

Ιόνιο Πανεπιστήμιο

Τμήμα Πληροφορικής



Πτυχιακή Εργασία

**Ανίχνευση Ψευδών Ειδήσεων στο Twitter
μέσω Μηχανικής Μάθησης**

Ευάγγελος Νικολάου

Επιβλέπουσα: Αικατερίνη Κερμανίδου

13 Φεβρουαρίου 2023

Επιβλέπουσα

Αικατερίνη Κερμανίδου, *Αν. Καθηγήτρια,*
Ιόνιο Πανεπιστήμιο

Τριμελής Επιτροπή

Αικατερίνη Κερμανίδου, *Αν. Καθηγήτρια,*
Ιόνιο Πανεπιστήμιο

Ανδρέας Καναβός, *Αν. Καθηγητής,*
Ιόνιο Πανεπιστήμιο

Θεμιστοκλής Έξαρχος, *Αν. Καθηγητής,*
Ιόνιο Πανεπιστήμιο

Αφιερώνεται στους γονείς μου

Περίληψη

Η χρήση του διαδικτύου έχει επιφέρει νέους τρόπους επικοινωνίας μεταξύ των ανθρώπων. Τα μέσα κοινωνικής δικτύωσης έχουν ενσωματωθεί στην καθημερινότητα εκατοντάδων εκατομμυρίων ανθρώπων καλύπτοντας ένα ευρύ φάσμα αναγκών τους και μεταξύ αυτών και την ανάγκη για ενημέρωση. Τα τελευταία χρόνια, σημαντικές διαστάσεις έχει λάβει το πρόβλημα της διάδοσης ψευδών ειδήσεων στο twitter εξαιτίας μιας σειράς πλεονεκτημάτων που αυτό που προσφέρει (χαμηλό κόστος και ευκολία). Ένας τρόπος αντιμετώπισής του είναι η επιστράτευση μεθόδων και τεχνικών μηχανικής μάθησης. Στα πλαίσια της παρούσας εργασίας δοκιμάστηκαν επιβλεπόμενοι αλγόριθμων μηχανικής μάθησης που προσφέρονται από το περιβάλλον του Weka που προσφέρει φιλικότητα προς τον χρήστη, μεγάλη ποικιλία αλγόριθμων και η ευκολία επεξεργασίας δεδομένων εκπαίδευσης.. Τα αρχικά δεδομένα στα οποία βασίστηκε η εργασία αφορούν τη διάδοση ψευδών ειδήσεων στο Twitter όπως αυτή έλαβε χώρα στις διαδηλώσεις του Hong Kong το 2019 [1].

Μετά από επιλογή των χαρακτηριστικών ώστε να επιτευχθεί μείωση της διάστασης, προέκυψαν δύο σύνολα εκπαίδευσης με 8 και 4 χαρακτηριστικά, αντίστοιχα. Όλοι οι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης πέτυχαν επιδόσεις ως προς την ακρίβεια και την ανάκληση πάνω από 90 % με τους καλύτερους να είναι το νευρωνικό δίκτυο Perceptron πολλαπλών στρωμάτων και το δένδρο απόφασης C4.5 (J-48) με μέση επίδοση ακρίβειας, ανάκλησης και F_1 στο 93%. Αυτό σημαίνει ότι η αποτελεσματική ανίχνευση ψευδών ειδήσεων είναι εφικτή εφόσον δημιουργηθούν πιο αντιπροσωπευτικά δεδομένα εκπαίδευσης ενισχυμένα με πληροφορίες σχετικά με τους αποστολείς.

Περιεχόμενα

A'	Εισαγωγή	1
A'.1	Κίνητρο και Περιγραφή Προβλήματος	1
A'.2	Αντικείμενο Πτυχιακής Εργασίας	2
A'.3	Οργάνωση Πτυχιακής Εργασίας	2
B'	Ανίχνευση Ψευδών Ειδήσεων στο Twitter	3
B'.1	Περιγραφή Προβλήματος	3
B'.2	Συναφή Προβλήματα	5
B'.3	Προσεγγίσεις Επίλυσης	5
B'.4	Σχετική Βιβλιογραφία	6
Γ'	Μηχανική Μάθηση	8
Γ'.1	Εισαγωγικά Στοιχεία	8
Γ'.2	Είδη Μηχανικής Μάθησης	10
Γ'.2.1	Μάθηση με επίβλεψη (supervised learning)	11
Γ'.2.2	Μάθηση χωρίς επίβλεψη (unsupervised learning)	12
Γ'.3	Εκπαίδευση	13
Γ'.3.1	Συλλογή δεδομένων και Προετοιμασία	13
Γ'.3.2	Κανονικοποίηση Κειμένου	14
Γ'.3.3	Επιλογή Χαρακτηριστικών	15
Γ'.4	Αλγόριθμοι Μάθησης	17
Γ'.4.1	Απλοϊκή Κατηγοριοποίηση κατά Bayes (Naive Bayes)	17
Γ'.4.2	Μηχανές Διανυσμάτων Στήριξης	18
Γ'.4.2.1	Ακολουθιακή Βελτιστοποίηση Ελαχίστου	18

Γ'.4.3	Λογιστική Παλινδρόμηση	19
Γ'.4.4	Κ Πλησιέστεροι Γείτονες	19
Γ'.4.5	Δένδρα Απόφασης	20
Γ'.4.5.1	Αλγόριθμοι J48 - ID3	21
Γ'.4.6	Κανόνες απόφασης και αλγόριθμος One Rule	22
Γ'.4.7	Νευρωνικά Δίκτυα Perceptron Πολλών Στρωμάτων	23
Γ'.4.8	Συνδυαστικοί Αλγόριθμοι Μάθησης	24
Γ'.4.8.1	Ο Αλγόριθμος Bagging	24
Γ'.4.8.2	Ο Αλγόριθμος Adaboost	25
Γ'.4.9	Αξιολόγηση Μοντέλων Ταξινόμησης	25
Δ'	Πείραμα	27
Δ'.1	TWITTER	27
Δ'.2	Προγραμματιστικό Περιβάλλον	30
Δ'.2.1	Αρχεία στο WEKA	32
Δ'.3	Δεδομένα	34
Δ'.4	Μεθοδολογία	38
Ε'	Συμπεράσματα	48
	Βιβλιογραφία	49
	Συντμήσεις	53
	Γλωσσάρι Ξενικών Όρων	54

Κατάλογος σχημάτων

Γ'.1 Παράδειγμα δένδρου απόφασης ψευδών ειδήσεων με χρήση δύο μόνο χαρακτηριστικών (εντροπίας H και μήκους κειμένου len).	21
Γ'.2 Τοπολογία δικτύου Perceptron πολλών στρωμάτων. Υπάρχει το στρώμα εισόδου (input layer), το στρώμα εξόδου (output layer) και τα ενδιάμεσα (κρυφά-hidden) στρώματα.	23
Δ'.3 Ο ιστότοπος της πλατφόρμας Twitter.	28
Δ'.4 Στιγμιότυπο εμφάνισης tweets.	29
Δ'.5 Αρχική οθόνη εμφάνισης WEKA.	31
Δ'.6 Οθόνη Explorer του WEKA με φορτωμένα τα fakes news δεδομένα. .	34
Δ'.7 Ενδεικτικό νέφος λέξεων με βάση τις συχνότητες εμφάνισης στα κείμενα των fake news tweets [1]	36
Δ'.8 Εφαρμογή του αλγόριθμου Naïve Bayes στα σύνολα δεδομένων μετά από επιλογή 8 και 4 χαρακτηριστικών, αντίστοιχα.	42
Δ'.9 Εφαρμογή του αλγόριθμου Logistic Regression.	43
Δ'.10 Εφαρμογή του αλγόριθμου SVM.	44
Δ'.11 Εφαρμογή του αλγόριθμου Δέντρου Απόφασης C4.5 (J48).	45
Δ'.12 Εφαρμογή Νευρωνικού Δικτύου Perceptron Πολλαπλών Στρωμάτων.	46

Κατάλογος πινάκων

Δ'.1 Συγκεντρωτικός πίνακας των χαρακτηριστικών των tweets.	35
Δ'.2 Συγκεντρωτικός πίνακας αξιολόγησης των ταξινομητών για το σύνολο δεδομένων εκπαίδευσης με τα 8 χαρακτηριστικά.	41
Δ'.3 Συγκεντρωτικός πίνακας αξιολόγησης των ταξινομητών για το σύνολο δεδομένων εκπαίδευσης με τα 4 χαρακτηριστικά.	47

Κεφάλαιο Α΄

Εισαγωγή

Α΄.1 Κίνητρο και Περιγραφή Προβλήματος

Η ραγδαία εξάπλωση της χρήσης του διαδικτύου, οδήγησε σε νέους τρόπους ψηφιακής επικοινωνίας μεταξύ των ανθρώπων με αποτέλεσμα την εμφάνιση των μέσων κοινωνικής δικτύωσης όπως το Facebook, το Twitter και άλλα. Με πολύ μικρό οικονομικό κόστος, καλύπτουν σημαντικές ανάγκες των χρηστών τους και γι' αυτό η εξάπλωσή τους είναι εκθετική. Οι χρήστες του διαδικτύου δεν είναι παθητικοί αποδέκτες πληροφοριών αλλά μέσω των κοινωνικών μέσων δικτύωσης μπορούν να κάνουν αναρτήσεις γενικού περιεχομένου, να ανταλλάσσουν απόψεις, καθώς και να αναρτούν κριτικές με την άποψή τους όσον αφορά συγκεκριμένα θέματα.

Με τη διάδοση των μέσων κοινωνικής δικτύωσης (ΜΚΔ), κάποια προϋπάρχοντα προβλήματα επιδεινώθηκαν προκαλώντας αυξημένη ανησυχία. Ένα τέτοιο πρόβλημα είναι αυτό της παραπληροφόρησης και της **διάδοσης ψευδών ειδήσεων (fake news)**. Ιδεολογικά και οικονομικά κίνητρα σε συνδυασμό με το χαμηλό κόστος χρήσης των ΜΚΔ έχουν οδηγήσει σε γιγάντωση του φαινομένου της διάδοσης ψευδών ειδήσεων. Ο τεράστιος όγκος των παραγόμενων δεδομένων, έχει κάνει επιτακτική την ανάγκη της αυτοματοποιημένης/αλγοριθμικής αντιμετώπισής του ώστε να περιοριστούν οι εξαιρετικά αρνητικές συνέπειες των στόχων των ψευδών ειδήσεων όπως η απαξίωση και η αποξένωση.

Α'.2 Αντικείμενο Πτυχιακής Εργασίας

Η παρούσα εργασία ασχολείται με την εξόρυξη γνώσης από κοινωνικά δίκτυα και συγκεκριμένα, με την αυτοματοποιημένη ανίχνευση ψευδών ειδήσεων (fake news) στο κοινωνικό δίκτυο twitter. Η καθιερωμένη προσέγγιση σε προβλήματα αυτής της μορφής βασίζεται σε μεθόδους και τεχνικές Μηχανικής Μάθησης όπου μετά τη συλλογή δεδομένων εκπαίδευσης σχετικών με την εγκυρότητα των ειδήσεων (αληθείς/ψευδείς), λαμβάνει χώρα η εκπαίδευση και αξιολόγηση των διαθέσιμων ταξινομητών (classifiers) για διάφορα σύνολα παραμέτρων.

Το προγραμματιστικό περιβάλλον/πλατφόρμα που χρησιμοποιείται είναι το Weka, το οποίο θεωρείται ιδιαίτερα πετυχημένο σε προβλήματα ανάλυσης δεδομένων με τεχνικές μηχανικής μάθησης και γι' αυτό είναι πολύ διαδεδομένο στους κόλπους της ερευνητικής κοινότητας. Το τελικό επιθυμητό αποτέλεσμα είναι η δημιουργία ενός ταξινομητή μηνυμάτων twitter (tweets) ο οποίος με μεγάλη πιθανότητα θα μπορεί να προβλέπει την εγκυρότητα νέων μηνυμάτων. Τέλος, τα αποτελέσματα για διάφορους ταξινομητές παρουσιάζονται και αναλύονται.

Α'.3 Οργάνωση Πτυχιακής Εργασίας

Η παρούσα πτυχιακή εργασία περιλαμβάνει πέντε (5) κεφάλαια. Στο πρώτο εισαγωγικό κεφάλαιο αναφέρονται τα κίνητρα της παρούσας μελέτης καθώς και οι αντίστοιχοι στόχοι. Ακολουθεί το δεύτερο κεφάλαιο όπου παρουσιάζεται αναλυτικά το πρόβλημα της ανίχνευσης ψευδών ειδήσεων στο κοινωνικό δίκτυο twitter. Στο τρίτο κεφάλαιο παρουσιάζονται συνοπτικά οι βασικές έννοιες και οι τεχνικές μηχανικής μάθησης που χρησιμοποιούνται για την επίλυση του παραπάνω προβλήματος ανίχνευσης. Το τέταρτο κεφάλαιο περιγράφει λεπτομερώς τα στάδια του πειράματος ανίχνευσης ψευδών ειδήσεων από τη επιλογή των δεδομένων εκπαίδευσης μέχρι την παραγωγή των σχετικών αποτελεσμάτων επεξεργασίας. Το τελευταίο κεφάλαιο παρουσιάζει τα συμπεράσματα που προέκυψαν από τα πειραματικά αποτελέσματα.

Κεφάλαιο Β΄

Ανίχνευση Ψευδών Ειδήσεων στο Twitter

Β.1 Περιγραφή Προβλήματος

Τα μέσα κοινωνικής δικτύωσης έχουν καταργήσει τα φυσικά σύνορα μεταξύ των χωρών και επιτρέπουν την επικοινωνία και την ταχύτατη ανταλλαγή πληροφοριών σε οποιοδήποτε μέρος του κόσμου. Αποτέλεσμα αυτής της δυνατότητας είναι η χρήση των περισσότερων από αυτά (Facebook, Twitter, κ.α.) για τη διάδοση τοπικών αλλά και διεθνών ειδήσεων.

Παραδοσιακά, οι ειδήσεις, συλλογή και διάδοση, είναι το αντικείμενο της δημοσιογραφίας, η οποία σε γενικές γραμμές ειδικεύεται στην συλλογή και παρουσίαση πληροφοριών που αφορούν δημόσια δρώμενα, εκδηλώσεις και γεγονότα με τέτοιο τρόπο ώστε οι ενδιαφερόμενοι πολίτες να είναι σε θέση να διαμορφώσουν ελεύθερα την δική τους άποψη και οπτική μετά από κριτική εξέταση των πληροφοριών [2, 3]. Βασικά χαρακτηριστικά των πληροφοριών πρέπει να είναι η ακρίβεια, η αξιοπιστία, η εγκυρότητα, η πληρότητα και η αντικειμενικότητα και επομένως, όλοι οι συντελεστές στον χώρο της δημοσιογραφίας θα πρέπει να προωθούν την αλήθεια την οποία και θα πρέπει να διασταυρώνουν με όποιο τρόπο μπορούν. Για να πραγματοποιηθούν τα παραπάνω, θα πρέπει το κοινωνικοπολιτικό περιβάλλον στο οποίο δραστηριοποιούνται οι φορείς της δημοσιογραφίας να χαρακτηρίζεται από ελευθεροτυπία χωρίς οποιουδήποτε είδους λογοκρισία ή εξωτερικό έλεγχο, ισότητα με

την έννοια της απουσίας μεροληψίας ή διακρίσεων και να υφίσταται η αρχή της διαφορετικότητας [4].

Παρόλο που τα φαινόμενα παραπληροφόρησης και διάδοσης ψευδών ειδήσεων υπάρχουν από τα αρχαία ακόμα χρόνια, η ψηφιοποίηση, το διαδίκτυο και τα μέσα κοινωνικής δικτύωσης έχουν λειτουργήσει ως πολλαπλασιαστές της έντασής τους. Η εγκυρότητα των ειδήσεων που μεταδίδονται με αυτόν τον τρόπο δεν είναι εγγυημένη και είναι δυνατή η αλλοίωσή τους από διάφορους παράγοντες που έχουν συμφέρον να το κάνουν για πολιτικούς, ιδεολογικούς ή οικονομικούς σκοπούς [1]. Οι αλλοιωμένες αυτές ειδήσεις αναφέρονται ως **ψευδείς ειδήσεις (fake news)** και αποτελούν φαινόμενο που έχει λάβει τόσο μεγάλες διαστάσεις ώστε να είναι επιτακτικός ο ακριβής ορισμός του αλλά και η αντιμετώπισή του. Η ανεξέλεγκτη διάδοση των ψευδών ειδήσεων μπορεί να υποβαθμίσει σημαντικά τις αξίες και τις αρχές των ιδεών που αλλοιώνουν.

Παρόλο που η εμπιστοσύνη των πολιτών στις πηγές πληροφόρησης είναι μειωμένη, η έκθεση σε ψευδείς ειδήσεις μπορεί να οδηγήσει σε στάσεις αποξένωσης, ανεπάρκειας και κυνισμού απέναντι σε υποψήφιους πολιτικούς [5]. Επίσης, οι ψευδείς ειδήσεις συνδέονται με πράξεις βίας που απειλούν την δημόσια ασφάλεια [6, 7].

Πιο ειδικά, η διάδοση ψευδών ειδήσεων σε θέματα πολιτικού περιεχομένου μέσω Twitter έχει λάβει μεγάλες διαστάσεις κυρίως λόγω της ευκολίας, του χαμηλού κόστους, αλλά και της ταχύτητας αναπαραγωγής. Μελέτες έχουν δείξει ότι η αναπαραγωγή αυτή είναι πολύ ταχύτερη από την αντίστοιχη των αληθών (πραγματικών) ειδήσεων με προφανές αποτέλεσμα την υποβάθμισή τους [8].

Η παραδοσιακή μέθοδος εντοπισμού ψευδών ειδήσεων βασίζεται στην επιστροφή εμπειρών και πιστοποιημένων δημοσιογράφων. Ωστόσο, η μέθοδος αυτή πέραν του ότι συνοδεύεται από υψηλό οικονομικό κόστος, δεν είναι αρκετά γρήγορη ώστε να αντιμετωπίσει την εκρηκτική ταχύτητα διάδοσης των ψηφιακών μέσων δικτύωσης. Ενδεικτικά αναφέρουμε ότι συχνά απαιτούνται πάνω από δύο έμπειροι δημοσιογράφοι οι οποίοι ανεξάρτητα θα πρέπει να αποφανθούν για την εγκυρότητα μιας είδησης [6].

Για τον λόγο αυτό, είναι επιτακτική ανάγκη να αυτοματοποιηθεί η παραπάνω διαδικασία. Αυτό γίνεται με την ανάπτυξη προγραμμάτων που βασίζονται στην Τεχνητή Νοημοσύνη (TN), την Μηχανική Μάθηση (MM) και την Επεξεργασία Φυσικής Γλώσ-

σας (Natural Language Processing - NLP).

Β'.2 Συναφή Προβλήματα

Προβλήματα που είναι σχετικά με το πρόβλημα της ανίχνευσης ψευδών ειδήσεων είναι:

- Έλεγχος-Επαλήθευση Γεγονότος (Fact Checking). Αναφέρεται στην διαδικασία επαλήθευσης ισχυρισμών που διατυπώνονται από δημόσια πρόσωπα [9]. Συγγενεύει με την ανίχνευση ψευδών ειδήσεων αλλά είναι γενικότερο.
- Ανίχνευση Φημολογίας (Rumor Detection). Είναι ο εντοπισμός ανεπιβεβαιωτών πληροφοριών τη στιγμή της δήλωσης [10].
- Ανίχνευση Πολιτικής Θέσης (Stance Detection). Αναφέρεται στον εντοπισμό της πολιτικής θέσης κάποιου αναλύοντας κείμενά που δημοσιεύει και μπορεί να θεωρηθεί υποκατηγορία του προβλήματος ανίχνευσης ψευδών ειδήσεων [11].
- Ανάλυση Συναισθήματος (Sentiment Analysis). Είναι η διαδικασία εντοπισμού προτιμήσεων και συναισθημάτων και όχι επιβεβαίωσης της αλήθειας γεγονότων ή ειδήσεων.

Β'.3 Προσεγγίσεις Επίλυσης

Όπως υποδηλώνει και το όνομά του, το πρόβλημα της ανίχνευσης ψευδών ειδήσεων είναι πρόβλημα δυαδικής κατηγοριοποίησης με την έννοια ότι μια είδηση θα πρέπει να αποφασισθεί αν είναι πραγματική (αληθής) ή όχι (ψευδής). Αρκετά συχνά όμως κάποια είδηση μπορεί να είναι μερικώς αληθής. Για τον λόγο αυτό, επιπρόσθετες κατηγορίες χρησιμοποιούνται, ανάλογα με τα διαθέσιμα δεδομένα εκπαίδευσης και έτσι προκύπτουν multi-class ταξινομητές.

Υπάρχουν όμως και προσεγγίσεις που μπορούμε να το θεωρήσουμε ως πρόβλημα παλινδρόμησης (regression) όπου η έξοδος είναι ένας αριθμός που παριστάνει την εγγύτητα της είδησης στην αλήθεια [12]. Το μειονέκτημα αυτής της προσέγγισης

είναι ότι απαιτεί δεδομένα εκπαίδευσης με αριθμητική εκτίμηση της αλήθειας των ειδήσεων που είναι πολύ λίγα [6].

Στην παρούσα εργασία ενδιαφερόμαστε για ειδήσεις με τη μορφή κειμένου με τους περιορισμούς που επιβάλλει η πλατφόρμα Twitter.

Β'.4 Σχετική Βιβλιογραφία

Στα πλαίσια της χρήσης τεχνικών NLP, πολύ καλές επιδόσεις ανίχνευσης ψευδών ειδήσεων επιτεύχθηκαν μέσω της ανάλυσης του κειμένου με τη βοήθεια των N-gram χαρακτηριστικών και στη συνέχεια εφαρμογή τεχνικών ML για κατηγοριοποίηση [13]. Υπενθυμίζεται ότι τα N-grams είναι ακολουθίες N διαδοχικών στοιχείων κειμένου (γράμματα, συλλαβές, λέξεις κ.α.). Τα πιο συχνά χρησιμοποιούμενα N-grams στην ανάλυση κειμένων βασίζονται σε γράμματα ή λέξεις και στην συγκεκριμένη εργασία χρησιμοποιήθηκαν λέξεις. Η κύρια ιδέα της χρήσης των N-grams έγκειται στο ότι οι συχνότητες εμφάνισής τους στα διάφορα tweets μπορούν να διαφοροποιήσουν τα αληθή από τα ψευδή. Προαπαιτούμενη της επιτυχούς χρήσης της μεθόδου αυτής είναι η κατάλληλη προ-επεξεργασία των κειμένων ώστε να αφαιρεθούν οι συχνές λέξεις (stop words) και να μετασχηματισθούν οι λέξεις σε κανονική μορφή (stemming). Το μοντέλο με τις καλύτερες επιδόσεις είναι η γραμμική SVM με χρήση Term Frequency-Inverted Document Frequency (TF-IDF) χαρακτηριστικών. Χρησιμοποιήθηκαν N-grams μεγέθους μέχρι 4 και δοκιμάστηκαν 6 διαφορετικές τεχνικές κατηγοριοποίησης: SVM, γραμμική SVM, k-NN, Decision Trees, Stochastic Gradient Descent (SGD) και Logistic Regression. Αξιοσημείωτο είναι το εύρημα ότι οι επιδόσεις των ταξινομητών φθίνουν καθώς το μέγεθος των N-grams αυξάνει και επομένως η βέλτιστη τιμή είναι $N=1$.

Επίσης, υβριδικές προσεγγίσεις ανίχνευσης ψευδών ειδήσεων που συνδυάζουν γλωσσολογικές τεχνικές με τεχνικές ανάλυσης κοινωνικών δικτύων δίνουν καλά αποτελέσματα [14]. Στηρίζονται στο σκεπτικό ότι τα δικτυακά μετα-δεδομένα που είναι διαθέσιμα μπορούν να συνεισφέρουν χαρακτηριστικά τα οποία καθιστούν τις κλάσεις των ψευδών και αληθινών ειδήσεων καλύτερα διαχωρίσιμες.

Οι Buntain και Golbeck επίδιωξαν την αυτοματοποίηση της ανίχνευσης ψευδών ειδήσεων χρησιμοποιώντας 2 υπάρχοντα σύνολα δεδομένων για να αναλύσουν και

να εκτιμήσουν την εγγύτητα δυνητικών απειλών στο Twitter μέσω του ελεύθερα διαθέσιμου συνόλου δεδομένων Buzzfeed [1, 15]. Τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι τα PHEME και CREDBANK. Η επιλογή των χαρακτηριστικών έγινε μέσω επαναληπτικής αφαίρεσης. Αρχικά χρησιμοποιήθηκαν όλα τα χαρακτηριστικά και ακολούθως, κάθε φορά εξαιρούνταν εκείνο με τη μικρότερη επίδραση στις επιδόσεις των μοντέλων. Κατέληξαν να κρατήσουν 7 και 12 από τα 45 χαρακτηριστικά για τα σύνολα PHEME και CREDBANK, αντίστοιχα.

Παρόλο που υπάρχουν διαθέσιμα αρκετά καθιερωμένα σύνολα δεδομένων εκπαίδευσης, παρατηρείται ότι υπάρχει η τάση των ερευνητών να πειραματίζονται και να δοκιμάζουν μοντέλα σε συγκεκριμένα γεγονότα/συμβάντα στα οποία παρατηρήθηκε διάδοση ψευδών ειδήσεων. Αυτό οφείλεται στο ότι για νέα συμβάντα είναι δύσκολο να βρεθούν χαρακτηρισμένα δεδομένα εκπαίδευσης. Αποτέλεσμα αυτής της έλλειψης είναι η προσπάθεια ανάπτυξης data-driven τεχνικών στις οποίες η ύπαρξη δεδομένων εκπαίδευσης δεν είναι απαραίτητη. Ανάλογα με την διαθέσιμη επιπρόσθετη πληροφορία καλές επιδόσεις μπορούν να προκύψουν [16]. Στην συγκεκριμένη προσέγγιση, τα tweets χαρακτηρίζονται με αυτόματο τρόπο με τον βαθμό αξιοπιστίας (credibility) της πηγής. Με τον τρόπο αυτό αυτοματοποιείται μερικώς η δημιουργία δεδομένων εκπαίδευσης μέσω επιλογής αξιόπιστων πηγών. Η όλη διαδικασία αναφέρεται ως εκπαίδευση με μερική επίβλεψη (weakly-supervised learning) και οι συγγραφείς αναφέρουν ότι η απουσία δεδομένων εκπαίδευσης που έχουν παραχθεί από εξειδικευμένους δημοσιογράφους μπορεί να αντισταθμιστεί αποτελεσματικά από την μέθοδο που προτείνουν.

Κεφάλαιο Γ΄

Μηχανική Μάθηση

Σε αυτό το κεφάλαιο παρουσιάζονται συνοπτικά οι καθιερωμένες μέθοδοι κατηγοριοποίησης κειμένων με χρήση τεχνικών μηχανικής μάθησης. Μετά από μια σύντομη εισαγωγή στις βασικές έννοιες της μηχανικής μάθησης, παρουσιάζονται οι παραδοσιακές προσεγγίσεις κατηγοριοποίησης οι οποίες χαρακτηρίζονται από το ότι απαιτούν την διαδικασία της μηχανικής χαρακτηριστικών (feature engineering), όπου τα χαρακτηριστικά θα πρέπει να σχεδιαστούν εκ των προτέρων από τον άνθρωπο και στη συνέχεια να δοθούν στον αλγόριθμο μηχανικής μάθησης. Ενδεικτικά αναφέρουμε ότι στον αντίποδα βρίσκονται τα τεχνητά νευρωνικά δίκτυα όπου ο εντοπισμός των καταλληλότερων χαρακτηριστικών είναι μέρος της εκπαίδευσης του δικτύου. Εκτός από τους κύριους αλγορίθμους μηχανικής μάθησης, θα περιγραφούν και ορισμένες τεχνικές μηχανικής χαρακτηριστικών, οι οποίες είναι απαραίτητες για την δημιουργία των καλών (αντιπροσωπευτικών) χαρακτηριστικών για πολλά από τα μοντέλα. Στόχος αυτού του κεφαλαίου είναι η παρουσίαση των βασικών χαρακτηριστικών και των ιδιοτήτων κάθε αλγορίθμου.

Γ.1 Εισαγωγικά Στοιχεία

Για την κατανόηση του εξωτερικού περιβάλλοντος, μια συνήθης πρακτική είναι η δημιουργία απλουστευμένων αναπαραστάσεων οι οποίες διατηρούν τα κύρια μόνο χαρακτηριστικά του υπό μελέτη φαινομένου και αγνοούν τα επουσιώδη και δευτερεύοντα. Οι απλουστευμένες αυτές αναπαραστάσεις ονομάζονται *μοντέλα (models)*

και ένας από τους τρόπους δημιουργίας τους είναι η *επαγωγική μάθηση* (*inductive learning*) η οποία στηρίζεται στην υπόθεση ότι τα χαρακτηριστικά που έχουμε στη διάθεσή μας για την δημιουργία του μοντέλου είναι αντιπροσωπευτικά (*επαγωγή - induction*) [17, 18]. Στην παραπάνω διαδικασία, σημαντική είναι η δυνατότητα εντοπισμού ομοιοτήτων μεταξύ αντικειμένων, μέσω των χαρακτηριστικών τους, ώστε να προκύψουν ομαδοποιήσεις ή συστάδες αντιπροσωπευτικές της ιεραρχικής δομής τους. Η όλη διαδικασία αναφέρεται ως *ταίριασμα ή αναγνώριση προτύπων* (*pattern matching ή recognition*).

Η *Μηχανική Μάθηση* (*Machine Learning*) ως κλάδος της Τεχνητής Νοημοσύνης, στοχεύει στην επινόηση αποτελεσματικών αλγόριθμων δημιουργίας μοντέλων και εύρεσης προτύπων δοσμένων αντιπροσωπευτικών δειγμάτων δεδομένων. Ένας διαδεδομένος και εύληπτος ορισμός της Μηχανικής Μάθησης, που οφείλεται στον A. Samuel (1959) είναι ο παρακάτω:

«Η μηχανική μάθηση είναι ένα πεδίο μελέτης, το οποίο στοχεύει στο να καταστήσει τους υπολογιστές ικανούς να μαθαίνουν χωρίς να έχουν προγραμματιστεί αυστηρά.»

Ένας πιο αυστηρός ορισμός για το τι εννοούμε όταν λέμε ότι ένα πρόγραμμα μαθαίνει, δόθηκε από τον Tom M. Mitchell [19]:

«Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο επίδοσης P , αν η επίδοσή του σε εργασίες της κλάσης T , όπως αποτιμάται από το μέτρο P , βελτιώνεται με την εμπειρία E .»

Με βάση τα παραπάνω, κατά την διαδικασία της μηχανικής μάθησης, υπάρχει ένα σύνολο δεδομένων εκπαίδευσης (*training set*), και ένας αλγόριθμος μάθησης (*learning algorithm*) ο οποίος επεξεργαζόμενος τα δεδομένα παράγει ένα μοντέλο το οποίο μπορεί να κάνει επιτυχημένες προβλέψεις πάνω στα δεδομένα αυτά. Με βάση την επαγωγική υπόθεση, αναμένεται το μοντέλο αυτό να μπορεί να δώσει επιτυχημένες προβλέψεις και για νέα δεδομένα που δεν χρησιμοποιήθηκαν στο σύνολο δεδομένων εκπαίδευσης.

Υπάρχει πληθώρα προβλημάτων στα οποία η μηχανική μάθηση μπορεί να εφαρμοστεί: αναγνώριση φωνής, αναγνώριση προσώπου, αναγνώριση χαρακτηριστικών κινήσεων σε video, κατάταξη ακτινογραφιών σε παθολογικές ή μη, αναγνώριση συναισθήματος από εικόνες ή κείμενα, εντοπισμός ψευδών ειδήσεων σε κείμενα και

πολλά άλλα.

Για την επίτευξη των στόχων που περιγράψαμε παραπάνω, η μηχανική μάθηση βασίζεται σε μεγάλο βαθμό στη θεωρία της στατιστικής. Ο κλάδος αναφέρεται ως *Στατιστική Μηχανική Μάθηση (Statistical Machine Learning)* και μελετά τις στατιστικές ιδιότητες των μοντέλων που παράγονται μέσω της επεξεργασίας δειγμάτων εκπαίδευσης [18]. Τι ιδιότητες πρέπει να έχει το δείγμα; Ποιο πρέπει να είναι το μέγεθός του; Πόσο καλές προβλέψεις μπορούμε να αναμένουμε; Με ποια κριτήρια αξιολογούνται οι προβλέψεις;

Σημαντικό ρόλο διαδραματίζει και η Επιστήμη των Υπολογιστών (Computer Science). Πολλά από τα εμπλεκόμενα προβλήματα είναι δυσεπίλυτα και επομένως, απαιτείται η χρήση όλων των διαθέσιμων αλγοριθμικών τεχνικών. Επίσης, σε αντίθεση με τα πρώτα προβλήματα της μηχανικής μάθησης κατά τις δεκαετίες του 70 και 80, ο όγκος των δεδομένων είναι τεράστιος και σε κάποιες περιπτώσεις ακόμα και η τετραγωνική πολυπλοκότητα δεν είναι αποδεκτή, αφού καθιστά την διαδικασία της μάθησης μη εφαρμόσιμη λόγω του χρόνου ή της μνήμης που απαιτείται. Συχνά είναι επιτακτική ανάγκη να επιστρατευθούν και να εφαρμοστούν παράλληλοι αλγόριθμοι εισάγοντας έτσι επιπλέον δυσκολίες για θέματα συγχρονισμού και επικοινωνίας.

Τέλος, σημαντική γνώση έχει αντληθεί από την Ανάκτηση Πληροφοριών (Information Retrieval), αντικείμενο της οποίας είναι η αποδοτική ανάκτηση πληροφορίας από συλλογές δεδομένων κειμένου (διαδίκτυο, ψηφιακές βιβλιοθήκες). Τυπικό πρόβλημα εδώ είναι η ανάκτηση όλων των κειμένων (εγγράφων) που περιέχουν συγκεκριμένες λέξεις-κλειδιά. Εδώ εμφανίστηκαν για πρώτη φορά, δύο βασικά μέτρα αξιολόγησης της αποτελεσματικότητας ανάκτησης των εγγράφων, η *ακρίβεια (precision)* και η *ανάκληση (recall)* καθώς και μέτρα ομοιότητας όπως αυτό της *αντίστροφης συχνότητας εγγράφου (inverse document frequency)*.

Γ'.2 Είδη Μηχανικής Μάθησης

Η μηχανική μάθηση χωρίζεται σε δύο κύρια είδη : στη μάθηση με επίβλεψη και τη μάθηση χωρίς επίβλεψη.

Γ'.2.1 Μάθηση με επίβλεψη (supervised learning)

Ο αλγόριθμος μάθησης βασίζεται σε παραδείγματα (examples) ή δεδομένα εκπαίδευσης (training data) για τα οποία είναι γνωστή η ορθή έξοδος του μοντέλου. Ο αλγόριθμος μάθησης προσπαθεί να βρει την καλύτερη αντιστοίχιση μεταξύ εισόδου και εξόδου ώστε να ελαχιστοποιηθεί κάποιο κριτήριο σφάλματος.

Ουσιαστικά, ο αλγόριθμος μάθησης πρέπει να μάθει επαγωγικά μία συνάρτηση που αναφέρεται ως συνάρτηση στόχος (target function) και αποτελεί την μαθηματική έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η συνάρτηση αυτή χρησιμοποιείται για την πρόβλεψη της τιμής μιας μεταβλητής (συνήθως πολυδιάστατης), που ονομάζεται μεταβλητή εξόδου, βάσει των τιμών ενός συνόλου μεταβλητών, που ονομάζονται μεταβλητές εισόδου ή χαρακτηριστικά (attributes ή features). Όπως αναφέρθηκε παραπάνω, η επαγωγική μάθηση βασίζεται στην επαγωγική υπόθεση: Η υπόθεση Η που προσεγγίζει καλά την συνάρτηση στόχο για ένα αρκετά μεγάλο σύνολο δοσμένων παραδειγμάτων, θα την προσεγγίζει το ίδιο καλά και για νέες τιμές εισόδου.

Στη μάθηση με επίβλεψη διακρίνονται δύο είδη προβλημάτων (learning tasks):

- i Την ταξινόμηση ή κατηγοριοποίηση (classification) όπου ο στόχος είναι η δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων / κατηγοριών) (π.χ. αληθής ή ψευδής είδηση).
- ii Την παρεμβολή ή παλινδρόμηση (regression) όπου το μοντέλο προβλέπει αριθμητικές τιμές (π.χ. πρόβλεψη βαθμού αλήθειας μιας είδησης).

Η καθιερωμένη μεθοδολογία στην επιβλεπόμενη μάθηση ακολουθεί τα εξής βήματα:

1. Επιλογή των δεδομένων. Στόχος είναι η δημιουργία του συνόλου δεδομένων εισόδου που θα αποτελέσουν τμήμα του συνόλου εκπαίδευσης (training set).
2. Επισήμανση-Χαρακτηρισμός Δεδομένων. Στο βήμα αυτό θα πρέπει τα δεδομένα του προηγούμενου βήματος να αντιστοιχιστούν με την σωστή έξοδο. Συνήθως επιστρατεύονται ειδικοί (experts) στο εκάστοτε αντικείμενο (π.χ. ιατροί για ακτινογραφίες ή δημοσιογράφοι για χαρακτηρισμό ειδήσεων). Έτσι ολοκληρώνεται η δημιουργία του συνόλου εκπαίδευσης.

3. Επιλογή χαρακτηριστικών. Κάθε αντικείμενο εισόδου παριστάνεται από ένα διάνυσμα χαρακτηριστικών το οποίο το περιγράφει. Είναι ουσιώδους σημασίας, τα χαρακτηριστικά αυτά να περιέχουν αρκετή πληροφορία ώστε να διαφοροποιούν τα αντικείμενα μεταξύ τους.
4. Επιλογή μεθόδου εκπαίδευσης. Υπάρχει πληθώρα διαθέσιμων τεχνικών μάθησης με επίβλεψη και ανάλογα με τη φύση και τις ιδιαιτερότητες του προβλήματος πρέπει να επιλεγούν οι καταλληλότερες.
5. Εκπαίδευση Μοντέλου. Ο αλγόριθμος μάθησης τρέχει με είσοδο τα δεδομένα εκπαίδευσης και παράγει ως έξοδο το μοντέλο για την συνάρτηση αντιστοίχισης δηλαδή τον ταξινομητή (classifier).
6. Αξιολόγηση του μοντέλου εκπαίδευσης. Υπολογίζεται η απόδοση του μοντέλου με εφαρμογή της σε ένα σύνολο δεδομένων ελέγχου (δείγμα ελέγχου - test set), που είναι διαφορετικό από το σύνολο εκπαίδευσης.

Γ'.2.2 Μάθηση χωρίς επίβλεψη (unsupervised learning)

Σε αυτό το είδος μάθησης, δεν είναι διαθέσιμα δεδομένα εκπαίδευσης και ο αλγόριθμος μάθησης βασίζεται σε κατάλληλες συναρτήσεις ομοιότητας ή απόστασης μεταξύ των δεδομένων εισόδου ώστε να ανακαλύψει συσχετίσεις ή ομάδες στο σύνολο των δεδομένων, αποκαλύπτοντας έτσι την άγνωστη ιεραρχική δομή τους. Σε αυτή την περίπτωση, η έξοδος είναι πρότυπα, που περιγράφουν ένα μέρος από τα δεδομένα. Ενδεικτικά παραδείγματα είναι οι κανόνες συσχέτισης και οι ομάδες, οι οποίες προκύπτουν από διαδικασίες συσταδοποίησης ή ομαδοποίησης.

Γενικά, οι **κανόνες συσχέτισης (association rules)** χρησιμοποιούνται για την εύρεση συχνών μοτίβων και συσχετίσεων μεταξύ των στοιχείων του συνόλου δεδομένων. Συνήθως, ένας κανόνας συσχέτισης έχει τη μορφή $X \rightarrow Y$, όπου τα X και Y αποτελούν δύο διαφορετικά σύνολα στοιχείων (στοιχειοσύνολα itemsets). Η ισχύς ενός κανόνα συσχέτισης μπορεί να μετρηθεί με βάση την υποστήριξη (support) και την εμπιστοσύνη (confidence) που ορίζονται ως εξής: η υποστήριξη είναι η συχνότητα με την οποία εφαρμόζεται ένας κανόνας σε ένα συγκεκριμένο σύνολο δεδομένων ενώ η εμπιστοσύνη είναι η συχνότητα εμφάνισης των στοιχείων του συνόλου Y , σε συναλλαγές που περιέχουν στοιχεία του συνόλου X .

Τέλος, η **συσταδοποίηση ή ομαδοποίηση (clustering)** επιδιώκει και αυτή όπως η ταξινόμηση να χωρίσει τα δεδομένα εισόδου σε συστάδες ή ομάδες. Η διαφορά με την ταξινόμηση είναι ότι δεν υπάρχουν δεδομένα εκπαίδευσης. Η ομαδοποίηση βασίζεται στις ιδιότητες των αντικειμένων όπως αυτές εκφράζονται από τις συναρτήσεις ομοιότητας ή απόστασης που χρησιμοποιούνται. Επίσης συνήθως το πλήθος των ομάδων είναι άγνωστο και γι' αυτό παράγονται ιεραρχίες ομάδων με την ελπίδα να αποκαλυφθεί η άγνωστη δομή των δεδομένων. Ενδεικτικά παραδείγματα τεχνικών ομαδοποίησης είναι ο αλγόριθμος K-means, η ιεραρχική ομαδοποίηση και η ομαδοποίηση με βάση την πυκνότητα. Ο αλγόριθμος K-means προϋποθέτει πρότερη γνώση του πλήθους K των ομάδων και προσπαθεί επαναληπτικά να βρει εκείνη τη διαμέριση του συνόλου δεδομένων σε K ομάδες ομάδες που ελαχιστοποιεί τη συνάρτηση κόστους. Η ιεραρχική ομαδοποίηση μπορεί να είναι διαιρετική (top-down) ή συγχωνευτική (bottom-up) και δεν απαιτεί γνώση του πλήθους των ομάδων αφού παράγει ιεραρχίες διαμερίσεων (δενδρογράμματα).

Γ'.3 Εκπαίδευση

Γ'.3.1 Συλλογή δεδομένων και Προετοιμασία

Στο πρώτο αυτό στάδιο γίνεται η συγκέντρωση, επιλογή και προετοιμασία των δεδομένων που θα χρησιμοποιηθούν για τη δημιουργία του συνόλου εκπαίδευσης. Γενικά, αυτό μπορεί να σημαίνει είτε ότι τα δεδομένα συλλέγονται από την αρχή, είτε ότι υπάρχουν έτοιμα σύνολα σχετικών δεδομένων τα οποία απαιτούν συναρμολότητα και επεξεργασία. Αν το πρόβλημα είναι καινούριο και πρέπει να επιλεγθούν τα κατάλληλα δεδομένα, τότε αυτή η διαδικασία μπορεί να συνδυαστεί με τη διαδικασία επιλογής των χαρακτηριστικών. Κατασκευάζεται αρχικά ένα σχετικά μικρό σύνολο δεδομένων με όλα τα χαρακτηριστικά που πιθανόν να έχουν αξία στην διαδικασία της μάθησης και ακολούθως, στα πλαίσια εξερευνητικής ανάλυσης, επιλέγονται τα πιο σημαντικά (feature selection) ώστε στη συνέχεια να συλλεχθεί και αναλυθεί το πλήρες σύνολο δεδομένων. Συχνά η δυσκολία βρίσκεται στο ότι υπάρχει ένας τεράστιος όγκος σχετικών δεδομένων, αλλά είναι δύσκολο να συλλεχθούν, είτε επειδή απαιτούνται πολλές μετρήσεις είτε επειδή βρίσκονται σε διάφορες μορφές και είναι δύσκολη η συγχώνευσή τους όπως και η διασφάλιση ότι δεν περιέχουν

σημαντικά λάθη ή ελλιπή δεδομένα.

Επιπλέον, στη μάθηση με επίβλεψη, η σωστή κατηγοριοποίηση των δεδομένων εκπαίδευσης, συχνά απαιτεί τη συμμετοχή ειδικών στον συγκεκριμένο τομέα και σημαντικές επενδύσεις χρόνου και κόστους.

Τέλος, πρέπει να ληφθεί υπόψη και ο όγκος των δεδομένων. Οι αλγόριθμοι μηχανικής μάθησης χρειάζονται ικανό όγκο δεδομένων, κατά προτίμηση δεδομένων που δεν περιέχουν θόρυβο. Όμως αυξημένος όγκος δεδομένων σημαίνει αυξημένο υπολογιστικό κόστος.

Γ'.3.2 Κανονικοποίηση Κειμένου

Στις περιπτώσεις που τα δεδομένα εκπαίδευσης είναι κείμενα, απαιτείται κατάλληλη προετοιμασία τους η οποία λαμβάνει χώρα στο στάδιο της προ-επεξεργασίας. Ουσιαστικά το κείμενο προσαρμόζεται στην κατάλληλη μορφή ώστε να αφαιρεθεί ο θόρυβος που μπορεί να υπάρχει. Στα κείμενα του Twitter (tweets) υπάρχουν αναφορές σε άλλους χρήστες, hashtags, σημεία στίξης, διευθύνσεις ιστοσελίδων και άλλα, τα οποία πρέπει να αφαιρούνται καθώς είναι περιττά και η παρουσία τους λειτουργεί ως θόρυβος που υποβαθμίζει την αποτελεσματικότητα των αλγόριθμων εκπαίδευσης. Η διαδικασία αυτή γίνεται με την βοήθεια των κανονικών εκφράσεων (regular expressions), οι οποίες είναι υπεύθυνες για τον εντοπισμό και την αφαίρεση του θορύβου. Κάποιες από τις πιο διαδεδομένες τεχνικές είναι οι εξής :

- Αφαίρεση αριθμών.
- Αφαίρεση σημείων στίξης.
- Μετατροπή κεφαλαίων γραμμάτων σε πεζά και το αντίστροφο.
- Αφαίρεση συνδέσμων (hyperlinks).
- Λημματοποίηση (tokenization).
- Αποκατάληξη (steming)
- Αφαίρεση συχνών λέξεων (stop-words).

Με αυτό τον τρόπο, όλες οι περιττές πληροφορίες αφαιρούνται και ο αλγόριθμος είναι έτοιμος να επεξεργαστεί το καθαρό κείμενο.

Γ.3.3 Επιλογή Χαρακτηριστικών

Τα χαρακτηριστικά (features ή attributes) είναι εκείνα που κάνουν τις κλάσεις των διαφόρων αντικειμένων να διαφέρουν και έτσι να είναι διαχωρίσιμες. Αν τα χαρακτηριστικά δεν έχουν υψηλό πληροφοριακό περιεχόμενο τα αντικείμενα δεν θα διαφοροποιούνται επαρκώς. Το πλήθος των χαρακτηριστικών καθορίζει τη διάσταση του προβλήματος κατηγοριοποίησης και είναι σημαντικό να είναι μικρό γιατί η πολυπλοκότητα πολλών αλγόριθμων είναι εκθετική ως προς τη διάσταση. Για την επεξεργασία των χαρακτηριστικών είναι απαραίτητη η μετατροπή τους σε μορφή που μπορεί να αναγνωρίσει ο κάθε ταξινομητής, ανάλογα με τον τύπο του. Η ανάγκη αυτή προκύπτει γιατί κάθε ταξινομητής αναπαριστά τα χαρακτηριστικά με διαφορετικούς τρόπους (διανύσματα ή δυαδική μορφή). Για παράδειγμα, σε κάποιους ταξινομητές όπως ο Naive Bayes, τα χαρακτηριστικά που θα επιλεγούν πρέπει να είναι σε δυαδική μορφή, έτσι ώστε ο ταξινομητής να μπορέσει να καταλάβει αν μία έξοδος έχει το συγκεκριμένο χαρακτηριστικό ή όχι. Η διαδικασία εξαγωγής και επιλογής των χαρακτηριστικών είναι αποτέλεσμα αναλυτικής στατιστικής ανάλυσης και επιτήρησης των δεδομένων, και ονομάζεται εξαγωγή χαρακτηριστικών (feature extraction ή feature engineering).

Δοσμένου του συνόλου εκπαίδευσης, οι διαφορετικές λέξεις με αξιόλογο πληροφοριακό περιεχόμενο εντοπίζονται και τοποθετούνται σε μια λίστα λέξεων. Ακολουθώς, ο ταξινομητής ελέγχει ποιες λέξεις του συνόλου εκπαίδευσης περιέχονται στο κάθε tweet. Αυτό το χαρακτηριστικό ονομάζεται unigram χαρακτηριστικό.. Γενικότερα, κάποια χαρακτηριστικά που προκύπτουν είναι τα N-grams, τα οποία είναι είτε μονογράμματα είτε διγράμματα, είτε ο συνδυασμός τους. Καταλληλότερος ταξινομητής για τη χρήση των μονογραμμάτων είναι ο Support Vector Machine, σύμφωνα με έρευνα των Alec Go, Richa Bhayani και Lei Huang το 2009.

Συμπεραίνοντας φαίνεται πως υπάρχει ένα μοντέλο το TF-IDF (Term Frequency-Inverse Document Frequency) το οποίο επιλέγει εκείνα τα χαρακτηριστικά που θα περιγράψουν καλύτερα ένα κείμενο, δηλαδή παρουσιάζει πόσο σημαντική είναι μία λέξη λόγω της συχνότητας εμφάνισής της μέσα στο κείμενο αλλά κοιτώντας και το

σύνολο. Το TF-IDF αποτελείται από δύο μετρικές : την συχνότητα όρου (TF) και την αντίστροφη συχνότητα του εγγράφου (IDF). Η συχνότητα όρου TF δηλώνει την συχνότητα που εμφανίζεται ο όρος μέσα στο κείμενο. Η χρήση μόνο της ποσότητας TF θα είχε σαν καλύτερους όρους, λέξεις που δεν περιέχουν καμία πληροφορία. Άρα το IDF ισοσταθμίζει την κατάσταση, λειτουργώντας σαν βάρος σημαντικότητας που όρου ως προς το κείμενο. Τέλος, το TF-IDF βασίζεται περισσότερο στους όρους που εμφανίζονται συχνότερα στο κείμενο. Πέρα από το μοντέλο TF-IDF και το N-grams, υπάρχουν ωστόσο και κάποια άλλα διαδεδομένα μοντέλα εξαγωγής χαρακτηριστικών που χρησιμοποιούνται σε δεδομένα κειμένου. Αυτά είναι :

- Το μοντέλο Διανυσματικού Χώρου (Vector Space).
- Το μοντέλο Επεξεργασίας Φυσικής Γλώσσας (NLP).
- Το μοντέλο του Σάκου Λέξεων (Bag of Words-BoW).

Στο μοντέλο Διανυσματικού Χώρου, τα κείμενα παριστάνονται ως δυαδικά διανύσματα όπου η τιμή 1 δηλώνει την εμφάνιση της αντίστοιχης λέξης στο κείμενο ενώ η τιμή 0 την απουσία της. Το μειονέκτημά του είναι ότι οδηγεί σε πολύ αραιά διανύσματα.

Το μοντέλο BoW είναι το πιο διαδεδομένο όσον αφορά την εξαγωγή χαρακτηριστικών από κείμενα. Στο μοντέλο αυτό τα κείμενα μετατρέπονται σε διανύσματα έτσι ώστε να γίνει η μοντελοποίηση των δεδομένων που έχουν σκοπό να χρησιμοποιηθούν σε διάφορους αλγορίθμους μηχανικής μάθησης. Ένα κείμενο (πρόταση, έγγραφο) αναπαρίσταται ως σάκος (bag) των λέξεων της, αγνοώντας τη γραμματική και λεξικολογική πληροφορία. Το σημαντικό χαρακτηριστικό που διατηρείται είναι η πολλαπλότητα της κάθε λέξης. Με άλλα λόγια, η BoW αναπαράσταση ορίζεται ως η αναπαράσταση ενός κειμένου που περιγράφει την συχνότητα εμφάνισης των λέξεων μέσα σε αυτό. Ισοδύναμα, κάθε BoW-διάνυσμα (σάκος) έχει διάσταση ίση με το πλήθος των λέξεων και οι αντίστοιχες συντεταγμένες παριστάνουν τις συχνότητες εμφάνισής τους. Στην πράξη, το μοντέλο BoW χρησιμοποιείται κυρίως ως εργαλείο παραγωγής χαρακτηριστικών. Όταν γίνει η μετατροπή του κειμένου σε σάκο (διάνυσμα συχνοτήτων) λέξεων, τότε είναι δυνατός ο υπολογισμός διαφόρων διανυσματικών μετρικών για τον χαρακτηρισμό του κειμένου. Όπως αναφέρθηκε παραπάνω, συνηθισμένος τύπος χαρακτηριστικών που υπολογίζονται από το μοντέλο BoW εί-

ναι η μακροχρόνια συχνότητα, δηλαδή ο αριθμός των φορών που εμφανίζεται ένας όρος στο κείμενο.

Γ'.4 Αλγόριθμοι Μάθησης

Όπως αναφέρθηκε σε προηγούμενη παράγραφο, οι αλγόριθμοι εκπαίδευσης ταξινομητών είναι αλγόριθμοι μηχανικής μάθησης με επίβλεψη. Αυτό σημαίνει ότι είναι απαραίτητη η διαθεσιμότητα δεδομένων εκπαίδευσης στα οποία η σωστή κλάση να έχει εντοπισθεί και καταγραφεί.

Οι ταξινομητές είναι μοντέλα των οποίων οι παράμετροι έχουν προκύψει μετά από εκπαίδευση και τα οποία λαμβάνουν τις κατάλληλες αποφάσεις κατηγοριοποίησης με βάση τα χαρακτηριστικά (features).

Γ'.4.1 Απλοϊκή Κατηγοριοποίηση κατά Bayes (Naive Bayes)

Οι ταξινομητές αυτής της κατηγορίας ανήκουν στην οικογένεια των πιθανοτικών ταξινομητών που βασίζονται στην εφαρμογή του θεωρήματος Bayes με υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών. Στον απλοϊκό ταξινομητή Bayes γίνεται η υπόθεση ότι η παρουσία ενός συγκεκριμένου χαρακτηριστικού σε μία τάξη δεν σχετίζεται με την παρουσία οποιουδήποτε άλλου χαρακτηριστικού (ανεξαρτησία χαρακτηριστικών). Στην πλειοψηφία των περιπτώσεων, η παραπάνω υπόθεση ανεξαρτησίας δεν ισχύει. Παρόλα αυτά όμως, οι επιδόσεις του ταξινομητή αυτού είναι αξιοσημείωτα καλές στις περισσότερες περιπτώσεις. Γι' αυτό οι επιδόσεις του χρησιμοποιούνται ως τιμές αναφοράς για τους υπόλοιπους ταξινομητές. Η εκπαίδευση ενός απλοϊκού ταξινομητή Bayes για την περίπτωση δύο κλάσεων έγκειται στον υπολογισμό των διαφόρων δεσμευμένων πιθανοτήτων που απαιτούνται για την εφαρμογή του κανόνα του Bayes:

$$P(C_1 | x_i) = \frac{P(x_i | C_1)P(C_1)}{P(x_i)} = \frac{P(x_i | C_1)P(C_1)}{P(x_i | C_1)P(C_1) + P(x_i | C_2)P(C_2)}$$

όπου $(x_1, x_2, \dots, x_n)$ είναι το σύνολο δεδομένων εκπαίδευσης, C_1, C_2 οι κλάσεις και $P(C_1 | x_i), P(C_1), P(x_i | C_1)$ η εκ των υστέρων πιθανότητα, η εκ των προτέρων πιθανότητα και η υπό συνθήκη πιθανότητα να προκύψει η x_i από την C_1 , αντίστοιχα.

Μετά την εκπαίδευση, ο ταξινομητής μπορεί να εφαρμοστεί σε νέα δεδομένα υπολογίζοντας τις εκ των υστέρων πιθανότητες των κλάσεων. Επιλέγεται η κλάση που μεγιστοποιεί αυτήν την πιθανότητα.

Μια παραλλαγή του απλοϊκού ταξινομητή Bayes είναι ο Gaussian απλοϊκός ταξινομητής Bayes. Εφαρμόζεται όταν τα χαρακτηριστικά λαμβάνουν συνεχείς τιμές και ακολουθούν την κανονική κατανομή η οποία προσδιορίζεται μέσω της εκτίμησης των δύο παραμέτρων της, της μέσης τιμής και της διασποράς.

Γ.4.2 Μηχανές Διανυσμάτων Στήριξης

Οι Μηχανές Διανυσμάτων Στήριξης (Support Vector Machines-SVM) είναι ταξινομητές που χρησιμοποιούνται σε πολλές κατηγορίες δεδομένων και μεταξύ αυτών και στην κατηγοριοποίηση κειμένων. Μπορούν να χρησιμοποιηθούν τόσο σε γραμμικά όσο και σε μη-γραμμικά δεδομένα (μέσω κατάλληλων kernels) και μέχρι πρόσφατα, πριν την εμφάνιση των μοντέλων βαθιάς μάθησης (deep learning), ήταν από τις δημοφιλέστερες τεχνικές ταξινόμησης. Υποθέτοντας ότι τα πολυδιάστατα δεδομένα εκπαίδευσης προέρχονται από δύο κλάσεις, ο στόχος είναι ο διαχωρισμός τους μέσω της εύρεσης του βέλτιστου διαχωριστικού υπερεπιπέδου το οποίο εκφράζεται μέσω των διανυσμάτων υποστήριξης για τις δοσμένες κλάσεις [20]. Το πρόβλημα αυτό ανάγεται με κατάλληλη διατύπωση σε πρόβλημα τετραγωνικού προγραμματισμού (quadratic programming).

Γ.4.2.1 Ακολουθιακή Βελτιστοποίηση Ελαχίστου

Ο αλγόριθμος της Ακολουθιακής Βελτιστοποίησης Ελαχίστου (Sequential Minimal Optimization-SMO) επιλύει του πρόβλημα τετραγωνικού προγραμματισμού που προκύπτει κατά την εκπαίδευση των μηχανών διανυσμάτων στήριξης [21]. Ο αλγόριθμος SMO είναι απλός, εύκολος στην εφαρμογή, γρήγορος και έχει καλές ιδιότητες κλιμάκωσης για δύσκολα προβλήματα από ότι ο τυπικός αλγόριθμος εκπαίδευσης SVM. Ο SMO διασπά το πρόβλημα του τετραγωνικού προγραμματισμού σε μία σειρά από μικρότερα υποπροβλήματα. Αυτά επιλύονται αναλυτικά, γεγονός που οδηγεί στην αποφυγή της χρήσης μιας χρονοβόρας αριθμητικής βελτιστοποίησης τετραγωνικού προγραμματισμού. Η μνήμη που απαιτείται είναι γραμμική οπότε επιτρέπει

στον αλγόριθμο να διαχειριστεί πολύ μεγάλα σύνολα εκπαίδευσης. Επειδή ο άμεσος υπολογισμός των εμπλεκόμενων πινάκων αποφεύγεται, ο SMO βρίσκεται κάπου μεταξύ γραμμικής και τετραγωνικής διάστασης στο μέγεθος του συνόλου εκπαίδευσης για διάφορα προβλήματα δοκιμής, ενώ ο τυπικός SVM αλγόριθμος βρίσκεται κάπου μεταξύ γραμμικής και κυβικής διάστασης. Πειραματικές δοκιμές έχουν δείξει ότι ο SMO είναι ταχύτερος για γραμμικούς SVM και αραιά σύνολα δεδομένων.

Γ'.4.3 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic Regression) είναι ένα στατιστικό μοντέλο που βασίζεται στην υπόθεση στο σύνολο δεδομένων εκπαίδευσης υπάρχουν ανεξάρτητες μεταβλητές που καθορίζουν το αποτέλεσμα της κατάταξης. Στο μοντέλο αυτό, η εξαρτώμενη μεταβλητή (κλάση) είναι κατηγορική (nominal). Αυτό σημαίνει πως η μεταβλητή μπορεί να πάρει διακριτές μόνο τιμές από ένα συγκεκριμένο σύνολο πιθανών τιμών. Στόχος του αλγόριθμου είναι η εύρεση του καταλληλότερου μοντέλου που θα περιγράψει τη σχέση του διχοτόμου χαρακτηριστικού του συνόλου δεδομένων με ένα σύνολο ανεξάρτητων μεταβλητών.

Γ'.4.4 Κ Πλησιέστεροι Γείτονες

Οι ταξινομητές αυτής της κατηγορίας βασίζονται στο ότι είναι δυνατός ο ορισμός απόστασης μεταξύ των εγγραφών των δεδομένων. Η κύρια ιδέα στην οποία στηρίζονται είναι ότι εγγραφές που είναι κοντά με την έννοια ότι η μεταξύ τους απόσταση είναι μικρή, είναι πολύ πιθανό να ανήκουν στην ίδια κλάση. Ο πιο απλός αλγόριθμος αυτού του τύπου, υπολογίζει τα κέντρα των κλάσεων με βάση το σύνολο δεδομένων εκπαίδευσης και στη συνέχεια οι εγγραφές ταξινομούνται στην κλάση της οποίας το κέντρο είναι πιο κοντά σε αυτές.

Στην περίπτωση της ταξινόμησης με βάση του k πλησιέστερους γείτονες (k Nearest Neighbors, k -NN), οι εγγραφές των δεδομένων εκπαίδευσης χρησιμοποιούνται όπως είναι για την κατάταξη των νέων εγγραφών, χωρίς τον υπολογισμό των κέντρων των κλάσεων. Συγκεκριμένα, όταν έρχεται μια νέα εγγραφή για κατάταξη υπολογίζεται η απόστασή της από τις υπόλοιπες εγγραφές. Στη συνέχεια, επιλέγονται οι k πλησιέστερες προς την νέα εγγραφή και μετά υπολογίζεται η κλάση με τα περισσό-

τερα μέλη μεταξύ αυτών των k εγγραφών. Η νέα εγγραφή ταξινομείται στην κλάση αυτή. Όπως γίνεται αντιληπτό, οι ταξινομητές αυτής της κατηγορίας δεν απαιτούν εκπαίδευση αφού τα ίδια τα δεδομένα εκπαίδευσης χρησιμοποιούνται για τον υπολογισμό των πλησιέστερων γειτόνων. Η πιο σημαντική παράμετρος του k -NN ταξινομητή είναι το k δηλαδή το πλήθος των πλησιέστερων γειτόνων που πρέπει να εξεταστούν για να ληφθεί η απόφαση κατάταξης. Οι συνηθισμένες τιμές είναι $k=3$ και $k=5$.

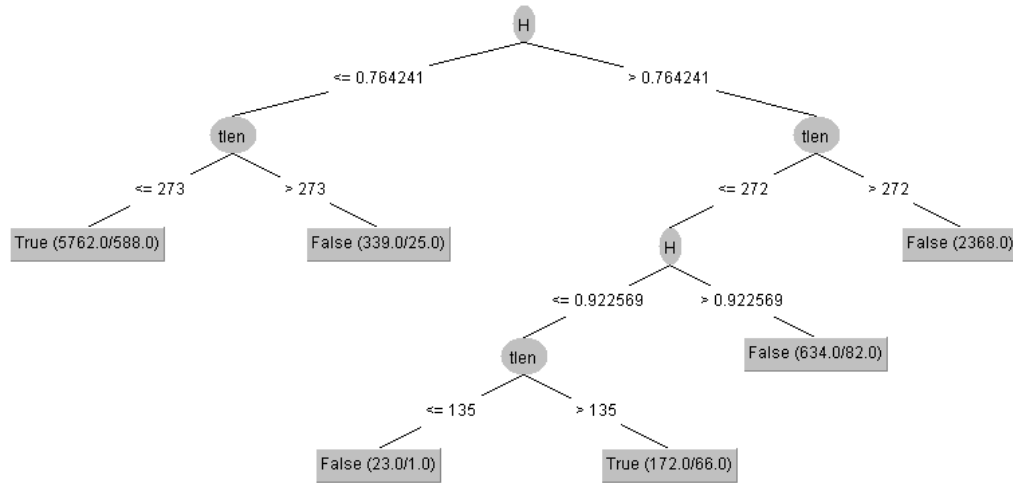
Γ'.4.5 Δένδρα Απόφασης

Μια σημαντική κατηγορία ταξινομητών είναι τα δένδρα απόφασης (decision trees). Στην τεχνική αυτή χρησιμοποιείται η μέθοδος 'διαίρει και βασίλευε' για να διασπαστεί ο χώρος αναζήτησης του προβλήματος σε υποσύνολα. Ένα δένδρο απόφασης είναι ένα δένδρο όπου η ρίζα και οι εσωτερικοί κόμβοι αντιπροσωπεύουν μια ερώτηση σχετική με κάποια χαρακτηριστικό των δεδομένων. Οι ακμές που ξεκινούν από κάθε κόμβο αντιπροσωπεύουν τις πιθανές απαντήσεις στην ερώτηση του κόμβου. Κάθε φύλλο παριστάνει μια πρόβλεψη (εκτίμηση) της λύσης που ουσιαστικά σημαίνει κατάταξη σε κάποια από τις κλάσεις.

Ένα μοντέλο δένδρου απόφασης αποτελείται από τρία μέρη: i) το δένδρο απόφασης όπως ορίστηκε προηγουμένως, ii) έναν αλγόριθμο μάθησης ο οποίος δημιουργεί το δένδρο με βάση τα δεδομένα εκπαίδευσης και iii) έναν αλγόριθμο εφαρμογής του δένδρου απόφασης στα δεδομένα ώστε να επιλύεται το πρόβλημα κατηγοριοποίησης.

Ενδεικτικό παράδειγμα ενός δένδρου απόφασης φαίνεται στο Σχήμα Γ'.1 όπου επελέγησαν δύο μόνο χαρακτηριστικά (εντροπία και μήκος κειμένου) από τις εγγραφές της βάσης των fake news.

Η εφαρμογή του δένδρου απόφασης σε δεδομένα εισόδου έγκειται στην εύρεση μονοπατιού από τη ρίζα του δένδρου μέχρι κάποιο φύλλο όπου σε κάθε κόμβο επιλέγεται εκείνο το παιδί που αντιπροσωπεύει θετική απάντηση στην ερώτηση που παριστάνει ο κόμβος. Η διαδικασία αυτή δεν είναι υπολογιστικά απαιτητική αφού το ύψος των δένδρων απόφασης είναι κατά μέσο όρο ανάλογο του λογάριθμου του πλήθους των κόμβων. Ως παράδειγμα, αναφέρουμε ότι για την ταξινόμηση ενός tweet με εντροπία H 0.77 και μήκος len 140, η εφαρμογή του δένδρου απόφασης της σχή-



Σχήμα Γ'.1: Παράδειγμα δένδρου απόφασης ψευδών ειδήσεων με χρήση δύο μόνο χαρακτηριστικών (εντροπίας H και μήκους κειμένου $tlen$).

ματος Γ'.1 οδηγεί στην αποδοχή του tweet ως αληθούς με βάση το μονοπάτι (ρίζα → δεξί παιδί → αριστερό παιδί → αριστερό παιδί → δεξί παιδί (φύλλο)).

Γ.4.5.1 Αλγόριθμοι J48 - ID3

Ο αλγόριθμος ID3 προσπαθεί να κατασκευάσει ένα δένδρο απόφασης που ελαχιστοποιεί τον αναμενόμενο αριθμό συγκρίσεων επιτυγχάνοντας έτσι μικρά μήκη μονοπατιών για την ταξινόμηση των δεδομένων. Για τον σκοπό αυτό χρησιμοποιεί την εντροπία (entropy) με βάση την οποία το αρχικό σύνολο δεδομένων εκπαίδευσης διασπάται επαναληπτικά. Συγκεκριμένα, για κάθε υποψήφια διάσπαση του συνόλου δεδομένων, ο αλγόριθμος ID3 υπολογίζει το πληροφοριακό κέρδος της με βάση τον τύπο

$$Gain(D, S) = H(D) - \sum_{i=1}^s P(D_i)H(D_i) ,$$

όπου D είναι το σύνολο δεδομένων, S το σύνολο καταστάσεων που αντιπροσωπεύει τις διασπάσεις, $S = \{D_1, D_2, \dots, D_s\}$ και $H(D)$ η εντροπία του συνόλου D .

Ο αλγόριθμος J48 (C4.5) αποτελεί βελτιωμένη επέκταση του αλγορίθμου ID3. Τα επιπλέον χαρακτηριστικά του είναι ο χειρισμός των ελλειπουσών τιμών (missing

values), το κλάδεμα (prunning) των παραγόμενων δέντρων απόφασης, το εύρος των χαρακτηριστικών συνεχών τιμών και η εξαγωγή κανόνων. Στο περιβάλλον του Weka, ο J48 είναι μία υλοποίηση σε Java του αλγορίθμου C4.5.

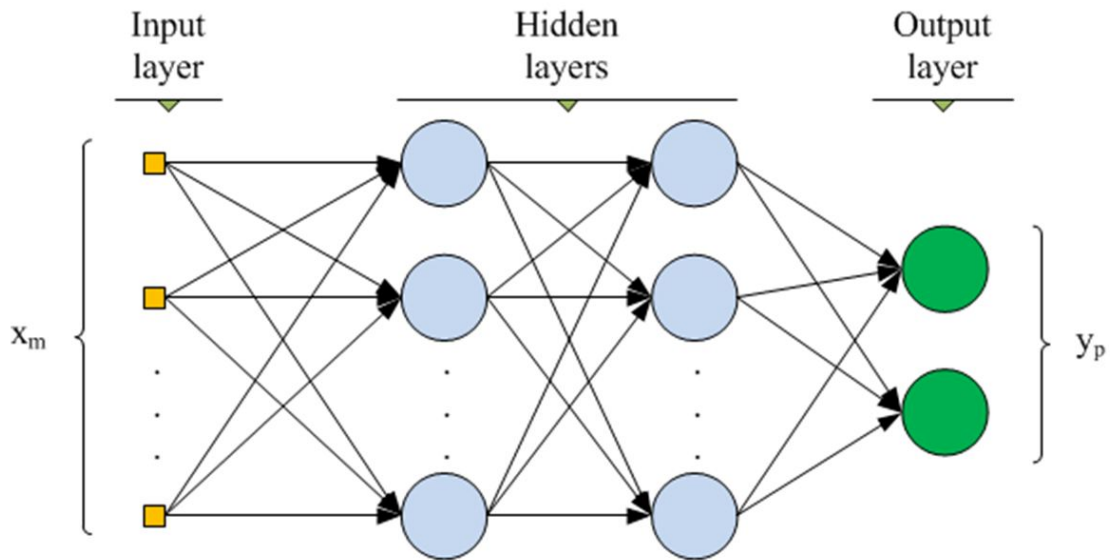
Γ'.4.6 Κανόνες απόφασης και αλγόριθμος One Rule

Η ιδέα αυτής της μεθόδου κατηγοριοποίησης βρίσκεται στη δημιουργία κανόνων τύπου if-then για την επιλογή της τάξης των δεδομένων εισόδου. Συγκεκριμένα, ένας κανόνας απόφασης ή κατηγοριοποίησης (classification rule), $r = a, c$ αποτελείται από το τμήμα if (πρότερο τμήμα) a και το τμήμα then, c . Αν το κατηγορήμα a αξιολογηθεί ως αληθές τότε επιλέγεται η κατηγορία (κλάση) c . Παρόλο που οι κανόνες απόφασης μπορούν εύκολα να προκύψουν από δένδρα απόφασης χρησιμοποιώντας τα μονοπάτια από τη ρίζα στα φύλλα υπάρχουν τεχνικές που εστιάζουν εξ' αρχής στη δημιουργία κανόνων απόφασης και ονομάζονται αλγόριθμοι κάλυψης (covering). Βασίζονται στην εύρεση του βέλτιστου κανόνα απόφασης με κριτήριο τη μεγιστοποίηση της πιθανότητας σωστής ταξινόμησης.

Χαρακτηριστικός εκπρόσωπος αυτής της κατηγορίας αλγόριθμων είναι ο **αλγόριθμος One Rule (OneR ή 1R)**. Ο αλγόριθμος αυτός ψάχνει ένα συγκεκριμένο χαρακτηριστικό του συνόλου δεδομένων, το οποίο είναι ικανό να ταξινομήσει τα στοιχεία του με τη μέγιστη δυνατή ακρίβεια. Είναι ένας απλός αλλά και σχετικά ακριβής αλγόριθμος ταξινόμησης, ο οποίος δημιουργεί ένα κανόνα για κάθε χαρακτηριστικό στα δεδομένα και έπειτα επιλέγει τον κανόνα που φέρει το μικρότερο συνολικό σφάλμα ως ένα κανόνα απόφασης. Η επιλογή των χαρακτηριστικών γίνεται με την βοήθεια ενός πίνακα συχνότητας για κάθε πρόβλεψη έναντι του στόχου. Είναι αποδεδειγμένο ότι ο αλγόριθμος OneR παράγει κανόνες μόνο ελαφρώς λιγότερο ακριβείς από τους πλέον σύγχρονους αλγορίθμους ταξινόμησης. Ακόμη, το συνολικό σφάλμα που υπολογίζεται από τους πίνακες συχνότητας είναι το μέτρο κάθε συμβολής χαρακτηριστικού. Ένα μικρό συνολικό σφάλμα σημαίνει μεγαλύτερη συμβολή στην προβλεψιμότητα του μοντέλου.

Γ.4.7 Νευρωνικά Δίκτυα Perceptron Πολλών Στρωμάτων

Στα νευρωνικά δίκτυα Perceptron πολλών στρωμάτων (Multilayer Perceptron-MLP), οι νευρώνες οργανώνονται σε επίπεδα ή στρώματα (layers) και η τοπολογία του δικτύου είναι αυτή που φαίνεται στην εικόνα Γ'.2.



Σχήμα Γ'.2: Τοπολογία δικτύου Perceptron πολλών στρωμάτων. Υπάρχει το στρώμα εισόδου (input layer), το στρώμα εξόδου (output layer) και τα ενδιάμεσα (κρυφά-hidden) στρώματα.

Ένα δίκτυο MLP έχει $L + 1$ στρώματα με το στρώμα 0 να είναι το στρώμα εισόδου, το στρώμα L να είναι το στρώμα εξόδου και τα στρώματα $1, 2, \dots, L - 1$ να είναι τα κρυφά (hidden) στρώματα. Οι νευρώνες του στρώματος l λαμβάνουν είσοδο από τους νευρώνες του στρώματος $l - 1$ και τροφοδοτούν τους νευρώνες του στρώματος $l + 1$ με $l = 1, 2, \dots, L - 1$. Στα δίκτυα MLP, προτιμάται η σιγμοειδής συνάρτηση ενεργοποίησης για την έξοδο των νευρώνων

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

όπου u το εσωτερικό γινόμενο των συναπτικών βαρών $w_1, w_2, \dots, w_n$ του νευρώνα με τις αντίστοιχες εισόδους $x_1, x_2, \dots, x_n$ κατάλληλα διορθωμένο με το κατώφλι θ :

$$u = w_1x_1 + w_2x_2 + \dots + w_nx_n - \theta.$$

Το πολύ σημαντικό θεωρητικό αποτέλεσμα για τα δίκτυα MLP είναι ότι μπορούν να προσεγγίσουν οποιαδήποτε ομαλή συνάρτηση με όση ακρίβεια είναι επιθυμητή και γι' αυτό ονομάζονται καθολικοί προσεγγιστές (universal approximators).

Η εκπαίδευση ενός MLP γίνεται με τη βοήθεια δεδομένων εκπαίδευσης και ουσιαστικά πρόκειται για την εύρεση κατάλληλων τιμών για να συναπτικά βάρη των νευρώνων του δικτύου ώστε να ελαχιστοποιείται κάποια συνάρτηση σφάλματος. Ο πλέον διαδεδομένος αλγόριθμος εκπαίδευσης είναι ο αλγόριθμος της οπισθοδιάδοσης (backpropagation) [22]. Το περιβάλλον Weka προσφέρει υλοποίηση του backpropagation για την εκπαίδευση του MLP. Μια βασική παράμετρος του MLP είναι το πλήθος των κρυφών στρωμάτων. Ένας πρακτικός κανόνας που πειραματικά δίνει καλά αποτελέσματα είναι η επιλογή πλήθους κρυφών στρωμάτων ίσο με το μισό του πλήθους των χαρακτηριστικών συν το πλήθος των κλάσεων.

Γ'.4.8 Συνδυαστικοί Αλγόριθμοι Μάθησης

Στην κατηγορία αυτή έχουμε αλγόριθμους μάθησης που συνδυάζουν δύο ή περισσότερους αλγόριθμους μάθησης που εκπαιδεύονται και εκτελούνται ανεξάρτητα και στη συνέχεια, οι προβλέψεις τους για τα νέα δεδομένα συνδυάζονται με κατάλληλη μέθοδο. Οι συνδυαστικοί αλγόριθμοι μάθησης που χρησιμοποιήθηκαν στην εργασία είναι δύο: i) ο αλγόριθμος Bagging, και ii) ο αλγόριθμος Adaboost

Γ'.4.8.1 Ο Αλγόριθμος Bagging

Είναι μια τεχνική η οποία αναδειγματοληπτεί επαναληπτικά ένα αρχικό σύνολο δεδομένων εκπαίδευσης δημιουργώντας έτσι πολλά διαφορετικά σύνολα εκπαίδευσης τα οποία στη συνέχεια χρησιμοποιούνται για την εκπαίδευση των επιμέρους ταξινομητών [23]. Πιο συγκεκριμένα, στο πρώτο στάδιο (bootstrap) δημιουργούνται τα επιμέρους σύνολα δεδομένων εκπαίδευσης εφαρμόζοντας αναδειγματοληψία με βάση την ομοιόμορφη κατανομή με τα οποία γίνεται η εκπαίδευση των ταξινομητών. Συνηθισμένοι επιμέρους ταξινομητές είναι τα νευρωνικά δίκτυα, οι μηχανές διανυσμάτων στήριξης και τα δένδρα αποφάσεων. Στο δεύτερο στάδιο (aggreagating) συνδυάζονται τα αποτελέσματα των επιμέρους ταξινομητών για την παραγωγή του τελικού αποτελέσματος. Πειραματικά αποτελέσματα έδειξαν ότι ο αλγόριθμος Bagging

παράγει ταξινομητές που δεν είναι επιρρεπείς σε υπερπροσαρμογή και επίσης, βελτιώνει τις επιδόσεις ασταθών επιμέρους ταξινομητών.

Γ'.4.8.2 Ο Αλγόριθμος Adaboost

Στους αλγόριθμους Boosting (ενίσχυσης) πραγματοποιείται η εκπαίδευση μιας σειράς ταξινομητών όπου το σύνολο εκπαίδευσης για κάθε ταξινομητή προκύπτει από το αρχικό σύνολο εκπαίδευσης με βάση την επίδοση του προηγούμενου ταξινομητή. Ειδικότερα, στο αλγόριθμο προσαρμοστικής ενίσχυσης AdaBoost, η κύρια ιδέα είναι η προσαρμογή της κατανομής των δεδομένων εκπαίδευσης μέσω μιας επαναληπτικής διαδικασίας κατά την οποία η κατανομή των δεδομένων εκπαίδευσης τροποποιείται έτσι ώστε να δίνεται μεγαλύτερη βαρύτητα στα δεδομένα που είναι δυσκολότερο να ταξινομηθούν σωστά. Στο τέλος κάθε επανάληψης υπολογίζεται το εμπειρικό σφάλμα ώστε να τροποποιηθεί η κατανομή των δεδομένων εκπαίδευσης ανάλογα. Δεδομένα με μεγάλο βάρος είναι πιο πιθανό να επιλεγούν [24].

Γ'.4.9 Αξιολόγηση Μοντέλων Ταξινόμησης

Η μέτρηση των επιδόσεων των διαφόρων ταξινομητών είναι απαραίτητη προϋπόθεση για την σύγκριση και αξιολόγησή τους. Το πιο απλό μέτρο είναι το ποσοστό επιτυχίας (success rate) το οποίο είναι ένας απλός αριθμός και δίνει το ποσοστό των σωστών ταξινομήσεων του ταξινομητή για ένα δοσμένο σύνολο δεδομένων ελέγχου. Το απλό αυτό κριτήριο αξιολόγησης μπορεί να εφαρμοστεί τόσο σε δυαδικά προβλήματα όσο και προβλήματα πολλαπλών κλάσεων. Απλά απαριθμούνται οι λανθασμένες ταξινομήσεις.

Πιο προχωρημένα μέτρα αξιολόγησης, βασίζονται στις παρακάτω κατηγορίες ταξινομήσεων:

- *Αληθώς Θετικό (True Positive - TP).*
Η εγγραφή ορθά ταξινομείται ως θετική.
- *Ψευδώς Θετικό (False Positive - FP).*
Η εγγραφή εσφαλμένα ταξινομείται ως θετική.

- *Αληθώς Αρνητικό (True Negative - TN).*
Η εγγραφή ορθά ταξινομείται ως αρνητική.
- *Ψευδώς Αρνητικό (False Negative - FN).*
Η εγγραφή εσφαλμένα ταξινομείται ως αρνητική.

Με βάση τα παραπάνω ορίζεται η **ορθότητα (accuracy)** ως το ποσοστό των σωστών προβλέψεων/ταξινομήσεων του μοντέλου

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} .$$

όπου στον αριθμητή υπάρχουν όλες οι σωστές προβλέψεις ενώ στον παρονομαστή βρίσκεται το άθροισμα και των σωστών και των λανθασμένων προβλέψεων.

Η **ευστοχία ή ακρίβεια (precision)** ορίζεται ως το ποσοστό των σωστών θετικών προβλέψεων/ταξινομήσεων του μοντέλου

$$precision = \frac{TP}{TP + FP} .$$

Ανάλογα έχουμε την **ανάκληση (recall)** ως το ποσοστό των θετικών προβλέψεων

$$recall = \frac{TP}{TP + FN} .$$

Τα κριτήρια precision και recall συνδυάζονται στο κριτήριο F_1 που ορίζεται ως

$$F_1 = 2 \frac{precision \cdot recall}{precision + recall} .$$

Ισχύει $0 \leq F_1 \leq 1$ με την ιδανική τιμή να είναι το 1 και χειρότερη το 0.

Κεφάλαιο Δ΄

Πείραμα

Δ΄.1 Twitter

Το Twitter είναι ένα μέσο κοινωνικής δικτύωσης που επιτρέπει στους χρήστες να ανταλλάσσουν μηνύματα μέχρι 280 χαρακτήρες, τα οποία ονομάζονται tweets (Εικόνα Δ΄.3). Αρχικά το όριο ήταν το 140 χαρακτήρες, αλλά αυτός ο αριθμός διπλασιάστηκε το 2017. Η πλατφόρμα του Twitter δημιουργήθηκε το Μάρτιο του 2006 από τον Jack Dorsey και πολύ γρήγορα απέκτησε φήμη και έγινε ευρέως διαδεδομένο προσελκύοντας μεγάλο πλήθος χρηστών. Από την εμφάνισή του μέχρι σήμερα, αναπτύσσεται συνεχώς και οι ενεργοί χρήστες του ανέρχονται σε δεκάδες εκατομμύρια. Το 2012 περισσότεροι από 100 εκατομμύρια χρήστες δημοσίευσαν 340 εκατομμύρια tweet την ημέρα. Το 2022 οι χρήστες εκτιμώνται σε πάνω από 200 εκατομμύρια.

Όλοι οι χρήστες, εγγεγραμμένοι ή μη, που βρίσκονται σε αυτό το κοινωνικό μέσο έχουν τη δυνατότητα να διαβάσουν όσα σχόλια δημοσιεύονται, αλλά μόνο οι εγγεγραμμένοι μπορούν να εκφράζουν τη γνώμη τους. Στο Twitter πέρα από τους μεμονωμένους απλούς χρήστες και τις μικρές κοινωνικές ομάδες, υπάρχουν και εταιρικά προφίλ, διάσημες προσωπικότητες από το χώρο του θεάματος, των επιχειρήσεων και της πολιτικής. Επομένως είναι σαφές πως τα θέματα που διαμορφώνονται καλύπτουν πάρα πολλά θέματα κάθε είδους και έτσι ικανοποιούν όλο το κοινό. Είναι δύσκολο να μην βρει ο οποιοσδήποτε κάποιο θέμα που τον ενδιαφέρει.

Η δημοτικότητα των κοινωνικών μέσων δικτύωσης και η ευκολία διαθεσιμότητας και διαχείρισης των δεδομένων σημαίνει ότι όλο και περισσότερο αυτά τα εργαλεία

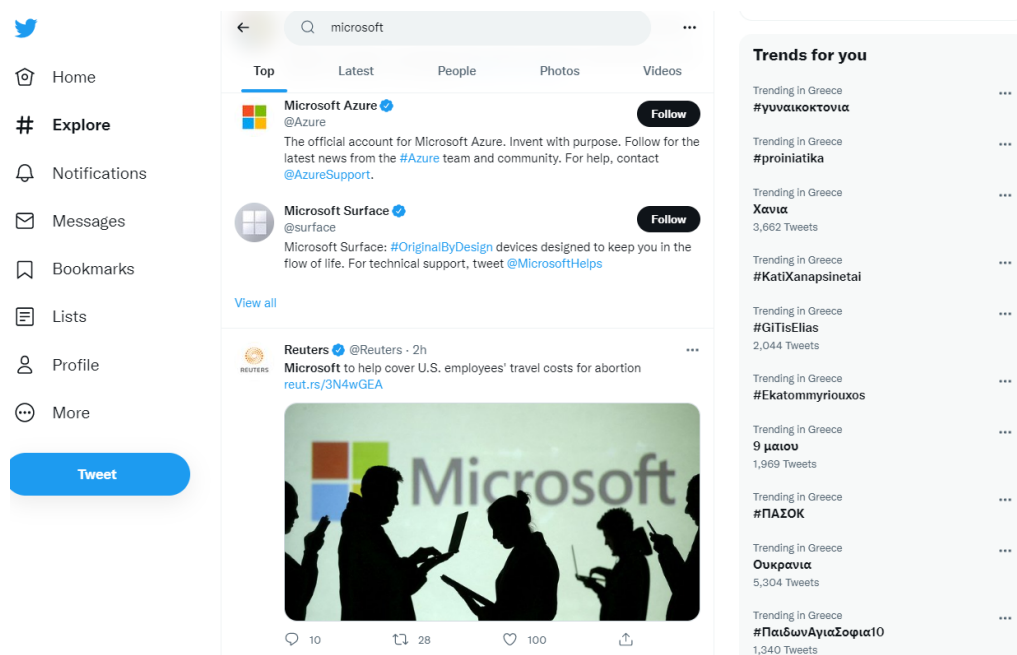


Σχήμα Δ'.3: Ο ιστότοπος της πλατφόρμας Twitter.

αποτελούν κύριες πηγές εξόρυξης δεδομένων [25]. Ωστόσο επισημαίνεται πως οι λόγοι που επιλέγεται το Twitter είναι γιατί:

1. Είναι δημοφιλές όσον αφορά την προσοχή των μέσων ενημέρωσης που λαμβάνει και έτσι προσελκύει περισσότερη έρευνα εξαιτίας της πολιτιστικής του κατάστασης.
2. Καθιστά εύκολη την εύρεση και παρακολούθηση συζητήσεων και από την αναζήτηση που υπάρχει μέσα στην εφαρμογή αλλά και από τα αποτελέσματα που προκύπτουν αν αναζητηθεί σε κάποια μηχανή αναζήτησης.
3. Τα hashtags διευκολύνουν τη συλλογή, την ταξινόμηση και την επέκταση αναζητήσεων στη συλλογή δεδομένων.
4. Τα δεδομένα ανακτώνται πολύ εύκολα και γρήγορα εξαιτίας της φυσικής ομαδοποίησης με τη βοήθεια των hashtags.
5. Το Twitter API είναι πιο προσβάσιμο σε σχέση με άλλα κοινωνικά μέσα άρα οι προγραμματιστές μπορούν να το χρησιμοποιήσουν.

Τέλος, τα βασικά χαρακτηριστικά που υπάρχουν στη δομή των tweets είναι:



Σχήμα Δ'.4: Στιγμιότυπο εμφάνισης tweets.

1. Μικρό μέγεθος με περιεκτικό περιεχόμενο (270 χαρακτήρες το πολύ).
2. Συντακτικά και γραμματικά απλό περιεχόμενο (λόγω μεγέθους).
3. Συχνή χρήση συντομεύσεων.
4. Hashtags.
5. URLs.
6. Retweet.
7. Usernames.

Τα hashtags δηλώνονται με το # στην αρχή της φράσης και δεν υποστηρίζονται τα κενά ανάμεσα στις λέξεις. Χρησιμοποιούνται για να δηλωθεί ένα θέμα ή μία κατάσταση με σκοπό να διευκολυνθεί η αναζήτηση όμοιων tweets. Για παράδειγμα το hashtag #savetheplanet εμφανίζει αποτελέσματα σχετικά με δράσεις και ενέργειες για την αποτροπή της οικολογικής καταστροφής του πλανήτη. Δηλώνονται μόνο με πεζά γράμματα ενώ η χρήση κάποιων κεφαλαίων συμβολίζει τον διαχωρισμό των

λέξεων και έτσι τα καθιστά πιο ευανάγνωστα. Τα retweets είναι κοινοποιήσεις ήδη υπαρχόντων tweet. Τα usernames αρχικά δηλώνονται με το @ και χρησιμοποιούνται για να επισημάνουν κάποιον εγγεγραμμένο στην εφαρμογή χρήστη.

Δ'.2 Προγραμματιστικό Περιβάλλον

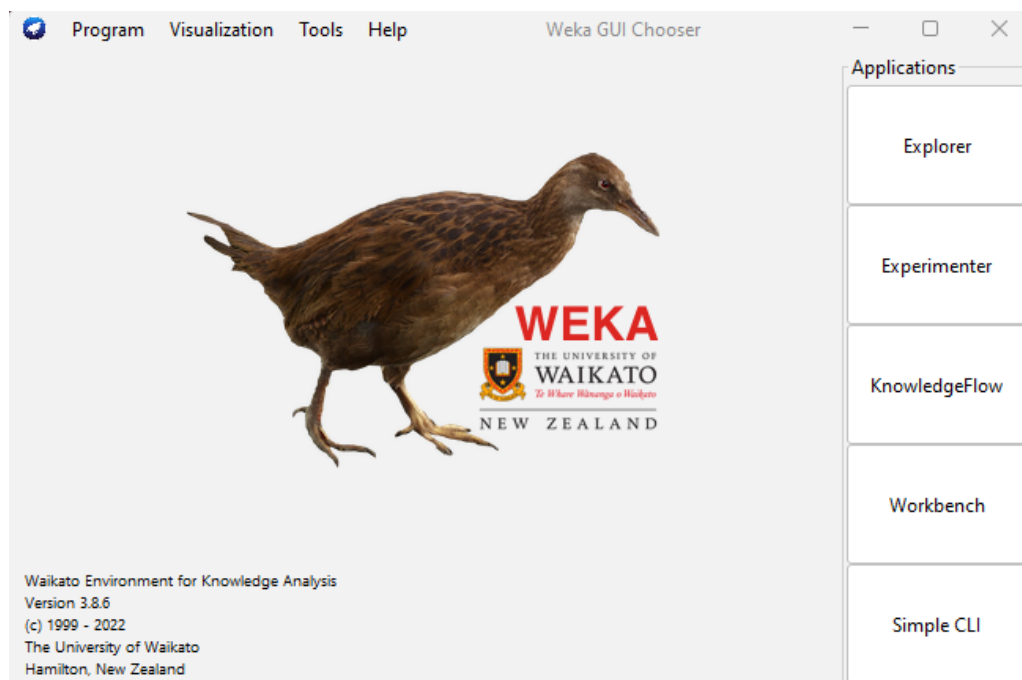
Το Weka (Waikato Environment for Knowledge Analysis) είναι μία πλατφόρμα μηχανικής μάθησης με σκοπό την παροχή ενός φιλικού και πλήρους περιβάλλοντος για την εφαρμογή τεχνικών μηχανικής μάθησης σε ποικιλία πραγματικών προβλημάτων [26]. Αναπτύχθηκε από το πανεπιστήμιο του Waikato στη Νέα Ζηλανδία και ανήκει στην κατηγορία ελεύθερου λογισμικού (freeware) αλλά και έχει ως στόχο την εξόρυξη γνώσης από δεδομένα. Διατίθεται δημοσίως σε όλο τον κόσμο επιτρέποντας στους χρήστες να κάνουν τροποποιήσεις στο λογισμικό όσο το χρησιμοποιούν. Σε αντίθεση με άλλα προγράμματα μηχανικής μάθησης, η έμφαση δόθηκε στην σχεδίαση ενός γραφικού περιβάλλοντος στο οποίο υπάρχει πλήθος διαδραστικών εργαλείων για τον χειρισμό των δεδομένων, την απεικόνιση αποτελεσμάτων καθώς και την σύγκριση των συνόλων κανόνων, ώστε να συμπληρωθούν τα βασικά εργαλεία της μηχανικής μάθησης.

Το Weka υποστηρίζει διάφορες εργασίες της εξόρυξης δεδομένων όπως προεπεξεργασία δεδομένων, κατηγοριοποίηση, συσταδοποίηση και ανακάλυψη συσχετιστικών κανόνων. Πιο συγκεκριμένα, οι λειτουργίες που υλοποιούνται στο Weka είναι:

- Προ-επεξεργασία των δεδομένων με την χρήση φίλτρων (filters).
- Δημιουργία μοντέλων μέσω συνόλων δεδομένων εκπαίδευσης.
- Αξιολόγηση αλγορίθμων με πληθώρα υλοποιημένων μετρικών.
- Απεικόνιση των αποτελεσμάτων που προέκυψαν με την ολοκλήρωση της διαδικασίας εκπαίδευσης τόσο με κείμενο όσο και με γραφικά.

Είναι διαθέσιμο δωρεάν στο διαδίκτυο στην ιστοσελίδα

<https://sourceforge.net/projects/weka/>



Σχήμα Δ'.5: Αρχική οθόνη εμφάνισης WEKA.

και είναι γραμμένο στη γλώσσα προγραμματισμού Java. Η μηχανική μάθηση είναι μία τεχνολογία με ευρύ φάσμα εφαρμογών. Έχει την προοπτική να γίνει ένα από τα βασικά στοιχεία του έξυπνου πληροφοριακού συστήματος που επιτρέπει την συμπίκνωση των γενικευμένων δεδομένων που προέρχονται από μεγάλες βάσεις δεδομένων από καταγεγραμμένες πληροφορίες, οι οποίες εφαρμόζονται σαν γνώση σε διάφορους πρακτικούς τρόπους όπως η ενσωμάτωση σε αυτόματες διαδικασίες όπως τα ειδικά συστήματα, ή χρησιμοποιούνται απευθείας για επικοινωνία με ειδικούς ή για εκπαιδευτικούς σκοπούς. Σήμερα ωστόσο, το πεδίο δεν είναι αρκετά ώριμο ώστε να το κάνει αυτό. Οι περισσότερες ερευνητικές προσπάθειες εστιάζουν στην εύρεση νέων αλγορίθμων για εκμάθηση, παρά για απόκτηση εμπειριών που εφαρμόζονται σε υπάρχουσες τεχνικές σε πραγματικά προβλήματα. Το πρόγραμμα Weka αποκαθιστά την ισορροπία εφαρμόζοντας τις σίγουρες τεχνικές της μηχανικής μάθησης σε μία ποικιλία προβλημάτων.

Η φιλοσοφία πίσω από το Weka είναι να απομακρυνθεί από την υποστήριξη της επιστήμης των υπολογιστών και ταυτόχρονα από την υποστήριξη του τελικού χρήστη

της μηχανικής μάθησης. Ο τελικός χρήστης είναι κάποιος που καταλαβαίνει τα δεδομένα και γνωρίζει τις δυνατότητες της μηχανικής μάθησης για να επιλέξει και να ερευνήσει την εφαρμογή των διάφορων τεχνικών. Για να διατηρηθεί αυτή η φιλοσοφία επικεντρωθήκαμε στην εξασφάλιση ότι οι λεπτομέρειες της υλοποίησης των αλγορίθμων της μηχανικής μάθησης και οι μορφές εισόδου που απαιτούνται είναι κρυμμένες από τον χρήστη.

Δ'.2.1 Αρχεία στο WEKA

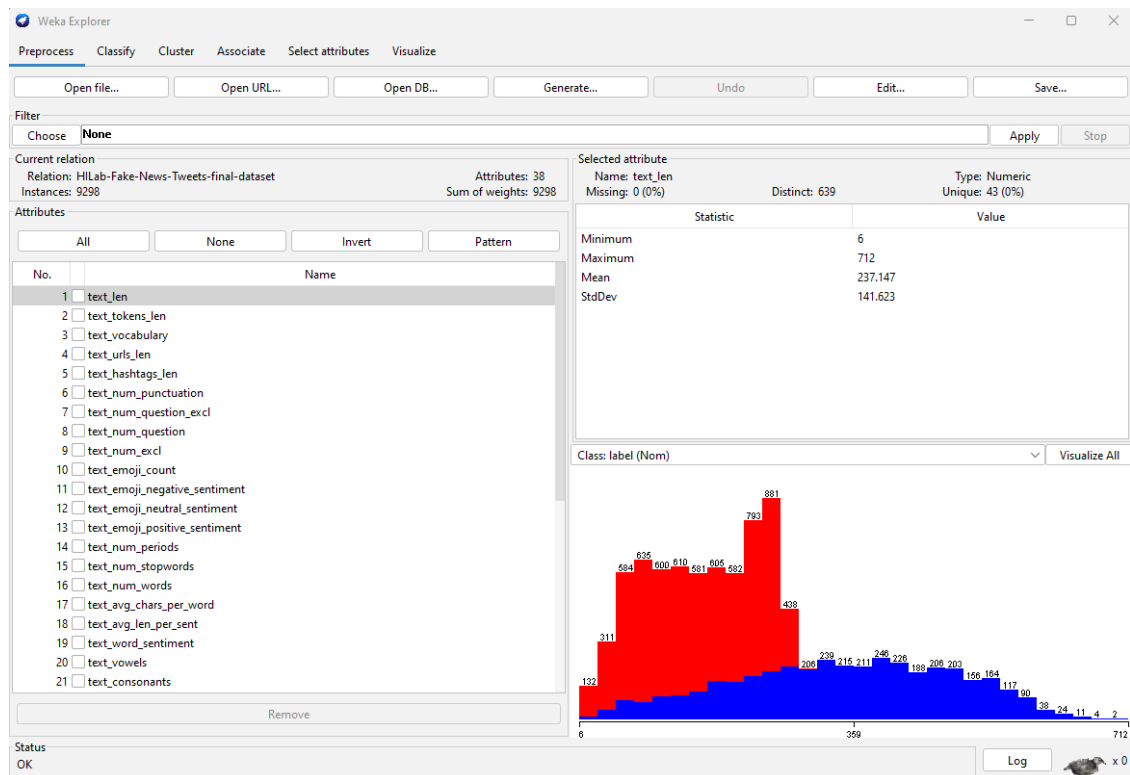
Η είσοδος των δεδομένων στο περιβάλλον Weka γίνεται μέσω αρχείων τα οποία μπορούν να φορτωθούν στη μνήμη είτε διαλογικά είτε μέσω κατάλληλων συναρτήσεων. Υποστηρίζονται διάφορες μορφές αρχείων. Η βασική μορφή είναι τα αρχεία ARFF (Attribute-Relation File Format) τα οποία είναι αρχεία χαρακτήρων (text files) όπου οι τιμές διαχωρίζονται με κόμμα (Coma Separated Values (CSV)). Κάθε ARFF αρχείο περιέχει μια επικεφαλίδα, στην οποία ορίζονται το όνομα της σχέσης (relation) και τα πεδία (attributes). Παράδειγμα αρχείου ARFF που συνοδεύει το λογισμικό WEKA, σχετικό με το είδος των φακών επαφής που πρέπει να φορέσει ένας ασθενής με βάση τα χαρακτηριστικά του (ηλικία, είδος φακού, αστιγματισμό και δάκρυσμα) είναι το παρακάτω:

```
1:% Title: Database for fitting contact lenses
2:% Attribute Information:
3:% -- 3 Classes
4:% hard : the patient should be fitted with hard contact lenses,
5:% soft : the patient should be fitted with soft contact lenses,
6:% none : the patient should not be fitted with contact lenses.
7:%
8:% 1. age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic
9:% 2. spectacle prescription: (1) myope, (2) hypermetrope
10:% 3. astigmatic: (1) no, (2) yes
11:% 4. tear production rate: (1) reduced, (2) normal
12:
13:@relation contact-lenses
14:
15:@attribute age {young, pre-presbyopic, presbyopic}
16:@attribute spectacle-prescrip {myope, hypermetrope}
```

```
17:@attribute astigmatism {no, yes}
18:@attribute tear-prod-rate {reduced, normal}
19:@attribute contact-lenses {soft, hard, none}
20:
21:@data
22:young,myope,no,reduced,none
23:young,myope,no,normal,soft
24:young,myope,yes,normal,hard
25:young,hypermetrope,no,reduced,none
26:young,hypermetrope,yes,normal,hard
27:pre-presbyopic,myope,no,reduced,none
28:pre-presbyopic,myope,no,normal,soft
29:pre-presbyopic,hypermetrope,yes,normal,none
30:presbyopic,hypermetrope,no,normal,soft
31:presbyopic,hypermetrope,yes,reduced,none
```

Ο δημιουργός του ARFF αρχείου μπορεί να ενσωματώσει διάφορες πληροφορίες μέσω σχολίων (γραμμές 1-12). Οι γραμμές με σχόλια πρέπει να ξεκινάνε με τον χαρακτήρα %. Στη συνέχεια, υπάρχει η λέξη *@relation* και ακολουθεί το όνομα του πίνακα δεδομένων που στο παράδειγμα είναι *contact-lenses* (γραμμή 13). Ακολουθεί η δήλωση των πεδίων (γραμμές 15-19). Για κάθε πεδίο απαιτείται μια γραμμή, στην αρχή της οποίας υπάρχει η λέξη *@attribute*, ακολουθεί το όνομα του πεδίου (π.χ. *age*), και κατόπιν δηλώνεται ο τύπος του πεδίου. Αν το πεδίο είναι αριθμητικό, χρησιμοποιείται η λέξη *numeric*. Αν είναι ονομαστικό (nominal), δηλώνονται οι δυνατές τιμές μέσα σε αγκύλες. Αν είναι συμβολοσειρά, χρησιμοποιείται η λέξη *string*. Στο παραπάνω παράδειγμα όλα τα πεδία είναι ονομαστικά. Μετά τη δήλωση των πεδίων ακολουθεί το τμήμα με τα δεδομένα, που δηλώνεται με σε μια γραμμή με τη λέξη *@data* (γραμμή 21). Κάθε γραμμή των δεδομένων περιγράφει έναν ασθενή ως διάνυσμα όπου οι συντεταγμένες είναι τα χαρακτηριστικά (attributes) που δίνονται ως τιμές που χωρίζονται με κόμμα. Στο παραπάνω παράδειγμα, στη γραμμή 22, βρίσκεται η περιγραφή ενός ασθενή για τον οποίο τα πεδία *age*, *spectacle-prescrip*, *astigmatism* και *tear production rate* έχουν τιμές *young,myope,no,reduced,none* αντίστοιχα.

Πέρα από το αρχείο ARFF, το Weka μπορεί να δεχτεί και άλλους τύπους αρχείων όπως τα πολύ διαδεδομένα CSV αρχεία, αρχεία xls καθώς και Matlab ASCII αρχεία.



Σχήμα Δ'.6: Οθόνη Explorer του WEKA με φορτωμένα τα fakes news δεδομένα.

Με τη βοήθεια εργαλεία που παρέχει το περιβάλλον Weka, κάθε ένα από τα παραπάνω αρχεία μπορεί να μετατραπεί σε αρχείο ARFF.

Δ'.3 Δεδομένα

Το σύνολο δεδομένων εκπαίδευσης που χρησιμοποιήθηκαν στην παρούσα εργασία είναι ελεύθερα διαθέσιμα και δημιουργήθηκαν στα πλαίσια των εργασιών [1, 27]. Παρακάτω περιγράφεται συνοπτικά η διαδικασία συλλογής και προετοιμασίας τους. Το αρχικό σύνολο των δεδομένων ψευδών ειδήσεων ανακτήθηκε το 2019 από το Twitter's Election Integrity Hub, περιείχε συνολικά 13,856,454 tweets και κάθε tweet περιελάμβανε 31 πεδία σχετικά με το περιεχόμενο του tweet και τον αποστολέα του.

Πίνακας Δ'.1: Συγκεντρωτικός πίνακας των χαρακτηριστικών των tweets.

Χαρακτηριστικά	
1 Μήκος Tweet	20 Πλήθος φωνηέντων
2 Πλήθος tokens	21 Πλήθος συμφώνων
3 TTR	22 Πλήθος κεφαλαίων
4 Πλήθος URLs	23 Πλήθος πεζών
5 Πλήθος hashtags	24 Μέγιστο μήκος ακολ φωνηέντων
6 Πλήθος σημείων στίξης	25 Μέγιστο μήκος ακολ συμφώνων
7 Πλήθος “?!”	26 Ύπαρξη >3 διαδοχ ίδιων χαρακτ
8 Πλήθος “?”	27 Πλήθος ψηφίων
9 Πλήθος “!”	28 Πλήθος γραμμάτων
10 Πλήθος emojis	29 Λόγος γραμμάτων προς ψηφία
11 Αθροισμα αρν emoji συναισθ	30 Πλήθος αντωνυμιών
12 Αθροισμα ουδετ emoji συναισθ	31 Πλήθος προσδιορισμών
13 Αθροισμα θετ emoji συναισθ	32 Πλήθος ουσιαστικών
14 Πλήθος περιόδων	33 Πλήθος επιρρημάτων
15 Πλήθος κοινών λέξεων	34 Πλήθος “to”
16 Πλήθος λέξεων	35 Πλήθος ρημάτων
17 Μέσο πλήθος χαρακτήρων ανά λέξη	36 Πλήθος entities
18 Μέσο πλήθος χαρακτ ανά πρόταση	37 Εντροπία των tweets
19 Συνολικό συναισθ βάρος κειμένου	38 Ετικέτα (χαρακτηρισμός)

Συγκεκριμένα, όπως φαίνεται και από τον πίνακα Δ'.1, τα πεδία μπορούν να ταξινομηθούν σε 3 κατηγορίες:

- 11 περιείχαν πληροφορίες σχετικές με τον λογαριασμό (account) όπως χρησιμοποιούμενο όνομα, πλήθος ακόλουθων (followers), πλήθος λογαριασμών που ακολουθούσε ο χρήστης, κ.α.
- 14 περιείχαν πληροφορίες σχετικές με το μήνυμα tweet όπως το κείμενο, τα hashtags, σχετικά URLs, κ.α.
- 6 περιείχαν πληροφορίες σχετικές με το δίκτυο όπως πλήθος θετικών αξιολογήσεων (likes), προωθήσεων (retweets) κ.α.

[illegible]

Σχήμα Δ'.7: Ενδεικτικό νέφος λέξεων με βάση τις συχνότητες εμφάνισης στα κείμενα των fake news tweets [1]

Για τον εντοπισμό των σχετικών με τη διάδοση ψευδών ειδήσεων tweets, οι Zervopoulos et al [1] πειραματίστηκαν με τις συχνότητες των λέξεων τόσο στο Αγγλικό όσο και στο μεταφρασμένο από τη Κινεζική κείμενο. Ενδεικτικό παράδειγμα των νεφών λέξεων (word clouds) που προέκυψαν φαίνεται στην εικόνα Δ'.7. Τα βήματα της προεπεξεργασίας ήταν τα παρακάτω: Αρχικά, δημιουργήθηκε λίστα με hashtags που σχετίζονται με τις διαδηλώσεις στο Hong Kong. Στη συνέχεια, hashtags σχετικά με αυτά της προηγούμενης λίστας ιχνηλατήθηκαν και υπολογίστηκαν οι συχνότητες συνύπαρξης (co-occurrence) τους για όλα τα διαθέσιμα tweets. Τέλος, τα tweets με χαμηλές συχνότητες απορρίφθηκαν ως μη σχετικά με τις διαδηλώσεις του Hong Kong. Το μειονέκτημα αυτής της διαδικασίας επιλογής είναι ότι απαιτεί την ύπαρξη των hashtags που αναφέρθηκαν. Μετά την ολοκλήρωση της διαδικασίας επιλογής απέμειναν 3908 tweets σχετικά με τις διαδηλώσεις του Hong Kong.

Έχοντας ολοκληρώσει το στάδιο επιλογής των δεδομένων εκπαίδευσης ως προς το κομμάτι των ψευδών ειδήσεων, απομένει το στάδιο εύρεσης δεδομένων που αφορούν αληθείς ειδήσεις. Η δυσκολία εδώ έγκειται στο ότι δεν υπάρχουν διαθέσιμα σύνολα δεδομένων πραγματικών (αληθών) ειδήσεων από το Twitter. Επίσης, η επιστράτευση έμπειρων δημοσιογράφων για την επεξεργασία και τον χαρακτηρισμό των διαθέσιμων tweets ώστε να εντοπιστούν αυτά που είναι φορείς πραγματικών ειδήσεων είναι χρονοβόρα και δαπανηρή. Για τους λόγους αυτούς, η σύνθεση του συνόλου των δεδομένων εκπαίδευσης για τις πραγματικές ειδήσεις έγινε επιλέγοντας εκείνα τα tweets των διαδηλώσεων του Hong Kong που έχουν ως πηγή (λογαριασμό Twitter) γνωστά και αξιόπιστα πρακτορεία ειδήσεων (news agencies) ή πιστοποιημένους δημοσιογράφους. Τα κύρια πρακτορεία ειδήσεων που θεωρήθηκαν αξιόπιστα είναι τα παρακάτω: BBC News, Reuters, Bloomberg, BuzzFeed, Channel News Asia, CNN, Agence France-Presse, South China Morning Post, Wall Street Journal, The New York Times, The Associated Press, The Washington Post και Quartz. Συνολικά από τους 120 λογαριασμούς (13 από πρακτορεία και 107 από δημοσιογράφους) συγκεντρώθηκαν 41996 και 103359 tweets, αντίστοιχα.

Η στατιστική επεξεργασία των tweets για πραγματικές ειδήσεις έδειξε ότι τα κείμενα των ψευδών tweets μοιάζουν περισσότερο στο κείμενο των tweets των δημοσιογράφων παρά σε αυτό των πρακτορείων ειδήσεων [1].

Η επιλογή των πραγματικών tweets για την ενσωμάτωσή τους στο σύνολο δεδομένων εκπαίδευσης έγινε με την ίδια μέθοδο που εφαρμόστηκε στα ψευδή tweets.

Σχετικά θεωρήθηκαν τα 5388 και 666 από τα 103359 και 41996 tweets των δημοσιογράφων και πρακτορείων, αντίστοιχα. Συνολικά, το σύνολο δεδομένων εκπαίδευσης που δημιουργήθηκε περιέχει 3908 και 5388 ψευδή και αληθή tweets.

Η επιλογή των χαρακτηριστικών για την αναπαράσταση του κάθε tweet έγινε με γλωσσολογικά κριτήρια και οδήγησε στην δημιουργία 38 χαρακτηριστικών συμπεριλαμβανομένου του χαρακτηριστικού της κλάσης [1]. Τα χαρακτηριστικά με τη μεγαλύτερη συσχέτιση την κλάση των tweets αναφέρονται να είναι αυτά της εντροπίας του tweet, του μήκους του και του λόγου TTR (type to token ratio).

Δ'.4 Μεθοδολογία

Στην εργασία αυτή χρησιμοποιήθηκαν οι εξής αλγόριθμοι: Απλοϊκός Bayes, Λογιστική Παλινδρόμηση, k Πλησιέστεροι Γείτονες ($k=1,3,5,7$), Κανόνες Απόφασης OneR, Μηχανές Διανυσμάτων Στήριξης, Νευρωνικό Δίκτυο Perceptron πολλαπλών στρωμάτων (MLP), και Δένδρα Απόφασης-J48 που αναλύθηκαν προηγουμένως.

Δοσμένου του συνόλου εκπαίδευσης που περιγράφηκε προηγουμένως, στην πρώτη φάση του πειράματος εφαρμόζεται το φίλτρο επιλογής χαρακτηριστικών του WEKA, AttributeSelection, ώστε ο αρχικός αριθμός των 38 χαρακτηριστικών να περιορισθεί και να αφαιρεθούν αυτά χωρίς αξιόλογο πληροφοριακό περιεχόμενο.

Παράμετροι του φίλτρου αυτού είναι η μέθοδος αξιολόγησης των χαρακτηριστικών και η στρατηγική αναζήτησης του βέλτιστου υποσυνόλου χαρακτηριστικών. Είναι εμφανές πως ο αριθμός των attributes μειώνεται αρκετά μετά την εφαρμογή του φίλτρου.

Στα πλαίσια της εργασίας αυτής, δοκιμάστηκε η μέθοδος επιλογής χαρακτηριστικών του Weka, CfsSubsetEval. Οι διάφοροι συνδυασμοί των χαρακτηριστικών δοκιμάζονται για να βρεθεί εκείνος που πετυχαίνει την βέλτιστη πρόβλεψη της κλάσης των δεδομένων εκπαίδευσης. Η CfsSubsetEval εκτιμά την αξία κάθε υποσυνόλου χαρακτηριστικών υπολογίζοντας την δυνατότητα πρόβλεψης του κάθε χαρακτηριστικού συνεκτιμώντας τον βαθμό πλεονασμού μεταξύ τους. Έτσι, υποσύνολα χαρακτηριστικών με υψηλή συσχέτιση με την κλάση αλλά και χαμηλή συσχέτιση μεταξύ τους προτιμώνται [28]. Μεταξύ των 38 χαρακτηριστικών, προκρίθηκαν τα παρακάτω:

text_len, text_vocabulary, text_urls_len, text_hashtags_len,
text_num_question, text_num_excl, text_emoji_negative_sentiment,
text_sum_word_entropy

Επίσης, με επαναληπτική εφαρμογή του OneR, επελέγησαν τα παρακάτω 4 χαρακτηριστικά

text_sum_word_entropy, text_len, text_vocabulary, text_num_entities

Προέκυψαν έτσι δύο σύνολα εκπαίδευσης, ένα με 8 χαρακτηριστικά και ένα με 4.

Η εφαρμογή του απλοϊκού κατηγοριοποιητή Bayes στα δύο σύνολα εκπαίδευσης έδωσε πολύ καλά αποτελέσματα όπως φαίνεται στην εικόνα Δ'.8 με ποσοστά επιτυχίας 91.99% και 90.83%, αντίστοιχα. Και εδώ, όπως και σε άλλες περιπτώσεις, ο απλοϊκός κατηγοριοποιητής Bayes επιτυγχάνει πολύ καλές επιδόσεις και δικαιολογεί τον λόγο που θεωρείται σημείο αναφοράς και σύγκρισης με άλλους πιο σύνθετους κατηγοριοποιητές.

Η εφαρμογή της μεθόδου της λογιστικής παλινδρόμησης έδωσε τα αποτελέσματα που φαίνονται στην εικόνα Δ'.9 με ποσοστό επιτυχίας 91.44%. Για το δεύτερο σύνολο, το ποσοστό επιτυχίας ήταν 90.46%.

Τα αποτελέσματα για τη μέθοδο Μηχανών Διανυσμάτων Στήριξης με χρήση SMO, των Δένδρων Απόφασης J48 και των Νευρωνικών Δικτύων Perceptron Πολλαπλών Στρωμάτων φαίνονται στις Εικόνες Δ'.10, Δ'.11 και Δ'.12 αντίστοιχα.

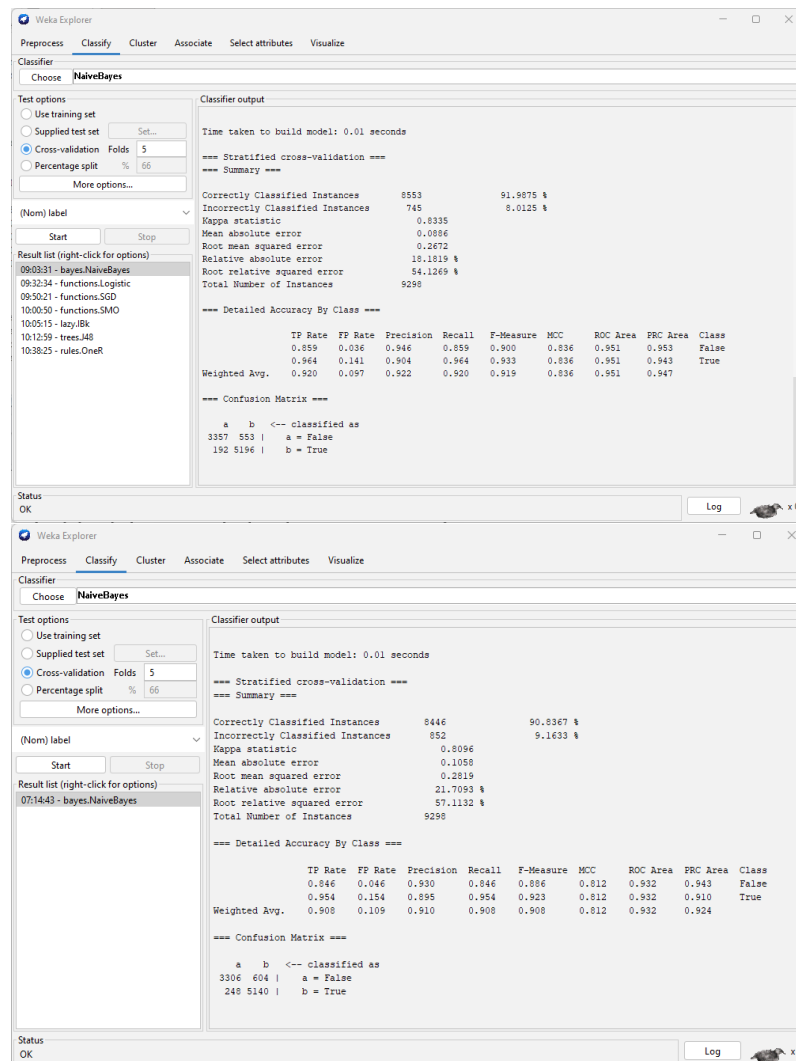
Παρατηρούμε ότι όλοι οι κλασσικοί αλγόριθμοι μηχανικής μάθησης έδωσαν υψηλά ποσοστά επιτυχίας στην ανίχνευση ψευδών ειδήσεων (πάνω από 90%). Οι δύο καλύτεροι με επιδόσεις σχεδόν ίσες είναι το νευρωνικό δίκτυο Perceptron Πολλαπλών Στρωμάτων και το Δένδρο Απόφασης J48 με μέσες επιδόσεις ως προς την ακρίβεια, την ανάκληση και το F_1 93%, 92.9%, 92.8% και 92.9%, 92.8%, 92.8%, αντίστοιχα. Αξιοσημείωτες είναι οι επιδόσεις του απλοϊκού Bayes και των 5 πλησιέστερων γειτόνων που είναι στο 92% περίπου, πολύ κοντά στους καλύτερους. Υπενθυμίζεται ότι στην περίπτωση του κατηγοριοποιητή πλησιέστερων γειτόνων δεν απαιτείται καν εκπαίδευση με την έννοια ότι τα ίδια τα δεδομένα εκπαίδευσης συνιστούν τον κατηγοριοποιητή.

Τα αποτελέσματα των συνδυαστικών μεθόδων Bagging και AdaBoost συνοψίζονται στους συγκεντρωτικούς πίνακες για τα δύο σύνολα χαρακτηριστικών που χρησιμο-

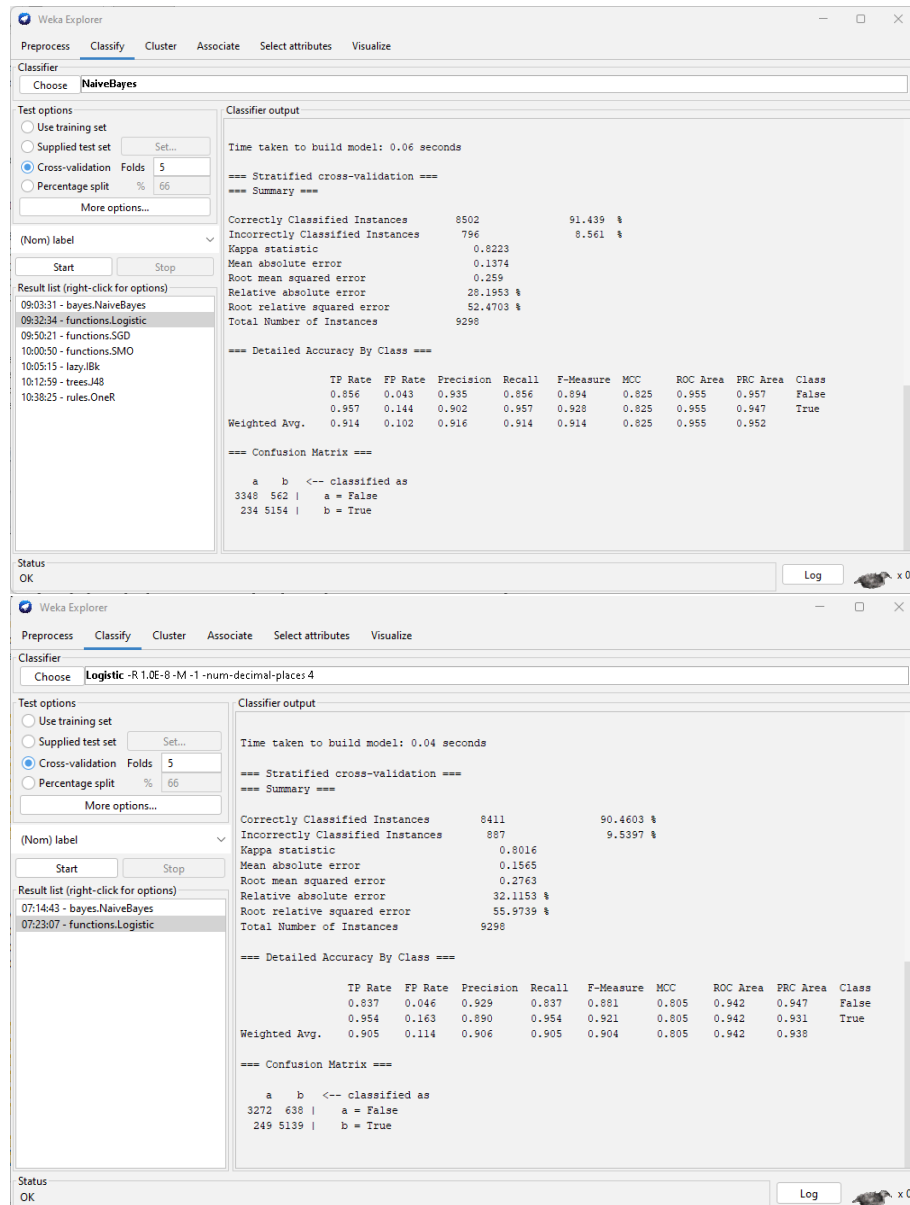
ποιούνται. Και για τις δύο συνδυαστικές μεθόδους, δοκιμάστηκαν ως παράμετροι, ταξινομητές μηχανών διανυσμάτων στήριξης, δέντρων απόφασης και νευρωνικών δικτύων multiperceptron (MLP). Τα αποτελέσματα είναι όμοια με μικρή υπεροχή του MLP όπως φαίνεται και στους συγκεντρωτικούς πίνακες.

Πίνακας Δ'.2: Συγκεντρωτικός πίνακας αξιολόγησης των ταξινομητών για το σύνολο δεδομένων εκπαίδευσης με τα 8 χαρακτηριστικά.

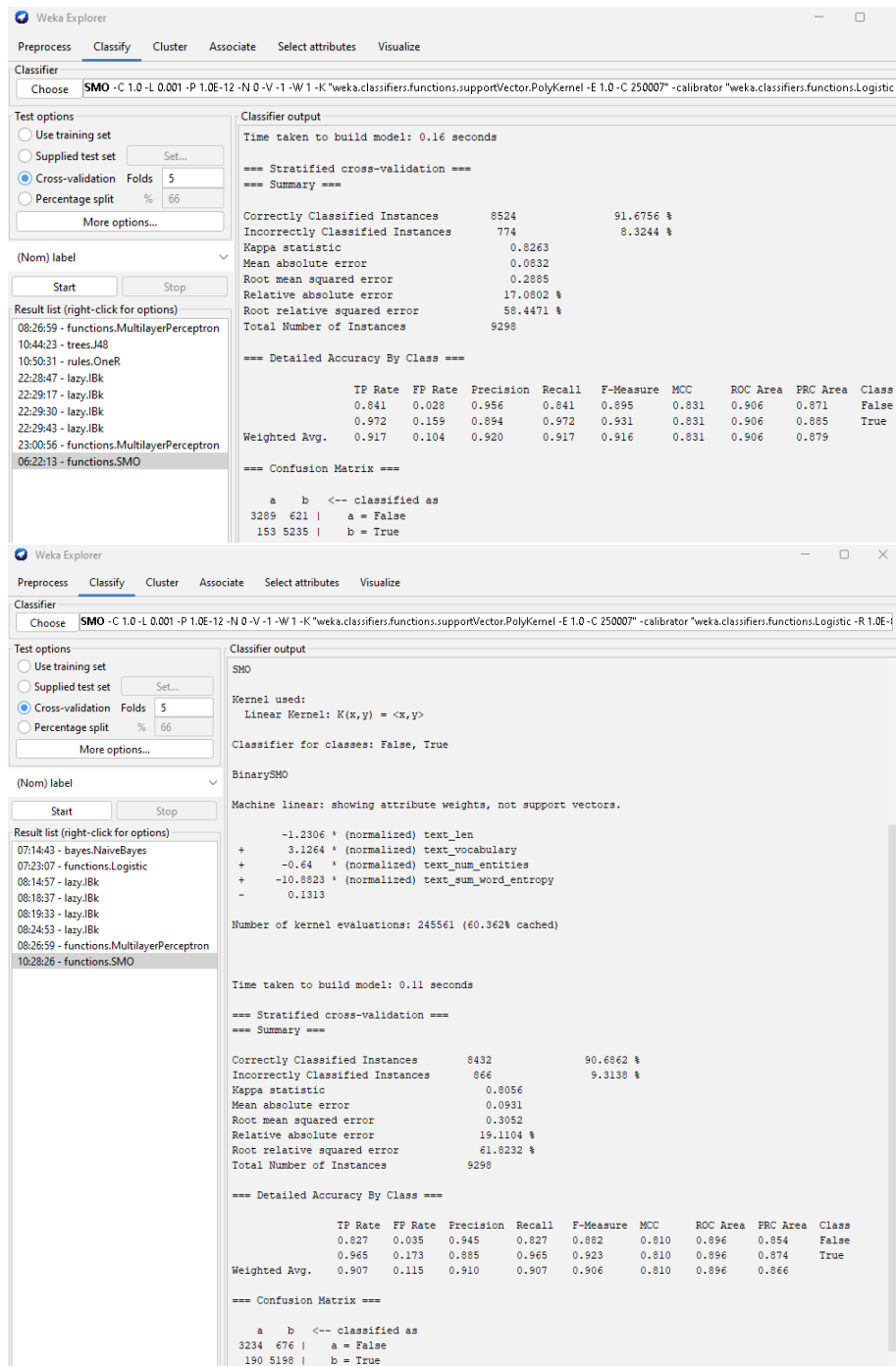
Αλγόριθμος	Τάξη	Ακρίβεια	Ανάκληση	F_1
Naive Bayes	Ψευδής	94.6 %	85.9 %	90.0 %
	Αληθής	90.4 %	96.4 %	93.3 %
	Μέση Τιμή	92.2 %	92.0 %	91.9 %
Logistic Regression	Ψευδής	93.5 %	85.6 %	89.4 %
	Αληθής	90.2 %	95.7 %	92.8 %
	Μέση Τιμή	91.6 %	91.4 %	91.4 %
5-Nearest Neighbour	Ψευδής	94.5 %	87.3 %	90.8 %
	Αληθής	91.3 %	96.3 %	93.7 %
	Μέση Τιμή	92.6 %	92.5 %	92.5 %
Support Vector Machine	Ψευδής	95.6 %	84.1 %	89.5 %
	Αληθής	89.4 %	97.2 %	93.1 %
	Μέση Τιμή	92.0 %	91.7 %	91.6 %
Κανόνας Απόφασης OneR	Ψευδής	88.9 %	80.1 %	84.3 %
	Αληθής	86.5 %	92.7 %	89.5 %
	Μέση Τιμή	87.5 %	87.4 %	87.3 %
Δένδρο Απόφασης (C4.5-J.48)	Ψευδής	95.0 %	87.5 %	91.1 %
	Αληθής	91.4 %	96.7 %	94.0 %
	Μέση Τιμή	92.9 %	92.8 %	92.8 %
Νευρωνικό Δίκτυο MLP	Ψευδής	95.5 %	87.1 %	91.1 %
	Αληθής	91.2 %	97.0 %	94.0 %
	Μέση Τιμή	93.0 %	92.9 %	92.8 %
Bagging MLP	Ψευδής	96.0 %	88.0 %	91.8 %
	Αληθής	91.8 %	97.3 %	94.5 %
	Μέση Τιμή	93.6 %	93.4 %	93.4 %
AdaBoost MLP	Ψευδής	95.1 %	87.8 %	91.3 %
	Αληθής	91.6 %	96.8 %	94.1 %
	Μέση Τιμή	93.1 %	93.0 %	92.9 %



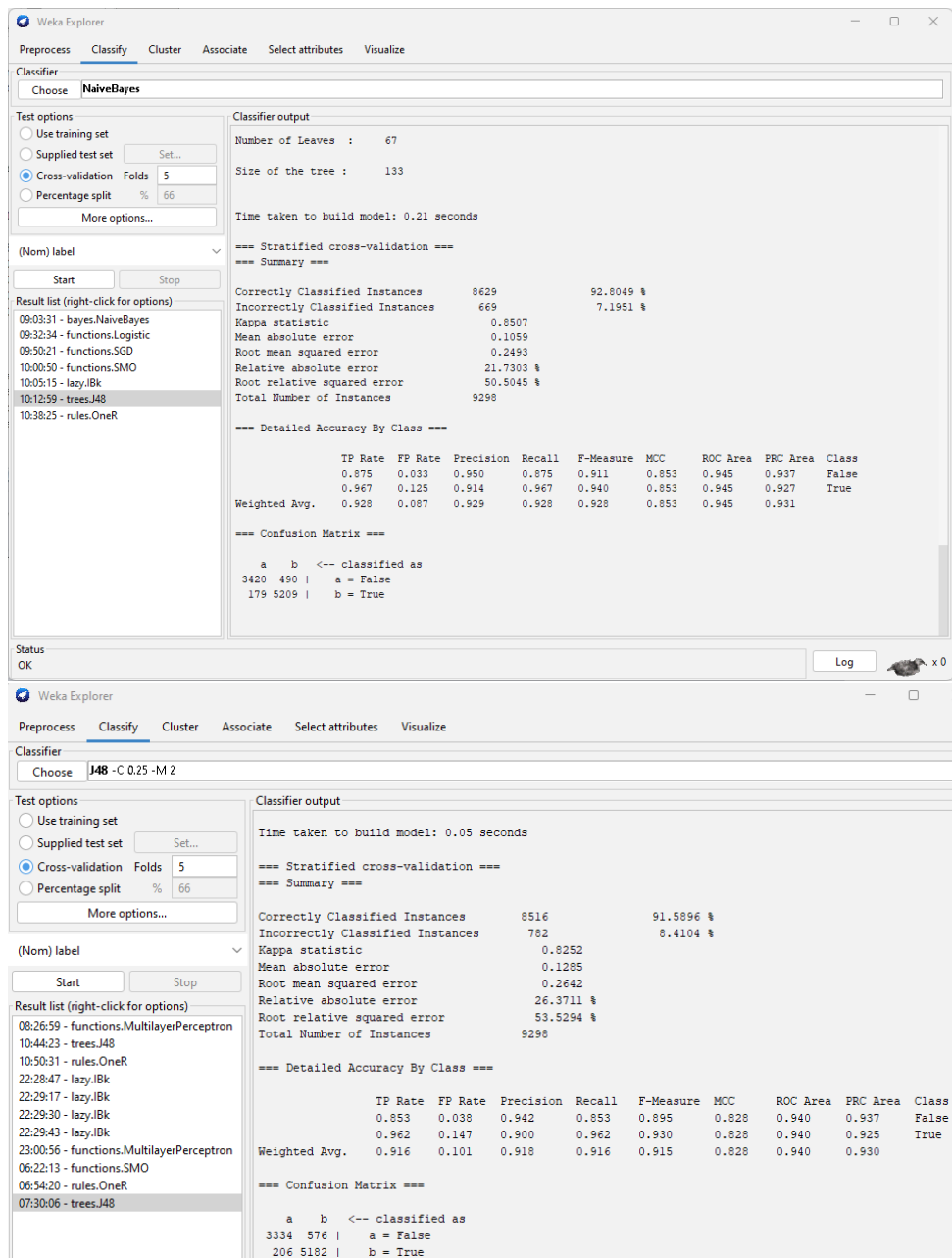
Σχήμα Δ'.8: Εφαρμογή του αλγόριθμου Naive Bayes στα σύνολα δεδομένων μετά από επιλογή 8 και 4 χαρακτηριστικών, αντίστοιχα.



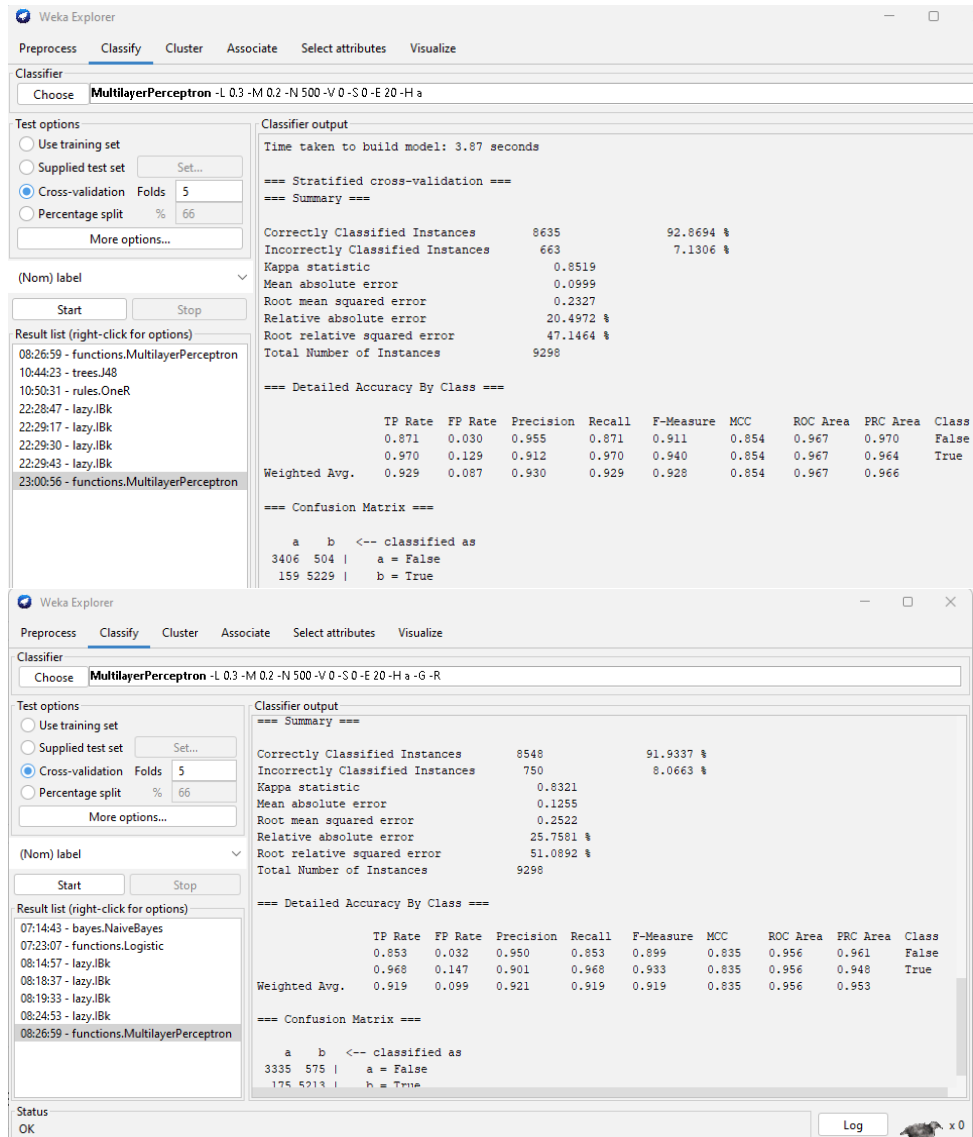
Σχήμα Δ'.9: Εφαρμογή του αλγόριθμου Logistic Regression.



Σχήμα Δ'.10: Εφαρμογή του αλγόριθμου SVM.



Σχήμα Δ'.11: Εφαρμογή του αλγόριθμου Δέντρου Απόφασης C4.5 (J48).



Σχήμα Δ'.12: Εφαρμογή Νευρωνικού Δικτύου Perceptron Πολλαπλών Στρωμάτων.

Πίνακας Δ'.3: Συγκεντρωτικός πίνακας αξιολόγησης των ταξινομητών για το σύνολο δεδομένων εκπαίδευσης με τα 4 χαρακτηριστικά.

Αλγόριθμος	Τάξη	Ακρίβεια	Ανάκληση	F_1
Naive Bayes	Ψευδής	93.0 %	84.6 %	88.6 %
	Αληθής	89.5 %	95.4 %	92.3 %
	Μέση Τιμή	91.0 %	90.8 %	90.8 %
Logistic Regression	Ψευδής	92.9 %	83.7 %	88.1 %
	Αληθής	89.0 %	95.4 %	92.1 %
	Μέση Τιμή	90.6 %	90.5 %	90.4 %
5-Nearest Neighbour	Ψευδής	93.1 %	85.2 %	88.9 %
	Αληθής	89.9 %	95.4 %	92.5 %
	Μέση Τιμή	91.2 %	91.1 %	91.0 %
Support Vector Machine	Ψευδής	94.5 %	82.7 %	88.2 %
	Αληθής	88.5 %	96.5 %	92.3 %
	Μέση Τιμή	91.0 %	90.7 %	90.6 %
Κανόνας Απόφασης OneR	Ψευδής	88.9 %	80.1 %	84.3 %
	Αληθής	86.5 %	92.7 %	89.5 %
	Μέση Τιμή	87.5 %	87.4 %	87.3 %
Δένδρο Απόφασης (C4.5-J.48)	Ψευδής	94.2 %	85.3 %	89.5 %
	Αληθής	90.0 %	96.2 %	93.0 %
	Μέση Τιμή	91.8 %	91.6 %	91.5 %
Νευρωνικό Δίκτυο MLP	Ψευδής	95.0 %	85.3 %	89.9 %
	Αληθής	90.1 %	96.8 %	93.3 %
	Μέση Τιμή	92.1 %	91.9 %	91.9 %
Bagging MLP	Ψευδής	95.5 %	84.9 %	89.9 %
	Αληθής	89.9 %	97.1 %	93.3 %
	Μέση Τιμή	92.2 %	92.0 %	91.9 %
AdaBoost MLP	Ψευδής	95.4 %	84.7 %	89.7 %
	Αληθής	89.7 %	97.0 %	93.2 %
	Μέση Τιμή	92.1 %	91.8 %	91.8 %

Κεφάλαιο Ε΄

Συμπεράσματα

Η καθολική χρήση του διαδικτύου και των υπηρεσιών που προσφέρει έχει επιφέρει νέους τρόπους επικοινωνίας μεταξύ των ανθρώπων. Είναι αναμφισβήτητο το γεγονός ότι τα μέσα κοινωνικής δικτύωσης αποτελούν αναπόσπαστο κομμάτι της καθημερινότητάς εκατοντάδων εκατομμυρίων ανθρώπων καλύπτοντας ένα ευρύ φάσμα αναγκών τους. Στο θέμα της διάδοσης ειδήσεων, έχει λάβει σημαντικές διαστάσεις το πρόβλημα της διάδοσης ψευδών ειδήσεων στο twitter εξαιτίας μιας σειράς πλεονεκτημάτων που προσφέρει. Ένας πολλά υποσχόμενος τρόπος αντιμετώπισής του είναι η επιστράτευση μεθόδων και τεχνικών μηχανικής μάθησης.

Στα πλαίσια της παρούσας εργασίας δοκιμάστηκαν και εφαρμόστηκαν πλήθος επιβλεπόμενων αλγόριθμων μηχανικής μάθησης που προσφέρονται από το περιβάλλον του Weka. Το πολύ μεγάλο πλεονέκτημα του Weka είναι η φιλικότητα προς τον χρήστη, η δυνατότητα εφαρμογής πληθώρας αλγόριθμων μέσα από ένα γραφικό περιβάλλον επικοινωνίας με τον χρήστη καθώς και η ευκολία επεξεργασίας δεδομένων εκπαίδευσης ανεξάρτητα από τη μορφή των αρχείων (xls, csv, arff).

Τα αρχικά δεδομένα στα οποία βασίστηκε η εργασία αφορούν τη διάδοση ψευδών ειδήσεων στο Twitter όπως αυτή έλαβε χώρα στις διαδηλώσεις του Hong Kong το 2019. Η συλλογή και προεπεξεργασία των δεδομένων εκπαίδευσης έγινε στα πλαίσια της εργασίας [1] και είναι ελεύθερα διαθέσιμα για ερευνητικούς σκοπούς.

Στα πειράματα που πραγματοποιήθηκαν έγινε επιλογή των χαρακτηριστικών ώστε να επιτευχθεί μείωση της διάστασης και προέκυψαν δύο σύνολα εκπαίδευσης με 8 και 4 χαρακτηριστικά, αντίστοιχα. Στη συνέχεια, εφαρμόστηκαν διάφοροι αλγό-

ριθμοί επιβλεπόμενης μηχανικής μάθησης ώστε να δημιουργηθούν και να αξιολογηθούν μοντέλα αυτόματης ανίχνευσης ψευδών ειδήσεων στο Twitter. Τα συγκριτικά αποτελέσματα από τα πειράματα φαίνονται πίνακα Δ'.2. Παρατηρούμε ότι όλοι οι κλασσικοί αλγόριθμοι μηχανικής μάθησης έδωσαν υψηλά ποσοστά επιτυχίας στην ανίχνευση ψευδών ειδήσεων (πάνω από 90%). Οι δύο καλύτεροι με επιδόσεις σχεδόν ίσες είναι το νευρωνικό δίκτυο Perceptron Πολλαπλών Στρωμάτων και το Δένδρο Απόφασης J48 με μέσες επιδόσεις ως προς την ακρίβεια, την ανάκληση και το F_1 93%, 92.9%, 92.8% και 92.9%, 92.8%, 92.8%, αντίστοιχα. Αξιοσημείωτες είναι οι επιδόσεις του απλοϊκού Bayes και των 5 πλησιέστερων γειτόνων που είναι στο 92% περίπου, πολύ κοντά στους καλύτερους.

Συνοψίζοντας, σύμφωνα με όλα τα παραπάνω διαπιστώνεται πως η αυτοματοποίηση της ανίχνευσης ψευδών ειδήσεων στο Twitter είναι εφικτή και δεν είναι υπερβολή να αναμένονται πολύ υψηλά ποσοστά επιτυχούς ανίχνευσης. Ωστόσο αναμένεται αξιοποίηση αυτών των αποτελεσμάτων από τους παραγωγούς των ψευδών ειδήσεων ώστε αυτές να γίνουν δυσκολότερα ανιχνεύσιμες και γι' αυτό είναι ανάγκη να δημιουργηθούν και να δοκιμαστούν μοντέλα με πολύ μεγαλύτερα σύνολα δεδομένων εκπαίδευσης που περιέχουν επιπρόσθετη πληροφορία για τις δικτυακές πηγές των ειδήσεων.

Βιβλιογραφία

- [1] Alexandros Zervopoulos, Aikaterini Georgia Alvanou, Konstantinos Bezas, Asterios Papamichail, Manolis Maragoudakis, and Katia Kermanidis. Hong kong protests: Using natural language processing for fake news detection on twitter. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 408–419. Springer, 2020.
- [2] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. Defining “fake news” a typology of scholarly definitions. *Digital journalism*, 6(2):137–153, 2018.
- [3] Bill Kovach and Tom Rosenstiel. *The Elements of Journalism, Revised and Updated 4th Edition: What Newspeople Should Know and the Public Should Expect*. Crown, 2021.
- [4] Denis McQuail. *McQuail’s mass communication theory*. Sage publications, 2010.
- [5] Meital Balmas. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication research*, 41(3):430–454, 2014.
- [6] Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. *arXiv preprint arXiv:1811.00770*, 2018.
- [7] Cecilia Kang and Adam Goldman. In washington pizzeria attack, fake news brought real guns. *New York Times*, 5, 2016.
- [8] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

- [9] Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 workshop on language technologies and computational social science*, pages 18–22, 2014.
- [10] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36, 2018.
- [11] William Ferreira and Andreas Vlachos. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL, 2016.
- [12] Ndapandula Nakashole and Tom Mitchell. Language-aware truth assessment of fact candidates. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1019, 2014.
- [13] Hadeer Ahmed, Issa Traore, and Sherif Saad. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer, 2017.
- [14] Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4, 2015.
- [15] Cody Buntain and Jennifer Golbeck. Automatically identifying fake news in popular twitter threads. In *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, pages 208–215. IEEE, 2017.
- [16] Stefan Helmstetter and Heiko Paulheim. Weakly supervised learning for fake news detection on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 274–277. IEEE, 2018.
- [17] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.
- [18] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

- [19] Ryszard Stanislaw Michalski, Jaime Guillermo Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.
- [20] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.
- [21] John Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.
- [22] Paul John Werbos. *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*, volume 1. John Wiley & Sons, 1994.
- [23] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [24] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.
- [25] Wasim Ahmed. Using twitter as a data source: An overview of current social media research tools. *LSE Impact of Social Sciences*, 2015.
- [26] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.
- [27] Alexandros Zervopoulos, Aikaterini Georgia Alvanou, Konstantinos Bezas, Asterios Papamichail, Manolis Maragoudakis, and Katia Kermanidis. Deep learning for fake news detection on twitter regarding the 2019 hong kong protests. *Neural Computing and Applications*, 34(2):969–982, 2022.
- [28] Mark A Hall and Lloyd A Smith. Practical feature subset selection for machine learning. 1998.

Συντμήσεις

API	Application Programming Interface
ARFF	(Attribute- Relation File Format
CSV	Comma Separated Values
k-NN	k Nearest Neighbors
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Network
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverted Document Frequency
URL	Uniform Resource Locator

Γλωσσάρι Ξενικών Όρων

Association Rules	Κανόνες Συσχέτισης
Classification	Κατηγοριοποίηση ή Ταξινόμηση
Classifier	Ταξινομητής
Clustering	Ομαδοποίηση ή Συσταδοποίηση
Confusion Matrix	Πίνακας Σύγχυσης
Fake News	Ψευδείς Ειδήσεις
Feature	Χαρακτηριστικό
Feature Vector	Διάνυσμα Χαρακτηριστικών
Learning	Μάθηση
Machine Learning	Μηχανική Μάθηση
Model	Μοντέλο

Natural Language Processing	Επεξεργασία Φυσικής Γλώσσας
Precision	Ακρίβεια
Regression	Παλινδρόμηση
Similarity	Ομοιότητα
Specificity	Εξειδίκευση
Supervised Learning	Επιβλεπόμενη Μάθηση
Training	Εκπαίδευση
Training Data	Δεδομένα Εκπαίδευσης
Unsupervised Learning	Μάθηση χωρίς Επίβλεψη