ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ

ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΚΤΙΜΗΣΗ ΚΑΙ ΕΠΙΚΥΡΩΣΗ ΕΓΚΥΡΟΤΗΤΑΣ ΠΕΡΙΕΧΟΜΕΝΟΥ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΓΚΡΙΛΛΑΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

Κ. ΔΕΜΕΣΤΙΧΑΣ

ΕΠ. ΚΑΘΗΓΗΤΗΣ Γ.Π.Α ΚΑΙ ΕΠΙΣΤΗΜΟΝΙΚΟΣ ΣΥΝΕΡΓΑΤΗΣ Ε.Μ.Π

ΦΕΒΡΟΥΑΡΙΟΣ 2022

ΕΚΤΙΜΗΣΗ ΚΑΙ ΕΠΙΚΥΡΩΣΗ ΕΓΚΥΡΟΤΗΤΑΣ ΠΕΡΙΕΧΟΜΕΝΟΥ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ Μ/Υ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΝΑΥΤΙΛΙΑΣ ΚΑΙ ΒΙΟΜΗΧΑΝΙΑΣ

ΤΜΗΜΑΤΟΣ ΒΙΟΜΗΧΑΝΙΚΗΣ ΔΙΟΙΚΗΣΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΑΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΤΕΧΝΟ-ΟΙΚΟΝΟΜΙΚΑ ΣΥΣΤΗΜΑΤΑ»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΕΚΤΙΜΗΣΗ ΚΑΙ ΕΠΙΚΥΡΩΣΗ ΕΓΚΥΡΟΤΗΤΑΣ ΠΕΡΙΕΧΟΜΕΝΟΥ ΜΕ ΤΗ ΧΡΗΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΓΚΡΙΛΛΑΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

Εξεταστική Επιτροπή

Δεμέστιχας Κωνσταντίνος,
Επίκουρος Καθηγητής Γ.Π.Α και Επιστημονικός Συνεργάτης Ε.Μ.Π
(Επιβλέπων Καθηγητής)

Αδαμοπούλου Ευγενία,
Επιστημονική συνεργάτης Ε.Μ.Π

Σύκας Ευστάθιος,
Καθηγητής Ε.Μ.Π

Ρουσσάκη Ιωάννα,
Επίκουρη Καθηγήτρια Ε.Μ.Π

ΠΕΡΙΛΗΨΗ

Στη σύγχρονη εποχή, η ραγδαία εξέλιξη του διαδικτύου και η καθολική χρήση των πλατφορμών κοινωνικής δικτύωσης έχουν σηματοδοτήσει την έλευση μιας νέας εποχής στη διαθεσιμότητα και διασπορά των ειδήσεων σε παγκόσμιο επίπεδο. Προκειμένου να χαρακτηρίζεται η ενημέρωση από εγκυρότητα και αμεροληψία, απαραίτητη προϋπόθεση είναι όσοι παράγουν τις ειδήσεις να προασπίζονται την αλήθεια και να ελέγχουν την αξιοπιστία της πληροφορίας που δημιουργούν ή αναπαράγουν. Ωστόσο, οι διαρκείς εξελίξεις (τεχνολογικές, κοινωνικές, πολιτικές, κτλ.) και η δυνατότητα του καθενός να δημιουργεί ή να διασπείρει την πληροφορία, οδηγούν πολλές φορές στην απουσία ελέγχου της εγκυρότητάς της, γεγονός που πυροδοτεί την παραπληροφόρηση και τη διασπορά των ψευδών ειδήσεων («fake news»). Προκειμένου να αντιμετωπιστεί αποτελεσματικά το φαινόμενο της παραπληροφόρησης, καθίσταται αναγκαία η μετάβαση από την ανθρώπινη προσπάθεια στην άμεση ανίχνευση των ψευδών ειδήσεων μέσω αυτόματων συστημάτων, καθώς ο όγκος των δεδομένων τη σύγχρονη εποχή είναι τεράστιος και η ταχύτητα διάδοσής τους στο διαδίκτυο καθιστά αναποτελεσματικό τον εντοπισμό παραπληροφόρησης μέσω των παραδοσιακών δημοσιογραφικών τεχνικών. Η αυτοματοποίηση στην ανίχνευση των ψευδών ειδήσεων επιτυγχάνεται μέσω της Μηχανικής Μάθησης και της πληθώρας των μοντέλων που παρέχει, τα οποία μπορούν να εφαρμοστούν στο μεγάλο όγκο δεδομένων που είναι διαθέσιμα στο διαδίκτυο και μέσω αυτών να εξαχθεί η απαραίτητη γνώση που θα εξυπηρετήσει το σκοπό μας. Αυτό ακριβώς θα μελετηθεί μέσω της παρούσας διπλωματικής εργασίας, όπου αφού αναζητηθούν οι πιο σύγχρονες τεχνικές στη βιβλιογραφία, θα εφαρμοστούν διάφορα μοντέλα Μηχανικής Μάθησης σε δεδομένα που θα εξαχθούν από το διαδίκτυο και τα οποία θα αξιολογηθούν ως προς την απόδοσή τους. Μέσω των μετρικών αξιολόγησης, θα αναδειχθεί η αποδοτικότερη μέθοδος που θα δώσει λύση στο πρόβλημά μας, δηλαδή στην εύρεση και ανάπτυξη συστήματος για την εκτίμηση και επικύρωση διαδικτυακού περιεχομένου (online content verification).

ABSTRACT

In modern times, the rapid evolution of the internet and the universal use of social networking platforms have marked the advent of a new era in the availability and dispersion of news globally. In order for information to be characterized by validity and impartiality, it is essential that those who produce the news defend the truth and check the reliability of the information they create or reproduce. However, constant developments (technological, social, political, etc.) and the ability of each to create or disperse information, often lead to the absence of verification of its validity, which triggers misinformation and the spread of false news ("fake news"). In order to effectively deal with the phenomenon of misinformation, it becomes necessary to switch from human effort to direct detection of false news through automatic systems, since the volume of data nowadays is huge and the speed of its dissemination on the internet makes it ineffective to detect misinformation through traditional journalistic techniques. Automation in the detection of false news is achieved through Machine Learning and the multitude of models it provides, which can learn from the data and extract the necessary knowledge from them. This is exactly what will be studied through this thesis, where after searching for the most modern techniques in the literature, various models of Machine Learning will be applied on different datasets and will be evaluated on their performance. Through the evaluation metrics, the most efficient method will emerge that will provide a solution to the problem of online content verification and fake news detection.

Keywords: Fake news, Content verification, Machine Learning, Python, Data Mining.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΕΙΣΑΓΩΓΗ

Η έλευση του διαδικτύου και η ταχεία εδραίωση και χρήση των πλατφορμών κοινωνικής δικτύωσης (όπως το Facebook και το Twitter) έχουν συμβάλλει στην αδιάκοπη και απρόσκοπτη διάθεση πληροφοριών σε σχεδόν πραγματικό χρόνο. Πλέον, τα παραδοσιακά μέσα ειδήσεων, όπως οι εφημερίδες και τα περιοδικά σε έντυπη μορφή, έχουν αντικατασταθεί σε μεγάλο βαθμό από τα ψηφιακά μέσα, όπως για παράδειγμα τις διαδικτυακές ειδησεογραφικές πλατφόρμες, τα μέσα κοινωνικής δικτύωσης, τα ιστολόγια ή «blogs» και άλλες μορφές ψηφιακών μέσων.

Τα μέσα αυτά είναι ιδιαίτερα ισχυρά και χρήσιμα, καθώς επιτρέπουν στους χρήστες να βρίσκουν κάθε είδους πληροφορία και παράλληλα να μπορούν να εκφράζουν τη γνώμη τους και να μοιράζονται τις ιδέες τους πάνω σε συγκεκριμένα ζητήματα. Ωστόσο, μαζί με τα οφέλη, υποκρύπτονται και κίνδυνοι, καθώς πολλές φορές οι ειδήσεις που βρίσκουμε στο διαδίκτυο δεν έχουν ελεγχθεί ως προς την εγκυρότητά τους, εμπεριέχουν μεροληπτικές απόψεις ή προσπαθούν να χειραγωγήσουν, οδηγώντας με αυτόν τον τρόπο στην παραπληροφόρηση. Το φαινόμενο των ψευδών ειδήσεων («fake news») είναι ιδιαίτερα αισθητό στις μέρες μας και μελετάται εκτενώς, καθώς η ικανότητά μας να λαμβάνουμε αποφάσεις βασίζεται κυρίως στο είδος της πληροφορίας που καταναλώνουμε. Χαρακτηριστικό και επίκαιρο παράδειγμα είναι η περίπτωση του ιού Covid-19, όπου ψευδείς αναφορές γύρω από την προέλευση, τα χαρακτηριστικά και τη συμπεριφορά του ιού διαχέονται συνέχεια στο διαδίκτυο και κυρίως στα μέσα κοινωνικής δικτύωσης, διαμορφώνοντας με αυτόν τον τρόπο εσφαλμένες και μη επιστημονικά τεκμηριωμένες απόψεις [6].

Όλα τα παραπάνω κρίνουν απαραίτητη την ανάπτυξη συστημάτων που θα ανιχνεύουν τις ψευδείς ειδήσεις πριν τη διασπορά τους και θα αντικαταστήσουν την ανθρώπινη προσπάθεια, η οποία με τον σύγχρονο όγκο των διαθέσιμων πληροφοριών αποδεικνύεται ανεπαρκής. Άλλωστε, τα δεδομένα μεγάλου όγκου (big data) έχουν ήδη εισβάλλει και στο χώρο της δημοσιογραφίας, οδηγώντας την από την παραδοσιακή της μορφή στη δημοσιογραφία δεδομένων (data journalism).

Στόχος της παρούσας διπλωματικής εργασίας είναι να αναζητήσει και να εφαρμόσει state-of-the-art μεθόδους Μηχανικής Μάθησης σε δεδομένα που είναι δημόσια διαθέσιμα σε διαδικτυακούς τόπους και έχουν χρησιμοποιηθεί σε παρόμοιες έρευνες,

ώστε να δώσει λύση στο πρόβλημα της επαλήθευσης διαδικτυακού περιεχομένου και του εντοπισμού ψευδών ειδήσεων άμεσα και αποτελεσματικά.

Στο πρώτο κεφάλαιο, λοιπόν, αφού αναλυθούν ορισμένες βασικές έννοιες γύρω από την Εξόρυξη Γνώσης (Data Mining) και τη Μηχανική Μάθηση (Machine Learning), οι οποίες θα χρησιμοποιηθούν στη συνέχεια και θα πρέπει να γνωρίζει ο αναγνώστης, θα δοθούν οι λεπτομέρειες σχετικά με τα είδη και τους αλγορίθμους Μηχανικής Μάθησης καθώς και οι μετρικές που χρησιμοποιούνται για την αξιολόγηση των διαφορετικών μοντέλων.

Στο δεύτερο κεφάλαιο θα δοθούν οι βασικές έννοιες του προβλήματος της επικύρωσης της εγκυρότητας του διαδικτυακού περιεχομένου και θα μελετηθεί η σύγχρονη βιβλιογραφία και οι διαφορετικές μέθοδοι Μηχανικής Μάθησης που έχουν προταθεί για την επίλυση του προβλήματος.

Στη συνέχεια, στο τρίτο κεφάλαιο, θα γίνει η ανάλυση της μεθοδολογίας, όπου θα περιγραφεί η διαδικασία της συλλογής, προ-επεξεργασίας και μετασχηματισμού των δεδομένων καθώς και η επιλογή και παραμετροποίηση των διαφορετικών μοντέλων μηχανικής μάθησης που θα εφαρμοστούν στα διαφορετικά σύνολα δεδομένων.

Στο τέταρτο κεφάλαιο θα παρουσιαστούν τα αποτελέσματα της εφαρμογής των διαφορετικών αλγορίθμων Μηχανικής Μάθησης στα σύνολα δεδομένων. Αφού γίνει η ανάλυση των αποτελεσμάτων, θα αξιολογηθούν οι διαφορετικοί αλγόριθμοι και θα γίνει η ανάδειξη της αποδοτικότερης μεθόδου.

Στο πέμπτο και τελευταίο κεφάλαιο θα δοθούν τα συμπεράσματα, όποιοι περιορισμοί παρουσιάστηκαν κατά την πειραματική διαδικασία και προτάσεις για επέκταση του θέματος και περαιτέρω έρευνα.

ΚΕΦΑΛΑΙΟ 1. Εξόρυξη Γνώσης και Μηχανική Μάθηση

Στο κεφάλαιο αυτό, αφού αναλυθούν ορισμένες βασικές έννοιες γύρω από την Εξόρυξη Γνώσης (Data Mining) και τη Μηχανική Μάθηση (Machine Learning), θα δοθούν οι λεπτομέρειες σχετικά με τα είδη και τους αλγορίθμους Μηχανικής Μάθησης καθώς και οι μετρικές που χρησιμοποιούνται για την αξιολόγηση των διαφορετικών μοντέλων.

1.1 Εξόρυξη δεδομένων (Data Mining)

Η εξόρυξη δεδομένων αποτελεί έναν κλάδο της πληροφορικής που τα τελευταία χρόνια έχει γνωρίσει μεγάλη άνθιση. Ο τομέας της εξόρυξης δεδομένων συνεργάζεται με πολλούς επιστημονικούς κλάδους, όπως αυτούς της στατιστικής (statistics), της τεχνητής νοημοσύνης (artificial intelligence), της μηχανικής μάθησης (machine learning) και των βάσεων δεδομένων (databases).

Η εξόρυξη δεδομένων εκμεταλλεύεται τον τεράστιο όγκο των διαθέσιμων δεδομένων και καταφέρνει, μέσω της επεξεργασίας τους, να ανακαλύψει χρήσιμη γνώση. Τη σύγχρονη εποχή, η καταγραφή και αποθήκευση των δεδομένων γίνεται με καταιγιστικούς ρυθμούς. Η έξαρση αυτού του φαινομένου οφείλεται στην αυξημένη υπολογιστική ισχύ, στην ενσωμάτωση της τεχνολογίας σε όλους τους τομείς της σύγχρονης κοινωνίας και στην ευρύτατη χρήση του διαδικτύου. Ο ανθρώπινος νους έχει περιορισμένες αναλυτικές δυνατότητες, ανεπαρκείς για την αντιμετώπιση του μεγάλου όγκου των δεδομένων, γεγονός που καθιστά την επεξεργασία των δεδομένων αυτών αργή, ανακριβή και αναξιόπιστη. Διαφορετικοί επιστημονικοί κλάδοι, όπως η στατιστική και η μηχανική μάθηση, δεν παρέχουν τη δυνατότητα διαχείρισης του τεράστιου όγκου των δεδομένων, ενώ ο κλάδος των βάσεων δεδομένων, ο οποίος είναι και ο κύριος αρμόδιος για την αποθήκευση μεγάλου όγκου δεδομένων, δεν παρέχει τα εργαλεία για την ανάλυση αυτών [3].

1.1.1 Κατηγορίες εξόρυξης δεδομένων

1.1.1.1 Κατηγοριοποίηση (classification)

Στην περίπτωση της κατηγοριοποίησης παράγεται ένα μοντέλο, το οποίο ονομάζεται ταξινομητής (classifier), το οποίο εκπαιδεύεται πάνω στα διαθέσιμα δεδομένα. Ουσιαστικά, η εκπαίδευση του μοντέλου γίνεται μέσω της εκμάθησης μιας

συνάρτησης, η οποία αντιστοιχίζει ένα αντικείμενο εισόδου, που αναπαρίσταται ως ένα διάνυσμα τιμών, στη λεγόμενη κλάση εξόδου, η οποία λαμβάνει διακριτές τιμές.

1.1.1.2 Συσταδοποίηση (clustering)

Η συσταδοποίηση αποτελεί μία περιγραφική μέθοδο η οποία στοχεύει στη δημιουργία συστάδων (clusters), δηλαδή ομάδων, οι οποίες θα περιέχουν δείγματα με παρόμοιες ιδιότητες. Η ομοιότητα των δεδομένων εξαρτάται κάθε φορά από το υπό μελέτη πρόβλημα. Ουσιαστικά, πρέπει να βρεθεί ένα πεπερασμένο σύνολο ομάδων, το οποίο θα περιγράφει με βέλτιστο τρόπο τα δεδομένα.

1.1.1.3 Παλινδρόμηση (regression)

Η παλινδρόμηση αποτελεί μία μέθοδο πρόβλεψης, μέσω της οποίας μοντελοποιείται η σχέση μεταξύ μίας βαθμωτής εξαρτημένης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών. Η παλινδρόμηση ανακαλύπτει μία συνάρτηση η οποία μοντελοποιεί τα δεδομένα με τον καλύτερο δυνατό τρόπο.

1.1.1.4 Εξαγωγή κανόνων συσχέτισης (association rule learning)

Οι κανόνες συσχέτισης καταφέρνουν να ανακαλύψουν κρυμμένες και ενδιαφέρουσες συσχετίσεις μεταξύ των μεταβλητών ενός συνόλου δεδομένων. Αυτό επιτυγχάνεται με την ανακάλυψη κανόνων χρησιμοποιώντας διάφορες μετρικές ενδιαφέροντος. Οι συσχετίσεις μεταξύ των μεταβλητών υποδηλώνονται με τη χρήση του τελεστή →, οπότε ένας κανόνας συσχέτισης A→B προβλέπει την εμφάνιση των χαρακτηριστικών του συνόλου B δεδομένης της εμφάνισης των χαρακτηριστικών του συνόλου A.

1.1.1.5 Ανίχνευση ανωμαλιών (anomaly detection)

Η ανίχνευση ανωμαλιών ασχολείται με την ανακάλυψη αποκλίσεων στα δεδομένα συγκριτικά με τυπικές τιμές των δεδομένων που έχουν ήδη συλλεχθεί. Συνήθως, τα ανώμαλα σημεία μαρτυρούν κάποιο πρόβλημα, όπως είναι μία τραπεζική απάτη, δομικές ατέλειες, προβλήματα υγείας ή λάθη σε ένα κείμενο.

1.1.2 Ανακάλυψη γνώσης από βάσεις δεδομένων

Η ανακάλυψη γνώσης από βάσεις δεδομένων (knowledge discovery in databases – KDD) αφορά την αποκάλυψη ή παραγωγή λειτουργικής γνώσης μέσα από την ανάλυση των δεδομένων. Πρόκειται για μία ολοκληρωμένη διαδικασία, από τη συλλογή δεδομένων μέχρι την αξιοποίηση των αποτελεσμάτων σε πιο πρακτικό επίπεδο. Τα

βασικά στάδια της ανακάλυψης γνώσης από βάσεις δεδομένων περιγράφονται παρακάτω [4]:

- ***Συλλογή δεδομένων (data collection)***

Η συλλογή των δεδομένων γίνεται είτε αυτόματα, πχ με τη χρήση αισθητήρων, είτε μη αυτόματα, πχ χρησιμοποιώντας ερωτηματολόγια. Σε περιπτώσεις που οι αισθητήρες δυσλειτουργούν ή δεν απαντηθούν κάποιες ερωτήσεις στα ερωτηματολόγια, τα δεδομένα μπορεί να περιέχουν αρκετό θόρυβο, ο οποίος επηρεάζει αρνητικά τις αναλύσεις που ακολουθούν στα επόμενα στάδια.

- ***Προεπεξεργασία δεδομένων (preprocessing)***

Κατά τη συλλογή των δεδομένων υπάρχει περίπτωση ορισμένα δεδομένα να περιέχουν μη αποδεκτές τιμές (πχ Εισόδημα: -100), ανέφικτους συνδυασμούς δεδομένων (πχ Φύλο: Άντρας, Έγκυος: Ναι), τιμές που απουσιάζουν, κτλ. Επομένως, η ανάλυση τέτοιων δεδομένων μπορεί να οδηγήσει σε παραπλανητικά αποτελέσματα, γεγονός που καθιστά αναγκαία τη διασφάλιση της ποιότητας των δεδομένων πριν να διεξαχθούν οι αναλύσεις. Η διαδικασία της προεπεξεργασίας περιλαμβάνει, κυρίως, τον καθαρισμό των δεδομένων (cleaning), την κανονικοποίηση (normalization) και την εξαγωγή χαρακτηριστικών (feature extraction).

- ***Μετασχηματισμός δεδομένων (transformation)***

Τα δεδομένα μετασχηματίζονται, με σκοπό να αποκτήσουν κατάλληλη μορφή για τη μετέπειτα επεξεργασία τους. Πρακτικά, ο μετασχηματισμός πραγματοποιείται μέσω της χρήσης μίας ντετερμινιστικής μαθηματικής συνάρτησης σε όλα τα στιγμιότυπα του συνόλου δεδομένων, ώστε η τιμή κάθε στιγμιότυπου να αντικατασταθεί με την αντίστοιχη μετασχηματισμένη τιμή. Ο μετασχηματισμός είναι απαραίτητος, κυρίως, για την εξομάλυνση των δεδομένων και την απομάκρυνση του θορύβου.

- ***Εξόρυξη δεδομένων (data mining)***

Σε αυτό το στάδιο εφαρμόζεται κάποιος αλγόριθμος εξόρυξης δεδομένων ώστε να παραχθεί το τελικό μοντέλο. Τα πλέον προεπεξεργασμένα και μετασχηματισμένα δεδομένα τροφοδοτούνται στον αλγόριθμο, ώστε να κατασκευαστεί το επιθυμητό μοντέλο, το οποίο, συνήθως, είναι κάποιο μοντέλο κατηγοριοποίησης ή πρόβλεψης. Το

παραγόμενο μοντέλο, το οποίο κατασκευάστηκε με βάση γνωστά δεδομένα, θέλουμε να μας δώσει απάντηση για την τιμή μιας άγνωστης μεταβλητής στόχου.

- ***Διερμηνεία και Αξιολόγηση (interpretation - evaluation)***

Σε αυτό το στάδιο λαμβάνει χώρα η διερμηνεία και αξιολόγηση των αποτελεσμάτων που εξάγει το μοντέλο, μέσω της χρήσης κατάλληλων μετρικών.

1.2 Μηχανική Μάθηση

Η μηχανική μάθηση (machine learning) αποτελεί ένα πεδίο της επιστήμης των υπολογιστών στο οποίο, τα τελευταία χρόνια, έχει επικεντρωθεί το ενδιαφέρον των επιστημόνων. Το 1959, ο Arthur Samuel αναφέρει ότι ‘η μηχανική μάθηση δίνει στους υπολογιστές τη δυνατότητα να μάθουν χωρίς να είναι ρητά προγραμματισμένοι’ [37]. Ένας γενικός ορισμός για τη μηχανική μάθηση δίνεται από τον Mitchell το 1997 [7]:

> *‘Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E, όπου T η κλάση εργασιών και P το μέτρο απόδοσης, αν η απόδοση P στις εργασίες T βελτιώνεται μέσω της εμπειρίας E’.*

Προερχόμενη από τους χώρους της αναγνώρισης προτύπων και της υπολογιστικής μάθησης που χρησιμοποιείται στην τεχνητή νοημοσύνη, η μηχανική μάθηση σχετίζεται με την ανάλυση και σχεδίαση αλγορίθμων που μπορούν να μάθουν από τα δεδομένα και να κάνουν προβλέψεις σχετικές με αυτά. Η εφαρμογή των τεχνικών μηχανικής μάθησης σε τεράστιους όγκους δεδομένων δημιούργησε τον κλάδο της εξόρυξης δεδομένων (data mining). Στην εξόρυξη δεδομένων, τα πολυάριθμα δεδομένα επεξεργάζονται, ώστε να κατασκευαστεί ένα απλό μοντέλο με χρήσιμη λειτουργικότητα και υψηλή απόδοση. Οι εφαρμογές της είναι ποικίλες, όπως για παράδειγμα, στον τραπεζικό τομέα, όπου χρησιμοποιείται για την ανάλυση παρελθοντικών δεδομένων και την ανίχνευση απάτης ή τη διαχείριση αποθέματος. Στην ιατρική, τα μοντέλα μάθησης χρησιμοποιούνται στη διάγνωση, όπως για παράδειγμα κατά τον εντοπισμό καρκινικών και φυσιολογικών κυττάρων. Στις τηλεπικοινωνίες, τα μοντέλα μηχανικής μάθησης χρησιμοποιούνται για τη μεγιστοποίηση της παρεχόμενης ποιότητας υπηρεσιών, αλλά και στο πρόβλημα της πρόβλεψης απώλειας πελατών. Στην επιστήμη, οι τεράστιοι όγκοι δεδομένων στους κλάδους της φυσικής, αστρονομίας και βιολογίας είναι δυνατόν να αναλυθούν γρήγορα μόνο από τους υπολογιστές [9].

Ωστόσο, η μηχανική μάθηση, πέρα από την εξόρυξη δεδομένων, αποτελεί και μέρος της τεχνητής νοημοσύνης. Για να είναι έξυπνο, ένα σύστημα που βρίσκεται σε ένα μεταβαλλόμενο περιβάλλον, πρέπει να έχει τη δυνατότητα να μαθαίνει και να προσαρμόζεται στις αλλαγές, ώστε να παρέχει αποτελεσματικά λύσεις για όλες τις δυνατές περιπτώσεις.

Συνδυάζοντας τις έννοιες της εξόρυξης δεδομένων και της τεχνητής νοημοσύνης, μπορούμε να πούμε ότι η μηχανική μάθηση αφορά τον προγραμματισμό των υπολογιστών με τη χρήση παρελθοντικών δεδομένων ή προηγούμενης εμπειρίας, με απώτερο στόχο τη βελτιστοποίηση ενός κριτηρίου απόδοσης. Το μοντέλο που κατασκευάζεται πρέπει να βελτιστοποιεί τις παραμέτρους του μέσω της διαδικασίας της εκπαίδευσης και μπορεί να είναι προγνωστικό μοντέλο, οπότε να εκτελεί προβλέψεις για μελλοντικές καταστάσεις, περιγραφικό, όπου αποκτά γνώση από τα δεδομένα ή να συνδυάζει και τα δύο. Η μηχανική μάθηση χρησιμοποιεί τη στατιστική για να κατασκευάζει τα μαθηματικά μοντέλα, καθώς η βασική διεργασία βασίζεται στην εξαγωγή συμπεράσματος από ένα δείγμα.

Στη μηχανική μάθηση, πρέπει να δοθεί, αρχικά, ιδιαίτερη προσοχή στη διαδικασία εκπαίδευσης, καθώς απαιτείται τόσο η χρήση των κατάλληλων αλγορίθμων που θα επιλύσουν το πρόβλημα βελτιστοποίησης όσο και η αποτελεσματική διαχείριση και επεξεργασία των τεράστιων όγκων δεδομένων. Δεύτερον, αφού εκπαιδευτεί το μοντέλο, πρέπει αυτό να έχει υψηλή απόδοση, όχι μόνο αναφορικά με την πρόβλεψη, αλλά και με τη χωρική και χρονική πολυπλοκότητα του αλγορίθμου.

Η μηχανική μάθηση χρησιμοποιείται σε μία πληθώρα προβλημάτων και υπολογιστικών διαδικασιών, όπου συχνά είναι ανέφικτος ο σχεδιασμός και η ανάπτυξη αναλυτικών αλγορίθμων. Ορισμένα παραδείγματα εφαρμογών της μηχανικής μάθησης είναι το φιλτράρισμα της ηλεκτρονικής αλληλογραφίας (spam filtering), ο εντοπισμός εισβολέων σε ιδιωτικά δίκτυα, η οπτική αναγνώριση χαρακτήρων (OCR), η υπολογιστική όραση, η επεξεργασία φυσικής γλώσσας, κτλ.

1.2.1 Είδη Μηχανικής Μάθησης

Η μηχανική μάθηση διακρίνεται στα τρία παρακάτω βασικότερα είδη, ανάλογα με τη φύση των σημάτων ή της ανατροφοδότησης σε ένα σύστημα μάθησης [37]:

1.2.1.1 Επιβλεπόμενη μάθηση (supervised learning)

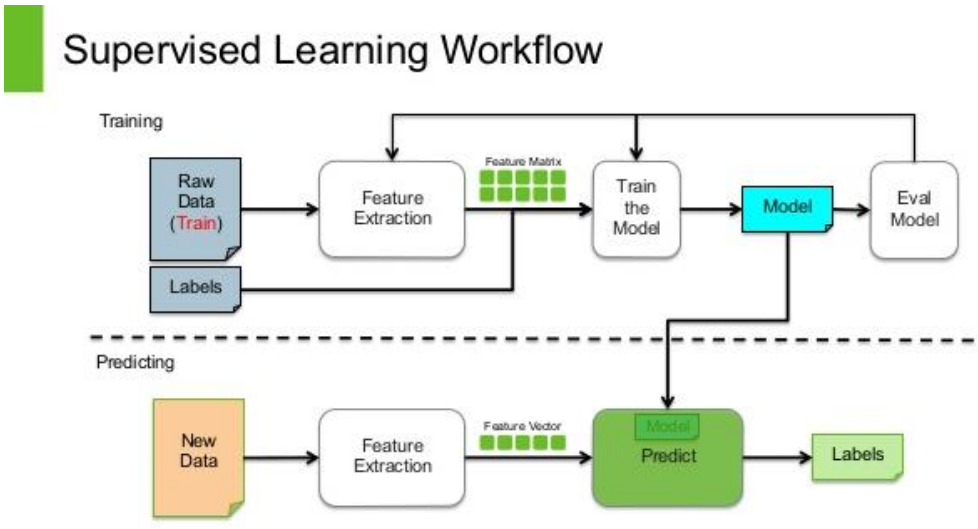
Στην παρούσα διπλωματική εργασία θα χρησιμοποιηθούν μοντέλα επιβλεπόμενης μάθησης, τα οποία τροφοδοτούνται με δεδομένα, τα οποία ονομάζονται δεδομένα εκπαίδευσης (training set), ώστε να κατασκευαστεί μία συνάρτηση η οποία θα απεικονίζει δεδομένες εισόδους σε συγκεκριμένες εξόδους. Τα δεδομένα εκπαίδευσης αποτελούνται από ένα σύνολο εγγραφών, όπου κάθε εγγραφή χαρακτηρίζεται από ένα διάνυσμα εισόδου και την επιθυμητή τιμή της εξόδου. Ένας, λοιπόν, αλγόριθμος επιβλεπόμενης μάθησης αναλύει τα δεδομένα εισόδου και παράγει μία συνάρτηση, η οποία, στη συνέχεια, χρησιμοποιείται για την απεικόνιση νέων δεδομένων, όπου πλέον η έξοδος είναι άγνωστη. Ο βασικός στόχος κατά τη διαδικασία της εκπαίδευσης είναι η γενίκευση του αλγορίθμου, ώστε η αντιστοίχιση των νέων εισόδων στις αντίστοιχες εξόδους να γίνεται με τη μέγιστη δυνατή ακρίβεια.

Κατά την επίλυση ενός προβλήματος επιβλεπόμενης μάθησης πρέπει να ακολουθούνται τα παρακάτω βήματα [38]:

1. Θα πρέπει να οριστεί ο τύπος των δεδομένων εκπαίδευσης. Για παράδειγμα, στην περίπτωση της ανάλυσης γραφικού χαρακτήρα, οι εγγραφές μπορεί να είναι ένας απλός χαρακτήρας γραμμένος στο χέρι, μία λέξη από τέτοιους χαρακτήρες ή μια γραμμή από λέξεις.
2. Θα πρέπει να γίνει η συλλογή των δεδομένων εκπαίδευσης. Όπως προαναφέρθηκε, στην περίπτωση της επιβλεπόμενης μάθησης, το σύνολο των δεδομένων εκπαίδευσης πρέπει να αποτελείται τόσο από τα διανύσματα εισόδου όσο και από τις εξόδους που αντιστοιχούν σε αυτά. Τα δεδομένα αυτά μπορεί να συλλεχθούν από ειδικούς ή να προκύψουν από διάφορες μετρήσεις.
3. Στο βήμα αυτό θα πρέπει να αποφασιστεί ο τρόπος με τον οποίο θα αναπαρασταθούν τα χαρακτηριστικά εισόδου, τα οποία θα τροφοδοτηθούν στον αλγόριθμο μάθησης. Συνήθως, το διάνυσμα εισόδου μετατρέπεται σε ένα διάνυσμα χαρακτηριστικών, το οποίο περιγράφει το αντικείμενο. Ο αριθμός των χαρακτηριστικών δεν πρέπει να είναι ιδιαίτερα μεγάλος, λόγω του φαινομένου της κατάρας της διαστατικότητας (curse of dimensionality), ωστόσο, θα πρέπει τα χαρακτηριστικά που θα επιλεγούν να εμπεριέχουν αρκετή πληροφορία, ώστε να μπορεί να προβλεφθεί με ακρίβεια η έξοδος.

4. Θα πρέπει να αποφασιστεί το είδος του αλγορίθμου που θα χρησιμοποιηθεί. Ο αλγόριθμος αυτός προέρχεται από ένα ευρύ φάσμα τεχνικών μηχανικής μάθησης, όπως είναι τα δέντρα απόφασης, τα νευρωνικά δίκτυα, τα μοντέλα SVM, κτλ.
5. Αφού ολοκληρωθεί η σχεδίαση, στον αλγόριθμο τροφοδοτούνται τα δεδομένα εκπαίδευσης. Αρκετοί αλγόριθμοι επιβλεπόμενης μάθησης απαιτούν από το χρήστη να ορίσει ορισμένες παραμέτρους ελέγχου. Οι παράμετροι αυτές μπορούν να αναπροσαρμόζονται, μέχρις ότου βελτιστοποιηθεί η απόδοση του αλγορίθμου σε ένα υποσύνολο των δεδομένων εκπαίδευσης, το οποίο λέγεται σύνολο επαλήθευσης (validation set), ή μέσω της διαδικασίας cross-validation.
6. Τέλος, θα πρέπει να ελεγχθεί η ακρίβεια του αλγορίθμου. Αφού γίνει η αναπροσαρμογή των παραμέτρων και ολοκληρωθεί η εκπαίδευση, θα πρέπει να μετρηθεί η απόδοση του αλγορίθμου μάθησης στο σύνολο ελέγχου (test set).

Αφού ολοκληρωθούν όλα τα παραπάνω βήματα, το τελικό μοντέλο είναι έτοιμο να εφαρμοστεί σε νέα δεδομένα, με άγνωστες εξόδους. Η διαδικασία που ακολουθείται έχει ως εξής: το μοντέλο τροφοδοτείται με τις νέες εισόδους, εξάγεται το διάνυσμα των χαρακτηριστικών και γίνεται η πρόβλεψη, οπότε, τελικά, επιστρέφονται οι προβλέψεις των κλάσεων της μεταβλητής εξόδου.



Εικόνα 1 Η ροή εργασιών κατά την επιβλεπόμενη μάθηση [Πηγή: http://en.proft.me/2015/12/24/types-machine-learning-algorithms/]

1.2.1.2 Μη επιβλεπόμενη μάθηση (unsupervised learning)

Στην περίπτωση της μη επιβλεπόμενης μάθησης ο αλγόριθμος προσπαθεί να εξηγήσει την κρυμμένη δομή των δεδομένων, καθώς δεν είναι γνωστές οι κλάσεις της εξόδου. Αφού τα διαθέσιμα δεδομένα δεν περιέχουν πληροφορίες για την έξοδο, δεν μπορεί να αξιολογηθεί η απόδοση του αλγορίθμου, όπως γίνεται στην επιβλεπόμενη μάθηση. Η μη επιβλεπόμενη μάθηση χρησιμοποιείται στη στατιστική για την εκτίμηση πυκνότητας, καθώς και για να εντοπίσει τα σημαντικότερα χαρακτηριστικά των δεδομένων [39].

1.2.1.3 Ενισχυτική μάθηση (reinforcement learning)

Η ενισχυτική μάθηση αφορά τον τρόπο με τον οποίο οι πράκτορες λογισμικού εκτελούν ενέργειες σε ένα περιβάλλον, με σκοπό να μεγιστοποιήσουν μία αθροιστική επιβράβευση. Αυτό το είδος μάθησης χρησιμοποιείται σε διάφορους τομείς, όπως στη θεωρία παιγνίων, στη θεωρία πληροφοριών, στα συστήματα πολλών πρακτόρων, στη νοημοσύνη σμήνους, στη στατιστική και στους γενετικούς αλγορίθμους.

Το περιβάλλον κατά την ενισχυτική μάθηση συντίθεται ως μια διαδικασία απόφασης Markov (Markov Decision Process – MDP), καθώς πολλοί αλγόριθμοι ενισχυτικής μάθησης χρησιμοποιούν τεχνικές δυναμικού προγραμματισμού. Η βασική διαφορά των κλασικών τεχνικών από αυτές της ενισχυτικής μάθησης είναι ότι οι δεύτερες δε χρειάζονται τη γνώση σχετικά με τις διαδικασίες MDP.

Η ενισχυτική μάθηση διαφέρει από την επιβλεπόμενη ως προς το γεγονός ότι δε διατίθενται τα σωστά ζευγάρια εισόδου/εξόδου, αλλά η εκπαίδευση γίνεται μέσω μιας σειράς ποινών και επιβραβεύσεων. Εδώ γίνεται προσπάθεια να εξασφαλιστεί η ισορροπία μεταξύ της εξερεύνησης (exploration) και της εκμετάλλευσης (exploitation). Η εξερεύνηση αναφέρεται στην ανακάλυψη χώρου άγνωστου μέχρι τώρα και η εκμετάλλευση στην αξιοποίηση της ήδη υπάρχουσας γνώσης [40].

1.2.2 Κατηγοριοποίηση κειμένου (text classification)

Η κατηγοριοποίηση κειμένου αποτελεί μία διαδικασία κατά την οποία αναλύεται μία σειρά εγγράφων και ανατίθενται κατηγορίες σε αυτά. Για παράδειγμα, οι ειδήσεις οργανώνονται σε θεματικές ενότητες και οι ακαδημαϊκές εργασίες ταξινομούνται με βάση το υπό μελέτη αντικείμενο. Ένα ακόμη χαρακτηριστικό παράδειγμα της

κατηγοριοποίησης κειμένου είναι ο διαχωρισμός της ηλεκτρονικής αλληλογραφίας σε επιθυμητή και ανεπιθύμητη (spam/non-spam) ή στην περίπτωση της παρούσας διπλωματικής εργασίας ο χαρακτηρισμός μιας είδησης ως ψευδούς ή αληθούς.

Μία από τις πιο δημοφιλείς τεχνικές κατηγοριοποίησης κειμένου είναι η μέθοδος bag-of-words. Σε αυτήν τη μέθοδο, ένα κείμενο, το οποίο μπορεί να αντιστοιχεί σε μία πρόταση ή ένα έγγραφο, αναπαρίσταται ως ένα σύνολο (bag) από λέξεις, όπου το μόνο που μας ενδιαφέρει είναι η συχνότητα εμφάνισης των λέξεων. Πρακτικά, η μέθοδος bag-of-words χρησιμοποιείται σαν εργαλείο για την εξαγωγή χαρακτηριστικών. Δηλαδή, αφού μετασχηματισθεί το κείμενο σε ένα σύνολο από λέξεις, υπολογίζονται διάφορες μετρικές με στόχο να χαρακτηρισθεί το κείμενο. Ο πιο συνηθισμένος τύπος χαρακτηριστικών που παράγονται από το μοντέλο είναι η συχνότητα εμφάνισης των λέξεων (term frequency), δηλαδή πόσες φορές εμφανίζεται κάθε λέξη μέσα στο κείμενο.

Ωστόσο, το κείμενο δεν αναπαρίσταται αποτελεσματικά μέσω των συχνοτήτων των λέξεων. Αυτό οφείλεται στο γεγονός ότι μέρη του λόγου που δεν προσδίδουν νόημα, όπως είναι τα άρθρα και οι αντωνυμίες, εμφανίζονται πολύ συχνά μέσα στο κείμενο. Οπότε, αυτές οι λέξεις θα έχουν μεγαλύτερο βαθμό βαρύτητας σύμφωνα με τη μέθοδο, γεγονός που δεν ισχύει. Για να αντιμετωπιστεί το συγκεκριμένο πρόβλημα, χρησιμοποιούνται μέθοδοι ώστε να ‘κανονικοποιηθούν’ οι συχνότητες εμφάνισης των λέξεων, η πιο δημοφιλής εκ των οποίων είναι η TF-IDF (term frequency – inverse document frequency), όπου σε κάθε λέξη ανατίθεται ένα βάρος. Περισσότερες λεπτομέρειες γύρω από την προ-επεξεργασία και το μετασχηματισμό κειμένου θα δοθούν στην αμέσως επόμενη ενότητα.

1.2.3 Προ-επεξεργασία και Μετασχηματισμός δεδομένων

Στην ενότητα αυτή γίνεται παρουσίαση των βασικών τεχνικών προ-επεξεργασίας των δεδομένων και του μετασχηματισμού αυτών, πριν δοθούν σαν είσοδος στα διαφορετικά μοντέλα μηχανικής μάθησης που θα χρησιμοποιηθούν κατά την πειραματική ανάλυση [1]. Τα δεδομένα που χρησιμοποιούνται για το πρόβλημα της εκτίμησης και επικύρωσης διαδικτυακού περιεχομένου είναι σε μορφή κειμένου, οπότε θα εφαρμοστούν σε αυτό τεχνικές επεξεργασίας φυσικής γλώσσας.

1.2.3.1 Λειτουργίες Tokenization, Stemming, Lemmatization και LowerCasing

Η διαδικασία tokenization αφορά τη διαδικασία του κατακερματισμού του κειμένου σε λέξεις, φράσεις, σύμβολα ή άλλα στοιχεία που καλούνται tokens. Τυπικά, το tokenization λαμβάνει χώρα στο επίπεδο των λέξεων. Ωστόσο, μερικές φορές, λόγω της πολυπλοκότητας της γλώσσας, χρησιμοποιούνται οι παρακάτω ευρετικές τεχνικές (heuristics) [1]:

- Τα σημεία στίξης και τα κενά μπορεί να περιέχονται στη λίστα που περιλαμβάνει τα tokens.
- Όλοι οι συνεχόμενοι χαρακτήρες αποτελούν μέρος ενός token.
- Τα διαφορετικά tokens διαχωρίζονται με κενούς χαρακτήρες ή με σημεία στίξης.

Η τεχνική Stemming χρησιμοποιείται για την εύρεση της ρίζας λέξεων με παρόμοια μορφολογία επιτυγχάνοντας με αυτό τον τρόπο τη μείωση των διαστάσεων του χώρου των χαρακτηριστικών. Για παράδειγμα οι λέξεις ψάρι, ψάρια και ψαρεύω έχουν όλες τη ρίζα “ψάρι”, οπότε αυτόματα αντί για τρεις λέξεις καταλήγομε σε μία βασική με το ίδιο σημασιολογικό περιεχόμενο. Το πιο γνωστό παράδειγμα τέτοιου αλγορίθμου είναι ο “Porter Stemmer” ο οποίος κατασκευάστηκε στη δεκαετία του 1980 και χρησιμοποιείται στις μέρες μας στις βιβλιοθήκες πολλών γλωσσών προγραμματισμού [43].

Επιπλέον, η τεχνική Lemmatization, η οποία είναι παρόμοια με αυτήν του Stemming, χρησιμοποιείται για την αντικατάσταση των λέξεων με τη μορφολογική τους ρίζα. Η διαφορά από το Stemming είναι ότι όλες οι ρίζες που προκύπτουν αντιστοιχούν σε πραγματικές λέξεις, για αυτό το λόγο είναι απαραίτητη και η πληροφορία των συμφραζόμενων της λέξης, δηλαδή αν πρόκειται για ρήμα, ουσιαστικό, κτλ. [43].

Τέλος, μία επιπλέον απλή τεχνική που εφαρμόζεται σε δεδομένα κειμένου είναι η μετατροπή όλων των λέξεων σε πεζούς χαρακτήρες (lowercasing). Στην περίπτωση που παραλειφθεί αυτό το βήμα, τότε ο αλγόριθμος θα αστοχεί να διακρίνει δύο όμοιες λέξεις μία εκ των οποίων θα ξεκινά με κεφαλαίο γράμμα γιατί θα βρίσκεται στην αρχή της πρότασης [2].

1.2.3.2 Λειτουργία Stop word removal

Μία από τις σημαντικότερες διαδικασίες κατά την προ-επεξεργασία κειμένου είναι το stop word removal, δηλαδή η αφαίρεση λέξεων που δεν προσδίδουν νόημα και αξία κατά την επεξεργασία της φυσικής γλώσσας. Τα πιο συνηθισμένα stop words είναι τα άρθρα και οι αντωνυμίες, τα οποία εμφανίζονται με πολύ μεγάλη συχνότητα στις συλλογές κειμένων. Για τον εντοπισμό των stop words είναι δυνατό να χρησιμοποιηθούν γνωστές, έτοιμες λίστες οι οποίες περιέχουν τα πιο συνηθισμένα stop words ή να εξετασθεί η συχνότητα εμφάνισης των λέξεων στο κείμενο, όπου σε αυτήν την περίπτωση τα stop words θα έχουν πολύ μεγάλες συχνότητες [1].

1.2.3.3 Λειτουργία Part-of-speech Tagging

Η διαδικασία αυτή αφορά τον προσδιορισμό των μερών του λόγου των λέξεων που ανήκουν στην υπό μελέτη συλλογή κειμένου. Συναντάται με τις ονομασίες part-of-speech-tagging, POS-tagging ή POST. Το part-of-speech tagging πραγματοποιείται στα πλαίσια της υπολογιστικής γλωσσολογίας μέσω της χρήσης αλγορίθμων που συσχετίζουν διαφορετικές λέξεις και κρυμμένα μέρη του λόγου, με βάση ενός συνόλου περιγραφικών ετικετών (tags). Υπάρχουν δύο είδη POS-tagging αλγορίθμων, αυτοί που βασίζονται σε κανόνες και οι στοχαστικοί [1][43].

1.2.3.4 Μετασχηματισμός TF-IDF

Η διαδικασία TF-IDF είναι μία στατιστική μετρική που χρησιμοποιείται για την εξαγωγή χαρακτηριστικών σε δεδομένα φυσικής γλώσσας και αξιολογεί πόσο σημαντική είναι μια λέξη σε ένα έγγραφο ή σε μία συλλογή εγγράφων (corpus). Αυτό το επιτυγχάνει πολλαπλασιάζοντας δύο μετρικές: 1) τη συχνότητα της λέξης μέσα στο κείμενο και 2) την αντίστροφη συχνότητα κειμένου της λέξης σε ένα σύνολο εγγράφων. Η αντίστροφη συχνότητα συμβολίζει την τιμή που είναι αντιστρόφως ανάλογη του πλήθους των κειμένων που περιέχουν τον όρο. Αυτό σημαίνει ότι η μετρική TF-IDF αυξάνεται αναλογικά με το πόσες φορές εμφανίζεται μία λέξη μέσα σε ένα κείμενο, αλλά παράλληλα αντισταθμίζεται από τη συχνότητα της λέξης μέσα σε ένα σύνολο εγγράφων. Αυτό που κάνει ουσιαστικά είναι να αποδίδει μεγαλύτερο βάρος σε λέξεις που έχουν μικρότερη συχνότητα μέσα στο κείμενο.

Προκειμένου να κατασκευαστεί το TF-IDF πλέγμα, πρέπει πρώτα να μετατραπεί το κείμενο σε ένα Term Document Matrix και να μετρηθεί η συχνότητα των όρων μέσα

στο έγγραφο. Έπειτα, μέσω της συνάρτησης βάρους (weighting) αποδίδεται το κατάλληλο βάρος σε κάθε λέξη, όπου τα υψηλά βάρη στο TF-IDF αντιστοιχούν σε λέξεις με υψηλή συχνότητα στο συγκεκριμένο κείμενο και χαμηλή συχνότητα σε ολόκληρη τη συλλογή κειμένων, γεγονός που σημαίνει ότι τα βάρη τείνουν να φιλτράρουν τις πολύ κοινές λέξεις [2]. Παρακάτω δίνονται οι τύποι υπολογισμού των TF-IDF συχνοτήτων:

$$TF(T, D) = \frac{\text{Συχνότητα του } T \text{ στο } D}{\text{Σύνολο λέξεων στο } D}$$

$$DF(T) = \text{Συχνότητα του } T \text{ σε } N \text{ έγγραφα}$$

$$TF - IDF(T, D) = TF(T, D) * \log\left(\frac{N}{DF(T)}\right),$$

όπου T: λέξη, D: έγγραφο, N: αριθμός δειγμάτων/εγγράφων

Στο σημείο αυτό αξίζει να σημειωθεί ότι η μέθοδος TF-IDF έχει προτιμηθεί έναντι της μεθόδου Bag-of-Words για την εξαγωγή των χαρακτηριστικών, καθώς η δεύτερη λαμβάνει υπόψη μόνο τη συχνότητα εμφάνισης της λέξης μέσα στο κείμενο, ενώ μέσω της TF-IDF καταλήγουμε στις πιο σημαντικές λέξεις, γεγονός που θα αυξήσει την ακρίβεια των μοντέλων μηχανικής μάθησης που θα χρησιμοποιηθούν στη συνέχεια.

1.2.4 Αλγόριθμοι Μηχανικής Μάθησης

Στην ενότητα αυτή θα παρουσιαστούν όλα τα μοντέλα μηχανικής μάθησης που θα χρησιμοποιηθούν για την αξιολόγηση και επαλήθευση διαδικτυακού περιεχομένου, ώστε να εντοπιστεί ψευδές και παραπλανητικό περιεχόμενο. Σε αυτό το σημείο αξίζει να σημειωθεί ότι το πρόβλημα που καλούμαστε να επιλύσουμε είναι η δυαδική ταξινόμηση (binary classification), καθώς ο στόχος μας είναι η κατηγοριοποίηση της κλάσης εξόδου σε Αληθή ή Ψευδή είδηση.

1.2.4.1 Λογιστική Παλινδρόμηση (Logistic Regression - LR)

Το μοντέλο Λογιστικής Παλινδρόμησης (LR) χρησιμοποιείται σε προβλήματα κατηγοριοποίησης δύο ή περισσότερων διακριτών μεταβλητών εξόδου, σε αντίθεση με τη γραμμική παλινδρόμηση όπου η μεταβλητή εξόδου λαμβάνει συνεχείς τιμές και αποτελεί μία στατιστική διαδικασία όπου γίνεται εκτίμηση της σχέσης μεταξύ μιας εξαρτημένης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Πιο

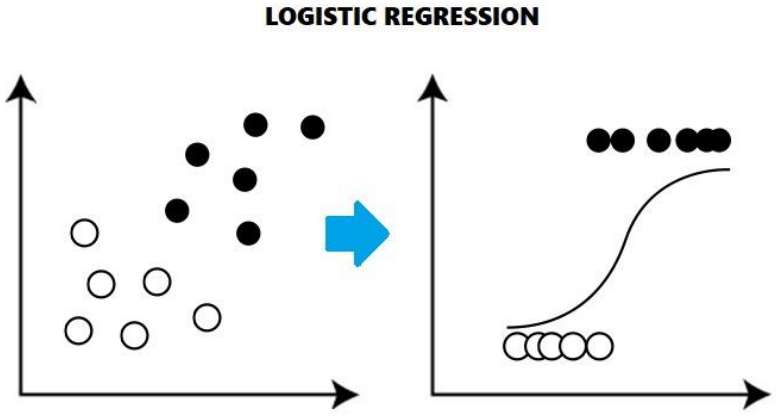
συγκεκριμένα, το λογιστικό μοντέλο είναι ένα μη γραμμικό μοντέλο, του οποίου η μεταβλητή απόκρισης είναι διακριτή και τα σφάλματα δεν ακολουθούν κανονική κατανομή. Η λογιστική παλινδρόμηση αποτελεί γενίκευση της γραμμικής παλινδρόμησης, όπου η εξαρτημένη μεταβλητή λαμβάνει τις τιμές 0 (απουσία χαρακτηριστικού) και 1 (παρουσία χαρακτηριστικού) και χρησιμοποιείται όταν επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού.

Μαθηματικά, η συνάρτηση υπόθεσης της λογιστικής παλινδρόμησης ορίζεται ως εξής [8]:

$$h_\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Η λογιστική παλινδρόμηση χρησιμοποιεί μια σιγμοειδή συνάρτηση για να μετασχηματίσει την έξοδο σε πιθανοτική τιμή. Ο στόχος είναι να ελαχιστοποιηθεί η συνάρτηση κόστους ώστε να επιτευχθεί η βέλτιστη δυνατή πιθανότητα. Η συνάρτηση κόστους υπολογίζεται όπως παρακάτω:

$$Cost(h_\theta(X), y) = \begin{cases} \log\big(h_\theta(X)\big), & y = 1 \\ -\log\big(1 - h_\theta(X)\big), & y = 0 \end{cases}$$



Εικόνα 2 Παράδειγμα μοντέλου Logistic Regression [Πηγή: https://www.equiskill.com/understanding-logistic-regression/]

Η κανονικοποίηση (regularization) συνιστά μία από τις σημαντικότερες τεχνικές της μηχανικής μάθησης που χρησιμοποιείται για να αποφεύγουμε την υπερεκπαίδευση των μοντέλων. Αν το δούμε από μαθηματική σκοπιά, ένας επιπλέον όρος προστίθεται στη συνάρτηση κόστους, έτσι ώστε οι συντελεστές της εξίσωσης της παλινδρόμησης να μην ταιριάσουν (fit) με απόλυτο τρόπο στα δεδομένα εκπαίδευσης, επομένως να μη

γίνει υπερεκπαίδευση. Τα διαφορετικά είδη κανονικοποίησης είναι δύο, τα οποία συμβολίζονται με L1 και L2 αντίστοιχα, τα οποία διαφοροποιούνται ως προς το γεγονός ότι στην L2 κανονικοποίηση, ο όρος regularization είναι ίσος με το άθροισμα του τετραγώνου των συντελεστών, ενώ στην L1 αντίστοιχα ισοδυναμεί με το άθροισμα των συντελεστών της εξίσωσης της γραμμικής παλινδρόμησης.

1.2.4.2 Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines - SVM)

Τα μοντέλα SVM έχουν σαν βασικό στόχο να βρουν το βέλτιστο υπερεπίπεδο διαχωρισμού, το οποίο ταξινομεί όσο το δυνατόν με μεγαλύτερη ακρίβεια τα σημεία των δεδομένων, διαχωρίζοντάς τα σε δύο κλάσεις [11]. Τα σημεία εκπαίδευσης που είναι πιο κοντά στο βέλτιστο υπερεπίπεδο ονομάζονται διανύσματα υποστήριξης και είναι αυτά που χρησιμοποιούνται για να αποφασιστεί η κλάση της εξόδου. Προκειμένου να διαχειριστούν προβλήματα μη γραμμικότητας, τα μοντέλα SVM προβάλλουν, αρχικά, τα δεδομένα σε ένα χώρο υψηλότερων διαστάσεων, μέσω μιας συνάρτησης κελύφους (kernel), όπου, πλέον, προσπαθούν να βρουν το γραμμικό περιθώριο (margin) που θα διαχωρίζει τα δεδομένα [12]. Καθώς υπάρχουν πολλαπλά διαφορετικά υπερεπίπεδα σε έναν N-διάστατο χώρο, ο στόχος είναι να βρεθεί το υπερεπίπεδο που διαχωρίζει τα σημεία δύο κλάσεων με το βέλτιστο περιθώριο (margin). Μια μαθηματική αναπαράσταση της συνάρτησης κόστους του SVM μοντέλου δίνεται παρακάτω:

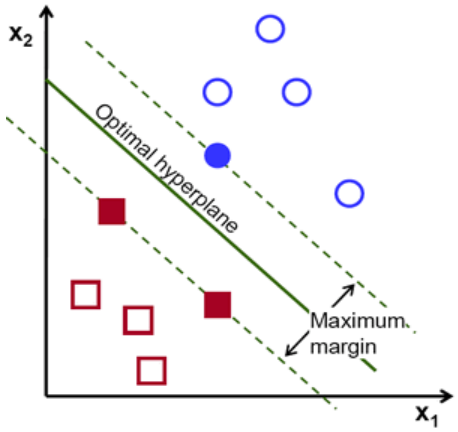
$$J(\theta) = \frac{1}{2}\sum_{j=1}^{n} \theta_j^2,$$

όπου

$$\theta^T\left(x^{(i)}\right) \geq 1, \qquad y^{(i)} = 1$$

$$\theta^T\left(x^{(i)}\right) \leq -1, \quad y^{(i)} = 0$$

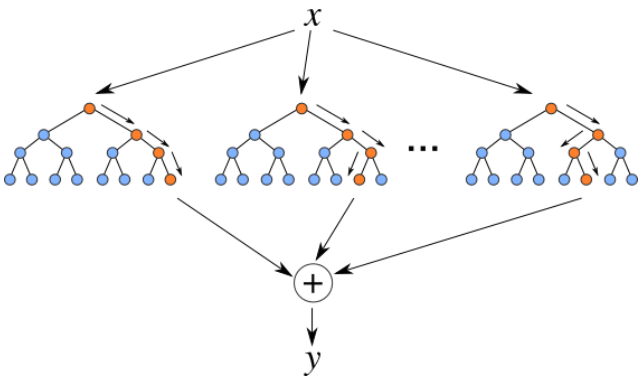
Η παραπάνω συνάρτηση χρησιμοποιεί γραμμική συνάρτηση kernel. Τα κελύφη (kernel) χρησιμοποιούνται για να είναι δυνατός ο διαχωρισμός σημείων που δεν είναι εύκολα διαχωρίσιμα ή βρίσκονται σε χώρους πολλών διαστάσεων, όπου αυξάνεται η πολυπλοκότητα του προβλήματος. Για την εκτέλεση των πειραμάτων στην παρούσα διπλωματική εργασία χρησιμοποιήθηκαν οι ακόλουθες συναρτήσεις kernel: sigmoid SVM, polynomial SVM, Gaussian SVM και το βασικό γραμμικό μοντέλο SVM [6].



Εικόνα 3 Παράδειγμα μοντέλου SVM [Πηγή: https://emilemathieu.fr/posts/2018/08/svm/]

1.2.4.3 Τυχαία Δάση (Random Forests)

Η μέθοδος Random Forests [16] αποτελεί παράδειγμα συλλογικής μάθησης και χρησιμοποιείται σε προβλήματα ταξινόμησης, όπου κατασκευάζει ένα πλήθος δέντρων απόφασης κατά τη διαδικασία της εκπαίδευσης και εξάγει σαν αποτέλεσμα την κλάση που εμφανίζεται πιο συχνά στο σύνολο των παραγόμενων δέντρων. Η μέθοδος αυτή χρησιμοποιεί τη bootstrap συνάθροιση, έναν αλγόριθμο συλλογικής μάθησης, που χρησιμοποιείται για τη βελτίωση της σταθερότητας και της ακρίβειας των αλγορίθμων μηχανικής μάθησης. Τα τυχαία δάση έρχονται να συμπληρώσουν τα δέντρα απόφασης, καθώς εξαλείφουν την τάση των δεύτερων να υπερεκπαιδεύονται.



Εικόνα 4 Παράδειγμα της μεθόδου Random Forest [Πηγή: https://kgpdag.wordpress.com/]

1.2.4.4 Μέθοδος Naive Bayes

Οι ταξινομητές Naive Bayes ανήκουν στην οικογένεια των πιθανοτικών μοντέλων, τα οποία βασίζονται στο θεώρημα Μπέυζ, σύμφωνα με το οποίο τα χαρακτηριστικά που περιγράφουν τα δεδομένα πρέπει να είναι ανεξάρτητα μεταξύ τους. Αυτό σημαίνει ότι

η παρουσία (ή απουσία) ενός χαρακτηριστικού μίας κλάσης δε σχετίζεται με την παρουσία (ή απουσία) οποιουδήποτε άλλου χαρακτηριστικού [15].

Από μαθηματική σκοπιά, ο Naive Bayes ταξινομητής υπολογίζει την πιθανότητα ένα δείγμα εισόδου να ανήκει σε μία συγκεκριμένη κλάση [15]. Δεδομένου ενός δείγματος X, το οποίο αποτελείται από ένα διάνυσμα χαρακτηριστικών $\{x_1, \dots, x_n\}$, η πιθανότητα της κλάσης y_j μπορεί να δοθεί από την εξίσωση:

$$p(y_j|X) = p(X|y_j)p(y_j) = p(x_1, \dots, x_n|y_j)p(y_j)$$

όπου $p(y_j)$ η προηγούμενη πιθανότητα του y_j. Ωστόσο, η μέθοδος Naïve Bayes υποθέτει, όπως είπαμε, ότι οι δεσμευμένες πιθανότητες των ανεξάρτητων μεταβλητών είναι στατιστικά ανεξάρτητες, οπότε προκύπτει ότι:

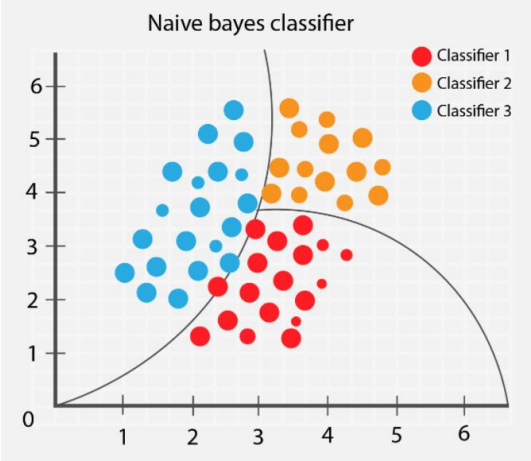
$$p(X|y_j) = \prod_{i=1}^{n} p(x_i|y_j)$$

και η πιθανότητα της κλάσης y_j δίνεται, τελικά, από την εξίσωση:

$$p(y_j|X) = p(y_j) \prod_{i=1}^{n} p(x_i|y_j)$$

Δεδομένου ενός αριθμού κλάσεων $Y = \{y_1, \dots, y_k\}$, ο Naïve Bayes ταξινομεί ένα νέο δείγμα X με βάση τη σχέση:

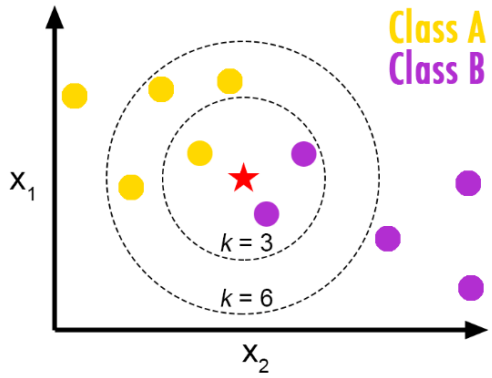
$$c = \underset{y_j \in Y}{\operatorname{argmax}}\, p(y_j|X)$$



Εικόνα 5 Παράδειγμα της μεθόδου Naïve Bayes [Πηγή: https://kdagiit.medium.com/naive-bayes-algorithm-4b8b990c7319]

1.2.4.5 Αλγόριθμος κ-κοντινότερων γειτόνων (k-Nearest Neighbors - k-NN)

Ο αλγόριθμος k-NN [17] είναι μία μη-παραμετρική μέθοδος, όπου για κάθε νέα παρατήρηση, η οποία περιγράφεται σαν ένα διάνυσμα N χαρακτηριστικών στο N-διάστατο χώρο, υπολογίζονται τα k κοντινότερα σημεία, με βάση τις ευκλείδειες αποστάσεις. Αυτά τα σημεία συνιστούν τους k κοντινότερους γείτονες. Τελικά, η κλάση εξόδου που αποδίδεται στο αντικείμενο αντιστοιχεί στην κλάση της πλειοψηφίας των γειτόνων. Γι' αυτό το λόγο, το k συνήθως λαμβάνεται ως ένας μικρός περιττός αριθμός, ώστε να αποφευχθεί το ενδεχόμενο της ισοψηφίας. Βέβαια, η τιμή που θα έχει το k εξαρτάται και από τα δεδομένα. Γενικά, μεγαλύτερες τιμές του k μειώνουν την επίδραση θορύβου, ωστόσο κάνουν λιγότερο διακριτά τα σύνορα μεταξύ των κλάσεων. Ένας αποτελεσματικός τρόπος για την εύρεση του k είναι η χρήση ευρετικών αλγορίθμων (heuristics).



Εικόνα 6 Παράδειγμα αλγορίθμου k-NN [Πηγή: http://bdewilde.github.io/blog/blogger/2012/10/26/classification-of-hand-written-digits-3/]

1.2.4.6 Συνδυαστικές Μέθοδοι (Ensemble)

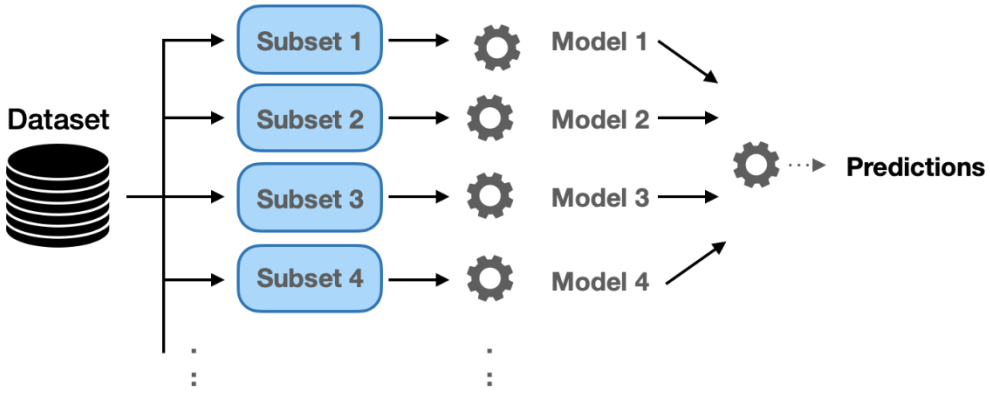
Ο βασικός στόχος των συνδυαστικών μοντέλων μηχανικής μάθησης σε προβλήματα ταξινόμησης είναι η βελτίωση της απόδοσής τους στα σύνολα δεδομένων, μέσω του συνδυασμού των προβλέψεων πολλών διαφορετικών μοντέλων, τα οποία ονομάζονται αδύναμοι ταξινομητές (weak classifiers).

Οι αδύναμοι ταξινομητές χρησιμοποιούνται σαν υπορουτίνες, οι οποίες συνδυάζονται ώστε να συνθέσουν έναν απίστευτα ακριβή ταξινομητή. Για κάθε αδύναμο ταξινομητή, ο συνδυαστικός αλγόριθμος διατηρεί μία κατανομή των βαρών στα πρότυπα του συνόλου εκπαίδευσης, ώστε να έχει κάθε πρότυπο τη δυνατότητα να συνεισφέρει με διαφορετικό τρόπο στο τελικό σφάλμα εκπαίδευσης. Αρχικά, όλα τα βάρη λαμβάνονται ίσα, ωστόσο, σε κάθε επανάληψη τα βάρη των δειγμάτων που

ταξινομούνται λανθασμένα αυξάνονται, έτσι ώστε να αναγκάσουν τον αδύναμο ταξινομητή να επικεντρωθεί σε αυτές τις δύσκολες περιπτώσεις. Μόλις ολοκληρωθεί η διαδικασία, οι απλοί ταξινομητές συνδυάζονται για να προκύψει ένα τελικό μοντέλο, το οποίο, συνήθως, επιτυγχάνει μεγάλο ποσοστό ακρίβειας (accuracy) στο σύνολο των δεδομένων ελέγχου [49].

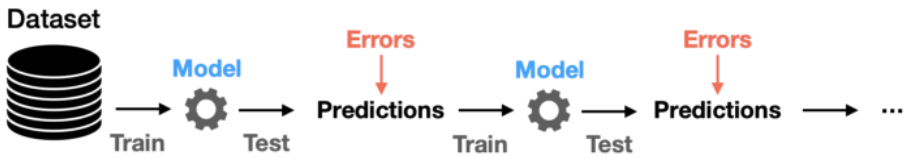
Οι συνδυαστικές μέθοδοι ή μέθοδοι "Ensemble" διακρίνονται σε δύο βασικές κατηγορίες:

- "Bagging": Στην περίπτωση αυτή η διαδικασία της εκπαίδευσης γίνεται παράλληλα για όλα τα μοντέλα που συμμετέχουν για τη δημιουργία του τελικού μοντέλου Ensemble. Κάθε μοντέλο εκπαιδεύεται χρησιμοποιώντας ένα τυχαίο υποσύνολο των δεδομένων.



Εικόνα 7 Μέθοδος Bagging για την εκτέλεση των τελικών προβλέψεων συνδυάζοντας ξεχωριστές προβλέψεις από διαφορετικά μοντέλα [Πηγή: https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c]

- "Boosting": Τα μοντέλα αυτής της κατηγορίας εκπαιδεύονται με σειριακό τρόπο. Κάθε ανεξάρτητο μοντέλο μαθαίνει από τις λάθος προβλέψεις που έχει κάνει το προηγούμενο μοντέλο.



Εικόνα 8 Η σειριακή εκπαίδευση του μοντέλου Adaboost [Πηγή: https://towardsdatascience.com/ensemble-models-5a62d4f4cb0c]

Στα πλαίσια των πειραμάτων της παρούσας διπλωματικής εργασίας χρησιμοποιήθηκε ο αλγόριθμος Random Forest ως "Bagging" μέθοδος και ο αλγόριθμος Adaboost ως "Boosting". Και στις δύο κατηγορίες μοντέλων χρησιμοποιούνται δέντρα απόφασης

και, κατά τη διαδικασίας της εκπαίδευσης, επιλέγονται μόνο τα χαρακτηριστικά που θα αυξήσουν την προβλεπτική ικανότητα των μοντέλων, γεγονός που οδηγεί σε μείωση των διαστάσεων και στη βελτίωση του χρόνου εκτέλεσης, καθώς τα ασήμαντα χαρακτηριστικά δε συμμετέχουν στην υπολογιστική διαδικασία. Οι μέθοδοι Ensemble βελτιώνουν την απόδοση των μοντέλων, καθώς είναι λιγότερο επιρρεπείς στο πρόβλημα της υπερεκπαίδευσης, συγκριτικά με ένα απλό δέντρο απόφασης, καθώς το τελικό αποτέλεσμα είναι ο μέσος όρος όλων των δέντρων απόφασης που χρησιμοποιήθηκαν για τη σύνθεση του μοντέλου. Η μέθοδος Adaboost, ωστόσο, είναι ευαίσθητη στο θόρυβο και στις περιθωριακές τιμές (outliers), ενώ η Random Forest παρουσιάζει πλεονέκτημα, καθώς έχει χαμηλή ευαισθησία σε ακραίες τιμές.

1.2.5 Το φαινόμενο της υπερεκπαίδευσης

Ένα από τα σημαντικότερα προβλήματα που καλούνται να ξεπεράσουν οι τεχνικές μηχανικής μάθησης είναι αυτό της υπερεκπαίδευσης. Η υπερεκπαίδευση συμβαίνει όταν ένα μοντέλο είναι υπερβολικά σύνθετο, όπως, για παράδειγμα, όταν έχει πάρα πολλές παραμέτρους σε σχέση με τον αριθμό των παρατηρήσεων. Στην περίπτωση αυτή το μοντέλο που παράγεται έχει πολύ μικρή προβλεπτική ικανότητα, καθώς αντιδρά με υπερβολικό τρόπο σε μικρές διακυμάνσεις των δεδομένων εκπαίδευσης. Η υπερεκπαίδευση γίνεται αντιληπτή μέσω της παρατήρησης των σφαλμάτων, όπου το σφάλμα κατά την εκπαίδευση του μοντέλου είναι πολύ μικρό, αλλά όταν το μοντέλο εφαρμόζεται σε νέα δεδομένα κατά τη διαδικασία ελέγχου, η τιμή του σφάλματος λαμβάνει υπερβολικά μεγάλη τιμή [51].

1.2.6 Βελτιστοποίηση υπερ-παραμέτρων μοντέλων μηχανικής μάθησης

Στη Μηχανική Μάθηση, η βελτιστοποίηση των υπερ-παραμέτρων (hyperparameter optimization) αποτελεί το πρόβλημα της επιλογής ενός συνόλου βέλτιστων υπερ-παραμέτρων για τον εκάστοτε αλγόριθμο μάθησης. Η υπερ-παράμετρος δεν είναι άλλο από μία παράμετρο που χρησιμοποιείται για τον έλεγχο της διαδικασίας μάθησης.

Ο παραδοσιακός τρόπος βελτιστοποίησης υπερ-παραμέτρων είναι η μέθοδος “Grid Search” η οποία εκτελεί εξαντλητική αναζήτηση για την εύρεση των καλύτερων υπερ-παραμέτρων ενός αλγορίθμου μάθησης σε ένα συγκεκριμένο υποσύνολο του χώρου, το οποίο δηλώνεται από το χρήστη. Η επιλογή του υποσυνόλου πάνω στο οποίο θα εκτελεστεί η αναζήτηση θα πρέπει να γίνεται προσεκτικά, καθώς οι τιμές των υπερ-

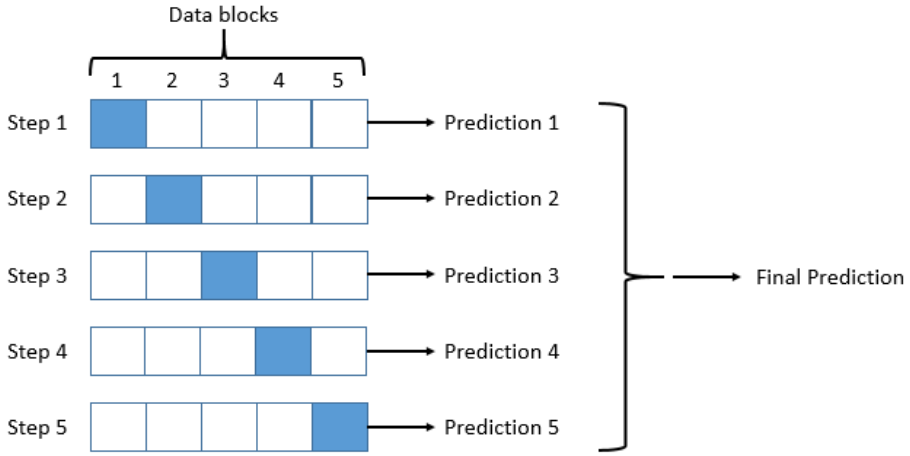
παραμέτρων μπορεί να είναι διακριτές ή συνεχείς. Η επιλογή της βέλτιστης υπερ-παραμέτρου γίνεται μέσω cross-validation, όπου ο αλγόριθμος αξιολογείται σε διαφορετικούς συνδυασμούς υπερ-παραμέτρων, οπότε και η επιλογή της βέλτιστης τιμής γίνεται βάσει κάποιας μετρική απόδοσης, όπως για παράδειγμα της μετρικής "Accuracy" για προβλήματα ταξινόμησης.

Παρόλο που η μέθοδος Grid Search αποτελεί μία απλή και υπολογιστικά ακριβή μέθοδο, προτιμάται συχνά για την εύρεση των βέλτιστων υπερ-παραμέτρων, καθώς δίνει πιο ακριβή αποτελέσματα σε σύγκριση με ευριστικούς αλγορίθμους (heuristics) ή εκτιμητές. Επίσης, καθώς τις περισσότερες φορές οι παράμετροι που δίνονται σαν είσοδος για τη βελτιστοποίηση είναι λίγες σε αριθμό, ο χρόνος που απαιτείται για την εκτέλεση των υπολογισμών δεν είναι σημαντικά χειρότερος από το χρόνο που καταγράφουν άλλες μέθοδοι [22].

1.2.7 Αξιολόγηση των μοντέλων Μηχανικής Μάθησης

Μία από τις μεθόδους που χρησιμοποιούνται για την αξιολόγηση των μοντέλων μηχανικής μάθησης κατά τη διαδικασία εκπαίδευσής τους είναι η μέθοδος «cross-validation». Παραδοσιακά, το μοντέλο μηχανικής μάθησης εκπαιδεύεται και αξιολογείται σε συγκεκριμένα υποσύνολα των δεδομένων, τα οποία καθορίζονται εξαρχής από το χρήστη, όπου συνήθως το 70% ή 80% των αρχικών δεδομένων χρησιμοποιούνται για την εκπαίδευση και το υπόλοιπο 30% ή 20% αντίστοιχα για την αξιολόγηση του μοντέλου, όπου γίνεται τυχαία επιλογή των δειγμάτων που θα περιέχονται σε κάθε υποσύνολο.

Ωστόσο, όταν χρησιμοποιείται η μέθοδος cross-validation, το σύνολο εκπαίδευσης διασπάται σε ανεξάρτητα και ισομεγέθη υποσύνολα, όπου το ένα υποσύνολο χρησιμοποιείται για την αξιολόγηση και η ένωση των υπολοίπων συμμετέχει στην εκπαίδευση του μοντέλου. Το σφάλμα εκπαίδευσης δίνεται λαμβάνοντας το μέσο όρο των σφαλμάτων που έχουν προκύψει από την εκπαίδευση του μοντέλου στα διαφορετικά υποσύνολα. Η μέθοδος ονομάζεται k-fold cross-validation, όπου η παράμετρος k αντιστοιχεί στον αριθμό των ανεξάρτητων υποσυνόλων που θα δημιουργηθούν. Στην εικόνα που ακολουθεί δίνεται ένα παράδειγμα 5-fold cross-validation:



Εικόνα 9 Παράδειγμα 5-fold cross-validation [Πηγή: https://www.researchgate.net/figure/Diagram-of-the-5-fold-cross-validation-method-blocks-in-blue-represent-the-testing-folds_fig1_337447405]

Προκειμένου να αξιολογηθεί η απόδοση των μοντέλων μηχανικής μάθησης που θα αναπτυχθούν στα πλαίσια της παρούσας διπλωματικής εργασίας, θα χρησιμοποιηθούν οι μετρικές Accuracy, Precision, Recall και F-measure, οι οποίες υπολογίζονται με βάση τον πίνακα σύγχυσης (confusion matrix) [40] που ακολουθεί:

		Κλάση πρόβλεψης	
		Κλάση 1	*Κλάση 2*
Πραγματική κλάση	*Κλάση 1*	TP	FN
	Κλάση 2	FP	TN

Πίνακας 1 Πίνακας σύγχυσης για την αξιολόγηση μοντέλων Μηχανικής Μάθησης

Ο πίνακας σύγχυσης βοηθά στην οπτικοποίηση της απόδοσης ενός αλγορίθμου μηχανικής μάθησης, όπου κάθε στήλη του αναπαριστά τα δείγματα της κλάσης πρόβλεψης, ενώ κάθε γραμμή του αντιστοιχεί στα δείγματα της πραγματικής κλάσης. Ο πίνακας σύγχυσης χρησιμοποιείται σε προβλήματα κατηγοριοποίησης, όπου η έξοδος μπορεί να αποτελείται από δύο (binary classification) ή περισσότερες (multi-class classification) κλάσεις. Η γενική μορφή του πίνακα σύγχυσης που δόθηκε παραπάνω αποτελείται από δύο κλάσεις εξόδου, καθώς το υπό μελέτη πρόβλημα αφορά το χαρακτηρισμό μιας είδησης ως ψευδούς ή αληθούς (πχ Κλάση 1 = Ψευδής και Κλάση 2 = Αληθής είδηση). Οι συμβολισμοί TP, FN, FP και TN συμβολίζουν τα εξής:

TP = όσα παραδείγματα ανήκουν στην κλάση 1 και ταξινομήθηκαν ορθά στην κλάση 1

FN = όσα παραδείγματα ανήκουν στην κλάση 1, αλλά ταξινομήθηκαν εσφαλμένα στην κλάση 2

FP = όσα παραδείγματα ανήκουν στην κλάση 2, αλλά ταξινομήθηκαν εσφαλμένα στην κλάση 1

TN = όσα παραδείγματα ανήκουν στην κλάση 2 και ταξινομήθηκαν ορθά στην κλάση 2

Με βάση αυτούς τους δείκτες, υπολογίζονται οι μετρικές αξιολόγησης των μοντέλων, η περιγραφή των οποίων δίνεται παρακάτω:

- Η μετρική Precision δηλώνει το ποσοστό των θετικών παραδειγμάτων που ταξινομήθηκαν σωστά και υπολογίζεται με βάση τον τύπο:

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

Προφανώς, όσο αυξάνεται ο δείκτης Precision, τόσο μειώνεται ο αριθμός των FP, δηλαδή των αρνητικών παραδειγμάτων που ταξινομούνται σωστά.

- Η μετρική Recall είναι το ποσοστό των θετικών παραδειγμάτων που αναγνωρίστηκαν σωστά από τον ταξινομητή και δίνεται από τον παρακάτω τύπο:

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

Όσο μεγαλύτερη η τιμή του Recall, τόσο λιγότερα θετικά παραδείγματα έχουν ταξινομηθεί λάθος.

- Η μετρική Accuracy, που είναι και η πιο συνηθισμένη σε προβλήματα ταξινόμησης, δηλώνει το ποσοστό του συνολικού αριθμού των σωστών προβλέψεων και ο αντίστοιχος τύπος της είναι:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

- Οι μετρικές Precision και Recall δεν μπορούν από μόνες τους να περιγράψουν την αποτελεσματικότητα ενός ταξινομητή, καθώς αν η μία μαρτυρά καλή απόδοση δε σημαίνει ότι θα συμβαίνει το ίδιο και με την άλλη. Για να ξεπεραστεί αυτό το πρόβλημα, χρησιμοποιείται μία τέταρτη μετρική, η F-measure, η οποία ορίζεται ως ο αρμονικός μέσος των μετρικών precision και recall [41], οπότε ο αντίστοιχος τύπος της είναι:

$$F-measure = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall}$$

Όταν η τιμή της μετρικής F-measure προσεγγίζει τη μονάδα, τότε έχει επιτευχθεί ένας καλός συνδυασμός των Precision και Recall [41].

ΚΕΦΑΛΑΙΟ 2. Βιβλιογραφική Ανασκόπηση. Μελέτη εργαλείων και μεθόδων για την εκτίμηση και επικύρωση διαδικτυακού περιεχομένου

Στο κεφάλαιο αυτό θα δοθούν οι βασικές έννοιες του προβλήματος της επικύρωσης της εγκυρότητας του διαδικτυακού περιεχομένου και θα μελετηθεί η σύγχρονη βιβλιογραφία και οι διαφορετικές μέθοδοι Μηχανικής Μάθησης που έχουν προταθεί για την επίλυση του προβλήματος.

2.1 Η ποιότητα των ειδήσεων στο διαδίκτυο

Αν μπορούσαμε να δώσουμε έναν γενικό ορισμό της δημοσιογραφίας αυτός θα ήταν ο εξής [5][24]:

«Η αδιάλειπτη παρουσίαση πληροφοριών και συζητήσεων σχετικά με τις δημόσιες εκδηλώσεις, τάσεις και ζητήματα που διανέμονται μέσω διάφορων μέσων με πρωταρχικό σκοπό την ενημέρωση, την ψυχαγωγία και τη σύνδεση των πολιτών σε κοινότητες».

Ο πρωταρχικός σκοπός της δημοσιογραφίας θα πρέπει να είναι η διάθεση αμερόληπτης και αντικειμενικής πληροφορίας στους πολίτες, έτσι ώστε να είναι οι ίδιοι ελεύθεροι να επεξεργαστούν την πληροφορία και να διαμορφώσουν τη δική τους άποψη. Αυτό συνεπάγεται ότι όσοι παράγουν και διανέμουν τις ειδήσεις πρέπει να υπερασπίζονται και να προωθούν με κάθε τρόπο την αλήθεια [25].

Τη σύγχρονη εποχή, τα μέσα κοινωνικής δικτύωσης ευθύνονται για την ψηφιοποίηση της πληροφορίας και τη συνεχή διάθεσή της ανά πάσα ώρα και στιγμή. Οι πολίτες βασίζονται πολύ στα συγκεκριμένα εργαλεία, καθώς η χρήση είναι πιο φιλική, το κόστος χαμηλό και η αλληλεπίδραση άμεση. Η ποιότητα των ειδήσεων συχνά συνδέεται άμεσα με οικονομικές σκοπιμότητες, καθώς υποστηρίζεται ότι η παραγωγή ορθών ειδήσεων είναι δαπανηρή, ενώ οι ανακριβείς ή αλλιώς ψευδείς πληροφορίες διαχέονται με μεγαλύτερη ταχύτητα και με μικρότερο κόστος, καθώς δε γίνεται διασταύρωση της εγκυρότητάς τους. Επίσης, η έλλειψη ποιότητας στις ειδήσεις και η διασπορά του ψευδούς περιεχομένου ενισχύεται από τη δυνατότητα που έχει ο καθένας να διαχέει ή να αναπαράγει μία είδηση σε διαδικτυακές πλατφόρμες, χωρίς να γίνεται κάποιος έλεγχος της ορθότητας πρώτα.

2.2 Η παραπληροφόρηση στα σύγχρονα μέσα κοινωνικής δικτύωσης

Υπήρξε μια εποχή, που αν κάποιος ήθελε να μάθει τα νέα, τότε θα έπρεπε να περιμένει την εφημερίδα της επόμενης ημέρας. Ωστόσο, τη σύγχρονη εποχή, η ανάπτυξη των ηλεκτρονικών εφημερίδων, τα μέσα κοινωνικής δικτύωσης και άλλα διαδικτυακά μέσα ενημέρωσης έχουν γίνει οι κύριες πηγές που τροφοδοτούν ειδήσεις συνεχώς σε οποιαδήποτε ώρα της ημέρας.

Είτε πρόκειται για τις προεδρικές εκλογές, την αλλαγή του κλίματος ή τον ιό Covid-19, η παραπληροφόρηση συνεχίζει να εξαπλώνεται ανεξέλεγκτα στα κοινωνικά μέσα. Ενώ μπορούμε να περιμένουμε, ακόμη και να απαιτήσουμε από τις κοινωνικές πλατφόρμες να πατάξουν την παραπληροφόρηση, η πιθανότητα να εξαλειφθεί αυτή είναι πολύ μικρή. Ένας λόγος είναι ότι θα χρειαζόταν πλήρης έλεγχος σχεδόν όλου του περιεχομένου, το οποίο μπορεί να επιτευχθεί μόνο με αυτοματοποιημένο τρόπο [54].

Άλλωστε, η παραπληροφόρηση παίρνει ‘κλικ’, για αυτό και βλέπουμε τόση παραπληροφόρηση στις σύγχρονες πλατφόρμες, οι οποίες επωφελούνται από αυτό, καθώς όσο πιο ακραίο είναι το περιεχόμενο, τόσο περισσότεροι άνθρωποι αλληλοεπιδρούν με αυτό, χωρίς να έχει σημασία αν είναι αληθές ή ψευδές.

Υπάρχουν, επίσης, πολλά άτομα που διανέμουν σκόπιμα ανακριβείς πληροφορίες σε μια προσπάθεια να επηρεάσουν την κοινή γνώμη, όπως για παράδειγμα στην περίπτωση των εκλογών. Η υπονόμευση της κοινωνικής εμπιστοσύνης μπορεί να διαταράξει την οικονομία, να επηρεάσει την απασχόληση, καθώς και άλλους τομείς, όπως είναι η δημόσια υγεία. Ορισμένες φορές μάλιστα οι ψευδείς ειδήσεις κατασκευάζονται για να προκαλέσουν σύγχυση και να αποπροσανατολίσουν ή διαστρεβλώσουν την ορθή κρίση των αναγνωστών [5].

2.3 Η έννοια «ψευδείς ειδήσεις»

Οι ψευδείς ειδήσεις είναι φαινόμενο που υπάρχει εδώ και πολύ καιρό, σχεδόν το ίδιο χρονικό διάστημα που οι ειδήσεις άρχισαν να κυκλοφορούν ευρέως μετά την εφεύρεση της τυπογραφίας το 1439. Ωστόσο, δεν υπάρχει καμία συμφωνημένη ονομασία του όρου ‘ψευδείς ειδήσεις’.

Μια στενή έννοια των ψευδών ειδήσεων είναι τα ειδησεογραφικά άρθρα που είναι σκόπιμα και αποδεδειγμένα ψευδή και θα μπορούσαν να παραπλανήσουν τους

αναγνώστες. Υπάρχουν δύο βασικά χαρακτηριστικά αυτής της έννοιας: η αυθεντικότητα και η πρόθεση. Πρώτον, οι ψευδείς ειδήσεις περιλαμβάνουν μη αληθείς πληροφορίες που μπορούν να επαληθευτούν ως τέτοιες. Δεύτερον, οι ψευδείς ειδήσεις δημιουργούνται με ανέντιμη πρόθεση να παραπλανήσουν τους καταναλωτές. Βάσει αυτών, θα δώσουμε τον παρακάτω ορισμό για τις ψευδείς ειδήσεις [26]:

Ψευδής είδηση είναι ένα ειδησεογραφικό άρθρο που είναι σκόπιμα και αποδεδειγμένα ψευδές.

Σύμφωνα με τον παραπάνω ορισμό, οι παρακάτω ειδήσεις δε θεωρούνται ψευδείς:

- Ειδήσεις σάτιρας, οι οποίες δεν έχουν πρόθεση να παραπλανήσουν ή να εξαπατήσουν καταναλωτές και είναι απίθανο να εκληφθούν εσφαλμένα ως κάποιο γεγονός.
- Φήμες που δεν προέρχονται από ειδησεογραφικά γεγονότα.
- Θεωρίες συνωμοσίας, οι οποίες είναι δύσκολο να επαληθευτούν ως αληθείς ή ψευδείς.
- Παραπληροφόρηση που δημιουργείται ακούσια.
- Φάρσες που έχουν ως μόνο κίνητρο τη διασκέδαση ή την εξαπάτηση στοχευμένων ατόμων.

2.4 State-of-the-art μέθοδοι Μηχανικής Μάθησης για την ανίχνευση ψευδών ειδήσεων

Η ανίχνευση των ψευδών ειδήσεων αποτελεί ένα ενδιαφέρον πρόβλημα για μελέτη, καθώς ο χώρος της Μηχανικής Μάθησης προσφέρει πολλές διαφορετικές μεθόδους που μπορούν να χρησιμοποιηθούν προς αυτήν την κατεύθυνση και τα δεδομένα που χρειάζονται για την εκπαίδευση των μοντέλων είναι άφθονα. Ουσιαστικά, το πρόβλημα που καλούνται να επιλύσουν οι αλγόριθμοι μηχανικής μάθησης είναι αυτό της κατηγοριοποίησης κειμένου (text classification). Παρακάτω δίνονται οι πιο βασικές μέθοδοι που συναντιούνται στη βιβλιογραφία για το υπό μελέτη πρόβλημα:

- **Μηχανή διανυσμάτων υποστήριξης (SVM)**: Οι ερευνητές στο [27] κατόπιν δοκιμών με διάφορους ταξινομητές μηχανικής μάθησης, διαπίστωσαν ότι η μέθοδος SVM τους έδωσε τα καλύτερα αποτελέσματα στον εντοπισμό των ψευδών ειδήσεων.

- **Naive Bayes**: Οι ερευνητές στο [28] χρησιμοποίησαν αυτόν τον ταξινομητή μηχανικής μάθησης για να κατηγοριοποιήσουν μία είδηση ως ψευδή ή αληθή. Όπως διαπιστώθηκε και στο [33], μία παραδοσιακή αρχιτεκτονική μηχανικής μάθησης, όπως ο Naive Bayes, μπορεί να επιτύχει πολύ υψηλή ακρίβεια με κατάλληλη επιλογή χαρακτηριστικών. Σε ένα μικρό σύνολο δεδομένων με λιγότερα από 100 χιλιάδες ειδησεογραφικά άρθρα, το Naive Bayes (με n-gram διανύσματα) μπορεί να αποτελέσει πρωταρχική επιλογή, καθώς επιτυγχάνει παρόμοια απόδοση σε σύγκριση με τα νευρωνικά δίκτυα που βασίζονται σε μοντέλα υψηλής επιβάρυνσης.

- **Λογιστική παλινδρόμηση**: Αυτός ο ταξινομητής χρησιμοποιείται όταν η τιμή που πρέπει να προβλεφθεί είναι κατηγορική. Οι ερευνητές στο [29] χρησιμοποίησαν αυτόν τον ταξινομητή για να ανιχνεύσουν ψευδείς ειδήσεις.

- **Τυχαία δάση**: Στο [33] οι ερευνητές χρησιμοποίησαν διαφορετικούς ταξινομητές μηχανικής μάθησης για να ανιχνεύσουν τις ψευδείς ειδήσεις, μία εκ των οποίων ήταν τα τυχαία δάση που παρουσίασαν καλή απόδοση.

- **Νευρωνικό δίκτυο**: Οι ερευνητές στο [30] χρησιμοποίησαν νευρωνικά δίκτυα για την ανίχνευση ψευδών ειδήσεων. Τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks) είναι επίσης αρκετά δημοφιλή σε αντίστοιχα προβλήματα, ιδίως όσο αυξάνεται η πολυπλοκότητα, καθώς σε μικρά σύνολα δεδομένων θα έχουν φτωχή απόδοση.

- **k-Nearest Neighbor**: Πρόκειται για έναν επιβλεπόμενο αλγόριθμο μηχανικής μάθησης που χρησιμοποιείται για την επίλυση των προβλημάτων ταξινόμησης ο οποίος έχει χρησιμοποιηθεί ευρέως και σε προβλήματα εντοπισμού ψευδών ειδήσεων στα μέσα κοινωνικής δικτύωσης [31].

- **Δέντρο αποφάσεων**: Αυτός ο επιβλεπόμενος αλγόριθμος μηχανικής μάθησης μπορεί να βοηθήσει στον εντοπισμό των ψευδών ειδήσεων. Το βασικό του χαρακτηριστικό είναι ότι διαχωρίζει το σύνολο δεδομένων σε μικρότερα υποσύνολα [32].

Ένα βασικό χαρακτηριστικό των υπαρχουσών μελετών είναι ότι επικεντρώνονται στην ανίχνευση ειδήσεων συγκεκριμένων τύπων (όπως για παράδειγμα ειδήσεις από τον πολιτικό χώρο), οπότε κατά συνέπεια, αναπτύσσουν μοντέλα και σχεδιάζουν χαρακτηριστικά για συγκεκριμένα σύνολα δεδομένων που ίσως δεν έχουν την ίδια

προβλεπτική ικανότητα αν εφαρμοστούν σε δεδομένα διαφορετικού ειδησεογραφικού περιεχομένου. Ένας άλλος περιορισμός των συγκριτικών μελετών είναι ότι αυτές έχουν επικεντρωθεί σε έναν περιορισμένο αριθμό χαρακτηριστικών που έχουν ως αποτέλεσμα την ελλιπή διερεύνηση πιθανών σημαντικών χαρακτηριστικών πάνω στο εκάστοτε σύνολο δεδομένων [33].

Ο Wang [14] δημιούργησε ένα σύνολο δεδομένων αναφοράς, το Liar, και πειραματίστηκε με ορισμένα υπάρχοντα μοντέλα σε αυτό. Το αποτέλεσμα της σύγκρισης υποδεικνύει το βαθμό απόδοσης διάφορων μοντέλων σε ένα δομημένο σύνολο δεδομένων, όπως το Liar. Ωστόσο, το μέγεθος αυτού του συνόλου δεδομένων δεν είναι επαρκές για την ανάλυση νευρωνικών δικτύων και διαπιστώθηκε ότι ορισμένα μοντέλα υποφέρουν από το φαινόμενο της υπερ-εκπαίδευσης.

Αρκετές ερευνητικές εργασίες δείχνουν πολλά υποσχόμενα αποτελέσματα στην ανίχνευση ψευδών ειδήσεων μέσω νευρωνικών δικτύων. Ο Wang [14] στην έρευνά του έχει κατασκευάσει ένα υβριδικό μοντέλο νευρωνικού δικτύου με συνελικτικό τρόπο (convolutional neural network) που υπερτερεί έναντι άλλων παραδοσιακών μοντέλων μηχανικής μάθησης. Οι ερευνητές στο [35] έχουν πραγματοποιήσει μια εκτεταμένη ανάλυση των γλωσσολογικών χαρακτηριστικών και ανέδειξαν το εντυπωσιακό αποτέλεσμα των LSTM συνελικτικών νευρωνικών δικτύων.

Η εκπαίδευση των ταξινομητών της μηχανικής μάθησης είναι ένα σημαντικό έργο, καθώς κάθε μοντέλο πρέπει να εκπαιδεύεται με τα κατάλληλα δεδομένα, ώστε να μπορεί να έχει καλή ικανότητα γενίκευσης σε διαφορετικά δεδομένα που του τροφοδοτούνται. Το κύριο πρόβλημα που εμφανίζεται κατά την εκπαίδευση των μοντέλων είναι ότι σε αρκετές περιπτώσεις οι κλάσεις εξόδου μπορεί να μην έχουν καλή αναλογία ή οι ειδήσεις να προέρχονται από την ίδια πηγή, όπου το στυλ γραφής θα είναι παρόμοιο, οπότε δεν υπάρχει ποικιλομορφία στην πληροφορία που χρησιμοποιείται για την εκμάθηση του ταξινομητή [33].

Πολλοί ερευνητές δίνουν βαρύτητα στην εξαγωγή χαρακτηριστικών έτσι ώστε να ενισχύσουν την απόδοση των μοντέλων τους, όπως στο [36], όπου χρησιμοποίησαν κατά την εκπαίδευση των ταξινομητών τρεις διαφορετικές μεθόδους για την εξαγωγή χαρακτηριστικών: την TF-IDF, τη μέθοδο N-Gram και το μοντέλο Bag of Words, τα οποία τροφοδοτούνται σαν είσοδος στους ταξινομητές. Τέλος, πολλές ερευνητικές εργασίες προτείνουν, επίσης, τη χρήση της ανάλυσης συναισθήματος για την εξαγωγή

χαρακτηριστικών, καθώς μπορεί να βρεθεί συσχέτιση μεταξύ του συναισθήματος που αναδύεται από κάποιο άρθρο και τον τύπο αυτού [33]. Ορισμένοι, μάλιστα, χρησιμοποιούν σε συνδυασμό με τα λεξιλογικά χαρακτηριστικά και χαρακτηριστικά συναισθήματος, όπως για παράδειγμα αν μία είδηση είναι θετικά ή αρνητικά πολωμένη [34].

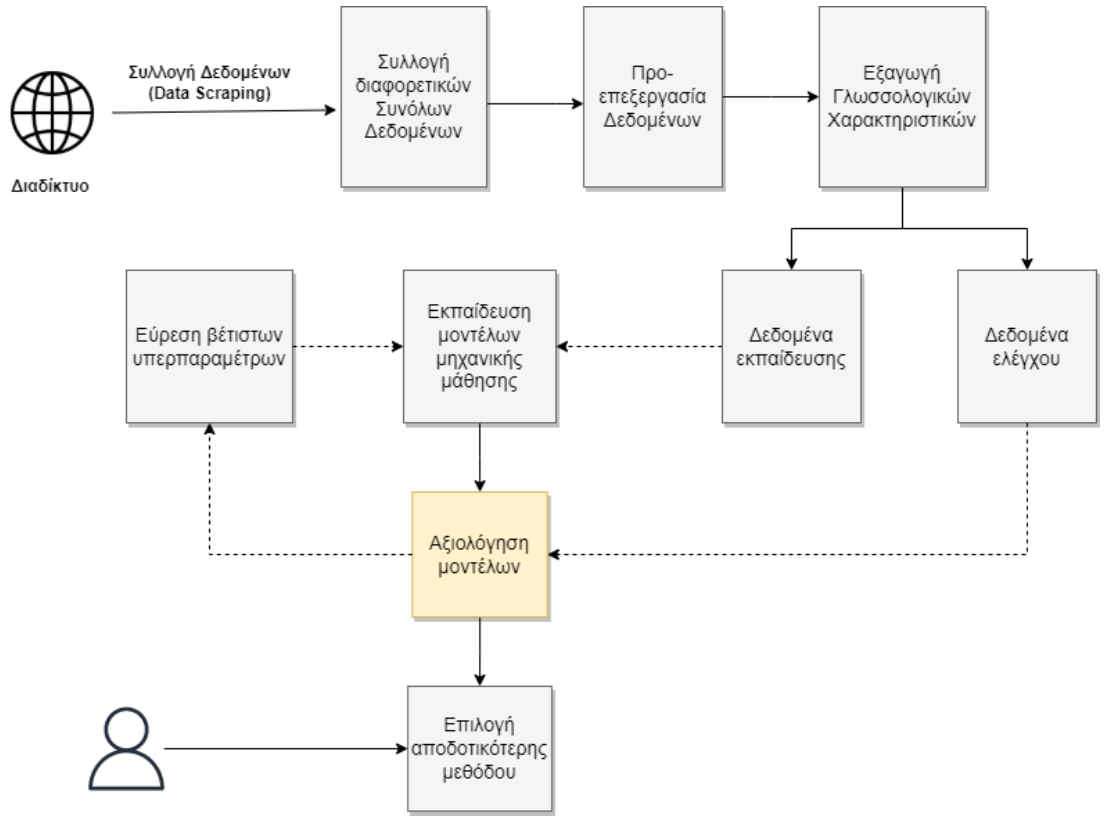
Όπως μπορεί να γίνει αντιληπτό, το πρόβλημα ανίχνευσης ψευδών ειδήσεων έχει μελετηθεί εκτενώς στη βιβλιογραφία και, όπως φαίνεται, οι διαφορετικές μέθοδοι μηχανικής μάθησης δίνουν πολύ υποσχόμενα αποτελέσματα. Οι πιο δημοφιλείς μέθοδοι θα χρησιμοποιηθούν και στην παρούσα διπλωματική εργασία όπου συνδυαστικά με διαφορετικά σύνολα δεδομένων θα αξιολογηθούν ως προς την απόδοσή τους.

ΚΕΦΑΛΑΙΟ 3. Συλλογή Δεδομένων και Μεθοδολογία

Το πειραματικό κομμάτι της παρούσας διπλωματικής εργασίας θα εκτελεστεί στη γλώσσα προγραμματισμού Python. Η Python χαρακτηρίζεται από ευκολία χρήσης και ευελιξία, καθώς διαθέτει πολυάριθμες βιβλιοθήκες, μέσω των οποίων είναι η δυνατή η δημιουργία ολοκληρωμένων προγραμμάτων μέσα σε λίγες γραμμές κώδικα. Στα πλαίσια της μηχανικής μάθησης, διαθέτει τη βιβλιοθήκη scikit-learn, η οποία περιλαμβάνει έτοιμες υλοποιήσεις αλγορίθμων ταξινόμησης, παλινδρόμησης και ομαδοποίησης, συμπεριλαμβανομένων των SVM, Random Forests, τεχνητών νευρωνικών δικτύων, k-means, DBSCAN, κτλ. Η βιβλιοθήκη scikit-learn έχει σχεδιαστεί με τέτοιο τρόπο ώστε να είναι συμβατή με τις υπόλοιπες βιβλιοθήκες της Python και αποτελεί ένα ισχυρό εργαλείο στα χέρια των αναλυτών και των επιστημόνων μηχανικής μάθησης.

Ο κώδικας υλοποιήθηκε σε Python 3.8, χρησιμοποιώντας το περιβάλλον Anaconda, μία ανοικτού κώδικα πλατφόρμα, η οποία παρέχει τη δυνατότητα άμεσης εγκατάστασης και χρήσης μίας πληθώρας έτοιμων βιβλιοθηκών.

Τα παρακάτω βήματα ακολουθήθηκαν κατά την εκτέλεση των πειραμάτων, τα οποία θα αναλυθούν στις επόμενες ενότητες:



Εικόνα 10 Διάγραμμα μεθοδολογίας

- Συλλογή δεδομένων που βρίσκονται διαθέσιμα διαδικτυακά (publicly available) και έχουν χρησιμοποιηθεί στο παρελθόν σε παρόμοιες μελέτες επικύρωσης της εγκυρότητας διαδικτυακού περιεχομένου.
- Προ-επεξεργασία των δεδομένων και μετασχηματισμός αυτών (εξαγωγή χαρακτηριστικών, κτλ.) έτσι ώστε να έρθουν σε μορφή που να εξυπηρετεί τους αλγορίθμους ανάλυσής τους.
- Επιλογή και παραμετροποίηση των διαφορετικών αλγορίθμων μηχανικής μάθησης.
- Εκπαίδευση των διαφορετικών μοντέλων πάνω στα διάφορα σύνολα δεδομένων και επιλογή των βέλτιστων υπερ-παραμέτρων (hyper parameter tuning) κατά τη διαδικασία της εκπαίδευσης.
- Αξιολόγηση των μοντέλων ως προς την απόδοσή της και στα τέσσερα σύνολα δεδομένων χρησιμοποιώντας διαφορετικές μετρικές. Επιλογή της αποδοτικότερης μεθόδου για την αξιολόγηση και επαλήθευση διαδικτυακού περιεχομένου.

3.1 Σύνολα δεδομένων

Για την εκτέλεση του πειραματικού μέρους, έχουν επιλεγεί τέσσερα διαφορετικά σύνολα δεδομένων τα οποία έχουν χρησιμοποιηθεί σε προηγούμενες μελέτες ανίχνευσης ψευδών ειδήσεων με μεθόδους Μηχανικής Μάθησης. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν περιέχουν τόσο ψευδείς όσο και αληθείς ειδήσεις που προέρχονται από άρθρα ποικίλων περιοχών ενδιαφέροντος, όπως οικονομία, πολιτική, κτλ. Οι αληθείς ειδήσεις περιέχουν την περιγραφή γεγονότων του πραγματικού κόσμου, ενώ οι ψευδείς ειδήσεις περιέχουν δηλώσεις που δε συσχετίζονται με γεγονότα, οπότε δεν έχει αποδειχτεί η εγκυρότητά τους.

Σύνολο Δεδομένων 1 (DS1)

Το πρώτο σύνολο δεδομένων καλείται “ISOT Fake News Dataset” [13] το οποίο περιέχει αληθείς και ψευδείς ειδήσεις που έχουν ανακτηθεί από τον Παγκόσμιο Ιστό. Οι αληθείς ειδήσεις προήλθαν από τον αναγνωρισμένο διαδικτυακό ιστότοπο *reuters.com*, ενώ οι ψευδείς ειδήσεις ανακτήθηκαν από πολλαπλές πηγές και κυρίως, από διαδικτυακούς τόπους που έχουν σημανθεί από τον ιστότοπο *politifact.com*. Το σύνολο δεδομένων περιέχει συνολικά 44.898 άρθρα, εκ των οποίων 21.417 είναι άρθρα με αληθές περιεχόμενο, ενώ 23.481 ειδήσεις είναι ψευδείς. Το σύνολο των άρθρων περιλαμβάνει ειδήσεις από διάφορες περιοχές, αν και το μεγαλύτερο ποσοστό αφορά ειδήσεις πολιτικού περιεχομένου.

Σύνολο Δεδομένων 2 (DS2)

Το δεύτερο σύνολο δεδομένων είναι διαθέσιμο στη διαδικτυακή πλατφόρμα Kaggle [44] και περιέχει στο σύνολο 20.386 άρθρα για την εκπαίδευση των μοντέλων μηχανικής μάθησης (training set) και 5.126 άρθρα για την επικύρωση (testing set). Το συγκεκριμένο σύνολο δεδομένων έχει χτιστεί από πολλαπλές πηγές του διαδικτύου και περιέχει τόσο αληθή όσο και ψευδή άρθρα από διαφορετικές περιοχές ειδησεογραφικού ενδιαφέροντος.

Σύνολο Δεδομένων 3 (DS3)

Το τρίτο σύνολο δεδομένων είναι επίσης διαθέσιμο στον ιστότοπο Kaggle [45] και περιέχει συνολικά 3.352 άρθρα, τόσο αληθή όσο και ψευδή. Τα αληθή άρθρα προέρχονται από αξιόπιστες ειδησεογραφικές πηγές όπως είναι το CNN, Reuters, New York Times, κτλ., ενώ οι ψευδείς ειδήσεις προέρχονται από μη-αξιόπιστους

διαδικτυακούς τόπους ειδήσεων. Οι περιοχές που καλύπτουν είναι τα αθλητικά, η διασκέδαση και η πολιτική.

Σύνολο Δεδομένων 4 (DS4)

Το τελευταίο σύνολο δεδομένων είναι το FakeNewsNet [21] το οποίο έχει συλλέξει τόσο ψευδείς όσο και αληθείς ειδήσεις από τις πλατφόρμες PolitiFact και GossipCop. Στα πλαίσια της παρούσας διπλωματικής θα χρησιμοποιηθούν μόνο τα δεδομένα από το PolitiFact τα οποία παρέχονται σε δύο διαφορετικά αρχεία (.csv), ένα για τις αληθείς και ένα για τις ψευδείς ειδήσεις [48]. Ο αριθμός των συνολικών ειδήσεων είναι 1.056.

3.2 Προ-επεξεργασία και Μετασχηματισμός των δεδομένων

Τα αρχικά σύνολα δεδομένων υποβλήθηκαν σε προ-επεξεργασία πριν τροφοδοτηθούν στα μοντέλα πρόβλεψης. Οι τεχνικές που χρησιμοποιήθηκαν έχουν αναλυθεί σε προηγούμενο κεφάλαιο και εφαρμόστηκαν με βάση υπάρχουσες βιβλιοθήκες της Python. Για την παρουσίαση των τεχνικών προ-επεξεργασίας χρησιμοποιήθηκαν ενδεικτικά τέσσερις ειδήσεις από το Σύνολο Δεδομένων (1), η αρχική μορφή των οποίων είναι η παρακάτω:

```
Είδηση 1:  As U.S. budget fight looms, Republicans flip their fiscal script
Είδηση 2:  U.S. military to accept transgender recruits on Monday: Pentagon
Είδηση 3:  Senior U.S. Republican senator: 'Let Mr. Mueller do his job'
Είδηση 4:  FBI Russia probe helped by Australian diplomat tip-off: NYT
```

- **LowerCasing**: Προκειμένου να μετατραπούν όλοι οι χαρακτήρες του κειμένου σε πεζούς χαρακτήρες χρησιμοποιήθηκε η συνάρτηση lower() της Python.

```
Είδηση 1:  as u.s. budget fight looms, republicans flip their fiscal script
Είδηση 2:  u.s. military to accept transgender recruits on monday: pentagon
Είδηση 3:  senior u.s. republican senator: 'let mr. mueller do his job'
Είδηση 4:  fbi russia probe helped by australian diplomat tip-off: nyt
```

- **Tokenization**: Για την τεχνική αυτή χρησιμοποιήθηκε το πακέτο **Word_Tokenize** της βιβλιοθήκης NTLK. Παρακάτω δίνεται το αποτέλεσμα της εφαρμογής του tokenization όπου κάθε είδηση αποτελεί ένα διάνυσμα με λέξεις:

```
Είδηση 1:  ['as', 'u.s.', 'budget', 'fight', 'looms', ',', 'republicans', 'flip', 'their', 'fiscal', 'script']
Είδηση 2:  ['u.s.', 'military', 'to', 'accept', 'transgender', 'recruits', 'on', 'monday', ':', 'pentagon']
Είδηση 3:  ['senior', 'u.s.', 'republican', 'senator', ':', "'let", 'mr.', 'mueller', 'do', 'his', 'job', "'"]
Είδηση 4:  ['fbi', 'russia', 'probe', 'helped', 'by', 'australian', 'diplomat', 'tip-off', ':', 'nyt']
```

- **Stemming**: Στην περίπτωση αυτή χρησιμοποιήθηκε το πακέτο **PorterStemmer** της βιβλιοθήκης NTLK. Παρακάτω δίνεται ξανά το παράδειγμα του tokenization, αφού εφαρμόστηκε και η τεχνική stemming. Παρατηρούμε ότι σε κάποιες λέξεις έχουν δοθεί πλέον οι ρίζες τους γεγονός που θα αποτρέψει και την ύπαρξη πολλαπλών λέξεων με το ίδιο νόημα και κατά συνέπεια την ελάττωση του θορύβου στην τελική πληροφορία που θα τροφοδοτηθεί στα μοντέλα:

```
Είδηση 1:  ['as', 'u.s.', 'budget', 'fight', 'loom', ',', 'republican', 'flip', 'their', 'fiscal', 'script']
Είδηση 2:  ['u.s.', 'militari', 'to', 'accept', 'transgend', 'recruit', 'on', 'monday', ':', 'pentagon']
Είδηση 3:  ['senior', 'u.s.', 'republican', 'senat', ':', "'let", 'mr.', 'mueller', 'do', 'hi', 'job', "'"]
Είδηση 4:  ['fbi', 'russia', 'probe', 'help', 'by', 'australian', 'diplomat', 'tip-off', ':', 'nyt']
```

Στο σημείο αυτό να σημειωθεί ότι η τεχνική Lemmatization δε χρησιμοποιήθηκε στα δεδομένα μας γιατί διαπιστώθηκε ότι κατόπιν εφαρμογής του πακέτου WordNetLemmatizer της βιβλιοθήκης NTLK, τα αποτελέσματα ήταν σχεδόν τα ίδια, χωρίς να λαμβάνουμε κάποια βελτίωση στην απόδοση των μοντέλων. Άλλωστε, αυτές οι δύο τεχνικές είναι σχεδόν ίδιες, όπως αναφέρθηκε σε προηγούμενο κεφάλαιο, οπότε έγινε η επιλογή της τεχνικής Stemming, καθώς είναι γρηγορότερη και δίνει πιο εύρωστα διανύσματα λέξεων.

- **Stop word removal**: Το επόμενο βήμα στην προ-επεξεργασία των δεδομένων είναι το Stop word removal, όπου αφαιρέθηκαν ειδικοί χαρακτήρες, όπως σημεία στίξης, και λέξεις, όπως άρθρα και αντωνυμίες, που δεν προσδίδουν καμία αξία στην εκπαίδευση των μοντέλων.

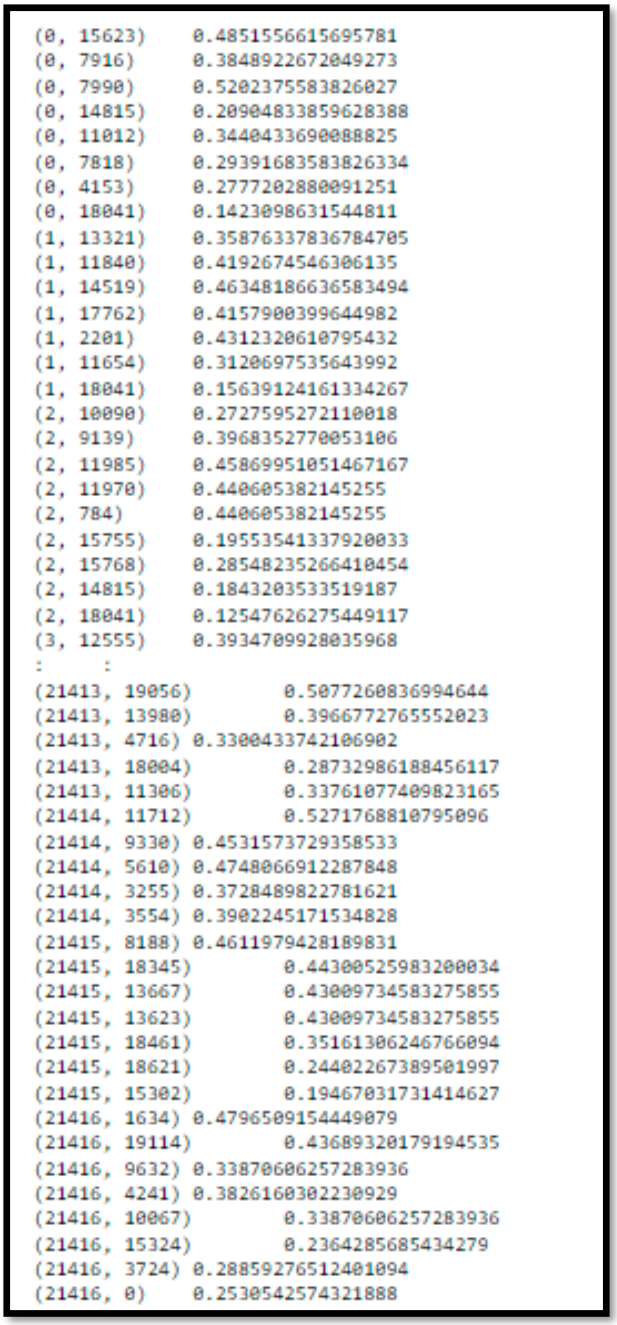
```
Είδηση 1:  ['u.s.', 'budget', 'fight', 'loom', 'republican', 'flip', 'fiscal', 'script']
Είδηση 2:  ['u.s.', 'militari', 'accept', 'transgend', 'recruit', 'monday', 'pentagon']
Είδηση 3:  ['senior', 'u.s.', 'republican', 'senat', "'let", 'mr.', 'mueller', 'hi', 'job']
Είδηση 4:  ['fbi', 'russia', 'probe', 'help', 'australian', 'diplomat', 'tip-off', 'nyt']
```

- **Part-of-speech Tagging**: Η διαδικασία αυτή χρησιμοποιήθηκε για τον προσδιορισμό των μερών του λόγου των λέξεων που αντιστοιχούν σε κάθε διάνυσμα είδησης, όπως φαίνεται στο παρακάτω παράδειγμα:

```
Είδηση 1:  [('u.s.', 'JJ'), ('budget', 'NN'), ('fight', 'NN'), ('loom', 'NN'), ('republican', 'JJ'), ('flip', 'NN'), ('fiscal',
'JJ'), ('script', 'NN')]

Είδηση 2:  [('u.s.', 'JJ'), ('militari', 'NN'), ('accept', 'IN'), ('transgend', 'NN'), ('recruit', 'NN'), ('monday', 'NN'), ('p
entagon', 'NN')]

Είδηση 3:  [('senior', 'JJ'), ('u.s.', 'JJ'), ('republican', 'JJ'), ('senat', 'NN'), ("'let", 'POS'), ('mr.', 'NN'), ('muelle
r', 'NNP'), ('hi', 'VBD'), ('job', 'NN')]

Είδηση 4:  [('fbi', 'NN'), ('russia', 'NN'), ('probe', 'NN'), ('help', 'NN'), ('australian', 'JJ'), ('diplomat', 'NN'), ('tip-o
ff', 'NN'), ('nyt', 'NN')]
```

Παρατηρούμε ότι σε κάθε λέξη έχει αντιστοιχηθεί και ένας χαρακτηρισμός, όπως για παράδειγμα η λέξη ‘budget’ έχει χαρακτηριστεί ως ‘NN’, δηλαδή ως ουσιαστικό. Ο αναλυτικός πίνακας με το τι σημαίνει κάθε συμβολισμός μπορεί να βρεθεί στο Παράρτημα.

- **Μετασχηματισμός TF-IDF**: Το τελευταίο βήμα στην επεξεργασία των δεδομένων αφορά την εξαγωγή των χαρακτηριστικών μέσω της TF-IDF στατιστικής. Τα διανύσματα χαρακτηριστικών που θα υπολογιστούν θα αποτελέσουν την είσοδο στα μοντέλα πρόβλεψης. Όπως φαίνεται στο παρακάτω παράδειγμα, το οποίο έχει ληφθεί από το περιβάλλον Jupyter όπου έχει τρέξει το πειραματικό μέρος, επιστρέφεται ένας πίνακας μεγέθους (NxM) όπου N: ο αριθμός των δειγμάτων/εγγράφων στο σύνολο δεδομένων και M: ο αριθμός των διαφορετικών χαρακτηριστικών/λέξεων που εντοπίζονται σε όλο το σώμα κειμένου. Κάθε τιμή του πίνακα αντιστοιχεί στην TF-IDF μετρική που έχει αποδοθεί σε κάθε λέξη στο σύνολο των δειγμάτων.



```
(0, 15623)    0.4851556615695781
(0, 7916)     0.3848922672049273
(0, 7990)     0.5202375583826027
(0, 14815)    0.20904833859628388
(0, 11012)    0.3440433690088825
(0, 7818)     0.29391683583826334
(0, 4153)     0.2777202880091251
(0, 18041)    0.1423098631544811
(1, 13321)    0.35876337836784705
(1, 11840)    0.4192674546306135
(1, 14519)    0.46348186636583494
(1, 17762)    0.4157900399644982
(1, 2201)     0.4312320610795432
(1, 11654)    0.3120697535643992
(1, 18041)    0.15639124161334267
(2, 10090)    0.2727595272110018
(2, 9139)     0.3968352770053106
(2, 11985)    0.45869951051467167
(2, 11970)    0.440605382145255
(2, 784)      0.440605382145255
(2, 15755)    0.19553541337920033
(2, 15768)    0.28548235266410454
(2, 14815)    0.1843203533519187
(2, 18041)    0.12547626275449117
(3, 12555)    0.3934709928035968
:     :
(21413, 19056)        0.5077260836994644
(21413, 13980)        0.3966772765552023
(21413, 4716) 0.3300433742106902
(21413, 18004)        0.28732986188456117
(21413, 11306)        0.33761077409823165
(21414, 11712)        0.5271768810795096
(21414, 9330) 0.4531573729358533
(21414, 5610) 0.4748066912287848
(21414, 3255) 0.3728489822781621
(21414, 3554) 0.3902245171534828
(21415, 8188) 0.4611979428189831
(21415, 18345)        0.44300525983200034
(21415, 13667)        0.43009734583275855
(21415, 13623)        0.43009734583275855
(21415, 18461)        0.35161306246766094
(21415, 18621)        0.24402267389501997
(21415, 15302)        0.19467031731414627
(21416, 1634) 0.4796509154449079
(21416, 19114)        0.43689320179194535
(21416, 9632) 0.33870606257283936
(21416, 4241) 0.3826160302230929
(21416, 10067)        0.33870606257283936
(21416, 15324)        0.2364285685434279
(21416, 3724) 0.28859276512401094
(21416, 0)    0.2530542574321888
```

3.3 Επιλογή και παραμετροποίηση διαφορετικών μοντέλων Μηχανικής Μάθησης

Στην ενότητα αυτή θα δοθούν οι λεπτομέρειες των μοντέλων μηχανικής μάθησης που χρησιμοποιήθηκαν στη γλώσσα προγραμματισμού Python. Η θεωρητική ανάλυση των αλγορίθμων έχει ήδη δοθεί σε προηγούμενο κεφάλαιο. Τα μοντέλα που χρησιμοποιήθηκαν προέρχονται από τη βιβλιοθήκη sklearn της Python και οι πληροφορίες τους μπορούν να βρεθούν συγκεντρωτικά στον παρακάτω πίνακα. Για

περισσότερες λεπτομέρειες ο αναγνώστης μπορεί να ανατρέξει στην επίσημη σελίδα του scikit-learn της python [52].

Μοντέλο Μηχανικής Μάθησης	**Python βιβλιοθήκη**	**Παράμετροι**	**Εύρεση βέλτιστων υπερ-παραμέτρων**
Λογιστική Παλινδρόμηση (Logistic regression)	*sklearn.linear_model.LogisticRegression*	• **Παράμετρος penalty:** Προεπιλεγμένη τιμή = L2 Η τιμή πέναλτι L2 δηλώνει την "Ridge" παλινδρόμηση, όπου χρησιμοποιούνται όλες οι ανεξάρτητες μεταβλητές της εξίσωσης, σε αντίθεση με την "Lasso" παλινδρόμηση, η οποία επιλέγει τις μεταβλητές που έχουν τη σημαντικότερη επίδραση στην εξαρτημένη μεταβλητή (L1 τιμή πέναλτι). Ωστόσο, υπάρχει και η δυνατότητα επιλογής της "Elastic" τιμής, όπου και οι δύο τιμές πέναλτι L1 και L2 χρησιμοποιούνται στο μοντέλο. • **Παράμετρος C:** Προεπιλεγμένη τιμή = 1.0 Η παράμετρος αυτή όσο μικρότερη είναι τόσο πιο ισχυρή κανονικοποίηση (regularization) υποδηλώνει. Με την τιμή 1 προσπαθούμε να αποφύγουμε μία επιθετική κανονικοποίηση των παραμέτρων. • **Παράμετρος solver:** Προεπιλεγμένη τιμή = 'lbfgs' Η παράμετρος αυτή υποδηλώνει τον αλγόριθμο που χρησιμοποιείται για το πρόβλημα βελτιστοποίησης. Για μικρά σύνολα δεδομένων χρησιμοποιείται ο αλγόριθμος 'liblinear', ενώ για μεγάλο όγκο δεδομένων προτιμώνται οι 'sag' and 'saga'. Ο αλγόριθμος 'lbfgs' χρησιμοποιείται ευρέως σε προβλήματα αριθμητικής βελτιστοποίησης και, κατά συνέπεια, στη λογιστική παλινδρόμηση προβλημάτων μηχανικής μάθησης.	Χρησιμοποιήθηκαν οι προ-καθορισμένες (default) τιμές παραμέτρων, καθώς η απόδοση του βασικού μοντέλου στην πλειονότητα των συνόλων δεδομένων ήταν πολύ υψηλή.

Μηχανές Διανυσμάτων Υποστήριξης (SVM)	*sklearn.svm.LinearSVC*	Το Linear SVC μοντέλο που χρησιμοποιήθηκε είναι ουσιαστικά μοντέλο SVM με γραμμικό kernel, δηλαδή το υπερεπίπεδο διαχωρισμού για την ταξινόμηση των σημείων είναι γραμμή. Η επιλογή του Linear SVM βασίστηκε στη βιβλιογραφία, όπου το μοντέλο αυτό έχει διαπιστωθεί ότι δίνει αρκετά καλά αποτελέσματα σε προβλήματα ανίχνευσης ψευδών ειδήσεων [6]. • **Παράμετρος C:** Προεπιλεγμένη τιμή = 1.0 Πρόκειται για την παράμετρο κανονικοποίησης (regularization), η οποία δηλώνει σε τι βαθμό επιθυμούμε να αποφεύγεται η λανθασμένη ταξινόμηση κάθε δείγματος του συνόλου εκπαίδευσης. Η τιμή της παραμέτρου είναι αντιστρόφως ανάλογη του μεγέθους του υπερ-επιπέδου που επιλέγεται, οπότε μεγάλες τιμές του C θα οδηγήσουν στην επιλογή ενός υπερ-επιπέδου με μικρό περιθώριο (margin), ενώ το αντίθετο ισχύει για μικρές τιμές του C.	Χρησιμοποιήθηκαν οι προ-καθορισμένες (default) τιμές παραμέτρων, καθώς η απόδοση του βασικού μοντέλου στην πλειονότητα των συνόλων δεδομένων ήταν πολύ υψηλή.
Naïve Bayes	*sklearn.naive_bayes.MultinomialNB*	Για την εφαρμογή του μοντέλου Naïve Bayes χρησιμοποιήθηκε η Multinomial εκδοχή, η οποία είναι κατάλληλη για δεδομένα που ακολουθούν πολυωνυμική κατανομή και χρησιμοποιείται ευρέως σε προβλήματα ταξινόμησης κειμένου, όπου τα δεδομένα αναπαρίστανται σαν διανύσματα λέξεων (στη δική μας περίπτωση έχουν υπολογιστεί τα TF-IDF διανύσματα τα οποία δίνονται σαν είσοδος στο μοντέλο). • **Παράμετρος alpha:** Προεπιλεγμένη τιμή = 1.0 Πρόκειται για την “Laplace Smoothing” τεχνική του αλγορίθμου η οποία χρησιμοποιείται για την αντιμετώπιση του προβλήματος της μηδενικής πιθανότητας.	Χρησιμοποιήθηκαν οι προ-καθορισμένες (default) τιμές παραμέτρων, καθώς η απόδοση του βασικού μοντέλου στην πλειονότητα των συνόλων δεδομένων ήταν πολύ υψηλή.

Αλγόριθμος κ-κοντινότερων γειτόνων (k-NN)	*sklearn.neighbors.KNeighborsClassifier*	Για τον αλγόριθμο k-NN χρησιμοποιήθηκαν οι παρακάτω τιμές παραμέτρων: • **Παράμετρος n_neighbors:** Έγινε επιλογή μέσω grid search Δε χρησιμοποιήθηκε η προ-επιλεγμένη τιμή, καθώς έγινε αναζήτηση των βέλτιστων τιμών για κάθε σύνολο δεδομένων. Πρόκειται για τον αριθμό των κοντινότερων γειτόνων μέσω των οποίων θα γίνει η τοποθέτηση κάθε νέου σημείου μέσα στο χώρο. • **Παράμετρος n_neighbors:** Προεπιλεγμένη τιμή = uniform Η τιμή αυτή δηλώνει ότι όλα τα σημεία στη γειτονιά θα χρησιμοποιηθούν με την ίδια βαρύτητα. • **Παράμετρος algorithm:** Προεπιλεγμένη τιμή = auto Αυτό σημαίνει ότι το μοντέλο θα βρει τον πιο κατάλληλο αλγόριθμο που μαθαίνει καλύτερα τα δεδομένα εισόδου. Οι αλγόριθμοι μπορεί να είναι οι: ‘ball_tree’, ‘kd_tree’, ‘brute’.	Χρησιμοποιήθηκε η μέθοδος grid search για την αναζήτηση της βέλτιστης τιμής k για κάθε σύνολο δεδομένων χωριστά. Το υποσύνολο των τιμών που δόθηκε στη μέθοδο grid search ήταν τιμές από το 1 μέχρι το 20, οπότε η βέλτιστη τιμή για κάθε σύνολο δεδομένων ήταν: **Σύνολο Δεδομένων 1** k = 18 **Σύνολο Δεδομένων 2** k = 17 **Σύνολο Δεδομένων 3** k = 12 **Σύνολο Δεδομένων 4** k = 8
Τυχαία Δάση (Random Forest)	*sklearn.ensemble.RandomForestClassifier*	Η μέθοδος αυτή όπως έχει αναφερθεί είναι συνδυαστική μέθοδος (ensemble), όπου συνδυάζει πολλά δέντρα απόφασης για να εξάγει την τελική πρόβλεψη. Οι δύο βασικές παράμετροι της μεθόδου δίνονται παρακάτω, όπου και για τις δύο έγινε αναζήτηση των βέλτιστων τιμών. • **Παράμετρος max_features:** Έγινε επιλογή μέσω grid search Η παράμετρος αυτή δηλώνει τον αριθμό των χαρακτηριστικών που πρέπει να ληφθούν υπόψη κατά το διαχωρισμό των δέντρων. • **Παράμετρος n_estimators:** Έγινε επιλογή μέσω grid search Η παράμετρος αυτή δηλώνει τον αριθμό των μεμονωμένων δέντρων απόφασης που θα συνθέσουν το τελικό μοντέλο.	Χρησιμοποιήθηκε η μέθοδος grid search για την αναζήτηση των βέλτιστων τιμών των παραμέτρων max_features και n_estimators σε κάθε σύνολο δεδομένων. Τα υποσύνολα που χρησιμοποιήθηκαν ήταν: n_estimators: [100, 150, 200, 250, 300] max_features: ['auto', 'sqrt', 'log2'] Οι αντίστοιχες καλύτερες τιμές για κάθε σύνολο δεδομένων ήταν: **Σύνολο Δεδομένων 1** max_features = 'log2' (δηλαδή παίρνουμε το λογάριθμο του συνόλου των χαρακτηριστικών) n_estimators = 250

			Σύνολο Δεδομένων 2 max_features = 'auto' (τετραγωνική ρίζα του συνόλου των χαρακτηριστικών – ίδιο με το 'sqrt') n_estimators = 100 **Σύνολο Δεδομένων 3** max_features = 'auto' (τετραγωνική ρίζα του συνόλου των χαρακτηριστικών – ίδιο με το 'sqrt') n_estimators = 150 **Σύνολο Δεδομένων 4** max_features = 'log2' (δηλαδή παίρνουμε το λογάριθμο του συνόλου των χαρακτηριστικών) n_estimators = 100
Adaboost	*sklearn.ensemble.AdaBoostClassifier*	Ο αλγόριθμος αυτός, όπως και ο Random Forest, είναι Ensemble και χρησιμοποιεί ως προεπιλογή δέντρα απόφασης. Η διαφορά με τον RF, όπως αναλύθηκε σε προηγούμενο κεφάλαιο, είναι ότι η εκπαίδευση γίνεται με τη μέθοδο Boosting, ενώ στο RF μοντέλο χρησιμοποιείται η μέθοδος Bagging. Οι βασικές παράμετροι που χρησιμοποιήθηκαν ήταν οι παρακάτω: • **Παράμετρος base_estimator:** Προεπιλεγμένη τιμή = None Αυτό σημαίνει ότι τα μοντέλα που χρησιμοποιήθηκαν στη συνδυαστική μέθοδο είναι τα δέντρα απόφασης. Έχει διατηρηθεί η προεπιλεγμένη τιμή, έτσι ώστε να συγκριθεί η μέθοδος Adaboost με την Random Forest, η οποία χρησιμοποιεί, επίσης, δέντρα απόφασης. • **Παράμετρος n_estimators:** Προεπιλεγμένη τιμή = 50 Πρόκειται για το μέγιστο αριθμό των ταξινομητών στον οποίο ολοκληρώνεται η 'boosting' διαδικασία. Στην περίπτωση που γίνεται καλή εφαρμογή στα δεδομένα, η διαδικασία μάθησης τερματίζει γρηγορότερα.	Χρησιμοποιήθηκαν οι προ-καθορισμένες (default) τιμές παραμέτρων.

3.4 Εφαρμογή των μοντέλων στα σύνολα δεδομένων

Στην ενότητα αυτή θα περιγραφεί αναλυτικά η διαδικασία εκτέλεσης των πειραμάτων στη γλώσσα προγραμματισμού Python. Όπως έχει προαναφερθεί, το πρόβλημα που καλούμαστε να επιλύσουμε είναι η κατηγοριοποίηση μιας είδησης ως ‘Αληθούς’ ή ‘Ψευδούς’ (πρόβλημα δυαδικής ταξινόμησης).

Βήμα 1

Τα διαφορετικά σύνολα δεδομένων τα οποία βρίσκονται σε μορφή αρχείων “csv” διαβάζονται μέσω της “pandas” βιβλιοθήκης της Python και αποθηκεύονται σε “dataframes”, τα οποία μπορούμε να φανταστούμε σαν πίνακες μιας σχεσιακής βάσης.

	id	news_url	title	tweet_ids	label
0	politifact224	NaN	Biden Issues Statement Following Conversation ...	399365492	TRUE
1	politifact8846	http://www.greentechmedia.com/research/ussmi	U.S. Solar Market Insight Report	27152919999\t27254128646\t27623758097\t4540637...	TRUE
2	politifact13827	https://web.archive.org/web/20170307175331/htt...	Whoopi Goldberg: Navy SEAL Widow was “Looking ...	837096266798665728\t837099832263192578\t837103...	FAKE
3	politifact462	https://web.archive.org/web/20080212000046/htt...	Transcripts	377729302\t945210227\t1217912078\t1270179527\t...	TRUE
4	politifact15291	https://web.archive.org/web/20180424001608/htt...	Archbishop Desmond Tutu dies while holidaying ...	NaN	FAKE

Εικόνα 11 Παράδειγμα pandas data frame στη γλώσσα προγραμματισμού Python

Στη συνέχεια γίνεται μία αρχική προετοιμασία των δεδομένων (πχ εύρεση τιμών ‘NULL’ και αντικατάστασή τους με κενό κείμενο, επιλέγεται η στήλη που περιέχει το κείμενο της είδησης και ενσωματώνεται η κλάση εξόδου σε περιπτώσεις συνόλων δεδομένων όπου οι τιμές εξόδου δίνονται σε ξεχωριστό αρχείο).

Βήμα 2

Στο βήμα αυτό εκτελούνται όλες οι τεχνικές προ-επεξεργασίας οι οποίες παρουσιάστηκαν λεπτομερώς στην προηγούμενη ενότητα. Περιληπτικά, εφαρμόζονται οι παρακάτω τεχνικές προ-επεξεργασίας, με τη σειρά που δίνονται:

- LowerCasing
- Tokenization
- Stemming
- Stop word removal
- Part-of-speech Tagging

Βήμα 3

Αφού γίνει η προ-επεξεργασία των δεδομένων, ακολουθεί ο μετασχηματισμός τους μέσω της TF-IDF στατιστικής, όπου εξάγονται τα διανύσματα χαρακτηριστικών που θα τροφοδοτηθούν στα μοντέλα πρόβλεψης.

Βήμα 4

Στη συνέχεια, τα μετασχηματισμένα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης και δεδομένα ελέγχου, με ποσοστό 75% για την εκπαίδευση και 25% για την αξιολόγηση των μοντέλων.

Δεδομένα εκπαίδευσης	0.75 * Σύνολο δεδομένων
Δεδομένα ελέγχου	0.25 * Σύνολο δεδομένων

Βήμα 5

Τα δεδομένα εκπαίδευσης δίνονται σαν είσοδος στο εκάστοτε μοντέλο μηχανικής μάθησης, το οποίο καλείται με τις κατάλληλες τιμές παραμέτρων (είτε τις προ-επιλεγμένες είτε αυτές που προέκυψαν από τη διαδικασία βελτιστοποίησης), και εκπαιδεύεται ώστε να μάθει τα δεδομένα (fit).

Βήμα 6

Στο σημείο αυτό, τα εκπαιδευμένα μοντέλα παίρνουν σαν είσοδο τα δεδομένα ελέγχου (χωρίς την κλάση εξόδου) και παράγουν την πρόβλεψή τους. Η πρόβλεψη αυτή συγκρίνεται με τις πραγματικές τιμές εξόδου και υπολογίζονται οι τέσσερις μετρικές αξιολόγησης, Accuracy, Precision, Recall και F-Measure μέσω του πίνακα σύγχυσης. Οι τιμές των μετρικών αυτών χρησιμοποιούνται για την αξιολόγηση των αποτελεσμάτων στο επόμενο κεφάλαιο.

ΚΕΦΑΛΑΙΟ 4. Εφαρμογή και Αποτελέσματα

Στο κεφάλαιο αυτό θα παρουσιαστούν τα αποτελέσματα της εφαρμογής των διαφορετικών αλγορίθμων Μηχανικής Μάθησης στα τέσσερα σύνολα δεδομένων. Η αξιολόγηση έχει γίνει με βάση τις μετρικές Accuracy, Precision, Recall και F-Measure. Αφού γίνει η ανάλυση των αποτελεσμάτων, θα αξιολογηθούν οι διαφορετικοί αλγόριθμοι και θα γίνει η ανάδειξη της αποδοτικότερης μεθόδου.

4.1 Αποτελέσματα εφαρμογής των μοντέλων Μηχανικής Μάθησης στα σύνολα των δεδομένων

Παρακάτω δίνονται οι πίνακες με τις τιμές των διαφορετικών μετρικών για κάθε αλγόριθμο και για κάθε σύνολο δεδομένων. Τα σύνολα δεδομένων συμβολίζονται με ΣΔ και τον αντίστοιχο αριθμό 1-4 (περισσότερες λεπτομέρειες μπορούν να βρεθούν στην αντίστοιχη ενότητα). Επίσης, για κάθε διαφορετικό σύνολο δεδομένων (ανά στήλη των πινάκων) έχει σημανθεί με έντονη γραφή η μεγαλύτερη τιμή για να γίνει εύκολη διάκριση των μοντέλων που έχουν την καλύτερη απόδοση.

Στον παρακάτω πίνακα δίνεται η μετρική Accuracy, όπου είναι εμφανές ότι οι μέθοδοι Linear SVM και Random Forest δίνουν την καλύτερη ακρίβεια προβλέψεων.

Αλγόριθμος	**DS1**	**DS2**	**DS3**	**DS4**
Logistic Regression	0.96	0.96	0.96	0.74
Linear SVM	**0.97**	0.98	**0.98**	0.82
Naïve Bayes	0.95	0.95	0.94	0.78
k-Nearest Neighbors (k-NN)	0.92	0.94	0.93	0.75
Random Forest	**0.97**	**0.99**	0.96	**0.83**
Adaboost	0.92	0.98	0.97	0.76

Πίνακας 2 Μετρική Accuracy των διαφορετικών αλγορίθμων στα 4 σύνολα δεδομένων

Στον παρακάτω πίνακα δίνεται η μετρική Precision, όπου και πάλι οι μέθοδοι Linear SVM και Random Forest υπερτερούν έναντι των υπολοίπων.

Αλγόριθμος	**ΣΔ1**	**ΣΔ2**	**ΣΔ3**	**ΣΔ4**
Logistic Regression	0.96	0.96	0.96	0.76
Linear SVM	**0.97**	0.98	**0.98**	0.82

Naïve Bayes	0.95	0.96	0.94	0.80
k-Nearest Neighbors (k-NN)	0.92	0.95	0.93	0.75
Random Forest	**0.97**	**0.99**	0.96	**0.84**
Adaboost	0.92	0.98	0.97	0.77

Πίνακας 3 Μετρική Precision των διαφορετικών αλγορίθμων στα 4 σύνολα δεδομένων

Η μετρική Precision όπως φαίνεται παρακάτω αναδεικνύει τις μεθόδους Linear SVM και Random Forest.

Αλγόριθμος	**ΣΔ1**	**ΣΔ2**	**ΣΔ3**	**ΣΔ4**
Logistic Regression	0.96	0.96	0.96	0.74
Linear SVM	**0.97**	0.98	**0.98**	0.82
Naïve Bayes	0.95	0.95	0.94	0.78
k-Nearest Neighbors (k-NN)	0.92	0.94	0.93	0.75
Random Forest	**0.97**	**0.99**	0.96	**0.83**
Adaboost	0.92	0.98	0.97	0.76

Πίνακας 4 Μετρική Recall των διαφορετικών αλγορίθμων στα 4 σύνολα δεδομένων

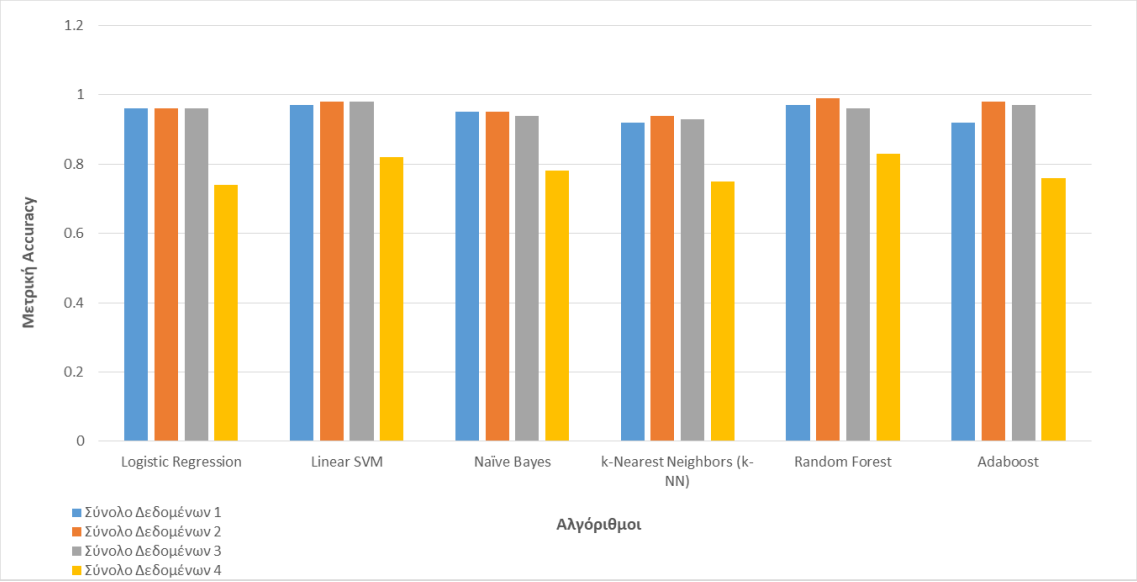
Την ίδια εικόνα λαμβάνουμε και από τον παρακάτω πίνακα, όπου παρουσιάζεται η μετρική F-Measure και οι αλγόριθμοι Linear SVM και Random Forest ως οι επικρατέστεροι.

Αλγόριθμος	**ΣΔ1**	**ΣΔ2**	**ΣΔ3**	**ΣΔ4**
Logistic Regression	0.96	0.96	0.96	0.72
Linear SVM	**0.97**	0.98	**0.98**	0.81
Naïve Bayes	0.95	0.95	0.94	0.77
k-Nearest Neighbors (k-NN)	0.92	0.95	0.93	0.75
Random Forest	**0.97**	**0.99**	0.96	**0.82**
Adaboost	0.92	0.98	0.97	0.74

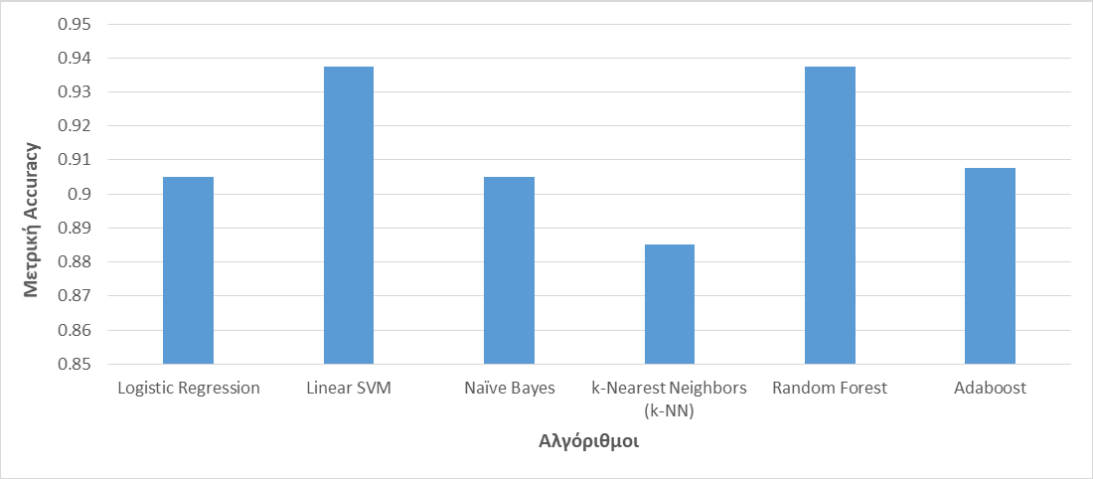
Πίνακας 5 Μετρική F-Measure των διαφορετικών αλγορίθμων στα 4 σύνολα δεδομένων

Προκειμένου να γίνει καλύτερα αντιληπτή η απόδοση όλων των αλγορίθμων στα διαφορετικά σύνολα δεδομένων, δημιουργήθηκαν διάφορα γραφήματα με τις τιμές των μετρικών που παρουσιάστηκαν στους παραπάνω πίνακες.

Στα παρακάτω δύο γραφήματα δίνονται οι τιμές της μετρικής Accuracy για κάθε αλγόριθμο μηχανικής μάθησης, με τη μόνη διαφορά ότι στο πρώτο διάγραμμα παρατίθενται οι τιμές για κάθε σύνολο δεδομένων ξεχωριστά, ενώ στο δεύτερο έχει ληφθεί η μέση τιμή των τιμών Accuracy όλων των συνόλων δεδομένων ανά αλγόριθμο.

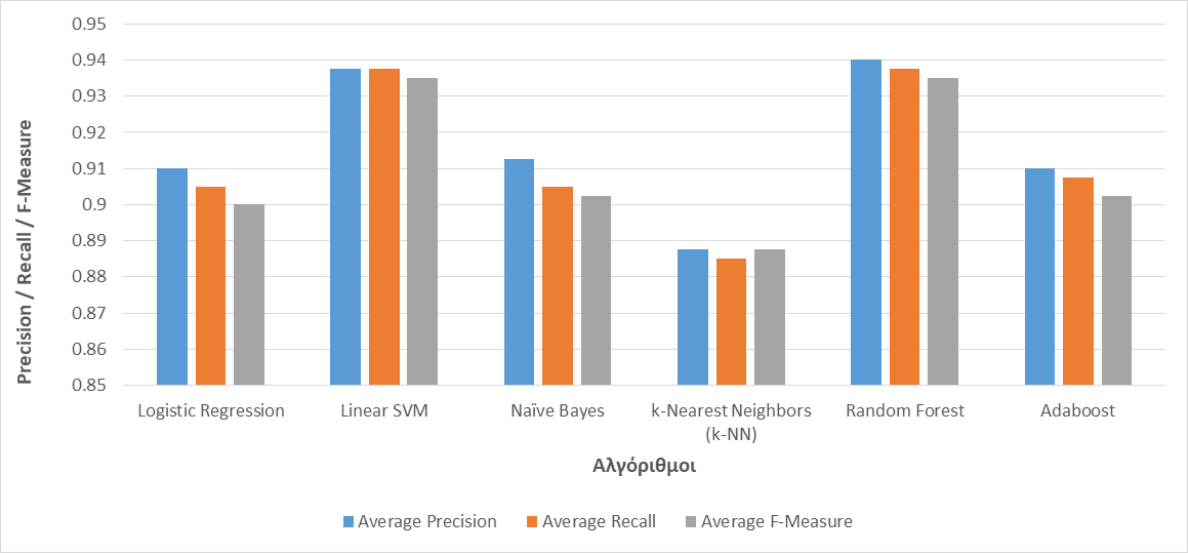


Εικόνα 12 Μετρική Accuracy ανά σύνολο δεδομένων και ανά αλγόριθμο



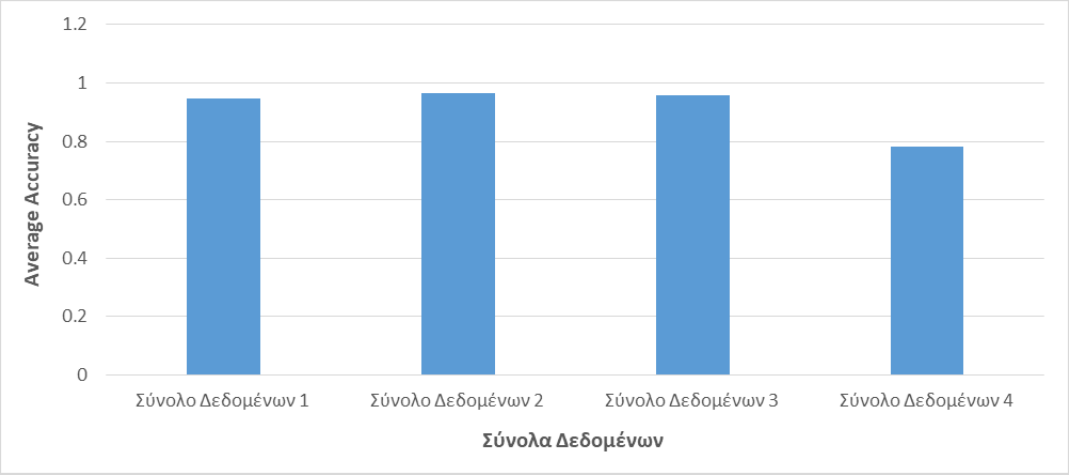
Εικόνα 13 Μέση τιμή της μετρικής Accuracy όλων των συνόλων δεδομένων για κάθε αλγόριθμο χωριστά

Στη συνέχεια δίνονται οι μέσες τιμές των μετρικών Precision, Recall και F-Measure σε όλα τα σύνολα δεδομένων για κάθε αλγόριθμο ξεχωριστά.

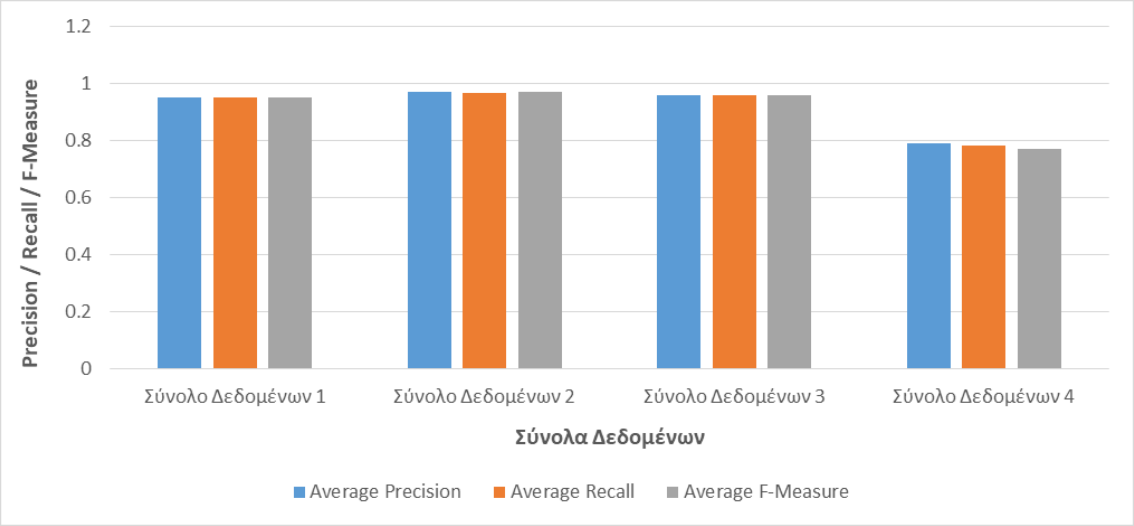


Εικόνα 14 Μέση τιμή των μετρικών Precision, Recall και F-Measure όλων των συνόλων δεδομένων για κάθε αλγόριθμο χωριστά

Τέλος, δίνονται οι μέσες τιμές όλων των μετρικών (Accuracy στο διάγραμμα που ακολουθεί και Precision, Recall και F-Measure στο δεύτερο που ακολουθεί) ως προς τα σύνολα δεδομένων, ώστε να μελετηθεί και η απόδοση των μοντέλων στα δεδομένα που έχουν επιλεγεί. Άλλωστε η έξοδος του κάθε αλγορίθμου μηχανικής μάθησης είναι αλληλένδετα συνδεδεμένη με τη φύση των δεδομένων που δίνουμε σαν είσοδο, οπότε στην επόμενη ενότητα θα γίνει η αξιολόγηση και ως προς αυτήν την κατεύθυνση.



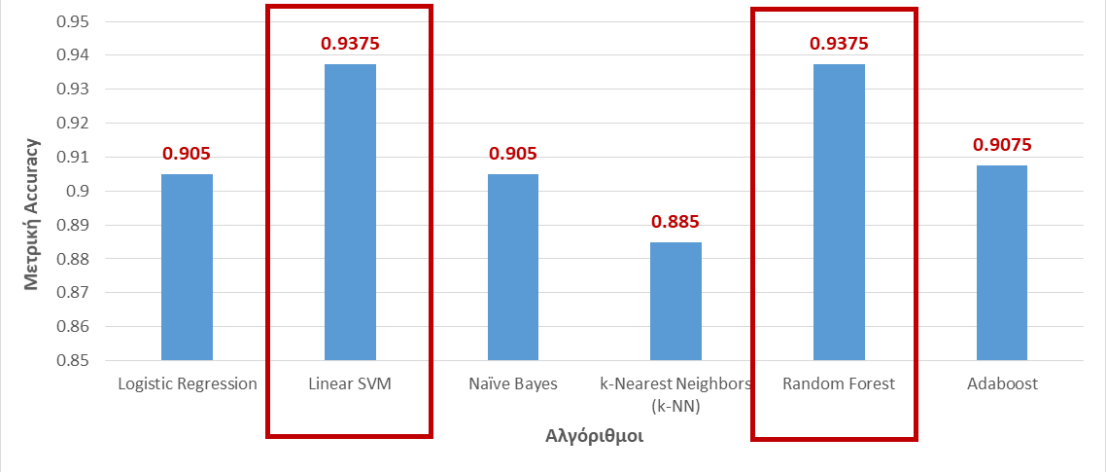
Εικόνα 15 Μέση τιμή της μετρικής Accuracy όλων των αλγορίθμων για κάθε σύνολο δεδομένων χωριστά



Εικόνα 16 Μέση τιμή των μετρικών Precision, Recall και F-Measure όλων των αλγορίθμων για κάθε σύνολο δεδομένων χωριστά

4.2 Σύγκριση και αξιολόγηση των αποτελεσμάτων. Επιλογή της αποδοτικότερης μεθόδου

Όπως διαπιστώθηκε από την προηγούμενη ενότητα, οι δύο αλγόριθμοι που είχαν πολύ καλή απόδοση στα διαφορετικά σύνολα δεδομένων είναι οι Linear SVM και Random Forest. Μάλιστα, αν δούμε τις μέσες τιμές του Accuracy στο παρακάτω διάγραμμα, παρατηρούμε ότι έχουν ακριβώς την ίδια τιμή, οπότε δεν μπορούμε να επιλέξουμε με ασφάλεια μία από τις δύο μεθόδους.



Εικόνα 17 Επιλογή του καλύτερου μοντέλου βάσει της μέσης τιμής του Accuracy

Για το Random Forest ήταν αναμενόμενο το αποτέλεσμα, καθώς αποτελεί μέθοδο Ensemble, που σε γενικές γραμμές οι μέθοδοι αυτές ξεπερνούν το πρόβλημα της υπερ-εκπαίδευσης, οπότε οδηγούν σε πιο εύρωστα μοντέλα. Επίσης, σε παρόμοια

προβλήματα στη βιβλιογραφία, η μέθοδος αυτή δίνει πολύ υψηλές τιμές της μετρικής Accuracy [6]. Ωστόσο, το μειονέκτημα της μεθόδου Random Forest είναι ότι είναι υπολογιστικά ακριβή, καθώς έγινε πρώτα η εύρεση των βέλτιστων υπερ-παραμέτρων και έπειτα η εκπαίδευση των μοντέλων με τις καλύτερες τιμές αυτών των παραμέτρων. Αυτό είχε σαν αποτέλεσμα ο χρόνος εκτέλεσης να κυμαίνεται από 15 μέχρι 30 λεπτά για κάθε σύνολο δεδομένων.

Από την άλλη μεριά, ο αλγόριθμος SVM με γραμμικό υπερ-επίπεδο (Linear SVM) προτείνεται σαν υποσχόμενη μέθοδος σε προβλήματα ταξινόμησης κειμένου (text classification) για τους παρακάτω λόγους [53]:

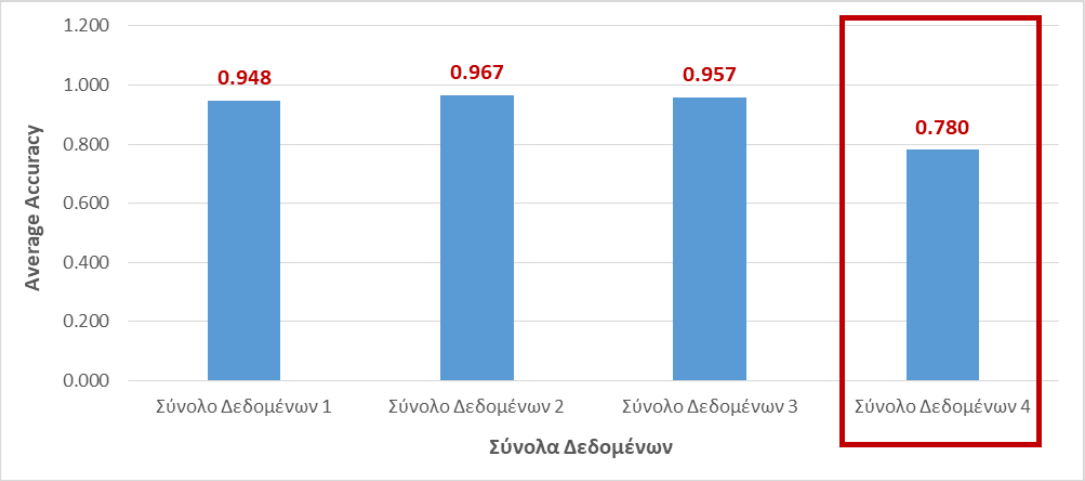
- Τα δεδομένα κειμένου είναι συχνά εύκολα διαχωρίσιμα, οπότε ο γραμμικός διαχωρισμός επιλύει κατευθείαν το πρόβλημα.
- Τα δεδομένα κειμένου περιέχουν πολλά χαρακτηριστικά και το γραμμικό Kernel φαίνεται ότι λειτουργεί καλά σε αυτήν την περίπτωση. Αυτό συμβαίνει επειδή όταν αντιστοιχίζουμε τα δεδομένα σε χώρο περισσότερων διαστάσεων (πχ με RBF kernel) δε βελτιώνουμε απαραίτητα την απόδοση του μοντέλου [23].

Συμπερασματικά, μπορούμε να αποφανθούμε ότι οι μέθοδοι Random Forest και Linear SVM έχουν παρόμοια απόδοση στα διαφορετικά σύνολα δεδομένων, ωστόσο η μέθοδος Linear SVM υπερτερεί διότι είναι υπολογιστικά γρήγορη. Επίσης, σε περίπτωση που χρειάζεται να γίνει βελτιστοποίηση των υπερ-παραμέτρων με τη μέθοδο grid search, η μόνη παράμετρος στην περίπτωση του Linear SVM είναι η παράμετρος C (regularization) (για άλλα kernel υπάρχει και η παράμετρος gamma), ενώ στην περίπτωση του Random Forest έχουμε δύο παραμέτρους (n_estimators και max_features).

Πέρα από τις δύο καλύτερες μεθόδους, παρατηρούμε ότι οι τρεις επόμενες, Adaboost, Naïve Bayes και Logistic Regression έχουν σχεδόν την ίδια μέση τιμή της μετρικής Accuracy ~ 91%, γεγονός που τις καθιστά αρκετά αποδοτικές για το υπό μελέτη πρόβλημα. Η μέθοδος k-NN έρχεται τελευταία, για την οποία, αν και εκτελέστηκε grid search για την εύρεση του βέλτιστου k, δεν κατάφερε να φτάσει στα ίδια επίπεδα με τις υπόλοιπες.

Ένα ακόμη συμπέρασμα που έχει εξαχθεί από την αξιολόγηση των αποτελεσμάτων είναι ότι το τελευταίο σύνολο δεδομένων έχει φτωχότερη απόδοση από τα υπόλοιπα

τρία, τα οποία παρουσιάζουν ακρίβεια πάνω από 90% με όλους τους αλγορίθμους πρόβλεψης, όπως φαίνεται και στην παρακάτω εικόνα:



Εικόνα 18 Σύγκριση της συνολικής απόδοσης των μοντέλων στα διαφορετικά σύνολα δεδομένων

Η εξήγηση της χαμηλότερης απόδοσης βασίζεται στο γεγονός ότι το σύνολο αυτό έχει μικρό αριθμό δειγμάτων αναλογικά με τον αριθμό των χαρακτηριστικών του, οπότε ο αλγόριθμος δεν έχει καλή ικανότητα γενίκευσης.

ΚΕΦΑΛΑΙΟ 5. Συμπεράσματα και Προτάσεις

Στο κεφάλαιο αυτό θα δοθούν τα συμπεράσματα, όποιοι περιορισμοί παρουσιάστηκαν κατά την πειραματική διαδικασία και προτάσεις για επέκταση του θέματος και περαιτέρω έρευνα.

5.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της ανίχνευσης ψευδών ειδήσεων στο διαδίκτυο με τη χρήση μεθόδων Μηχανικής Μάθησης. Πιο συγκεκριμένα, πρόκειται για πρόβλημα δυαδικής ταξινόμησης (binary classification), όπου τα μοντέλα πρέπει να προβλέψουν αν μια είδηση είναι αληθής ή ψευδής. Τα δεδομένα που συλλέχθηκαν έχουν χρησιμοποιηθεί σε προηγούμενες μελέτες ανίχνευσης ψευδών ειδήσεων με μεθόδους Μηχανικής Μάθησης και περιέχουν τόσο ψευδείς όσο και αληθείς ειδήσεις που προέρχονται από άρθρα ποικίλων περιοχών ενδιαφέροντος, όπως οικονομία, πολιτική, κτλ. Επίσης, έγινε η επιλογή συνόλων με διαφορετική τάξη μεγέθους, ώστε να αξιολογηθεί η απόδοση των μοντέλων και ως προς αυτήν την κατεύθυνση.

Οι αλγόριθμοι που επιλέχθηκαν είναι αρκετά δημοφιλείς στη βιβλιογραφία, όσον αφορά προβλήματα ανίχνευσης ψευδών ειδήσεων, και προτείνεται η εφαρμογή τους λόγω της υψηλής απόδοσης που παρουσιάζουν. Η ακρίβεια των μοντέλων ήταν αρκετά υψηλή στα τρία από τα τέσσερα σύνολα δεδομένων, καθώς το τέταρτο ήταν το μικρότερο σε μέγεθος και συνδυαστικά με το διάνυσμα χαρακτηριστικών, οι εγγραφές δεν ήταν αρκετές, ώστε τα μοντέλα να αποκτήσουν καλή ικανότητα γενίκευσης. Οι αλγόριθμοι Linear SVM και Random Forest είχαν πολύ καλή και σταθερή απόδοση βάσει όλων των μετρικών αξιολόγησης, γεγονός που επιβεβαιώνεται και στη βιβλιογραφία. Τα μοντέλα Linear SVM προτείνονται για προβλήματα ταξινόμησης κειμένου, καθώς ο γραμμικός διαχωρισμός λειτουργεί αρκετά καλά σε δεδομένα κειμένου. Από την άλλη μεριά, ο Random Forest είναι ένας ισχυρός συνδυαστικός αλγόριθμος (Ensemble) ο οποίος ξεπερνά το πρόβλημα της υπερ-εκπαίδευσης, χωρίς να επηρεάζεται από ακραίες τιμές (outliers), οι οποίες είναι από τα σοβαρότερα προβλήματα στην απόδοση των ταξινομητών. Επιπλέον, η αναζήτηση των βέλτιστων υπερ-παραμέτρων του αλγορίθμου αύξησε την ικανότητα πρόβλεψης του μοντέλου.

Πέρα από τις δύο καλύτερες μεθόδους, παρατηρούμε ότι οι τρεις επόμενες, Adaboost, Naïve Bayes και Logistic Regression έχουν αρκετά υψηλή τιμή Accuracy ~ 91%, γεγονός που τις καθιστά δυνατούς υποψηφίους για το υπό μελέτη πρόβλημα. Η μέθοδος k-NN έρχεται τελευταία, για την οποία, αν και εκτελέστηκε grid search για την εύρεση του βέλτιστου k, δεν κατάφερε να φτάσει στα ίδια επίπεδα με τις υπόλοιπες.

Ένας ακόμη λόγος που οι διαφορετικοί αλγόριθμοι είχαν πολύ υψηλή απόδοση στην πλειονότητα των συνόλων δεδομένων είναι η προ-επεξεργασία και ο μετασχηματισμός στην οποία υποβλήθηκαν τα δεδομένα, ώστε να καταλήξουμε μόνο με σημαντική πληροφορία, απαλλαγμένη από θόρυβο, ώστε να βοηθήσουμε τα μοντέλα να μάθουν τα δεδομένα χωρίς να υπερ-εκπαιδεύονται, επομένως να έχουν καλή ικανότητα γενίκευσης όταν εφαρμόζονται σε άγνωστα σύνολα δεδομένων.

5.2 Προβληματισμοί και περιορισμοί

Το πρόβλημα που μελετήθηκε στην παρούσα διπλωματική εργασία είναι τόσο ευρύ, που ο κάθε ερευνητής έχει πληθώρα επιλογών τόσο σε μεθόδους μηχανικής μάθησης όσο και μεθόδους επεξεργασίας του κειμένου και εξαγωγής σημαντικών χαρακτηριστικών που θα ενισχύσουν την απόδοση των μοντέλων.

Ένας περιορισμός που προέκυψε κατά την εκτέλεση των πειραμάτων ήταν ότι κατά την προσπάθεια εφαρμογής συνελικτικού νευρωνικού δικτύου στα σύνολα δεδομένων, τα αποτελέσματα που προέκυψαν δεν ήταν τα αναμενόμενα, καθώς απαιτούνταν πολύ μεγαλύτερος αριθμός εγγραφών για την εκπαίδευση του μοντέλου. Άλλωστε, τα πειράματα με νευρωνικά δίκτυα απαιτούν και υπολογιστική ισχύ αλλά και αρκετό χρόνο, συγκριτικά με άλλες μεθόδους που είναι λιγότερο ακριβές υπολογιστικά.

Επίσης, οι τεχνικές εξαγωγής χαρακτηριστικών που παρουσιάζονται στη βιβλιογραφία είναι πολυάριθμες, ωστόσο έγινε η επιλογή της τεχνικής TF-IDF γιατί προτείνεται σε μεγάλο μέρος της βιβλιογραφίας, καθώς η δοκιμή άλλων τεχνικών σε συνδυασμό με τους έξι συνολικά αλγορίθμους και τα τέσσερα διαφορετικά σύνολα δεδομένων θα αύξανε αρκετά την πολυπλοκότητα του πειραματικού μέρους.

5.3 Προτάσεις για περαιτέρω έρευνα

Το πρόβλημα της ανίχνευσης ψευδών ειδήσεων έχει πολλές προοπτικές για εξέλιξη των μεθόδων και την εφαρμογή πιο προηγμένων τεχνικών, καθώς τα διαθέσιμα

δεδομένα είναι ανεξάντλητα και η πληροφορία που μπορεί να εξαχθεί από αυτά μπορεί να οδηγήσει σε πιο σύνθετες προβλέψεις, όπως για παράδειγμα εκτός από δυαδική ταξινόμηση σε «Αληθή» ή «Ψευδή» είδηση να γίνει η αναζήτηση περισσότερων κλάσεων στην έξοδο (multiclass classification). Επίσης, θα είχε ενδιαφέρον να ερευνάται και η αξιοπιστία του αρθρογράφου, μέσα από την αναζήτηση και αξιολόγηση του έργου του και την παρουσία του μέσα στα μέσα κοινωνικής δικτύωσης.

Μία ακόμη αρκετά ενδιαφέρουσα κατεύθυνση στο υπό μελέτη πρόβλημα είναι η εξαγωγή χαρακτηριστικών που δε θα εστιάζουν μόνο σε γλωσσολογικές τεχνικές αλλά και στην ανάλυση συναισθήματος. Για παράδειγμα, το συναίσθημα που αναδύεται από μία είδηση (αν είναι για παράδειγμα αρνητικά ή θετικά πολωμένη) θα μπορέσει να βοηθήσει στην επικύρωση της εγκυρότητάς της.

Τέλος, πέρα από τον έλεγχο του κειμένου και των γεγονότων στα ειδησεογραφικά άρθρα, θα μπορούσαν να χρησιμοποιούνται σαν είσοδος στα μοντέλα και οι εικόνες που συνοδεύουν την είδηση. Μέσα από τις εικόνες θα μπορούσε να εξαχθεί πολύ σημαντική πληροφορία που συνδυαστικά με το κείμενο θα βελτίωνε ακόμη περισσότερο την προβλεπτική ικανότητα των μοντέλων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Καρποδίνης, Κ. Ανάλυση συναισθημάτων σε δεδομένα από το Twitter χρησιμοποιώντας εργαλεία της R και μοντέλα μηχανικής μάθησης. Τμήμα Ηλεκτρολόγων Μηχανικών και Τεχνολογίας Υπολογιστών, Πανεπιστήμιο Πατρών, 2016

[2] Χρυσίνα, Δ. Ανάλυση κειμένου αγγελιών εργασίας με εφαρμογή του πακέτου tm της στατιστικής γλώσσας R. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Τμήμα Πληροφορικής, 2018

[3] Κύρκος, Ε. Επιχειρηματική Ευφυΐα και Εξόρυξη Δεδομένων. Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα, www.kallipos.gr, 2015

[4] Βερύκιος, Β., Καγκλής, Β., Σταυρόπουλος, Η. Η επιστήμη των δεδομένων μέσα από τη γλώσσα R. Ελληνικά Ακαδημαϊκά Ηλεκτρονικά Συγγράμματα και Βοηθήματα, www.kallipos.gr, 2015

[5] Καράμπελα, Α. Ανίχνευση ψευδών ειδήσεων με την αξιοποίηση τεχνικών Μηχανικής Μάθησης και επεξεργασίας δεδομένων μεγάλου όγκου. Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Σχολή Οικονομικών και Πολιτικών Επιστημών, 2021

[6] Ahmad, I., Yousaf, M., Yousaf, S., Ahmad, M.O. Fake News Detection Using Machine Learning Ensemble Methods. Complexity 2020, 2020, 8885861

[7] Mitchell, T.M. Machine learning and data mining. Communications of the ACM, 42(11):30–36, 1999

[8] Mitchell, T. M. Discipline of Machine Learning. Carnegie Mellon University, Pittsburgh, PA, USA, 2006

[9] Alpaydin, E. Introduction to Machine Learning. Massachusetts Institute of Technology, 2010

[10] Mokhtar, M.S., Jusoh, Y.Y., Admodisastro, N., Pa, N.C., Amruddin, A.Y. Fakebuster: Fake News Detection System Using Logistic Regression Technique In Machine Learning. Int. J. Eng. Adv. Technol. (IJEAT) 2019, 9, 2407–2410

[11] Ahmed, H., Traore, I., Saad, S. Detection of Online Fake News

Using N-Gram Analysis and Machine Learning Techniques. In Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, ser. Lecture Notes in Computer Science. Springer, Cham, Oct. 2017, pp. 127–138

[12] Hofmann, T., Scholkopf, B., Smola, A.J. Kernel methods in machine learning. 7e Annals of Statistics, vol. 36, no. 3, pp. 1171–1220, 2008

[13] Ahmed, H., Traore, I., Saad, S. Detecting opinion spams and fake news using text classification. Security and Privacy, vol. 1, no. 1, 2018

[14] Wang, W.Y. Liar, Liar Pants on Fire": A new benchmark dataset for fake news detection. arXiv preprint arXiv:1705.00648, 2017

[15] Murphy, K.P. Machine learning: a probabilistic perspective. In: Chapter 1: Introduction. MIT Press, pp 1–26, 2012

[16] Cusmuliuc, C.G., Coca, L.G., Iftene, A. Identifying Fake News on Twitter using Naive Bayes, SVM and Random Forest Distributed Algorithms. In Proceedings of The 13th Edition of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR-2018), 2018, pp. 177-188

[17] Kesarwani, A., Chauhan, S. S., Nair, A. R. Fake News Detection on Social Media using K-Nearest Neighbor Classifier. 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), Las Vegas, NV, USA, pp.1-4, 2020

[18] Gaonkar, S., Itagi, S., Chalippatt, R., Gaonkar, A., Aswale, S., Shetgaonkar, P. Detection of Online Fake News: A Survey. In Proceedings of the 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 30–31 March 2019; pp. 1–6

[19] Dutta P.S, Das M, Biswas S, Bora M, Saikia S.S. Fake news prediction: a survey. International Journal of Scientific Engineering and Science. 2019;3(3):1–3

[20] Zhou X, Zafarani R. A survey of fake news: fundamental theories, detection methods, and opportunities. ACM Computing Surveys. 2020;53(5):1–40. doi: 10.1145/3395046, 2020

[21] Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. FakeNewsNet: a data repository with news content, social context and dynamic information for

studying fake news on social media. https://arxiv.org/abs/1809.01286 arxiv. 2018

[22] Yang, C. Evaluating unsupervised and supervised image classification methods for mapping cotton root rot. Precision Agriculture, 2014

[23] Hsu, C., CHANG, W., LIN, C.J. A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University, 2003

[24] Lacy, S., Rosenstiel, T. Defining and measuring quality journalism. Rutgers School of Communication and Information, 2005

[25] Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. Digital journalism, 6(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

[26] Shu, K., et al. Fake news detection on social media: A data mining perspective. ACM SIGKDD Explorations Newsletter 19, 1 (2017), 22–36

[27] Singh, V., et al. Automated Fake News Detection Using Linguistic Analysis and Machine Learning. International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), 2017, http://doi.org/10.13140/RG.2.2.16825.67687

[28] Pratiwi, I. Y. R., Asmara, R. A., & Rahutomo, F. Study of hoax news detection using naïve bayes classifier in Indonesian language. 2017 11th International Conference on Information & Communication Technology and System (ICTS), Surabaya, pp.73-78, 2017, https://doi.org/10.1109/ICTS.2017.8265649

[29] Kaur, S., Kumar, P. & Kumaraguru, P. (2020). Automating fake news detection system using multi-level voting model. Soft Computing, 24 (12), 9049–9069. https://doi.org/10.1007/s00500-019-04436-y

[30] Ni, B., Guo, Z., Li, J., & Jiang, M. (2020). Improving Generalizability of Fake News Detection Methods using Propensity Score Matching. Social and Information Networks. https://arxiv.org/abs/2002.00838

[31] Kesarwani, A., Chauhan, S. S., & Nair, A. R. (2020). Fake News Detection on Social Media using K-Nearest Neighbor Classifier. 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE), Las Vegas, NV, USA, pp.1-4, https://doi.org/10.1109/ICACCE49060.2020.9154997

[32] Kotteti, C. M. M., Dong, X., Li, N., & Qian, L. (2018). Fake news detection enhancement with data imputation. 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, Athens, 2018, pp.187-192. https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00042

[33] Khan, J.Y., et al. A benchmark study on machine learning methods for fake news detection. Arxiv Preprint Arxiv:1905.04749, 2019

[34] Rubin, V., Conroy, N., Chen, Y., Cornwell, S. Fake news or truth? Using satirical cues to detect potentially misleading news. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection, pages 7–17, 2016.

[35] Rashkin, H., et al. Truth of varying shades: Analyzing language in fake news and political fact-checking. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2931–2937, 2017.

[36] Asaad, B., Erascu, M. A Tool for Fake News Detection. 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, Romania, 2018, pp.379-386. https://doi.org/10.1109/SYNASC.2018.00064

[37] https://en.wikipedia.org/wiki/Machine_learning

[38] https://en.wikipedia.org/wiki/Supervised_learning

[39] https://en.wikipedia.org/wiki/Unsupervised_learning

[40] https://en.wikipedia.org/wiki/Reinforcement_learning

[41] https://en.wikipedia.org/wiki/Confusion_matrix

[42] https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

[43] https://marcobonzanini.com/2015/01/26/stemming-lemmatisation-and-pos-tagging-with-python-and-nltk/

[44] Kaggle, Fake News, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/c/fake-news

[45] Kaggle, Fake News Detection, Kaggle, San Francisco, CA, USA, 2018, https://www.kaggle.com/jruvika/fake-news-detection

[46] https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

[47] https://medium.com/mlearning-ai/nlp-tokenization-stemming-lemmatization-and-part-of-speech-tagging-9088ac068768

[48] https://paperswithcode.com/dataset/fakenewsnet

[49] https://bdtechtalks.com/2020/11/12/what-is-ensemble-learning/

[50] https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/

[51] https://www.ibm.com/cloud/learn/overfitting

[52] https://scikit-learn.org/stable/supervised_learning.html#supervised-learning

[53] https://www.svm-tutorial.com/2014/10/svm-linear-kernel-good-text-classification/

[54] https://www.forbes.com/sites/petersuciu/2021/08/02/spotting-misinformation-on-social-media-is-increasingly-challenging/?sh=437996892771

[55] https://ieeexplore.ieee.org/abstract/document/9049290

[56] https://dl.acm.org/doi/abs/10.1145/3339252.3341497

[57] https://www.sciencedirect.com/science/article/abs/pii/S1568494621007006

[58] https://www.sciencedirect.com/science/article/abs/pii/S1568494620309881

ΠΑΡΑΡΤΗΜΑ

Αλφαβητική λίστα με τους συμβολισμούς που χρησιμοποιούνται στην τεχνική προ-επεξεργασίας δεδομένων Part-of-speech Tagging [46][47].

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb