ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΑΠΟΦΑΣΕΩΝ – ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Ανίχνευση bots στο Twitter με χρήση Συνελικτικών Δικτύων Γραφημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κομβόπουλου Ελευθέριου

Επιβλέπων : Ασκούνης Δημήτριος

Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2024

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ ΚΑΙ ΣΥΣΤΗΜΑΤΩΝ
ΑΠΟΦΑΣΕΩΝ – ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΣΤΗΡΙΞΗΣ ΑΠΟΦΑΣΕΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Ανίχνευση bots στο Twitter με χρήση Συνελικτικών Δικτύων Γραφημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κομβόπουλου Ελευθέριου

Επιβλέπων : Ασκούνης Δημήτριος
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιουλίου 2024.

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

Αθήνα, Ιούνιος 2024

.....

Κομβόπουλος Ελευθέριος

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

Περίληψη

Η διπλωματική αυτή εργασία πραγματεύεται την ανάπτυξη ενός συστήματος μηχανικής μάθησης, με στόχο την ανίχνευση ψευδών λογαριασμών στην πλατφόρμα του Twitter. Το Twitter αποτελεί ένα μέσο κοινωνικής δικτύωσης που επιτρέπει την αλληλεπίδραση των χρηστών μέσω σύντομων δημοσιεύσεων που ονομάζονται tweets. Το κυρίαρχο πλεονέκτημα της συγκεκριμένης εφαρμογής είναι η ελεύθερη παράθεση απόψεων και ιδεών, οι οποίες μάλιστα έχουν τη δυνατότητα να οργανωθούν και να ομαδοποιηθούν με βάση το θέμα και τους συμμετέχοντες. Με αυτόν τον τρόπο δημιουργούνται νήματα και λίστες βαθιών και εκτενών, ή και μη, συζητήσεων με χαρακτηριστικά όπως τα likes, τα mentions, τα replies και τα hashtags να πρωταγωνιστούν.

Όλα όσα αναφέραμε ως θετικές δυνατότητες της πλατφόρμας του Twitter, έχουν την έμφυτη τάση να μετατρέπονται ανά πάσα στιγμή σε κρίσιμα μειονεκτήματά του, όχι προς το ίδιο το μέσο, αλλά φυσικά για τους ανθρώπους που το αξιοποιούν. Συγκεκριμένα, η δύναμη της επιρροής, που ένα τέτοιο μέσο προσεγγίζει, δε θα μπορούσε να μη συνοδευτεί από ζητήματα ασφάλειας και πιστότητας, όσον αφορά τις ειδήσεις και τις ιδέες που εκπέμπει. Με απλά λόγια, το Twitter, εδώ και πολλά χρόνια, έχει αποτελέσει ένα μέσο στρατευμένης διακίνησης απόψεων και ιδεών, με στόχο την κατεύθυνση ή και την παραπλάνηση μεγάλων ομάδων ανθρώπων, για κακοπροαίρετους σκοπούς. Η ελευθερία και η ανεξαρτησία μετατρέπεται έμμεσα σε υποδόρια χειραγώγηση και η ανάγκη για περιορισμό της εξάπλωσης των fake news και των bot λογαριασμών κρίνεται κάτι παραπάνω από επιτακτική.

Η εργασία αρχικά καταπιάνεται με όλο το θεωρητικό υπόβαθρο των μοντέλων που απασχολούν το συγκεκριμένο πρόβλημα, αλλά και με τις πιο θεμελιώδεις έννοιες της Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης, ενώ στη συνέχεια περιγράφει όλες τις κατηγορίες προϋπάρχουσων μεθόδων που επιχείρησαν να δώσουν λύση. Τέλος, στην αρχιτεκτονική που υλοποιούμε και προτείνουμε, αξιοποιούμε πολυτροπικές μεθόδους επεξεργασίας δεδομένων, τις οποίες εν τέλει συνδυάζουμε προκειμένου να καταλήξουμε στις τελικές προβλέψεις. Κύριοι πρωταγωνιστές του συστήματος είναι τα κατάλληλα προσαρμοσμένα Graph Convolutional Networks, τα οποία μεταφέρουν πληροφορία και εκτελούν αλληλεπιδράσεις στις γειτονιές των χρηστών, καθιστώντας τις σχέσεις μεταξύ τους πλήρως καθοριστικές.

Τέλος, συγκρίνουμε τις επιδόσεις του συστήματός μας με τις προγενέστερες της επιστημονικής κοινότητας, τονίζουμε τα δυνατά της σημεία αλλά και επισημάνουμε κάποιες μελλοντικές βελτιωτικές κινήσεις, οι οποίες μπορούν να εκτοξεύσουν περισσότερο την ακρίβεια, τη βαθύτητα αλλά και την ανθεκτικότητα του μοντέλου μας.

Λέξεις κλειδιά – Twitter, Μέσο Κοινωνικής Δικτύωσης, Ψευδείς Λογαριασμοί, Τεχνητή Νοημοσύνη, Μηχανική Μάθηση, Συνελικτικά Δίκτυα Γραφημάτων, Γειτονιές Χρηστών

Abstract

This thesis aims to the development of a bot detection model on Twitter, using Graph Convolutional Networks. Twitter is one of the most famous social media platforms, counting more than 350 million global users. Although Twitter was initially built in the sphere of communication, like the rest social media applications, its latest purposes concern the fields of information and advertising. Specifically, Twitter nowadays disposes a heavy impact in the information and the spread of ideas and opinions, which are mainly connected with socio-political issues, and organized with “hashtags”. As a result, the individual desire for strategic promotion of fake news, led Twitter to experience the rise of copious “bot” accounts, generated by automated software. The detection of those non-genuine accounts is vital as the need for their limitation and elimination is urgent.

The most dangerous part in the detection of those bot accounts is the fact that they are not static and indolent entities, but they progressively adapt their behavior with divergent characteristics. These characteristics, known as “user features”, usually include profile information, interaction with other accounts, and tweets. In other words, while the implemented models attempt to distinguish the authenticity of Twitter users, the bots dynamically evolve their actions and presence in the Twitter community, shaping a confusing landscape in the detection procedure.

There is a variety of existing deployed systems with dedication for users’ classification, utilizing multiple techniques in their models. Some of them focus more in user’s data and statistics, while others process only the tweets or the interaction links among the users. Both simple algorithms for clustering and machine learning methods have managed to achieve remarkable results, whereas the accuracy was better increased when multimodal perception of data and Natural Language Process came to the surface. However, the successful generalization of the parameters of the problem remains an open demanding question.

Our approach is based on three essential principles. At first, we analyse the user stats and appropriately create a balanced and realistic community graph. Secondly, we split the total user features into four fundamental categories and process them separately. Finally, we highlight the procedure of combining the results of the four Graph Convolutional Networks before exposing the results. Our system provides heterogeneity in data retrieval and processing, while it also underscores divergence’s inclusiveness and scalability for further future versions. All in all, our model achieves high metrics in comparison with other state-of-art architectures.

Keywords – Twitter, Social Media, Bot, Detection, Machine Learning, Natural Language Process, Graph Convolutional Networks, User Features, Tweets

Περιεχόμενα

Εικόνες και Πίνακες

Κεφάλαιο 1

Εισαγωγή

1.1 Μέσα Κοινωνικής Δικτύωσης

Social media, ή αλλιώς μέσα κοινωνικής δικτύωσης ονομάζονται οι τεχνολογικές πλατφόρμες που εξυπηρετούν τη δημιουργία, την κοινοποίηση και τη διαχείριση περιεχομένου, ιδεών, ενδιαφερόντων, και άλλων μορφών έκφρασης, μέσω εικονικών κοινοτήτων και δικτύων.

Απαρτίζουν μέσα διαδραστικής συμμετοχής και ευρείας επικοινωνίας ανθρώπων, δίχως γεωγραφικούς περιορισμούς. Τα μέσα κοινωνικής δικτύωσης θεωρούνται Web 2.0 Internet applications και είναι προσβάσιμα μέσω desktop και mobile εφαρμογών [36].

Παρέχουν τη δυνατότητα στους χρήστες να παράγουν, να επεξεργάζονται και να διαμοιράζονται προσωπικά τους κείμενα, φωτογραφίες, βίντεο, όχι μόνο ανά μεταξύ τους, δηλαδή με τη μορφή προσωπικής συζήτησης, αλλά και σε ιδιωτικές ή δημόσιες κοινότητες χρηστών, οι οποίες με τη σειρά τους δημιουργούν αλυσίδες αλληλεπίδρασης πάνω σε αυτό το γεγονός [28].

Το 1991, όταν ο Tim Berners-Lee ενσωμάτωσε hypertext λογισμικό με το Internet, δημιουργώντας το World Wide Web, θεμελίωσε τη βάση όλων των μεταγενέστερων διαδικτυακών επικοινωνιών. Αυτή η ανακάλυψη διευκόλυνε το σχηματισμό online κοινοτήτων και offline support groups, μέσω της χρήσης weblogs και emails. Έτσι, στα μέσα της συγκεκριμένης δεκαετίας, πλατφόρμες όπως GeoCities, Classmates και SixDegrees ήρθαν για να υλοποιήσουν τα πρώτα clients άμεσων μηνυμάτων και συνομιλιών, καθώς και καινοτόμα features όπως profiles και λίστα φίλων. Έπειτα, η ραγδαία έξαρση παρατηρήθηκε στις αρχές της επόμενης δεκαετίας, όταν τη σκυτάλη ανέλαβαν πιο ανεπτυγμένες πλατφόρμες , όπως το Myspace(2003), το Facebook(2004), το Youtube(2005) και το Twitter(2006), δίνοντας άλλες διαστάσεις στην κοινωνική δικτύωση. Πέρα από το ότι αποτέλεσαν τον ακρογωνιαίο λίθο της διαδικτυακής και εικονικής πραγματικότητας του σήμερα, έθεσαν τα θεμέλια και τα βασικά συστατικά όλων των μεταγενέστερων αντίστοιχων εφαρμογών που ακολούθησαν [43].

Εικόνα 1: Social Media

Τα social media - ειδικά στις μέρες μας - δεν αποτελούν μονάχα μέσο κοινωνικοποίησης και επικοινωνίας, ακόμα και αν ο αρχικός τους σχεδιασμός και η εναρκτήρια τους σύλληψη στόχευε αποκλειστικά σε αυτό. Πλέον οι δημοφιλείς αυτές πλατφόρμες εμπεριέχουν και παράπλευρες λειτουργίες, όπως η ενημέρωση, η ομαδοποίηση, η διαφήμιση προϊόντων και η προώθηση υπηρεσιών. Όπως είναι άμεσα αντιληπτό, πλέον ο κόσμος της κοινωνικής δικτύωσης, παρότι είναι χτισμένος στη σφαίρα της επικοινωνίας, ουσιαστικά έχει μεταποιηθεί σε έναν εμπορικό ή ιδεολογικό σύμμαχο επιχειρήσεων και συμφερόντων. Θα μπορούσε κάποιος να πει πως ναι μεν έχουν διατηρήσει και εξελίξει την ταυτότητά τους, αλλά πλέον διαθέτουν πολύπλευρη οπτική και αυξημένη συμμετοχή στην καθημερινότητα των ανθρώπων.

Σε γενικές γραμμές, οι νέες τεχνολογίες έχουν διογκώσει τις δυνατότητες των social media, και κατ’ επέκταση των χρηστών τους. Ωστόσο η υπέρμετρη χρήση τους, η εμπλουτισμένη διείσδυσή τους σε προσωπικά δεδομένα και η εκτεταμένη εμβέλεια τους έχουν σχηματίσει έναν άξονα που ελλοχεύει κινδύνους και απειλές. Η φύση των απειλών αυτών δεν είναι απαραίτητα άμεση, αλλά έμμεση.

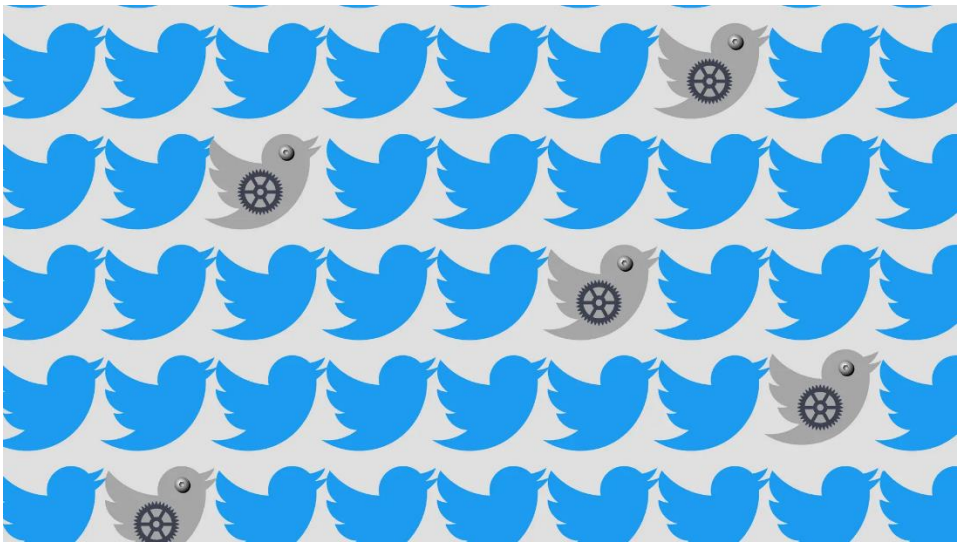
Οι απειλές ανήκουν σε διάφορες κατηγορίες, όπως εξαπάτηση, υποκλοπή προσωπικών δεδομένων, κλπ. Στα πλαίσια αυτής της εργασίας βέβαια θα εμβαθύνουμε σε πιο βαθιά μονοπάτια των επιπτώσεων των social media στην εποχή μας, οι οποίες δεν αρέσκονται στις βραχυπρόθεσμες εκμεταλλεύσεις που αναφέρθηκαν προηγουμένως, αλλά στοχεύουν σε πιο μακροπρόθεσμα αποτελέσματα στις ζωές των ανθρώπων [37].

1.2 Twitter

Το Twitter είναι μια από τις μεγαλύτερες παγκοσμίως πλατφόρμες κοινωνικής δικτύωσης, διαθέτοντας πλέον παραπάνω από 500 εκατομμύρια χρήστες. Δημιουργήθηκε το 2006 στην Αμερική, από τους Jack Doresy, Noah Glass, Biz Stone και Evan Williams. Η έδρα της βρίσκεται στο Σαν Φρανσίσκο της Καλιφόρνια, αλλά απαρτίζεται από περισσότερα από 25 παραρτήματα ανά τον κόσμο.

Οι χρήστες του Twitter έχουν τη δυνατότητα μέσω του προφίλ τους να στέλνουν και κοινοποιούν μηνύματα, με τη μορφή γραπτού κειμένου, εικόνων και βίντεο. Τα μηνύματα αυτά έχουν λάβει την ιστορική ονομασία «tweets». Μια ιδιαιτερότητα των tweets είναι ότι μπορούν να περικλείουν περιορισμένο πλήθος χαρακτήρων, το όριο του οποίου έχει μεταβληθεί από 140 που ήταν αρχικά, σε 280 (2017). Από το 2012 παραπάνω από 100 εκατομμύρια χρήστες παρήγαγαν 340 εκατομμύρια tweets κάθε μέρα, ενώ το 2019 η πλατφόρμα έφτασε τους 330 εκατομμύρια μηνιαία ενεργούς χρήστες, αποτελώντας έτσι - μέχρι το Φεβρουάριο το 2024 – την έκτη σε επισκεψιμότητα [44] ιστοσελίδα στον κόσμο, η οποία από τον Οκτώβριο του 2022 ανήκει στον δισεκατομμυριούχο Elon Musk.

Το ενδιαφέρον, αλλά ταυτόχρονα και επικίνδυνο στατιστικό του Twitter είναι πως με βάση εκτιμήσεων που πραγματοποιήθηκαν το 2020, το 15% των λογαριασμών (48 εκατομμύρια λογαριασμοί) δεν εκπροσωπούσαν γνήσιους ανθρώπους. Το στατιστικό αυτό συνδέθηκε άμεσα και με τις κατηγορίες κατά της πλατφόρμας για εξάπλωση της παραπληροφόρησης και εχθρικού λόγου, οι οποίες χρονικά εκκίνησαν με την εξαγορά από τον Elon Musk [47].



Εικόνα 2: Twitter

Συνδέοντας όλα τα παραπάνω με όσα αναφέρθηκαν προηγουμένως για τις αρνητικές επιπτώσεις των social media, ένα ισχυρό αντίκτυπο που έχει το Twitter σαν κοινωνικό δίκτυο είναι η διαμόρφωση και η εξάπλωση ιδεών και απόψεων, στο κοινωνικό πλέγμα των χρηστών της, οι οποίοι δυστυχώς δεν αποτελούν πάντα παραδείγματα άψογης και σωστά εκπαιδευμένης κριτικής ικανότητας. Οι ιδέες και οι απόψεις αυτές αφορούν κοινωνικοπολιτικά κυρίως ζητήματα, τα οποία πολλές φορές οργανώνονται και με τα λεγόμενα «tags», τα οποία είναι γνωστά και σε άλλα social media και είναι υπεύθυνα για την ομαδοποίηση συζητήσεων και γενικά κοινοποιήσεων σχετικά με ένα συγκεκριμένο και εξειδικευμένο θέμα.

Ο όρος «fake news» συνοψίζει την παραπάνω κατάσταση και της προσδίδει την ανάγκη για πρόβλεψη και αντιμετώπιση. Η υπόσταση των «fake news» γίνεται ακόμα πιο επικίνδυνη και απειλητική εάν η προέλευσή τους ανήκει σε ένα ολοένα και αυξανόμενο υποσύνολο χρηστών του Twitter, τα «bots». Συνεπώς ένα από τα βασικά και κύρια μέσα πρόληψης και διαχείρισης του προβλήματος είναι η ανίχνευση αυτών των μη γνήσιων, αλλά ψευδών λογαριασμών, με στόχο τον περιορισμό και την εξάλειψή τους.

Το συγκεκριμένο πρόβλημα δεν υφίσταται μονάχα στην πλατφόρμα του Twitter, αλλά και σε αντίστοιχες ιστοσελίδες κοινωνικής δικτύωσης, δίνοντας έτσι την αφορμή σε πολλούς επιστήμονες και μηχανικούς να ασχοληθούν με το ζήτημα. Στην επικείμενη εργασία, θα επιχειρήσουμε να ανιχνεύσουμε την παρουσία των bots στο Twitter, συγκρίνοντας πάντα την απόδοση και την ακρίβεια των αποτελεσμάτων και με προϋπάρχοντα συστήματα, τα οποία υλοποιήθηκαν για τον ίδιο σκοπό.

1.3 Αντικείμενο της διπλωματικής

Σκοπός της παρούσας εργασίας είναι η εμβάθυνση στα μοντέλα Μηχανικής Μάθησης όσον αφορά την αξιοποίηση τους στην ανίχνευση των bots στην πλατφόρμα του Twitter. Όπως έχει ήδη τονιστεί με εμφατικό τρόπο, το Twitter πλέον απαρτίζεται – σε συνεχώς αυξανόμενα ποσοστά – από μη γνήσιους προσωπικούς λογαριασμούς, των οποίων η ανίχνευση κρίνεται επιτακτική και απαραίτητη, προκειμένου η εξάπλωση των fake news να περιοριστεί σε όσο μεγαλύτερο βαθμό γίνεται.

Στη συγκεκριμένη διπλωματική, προτείνουμε την άντληση και αξιοποίηση πολύπλευρων χαρακτηριστικών των χρηστών, όπως την περιγραφή του προφίλ του (descriptions), τα στατιστικά του προφίλ του (numerical attributes), τις δημοσιεύσεις του (tweets), καθώς επίσης και ένα υποσύνολο άλλων παράπλευρων στοιχείων (κατηγορικά χαρακτηριστικά, tweets stats και flags). Δίνεται ιδιαίτερη έμφαση στον τρόπο επεξεργασίας των παραπάνω δεδομένων, έτσι ώστε η τελική μορφή τους να είναι και μαθηματικά διαχειρίσιμη αλλά επίσης και ερμηνευτικά ισχυρή, ως προς την ταυτότητα του χρήστη.

Εμπνευσμένοι από το υπάρχον σύστημα RGCN, χωρίζουμε τα παραπάνω χαρακτηριστικά των χρηστών σε κατηγορίες, με βάση τη φύση τους και την προέλευσή τους, και τα προωθούμε σε τέσσερα ξεχωριστά συνελικτικά νευρωνικά δίκτυα γραφημάτων (GCNs), εκμεταλλευόμενοι τις σχέσεις following στην κοινότητα του Twitter. Ο συνδυασμός των πορισμάτων των τεσσάρων μοναδικών και ατομικών υποσυστημάτων οδηγεί και στην τελική πρόβλεψη σχετικά με την κλάση (human or bot) του χρήστη. Η οποία πραγματοποιείται μέσω ενός τερματικού Multi Layer Perceptron.

Το σύστημα μας παρουσιάζει διάφορα πλεονεκτήματα, τα οποία κατορθώνουν όχι μόνο να το καθιστούν ανταγωνιστικό, συγκριτικά με άλλες state of the art αρχιτεκτονικές, αλλά επίσης να θέτουν τα θεμέλια για μια συνεχή προσαρμοστικότητα και ανταπόκριση στις απαιτήσεις ενός πολύπλοκου classification task, όπως αυτό του Twitter Bot Detection. Συγκεκριμένα :

- Όλη η πληροφορία – που αφορά τους χρήστες – από το Twitter ταξινομείται με σημασιολογικό τρόπο σε ξεχωριστές και ατομικές ροές στο σύστημα μας. Με αυτόν τον τρόπο, επιμερίζεται η προετοιμασία και η επεξεργασία των χαρακτηριστικών, δίνοντας ταυτόχρονα πλάτος στην αποτυπωμένη αρχική «εικόνα» του χρήστη.
- Τα γλωσσολογικά χαρακτηριστικά αξιοποιούνται μέσω μεταφοράς μάθησης γνωστών και εξειδικευμένων LLMs (RoBERTa), των οποίων η επιμέρους ακρίβεια συνεχώς βελτιώνεται και αναπτύσσεται. Με αυτόν τον τρόπο το σύστημα μας μπορεί να χαρακτηριστεί επίκαιρο, ευέλικτο, αλλά και συγχρονισμένο με τις εξελίξεις που λαμβάνουν χώρα στην επεξεργασία της φυσικής γλώσσας.

- Στην αρχιτεκτονική μας συμπεριλαμβάνουμε και τοποθετούμε πολύ υψηλά στην ιεραρχία τις σχέσεις μεταξύ των χρηστών, καθώς αυτές διαδραματίζουν κομβικό ρόλο στην ενδιάμεση επεξεργασία των χαρακτηριστικών. Γίνεται προσεγμένη επιλογή των χρηστών που θα συμμετέχουν στον κοινωνικό γράφο. Αναλύουμε διεξοδικά τις παραμέτρους που επηρεάζουν τη δομή ενός γράφου και τις καθορίζουμε με τέτοιον τρόπο που θα εξασφαλιστεί η βέλτιστη λειτουργία ενός GCN.
- Σε κάθε περίπτωση τονίζουμε τα δυνατά χαρακτηριστικά και τις υψηλές αποδόσεις των πειραμάτων μας, ωστόσο δεν παραλείπουμε να επισημάνουμε τα μονοπάτια βελτίωσης που σχηματίζονται στον ορίζοντα, έτσι ώστε το σύστημα να παραμείνει αποδοτικό στις εξελίξεις του περιβάλλοντος αλλά και να αυξήσει ακόμα περισσότερο το βάθος και την αξία του.

1.4 Περιορισμοί των υφιστάμενων συστημάτων

Όπως θα αναφερθούμε και εκτενέστερα στη συνέχεια της παρούσας διπλωματικής, έχουν αναπτυχθεί ποίκιλες προσεγγίσεις στο πρόβλημα ανίχνευσης των ψευδών λογαριασμών στην πλατφόρμα του Twitter. Εκκινώντας από απλούς αλγόριθμους και ταξινομητές, συνεχίζοντας με συνελικτικά δίκτυα γραφημάτων και καταλήγοντας σε βαθιά πολύ-επίπεδα συνδυαστικά μοντέλα, τα υπάρχοντα συστήματα έχουν κατορθώσει να παρέχουν ακριβείς προβλέψεις για τους χρήστες του Twitter, όσον αφορά την αυθεντικότητά τους. Ωστόσο, δεν παύουν να διαθέτουν διάφορους περιορισμούς.

Ένας βασικός και αναπόφευκτος περιορισμός σχετίζεται με την επικαιροποίηση του συνόλου δεδομένων των μοντέλων. Στο φάσμα του χρόνου, οι κακόβουλες πηγές παραγωγής ψευδών λογαριασμών προσαρμόζουν τη λειτουργία και τη δραστηριότητα τους, προκειμένου να αποφεύγουν τα detection filters, προσομοιάζοντας με μεγαλύτερη αποτελεσματικότητα τη συμπεριφορά γνήσιων λογαριασμών. Με άλλα λόγια, νέα βελτιωμένα bots συνεχώς εμφανίζονται ανανεωμένα στο προσκήνιο, δυσχεραίνοντας την ακρίβεια όλων των μοντέλων και σχηματίζοντας έναν χάρτη λογαριασμών με αποκλίνοντα χαρακτηριστικά (divergent users).

Μία σημαντική έλλειψη – στην πλειοψηφία των υφιστάμενων συστημάτων – είναι η συλλογή και αξιοποίηση στατιστικών που αφορούν τα ίδια τα tweets των χρηστών. Χαρακτηριστικά όπως likes, comments, mentions, και άλλα, προσδίδουν επιπρόσθετη χρήσιμη πληροφορία για τα tweets, και είναι ικανά να βελτιώσουν την προβλεπτική ικανότητα ενός μοντέλου.

Μια κυρίαρχη παράμετρος της εκπαίδευσης ενός συνελικτικού νευρωνικού δικτύου, στην περίπτωση του bot detection task στο Twitter, είναι ο κοινωνικός γράφος των χρηστών. Συγκεκριμένα, η παρουσία ενός αρχικού component, το οποίο θα εκτελεί μια βασική ανάλυση σχετικά με τα χαρακτηριστικά του γράφου, όπως :

- Πυκνότητα γραφήματος
- Εξωτερικός βαθμός κορυφών (outbound degree)
- Ποσοστό ψευδών λογαριασμών

αλλά και ενός δεύτερου component το οποίο θα είναι υπεύθυνο για την κατάλληλη εξισορρόπηση του training set, μέσω του oversampling, κρίνεται απαραίτητη έτσι ώστε η εκπαίδευση του μοντέλου μας να είναι όσο πιο αμερόληπτη και αντικειμενική γίνεται, αποφεύγοντας καταστάσεις overfitting και λανθάνουσες τετριμμένες προβλέψεις. Αυτή η διαδικασία προετοιμασίας και προεπεξεργασίας του κοινωνικού γράφου – που θα εισαχθεί στο νευρωνικό δίκτυο – απουσιάζει από πολλά υπάρχοντα συστήματα.

1.5 Οργάνωση κειμένου

Η εργασία αυτή οργανώνεται και εκτείνεται σε έξι επιμέρους κεφάλαια. Το Κεφάλαιο 1 δημιουργεί μια πρώτη εικόνα για τον κόσμο του Twitter, καθώς και για το πρόβλημα της ανίχνευσης των bots. Το Κεφάλαιο 2 παρέχει στον αναγνώστη το απαραίτητο θεωρητικό υπόβαθρο αναφορικά με τις θεμελιώδεις έννοιες και αρχές της Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης, περιγράφοντας αναλυτικά το σκοπό και τη λειτουργία όλων των διαφορετικών τύπων νευρωνικών δικτύων. Στο Κεφάλαιο 3 αποτυπώνονται και παρουσιάζονται υπάρχουσες μέθοδοι και αρχιτεκτονικές που στοχεύουν στην ανίχνευση των bots στο Twitter, επισημάνοντας ταυτόχρονα και τη λογική και τους περιορισμούς της κάθε προσέγγισης. Το Κεφάλαιο 4 αναφέρεται στο σύνολο δεδομένων που επιλέχθηκε να αξιοποιηθεί στη συγκεκριμένη εργασία κατά την εκπαίδευση του μοντέλου μας. Στο Κεφάλαιο 5 περιγράφεται η μεθοδολογία που ακολουθήσαμε για την υλοποίηση του δικού μας συστήματος, του οποίου τα πειράματα και τα αποτελέσματα καταγράφονται και σχολιάζονται στο Κεφάλαιο 6. Τέλος, το Κεφάλαιο 7 περιέχει ένα τελικό συμπέρασμα της εργασίας, αλλά προτείνονται και μελλοντικές πιθανές κινήσεις που θα βελτιώσουν περαιτέρω την ακρίβεια του μοντέλου μας.

Κεφάλαιο 2

Θεωρία

2.1 Τεχνητή Νοημοσύνη

Τεχνητή Νοημοσύνη είναι το επιστημονικό πεδίο το οποίο συνδυάζει την επιστήμη των υπολογιστών και ισχυρά σύνολα δεδομένων, με στόχο την επίλυση προβλημάτων.

Οι πρώτες συζητήσεις γύρω από την Τεχνητή Νοημοσύνη γεννήθηκαν κατά τη δημοσίευση της εργασίας «Computing Machinery and Intelligence» από τον Alan Turing το 1950. Κεντρικό περιεχόμενο της συγκεκριμένης εργασίας αποτελεί η αμφιβολία σχετικά με το εάν οι μηχανές είναι ικανές να σκεφτούν. Επέκταση αυτού του ερωτήματος είναι το διάσημο «Turing test», κατά το οποίο ένας άνθρωπος καλείται - μέσω ερωτήσεων - να διακρίνει την φύση του συνομιλητή του : άνθρωπος ή μηχανή. Παρόλο που το συγκεκριμένο πρόβλημα έχει ερευνηθεί και αναλυθεί ενδελεχώς, δεν παύει να αποτελεί σημαντικό κομμάτι της ιστορίας της Τεχνητής Νοημοσύνης, και να αγγίζει όχι μόνο τεχνολογικά αλλά και φιλοσοφικά μονοπάτια.

Κύριο μέλημα της Τεχνητής Νοημοσύνης είναι η κατάλληλη χρήση αλγορίθμων με την οποία θα επιτευχθεί η ακριβέστερη πρόβλεψη ή κατηγοριοποίηση δεδομένων. Μερικές εφαρμογές της είναι οι εξής [25]:

1. Speech recognition. Σκοπός της είναι η αυτόματη αναγνώριση ανθρώπινης ομιλίας (ASR), δηλαδή η μετατροπή προφορικής ομιλίας σε γραπτό κείμενο. Βασίζεται σε μοντέλα επεξεργασίας φυσικής γλώσσας (NLP) και παρέχεται σε συστήματα που επιθυμούν να βελτιώσουν την προσβασιμότητα τους (Siri).
2. Computer vision. Σκοπός της είναι η εξαγωγή χρήσιμης και σημαντικής πληροφορίας από εικόνες, βίντεο και γενικά οπτικά δεδομένα, η οποία στη συνέχεια αξιοποιείται κατά τη λήψη αποφάσεων και την διεξαγωγή αντίστοιχων ενεργειών. Παραδείγματα χρήσης computer vision μοντέλων υπάρχουν στις επιστήμες υγείας (ακτινολογική απεικόνιση), στη βιομηχανία αυτοκινήτων (αυτόματη οδήγηση), και σε πολλά άλλα πεδία.
3. Recommendation system. Σκοπός του είναι η αξιοποίηση παρελθοντικών συμπεριφορών, με στόχο την ανακάλυψη τάσεων και συσχετίσεων, που οδηγούν σε αποτελεσματικές στρατηγικές πωλήσεων. Χρησιμοποιείται κατά κόρον σε πλατφόρμες που συνδέουν προϊόντα και πελάτες, όπως e-shops. Το ιδανικό αποτέλεσμα των recommendation systems είναι η πρόταση του κατάλληλου προϊόντος στον κατάλληλο πελάτη την κατάλληλη χρονική στιγμή.

Λαμβάνοντας υπόψιν τις εφαρμογές της Τεχνητής Νοημοσύνης, εύκολα γίνεται αντιληπτό πως η πηγή της δημιουργίας της από τον άνθρωπο είναι βασισμένη πάνω σε δύο βασικούς πυλώνες, την αποκωδικοποίηση ερεθισμάτων και την λήψη περίπλοκων αποφάσεων.

2.2 Μηχανική Μάθηση

Μηχανική μάθηση είναι ο κλάδος της Τεχνητής Νοημοσύνης ο οποίος επικεντρώνεται στη χρήση δεδομένων και αλγορίθμων για να μιμηθεί τον τρόπο με τον οποίο οι άνθρωποι μαθαίνουν και σταδιακά βελτιώνουν την ακρίβεια τους.

Η Μηχανική Μάθηση αποτελεί ένα σημαντικό συστατικό του συνεχώς αναπτυσσόμενου πεδίου της επιστήμης των δεδομένων (Data Science), στο οποίο στατιστικές μέθοδοι και μαθηματικοί αλγόριθμοι συνθέτουν συστήματα παραγωγής προβλέψεων και ανάδειξης χρήσιμης πληροφορίας. Η αξία αυτών των recommendations και των insights είναι κομβική στην βελτίωση μετρικών ανάπτυξης επιχειρήσεων και εφαρμογών, καθώς οι απαιτήσεις της αγοράς εξαπλώνονται ραγδαία και τα μεγέθη των δεδομένων αυξάνονται εκθετικά (Big Data).

Από τεχνική άποψη, η Μηχανική Μάθηση βασίζεται στην έννοια της εκπαίδευσης των αλγορίθμων που πλαισιώνουν ένα σύστημα, η οποία πραγματοποιείται στο χώρο των αντίστοιχων εισερχόμενων δεδομένων. Εισερχόμενα δεδομένα μπορεί να είναι διάφορα χαρακτηριστικά οντοτήτων, όπως προσωπικά στοιχεία, αριθμητικοί δείκτες, περιγραφές αντικειμένων, κλπ. Η εκπαίδευση, σα διαδικασία, ορίζεται από τρία επιμέρους συστατικά [23] :

1. Τη μαθηματική διεργασία που καθορίζει την τελική απόφαση του συστήματος. Με λίγα λόγια τον τρόπο με τον οποίο τα εισερχόμενα δεδομένα συνδυάζονται για να προβλέψουν ή να κατηγοριοποιήσουν μια οντότητα.
2. Μια συνάρτηση σφάλματος που υπολογίζει τη διαφορά ανάμεσα στο πόρισμα του συστήματος και το επιθυμητό αποτέλεσμα. Ως επιθυμητό αποτέλεσμα ορίζεται είτε η ίδια η πραγματικότητα, στην περίπτωση που γνωρίζουμε εξαρχής την κατηγοριοποίηση των εισερχόμενων δεδομένων (supervised learning), είτε η βελτιστοποίηση ενός κριτηρίου που εμείς έχουμε ορίσει, στην περίπτωση που δε γνωρίζουμε εξαρχής την κατηγοριοποίηση των εισερχόμενων δεδομένων (unsupervised learning).
3. Την προσαρμογή των παραμέτρων του αλγορίθμου προκειμένου να μειωθεί το σφάλμα που προκλήθηκε και να βελτιωθεί η απόδοση του συστήματος.

Η παραπάνω διαδικασία επαναλαμβάνεται συνεχώς έως ότου το σφάλμα γίνει αποδεκτό, δηλαδή μέχρι να ξεπεράσει το κατώφλι που εμείς έχουμε ορίσει. Βέβαια, η φύση και η γεωμετρία των προβλημάτων πρόβλεψης και κατηγοριοποίησης, συχνά θέτουν εκείνες τα όρια βελτιστοποίησης ενός συστήματος. Συνεπώς, το σημείο κατά το οποίο η συνέχεια της εκπαίδευσης δεν επιφέρει περαιτέρω βελτίωση στην απόδοση του συστήματος, καθορίζει και τον τερματισμό της εκπαίδευσης.

Όπως έχει υπονοηθεί και παραπάνω, τα συστήματα Μηχανικής Μάθησης χαρακτηρίζονται άμεσα από το εάν τα εισερχόμενα δεδομένα περιέχουν ετικέτα ή όχι. Δηλαδή από το εάν γνωρίζουμε εξαρχής την κατηγοριοποίηση τους ή όχι. Ωστόσο, στη σύγχρονη εποχή των Big Data, ιδιαίτερη αξία αποκτά η αντιμετώπιση συνόλων δεδομένων που αποτελούνται από μεικτές - όσον αφορά την ετικέτα – οντότητες (semi-supervised learning).

2.3 Βαθιά Μάθηση

Η Βαθιά Μάθηση είναι ένα υποσύνολο της Μηχανικής Μάθησης, και μπορεί ουσιαστικά να περιγραφεί ως ένα νευρωνικό δίκτυο με τρία ή παραπάνω επίπεδα. Το νευρωνικό δίκτυο επιχειρεί να προσομοιάσει τη συμπεριφορά και τη λειτουργία του ανθρώπινου εγκεφάλου, χωρίς βέβαια να υπονοεί ότι καλύπτει όλες τις δυνατότητές του [22].

Αν και Μηχανική και Βαθιά Μάθηση έχουν ως κοινούς πυρήνες την ενασχόληση με μεγάλα σε εύρος σύνολα δεδομένων και την προσπάθεια παραγωγής προβλέψεων και insights, υποβόσκει μια βασική διαφορά που πρακτικά τις διαχωρίζει. Συγκεκριμένα, στη Βαθιά Μάθηση μπορεί να ληφθεί ως είσοδος ένα μη δομημένο και προεπεξεργασμένο σύνολο δεδομένων, ενώ ταυτόχρονα το σύστημα είναι ικανό - μέσω εκπαίδευσης – να καθορίσει τα πηγαία χαρακτηριστικά που διαμορφώνουν μια οντότητα και την ξεχωρίζουν από κάποια άλλη [5].

Για παράδειγμα, έστω ότι έχουμε ένα σύνολο από φωτογραφίες και στόχος του αναπτυσσόμενου συστήματος είναι να τις κατηγοριοποιούμε σε είδος αντικειμένου (αυτοκίνητο, τηλεόραση, τραπέζι κλπ). Μέσω Βαθιάς Μάθησης, δηλαδή χρήσης νευρωνικών δικτύων, τα χαρακτηριστικά της εικόνας, που συμβάλουν ενεργά στην λήψη της παραπάνω απόφασης, εξάγονται αυτόματα έπειτα από προπόνηση των επιμέρους βαρών του νευρωνικού δικτύου.

Στην περίπτωση της Μηχανικής Μάθησης, η ιεραρχία των χαρακτηριστικών που διέπουν ένα σύστημα, είναι καθιερωμένη από τον άνθρωπο, μειώνοντας έτσι τον αυτοματισμό και την αυτοβουλία του συστήματος. Θα μπορούσε κάποιος να πει πως τα συστήματα

Βαθιάς Μάθησης είναι γεννημένα να ανταπεξέρχονται και να προσαρμόζονται μόνα τους στα σύνολα δεδομένων που τους ανατίθενται.

2.4 Ανατροφοδοτούμενα Νευρωνικά Δίκτυα και Backpropagation

Κατά τη συζήτηση για νευρωνικά δίκτυα, δε γίνεται να παραληφθεί η διαδικασία backpropagation, στην οποία οφείλεται η συνολική λειτουργία και αποτελεσματικότητα συστημάτων Μηχανικής και Βαθιάς Μάθησης.

Καταρχάς, υπενθυμίζουμε πως το νευρωνικό δίκτυο έχει ως στόχο την προσομοίωση των λογικών διεργασιών που πραγματοποιεί ο ανθρώπινος εγκέφαλος. Αποτελείται από στρώματα νευρώνων, τα οποία συνδέονται εξαντλητικά μεταξύ τους με βάρη. Ο αριθμός των συνολικών νευρώνων και των στρωμάτων αυτών σχετίζεται με τη φύση και το είδος του εισερχόμενου συνόλου δεδομένων και των χαρακτηριστικών του, και αποτελεί αντικείμενο βελτιστοποίησης στο εκάστοτε πρόβλημα.

Το input ενός νευρωνικού δικτύου δεν είναι άλλο παρά το σύνολο των χαρακτηριστικών των οντοτήτων που μας ενδιαφέρουν. Για παράδειγμα, εάν οι οντότητες που μας ενδιαφέρουν είναι οι άνθρωποι και στόχος του συστήματος είναι η κατηγοριοποίησή τους σε επιζώντες ή μη σε έναν πόλεμο, τότε η είσοδος του νευρωνικού δικτύου θα ήταν ένας πίνακας που θα περιείχε την ηλικία, το φύλο, το επάγγελμα, την καταγωγή, καθώς και άλλα αντίστοιχα προσωπικά στοιχεία για κάθε άνθρωπο.

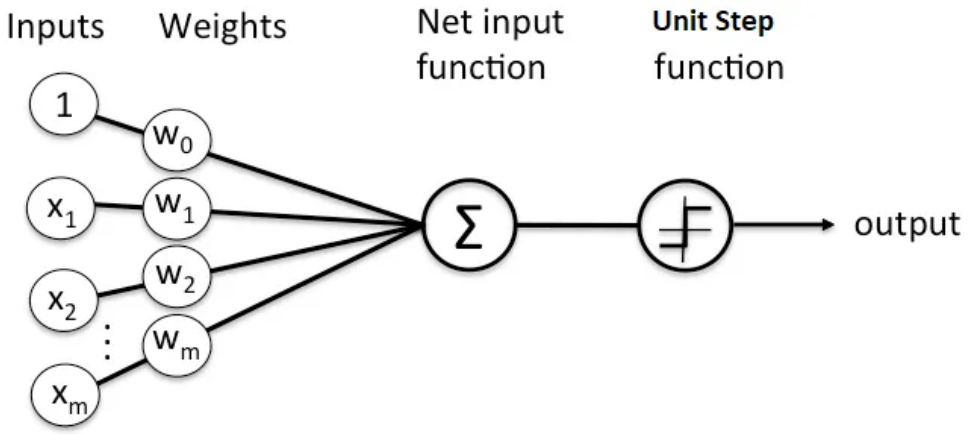
Το νευρωνικό δίκτυο συνδυάζει όλα τα παραπάνω χαρακτηριστικά, καθώς και τα αποτελέσματα των συνδυασμών τους, κατά στρώματα, δημιουργώντας έτσι έναν γράφο νευρώνων, οι οποίοι συνδέονται μεταξύ τους μέσω βεβαρημένων ακμών. Το αποτέλεσμα του παραπάνω συστήματος είναι το τελικό στρώμα νευρώνων, του οποίου το πλήθος ταυτίζεται με τον αριθμό των κατηγοριών (πχ το πρόβλημα των επιζώντων περιέχει δύο κατηγορίες), ενώ οι τιμές του οποίου καθορίζουν το αποτέλεσμα της κατηγοριοποίησης.

Η διαδικασία backpropagation, λοιπόν, εμπλέκεται στην εκπαίδευση του παραπάνω συστήματος, καθώς κατά την πλήρη εκτέλεση της παραπάνω διαδικασίας, μέσω των αποτελεσμάτων παράγεται ένα σφάλμα (loss), το οποίο δείχνει τη διαφορά της πρόβλεψης και την πραγματικότητας. Μέσω του backpropagation, το loss επιμερίζεται και τροφοδοτείται πίσω στο δίκτυο. Έτσι, σε βάθος επαναλήψεων και εποχών, οι βεβαρημένες ακμές, που ευθύνονται για όλες τις ενδιάμεσες μαθηματικές πράξεις που οδηγούν στο αποτέλεσμα, ανανεώνονται, ρυθμίζονται και προσαρμόζονται ολοένα και ορθότερα, βελτιώνοντας έτσι την απόδοση του συστήματος [30].

2.5 Single Layer Perceptron

Αποτελεί την ειδική περίπτωση ενός feed-forward νευρωνικού δικτύου, το οποίο αποτελείται αποκλειστικά από ένα στρώμα νευρώνων. Έχει την ικανότητα να κατηγοριοποιήσει μόνο γραμμικά διαχωρίσιμες οντότητες σε δύο διακριτές ομάδες [4].

Δεν περιέχει κρυμμένα ενδιάμεσα στρώματα, καταντώντας έτσι ένα επιφανειακό και ρηχό νευρωνικό δίκτυο, με απλή μαθηματική μορφή αλλά και γραμμικούς περιορισμούς όσον αφορά τα αποτελέσματά του. Θα μπορούσαμε να πούμε πως πρακτικά διαχωρίζει το σύνολο των δεδομένων στο χώρο σε δύο κατηγορίες, μέσω μιας ευθείας γραμμής.



Εικόνα 3: Single Layer Perceptron

Το σύστημα αποτελείται από δύο μέρη, τον πολλαπλασιασμό της εισόδου με τα βάρη του δικτύου και την συνάρτηση ενεργοποίησης, η οποία οδηγεί το αποτέλεσμα σε πεδίο δύο τιμών, έστω 1 και -1, τα οποία αντιστοιχούν στις δύο κλάσεις της κατηγοριοποίησης.

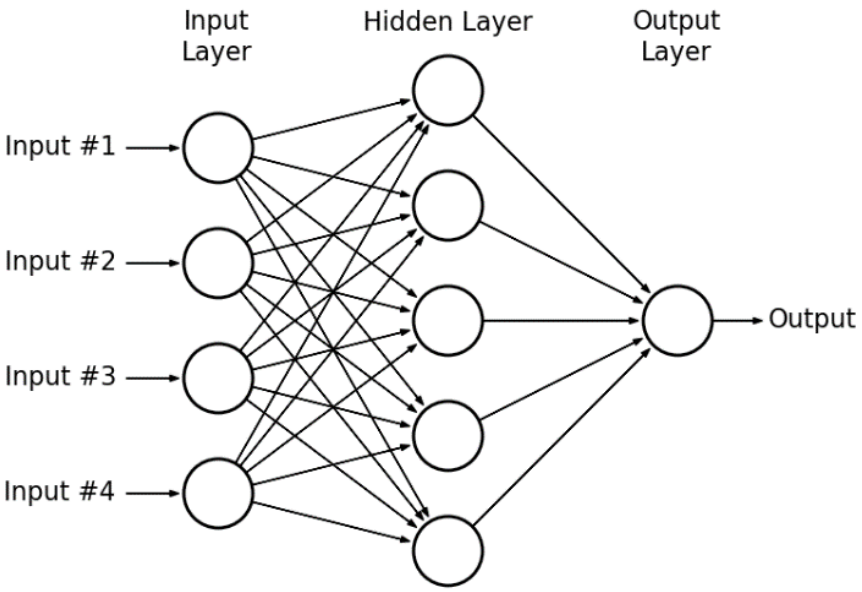
$$z = \sum_{j=1}^{m} x_j w_j = w^T x$$

Για τη διάκριση αυτών των δύο τιμών χρησιμοποιείται πάντα ένα κατώφλι θ.

$$\mathrm{g(z)} = \begin{cases} 1, & \text{if } z \geq \theta \\ -1, & \text{otherwise} \end{cases}$$

2.6 Multi Layer Perceptron

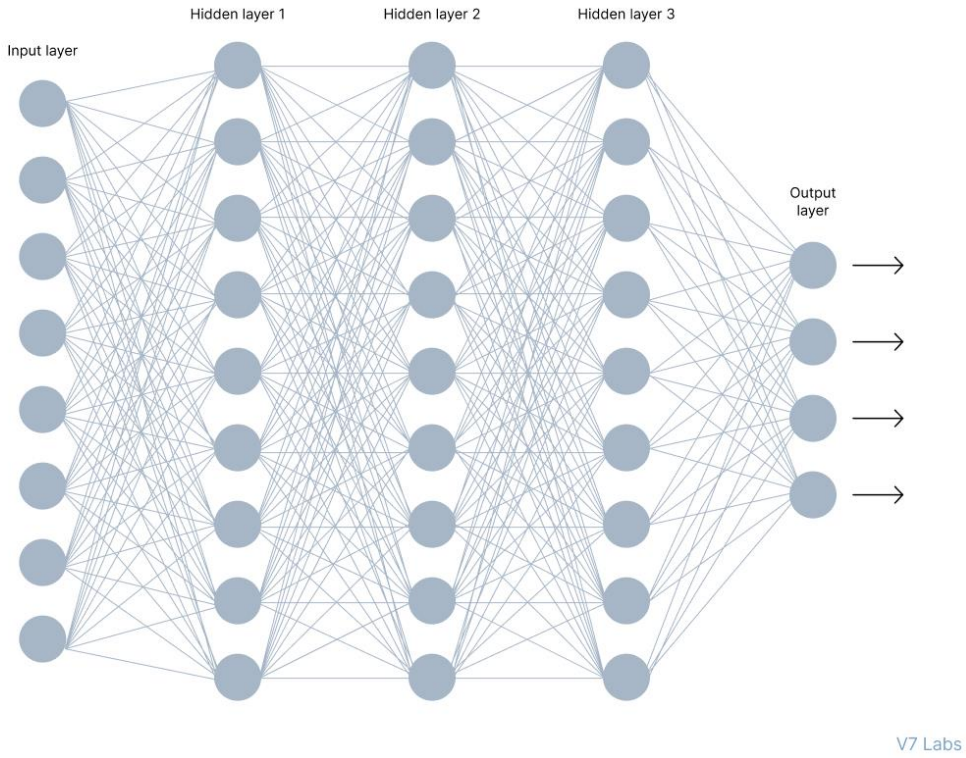
Σε αντίθεση με την περίπτωση του SLP, η συγκεκριμένη γενικευμένη κατηγορία νευρωνικού δικτύου διέπεται από την ύπαρξη κρυμμένων ενδιάμεσων στρωμάτων νευρώνων. Κάθε νευρώνας ενός κρυμμένου στρώματος λαμβάνει ως είσοδο την έξοδο των νευρώνων του προηγούμενου στρώματος, υπολογίζει – με βάση τα εκπαιδευόμενα βάρη του - ένα σταθμισμένο άθροισμα τους, και τελικά παράγει τη δική του έξοδο μέσω μιας συνάρτησης ενεργοποίησης [10].



Εικόνα 4: Multi Layer Perceptron

Στην παραπάνω εικόνα φαίνεται το παράδειγμα ενός νευρωνικού δικτύου με ένα ενδιάμεσο στρώμα. Έχοντας αναλύσει την περίπτωση του SLP, μπορούμε να πούμε πως κάθε νευρώνας ενός κρυμμένου στρώματος, σε συνδυασμό με τα inputs και τα βάρη του, είναι ένα τοπικό SLP.

Προφανώς αναλόγως με τη φύση και τις απαιτήσεις του κάθε προβλήματος, ένα MLP μπορεί να παραμετροποιηθεί, όσον αφορά το μέγεθος και τα επίπεδά του, αυξάνοντας την πολυπλοκότητα των υπολογισμών, αλλά και το βάθος των αποτελεσμάτων.

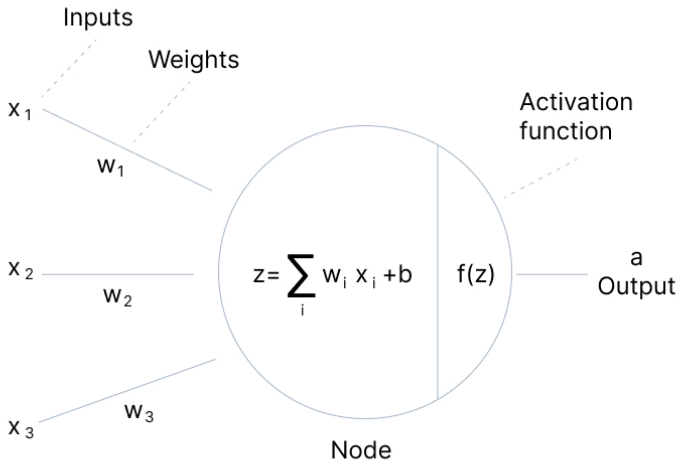


Εικόνα 5: Hidden Layers

2.7 Συναρτήσεις Ενεργοποίησης

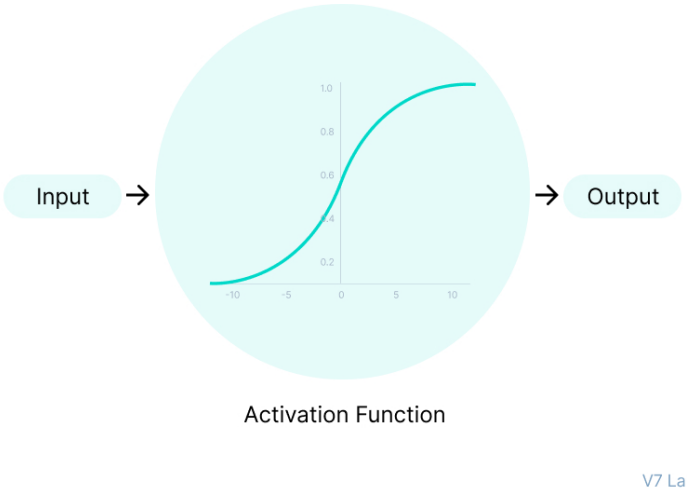
Η συνάρτηση ενεργοποίησης είναι μια καθοριστική έννοια κι ένα αναπόσπαστο κομμάτι στην ύπαρξη και τη λειτουργία των νευρωνικών δικτύων. Ο ρόλος της είναι η απόφαση σχετικά με το εάν ένας νευρώνας θα ενεργοποιηθεί ή όχι, δηλαδή το κατά πόσο σημαντική είναι μαθηματικά η είσοδος του νευρώνα στην εξέλιξη της διαδικασίας πρόβλεψης του νευρωνικού δικτύου.

Πρακτικά, ο πρωταρχικός ρόλος της συνάρτησης ενεργοποίησης είναι η μετατροπή του σταθμισμένου αθροίσματος του προηγούμενου στρώματος νευρώνων, σε μια τιμή εξόδου, η οποία θα διοχετευτεί στο επόμενο στρώμα.



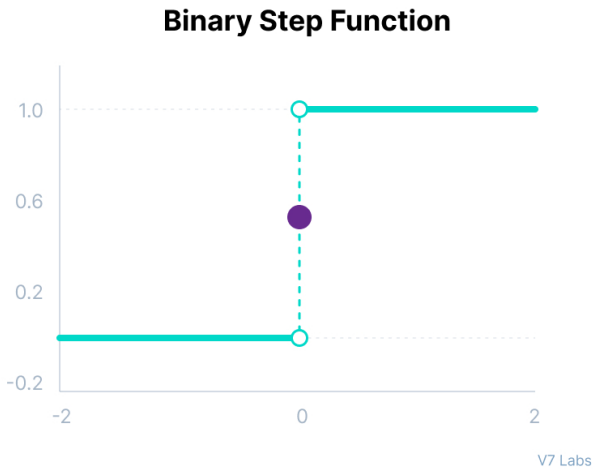
Εικόνα 6: Συνάρτηση Ενεργοποίησης

Χωρίς τη συνάρτηση ενεργοποίησης, το μοντέλο των Perceptrons θα αποτελούσε απλά μια εφαρμογή γραμμικής παλινδρόμησης (Linear Regression Model), αφού κάθε νευρώνας θα υπολόγιζε απλά ένα γραμμικό συνδυασμό των εισόδων του, με βάση κάποια βάρη. Συνεπώς η ύπαρξη πολλαπλών νευρώνων και στρωμάτων θα έχανε το νόημα της και θα εκφυλιζόταν, αφού η σύνθεση δύο γραμμικών συναρτήσεων είναι από μόνη της μια γραμμική συνάρτηση. Με αυτόν τον τρόπο το μοντέλο θα ήταν μεν απλό, αλλά ταυτόχρονα αδύναμο να αντιμετωπίσει σύνθετα σύνολα δεδομένων.



Υπάρχουν διάφορα είδη συναρτήσεων ενεργοποίησης, αναμεσά τους και απλές και σύνθετες, διάσημες και λιγότερο γνωστές [38].

2.7.1 Binary Step Function



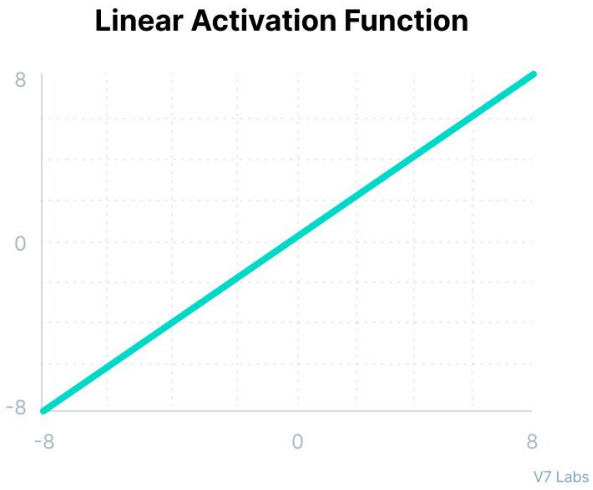
Εικόνα 7: Binary Step Function

Η παρούσα συνάρτηση εξαρτάται αποκλειστικά από το κατώφλι, το οποίο καθορίζει εάν ο νευρώνας θα ενεργοποιηθεί ή όχι. Στην αρνητική περίπτωση, η τιμή της εξόδου του νευρώνα δε θα περάσει στο επόμενο στρώμα.

$$\textit{Binary step}$$

$$f(x) = \begin{cases} 0 & for\ x < 0 \\ 1 & for\ x \geqslant 0 \end{cases}$$

2.7.2 *Linear Activation Function*



Εικόνα 8: Linear Activation Function

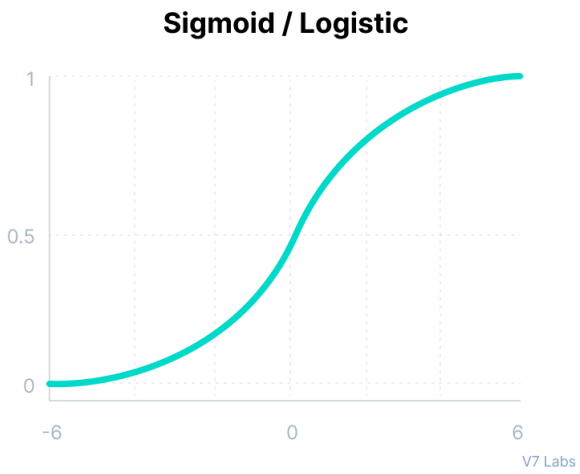
Η γραμμική συνάρτηση μεταφοράς απλά μεταφέρει ακέραια την είσοδό της στο επόμενο στρώμα, χωρίς να την παραποιεί.

$$Linear$$

$$f(x) = x$$

Η χρήση της εκφυλίζει την πολυεπίπεδη φύση ενός νευρωνικού δικτύου, και πρακτικά μετατρέπει ένα MLP σε SLP. Η παράγωγός της δεν εξαρτάται από την είσοδο, κι έτσι η backpropagation λογική είναι αδύνατο να εφαρμοστεί. Έτσι, όλα τα στρώματα του δικτύου καταρρέουν σε ένα.

2.7.3 Sigmoid / Logistic



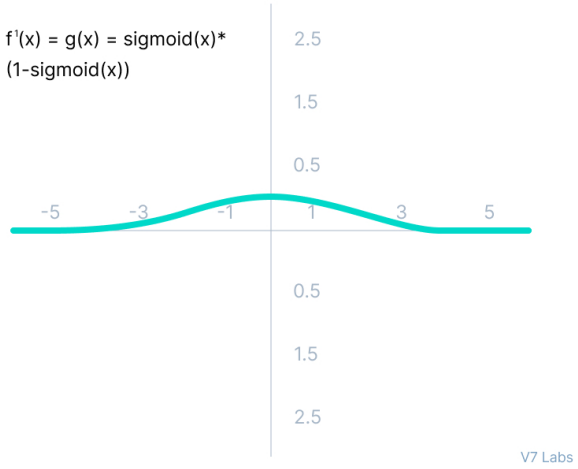
Εικόνα 9: Sigmoid Activation Function

Η συγκεκριμένη συνάρτηση λαμβάνει ως είσοδο οποιαδήποτε πραγματική τιμή και την τοποθετεί στο διάστημα από 0 έως 1. Όσο μικρότερη είναι η τιμή τόσο πιο κοντά στο 0 αυτή κυμαίνεται, ενώ όσο μεγαλύτερη είναι η τιμή τόσο πιο πολύ προσεγγίζει το 1.

$$Sigmoid\ /\ Logistic$$

$$f(x) = \frac{1}{1 + e^{-x}}$$

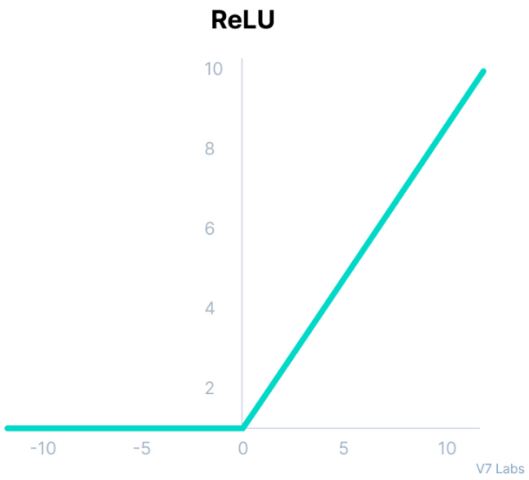
Η ομαλή παραγωγισιμότητα της, σε συνδυασμό με την πιθανολογική μορφή της, της δίνει τον τίτλο της πιο ευρέως γνωστής και χρησιμοποιημένης συνάρτησης μεταφοράς. Ωστόσο δεν παύει να παρουσιάζει περιορισμούς.



Συγκεκριμένα, όπως φαίνεται παραπάνω, η παράγωγος της είναι σημαντική μόνο σε μια μικρή γειτονιά κοντά στο 0. Για τιμές εκτός του διαστήματος [-3, 3], η παράγωγος μηδενίζεται, με αποτέλεσμα η εκπαίδευση που λαμβάνει χώρα στο νευρωνικό δίκτυο μέσω του backpropagation, να μην επιφέρει βελτιώσεις στα βάρη του δικτύου. Το συγκεκριμένο φαινόμενο ονομάζεται «Vanishing gradient problem» και συνοδεύεται με αδυναμία του νευρωνικού δικτύου να εκπαιδευτεί.

Τέλος, ένα άλλο μειονέκτημα της Σιγμοειδούς συνάρτησης μεταφοράς είναι ότι η έξοδος της δεν είναι συμμετρική κοντά στο 0, θέτοντας έτσι μια αστάθεια στο νευρωνικό δίκτυο.

2.7.4 Relu



Εικόνα 10: ReLU Activation Function

Παρόλο που η παρουσα συνάρτηση δίνει την εντύπωση μιας γραμμικής ενεργοποίησης, παρουσιάζει παράγωγο κι έτσι κρίνεται κατάλληλη και αποτελεσματική σε συστήματα που περιέχουν backpropagation.

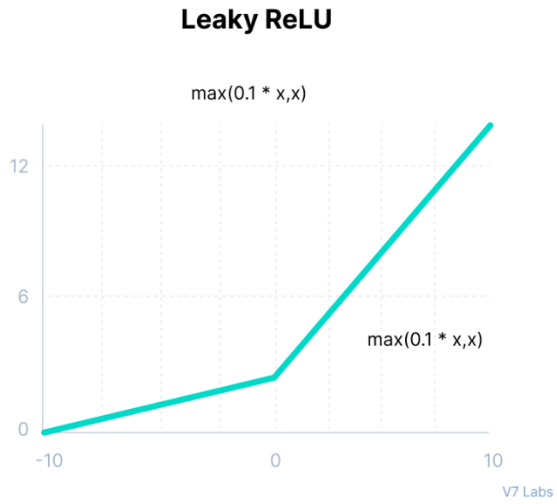
$$ReLU$$

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

Η ταυτόχρονη απενεργοποίηση νευρώνων, λόγω της αρνητικής τους εισόδου, αποτελεί κύριο πλεονέτημα της χρήσης της, ενώ η μη κορεστική γραμμική ιδιότητά της επιταχύνει τη σύγκλιση της συνάρτησης απώλειας (loss function) προς το συνολικό ελάχιστο της.

Από της άλλη μεριά, το γεγονός πως κάποιοι νευρώνες μπορεί να απενεργοποιηθούν αρχικά – λόγω αρνητικής εισόδου - και να μην ενημερωθούν ποτέ ξανά, καθιστά την εκπαίδευση του νευρωνικού δικτύου ανορθόδοξη σε μερικές περιπτώσεις. Το φαινόμενο αυτό ονομάζεται «Dying ReLU problem».

Λύση στο παραπάνω κύριο μειονέκτημα δίνει μια αναβαθμισμένη έκδοση της ReLU :



Εικόνα 11: Leaky ReLU Activation Function

Η διαφοροποίηση έγκειται στην ύπαρξη θετικής και μη όχι μηδενικής παραγώγου σε αρνητικές εισόδους, Έτσι οι «νεκροί» νευρώνες παύουν πλέον να υπάρχουν και το backpropagation πραγματοποιείται ορθά σε κάθε περίπτωση.

$$Leaky\ ReLU$$

$$f(x) = max\ (0.1x, x)$$

Υπάρχουν πολλές επιπρόσθετες παραλλαγές της ReLU, όπως :

1. Parametric ReLU, η οποία γενικεύει την παραπάνω έκδοση της Leaky ReLU, για συντελεστές διαφορετικούς του 0.1
2. ELU, η οποία δίνει ένα εκθετικό – και όχι γραμμικό - χαρακτήρα σε αρνητικές τιμές εισόδου.

2.7.5 Softmax

Softmax

$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

V7 Labs

Η συνάρτηση Softmax διαθέτει τον ίδιο λογικό πυρήνα με την Σιγμοειδή, παραθέτοντας πιθανολογικά αποτελέσματα. Ωστόσο δεν αποτελεί απλά ένα δείκτη από 0 έως 1, αλλά φροντίζει τα αποτελέσματα όλων των νευρώνων του ίδιου στρώματος να αθροίζουν σε 1.

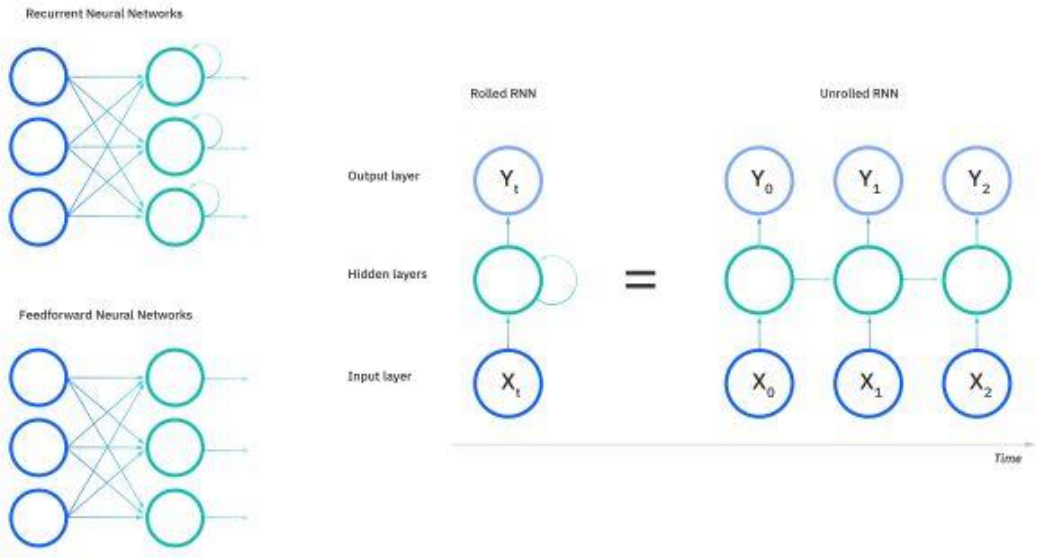
Οι διαφορετικοί νευρώνες του ίδιου στρώματος θεωρούνται ανταγωνιστικά ενδεχόμενα κι έτσι προφανώς πρέπει οι πιθανότητες τους ως προς την ενεργοποίηση να αθροίζουν σε 1.

Η συγκεκριμένη συνάρτηση ενεργοποίησης εφαρμόζεται συνήθως μόνο στο τελικό στρώμα ενός νευρωνικού δικτύου, το οποίο ταυτίζεται με το αποτέλεσμα της κατηγοριοποίησης – πρόβλεψης μιας εισόδου. Με αυτόν τον τρόπο το σύστημα μοιράζει τις πιθανότητες σχετικά με το εάν μια οντότητα ανήκει σε μια κατηγορία ή όχι. Για παράδειγμα σε ένα classification problem αναγνώρισης ζώων μέσω εικόνων, μπορεί το σύστημα να αποφανθεί ότι μια εικόνα παραπέμπει κατά 90% σε ελέφαντα, κατά 8% σε ρινόκερο και κατά 2% σε ιπποπόταμο.

2.8 Επαναληπτικά Νευρωνικά Δίκτυα

Επαναλαμβανόμενο νευρωνικό δίκτυο (RNN) είναι ένας ιδιαίτερος τύπος νευρωνικού δικτύου, στον οποίο συμμετέχουν ακολουθιακά δεδομένα ή χρονοσειρές. Σε αντίθεση με τα παραδοσιακά νευρωνικά δίκτυα που διαθέτουν ανεξάρτητες εισόδους και εξόδους, τα RNNs διακρίνονται για τη μνήμη τους, καθώς πληροφορίες προηγούμενων εισόδων επηρεάζουν το τρέχον αποτέλεσμα [21].

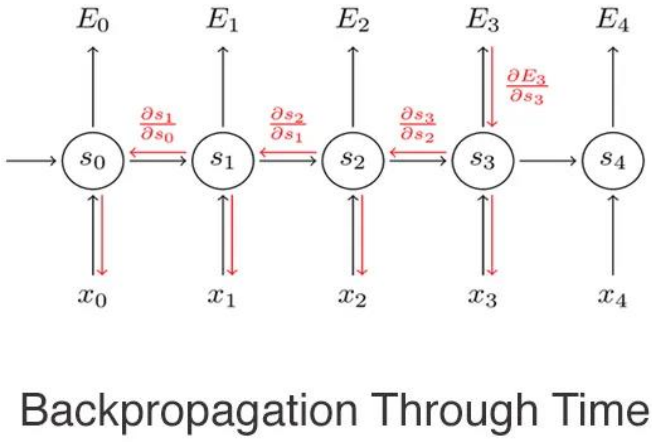
Γενικά τα συγκεκριμένα δίκτυα εκπαιδεύονται με τη σειρά τους σε μεγάλα σύνολα δεδομένων, των οποίων όμως τα χαρακτηριστικά διαθέτουν μια συγκεκριμένη σειρά ή ακολουθία, η οποία είναι καθοριστική κατά την επεξεργασία τους. Εφαρμόζονται κυρίως σε συστήματα speech recognition (Siri), image captioning και natural language processing (Google Translate).



Εικόνα 12: Recurrent Neural Networks

Εκτός από τη θέση ενός χαρακτηριστικού, για παράδειγμα μιας λέξης, η οποία παίζει ρόλο στη σημασία μιας εισόδου, άλλη μια διαφορά των RNNs είναι ότι μοιράζονται τις παραμέτρους κατά μήκος των στρωμάτων του δικτύου. Υπενθυμίζουμε πως στα παραδοσιακά feedforward δίκτυα κάθε κόμβος έχει τα δικά του βάρη. Έτσι, λοιπόν, τα RNNs σε κάθε επανάληψη αναγκάζονται όχι απλά να υπολογίζουν και να αθροίζουν τα σφάλματα, αλλά και να ανανεώνουν και τα συνολικά κοινά βάρη. Συνεπώς, ο αλγόριθμος Backpropagation εκτελείται σε μια τροποποιημένη λογική, αφού το σφάλμα σε κάθε νευρώνα είναι πλέον συσσωρευμένο από τρέχοντα αλλά και παρελθοντικά inputs. Σε κάθε βήμα παραγωγής εξόδου, προσδιορίζεται το σφάλμα και τροφοδοτείται μέσω των

παραγώγων του πίσω σε κάθε προηγούμενο κατάσταση (state). Έτσι τα βάρη συνεχώς ανανεώνονται σε κάθε επανάληψη και κάθε σημείο της αλυσίδας.

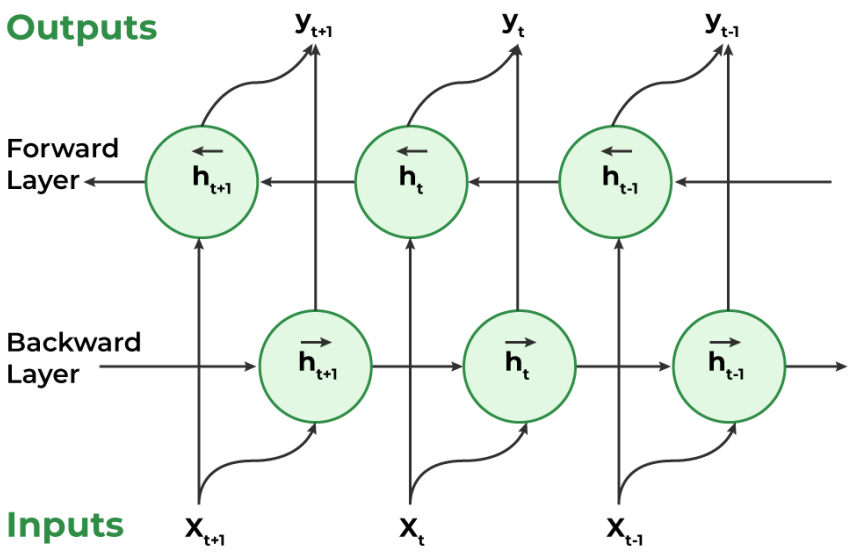


Εικόνα 13: Backpropagation

Εκτός από την κλασική έκδοση ενός RNN, υπάρχουν και πιο εξελιγμένες αρχιτεκτονικές, οι οποίες υπερτερούν σε ακρίβεια, σε πολυπλοκότητα και σε απόδοση. Πάντα βέβαια κατά την επιλογή του καταλληλότερου μοντέλου πρέπει να λαμβάνονται υπόψιν οι ιδιαιτερότητες του εκάστοτε προβλήματος.

2.8.1 Bidirectional RNN

Ουσιαστικά σε αυτήν την εκδοχή προστίθεται η εξάρτηση ενός αποτελέσματος και από μελλοντικές εισόδους. Στην κλασική περίπτωση οι υπολογισμοί αφορούσαν μόνο παροντικές και παρελθοντικές καταστάσεις.



Εικόνα 14: Bidirectional Recurrent Neural Networks

Ουσιαστικά είναι ένας συνδυασμός δύο RNNs, εκείνου που έχει την αρχική φυσιολογική κατεύθυνση, κι εκείνου που έχει την ακριβώς αντίθετη. Η παραγωγή των αποτελεσμάτων στο hidden layer πραγματοποιείται ξεχωριστά για κάθε κατεύθυνση, κι έπειτα οi έξοδοι συνδυάζονται προκειμένου να προκύψει η τελική έξοδος του παρόντος state. Όπως γίνεται αντιληπτό, τα εκπαιδευόμενα βάρη ανήκουν πλέον σε τρεις υποκατηγορίες :

1. Forward path, εκείνο που διατρέχει τα inputs κατά τη φυσιολογική τους φορά
2. Backward path, εκείνο που διατρέχει τα inputs κατά την ανάποδη φορά
3. Paths Combination, εκείνο που συνδυάζει τις επιμέρους ενδιάμεσες εξόδους για να υπολογίσει το τελικό αποτέλεσμα του state.

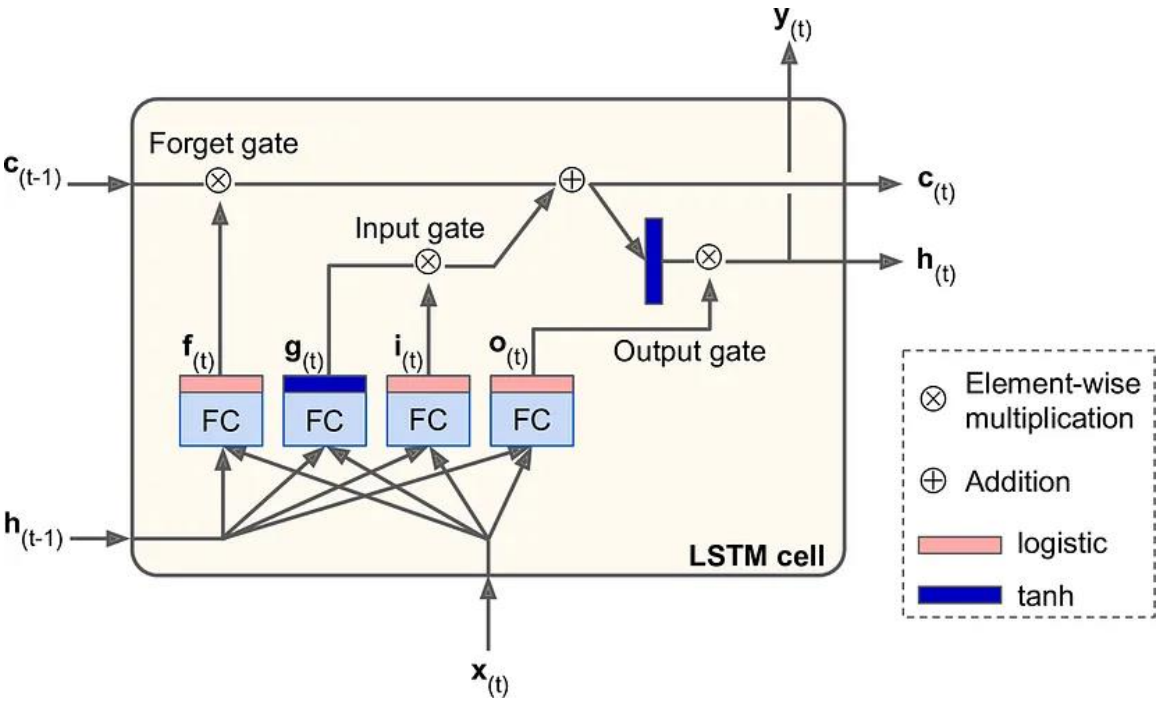
Σαν πλεονεκτήματα, ένα Bidirectional RNN διαθέτει την εμπλουτισμένη ακρίβεια, αφού στα παρελθοντικά στοιχεία συνυπολογίζει και μελλοντικά γεγονότα, ενώ επίσης και την ικανότητα να χειρίζεται πιο πολύπλοκες ακολουθιακές εξαρτήσεις. Από την άλλη σαν μειονεκτήματα εμφανίζεται το αυξημένο υπολογιστικό κόστος κατά την εκπαίδευσή του, η δυσκολία στην παραλληλοποίηση του, καθώς και η μεγαλύτερη πιθανότητα να προκαλέσει προβλήματα overfitting, λόγω ανάπτυξης πολλαπλών παραμέτρων [17].

Overfitting ονομάζεται η κατάσταση κατά την οποία ένα σύστημα μηχανικής μάθησης έχει προπονηθεί με τέτοιον τρόπο έτσι ώστε να αναδεικνύει μεγάλη αποδοτικότητα στο σύνολο δεδομένων που χρησιμοποίησε, αλλά αδυναμία να ανταπεξέλθει σε καινούργιες εισόδους. Με άλλα λόγια, έχει προσαρμοστεί υπέρ το δέον στις παρεχόμενες εισόδους, χάνοντας την ιδιότητα να γενικεύσει τη λειτουργία και την αποτελεσματικότητα του σε νέα δεδομένα.

2.8.2 Long short-term memory

Στη λειτουργία της ανθρώπινης λογικής, πολλές φορές συναντάμε την κατάσταση κατά την οποία η τρέχουσα πραγματικότητα δεν καθορίζεται απαραίτητα από την αμέσως προηγούμενη της, αλλά επηρεάζεται άμεσα από κάποια που ανήκει στο πιο απομακρυσμένο παρελθόν. Η LSTM αρχιτεκτονική έρχεται να δώσει λύσει στο «vanishing gradient» πρόβλημα, μέσω της προσθήκης μακροπρόθεσμων εξαρτήσεων.

Συγκεκριμένα, σχηματίζει κελιά στα hidden layers του νευρωνικού δικτύου, τα οποία έχουν μια πύλη εισόδου, μια πύλη εξόδου και μια πύλη forget, οι οποίες από κοινού ελέγχουν τη ροή της πληροφορίας που είναι απαραίτητη για να παραχθεί η πρόβλεψη του δικτύου. Πέρα από το hidden state και τα παραπάνω gates, εμπεριέχεται και ένα cell state, στο οποίο αποθηκεύεται μακροχρόνια πληροφορία, η οποία τσεκάρεται, αξιοποιείται και εμπλουτίζεται σε κάθε επανάληψη.



Εικόνα 15: Long Short-Term Memory

Κάθε πύλη διαθέτει ενσωματωμένα εκπαιδευόμενα βάρη, ενώ επίσης είναι υπεύθυνη για τη μεταφορά πληροφορίας μεταξύ του hidden και του cell state. Συγκεκριμένα, η πύλη forget είναι υπεύθυνη για την πληροφορία που η παρούσα κρυμμένη κατάσταση οφείλει να «ξεχάσει» ή να «θυμηθεί» από το προηγούμενο cell state. Η πύλη input είναι υπεύθυνη για την πληροφορία που πρέπει να προστεθεί ή να παραμεληθεί στο cell state.

Τέλος, η πύλη output είναι υπεύθυνη για την πληροφορία που πρέπει να διοχετευτεί από το παρόν cell state στο παρόν hidden state [3].

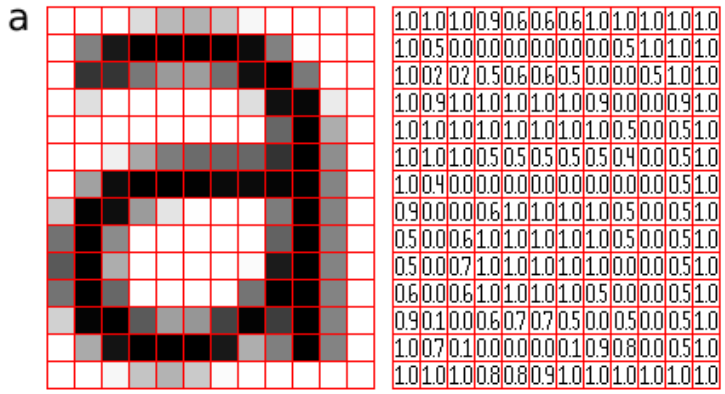
Τα θετικά ενός LSTM επαναλαμβανόμενου νευρωνικού δικτύου περικλείονται γύρω από τις αρετές της διαλειτουργικότητας, μέσω των κελιών μνήμης, της ευελιξίας, λόγω της πληθώρας προβλημάτων που μπορεί να αντιμετωπίσει, και της ανθεκτικότητας σε θορυβώδη δεδομένα. Ωστόσο, η πολυπλοκότητα αυτών των δικτύων πρέπει να συμβαδίζει με το μέγεθος του συνόλου δεδομένων στο οποίο προπονούνται, καθώς τα υπολογιστικά κόστη κρίνονται ακριβά, ενώ φαινόμενα overfitting μπορούν εύκολα να παρουσιαστούν σε μικρά datasets.

2.9 Συνελικτικά Νευρωνικά Δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα είναι μια κλάση νευρωνικών δικτύων που εξειδικεύεται στην επεξεργασία δεδομένων που διαθέτουν τοπολογία πλέγματος, όπως μια εικόνα [33].

Σε αντίθεση με τα feedforward νευρωνικά δίκτυα, στα οποία τοποθετούνται χαρακτηριστικά ανεξάρτητα μεταξύ τους ως προς τη στοίχιση, και με τα recurrent νευρωνικά δίκτύα, στα οποία τοποθετείται ταξινομημένη είσοδος με γραμμική συνεκτικότητα, στα συνελικτικά νευρωνικά δίκτυα τα δεδομένα εισόδου έχουν επιφανειακή ή χωρική αλληλεξάρτηση.

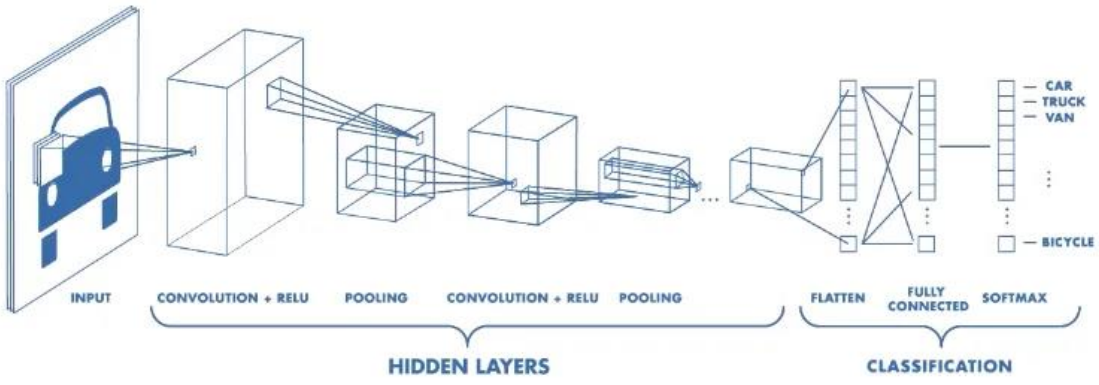
Για παράδειγμα, μια ψηφιακή εικόνα αποτελείται από μια σειρά από pixels, πλήρως διατεταγμένα σε ένα ορθογώνιο πλέγμα. Το κάθε ένα από αυτά εκπροσωπεί ένα σημείο της εικόνας, και ταυτίζεται με μια κανονικοποιημένη αριθμητική τιμή, η οποία δείχνει την φωτεινότητα ή την απόχρωση του συγκεκριμένου σημείου. Από εκεί και πέρα, οι διαστάσεις κάθε δεδομένου εισόδου μπορεί να είναι θεωρητικά άπειρες, αλλά στην περίπτωση των εικόνων, η οποία είναι και η δημοφιλέστερη, συνήθως χρησιμοποιούνται οι κλίμακες Grayscale και GBR, για ασπρόμαυρες και έγχρωμες εικόνες αντίστοιχα.



Εικόνα 16: Image in pixels

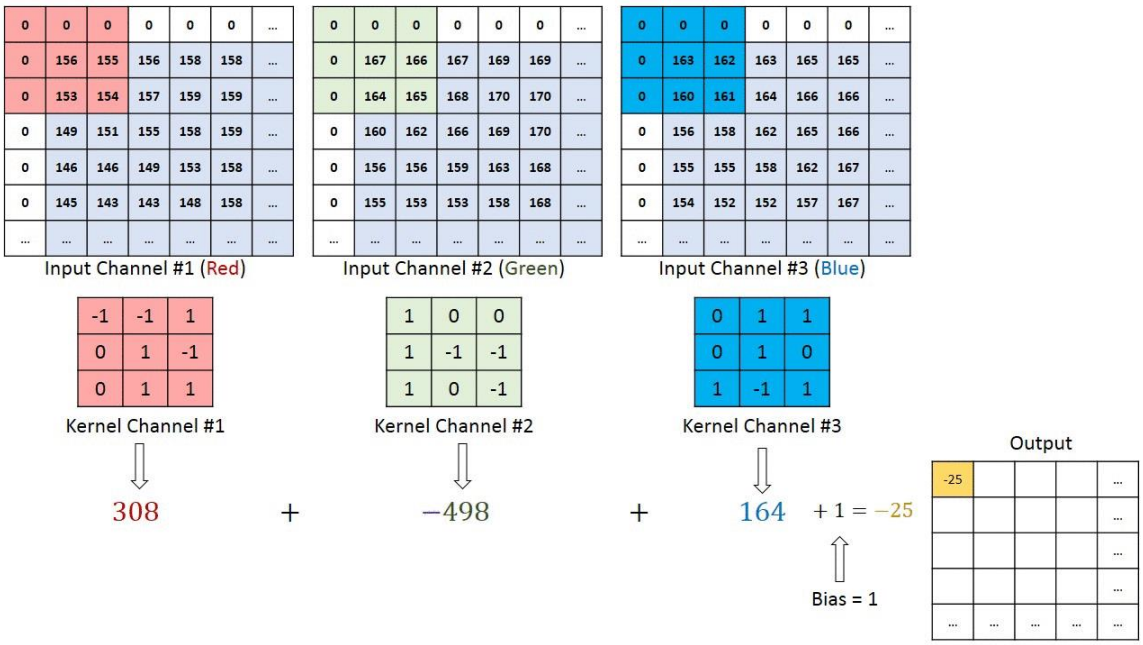
Τα συνελικτικά νευρωνικά δίκτυα αποκτούν άμεση σημασία και εφαρμογή σε computer vision προβλήματα και σε κατηγοριοποιήσεις εικόνων, ενώ η αρχιτεκτονική τους είναι εμπνευσμένη από τον τρόπο με τον οποίο οι νευρώνες του ανθρώπινου εγκεφάλου συνδέονται και συνεργάζονται με τον οπτικό φλοιό. Συγκεκριμένα, το ανθρώπινο βιολογικό νευρολογικό σύστημα βασίζεται αφενός στη μεμονωμένη λήψη του ερεθίσματος που προέρχεται από μια περιορισμένη οπτικά περιοχή (Receptive Field), και αφετέρου στη συλλογή και σύνθεση όλων των δυνατών οπτικών ερεθισμάτων που καλύπτουν μια ολόκληρη όψη. Με αυτόν τον τρόπο, οι επιμέρους επεξεργασίες συναληθεύονται και οδηγούν στην τελική ολοκληρωμένη εξαγωγή πληροφορίας.

Η αρχιτεκτονική ενός CNN αποτελείται από τρία στρώματα : το συνελικτικό στρώμα (convolutional layer), το στρώμα ομαδοποίησης (pooling layer) και ένα πλήρως συνδεδεμένο στρώμα (fully connected layer) [20].



Εικόνα 17: Αρχιτεκτονική ενός Convolutional Network

Το συνελικτικό στρώμα αποτελεί το πιο φορτισμένο υπολογιστικά τμήμα του δικτύου, αφού η κύρια λειτουργία του είναι ο πολλαπλασιασμός πινάκων. Συγκεκριμένα υπολογίζεται το εσωτερικό γινόμενο δύο πινάκων. Ο πρώτος, γνωστός και ως πυρήνας (kernel), απαρτίζεται από τα εκπαιδευόμενα βάρη του δικτύου, ενώ ο δεύτερος ταυτίζεται με τα τις τιμές των pixels μιας συγκεκριμένης περιοχής της εικόνας. Ο πολλαπλασιασμός πραγματοποιείται για κάθε περιοχή της συνολικής εικόνας και ως αποτέλεσμα λαμβάνεται ο χάρτης ενεργοποίησης (activation map), δηλαδή ένας διδιάστατος πίνακας, ο οποίος θα τροφοδοτηθεί στο επόμενο στρώμα του δικτύου [46].



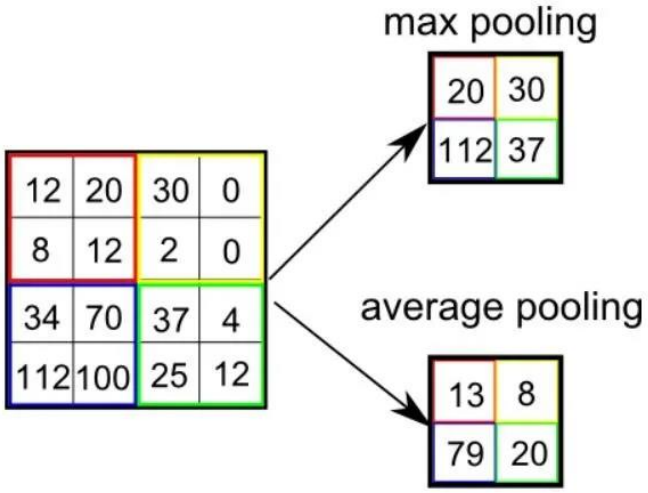
Εικόνα 18: Συνέλιξη της πληροφορίας

Με αυτόν τον τρόπο αναδεικνύονται και εξάγονται κάποια χαρακτηριστικά της εικόνας, τα οποία μπορούν έπειτα να χρησιμοποιηθούν και να τεθούν σε επεξεργασία. Κάθε καινούργιο επίπεδο συνέλιξης που προστίθεται στο συνελικτικό στρώμα αυξάνει το βάθος της σημασίας των χαρακτηριστικών που εξάγονται. Με άλλα λόγια, το πρώτο και αρχικό επίπεδο είναι υπεύθυνο για τα low level features, όπως το χρωματισμό και την αντίθεση των τμημάτων της εικόνας, ενώ τα επόμενα επίπεδα εμβαθύνουν ολοένα και περισσότερο στην ουσία και το περιεχόμενο της εικόνας, καταλήγοντας έτσι να παρέχουν high level features σχετικά με αυτήν. Έτσι επιτυγχάνεται ολοκληρωμένη και ποιοτική κατανόηση των εικόνων, η οποία μπορεί να οδηγήσει σε ορθότερη κατηγοριοποίηση όσον αφορά το περιεχόμενό της.

Επόμενο τμήμα επεξεργασίας ενός συνελκτικού δικτύου CNN είναι το pooling layer, του οποίου η προσφορά έγκειται σε δύο στοιχεία. Πρώτον, τη μείωση των διαστάσεων των activation maps μέσω στατιστικών συναρτήσεων. Έτσι μετριάζεται η πολυπλοκότητα και

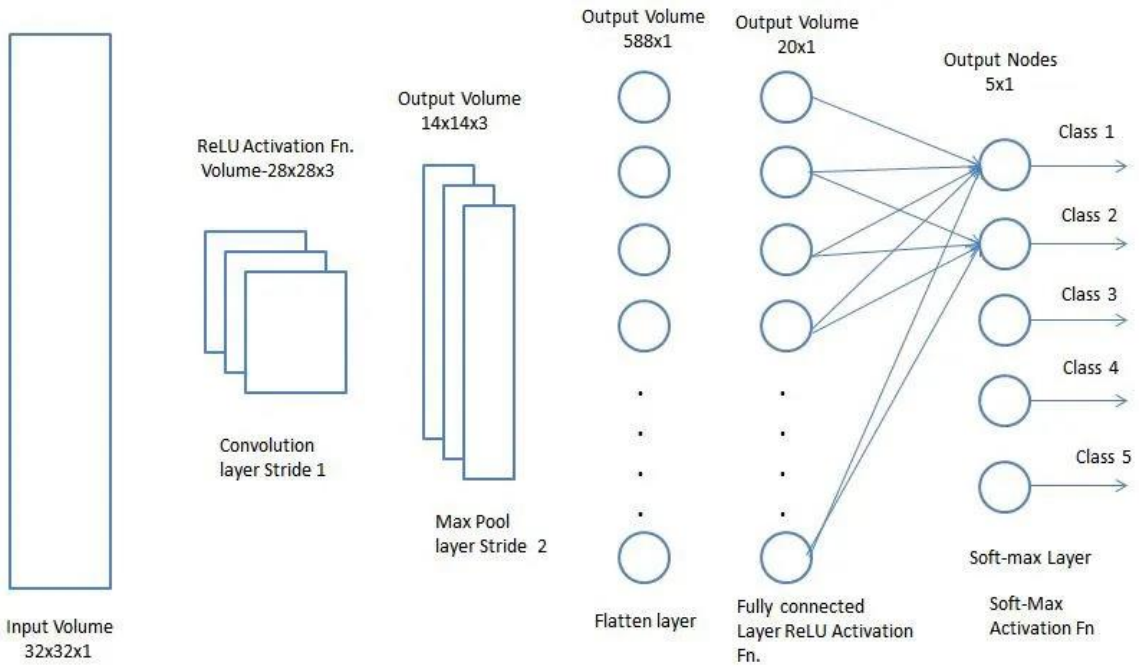
το υπολογιστικό κόστος του συστήματος, τα οποία δεν είναι απαραίτητα συνδεδεμένα και με μεγαλύτερη απόδοση και ακρίβεια. Με άλλα λόγια μειώνεται το μέγεθος της εισόδου που θα τροφοδοτηθεί στο fully connected νευρωνικό δίκτυο, χωρίς να επηρεάζεται αρνητικά η αποτελεσματικότητα του. Δεύτερον, καταστέλλεται ο θόρυβος πληροφορίας που είναι ενδεχομένως να είναι διασκορπισμένος στο χάρτη ενεργοποίησης κατά γειτονιές [34].

Σαν στατιστικές συναρτήσεις χρησιμοποιούνται οι max, min, average και παραλλαγές τους. Κάθε μια διαθέτει θετικά και αρνητικά, ενώ η επιλογή ανάμεσα τους συχνά εξαρτάται από το επιμέρους πρόβλημα και αποτελεί αντικείμενο βελτιστοποίησης του συνολικού νευρωνικού δικτύου. Τα επιμέρους στοιχεία κάθε περιοχής του activation map συμπτύσσονται σε μια τιμή.



Εικόνα 19: Pooling Layer in Convolutional Networks

Σαν τελικό στάδιο του δικτύου τοποθετείται ένα πλήρως συνδεδεμένο νευρωνικό δίκτυο, το οποίο λαμβάνει είσοδο τα high level features της εικόνας, όπως αυτά παρουσιάζονται στον activation map. Αφού πρώτα ισοπεδωθεί (flatten) ο τελικός πίνακας, πλέον η αρχική εικόνα έχει μετατραπεί πλήρως σε μια κατάλληλη κανονική μορφή εισόδου ενός feedforward νευρωνικού δικτύου. Μέσω αυτού του δικτύου, του αλγορίθμου backpropagation, και των συναρτήσεων ενεργοποίησης, το σύστημα θα προπονηθεί και θα είναι τελικά ικανό να αποφανθεί σχετικά με την κατηγοριοποίηση των εικόνων του dataset.



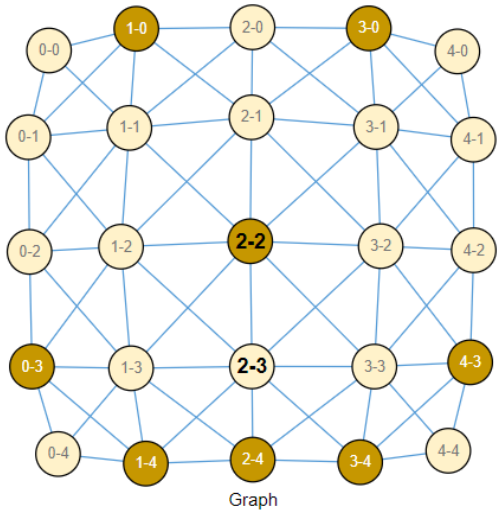
Εικόνα 20: Datapath in Convolutional Networks

2.10 Νευρωνικά Δίκτυα Γράφων

Νεωρωνικά δίκτυα γράφων είναι εκείνα τα νευρωνικά δίκτυα που εξειδικεύονται στην επεξεργασία δεδομένων των οποίων η αναπαράσταση παραπέμπει σε γράφημα. Ο πρωταρχικός στόχος της αρχιτεκτονικής ενός GNN είναι η εκμάθηση και η αξιοποίηση πληροφορίας που προέρχεται από την ευρύτερη γειτονιά του συνόλου δεδομένων [39].

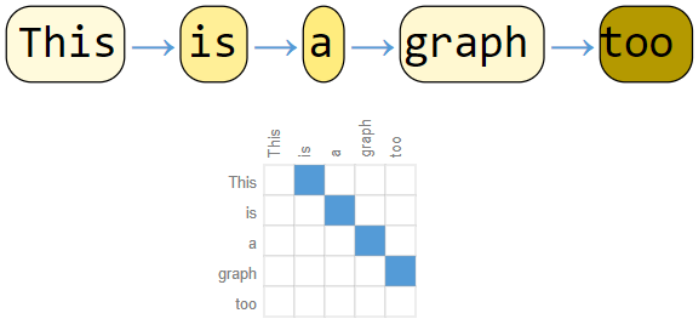
Η λειτουργικότητα άλλων νευρωνικών δικτύων, όπως των feedforward, recurrent και concolutional neural networks, δεν είναι ικανή να καλύψει τις ανάγκες όλων των συνόλων δεδομένων. Συγκεκριμένα τα δεδομένα που διαθέτουν τη μορφή γράφων έχουν δομημένη μορφή, που αφενός δεν συμπεριλαμβάνεται στην απλοϊκή λογική των feedforward και αφετέρου δε προσαρμόζεται στη γραμμική προσέγγιση των recurrent. Ταυτόχρονα, είναι υπολογιστικά ασύμφορο για ένα CNN να επιχειρήσει να ανταπεξέλθει σε γραφήματα. Αυτό συμβαίνει διότι οι γράφοι από τη φύση τους έχουν μια πολύπλοκη και αυθαίρετη τοπολογία, της οποίας η ιδιομορφία είναι η απουσία χωροταξίας (spatial locality).

Τα δεδομένα που παρουσιάζονται με τη μορφή γράφων βρίσκονται παντού στην καθημερινότητά μας. Μια εικόνα είναι ουσιαστικά ένα γράφημα, του οποίου τα στοιχεία pixels διαθέτουν μια προσωπική τιμή (σε μια ή τρεις διαστάσεις) και συνδέονται μέσω ακμών με τα γειτονικά τους αντίστοιχα στοιχεία pixels [40].



Εικόνα 21: Τετριμμένος γράφος

Μια πρόταση, ή γενικά ένα γλωσσικό κείμενο, είναι επίσης ένα γράφημα, του οποίου οι κόμβοι ταυτίζονται με τις λέξεις και του οποίου οι ακμές είναι συνδέσεις μεταξύ διαδοχικών λέξεων [41].



Εικόνα 22: Μια πρόταση ως απλός γράφος

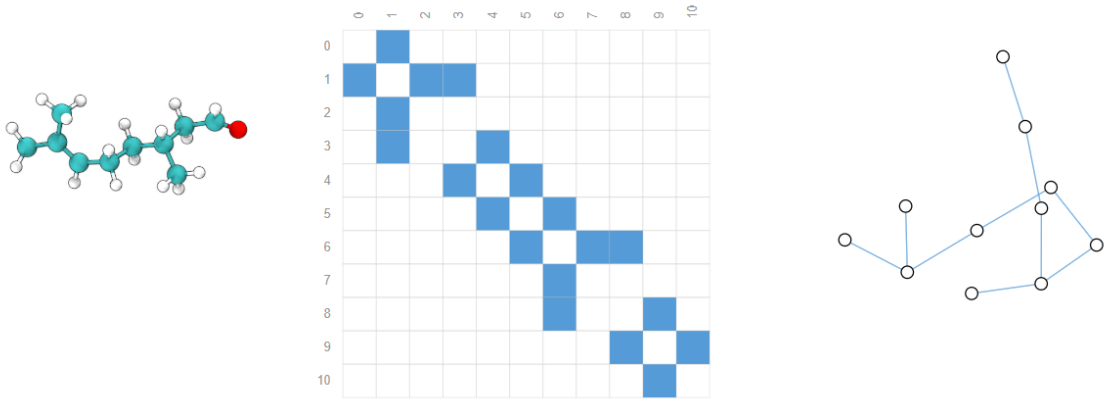
Πέρα όμως από τις παραπάνω απλές μορφές γράφων, στην καθημερινότητα και στη φύση κυριαρχούν κυρίως πιο σύνθετα και απρόβλεπτα σχήματα, όπως κοινωνικά δίκτυα, χημικές ενώσεις, μαθηματικές εξισώσεις, συνδέσεις πηγών πληροφορίας, και άλλα. Η ετερογένεια, η ποικιλομορφία και η απουσία συμμετρικότητας στις παραπάνω περιπτώσεις είναι κοινά γνωρίσματα που καθιστούν τη συμμετοχή των GNN αναγκαία.

Γενικά, υπάρχουν τρία είδη προβλημάτων σε δομημένα σύνολα δεδομένων :

1. Graph-level classifications, στα οποία έχουμε ως στόχο την κατηγοριοποίηση όλου του γράφου

2. Node-level classifications, στα οποία έχουμε ως στόχο την τιτλοφόρηση των κορυφών του γράφου, δηλαδή των στοιχείων του
3. Edge-level classifications (link predictions), στα οποία έχουμε ως στόχο την πρόβλεψη της ύπαρξης και της αξίας των ακμών του γράφου

Ένα χαρακτηριστικό παράδειγμα για την πρώτη περίπτωση είναι η εξαγωγή συμπερασμάτων για ένα μόριο, του οποίο τα στοιχεία και οι ενώσεις μπορούν να αναπαρασταθούν ως ένα συνεκτικό γράφημα. Με λίγα λόγια, μπορεί να χρειάζεται να υλοποιηθεί ένα σύστημα που κρίνει την καταλληλότητα του μορίου όσον αφορά τη σύνδεσή του με τους υποδοχείς που εμπλέκονται σε μια ασθένεια. Το συγκεκριμένο πρόβλημα είναι ανάλογο της κατηγοριοποίησης εικόνων, στο οποίο θέλουμε να αποφανθούμε σχετικά με το περιεχόμενο μιας ολόκληρης εικόνας.

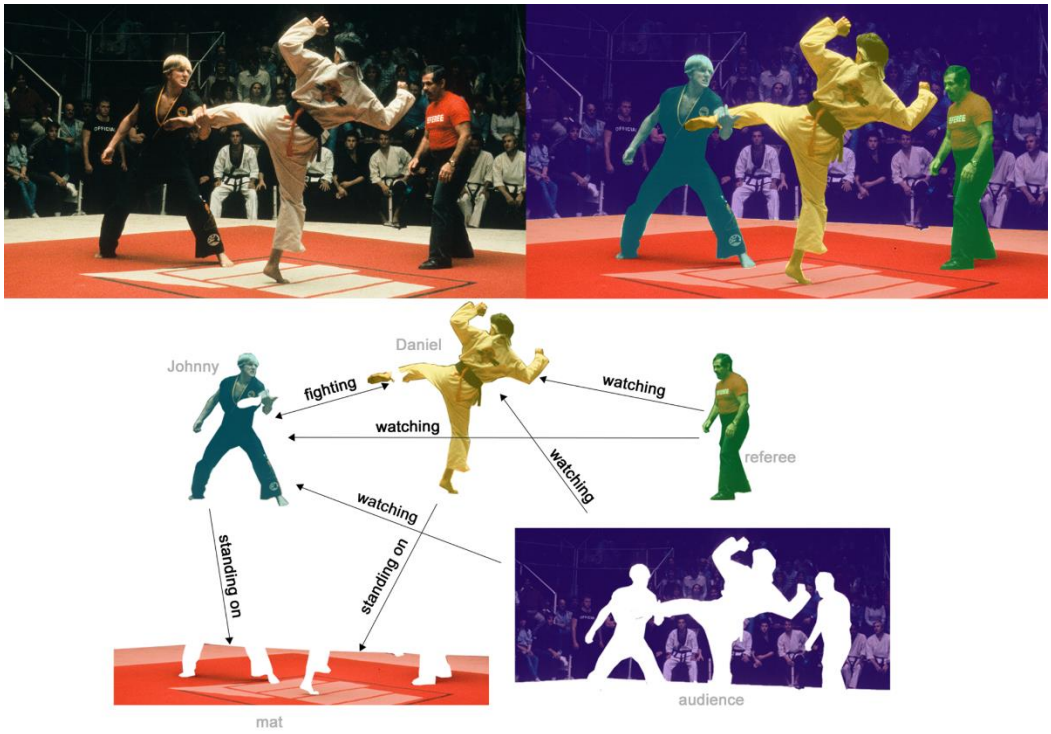


Εικόνα 23: Πίνακας γειτνίασης γράφου που αναπαριστά ένα μόριο

Ένα παράδειγμα για τη δεύτερη περίπτωση αποτελεί η πρόβλεψη μιας ιδιότητας που κατέχει ένας άνθρωπος σε έναν κοινωνικό κύκλο. Συγκεκριμένα, μια κοινότητα κατοίκων σε ένα χωριό σχηματίζει ένα γράφημα, έχοντας ως κόμβους τους κατοίκους και ως ακμές τις συμπάθειες μεταξύ τους. Με βάση και κάποια επιπρόσθετα προσωπικά χαρακτηριστικά των κατοίκων, όπως φύλο, ηλικία, επάγγελμα, κλπ, μπορεί να υλοποιηθεί ένα σύστημα το οποίο να προβλέπει το μορφωτικό τους επίπεδο, ή να τους κατηγοριοποιεί σε ευτυχισμένους και μη.

Εικόνα 24: Κοινωνικό γράφημα

Ένα παράδειγμα της τρίτης περίπτωσης δεν είναι άλλο παρά η αναγνώριση των συνδέσεων μεταξύ αντικειμένων που παρουσιάζονται σε μια εικόνα. Με λίγα λόγια, αφού προκαθοριστούν τα αντικείμενα που συμμετέχουν σε μια εικόνα, καθώς και τα χαρακτηριστικά τους, σκοπός του συστήματος είναι να προβλέψει τις μεταξύ τους σημασιολογικές και οπτικές σχέσεις. Δηλαδή ένας κόμβος που εκπροσωπείται από έναν άνθρωπο μπορεί να συνδέεται με έναν άλλον κόμβο του γράφου που εκπροσωπείται από το πάτωμα του χώρου, μέσω της ακμής «standing on».



Εικόνα 25: Σημασιολογικό γράφημα

Σε όλες τις παραπάνω περιπτώσεις, ο γράφος που σχηματίζεται από τα δεδομένα του προβλήματος έχει ως βασική παράμετρο το επίπεδο συνεκτικότητας του. Το πλήθος των ακμών που συνυπάρχουν σε έναν γράφο επηρεάζουν άμεσα τη λειτουργία ενός GNN, και κατ' επέκταση αποτελούν σημείο αναφοράς ως προς την βελτιστοποίηση του.

Σαν input σε ένα GNN παρατίθενται τα χαρακτηριστικά κάθε κορυφής του γράφου, καθώς και όλες οι συνδέσεις μεταξύ τους. Το βασικό γνώρισμα της λειτουργίας ενός GNN είναι ο συνδυασμός των ατομικών χαρακτηριστικών – embeddings κάθε κόμβου με όλες τις αντίστοιχες πληροφορίες των άμεσων γειτόνων του, με στόχο την τελική αναπαράσταση του κόμβου. Αυτή η διαδικασία πραγματοποιείται επαναληπτικά, κι έτσι τα embeddings όλων των κορυφών του γράφου ανανεώνονται συνεχώς, αξιοποιώντας πληροφορία όλο και πιο απομακρυσμένων γειτόνων. Κατά μια έννοια τα χαρακτηριστικά κάθε κορυφής του γραφήματος «ταξιδεύουν» κατά μήκος όλων των πιθανών μονοπατιών και επηρεάζουν διαβαθμισμένα τις αναπαραστάσεις ξένων κορυφών. Αυτή η δυναμική συνεχής ενημέρωση και ανανέωση του γράφου οδηγεί σε τελικά συμπεράσματα σχετικά με τις κορυφές του. Φυσικά αναλύσαμε την περίπτωση των node level classifications, ωστόσο η λογική και των άλλων classifications δεν απέχει αρκετά.

Από μαθηματική οπτική, τα embeddings υπολογίζονται με βάση την εξής έκφραση :

$$\mathbf{H} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\mathbf{\Theta}\right)$$

όπου :

H = nodes representations matrix (οι υπολογισμένες αναπαραστάσεις των κορυφών),

X = nodes features (τα χαρακτηριστικά των κορυφών),

A = graph adjency matrix (πίνακας γειτνίασης του γραφήματος),

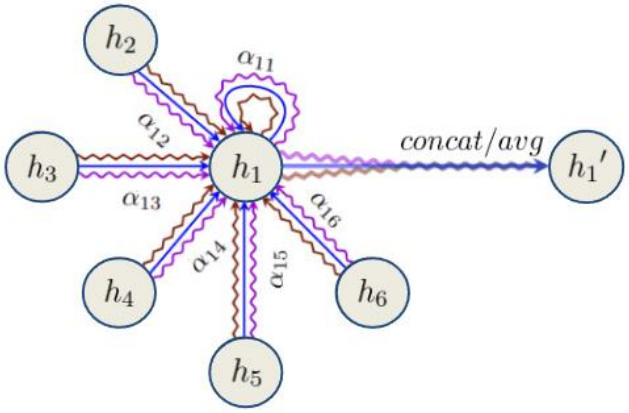
D = graph degree matrix (πίνακας βαθμών των κορυφών του γραφήματος)

Θ = trainable parameters matrix (πίνακας εκπαιδευόμενων παραμέτρων)

σ = activation function (συνάρτηση ενεργοποίησης)

Ένα βασικό πλεονέκτημα των GNNs είναι η ικανότητα τους να ανταπεξέρχονται σε supervised αλλά και unsupervised σύνολα δεδομένων, όντας έτσι ευέλικτα και πολύπλευρα σε πολλά προβλήματα κατηγοριοποίησης. Από την άλλη, ένας περιορισμός τους είναι το υπολογιστικό τους κόστος, καθώς οι γράφοι σαν δομές δεδομένων έχουν εκθετική κλιμάκωση ως προς την πολυπλοκότητα τους. Ταυτόχρονα, όταν τα δεδομένα διαθέτουν θόρυβο ή είναι ελλιπή, τότε αυξάνεται η πιθανότητα του overfitting κατά την εκπαίδευση του μοντέλου.

Μια εξελιγμένη παραλλαγή των Graph Neural Networks (GNNs) είναι τα Graph Attention Networks (GATs). Η διαφοροποίηση βρίσκεται στον τρόπο με τον οποίο ο κάθε κόμβος αξιοποιεί τα δεδομένα που του προσφέρουν οι γείτονές του. Στην default περίπτωση των GNNs, η ανανέωση των embeddings πραγματοποιείται μέσω της λογικής του μέσου όρου, δηλαδή μέσω του υπολογισμού του μέσου όρου των χαρακτηριστικών της άμεσης γειτονιάς. Στην περίπτωση των GATs, ωστόσο, τα βάρη που καθορίζουν τη συμβολή κάθε γείτονα στον current κόμβο, δεν είναι σταθερά και ισορροπημένα, αλλά εκπαιδευόμενα και μεταβλητά. Με αυτόν τον τρόπο, βελτιώνεται ο τρόπος με τον οποίο οι συνδέσεις – ακμές επηρεάζουν και ανανεώνουν τα embeddings των κορυφών [12].



Εικόνα 26: Μηχανισμός προσοχής σε ένα γράφημα

Από σημασιολογική άποψη η συγκεκριμένη τακτική έχει έντονο νόημα, αφού δεν είναι πάντα όλες οι συνδέσεις ισότιμες μεταξύ τους, είτε μιλώντας για κοινωνικά, είτε για γενικότερα δίκτυα οντοτήτων. Τα νευρωνικά δίκτυα που διαθέτουν τέτοιους self attention μηχανισμούς παρουσιάζουν μια ευελιξία και αυτονομία ως προς την προσαρμογή τους στα δεδομένα εισόδου, αν και φυσικά αυξάνουν την πολυπλοκότητα και τα υπολογιστικά κόστη του συστήματος.

2.11 Επεξεργασία Φυσικής Γλώσσας

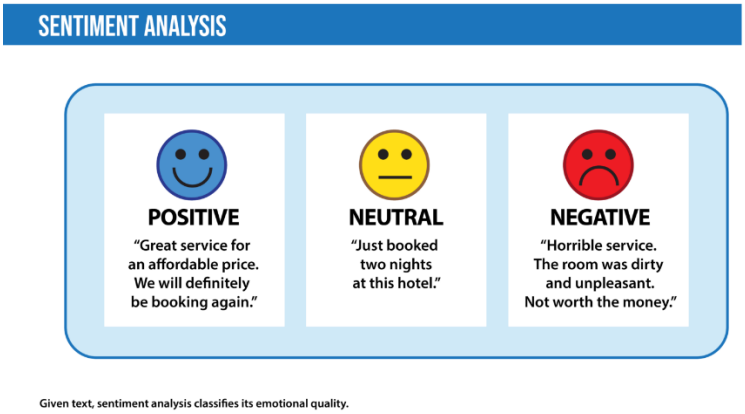
Η επεξεργασία φυσικής γλώσσας (NLP) αναφέρεται στον τομέα της Τεχνητής Νοημοσύνης που σχετίζεται με την ικανότητα ενός υπολογιστικού συστήματος να καταλαβαίνει γραπτό και προφορικό κείμενο με τρόπο αντίστοιχο του ανθρώπου.

Συγκεκριμένα, ένα σύστημα NLP χρησιμοποιεί γλωσσολογικά μοντέλα της ανθρώπινης γλώσσας, και συνδυάζοντάς τα με στατιστικές μεθόδους και αρχιτεκτονικές μηχανικής και βαθιάς μάθησης, καταφέρνει να αντιλαμβάνεται voice ή text δεδομένα στο πλήρες νόημα τους, αποκωδικοποιώντας τις προθέσεις και τα συναισθήματα του χρήστη [24].

Τα προγράμματα που χρησιμοποιούν NLP βρίσκονται παντού ανάμεσά μας, όπως για παράδειγμα σε GPS συστήματα, σε εφαρμογές που περιέχουν speech-to-text εντολές, σε chatbots εξυπηρέτησης πελατών, και πολλά άλλα. Ωστόσο πλέον τα NLP επιχειρούν να δώσουν λύσεις και αυτοματισμούς και σε λειτουργίες επιχειρήσεων, όσον αφορά την οργάνωση και την ανάλυση εταιρικών διαδικασιών, αυξάνοντας έτσι την παραγωγικότητα των εργαζομένων και την διοικητική απόδοση.

Πιο συγκεκριμένα, η επεξεργασία φυσικής γλώσσας (NLP) αποκτά ουσία και υπόσταση στα παρακάτω tasks [9]:

1. Sentiment analysis, δηλαδή η διαδικασία κατηγοριοποίησης ενός γραπτού κειμένου σε κάποιο συναισθηματικό υπόβαθρο. Η πιο απλή κλίμακα που μπορεί να χρησιμοποιηθεί είναι εκείνη που δίνει positive, negative και neutral υπόσταση σε κάποιο γλωσσικό δεδομένο.



Εικόνα 27: Ανύλυση συναισθήματος

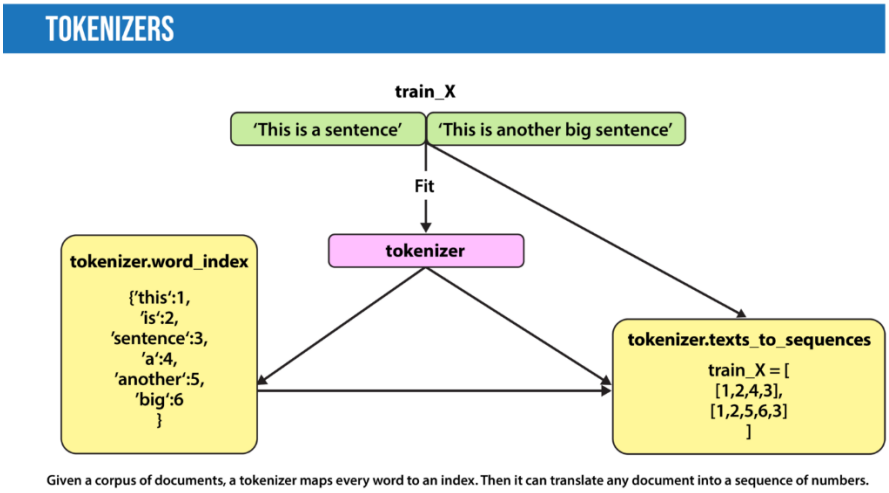
Μια πιο σύνθετη ανάλυση που μπορεί να πραγματοποιηθεί είναι εκείνη που ανιχνεύει την επιθετικότητα, την τοξικότητα και την αισχρότητα σε σχόλια ή μηνύματα, με σκοπό να τα αποκλείσει από μια συζήτηση ή ένα αρχείο.

2. Machine translation, δηλαδή αυτοματοποιημένη μετάφραση από γλώσσα σε γλώσσα. Έχει ως στόχο την σύνδεση ή την αναγνώριση λέξεων όσον αφορά το νόημα τους. Οι πιο γνωστές εφαρμογές της συναντώνται σε πλατφόρμες κοινωνικής δικτύωσης, αλλά και μηχανές αναζήτησης, βελτιώνοντας την επικοινωνία ανάμεσα στους ανθρώπους.
3. Named entity recognition, δηλαδή η εξαγωγή ονόματος ή ετικέτας ή τίτλου σε ένα γλωσσικό δεδομένο. Λαμβάνει συνήθως ως είσοδο μια λέξη ή πρόταση και δίνει σαν έξοδο το tag της, δηλαδή μια προκαθορισμένη κατηγορία στην οποία ανήκει, όπως Location, Quantity, Name και άλλα.
   Σε εντελώς αντίστοιχη λογική κυμαίνεται και το Topic Modeling, δηλαδή το task το οποίο ανακαλύπτει αφηρημένα topics σε μια λίστα από αρχεία. Με άλλα λόγια δίνει μια γκάμα από topics, κάθε ένα από τα οποία καθορίζεται από αντιπροσωπευτικές λέξεις και αντιστοιχεί αναλογικά σε ένα κείμενο. Τέτοιου είδους μοντέλα αξιοποιούνται κατά κόρον σε δικηγορικές επιχειρήσεις, με σκοπό την ανίχνευση τεκμηρίων.
4. Spam detection. Είναι το διαδεδομένο πρόβλημα δυαδικής ταξινόμησης, το οποίο επιχειρεί να κατηγοριοποιήσει τα emails σε normal ή spam. Λαμβάνουν ως παραμέτρους όχι μόνο το περιεχόμενο, αλλά και τα χαρακτηριστικά ενός email, έτσι ώστε να προστατεύσουν το χρήστη από κακόβουλα μηνύματα και ανεπιθύμητες αλληλογραφίες.
5. Grammatical error correction. Σκοπεύει να δώσει μια αυξημένη εμπειρία στο χρήστη που γράφει, επεξεργάζεται και δημοσιεύει ένα κείμενο. Προπονημένο με δοσμένες σωστές γραμματικά προτάσεις, είναι ικανό να διορθώνει τη γραμματική σε ένα κείμενο.
6. Text generation (NLG), το οποίο παράγει text που προσομοιάζει την ανθρώπινη ομιλία, είτε γραπτή είτε προφορική. Εφαρμόζεται σε συστήματα autocomplete, δηλαδή πρόβλεψης ολοκλήρωσης ή διαδοχικής λέξης, καθώς και σε chatbots τα οποία έχουν δημιουργηθεί για να παρέχουν ζωντανή εξυπηρέτηση σε πελάτες μέσω διαλόγου εφάμιλλου του ανθρώπου.

Υπάρχουν και περαιτέρω εφαρμογές, όπως Summarization, Information Retrieval και Question Answering, τα οποία όμως κινούνται όλα στο ίδιο μήκος κύματος με τα προηγούμενα, δηλαδή στην κατανόηση γραπτού κειμένου και την παραγωγή χρήσιμης πληροφορίας σχετικά με αυτά. Αυτή η χρήσιμη πληροφορία πολλές φορές δεν είναι άλλη από την παραγωγή καινούργιου γλωσσικού δεδομένου, του οποίου η προέλευση (ανθρώπινη ή μηχανική) δεν μπορεί να προσδιοριστεί με ευκολία. Με βάση το τελευταίο κρίνεται και η απόδοση ενός NLP – NLG.

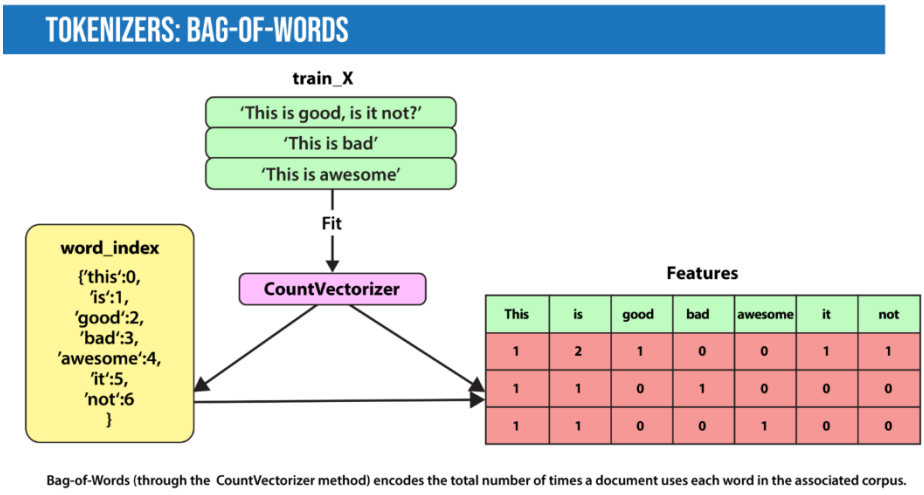
Η λειτουργία ενός μοντέλου NLP βασίζεται στην εύρεση συσχετίσεων ανάμεσα στα κύρια συστατικά ενός γλωσσικού συνόλου δεδομένων, δηλαδή γράμματα, λέξεις και προτάσεις. Η αρχιτεκτονική τους επιμερίζεται σε τρία κύρια μέρη :

1. Data preprocessing, δηλαδή την προετοιμασία των δεδομένων που πρόκειται να εισέλθουν στο μοντέλο, μορφοποιώντας τα κατάλληλα έτσι ώστε να είναι συμβατά με τις ανάγκες του μοντέλου και να βελτιώσουν την απόδοσή του.
   Αρχικά οι λέξεις και οφείλουν να συμπτυχθούν στην αρχική τους γραμματική βάση, προκειμένου να μην έχουμε σύνθετες, παράγωγες ή παραλλαγές της ίδιας κατά βάση λέξης. Για παράδειγμα οι λέξεις «σχολείο», «σχολικός» και «σχολείων» έχουν την ίδια προέλευση και δεν πρέπει να θεωρηθούν ως διαφορετικές. Η σύμπτυξη αυτή των λέξεων λαμβάνει χώρα μέσω δύο τρόπων, του Stemming και του Lemmatization. Ο πρώτος επικεντρώνεται στα αρχικά κοινά γράμματα των λέξεων, τα οποία σηματοδοτούν πως διαθέτουν την ίδια προέλευση, ενώ ο δεύτερος και πιο επίσημος τρόπος ουσιαστικά ανιχνεύει την γλωσσική ρίζα αυτή κάθε αυτή, χρησιμοποιώντας λεξιλόγιο. Είναι προφανές πως με το Lemmatization αποφεύγονται προβλήματα ψευδών συμπτύξεων μέσω Stemming, όπως των λέξεων «παρακάμπτω» και «παρακάμερα», των οποίων η αρχική ακολουθία γραμμάτων «παρακαμ» δεν αντιπροσωπεύει την βασική ταυτότητα τους, η οποία είναι διαφορετική.
   Έπειτα ακολουθείται η διαγραφή και αποφυγή λέξεων που δεν προσφέρουν ιδιαίτερα πολλά σε μια πρόταση ή γενικά σε ένα γλωσσικό κείμενο. Φυσικά αναφερόμαστε κυρίως στα άρθρα (πχ «το», «η», «των»), τα οποία δεν προσθέτουν νόημα και πληροφορία στα δεδομένα. Το σύνολο αυτών των λέξεων αποκαλείται «stop words».
   Τέλος, έχουμε το Tokenization, μέσω του οποίου κάθε λέξη αποκτά ένα ξεχωριστό αναγνωριστικό σε αριθμητική μορφή. Αυτό το token θα εκπροσωπεί την λέξη στην μετέπειτα επεξεργασία της στους αλγορίθμους μηχανικής και βαθιάς μάθησης.



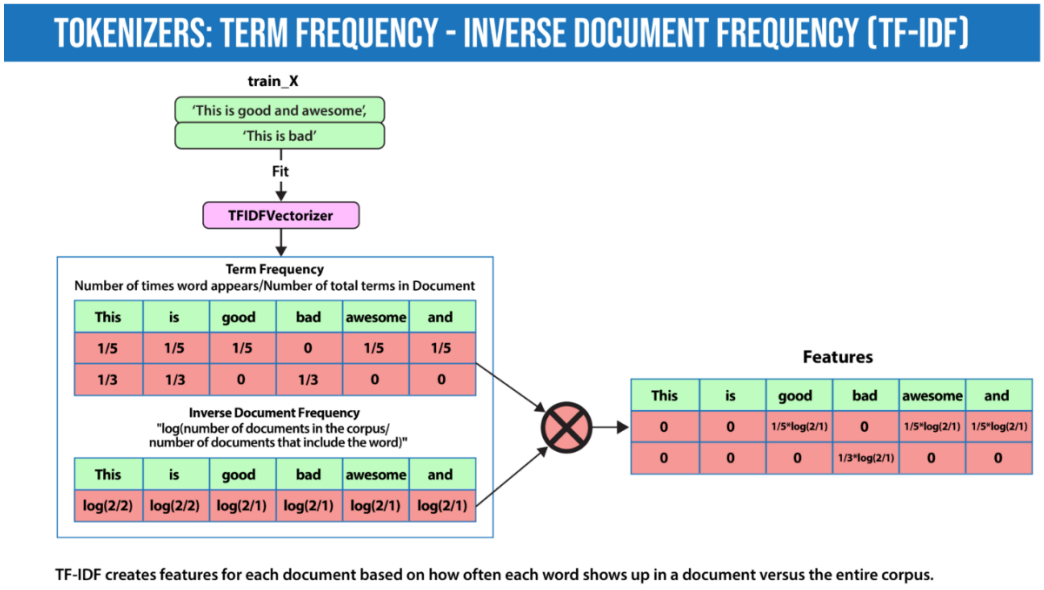
Εικόνα 28: Tokenizer

2. Feature extraction, ή αλλιώς η εξαγωγή μετρικών που πλαισιώνουν ένα document. Πρακτικά οι μετρικές αυτές είναι αριθμητικοί δείκτες, οι οποίοι περιγράφουν με στατιστικό τρόπο το λεξιλόγιο ενός κειμένου. Αναλυτικότερα, διάσημοι τέτοιοι δείκτες είναι το Bag-of-Words, το οποίο πληροφορεί σχετικά με τη συχνότητα λέξεων (ή ακολουθία λέξεων) σε ένα κείμενο :



Εικόνα 29: Features of tokens

και το TF-IDF, το οποίο αποτελεί τον πολλαπλασιασμό του Term Frequency (TF) και του Inverse Document Frequency (IDF). Ο πρώτος όρος αναφέρεται στην ποσοστιαία συχνότητα της λέξης σε ένα συγκεκριμένο κείμενο και επικεντρώνεται στη σημαντικότα της λέξης αυτής στο κείμενο. Ο δεύτερος όρος ταυτίζεται με το λογάριθμο του αντίστροφου ποσοστού των συνολικότερων κειμένων (documents) που περιέχουν τη λέξη.



Εικόνα 30: Μετρική TF-IDF

Μια πιο σύγχρονη προσέγγιση εξαγωγής χαρακτηριστικών είναι η Word2Vec, η οποία αξιοποιεί νευρωνικά δίκτυα προκειμένου να αναπαραστήσει μια λέξη μέσω ενός κανονικοποιημένου αριθμητικού πίνακα υψηλών διαστάσεων. Το embedding αυτό, που αντιστοιχεί σε μια λέξη, προσεγγίζει την σημασιολογική τοποθέτηση αυτής της λέξης ανάμεσα στις αντίστοιχες που συνυπάρχουν σε ένα κείμενο. Συνεπώς, η ομοιότητα δύο λέξεων κρίνεται από τη γεωμετρική απόσταση των embeddings τους.

3. Modeling. Σαν τελικό βήμα παρουσιάζεται η επιλογή του μαθηματικού μοντέλου που θα αντλήσει τα προεπεξεργασμένα δεδομένα και θα παράγει χρήσιμα συμπεράσματα. Naïve Bayes, Decision trees, logistic regression είναι μερικά παραδείγματα αλγόριθμων που μπορούν να χρησιμοποιηθούν σε classification tasks.

Ωστόσο, όπως θα δούμε και παρακάτω, τα συστήματα βαθιάς μάθησης είναι συνήθως ανεξάρτητα και δεν προϋποθέτουν εξαγωγή χαρακτηριστικών.

Οι πιο διάσημοι αλγόριθμοι NLP μοντέλων είναι διασκορπισμένοι στο φάσμα της πολυπλοκότητας. Πολλοί εξ αυτών διαθέτουν πιθανολογική και στατιστική προσέγγιση, ενώ άλλοι αποτελούν βαθιά νευρωνικά δίκτυα. Η επιλογή ανάμεσα τους είναι καθαρά υποκειμενική και εξαρτάται πλήρως από τη φύση του εκάστοτε προβλήματος.

2.11.1 Logistic regression

Αποτελεί έναν αλγόριθμο που συναντάται σε supervised classification tasks και επιχειρεί να αποδώσει την πιθανότητα ενός ενδεχομένου να πραγματοποιηθεί. Σε NLP συστήματα εφαρμόζεται κυρίως με spam detection και toxicity classification σκοπούς.

2.11.2 Naive Bayes

Μπορεί να βοηθήσει στην εύρεση σφαλμάτων σε γλωσσικό κείμενο και πρακτικά υπολογίζει τις ανταγωνιστικές πιθανότητες ενός κειμένου να ανήκει σε διάφορες κατηγορίες. Απλοποιεί τη γενική ενδιαφέρουσα πιθανότητα P(label / text) σε επιμέρους πιθανότητες που είναι ευκολότερο να υπολογιστούν. Με άλλα λόγια :

$$P(label / text) = P(label) * P(text / label)$$

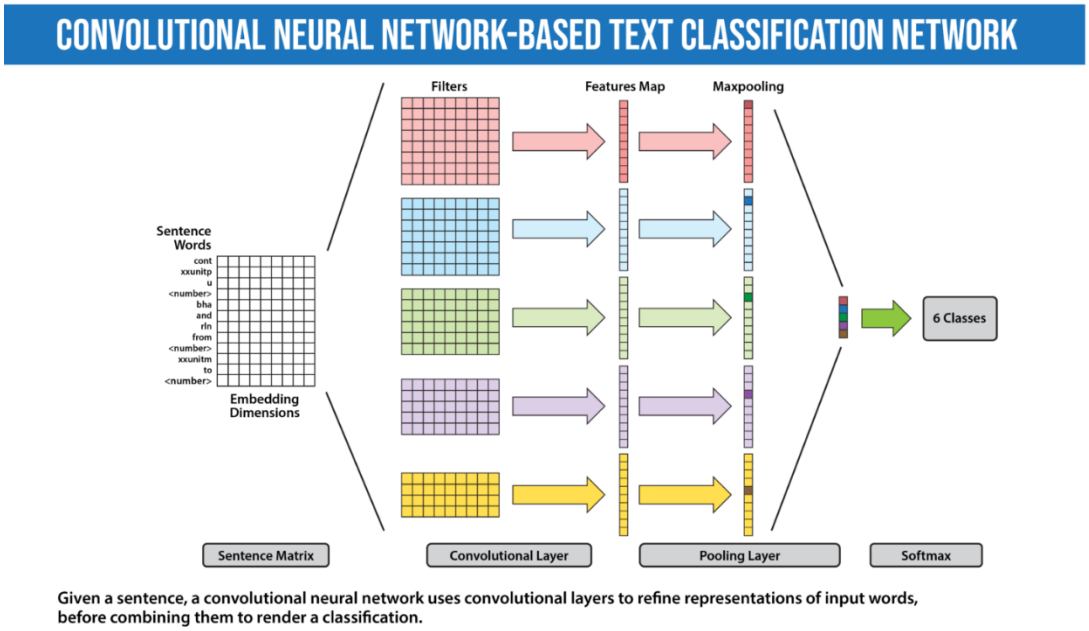
όπου :

$$P(text / label) = P(word\_1 / label) * P(word\_2 / label) * \ldots$$

Να σημειωθεί πως στους παραπάνω υπολογισμούς υποβόσκει η παραδοχή της ανεξαρτησίας ανάμεσα στις λέξεις του κειμένου.

2.11.3 Convolutional Neural Networks

Όπως γνωρίζουμε, τα CNNs χρησιμοποιούνται στην επεξεργασία και ταξινόμηση εικόνων. Ωστόσο, υπενθυμίζουμε πως η είσοδος ενός CNN δεν παύει να είναι ένας πίνακας τιμών, ο οποίος μεν αντιπροσωπεύει pixels στην περίπτωση της εικόνας, αλλά μπορεί δε να ενσωματώσει τα χαρακτηριστικά των λέξεων – δηλαδή τα embeddings - στην περίπτωση των γλωσσικών δεδομένων.



Εικόνα 31: Συνελικτικό Νευρωνικό Δίκτυο σε γλωσσολογικά δεδομένα

2.11.4 Recurrent Neural Networks

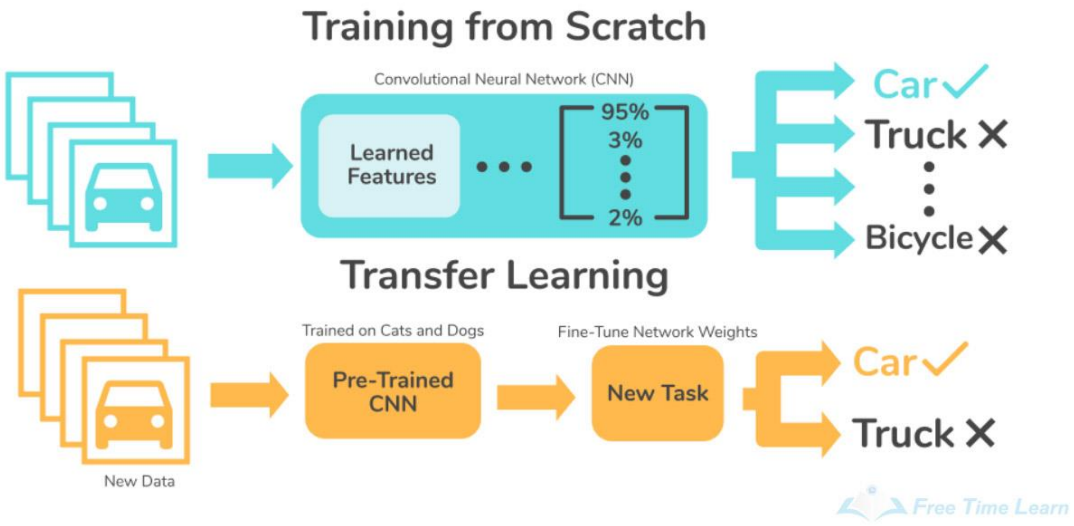
Τα ανατροφοδοτούμενα νευρωνικά δίκτυα μπορούν να αξιοποιηθούν στην πρόβλεψη επόμενων λέξεων και φράσεων σε ένα κείμενο. Παρά τα μειονεκτήματα τους στην προκείμενη περίπτωση, δηλαδή την αδυναμία του μοντέλου να ανιχνεύσει το ουσιαστικό νόημα των λέξεων, διαθέτει ενισχυμένη απόδοση στην πρόβλεψη ακολουθιών λόγω της ικανότητας να «θυμάται» αρκετά παρελθοντική πληροφορία.

2.12 Μεταφορά Μάθησης

Η μεταφορά μάθησης είναι μια πολύτιμη τεχνική στα συστήματα Μηχανικής και Βαθιάς Μάθησης. Η αξία της βρίσκεται στο γεγονός πως υπάρχοντα και προεκπαιδευμένα μοντέλα μπορούν να αξιοποιηθούν έτσι ώστε να ενισχύσουν με γνώση νευρωνικά δίκτυα που διαθέτουν περιορισμένα δεδομένα [45].

Η μεταφορά και επαναχρησιμοποίηση υπάρχοντων έτοιμων μοντέλων λαμβάνει χώρα συνήθως με σκοπό να επιταχύνει και να ενισχύσει την απόδοση ενός υπό ανάπτυξη νέου μοντέλου. Δεν αντικαθιστά τη λειτουργία του, ούτε αλλοιώνει την ταυτότητά του, αλλά βοηθάει στην διεκπεραίωση υποπροβλημάτων, των οποίων η επίλυση μπορεί να κριθεί αναγκαία και απαραίτητη ή απλώς ενισχυτική. Για παράδειγμα, η αναγνώριση μιας γάτας σε ένα computer vision πρόβλημα ταξινόμησης, μπορεί να απλοποιηθεί σε δύο επιμέρους tasks. Το πρώτο είναι η αναγνώριση των εικόνων που αναπαριστούν ζώο και όχι κάτι διαφορετικό, και το δεύτερο είναι η εξειδικευμένη κατηγοριοποίηση του ζώου το είδος της γάτας. Σε αυτό το pipeline μπορεί να χρησιμοποιηθεί μεταφορά μάθησης, προκείμενου να ξεπεραστεί το πρώτο task και το νευρωνικό δίκτυο να επικεντρωθεί εν τέλει στο δεύτερο task. Αντιλαμβανόμαστε λοιπόν πως πέρα από την έλλειψη δεδομένων πολλές φορές το transfer learning αποφορτίζει το σύστημα από υπολογιστικά κόστη, καθώς πολλαπλά layers προκαλούν αυξημένη πολυπλοκότητα. Επομένως η παράκαμψη κάποιων layers μειώνει το κόστος και αυξάνει την απόδοση του αναπτυσσόμενου και εκπαιδευόμενου νευρωνικού δικτύου.

Εντελώς αντίστοιχα, σε ένα πρόβλημα ταξινόμησης ανθρώπων με βάση κάποια χαρακτηριστικά τους, το input ενός νευρωνικού δικτύου μπορεί να μην είναι επαρκές αποκλειστικά και μόνο μέσω του παρεχόμενου dataset. Δηλαδή το dataset μπορεί να μας δίνει – μεταξύ άλλων - την περιγραφή της προσωπικότητας ενός ανθρώπου (σε ένα κείμενο), αλλά το νευρωνικό δίκτυο μπορεί να απαιτεί την ύπαρξη ενός χαρακτηριστικού της προσωπικότητας του ανθρώπου. Η παραπάνω απαίτηση είναι μεν προαιρετική, καθώς η λειτουργία του νευρωνικού δικτύου δεν εξαρτάται πλήρως από το χαρακτηριστικό της προσωπικότητας, αλλά σε γενικό επίπεδο ενδέχεται να κριθεί καίρια ως προς τη απόδοση του συνολικού συστήματος. Συνεπώς το εξαγόμενο χαρακτηριστικό μπορεί άμεσα να προκύψει από τη μεταφορά μάθησης ενός ξένου μοντέλου NLP, το οποίο έχει εξειδικευμένα προπονηθεί σε αυτό το συγκεκριμένο concept. Έτσι εξοικονομείται χρόνος, αλλά καλύπτονται και πιθανές ανεπάρκειες (σε πλήθος) του διαθέσιμου dataset.



Εικόνα 32: Μεταφορά Μάθησης

2.13 Multi Modal Transformation

Μέχρι στιγμής εξηγήσαμε την πορεία που ακολουθήθηκε κατά την εξαγωγή (extraction) και την προεπεξεργασία (preprocessing) των χαρακτηριστικών των χρηστών. Ως κατάληξη έχουμε την οργάνωση των χαρακτηριστικών σε δύο κατηγορίες, description και numerical. Η πρώτη κατηγορία ταυτίζεται με τα embeddings που παράγονται από το εξωτερικό γλωσσικό μοντέλο, όταν αυτό λαμβάνει ως είσοδο τις λεκτικές περιγραφές των χρηστών του δικτύου. Από την άλλη, η δεύτερη κατηγορία περιέχει όλα τα αριθμητικά στοιχεία και μετρικές που αναλύθηκαν παραπάνω.

Όπως γίνεται εύκολα αντιληπτό, η φύση των δύο παραπάνω «πακέτων» των χαρακτηριστικών δεν είναι ίδια, ακόμα και αν ο μαθηματικός τύπος των μεταβλητών τους είναι κοινός (scaled float). Συνεπώς ο τελικός στόχος της συνένωσης και των δύο κατηγοριών – προκειμένου να σχηματιστούν οι τερματικοί πίνακες / vectors που θα εκπροσωπήσουν τους χρήστες στο Graph Neural Network μέσω των χαρακτηριστικών τους – δεν είναι προφανής.

Τα embeddings, δηλαδή το description, είναι πρακτικά ένας πίνακας 384 μεταβλητών κανονικοποιημένων πραγματικών αριθμητικών τιμών, ενώ οι μετρικές, δηλαδή το numerical, είναι ένας πίνακας 9 αντίστοιχων τιμών. Παρόλο που οι τιμές διαθέτουν τον ίδιο τύπο και μορφή και στις δύο περιπτώσεις, η απλοϊκή συνένωση των δύο αυτών

πινάκων δεν κρίνεται ορθή σε καμία περίπτωση. Το πρόβλημα έγκειται στο γεγονός πως κάθε ένα dimension (από τα 384) του description δεν μπορεί και δεν πρέπει να εξομοιωθεί με τα αντίστοιχα αριθμητικά χαρακτηριστικά. Με άλλα λόγια, δεν πρέπει να σταθεί σαν μια μεμονωμένη και ανεξάρτητη αριθμητική τιμή, στο πεδίο ορισμού των attributes ενός χρήστη. Η σημασία και η αξία ενός embedding αναδεικνύεται μέσω του συνδυασμού όλων των 384 dimensions ως πακέτο τιμών.

Τη λύση στο παραπάνω πρόβλημα δίνουν συγκεκριμένες τεχνικές συγχώνευσης διαφορετικών τύπων δεδομένων, με τις οποίες multi modal χαρακτηριστικά μπορούν να μετασχηματιστούν σε ένα ενιαίο πίνακα τιμών [6]. Αυτός ο πίνακας θα αποτελέσει και την είσοδο στο Graph Neural Network.

Ανάλογα τον τύπο των δεδομένων που συγχωνεύονται, προτείνονται και διαφορετικοί αλγόριθμοι μετασχηματισμού [26]. Στην παρούσα εργασία συγκεντρώνονται οι πιο διάσημοι εξ αυτών και δοκιμάζονται ξεχωριστά για την επίδοσή τους, αποτελώντας έτσι μια από τις global παραμέτρους του συστήματός μας. Συγκεκριμένα οι μέθοδοι που δοκιμάζονται είναι οι εξής :

1. *BLOCK*
2. *LinearSum*
3. *Tucker*
4. *MFB*

Η βιβλιοθήκη που αξιοποιείται για τις παραπάνω μεθόδους είναι η εξής :

'block.fusions' by ***Cadene*** (block.bootstrap.pytorch)

2.14 Ισορροπία στο Dataset

Στη Μηχανική Μάθηση και γενικά την Τεχνητή Νοημοσύνη, το σύνολο δεδομένων που αξιοποιείται προκειμένου να εκπαιδευτεί και να αξιολογηθεί αποτελείται από στοιχεία και χαρακτηριστικά οντοτήτων. Στην περίπτωση των classification tasks, οι οντότητες αυτές χωρίζονται σε δύο ή περισσότερες ομάδες. Οι ομάδες αυτές καθορίζονται από τη γνωστή ετικέτα (label) ενός στοιχείου.

Για παράδειγμα σε ένα dataset που περιέχει στοιχεία ασθενών, η κύρια οντότητα είναι ο ασθενής, ενώ οι επιμέρους ομάδες καθορίζονται από την ταξινόμηση που θέλουμε να εφαρμόσουμε. Εάν επιθυμούμε να ταξινομήσουμε τους ασθενείς σε επίπεδα σοβαρότητας και επικινδυνότητας της υγείας τους, τότε οι ομάδες ταυτίζονται με αυτά τα επίπεδα. Η ένδειξη που μας ενημερώνει σχετικά με το επίπεδο που ανήκει ο κάθε

ασθενής προέρχεται από την ετικέτα του και είναι αναπόσπαστο μέρος των επιμέρους χαρακτηριστικών του χρήστη, όταν επικεντρωνόμαστε σε supervised ή semi-supervised learning.

Είναι εύλογο ένα σύνολο δεδομένων να μην περιέχει ισοπληθή ομάδες. Στο παραπάνω παράδειγμα αυτό θα συνέβαινε στην περίπτωση όπου οι ασθενείς με ήπια συμπτώματα ανήκαν στο 50% του ευρύτερου πληθυσμού, οι ασθενείς με αυξημένα συμπτώματα ανήκαν στο 40% του γενικού πληθυσμού, ενώ το υπόλοιπο 10% εκπροσώπευε την τρίτη και τελευταία ομάδα ασθενών, των οποίων η κατάσταση έχει κρίσιμη επικινδυνότητα.

Η κατάσταση κατά την οποία ένα σύνολο δεδομένων δεν είναι ισορροπημένο – όσον αφορά την κατανομή των κλάσεων/ομάδων του – αποτελεί τον κανόνα και όχι την εξαίρεση στο χώρο της επιστήμης των δεδομένων [19]. Στη Μηχανική Μάθηση, η εκπαίδευση του μοντέλου επηρεάζεται άμεσα και η απόδοση του συστήματος καθορίζεται έμμεσα από το φαινόμενο αυτό. Ένα νευρωνικό δίκτυο, προκειμένου να εκπαιδευτεί με δίκαιο και μαθηματικά ορθόδοξο τρόπο, οφείλει να διαθέτει ένα ισορροπημένο σύνολο δεδομένων ως πεδίο δράσης του. Οποιοδήποτε ποσοστό ανισορροπίας προκαλεί ένα είδος «προκατάληψης» και μια λανθασμένη τάση πρόβλεψης στο μοντέλο. Μόνο με ένα ισορροπημένο dataset το μοντέλο εξαναγκάζεται να εκπαιδευτεί κατάλληλα, να αποκωδικοποιήσει με το βέλτιστο δυνατό τρόπο τα χαρακτηριστικά των στοιχείων του και εν τέλει να ανιχνεύσει τις ουσιώδεις και βαθιές διαφορές ανάμεσα στις κλάσεις που επιχειρεί να προβλέψει.

Οι μεγάλες ανισορροπίες σε ένα dataset «διευκολύνουν» το έργο και την εκπαίδευση ενός νευρωνικού δικτύου, το οποίο προσεγγίζει εύκολα μεγάλες αποδόσεις, έχοντας την τάση να προβλέπει σωστά την major κλάση και ταυτόχρονα να μην αποκτά ποτέ την ικανότητα να ξεχωρίζει τα χαρακτηριστικά της minor κλάσης, πράγμα που είναι και το αληθινό γενικό ζητούμενο. Θα μπορούσε να πει κανείς πως η ποικιλία και το diversity των στοιχείων ενός dataset, ως προς τα χαρακτηριστικά τους, οφείλει πάντα να συνοδεύεται και με ισορροπία, ως προς τις ετικέτες τους, καθώς μόνο με αυτόν τον τρόπο το νεωρωνικό δίκτυο εξαναγκάζεται να «μάθει» και να «διακρίνει» την πηγή της διαφορετικότητας των επιμέρους κλάσεων.

Υπάρχουν διάφορες επικρατούσες τεχνικές στην προσπάθεια εξισορρόπησης ενός dataset :

1. Χρήση των κατάλληλων μετρικών απόδοσης. Αποτελεί την πιο απλή λύση και ουσιαστικά αναφέρεται στην αξιοποίηση των δεικτών Precision, Recall και όχι μόνο του κλασικού Accuracy. Ο χάρτης σύγχυσης, γνωστός διεθνώς ως Confusion Matrix, είναι επίσης μια κατάλληλη προσέγγιση στην απεικόνιση των προβλέψεων του συστήματος, καθώς πληροφορεί σχετικά με την επιμέρους (ως προς την κλάση) απόδοση, και όχι τη γενική.

2. Oversampling. Είναι η διαδικασία επαύξησης των στοιχείων ενός dataset με στόχο την εξισορρόπηση των κλάσεων. Πραγματοποιείται είτε μέσω της γενικής τυχαίας παραγωγής καινούργιων μελών του συνόλου δεδομένων, είτε μέσω της στοχευμένης παραγωγής νέων μελών μόνο της minor κλάσης. Στη δεύτερη περίπτωση η παραγωγή μπορεί να ταυτιστεί με αντιγραφή στοιχείων στην απλή της μορφή, αλλά η βέλτιστη μέθοδος είναι η παραγωγή νέων μοναδικών στοιχείων με χρήση ευριστικών αλγορίθμων. Σε κάθε περίπτωση, το κύριο πλεονέκτημα του oversampling είναι η διατήρηση όλων των αρχικών δεδομένων ενός dataset. Με άλλα λόγια η αρχική πληροφορία δε χάνεται. Από την άλλη, ένα βασικό μειονέκτημα είναι η ευαλλωτότητα στο ενδεχόμενο του overfitting.
3. Undersampling. Είναι η διαδικασία μείωσης των στοιχείων του dataset που ανήκουν στην major κλάση. Αυτή η μείωση των στοιχείων πραγματοποιείται είτε με τυχαίο τρόπο στην απλή της μορφή, είτε μέσω της εξαγωγής αντιπροσώπων. Συγκεκριμένα, η major κλάση εξετάζεται μεμονωμένα και μέσω αλγορίθμων που προσομοιάζουν τον ‘K-means’, επιτυγχάνεται η εξαγωγή των στοιχείων που μπορούν να αντιπροσωπεύσουν και να αντικαταστήσουν ένα υποσύνολο «παρόμοιων» γειτονικών στοιχείων. Με αυτόν τον τρόπο μειώνεται το πλήθος της major κλάσης και το dataset εξισορροπείται. Το πλεονέκτημα αυτής της στρατηγικής είναι η ταυτόχρονη μείωση του run time, αλλά το κρίσιμο μειονέκτημα της – που την καθιστά ακατάλληλη σε πολλές περιπτώσεις – είναι η απόρριψη και η μη αξιοποίηση χρήσιμης πληροφορίας, η οποία «χάνεται» κατά τη μείωση των στοιχείων.

Σε γενικές γραμμές, η εξισορρόπηση ενός dataset στον κόσμο της Μηχανικής Μάθησης είναι ένα βασικό συστατικό της συνολικής προ-επεξεργασίας των δεδομένων του συστήματος. Επισημαίνεται πως η έννοια της ισορροπίας αποκτά και έχει νόημα μόνο στο κομμάτι της εκπαίδευσης (training) του μοντέλου. Συνεπώς πρέπει πάντα να λαμβάνει χώρα μόνο στο training υποσύνολο και μετά τη διαδικασία του διαχωρισμού σε train και test set. Το validation και το test set οφείλουν να παραμείνουν αναλλοίωτα και να αποτυπώνουν την πραγματική αληθινή κατανομή.

Κεφάλαιο 3

Σχετική Έρευνα

Υπάρχει πληθώρα προγενέστερων υπαρκτών λύσεων στο συγκεκριμένο classification task. Μερικά συστήματα επιχειρούν να ταξινομήσουν τους χρήστες αξιοποιώντας περισσότερο τα γλωσσικά και τα αριθμητικά τους χαρακτηριστικά, κάποια άλλα επικεντρώνονται στα tweets, ενώ μερικά καταφέρνουν να συνδυάζουν όλα τα παραπάνω.

Η αναζήτηση εκκίνησε από μοντέλα classifiers και clustering που εμπεριείχαν γνωστούς μαθηματικούς αλγόριθμους προκειμένου να αντιμετωπίσουν τους χρήστες και τα χαρακτηριστικά τους. Από ένα σημείο και μετά, η ανάπτυξη νευρωνικών δικτύων – και κυρίως των Recurrent Neural Networks, ήρθε για να βοηθήσει στην ενσωμάτωση και τη συμμετοχή των tweets. Βέβαια, οι αρχιτεκτονικές, που ακολούθησαν και εκτόξευσαν την ακρίβεια και την απόδοση των προβλέψεων, ήταν εκείνες των Graph Convolutional Networks, οι οποίες πρόσθεσαν τη διάσταση των σχέσεων και συνδέσεων μεταξύ των users. Τέλος, οι μεταγενέστερες και πιο πρόσφατες απόπειρες ανίχνευσης των bots στο Twitter, απαρτίζονται από πολυεπίπεδα δίκτυα και πολλαπλά modals, τα οποία εκπαιδεύονται με attention μηχανισμούς. Με λίγα λόγια, αυτά τα εξελιγμένα συστήματα συσσωρεύουν βαθιά πληροφορία από διαφορετικά μοντέλα και τελικά τα συνδυάζουν κατάλληλα προκειμένου να παράξουν το τελικό αποτέλεσμα. Τα διαφορετικά αυτά μοντέλα συνήθως απασχολούν και ξεχωριστό κομμάτι του dataset.

Σε αυτό το σημείο, θα περιγράψουμε σύντομα μερικά υπάρχοντα συστήματα, καθώς επίσης και θα τα ομαδοποιήσουμε σε γενικότερες κατηγορίες λύσεων, όπως αναλύσαμε παραπάνω.

3.1 Ανιχνεύοντας με Αλγορίθμους και Ταξινομητές

Οι ακόλουθες προσεγγίσεις αποτέλεσαν κάποιες από τις πρώτες λύσεις στο πρόβλημα της ταξινόμησης λογαριασμών στο Twitter, και συνείσφεραν ως διάσημα baselines στις επερχόμενες αρχιτεκτονικές.

Το **Botometer** [8] επικεντρώθηκε περισσότερο στο πλήθος και των ποικιλία των εξαγόμενων χαρακτηριστικών των χρηστών, τα οποία προήλθαν από προσωπικές

πληροφορίες και περιεχόμενο μέσω του Twiter API. Συγκεκριμένα, πάνω από 1000 features, μεταξύ αυτών και metadata, δρομολογήθηκαν σε Random Forest ταξινομητή, ο οποίος χωρίζοντας το πρόβλημα σε τυχαία και επικαλυπτόμενα υποσύνολα χαρακτηριστικών και χρηστών, οδηγεί τελικά στα decision trees, τα οποία συναποφασίζουν για την τελική πρόβλεψη. Παρόλο που ο συγκεκριμένος αλγόριθμος παρουσιάζει αξιοσημείωτη ακρίβεια, δίχως ιδιαίτερη υπολογιστική πολυπλοκότητα, η τελική δομή των δέντρων αποφάσεων κρίνεται πολύ ευαίσθητη στην περίπτωση προσθήκης νέων δεδομένων.

Μια εντελώς αποκλίνουσα και ιδιάζουσα λύση στο χώρο του Twitter ονομάζεται **Cresci** [7]. Σε αυτήν την προσέγγιση, οι διαδικτυακές κινήσεις των χρηστών κωδικοποιούνται και προσομοιάζουν τις αλυσίδες DNA. Με λίγα λόγια, τα ατομικά actions αποτυπώνονται σαν χρονικές συμβολοσειρές (P-L-P-C-F-F), στις οποίες κάθε γράμμα διαθέτει και μια διαφορετική ερμηνεία, ακριβώς όπως και στις αλυσίδες DNA. Δηλαδή στο παραπάνω παράδειγμα :

P = post

L = like

C = comment

F = follow

Με αυτόν τον τρόπο, όλοι οι χρήστες καταλήγουν να αντιπροσωπεύονται από μοναδικές συμβολοσειρές, οι οποίες υπόκεινται σε περαιτέρω υπολογιστικές διαδικασίες, όπως Longest Common Substring (LCS), που θα καθορίσουν τις ομοιότητες μεταξύ αυτών.

Η λύση του **Miller** [35] εξειδικεύεται αποκλειστικά σε clustering μεθόδους, οι οποίες συνήθως αξιοποιούνται σε unsupervised learning tasks. Οι δύο καθοριστικοί αλγόριθμοι που επιλέγονται για να επιδράσουν στα δεδομένα είναι :

- DenStream
- StreamKM++

Και οι δύο εκπροσωπούν εξελιγμένες εκδόσεις των διάσημων θεμελιωδών αλγορίθμων :

- DBSCAN
- kNN

αντίστοιχα, οι οποίοι τροποποιούνται κατάλληλα έτσι ώστε να ταιριάξουν καλύτερα στο συγκεκριμένο πρόβλημα. Το σύνολο δεδομένων περιέχει κάποια κλασικά αριθμητικά χαρακτηριστικά των χρηστών, όπως πλήθος likes, tweets, follows κλπ. Ο γενικότερος

σκοπός του συστήματος είναι η διερεύνηση και ανίχνευση ανώμαλων μοτίβων (anomaly patterns), τα οποία υποβόσκουν στα διάφορα είδη user metadata. Ο DenStream χρησιμοποιείται σε μια επαναληπτική λούπα για να καθορίσει τα p-micro-clusters, τα οποία ομαδοποιούν όλους τους γνήσιους λογαριασμούς, εξαιρώντας όλους τους ψεύτικους. Από την άλλη, ο StreamKM++ εκτελεί τον kNN με μια πιθανολογική και όχι τυχαία αρχικοποίηση των κεντρικών σημείων, αποφεύγοντας έτσι τις αναπόφευκτες ανακρίβειες που μπορεί να ελλοχεύει μια αρχικά λανθασμένη επιλογή. Ακολουθεί τελικά ο συνδυασμός των δύο αλγορίθμων, προτού παραχθεί η τελική πρόβλεψη.

3.2 Ανιχνεύοντας με Νευρωνικά Δίκτυα

Τα Recurrent Neural Networks απέκτησαν ένα σημαντικό ρόλο στην ενδυνάμωση των συστημάτων ανίχνευσης bots στο Twitter, καθώς παρέθεσαν τον τρόπο επεξεργασίας και αξιοποίησης των tweets.

Η λύση του **Wei** [48] εκμεταλλεύεται τη λειτουργικότητα των Long Short-Term Memory μοντέλων. Κωδικοποιώντας (tokenization) τα αρχικά tweets – τα οποία ταυτίζονται με την μοναδική και πραγματική πηγή εισόδου του συστήματος – και σχηματίζοντας τα αντίστοιχα embeddings αυτών, δημιουργεί τις προϋποθέσεις για την ανίχνευση ψευδούς περιεχομένου μέσω ενός bidirectional LSTM. Το μοντέλο, διαθέτοντας τρία hidden layers και μια τελική συνάρτηση ενεργοποίησης Softmax, εκπαιδεύεται προκειμένου να ελαττώσει την αρνητική λογαριθμική πιθανοφάνεια και προς τις δύο κατευθύνσεις, και εν τέλει παράγει μια πιθανολογική κατηγοριοποίηση του tweet. Το βασικό πλεονέκτημα του είναι πως δεν εμπλέκεται με τεχνικές feature engineering, δηλαδή δέχεται ως είσοδο το πραγματικό content του tweet, το οποίο σε συνδυασμό με το μειωμένο υπολογιστικό του κόστος και τις αυξημένες του αποδόσεις, επαρκεί προκειμένου να μετατρέψει το σύστημα σε ένα πολύ δυνατό και ενδιαφέρον baseline.

Η προσέγγιση του **Kudugunta** [29] αγκαλιάζει δύο θεμελιώδεις πτυχές των δεδομένων που προέρχονται από τη σφαίρα του Twitter, των user features και των tweets. Απαρτίζεται από δύο ισχυρά components :

- Tweet-Level Classification
- Account-Level Classification

τα οποία επεξεργάζονται τις δύο παραπάνω κατηγορίες εισόδου. Το πρώτο υποσύστημα συμβάλλει στην ανίχνευση των bots μέσω του περιεχομένου και των χαρακτηριστικών των tweets. Η κωδικοποίηση και ο σχηματισμός των embeddings των tweets, ο οποίος λαμβάνει χώρα με τη μεταφορά και συνδρομή του 'GloVE' (pretrained and specialized in Twitter data), ακολουθείται από την ενεργοποίηση ενός τυπικού LSTM μοντέλου, του οποίου η έξοδος αθροίζεται με τα metadata των tweets και τοποθετείται ως είσοδος σε ένα νευρωνικό δίκτυο δύο στρωμάτων, το οποίο είναι και υπεύθυνο για τα τελικά αποτελέσματα. Από την άλλη, το δεύτερο υποσύστημα αξιοποιεί κάποια υψηλά ερμηνεύσιμα χαρακτηριστικά των tweets, εναποθέτοντας τα σε πέντε διαφορετικούς ταξινομητές – SGD και AdaBoost Classifier μεταξύ αυτών – και συγκρίνοντας τις ανταγωνιστικές του επιδόσεις, λαμβάνοντας υπόψιν και τις διαδικασίες undersampling SMOTENN και SMOTOMEK. Εκτός από την ανάγκη για εξισορρόπηση του dataset, τα εντυπωσιακά του αποτελέσματα αποδεικνύουν πως η εκτενής χρήση πολυπληθών χαρακτηριστικών κρίνεται περιττή.

3.3 Ανιχνεύοντας με Συνελικτικά Δίκτυα Γράφων

Τα Graph Convolutional Networks μπορούν να χαρακτηριστούν και ως game changers στην ιστορία της ανίχνευσης ψευδών λογαριασμών στην κοινότητα του Twitter. Ο γράφος που χτίζεται και αξιοποιείται κάθε φορά, δεν είναι τίποτε άλλο παρά η απεικόνιση των διάφορων σχέσεων μεταξύ των χρηστών.

Η απλή προσέγγιση του **Alhosseini** [1] εκμεταλλεύεται μερικά απλά χαρακτηριστικά των χρηστών, προερχόμενα από το Twitter API, και εφαρμόζοντας συνελικτικά δίκτυα στις γειτονιές των χρηστών, καταφέρνει να πετύχει υψηλές μετρικές στις αρχικές και πρωτοεμφανιζόμενες κατηγορίες bots. Είναι αδιαμφησβήτητο το γεγονός πως το συγκεκριμένο σύστημα, παρά την απλότητα και το μινιμαλισμό του, αποτέλεσε ένα βασικό πυλώνα σύγκρισης για τις επερχόμενες αρχιτεκτονικές.

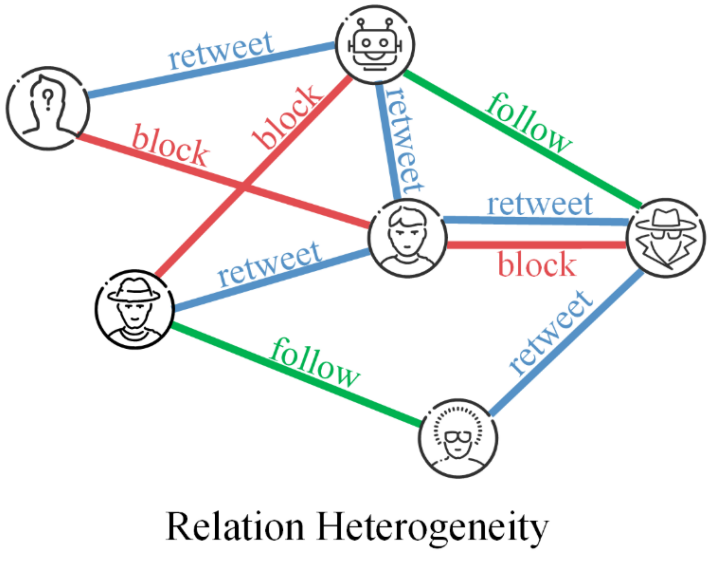
Το **BGSRD** [18] κυρίως επικεντρώνει τις προσπάθειες του στην πλήρη αξιοποίηση των tweets των χρηστών. Στην πραγματικότητα, το σύστημα χωρίζεται σε δύο κρίσιμα διαδραστικά sections :

- Text representations via BERT model
- Graph Convolutional Network in text data

Το πρώτο λαμβάνει ως input τις λέξεις των tweets, και αφού διαχωρίσει τις συνεχόμενες προτάσεις, είναι υπεύθυνο για την παραγωγή των Next Sequence Predictions (NSP). Η εκμάθηση του μοντέλου συμπεριλαμβάνεται στην αποτύπωση των sentences pairs, αλλά και στην δειγματοληψία «αρνητικών» προτάσεων. Το δεύτερο section προορίζεται για το σχηματισμό ενός καινοτόμου μοναδικού γράφου, στον οποίο τα tweets (documents) και οι συμμετέχουσες λέξεις (words) διαδραματίζουν το ρόλο των κόμβων, ενώ οι ακμές διαθέτουν βαθμό και καθορίζονται από τις συνδέσεις μεταξύ των δύο προαναφερθέντων entities. Συγκεκριμένα, τα documents συσχετίζονται με ένα words βάση της TF-IDF μετρικής, ενώ η PPMI μετρική αντίστοιχα χρησιμεύει στη σύνδεση μεταξύ δύο words. Το πρώτο section παρέχει τις αναπαραστάσεις του ως input στο δεύτερο, ενώ και τα δύο μαζί βελτιστοποιούνται και συνδυάζονται για να γεννήσουν τις τελικές προβλέψεις.

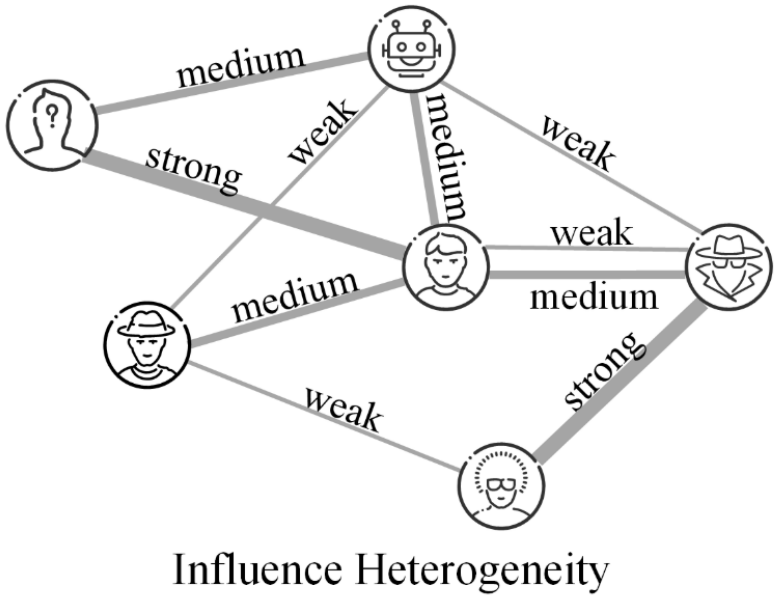
3.4 Ανιχνεύοντας με Attention Δίκτυα Γράφων

Η έμφυτη και μοιραία ανάγκη για ενσωμάτωση παραπάνω από ένα είδους σύνδεσης στον κατασκευασμένο γράφο της κοινότητας των users, οδήγησε τους επιστήμονες να συμπεριλαμβάνουν μηχανισμούς προσοχής (attention mechanisms) στα υλοποιημένα Graph Convolutional Networks, καθώς μόνο με αυτόν τον τρόπο θα μπορούσε να κριθεί και να υπολογιστεί η συνεισφορά και η σημασία κάθε είδους link.



Εικόνα 33: Ετερογένεια συσχέτισης

Φυσικά να τονίσουμε πως οι μηχανισμοί προσοχής υφίστανται και στις τοπικές επιρροές της γειτονιάς ενός χρήστη. Με άλλα λόγια, κάθε χρήστης έχει ρεαλιστικά και πραγματικά τη δυνατότητα να διαθέτει διαφορετική – σε βαθμό – επίδραση στο «φίλο» του, και το αντίστροφο.



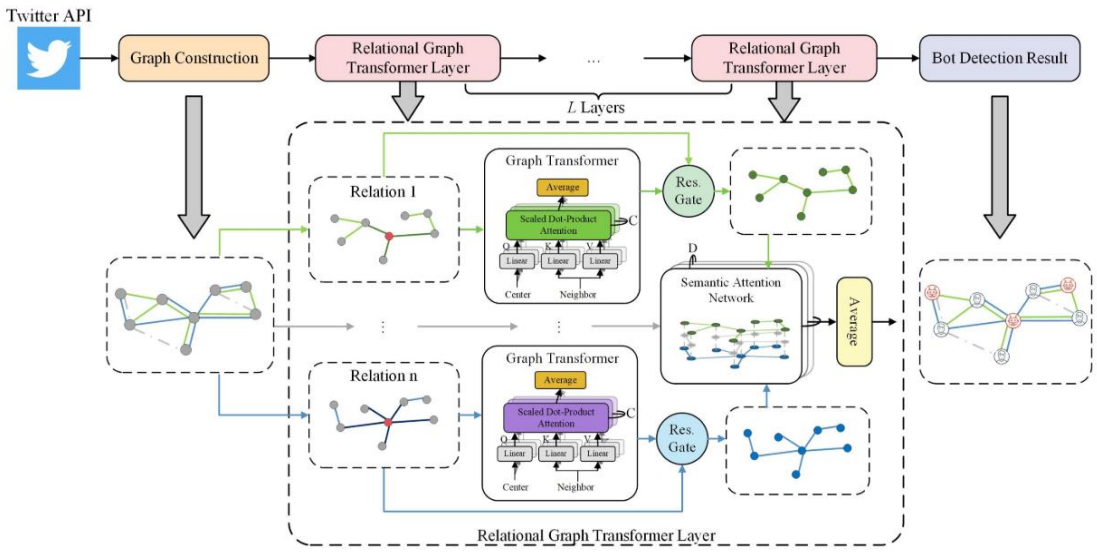
Εικόνα 34: Ετερογένεια επιρροής

Η αρχιτεκτονική του **BotMGAT** [2] αξιοποιεί αριθμητικά και boolean στοιχεία του χρήστη και παρουσιάζει μια πολύ ενδιαφέρουσα προσέγγιση του προβλήματος ανίχνευσης των bots στο Twitter. Πρακτικά, το σύστημα περικλείει πολλαπλά διαφορετικά views του Graph Convolutional Network, κάθε ένα από τα οποία αναπτύχθηκε με βάση ένα μοναδικό είδος σύνδεσης μεταξύ των users (following links, interaction links, etc ..). Ένα attention layer παρεμβάλλεται μεταξύ των επιμέρους outputs αυτών των ξεχωριστών GCNs και του τελικού αποτελέσματος, προκειμένου να εκπαιδευτεί και να καθορίσει τα ποσοστά συμβολής του κάθε ένα. Βέβαια, αυτό που κάνει το συγκεκριμένο σύστημα καινοτόμο είναι η transfer learning τακτική του, κατά την οποία το ολοκληρωμένο προπονημένο σύστημα που προαναφέρθηκε, χρησιμοποιείται επίσης για την κατηγοριοποίηση νέων δεδομένων, που συχνά αφορούν ειδικές dedicated κοινότητες του Twitter, όπως χρήστες ενός ορισμένο γκρουπ ή χρήστες που συμμετέχουν σε ένα απομονωμένο hashtag. Συνεπώς, οι εξωτερικοί αυτοί χρήστες αποκτούν μια προσωρινή ετικέτα μέσω του MGAT Network κι έπειτα δρομολογούνται σε έναν κατάλληλο Machine Learning Classifier, όπως ο Random Forest, ο οποίος θα παράξει την τελικό πρόβλεψη σχετικά με την κατηγορία τους.

Το **HIN** [13] σύστημα αφιερώνει την αρχιτεκτονική του στην ετερογένεια των συνδέσεων στο σύνολο των users. Εξατομικευμένοι και ξεχωριστοί γράφοι κατασκευάζονται και βασίζονται στο ίδιο dataset, όμως προσαρμόζοντας τις ακμές τους σύμφωνα με ένα διαφορετικό είδος συσχέτισης μεταξύ των users :

- Following
- Blocking
- Liking
- Mentioning
- Quoting
- Retweeting

Κάθε γράφος αποστέλλεται σε έναν ατομικό Graph Attention Transformer, ο οποίος κατέχει την ιδιότητα να δίνει βάρος στη γειτνίαση των users, με τη λογική πως κάθε χρήστης επηρεάζεται σε διαφορετικό βαθμό από τους γύρω του. Έπειτα από πολλαπλά συνελικτικά και βελτιωτικά επίπεδα, τα επιμέρους αποτελέσματα κάθε γράφου συναθροίζονται σε έναν τελικό γράφο. Η τερματική μονάδα επεξεργασίας του συστήματος ονομάζεται Semantic Attention Networks και υλοποιείται προκειμένου να αποδώσει τη σχετική σημασία και επίδραση μεταξύ των διαφορετικών links. Έτσι η τελική πρόβλεψη, αναφορικά με έναν χρήστη, προέρχεται από τον ποσοστιαίο συνδυασμό όλων των διαφορετικών γράφων.



Εικόνα 35: Αρχιτεκτονική HIN

3.5 Ανιχνεύοντας με Βαθιά Multimodal Δίκτυα Γράφων

Η ανάκτηση διαφορετικών μορφών πληροφορίας από το Twitter, σχετικά με έναν χρήστη, καθώς επίσης και η συνεχής ανάγκη για αναβάθμιση της απόδοσης των συστημάτων ανίχνευσης των bots, οδήγησε την επιστημονική κοινότητα να συμπεριλάβει πολλαπλές υπολογιστικές ενότητες στις αρχιτεκτονικές του;.

Το **SATAR** [15] παρουσιάζει τρία ατομικά υποσυστήματα, τα οποία περιεργάζονται φυσικά διαφορετική μορφή πληροφορίας, η οποία εν τέλει αθροίζεται για να μεγιστοποιήσει την ποιότητα των παραγόμενων αποτελεσμάτων. Συγκεκριμένα :

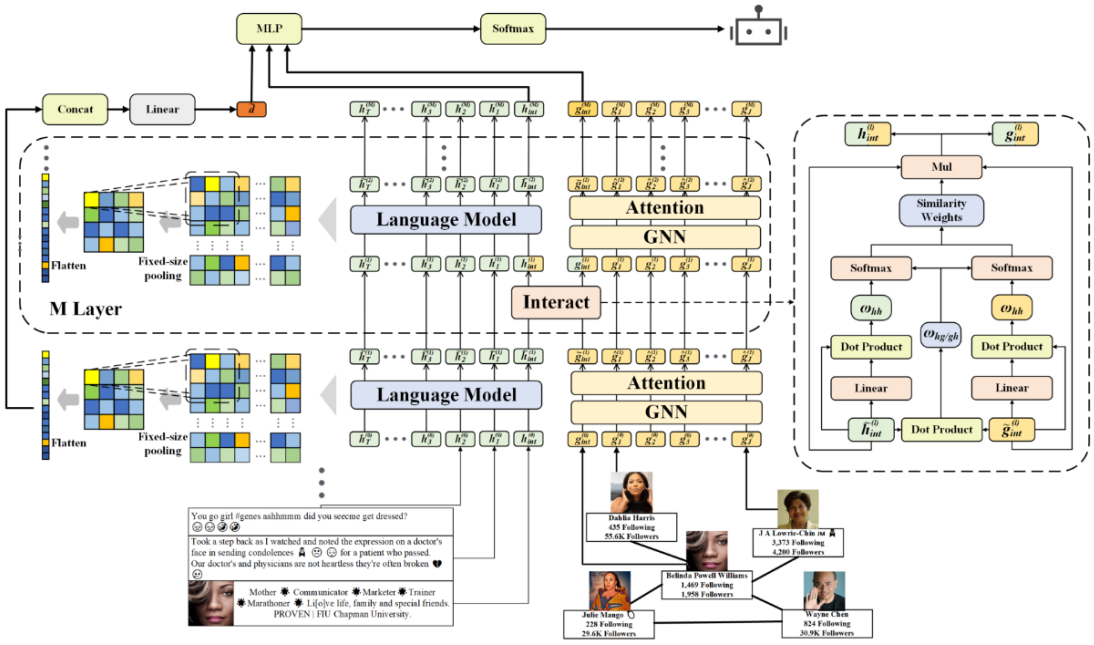
- Profile-Property Sub-Network
- Tweet-Semantic Sub-Network
- Following-Follower Sub-Network

Στο πρώτο σενάριο, 15 κατηγορικά, 5 αριθμητικά και 1 ειδικά κωδικοποιημένο feature παρατίθενται σε ένα Fully Connected Layer που ακολουθείται από μια συνάρτηση ενεργοποίησης ReLU. Με αυτόν τον τρόπο, προκύπτει ο ‘rp’ vector. Στο δεύτερο σενάριο, υλοποιούνται δύο ξεχωριστοί Encoders, Tweet-Level και Word-Level, και οι δύο εκ των οποίων περιέχουν ένα bidirectional Recurrent Neural Network. Στην πρώτη περίπτωση χρησιμοποιούνται γειτονικά tweets των οποίων το περιεχόμενο αθροίζεται, ενώ στη δεύτερη το input ταυτίζεται με τα embeddings των περικλειόμενων λέξεων των tweets, τα οποία παράγονται με Word2Vec. Και στις δύο οπτικές, ωστόσο, η κωδικοποίηση και η ανατροφοδότηση των δικτύων συνοδεύεται με αντίστοιχους attention μηχανισμούς, ενώ η τελική αναπαράσταση του χρήστη (user representation) καλείται ως ‘rs’ vector. Τέλος, το τρίτο υποσύστημα παίζει έναν καταλυτικό ρόλο ενώνοντας τα δύο προηγούμενα υποσυστήματα με την έννοια της γειτνίασης των χρηστών, η οποία συνοψίζεται στον ‘rn’ vector. Έπειτα από επαναληπτικές υπολογιστικές επαναλήψεις, στις οποίες τα διανύσματα ‘rp’ και ‘rs’ υπολογίζονται για κάθε χρήστη και συμβάλλουν στον καθορισμό του συνολικού vector ‘rn’, παρεμβάλλεται και ο Co-Influence Aggregator, ο οποίος είναι υπεύθυνος να χειριστεί τις επιμέρους επιρροές και να παρέχει το συνδυαστικό τελικό αποτέλεσμα.

Το **BIC** [32] διαιρεί τα αρχικά χαρακτηριστικά του χρήστη σε δύο γενικές μεγάλες κατηγορίες, οι οποίες είναι οι εξής :

- Numerical and Categorical
- Personal Description and Tweets

Η πρώτη τροφοδοτείται σε ένα Graph Attention Network, χτισμένο στα “following” relations μεταξύ των users. Η δεύτερη δρομολογείται στο γλωσσικό μοντέλο RoBERTa, το οποίο μετατρέπει τις αρχικές λεκτικές περιγραφές σε αριθμητικά embeddings. Τα δύο παραχθέντα αποτελέσματα, από το Graph και το Text propagation, συνδυάζονται σε ένα τρίτο interactive module, το οποίο εκπαιδεύεται με σκοπό να ανιχνεύσει τη σχετική σημασία των δύο components, υπολογίζοντας τις συσχετίσεις των δύο representative vectors. Τα δύο διαβαθμισμένα vectors που προκύπτουν, λοιπόν, επιστρέφουν και τροφοδοτούν το επόμενο layer των δύο components. Η πινελιά του BIC, ωστόσο, έγκειται σε ένα τέταρτο module, μέσω του οποίου ανιχνεύεται η σημασιολογική συνέπεια των tweets των users. Συγκεκριμένα, τα attention weights από τον Language Transformer χρησιμοποιούνται για να σχηματίσουν έναν matrix συσχετίσεων μεταξύ των tweets ενός user, κι έπειτα από κάποια pooling και flattening layers, να εξάγονται τυχόντα ανώμαλα μοτίβα στις δημοσιεύσεις των χρηστών.

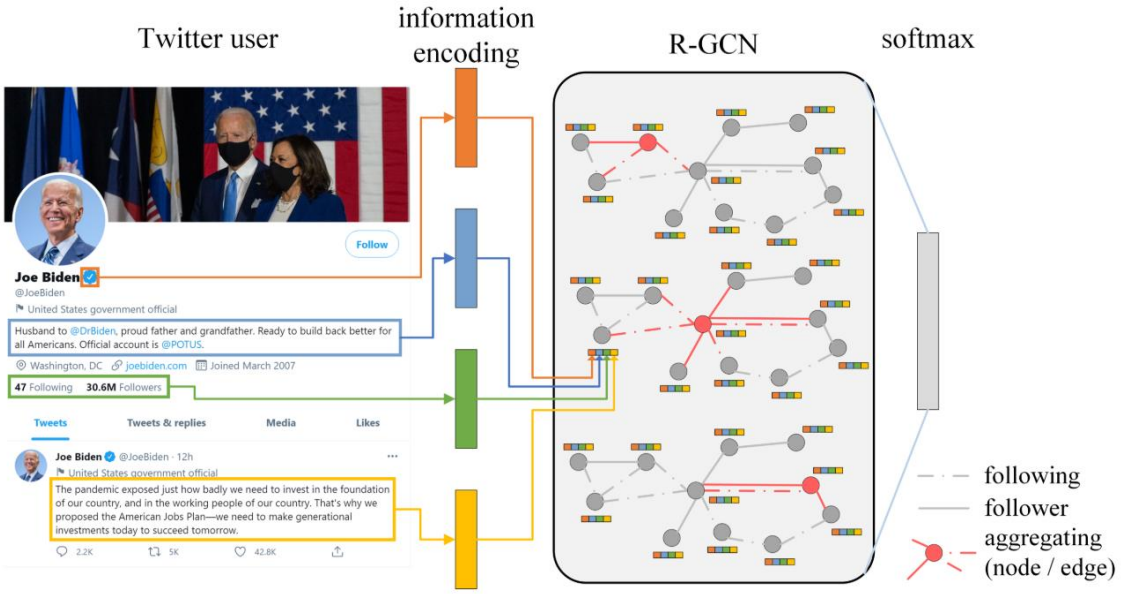


Εικόνα 36: Αρχιτεκτονική BIC

Το **BotRGCN** [16] χωρίζει τα γενικά στοιχεία του χρήστη σε τέσσερις γενικές κατηγορίες :

- Description
- Tweets
- Numerical
- Categorical

Υλοποιεί μια ξεχωριστή υπολογιστική διαδικασία για κάθε κατηγορία, η οποία οδηγεί και σε τέσσερα τελικά αντιπροσωπευτικά διανύσματα. Πιο λεπτομερειακά, τα πρώτα δύο τίθενται ως είσοδοι στο Natural Language Model 'RoBERTa', ενώ τα άλλα δύο τροφοδοτούνται σε δύο ατομικά Multi Layer Perceptrons. Με αυτή τη λογική, τα τέσσερα παραγόμενα vectors συνδυάζονται και αποτελούν το summary των χρηστών της κοινότητας, το οποίο και εκφράζει το χρήστη στην μετέπειτα πορεία του στο τερματικό και καθοριστικό Graph Neural Network. Το δίκτυο αυτό είναι ετερογενές και σχηματίζεται μέσω των following και followers relations των users. Αυτή η στρατηγική προσθέτει βαθύτητα στην συνεκτικότητα και την συνδεσιμότητα των χρηστών.



Εικόνα 37: Αρχιτεκτονική RGCN

3.6 Μειονεκτήματα και Περιορισμοί

Όπως έχουμε ήδη αναφέρει, κάθε υλοποιημένη μέθοδος ανίχνευσης bots στο Twitter παρέχει και μια ιδιαίτερη και διαφορετική προστιθέμενη αξία στο γενικό classification task. Μοναδικές αρχιτεκτονικές, με εξειδικευμένες προσεγγίσεις και στοχευμένες στρατηγικές, αξιοποιούν διαφορετικά χαρακτηριστικά και εκμεταλλεύονται με ξεχωριστό τρόπο τα διαθέσιμα μοντέλα και αλγορίθμους, προκειμένου να καταλήξουν σε ένα δυαδικό – αν μη τι άλλο – αποτέλεσμα. Παρόλο που όλα αυτά τα διάφορα συστήματα επιτυγχάνουν αξιόλογες βαθμολογίες και μετρικές στα πειράματα τους, από την άλλη

διαθέτουν επίσης και κάποιους περιορισμούς, οι οποίοι γεννούν την ανάγκη για αντιμετώπιση από μελλοντικά συστήματα.

Μια από τις μεγαλύτερες και πιο προφανείς παραλείψεις είναι η συμμετοχή αριθμητικών στατιστικών για τα tweets, τα οποία, εκτός από το αδιαμφησβήτητα απαραίτητο λεκτικό περιεχόμενο τους, παρέχουν επίσης και άκρως ενδιαφέροντα αριθμητικά στοιχεία, προς επεξεργασία, όπως :

- Likes
- Mentions
- Quotes
- Comments
- Retweets

και άλλα.

Ένα κοινό επιπρόσθετο μειονέκτημα των προαναφερθόντων υπάρχοντων συστημάτων είναι ότι όλα έχουν βασιστεί και εκπαιδευτεί σε σχετικά παλιά dataset, όπως το «**TwiBot-22**» ή και ακόμα παλαιότερες εκδόσεις αυτού. Τα bots δεν είναι παθητικές οντότητες οι οποίες παραμένουν στατικές στην πάροδο του χρόνου. Αντιθέτως, γίνονται όλα και πιο «έξυπνα», συνέχεια εξελίσσονται, με αποτέλεσμα να ανταπεξέρχονται και να προσαρμόζονται όλο και πιο αποδοτικά στα detection systems και filters. Καινούργια special bots με αποκλίνουσα συμπεριφορά είναι πλέον ευρέως διαδεδομένα και εξαπλωμένα στην πλατφόρμα του Twitter, γεγονός το οποίο καθιστά την ανάγκη για βελτίωση και επικαιροποίηση των συστημάτων όλο και πιο κρίσιμη στις μέρες μας.

Βέβαια, πέρα από αυτές τις βασικές παρατηρήσεις, οι οποίες θέτουν τις θεμελιώδεις κατευθυντήριες γραμμές, σχετικά με την ανάπτυξη νέων bot detection systems, πάντα θα υπάρχει και το αλγοριθμικό επίπεδο αλλαγών. Με άλλα λόγια, η υλοποίηση και η δοκιμή νέων μαθηματικών μοντέλων, ή η διαφορετική αξιοποίηση των υπάρχοντων, πάντα θα αποτελεί μια επαρκή κινητήριο δύναμη για εμφάνιση νέων προσεγγίσεων και σχετικών συστημάτων.

Όπως εύκολα γίνεται αντιληπτό, οι ανατροφοδοτούμενοι περιορισμοί και οι συνεχώς αυξημένες ανάγκες, γεννούν ένα ευρύ χώρο για βελτιώσεις και νέες προσεγγίσεις, οι οποίες – λαμβάνοντας υπόψιν και τις προηγούμενες – θα στενέψουν τα περιθώρια εξάπλωσης καινούργιων bots στο Twitter και θα οδηγήσουν σε μια συνολικότερη και επικαιροποιημένη αντιμετώπιση τους.

Πέρα από την πρακτική πτυχή της υπόθεσης, το ενδιαφέρον του προβλήματος μπορεί να προσδιοριστεί και ως επιστημονικό, αφού η διαχείριση και η επεξεργασία κοινωνικών γράφων, καθώς και η εφαρμογή συνελικτικών μοντέλων πάνω σε αυτούς, αποτελεί μια στρατηγική που δεν περιορίζεται αποκλειστικά και μόνο στην ανίχνευση bots στο Twitter, αλλά έχει τη δυνατότητα να αποτελέσει σημαντικό εργαλείο και σε άλλα social fields.

Κεφάλαιο 4

Σύνολο Δεδομένων

4.1 TwiBot-22

Το σύνολο δεδομένων που χρησιμοποιήθηκε κατά την υλοποίηση του συστήματος είναι το «TwiBot-22», το οποίο αποτελεί ένα ευρέως διαδεδομένο σημείο αναφοράς για την ανάπτυξη και τη σύγκριση πολλών αντίστοιχων συστημάτων ανίχνευσης Bot στο Twitter.

Η διαδικασία συλλογής των δεδομένων συνοψίζεται σε δύο στάδια. Προηγούμενες χρονικά εκδόσεις φανέρωσαν αδυναμία στη συλλογή διαφορετικών τύπων γνήσιων και μη αυθεντικών λογαριασμών. Έτσι, στο πρώτο στάδιο, βασισμένοι στην αποδοχή της διαφορετικότητας και της απόκλισης των χαρακτηριστικών των χρηστών του Twitter, πραγματοποιείται αναζήτηση κατά πλάτος στον κοινωνικό του γράφο, εκκινώντας από χρήστες «σπόρους» και εξαπλώνοντας μέσω των ακολούθων τους. Οι κύριες στρατηγικές που λαμβάνονται υπόψιν και εφαρμόζονται κατά αυτήν την διαδικασία του BFS (Breadth First Search) είναι οι εξής :

1. Distribution diversity. Έχοντας συμπεριλάβει τον current user στην κατά βάθος αναζήτηση, το κριτήριο με το οποίο επιλέγονται οι γείτονές του για συλλογή στο dataset, είναι τα metadata και η διακύμανσή τους (distribution). Συγκεκριμένα εκλέγονται οι k χρήστες με την υψηλότερη τιμή, οι k χρήστες με τη χαμηλότερη τιμή, καθώς και άλλοι k τυχαίοι χρήστες, των οποίων οι τιμές κυμαίνονται στο ενδιάμεσο. Metadata θεωρούνται τα χαρακτηριστικά ενός προφίλ χρήστη, όπως για παράδειγμα το πλήθος ακολούθων του, το πλήθος των likes του, και άλλα αριθμητικά δεδομένα. True or False χαρακτηριστικά μετατρέπονται άμεσα σε αριθμητικά τα οποία διαθέτουν αποκλειστικά και μόνο μια υψηλή και μια χαμηλή τιμή.
2. Value diversity. Η πιθανότητα να επιλεχτεί ένας χρήστης, έστω X, είναι αυξημένη αναλογικά με την «απόσταση», ή αλλιώς «απόκλιση», που διαθέτουν τα χαρακτηριστικά του (numerical metadata) σε σχέση με τα χαρακτηριστικά του «πατέρα» του. «Πατέρας» του θεωρείται ο χρήστης που μόλις συμπεριλήφθηκε στη συλλογή και διαθέτει ως γείτονά του τον X.

Στηριζόμενοι, λοιπόν, σε αυτές τις δύο στρατηγικές δειγματοληψίας και συλλογής χρηστών, γεννιέται το συνολικό κοινωνικό δίκτυο λογαριασμών, αυθεντικών και μη. Σημειώνεται πως για κάθε γειτονική εξάπλωση στη διαδικασία του BFS, επιστρατεύεται τυχαία μονάχα ένα metadata και μια από τις παραπάνω δύο οπτικές.

Στο δεύτερο στάδιο σκοπός είναι ο περαιτέρω εμπλουτισμός της ετερογένειας του δικτύου, καθώς συλλέγονται επιπρόσθετες σχέσεις και οντότητες.

Σαν αποτέλεσμα, δημιουργείται ένας εκτενής ετερογενής γράφος, ο οποίος περιέχει ως κόμβους τους συμμετέχοντες χρήστες, και ως ακμές τις σχέσεις «follow» μεταξύ τους. Ως νέα entities, βασιζόμενοι στον παραπάνω γράφο, εμπεριέχονται επίσης και tweets, lists, και mentioned hashtags, καθώς επίσης και 12 άλλα είδη συνδέσεων μεταξύ τους.

Ένα κύριο και απαραίτητο συστατικό της σύνθεσης ενός ολοκληρωμένου dataset είναι και η κατηγοριοποίηση των οντοτήτων. Εν προκειμένω, η ενδιαφέρουσα ως προς ταξινόμηση οντότητα είναι εκείνη των χρηστών, δηλαδή των real και bot users. Η διαδικασία «Data Annotation», λοιπόν, πλαισιώνει τη διαδικασία συλλογής των χρηστών και απαρτίζεται από τρία βήματα :

1. Labelization from experts. Τυχαία επιλέγονται 1000 χρήστες και τα στοιχεία τους παρατίθενται σε ειδικούς προκειμένου να αξιολογηθούν. Σκοπός είναι να ταξινομηθούν αυτοί οι χρήστες σε αληθινούς και ψεύτικους.
2. Generate noisy labels. Αξιοποιώντας μερικά στοχευμένα χαρακτηριστικά των χρηστών, όπως ευαίσθητα tweets και spam keywords, καθώς και υποσύνολο των αριθμητικών metadata, εφαρμόζονται 8 χειρωνακτικές συναρτήσεις και 7 μοντέλα νευρωνικών δικτύων αντίστοιχα. Η εκπαίδευση λαμβάνει χώρα στο σύνολο των χρηστών των οποίων η ετικέτα είναι γνωστή μέσω της γνωμάτευσης των ειδικών που προηγήθηκε. Ως αποτέλεσμα έχουμε την αβέβαιη ταξινόμηση όλων των χρηστών του «TwiBot-22» (1000000 σε πλήθος), μέσω ετικετών που διαθέτουν θόρυβο (noisy labels). Έπεται κατάλληλο φιλτράρισμα των τεσσάρων μοντέλων που χρησιμοποιήθηκαν (MLP, GAT, GCN και R-GCN), καθώς και των προβλέψεών τους, προκειμένου να αφαιρεθούν εκείνες με το μεγαλύτερο ποσοστό αβεβαιότητας.
3. Majority voting. Αφού αποκτηθούν τα noisy labels των χρηστών, ακολουθεί η εκτίμηση της αληθοφάνειας του κατά Snorkel, καθώς και ο ταυτόχρονος καθαρισμός τους. Η έξοδος του συστήματος Snorkel είναι οι πιθανολογικές ετικέτες των χρηστών, οι οποίες εν τέλει θα χρησιμοποιηθούν, σε συνδυασμό με τα χαρακτηριστικά των χρηστών, σε έναν τελικό MLP classifier. Έτσι προκύπτουν και τα τελικά annotations του «TwiBot-22», τα οποία παρέχουν αυξημένη ακρίβεια σε σύγκριση με το «TwiBot-20», λαμβάνοντας υπόψιν τις αρχικές τοποθετήσεις των ειδικών.

Το σύνολο δεδομένων «TwiBot-22» καταλήγει να παρέχει πιο εμπλουτισμένα και πιο ακριβή χαρακτηριστικά χρηστών του Twitter, σε σχέση με τους ανταγωνιστές του. Διακρίνεται βέβαια και για την πληθώρα οντοτήτων κα συσχετίσεων, προσφέροντας έτσι, πέρα από αξιοπιστία και συνέπεια, ποικιλία και πολυμορφικότητα.

Dataset	C-15	G-17	C-17	M-18	C-S-18	C-R-19	B-F-19	TwiBot-20	TwiBot-22
# Human	1,950	1,394	3,474	8,092	6,174	340	380	5,237	860,057
# Bot	3,351	1,090	10,894	42,446	7,102	353	138	6,589	139,943
# User	5,301	2,484	14,368	50,538	13,276	693	518	229,580	1,000,000
# Tweet	2,827,757	0	6,637,615	0	0	0	0	33,488,192	88,217,457
# Human Tweet	2,631,730	0	2,839,361	0	0	0	0	33,488,192	81,250,102
# Bot Tweet	196,027	0	3,798,254	0	0	0	0	33,488,192	6,967,355
# Edge	7,086,134	0	6,637,615	0	0	0	0	33,716,171	170,185,937

Πίνακας 1: Πληροφορίες σχετικά με τα Twitter datasets

4.2 Αξιοποιημένο Σύνολο Δεδομένων

Το αρχικό μας dataset «TwiBot-22», όπως έχει αναφερθεί κατά την περιγραφή του, παρέχει στοιχεία για 1 εκατομμύριο χρήστες. Ως στοιχεία θεωρούμε αριθμητικά και κατηγορικά τους χαρακτηριστικά, περιγραφές, tweets, καθώς και πλήθος διαφορετικών συνδέσεων μεταξύ τους. Συγκεκριμένα :

user.json

Το αρχείο περιλαμβάνει τα προσωπικά χαρακτηριστικά όλων των συμμετεχόντων χρηστών, τα οποία έχουν εξαχθεί από τα αντίστοιχα προφίλ των λογαριασμών. Αναφέρουμε επιγραμματικά εκείνα τα οποία αξιοποιούμε στο σύστημα μας, αλλά και εκείνα τα οποία θα μπορούσαμε να εκμεταλλευτούμε μελλοντικά :

ID χρήστη

Ημερομηνία δημιουργίας λογαριασμού

Περιγραφή χρήστη

Τοποθεσία

Όνομα

Username

Protected (OR NOT) λογαριασμός

Verified (OR NOT) λογαριασμός

Αριθμός followers

Αριθμός following

Πλήθος tweets

Πλήθος λιστών

Προσωπικά links

Φωτογραφία προφίλ

Προσωπικά widgets - entities

Σχετικά με τα προσωπικά links και widgets – entities του χρήστη, να διευκρινίσουμε πως ένα προφίλ στο Twitter συχνά διαθέτει και extra παραμέτρους στην περιγραφή του. Για παράδειγμα διάφορους συνδέσμους που παραπέμπουν σε εξωτερικές σελίδες, οι οποίες είτε μπορούν να ανήκουν σε τρίτους λογαριασμούς του χρήστη είτε όχι. Άλλες περιπτώσεις είναι διάφορα hashtags ή εικονίδια, τα οποία έχει ενσωματώσει ο χρήστης σαν μέρος της περιγραφής του. Γενικά το σύστημά μας επικεντρώνεται αποκλειστικά στο λεκτικό μέρος της περιγραφής, αγνοώντας όλα τα παραπάνω.

Σε μια μελλοντική έκδοση, όλα τα παραπάνω widgets, αλλά και η φωτογραφία προφίλ του χρήστη, μπορούν να αξιοποιηθούν μέσω εξωτερικών μοντέλων και της transfer learning στρατηγικής, προκειμένου να αποδώσουν embeddings που θα εμπλουτίσουν το πεδίο ορισμού των χαρακτηριστικών του χρήστη κι έτσι θα αυξήσουν την ακρίβεια του συστήματός μας.

label.csv

Το αρχείο περιλαμβάνει τις ετικέτες των συμμετεχόντων χρηστών, πληροφορώντας για το εάν ένας χρήστης είναι Human OR Bot. Η σύνδεση γίνεται μέσω του ID του χρήστη.

tweet.json

Το αρχείο περιλαμβάνει τα χαρακτηριστικά των tweets των συμμετεχόντων χρηστών. Αναφέρουμε επιγραμματικά εκείνα τα οποία αξιοποιούμε στο σύστημα μας, αλλά και εκείνα που θα μπορούσαμε να εκμεταλλευτούμε μελλοντικά :

ID tweet

ID συγγραφέα

Text Content

ID παραλήπτη

Πλήθος retweets

Πλήθος replies

Πλήθος quotes

Αριθμός likes

Reference links

Ημερομηνία δημιουργίας tweet

Ενσωματωμένα widgets - media

Γλώσσα

Η ημερομηνία δημιουργίας και η γλώσσα του tweet είναι δύο απλά χαρακτηριστικά που θα μπορούσαν να συμπεριληφθούν στο υπάρχον σύστημα, αλλά επιλεκτικά απορρίφθηκαν. Όσον αφορά τα reference links και τα ενσωματωμένα widgets – media είναι extra παράμετροι ενός tweet που συχνά παίζουν καθοριστικό ρόλο στη φύση και την προέλευσή τους. Συνεπώς σε μια επόμενη version του συστήματος, μπορούν να αξιοποιηθούν μέσω γλωσσικών (για links) και computer vision (για εικονίδια) μοντέλων και με μεταφορά μάθησης να ενισχύσουν το πεδίο ορισμού των χαρακτηριστικών των tweets, αυξάνοντας τη χρήσιμη πληροφορία σχετικά με αυτά και κατ' επέκταση βελτιώνοντας την απόδοση του συνολικότερου συστήματος.

edge.csv

Το αρχείο αυτό διαθέτει όλες τις υφιστάμενες συνδέσεις μεταξύ των δύο κύριων οντοτήτων του συνόλου δεδομένων. Ως κύριες οντότητες ορίζουμε φυσικά τους χρήστες και τα tweets. Υπάρχουν συνολικά 14 πιθανές συνδέσεις ανάμεσα σε αυτές τις δύο

οντότητες, τις οποίες σημειώνουμε επιγραμματικά παρακάτω και τις διαχωρίζουμε σε εκείνες που χρησιμοποιούμε ενεργά στο σύστημά μας και σε εκείνες που θα μπορούσαμε να λάβουμε υπόψιν μελλοντικά :

following (user – user)

post (user – tweet)

followers (user – user)

pinned (user – tweet)

own (user – list)

membership (list – user)

retweeted (tweet – tweet)

like (user – tweet)

followed (user – user)

quoted (tweet – tweet)

discuss (tweet – hashtag)

replied (tweet – tweet)

mentioned (tweet – user)

contain (list – tweet)

Εύκολα κάποιος παρατηρεί πως υποβόσκουν και δύο επιπρόσθετες οντότητες στις παραπάνω συσχετίσεις : list και hashtag. Η πρώτη δημιουργείται και τροποποιείται από χρήστες, ανήκει σε αυτούς και περιέχει tweets που παρουσιάζουν έντονο υποκειμενικό ενδιαφέρον. Η δεύτερη ουσιαστικά οργανώνει τα tweets σε conversations γύρω από ένα συγκεκριμένο και εξειδικευμένο θέμα – topic.

Η πλειονότητα των συσχετίσεων δε χρησιμοποιείται ενεργά από το σύστημα μας, είτε λόγω αυξημένης πολυπλοκότητας και περιττής πληροφορίας (προσθήκη οντοτήτων hashtag και list, πιο ιδιαίτερες συνδέσεις μεταξύ των υπάρχοντων οντοτήτων user και tweet), είτε επειδή εμπεριέχεται με έμμεσο τρόπο στα ίδια τα χαρακτηριστικά των βασικών οντοτήτων, τα οποία περιγράψαμε προηγουμένως (replied, mentioned). Δηλαδή με άλλα λόγια δε θα είχε νόημα η διπλή άντλησή τους από το σύνολο δεδομένων.

Κεφάλαιο 5

Προτεινόμενη Μέθοδος

5.1 Εισαγωγή

5.1.1 Συνελικτικά Δίκτυα Γράφων

Η διαφοροποίηση σε ένα Convolutional Network, σε σύγκριση με ένα απλό Neural Network, έγκειται στο γεγονός πως η φύση των εισερχόμενων δεδομένων παρουσιάζει τοπολογική συμπεριφορά. Με άλλα λόγια, οι οντότητες, των οποίων τα χαρακτηριστικά καλούνται να εισαχθούν και να αξιολογηθούν από το δίκτυο, δεν είναι πλέον εντελώς ανεξάρτητες μεταξύ τους, αλλά διαθέτουν μια τοποθέτηση και ταξινόμηση στο χώρο των δεδομένων. Κατ' επέκταση, κάθε οντότητα εμπεριέχεται σε μια γειτονιά αντίστοιχων οντοτήτων, από την οποία μοιραία επηρεάζεται, αλλά επηρεάζει κιόλας.

Στην ειδική περίπτωση των Graph Convolutional Networks, η τοπολογική ταξινόμηση των οντοτήτων δεν έρχεται με μια ευθύγραμμη και τετριμμένη λογική. Συγκεκριμένα, ένας γράφος απαρτίζεται από αυθαίρετες και γεωμετρικά ελεύθερες συσχετίσεις, ανάμεσα στις κορυφές του. Δεν περιορίζεται στις ιδιότητες των στοιχείων μιας εικόνας, της οποίας τα pixels είναι στοιχισμένα σε μια καθορισμένη σειρά.

5.1.2 Κύρια ιδέα

Στο συγκεκριμένο κεφάλαιο θα παρουσιάσουμε μια νέα ανταγωνιστική προσέγγιση στο πρόβλημα του bot detection στην πλατφόρμα του Twitter. Η μέθοδος που προτείνεται βασίζεται στην έννοια της αλληλεπίδρασης των χρηστών και στη δύναμη την επιρροής που διαθέτουν οι συσχετίσεις και οι συνδέσεις ενός κοινωνικού γραφήματος.

Το σύστημα αξιοποιεί συνελικτικά δίκτυα γράφων τα οποία αρχικοποιούνται και προσαρμόζονται κατάλληλα έτσι ώστε να αποδίδουν με βέλτιστο τρόπο χρήσιμα deep insights, ανάλογα με το είδος πληροφορίας που τους διοχετεύεται. Η πληροφορία που ορίζει, καθορίζει και περιγράφει ένα χρήστη ταυτίζεται με ένα αντιπροσωπευτικό διάνυσμα, το οποίο συμπεριλαμβάνει αριθμητικές τιμές και τοποθετεί το στίγμα του χρήστη στο γενικότερο χάρτη των πιθανών χαρακτηριστικών. Ως χαρακτηριστικά ενός χρήστη τίθενται προσωπικά του στοιχεία που λαμβάνονται μέσω του Twitter profile του, όπως κάποια general stats (followers, likes κλπ), η περιγραφή ή τα tweets του.

Μέσω των GCNs η πληροφορία των χρηστών ταξιδεύει σε όλο το φάσμα του κοινωνικού γραφήματος, επιτρέποντας την αναπροσαρμογή των χαρακτηριστικών τους και κατ' επέκταση της τελικής τους προβλεπόμενης ταυτότητας (real or bot).

Από την άλλη μεριά, μέσω της ενσωμάτωσης διαφορετικών ειδών πληροφορίας, εξασφαλίζεται και προωθείται η ιδιότητα της γενίκευσης, καθώς ο χαρακτηρισμός ενός χρήστη δεν περιορίζεται σε μια μονόπλευρη διάσταση, αλλά διαθέτει πολλές μορφές. Η ανάγκη για γενίκευση είναι κρίσιμη καθώς η παρουσία και οι ενέργειες των bots στην πλατφόρμα του Twitter συνεχώς προσαρμόζονται και μεταμορφώνονται, δυσχεραίνοντας τη λειτουργία των ανιχνευτών.

5.2 Ορισμός προβλήματος

Στη συγκεκριμένη παράγραφο θα παρουσιαστεί συνοπτικά ο ορισμός του προβλήματος της ανίχνευσης του Twitter, βασισμένος στο σύνολο των παραμέτρων που αξιοποιεί το μοντέλο μας.

Έστω U ένας τυχαία επιλεγμένος χρήστης του Twitter. Η πληροφορία του διαχωρίζεται σε τρία πλαίσια.

- Ιδιωματική πληροφορία (P) : Αφορά όλα εκείνα τα στοιχεία (αριθμητικά, κατηγορικά, περιγραφικά) που προέρχονται απευθείας από το προφίλ του. Επιμερίζεται σε p_num και p_desc.
- Σημασιολογική πληροφορία (T) : Αφορά όλα εκείνα τα στοιχεία που αντλούνται από τις δημοσιεύσεις του χρήστη. Χωρίζεται σε t_cont και t_stats. Στο πρώτο αποτυπώνεται το περιεχόμενο των επιμέρους tweets, ενώ στο δεύτερο αποθηκεύονται τα αντίστοιχα αριθμητικά στατιστικά των tweets.
- Κοινωνική πληροφορία (N) : Αφορά τις επαφές – συνδέσεις του χρήστη, οι οποίες καθορίζονται από τις σχέσεις "following" που υφίστανται στο Twitter.

Συνεπώς, θεωρώντας το πρόβλημα της ανίχνευσης bots στο Twitter ως ένα πρόβλημα δυαδικής ταξινόμησης (0 for real και 1 for bots), έχουμε :

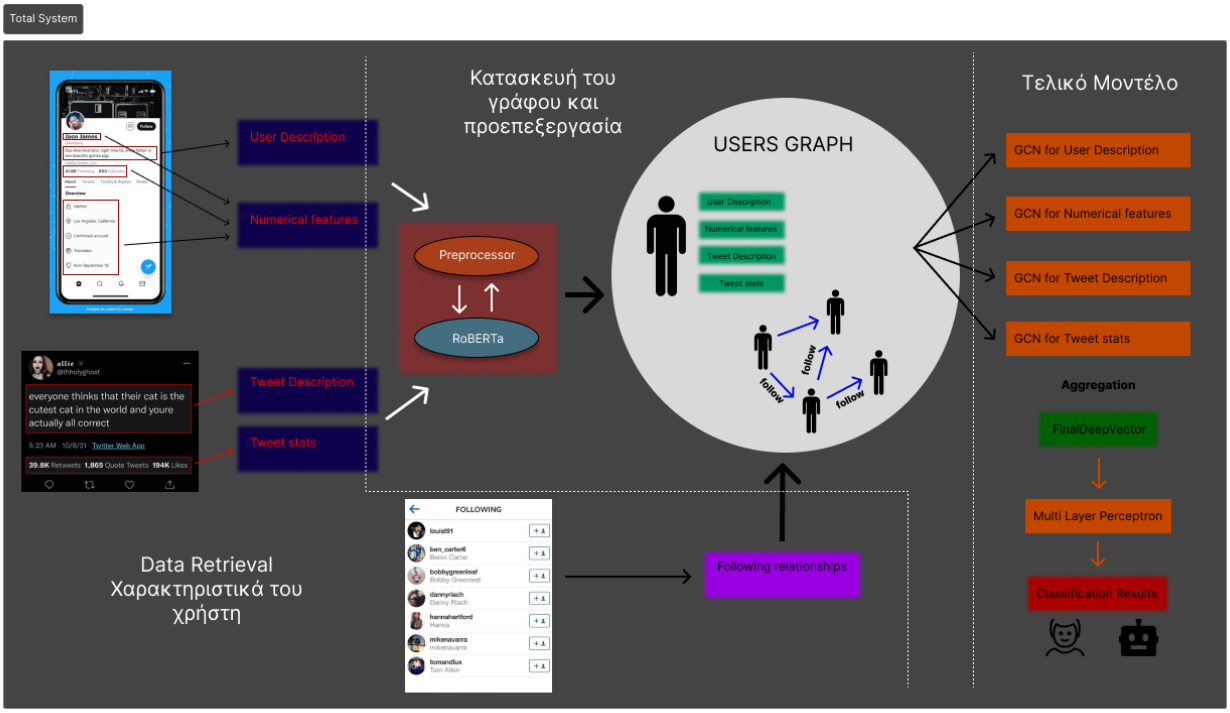
Problem: ***Twitter Bot Detection***

Given a Twitter User ***U*** *and its information* ***P***, ***T***, ***N***, *learn a bot detection function*

$$f : f(P(U), T(U), N(U)) \longrightarrow \ddot{Y}$$

such that $\ddot{Y}$ *approximates ground truth Y to maximize prediction accuracy.*

5.3 Μεθοδολογία



Εικόνα 38: Συνολικό σύστημα mGCN

Στην παραπάνω εικόνα φαίνεται η συνολική αρχιτεκτονική του συστήματος μας ‘mGCN’. Το pipeline χωρίζεται σε τρία επιμέρους τμήματα :

1. Χαρακτηριστικά του χρήστη
2. Κατασκευή του γράφου και προεπεξεργασία
3. Τελικό Μοντέλο

τα οποία αναλύονται εξαντλητικά στις παρακάτω παραγράφους. Το πρώτο σχετίζεται με τη συλλογή και αρχικοποίηση των features του χρήστη. Το δεύτερο είναι υπεύθυνο για την επεξεργασία των χαρακτηριστικών και το σχηματισμό του τελικού γράφου. Το τρίτο περιέχει τα τέσσερα συνελικτικά νευρωνικά δίκτυα GCN, καθώς και το τερματικό MLP, μέσω των οποίων θα καταλήξουμε στις τελικές προβλέψεις.

5.3.1 Χαρακτηριστικά του χρήστη

Στο πρόβλημα μας θεωρούμε ως κορυφές του γράφου τους χρήστες του Twitter, και ως ακμές του τις συνδέσεις ανάμεσα τους. Η σύνδεση ανάμεσα στον χρήστη U1 και το χρήστη U2 καθορίζεται από το εάν ο U1 ακολουθεί τον U2.

Αρχικά, σαν πρώτο βήμα αποθηκεύουμε τα ενδιαφέροντα χαρακτηριστικά όλων των χρηστών που μας παρέχει το Dataset (1 εκατομμύριο σε σύνολο). Το σύνολο των χαρακτηριστικών, που περιγράφει και εκπροσωπεί τον χρήστη U στο γράφο, μοιράζεται σε τέσσερα «πακέτα» :

1. numerical
2. description
3. tweets
4. tweets_numerical

Ο γενικός πίνακας που περιέχει όλα τα επιμέρους χαρακτηριστικά που συμμετέχουν στις δύο πρώτες ομάδες αντλείται από το 'user.json' και είναι :

[Description, Name, Username, Location, EntitiesCounter, YearsOfExistence, Protected, FollowersCount, FollowingCount, TweetsCount, ListedCount, Verified]

όπου :

Description = Λεκτική περιγραφή του χρήστη (string)

Name = Όνομα του χρήστη (string)

Username = Username του χρήστη (string)

Location = Τοποθεσία του χρήστη (string)

EntitiesCounter = Πλήθος entities – widgets στη συνολική περιγραφή του χρήστη (int)

YearsOfExistence = Χρόνια ύπαρξης του χρήστη (int)

Protected = Flag για το εάν ο λογαριασμός του χρήστη είναι προστατευμένος (bool)

FollowersCount = Αριθμός followers του χρήστη (int)

FollowingCount = Αριθμός following του χρήστη (int)

TweetsCount = Πλήθος tweets του χρήστη (int)

ListedCount = Πλήθος λιστών του χρήστη (int)

Verified = Flag για το εάν ο λογαριασμός του χρήστη είναι επικυρωμένος (bool)

Ο γενικός πίνακας που περιέχει όλα τα επιμέρους χαρακτηριστικά που συμμετέχουν στις δύο τελευταίες ομάδες αντλείται από το 'tweet.json' και είναι :

[TweetText, Retweets, Replies, Quotes, Likes]

όπου :

TweetText = Λεκτικό περιεχόμενο του tweet (string)

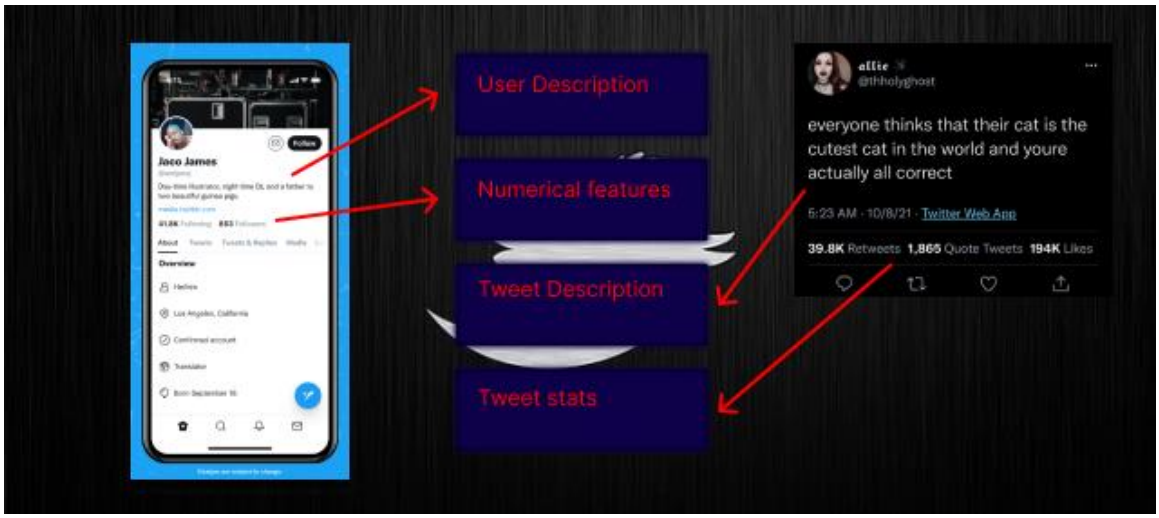
Retweets = Πλήθος των αναδημοσιεύσεων του tweet (int)

Replies = Πλήθος απαντήσεων του tweet (int)

Quotes = Πλήθος σχολίων του tweet (int)

Likes = Πλήθος likes του tweet (int)

Σε αντίθεση με τον προηγούμενο πίνακα – που αφορά τις δύο πρώτες ομάδες – ο συγκεκριμένος πίνακας απαρτίζεται από χαρακτηριστικά που συνδέονται με tweets και όχι users. Η αντιστοίχιση των tweets με users, η οποία θα μας οδηγήσει και στην εξαγωγή των χαρακτηριστικών που εκπροσωπούν χρήστες και όχι tweets, εξηγείται και αναλύεται παρακάτω.



Εικόνα 39: Άντληση δεδομένων από το Twitter

5.3.2 Κατασκευή του γράφου και προεπεξεργασία

Αφού έχουν αποθηκευτεί οι vectors όλων των users, ακολουθεί σαν επόμενο βήμα η ανίχνευση των συνδέσεων μεταξύ τους. Όπως αναφέρθηκε και προηγουμένως, οι συνδέσεις ανάμεσα στους χρήστες καθορίζονται από τις σχέσεις “following” και “followers” μεταξύ τους. Εάν η βάση δεδομένων μας ενημερώνει πως υπάρχει σχέση “following” από τον U1 στον U2, τότε η ακμή [U1, U2] προστίθεται στο γράφο. Εάν η βάση δεδομένων μας ενημερώνει πως υπάρχει σχέση “followers” από τον U1 στον U2, τότε η ακμή [U2, U1] προστίθεται στο γράφο. Τονίζουμε πως στο Twitter δεν εξασφαλίζεται συμμετρία στα follows. Το ότι ο χρήστης U1 ακολουθεί το χρήστη U2 δεν συνεπάγεται πως ισχύει και το αντίστροφο, δηλαδή ότι ο χρήστης U2 ακολουθεί το χρήστη U1. Επομένως ο γράφος σχηματίζεται με κατευθυνόμενες ακμές.

Προκειμένου να σχηματιστεί ένας υπολογιστικά διαχειρίσιμος γράφος, το σύστημά μας δε θα συνεργαστεί με όλους τους χρήστες (1 εκατομμύριο σε σύνολο). Αντιθέτως, οι χρήστες που θα συμμετάσχουν στο γράφο εκλέγονται με τρια κριτήρια :

1. Τη χωρητικότητα τους σε ακμές (neighbourhood capacity)
2. Την ετικέτα τους (normal or bot)
3. Την διαθεσιμότητα τους σε tweets

Το πρώτο κριτήριο μας βοηθά να ελέγχουμε την πυκνότητα του σχηματιζόμενου γράφου. Εύλογα, λοιπόν, διαμορφώνονται διάφορα σενάρια, τρία εκ των οποίων παρουσιάζουν το πιο έντονο ενδιαφέρον :

Α. Γράφος με υψηλή πυκνότητα

Β. Γράφος με χαμηλή πυκνότητα

Γ. Γράφος με μέση πυκνότητα

Στην πρώτη περίπτωση επιλέγονται με προτεραιότητα χρήστες που διαθέτουν μεγάλο outbound βαθμό, δηλαδή έχουν πολλούς γείτονες. Αντιθέτως, στη δεύτερη περίπτωση επιλέγονται με προτεραιότητα χρήστες που διαθέτουν μικρό outbound βαθμό. Τέλος, στην τρίτη περίπτωση επιλέγονται χρήστες τυχαία, χωρίς να λαμβάνεται υπόψιν ο outbound βαθμός τους.

Το δεύτερο κριτήριο μας βοηθά να ελέγχουμε το ποσοστό των bots που ο γράφος ελλοχεύει. Είναι ένα κριτήριο που βαφτίζεται αναγκαίο σε κάθε classification problem, καθώς το σύνολο δεδομένων πρέπει να είναι φυσικά ισορροπημένο έως ένα επίπεδο. Ένα σύστημα δεν μπορεί και δεν πρόκειται να παράξει σωστές προβλέψεις όταν είναι προπονημένο σε ένα ανισόρροπο dataset, στο οποίο η κλάση Α υπερισχύει δραματικά της κλάσης Β, αναφερόμενοι σε πληθυσμό. Έτσι στο πρόβλημά μας γεννούνται ξανά πολλά σενάρια, αν και εμείς θα κυμανθούμε σε ποσοστά bots 40 – 50%.

Το τρίτο κριτήριο αποτελεί μια βασική προϋπόθεση ως προς τη συμμετοχή ενός χρήστη στο γράφο. Μιας και το σύστημά μας αναπτύσσεται με βάση divergent χαρακτηριστικά χρηστών, συμπεριλαμβάνουμε επί σκοπού μόνο χρήστες που διαθέτουν tweets. Σε μια μελλοντική version θα αντιμετωπίζονται και πιο «κενοί» χρήστες, ωστόσο προς το παρόν για λόγους πληρότητας παραμένουμε στο αρχικό πλάνο.

Όλα τα παραπάνω συναληθεύονται με τον αυστηρό περιορισμό του συνολικού αριθμού χρηστών που θα απαρτίσουν το γράφο. Συγκεκριμένα, το σύστημα μας ανταπεξέρχεται υπολογιστικά και λειτουργεί με 4-8 χιλιάδες χρήστες, αριθμός ικανός για να προσδώσει χρήσιμα αποτελέσματα.

Ο σχηματισμός του γράφου ολοκληρώνεται με την μετατροπή των vectors των users στην κατάλληλη (για ένα νευρωνικό δίκτυο) μορφή. Παραπάνω περιγράψαμε αναλυτικά τη σημασία των χαρακτηριστικών των χρηστών, ωστόσο δεν εμβαθύναμε επαρκώς στον

τύπο τους σαν δεδομένα. Για την ακρίβεια, προκειμένου ένα νευρωνικό δίκτυο να εφαρμοστεί σωστά, πρέπει η είσοδος του να παρουσιάζει συνοχή και είναι διαχειρίσιμη σε μαθηματικά μοντέλα και αλγόριθμους. Συνεπώς, ονομάζουμε graph_vector[U] τον πίνακα τον οποίο διαθέτει τις τελικές τιμές των χαρακτηριστικών των χρηστών που θα εισαχθούν στο σύστημα μας, και ο οποίος προέρχεται από τον αρχικό vector[U] μέσω των εξής μετατροπών :

Πρώτη φάση

Description (string) → *DescEmbedding* (type = float, dimensions = 768)

TweetText (*string*) → *TweetEmbedding* (type = float, dimensions = 768)

TweetEmbedding → *Merge by author* (user) → *Select the most representative tweets for every user* → *FinalEmbedding*

(type = float, dimensions = 768)

Name (string), *Username* (string), *Location* (string) → *NameLocFields* (type = int)

Δεύτερη φάση

description	=	*DescEmbedding*
numerical	=	[*NameLocFields*, *EntitiesCounter*, *YearsOfExistence*, *Protected*, *FollowersCount*, *FollowingCount*, *TweetsCount*, *ListedCount*, *Verified*]
tweets	=	*FinalEmbedding*
tweets_numerical	=	[*Retweets, Replies, Quotes, Likes*]

Τρίτη φάση

numerical → ***scaled_numerical***

tweets_numerical → ***scaled_tweets_numerical***

Στην πρώτη φάση ουσιαστικά η λεκτική περιγραφή ενός χρήστη και το κείμενο ενός tweet αποκτά αριθμητική υπόσταση μέσω του γλωσσικού μοντέλου :

'all-MiniLM-L6-v2' (βιβλιοθήκη = ***sentence_transformers***)

'twitter-roberta-base' (βιβλιοθήκη = **sentence_transformer**)

Εφαρμόζεται δηλαδή μεταφορά μάθησης, προκειμένου να αποτυπώσουμε με βάθος το περιεχόμενο της περιγραφής κάθε χρήστη σε έναν κανονικοποιημένο πίνακα float με 768 dimensions. Σκοπός αυτής της μετατροπής είναι η αριθμητική φύση του συγκεκριμένου χαρακτηριστικού, η οποία κρίνεται απαραίτητη κατά την επεξεργασία του σε ένα μαθηματικό μοντέλο του οποίου ο αλγόριθμος εμπεριέχει πράξεις και συγκρίσεις. Το ίδιο εφαρμόζεται και στα περιεχόμενα των tweets, μόνο που σε αυτήν την περίπτωση, προκειμένου να καταλήξουμε στον τελικό χαρακτηριστικό πίνακα που θα εκπροσωπήσει το χρήστη, ακολουθούμε πρώτα δύο κινήσεις :

Α. Ομαδοποίηση κατά συγγραφέα (user)

Β. Εύρεση των πιο αντιπροσωπευτικών tweets κάθε χρήστη

Γ. Συγχώνευση των tweets που ανιχνεύτηκαν στο Γ σε ένα συνολικό κείμενο

Δ. Εκ νέου υπολογισμό του embedding του συνολικού κειμένου

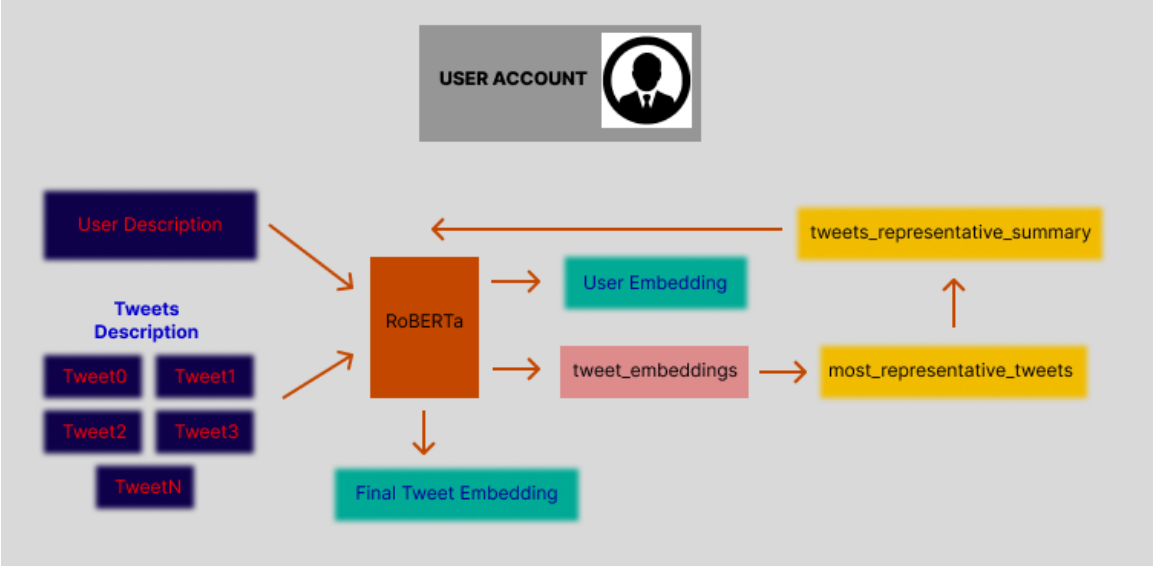
Από την άλλη, όσον αφορά τις άλλες τρεις λεκτικές μεταβλητές, αυτές συμπτύσσονται σε μια αριθμητική με βάση την ύπαρξή τους. Επειδή πολλά προφίλ χρηστών δεν παρουσιάζουν απαραίτητα λεπτομέρειες σχετικά με προσωπικά ονόματα και τοποθεσία, κρίνεται καίρια η καταγραφή του πλήθους (από 0 έως 3) των συγκεκριμένων στοιχείων. Έτσι για παράδειγμα ο χρήστης που παρέχει μόνο Username και Location αποκτά την τιμή 2 στη συγκεκριμένη μετρική, ενώ κάποιος άλλος που δεν παρέχει τίποτα από τα τρία αποκτά τιμή 0.

Στη δεύτερη φάση τα συνολικά χαρακτηριστικά οργανώνονται κα χωρίζονται σε τέσσερις επιμέρους κατηγορίες, τα description και τα numerical. Η πρώτη και η τρίτη ταυτίζονται με τα embeddings που λαμβάνονται ως output από το γλωσσικό μοντέλο, ενώ η δεύτερη και η τέταρτη αποτελούνται από όλες τις μεμονωμένες αριθμητικές μετρικές.

Στην τρίτη και τελευταία φάση, τα αριθμητικά χαρακτηριστικά κανονικοποιούνται μέσω του μοντέλου :

'MinMaxScaler' (βιβλιοθήκη = ***sklearn***, τομέας = ***preprocessing***)

Με αυτόν τον τρόπο, το σύνολο τιμών των χαρακτηριστικών αποκτά κοινή βάση και κλίμακα, με αποτέλεσμα η επεξεργασία αυτών να καθίσταται πιο ομαλή και αποδοτική. Υπενθυμίζεται πως η διαδικασία αυτή δε χρειάζεται να εφαρμοστεί και στο description, αφού έχει ήδη πραγματοποιηθεί άμεσα από το γλωσσικό μοντέλο.



Εικόνα 40: Προεπεξεργασία γλωσσικών δεδομένων

5.3.3 Τελικό μοντέλο του συστήματος

Το σύνολο δεδομένων μας αποτελείται από divergent users που διαθέτουν κάποια – διαφορετικών μορφών – χαρακτηριστικά, όπως αναλυτικά περιγράφτηκε παραπάνω. Η ιδιαιτερότητα του συνόλου δεδομένων, η οποία μας οδήγησε στο να χρησιμοποιήσουμε Συνελικτικά Νευρωνικά Δίκτυα Γράφων (GCN) και όχι απλούς ταξινομητές πολλαπλών στρωμάτων (MLP), είναι η ύπαρξη σχέσεων μεταξύ των χρηστών, η οποία δημιουργεί ένα γράφο, ή αλλιώς μια τοπολογική ταξιθέτηση στο χώρο των δεδομένων.

Όπως αναφέραμε νωρίτερα, τα χαρακτηριστικά των χρηστών επιμερίζονται σε τέσσερις κατηγορίες :

1. description
2. numerical
3. tweets
4. tweets_numerical

κάθε μια από τις οποίες θέτεται υπό διαφορετική επεξεργασία προτού αποτελέσει το τελικό διάνυσμα χαρακτηριστικών του χρήστη. Προκειμένου να αξιοποιήσει στο έπακρο την ποικιλομορφία και τη διαφορετικότητα των τεσσάρων παραπάνω κατηγοριών, το σύστημα μας εφαρμόζει τέσσερα αντίστοιχα επιμέρους GCN, εξάγοντας με αυτόν τον τρόπο τέσσερα outputs, τα οποία βαφτίζονται “deep”, καθώς αποτελούν βαθιά representations των χρηστών του δικτύου, έπειτα από πολλαπλούς κύκλους αναδρομικών αλληλεπιδράσεων με τους γείτονές τους. Υπενθυμίζουμε πως ως γείτονες ενός χρήστη U εννοούνται οι χρήστες N(U) για τους οποίους ισχύει :

following (U, N(U)) = True

Τα αριθμητικά χαρακτηριστικά (2 και 4) τοποθετούνται σαν inputs σε ένα GCN_num με τις δύο convolutional layers και τις αντίστοιχες συναρτήσεις ενεργοποίησης :

Conv0 (Input -> 16)

ReLU

Conv1 (16 -> 8)

ReLU

Από την άλλη τα embeddings, τα οποία ταυτίζονται με τους αντιπροσωπευτικούς αριθμητικούς πίνακες των γλωσσικών χαρακτηριστικών των χρηστών, όπως αυτοί αρχικοποιήθηκαν, υπολογίστηκαν και τροποποιήθηκαν παραπάνω, τοποθετούνται σαν inputs σε ένα GCN_desc με δύο convolutional layers και τις αντίστοιχες συναρτήσεις ενεργοποίησης :

Conv0 (Input -> 32)

ReLU

Conv1 (32 -> 8)

ReLU

Η λογική στην επιλογή διαφορετικών GCN έγκειται καθαρά στο πλήθος των διαστάσεων των διαφορετικών χαρακτηριστικών. Συγκεκριμένα, τα (εξαρχής) αριθμητικής μορφής στοιχεία των χρηστών είναι συγκριτικά πολύ πιο λίγα σε σχέση με το μέγεθος των embeddings. Συνεπώς οι αρχικές μεταβάσεις μεταξύ των στρωμάτων οφείλει να είναι ομαλή στη δεύτερη περίπτωση, προκειμένου να μην αλλοιωθεί η πολυδιάστατη φύση των γλωσσικών χαρακτηριστικών.

Για παράδειγμα μια επιλογή απότομης σύμπτυξης του input από 768 σε 8 διαστάσεις, σίγουρα θα «εξάτμιζε» χρήσιμη πληροφορία. Από την άλλη μια επιλογή υπερβολικής διεύρυνσης των διαστάσεων ενός input, από 9 σε 64, σίγουρα θα αποτελούσε πλεονασμό και θα πρόσθετε περισσότερο υπολογιστικό φόρτο παρά ακρίβεια στο μοντέλο.

Έχοντας, λοιπόν, εξάγει τους τέσσερις “deep” πίνακες χαρακτηριστικών, μέσω των παραπάνω επιμέρους GCN, σαν επόμενη κίνηση έρχεται ο συνδυασμός τους προκειμένου να αξιοποιηθούν συνδυαστικά. Υπάρχουν αρκετοί τρόποι με τους οποίους διανύσματα μπορούν να συνδυαστούν μεταξύ τους έτσι ώστε να κανονικοποιηθούν σε ένα τελικό βασικό πίνακα, ωστόσο εμείς χρησιμοποιούμε το απλό aggregation, ή αλλιώς concatenation. Έτσι καταλήγουμε σε ένα διάνυσμα 32 διαστάσεων (4 x 8dimensions), το οποίο το εναποθέτουμε ως input στο τελικό MLP, το οποίο θα μας γεννήσει και τις τελικές

προβλέψεις του συστήματος. Το MLP που χρησιμοποιείται, διαθέτει τα παρακάτω layers, τα οποία απαρτίζονται από γραμμικούς μετασχηματισμούς και συναρτήσεις ενεργοποίησης :

Linear0 (Input -> 128)

ReLU

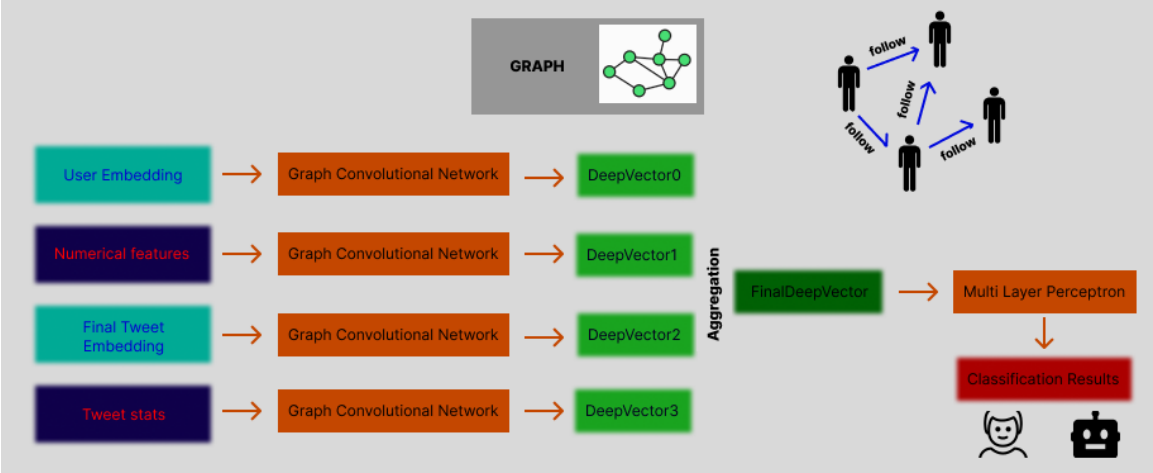
Linear1 (128 -> 64)

ReLU

Linear2 (64 -> 8)

ReLU

Linear3 (8 -> 2)



Εικόνα 41: Το pipeline του μοντέλου μας - Συνδυασμός 4ων GCN

5.4 Εκπαίδευση και Βελτιστοποίηση

Το τελικό στάδιο του συστήματος είναι η εφαρμογή του συνελικτικού αλγόριθμου στο γράφο χρηστών που έχει σχηματιστεί. Η υλοποίηση του συγκεκριμένου εγχειρήματος πραγματοποιείται με τη βοήθεια της βιβλιοθήκης :

'torch_geometric.nn'

Κατά τη χρήση της συνάρτησης GCNConv της παραπάνω βιβλιοθήκης, καθώς επίσης και κατά την εκπαίδευση του classifier σε 100 εποχές, επιστρατεύονται οι εξής απαραίτητοι παράμετροι :

- Adam optimizer with learning rate = 0.01 (**'torch.optim'**)
- weight_decay = 0.0005 (**'torch.optim'**)
- Cross Entropy Loss (**'torch.nn'**)

όπου στην παρένθεση σημειώνεται η αντίστοιχη βιβλιοθήκη.

Το μοντέλο περιέχει συνελικτικά στρώματα και υπολογιστικά επίπεδα, τα οποία περιγράφουν την πορεία και την κατεύθυνση των εισερχόμενων δεδομένων με στόχο την παραγωγή της τελικής εξόδου. Για την ακρίβεια, η προώθηση των τιμών λαμβάνει χώρα με την εξής ιεραρχία :

Convolutional Layer 0 (from initial number of attributes to 16)

ReLU Activation Function

Convolutional Layer 1 (from 16 to 4)

Linear Transformation (from 4 to 2)

Η διάταξη και οι παράμετροι των παραπάνω υπολογιστικών επιπέδων είναι αντικείμενο βελτιστοποίησης του συνολικού συστήματος, και διαφορετικοί συνδυασμοί φυσικά λαμβάνονται υπόψιν στα αποτελέσματα. Οι συναρτήσεις ReLU και Linear Transformation προέρχονται επίσης από τη βιβλιοθήκη **'torch.nn'**.

Σαν είσοδος στο συνελικτικό αλγόριθμο θέτεται ο συνολικά κατασκευασμένος και επεξεργασμένος γράφος, του οποίου η γενικότερη κλάση περιέχει τα εξής λειτουργικά πεδία :

X = οι τερματικοί πίνακες χαρακτηριστικών των χρηστών

Y = οι ετικέτες των χρηστών

Edge_index = οι συνδέσεις μεταξύ των χρηστών (ακμές του δικτύου)

Train_mask = η μάσκα που καθορίζει το ποιοι χρήστες απαρτίζουν στο train set

Val_mask = η μάσκα που καθορίζει το ποιοι χρήστες απαρτίζουν το validation set

Test_mask = η μάσκα που καθορίζει το ποιοι χρήστες απαρτίζουν το test set

Σαν επιπρόσθετα πεδία του γράφου μπορούν να οριστούν τα εξής :

Num_nodes = πλήθος χρηστών στο δίκτυο (κορυφές)

Num_bots = πλήθος bots στο δίκτυο

Num_attrs = πλήθος χαρακτηριστικών των χρηστών

Num_edges = πλήθος συνδέσεων στο δίκτυο (ακμές)

Τα λειτουργικά πεδία χρησιμοποιούνται ενεργά και οδηγούν στην παραγωγή των τελικών αποτελεσμάτων του συστήματος (classification results). Από την άλλη, τα επιπρόσθετα πεδία απλά προσδίδουν χρήσιμες ενημερωτικές πληροφορίες σχετικά με το γράφο και τις ιδιότητές του.

Σημειώνεται πως το σχήμα train – validation – test υλοποιείται σε split 70% / 10% / 20%.

5.5 Γενικές παράμετροι του συστήματος

Στην περιγραφή του συστήματος αναφέρθηκαν πολλοί παράμετροι, οι τιμές των οποίων καθορίζουν σε ένα βαθμό την αποτελεσματικότητα του συστήματος. Αυτοί οι παράμετροι ονομάζονται global parameters του συστήματος, και τα πεδία τιμών τους διαμορφώνουν ένα πλαίσιο διαφορετικών πιθανών σεναρίων, πάνω στο οποίο θα τοποθετηθούν τα στοιχεία του γραφήματος και θα εφαρμοστεί ξεχωριστά ο συνελικτικός αλγόριθμος. Συγκεκριμένα, οι global parameters του συστήματος είναι επιγραμματικά οι εξής :

1. ***Αρχικό πλήθος χρηστών***
   Ενδεχόμενα : 4000, 6000, 8000
2. ***Πυκνότητα γραφήματος χρηστών***
   Ενδεχόμενα : minimum, average, maximum
3. ***Ποσοστό ψευδών λογαριασμών (bots)***
   Ενδεχόμενα : 30%, 40%, 50%
4. ***Τελικό μέγεθος του Multi Modal Transformation***
   Ενδεχόμενα : 20, 50, 100
5. ***Είδος Multi Modal Transformation***
   Ενδεχόμενα : BLOCK, LinearSum, Tucker, MFB
6. ***Είδος γλωσσικού μοντέλου (NLP) :***
   Ενδεχόμενα : ‘all-MiniLM-L6-v2’, ‘twitter-roberta-base’

7. ***Computing Layers in the Network***

Ενδεχόμενα : Conv(N to 16) – ReLU – Conv(16 to 8) – ReLU – Conv(8 to 2)
Conv(N to 32) – ReLU – Conv(32 to 8) – ReLU – Conv(8 to 2)

Τα παραπάνω ενδεχόμενα σχηματίζουν ένα σύνολο διαφορετικών συνδυασμών, που ονομάζονται σενάρια. Είναι υπολογιστικά και χρονικά ασύμφορο να δοκιμαστούν και να εξεταστούν διεξοδικά όλα τα σενάρια. Σκοπός της ανάλυσης των σεναρίων είναι η κατανόηση της προσαρμογής και της συμπεριφοράς του μοντέλου σε αυτά. Η απόδοση του συστήματος τείνει να μεταβάλλεται με ένα λογικό αίτιο κατά τη δοκιμή και την εναλλαγή των παραπάνω σεναρίων. Αυτές τις τάσεις λοιπόν – που θα οδηγήσουν και στην βέλτιστη ακρίβεια του συστήματος – επιχειρούμε να καταγράψουμε στα πειράματα που ακολουθούν.

5.5.1 Αρχικό πλήθος χρηστών

Η συγκεκριμένη παράμετρος του συστήματος δεν καθορίζει ιδιαίτερα την απόδοση του συστήματος, αλλά ουσιαστικά θέτει το γενικότερο βάθος του και το ποσοστό αξιοπιστίας του. Με άλλα λόγια, όσο μεγαλύτερο είναι το δείγμα των χρηστών, το σύστημα αποκτά – μέσω της εκπαίδευσης – ισχυρότερη αντίληψη στο χώρο των χαρακτηριστικών των χρηστών, και κατ' επέκταση πιο ευρεία ικανότητα γενίκευσης σε νέα δεδομένα. Είναι λογικό ένα σύστημα με μικρό σύνολο δεδομένων να μην έχει την ίδια «εμπειρία» και να μην αποκτά την ίδια «γνώση». Αυτό καθιστά τα βάρη του περισσότερο «ρηχά» και την προβλεπτική του ικανότητα λιγότερο αξιόπιστη, όταν αυτό τεθεί αντιμέτωπο με νέα άγνωστα δεδομένα.

5.5.2 Πυκνότητα γραφήματος χρηστών

Μεταβάλλοντας την πυκνότητα του γράφου των χρηστών και διατηρώντας σταθερές όλες τις υπόλοιπες global parameters του συστήματος, παρατηρείται – έστω και μικρό βαθμό – πως το συνελικτικό νευρωνικό δίκτυο ανταπεξέρχεται καλύτερα σε σχετικά αρκούντως μεγάλο πλήθος ακμών. Όσο περισσότερες είναι οι ακμές ενός γραφήματος, τόσο περισσότερη πληροφορία μεταδίδεται κατά τη συνέλιξη, καθώς οι γείτονες μιας μέσης κορυφής αυξάνονται, με αποτέλεσμα να πληθαίνουν και οι επιρροές. Για παράδειγμα, στην ακραία τετριμμένη περίπτωση όπου ο γράφος εκφυλίζεται σε μια γραμμικά συνδεδεμένη λίστα, η πληροφορία περιορίζεται σε μικρές τοπικές γειτονιές, αφού η εξάπλωση της καθίσταται αργή και ασθενής μέσω δύο Convolutional Layers. Από την άλλη μεριά, η προσθήκη ολοένα και περισσότερων layers δεν αποτελεί λύση, καθώς έτσι αλλοιώνεται η έννοια της γειτνίασης των χρηστών. Στην ίδια αλλοίωση οδηγείται το μοντέλο το οποίο έχει ως είσοδο ένα γράφημα με υπερβολικά μεγάλη πυκνότητα. Η πιο ακραία περίπτωση σε αυτό το σενάριο ταυτίζεται με έναν πλήρες γράφο στον οποίο όλοι οι χρήστες συνδέονται μεταξύ τους.

Στα παραπάνω τρία πειράματα, το worst case scenario του πλήρους γράφου απέχει αρκετά, προκειμένου να αποφευχθεί. Συνεπώς η αύξηση των κορυφών του γραφήματος μόνο θετική επιρροή έχει στον συνελικτικό αλγόριθμο και γενικότερα στην απόδοση του συστήματος. Φυσικά το συμπέρασμα αυτό δεν κρίνεται απόλυτο και βέλτιστο, καθώς σημαντικό ρόλο παίζουν και οι υπόλοιποι global parameters. Σκοπός των πειραμάτων είναι η ανακάλυψη της τάσης της μεταβολής της πυκνότητας του γράφου, όπως και η θεωρητική της επεξήγηση.

5.5.3 Ποσοστό ψευδών λογαριασμών

Μεταβάλλοντας το πλήθος των bots ουσιαστικά ρυθμίζουμε την ισορροπία του dataset, όσον αφορά της δύο κλάσεις της ταξινόμησης. Όσο λιγότερα είναι τα bots, τόσο μεγαλύτερη απόδοση φαίνεται να έχει το σύστημα, γεγονός της που δεν το καθιστά απαραίτητα αποτελεσματικό και βέλτιστο.

Σε ένα classification task, προκειμένου το νευρωνικό δίκτυο να αποκωδικοποιήσει της ουσιώδεις διαφορές ανάμεσα της κλάσεις που επιχειρεί να προβλέψει, απαιτεί ως input ένα ισορροπημένο dataset. Οι μεγάλες ανισορροπίες σε ένα dataset, «διευκολύνουν» την εκπαίδευση της νευρωνικού δικτύου, το οποίο προσεγγίζει μεγάλες αποδόσεις έχοντας την τάση να προβλέπει σωστά την major κλάση, και ταυτόχρονα να μην αποκτά ποτέ την ικανότητα να ξεχωρίζει τα χαρακτηριστικά της minor κλάσης, πράγμα που είναι και το αληθινό γενικό ζητούμενο. Θα μπορούσε να πει κανείς πως η ποικιλία και το diversity των στοιχείων της dataset, ως της τα χαρακτηριστικά της, οφείλει πάντα να συνοδεύεται και με ισορροπία, ως της της ετικέτες της, καθώς μόνο με αυτό τον τρόπο το νευρωνικό δίκτυο εξαναγκάζεται να «μάθει» και να «διακρίνει» την πηγή της διαφορετικότητας των επιμέρους κλάσεων.

5.5.4 Είδος και τελικό μέγεθος του Multi Modal Transformation

Ο τρόπος με τον οποίο τα διανύσματα χαρακτηριστικών συνδυάζονται, καθώς και το τελικό μέγεθος του πίνακα, είναι παράμετροι που καθορίζουν την αρχιτεκτονική και την απόδοση του συστήματος. Κάποιοι μέθοδοι είναι περισσότερο κατάλληλοι σε δεδομένα που αντιπροσωπεύουν εικόνα, ενώ άλλα ειδικεύονται σε embeddings γλωσσικών μοντέλων. Από την άλλη, το μέγεθος του τελικού vector είναι και αυτός από μόνος του παράμετρος βελτιστοποίησης, αφού μικρά μεγέθη συμπτύσσουν πληροφορία, ενώ μεγάλα μεγέθη «απλώνουν» τα συνδυαζόμενα χαρακτηριστικά.

5.5.5 Είδος γλωσσικού μοντέλου

Η βιβλιοθήκη ‘sentence transformers’ διαθέτει ποικιλία γλωσσικών μοντέλων, τα οποία έχουν ως στόχο την αποτύπωση προτάσεων, παραγραφών και γενικά γλωσσολογικών

δεδομένων σε αριθμητικούς πίνακες πολλών διαστάσεων, οι οποίοι με τη σειρά τους μπορούν να αξιοποιηθούν σε μαθηματικά μοντέλα για υπολογισμούς και συγκρίσεις. Το κάθε NLP ωστόσο έχει εκπαιδευτεί με διαφορετική λογική σε διαφορετικά αρχικά δεδομένα, γεγονός που καθιστά την επιλογή του κάτι παραπάνω από κρίσιμη, αφού στην περίπτωση μας το σύστημα αποτελείται κατά 50% από tweets και descriptions. Επομένως, το 'all-MiniLM-L6-v2', το οποίο έχει προπονηθεί σε γενικής φύσεως λεκτικά περιεχόμενα, δεν μπορεί σε καμία περίπτωση να μας δώσει την βαθύτητα των embeddings του 'twitter-roberta-base', το οποίο έχει βασιστεί και κριθεί σε γλωσσολογικά δεδομένα που έχουν αντληθεί εξειδικευμένα από το API του Twitter. Κλείνοντας τη συγκεκριμένη global parameter, ας μην ξεχάσουμε να τονίσουμε τη σημασία του πλήθους των dimensions ενός γλωσσικού μοντέλου. Εξερευνώντας στις διάφορες εναλλακτικές της παραπάνω βιβλιοθήκης, παρατηρούνται αυξομειώσεις στο μέγεθος των embeddings που παράγονται. Το 'twitter-roberta-base' – το οποίο εν τέλει χρησιμοποιείται στο σύστημά μας – διαθέτει ένα capacity 768 αριθμητικών τιμών, το οποίο κρίνεται ταυτόχρονα και επαρκές και υπολογιστικά συμφέρον.

5.5.6 Computing Layers in the Network

Όπως και στην υλοποίηση των MLP, έτσι και στα Graph Convolutional Networks, το πλήθος και η χωρητικότητα των επιπέδων των νευρώνων του δικτύου αποτελούν μια πολύ σημαντική παράμετρο στην υπολογιστική διαδρομή των inputs και την προβλεπτική ακρίβεια των outputs. Για τα γλωσσικά μας χαρακτηριστικά, τα οποία αποτυπώνονται με τη μορφή embeddings, προτιμούμε αρχικά layers με συγκριτικά πιο πολλούς νευρώνες, σε σχέση με τα αντίστοιχα αριθμητικά χαρακτηριστικά, αφού τα πρώτα εκτείνονται σε μια περιοχή 768 διαστάσεων, ενώ τα δεύτερα είναι σε πλήθος πιο περιορισμένα. Σε κάθε περίπτωση πάντως, το συνολικό πλήθος των στρωμάτων που επιλέγεται είναι 3. Στην συγκεκριμένη απόφαση πρέπει να ληφθεί υπόψιν ότι ένα συνελικτικό δίκτυο με πολλά στρώματα αλλοιώνει εν μέρει την φύση του γραφήματος, καθώς σε μια εποχή εκπαίδευσης η πληροφορία «ταξιδεύει» κατά μήκος όλου του γράφου, μειώνοντας έτσι την αξία της γειτνίασης. Από την άλλη, η επιλογή μόνο ενός layer αποφέρει ακριβώς το ανάποδο, δηλαδή τον περιορισμό στην εξάπλωση της πληροφορίας. Έτσι, η γενική κοινότητα έχει καταλήξει στην χρήση 2-3 στρωμάτων, το οποίο βέβαια δεν αποτελεί αυστηρό κανόνα, αφού η κάθε περίπτωση οφείλεται πολλές φορές να αξιολογείται και μεμονωμένα.

Κεφάλαιο 6

Πειράματα και Αποτελέσματα

6.1 Εξισορρόπηση του training set

Το αρχικό balanced dataset των 6000 χρηστών χωρίζεται σε τρία υποσύνολα [11]:

- train_set : 60 - 70%
- validation_set : 10 - 20%
- test_set : 20%

Προκειμένου να διορθωθούν οι όποιες επιμέρους ανισορροπίες στο train_set, χρησιμοποιήθηκε η διαδικασία oversampling :

SMOTE (imblearn.over_sampling)

η οποία αξιοποιήθηκε στην γέννηση νέων τεχνητών χρηστών, με βάση τα χαρακτηριστικά των πραγματικών, καθώς και στην αντίστοιχη δημιουργία των ακμών των τεχνητών αυτών χρηστών. Η πρώτη περίπτωση δεν είναι παρά μια τετριμμένη και απλοϊκή διαδικασία, ωστόσο η δεύτερη διαθέτει μια πολυπλοκότητα και υλοποιήθηκε μέσω μιας ευριστικής στρατηγικής μετασχηματισμών.

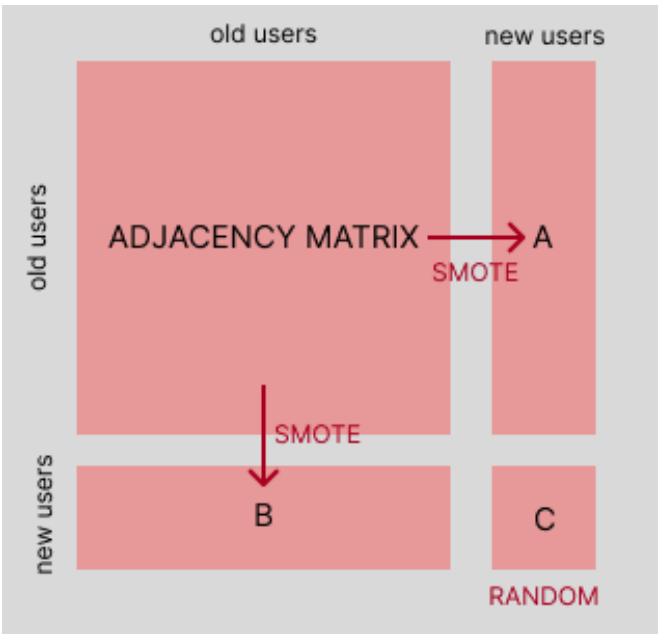
Συγκεκριμένα, ως πίνακας χαρακτηριστικών χρησιμοποιήθηκε ο adjacency matrix των συμμετέχοντων χρηστών, ο οποίος σταδιακά διευρύνθηκε και επεκτάθηκε – προκειμένου να συμπεριλάβει και τους νέους τεχνητούς χρήστες – σε τρεις φάσεις :

Α. Προσθέτοντας ακμές από τους παλιούς κόμβους στους νέους

Β. Προσθέτοντας ακμές από τους παλιούς κόμβους στους νέους

Γ. Προσθέτοντας ακμές από τους νέους κόμβους στους νέους

Στις φάσεις Α και Β εφαρμόστηκε ξεχωριστά oversampling μέσω '**SMOTE'**, ενώ η Γ φάση είχε χαρακτήρα ενός random generator, μιας και ο αρχικός πίνακας γειτνίασης δεν μπορούσε να συνεισφέρει ενεργά με χαρακτηριστικά στη συγκεκριμένη περίπτωση.



Εικόνα 42: Oversampling μέσω SMOTE

Για διευκρινιστικούς λόγους και σκοπούς πληρότητας, τονίζουμε πως το oversampling στην A φάση χρησιμοποίησε ως input τον αρχικό adjacency matrix, ενώ η B φάση τον transposed πίνακα αυτού, επιλογή λογική καθώς επιθυμούμε την ταυτόχρονη και ορθή επέκταση και προς τις δύο διαστάσεις.

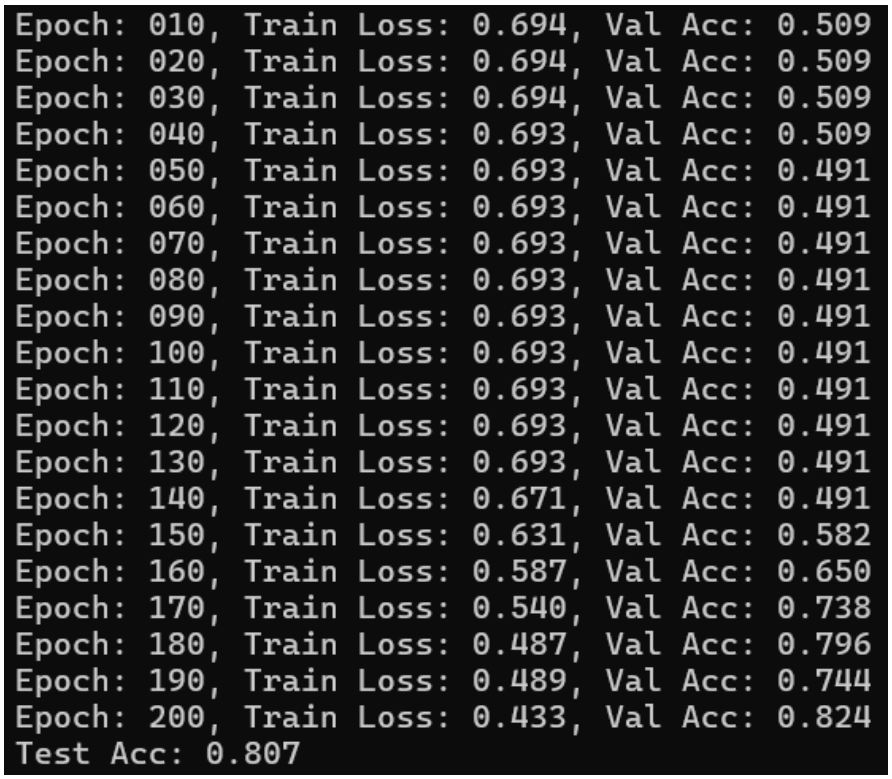
6.2 Εκτέλεση του συστήματος

Παρακάτω παρατίθενται τα αποτελέσματα – καθώς και τα ενδιάμεσα στάδια – μιας τυπικής και τυχαίας εφαρμογής του συστήματος. Ακολουθούν αποκρίσεις τερματικού, με κατάλληλα σχόλια και επεξηγήσεις.

```
C:\Users\lefte\Twitter>python main.py
Importing nodes, tweets and edges between them ..
Implementing MLP ..
Implementing GCN ..
Creating the graph ..
~~~ Graph with description attributes ~~~
Nodes            :6006
Training nodes   :3603
Bots             :3006
Training bots    :1791
Attributes       :768
Edges            :62205
~~~ Graph with numerical attributes ~~~
Nodes            :6006
Training nodes   :3603
Bots             :3006
Training bots    :1791
Attributes       :9
Edges            :62205
~~~ Graph with tweets attributes ~~~
Nodes            :6006
Training nodes   :3603
Bots             :3006
Training bots    :1791
Attributes       :768
Edges            :62205
~~~ Graph with tweets numerical attributes ~~~
Nodes            :6006
Training nodes   :3603
Bots             :3006
Training bots    :1791
Attributes       :4
Edges            :62205
```

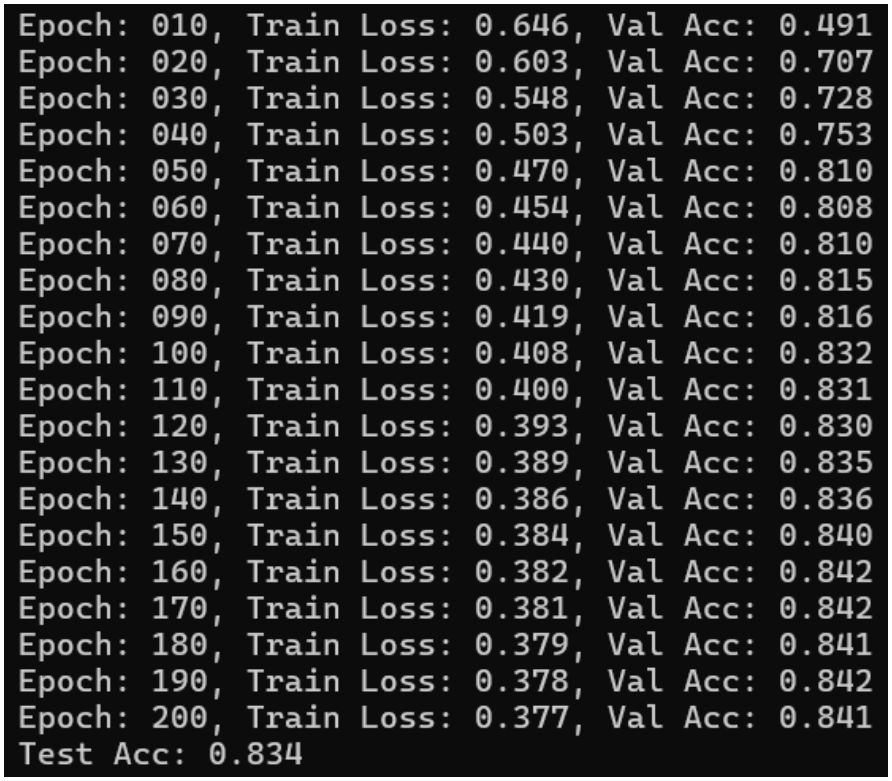
Εικόνα 43: Γενικές πληροφορίες του dataset

Εδώ βλέπουμε τους συνολικούς χρήστες, τα μεγέθη των train, validation και test sets, τα υποσύνολα των bots σε κάθε περίπτωση, καθώς επίσης και το πλήθος των attributes και των ακμών σε κάθε σενάριο.



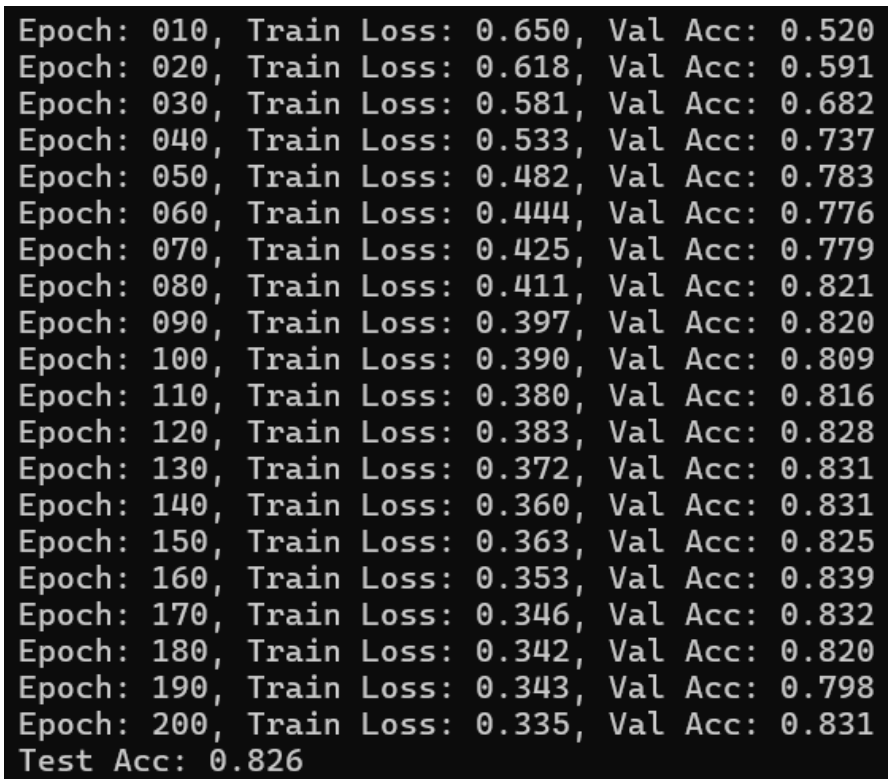
Εικόνα 44: GCN για την περιγραφή του χρήστη

Εδώ παραπάνω βλέπουμε την πορεία των αποτελεσμάτων, καθώς και της απόδοσης, του μεμονωμένου GCN που λαμβάνει ως input το description.



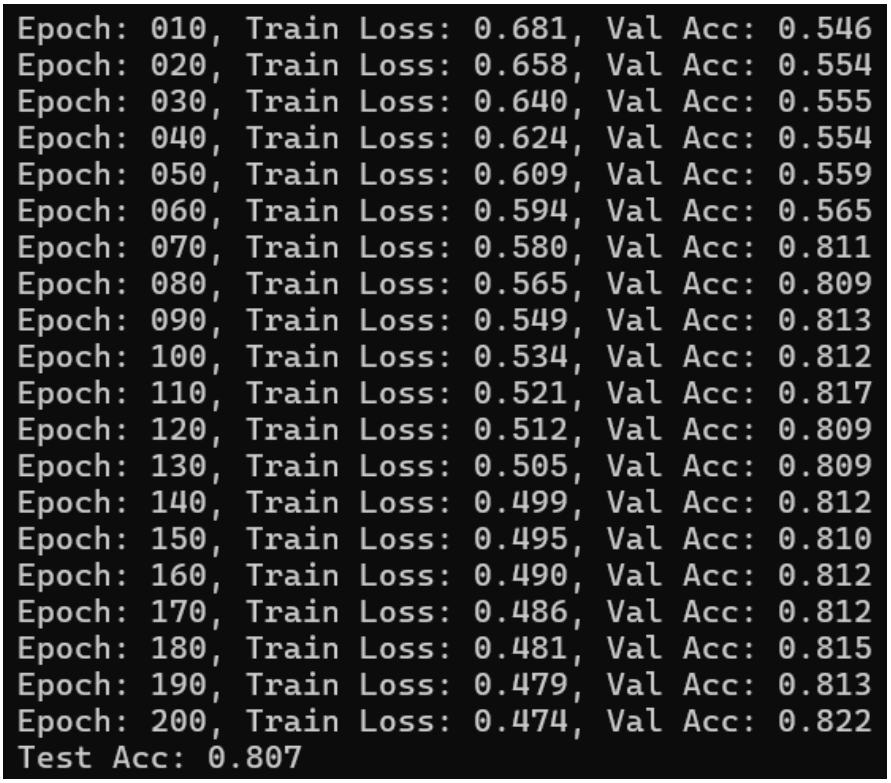
Εικόνα 45: GCN για τα αριθμητικά στοιχεία του χρήστη

Εδώ παραπάνω βλέπουμε την πορεία των αποτελεσμάτων, καθώς και της απόδοσης, του μεμονωμένου GCN που λαμβάνει ως input το numerical.



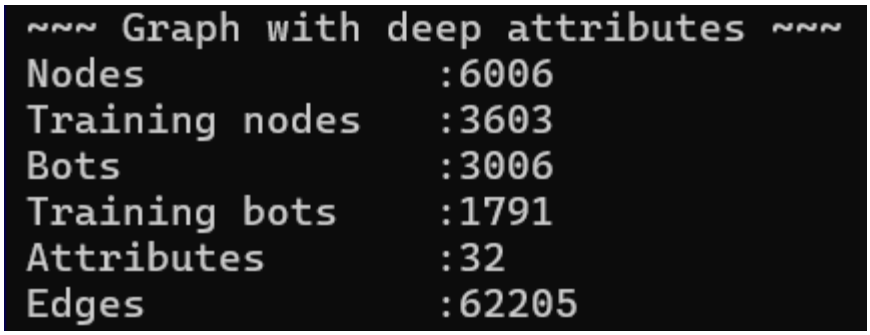
Εικόνα 46: GCN για τα tweets του χρήστη

Εδώ παραπάνω βλέπουμε την πορεία των αποτελεσμάτων, καθώς και της απόδοσης, του μεμονωμένου GCN που λαμβάνει ως input το tweets.



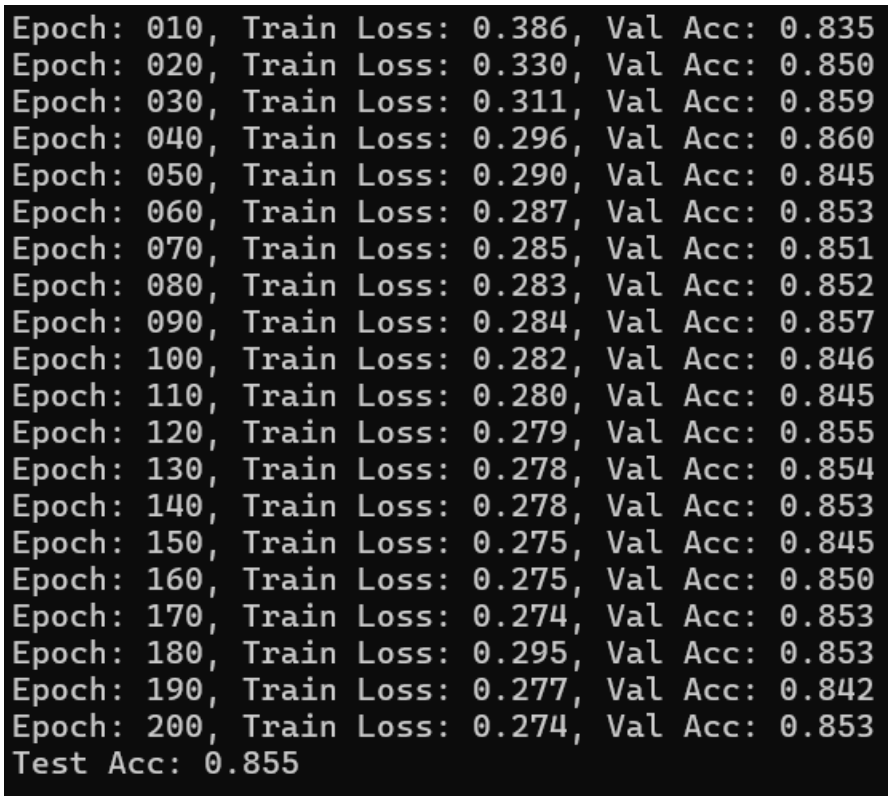
Εικόνα 47: GCN για τα αριθμητικά στοιχεία των tweets του χρήστη

Εδώ παραπάνω βλέπουμε την πορεία των αποτελεσμάτων, καθώς και της απόδοσης, του μεμονωμένου GCN που λαμβάνει ως input το tweets_numerical.



Εικόνα 48: Γενικές πληροφορίες των τελικών αναπαραστάσεων των χρηστών

Εδώ παραπάνω βλέπουμε τους συνολικούς χρήστες των train, validation, και test sets, τα υποσύνολα των bots σε κάθε περίπτωση, καθώς επίσης και το πλήθος των attributes και των ακμών σε κάθε σενάριο.



Εικόνα 49: Τελικό MLP για τα aggregated χαρακτηριστικά του χρήστη

Τέλος, εδώ παραπάνω αποτυπώνεται η πορεία των αποτελεσμάτων, καθώς και της απόδοσης, του τελικού MLP που λαμβάνει ως input τα συνδυαστικά deep outputs των τεσσάρων υποσυστημάτων.

Παρατηρείται, πως ο συνδυασμός των υποσυστημάτων συνεισφέρει φανερά στη βελτίωση της απόδοσης του συνολικού συστήματος, γεγονός που δικαιώνει την επιλογή μας να προσθέσουμε και να συνδυάσουμε στην αρχιτεκτονική μας και τις τέσσερις επιμέρους κατηγορίες.

Η εκπαίδευση, για λόγους αποφυγής overfitting και εξοικονόμησης πόρων, θα μπορούσε να τερματίζεται και στις 100 εποχές, καθώς ήδη από τότε προσεγγίζεται η μέγιστη τιμή του accuracy στο υποσύνολο του validation set.

6.3 Αποτελέσματα και Σχολιασμός

Προκειμένου να αξιολογηθούν τα αποτελέσματα του παραπάνω συστήματος μας, χρησιμοποιήθηκαν οι εξής μετρικές :

- Accuracy : σωστές προβλέψεις / συνολικές προβλέψεις
- Precision : σωστές προβλέψεις bots / συνολικές προβλέψεις bots
- Recall : σωστές προβλέψεις bots / συνολικά bots
- F1 Score : μαθηματικός συνδυασμός Precision και Recall

Ακολουθούν δύο πίνακες. Ο πρώτος αποτυπώνει τις μετρικές των τεσσάρων επιμέρους ατομικών components του συστήματος μας, όπως επίσης και του τελικού συνδυασμού τους, ενώ ο δεύτερος συγκρίνει την αρχιτεκτονική μας με τις προϋπάρχουσες διάσημες μεθόδους, οι οποίες έχουν περιγραφτεί και αναλυθεί παραπάνω.

	USER DESCRIPTION	***NUMERICAL ATTRIBUTES***	***TWEETS CONTENT***	***TWEETS STATS***	***MGCN DETECTOR***
ACCURACY	0.815	0.846	0.831	0.813	**0.851**
PRECISION	0.840	0.828	0.831	0.757	**0.865**
RECALL	0.786	0.876	0.853	**0.925**	0.835
F1SCORE	0.809	**0.851**	0.835	0.832	0.849

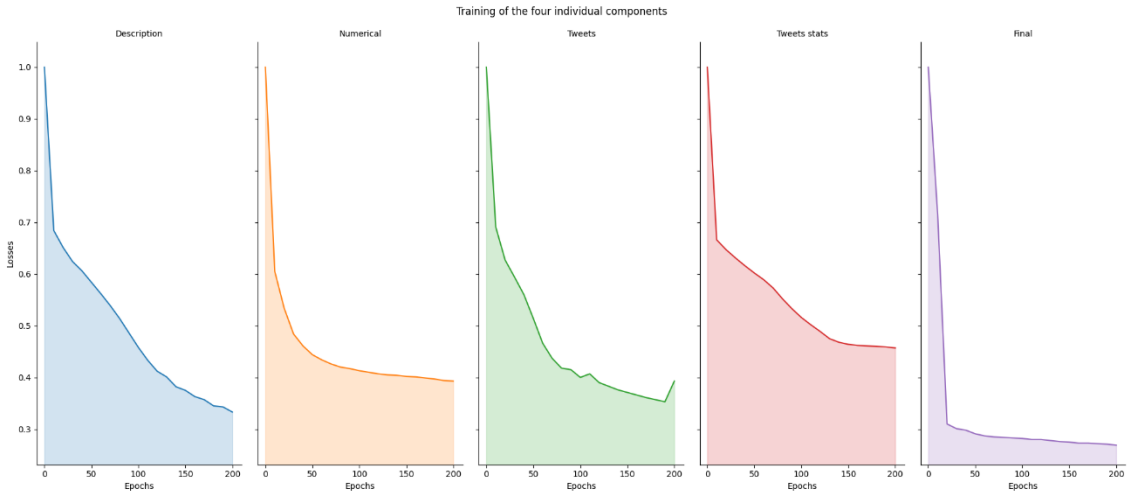
Πίνακας 2: Μετρικές και επιδόσεις των επιμέρους συστημάτων συγκριτικά με το τελικό μας μοντέλο 'MultiGCN Detector'

	ALHOSSEINI	***BOTOMETER***	***CRESCI***	***BIC***	***BOT RGCN***	***HIN***	***MGCN DETECTOR***
ACCURACY	0.681	0.558	0.479	**0.873**	0.846	0.866	0.851
PRECISION				0.847			**0.865**
RECALL				**0.934**			0.835
F1SCORE	0.731	0.489	0.107	**0.888**	0.870	0.882	0.849

Πίνακας 3: Μετρικές και επιδόσεις υπάρχοντων συστημάτων [1, 8, 7, 32, 16, 13] συγκριτικά με το δικό μας μοντέλο 'MultiGCN Detector'

Συγκριτικά με τις υπόλοιπες state of the art προσεγγίσεις, τοι σύστημα μας επιτυγχάνει πολύ υψηλές αποδόσεις, παρουσιάζοντας ένα άκρως ανταγωνιστικό performance σε ένα εξελιγμένο dataset με αυξημένο divergency στα χαρακτηριστικά και τις συνδέσεις των συμμετέχοντων χρηστών. Σημειώνουμε πως τα υπόλοιπα συστήματα έχουν εκπαιδευτεί και αξιολογηθεί στο **TwiBot-20**.

Ιδιαίτερη σημασία έχει το γεγονός πως τα επιμέρους components εξάγουν πολύ χρήσιμα και ακριβή αποτελέσματα και από μόνα τους, τονίζοντας έτσι την αξία και τη σημασία του αρχικού split των χαρακτηριστικών, η οποία αποτέλεσε βασικό θεμέλιο του συνολικού μας συστήματος. Με αυτόν τον τρόπο optimization του συστήματος μπορεί να πραγματοποιηθεί με μια πιο ορθολογική και επιμεριστική στρατηγική, βελτιώνοντας ξεχωριστά κάθε απομονωμένο component.



Εικόνα 50: Πορεία της εκπαίδευσης των υποσυστημάτων και του τελικού MLP

Φυσικά, ο τερματικός συνδυασμός των τεσσάρων components κάθε άλλο παρά περιττός μπορεί να κριθεί, αφού ναι μεν αυξάνει περισσότερο την τελική ακρίβεια των αποτελεσμάτων, αλλά προσδίδει δε και μια πιο βαθιά υπόσταση στη μέθοδο της επεξεργασίας και της αξιοποίησης των δεδομένων των χρηστών. Συγκεκριμένα, το γεγονός πως συλλέγει διαφορετικά είδη χαρακτηριστικών και τα δρομολογεί με κατάλληλο και εξειδικευμένο τρόπο, προσθέτει ένα οριζόντιο βάθος στη λογική του, το οποίο αυξάνει συγκλονιστικά την πιθανότητα πετυχημένης γενίκευσης σε καινούργια «φρέσκα» δεδομένα.

Τέλος, πολύ σημαντική επιτυχία του μοντέλου μας αποτελεί η ανάδειξη της υψηλότερης τιμής στη μετρική του Precision. Αυτό σημαίνει πως μια ανίχνευση ενός bot, στον

κοινωνικό κύκλο του Twitter, έχει συγκριτικά τη μεγαλύτερη πιθανότητα – σε σχέση με τα υπόλοιπα συστήματα – να είναι είναι πετυχημένη.

Κεφάλαιο 7

Επίλογος

7.1 Συμπέρασμα

Η ανίχνευση ψευδών λογαριασμών στο Twitter αποτελεί ένα πρόβλημα ταξινόμησης με ιδιαίτερο ενδιαφέρον, όχι μόνο λόγω της επιτακτικής ανάγκης για αξιοπιστία στις πλατφόρμες κοινωνικής δικτύωσης, αλλά επίσης για την ανταγωνιστική φιλοσοφία του, καθώς η εξέλιξη του και η επίδραση του κρίνεται αμφίδρομη.

Τα μαθηματικά μας μοντέλα επιχειρούν να ανιχνεύσουν τους πυρήνες της φύσης των bots στο Twitter, με απώτερο σκοπό να χτίσουν ένα συμπαγές σύστημα που θα περιορίσει αυστηρά τις δράσεις τους στην πλατφόρμα του Twitter (επιρροή μέσω εξάπλωσης fake news). Ταυτόχρονα όμως, από την άλλη μεριά, τα bots συνεχώς προσαρμόζονται σε αυτά τα ανιχνευτικά λογισμικά και δημιουργούν νέες και αναβαθμισμένες αποκλίνουσες συμπεριφορές προκειμένου να ξεφύγουν και να αποδράσουν.

Σαν αποτέλεσμα, το πρόβλημα διαθέτει μια αναπόφευκτη δυναμική πτυχή, λόγω της αμφίδρομης επίδρασης του ανιχνευτή και του ανιχνευόμενου κι έτσι η πολυπλοκότητα του concept αυξάνεται, όπως επίσης και η ανάγκη για συνεπή και συνεχή βελτίωση και επικαιροποίηση των συστημάτων που την εκπροσωπούν.

7.2 Μελλοντικές κινήσεις

Υπάρχουν πολλές πιθανές κινήσεις που μπορούν να ληφθούν υπόψιν στο μέλλον, προκειμένου το σύστημα να αναβαθμιστεί και να βελτιώσει την ακρίβεια του.

7.2.1 Επικαιροποίηση του dataset

Η συνεχής διατήρηση μιας επικαιροποιημένης έκδοσης του συνόλου δεδομένων μας διαδραματίζει έναν κομβικό ρόλο στην ανίχνευση των ψευδών λογαριασμών, γεγονός που είναι και το ζητούμενο. Όπως έχουμε ήδη επισημάνει, η συμπεριφορά των bots πάντα εξελίσσεται και μετασχηματίζεται, αποκτώντας αποκλίνοντα χαρακτήρα. Αυτό

καθιστά τη μορφή τους ολοένα και πιο σύνθετη, ενώ την ανίχνευσή τους σαφώς πολυπλοκότερη.

Τα bots προσαρμόζουν τη συμπεριφορά τους και την διαδικτυακή παρουσία τους στα υπάρχοντα φίλτρα και ανιχνευτικά λογισμικά, έτσι ώστε να αυξήσουν τη διαφάνεια και τη διάρκεια ζωής τους. Σαν αποτέλεσμα, είναι κάτι παραπάνω από καθοριστικό για εμάς να ανανεώνουμε τη βάση δεδομένων μας με καινούργια unseen data.

Το **TwiBot-22** είναι η τελευταία και πιο πρόσφατη έκδοση μαζικού συνόλου δεδομένων που προέρχεται από την πλατφόρμα του Twitter, ωστόσο είναι σίγουρο πως νέα updates ακολουθούν στο μέλλον.

7.2.2 Εμπλουτισμός των user features

Μέχρι στιγμής, η αρχιτεκτονική του συστήματος μας αξιοποιούσε αποκλειστικά και μόνο τα παρακάτω χαρακτηριστικά, όσον αφορά τους χρήστες του Twitter :

1. Description of the user
2. Numerical attributes of the account
3. Personal tweets of the user
4. Numerical stats of the tweets

Το σύστημα μας αποκτά, επεξεργάζεται και κατευθύνει τα τέσσερα παραπάνω είδη χαρακτηριστικών σε τέσσερα ατομικά και ξεχωριστά μονοπάτια, τα οποία ονομάζονται components και περιέχουν φυσικά ένα κατάλληλα προσαρμοσμένο Graph Neural Convolutional Network.

Μια ισχυρή και σημαντική ενότητα που λείπει από τα μοντέλα μας – και που αδιαμφησβήτητα θα ενισχύσει και θα επιταχύνει την αποδοτικότητα του συστήματος – είναι η αξιοποίηση της εικόνας προφίλ του χρήστη. Διαθέτουμε εύκολα τη δυνατότητα να συμπεριλάβουμε την εικόνα του χρήστη στο σύνολο δεδομένων μας, και έπειτα από κάποια απαραίτητα ενδιάμεσα βήματα, όπως για παράδειγμα τα pooling και τα flattening layers, να καταλήξουμε σε κάποια αριθμητικά embeddings, τα οποία μπορούμε λοιπόν να δρομολογήσουμε σε έναν ολοκαίνουργιο και ξεχωριστό πέμπτο component με ένα αντίστοιχο και παρόμοιο GCN. Με αυτόν τον τρόπο, εκμεταλλευόμαστε την ατομικότητα και την ανεξαρτησία μεταξύ των υποσυστημάτων και προσθέτουμε ένα νέο υπολογιστικό μονοπάτι, μια νέα διάσταση, που μέσω ενός πέμπτου deep vector – το οποίο συμμετέχει στο τελικό MLP – προσδίδει νέα χρήσιμη εξαγόμενη πληροφορία στο συνολικό σύστημα. Επομένως είναι εύκολα κατανοητό πως εμπλουτίζοντας τη σύσταση του συστήματος με καινούργια είδη χαρακτηριστικών, αυξάνεται και διευρύνεται η πολυτροπικότητα

(multimodality), κι έτσι υψώνονται και οι πιθανότητες για μεγαλύτερη ακρίβεια και απόδοση.

7.2.3 Εμπλοκή καινούργιων concatenating methods

Το υπάρχων σύστημα εν τέλει συνδυάζει τα τερματικά deep vectors – που προέρχονται από τα τέσσερα επιμέρους components – και σχηματίζει το input του τελικού Multi Layer Perceptron με έναν απλοϊκό και μινιμαλιστικό τρόπο. Συγκεκριμένα, τα τέσσερα εξαγόμενα vectors (μεγέθους 8) τοποθετούνται μαζί σε ένα κοινό διάνυσμα (μεγέθους 32). Προκειμένου να βελτιστοποιήσουμε την πολυτροπικότητα της αρχιτεκτονικής μας, μπορούμε να προσθέσουμε ένα Attention Layer ανάμεσα στις τέσσερις εξόδους των Graph Convolutional Networks και της εισόδου του MLP. Για να πραγματοποιηθεί κάτι τέτοιο, η συνολική υλοποίηση του συστήματος οφείλει να τροποποιηθεί και να γίνει πιο δυναμική. Με άλλα λόγια, η εκπαίδευση πρέπει να γίνει συνολική και να λαμβάνει χώρα σε όλο το pipeline, και όχι ξεχωριστά για κάθε component.

Τα αποτελέσματα των τεσσάρων GCNs θα συνδυάζονται με βάση τα επιμέρους βάρη του Attention Layer, τα οποία φυσικά θα αποτελούν εκπαιδευόμενη παράμετρο κατά την εκπαίδευση του συστήματος. Με αυτόν τον τρόπο, η συμβολή κάθε component θα είναι διαφορετική στον τελικό σχηματισμό του representative vector. Το σύστημα μας θα εκπαιδεύεται και μέσω του Backpropagation, θα ενημερώνει πλέον όχι μόνο τα layers των GCN και MLP, αλλά και τα βάρη του Attention Layer, το οποίο θα ερμηνεύει τα διαφορετικά επίπεδα σημασίας – όσον αφορά την υπολογιστική συμβολή στα αποτελέσματα – ανάμεσα στα υπάρχοντα είδη χαρακτηριστικών του χρήστη.

Η εξής υλοποίηση αποκτά προστιθέμενη και αυξημένη αξία και σημασία στην περίπτωση που το multimodality του συστήματος μας διευρυνθεί ακόμα παραπάνω στο μέλλον.

Βιβλιογραφία

1. Alhosseini, S. A., Najafi, P., Tareaf, R. Bin, & Meinel, C. (2019). Detect me if you can: Spam bot detection using inductive representation learning. *The Web Conference 2019 - Companion of the World Wide Web Conference, WWW 2019*. doi: 10.1145/3308560.3316504

2. Alothali, E., Salih, M., Hayawi, K., & Alashwal, H. (2022). Bot-MGAT: A Transfer Learning Model Based on a Multi-View Graph Attention Network to Detect Social Bots. *Applied Sciences (Switzerland)*, *12*(16). doi: 10.3390/app12168117

3. Anishnama. (2023, April 28). *Understanding LSTM: Architecture, Pros and Cons, and Implementation | by Anishnama | Medium*. Https://Medium.Com/@anishnama20/Understanding-Lstm-Architecture-Pros-and-Cons-and-Implementation-3e0cca194094. Retrieved from https://medium.com/@anishnama20/understanding-lstm-architecture-pros-and-cons-and-implementation-3e0cca194094

4. Ansh David. (2021, April 23). *Single Layer Perceptron and Activation Function | by Ansh David | CodeX | Medium*. Https://Medium.Com/Codex/Single-Layer-Perceptron-and-Activation-Function-B6b74b4aae66. Retrieved from https://medium.com/codex/single-layer-perceptron-and-activation-function-b6b74b4aae66

5. AWS. (n.d.). *What is Deep Learning?* Https://Aws.Amazon.Com/What-Is/Deep-Learning/.

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

7. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., & Tesconi, M. (2016). DNA-Inspired Online Behavioral Modeling and Its Application to Spambot Detection. *IEEE Intelligent Systems*, *31*(5). doi: 10.1109/MIS.2016.29

8. Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A System to Evaluate Social Bots. *WWW 2016 Companion - Proceedings of the 25th International Conference on World Wide Web*. doi: 10.1145/2872518.2889302

9. DeepLearning.AI. (2023, January 11). *Natural Language Processing*. Https://Www.Deeplearning.Ai/Resources/Natural-Language-Processing/.

10. Everton Gomede. (2023a). *Beyond the Surface: Unveiling the Depths of Multi-Layer Perceptrons | by Everton Gomede, PhD | Artificial Intelligence in Plain English*. Https://Ai.Plainenglish.Io/Multi-Layer-Perceptron-2ada79402956. Retrieved from https://ai.plainenglish.io/multi-layer-perceptron-2ada79402956

11. Everton Gomede. (2023b, August 28). *The Significance of Train-Validation-Test Split in Machine Learning | by Everton Gomede, PhD | Medium*. Https://Medium.Com/@evertongomede/the-Significance-of-Train-Validation-Test-Split-in-Machine-Learning-91ee9f5b98f3. Retrieved from https://medium.com/@evertongomede/the-significance-of-train-validation-test-split-in-machine-learning-91ee9f5b98f3

12. Farzad Karami. (2023, July 20). *Understanding Graph Attention Networks: A Practical Exploration*. Https://Medium.Com/@farzad.Karami/Understanding-Graph-Attention-Networks-a-Practical-Exploration-Cf033a8f3d9d.

13. Feng, S., Tan, Z., Li, R., & Luo, M. (2022). Heterogeneity-Aware Twitter Bot Detection with Relational Graph Transformers. *Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022*, *36*. doi: 10.1609/aaai.v36i4.20314

14. Feng, S., Tan, Z., Wan, H., Wang, N., Chen, Z., Zhang, B., Zheng, Q., Zhang, W., Lei, Z., Yang, S., Feng, X., Zhang, Q., Wang, H., Liu, Y., Bai, Y., Wang, H., Cai, Z., Wang, Y., Zheng, L., … Luo, M. (2022). TwiBot-22: Towards Graph-Based Twitter Bot Detection. *Advances in Neural Information Processing Systems*, *35*.

15. Feng, S., Wan, H., Wang, N., Li, J., & Luo, M. (2021). SATAR: A Self-supervised Approach to Twitter Account Representation Learning and its Application in Bot Detection. *International Conference on Information and Knowledge Management, Proceedings*. doi: 10.1145/3459637.3481949

16. Feng, S., Wan, H., Wang, N., & Luo, M. (2021). 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, ASONAM 2021. doi: 10.1145/3487351.3488336

17. GeeksforGeeks. (2023). *Bidirectional Recurrent Neural Network - GeeksforGeeks*. Https://Www.Geeksforgeeks.Org/Bidirectional-Recurrent-Neural-Network/. Retrieved from https://www.geeksforgeeks.org/bidirectional-recurrent-neural-network/

18. Guo, Q., Xie, H., Li, Y., Ma, W., & Zhang, C. (2022). Social bots detection via fusing bert and graph convolutional networks. *Symmetry*, *14*(1). doi: 10.3390/sym14010030

19. Himanshu Tripathi. (2019, September 24). *What Is Balanced And Imbalanced Dataset? | by Himanshu Tripathi | Analytics Vidhya | Medium*. Https://Medium.Com/Analytics-Vidhya/What-Is-Balance-and-Imbalance-Dataset-89e8d7f46bc5. Retrieved from https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5

20. IBM Cloud Education. (n.d.-a). *What are Convolutional Neural Networks? | IBM*. IBM Cloud Education. Retrieved from https://www.ibm.com/topics/convolutional-neural-networks

21. IBM Cloud Education. (n.d.-b). *What are Recurrent Neural Networks? | IBM*. IBM Cloud Education. Retrieved from https://www.ibm.com/topics/recurrent-neural-networks

22. IBM Cloud Education. (n.d.-c). *What is Deep Learning? | IBM*. IBM Cloud Education. Retrieved from https://www.ibm.com/topics/deep-learning

23. IBM Cloud Education. (n.d.-d). *What Is Machine Learning (ML)? | IBM*. IBM Cloud Education. Retrieved from https://www.ibm.com/topics/machine-learning

24. IBM Cloud Education. (n.d.-e). *What is NLP (natural language processing)?* IBM Cloud Education.

25. IBM Cloud Education. (2020). What is Artificial Intelligence (AI)? | IBM. In IBM Cloud Education.

26. Ilias, L., & Askounis, D. (2022). Multimodal Deep Learning Models for Detecting Dementia From Speech and Transcripts. *Frontiers in Aging Neuroscience*, *14*. doi: 10.3389/fnagi.2022.830943

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

28. Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media. *Business Horizons*, *54*(3), 241–251. doi: 10.1016/J.BUSHOR.2011.01.005

29. Kudugunta, S., & Ferrara, E. (2018). Deep neural networks for bot detection. *Information Sciences*, *467*. doi: 10.1016/j.ins.2018.08.019

30. Lakshmi Pallempati. (2023, May 2). *Backpropagation through time (BPTT) | by Lakshmi Pallempati | Medium*. Https://Medium.Com/@lakshmi.Pallempati22/Backpropagation-through-Time-Bptt-7f32037a2699. Retrieved from https://medium.com/@lakshmi.pallempati22/backpropagation-through-time-bptt-7f32037a2699

31. Lee, K., Eoff, B. D., & Caverlee, J. (2011). Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, ICWSM 2011*. doi: 10.1609/icwsm.v5i1.14106

32. 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. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*. doi: 10.18653/v1/2023.acl-long.575

33. Mayank Mishra. (2020a, August 26). *Convolutional Neural Networks, Explained | by Mayank Mishra | Towards Data Science*. Https://Towardsdatascience.Com/Convolutional-Neural-Networks-Explained-9cc5188c4939. Retrieved from https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

34. Mayank Mishra. (2020b, August 26). *Convolutional Neural Networks, Explained | by Mayank Mishra | Towards Data Science*.

Https://Towardsdatascience.Com/Convolutional-Neural-Networks-Explained-9cc5188c4939. Retrieved from https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

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

36. Obar, J. A., & Wildman, S. S. (2015). Social Media Definition and the Governance Challenge: An Introduction to the Special Issue. *SSRN Electronic Journal*. doi: 10.2139/SSRN.2647377

37. PimpernelGawkroger. (2019, January 20). *Advantages and disadvantages of social media*. Https://Medium.Com/@clinguen/Advantages-and-Disadvantages-of-Social-Media-47cd957b73d5.

38. Pragati Baheti. (2021, May 27). *Activation Functions in Neural Networks [12 Types & Use Cases]*. Https://Www.V7labs.Com/Blog/Neural-Networks-Activation-Functions. Retrieved from https://www.v7labs.com/blog/neural-networks-activation-functions

39. Prashant Sharma. (n.d.). *What are Graph Neural Networks, and how do they work?* Https://Www.Analyticsvidhya.Com/Blog/2022/03/What-Are-Graph-Neural-Networks-and-How-Do-They-Work/. Retrieved from https://www.analyticsvidhya.com/blog/2022/03/what-are-graph-neural-networks-and-how-do-they-work/

40. Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. B. (2021a). A Gentle Introduction to Graph Neural Networks. *Distill*, *6*(9), e33. doi: 10.23915/DISTILL.00033

41. Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. B. (2021b). A Gentle Introduction to Graph Neural Networks. *Distill*, *6*(9), e33. doi: 10.23915/DISTILL.00033

42. Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The graph neural network model. *IEEE Transactions on Neural Networks*, *20*(1), 61–80. doi: 10.1109/TNN.2008.2005605

43. similarweb. (n.d.-a). *Top Websites Ranking - Most Visited Websites in May 2024 | Similarweb*. Https://Www.Similarweb.Com/Top-Websites/. Retrieved from https://www.similarweb.com/top-websites/

44. similarweb. (n.d.-b). *twitter.com Traffic Analytics, Ranking & Audience [May 2024] | Similarweb*. Https://Www.Similarweb.Com/Website/Twitter.Com/#overview. Retrieved from https://www.similarweb.com/website/twitter.com/#overview

45. SKY ENGINE AI. (2023, September 14). *What is Transfer Learning?* Https://Skyengine.Ai/Se/Skyengine-Blog/128-What-Is-Transfer-Learning.

46. Sumit Saha. (2018, December 15). *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way | by Sumit Saha | Towards Data Science*. SaturnCloud. Retrieved from https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

47. Tauhid Zaman. (2022, August 1). *Is Elon Musk Right about the Bot Problem on Twitter? | Yale Insights*. YALE INSIGHTS. Retrieved from https://insights.som.yale.edu/insights/is-elon-musk-right-about-the-bot-problem-on-twitter

48. Wei, F., & Nguyen, U. T. (2019). Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. *Proceedings - 1st IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA 2019*. doi: 10.1109/TPS-ISA48467.2019.00021

49. Wu, L., Cui, P., Pei, J., & Zhao, L. (2022). Graph Neural Networks: Foundations, Frontiers, and Applications. *Graph Neural Networks: Foundations, Frontiers, and Applications*, 1–689. doi: 10.1007/978-981-16-6054-2

50. Yang, K. C., Varol, O., Hui, P. M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*. doi: 10.1609/aaai.v34i01.5460