται έχει 2-επίπεδα. Η αξιοποίηση τεχνητών νευρωνικών δικτύων σε γράφους πάνω στα κοινωνικά δίκτυα έφερε μία επανάσταση λόγω της μεγάλης εκφραστικότητας που έχουν στο να αναπαριστούν την τοπολογική δομή του δικτύου. Παράλληλα συνοδεύονται από πολύ υψηλές επιδόσεις και μικρή πολυπλοκότητα.

Οι Feng et al. [20] πρότειναν μία αρχιτεκτονική η οποία βελτιώνει την αντίστοιχη προσέγγιση του [71] και εμφανίζει μεγαλύτερη πληρότητα. Αναλυτικότερα αναβάθμισε τον γράφο ώστε να εμπερικλείει δεδομένα που ανταποκρίνονται πλησιέστερα στις ρεαλιστικές συνδέσεις που αναπτύσσονται στα μέσα κοινωνικής δικτύωσης μεταξύ των χρηστών. Βασίζεται λοιπόν σε έναν πλέον κατευθυνόμενο ετερογενή γράφο οι κόμβοι του οποίου αντικατοπτρίζουν τους χρήστες και οι συνδέσεις που αναπτύσσονται είναι δύο: followers και followings. Το διάνυσμα των χρηστών-κόμβων περικλείει διάφορα χαρακτηριστικά που περιγράφουν τον λογαριασμό των χρηστών αλλά και το περιεχόμενο των tweets. Για να επεξεργαστεί τον γράφο αποδοτικά Εφάρμοσε Συσχετιστικά Συνελικτικα Νευρωνικά Δίκτυα σε Γράφους 2 επιπέδων(Relational Graph Convolutional Networks- RGCN) .

Οι Lei et al. [72] πρότειναν μία μέθοδο η οποία επιτρέπει την βαθιά και γόνιμη αλληλεπίδραση μεταξύ των αρχιτεκτονικών επεξεργασίας κειμένου και γράφου οι οποίες προηγουμένως λειτουργούσαν απλά συμπληρωματικά χωρίς καμία συσχέτιση. Επιπλέον εντοπίζει την

σημασιολογική συνέπεια μεταξύ των tweets του κάθε χρήστη. Το BIC μοντέλο αποτελείται από M επίπεδα καθένα από τα οποία περιλαμβάνει δύο λειτουργικές μονάδες: (i) την αλληλεπίδραση μεταξύ κειμένου και γράφου και (ii) την ανίχνευση σημασιολογικής συνέπειας. Στο πρώτο κομμάτι λοιπόν το δίκτυο μαθαίνει τη σχετική σημασία των δύο μορφών κειμένου και γράφου εφαρμόζοντας αλληλεπιδραστικές αναπαραστάσεις με μηχανισμούς προσοχής. Το δεύτερο κομμάτι χρησιμοποιεί τα βάρη προσοχής από το προεκπαιδευμένο μοντέλο γλώσσας RoBERTa για να ανιχνεύσει τις αλληλοεξαρτήσεις και τη συνέπεια μεταξύ των αναρτήσεων στο Twitter. Ο συνδυασμός αυτός παρουσιάζει πολύ υψηλά επίπεδα επίδοσης και ταυτόχρονα μεγάλη εκφραστικότητα.

3.2.3 Μέθοδοι που χρησιμοποιούν transformers

Οι Garcia-Silva et al. [73] πρότειναν την πρώτη μέθοδο η οποία αντιμετώπιζε το ζήτημα της ανίχνευσης bot με χρήση του δικτύου transformers. Αναλυτικοτερα παρατήρησαν ότι η βέλτιστη ρύθμιση παραμέτρων σε γεννητικούς transformers (finetuning in generative transformers) μπορεί να οδηγήσει σε σημαντική βελτίωση των επιδόσεων του μοντέλου. Ανέλυσαν λοιπόν τα αρχιτεκτονικά στοιχεία των μοντέλων BERT_base και GPT και μελετησαν το αποτέλεσμα της ρύθμισης των παραμέτρων στις κρυφες τους καταστάσεις και στις αναπαραστάσεις εξόδου. Στην ουσία προσθέτουν ένα γραμμικό επίπεδο πάνω στην τελευταία κρυφή κατάσταση του τεκμηρίου ταξινόμησης για να εκπαιδεύσουνε έναν ταξινομητή softmax. Αναλύοντας τις κρυφές καταστάσεις και των δύο μοντέλων στην εργασία ανίχνευσης bot, διαπίστωσαν ότι οι κρυφές καταστάσεις του GPT συνέβαλαν στην εκμάθηση πιο ακριβών ταξινομητών από το BERT, σε όλα τα επίπεδα. Επιπλέον ανέδειξαν ότι οι κρυφές καταστάσεις του BERT κωδικοποιούν λιγότερες γραμματικές πληροφορίες από το προεκπαιδευμένο μοντέλο, ιδιαίτερα στα τελικά επίπεδα, ενώ τα γλωσσικά εμφυτεύματα ομογενοποιούνται αρκετά μετά το fine tuning.

Λίγο αργότερα οι Martín Gutiérrez et al. [74] μελέτησαν τεχνικές μεταφοράς μάθησης μέσω ισχυρών μοντέλων NLP όπως οι Transformers, για την εξαγωγή συμπαγών πολυγλωσσικών αναπαραστάσεων των χαρακτηριστικών βασισμένων σε κείμενα που σχετίζονται με τους λογαριασμούς των χρηστών όπως η περιγραφή τους και τα tweets. Η βασική τους υπόθεση ήταν να ξεπεράσουν τους περιορισμούς που εθεταν προηγούμενες μελέτες πάνω στην ανάπτυξη αντιπροσωπευτικών διανυσματικών γλωσσικών αναπαραστάσεων από πολυγλωσσικά κείμενα. Πειραματίστηκαν με διάφορους συνδυασμούς λεκτικών εμφυτευμάτων, εμφυτευμάτων εγγράφων και μοντέλων Transformers(BERT και RoBERTa), προκειμένου να βρουν την βέλτιστη κωδικοποίηση του κειμένου, τα οποία συνδύασαν με μεταδεδομένα των χρηστών σε ένα πυκνό νευρωνικό δίκτυο που ονομάστηκε Bot-DenseNet. Από τα αποτελέσματά τους αποδείχτηκε ότι το βέλτιστο μοντέλο όσον αφορά την απόδοση και την απλότητά του αλλά και την δημιουργία ανθεκτικών γλωσσικών αναπαραστάσεων ήταν το συνδυαζόμενο με το RoBERTa.

Οι Feng et al. [75] πρότειναν ένα καινοτόμο πλαίσιο ανίχνευσης bot στο Twitter η αρχιτεκτονική του οποίου βασίζεται στην εφαρμογή των transformers σε γράφους. Εκμεταλλεύεται την τοπολογική δομή του πραγματικού κόσμου του Twitter και την ποικιλομορφία των αλληλεπιδράσεων κατασκευάζοντας έναν ετερογενή γράφο με τους χρήστες ως κόμβους και

πολυμορφες σχέσεις ως ακμές. Αναλυτικότερα οι σχέσεις με τις οποίες πειραματίστηκε είναι οι σχέσεις ακολούθησης (‘follower’,’following’) , οι σχέσεις αναδημοσίευσης tweet(retweet) και ο αποκλεισμός άλλων χρηστών (block). Στη συνέχεια, αξιοποιούν τους transformers (Relational Graph Transformers-RGT) για να μοντελοποιήσουν την ένταση της επιρροής και να μάθουν τις αναπαραστάσεις κόμβων. Τέλος , χρησιμοποιούν δίκτυα με μηχανισμούς προσοχής προκειμένου να γίνει η διαβίβαση των μηνυμάτων κατά μήκος του γράφου και να καταλήξουμε στην τελική ταξινόμηση. Η μέθοδος αυτή εμφάνισε πρωτοφανή επίπεδα ακρίβειας και επιπλέον ήταν η πρώτη που αναβάθμισε τον γράφο σε σημείο που προσομοιάζει τον πραγματικο κόσμο του Twitter και τις σχέσεις που δημιουργούνται.

Οι Loukas et al. [76] εισήγαγαν την πρώτη μελέτη που χρησιμοποιεί μόνο το πεδίο περιγραφής του χρήστη και τις εικόνες τριών καναλιών που υποδηλώνουν τον τύπο και το περιεχόμενο των tweets που δημοσιεύουν οι χρήστες. Αρχικά, κατασκευάζουν ψηφιακές ακολουθίες DNA, τις οποίες μετατρέπουν σε τρισδιάστατες εικόνες . Στην συνέχεια, εφαρμόζουν προ-εκπαιδευμένα μοντέλα του τομέα της όρασης, συμπεριλαμβανομένων των EfficientNet, AlexNet, VGG16, κ.λπ. Ακολούθως, προτείνετε μια πολυτροπική προσέγγιση, όπου αξιοποιείται το TwHIN-BERT για την απόκτηση της κειμενικής αναπαράστασης του πεδίου περιγραφής του χρήστη και το VGG16 για την απόκτηση της οπτικής αναπαράστασης για την εικονική πλευρά. Τέλος , πειραματίζονται με τρεις διαφορετικές μεθόδους συγχώνευσης, δηλαδή συνένωση, ενότητα πολυτροπικής πύλης (gated multimodal unit) και διασταύρωση προσοχής (crossmodal attention) μεταξύ των διαφορετικών μεθόδων και συγκρίνουν τις επιδόσεις τους. Οι προτεινόμενες προσεγγίσεις εμφάνισαν πολύτιμα πλεονεκτήματα και πολύ υψηλή ακρίβεια.

Κεφάλαιο 4

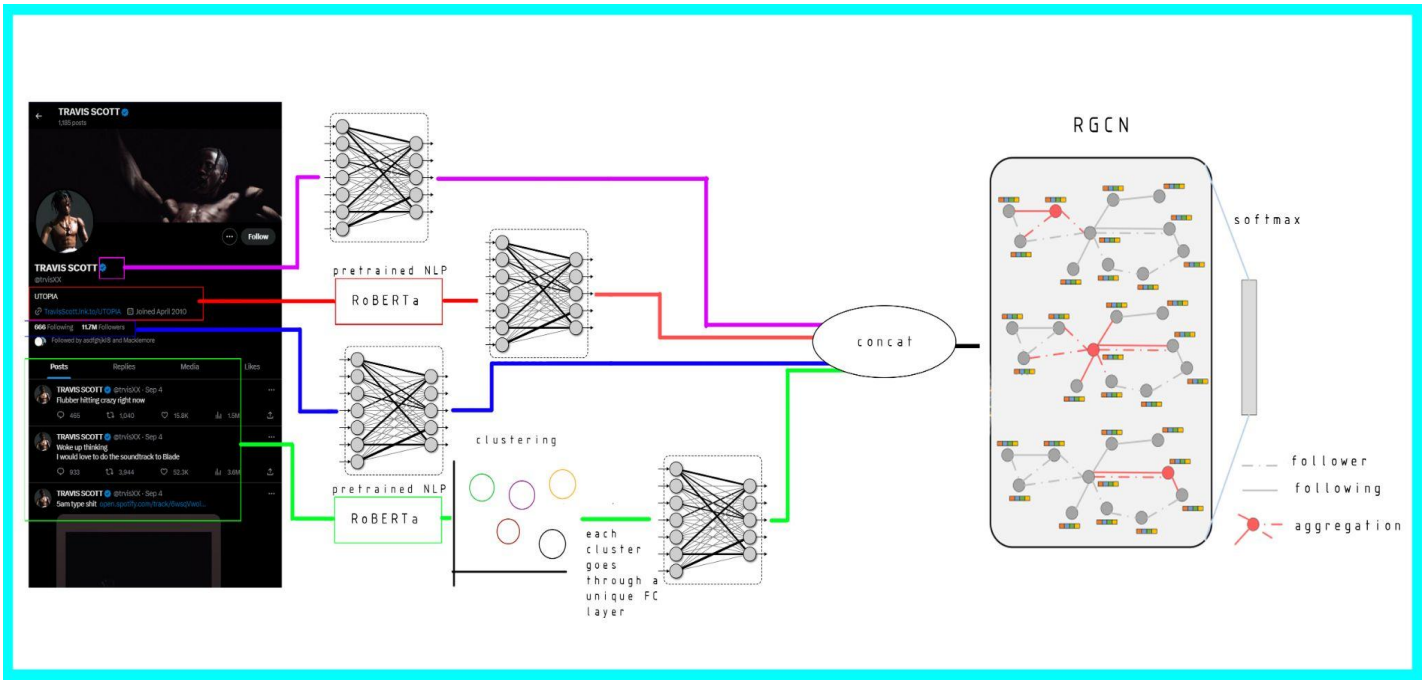
Προτεινόμενη Μέθοδος

4.1 Κύρια Ιδέα

Στην συγκεκριμένη ενότητα θα εισάγουμε και θα θεμελιώσουμε μια νέα προσέγγιση η οποία εμφανίζει ανταγωνιστική επίδοση στην ανίχνευση bot λογαριασμών στο Twitter. Η αρχιτεκτονική που προτείνουμε καλείται να αντιμετωπίσει δύο νέες προκλήσεις που έφεραν τα συνεχώς μεταβαλλόμενα μέσα κοινωνικής δικτύωσης: γενίκευση (generalization) και την τοποθέτηση της κοινωνικής δραστηριότητας των bot (community).

Η πρόκληση της γενίκευσης στην ανίχνευση των bot στα κοινωνικά μέσα απαιτεί από τους ανιχνευτές να αναγνωρίζουν ταυτόχρονα bot που επιτίθενται με πολλούς διαφορετικούς τρόπους και εκμεταλλεύονται διάφορα χαρακτηριστικά στο Twitter. Όπως περιγράψαμε στο υποκεφάλαιο 1.3 υπάρχουν διάφορες ομάδες bot οι οποίες προσπαθούν να επηρεάσουν την ακεραιότητα του Twitter αξιοποιώντας διάφορα μέσα όπως απάτες σε αναδημοσιεύσεις (retweet), κακόβουλη προώθηση hashtag και ανεπιθύμητων συνδέσμων URL, μίμηση αυθεντικών χρηστών κτλ. Επιπλέον, με την πάροδο του χρόνου και προκειμένου να αποφεύγουν τα υπάρχοντα μέτρα ανίχνευσης τα bot τείνουν να προσομοιάζουν την συμπεριφορά ρεαλιστικών χρηστών όλο και περισσότερο ενσωματώνοντας διάφορα χαρακτηριστικά όπως ψευδείς φωτογραφίες προφίλ, αυξημένες πληροφορίες προφιλ οι οποίες προκύπτουν από άλλους χρήστες κτλ. Επομένως η αρχιτεκτονική του δικτύου πρέπει να είναι ανθεκτική στην ποικιλομορφία αυτή και στην μεταμφίεση των bots ούτως ώστε να έχει ουσιαστική επίδραση στον πραγματικό κόσμο του Twitter. Παράλληλα, υπάρχουν πολλά bots τα οποία αν εξεταστούν μεμονωμένα προσομοιάζουν την συμπεριφορά ρεαλιστικών χρηστών αλλά στην πραγματικότητα λειτουργούν σε ομάδες προκειμένου να προωθήσουν τους κακόβουλους σκοπούς τους, αποφεύγοντας τα μέτρα ανίχνευσης. Επομένως, ο χαρακτηρισμός των χρηστών μέσω της κοινωνικής τους δραστηριότητας και η ενσωμάτωση των αλληλοεπιδράσεων τους αποτελεί απαραίτητο εφόδιο για έναν αποδοτικό ανιχνευτή.

Προκειμένου λοιπόν να αντιμετωπίσουμε τις προαναφερθείσες προκλήσεις προτείνουμε ένα νέο πλαίσιο το οποίο βασίζεται στο representation learning (μάθηση μέσω διανυσμάτων αναπαράστασης). Συγκεκριμένα, κωδικοποιεί ταυτόχρονα την πληροφορία που εμπεριέχεται στα tweet, την πληροφορία του χρήστη και τη γειτονική πληροφορία των χρηστών χωρίς μηχανική χαρακτηριστικών (feature engineering) για να προωθήσει τη γενίκευση της ανίχνευσης. Είναι ένα end-to-end μοντέλο που συνδυάζει μεθόδους επιβλεπόμενης και μη-επιβλεπόμενης μάθησης προκειμένου να αντιμετωπίσει αποτελεσματικά την ποικιλομορφία και την μεταβλητότητα που χαρακτηρίζουν την φύση της εργασίας. Η συνολική αρχιτεκτονική φαίνεται και στο ακόλουθο σχήμα:



Σχήμα 4.1.1: Επισκόπηση του μοντέλου μας. Οι μοβ,κόκκινες, μπλε και πράσινες μονάδες υποδηλώνουν κατηγορικά μεταδεδομένα, περιγραφή χρήστη, αριθμητικά μεταδεδομένα και tweets.

4.2 Ορισμός Προβλήματος

Στο συγκεκριμένο υποκεφάλαιο θα παρουσιάσουμε τα χαρακτηριστικά και θα θέσουμε τα θεμέλια του προβλήματος της ανίχνευσης bot λογαριασμών στο Twitter. Έστω ένας χρήστης ***U*** του Twitter, η πληροφορία του οποίου διαχωρίζεται σε τρία πλαίσια: σημασιολογική πληροφορία (semantic) ***T*** , ιδιωματική πληροφορία (property) ***P*** και πληροφορία γειτονιάς (neighborhood) ***N***. Η σημασιολογική πληροφορία περιλαμβάνει το σύνολο των M tweet κάθε χρήστη $T = \{t_i\}_{i=1}^{M}$. Κάθε

tweet $t_i = \{w_1^i, \ldots\ldots, w_{Q_i}^i\}$ περιέχει Q_i λέξεις. Η ιδιωματική πληροφορία του χρήστη $P = \{p_{num}, p_{cat}, p_{descr}\}$ περιλαμβάνει αριθμητικά και κατηγορηματικά χαρακτηριστικά όπως αυτά προκύπτουν από το API του Twitter καθώς και πληροφορία η οποία προκύπτει από την περιγραφή του χρήστη (user description). Τέλος η πληροφορία της γειτονιάς $N = \{N^f, N^t\}$ όπου $N^f = \{N_1^f, N_2^f, \ldots\ldots, N_u^f\}$ είναι οι χρήστες που ακολουθεί ο χρήστης και $N^t = \{N_1^t, N_2^t, \ldots\ldots, N_u^t\}$ είναι οι χρήστες που τον ακολουθούν. Επομένως μπορούμε να θεωρήσουμε το πρόβλημα της ανίχνευσης bot στο Twitter ως ένα πρόβλημα δυαδικής ταξινόμησης όπου κάθε χρήστης είναι είτε πραγματικός (y=0) είτε bot (y=1).

Problem: Twitter Bot Detection Given a Twitter user U and its information T, P and N, learn a bot detection function $f : f(U(T, P, N)) \rightarrow \hat{y}$, such that $\hat{y}$ approximates ground truth y to maximize prediction accuracy.

Σχήμα 4.2.1 : Ορισμός Προβλήματος

4.3 Αρχιτεκτονική Δικτύου - Μεθοδολογία

Στο συγκεκριμένο κεφάλαιο θα παρουσιάσουμε αναλυτικά την συνολική αρχιτεκτονική του μοντέλου μας. Εφόσον αξιοποιούμε και συνδυάζουμε πολυτροπικά χαρακτηριστικά διακρίνουμε τα ακόλουθα υποδίκτυα: (i)Σημασιολογικό υποδίκτυο (Semantic sub-network), (ii) Ιδιωματικό Υποδίκτυο (Property Sub-network), (iii) Υποδίκτυο Γειτονιάς (Neighborhood Sub-network)

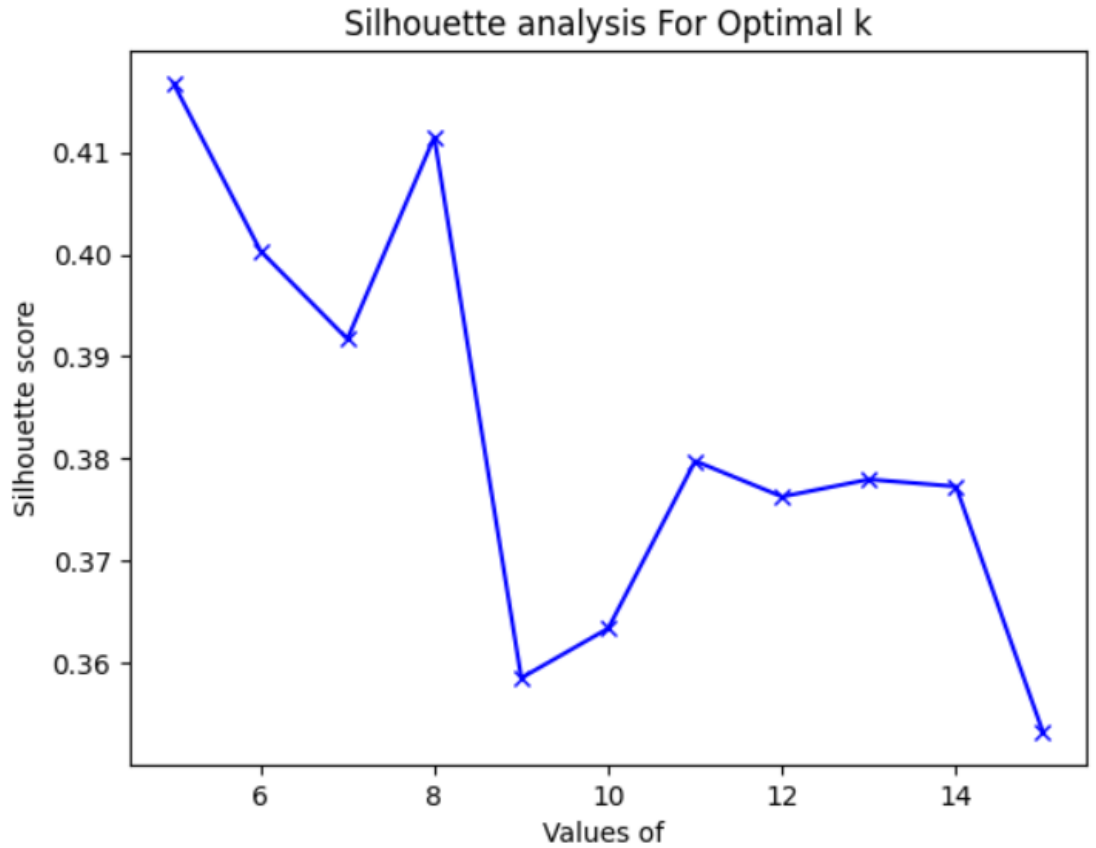
4.3.1 Semantic Sub-network

Στο συγκεκριμένο υποδίκτυο εκτελείται η επεξεργασία των tweets κάθε χρήστη. Για να κωδικοποιήσουμε τα tweets κάθε χρήστη και να βρούμε τα embeddings τους χρησιμοποιούμε το προεκπαιδευμένο μοντέλο RoBERTa. Συνεπώς αρχικά τα tweets όλων των χρηστών διέρχονται μέσω του RoBERTa προκειμένου να βρούμε την μαθηματική τους αναπαράσταση. Αναλυτικότερα κάθε λέξη αντιστοιχίζεται σε ένα μαθηματικό διάνυσμα και στην συνέχεια υπολογίζουμε την μέση τιμή τους.

$$\bar{t} = RoBERTa(\{w_i\}_{i=1}^{Qi}) \quad , \qquad \bar{t} \in \mathbb{R}^{Ds\,x\,1} \tag{4.3.1}$$

όπου $\bar{t}$ συμβολίζει την αναπαράσταση ενός tweet, Q_i το πλήθος των λέξεων που περιλαμβάνει ένα tweet και Ds είναι η διάσταση του διανύσματος των embedding.

Στην συνέχεια εκτελούμε τον αλγόριθμο συσταδοποίησης (clustering algorithm) KMeans ούτως ώστε να διαχωρίσουμε το σύνολο των tweets σε λογικές ομάδες. Καθορίζουμε το βέλτιστο πλήθος ομάδων (clusters) οι οποίες μας δίνουν την βέλτιστη διακριτότητα μεταξύ τους μέσω της ανάλυσης ΄΄σιλουέτας΄΄ (silhouette analysis). Ο συντελεστής σιλουέτας ή σκορ σιλουέτας του KMeans είναι ένα μέτρο που μετράει πόσο παρόμοιο είναι ένα δεδομένο σημείο μέσα σε μία ομάδα (cohesion) σε σύγκριση τις υπόλοιπες ομάδες (separation). Με βάση την παρακάτω γραφική βλέπουμε ότι το βέλτιστο πλήθος ομάδων είναι 5 .



Σχήμα 4.3.1 : Ανάλυση σιλουέτας για την εύρεση του βέλτιστου πλήθους ομάδων (k=5)

Συνεπώς εφαρμόζουμε τον αλγόριθμο KMeans για 5 clusters.

Για κάθε χρήστη υπολογίζουμε την μέση τιμή των tweets του ανάλογα με το cluster στο οποίο ανήκουν και τελικά τα τροφοδοτούμε σε ένα fully-connected layer, ξεχωριστό για κάθε cluster, προκειμένου να βρούμε την αναπαράστασή τους.

$$\overline{t_{ci}} = average(\{\bar{t} \in c_i\}) \ , \ \overline{t_{ci}} \in \mathbb{R}^{Ds\,x\,1} \qquad (4.3.2)$$

όπου $\overline{t_{ci}}$ η μέση τιμή των embeddings των tweets ενός χρήστη που ανήκουν στην κλάση c_i και c_i =(c_1,....,c_5) το πλήθος των κλάσεων

$$r_{t,ci} = \varphi(W_{t,ci} \cdot \overline{t_{ci}} + b_{t,ci}) \ , r_{t,ci} \in \mathbb{R}^{D/N} \qquad (4.3.3)$$

όπου $W_{t,ci}, b_{t,ci}$ είναι παράμετροι προς εκμάθηση για κάθε cluster και φ η μη γραμμική συνάρτηση ενεργοποίησης ReLU.

4.3.2 Property Sub-network

Στο συγκεκριμένο υποδίκτυο αξιοποιούμε την πληροφορία των χρηστών όπως αυτή προκύπτει απευθείας από το API του Twitter. Αναλυτικότερα έχουμε στην διάθεσή μας μία πληθώρα χαρακτηριστικών τα οποία περιγράφουν τόσο το προφίλ όσο και την δραστηριότητα των χρηστών στο Twitter. Σύμφωνα με τους Kudugunta et al. [68] η αξιοποίηση μεγάλου αριθμού τέτοιων στατικών χαρακτηριστικών όχι μόνο δεν βελτιώνει αισθητά την επίδοση του μοντέλου αλλά επιπλέον αυξάνει εκθετικά την πολυπλοκότητά του. Συνεπώς μπορούμε να σχηματίσουμε αποδοτικά μοντέλα με ένα ελάχιστο αριθμό χαρακτηριστικών. Τα χαρακτηριστικά που χρησιμοποιούμε διακρίνονται σε αριθμητικά , κατηγορηματικά και αυτά που εξάγουμε από την περιγραφή των χρηστών. Ακολούθως περιγράφουμε τον τρόπο ανάλυσής τους.

(i) Αριθμητικά χαρακτηριστικά

Τα αριθμητικά χαρακτηριστικά που επιλέγουμε προκύπτουν απευθείας από το API του Twitter χωρίς να απαιτείται κάποια μηχανική χαρακτηριστικών για την εξαγωγή τους (feature engineering) όπως φαίνεται και στον ακόλουθο πίνακα.

Πίνακας 4.3.1 : Το σύνολο των αριθμητικών χαρακτηριστικών που αξιοποιήσαμε στο μοντέλο μας

Feature Name	Description
#followers	number of followers
#followings	number of followings
#favorites	number of likes
#statuses	number of statuses
active days	number of active days
screen_name length	screen name character count

Στην συνέχεια τα κανονικοποιούμε μέσω του Z-score ούτως ώστε να έχουν μηδενική μέση τιμή και μοναδιαία διασπορά και ακολούθως τα τροφοδοτούμε σε ένα fully-connected layer για να βρούμε την αναπαράστασή τους.

$$r_{p,num} = \varphi(W_{p,num} \cdot z_{score}(p_{num}) + b_{p,num}), \quad r_{p,num} \in \mathbb{R}^{D/N} \qquad (4.3.4)$$

με $W_{p,num}$, $b_{p,num}$ παράμετροι προς εκμάθηση και φ η μη γραμμική συνάρτηση ενεργοποίησης ReLU.

(ii) Κατηγορηματικά χαρακτηριστικά

Τα κατηγορηματικά χαρακτηριστικά, παρόμοια με τα αριθμητικά, προκύπτουν απευθείας από το API του Twitter χωρίς να απαιτείται κάποια μηχανική χαρακτηριστικών για την εξαγωγή τους (feature engineering) όπως φαίνεται και στον ακόλουθο πίνακα.

Πίνακας 4.3.2 : Το σύνολο των κατηγορηματικών χαρακτηριστικών που αξιοποιήσαμε στο μοντέλο μας

Feature Name	Description
protected	protected or not
geo_enabled	enable geo location or not
verified	verified or not
contributors_enabled	enable contributors or not
is_translator	translator or not
is_translation_enabled	translation or not
profile_user_background_image	have background image or not
has_extended_profile	have extended profile or not
default_profile	the default profile
default_profile_image	the default profile image
profile_background_tile	the background tile

Χρησιμοποιούμε one-hot-encoding για να τα κωδικοποιήσουμε και στην συνέχεια τα τροφοδοτούμε σε ένα fully-connected layer για να βρούμε την αναπαράστασή τους.

$$r_{p,cat} = \varphi(W_{p,cat} \cdot p_{cat} + b_{p,cat}), \qquad r_{p,cat} \in \mathbb{R}^{D/N} \qquad (4.3.5)$$

με $W_{p,cat}$, $b_{p,cat}$ παράμετροι προς εκμάθηση και φ η μη γραμμική συνάρτηση ενεργοποίησης ReLU.

(iii) Χαρακτηριστικά που προκύπτουν από την περιγραφή προφίλ

Παρόμοια με την επεξεργασία των tweets χρησιμοποιούμε το προεκπαιδευμένο μοντέλο RoBERTa για να βρούμε τα εμφυτευματα (embeddings) της περιγραφής των χρηστών.

$$\overline{p_{desc}} = RoBERTa(\{w_{desc_i}\}_{i=1}^{Qi}) \ , \qquad \overline{p_{desc}} \in \mathbb{R}^{Ds\,x\,1} \qquad (4.3.6)$$

όπου w_{desc} είναι οι λέξεις που περιλαμβάνει η περιγραφή του κάθε χρήστη και Qi το πλήθος τους. Στην συνέχεια προωθούμε τα διανύσματα αυτά μέσω ενός fully-connected layer για να βρούμε την αναπαράστασή τους.

$$r_{p,desc} = \varphi(W_{p,dec} \cdot \overline{p_{desc}} + b_{p,desc}) , \qquad r_{p,desc} \in \mathbb{R}^{D/N} \qquad (4.3.7)$$

με $W_{p,desc}$, $b_{p,desc}$ παράμετροι προς εκμάθηση και φ η μη γραμμική συνάρτηση ενεργοποίησης ReLU.

4.3.3 Neighborhood Sub-network

Η προτεινόμενη αρχιτεκτονική είναι σχεδιασμένη στο να αξιοποιεί και να ενσωματώνει την πληροφορία που προκύπτει από την κοινωνική γειτονιά των χρηστών δηλαδή τις αλληλεπιδράσεις τους και την κοινωνική τους δραστηριότητα. Άλλωστε σύμφωνα με τον Cresci, υπάρχουν bot λογαριασμοί οι οποίοι εάν αναλυθούν ατομικά προσομοιάζουν την συμπεριφορά ρεαλιστικών χρηστών αλλά στην πραγματικότητα ενεργούν σε ομάδες για να πετύχουν κακόβουλους σκοπούς. Κατασκευάζουμε λοιπόν έναν κατευθυνόμενο ετερογενή γράφο από το δίκτυο του Twitter και ακολούθως εφαρμόζουμε Σχεσιακά Συνελικτικά Νευρωνικά Δίκτυα σε Γράφους (Relational Graph Convolutional Networks - RGCN) για να εξάγουμε τις αναπαραστασεις των χρηστών .

Στα πλαίσια του κατευθυνόμενου ετερογενή γράφου αντιμετωπίζουμε τους χρήστες σαν κόμβους (nodes) και οι συνδέσεις που αναπτύσσονται μεταξύ τους (links) είναι του ακόλουθου(follower) και αυτου που ακολουθώ (following).

$$R=\{ r1 , r2 \}=\{ \text{‘following’} , \text{‘follower’} \}$$

Εφόσον οι ακμές του γράφου είναι δύο ειδών κάθε χρήστης σχηματίζει δύο γειτονιές:
$Nr_1(u) = N^f(u)$ για την σχέση following και $Nr_2(u) = N^t(u)$ για την σχέση follower.

Προτού προχωρήσουμε στην εφαρμογή του RGCN κατασκευάζουμε το συνολικό διάνυσμα χαρακτηριστικών για κάθε χρήστη συνενώνοντας τα διάφορα είδη χαρακτηριστικών σε ένα κοινό διάνυσμα το οποίο στην συνέχεια τροφοδοτούμε σε ένα πολυεπίπεδο perceptron (MLP) για να βρούμε την συνολική αναπαράστασή τους.

$$r = concat(r_{p,num}, r_{p,cat}, r_{p,desc}, r_{t,c0}, \ldots\ldots, r_{t,cN}) \quad (4.3.8)$$

$$x_i^{(0)} = \varphi(W_1 \cdot r_i + b_1), \qquad x_i^{(0)} \in \mathbb{R}^{D\times 1} \quad (4.3.9)$$

με W_1, b_1 παράμετροι προς εκμάθηση και φ η μη γραμμική συνάρτηση ενεργοποίησης ReLU. Στην συνέχεια εφαρμόζουμε το l-th επίπεδο του RGCN.

$$x_i^{(l+1)} = \Theta_{self} \cdot x_i^{(l)} + \sum_{r\in R} \sum_{j\in N_r(i)} \frac{1}{|N_r(i)|} \Theta_r \cdot x_j^{(l)}, \qquad x_i^{(l+1)} \in \mathbb{R}^{D\times 1} \quad (4.3.10)$$

όπου Θ είναι ο πίνακας προβολής (projection matrix). Μετά απο L επίπεδα RGCN μετασχηματίζουμε την αναπαράσταση των χρηστών μέσω ενός MLP(Multi layer Perceptron).

$$h_i = \varphi(W_2 \cdot x_i^{(l)} + b_2), \qquad h_i \in \mathbb{R}^{D\times 1} \quad (4.3.11)$$

με W_2, b_2 παράμετροι προς εκμάθηση και h_i η τελική αναπαράσταση των χρηστών .

4.4 Εκπαίδευση και Βελτιστοποίηση

Προκειμένου να διεξάγουμε την τελική ταξινόμηση των χρηστών σε πραγματικούς ή bot βασισμένοι στις αναπαραστάσεις των χρηστών που προέκυψαν απο το R-GCN , εφαρμόζουμε ένα fully-connected layer με συνάρτηση ενεργοποίησης την εκθετική (softmax layer).

$$\hat{y}_i = softmax(W_0 \cdot h_i + b_0) \quad (4.4.1)$$

με W_0, b_0 παράμετροι προς εκμάθηση.

Η συνάρτηση κόστους που χρησιμοποιούμε είναι το Σφάλμα Διασταυρούμενης Εντροπίας (Cross Entropy Loss) που διέπεται από την ακόλουθη μαθηματική σχέση και ώς αλγόριθμο βελτιστοποίησης επιλέξαμε τον Adam.

$$L = -\sum_{i\in Y} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] - \lambda \sum_{w\in\theta} w^2 \quad (4.4.2)$$

όπου y_i είναι η πραγματική ετικέτα των χρηστών, και θ το σύνολο των υπο εκμάθηση παραμέτρων που έχει το δίκτυο.

Κεφάλαιο 5

Πειράματα και Αποτελέσματα

5.1 Σύνολο δεδομένων Twibot-20

Η έρευνά μας διεξάγεται πάνω σε ένα ολοκληρωμένο σύνολο δεδομένων που καλείται Twibot-20 [77]. Το Twibot-20 αποτελεί σημείο αναφοράς όλων των ερευνητικών εργασιών πάνω στην ανίχνευση bot λογαριασμών από το 2020 και μετέπειτα λόγω της πληρότητας που το χαρακτηρίζει. Περιλαμβάνει 229.580 χρήστες Twitter, 33.488.192 tweets και 33.716.171 συνδέσεις και καλύπτει σχεδόν όλες τις ποικιλίες bot που ενεργούν στα κοινωνικά δίκτυα. Παράλληλα είναι το μοναδικό υψηλής ποιότητας σύνολο δεδομένων που παρέχει πληροφορίες γραφημάτων ούτως ώστε οι αρχιτεκτονικές που προτείνουμε να είναι εφαρμόσιμες. Ακολουθούμε τις ίδιες διαιρέσεις που προβλέπονται στο Twibot-20, ώστε τα αποτελέσματα μας να είναι άμεσα συγκρίσιμα με προηγούμενες εργασίες.

Σύμφωνα λοιπόν με τις διαιρέσεις που ορίζει το σύνολο δεδομένων έχουμε: το train_set για την εκπαίδευση του μοντέλου, το dev_set για να παρακολουθούμε την διαδικασία της εκπαίδευσης και να βελτιστοποιήσουμε τις παραμέτρους του μοντέλου, το test_set για να αξιολογούμε το μοντέλο και το support_set το οποίο περιλαμβάνει χρήστες χωρίς ετικέτες (unlabeled) οι οποίοι απλά πλαισιώνουν τον γράφο. Αναλυτικές λεπτομέρειες για τα υποσύνολα αυτά φαίνονται στις ακόλουθες εικόνες.

Info of the df_train

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8278 entries, 0 to 8277
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ID        8278 non-null   int64
 1   profile   8278 non-null   object
 2   tweet     8223 non-null   object
 3   neighbor  7524 non-null   object
 4   label     8278 non-null   int64
dtypes: int64(2), object(3)
```

Info of the df_dev

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2365 entries, 0 to 2364
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ID        2365 non-null   int64
 1   profile   2365 non-null   object
 2   tweet     2350 non-null   object
 3   neighbor  2141 non-null   object
 4   label     2365 non-null   int64
dtypes: int64(2), object(3)
```

Σχήμα 5.1.1: Αναλυτικές πληροφορίες για το περιεχόμενο των συνόλων δεδομένων df_train και df_dev

Info of the df_test

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1183 entries, 0 to 1182
Data columns (total 5 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ID        1183 non-null   int64
 1   profile   1183 non-null   object
 2   tweet     1173 non-null   object
 3   neighbor  1074 non-null   object
 4   label     1183 non-null   int64
dtypes: int64(2), object(3)
```

Info of the df_support

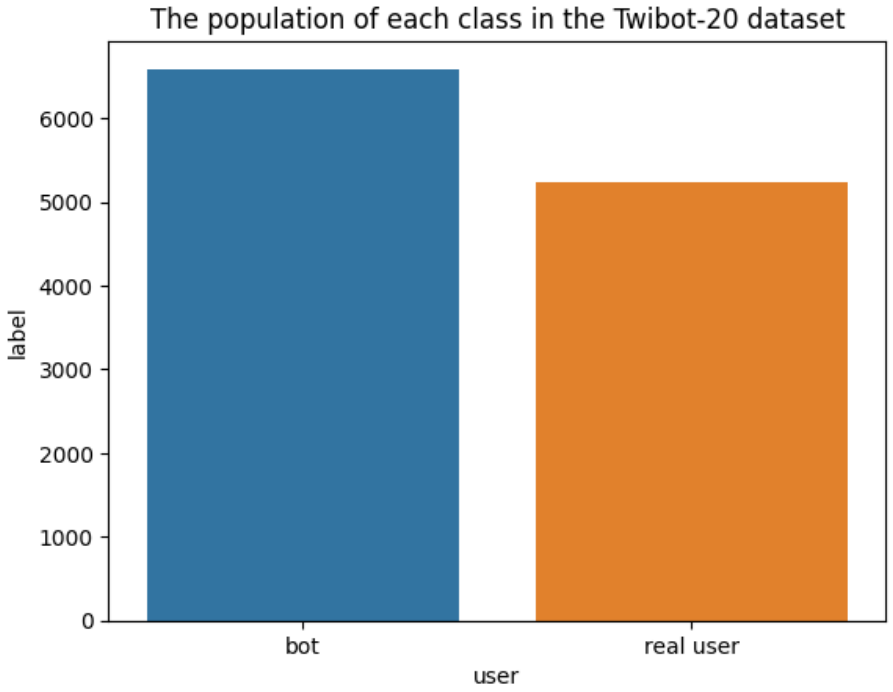
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 217754 entries, 0 to 217753
Data columns (total 5 columns):
 #   Column    Non-Null Count   Dtype
---  ------    --------------   -----
 0   ID        217754 non-null  object
 1   profile   217746 non-null  object
 2   tweet     194297 non-null  object
 3   neighbor  1185 non-null    object
 4   label     0 non-null       object
dtypes: object(5)
```

Σχήμα 5.1.2: Αναλυτικές πληροφορίες για το περιεχόμενο των συνόλων δεδομένων df_test και df_support

Ακολούθως παραθέτουμε κάποια παραδείγματα χρηστών ούτως ώστε να έχουμε μία πληρέστερη επισκόπηση του συνόλου δεδομένων και των τύπων των χαρακτηριστικών που παρέχει. Να σημειώσουμε ότι στην στήλη ‘profile’ ανήκουν μία πληθώρα χαρακτηριστικών από τα οποία εμείς επιλέγουμε τα αριθμητικά, κατηγορηματικά χαρακτηριστικά καθώς και την περιγραφή του χρήστη. Στην στήλη ‘neighbor’ περιέχεται η πληροφορία για την κοινωνική γειτονιά του κάθε χρήστη δηλαδή τόσο οι ακόλουθοί του όσο και αυτους που ακολουθεί πράγμα που επιτρέπει την δημιουργία του ετερογενούς γραφήματος.

ID	profile	tweet	neighbor
1630890068	{'id': '1630890068 ', 'id_str': '1630890068 ',...	[@sethgoldberg17 @jaysonst Fan interference? I...	{'following': ['237453978', '462581299', '1706...
9580757536768	{'id': '713519580757536769 ', 'id_str': '71351...	[@C130Matt I think I heard a voice from out in...	{'following': ['36991422', '32567081', '133983...
93345260	{'id': '93345260 ', 'id_str': '93345260 ', 'na...	[@savage_esquire That's unfuckingbelievable.\n...	{'following': ['714636670268792832', '23341114...
1749309397	{'id': '1749309397 ', 'id_str': '1749309397 ',...	[@Jomboy_ Doesn't want to pull anymore Hammys\...	{'following': ['3124065581', '413364940', '211...
50471224	{'id': '50471224 ', 'id_str': '50471224 ', 'na...	[The sports card market is unreal right now. P...	{'following': ['4202878276', '637216245', '129...

Σχήμα 5.1.3: Τα χαρακτηριστικά που περιέχει το σύνολο δεδομένων για τους πρώτους πέντε χρήστες



Σχήμα 5.1.4: Bar_plot που αναδεικνύει το πλήθος των ρεαλιστικών χρηστών και των bot στο σύνολο δεδομένων

5.2 Αποτελέσματα και Επίδοση Μοντέλου

Εκπαιδεύουμε λοιπόν το μοντέλο μας και αξιολογούμε την επίδοσή του πάνω στο σύνολο δεδομένων Twibot-20. Αν θεωρηθούν ως True Positives τα δείγματα που ταξινομήθηκαν ορθώς σε μια κλάση, ως False Positives τα δείγματα που ταξινομήθηκαν σε μια κλάση αλλά εσφαλμένα, ως False Negatives τα δείγματα που έπρεπε να ταξινομηθούν σε μια κλάση αλλά δεν το έκαναν

και ως True Negatives τα δείγματα που ορθώς δεν ταξινομήθηκαν σε μια κλάση, τότε η αξιολόγηση της επίδοσης του μοντέλου γίνεται με βάση τις κάτωθι μετρικές:

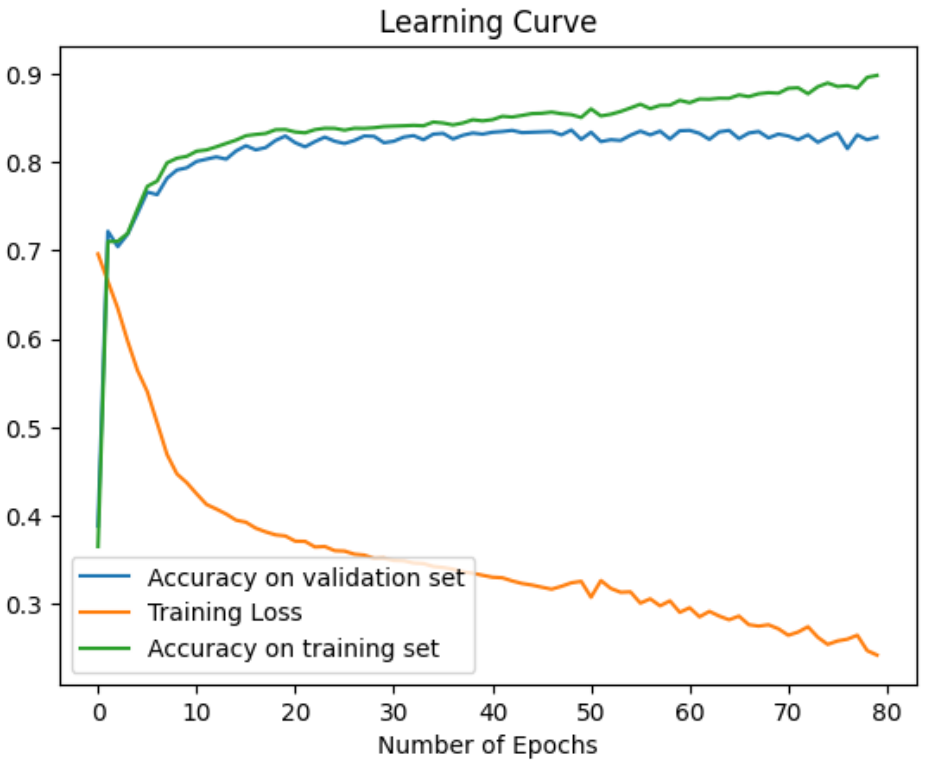
$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5.2.1)$$

$$Precision = \frac{TP}{TP + FP} \quad (5.2.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (5.2.3)$$

$$F1\ score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5.2.4)$$

Παρουσιάζουμε αρχικά την γραφική καμπύλη εκμάθησης ούτως ώστε να οπτικοποιήσουμε την ροή της εκπαίδευσης.



Σχήμα 5.2.1: Η καμπύλη εκμάθησης του μοντέλου μας.

Η καμπύλη αυτή αναδεικνύει τον τρόπο με τον οποίο μεταβάλλονται το κόστος εκπαίδευσης (training_loss), η ακρίβεια πάνω στο σύνολο εκπαίδευσης (training_accuracy) και η ακρίβεια πάνω στο σύνολο επικύρωσης(validation_accuracy). Είναι σημαντικό να τονίσουμε πως το μοντέλο που αναμένεται να γενικεύει βέλτιστα στο test set είναι αυτό με την μεγαλύτερη ακρίβεια στο σύνολο επικύρωσης. Όπως φαίνεται στο διάγραμμα τόσο το training_loss όσο και το training_accuracy τείνουν να βελτιώνονται συνεχώς κατά την πορεία της εκπαίδευσης . Παρόλα αυτά το επιθυμητό μοντέλο δεν είναι αυτό με τις ελάχιστες προαναφερθείσες μετρικές καθώς τότε θα είχαμε υπερεκπαιδευτεί πάνω στα δεδομένα εκπαίδευσης και παράλληλα θα είχαμε χαμηλή επίδοση πάνω στο σύνολο δοκιμών. Έτσι λοιπόν κατά την διαδικασία της εκπαίδευσης αποθηκεύουμε κάθε φορά το μοντέλο με το υψηλότερο validation accuracy το οποίο φαίνεται να προσεγγίζουμε μέσα σε περίπου 60 εποχές.

Στην συνέχεια παρουσιάζουμε την επίδοση του μοντέλου μας και την συγκρίνουμε με τις ακόλουθες υλοποιήσεις:

• Οι Miller et al. [64] εξάγουν 107 χαρακτηριστικά από τα tweet ενός χρήστη και τα μεταδεδομένα τους και πραγματοποιούν ανίχνευση bot στο Twitter ως ανίχνευση ανωμαλίας.

• Οι Cresci et al. [55] χρησιμοποιούν μία ψηφιακή σειρά DNA για να κωδικοποιήσουν την ακολουθία της online δραστηριότητας ενός χρήστη.

• Το Botometer [56] είναι ένα δημόσια διαθέσιμο εργαλείο που εκμεταλλεύεται περισσότερα από χίλια χαρακτηριστικά.

• Το SATAR [70] προτείνει ένα αυτοεπιβλεπόμενο μοντέλο αναπαράστασης με την κοινή αξιοποίηση πολυτροπικών χαρακτηριστικών του χρήστη. Στη συνέχεια, κατηγοριοποιεί τα bot με λεπτομέρεια.

• Οι Kugugunta et al. [68] προτείνουν μια αρχιτεκτονική που εκμεταλλεύεται συνδυαστικά τα tweet ενός χρήστη και τις πληροφορίες του λογαριασμού.

• Οι Wei et al. [69] προτείνουν ένα μοντέλο ανίχνευσης bot μέσω ενός BiLSTM τριών επιπέδων για την κωδικοποίηση των tweet.

• Οι Alhosseini et al. [71] χρησιμοποιούν GCN για την μάθηση των αναπαραστάσεων των χρηστών και την κατηγοριοποίηση των bot.

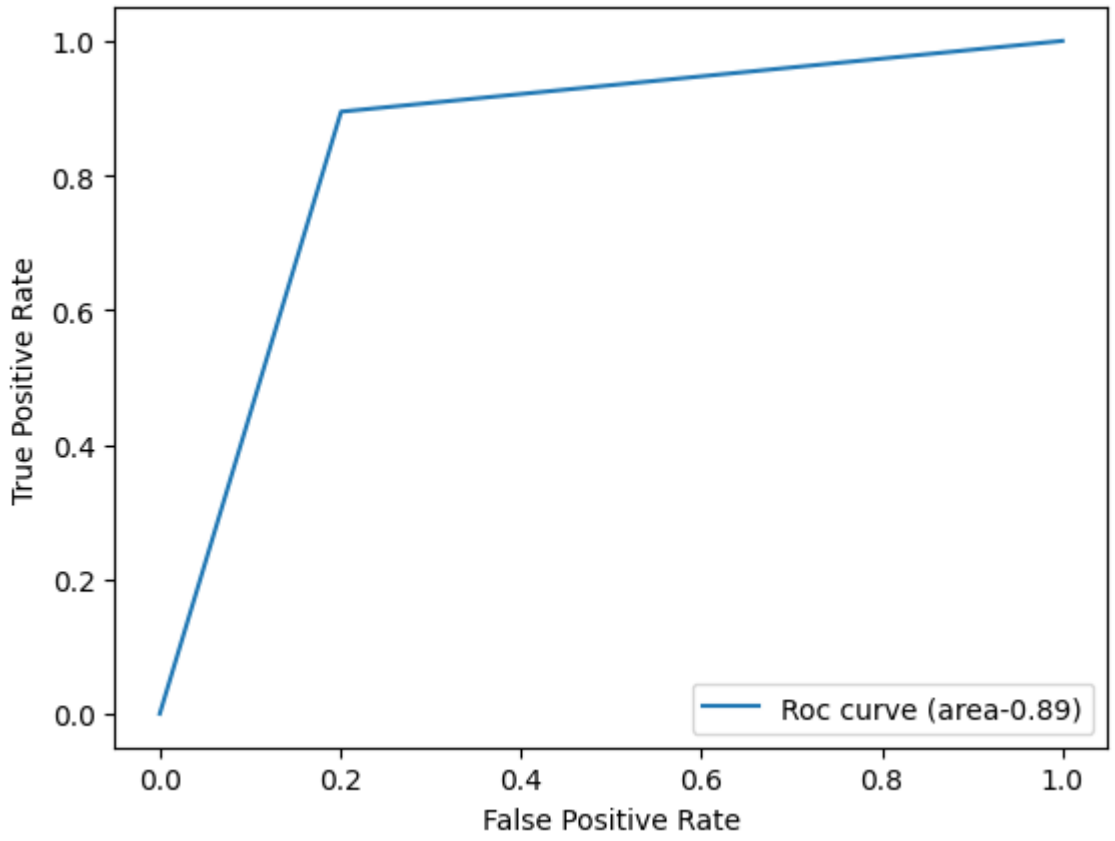
• Το BotRGCN [20] δημιουργεί μία αρχιτεκτονική βασισμένη σε Σχεσιακά Νευρωνικά Δίκτυα σε γράφους και αξιοποιεί από κοινού τα tweet των χρηστών και τρία είδη μεταδεδομένων.

• Οι Lei et al. [72] πρότειναν το BIC μοντέλο το οποίο αναπτύσσει μία βαθιά αλληλεπίδραση μεταξύ κειμένου και γράφου και ανιχνεύει την σημασιολογική συνέπεια μεταξύ των tweet των χρηστών.

• Το RGT [75] εκμεταλλεύεται το γράφο σχέσεων και επιρροής για να πραγματοποιήσει ανίχνευση bot.

Πίνακας 5.2.1: Οι μετρικές αξιολόγησης του μοντέλου μας.

Evaluation Metrics	Values(%)
Accuracy	85.55
Precision	81.31
Recall	95.16
F1-score	87.69



Σχήμα 5.2.2: Η καμπύλη Roc του μοντέλου μας.

Πίνακας 5.2.2: Πίνακας σύγκρισης των αποτελεσμάτων μας με ήδη υπάρχουσες μεθόδους.

Method	Accuracy	F1-Score	Precision	Recall
Miller et al.	64.50	74.81	60.71	97.44
Cresci et al.	47.76	13.69	7.66	64.47
BotOrNot	53.09	55.13	55.67	50.82
Kudugunta et al.	59.59	47.26	80.40	33.47
Alhossini et al.	59.92	72.09	57.83	95.72
BiLSTM	70.23	53.61	62.74	46.83
BotRGCN	83.27	85.26	81.39	89.53
SATAR	84.02	86.07	81.50	91.22)
RGT	86.57	88.01	85.15	91.06
BIC	87.36	88.88	84.76	93.44
My_Model	85.55	87.69	81.31	95.16

Από τον παραπάνω πίνακα ο οποίος συγκρίνει την επίδοση του μοντέλου μας με άλλα σύγχρονα μοντέλα καταλήγουμε σε κάποια συμπεράσματα:

- Το μοντέλο μας παρουσιάζει ανταγωνιστική επίδοση στην ανίχνευση bot λογαριασμών συγκριτικά με άλλες σύγχρονες μεθόδους. Αναλυτικότερα σημειώνει την τριτη καλύτερη επίδοση ξεπερνώντας όλες τις υπάρχουσες μεθόδους εκτός από το BIC μοντέλο και το RGT.

- Συγκριτικά με μεθόδους οι οποίες βασίζονται σε στατικά χαρακτηριστικά τα οποία προκύπτουν μέσω μηχανικής εξαγωγής χαρακτηριστικών (feature engineering) όπως του Cresci και BotOrNot σημειώνει σαφώς ανώτερη επίδοση. Αυτό δίνει έμφαση στο γεγονός ότι παραδοσιακά μοντέλα μηχανικής μάθησης τα οποία βασίζονται στην μηχανική χαρακτηριστικών αδυνατούν να πλαισιώσουν την ποικιλομορφία που χαρακτηρίζει το ζήτημα της ανίχνευσης bot.

- Συγκριτικά με το BotRGCN, το οποίο χρησιμοποιεί ομοίως Σχεσιακά Συνελικτικά Νευρωνικά Δίκτυα σε Γράφους για να αφομοιώσει την κοινωνική πληροφορία κάθε χρήστη, εμφανίζει σημαντική βελτίωση. Ειδικότερα το μοντέλο μας ξεπερνάει το BotRGCN

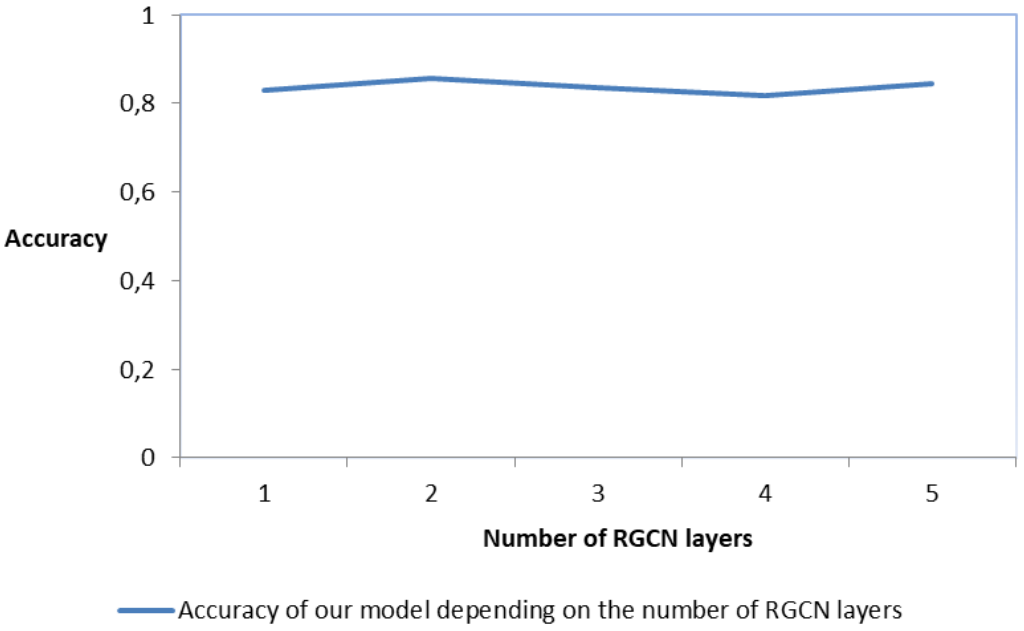
σε όλες τις μετρικές εμφανίζοντας 2,3% βελτίωση στην ακρίβεια (accuracy) γεγονός που αναδεικνύει την αποδοτικότητα και την καθοριστική συνεισφορά του αλγορίθμου συσταδοποίησης (clustering) στην επεξεργασία των tweet.

- Τέλος, συγκριτικά με μεθόδους οι οποίες βασίζονται σε ανατροφοδοτούμενα νευρωνικά δίκτυα(RNN) και ιδιαίτερα σε LSTM όπως του Kudugunta [68] και του Alhosseini [71] εμφανίζει πάλι ανώτερες επιδόσεις πράγμα που θεμελιώνει την συνεισφορά των τεχνητών νευρωνικών δικτύων σε γράφους (GNN's) στην αναβάθμιση του μοντέλου. Ιδιαίτερη προσοχή αξίζει να δώσουμε στο μοντέλο SATAR το οποίο βασίζεται πάλι σε LSTM αλλά εμφανίζει ανώτερη επίδοση από το BotRGCN το οποίο βασίζεται όπως προαναφέραμε σε τεχνητά νευρωνικά δίκτυα σε γράφους. Το μοντέλο μας ξεπερνάει και το SATAR δίνοντας έμφαση για άλλη μία φορά στην μεγάλη αποδοτικότητα του συνδυασμού των δύο μεθόδων: Clustering και GNN's.

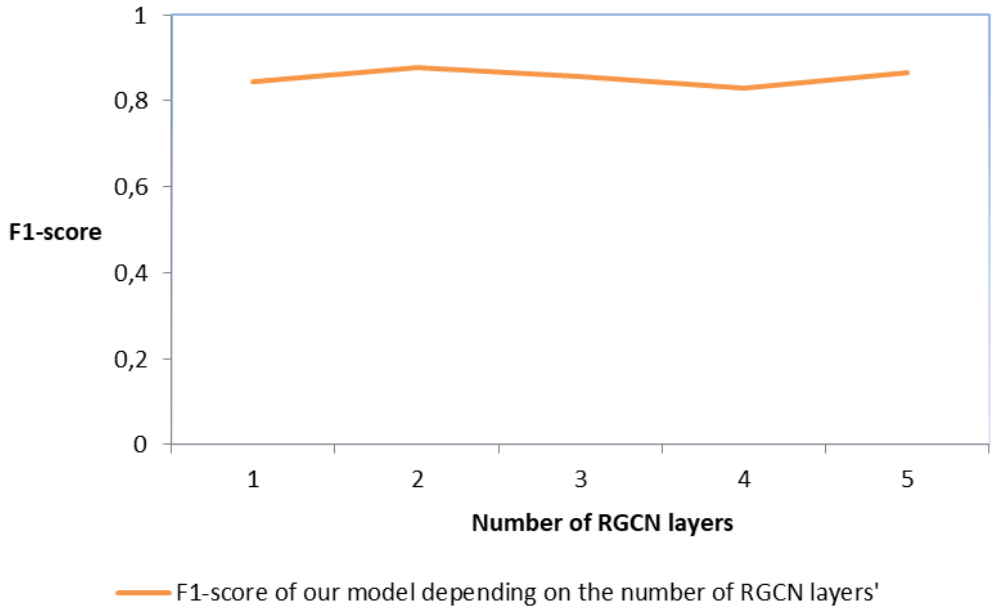
5.3 Πειράματα

Στο συγκεκριμένο υποκεφάλαιο παρουσιάζουμε την πειραματική διαδικασια που ακολουθήσαμε για να ορίσουμε κάποιες υπερπαραμέτρους του μοντέλου αλλα και για να διερευνήσουμε περαιτέρω τον τρόπο λειτουργίας του.

Αρχικά πειραματιζόμαστε με διαφορετικό αριθμό επιπέδων RGCN και ελέγχουμε την επίδρασή τους στην συνολική απόδοση του μοντέλου. Γνωρίζουμε πως ο αριθμός των επιπέδων ορίζει το εύρος της γειτονιάς του κάθε χρήστη. Για παράδειγμα με 1-layer ο χρήστης αξιοποιεί μόνο πληροφορία απο τους απευθείας γείτονές του ενώ για 2-layer αξιοποιείται επιπλέον η πληροφορία από τους γείτονες των γειτόνων. Συνεπώς ο συνολικός αριθμός επιπέδων πρέπει να κινηθεί σε μικρές τιμές αλλιώς υπάρχει ο κίνδυνος αλλοίωσης της πληροφορίας του γράφου και ομογενοποίησης όλων των χρηστών-κόμβων. Τα αποτελέσματά μας φαίνονται στο ακόλουθο διάγραμμα σύμφωνα με το οποίο έχουμε βέλτιστη απόδοση αν χρησιμοποιήσουμε ακολουθιακά 2 επίπεδα RGCN τα οποία μας προσφέρουν παράλληλα μικρό αριθμό παραμέτρων και άρα μικρή πολυπλοκότητα.



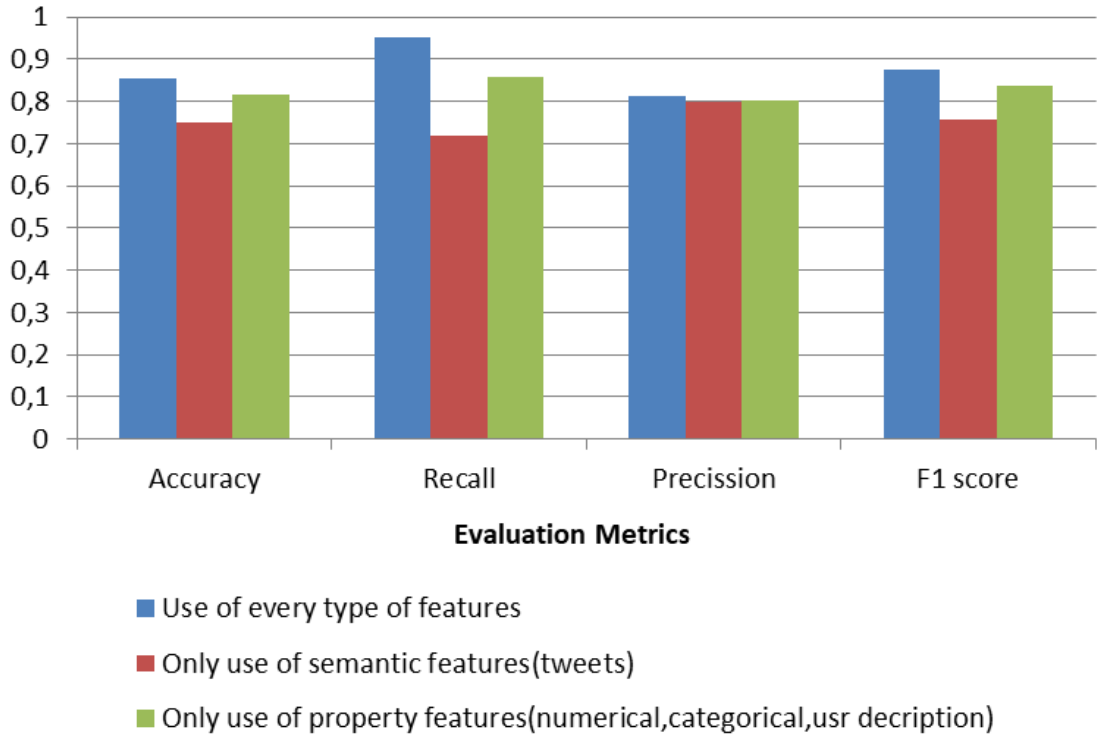
Σχήμα 5.3.1: Διάγραμμα ακρίβειας (accuracy) του μοντέλου μας συναρτήση των επιπέδων RGCN



Σχήμα 5.3.2: Διάγραμμα της μετρικής F1-score του μοντέλου μας συναρτήση των επιπέδων RGCN

Στην συνέχεια διερευνούμε εάν πράγματι η συνολική κωδικοποίηση πολυτροπικών χαρακτηριστικών για κάθε χρήστη βελτιώνει το μοντέλο και οδηγεί σε πιο ανθεκτικούς ταξινομητές.

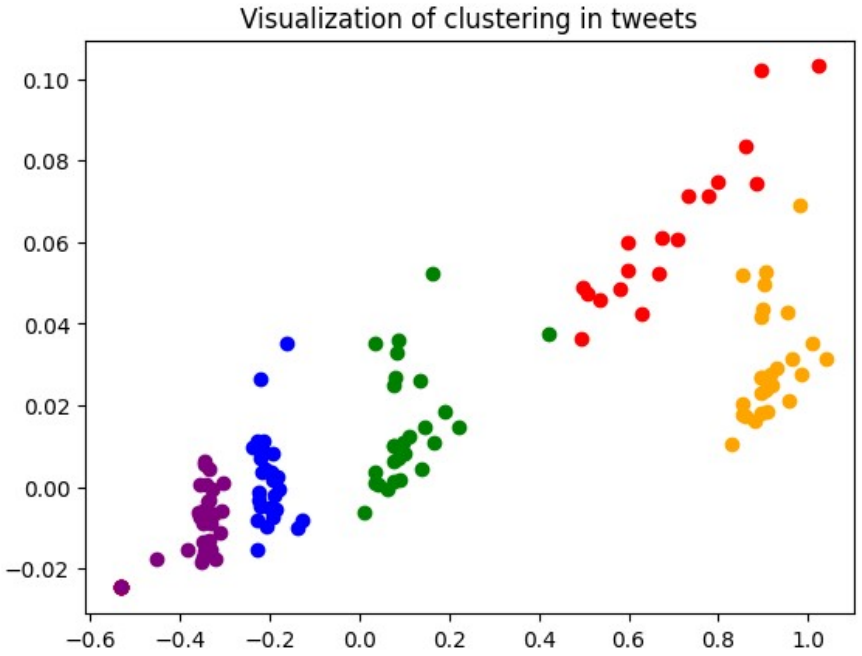
Εκπαιδεύουμε λοιπόν μοντέλα με μειωμένα σετ χαρακτηριστικών και παρουσιάζουμε την επίδοσή τους στο ακόλουθο διάγραμμα. Όπως είναι εμφανές επιβεβαιώνετε ότι κάθε κομμάτι της πληροφορίας του χρήστη είναι απαραίτητο για την ανάπτυξη μοντέλων με υψηλή απόδοση.



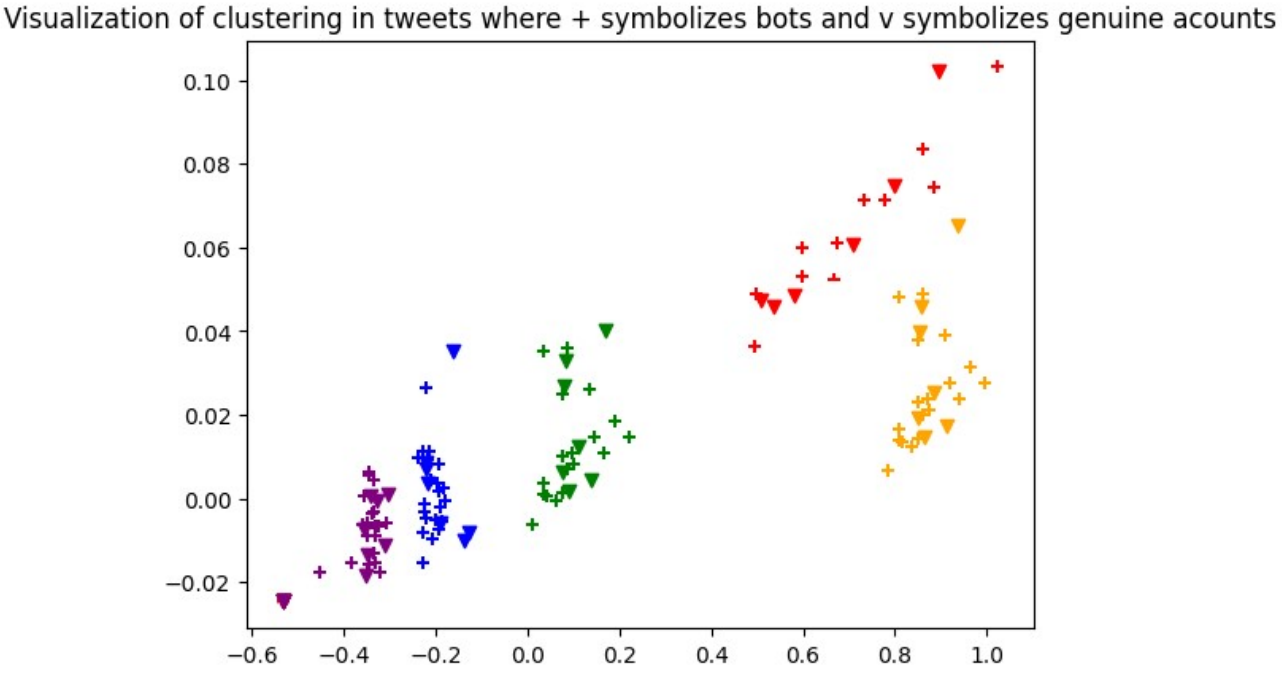
Σχήμα 5.3.3: Bar plot που παρουσιάζει τις μετρικές του μοντέλου μας για διαφορετικό συνδυασμό χαρακτηριστικών

Στην συνέχεια οπτικοποιούμε την λειτουργία του αλγορίθμου συσταδοποίησης KMeans που έχουμε εφαρμόσει στα tweets. Γνωρίζουμε ότι αποτελεί αναπόσπαστο κομμάτι της αρχιτεκτονικής του δικτύου, βελτιώνοντας την συνολική επίδοσή του εφόσον παρόμοιες αρχιτεκτονικές οι οποίες βασίζονται σε RGCN αλλά απλά περιορίζονται στην κωδικοποίηση των tweets μέσω ενός προεκπαιδευμένου μοντέλου επεξεργασίας φυσικής γλώσσας (όπως το BotRGCN) παρουσιάζουν κατώτερη επίδοση. Μάλιστα το παραπάνω συμπέρασμα επιβεβαιώνεται και από το Σχήμα 5.3.6 όπου συγκρίνουμε την επίδοση του μοντέλου μας με την αντίστοιχη επίδοση που θα είχαμε εάν είχαμε αφαιρέσει το clustering. Διαισθητικά μέσω του clustering χωρίζουμε το σύνολο των tweets σε 5 θεματικές κλάσεις και για κάθε χρήστη βρίσκουμε την μέση άποψή του για κάθε θεματική ομάδα. Η ασυνέπεια που χαρακτηρίζει τα bot εφόσον συνήθως περιορίζονται σε κάποιο θεματικό εύρος προκειμένου να πετύχουν τους κακόβουλους στόχους τους καθιστά την παραπάνω μέθοδο αποδοτική. Επιπλέον σύμφωνα με τους Hansen και Larsen [78] η χρήση μη επιβλεπόμενων μεθόδων όπως το clustering συμβάλλει καθοριστικά στην αναβάθμιση του επιπέδου γενίκευσης του μοντέλου και στην δυνατότητα επέκτασης του σε νέα δεδομένα. Παρακάτω παραθέτουμε τις

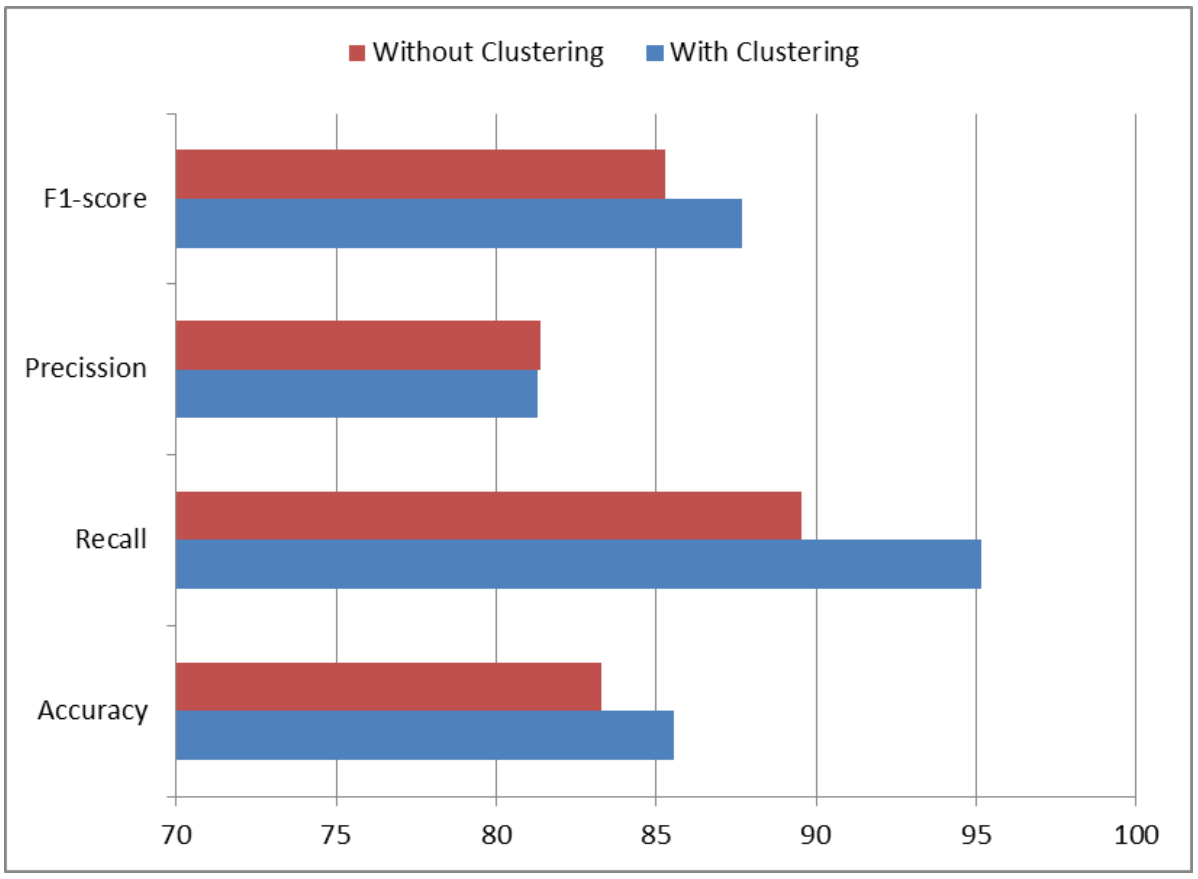
γραφικές οι οποίες οπτικοποιούν την λειτουργία αλγορίθμου συσταδοποίησης KMeans. Ιδιαίτερα στην δεύτερη γραφική φαίνεται να υπάρχει διακριτότητα μεταξύ των bot και των ρεαλιστικών χρηστών.



Σχήμα 5.3.4: Οπτικοποίηση της συσταδοποίησης (clustering) στα tweets



Σχήμα 5.3.5: Οπτικοποίηση της συσταδοποίησης (clustering) στα tweets όπου ο '+' συμβολίζει τα bots και ο 'v' τους ρεαλιστικούς χρήστες.



Σχήμα 5.3.6: Bar plot των μετρικών αξιολόγησης του μοντέλου μας με χρήση Clustering αλλά και χωρίς

Πίνακας 5.3.1: Ρύθμιση των Υπερπαραμέτρων του μοντέλου μας

Hyperparameters	Values
number of RGCN layers	2
graph module input size	416
text module input size	768
epochs	100
early stop epoch	68
dropout	0.5
learning rate	$1e^{-4}$

L2 regularization	$1e^{-5}$
optimizer	RAdamW

Κεφάλαιο 6

Επίλογος

6.1 Σύνοψη και Συμπεράσματα

Συνοψίζοντας η ανίχνευση bot λογαριασμών στα μέσα κοινωνικής δικτύωσης προσελκύει όλο και μεγαλύτερη προσοχή. Είναι ένα ζήτημα αυξανόμενης σημασίας και βαρύτητας το οποίο καλείται να προστατεύσει την ανεξαρτησία και την αξιοπιστία των μέσων κοινωνικής δικτύωσης. Η έρευνά μας επικεντρώνεται γύρω από το Twitter, μία κυρίαρχη πλατφόρμα κοινωνικής δικτύωσης που ασκεί άμεση κοινωνική επιρροή και διαμορφώνει την κοινή γνώμη. Η θεμελιώδης θέση του και η κοινωνική του δύναμη οδήγησε στην εμφάνιση bot λογαριασμών οι οποίοι προσπαθούν να εκμεταλλευτούν τους πόρους του για να πετύχουν ιδιοτελής σκοπούς.

Όπως προαναφέραμε, τα bots είναι λογαριασμοί οι οποίοι προκύπτουν από αυτοματοποιημένα λογισμικά. Προκειμένου να αποκτήσουμε βαθύτερη γνώση στον τρόπο λειτουργίας του παρουσιάσαμε τα βασικά χαρακτηριστικά τους αλλά και τα μέσα που χρησιμοποιούν στις επιθέσεις τους. Η απαιτητικότητα της ανίχνευσής τους στηρίζεται στην ποικιλομορφία και στην εξελισσιμότητά τους. Διαφορετικά είδη Bot αξιοποιούν διαφορετικά χαρακτηριστικά και μέσα του Twitter προκειμένου να προωθήσουν την ατζέντα τους. Κάποιοι τύποι επιθέσεων είναι οι ακόλουθοι: διανέμουν κακόβουλους συνδέσμους, προσποιούνται τους κοινωνικούς φίλους σε χρήστες για να αποσπάσουν επικίνδυνες και ζημιογόνες πληροφορίες, αναδημοσιεύουν ειδήσεις με μεροληπτικό περιεχόμενο προκειμένου να επηρεάσουν την κοινή γνώμη κτλ. Παράλληλα τα bot εξελίσσονται συνέχεια ώστε να ξεπερνούν τα είδη υπάρχοντα μέτρα ανίχνευσης αλλά και για να αναβαθμίσουν την αληθοφάνεια τους και ως επακόλουθο να αυξήσουν την επιρροή τους. Με τον καιρό λοιπόν γίνονται όλο και πιο ευφυή, προσομοιώνοντας την συμπεριφορά ρεαλιστικών χρηστών.

Ακολούθως προτείνουμε το δικό μας μοντέλο το οποίο συνδυάζει τεχνικές επιβλεπόμενης και μη επιβλεπόμενης μάθησης προκειμένου να ανιχνεύει τα bot στο Twitter. Έχοντας κάνει πρώτα μία εκτενή ανάλυση των ήδη υπαρχουσών μεθόδων οι αρχιτεκτονικές των οποίων βασίζονται είτε στην μηχανική χαρακτηριστικών είτε στην βαθιά μάθηση προσπαθούμε να αναπτύξουμε ένα πιο ανθεκτικό μοντέλο το οποίο θα ξεπερνά τις αδυναμίες που εμφανίζουν οι πρότερες δουλειές. Η αρχιτεκτονική του δικτύου μας συνδυάζει πολυτροπικά χαρακτηριστικά για κάθε χρήστη ενσωματώνοντας την πληροφορία από την κοινωνική του δραστηριότητα. Αναλυτικότερα κωδικοποιούμε τα tweets κάθε χρήστη μέσω του προεκπαιδευμένου μοντέλου φυσικής επεξεργασίας γλώσσας RoBERTa και στην συνέχεια εφαρμόζουμε έναν αλγόριθμο clustering για να τα χωρίσουμε σε θεματικές ενότητες. Ακολούθως συνδυάζουμε τα σημασιολογικά χαρακτηριστικά που αναφέραμε προηγουμένως με άλλα χαρακτηριστικά τα οποία προκύπτουν απευθείας απο το API του Twitter, χωρίς μηχανική χαρακτηριστικών, και περιγράφουν τον λογαριασμό των χρηστών. Τροφοδοτούμε το σύνολο των χαρακτηριστικών σε ενα Σχεσιακό Συνελικτικό Νευρωνικό Δίκτυο 2- επιπέδων (RGCN) το οποίο ενεργεί πάνω σε έναν ετερογενή γράφο που καλύπτει τις σχέσεις ακολούθησης. Τέλος, εκτελούμε διάφορα πειράματα τα οποία αναδεικνύουν την αποδοτικότητα του μοντέλου μας καθώς και την ανταγωνιστική του επίδοση. Παράλληλα, αποτυπώνουν την αποτελεσματικότητα του συνδυασμού clustering και GNNs στην αναβάθμιση του μοντέλου.

6.2 Μελλοντικές επεκτάσεις

Το έργο της παρούσας διπλωματικής αφήνει περιθώρια για μελλοντικές επεκτάσεις. Ενδεικτικά αναφέρονται οι ακόλουθες κατευθύνσεις που μπορούν να αποτελέσουν σημείο αναφοράς για επόμενους ερευνητές:

- Ο ετερογενής γράφος που κατασκευάζουμε καλύπτει μόνο τις σχέσεις ακολούθησης. Οι συνδέσεις που αναπτύσσονται μεταξύ των χρηστών-κόμβων είναι του ακόλουθου (‘follower’) και αυτή των χρηστών που ακολουθούν (‘following’). Παρόλα αυτά για να αναπαριστούμε πλήρως την κοινωνική δραστηριότητα κάθε χρήστη πρέπει ο γράφος να χαρακτηρίζεται από μεγαλύτερη πληρότητα. Με άλλα λόγια μπορεί να καλύπτει και άλλες σχέσεις που αναπτύσσονται μεταξύ των χρηστών όπως το retweet, οι αναφορές (‘mention’) κτλ. Έτσι η αναβάθμιση του γράφου θα οδηγήσει σε μία πληρέστερη εικόνα της δραστηριότητας των χρηστών.

- Στην αρχιτεκτονική που προτείνουμε εφαρμόζουμε ένα αλγόριθμο συσταδοποίησης (clustering) στα tweets ώστε να τα χωρίσουμε σε θεματικές ενότητες και στην συνέχεια αφού υπολογίσουμε την μέση τιμή τους τα προωθούμε στον γράφο στο διάνυσμα χαρακτηριστικών του κάθε χρήστη. Παρόλα αυτά μπορεί να αναπτυχθεί μία πιο βαθιά και ουσιώδης αλληλεπίδραση μεταξύ clustering και GNNs ώστε να χρησιμοποιήσουμε στο έπακρο τα οφέλη τους. Ως μελλοντική επέκταση λοιπόν ενδείκνυται η ανάπτυξη ενός μοντέλου που θα συνδέει σε βάθος τις δύο μεθόδους με διαδραστικές αναπαραστάσεις.

- Η μεθοδολογία που προτείναμε δύναται να εφαρμοστεί και σε άλλα μέσα κοινωνικής δικτύωσης όπως Facebook και Instagram. Μάλιστα ενδείκνυται η ανάπτυξη ενός μοντέλου

εκπαιδευμένο πάνω και στα τρία κυρίαρχα μέσα κοινωνικής δικτύωσης ώστε να εξαχθεί ένας καθολικός ανιχνευτής bot λογαριασμών .

- Απαραίτητο εφόδιο που πρέπει να πληρεί κάθε μοντέλο μηχανικής μάθησης είναι να εμφανίζει υψηλά επίπεδα εξηγησιμότητας (explainability) ούτως ώστε να έχουμε μία βαθιά επισκόπηση της εσωτερικής λειτουργίας του και παράλληλα να υπάρχει η δυνατότητα ενδελεχούς ανάλυσης των αδυναμιών του προς ανάπτυξη μοντέλων με μεγαλύτερη αποδοτικότητα. Για τον λόγο αυτό ενδείκνυται ο συνδυασμός του μοντέλου με μεθόδους εξηγησιμότητας (explainability methods) όπως LIME και Grad-CAM [79],[80].

- Εξετάζοντας το πώς λαμβάνονται πολλές αποφάσεις από ανθρώπους, σπανίως εξαρτώνται από ένα μόνο σημείο δεδομένων ή μια μοναδική πηγή πληροφοριών. Αυτό νοηματοδοτεί την έννοια της συνένωσης πληροφοριών από διάφορες πηγές στο πλαίσιο της πολυτροπικής μηχανικής μάθησης. Στην μεθοδολογία που προτείνουμε απλά συνενωνουμε τα πολυτροπικά χαρακτηριστικά των χρηστών και βρίσκουμε την κοινή αναπαράστασή τους με την χρήση νευρωνικών δικτύων. Έτσι, ενδείκνυται η ανάπτυξη ενός πολυτροπικού μοντέλου με διάφορες μεθόδους και επίπεδα συνένωσης των πηγών πληροφορίας ώστε να υπάρχει μία πιο ουσιώδη και βαθιά από κοινού αξιοποίησή τους [81], [82], [83], [84], [85], [86].

Παράρτημα Α

Βιβλιογραφία

[1] Hess, A. (2001). *Emile Durkheim, Georg Simmel and Ferdinand Tönnies: Social Differentiation and Functionalist Sociology* (pp. 36–49). https://doi.org/10.1057/9780230629219_4

[2] Walker, K. N., MacBride, A., & Vachon, M. L. S. (1977). Social support networks and the crisis of bereavement. *Social Science & Medicine (1967)*, *11*(1), 35–41. https://doi.org/https://doi.org/10.1016/0037-7856(77)90143-3

[3] Ilias, L., & Askounis, D. (2023). *Multitask learning for recognizing stress and depression in social media*.

[4] Ilias, L., Mouzakitis, S., & Askounis, D. (2023). Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media. *IEEE Transactions on Computational Social Systems*, 1–12. https://doi.org/10.1109/TCSS.2023.3283009

[5] Geiger, R. S. (2016). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society*, *19*(6), 787–803. https://doi.org/10.1080/1369118X.2016.1153700

[6] Ahsan, M., & Sharma, T. (2020). Spams classification and their diffusibility prediction on Twitter through sentiment and topic models. *International Journal of Computers and Applications*, *44*, 1–11. https://doi.org/10.1080/1206212X.2020.1758430

[7] Forelle, M., Howard, P. N., Monroy-Hernández, A., & Savage, S. (2015). Political Bots and the Manipulation of Public Opinion in Venezuela. *ArXiv*, *abs/1507.07109*. https://api.semanticscholar.org/CorpusID:3441751

[8] Grimme, C., Preuss, M., Clever, L., & Trautmann, H. (2017). Social Bots: Human-Like by Means of Human Control? *Big Data*, *5*. https://doi.org/10.1089/big.2017.0044

[9] Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The Rise of Social Bots. *Commun. ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

[10] Wang, A. H. (2010). Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach. In S. Foresti Sara and Jajodia (Ed.), *Data and Applications Security and Privacy XXIV* (pp. 335–342). Springer Berlin Heidelberg.

[11] Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., Quinn, S. C., & Dredze, M. (2018). Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate. *American Journal of Public Health*, *108*(10), 1378–1384. https://doi.org/10.2105/AJPH.2018.304567

[12] Abokhodair, N., Yoo, D., & McDonald, D. (2015). *Dissecting a Social Botnet*. 839–851. https://doi.org/10.1145/2675133.2675208

[13] Ratkiewicz, J., Conover, M., Meiss, M., Goncalves, B., Flammini, A., & Menczer, F. (2021). Detecting and Tracking Political Abuse in Social Media. *Proceedings of the International AAAI Conference on Web and Social Media*, *5*(1), 297-304. https://doi.org/10.1609/icwsm.v5i1.14127

[14] Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2018). Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter. *ACM Transactions on the Web*, *13*. https://doi.org/10.1145/3313184

[15] Cresci, S., Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2017, September). *The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race*. https://doi.org/10.1145/3041021.3055135

[16] Elyashar, A., Fire, M., Kagan, D., & Elovici, Y. (2014). Guided Socialbots: Infiltrating the Social Networks of Specific Organizations' Employees. *Ai Communications*, *29*. https://doi.org/10.3233/AIC-140650

[17] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, *21*(11). https://doi.org/10.5210/fm.v21i11.7090

[18] Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., Zhu, L., Ferrara, E., Flammini, A., Menczer, F., Waltzman, R., Stevens, A., Dekhtyar, A., Gao, S., Hogg, T., Kooti, F., Liu, Y., Varol, O., Shiralkar, P., … Huang, T. (2016). The DARPA Twitter Bot Challenge. *CoRR*, *abs/1601.05140*. http://arxiv.org/abs/1601.05140

[19] Varol, O., Ferrara, E., Davis, C., Menczer, F., & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. *Proceedings of the International AAAI Conference on Web and Social Media*, *11*. https://doi.org/10.1609/icwsm.v11i1.14871

[20] Feng, S., Wan, H., Wang, N., & Luo, M. (2022). 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*, 236–239. https://doi.org/10.1145/3487351.3488336

[21] Samuel, A. L. (1967). Some Studies in Machine Learning Using the Game of Checkers. *IBM J. Res. Dev.*, *44*, 206–227. https://api.semanticscholar.org/CorpusID:2126705

[22] Cunningham Pádraig and Cord, M. and D. S. J. (2008). Supervised Learning. In P. Cord Matthieu and Cunningham (Ed.), *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval* (pp. 21–49). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

[23] Barlow, H. B. (1989). Review: *Neural Comput.*, *1*(3), 295–311. https://doi.org/10.1162/neco.1989.1.3.295

[24] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

[25] Hady Mohamed Farouk Abdel and Schwenker, F. (2013). Semi-supervised Learning. In M. and J. L. C. Bianchini Monica and Maggini (Ed.), *Handbook on Neural Information Processing* (pp. 215–239). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36657-4_7

[26] Kanal, L. N. (2003). Perceptron. In *Encyclopedia of Computer Science* (pp. 1383–1385). John Wiley and Sons Ltd.

[27] ROSENBLATT F. (1958). The perceptron: a probabilistic model for information storage and organization the brain . *Psychological review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

[28] Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14), 2627–2636. https://doi.org/https://doi.org/10.1016/S1352-2310(97)00447-0

[29] Janocha, K., & Czarnecki, W. M. (2017). On Loss Functions for Deep Neural Networks in Classification. *CoRR*, *abs/1702.05659*. http://arxiv.org/abs/1702.05659

[30] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations* (pp. 318–362). MIT Press.

[31] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, *86*, 2278–2324. https://doi.org/10.1109/5.726791

[32] Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR*, *abs/1609.04747*. http://arxiv.org/abs/1609.04747

[33] Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2019). A Comprehensive Survey on Transfer Learning. *CoRR*, *abs/1911.02685*. http://arxiv.org/abs/1911.02685

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

[35] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *CoRR*, *abs/1512.03385*. http://arxiv.org/abs/1512.03385

[36] Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, L. (2018). Universal Transformers. *CoRR*, *abs/1807.03819*. http://arxiv.org/abs/1807.03819

[37] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423

[38] Christiansen, M., & Chater, N. (1999). Connectionist Natural Language Processing: The State of the Art. *Cognitive Science. A Multidisciplinary Journal*, *vol 23*(4), 417-433. https://doi.org/10.1207/s15516709cog2304_2

[39] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*.

[40] Shi, T., & Liu, Z. (2014). Linking GloVe with word2vec. *CoRR*, *abs/1411.5595*. http://arxiv.org/abs/1411.5595

[41] Dubey, S. R., Singh, S. K., & Chaudhuri, B. B. (2021). A Comprehensive Survey and Performance Analysis of Activation Functions in Deep Learning. *CoRR*, *abs/2109.14545*. https://arxiv.org/abs/2109.14545

[42] Wu, Y., Lian, D., Xu, Y., Wu, L., & Chen, E. (2020). Graph Convolutional Networks with Markov Random Field Reasoning for Social Spammer Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*, 1054–1061. https://doi.org/10.1609/aaai.v34i01.5455

[43] Sanchez-Gonzalez, A., Heess, N., Springenberg, J. T., Merel, J., Riedmiller, M., Hadsell, R., & Battaglia, P. (2018). Graph Networks as Learnable Physics Engines for Inference and Control. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 4470–4479). PMLR. https://proceedings.mlr.press/v80/sanchez-gonzalez18a.html

[44] Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein Interface Prediction using Graph Convolutional Networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/f507783927f2ec2737ba40afbd17efb5-Paper.pdf

[45] Hamaguchi, T., Oiwa, H., Shimbo, M., & Matsumoto, Y. (2017). *Knowledge Transfer for Out-of-Knowledge-Base Entities: A Graph Neural Network Approach*.

[46] Ahmad, S. B. S., Rafie, M., & Ghorabie, S. M. (2021). Spam Detection on Twitter Using a Support Vector Machine and Users' Features by Identifying Their Interactions. *Multimedia Tools Appl.*, *80*(8), 11583–11605. https://doi.org/10.1007/s11042-020-10405-7

[47] Nettleton, D. (2013). Data Mining of Social Networks Represented as Graphs. *Computer Science Review*, *7*, 1–34. https://doi.org/10.1016/j.cosrev.2012.12.001

[48] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

[49] Kipf, T. N., & Welling, M. (2016). Semi-Supervised Classification with Graph Convolutional Networks. *CoRR*, *abs/1609.02907*. http://arxiv.org/abs/1609.02907

[50] Loyola-González, O., Monroy, R., Rodríguez, J., López-Cuevas, A., & Mata-Sánchez, J. I. (2019). Contrast Pattern-Based Classification for Bot Detection on Twitter. *IEEE Access*, *7*, 45800–45817. https://doi.org/10.1109/ACCESS.2019.2904220

[51] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks*.

[52] Lee, K., Caverlee, J., & Webb, S. (2010). Uncovering Social Spammers: Social Honeypots + Machine Learning. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 435–442. https://doi.org/10.1145/1835449.1835522

[53] Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. (2012). Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, *9*(6), 811 – 824. https://doi.org/10.1109/TDSC.2012.75

[54] Yang, C., Harkreader, R., & Gu, G. (2013). Empirical evaluation and new design for fighting evolving twitter spammers. *IEEE Transactions on Information Forensics and Security*, *8*(8), 1280 – 1293. https://doi.org/10.1109/TIFS.2013.2267732

[55] Cresci, S., Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2016). DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection. *IEEE Intelligent Systems*, *31*, 58–64. https://doi.org/10.1109/MIS.2016.29

[56] Davis, C., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. In *arXiv preprint arXiv:1602.00975*.

[57] Gilani, Z., Kochmar, E., & Crowcroft, J. A. (2017). Classification of Twitter Accounts into Automated Agents and Human Users. *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 489–496. https://api.semanticscholar.org/CorpusID:820775

[58] Pasricha, N., & Hayes, C. (2019). Detecting Bot Behaviour in Social Media using Digital DNA Compression. *Irish Conference on Artificial Intelligence and Cognitive Science*. https://api.semanticscholar.org/CorpusID:210964460

[59] Bacciu, A., Morgia, M. La, Mei, A., Nemmi, E. N., Neri, V., & Stefa, J. (2019). Bot and Gender Detection of Twitter Accounts Using Distortion and LSA. *Conference and Labs of the Evaluation Forum*. https://api.semanticscholar.org/CorpusID:198490029

[60] Loyola-González, O., Monroy, R., Rodríguez, J., López-Cuevas, A., & Mata-Sánchez, J. I. (2019). Contrast Pattern-Based Classification for Bot Detection on Twitter. *IEEE Access*, *7*, 45800–45817. https://doi.org/10.1109/ACCESS.2019.2904220

[61] Rodríguez-Ruiz, J., Mata-Sánchez, J. I., Monroy, R., Loyola-González, O., & López-Cuevas, A. (2020). A one-class classification approach for bot detection on Twitter. *Computers & Security*, *91*, 101715. https://doi.org/https://doi.org/10.1016/j.cose.2020.101715

[62] Shevtsov, A., Tzagkarakis, C., Antonakaki, D., & Ioannidis, S. (2021). *Identification of Twitter Bots based on an Explainable ML Framework: the US 2020 Elections Case Study*.

[63] Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in Online Social Networks. *Computer Communications*, *36*(10), 1120–1129. https://doi.org/https://doi.org/10.1016/j.comcom.2013.04.004

[64] Miller, Z., Dickinson, B., Deitrick, W., Hu, W., & Wang, A. H. (2014). Twitter spammer detection using data stream clustering. *Information Sciences*, *260*, 64–73. https://doi.org/https://doi.org/10.1016/j.ins.2013.11.016

[65] Chavoshi Nikan and Hamooni, H. and M. A. (2016). Identifying Correlated Bots in Twitter. In Y.-Y. Spiro Emma and Ahn (Ed.), *Social Informatics* (pp. 14–21). Springer International Publishing.

[66] Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. https://doi.org/10.1109/CTS.2013.6567202

[67] Ilias, L., & Roussaki, I. (2021). Detecting malicious activity in Twitter using deep learning techniques. *Applied Soft Computing*, *107*, 107360. https://doi.org/https://doi.org/10.1016/j.asoc.2021.107360

[68] Kudugunta, S., & Ferrara, E. (2018). Deep Neural Networks for Bot Detection. *CoRR*, *abs/1802.04289*. http://arxiv.org/abs/1802.04289

[69] Wei, F., & Nguyen, U. T. (2020). Twitter Bot Detection Using Bidirectional Long Short-term Memory Neural Networks and Word Embeddings. *CoRR*, *abs/2002.01336*. https://arxiv.org/abs/2002.01336

[70] Feng, S., Wan, H., Wang, N., Li, J., & Luo, M. (2021a). *SATAR: A Self-supervised Approach to Twitter Account Representation Learning and its Application in Bot Detection*.

[71] Alhosseini, S., Bin Tareaf, R., Najafi, P., & Meinel, C. (2019). *Detect Me If You Can: Spam Bot Detection Using Inductive Representation Learning*. 148–153. https://doi.org/10.1145/3308560.3316504

[72] Lei, Z., Wan, H., Zhang, W., Feng, S., Chen, Z., Li, J., Zheng, Q., & Luo, M. (2023). *BIC: Twitter Bot Detection with Text-Graph Interaction and Semantic Consistency*.

[73] Garcia-Silva, A., Berrio, C., & Gómez-Pérez, J. M. (2021). Understanding Transformers for Bot Detection in Twitter. *CoRR*, *abs/2104.06182*. https://arxiv.org/abs/2104.06182

[74] Martín Gutiérrez, D., Hernández-Peñaloza, G., Belmonte Hernández, A., Lozano-Diez, A., & Alvarez, F. (2021). A Deep Learning Approach for Robust Detection of Bots in Twitter Using Transformers. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2021.3068659

[75] Feng, S., Tan, Z., Li, R., & Luo, M. (2021c). Heterogeneity-aware Twitter Bot Detection with Relational Graph Transformers. *CoRR*, *abs/2109.02927*. https://arxiv.org/abs/2109.02927

[76] Ilias, L., Kazelidis, I. M., & Askounis, D. (2023). *Multimodal Detection of Social Spambots in Twitter using Transformers*.

[77] Feng, S., Wan, H., Wang, N., Li, J., & Luo, M. (2021b). *TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark*. 4485–4494. https://doi.org/10.1145/3459637.3482019

[78] Hansen, L. K., & Larsen, J. (1996). Unsupervised Learning and Generalization. In *Proceedings of IEEE International Conference on Neural Networks* (pp. 25-30). IEEE. https://doi.org/10.1109/ICNN.1996.548861

[79] Ilias, L., & Askounis, D. (2022). Explainable Identification of Dementia From Transcripts Using Transformer Networks. *IEEE Journal of Biomedical and Health Informatics*, *26*(8), 4153–4164. https://doi.org/10.1109/JBHI.2022.3172479

[80] Ilias, L., Soldner, F., & Kleinberg, B. (2022). *Explainable Verbal Deception Detection using Transformers*.

[81] Ilias L and Askounis D (2022) Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts. *Front. Aging Neurosci.* 14:830943. doi: 10.3389/fnagi.2022.830943

[82] Chatzianastasis, M., Ilias, L., Askounis, D., & Vazirgiannis, M. (2023). Neural Architecture Search with Multimodal Fusion Methods for Diagnosing Dementia. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096579

[83] Ilias, L., & Askounis, D. (2023). Context-aware attention layers coupled with optimal transport domain adaptation and multimodal fusion methods for recognizing dementia from spontaneous speech. *Knowledge-Based Systems*, *277*, 110834. https://doi.org/https://doi.org/10.1016/j.knosys.2023.110834

[84] Ilias, L., Askounis, D., & Psarras, J. (2022). A Multimodal Approach for Dementia Detection from Spontaneous Speech with Tensor Fusion Layer. *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 1–5. https://doi.org/10.1109/BHI56158.2022.9926818

[85] Ilias, L., Askounis, D., & Psarras, J. (2023a). Detecting dementia from speech and transcripts using transformers. *Computer Speech & Language*, *79*, 101485. https://doi.org/https://doi.org/10.1016/j.csl.2023.101485

[86] Ilias, L., Askounis, D., & Psarras, J. (2023b). Multimodal detection of epilepsy with deep neural networks. *Expert Systems with Applications*, *213*, 119010. https://doi.org/https://doi.org/10.1016/j.eswa.2022.119010