ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Πτυχιακή εργασία

Ανίχνευση ψευδών ειδήσεων με τεχνικές Μηχανικής Μάθησης σε κοινωνικά δίκτυα

Γραμματικόπουλος Λάμπρος
2022201800038

Επιβλέπουσα:

Ραυτοπούλου Παρασκευή
ΕΔΙΠ Α'

Τρίπολη, Απρίλιος 2022

Περίληψη

Αδιαμφισβήτητα, η πανδημία του Κορωνοϊού Covid-19, αποτελεί ένα παγκόσμιο φαινόμενο, που η γενιά μας δεν έχει ξαναζήσει. Τόσο όλοι μας ανεξαιρέτως ως άτομα, όσο και οι επιχειρήσεις και οι οργανισμοί καλούνται να λάβουν πρωτόγνωρα μέτρα. Όπως για παράδειγμα, ο κατ΄οίκον περιορισμός και το κλείσιμο των εμπορικών καταστημάτων, τα οποία έχουν επιφέρει μεγάλες αλλαγές στην καθημερινότητά μας. Η πανδημία του Κορωνοϊού δείχνει να έχει ενδυναμώσει περαιτέρω τη σχέση των πολιτών με τα μέσα κοινωνικής δικτύωσης (social media). Ένας από τους λόγους, που οι χρήστες άρχισαν να παρακολουθούν πιο στενά τα social media την περίοδο της πανδημίας του Κορωνοϊού, ήταν η ανάγκη για διαρκή ενημέρωση. Η κατανάλωση ειδήσεων σε αυτές τις πλατφόρμες αυξάνεται σταθερά, αλλά το ξέσπασμα της πανδημίας έδωσε άλλη διάσταση σε αυτήν τη συνήθεια. Μαζί όμως με την απομακρυσμένη επικοινωνία και ενημέρωση των χρηστών που παρέχουν τα κοινωνικά δίκτυα, όπως και σε άλλα μέσα ενημέρωσης, ελλοχεύουν κίνδυνοι παραπληροφόρησης και ψευδών ειδήσεων. Η εργασία αυτή λοιπόν στοχεύει στην ανίχνευση αυτών των ψευδών ειδήσεων.

Στα πλαίσια της παρούσας εργασίας, χρησιμοποιήθηκαν ειδήσεις που απαρτίζονται από tweets (δημοσιεύσεις) στην αγγλική γλώσσα με θέμα τον Κορωνοϊό (Covid-19) που λήφθηκαν από το Twitter με το Web-Scraping εργαλείο TWINT [ZPL17]. Στόχος της εργασίας, αποτέλεσε η ανίχνευση ψευδών ειδήσεων από αυτά τα tweets. Για την ανίχνευση των ψευδών ειδήσεων δημιουργήθηκε ένα δυαδικό μοντέλο ταξινόμησης (binary classification model), το οποίο διαχώρισε τις ειδήσεις σε αληθείς και ψευδείς. Στο μοντέλο αυτό, εισήχθηκαν ως δεδομένα τα tweets που λήφθηκαν και έπειτα χρησιμοποιήθηκαν επτά διαφορετικοί αλγόριθμοι Μηχανικής Μάθησης (Machine Learning - ML) προκειμένου να γίνει ο διαχωρισμός των ειδήσεων αυτών. Η δωρεάν βιβλιοθήκη Μηχανικής Μάθησης Scikit-learn [Ped+11] για τη γλώσσα προγραμματισμού Python, χρησιμοποιήθηκε στην εργασία αυτή και παρείχε όλους τους παραπάνω αλγορίθμους. Τέλος, έγινε λεπτομερής σύγκριση των επιδόσεων των αλγορίθμων προκειμένου να εντοπιστεί ο καταλληλότερος αλγόριθμος για μια διαδικασία δυαδικής ταξινόμησης κειμένου σαν και αυτή.

Abstract

Undoubtedly, the Covid-19 coronavirus pandemic is a global phenomenon that our generation has never experienced before. All of us, without exception, as individuals, as well as companies and organizations are called to take unprecedented measures. For example, the home restriction and the closure of commercial stores, which have brought about great changes in our daily lives. The coronavirus pandemic seems to have further strengthened the relationship of citizens with social media. One of the reasons that users began to follow social media more closely during the coronavirus pandemic was the need for constant information. Consumption of news on these platforms is growing steadily, but the outbreak of the pandemic gave another dimension to this habit. Along with the remote communication and information of the users provided by the social networks, as in other media, misinformation and false news lurks. Therefore, this thesis aims to detect false news.

In the context of this thesis, news was captured from english coronavirus (Covid-19) tweets (posts) obtained from Twitter with the Web-Scraping TWINT tool [ZPL17]. The aim of the thesis was to detect fake news from these tweets. In order to achieve that, a binary classification model was created, which separated the news into true and false. In this model, the received tweets were entered as data and then seven different Machine Learning (ML) algorithms were used in order to separate these news into these two categories. The Scikit-learn [Ped+11] Machine Learning library, is a free library for the Python programming language and was used in this thesis to provide all of the above algorithms. Finally, the performance of the algorithms was compared in detail in order to identify the most suitable algorithm for a binary text classification process like this.

Περιεχόμενα

Πρόλογος

Στόχος της εργασίας και βήματα για την υλοποίησή της

Στα πλαίσια αυτής της εργασίας έγινε προσπάθεια αντιμετώπισης του προβλήματος της ανεξέλεγκτα αυξανόμενης δημιουργίας και διασποράς ψευδών ειδήσεων στα Διαδικτυακά Μέσα Μαζικής Ενημέρωσης (ΔΜΜΜ). Για να αντιμετωπιστεί το πρόβλημα αυτό τέθηκε ως στόχος η ανίχνευση όλων των ψευδών ειδήσεων μιας χρονικής περιόδου 6 μηνών στο ΔΜΜΜ Twitter γύρω από το θέμα του Κορονοϊού. Προκειμένου όμως να γίνει μια τέτοια ανίχνευση σε μια τόσο μεγάλη κλίμακα, έπρεπε να κατανοηθεί και να αναλυθεί το πως είναι μια ψευδή είδηση σε ένα ΔΜΜΜ και ο καλύτερος τρόπος για να συμβεί αυτό ήταν η επαφή με ειδήσεις ή αλλιώς τα tweets.

Συνεπώς, η πρώτη ενέργεια ήταν να ληφθούν από το διαδίκτυο tweets από το Twitter χρησιμοποιώντας κάποιο εργαλείο «Απόξεσης» Ιστού (Web Scraper). Το εργαλείο που χρησιμοποιήθηκε ονομάζεται TWINT [ZPL17] και μέσω αυτού χρησιμοποιώντας 10 ετικέτες με θέμα τον Κορονοϊό και αφαιρώντας οποιοδήποτε tweet δεν ήταν στην αγγλική γλώσσα, λήφθηκαν 54.243.060 tweets. Μαζί με τα tweets λήφθηκαν για το καθένα περισσότερες λεπτομέρειες, όπως το όνομα χρήστη που έγραψε το tweet, τα likes του, τα retweets του, τα replies του, την τοποθεσία του χρήστη, το αν το tweet ήταν απάντηση σε άλλο tweet κ.λπ.

Έχοντας συγκεντρώσει όλα αυτά τα δεδομένα έπρεπε με κάποιο τρόπο που να είναι όσο το δυνατόν λιγότερο υποκειμενικός να αποφασιστούν τα κριτήρια διαχώρισης μιας είδησης σε αληθής ή ψευδή. Σε αυτό το κρίσιμο, για την προσπάθεια που έγινε, σημείο υπήρχαν δύο επιλογές για τη αυτή τη διαχώριση μιας είδησης. Είτε θα αποφασιζόταν ρητά αν το tweet αποτελούσε αληθή ή ψευδή είδηση είτε θα γινόταν με βάση κάποια κατώφλια αριθμών σε ένα ή περισσότερα κριτήρια που θα αποτελούσαν οι πληροφορίες που λήφθηκαν μαζί με το κάθε tweet. Έτσι, με στόχο την αντικειμενικότητα, προκειμένου να ληφθεί μαζί με την προσωπική άποψη του συγγραφέα και η άποψη του διαδικτύου για την φύση ενός tweet επιλέχθηκε ο δεύτερος τρόπος και τα κατώφλια αριθμών εισήχθηκαν στα likes, retweets και replies του κάθε tweet.

Το επόμενο βήμα στην προσπάθεια ανίχνευσης των ψευδών ειδήσεων, από αυτές που λήφθηκαν, ήταν να βρεθεί ένας γρήγορος και αποτελεσματικός τρόπος να επεξεργαστεί και να αναλυθεί όλη αυτή η πληροφορία. Αυτό πραγματοποιήθηκε με την χρήση τεχνικών και ειδικών αλγορίθμων Μηχανικής Μάθησης, οι οποίοι αφού παραμετροποιήθηκαν, μπορούσαν να εκπαιδευτούν με ένα μικρό (σε σχέση με το ολικό) κομμάτι των δεδομένων, στο οποίο τα tweets είχαν διαχωριστεί σε κατηγορίες (αληθή/ψευδή) και με βάση αυτό να μάθουν να αποφαίνονται (ή αλλιώς προβλέπουν) για την κατηγορία όλων των υπόλοιπων tweets, ώστε να πραγματοποιηθεί η ανίχνευση των ψευδών ειδήσεων. Σημειώνεται ότι, ο διαχωρισμός αυτός των tweets σε δύο κατηγορίες ονομάζεται δυαδική ταξινόμηση κειμένου (binary text classification).

Για την χρήση όμως τεχνικών και ειδικών αλγορίθμων Μηχανικής Μάθησης έπρεπε να κατανοηθεί η έννοια Μηχανική Μάθηση, ο τρόπος λειτουργίας της και η μεθοδολογία που ακολουθείται για την υλοποίηση ενός μοντέλου της. Έπειτα από την επίτευξη των παραπάνω, παρουσιάστηκε η ανάγκη για το περιβάλλον που θα χρειαζόταν για την χρήση αυτών των τεχνικών και αλγορίθμων. Την λύση έδωσε η βιβλιοθήκη της γλώσσας Python ονόματι Scikit-learn [Ped+11] (επίσης γνωστή και ως sklearn). Χρησιμοποιώντας λοιπόν την γλώσσα

Python με το περιβάλλον JupyterLab, η βιβλιοθήκη Scikit-learn [Ped+11] παρείχε όλους τους αλγόριθμους και τις τεχνικές για Μηχανικής Μάθησης προκειμένου να υλοποιηθεί τόσο το κομμάτι της παραμετροποίησης, της εκπαίδευσης των αλγορίθμων όσο και της πρόβλεψης των κατηγοριών των tweets. Ας σημειωθεί στο σημείο αυτό ότι για την ρύθμιση των παραμέτρων των αλγορίθμων και έχοντας ως στόχο έναν σωστό αλλά γρήγορο και αποτελεσματικό τρόπο, έγινε χρήση του εργαλείου Αναζήτησης Πλέγματος (Grid Search), που επίσης παρείχε η βιβλιοθήκη Scikit-learn [Ped+11], το οποίο με είσοδο 100.000 tweets (για περισσότερα οι χρόνοι αναμονής ήταν τεράστιοι) πρότεινε τις καλύτερες παραμέτρους για τους αλγόριθμους με βάση και το είδος των δεδομένων.

Έτσι, έπειτα από όλα αυτά τα βήματα και με την χρήση αυτών των πολύ ισχυρών αλγορίθμων Μηχανικής Μάθησης πραγματοποιήθηκε η ανίχνευση ψευδών ειδήσεων για μέχρι και 5.000.000 tweets και ο στόχος της εργασίας επιτεύχθηκε. Το σύνολο των tweets ήταν μόνο 5.000.000 σε σχέση με τα 54.243.060 που λήφθηκαν, διότι παρόλο που οι αλγόριθμοι Μηχανικής Μάθησης είναι πολύ ισχυροί είναι επίσης και πολύ δαπανηροί σε υπολογιστικούς πόρους σε ένα σύστημα με τέτοιο όγκο πληροφορίας. Επειδή λοιπόν κατά την διάρκεια της υλοποίησης της εργασίας οι υπολογιστικοί πόροι ήταν περιορισμένοι έτσι ήταν περιορισμένος και ο αριθμός των tweets που μπορούσαν να αναλυθούν και να επεξεργαστούν παράλληλα από τους αλγορίθμους.

Μετά το πέρας της επίτευξης του στόχου της εργασίας, υπήρξε και η ανάγκη για βελτίωση της επίδοσης του μοντέλου. Μέχρι εδώ έχει γίνει αναφορά στην χρήση πολλών αλγορίθμων έναντι ενός αλγορίθμου Μηχανικής Μάθησης. Αυτό συνέβη διότι η ανάλυση των ψευδών ειδήσεων δεν πραγματοποιήθηκε με έναν (1) αλλά με επτά (7) στο σύνολο αλγορίθμους και αυτό όχι γιατί οι αλγόριθμοι συνεργάστηκαν, αλλά διότι έγινε προσπάθεια σύγκρισης όλων αυτών προκειμένου να βρεθεί ανάμεσά τους εκείνος που θα επέφερε τα καλύτερα αποτελέσματα.

Προκειμένου να συγκριθούν οι αλγόριθμοι, χρησιμοποιήθηκαν διάφορες μετρικές αξιολόγησης αλγορίθμων Μηχανικής Μάθησης. Αυτές ήταν η Ακρίβεια (Accuracy), η Ακρίβεια (Precision), η Ανάκληση (Recall), η Καμπύλη Ακριβείας-Ανάκλησης (Precision-Recall Curve), το F1-Σκορ (F1-Score), Πίνακας Σύγχυσης (Confusion Matrix) αλλά και οι χρόνοι εκτέλεσης των αλγορίθμων. Τέλος, με βάση αυτές τις μετρικές και τους χρόνους εκτέλεσης προτάθηκε ο καλύτερος αλγόριθμος για δυαδική ταξινόμηση κειμένου.

Οργάνωση της εργασίας

Στην εργασία αυτή χρησιμοποιούνται τεχνικές Μηχανικής Μάθησης και γι' αυτό, στο Κεφάλαιο 1 γίνεται μια εισαγωγή στην Μηχανική Μάθηση, εξηγείται ο τρόπος λειτουργίας της, οι σύγχρονες εφαρμογές της, οι δυσκολίες που υπάρχουν κατά την υιοθεσία της και η έννοια του μοντέλου Μηχανικής Μάθησης. Στη συνέχεια, στο Κεφάλαιο 2, υποδεικνύονται και επεξηγούνται οι κατηγορίες της Μηχανικής Μάθησης, τα προβλήματα που επιλύει κάθε μια από αυτές και οι κυριότεροι αλγόριθμοι μηχανικής μάθησης που υπάρχουν ανά κατηγορία. Έπειτα παρουσιάζεται στο Κεφάλαιο 3, ένας αυτοματοποιημένος τρόπος δωρεάν συλλογής μεγάλου όγκου δεδομένων από το διαδίκτυο με χρήση της τεχνικής «Απόξεσης» Ιστού (Web Scraping) η οποία και αναλύεται. Στο Κεφάλαιο 4 παρατίθεται ένας δωρεάν τρόπος εύρεσης και χρήσης αλγορίθμων Μηχανικής Μάθησης, αλλά και άλλων μεταξύ αυτών. Ενώ, στο Κεφάλαιο 5, αναδεικνύεται εξ ολοκλήρου και λεπτομερώς η διαδικασία δημιουργίας ενός μοντέλου Μηχανικής Μάθησης για δυαδική ταξινόμηση κειμένου, με στόχο την εύρεση ψευδών ειδήσεων. Έπειτα, στο Κεφάλαιο 6, παρουσιάζονται διάφορες δημοφιλείς μετρικές που χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου Μηχανικής Μάθησης και γίνεται χρήση αυτών για την αξιολόγηση του μοντέλου που υλοποιήθηκε και τελικά προτείνεται ο καλύτερος αλγόριθμος Μηχανικής Μάθησης από εκείνους που συγκρίθηκαν.

Κεφάλαιο 1

Εισαγωγή στην Μηχανική Μάθηση

Ως επιστημονικό εγχείρημα, η Μηχανική Μάθηση (ΜΜ) αναπτύχθηκε από την αναζήτηση για την Τεχνητή Νοημοσύνη (ΤΝ) [Wik21]. Ήδη από την πρώιμη περίοδο της έρευνας στον τομέα της Τεχνητής Νημοσύνης σε ακαδημαϊκό επίπεδο, το ζήτημα της κατασκευής μηχανών που θα μάθαιναν από δεδομένα απασχόλησε τους ερευνητές. Προσπάθησαν να προσεγγίσουν το πρόβλημα με διάφορες συμβολικές μεθόδους, καθώς και με τα λεγόμενα νευρωνικά δίκτυα [Wik21]. Αυτά ήταν ως επί το πλείστον perceptrons και μοντέλα, που όπως διαπιστώθηκε αργότερα ήταν επανεφευρέσεις των γενικευμένων γραμμικών μοντέλων της στατιστικής. Επίσης χρησιμοποιήθηκε η πιθανοθεωρητική λογική, ιδιαίτερα στην αυτοματοποιημένη ιατρική διάγνωση [Wik21].

Ωστόσο, μια αυξανόμενη έμφαση σε προσεγγίσεις που βασίζονται στην λογική γνώση προκάλεσε ένα ρήγμα μεταξύ Τεχνητής Νοημοσύνης (ΤΝ) και Μηχανικής Μάθησης. Καθώς τα πιθανοθεωρητικά συστήματα μαστίζονταν από θεωρητικά και πρακτικά προβλήματα απόκτησης δεδομένων και αναπαράστασής τους. Από το 1980, έμπειρα συστήματα επικράτησαν στο πεδίο της Τεχνητής Νοημοσύνης (Artificial Intelligence - AI), και ο ρόλος της στατιστικής υποχώρησε. Η εργασία σε συμβολική, ή αλλιώς βασισμένη σε γνώση, εκμάθηση συνεχίστηκε εντός της ΤΝ, οδηγώντας στον επαγωγικό λογικό προγραμματισμό, αλλά οι κατευθυντήριες γραμμές της στατιστικής ήταν τώρα έξω από το χώρο της Τεχνητής Νοημοσύνης, στην αναγνώριση προτύπων και στην ανάκτηση πληροφοριών. Η έρευνα για νευρωνικά δίκτυα εγκαταλείφθηκε από την ΤΝ και την Επιστήμη Υπολογιστών τον ίδιο περίπου καιρό. Η ίδια επίσης κατεύθυνση ακολουθήθηκε πέρα από την ΤΝ και την πληροφορική, από ερευνητές άλλων ειδικοτήτων, συμπεριλαμβανομένων των Hopfield, Rumelhart και Hinton. Η επιτυχία ήρθε στα μέσα της δεκαετίας του 1980 με την επανεφεύρεση της μεθόδου ανάστροφης μετάδοσης (backpropagation).

Η Μηχανική Μάθηση (ΜΜ), αναδιοργανώθηκε ως ένα ξεχωριστό πεδίο, που άρχισε να ακμάζει κατά τη δεκαετία του 1990. Η προσοχή μετατοπίστηκε από τις συμβολικές προσεγγίσεις που κληρονόμησε από την Τεχνητή Νοημοσύνη, που στόχο είχαν την αντιμετώπιση επιλύσιμων προβλημάτων πρακτικής φύσης, και δόθηκε έμφαση σε μεθόδους και μοντέλα της στατιστικής και της θεωρίας πιθανοτήτων [Wik21]. Επίσης επωφελήθηκε από την διαθεσιμότητα ψηφιοποιημένων πληροφοριών και της δυνατότητας να διανεμηθούν μέσω του Διαδικτύου.

1.1 Τι είναι η Μηχανική Μάθηση;

Η Μηχανική Μάθηση είναι ένας κλάδος της Τεχνητής Νοημοσύνης (AI) και της επιστήμης των υπολογιστών που εστιάζει στη χρήση δεδομένων και αλγορίθμων για τη μίμηση του τρόπου με τον οποίο μαθαίνουν οι άνθρωποι. Όπως οι άνθρωποι, έτσι και οι αλγόριθμοι Μηχανικής Μάθησης βελτιώνουν σταδιακά την απόδοσή τους καθώς αυξάνεται ο αριθμός των δεδομένων που είναι διαθέσιμα για μάθηση.

Η Μηχανική Μάθηση αναπτύχθηκε από τη μελέτη της αναγνώρισης προτύπων και της υπολογιστικής θε-

ωρίας μάθησης στην Τεχνητή Νοημοσύνη [Wik21]. Διερευνά τη μελέτη και την κατασκευή αλγορίθμων που μπορούν, μέσω της χρήσης στατιστικών μεθόδων, να μαθαίνουν από τα δεδομένα (δηλαδή να εκπαιδεύονται) και να κάνουν ταξινομήσεις ή προβλέψεις σχετικά με αυτά. Τέτοιοι αλγόριθμοι λειτουργούν κατασκευάζοντας μοντέλα από πειραματικά δεδομένα, προκειμένου να κάνουν προβλέψεις βασιζόμενες στα δεδομένα ή να εξάγουν αποφάσεις που εκφράζονται ως το αποτέλεσμα. Στη συνέχεια, το αποτέλεσμα αυτό οδηγεί στη λήψη αποφάσεων εντός των εφαρμογών και των επιχειρήσεων.

Η Μηχανική Μάθηση εφαρμόζεται σε μια σειρά από υπολογιστικές εργασίες, όπου τόσο ο σχεδιασμός όσο και ο ρητός προγραμματισμός των αλγορίθμων είναι ανέφικτος. Παραδείγματα εφαρμογών αποτελούν τα φίλτρα spam (spam filtering), η οπτική αναγνώριση χαρακτήρων (OCR), οι μηχανές αναζήτησης και η υπολογιστική όραση.

Στο πεδίο της ανάλυσης δεδομένων, η Μηχανική Μάθηση είναι μια μέθοδος που χρησιμοποιείται για την επινόηση πολύπλοκων μοντέλων και αλγορίθμων που οδηγούν στην πρόβλεψη. Τα αναλυτικά μοντέλα επιτρέπουν στους ερευνητές, τους επιστήμονες δεδομένων, τους μηχανικούς και τους αναλυτές να παράγουν αξιόπιστες αποφάσεις και αποτελέσματα και να αναδείξουν αλληλοσυσχετίσεις μέσω της μάθησης από ιστορικές σχέσεις και τάσεις στα δεδομένα.

1.2 Εφαρμογές της Μηχανικής Μάθησης στην καθημερινή ζωή

Οι εφαρμογές της Μηχανικής Μάθησης είναι άφθονες. Μερικά παραδείγματα όπου χρησιμοποιούμε ήδη το αποτέλεσμα της Μηχανικής Μάθησης είναι τα εξής:

1. Smartphone που εντοπίζουν πρόσωπα ενώ τραβούν φωτογραφίες ή ξεκλειδώνονται.

2. Facebook, LinkedIn ή οποιοσδήποτε άλλος ιστότοπος μέσων κοινωνικής δικτύωσης προτείνει τους φίλους σας και τις διαφημίσεις που μπορεί να σας ενδιαφέρουν.

3. Τράπεζες που χρησιμοποιούν τη Μηχανική Μάθηση για τον εντοπισμό συναλλαγών απάτης (Fraud Detection) σε πραγματικό χρόνο.

4. Το πολυσυζητημένο αυτοκίνητο Google αυτόματης οδήγησης.

5. Διαδικτυακές προσφορές συστάσεων όπως αυτές από το Amazon και το Netflix.

6. Να γνωρίζετε τι λένε οι πελάτες για εσάς στο Twitter.

7. Αναγνώριση ομιλίας: Είναι επίσης γνωστή ως αυτόματη αναγνώριση ομιλίας (ASR), αναγνώριση ομιλίας υπολογιστή ή ομιλία σε κείμενο και είναι μια δυνατότητα που χρησιμοποιεί την επεξεργασία φυσικής γλώσσας (NLP) για την επεξεργασία της ανθρώπινης ομιλίας σε γραπτή μορφή. Πολλές κινητές συσκευές ενσωματώνουν αναγνώριση ομιλίας στα συστήματά τους για τη διεξαγωγή φωνητικής αναζήτησης, π.χ. Siri, ή παρέχεται μεγαλύτερη προσβασιμότητα σχετικά με την αποστολή μηνυμάτων.

8. Εξυπηρέτηση πελατών: Τα διαδικτυακά chatbots αντικαθιστούν τους ανθρώπινους πράκτορες κατά την εξυπηρέτηση των πελατών. Απαντούν σε συχνές ερωτήσεις (FAQ) γύρω από θέματα, όπως η αποστολή μιας παραγγελίας, παρέχουν εξατομικευμένες συμβουλές, ή διασταυρώνουν πωλήσεις προϊόντων, αλλάζοντας τον τρόπο με τον οποίο σκεφτόμαστε την ενασχόληση των πελατών με ιστότοπους και πλατφόρμες μέσων κοινωνικής δικτύωσης. Παραδείγματα περιλαμβάνουν bots ανταλλαγής μηνυμάτων σε ιστότοπους ηλεκτρονικού εμπορίου με εικονικούς πράκτορες, εφαρμογές ανταλλαγής μηνυμάτων, όπως το Slack και

το Facebook Messenger, και εργασίες που συνήθως εκτελούνται από εικονικούς βοηθούς και βοηθούς φωνής.

9. Όραση υπολογιστή (Computer vision): Αυτή η τεχνολογία Τεχνητής Νοημοσύνης επιτρέπει στους υπολογιστές και τα συστήματα να αντλούν σημαντικές πληροφορίες από ψηφιακές εικόνες, βίντεο και άλλες οπτικές εισόδους και με βάση αυτές τις εισόδους, μπορεί να αναλάβει δράση. Αυτή η ικανότητα παροχής συστάσεων την διακρίνει από τις εργασίες αναγνώρισης εικόνων. Με την υποστήριξη των συνελικτικών νευρωνικών δικτύων, η όραση υπολογιστή έχει εφαρμογές στην προσθήκη ετικετών φωτογραφιών στα μέσα κοινωνικής δικτύωσης, στην ακτινολογική απεικόνιση στην υγειονομική περίθαλψη και στα αυτοοδηγούμενα αυτοκίνητα στην αυτοκινητοβιομηχανία.

10. Μηχανές συστάσεων: Χρησιμοποιώντας δεδομένα συμπεριφοράς προηγούμενης κατανάλωσης, οι αλγόριθμοι Τεχνητής Νοημοσύνης μπορούν να βοηθήσουν στην ανακάλυψη τάσεων δεδομένων που μπορούν να χρησιμοποιηθούν για την ανάπτυξη πιο αποτελεσματικών στρατηγικών διασταύρωσης πωλήσεων. Αυτό χρησιμοποιείται για την παροχή σχετικών συστάσεων πρόσθετων στους πελάτες κατά τη διαδικασία ολοκλήρωσης αγοράς για διαδικτυακούς λιανοπωλητές. Για παράδειγμα, η Amazon σας προτείνει προϊόντα με βάση το ιστορικό περιήγησής σας.

11. Αυτοματοποιημένες συναλλαγές μετοχών: Σχεδιασμένες για τη βελτιστοποίηση των χαρτοφυλακίων μετοχών. Οι πλατφόρμες συναλλαγών υψηλής συχνότητας που βασίζονται σε Τεχνητή Νοημοσύνη, πραγματοποιούν χιλιάδες ή και εκατομμύρια συναλλαγές την ημέρα χωρίς ανθρώπινη παρέμβαση.

1.3 Γιατί η Μηχανική Μάθηση «τραβάει» τόσο μεγάλη προσοχή τελευταία;

Η ραγδαία εξέλιξη που παρατηρείται τα τελευταία χρόνια στον τομέα της Μηχανικής Μάθησης καθοδηγείται από ορισμένους παράγοντες:

1. Ο όγκος της παραγωγής δεδομένων αυξάνεται σημαντικά με τη μείωση του κόστους των αισθητήρων.

   (α) Κάθε φορά που κάνετε κάποια ενέργεια σε οποιονδήποτε ιστότοπο, συμπεριλαμβανομένου του Facebook και του YouTube, δημιουργείτε δεδομένα για αυτές τις εταιρείες.

   (β) Όλες οι συνδεδεμένες συσκευές, συμπεριλαμβανομένων των ζωνών γυμναστικής, των έξυπνων ρολογιών και του συνδεδεμένου εξοπλισμού, παράγουν δεδομένα.

2. Το κόστος αποθήκευσης αυτών των δεδομένων έχει μειωθεί σημαντικά.

3. Το κόστος των υπολογιστών έχει μειωθεί σημαντικά.

4. Το Cloud προσφέρει υπολογιστική ισχύ για όλους.

Αυτοί οι 4 παράγοντες συνδυάζονται για να δημιουργήσουν έναν κόσμο, όπου όχι μόνο δημιουργούμε περισσότερα δεδομένα, αλλά μπορούμε να τα αποθηκεύσουμε φθηνά και να εκτελέσουμε τεράστιους υπολογισμούς σε αυτά. Αυτό δεν ήταν δυνατό πριν, παρόλο που οι τεχνικές και οι αλγόριθμοι Μηχανικής Μάθησης ήταν γνωστοί.

1.4 Προκλήσεις στην υιοθεσία της Μηχανικής Μάθησης

Ενώ η Μηχανική Μάθηση έχει σημειώσει τεράστια πρόοδο τα τελευταία χρόνια, υπάρχουν μερικές μεγάλες προκλήσεις που πρέπει ακόμη να επιλυθούν. Είναι ένας τομέας ενεργούς έρευνας και πιθανότατα θα χρειαστεί μεγάλη προσπάθεια για την επίλυση αυτών των προβλημάτων στο επόμενο διάστημα. Μερικές από τις προκλήσεις αποτελούν:

1. Η απαίτηση για τεράστια δεδομένα: Απαιτείται τεράστιος όγκος δεδομένων για την εκπαίδευση ενός μοντέλου σήμερα. Για παράδειγμα, εάν θέλετε να ταξινομήσετε Γάτες και Σκύλους με βάση εικόνες (και δεν χρησιμοποιείτε υπάρχον μοντέλο), θα χρειαστεί το μοντέλο να εκπαιδευτεί σε χιλιάδες εικόνες. Αντιθέτως, συνήθως εξηγούμε τη διαφορά μεταξύ γάτας και σκύλου σε ένα παιδί χρησιμοποιώντας 2 ή 3 φωτογραφίες.

2. Η απαίτηση υψηλού υπολογισμού: Τώρα, τα μοντέλα Μηχανικής Μάθησης και βαθιάς μάθησης απαιτούν τεράστιους υπολογισμούς για την επίτευξη απλών εργασιών (απλές κατά τους ανθρώπους). Αυτός είναι ο λόγος για τον οποίο απαιτείται η χρήση ειδικού υλικού, συμπεριλαμβανομένων των GPU (Graphics Processing Unit) και των TPU (Tensor Processing Unit).

3. Η ερμηνεία των μοντέλων είναι δύσκολη μερικές φορές: Ορισμένες τεχνικές μοντελοποίησης μπορούν να μας δώσουν υψηλή ακρίβεια, αλλά είναι δύσκολο να εξηγηθούν. Αυτό μπορεί να αφήσει τους ιδιοκτήτες επιχείρησης απογοητευμένους. Φανταστείτε ότι είστε τράπεζα, αλλά δεν μπορείτε να πείτε γιατί αρνηθήκατε ένα δάνειο για έναν πελάτη!

4. Η απαίτηση νέων και καλύτερων αλγόριθμων: Οι ερευνητές αναζητούν συνεχώς νέους και καλύτερους αλγόριθμους για την αντιμετώπιση ορισμένων από τα προβλήματα που αναφέρονται παραπάνω.

5. Η ανάγκη για περισσότερους επιστήμονες δεδομένων: Επιπλέον, καθώς ο τομέας έχει αναπτυχθεί τόσο γρήγορα, δεν υπάρχουν πολλά άτομα με τις δεξιότητες που απαιτούνται για την επίλυση της τεράστιας ποικιλίας προβλημάτων. Αυτό αναμένεται να παραμείνει έτσι για τα επόμενα χρόνια.

1.5 Τι είναι ένα μοντέλο Μηχανικής Μάθησης;

Ένα μοντέλο Μηχανικής Μάθησης είναι ένα αρχείο που έχει εκπαιδευτεί να αναγνωρίζει ορισμένους τύπους μοτίβων. Εκπαιδεύετε ένα μοντέλο σε ένα σύνολο δεδομένων, παρέχοντάς του έναν αλγόριθμο που μπορεί να χρησιμοποιήσει για να μάθει από αυτά τα δεδομένα, η διαδικασία αυτή είναι η εκπαίδευση του μοντέλου.

Αφού εκπαιδεύσετε το μοντέλο, μπορείτε να το χρησιμοποιήσετε για να συλλογιστεί σε δεδομένα που δεν έχει δει στο παρελθόν και να κάνετε προβλέψεις σχετικά με αυτά τα δεδομένα. Για παράδειγμα, ας υποθέσουμε ότι θέλετε να δημιουργήσετε μια εφαρμογή που μπορεί να αναγνωρίσει τα συναισθήματα ενός χρήστη με βάση τις εκφράσεις του προσώπου του. Μπορείτε να εκπαιδεύσετε ένα μοντέλο παρέχοντάς του εικόνες προσώπων που το καθένα φέρει ετικέτα με ένα συγκεκριμένο συναίσθημα και, στη συνέχεια, μπορείτε να χρησιμοποιήσετε αυτό το μοντέλο σε μια εφαρμογή που μπορεί να αναγνωρίσει το συναίσθημα οποιουδήποτε χρήστη.

Κεφάλαιο 2

Κατηγορίες και αλγόριθμοι Μηχανικής Μάθησης

2.1 Κατηγορίες Μηχανικής Μάθησης και προβλήματα που επιλύουν

Τα είδη Μηχανικής Μάθησης συνήθως ταξινομούνται σε τρεις μεγάλες κατηγορίες, ανάλογα με τη φύση του εκπαιδευτικού «σήματος» ή την «ανατροφοδότηση» που είναι διαθέσιμα σε ένα σύστημα εκμάθησης. Αυτές είναι η Επιτηρούμενη Μάθηση, η Μη Επιτηρούμενη Μάθηση, η Ημι-Επιτηρούμενη Μάθηση και η Ενισχυτική Μάθηση.

2.1.1 Επιτηρούμενη Μάθηση (Supervised Learning)

Επιτηρούμενη Μάθηση ή αλλιώς Επιβλεπόμενη ή Εποπτευόμενη Μάθηση ή Μάθηση με Επίβλεψη λέγεται το υπολογιστικό πρόγραμμα που δέχεται τις παραδειγματικές εισόδους καθώς και τα επιθυμητά αποτελέσματα από έναν «δάσκαλο», και ο στόχος είναι να μάθει έναν γενικό κανόνα προκειμένου να αντιστοιχίσει τις εισόδους με τα αποτελέσματα [Wak22], [Hei18] και [Bur21]. Όταν έχετε προηγούμενα δεδομένα με αποτελέσματα (ετικέτες στην ορολογία Μηχανικής Μάθησης) και θέλετε να προβλέψετε τα αποτελέσματα για το μέλλον, θα χρησιμοποιούσατε αλγόριθμους Εποπτευόμενης Μηχανικής Μάθησης. Η Επιτηρούμενη Μηχανική Μάθηση χρησιμεύει στα εξής:

1. Ταξινόμηση (Classification): Όταν θέλετε να ταξινομήσετε τα αποτελέσματα σε διαφορετικές κατηγορίες. Τα δεδομένα εισόδου χωρίζονται σε δύο ή περισσότερες κλάσεις, και η μηχανή πρέπει να κατασκευάσει ένα μοντέλο, το οποίο θα αντιστοιχίζει τα δεδομένα σε μία (binary ταξινόμηση) ή περισσότερες (multi-label ταξινόμηση) κλάσεις. Τα φίλτρα Spam είναι ένα παράδειγμα δυαδικής ταξινόμησης, όπου οι είσοδοι είναι τα emails ή άλλα μηνύματα και οι κλάσεις είναι ”spam” και ”όχι spam”.

2. Παλινδρόμηση (Regression): Είναι χρήσιμη εάν υπάρχει σχέση μεταξύ της μεταβλητής εισόδου και της μεταβλητής εξόδου και χρησιμοποιείται για την πρόβλεψη συνεχών μεταβλητών, όπως πρόβλεψη καιρού, τάσεις αγοράς κ.λπ.

3. Πρόβλεψη (Forecasting): Η πρόβλεψη είναι η διαδικασία των προβλέψεων για το μέλλον με βάση τα δεδομένα του παρελθόντος και του παρόντος (χρονικές σειρές) και χρησιμοποιείται συνήθως για την ανάλυση των τάσεων των δεδομένων.

4. Συνδυασμός (Ensembling): Συνδυασμός των προβλέψεων πολλαπλών μοντέλων Μηχανικής Μάθησης για την παραγωγή ακριβούς πρόβλεψης.

2.1.2 Μη Επιτηρούμενη Μάθηση (Unsupervised Learning)

Ή αλλιώς Μη Επιβλεπόμενη ή Μη Εποπτευόμενη Μάθηση ή Μάθηση Χωρίς Επίβλεψη. Θεωρείται όταν δεν παρέχεται κάποια εμπειρία στον αλγόριθμο μάθησης, οπότε αυτός πρέπει να βρει την δομή των δεδομένων εισόδου [Wak22], [Hei18] και [Bur21]. Η Μη Επιτηρούμενη Μάθηση μπορεί να είναι αυτοσκοπός (ανακαλύπτοντας κρυμμένα μοτίβα σε δεδομένα) ή μέσο για ένα τέλος (χαρακτηριστικό της μάθησης). Υπάρχουν φορές που δεν θέλετε να προβλέψετε ακριβώς ένα αποτέλεσμα, αλλά θέλετε να εκτελέσετε μια τμηματοποίηση ή ομαδοποίηση. Για παράδειγμα, μια τράπεζα θα ήθελε να έχει μια τμηματοποίηση των πελατών της για να κατανοήσει τη συμπεριφορά τους. Αυτό είναι ένα πρόβλημα Μη Επιτηρούμενης Μηχανικής Μάθησης, καθώς δεν προβλέπουμε αποτελέσματα εδώ. Η μη επιτηρούμενη Μηχανική Μάθηση χρησιμεύει στα εξής:

1. Πρόβλημα ομαδοποίησης (Cluster analysis ή clustering): Διαχωρισμός του συνόλου δεδομένων σε ομάδες με βάση την ομοιότητα. Είναι χρήσιμη για την τμηματοποίηση δεδομένων σε πολλές ομάδες και την εκτέλεση ανάλυσης σε κάθε σύνολο δεδομένων για την εύρεση μοτίβων.

2. Ανίχνευση Ανωμαλιών (Anomaly Detection): Προσδιορισμός ασυνήθιστων σημείων δεδομένων σε ένα σύνολο δεδομένων.

3. Εξόρυξη Συσχέτισης (Association Mining): Προσδιορισμός συνόλων στοιχείων σε ένα σύνολο δεδομένων που εμφανίζονται συχνά μαζί.

4. Μείωση Διαστάσεων (Dimensionality Reduction): Μείωση του αριθμού των μεταβλητών σε ένα σύνολο δεδομένων.

2.1.3 Ημι-Επιτηρούμενη Μάθηση (Semi-Supervised Learning)

Η προσέγγιση της Ημι-Επιτηρούμενης Μάθησης στη Μηχανική Μάθηση περιλαμβάνει έναν συνδυασμό των δύο προηγούμενων τύπων [Wak22], [Hei18] και [Bur21]. Οι επιστήμονες δεδομένων μπορεί να τροφοδοτούν έναν αλγόριθμο που φέρει ως επί το πλείστον δεδομένα εκπαίδευσης, αλλά το μοντέλο είναι ελεύθερο να εξερευνήσει τα δεδομένα μόνο του και να αναπτύξει τη δική του κατανόηση του συνόλου δεδομένων. Η Ημι-Επιτηρούμενη Μηχανική Μάθηση χρησιμεύει στα εξής:

1. Μηχανική Μετάφραση (Machine Translation): Διδασκαλία αλγορίθμων για τη μετάφραση γλώσσας που βασίζονται σε ένα μη ολοκληρωμένο λεξικό.

2. Ανίχνευση Απάτης (Fraud Detection): Εντοπισμός περιπτώσεων απάτης όταν έχετε μόνο μερικά θετικά παραδείγματα.

3. Δεδομένα Επισήμανσης (Labelling Data): Οι αλγόριθμοι που έχουν εκπαιδευτεί σε μικρά σύνολα δεδομένων μπορούν να μάθουν να εφαρμόζουν αυτόματα ετικέτες δεδομένων σε μεγαλύτερα σύνολα.

2.1.4 Ενισχυτική Μάθηση (Reinforcement Learning)

Η Ενισχυτική Μάθηση αφορά ένα πρόγραμμα που υπολογιστή αλληλεπιδρά με ένα δυναμικό περιβάλλον στο οποίο πρέπει να επιτευχθεί ένας συγκεκριμένος στόχος (όπως η οδήγηση ενός οχήματος), χωρίς κάποιος δάσκαλος να του λέει ρητά αν έχει φτάσει κοντά στο στόχο του [Wak22], [Hei18] και [Bur21]. Αυτό το μοντέλο μαθαίνει καθώς προχωρά δοκιμάζοντας και αποτυγχάνωντας. Μια σειρά επιτυχών αποτελεσμάτων χρησιμοποιείται για να αναπτυχθεί η καλύτερη αντιμετώπιση για ένα δεδομένο πρόβλημα. Όπως για παράδειγμα, να μάθει να παίζει ένα παιχνίδι εναντίον κάποιου αντιπάλου. Η Ενισχυτική Μηχανική Μάθηση χρησιμεύει στα εξής:

1. Ρομποτική (Robotics): Τα ρομπότ μπορούν να μάθουν να εκτελούν εργασίες στον φυσικό κόσμο χρησιμοποιώντας αυτήν την τεχνική.

2. Βιντεοπαιχνίδια (Video Games): Η Ενισχυτική Μάθηση έχει χρησιμοποιηθεί για να διδάξει τις μηχανές να παίζουν διάφορα βιντεοπαιχνίδια.

3. Διαχείριση Πόρων (Resource Management): Δεδομένων των πεπερασμένων πόρων και ενός καθορισμένου στόχου, η Ενισχυτική Μάθηση μπορεί να βοηθήσει τις επιχειρήσεις να σχεδιάσουν τον τρόπο κατανομής των πόρων.

2.2 Οικογένειες αλγορίθμων ανά είδος Μηχανικής Μάθησης

Παρακάτω κατατάσσονται σε κατηγορίες οι δημοφιλέστεροι αλγόριθμοι ανά είδος Μηχανικής Μάθησης και δίνεται μία σύντομη περιγραφή για τον καθένα:

2.2.1 Naive Bayes

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει το **Classification** πρόβλημα [Sun17], [Wol20] και [Bro20]. Βασίζεται στο θεώρημα του Bayes και ταξινομεί κάθε τιμή ως ανεξάρτητη από οποιαδήποτε άλλη τιμή, υποθέτοντας ότι η παρουσία ενός συγκεκριμένου χαρακτηριστικού σε μια κλάση δεν σχετίζεται με την παρουσία οποιουδήποτε άλλου χαρακτηριστικού. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί ότι είναι μήλο εάν είναι κόκκινο, στρογγυλό και περίπου 3 ίντσες σε διάμετρο. Ακόμα κι αν αυτά τα χαρακτηριστικά εξαρτώνται το ένα από το άλλο ή από την ύπαρξη των άλλων χαρακτηριστικών, ένας αφελής ταξινομητής Bayes θα θεωρούσε ότι όλες αυτές οι ιδιότητες συμβάλλουν ανεξάρτητα στην πιθανότητα ότι αυτό το φρούτο είναι μήλο. Μας επιτρέπει να προβλέψουμε μια τάξη/κατηγορία, με βάση ένα δεδομένο σύνολο χαρακτηριστικών, χρησιμοποιώντας πιθανότητα. Το θεώρημα Bayes που χρησιμοποιεί ο αλγόριθμος παρατίθεται παρακάτω:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Η πιθανότητα του A, αν το B είναι αληθές, ισούται με την πιθανότητα του B, αν το A είναι αληθές, πολλαπλασιάζει την πιθανότητα το A να είναι αληθές, διαιρούμενο με την πιθανότητα το B να είναι αληθές. Στην εργασία χρησιμοποιήθηκε η υλοποίηση Gaussian Naive Bayes, η οποία υποθέτει ότι τα χαρακτηριστικά ακολουθούν μια κανονική κατανομή και ο τύπος της φαίνεται παρακάτω:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \times \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Όπου $P(x_i|y)$ είναι η πιθανότητα η είσοδος x_i να έχει την ετικέτα (label) y.

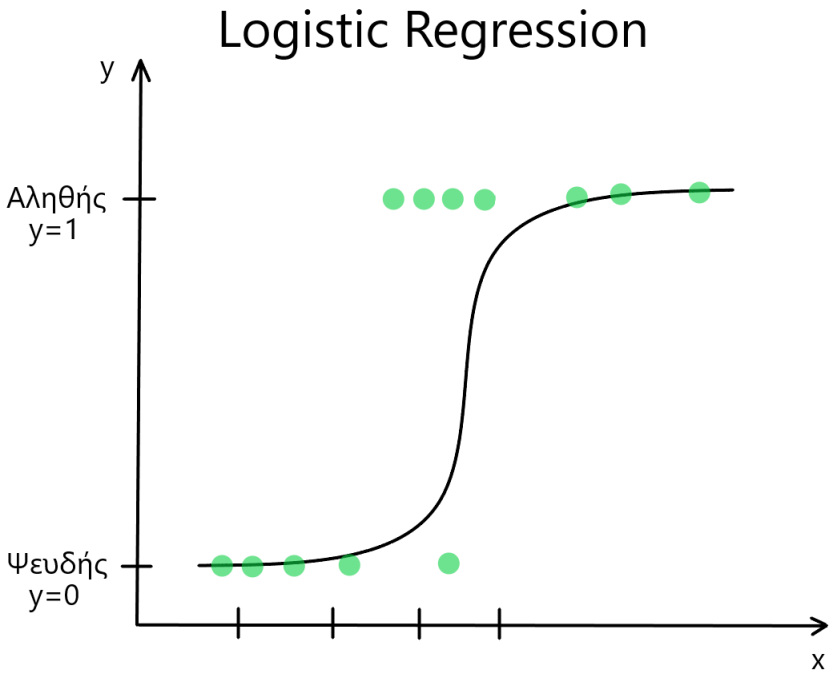
2.2.2 Linear Regression

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει το **Regression** πρόβλημα [Sun17], [Wol20] και [Bro20]. Είναι ο πιο βασικός τύπος παλινδρόμησης. Η απλή γραμμική παλινδρόμηση μας επιτρέπει να κατανοήσουμε τις σχέσεις μεταξύ δύο συνεχών μεταβλητών. Αν ο στόχος είναι η πρόβλεψη, η γραμμική

παλινδρόμηση μπορεί να χρησιμοποιηθεί για να προσαρμόσει ένα προγνωστικό μοντέλο σε ένα σύνολο παρατηρούμενων δεδομένων με τιμές απόκρισης και επεξηγηματικές μεταβλητές. Μετά από την ανάπτυξη ενός τέτοιου μοντέλου, εάν συλλεχθούν πρόσθετες τιμές των επεξηγηματικών μεταβλητών χωρίς συνοδευτική τιμή απόκρισης, το προσαρμοσμένο μοντέλο μπορεί να χρησιμοποιηθεί για να κάνει μια πρόβλεψη της απόκρισης.

2.2.3 Logistic Regression

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει το **Classification** πρόβλημα [Sun17], [Wol20] και [Bro20]. Εστιάζει στην εκτίμηση της πιθανότητας να συμβεί ένα συμβάν με βάση τα προηγούμενα δεδομένα. Χρησιμοποιείται για την πρόβλεψη ενός δυαδικού αποτελέσματος, είτε κάτι συμβαίνει είτε όχι, όπου μόνο δύο τιμές, 0 και 1, αντιπροσωπεύουν αποτελέσματα. Για παράδειγμα, Yes/No, Pass/Fail, Alive/Dead κ.λπ. Οι ανεξάρτητες μεταβλητές αναλύονται για να προσδιοριστεί το δυαδικό αποτέλεσμα με τα αποτελέσματα να εμπίπτουν σε μία από τις δύο κατηγορίες. Γράφεται ως εξής: P(Y=1|X) ή P(Y=0|X) και υπολογίζει την πιθανότητα της εξαρτημένης μεταβλητής Y, δεδομένης της ανεξάρτητης μεταβλητής X. Αυτό μπορεί να χρησιμοποιηθεί για τον υπολογισμό της πιθανότητας μιας λέξης να έχει θετική ή αρνητική σημασία (0, 1 ή σε μεταξύ τους μια κλίμακα). Μπορεί επίσης να χρησιμοποιηθεί για τον προσδιορισμό του αντικειμένου που περιέχεται σε μια φωτογραφία (δέντρο, λουλούδι, γρασίδι, κ.λπ.), με κάθε αντικείμενο να έχει μια πιθανότητα μεταξύ 0 και 1. Στο Σχήμα 2.1 φαίνεται η συνάρτηση διαχωρισμού $f(x) = \frac{1}{1+e^{-x}}$ των ετικετών των δεδομένων.



Σχήμα 2.1: Logistic Regression: Διαχωρισμός δεδομένων βάση ετικετών

2.2.4 K-Nearest Neighbors (kNN)

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει το **Classification** πρόβλημα [Sun17], [Wol20] και [Bro20]. Ο αλγόριθμος των κ-Πλησιέστερων Γειτόνων λειτουργεί ως εξής: έχοντας ένα σύνολο ομάδων, εκτιμά πόσο πιθανό είναι ένα σημείο δεδομένων να είναι μέλος κάποιας από τις ομάδες. Ουσιαστικά εξετάζει τα γειτονικά σημεία δεδομένων γύρω από ένα σημείο για να προσδιορίσει σε ποια ομάδα ανήκει αυτό.

Ταξινομεί τα νέα σημεία με την πλειοψηφία των κ γειτόνων τους. Κάθε σημείο εκχωρείται σε μία ομάδα που είναι πιο κοινή μεταξύ των κ πλησιέστερων γειτόνων του που μετρούνται με μια συνάρτηση απόστασης. Αυτή η συνάρτηση απόστασης μπορεί να είναι Ευκλείδεια, Μανχάταν, Minkowski και Hamming απόσταση. Οι πρώτες τρεις συναρτήσεις χρησιμοποιούνται για συνεχή συνάρτηση και η τέταρτη (Hamming) για κατηγορικές μεταβλητές (μεταβλητές που μπορούν να λάβουν μία τιμή από έναν περιορισμένο και συνήθως σταθερό αριθμό πιθανών τιμών). Αν κ = 1, τότε η περίπτωση απλώς εκχωρείται στην κλάση του πλησιέστερου γείτονά της. Συνήθως, η επιλογή του κ αποδεικνύεται πρόκληση κατά την εκτέλεση μοντελοποίησης kNN. Στο Σχήμα 2.2 φαίνεται αναλυτικά πως λειτουργεί ο αλγόριθμος καθώς αυξάνεται το κ.

Σχήμα 2.2: Παραδείγματα εισχώρησης αντικειμένων σε μία ομάδα βάση των κ-Πλησιέστερων Γειτόνων

2.2.5 Decision Trees

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει τα **Classification** και **Regression** προβλήματα [Sun17], [Wol20] και [Bro20]. Είναι μια δομή δέντρου που μοιάζει με διάγραμμα ροής που χρησιμοποιεί μια μέθοδο διακλάδωσης για να απεικονίσει κάθε πιθανό αποτέλεσμα μιας απόφασης. Κάθε κόμβος μέσα στο Δέντρο Απόφασης αντιπροσωπεύει μια δοκιμή σε μια συγκεκριμένη μεταβλητή και κάθε κλάδος είναι το αποτέλεσμα αυτής της δοκιμής. Χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης και λειτουργεί τόσο για κατηγορικές όσο και για συνεχείς εξαρτημένες μεταβλητές. Σε αυτόν τον αλγόριθμο, χωρίζουμε τα δεδομένα σε δύο ή περισσότερα ομοιογενή σύνολα. Αυτό γίνεται με βάση τα πιο σημαντικά χαρακτηριστικά/ανεξάρτητες μεταβλητές για να γίνουν όσο το δυνατόν πιο διακριτές ομάδες. Για τον διαχωρισμό σε διαφορετικές ετερογενείς ομάδες, χρησιμοποιούνται διάφορες τεχνικές όπως Gini, Information Gain, Chi-square, εντροπία. Εάν υπάρχουν M μεταβλητές εισόδου, καθορίζεται ένας αριθμός m«M έτσι ώστε σε κάθε κόμβο, m μεταβλητές επιλέγονται τυχαία από το M και ο καλύτερος διαχωρισμός σε αυτές τις m χρησιμοποιείται για τον διαχωρισμό του κόμβου. Η τιμή του m διατηρείται σταθερή κατά την ανάπτυξη του δάσους. Κάθε δέντρο αυξάνεται στον μεγαλύτερο βαθμό. Στο Σχήμα 2.3 φαίνεται η μορφή ενός Δένδρου Απόφασης (Decision Tree).

2.2.6 Random Forest

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει τα **Classification** και **Regression** προβλήματα [Sun17], [Wol20] και [Bro20]. Το τυχαίο δάσος (Random Forest), είναι μια μέθοδος εκμάθησης συνόλου, που συνδυάζει πολλαπλά Δένδρα Απόφασης (Desicion Trees) για να παράγει καλύτερα αποτελέσματα για ταξινόμηση, παλινδρόμηση και άλλες εργασίες. Κάθε μεμονωμένο Δένδρο Απόφασης είναι αδύναμο, αλλά

όταν συνδυάζεται και με άλλα, μπορεί να παράγει εξαιρετικά αποτελέσματα. Ο αλγόριθμος ξεκινά με ένα Δένδρο Απόφασης (ένα δέντρο τύπου γράφημα ή μοντέλο αποφάσεων) και μια είσοδος εισάγεται στην κορυφή του δέντρου. Στη συνέχεια ταξιδεύει προς τα κάτω στο δέντρο, με τα δεδομένα να τμηματοποιούνται σε όλο και μικρότερα σύνολα, με βάση συγκεκριμένες μεταβλητές.

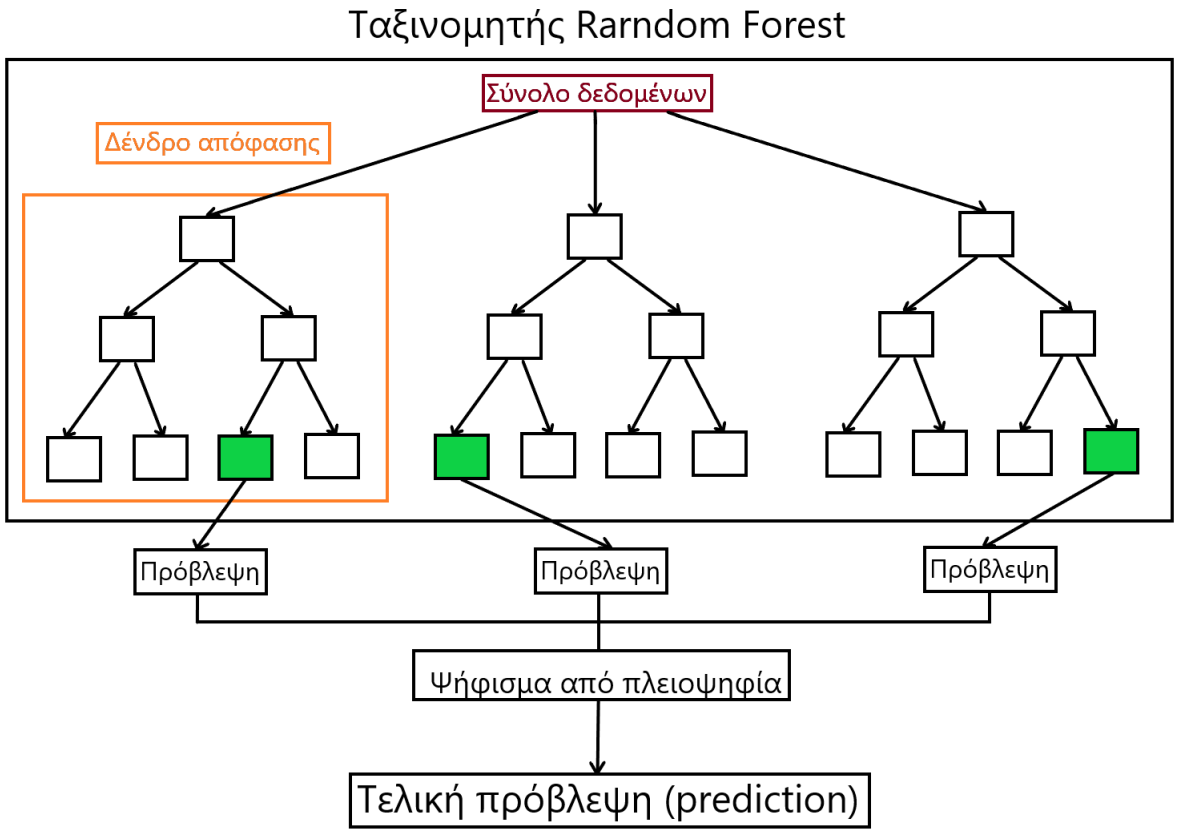
Πιο αναλυτικά, σε αυτόν τον αλγόριθμο έχουμε συλλογή από δέντρα αποφάσεων (γνωστά ως ”Δάσος”). Για να ταξινομήσουμε ένα νέο αντικείμενο με βάση τα χαρακτηριστικά, κάθε δέντρο δίνει μια ταξινόμηση και λέμε ότι το δέντρο «ψηφίζει» για αυτήν την κλάση. Το δάσος επιλέγει την ταξινόμηση με τις περισσότερες ψήφους (από όλα τα δέντρα του δάσους). Κάθε δέντρο δημιουργείται και αναπτύσσεται ως εξής:

1. Εάν ο αριθμός των περιπτώσεων στο σετ εκπαίδευσης είναι N, τότε το δείγμα N περιπτώσεων λαμβάνεται τυχαία αλλά με αντικατάσταση. Αυτό το δείγμα θα είναι το σετ εκπαίδευσης για την ανάπτυξη του δέντρου.

2. Εάν υπάρχουν M μεταβλητές εισόδου, καθορίζεται ένας αριθμός $m \ll M$ έτσι ώστε σε κάθε κόμβο, m μεταβλητές επιλέγονται τυχαία από το M και ο καλύτερος διαχωρισμός σε αυτές τις m χρησιμοποιείται για τον διαχωρισμό του κόμβου. Η τιμή του m διατηρείται σταθερή κατά την ανάπτυξη του δάσους.

3. Τέλος, κάθε δέντρο αναπτύσσεται στο μεγαλύτερο δυνατό βαθμό.

Στο Σχήμα 2.3 φαίνεται αναλυτικά πώς λειτουργεί ο αλγόριθμος.



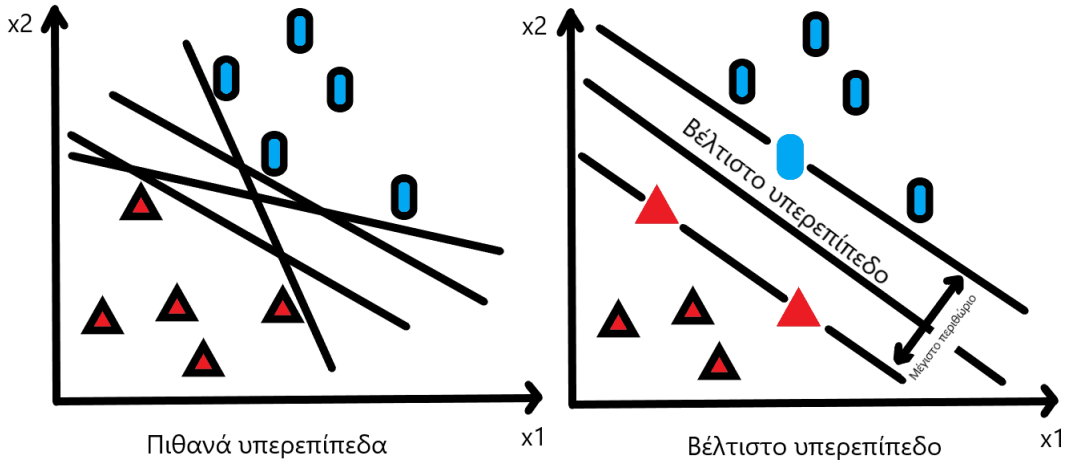
Σχήμα 2.3: Παράδειγμα ενός δάσους με πολλά δένδρα απόφασης

2.2.7 Gradient Boosting Machines (GBM)

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει το **Ensembling** πρόβλημα [Sun17], [Wol20] και [Bro20]. Ο GBM είναι ένας αλγόριθμος ενίσχυσης που χρησιμοποιείται όταν ασχολούμαστε με πολλά δεδομένα για να κάνουμε μια πρόβλεψη με υψηλή ακρίβεια. Το Boosting είναι στην πραγματικότητα ένα σύνολο αλγορίθμων εκμάθησης που συνδυάζει την πρόβλεψη πολλών βασικών εκτιμητών (estimators) προκειμένου να βελτιωθεί η ευρωστία σε έναν μόνο εκτιμητή. Συνδυάζει πολλούς αδύναμους ή μέσους προγνωστικούς παράγοντες δημιουργώντας έναν ισχυρότερο προγνωστικό παράγοντα δόμησης.

2.2.8 Support Vector Machine (SVM)

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και επιλύει το **Classification** πρόβλημα [Sun17], [Wol20] και [Bro20]. Οι αλγόριθμοι αυτοί είναι εποπτευόμενα μοντέλα εκμάθησης που αναλύουν δεδομένα που χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση. Ουσιαστικά φιλτράρουν τα δεδομένα σε κατηγορίες, κάτι που επιτυγχάνεται παρέχοντας ένα σύνολο παραδειγμάτων εκπαίδευσης, το κάθε σύνολο επισημαίνεται ότι ανήκει σε κάποια από τις κατηγορίες. Στη συνέχεια, ο αλγόριθμος εκχωρεί ένα υπερεπίπεδο (hyperplane) που διαχωρίζει καλύτερα τις ετικέτες. Σε δύο διαστάσεις αυτό είναι απλά μια γραμμή. Οτιδήποτε στη μία πλευρά της γραμμής είναι μια ομάδα και οτιδήποτε από την άλλη πλευρά είναι μια δεύτερη ομάδα. Στην δική μας περίπτωση, για παράδειγμα, αυτό θα ήταν αληθής είδηση και ψευδής είδηση. Έπειτα, ο αλγόριθμος λειτουργεί για να δημιουργήσει ένα μοντέλο που εκχωρεί νέες τιμές στη μία ή στην άλλη ομάδα. Προκειμένου να μεγιστοποιηθεί η Μηχανική Μάθηση, το καλύτερο υπερεπίπεδο είναι αυτό με τη μεγαλύτερη απόσταση μεταξύ κάθε ετικέτας. Στο Σχήμα 2.4 φαίνεται το βέλτιστο υπερπίπεδο που κρίνει ο αλγόριθμος. Ωστόσο, καθώς τα σύνολα δεδομένων γίνονται πιο πολύπλοκα, μπορεί να μην είναι δυνατό να σχεδιαστεί μια ενιαία γραμμή για την ταξινόμηση των δεδομένων σε δύο ομάδες. Χρησιμοποιώντας το SVM, όσο πιο σύνθετα είναι τα δεδομένα, τόσο πιο ακριβής θα γίνει ο προγνωστικός παράγοντας. Τέλος, το SVM επιτρέπει πιο ακριβή Μηχανική Μάθηση επειδή μπορεί να λειτουργήσει σε πολλές διαστάσεις.



Σχήμα 2.4: Βέλτιστο υπερεπίπεδο

2.2.9 K-Means

Η οικογένεια αυτή ανήκει στην **Μη επιτηρούμενη Μάθηση** και επιλύει το **Clustering** πρόβλημα [Sun17], [Wol20] και [Bro20]. Ο αλγόριθμος αυτός είναι ένας τύπος μάθησης χωρίς επίβλεψη, ο οποίος χρησιμοποιείται για την κατηγοριοποίηση δεδομένων χωρίς ετικέτα (label), δηλαδή δεδομένων χωρίς καθορισμένες κατηγορίες

ή ομάδες. Ο αλγόριθμος λειτουργεί βρίσκοντας ομάδες μέσα στα δεδομένα, με τον αριθμό των ομάδων που αντιπροσωπεύεται από τη μεταβλητή K. Στη συνέχεια λειτουργεί επαναληπτικά για να εκχωρήσει κάθε σημείο δεδομένων σε μία από τις ομάδες K με βάση τα χαρακτηριστικά που παρέχονται.

2.2.10 Hierarchical Clustering

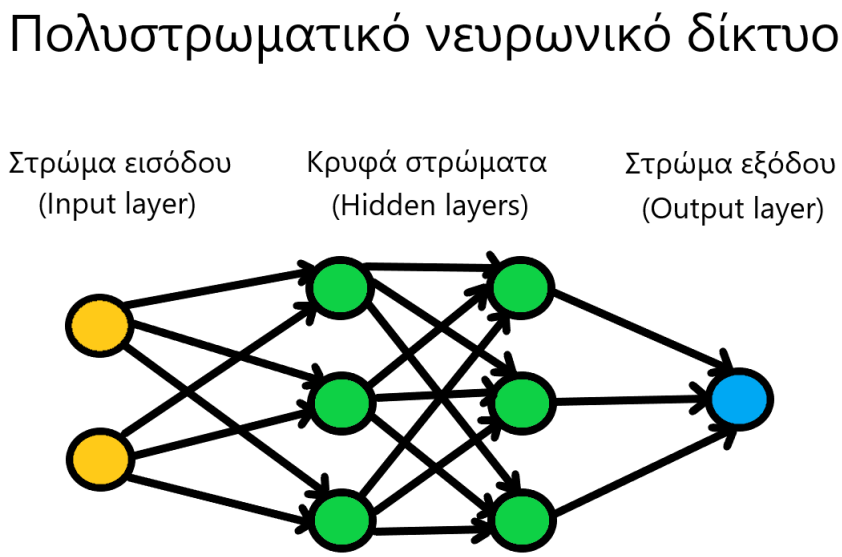
Η οικογένεια αυτή ανήκει στην **Μη επιτηρούμενη Μάθηση** και επιλύει το **Clustering** πρόβλημα [Sun17], [Wol20] και [Bro20]. Είναι μια μέθοδος ομαδοποίησης δεδομένων που επιδιώκει να οικοδομήσει μια ιεραρχία συστάδων (clusters). Οι στρατηγικές για την ιεραρχική ομαδοποίηση χωρίζονται γενικά σε δύο τύπους:

1. Συγκεντρωτική (Agglomerative): Αυτή είναι μια προσέγγιση ”από κάτω προς τα πάνω” (bottom-up), όπου κάθε παρατήρηση ξεκινά στο δικό της σύμπλεγμα και ζεύγη συστάδων συγχωνεύονται καθώς κάποιος κινείται προς τα πάνω στην ιεραρχία.

2. Διαιρετική (Divisive): Αυτή είναι μια προσέγγιση ”από πάνω προς τα κάτω” (top-down), όπου όλες οι παρατηρήσεις ξεκινούν σε ένα σύμπλεγμα και οι διαχωρισμοί εκτελούνται αναδρομικά καθώς κάποιος κινείται προς τα κάτω στην ιεραρχία. Γενικά, οι συγχωνεύσεις και οι διαχωρισμοί καθορίζονται με άπληστο τρόπο. Τα αποτελέσματα της ιεραρχικής ομαδοποίησης παρουσιάζονται συνήθως σε ένα δενδρόγραμμα.

2.2.11 Artificial Neural Networks (ANN)

Η οικογένεια αυτή ανήκει στην **Επιτηρούμενη Μάθηση** και στην **Ενισχυτική Μάθηση** και επιλύει τα **Classification**, **Regression**, **Clustering** και **Forecasting** προβλήματα [Sun17], [Wol20] και [Bro20]. Περιλαμβάνει «μονάδες» διατεταγμένες σε σειρές στρωμάτων, κάθε μία από τις οποίες συνδέεται με στρώματα εκατέρωθεν. Τα ANN εμπνέονται από βιολογικά συστήματα, όπως ο εγκέφαλος, και από τον τρόπο επεξεργασίας των πληροφοριών. Χρησιμοποιούνται κυρίως για αλγόριθμους Βαθιάς Μάθησης (Deep Learning) και επεξεργάζονται δεδομένα εκπαίδευσης (training data) μιμούμενοι τη διασυνδεσιμότητα του ανθρώπινου εγκεφάλου μέσω στρωμάτων κόμβων. Κάθε κόμβος αποτελείται από εισόδους, βάρη, μεροληψία (ή κατώφλι) και έξοδο. Εάν αυτή η τιμή εξόδου υπερβεί ένα δεδομένο όριο, «πυροδοτεί» ή ενεργοποιεί τον κόμβο, περνώντας δεδομένα στο επόμενο επίπεδο του δικτύου. Τα ANN είναι ουσιαστικά ένας μεγάλος αριθμός διασυνδεδεμένων στοιχείων επεξεργασίας, που λειτουργούν από κοινού για την επίλυση συγκεκριμένων προβλημάτων. Τα ANN είναι εξαιρετικά χρήσιμα για τη μοντελοποίηση μη γραμμικών σχέσεων σε πολυδιάστατα δεδομένα ή όπου η σχέση μεταξύ των μεταβλητών εισόδου είναι δύσκολο να κατανοηθεί. Υπάρχουν πολλοί αλγόριθμοι αυτής της οικογένειας, ένας από αυτούς που θα μας απασχολήσει είναι ο αλγόριθμος Πολυστρωματικού Νευρωνικού Δικτύου Perceptron (Multilayer Perceptron Neural Network) [Fuc21] και [Gil22]. Αυτός ο αλγόριθμος χρησιμοποιεί περισσότερα από ένα κρυφά στρώματα νευρώνων, σε αντίθεση με το perceptron ενός στρώματος και είναι γνωστός ως νευρωνικό δίκτυο βαθιάς τροφοδότησης (Deep Feedforward). Στο Σχήμα 2.5 φαίνεται ένα Πολυστρωματικό Νευρωνικό δίκτυο με δύο κρυφά στρώματα νευρώνων.



Σχήμα 2.5: Πολυστρωματικό Νευρωνικό δίκτυο

2.3 Οικογένειες αλγορίθμων που θα μας απασχολήσουν

Η ανίχνευση ψευδών ειδήσεων σε κοινωνικά δίκτυα που είναι το θέμα που θα μας απασχολήσει, αποτελεί σύμφωνα με τα παραπάνω ένα πρόβλημα ταξινόμησης (Classification), διότι θα πρέπει να ταξινομήσουμε ουσιαστικά τα δεδομένα, δηλαδή τις ειδήσεις, σε δύο κατηγορίες: «αληθή» και «ψευδή». Για τον λόγο αυτό λοιπόν, θα χρησιμοποιήσουμε οικογένειες αλγορίθμων της κατηγορίας Supervised Learning που επιλύουν το Classification πρόβλημα.

Κεφάλαιο 3

Συλλογή μεγάλου όγκου δεδομένων

Για να ξεκινήσουμε να ανιχνεύουμε ψευδείς ειδήσεις, με τους αλγόριθμους που αναφέρθηκαν στο προηγούμενο Κεφάλαιο, χρειάζεται αρχικά να συλλέξουμε έναν μεγάλο όγκο ειδήσεων, αληθών και ψευδών. Οι ειδήσεις, δηλαδή τα δεδομένα, χρειάζεται να είναι εκτεταμένου πλήθους διότι θα εισαχθούν στους αλγόριθμους και θα τους εκπαιδεύσουν ώστε να επιτευχθεί η ανίχνευση ψευδών ειδήσεων. Όσο μεγαλύτερος είναι ο όγκος των δεδομένων τόσο υψηλότερη ακρίβεια ανίχνευσης μπορεί να επιτευχθεί. Περισσότερα γι' αυτό θα βρείτε στο Κεφάλαιο 6.

3.1 Πόσα δεδομένα απαιτούνται για την εκπαίδευση ενός μοντέλου Μηχανικής Μάθησης;

Δεν υπάρχει απλή απάντηση σε αυτή την ερώτηση. Εξαρτάται από το πρόβλημα που προσπαθείτε να λύσετε, το κόστος συλλογής αυτών των δεδομένων και τα οφέλη που προέρχονται από τα επιπλέον δεδομένα. Αλλά εδώ είναι μερικές κατευθυντήριες οδηγίες:

1. Γενικά, θα θέλατε να συλλέξετε όσο το δυνατόν περισσότερα δεδομένα. Εάν το κόστος συλλογής των δεδομένων δεν είναι πολύ υψηλό, αυτό καταλήγει να λειτουργεί καλά.
2. Εάν το κόστος λήψης των δεδομένων είναι υψηλό, τότε θα πρέπει να κάνετε μια ανάλυση κόστους-οφέλους με βάση τα αναμενόμενα οφέλη που προέρχονται από μοντέλα Μηχανικής Μάθησης.
3. Τα δεδομένα που συλλέγονται πρέπει να είναι αντιπροσωπευτικά της αναμενόμενης συμπεριφοράς του μοντέλου και του περιβάλλοντος στο οποίο αναμένεται να λειτουργήσει.

3.2 Web Scraping

Προκειμένου να συλλεχθεί αυτός ο εκτεταμένος όγκος δεδομένων που απαιτούνται, για μια εργασία σαν και αυτή, έπρεπε να βρεθεί ένας αυτόματος τρόπος λήψης δεδομένων από το διαδίκτυο. Ένας τρόπος για να επιτευχθεί κάτι τέτοιο αναλύεται σε αυτήν την Ενότητα.

3.2.1 Τι είναι το web scraping;

Το **Web Scraping («Απόξεση» Ιστού)** ή αλλιώς web harvesting («Συγκομιδή» Ιστού), ή web data extraction (Εξαγωγή Δεδομένων Ιστού) είναι μια «απόξεση» δεδομένων (data scraping) δηλαδή, μια διαδικασία συλλογής

δομημένων δεδομένων ιστού με έναν αυτοματοποιημένο τρόπο. Γενικά, η εξαγωγή δεδομένων ιστού χρησιμοποιείται από άτομα και επιχειρήσεις που θέλουν να κάνουν χρήση του τεράστιου όγκου, διαθέσιμων στο κοινό, δεδομένων ιστού για να λάβουν πιο έξυπνες αποφάσεις.

Εάν έχετε αντιγράψει και επικολλήσει ποτέ πληροφορίες από έναν ιστότοπο, έχετε εκτελέσει την ίδια λειτουργία με κάθε web scraper, μόνο σε μικροσκοπική, χειροκίνητη κλίμακα. Σε αντίθεση με την τετριμμένη, ενοχλητική διαδικασία της μη αυτόματης εξαγωγής δεδομένων, το web scraping χρησιμοποιεί έξυπνη αυτοματοποίηση για να ανακτήσει εκατοντάδες, εκατομμύρια ή και δισεκατομμύρια σημεία δεδομένων από τα φαινομενικά ατελείωτα σύνορα του Διαδικτύου.

Το web scraping είναι δημοφιλής και αυτό επειδή, παρέχει κάτι πραγματικά πολύτιμο που τίποτα άλλο δεν μπορεί: σας δίνει δομημένα δεδομένα ιστού από οποιονδήποτε δημόσιο ιστότοπο. Περισσότερο από μια σύγχρονη ευκολία, η πραγματική δύναμη του δεδομένων έγκειται στην ικανότητά του να δημιουργεί και να τροφοδοτεί μερικές από τις πιο επαναστατικές επιχειρηματικές εφαρμογές στον κόσμο.

3.2.2 Τα βασικά του web scraping

Η διαδικασία του web scraping είναι εξαιρετικά απλή, λειτουργεί με δύο μέρη: το web crawler («ανιχνευτής» ιστού) και το web scraper («ξύστρα» ιστού). Ουσιαστικά, ο web crawler οδηγεί τον web scraper, μέσω του διαδικτύου, και εκείνος εξάγει τα δεδομένα που ζητήθηκαν. Παρακάτω φαίνεται πιο αναλυτικά η διαφορά μεταξύ του web crawler και του web scraper και του τρόπου λειτουργίας τους.

Ο **Web Crawler** γενικά αποκαλείται «αράχνη» (spider) και είναι μια Τεχνητή Νοημοσύνη που περιηγείται στο Διαδίκτυο για να αναζητήσει περιεχόμενο ακολουθώντας συνδέσμους και εξερευνώντας, όπως ένα άτομο με πολύ χρόνο στα χέρια του [Ken21]. Σε πολλά έργα, πρώτα «ανιχνεύετε» τον ιστό ή έναν συγκεκριμένο ιστότοπο για να ανακαλύψετε διευθύνσεις URL τις οποίες στη συνέχεια μεταβιβάζετε στον scraper σας.

Ο **Web Scraper** είναι ένα εξειδικευμένο εργαλείο που έχει σχεδιαστεί για την ακριβή και γρήγορη εξαγωγή δεδομένων από μια ιστοσελίδα [Ken21]. Οι web scrapers ποικίλλουν ευρέως ως προς το σχεδιασμό και την πολυπλοκότητα, ανάλογα με το έργο. Ένα σημαντικό μέρος κάθε scraper είναι οι εντοπιστές δεδομένων ή αλλιώς επιλογείς (data locators/selectors), που χρησιμοποιούνται για την εύρεση των δεδομένων που θέλετε να εξαγάγετε από το αρχείο HTML. Συνήθως, εφαρμόζονται επιλογείς XPath, CSS, regex ή συνδυασμός τους.

3.2.3 Τι είναι και πώς λειτουργει ένα web scraping tool;

Ένα web scraping tool (εργαλείο απόξεσης ιστού) είναι ένα πρόγραμμα λογισμικού που έχει σχεδιαστεί ειδικά για να εξάγει (ή να «ξύνει») σχετικές πληροφορίες από ιστότοπους [Ken21].

Ένα εργαλείο απόξεσης συνήθως κάνει αιτήματα HTTP σε έναν ιστότοπο-στόχο και εξάγει τα δεδομένα από μια σελίδα. Συνήθως, αναλύει περιεχόμενο που είναι δημόσια προσβάσιμο και ορατό στους χρήστες και αποδίδεται από τον διακομιστή ως HTML. Μερικές φορές κάνει επίσης αιτήματα σε εσωτερικές διεπαφές προγραμματισμού εφαρμογών (API) για ορισμένα σχετικά δεδομένα, όπως τιμές προϊόντων ή στοιχεία επικοινωνίας, που αποθηκεύονται σε μια βάση δεδομένων και παραδίδονται σε ένα πρόγραμμα περιήγησης μέσω αιτημάτων HTTP. Υπάρχουν διάφορα είδη εργαλείων web scrape, με δυνατότητες που μπορούν να προσαρμοστούν για να ταιριάζουν σε διαφορετικά έργα εξαγωγής. Για παράδειγμα, μπορεί να χρειαστείτε ένα εργαλείο απόξεσης που μπορεί να αναγνωρίσει μοναδικές δομές ιστοτόπων HTML ή να εξάγει, να διαμορφώσει ξανά και να αποθηκεύσει δεδομένα από API.

Τα εργαλεία απόξεσης μπορεί να είναι μεγάλα πλαίσια (frameworks) σχεδιασμένα για όλα τα είδη τυπικών εργασιών scraping, αλλά μπορείτε επίσης να χρησιμοποιήσετε βιβλιοθήκες προγραμματισμού γενικής χρήσης και να τις συνδυάσετε για να δημιουργήσετε έναν scraper.

Μια γενική διαδικασία web scraping φαίνεται παρακάτω:

1. Προσδιορίστε τον ιστότοπο-στόχο
2. Συλλέξτε διευθύνσεις URL των σελίδων από τις οποίες θέλετε να εξαγάγετε δεδομένα
3. Κάντε ένα αίτημα σε αυτές τις διευθύνσεις URL για να λάβετε το HTML της σελίδας
4. Χρησιμοποιήστε εντοπιστές για να βρείτε τα δεδομένα στο HTML
5. Αποθηκεύστε τα δεδομένα σε αρχείο JSON ή CSV ή σε κάποια άλλη δομημένη μορφή

3.3 Επιλογή Διαδικτυακού Μέσου Μαζικής Ενημέρωσης

Πριν γίνει χρήση της τεχνικής web scraping ώστε να ληφθούν δεδομένα έπρεπε να αποφασιστεί το Διαδικτυακό Μέσο Μαζικής Ενημέρωσης στο οποίο θα πραγματοποιούνταν η ανίχνευση ψευδών ειδήσεων. Το μέσο αυτό αποτέλεσε το Twitter το οποίο επιλέχθηκε γιατί είναι μια πλατφόρμα ενημέρωσης που στοχεύει στην κοινή χρήση ενημερωτικών και περιεκτικών ειδήσεων κυρίως μικρού μεγέθους που ονομάζονται tweets. Ταυτόχρονα, δίνει την δυνατότητα στους χρήστες να μοιράζονται ιδέες αλλά και πληροφορίες σε πραγματικό χρόνο.

3.3.1 Μορφή και χαρακτηριστικά των tweets

Ένα τυπικό tweet είναι συνήθως μια είδηση μικρού μεγέθους που περιέχει το όνομα του συντάκτη και την φωτογραφία του, το απεσταλμένο κείμενο, πιθανώς κάποια hashtags, μπορεί να περιέχει και συνδέσμους (links), ίσως περιέχει κάποιες αναφορές σε πρόσωπα (mentions). Ένα tweet μπορεί επίσης να περιέχει κάποια εικόνα ή κάποιο βίντεο, τα likes, τα retweets και τέλος τα replies του. Παρακάτω στο Σχήμα 3.1 φαίνονται δύο τέτοιο παραδείγματα τυχαίων tweets.

(α) 1ο Παράδειγμα (β) 2ο Παράδειγμα

Σχήμα 3.1: Παραδείγματα τυπικής μορφής των tweets

3.3.2 Εργαλεία που προσφέρει το Twitter για συλλογή δεδομένων

Το Twitter προσφέρει μια Διεπαφή Προγραμματισμού Εφαρμογών (Application Programming Interface - API) που ονομάζεται Twitter API. Χρησιμοποιώντας κάποια βιβλιοθήκη της Python (π.χ. Tweepy ή GetOldTweets3) μπορεί κάποιος να έχει πρόσβαση στο Twitter API και να συλλέξει δεδομένα από το Twitter εύκολα και συμβαδίζοντας με τους κανονισμούς του Twitter. Αυτή η μέθοδος όμως έχει κάποια μειονεκτήματα σε σχέση με την

τεχνική του web scraping που συζητήθηκε στην Ενότητα 3.2. Αρχικά είναι πιο αργή αλλά το κυριότερο είναι λιγότερο ευέλικτη όσον αφορά τη λήψη δεδομένων. Συγκεκριμένα, το Twitter API επιτρέπει την ανάκτηση των tweets έως και 7 ημέρες πριν και περιορίζεται στην λήψη 18.000 tweets ανά 15 λεπτά. Ενώ, σε αυτούς τους περιορισμούς προστίθενται και άλλοι από την εκάστοτε βιβλιοθήκη της Python που θα χρησιμοποιηθεί. Ωστόσο, το web scraping είναι πολύ ευαίσθητο στις αλλαγές του ιστότοπου [Pho22], [Don+20] και [Bec21]. Συνεπώς, θεωρώντας βέβαιη την ακεραιότητα του ιστότοπου του Twitter κατά την διεξαγωγή της εργασίας και με στόχο την λήψη πολύ μεγάλου όγκου πληροφορίας (ή ειδήσεων ή tweets) επιλέχθηκε η τεχνική του web scraping.

3.4 Web Scraping με το TWINT

Για την υλοποίηση της παρούσας εργασίας χρειάστηκε σύμφωνα με τα παραπάνω ένα web scraping tool προκειμένου να συλλεχθούν όσο το δυνατόν περισσότερα δεδομένα με έναν αυτοματοποιημένο τρόπο. Για την συγκεκριμένη εργασία, το Twitter αποτέλεσε την πηγή των ειδήσεων (ή δεδομένων ή tweets) που συλλέχθηκαν. Το Twitter API είναι ο τρόπος με τον οποίο κάποιος μπορεί να έχει πρόσβαση στα δεδομένα του Twitter. Όμως, ως εργαλείο έχει κάποιους περιορισμούς στο πόσα και κάθε πότε μπορεί κανείς να αποκτήσει δεδομένα από αυτό σύμφωνα με αυτά που συζητήθηκαν στην Υποενότητα 3.3.2. Εξαιτίας των παραπάνω, έχουν αναπτυχθεί διάφορες βιβλιοθήκες-εργαλεία που υλοποιούν την διαδικασία του web scraping σε δεδομένα από το Twitter. Αυτές οι βιβλιοθήκες, που στην πλειοψηφία τους είναι γραμμένες σε Python, δεν χρησιμοποιούν το Twitter API παρακάμπτοντας τους περιορισμούς που εκείνο θέτει. Μπορούν λοιπόν και να χρησιμοποιηθούν ανώνυμα χωρίς την δημιουργία λογαριασμού στο Twitter. Μία τέτοια δωρεάν βιβλιοθήκη είναι το TWINT [ZPL17] η οποία και χρησιμοποιήθηκε.

3.4.1 Κριτήρια για την συλλογή δεδομένων

Προκειμένου όμως να ληφθούν δεδομένα γύρω από το θέμα του κορονοϊού, έπρεπε να τεθούν κάποιοι περιορισμοί στο TWINT [ZPL17]. Συγκεκριμένα, χρησιμοποιήθηκαν κάποιες λέξεις-κλειδιά για την συλλογή των δεδομένων, ώστε το εργαλείο να μας επιστρέψει μόνο τις ειδήσεις (tweets) που περιέχουν μία ή περισσότερες από αυτές. Αυτές οι λέξεις-κλειδία φαίνονται στον Πίνακα 3.1.

Λέξεις-κλειδία
1. pandemic
2. quarantine
3. covidvariant
4. vaccination
5. coronavirus
6. corona
7. covid-19
8. covid19
9. covid
10. vaccine

Πίνακας 3.1: Λέξεις-κλειδία

Έπειτα, αφαιρέθηκαν όλες οι ειδήσεις (tweets) που δεν ήταν στην αγγλική γλώσσα προκειμένου να μπορεί να γίνει η επεξεργασία και η ανάλυση αυτών αργότερα, αλλά περισσότερες πληροφορίες για αυτό θα βρείτε

στο Κεφάλαιο 5. Τέλος, στα πλαίσια της εργασίας, ορίστηκε ένα χρονικό διάστημα των οποίων οι ειδήσεις θα ταξινομούνταν ως αληθείς ή ψευδείς και αυτό ήταν μεταξύ των ημερομηνιών 2021-06-01 και 2021-12-09 (συνολικά 6 μήνες και 9 μέρες). Το διάστημα αυτό επιλέχθηκε γιατί θεωρήθηκε αρκετά αντιπροσωπευτικό για το θέμα της εργασίας, δηλαδή ο Κορονοϊός, αλλά επίσης γιατί είναι αρκετά μεγάλο ώστε να δημιουργηθεί ένα μοντέλο Μηχανικής Μάθησης καλά εκπαιδευμένο (αφού θα εκπαιδεύεται με πολλά δεδομένα). Με την βοήθεια λοιπόν του εργαλείου και αρκετή υπομονή (αλλά όχι περισσότερη από το να γινόταν χειροκίνητα) πραγματοποιήθηκε η λήψη όλων αυτών των ειδήσεων που απαρτίστηκαν από συνολικά 54.243.060 tweets και σχηματίστηκε ένα CSV (comma-separated values) αρχείο μεγέθους 28.8 GB.

3.4.2 Μορφή των tweets μετά την λήψη τους από το TWINT και αλλαγές που έγιναν

Το CSV αρχείο που δημιουργήθηκε περιείχε όλα τα tweets και αλλά και αρκετές πληροφορίες ακόμα, για το καθένα. Αυτές μπορούν να εντοπιστούν και να κατανοηθούν από την επικεφαλίδα του αρχείου που εξηγεί το περιεχόμενο όλων των στηλών του αρχείου. Όπως φαίνεται στα παρακάτω Σχήματα 3.2, 3.3 και 3.4, η μορφή όλων των δεδομένων-στηλών είναι σε μορφή κειμένου (text) και έτσι είναι εύκολη τόσο η κατανόησή αυτών όσο και ο χειρισμός τους, κάτι που διακρίνεται και παρακάτω στην υποενότητα με τις αλλαγές που πραγματοποιήθηκαν.

Ξεκινώντας το διάβασμα και την κατανόηση του αρχείου συναντάται η πρώτη γραμμή του csv αρχείου που ονομάζεται επικεφαλίδα (header) η οποία στην δική μας περίπτωση είχε την παρακάτω μορφή:

id, conversation_id, created_at, date, time, timezone, user_id, username, name, place, tweet, language,mentions, urls, photos, replies_count, retweets_count, likes_count, hashtags, cashtags, link, retweet, quote_url, video, thumbnail, near, geo, source, user_rt_id, user_rt, retweet_id, reply_to, retweet_date, translate, trans_src, trans_dest

Το αρχείο που δημιουργήθηκε περιείχε την παραπάνω επικεφαλίδα, τα tweets, τις πληροφορίες τους και η μορφή του φαίνεται στα Σχήματα 3.2 και 3.3. Παρατηρείστε ότι στα tweets έχουν κρυφτεί τα πεδία username και name για λόγους ανωνυμοποίησης.

id	conversation_id	created_at	date	time	timezone	user_id	username	name	place	tweet	language	mentions	urls	photos	replies_count	retweets_count
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:59:06	200	15864446	[redacted]	[redacted]		The Omicron CO...	en	[]	['htt...	[]	0	1
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:57:20	200	327699059	[redacted]	[redacted]		@phi_nom3nal_i...	en	[]	[]	[]	1	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:54:48	200	3192562034	[redacted]	[redacted]		Pfizer Omicron P...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:53:10	200	2838876754	[redacted]	[redacted]		A 'sister' lineage...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:51:01	200	8.90E+17	[redacted]	[redacted]		In that it's a tale...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:50:41	200	8.90E+17	[redacted]	[redacted]		The defenders o...	en	[]	[]	[]	1	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:50:33	200	33106870	[redacted]	[redacted]		NZ and Australia...	en	[]	[]	[]	0	1
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:50:17	200	327699059	[redacted]	[redacted]		@YungChill6 @...	en	[]	[]	[]	1	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:50:13	200	9763482	[redacted]	[redacted]		New York State h...	en	[]	['htt...	[]	0	5
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:49:29	200	1.43E+18	[redacted]	[redacted]		Omicron, Pfizer:...	en	[]	['htt...	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:48:02	200	1.44E+18	[redacted]	[redacted]		Next will be 4 sh...	en	[{'screen...	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:46:36	200	1.40E+18	[redacted]	[redacted]		Alpha, Delta, an...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:46:36	200	2879714087	[redacted]	[redacted]		@MelissaMbarki...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:46:20	200	2771803146	[redacted]	[redacted]		San Francisco 20...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:44:45	200	1.42E+18	[redacted]	[redacted]		3 doses of Pfizer...	en	[]	[]	[]	53	18
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:43:56	200	8.90E+17	[redacted]	[redacted]		@Imani_Barbari...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:43:50	200	2373754790	[redacted]	[redacted]		Kitna mila @chet...	en	[{'screen...	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:43:32	200	274647532	[redacted]	[redacted]		my twitter feed i...	en	[]	[]	[]	1	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:43:16	200	1.46E+18	[redacted]	[redacted]		#Omicron #covi...	en	[]	[]	['https:...	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:42:49	200	1.24E+18	[redacted]	[redacted]		Omicron, Pfizer:...	en	[]	['htt...	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:42:34	200	1.34E+18	[redacted]	[redacted]		Pfizer says a 3rd...	en	[]	[]	[]	3	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:42:16	200	3013759560	[redacted]	[redacted]		@cloudy_cl &am...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:42:05	200	1.46E+18	[redacted]	[redacted]		Omicron Covid v...	en	[]	['htt...	[]	0	1
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:41:52	200	8.90E+17	[redacted]	[redacted]		Hello I want a w...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:40:32	200	1.05E+18	[redacted]	[redacted]		#WeWillNotCom...	en	[{'screen...	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:40:27	200	2533004783	[redacted]	[redacted]		@afneil Have u l...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:40:17	200	21524740	[redacted]	[redacted]		@CNN *Not resp...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:39:23	200	1.41E+18	[redacted]	[redacted]		Congo 5-Star?? I...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:39:07	200	1.16E+18	[redacted]	[redacted]		(2/12) The Fed h...	en	[]	[]	[]	1	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:37:30	200	8.90E+17	[redacted]	[redacted]		@hornyclerval S...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:37:12	200	23951401	[redacted]	[redacted]		@liamnorris1001...	en	[]	[]	[]	0	0
1.47E+18	1.47E+18	2021-12-09 0...	09-12-21	01:37:07	200	8.90E+17	[redacted]	[redacted]		@susanwrongta...	en	[]	[]	[]	0	0

Σχήμα 3.2: Αρχική μορφή των tweets αρχή

likes_count	hashtags	cashtags	link	retweet	quote_url	video	thumbnail	near	geo	source	user_rt_id	user_rt	retweet_id	reply_to	retweet_date	translate	trans_src	trans_dest
0	[]	[]	https://t...	FALSE		0								[]				
0	[]	[]	https://t...	FALSE		0								[{'scree...				
0	['omicro...	[]	https://t...	FALSE		1	https://p...							[]				
0	[]	[]	https://t...	FALSE		0								[]				
0	[]	[]	https://t...	FALSE		0								[]				
0	[]	[]	https://t...	FALSE		0								[]				
1	['nzpol',...	[]	https://t...	FALSE	https://t...	0								[]				
0	[]	[]	https://t...	FALSE		0								[{'scree...				
2	[]	[]	https://t...	FALSE		0								[]				
0	[]	[]	https://t...	FALSE		0								[]				
0	['vaccin...	[]	https://t...	FALSE	https://t...	0								[]				
0	['covidv...	[]	https://t...	FALSE		0								[]				
0	[]	[]	https://t...	FALSE		0								[{'scree...				
6	['covidv...	[]	https://t...	FALSE		0								[]				
43	[]	[]	https://t...	FALSE		1	https://p...							[]				
1	[]	[]	https://t...	FALSE		0								[{'scree...				
0	['covishi...	[]	https://t...	FALSE	https://t...	0								[]				
0	[]	[]	https://t...	FALSE		0								[]				
0	['omicro...	[]	https://t...	FALSE		1	https://p...							[]				
0	[]	[]	https://t...	FALSE		0								[]				
1	['nbc', '...	[]	https://t...	FALSE		1	https://p...							[]				
1	[]	[]	https://t...	FALSE		0								[{'scree...				
0	[]	[]	https://t...	FALSE		0								[]				
0	[]	[]	https://t...	FALSE		0								[]				
0	['wewill...	[]	https://t...	FALSE		0								[]				
2	[]	[]	https://t...	FALSE		0								[{'scree...				
2	[]	[]	https://t...	FALSE		0								[{'scree...				
0	[]	[]	https://t...	FALSE		0								[]				
7	[]	[]	https://t...	FALSE		0								[]				
1	[]	[]	https://t...	FALSE		0								[{'scree...				
2	[]	[]	https://t...	FALSE		0								[{'scree...				
2	[]	[]	https://t...	FALSE		0								[{'scree...				

Σχήμα 3.3: Αρχική μορφή των tweets συνέχεια

Πολλές όμως από αυτές τις στήλες περιείχαν πληροφορία που δεν μας χρησίμευε στην διαδικασία της

ανίχνευσης της κατηγορίας ενός tweet, αλλά και τα περισσότερα πεδία ορισμένων από αυτές τις στήλες ήταν κενά. Έτσι λοιπόν οι περισσότερες στήλες αφαιρέθηκαν. Κρίθηκαν απαραίτητες οι επικεφαλίδες των tweets που αναφερόντουσαν στον αριθμό των replies, των retweets, των likes και προφανώς τα ίδια τα tweets. Επιλέχθηκαν μόνο αυτές οι επικεφαλίδες διότι με βάση αυτές αργότερα στην Ενότητα 5.1 θα κρινόταν η κατηγορία ενός tweet. Έτσι, σχηματίστηκε ένα νέο CSV αρχείο μεγέθους 9.75 GB με λιγότερη πληροφορία αλλά πιο χρήσιμη για τον στόχο της εργασίας.

Επομένως, η επικεφαλίδα του νέου αρχείου που παράχθηκε ήταν η ακόλουθη:

tweet, replies_count, retweets_count, likes_count

Το νέο και τελικό αρχείο που δημιουργήθηκε περιείχε την κεφαλίδα, τα tweets, μόνο τις πληροφορίες που επιλέχθηκαν (δηλαδή τον αριθμό των replies στα tweets, των αριθμό των retweets του κάθε tweet και τέλος τα likes του κάθε tweet) και η μορφή του φαίνεται στο Σχήμα 3.4. Αυτά είναι τα στοιχεία των tweets που χρησιμοποιούνται στο Κεφάλαιο 5 για να κριθεί σε πολύ μεγάλη κλίμακα αν ένα tweet είναι αληθές ή ψευδές με στόχο την επίτευξη της αντικειμενικότητας του διαχωρισμού των tweet στον μεγαλύτερο δυνατό βαθμό. Συνεπώς, το αρχείο αυτό αποτέλεσε τον πυλώνα της εργασίας και πάνω σε αυτό βασίστηκε όλη η ανάλυση και η επεξεργασία που έγινε στην εργασία και καλύπτεται στα επόμενα κεφάλαια.

tweet	replies_count	retweets_co...	likes_count
The Omicron COVID variant is a growing cause of concern during the holiday season as New Yorkers anticipate...	0	1	0
@phi_nom3nal_ic3 But what's the reason?	1	0	0
Pfizer Omicron Protection Protection against Omicron coronavirus variant improves with three vaccine doses, Pf...	0	0	0
A 'sister' lineage of Omicron – the Covid variant rapidly sweeping the world – has been detected by scientists. 🤞‼	0	0	0
In that it's a tale as old as time w that type lol it sucks	0	0	0
The defenders of that person going basically "don't mob the mobber" is Interesting	1	0	0
NZ and Australia's COVID-19 strategy has shifted to suppression In the coming weeks we are looking at border...	0	1	1
@YungChill6 @CherriBerri21 since you a beginner I know you be asking	1	0	0
New York State has 20 confirmed omicron COVID variant cases. 13 of the 20 cases were identified in New York C...	0	5	2
Omicron, Pfizer: Three vaccines needed to protect against new Covid variant https://t.co/gPr9onhH8L	0	0	0
Next will be 4 shots 💉😳... do I hear 5 ..5... now we have 6 ... Sold to the multibillion dollar company @pfizer a...	0	0	0
Alpha, Delta, and Omicron walk into a bar and order a few Coronas. "That'll be $20.21 sir" the bartender says #c...	0	0	0
@MelissaMbarki maybe you have been hit by the next COVID variant LOL	0	0	0
San Francisco 2021. Out for a rainy stroll. A car passes by with no passengers. The driver has the wardrobe of a...	0	0	6
3 doses of Pfizer vaccine may be needed to protect against Omicron Covid variant https://t.co/MY0ZWbTA3u	53	18	43
@Imani_Barbarin Congratulations!!	0	0	1
Kitna mila @chetan_bhagat @kiranshaw in defaming #COVISHIELD & creating fear in peoples mind? "The i...	0	0	0
my twitter feed is nothing but new covid variant updates and yet if you placed a gun to my head and asked me...	1	0	0
#Omicron #covidvariant barely known yet already they tell you #Pfizer will work. No time for trials no time to te...	0	0	0
Omicron, Pfizer: Three vaccines needed to protect against new Covid variant https://t.co/wA6dAQbu4F	0	0	0
Pfizer says a 3rd shot is effective against Omicron, a pretty much harmless Covid variant. But the first 2 aren't. W...	3	0	1
@cloudy_cl & 99/100 people know about taper; FADE IT 100/100 know about Covid variant; FADE IT 99/10...	0	0	1
Omicron Covid variant poses greater risk for the unvaccinated, former White House advisor says https://t.co/W...	0	1	0
Hello I want a website that has fashion like Cider does but is less drop-shippy does anyone know any	0	0	0
#WeWillNotComply - unfortunately this could #backfire- as it us & our #lovedones we are #protecting by...	0	0	0
@afneil Have u lost any weight yet? I'm sure u want to try everything possible to beat this new Covid variant..!!	0	0	2
@CNN *Not responsible for side effects ** Number of required doses may change at any time ***Vaccine may n...	0	0	2
Congo 5-Star?? I'm experiencing agitation..omg. If they schmooze annnd..re-name their Covid variant 't'bogan'...	0	0	0
(2/12) The Fed has been clear that a taper is underway and most believe that not only will tapering continue, b...	1	0	7
@hornyclerval So sorry you're having to experience this	0	0	1
@liamnorris1001 And what, pray, is RELIGON? Please say it's not a new Covid variant	0	0	2
@susanwrongtag Action-packed cream cheese 🙄	0	0	2
London doctor warns parents of 'unusual' Omicron Covid variant symptom being seen in kids https://t.co/gla	0	0	0

Σχήμα 3.4: Τελική μορφή των tweets

Κεφάλαιο 4

Scikit-learn και classification αλγόριθμοι

Έχοντας χρησιμοποιήσει το TWINT [ZPL17] για την λήψη όλου του συνόλου των δεδομένων, το επόμενο βήμα ήταν η εύρεση αλγορίθμων που θα μας επέτρεπαν να λύσουμε το classification πρόβλημά μας. Υλοποιήσεις τέτοιων αλγορίθμων αλλά και πολλών ακόμα παρέχει η βιβλιοθήκη Scikit-learn [Ped+11].

4.1 Λίγα λόγια για το Scikit-learn

Το Scikit-learn [Ped+11] (επίσης γνωστό ως sklearn) είναι μια δωρεάν βιβλιοθήκη Μηχανικής Μάθησης λογισμικού για τη γλώσσα προγραμματισμού Python. Διαθέτει διάφορες οικογένειες αλγορίθμων ταξινόμησης, παλινδρόμησης και ομαδοποίησης, συμπεριλαμβανομένων Naive Bayes, Linear Regression, Logistic Regression, K-Nearest Neighbors (kNN), Decision Trees, Random Forest, Gradient Boosting Machines (GBM), Support Vector Machine (SVM), K-Means, Hierarchical Clustering, Artificial Neural Networks (ANN) κ.α. Επιπρόσθετα, η βιβλιοθήκη έχει σχεδιαστεί για να λειτουργεί με τις αριθμητικές και επιστημονικές βιβλιοθήκες Python NumPy και SciPy.

Η βιβλιοθήκη αυτή χρησιμοποιήθηκε στην εργασία προσφέροντας όλους τους αλγόριθμους Μηχανικής Μάθησης που παραμετροποιήθηκαν, εκπαιδεύτηκαν, χρησιμοποιήθηκαν και αξιολογήθηκαν στο μοντέλο που υλοποιήθηκε. Όλοι οι αλγόριθμοι που χρησιμοποιήθηκαν επιλύουν το πρόβλημα της δυαδικής Ταξινόμησης κειμένου (binary text Classification problem), αφού βοηθούν στην ταξινόμηση των κειμένων των tweets σε μία από τις δύο δυνατές κατηγορίες (αληθή ή ψευδή). Συνεπώς, αυτοί οι αλγόριθμοι υπόκεινται στην κατηγορία της Επιτηρούμενης Μάθησης που συζητήθηκε στην Υποενότητα 2.1.1 και διακρίνονται στον Πίνακα 4.2.

Τέλος, η βιβλιοθήκη αυτή παρείχε και την αυτοματοποιημένη παραμετροποίηση των αλγορίθμων μέσω του εργαλείου Αναζήτησης Πλέγματος (Grid Search), θα δούμε όμως περισσότερα για το εργαλείο αυτό στην Υποενότητα 5.5.1.

4.2 Classification αλγόριθμοι στο Scikit-learn

Συγκεκριμένα, οι οικογένειες Classification αλγορίθμων που χρησιμοποιήθηκαν για την υλοποίηση της εργασίας και τα αντίστοιχα ονόματα των υλοποιήσεών τους σε αλγορίθμους στο Scikit-learn [Ped+11] φαίνονται στον Πίνακα 4.2.

Οικογένειες classification αλγορίθμων	Υλοποιήσεις αλγορίθμων στο Scikit-learn
Support Vector Machine (SVM)	LinearSVC()
Decision Trees	DecisionTreeClassifier()
k-Nearest Neighbors (kNN)	KNeighborsClassifier()
Naive Bayes	GaussianNB()
Logistic Regression	LogisticRegression()
Random Forest	RandomForestClassifier()
Artificial Neural Networks (ANN)	MLPClassifier()

Πίνακας 4.1: Classification αλγόριθμοι

Κεφάλαιο 5

Δημιουργία classification μοντέλου για ανίχνευση ψευδών ειδήσεων

Έχουμε μέχρι τώρα εξηγήσει τι είναι η Μηχανική Μάθηση καθώς και τις κατηγορίες και τους αλγορίθμους γύρω από αυτήν. Επιλέξαμε επίσης, να πάρουμε δεδομένα από ένα μέσο κοινωνικής δικτύωσης, το Twitter, που συστηματικά χρησιμοποιείται για ενημέρωση και διασπορά ειδήσεων. Είδαμε τι χρειάζεται να κάνουμε προκειμένου να αποκτήσουμε, δωρεάν και με αυτόματο τρόπο, έναν ιδιαίτερα μεγάλο όγκο δεδομένων από το διαδίκτυο και βρήκαμε τρόπο να χρησιμοποιήσουμε τους αλγόριθμους που χρειαζόμαστε για την εργασία μας. Μένει να δημιουργήσουμε λοιπόν ένα μοντέλο Μηχανικής Μάθησης που θα χρησιμοποιεί αυτούς τους αλγόριθμους και θα ανιχνεύει ψευδές ειδήσεις. Σε αυτό το Κεφάλαιο θα δείξουμε ποιες ενέργειες και με ποια σειρά χρειάζεται να γίνουν ώστε να δημιουργηθεί ένα τέτοιο μοντέλο από την αρχή μέχρι το τέλος.

5.1 Φόρτωση των δεδομένων

Το πρώτο βήμα για την δημιουργία του μοντέλου που θα πρέπει να γίνει, είναι να «φορτώσουμε» τα δεδομένα μας από το CSV αρχείο που δημιουργήθηκε, σε κάποια δομή στην μνήμη του υπολογιστή. Ενώ, παράλληλα για κάθε ένα tweet με βάση κάποια κριτήρια να του «προσκολλούμε» και μια ετικέτα για το αν είναι αληθές ή ψευδές, δηλαδή να το χαρακτηρίζουμε. Για την εργασία αυτή μετά από πολλές δοκιμές σε τιμές με στόχο τις καλύτερες επιδόσεις του μοντέλου αλλά ταυτόχρονα και την εγκυρότητα της αξιολόγησης ενός tweet, τα κριτήρια αυτά ορίστηκαν όπως φαίνεται στον Πίνακα 5.1. Οπότε, ένα tweet θεωρείται **«αληθές»** εάν ισχύουν και οι τρείς αυτές συνθήκες ταυτόχρονα, δηλαδή ότι ο αριθμός των replies $\geq$ 200 ΚΑΙ ο αριθμός των retweets $\geq$ 300 ΚΑΙ ο αριθμός των likes $\geq$ 500 , αλλιώς θεωρείται **«ψευδές»**.

Ετικέτα	Αριθμός των replies	Αριθμός των retweets	Αριθμός των likes
Αληθές	≥200	≥300	≥500
Ψευδές	<200	<300	<500

Πίνακας 5.1: Κριτήρια επιλογής ετικέτας ενός tweet

Πιο συγκεκριμένα, τα κατώφλια αυτά επιλέχθηκαν, με στόχο τον αποκλεισμό των tweets που είτε ήταν spam, είτε προσωπική άποψη κάποιου τυχαίου και άγνωστου χρήστη πάνω στο θέμα, είτε που προέρχονταν από πηγές με πολύ περιορισμένο ενεργό αριθμό ακολούθων και συνεπώς με μικρό κύρος. Ο απώτερος στόχος της θεώρησης όλων αυτών των tweet ως ψευδείς ειδήσεις, ήταν μια αυτοματοποιημένη σε μεγάλη κλίμακα

εκπαίδευση αλλά και κυριότερα μια εκπαίδευση αλγορίθμων (βλ. Ενότητα 5.6) που θα ήταν αντικειμενική και δεν θα είχε γίνει για λίγα tweets από εμάς.

Έπειτα, για να εξασφαλίσουμε την τυχαιότητα με την οποία θα εκλάμβαναν τα δεδομένα οι αλγόριθμοι, ώστε να εκπαιδευτούν πιο «αντικειμενικά», προστέθηκε ένα τυχαίο ανακάτεμα στα δεδομένα. Τέλος, αφαιρέθηκαν μερικές (απειροελάχιστες στο πλήθος σε σύγκριση με τις συνολικές) γραμμές του αρχείου που τα δεδομένα είχαν εισαχθεί εσφαλμένα σε αυτές από το TWINT [ZPL17], ή στις οποίες μία ή περισσότερες στήλες ήταν κενές ή έλειπαν. Έτσι, καταλήξαμε να έχουμε στην μνήμη του υπολογιστή tweets χαρακτηρισμένα με ετικέτες και έτοιμα για επεξεργασία. Το μέγεθος που καταλάμβαναν τα tweets στην μνήμη διέφερε ανάλογα με το πόσα είχαν επιλεχθεί από το CSV αρχείο για την δημιουργία του μοντέλου (λόγω της περιορισμένης διαθέσιμης μνήμης επιλεγόταν έναν κομμάτι του συνόλου κάθε φορά, περισσότερες λεπτομέρειες για τις επιλογές του συνόλου αυτού αναλύονται στο Κεφάλαιο 6).

5.2 Επεξεργασία των δεδομένων

Το επόμενο βήμα κατά την δημιουργία ενός μοντέλου Μηχανικής Μάθησης είναι η επεξεργασία των tweets και συγκεκριμένα η μορφοποίηση τους σε μια απλούστερη και πιο εύκολα επεξεργάσιμη μορφή. Συγκεκριμένα, στα tweets έγινε αφαίρεση των σημείων στίξης (π.χ. -,!,?,...) και των ειδικών χαρακτήρων (π.χ. ,&,#,...), αφαίρεση όλων των μονών χαρακτήρων (π.χ. a,s,o,d,...) που πιθανώς μπορεί να υπάρχουν, μετατροπή σε πεζά γράμματα και τέλος πραγματοποιήθηκε λημματοποίηση (αναγωγή της λέξης στον πρώτο κλιτικό τύπο) σε ορισμένες από τις λέξεις των tweets χρησιμοποιώντας τον WordNetLemmatizer της βιβλιοθήκης nltk της Python για την δουλεία αυτή [Too]. Η απόφαση για το σε ποιες λέξεις θα πραγματοποιόταν αυτή η λημματοποίηση αποφασίστηκε από τον WordNetLemmatizer. Ένα παράδειγμα λημματοποίησης με τον WordNetLemmatizer είναι: το ”cats” μετατρέπεται σε ”cat”. Η λημματοποίηση γίνεται για να αποφευχθεί η δημιουργία χαρακτηριστικών που είναι σημασιολογικά παρόμοια αλλά συντακτικά διαφορετικά. Έτσι, ένα τυχαίο tweet που αρχικά είχε την εξής μορφή:

@monicaonairtalk We stopped at the mask. Before corona the Lord led me, showed me a plot to inoculate the world with a plant-based vaccine. My husband thought I was nuts, then the news of the “virus” came. Praise God he could see. We don’t know his status, but he hasn’t served since March 2020

Κατέληξε να έχει την εξής μορφή:

monicaonairtalk we stopped at the mask before corona the lord led me showed me plot to inoculate the world with plant based vaccine my husband thought wa nut then the news of the virus came praise god he could see we don know his status but he hasn served since march 2020

Όλη αυτή η επεξεργασία δεν είναι ανούσια και η αξία της φαίνεται σε επόμενη Ενότητα του Κεφαλαίου.

5.3 Δεδομένα εκπαίδευσης και δεδομένα δοκιμής

Οι αλγόριθμοι Μηχανικής Μάθησης θα χρειαστούν μια βάση, ή αλλιώς μια αρχική εκπαίδευση, προκειμένου να είναι σε θέση να χαρακτηρίσουν μια είδηση ως αληθής ή ψευδής. Έτσι, θα πρέπει τα δεδομένα να χωριστούν σε δύο ομάδες με κάποια αναλογία [Nel19]. Η πρώτη ομάδα συνήθως αποτελεί το μεγαλύτερο ποσο-

στό και περιέχει χαρακτηρισμένα δεδομένα εκπαίδευσης (train data) με τις ετικέτες «αληθές» («POSITIVE») και «ψευδές» («NEGATIVE»). Οι χαρακτηρισμένες ετικέτες των υπόλοιπων δεδομένων της δεύτερης ομάδας δεν λαμβάνονται υπόψιν και τα δεδομένα δοκιμής (test data) χρησιμοποιούνται από τους αλγορίθμους ώστε να χαρακτηριστούν οι ετικέτες τους εκ νέου με Μηχανική Μάθηση. Στην εργασία αυτή το 80% των δεδομένων αποτέλεσαν τα δεδομένα εκπαίδευσης και 20% των δεδομένων τα δεδομένα δοκιμής.

5.4 Μετατροπή κειμένου σε αριθμούς

Οι μηχανές, σε αντίθεση με τους ανθρώπους, δεν μπορούν να κατανοήσουν το ακατέργαστο κείμενο. Οι μηχανές μπορούν να αντιληφθούν μόνο αριθμούς. Ειδικότερα, οι στατιστικές τεχνικές όπως η Μηχανική Μάθηση μπορούν να αντιμετωπίσουν μόνο αριθμούς. Επομένως, πρέπει να μετατρέψουμε το κείμενό μας σε αριθμούς.

Υπάρχουν διαφορετικές προσεγγίσεις για τη μετατροπή του κειμένου στην αντίστοιχη αριθμητική μορφή. Το μοντέλο Bag of Words (Bag of Words Model) και το μοντέλο ενσωμάτωσης λέξεων (Word Embedding Model) είναι δύο από τις πιο συχνά χρησιμοποιούμενες προσεγγίσεις [Nel19]. Σε αυτήν την εργασία, θα χρησιμοποιήσουμε το μοντέλο του bag of words για να μετατρέψουμε τα tweets μας σε αριθμούς.

5.4.1 Μοντέλο Bag of Words

Για την υλοποίηση του Bag of Words μοντέλου, χρησιμοποιήθηκε η κλάση CountVetorizer από τη βιβλιοθήκη sklearn.feature_extraction.text του Scikit-learn [Ped+11]. Σε αυτήν την κλάση, υπήρχαν κάποιες σημαντικές παράμετροι που έπρεπε να ρυθμιστούν προκειμένου να επιτευχθεί η μετατροπή των λέξεων σε αριθμούς. Αρχικά, επειδή όταν μετατρέπετε λέξεις σε αριθμούς χρησιμοποιώντας την προσέγγιση Bag of Words, όλες οι μοναδικές λέξεις σε όλα τα tweets μετατρέπονται σε χαρακτηριστικά. Έτσι, όλα τα tweets μπορούν να περιέχουν δεκάδες χιλιάδες μοναδικές λέξεις. Αλλά οι λέξεις που έχουν πολύ χαμηλή συχνότητα εμφάνισης δεν αποτελούν καλή παράμετρο για την ταξινόμηση των tweets και απλά καταλαμβάνουν άσκοπη μνήμη και μειώνουν την επίδοση του μοντέλου. Επομένως, έπρεπε να βρεθούν τιμές για τις παραμέτρους που θα επέφεραν αρκετά υψηλή επίδοση στο μοντέλο, χωρίς όμως το επιστρεφόμενο αποτέλεσμα της κλάσης να είναι αδιαχείριστο για την διαθέσιμη μνήμη μας. Έτσι, όλες οι παρακάτω παράμετροι ορίστηκαν μετά από πολλές δοκιμές λαμβάνοντας υπόψιν όλα τα παραπάνω.

Ως εκ τούτου, η τιμή της πρώτης παραμέτρου max_features ορίστηκε σε 1500, πράγμα που σημαίνει ότι θέλουμε να χρησιμοποιήσουμε τις 1500 πιο συχνές λέξεις σε tweets ως χαρακτηριστικά για την εκπαίδευση του ταξινομητή (Classifier) μας. Η επόμενη παράμετρος είναι min_df και έχει οριστεί σε 20. Αυτό αντιστοιχεί στον ελάχιστο αριθμό tweets που πρέπει να περιέχουν ένα χαρακτηριστικό. Επομένως, συμπεριλήφθηκαν μόνο εκείνες οι λέξεις που εμφανίζονται σε τουλάχιστον 20 tweets. Έπειτα, για το χαρακτηριστικό max_df, η τιμή έχει οριστεί σε 0.3, δηλαδή ότι πρέπει να συμπεριληφθούν μόνο εκείνες οι λέξεις που εμφανίζονται στο 30% το πολύ όλων των tweets. Οι λέξεις που εμφανίζονται σχεδόν σε κάθε tweet συνήθως δεν είναι κατάλληλες για ταξινόμηση επειδή δεν παρέχουν μοναδικές πληροφορίες για το tweet.

Τέλος, αφαιρέθηκαν οι πιο κοινές λέξεις, όπως άρθρα, προθέσεις, αντωνυμίες, σύνδεσμοι κ.λπ. (αυτές οι λέξεις ονομάζονται stop-words) αφού δεν περιέχουν χρήσιμες πληροφορίες για το tweet. Λίγα παραδείγματα τέτοιων λέξεων στα αγγλικά είναι: «the», «a», «an», «so», «what». Για να αφαιρέσουμε τις stop-words χρησιμοποιήσαμε την λίστα με τις stop-words της βιβλιοθήκης nltk.corpus και την περάσαμε στην παράμετρο stop_words της κλάσης.

Σε αυτό το σημείο θα πρέπει να τονιστεί πως χωρίς το κομμάτι της επεξεργασίας σε κάθε tweet θα χρειαζόμασταν μεγαλύτερη μνήμη στο σύστημα, διότι θα πραγματοποιούνταν περισσότερες συγκρίσεις αφού θα

υπήρχαν παραπανίσιοι αριθμοί που όπως ήδη αναφέρθηκε θα είχαν παρόμοια σημασιολογική ερμηνεία και δεν θα πρόσφεραν κάποια περαιτέρω ακρίβεια στο μοντέλο μας. Αντιθέτως αυτές οι παραπάνω πράξεις θα μείωναν την απόδοση του μοντέλου μας.

Η προσέγγιση Bag of Words λειτουργεί καλά για τη μετατροπή κειμένου σε αριθμούς. Ωστόσο, έχει ένα μειονέκτημα. Εκχωρεί μια βαθμολογία σε μια λέξη με βάση την εμφάνισή της σε ένα συγκεκριμένο tweet. Δεν λαμβάνει υπόψη το γεγονός ότι η λέξη μπορεί επίσης να έχει υψηλή συχνότητα εμφάνισης και σε άλλα tweets. Θεωρώντας το κάθε tweet σαν ένα έγγραφο, το $TFIDF = TF \times IDF$ επιλύει αυτό το ζήτημα πολλαπλασιάζοντας τον όρο TF (term frequency): «συχνότητα μιας λέξης» με τον όρο IDF (inverse document frequency): «αντίστροφη συχνότητα εγγράφου» [Sco21].

Η συχνότητα μιας λέξης υπολογίζεται ως εξής:

Συχνότητα λέξης = Αριθμός εμφανίσεων της λέξης / Σύνολο λέξεων στο έγγραφο

και η Συχνότητα Αντίστροφου Εγγράφου υπολογίζεται ως:

Συχνότητα Αντίστροφου Εγγράφου(λέξη) = Συνολικός αριθμός εγγράφων / Αριθμός εγγράφων που περιέχουν τη λέξη.

Η τιμή TFIDF για μια λέξη σε ένα συγκεκριμένο tweet είναι υψηλότερη εάν η συχνότητα εμφάνισης αυτής της λέξης είναι μεγαλύτερη στο συγκεκριμένο tweet αλλά χαμηλότερη σε όλα τα άλλα tweets.

Έτσι, λοιπόν αντικαταστάθηκε η κλάση CountVetorizer με την κλάση TfidfVectorizer της ίδιας βιβλιοθήκης, η οποία υλοποιεί το Bag of Words μοντέλο αλλά κάνει χρήση και του TFIDF. Όλες οι τιμές των παραμέτρων έμειναν ως είχαν.

Παρακάτω φαίνεται ο κώδικας σε Python που έκανε όλη αυτήν την δουλειά:

```python
from sklearn.feature_extraction.text import TfidfVectorizer
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
import numpy as np

vectorizer = TfidfVectorizer(max_features=1500, min_df=20, max_df=0.3, stop_words=
    stopwords.words('english'))

train_x_vectors = np.asarray(vectorizer.fit_transform(train_x).todense())
test_x_vectors = np.asarray(vectorizer.transform(test_x).todense())
```

Έτσι ένα επεξεργασμένο tweet μαζί με την μετατροπή του σε αριθμούς που έχει πολύ λίγη πληροφορία με βάση τις παραμέτρους μας φαίνεται παρακάτω:

7newsaustralia died from corona they should wear mask
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Ενώ, ένα επεξεργασμένο tweet μαζί με την μετατροπή του σε αριθμούς που έχει αρκετή πληροφορία με βάση τις παραμέτρους μας φαίνεται πιο κάτω:

there is no confusion between influenza flu and covid coronavirus in any of our pcr or antigen test they are completely different family covid is new deadly pandemic virus covid19 covid coronavirus corona http co 2iqlpkr6qy
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.12146628 0. 0. 0.51024387 0. 0. 0.62271141 0.23764539 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.27669284 0. 0. 0.25512194 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.12243358 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.21004535 0. 0. 0. 0. 0. 0.24692106 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.13757594 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

5.5 Παραμετροποίηση classification αλγορίθμων

Σε αυτό το σημείο τα tweets έχουν αποκτήσει μορφή αριθμών και είναι έτοιμα για χρήση από τους αλγορίθμους Μηχανικής Μάθησης. Πριν όμως μπορέσουμε να χρησιμοποιήσουμε τους αλγόριθμους που αναφέραμε στο προηγούμενο Κεφάλαιο θα χρειαστεί να τους παραμετροποιήσουμε. Η παραμετροποίηση τους στοχεύει στα καλύτερα δυνατά αποτελέσματα του μοντέλου τόσο σε θέματα χρονικής διάρκειας εκτέλεσης όσο και σε θέματα απαίτησης μνήμης αλλά και επίδοσης στην ορθότητα της ανίχνευσης ψευδών ειδήσεων.

5.5.1 Ρύθμιση παραμέτρων με Grid Search

Η διαδικασία παραμετροποίησης των αλγορίθμων παρέχεται επίσης από την ίδια βιβλιοθήκη Scikit-learn [Ped+11] μέσω του εργαλείου Αναζήτησης Πλέγματος (Grid Search) [Mal20] και [Sci22]. Το Grid Search λειτουργεί ως εξής: για δεδομένες τιμές σε παραμέτρους που του έχουν δοθεί, εξετάζει εξαντλητικά όλους τους συνδυασμούς των παραμέτρων με όλες τις καταχωρημένες (από τον χρήστη) για αυτές τιμές και επιστρέφει τον καλύτερο συνδυασμό αυτών.

Έτσι, μετά από μελέτη όλων των παραμέτρων στους επτά αλγόριθμους που χρησιμοποιήθηκαν στο μοντέλο, ώστε να βρεθούν ποιες θα ήταν χρήσιμες για την ταξινόμηση, δόθηκαν κάποιες τιμές για αυτές στο Grid Search. Έπειτα, έγινε αναζήτηση του καλύτερου συνδυασμού των παραμέτρων και των τιμών αυτών με 100.000 tweets (με περισσότερα tweets για δοκιμή ο χρόνος απάντησης του Grid Search για αρκετές τιμές σε παραμέτρους ήταν υπερβολικά μεγάλος). Στον Πίνακα 5.2 παρατίθεται μια λίστα με την παραμετροποίηση που έγινε στους αλγορίθμους με βάση τα αποτελέσματα του Grid Search.

Αλγόριθμος	**Παράμετροι και τιμές**
LinearSVC()	C=1, max_iter=10000, random_state=42
DecisionTreeClassifier()	criterion='gini', splitter='random', random_state=42
KNeighborsClassifier()	n_neighbors=5, weights='distance', n_jobs=-1
GaussianNB()	
LogisticRegression()	solver='sag', random_state=42, n_jobs=-1
RandomForestClassifier()	max_features='log2', n_estimators=200, criterion='gini', random_state=42, n_jobs=-1
MLPClassifier()	max_iter=10, activation='relu', warm_start=True, verbose=False, random_state=42

Πίνακας 5.2: Παραμετροποίηση αλγορίθμων

Αυτές οι παράμετροι για κάθε αλγόριθμο αναλύονται στον Πίνακα 5.3.

Παράμετροι	Επεξήγηση Παραμέτρων
C	Παράμετρος κανονικοποίησης. Η ισχύς της τακτοποίησης είναι αντιστρόφως ανάλογη του C. Πρέπει να είναι αυστηρά θετική.
max_iter	Ο μέγιστος αριθμός επαναλήψεων του αλγορίθμου προς εκτέλεση.
random_state	Σπόρος που ελέγχει τη δημιουργία ψευδοτυχαίων αριθμών για το ανακάτεμα των δεδομένων.
criterion	Η λειτουργία για τη μέτρηση της ποιότητας ενός διαχωρισμού κόμβων στο δέντρο.
splitter	Η στρατηγική που χρησιμοποιήθηκε για την επιλογή του διαχωρισμού σε κάθε κόμβο.
n_neighbors	Προεπιλεγμένος αριθμός γειτόνων για χρήση για ερωτήματα (queries) του αλγορίθμου.
weights='distance'	Συνάρτηση βάρους που χρησιμοποιείται στην πρόβλεψη. Με αυτήν την τιμή, οι πιο κοντινοί γείτονες ενός σημείου ερωτήματος θα έχουν μεγαλύτερη επιρροή από τους γείτονες που βρίσκονται πιο μακριά.
n_jobs	Ο αριθμός των παράλληλων εργασιών προς εκτέλεση. Χρήση όλων των πυρήνων του συστήματος για ταχύτερη εκτέλεση αλγορίθμων.
solver='sag'	Αλγόριθμος προς χρήση στο πρόβλημα βελτιστοποίησης. Με αυτήν την τιμή έχουμε ταχύτερη εκτέλεση του αλγορίθμου για μεγάλα σύνολα δεδομένων.
max_features	Ο αριθμός των χαρακτηριστικών που πρέπει να ληφθούν υπόψη κατά την αναζήτηση του καλύτερου διαχωρισμού κόμβων.
n_estimators	Ο αριθμός των δέντρων στο δάσος.
activation='relu'	Συνάρτηση ενεργοποίησης για το κρυφό στρώμα (hidden layer).
warm_start	Επαναχρησιμοποίηση τη λύσης της προηγούμενης κλήσης του «fitted» μοντέλου για την καταχώρηση των δεδομένων (fit) ως αρχικοποίηση στην επόμενη κλήση του μοντέλου.
verbose	Εκτύπωση μηνυμάτων προόδου των επαναλήψεων του αλγορίθμου στην έξοδο.

Πίνακας 5.3: Επεξήγηση παραμέτρων για τους αλγόριθμους ΜΜ που χρησιμοποιήθηκαν

5.6 Εκπαίδευση classification αλγορίθμων

Έχοντας λοιπόν παραμετροποιήσει όσο το δυνατόν καλύτερα τους αλγορίθμους μένει να εκπαιδεύσουμε τους αλγορίθμους με τα δεδομένα εκπαίδευσης που έχουν μετατραπεί σε αριθμούς, ώστε να χαρακτηρίσουν τα δεδομένα δοκιμής που επίσης έχουν μετατραπεί σε αριθμούς. Η εκπαίδευση αυτή πραγματοποιείται μέσω της κλάσης fit(train_x, train_y) του Scikit-learn [Ped+11], η οποία δίνει τα δεδομένα εκπαίδευσης στους αλγορίθμους και αυτοί εκπαιδεύονται με αυτά ώστε να μπορούν να «προβλέπουν» νέα. Όπου, train_x είναι το tweet και train_y είναι η χαρακτηρισμένη από εμάς ετικέτα του («αληθές» ή «ψευδές»).

Εδώ όμως, πρέπει να τονιστεί πως τα δεδομένα εκπαίδευσης ισοσταθμίστηκαν από άποψη κλάσεων κατά την εισχώρησή τους στους αλγόριθμους Μηχανικής Μάθησης. Δηλαδή, ίδιος αριθμός από tweets τόσο με ετικέτα «αληθές» όσο και με ετικέτα «ψευδές» τροφοδοτήθηκαν στους αλγόριθμους ως δεδομένα εκπαίδευσης. Αυτή η ισοστάθμιση στα δεδομένα εκπαίδευσης είναι σημαντική και στοχεύει στην ομοιόμορφη εκπαίδευση των αλγορίθμων και για τις δύο ετικέτες ώστε να μπορούν να προβλέπουν και τις δύο με ίση πιθανότητα.

5.7 Πρόβλεψη με classification αλγόριθμους

Τέλος, το μόνο που μένει είναι να χρησιμοποιήσουμε τους αλγορίθμους για να προβλέψουμε για όλα τα tweets που αποτελούν τα δεδομένα δοκιμής μας, αν αποτελούν «αληθές» ή «ψευδές» ειδήσεις. Ο τρόπος με τον οποίο πραγματοποιείται αυτή η πρόβλεψη είναι μέσω της κλάσης predict(test_x) του Scikit-learn [Ped+11], η οποία χρησιμοποιώντας τον αλγόριθμο στον οποίο έχει εφαρμοστεί, προβλέπει/χαρακτηρίζει το tweet με μία ετικέτα από τις διαθέσιμες («αληθείς» ή «ψευδείς») και την επιστρέφει. Έτσι, ολοκληρώνεται όλη η διαδικασία ανίχνευσης ψευδών ειδήσεων και παραλαμβάνουμε τις ετικέτες για κάθε tweet. Επιπλέον, οι παραχθείσες ετικέτες από τους αλγόριθμους μπορούν να συγκριθούν με τις ετικέτες των tweets που είχαμε εμείς χαρακτηρίσει για να καταλάβουμε το πόσο ακριβής ήταν η ταξινόμηση των tweets από τους αλγορίθμους.

Στο σημείο αυτό πρέπει να σημειωθεί πως λόγω τις περιορισμένης μνήμης που είχε ο υπολογιστής στον οποίο υλοποιήθηκε η εργασία (32GB RAM) επιτεύχθηκε ανίχνευση ψευδών ειδήσεων μέχρι και σε 5 εκατομμύρια tweets. Το θέμα αυτό αναλύεται στο Κεφάλαιο 7.

Κεφάλαιο 6

Αξιολόγηση classification μοντέλου

Έχοντας πραγματοποιήσει την ανίχνευση ψευδών ειδήσεων χρησιμοποιώντας τους αλγόριθμους Μηχανικής Μάθησης, γεννάται ένα ερώτημα για το ποιος ή ποιοι από αυτούς είναι οι καλύτεροι σε θέματα επιδόσεων για την ανίχνευση αυτή. Σε αυτό το Κεφάλαιο λοιπόν, θα γίνει λεπτομερής σύγκριση όλων των αλγορίθμων που χρησιμοποιήθηκαν με βάση τους χρόνους εκτέλεσής τους αλλά και με κάποιες δημοφιλείς μετρικές αξιολόγησης μοντέλων Μηχανικής Μάθησης.

6.1 Χρόνοι εκτέλεσης αλγορίθμων

Οι αλγόριθμοι συγκρίθηκαν με βάση τους χρόνους εκτέλεσής τους καθώς τα tweets αυξάνονταν. Συγκεκριμένα, έγινε σύγκριση των χρόνων εκτέλεσης των αλγορίθμων για 1.000, 10.000, 100.000, 1.000.000, 2.000.000 και 5.000.000 tweets. Παρακάτω φαίνεται η γραφική αναπαράσταση της σύγκρισης αυτής φαίνεται στο Σχήμα 6.1.

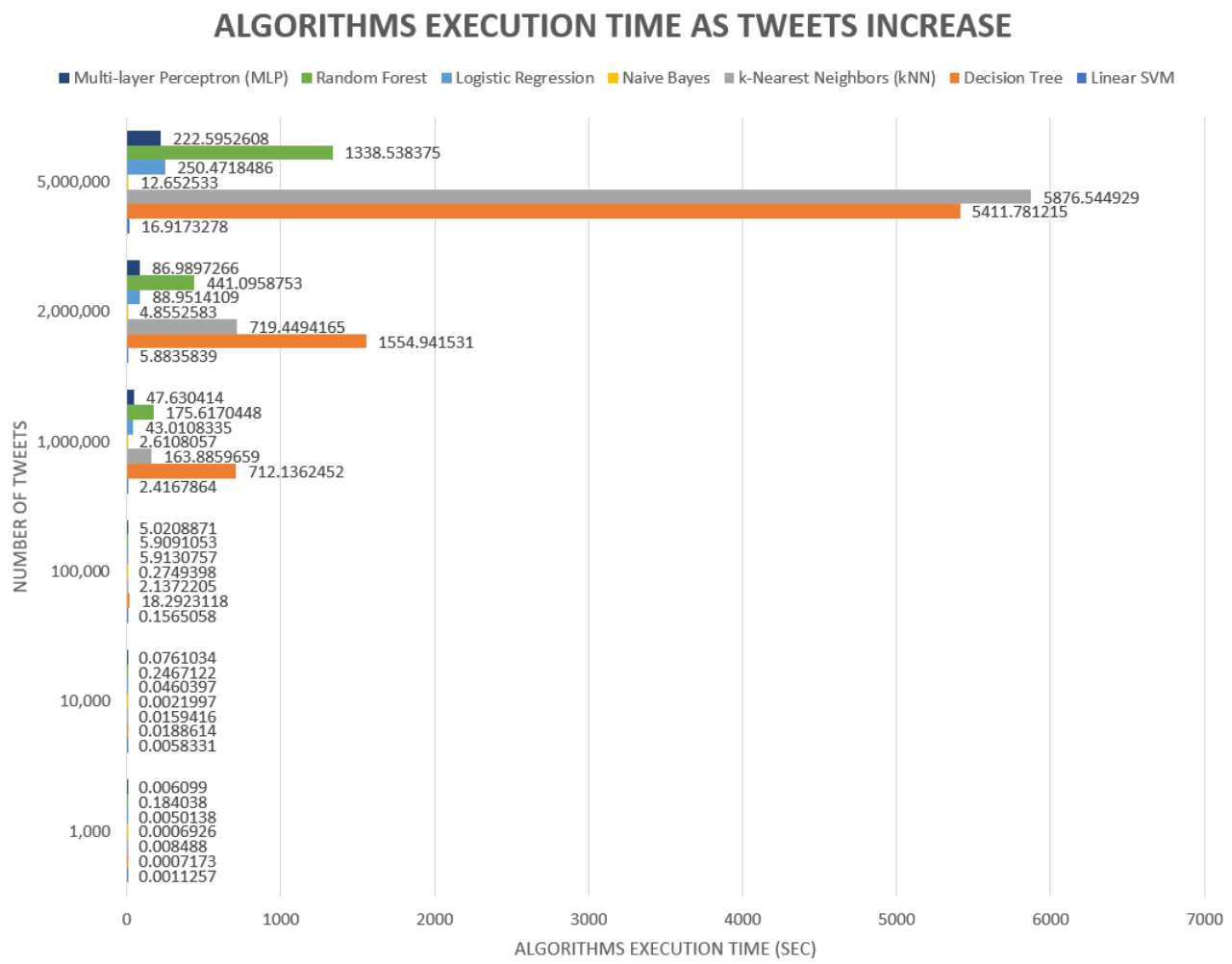
Από το Σχήμα 6.1, παρατηρούμε ότι για μεγάλου όγκου δεδομένα (5.000.000 tweets) ο ταχύτερος αλγόριθμος είναι ο Naive Bayes με ταχύτητα, σε δευτερόλεπτα (sec), της τάξης μεγέθους $1,2 \times 10^1$ sec. Ενώ, στην δεύτερη θέση χωρίς να απέχει πολύ είναι ο Linear SVM με ταχύτητα $1,6 \times 10^1$ sec. Μια τάξη μεγέθους πιο πάνω, βρίσκεται ο Multi-layer Perceptron (MLP) με $2,2 \times 10^2$ sec και ο Logistic Regression με $2,5 \times 10^2$ sec. Ενώ, ακόμα μια τάξη μεγέθους πιο πάνω βρίσκεται ο Random Forest με $1,3 \times 10^3$ sec, ο Desicion Tree με $5,4 \times 10^3$ sec και τελευταίος και αργότερος ο k-Nearest Neighbors (kNN) με $5,9 \times 10^3$ sec.

6.2 Δημοφιλείς μετρικές αξιολόγησης μοντέλων Μηχανικής Μάθησης

Προκειμένου να συγκρίνουμε την επίδοση της ανίχνευσης του κάθε αλγορίθμου θα χρησιμοποιήσουμε τις πιο δημοφιλείς μετρικές αξιολόγησης μοντέλων Μηχανικής Μάθησης.

6.2.1 Ακρίβεια (Accuracy)

Η ακρίβεια (Accuracy) είναι η απλούστερη από όλες τις μετρικές αξιολόγησης και η πιο συχνά χρησιμοποιούμενη λόγο του εύκολου υπολογισμού της. Η ακρίβεια ταξινόμησης ενός αλγορίθμου υπολογίζεται ως εξής:
Accuracy = αριθμός σωστών προβλέψεων / αριθμός συνολικών προβλέψεων

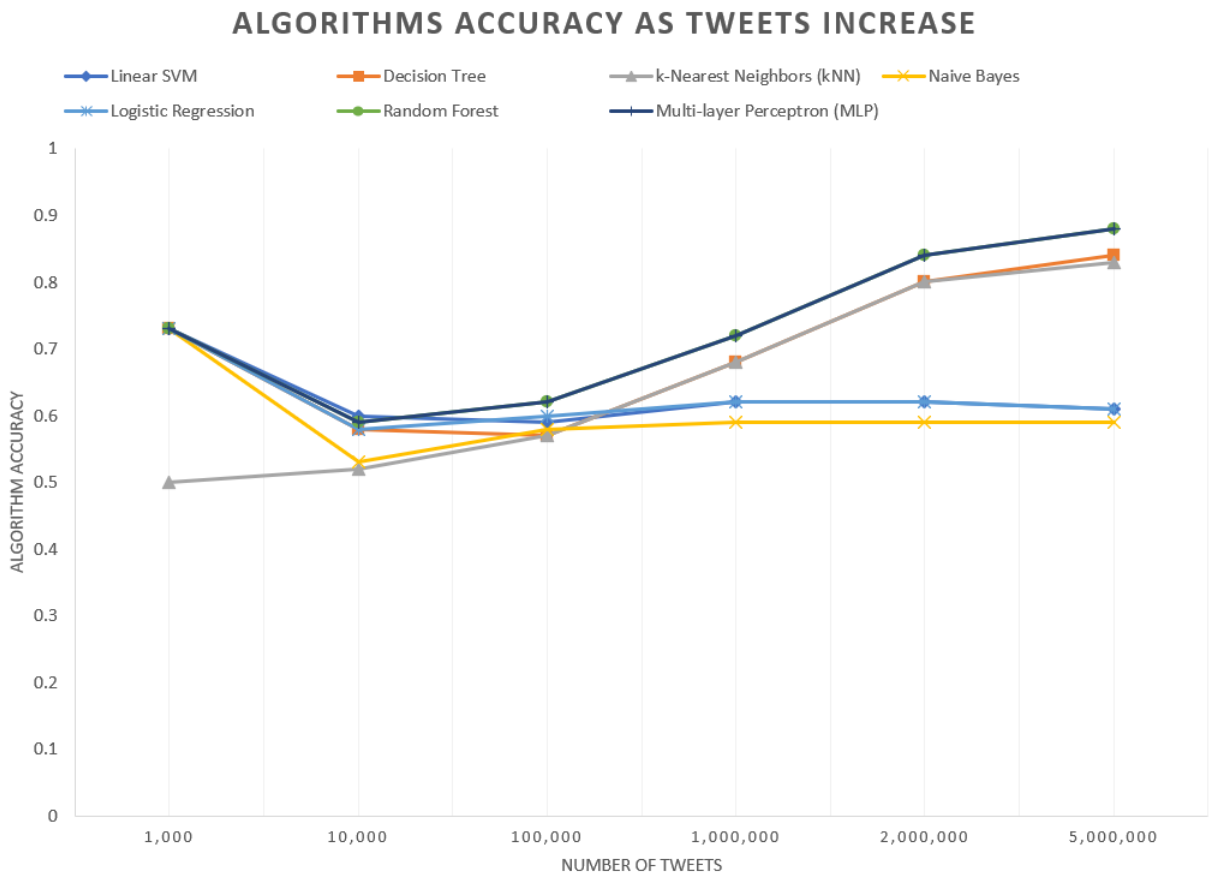


Σχήμα 6.1: Χρόνοι εκτέλεσης αλγορίθμων

Σωστή πρόβλεψη θεωρούμε το αν ένας αλγόριθμος χαρακτήρισε ένα tweet με την ίδια ετικέτα που είχε χαρακτηριστεί από εμάς είτε αυτή ήταν «αληθές» είτε «ψευδές». Για παράδειγμα, εάν το 70% των προβλέψεων είναι σωστές, τότε η ακρίβεια είναι 70%.

Στην εργασία αρχικά χρησιμοποιήθηκε η ακρίβεια (Accuracy) για την αξιολόγηση του μοντέλου και λήφθηκαν τα αποτελέσματα που φαίνονται στο Σχήμα 6.2. Παρατηρείται μια μικρή πτώση της ακρίβειας από τα 1.000 στα 10.000 tweets. Αυτό συμβαίνει διότι στα 1.000 tweets οι αλγόριθμοι δεν έχουν ακόμα εκπαιδευτεί με τον ίδιο αριθμό tweets με ετικέτα «αληθές» και «ψευδές», διότι δεν υπήρχαν αρκετά αληθή tweets. Αλλά, έχουν εκπαιδευτεί με δεδομένα εκπαίδευσης με περισσότερα ψευδή tweets επομένως, επειδή στα δεδομένα δοκιμής υπάρχουν περισσότερα tweets με ετικέτα «ψευδές» έχουν μεγαλύτερη πιθανότητα να χαρακτηριστούν από τους αλγορίθμους ως ψευδή και έτσι η ακρίβεια στα 1.000 tweets είναι πιο υψηλή, όχι όμως αντιπροσωπευτική ακόμα για το μοντέλο. Για το λόγο αυτό συγκρίνεται η ακρίβεια καθώς τα tweets αυξάνονται.

Από το Σχήμα 6.2, προκύπτει ότι την καλύτερη Ακρίβεια (Accuracy), για μεγάλου όγκου δεδομένα (5.000.000 tweets), επιτυγχάνουν οι αλγόριθμοι Multi-layer Perceptron (MLP) με ακρίβεια 88% και Random Forest επίσης με 88%. Αμέσως μετά, με λίγο χαμηλότερη ακρίβεια, είναι οι αλγόριθμοι Decision Tree με 84% και k-Nearest Neighbors (kNN) με 83%. Έπειτα, με πολύ χαμηλότερη ακρίβεια, βρίσκονται οι Logistic Regression με 61%



Σχήμα 6.2: Accuracy αλγορίθμων

και LinearSVM επίσης με 61%. Τέλος, όχι με πολύ χαμηλότερη ακρίβεια από τους δύο προηγούμενους, είναι ο αλγόριθμος Naive Bayes με 59%.

Όμως, η ακρίβεια είναι μια χρήσιμη μέτρηση μόνο όταν υπάρχει ίση κατανομή κλάσεων, ή αλλιώς ετικετών, σε μια ταξινόμηση. Αυτό σημαίνει ότι εάν παρατηρούνται περισσότερα δεδομένα μιας ετικέτας παρά μιας άλλης, η ακρίβεια δεν είναι πλέον χρήσιμη μέτρηση. Στην δική μας περίπτωση αυτό ισχύει γιατί περισσότερα από 90% των tweets αποτελούν «ψευδή» είδηση σε αντίθεση με εκείνα που αφορούν μια «αληθή» είδηση. Γι' αυτόν τον λόγο λοιπόν, θα αναζητήσουμε και άλλες μετρικές αξιολόγησης για το μοντέλο μας.

6.2.2 Αναζήτηση καλύτερων μετρικών αξιολόγησης

Ένας τρόπος επίλυσης προβλημάτων ανισορροπίας τάξης, είναι η χρήση καλύτερων μετρήσεων ακρίβειας οι οποίες λαμβάνουν υπόψη όχι μόνο τον αριθμό των σφαλμάτων πρόβλεψης που κάνει ένα μοντέλο, αλλά και τον τύπο των σφαλμάτων που γίνονται. Η ακρίβεια (Precision) και η ανάκληση (Recall) είναι οι δύο πιο κοινές μετρικές που λαμβάνουν υπόψη την ανισορροπία τάξης.

Ακρίβεια (Precision)

Θεωρώντας ότι ένα μοντέλο κατηγοριοποιεί τα δεδομένα σαν θετικές (POSITIVES) και αρνητικές (NEGATIVES) προβλέψεις, όταν θα κάνει μία πρόβλεψη σωστά, αυτή ονομάζεται σωστή (TRUE) ενώ όταν την κάνει λάθος ονομάζεται λανθασμένων (FALSE).

Η ακρίβεια είναι η αναλογία μεταξύ των σωστών θετικών προβλέψεων και όλων των θετικών προβλέψεων (σωστών και λανθασμένων) ενός μοντέλου. Μαθηματικά, υπολογίζεται ως εξής:

Precision = αριθμός σωστών θετικών προβλέψεων / (αριθμός σωστών θετικών προβλέψεων + αριθμός λανθασμένων θετικών προβλέψεων)

Ένα μη ακριβές μοντέλο μπορεί να βρει πολλές από τις θετικές περιπτώσεις (POSITIVES), αλλά επίσης ανιχνεύει και πολλές λανθασμένες θετικές περιπτώσεις (FALSE POSITIVES). Ενώ, ένα ακριβές μοντέλο ίσως δεν βρίσκει όλες τις θετικές (POSITIVES) περιπτώσεις, αλλά αυτές που το μοντέλο κατατάσσει ως θετικές (POSITIVES) είναι πολύ πιθανό να είναι σωστές (TRUE POSITIVES).

Ανάκληση (Recall)

Η ανάκληση είναι μια μετρική μας που προσδιορίζει σωστά τις αληθινές θετικές προβλέψεις ενός μοντέλου. Μαθηματικά, υπολογίζεται ως εξής:

Recall = αριθμός σωστών θετικών προβλέψεων / (αριθμός σωστών θετικών προβλέψεων + αριθμός λανθασμένων αρνητικών προβλέψεων)

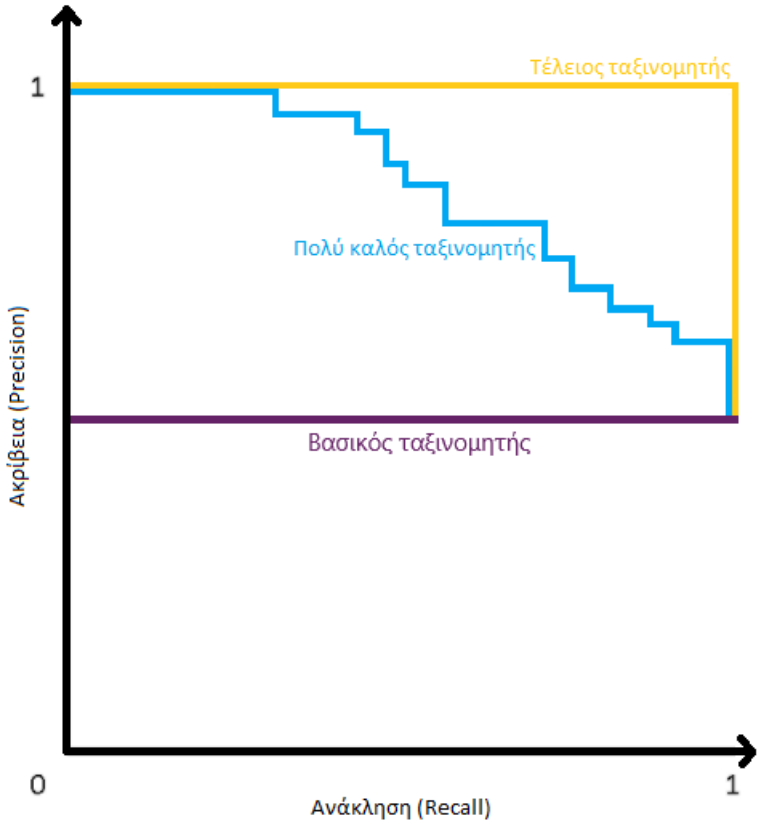
Ένα μοντέλο με υψηλή ανάκληση καταφέρνει καλά να βρει όλες τις θετικές περιπτώσεις (POSITIVES) στα δεδομένα, παρόλο που μπορεί επίσης να προσδιορίσει εσφαλμένα ορισμένες αρνητικές περιπτώσεις ως θετικές περιπτώσεις (FALSE POSITIVES). Ενώ, ένα μοντέλο με χαμηλή ανάκληση δεν είναι σε θέση να βρει όλες (ή ένα μεγάλο μέρος) από τις θετικές περιπτώσεις (POSITIVES) στα δεδομένα.

Καμπύλη Ακριβείας-Ανάκλησης

Ο συνδυασμός των Precision και Recall είναι μια χρήσιμη μετρική της επιτυχίας της πρόβλεψης όταν οι κλάσεις, «αληθής είδηση» (POSITIVES) ή «ψευδής είδηση» (NEGATIVES) είναι πολύ ανισόρροπες. Στην ανάκτηση πληροφοριών, η ακρίβεια είναι ένα μέτρο της συνάφειας των αποτελεσμάτων, ενώ η ανάκληση είναι ένα μέτρο του πόσα πραγματικά σχετικά αποτελέσματα επιστρέφονται.

Η καμπύλη ακριβείας-ανάκλησης δείχνει την αντιστάθμιση μεταξύ ακρίβειας και ανάκλησης. Μια υψηλή περιοχή κάτω από την καμπύλη αντιπροσωπεύει τόσο υψηλή ανάκληση όσο και υψηλή ακρίβεια, όπου η υψηλή ακρίβεια σχετίζεται με χαμηλό ποσοστό λανθασμένων θετικών περιπτώσεων και η υψηλή ανάκληση σχετίζεται με χαμηλό ποσοστό λανθασμένων αρνητικών περιπτώσεων. Οι υψηλές βαθμολογίες και για τα δύο δείχνουν ότι ο ταξινομητής επιστρέφει ακριβή αποτελέσματα (υψηλή ακρίβεια), καθώς και την πλειονότητα όλων των θετικών αποτελεσμάτων (υψηλή ανάκληση).

Ένα σύστημα με υψηλή ανάκληση αλλά χαμηλή ακρίβεια επιστρέφει πολλά αποτελέσματα, αλλά οι περισσότερες από τις προβλεπόμενες ετικέτες του είναι εσφαλμένες σε σύγκριση με τις ετικέτες εκπαίδευσης. Ένα σύστημα με υψηλή ακρίβεια αλλά χαμηλή ανάκληση είναι ακριβώς το αντίθετο, επιστρέφοντας πολύ λίγα αποτελέσματα, αλλά οι περισσότερες από τις προβλεπόμενες ετικέτες του είναι σωστές σε σύγκριση με τις ετικέτες εκπαίδευσης. Ένα ιδανικό σύστημα με υψηλή ακρίβεια και υψηλή ανάκληση θα επιστρέψει πολλά αποτελέσματα, με όλα τα αποτελέσματα να επισημαίνονται σωστά.

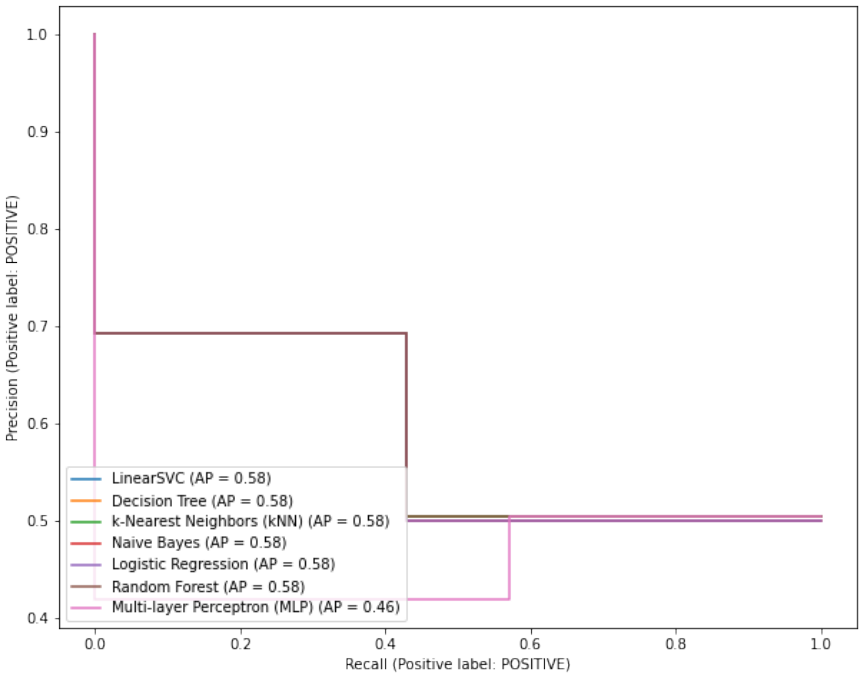


Σχήμα 6.3: Παράδειγμα καμπύλης Ακριβείας-Ανάκλησης

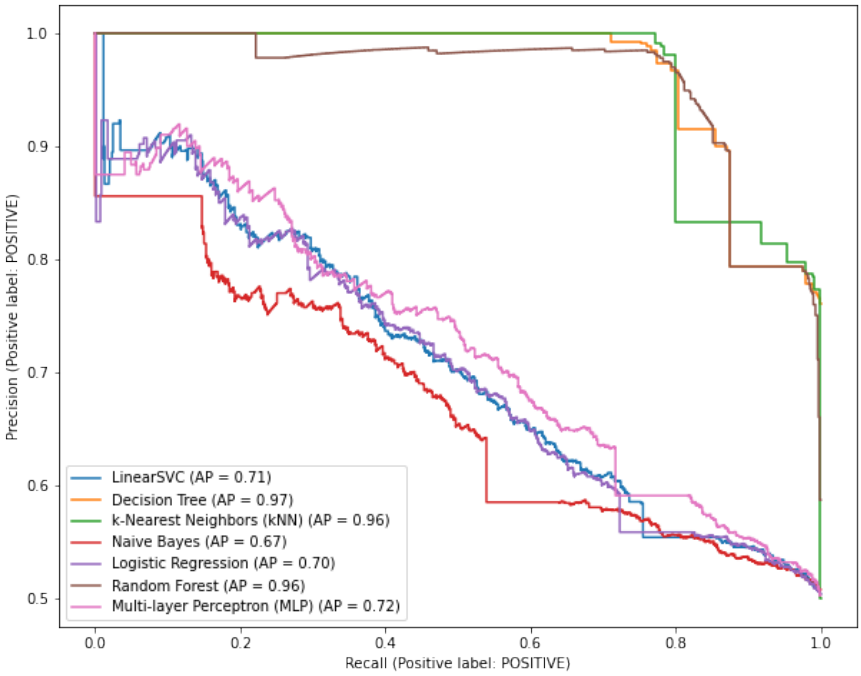
Στην πραγματική ζωή, δυστυχώς, έχουμε να αντιμετωπίσουμε το λεγόμενο Precision-Recall Trade-Off το οποίο αντιπροσωπεύει το γεγονός ότι σε πολλές περιπτώσεις, μπορείτε να τροποποιήσετε ένα μοντέλο για να αυξήσετε την ακρίβεια με κόστος χαμηλότερης ανάκλησης ή, από την άλλη πλευρά, να αυξήσετε την ανάκληση με κόστος χαμηλότερης ακρίβειας [Kor21].

Από την καμπύλη αυτών των δύο μετρικών μπορούμε εύκολα να καταλάβουμε κατά πόσο ένα μοντέλο πλησιάζει το ιδανικό. Στο Σχήμα 6.3 μπορούμε να δούμε την μορφή ενός τέλειου ταξινομητή (Classifier) καθώς και ενός καλού και ενός μέτριου. Παρατηρούμε ότι, έχουμε τέλειο ταξινομητή καθώς οι τιμές των μετρικών πλησιάζουν το 1.

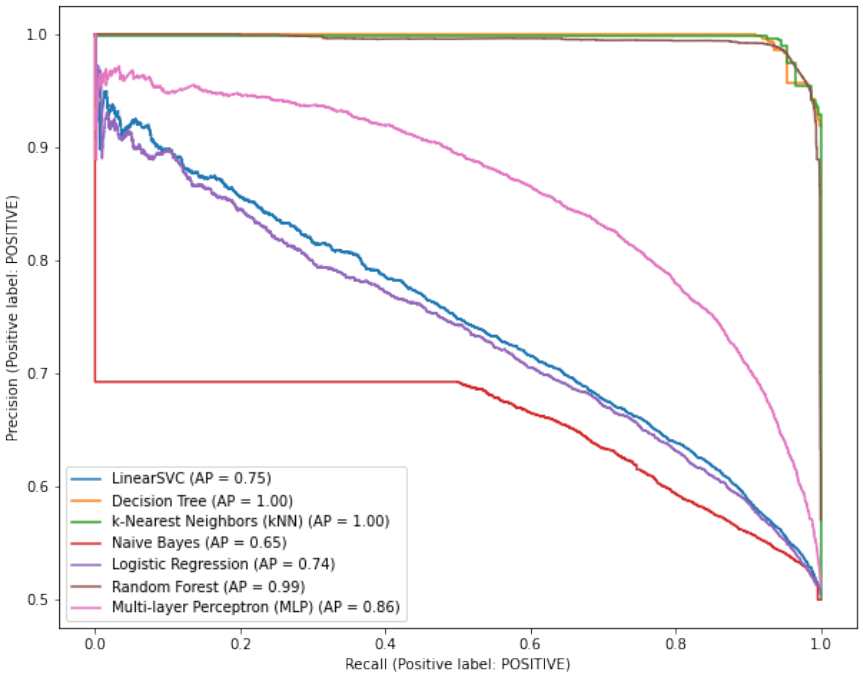
Για την περαιτέρω αξιολόγηση του μοντέλου ταξινόμησης που υλοποιήθηκε, χρησιμοποιήθηκε αυτή η καμπύλη και λήφθηκαν τα αποτελέσματα που φαίνονται στο Σχήμα 6.4. Στο οποίο παρατηρούμε ότι όλοι οι ταξινομητές μαθαίνουν όλο και περισσότερο με την αύξηση των tweets με αποτέλεσμα να έχουν όλο και υψηλότερα γραφήματα πλησιάζοντας το ιδανικό γράφημα στο σημείο (1,1), όπου η ακρίβεια (Precision) και η ανάκληση (Recall) είναι μέγιστες. Οι καλύτεροι ταξινομητές για το Σχήμα 6.4(ε) (με τα περισσότερα tweets) είναι ο Decision Tree, ο Random Forest και ο k-Nearest Neighboors (kNN) ακολουθούμενοι από τον Multilayer Perceptron (MLP). Έπειτα, θέση έχουν ο LinearSVC και ο Logistic Regression και στην τελευταία θέση βρίσκεται ο Naive Bayes.



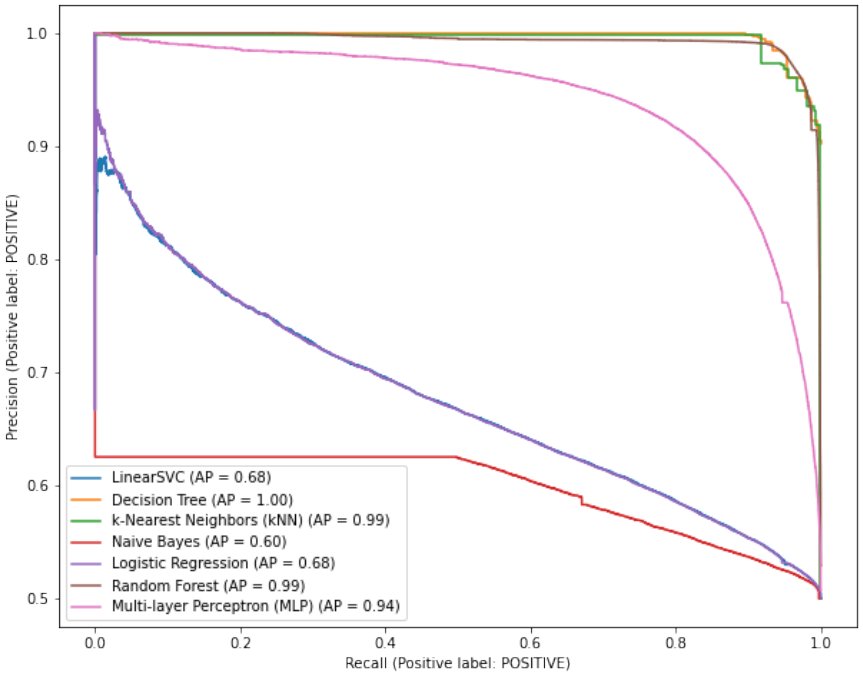
(α) 1.000 tweets



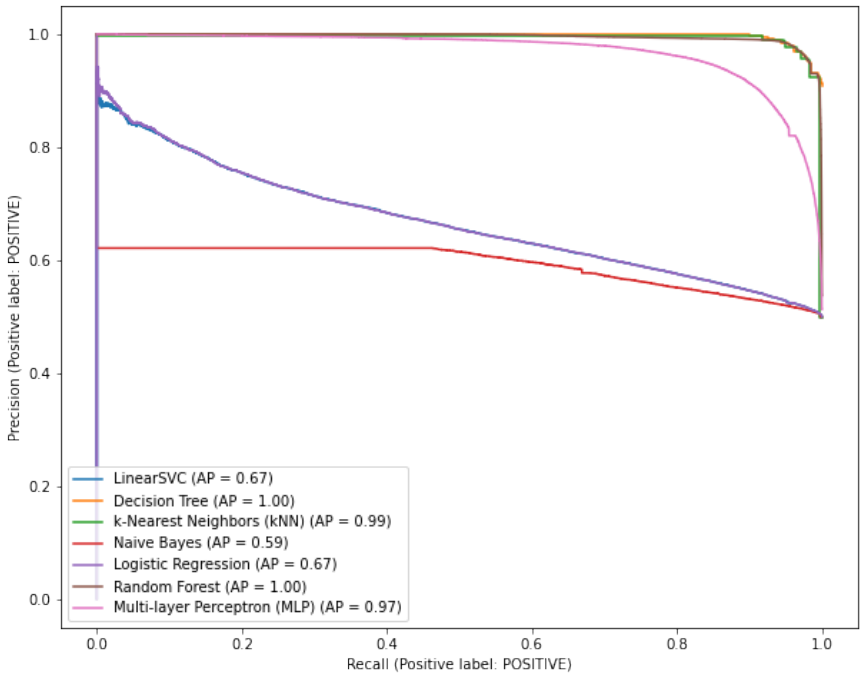
(β) 10.000 tweets



(γ) 100.000 tweets



(δ) 1.000.000 tweets



(ε) 2.000.000 tweets

Σχήμα 6.4: Καμπύλες Ακριβείας-Ανάκλησης καθώς αυξάνονται τα tweets

6.2.3 Συνδυάζοντας την Ακρίβεια και την Ανάκληση με το F1-Σκορ

Σε ορισμένες περιπτώσεις, μπορεί να θέλουμε να μεγιστοποιήσουμε είτε την ανάκληση είτε την ακρίβεια σε βάρος της άλλης μετρικής. Υπάρχουν όμως και πολλές περιπτώσεις όπου τόσο η ακρίβεια όσο και η ανάκληση είναι εξίσου σημαντικές. Σε τέτοιες περιπτώσεις, χρησιμοποιούμε μια μετρική που ονομάζεται F1-score (F1-Σκορ). Το F1-Score αποτελείται από την ακρίβεια και την ανάκληση και ορίζεται ως ο αρμονικός μέσος (Harmonic Mean) των δύο αυτών.

Ο αρμονικός μέσος όρος είναι ένα από τα πολλά είδη μέσου όρου, και συγκεκριμένα, ένα από τους Πυθαγόρειους μέσους. Μερικές φορές είναι κατάλληλο για καταστάσεις όπου ο μέσος ρυθμός είναι επιθυμητός. Ο αρμονικός μέσος ορίζεται ως εξής:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \left(\frac{\sum_{i=1}^{n} x_i^{-1}}{n} \right)^{-1}$$

Στο F1-Score, υπολογίζουμε το μέσο όρο της ακρίβειας (Precision) και της ανάκλησης (Recall). Επειδή όμως είναι και οι δύο ρυθμοί, χρησιμοποιείται ο αρμονικός μέσος όρος. Ο τύπος του F1-Score είναι ο εξής:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Εφόσον το F1-Score είναι ένας μέσος όρος των Precision and Recall, αυτό σημαίνει ότι το F1-Score δίνει ίση βαρύτητα στην Precision and Recall. Έτσι, προκύπτει ότι:

- Ένα μοντέλο θα έχει υψηλό F1-Score εάν τόσο η Ακρίβεια όσο και η Ανάκληση είναι υψηλές.
- Ένα μοντέλο θα έχει χαμηλό F1-Score εάν τόσο η Ακρίβεια όσο και η Ανάκληση είναι χαμηλές.
- Ένα μοντέλο θα έχει μεσαίο F1-Score εάν ένα από τα Precision και Recall είναι χαμηλό και το άλλο υψηλό.

Το F1-Score συνδυάζει την ακρίβεια και την ανάκληση σε μια ενιαία μετρική. Σε πολλές περιπτώσεις, όπως στο αυτοματοποιημένο benchmarking ή στην αναζήτηση πλέγματος (Grid Search), είναι πολύ πιο βολικό να υπάρχει μόνο μία μέτρηση απόδοσης και όχι πολλαπλές διότι, σε αυτές τις περιπτώσεις, θα πρέπει να οριστεί μια μεμονωμένη μετρική προς βελτιστοποίηση.

Συμπερασματικά, όταν υπάρχει η δυνατότητα εξέτασης πολλαπλών μετρικών για το μοντέλο, θα πρέπει οπωσδήποτε αυτή να γίνει. Κάθε μετρική έχει ορισμένα πλεονεκτήματα και μειονεκτήματα και καθένα από αυτά δίνει συγκεκριμένες πληροφορίες για τα δυνατά και τα αδύνατα σημεία ενός μοντέλου.

6.3 Πίνακας Σύγχυσης

Έχοντας συζητήσει για τις δημοφιλέστερες τεχνικές Μηχανικής Μάθησης δημιουργείται η ανάγκη εύρεσης ενός τρόπου αναπαράστασής τους. Αυτή ακριβώς την ανάγκη ικανοποιεί ο Πίνακας Σύγχυσης (Confusion Matrix), ο οποίος αποτελεί μια τεχνική για τη σύνοψη της απόδοσης ενός αλγορίθμου ταξινόμησης. Ο υπολογισμός ενός πίνακα σύγχυσης δίνει μια καλύτερη ιδέα για το τι «πετυχαίνει» σωστά το υλοποιημένο μοντέλο ταξινόμησης και τι είδους σφάλματα κάνει.

Πιο συγκεκριμένα, ένας πίνακας σύγχυσης είναι μια σύνοψη των αποτελεσμάτων πρόβλεψης σε ένα πρόβλημα ταξινόμησης. Ο αριθμός των σωστών και εσφαλμένων προβλέψεων συνοψίζεται με τιμές μετρικών και αναλύεται ανά ετικέτα δεδομένων. Ο πίνακας σύγχυσης δείχνει τους τρόπους με τους οποίους ένα μοντέλο ταξινόμησης πιθανώς να μπερδεύεται όταν κάνει προβλέψεις. Δίνει επίσης πληροφορίες όχι μόνο για τα σφάλματα που γίνονται από έναν ταξινομητή, αλλά και για τους τύπους των σφαλμάτων που γίνονται. Αυτή η ανάλυση είναι που ξεπερνά τον περιορισμό της χρήσης μόνο της ακρίβειας (accuracy) ταξινόμησης.

Στην εργασία αυτή χρησιμοποιήθηκε ο Πίνακας Σύγχυσης που παρεχόταν από το Scikit-learn [Ped+11] για 5.000.000 tweets ο οποίος φαίνεται στον Πίνακα 6.1. Στον πίνακα, φαίνονται μερικές μετρικές για τις οποίες δεν έχουμε μιλήσει καθώς επίσης και οι τιμές τους για κάθε κλάση-ετικέτα (POSITIVE και NEGATIVE) ξεχωριστά. Μια από αυτές, είναι μετρική Support η οποία είναι ο αριθμός των πραγματικών εμφανίσεων της αντίστοιχης κλάσης στο καθορισμένο σύνολο δεδομένων. Αν το Support των δεδομένων εκπαίδευσης δεν είναι ισορροπημένο μπορεί να υποδηλώνει λάθη στις αναφερόμενες βαθμολογίες του ταξινομητή. Το Support δεν αλλάζει μεταξύ των μοντέλων και αναλύει τη διαδικασία αξιολόγησης. Επίσης, το macro average, είναι ο απλός μέσος όρος των βαθμολογιών όλων των κλάσεων (POSITIVE και NEGATIVE). Συνεπώς, το macro average του Recall θα είναι ο μέσος όρος των Recall των κλάσεων POSITIVE και NEGATIVE. Τέλος, το weighted average είναι το άθροισμα των βαθμολογιών όλων των κλάσεων. αλλά η κάθε βαθμολογία της κλάσης πολλαπλασιάζεται με το αντίστοιχο ποσοστό της κλάσης αυτής. Παρατηρείστε ότι, τα macro average και weighted average είναι σχεδόν παντού ίδια, αυτό συμβαίνει διότι οι δύο κλάσεις έχουν ίσο αριθμό από δεδομένα (όπως φαίνεται από την μετρική Support).

Επιπροσθέτως, στον Πίνακα 6.1 διακρίνουμε τις τιμές του F1-Score για κάθε αλγόριθμο. Προκύπτει λοιπόν ότι, το καλύτερο F1-Score, για μεγάλου όγκου δεδομένα (5.000.000 tweets), επιτυγχάνουν οι αλγόριθμοι

Linear SVM				
	Precision	Recall	F1-Score	Support
NEGATIVE	0.61	0.64	0.63	92940
POSITIVE	0.62	0.58	0.6	92940
accuracy			0.61	185880
macro avg	0.61	0.61	0.61	185880
weighted avg	0.61	0.61	0.61	185880

Decision Tree				
	Precision	Recall	F1-Score	Support
NEGATIVE	0.86	0.82	0.84	92940
POSITIVE	0.83	0.86	0.85	92940
accuracy			0.84	185880
macro avg	0.84	0.84	0.84	185880
weighted avg	0.84	0.84	0.84	185880

k-Nearest Neighbors (kNN)				
	Precision	Recall	F1-Score	Support
NEGATIVE	0.83	0.83	0.83	92940
POSITIVE	0.83	0.83	0.83	92940
accuracy			0.83	185880
macro avg	0.83	0.83	0.83	185880
weighted avg	0.83	0.83	0.83	185880

Naive Bayes				
	Precision	Recall	F1-Score	Support
NEGATIVE	0.57	0.7	0.63	92940
POSITIVE	0.61	0.47	0.63	92940
accuracy			0.59	185880
macro avg	0.59	0.59	0.58	185880
weighted avg	0.59	0.59	0.58	185880

Logistic Regression				
	Precision	Recall	F1-Score	Support
NEGATIVE	0.61	0.64	0.63	92940
POSITIVE	0.62	0.58	0.6	92940
accuracy			0.61	185880
macro avg	0.61	0.61	0.61	185880
weighted avg	0.61	0.61	0.61	185880

Random Forest				
	Precision	Recall	F1-Score	Support
NEGATIVE	0.85	0.92	0.88	92940
POSITIVE	0.91	0.84	0.87	92940
accuracy			0.88	185880
macro avg	0.88	0.88	0.88	185880
weighted avg	0.88	0.88	0.88	185880

Multi-layer Perceptron (MLP)				
	Precision	Recall	F1-Score	Support
NEGATIVE	0.85	0.92	0.88	92940
POSITIVE	0.91	0.84	0.87	92940
accuracy			0.88	185880
macro avg	0.88	0.88	0.88	185880
weighted avg	0.88	0.88	0.88	185880

Πίνακας 6.1: Πίνακας Σύγχυσης (Confusion Matrix) του μοντέλου ταξινόμησης για 5.000.000 tweets

Multi-layer Perceptron (MLP) με 88% και Random Forest επίσης με 88%. Αμέσως μετά, με λίγο χαμηλότερο F1-Score, είναι οι αλγόριθμοι Decision Tree με 84% και k-Nearest Neighbors (kNN) με 83%. Έπειτα, με πολύ χαμηλότερο F1-Score, βρίσκονται οι Logistic Regression με 61% και LinearSVM επίσης με 61%. Τέλος, όχι με πολύ χαμηλότερο F1-Score από τους δύο προηγούμενους, είναι ο αλγόριθμος Naive Bayes με 58%.

6.4 Σύνοψη αποτελεσμάτων

Έχοντας λοιπόν αναλύσει όλα τα αποτελέσματα των μετρικών που χρησιμοποιήθηκαν για την αξιολόγηση του μοντέλου που υλοποιήθηκε θα κάνουμε μία σύνοψη αυτών σε αυτήν την Ενότητα. Σύμφωνα λοιπόν με τον Πίνακα 6.1 και τα Σχήματα 6.1, 6.2 και 6.4, θα συγκεντρώσουμε τα αποτελέσματα πιο συνοπτικά στον Πίνακα 6.2. Σε αυτόν, μπορεί να διακριθεί ο καλύτερος αλγόριθμος με βάση τις προαναφερθείσες μετρικές. Συγκεκριμένα, αν λάβουμε υπόψιν τον χρόνο εκτέλεσης των αλγορίθμων και ταυτόχρονα τις επιδόσεις στις μετρικές τότε την καλύτερη ισορροπία μεταξύ αυτών των δύο την προσφέρει ο αλγόριθμος Multi-layer Perceptron (MLP) από την οικογένεια τεχνητών νευρωνικών δικτύων ο οποίος επιτυγχάνει πολύ καλές επιδόσεις σε ένα αρκετά ικανοποιητικό χρονικό διάστημα.

Ταξινομητές	Χρόνος εκτέλεσης (5.000.000 tweets)	Acurracy (5.000.000 tweets)	Σκορ καμπύλης Precision-Recall (2.000.000 tweets)	F1-Score (5.000.000 tweets)
LinearSVM	0.3 λεπτά	61%	67%	61%
Decision Tree	1.5 ώρες	84%	100%	84%
k-Nearest Neighbors (kNN)	1.63 ώρες	83%	99%	83%
Naive Bayes	0.2 λεπτά	59%	59%	58%
Logistic Regression	4.2 λεπτά	61%	67%	61%
Random Forest	22.3 λεπτά	88%	100%	88%
Multi-layer Perceptron (MLP)	3.7 λεπτά	88%	97%	88%

Πίνακας 6.2: Σύνοψη επιδόσεων αλγορίθμων

Κεφάλαιο 7

Περαιτέρω βελτιώσεις και τροποποιήσεις του μοντέλου

Η εργασία αυτή συγκροτήθηκε με μεγάλη προσπάθεια και περιλάμβανε πολλές δυσκολίες οι οποίες αντιμετωπίστηκαν. Μερικές από αυτές ήταν, η δωρεάν συλλογή τεράστιου όγκου δεδομένων σε περιορισμένο χρόνο, η αντικειμενική εκπαίδευση των αλγορίθμων Μηχανικής Μάθησης με πάρα πολλά tweets (20% του συνόλου), η αναζήτηση και κατανόηση πολλών μεταβλητών των αλγορίθμων για την παραμετροποίηση του Grid Search, αλλά και η εύρεση του σωστού τρόπου εκπαίδευσης των αλγορίθμων με στόχο τις βέλτιστες επιδόσεις του μοντέλου.

Ένα πρόβλημα, από τα χαρακτηριστικά προβλήματα της Μηχανικής Μάθησης, το οποίο κατέστη μη επιλύσιμο (ολοκληρωτικά) ήταν η αδυναμία επεξεργασίας αυτού του τεράστιου όγκου δεδομένων στο σύνολό του. Η εργασία υλοποιήθηκε σε προσωπικό υπολογιστή με υψηλές επιδόσεις, ταχεία 32GB RAM και ταχύς 8-πύρηνος επεξεργαστής, οι οποίες όμως δεν ήταν επαρκείς για αυτήν την επεξεργασία. Μια λύση στο πρόβλημα αυτό, θα μπορούσε να αποτελέσει η ενοικίαση στο cloud μιας πλήρης σουίτας υπολογιστών υψηλής απόδοσης (HPC) σε κάποια εταιρία πιθανώς Microsoft, Amazon, Google προκειμένου να δημιουργηθεί το μοντέλο Μηχανικής Μάθησης με όλα τα δεδομένα που συλλέχθηκαν και πιθανώς ακόμα περισσότερα. Κατά αυτόν τον τρόπο θα πετυχαίναμε ακόμα καλύτερες επιδόσεις στο μοντέλο.

Μια επιπρόσθετη βελτίωση, θα μπορούσε να αποτελέσει η προσθήκη ακόμα περισσότερων αλγορίθμων που επιλύουν το Classification πρόβλημα στο μοντέλο. Κατά αυτόν τον τρόπο, πιθανώς να εντοπιζόταν ακόμα καλύτερος αλγόριθμος σε θέματα επιδόσεων. Μερικοί Classification αλγόριθμοι που δεν καλύφθηκαν σε αυτήν την εργασία και θα μπορούσαν να μελετηθούν είναι οι εξής: Stochastic Gradient Descent, Fisher's linear discriminant, Quadratic classifier, Learning vector quantization.

Βιβλιογραφία

[Bec21] Martin Beck. «How to scrape tweets from Twitter». Στο: *Medium* (Μάι. 2021). URL: https://towardsdatascience.com/how-to-scrape-tweets-from-twitter-59287e20f0f1.

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Επιμέλεια υπό M Jordan, J Kleinberg και B Schölkopf. 1η έκδοση. Springer, 2006, σ. 738. ISBN: 9780387310732. URL: https://link.springer.com/book/9780387310732.

[Bon21] Harika Bonthu. «Python tutorial: Working with CSV file for Data Science». Στο: *Analytics Vidhya* (Αύγ. 2021). URL: https://www.analyticsvidhya.com/blog/2021/08/python-tutorial-working-with-csv-file-for-data-science/.

[Bro20] Jason Brownlee. «A tour of machine learning algorithms». Στο: *Machine Learning Mastery* (Αύγ. 2020). URL: https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/.

[Bro21] Jason Brownlee. «Failure of classification accuracy for imbalanced class distributions». Στο: *Machine Learning Mastery* (Ιαν. 2021). URL: https://machinelearningmastery.com/failure-of-accuracy-for-imbalanced-class-distributions/.

[Bur21] Ed Burns. «What is machine learning and why is it important?» Στο: *SearchEnterpriseAI* (Μαρ. 2021). URL: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML.

[Don+20] Irvin Dongo, San Pablo, Yudith Cadinale, Ana Aguilera, Fabiola Martínez, Yuni Quintero και Sergio Barrios. «Web scraping versus Twitter API: Proceedings of the 22nd International Conference on Information Integration and web-based Applications & Services». Στο: *ACM Other conferences* (Νοέ. 2020). URL: https://dl.acm.org/doi/10.1145/3428757.3429104.

[Fuc21] Michael Fuchs. «NN - Multi-layer Perceptron Classifier (MLPClassifier) - Michael Fuchs Python». Στο: *MFuchs* (Φεβ. 2021). URL: https://michael-fuchs-python.netlify.app/2021/02/03/nn-multi-layer-perceptron-classifier-mlpclassifier/.

[Gil22] Navdeep Singh Gill. «Artificial Neural Networks Applications and algorithms». Στο: *XenonStack* (Μάι. 2022). URL: https://www.xenonstack.com/blog/artificial-neural-network-applications.

[Hei18] Hunter Heidenreich. «What are the types of machine learning?» Στο: *Medium* (Δεκ. 2018). URL: https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f.

[Ken21] Colm Kenny. «What is web scraping and how does it work?» Στο: *Zyte (formerly Scrapinghub) #1 Web Scraping Service* (Ιαν. 2021). URL: https://www.zyte.com/learn/what-is-web-scraping/.

[Kor21] Joos Korstanje. «The F1 score». Στο: *Medium* (Αύγ. 2021). URL: https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.

[Mal19] Usman Malik. «Text classification with Python and Scikit-Learn». Στο: *Stack Abuse* (Φεβ. 2019). URL: https://stackabuse.com/text-classification-with-python-and-scikit-learn/.

[Mal20] Farhad Malik. «What is grid search?» Στο: *Medium* (Φεβ. 2020). URL: https://medium.com/fintechexplained/what-is-grid-search-c01fe886ef0a.

[MRT18] M. Mohri, A. Rostamizadeh και A. Talwalkar. *Foundations of machine learning*. 2η έκδοση. Cambridge, Massachusetts: MIT Press, 2018. ISBN: 978-0-262-03940-6. URL: https://mitpress.mit.edu/books/foundations-machine-learning-second-edition.

[Nel19] Dan Nelson. «Overview of classification methods in python with scikit-learn». Στο: *Stack Abuse* (Μάι. 2019). URL: https://stackabuse.com/overview-of-classification-methods-in-python-with-scikit-learn/.

[Ped+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot και E. Duchesnay. «Scikit-learn: Machine Learning in Python». Στο: *Journal of Machine Learning Research* 12 (2011), σσ. 2825–2830. URL: https://scikit-learn.org/stable/.

[Pho22] James Phoenix. «How to scrape Twitter data». Στο: *Just Understanding Data* (Φεβ. 2022). URL: https://understandingdata.com/how-to-scrape-twitter-data/.

[Ran19] Vijaya Rani. «NLP tutorial for text classification in Python». Στο: *Medium* (Απρ. 2019). URL: https://medium.com/analytics-vidhya/nlp-tutorial-for-text-classification-in-python-8f19cd17b49e.

[RN05] Stuart Russell και Peter Norvig. *Τεχνητή νοημοσύνη: Μια σύγχρονη προσέγγιση*. 2η έκδοση. Κλειδάριθμος, 2005, σ. 1200. ISBN: 960-209-873-2. URL: https://www.klidarithmos.gr/texnhth-nohmosynh-2h-ekdosh?search=960-209-873-2.

[RW06] CE. Rasmussen και CKI. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, Ιαν. 2006, σ. 248. ISBN: 026218253X. URL: https://mitpress.mit.edu/books/gaussian-processes-machine-learning.

[Sci22] Scikit-learn. *Tuning the hyper-parameters of an estimator*. 2022. URL: https://scikit-learn.org/stable/modules/grid_search.html.

[Sco21] William Scott. «TF-IDF for document ranking from scratch in Python on Real World Dataset.» Στο: *Medium* (Σεπτ. 2021). URL: https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089.

[Sun17] Ray Sunil. «Commonly used machine learning algorithms: Data science». Στο: *Analytics Vidhya* (Σεπτ. 2017). URL: https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/#h2_17.

[Too] Natural Language Toolkit. *NLTK Wordnet word lemmatizer - API & demo: Text analysis online*. URL: https://textanalysisonline.com/nltk-wordnet-word-lemmatizer.

[Wak22] Katrina Wakefield. «A guide to the types of machine learning algorithms». Στο: *A guide to the types of machine learning algorithms | SAS UK* (2022). URL: https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html.

[Wik21] Wikipedia. «Μηχανική μάθηση». Στο: (Αύγ. 2021). URL: https://el.wikipedia.org/wiki/%CE%9C%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE_%CE%BC%CE%AC%CE%B8%CE%B7%CF%83%CE%B7.

[Wol20] Rachel Wolff. «5 types of classification algorithms in machine learning». Στο: *MonkeyLearn Blog* (Αύγ. 2020). URL: https://monkeylearn.com/blog/classification-algorithms/.

[ZPL17] Cody Zacharias, Francesco Poldi και Maxim Levin. *TWINT project*. https://www.patreon.com/twintproject, Ιούν. 2017. URL: https://github.com/twintproject/twint.

[Βλα+20] Ιωάννης Βλαχάβας, Πέτρος Κεφαλάς, Νίκος Βασιλειάδης, Φώτης Κόκκορας και Ηλίας Σακελλαρίου. *Τεχνητή Νοημοσύνη*. 4η έκδοση. ΕΚΔΟΣΕΙΣ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΜΑΚΕΔΟΝΙΑΣ, 2020, σ. 1000. ISBN: 978-618-5196-44-8. URL: https://aibook.gr/.