



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Ευφυής Ανάλυση Θεμάτων και Συχνότητας Hashtags σε Κοινωνικά Δίκτυα

Κιαμούρη Ελένη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία έτος 2026



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Ευφυής Ανάλυση Θεμάτων και Συχνότητας Hashtags σε Κοινωνικά Δίκτυα

Κιαμούρη Ελένη

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος
Αναπληρωτής Καθηγητής

Λαμία έτος 2026



UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

Intelligent Analysis of Topics and Hashtag Frequency on Social Media

Kiamouri Eleni

FINAL THESIS

ADVISOR

Dr Konstantinos Kolomvatsos, BSc MSc PhD Associate Professor in Intelligent
Systems for Pervasive Computing

Lamia year 2026

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις ⁽¹⁾, που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.

2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία: ...3.../...9.../20...26...

Η Δηλ.

ΚΙΑΜΟΥΡΗ ΕΛΕΝΗ

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.»

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία εξετάζει την ανάλυση και την αυτόματη θεματική κατηγοριοποίηση hashtags σε δεδομένα κοινωνικών δικτύων, με αξιοποίηση τεχνικών Μηχανικής και Βαθιάς Μάθησης. Για την πειραματική μελέτη χρησιμοποιήθηκε σύνολο 732 αναρτήσεων, στο οποίο τα hashtags υποβλήθηκαν σε διαδικασία εξαγωγής, κανονικοποίησης και θεματικής ομαδοποίησης. Με βάση προκαθορισμένες λέξεις-κλειδιά δημιουργήθηκαν οκτώ τελικές θεματικές κατηγορίες και προσδιορίστηκε μία κύρια κατηγορία για κάθε ανάρτηση, η οποία αποτέλεσε τη μεταβλητή-στόχο των μοντέλων ταξινόμησης. Παράλληλα, εξετάστηκε η αξιοποίηση διαφορετικών τύπων πληροφορίας, όπως οι δείκτες αλληλεπίδρασης των χρηστών, τα μεταδεδομένα των αναρτήσεων και η κειμενική πληροφορία. Υλοποιήθηκαν και συγκρίθηκαν οι αλγόριθμοι Random Forest, k-Nearest Neighbors και Support Vector Machine, καθώς και νευρωνικά μοντέλα LSTM, Bi-LSTM και υβριδικές αρχιτεκτονικές πολλαπλών εισόδων. Η αξιολόγηση πραγματοποιήθηκε με τις μετρικές Accuracy, Precision, Recall και F1-score, καθώς και με ανάλυση των πινάκων σύγχυσης. Τα αποτελέσματα έδειξαν ότι τα μοντέλα που βασίστηκαν αποκλειστικά στα Likes και Retweets παρουσίασαν περιορισμένη ικανότητα διάκρισης των θεματικών κατηγοριών, ενώ η ενσωμάτωση περισσότερων και διαφορετικών μορφών πληροφορίας οδήγησε σε σημαντική βελτίωση. Την υψηλότερη απόδοση παρουσίασε το υβριδικό μοντέλο Bi-LSTM Numeric + Text, με Accuracy 63,95% και macro F1-score 0,50. Συνολικά, τα ευρήματα αναδεικνύουν ότι η θεματική ταξινόμηση των hashtags δεν μπορεί να στηριχθεί αποτελεσματικά μόνο σε δείκτες αλληλεπίδρασης, αλλά ωφελείται σημαντικά από την ενσωμάτωση κειμενικής και συμπληρωματικής πληροφορίας. Παράλληλα, το περιορισμένο

μέγεθος και η ανισορροπία του συνόλου δεδομένων, καθώς και η παρουσία των `hashtags` στη μεταβλητή `metadata_text`, αποτελούν βασικούς παράγοντες που πρέπει να λαμβάνονται υπόψη κατά την ερμηνεία των αποτελεσμάτων.

ABSTRACT

This thesis examines the analysis and automatic thematic categorization of hashtags in social media data, utilizing machine learning and deep learning techniques. For the experimental study, a dataset of 732 posts were used, in which the hashtags underwent extraction, normalization, and thematic clustering. Based on predefined keywords, eight final thematic categories were created, and a primary category was identified for each post, which served as the target variable for the classification models. At the same time, the use of different types of information was examined, such as user interaction metrics, post metadata, and textual information. The Random Forest, k-Nearest Neighbors, and Support Vector Machine algorithms, as well as LSTM, Bi-LSTM neural networks, and hybrid multi-input architectures, were implemented and compared. Evaluation was performed using the metrics Accuracy, Precision, Recall, and F1-score, as well as through an analysis of confusion matrices. The results showed that models based exclusively on Likes and Retweets exhibited limited ability to distinguish between topic categories, while incorporating more and different types of information led to a significant improvement. The Bi-LSTM Numeric + Text hybrid model achieved the highest performance, with an accuracy of 63.95% and a macro F1-score of 0.50. Overall, the findings highlight that the thematic classification of hashtags cannot be effectively based solely on interaction metrics, but benefits significantly from the integration of textual and supplementary information. At the same time, the limited size and imbalance of the dataset are key factors that must be taken into account when interpreting the results.

Table of Contents

ΠΕΡΙΛΗΨΗ	I
ABSTRACT	IV
ΚΕΦΑΛΑΙΟ 1 ΕΙΣΑΓΩΓΗ	4
1.1 ΕΙΣΑΓΩΓΗ.....	4
1.2 ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ	4
1.3 ΣΚΟΠΟΣ ΚΑΙ ΣΤΟΧΟΙ ΤΗΣ ΕΡΓΑΣΙΑΣ	5
ΚΕΦΑΛΑΙΟ 2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ.....	7
2.1 ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ	7
2.1.1 ΟΡΙΣΜΟΣ ΤΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ	7
2.1.2 ΕΞΕΛΙΞΗ ΤΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ.....	7
2.1.3 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΚΑΙ ΣΥΓΧΡΟΝΟΣ ΡΟΛΟΣ ΤΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ	8
2.2 HASHTAGS ΚΑΙ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ	9
2.2.1 ΟΡΙΣΜΟΣ ΚΑΙ ΛΕΙΤΟΥΡΓΙΑ ΤΩΝ HASHTAGS.....	9
2.2.2 HASHTAGS ΩΣ ΔΕΙΚΤΕΣ ΘΕΜΑΤΙΚΟΥ ΠΕΡΙΕΧΟΜΕΝΟΥ.....	9
2.2.3 ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ	10
2.2.4 ΑΥΤΟΜΑΤΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΑΙ ΠΡΟΤΑΣΗ HASHTAGS	10
2.3 ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ.....	11
2.3.1 ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ	11
2.3.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΚΕΙΜΕΝΟΥ.....	12
2.3.3 ΑΡΙΘΜΗΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ ΚΕΙΜΕΝΟΥ ΚΑΙ EMBEDDINGS.....	12
2.3.4 ΕΠΕΞΕΡΓΑΣΙΑ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΠΕΡΙΕΧΟΜΕΝΟΥ ΚΟΙΝΩΝΙΚΩΝ ΔΙΚΤΥΩΝ	13
2.4 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ	13
2.4.1 ΕΙΣΑΓΩΓΗ ΣΤΗ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ	13
2.4.2 ΕΠΙΒΛΕΠΟΜΕΝΗ ΜΑΘΗΣΗ ΚΑΙ ΤΑΞΙΝΟΜΗΣΗ.....	14
2.4.3 ΤΑΞΙΝΟΜΗΣΗ ΚΕΙΜΕΝΟΥ.....	14
2.5 RANDOM FOREST.....	15
2.5.1 ΒΑΣΙΚΗ ΑΡΧΗ ΚΑΙ ΛΕΙΤΟΥΡΓΙΑ ΤΟΥ RANDOM FOREST	15
2.5.2 RANDOM FOREST ΣΤΗΝ ΤΑΞΙΝΟΜΗΣΗ.....	15
2.5.3 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΟΥ RANDOM FOREST.....	16
2.6 K-NEAREST NEIGHBORS (KNN)	16
2.6.1 ΒΑΣΙΚΗ ΑΡΧΗ ΚΑΙ ΛΕΙΤΟΥΡΓΙΑ ΤΟΥ K-NEAREST NEIGHBORS.....	16
2.6.2 ΑΠΟΣΤΑΣΗ, ΕΠΙΛΟΓΗ ΤΟΥ K ΚΑΙ ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ	17
2.6.3 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΟΥ KNN.....	18
2.7 SUPPORT VECTOR MACHINE (SVM)	18
2.7.1 ΒΑΣΙΚΗ ΑΡΧΗ ΚΑΙ ΛΕΙΤΟΥΡΓΙΑ ΤΟΥ SUPPORT VECTOR MACHINE.....	18
2.7.2 ΜΗ ΓΡΑΜΜΙΚΟΣ ΔΙΑΧΩΡΙΣΜΟΣ ΚΑΙ ΣΥΝΑΡΤΗΣΕΙΣ ΠΥΡΗΝΑ	18
2.7.3 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΟΥ SVM.....	19
2.8 ΒΑΘΙΑ ΜΑΘΗΣΗ ΚΑΙ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ	19
2.8.1 ΕΙΣΑΓΩΓΗ ΣΤΗ ΒΑΘΙΑ ΜΑΘΗΣΗ	19

2.8.2 ΤΕΧΝΗΤΑ ΚΑΙ ΒΑΘΙΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ	20
2.8.3 ΑΝΑΔΡΟΜΙΚΑ ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ (RNN)	20
2.8.4 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΗΣ ΒΑΘΙΑΣ ΜΑΘΗΣΗΣ	21
2.9 LONG SHORT-TERM MEMORY (LSTM)	22
2.9.1 ΒΑΣΙΚΗ ΑΡΧΗ ΚΑΙ ΛΕΙΤΟΥΡΓΙΑ ΤΟΥ LONG SHORT-TERM MEMORY (LSTM)	22
2.9.2 ΔΟΜΗ ΤΗΣ ΜΟΝΑΔΑΣ LSTM	22
2.9.3 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΟΥ LSTM	23
2.10 ΑΜΦΙΔΡΟΜΑ ΔΙΚΤΥΑ LSTM (BI-LSTM) ΚΑΙ ΥΒΡΙΔΙΚΕΣ ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ	23
2.10.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΚΑΙ ΛΕΙΤΟΥΡΓΙΑ ΤΟΥ BI-LSTM	23
2.10.2 BI-LSTM ΣΤΗΝ ΤΑΞΙΝΟΜΗΣΗ ΚΕΙΜΕΝΟΥ	24
2.10.3 ΥΒΡΙΔΙΚΕΣ ΚΑΙ ΠΟΛΥΚΑΝΑΛΙΚΕΣ ΑΡΧΙΤΕΚΤΟΝΙΚΕΣ	24
2.10.4 ΠΛΕΟΝΕΚΤΗΜΑΤΑ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΩΝ BI-LSTM ΚΑΙ ΥΒΡΙΔΙΚΩΝ ΑΡΧΙΤΕΚΤΟΝΙΚΩΝ	25
2.11 ΣΧΕΤΙΚΕΣ ΕΡΓΑΣΙΕΣ	25

ΚΕΦΑΛΑΙΟ 3 ΜΕΘΟΔΟΛΟΓΙΑ ΚΑΙ ΥΛΟΠΟΙΗΣΗ

3.1 ΣΥΝΟΛΟ ΔΕΔΟΜΕΝΩΝ.....	28
3.2 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ	29
3.2.1 ΕΞΑΓΩΓΗ ΚΑΙ ΚΑΝΟΝΙΚΟΠΟΙΗΣΗ ΤΩΝ HASHTAGS	29
3.2.2 ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΣΥΝΘΕΤΩΝ HASHTAGS.....	29
3.2.3 ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΤΩΝ HASHTAGS	30
3.2.4 ΚΑΘΟΡΙΣΜΟΣ ΤΗΣ ΚΥΡΙΑΣ ΚΑΤΗΓΟΡΙΑΣ	30
3.2.5 ΚΩΔΙΚΟΠΟΙΗΣΗ ΚΑΙ ΔΗΜΙΟΥΡΓΙΑ ΑΡΙΘΜΗΤΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ.....	31
3.2.6 ΔΗΜΙΟΥΡΓΙΑ ΚΕΙΜΕΝΙΚΗΣ ΑΝΑΠΑΡΑΣΤΑΣΗΣ ΜΕΤΑΔΕΔΟΜΕΝΩΝ	31
3.3 ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΕΙΣΟΔΟΥ	32
3.3.1 ΑΡΙΘΜΗΤΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΑΛΛΗΛΕΠΙΔΡΑΣΗΣ	32
3.3.2 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΜΕΤΑΔΕΔΟΜΕΝΩΝ	32
3.3.3 ΚΕΙΜΕΝΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ.....	33
3.3.4 ΣΥΝΔΥΑΣΜΟΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΑΝΑ ΜΟΝΤΕΛΟ	33
3.4 ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ RANDOM FOREST	33
3.4.1 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΕΙΣΟΔΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΗ-ΣΤΟΧΟΣ	33
3.4.2 ΔΙΑΧΩΡΙΣΜΟΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ.....	34
3.4.3 ΠΑΡΑΜΕΤΡΟΠΟΙΗΣΗ ΚΑΙ ΕΚΠΑΙΔΕΥΣΗ	34
3.4.4 ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ.....	35
3.5 ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ K-NEAREST NEIGHBORS.....	35
3.5.1 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΕΙΣΟΔΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΗ-ΣΤΟΧΟΣ.....	35
3.5.2 ΔΙΑΧΩΡΙΣΜΟΣ ΔΕΔΟΜΕΝΩΝ	35
3.5.3 ΠΑΡΑΜΕΤΡΟΠΟΙΗΣΗ ΚΑΙ ΕΚΠΑΙΔΕΥΣΗ	36
3.5.4 ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ.....	36
3.6 ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ SUPPORT VECTOR MACHINE.....	37
3.6.1 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΕΙΣΟΔΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΗ-ΣΤΟΧΟΣ	37
3.6.2 ΔΙΑΧΩΡΙΣΜΟΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ.....	37
3.6.3 ΠΑΡΑΜΕΤΡΟΠΟΙΗΣΗ ΚΑΙ ΕΚΠΑΙΔΕΥΣΗ	37
3.6.4 ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ.....	38
3.7 ΥΛΟΠΟΙΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ LSTM	39
3.7.1 ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΕΙΣΟΔΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΗ-ΣΤΟΧΟΣ.....	39
3.7.2 ΑΝΑΔΙΑΜΟΡΦΩΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΕΙΣΟΔΟΥ	39
3.7.3 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΤΟΥ ΔΙΚΤΥΟΥ	39

3.7.4 ΜΕΤΑΓΛΩΤΤΙΣΗ ΚΑΙ ΕΚΠΑΙΔΕΥΣΗ	40
3.7.5 ΑΞΙΟΛΟΓΗΣΗ ΤΟΥ ΜΟΝΤΕΛΟΥ.....	41
3.8 ΥΛΟΠΟΙΗΣΗ ΜΟΝΤΕΛΩΝ BIDIRECTIONAL LSTM.....	41
3.8.1 ΒΑΣΙΚΟ ΜΟΝΤΕΛΟ BI-LSTM.....	41
3.8.2 ΕΜΠΛΟΥΤΙΣΜΕΝΟ ΜΟΝΤΕΛΟ BI-LSTM.....	42
3.8.3 ΥΒΡΙΔΙΚΟ ΜΟΝΤΕΛΟ BI-LSTM ΠΟΛΛΑΠΛΩΝ ΕΙΣΟΔΩΝ.....	43
3.8.4 ΥΒΡΙΔΙΚΟ ΜΟΝΤΕΛΟ ΑΡΙΘΜΗΤΙΚΩΝ ΚΑΙ ΚΕΙΜΕΝΙΚΩΝ ΔΕΔΟΜΕΝΩΝ.....	44
3.8.5 ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ BI-LSTM	46

ΚΕΦΑΛΑΙΟ 4 ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ 47

4.1 ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ	47
4.1.1 ΔΙΑΧΩΡΙΣΜΟΣ ΤΟΥ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ	47
4.1.2 ΣΥΝΟΛΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ	47
4.1.3 ΡΥΘΜΙΣΕΙΣ ΕΚΠΑΙΔΕΥΣΗΣ.....	48
4.1.4 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ.....	48
4.1.5 ΔΙΑΔΙΚΑΣΙΑ ΣΥΓΚΡΙΣΗΣ	49
4.2 ΑΠΟΤΕΛΕΣΜΑΤΑ RANDOM FOREST	49
4.2.1 ΣΥΝΟΛΙΚΗ ΑΠΟΔΟΣΗ.....	49
4.2.2 ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ	49
4.2.3 ΑΝΑΛΥΤΙΚΕΣ ΜΕΤΡΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ	50
4.2.4 ΣΥΜΠΕΡΑΣΜΑ ΓΙΑ ΤΟ ΜΟΝΤΕΛΟ	51
4.3 ΑΠΟΤΕΛΕΣΜΑΤΑ K-NEAREST NEIGHBORS.....	51
4.3.1 ΣΥΝΟΛΙΚΗ ΑΠΟΔΟΣΗ.....	51
4.3.2 ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ	52
4.3.3 ΑΝΑΛΥΤΙΚΕΣ ΜΕΤΡΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ	53
4.3.4 ΣΥΜΠΕΡΑΣΜΑ ΓΙΑ ΤΟ ΜΟΝΤΕΛΟ	53
4.4 ΑΠΟΤΕΛΕΣΜΑΤΑ SUPPORT VECTOR MACHINE	53
4.4.1 ΣΥΝΟΛΙΚΗ ΑΠΟΔΟΣΗ.....	53
4.4.2 ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ	54
4.4.3 ΑΝΑΛΥΤΙΚΕΣ ΜΕΤΡΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ	55
4.4.4 ΣΥΜΠΕΡΑΣΜΑ ΓΙΑ ΤΟ ΜΟΝΤΕΛΟ	55
4.5 ΑΠΟΤΕΛΕΣΜΑΤΑ LSTM.....	56
4.5.1 ΣΥΝΟΛΙΚΗ ΑΠΟΔΟΣΗ.....	56
4.5.2 ΕΞΕΛΙΞΗ ΤΗΣ ΕΚΠΑΙΔΕΥΣΗΣ	56
4.5.3 ΠΙΝΑΚΑΣ ΣΥΓΧΥΣΗΣ	57
4.5.4 ΑΝΑΛΥΤΙΚΕΣ ΜΕΤΡΙΚΕΣ ΤΑΞΙΝΟΜΗΣΗΣ	58
4.5.5 ΣΥΜΠΕΡΑΣΜΑ ΓΙΑ ΤΟ ΜΟΝΤΕΛΟ	58
4.6 ΑΠΟΤΕΛΕΣΜΑΤΑ ΜΟΝΤΕΛΩΝ BI-LSTM.....	59
4.6.1 ΑΠΟΤΕΛΕΣΜΑΤΑ ΒΑΣΙΚΟΥ BI-LSTM	59
4.6.2 ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΜΠΛΟΥΤΙΣΜΕΝΟΥ BI-LSTM2.....	62
4.6.3 ΑΠΟΤΕΛΕΣΜΑΤΑ ΥΒΡΙΔΙΚΟΥ ΜΟΝΤΕΛΟΥ BI-LSTM	65
4.6.4 ΑΠΟΤΕΛΕΣΜΑΤΑ BI-LSTM NUMERIC + TEXT.....	69
4.7 ΣΥΓΚΡΙΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΜΟΝΤΕΛΩΝ	72
4.8 ΣΥΖΗΤΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ	75
4.9 ΣΥΓΚΡΙΣΗ ΜΕ ΠΡΟΗΓΟΥΜΕΝΕΣ ΜΕΛΕΤΕΣ	75
4.10 ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΕΦΑΛΑΙΟΥ	76

ΚΕΦΑΛΑΙΟ 5 ΣΥΜΠΕΡΑΣΜΑΤΑ..... 78

5.1 ΣΥΝΟΨΗ ΤΗΣ ΕΡΓΑΣΙΑΣ	78
5.2 ΣΥΜΠΕΡΑΣΜΑΤΑ.....	78
5.3 ΠΕΡΙΟΡΙΣΜΟΙ ΤΗΣ ΕΡΓΑΣΙΑΣ	80
5.4 ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ	80
<u>ΒΙΒΛΙΟΓΡΑΦΙΑ.....</u>	<u>82</u>

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή

1.1 Εισαγωγή

Τα κοινωνικά δίκτυα (Social Networks) αποτελούν πλέον αναπόσπαστο κομμάτι της καθημερινότητας δισεκατομμυρίων ανθρώπων παγκοσμίως, έχοντας μεταβάλει ριζικά τον τρόπο με τον οποίο επικοινωνούμε, ενημερωνόμαστε και αλληλεπιδρούμε. Πλατφόρμες όπως το X (πρώην Twitter), το Instagram, το Facebook και το TikTok λειτουργούν ως δυναμικά κανάλια παραγωγής και διαμοιρασμού πληροφοριών σε πραγματικό χρόνο.

Η εκρηκτική αυτή άνοδος των μέσων κοινωνικής δικτύωσης έχει οδηγήσει σε μια συνεχή, πρωτοφανή αύξηση του όγκου των παραγόμενων δεδομένων (Big Data). Καθημερινά παράγονται τεράστιες ποσότητες μη δομημένης πληροφορίας, η οποία περιλαμβάνει κείμενο, εικόνες, βίντεο και μεταδεδομένα (metadata). Η ανάγκη για τη γρήγορη αναζήτηση, κατηγοριοποίηση και ανάλυση αυτού του τεράστιου όγκου δεδομένων κατέστησε αναγκαία τη χρήση θεματικών ετικετών.

Στο πλαίσιο αυτό, τα hashtags (λέξεις ή φράσεις που ξεκινούν με το σύμβολο #) αναδείχθηκαν ως το πρωταρχικό εργαλείο για την οργάνωση της πληροφορίας. Τα hashtags λειτουργούν ως θεματικοί δείκτες που ομαδοποιούν τις αναρτήσεις γύρω από συγκεκριμένα θέματα, γεγονότα ή συζητήσεις, διευκολύνοντας τους χρήστες να εντοπίζουν περιεχόμενο που τους ενδιαφέρει. Παράλληλα, για τους ερευνητές και τις επιχειρήσεις, αποτελούν μια πολύτιμη πηγή δεδομένων για την ανάλυση των τάσεων (trends) και της κοινής γνώμης σε πραγματικό χρόνο.

1.2 Περιγραφή του προβλήματος

Η αυτόματη ανάλυση και κατηγοριοποίηση των hashtags στα σύγχρονα μέσα κοινωνικής δικτύωσης αποτελεί ένα σύνθετο πρόβλημα, το οποίο συνδέεται με την Επεξεργασία Φυσικής Γλώσσας, την Εξόρυξη Δεδομένων και τη Μηχανική Μάθηση. Η κύρια δυσκολία προκύπτει από το γεγονός ότι τα hashtags δημιουργούνται δυναμικά και αυθόρμητα από τους ίδιους τους χρήστες. Συνεπώς, δεν ακολουθούν πάντοτε συγκεκριμένους γραμματικούς ή συντακτικούς κανόνες και δεν περιορίζονται σε ένα σταθερό και προκαθορισμένο λεξιλόγιο.

Μία από τις σημαντικότερες δυσκολίες είναι η περιορισμένη πληροφορία που παρέχει ένα μεμονωμένο hashtag. Σε αντίθεση με ένα ολοκληρωμένο κείμενο, τα hashtags αποτελούνται συνήθως από μία λέξη ή από μία σύντομη φράση, με αποτέλεσμα να διαθέτουν ελάχιστο γλωσσικό πλαίσιο. Επομένως, η ερμηνεία τους μπορεί να είναι ασαφής, ιδιαίτερα όταν η ίδια λέξη χρησιμοποιείται με διαφορετική σημασία ανάλογα με την ανάρτηση ή τη διαδικτυακή συζήτηση στην οποία εμφανίζεται.

Πρόσθετη δυσκολία δημιουργούν οι μορφολογικές ιδιαιτερότητες των hashtags. Οι χρήστες συχνά συνδυάζουν πολλές λέξεις σε μία ενιαία συμβολοσειρά, χρησιμοποιώντας κεφαλαία γράμματα, κάτω παύλες ή χωρίς κανέναν διαχωριστικό χαρακτήρα, όπως στις περιπτώσεις #TravelAdventure, #PositiveVibes και #lifestyle_travel. Για την κατάλληλη επεξεργασία τους απαιτούνται τεχνικές προεπεξεργασίας, όπως η κανονικοποίηση των χαρακτήρων, η μετατροπή σε πεζά γράμματα, η τμηματοποίηση σε επιμέρους λέξεις και ο διαχωρισμός σύνθετων όρων.

Παράλληλα, τα δεδομένα των κοινωνικών δικτύων περιέχουν συχνά αργκό, συντομογραφίες, ορθογραφικές παραλλαγές και δημιουργικές εκφράσεις. Η μεγάλη ποικιλία των hashtags και η περιορισμένη επανάληψη πολλών από αυτά δυσκολεύουν την απευθείας πρόβλεψη κάθε μεμονωμένης ετικέτας. Για τον λόγο αυτό, στην παρούσα εργασία τα hashtags οργανώνονται σε ευρύτερες θεματικές κατηγορίες, ώστε να δημιουργηθεί ένα πιο συνεκτικό και διαχειρίσιμο πρόβλημα ταξινόμησης.

Μία ακόμη πρόκληση αφορά την ανισορροπία μεταξύ των θεματικών κατηγοριών. Ορισμένες κατηγορίες εμφανίζονται σε σημαντικά μεγαλύτερο αριθμό αναρτήσεων, ενώ άλλες εκπροσωπούνται από ελάχιστα δείγματα. Η ανισορροπία αυτή μπορεί να οδηγήσει τα μοντέλα στην ευνοϊκότερη πρόβλεψη των συχνότερων κατηγοριών και σε χαμηλότερη απόδοση στις σπανιότερες.

Στο πλαίσιο της εργασίας διερευνάται κατά πόσο χαρακτηριστικά αλληλεπίδρασης, όπως ο αριθμός των επισημάνσεων «μου αρέσει» και των αναδημοσιεύσεων, μπορούν να χρησιμοποιηθούν για την πρόβλεψη της κύριας θεματικής κατηγορίας των hashtags μιας ανάρτησης. Για τον σκοπό αυτό εφαρμόζονται και συγκρίνονται διαφορετικοί αλγόριθμοι Μηχανικής και Βαθιάς Μάθησης, όπως οι Random Forest, k-Nearest Neighbors, Support Vector Machine, LSTM και Bi-LSTM. Η απόδοσή τους αξιολογείται πειραματικά, ώστε να προσδιοριστεί ποιά προσέγγιση ανταποκρίνεται αποτελεσματικότερα στα χαρακτηριστικά και στους περιορισμούς του συγκεκριμένου συνόλου δεδομένων.

1.3 Σκοπός και στόχοι της εργασίας

Σκοπός της παρούσας πτυχιακής εργασίας είναι η μελέτη και η ανάπτυξη μεθόδων για την αυτόματη ταξινόμηση hashtags σε θεματικές κατηγορίες, αξιοποιώντας δεδομένα από αναρτήσεις σε μέσα κοινωνικής δικτύωσης. Η εργασία επικεντρώνεται στη διερεύνηση της σχέσης μεταξύ των hashtags και των χαρακτηριστικών που συνοδεύουν μια ανάρτηση, όπως ο αριθμός των επισημάνσεων «μου αρέσει», οι αναδημοσιεύσεις, η πλατφόρμα δημοσίευσης και το συναίσθημα του περιεχομένου.

Η μεγάλη ποικιλία και η περιορισμένη επανάληψη των hashtags καθιστούν δύσκολη την απευθείας πρόβλεψη κάθε μεμονωμένης ετικέτας. Για τον λόγο αυτό, προηγείται η προεπεξεργασία και η ομαδοποίησή τους σε ευρύτερες θεματικές κατηγορίες. Η διαδικασία αυτή περιλαμβάνει τον καθαρισμό των hashtags, την κανονικοποίησή τους και την αντιστοίχισή τους σε κατηγορίες με βάση ένα σύνολο λέξεων-κλειδιών. Για κάθε ανάρτηση προσδιορίζεται στη συνέχεια μία κύρια θεματική κατηγορία, η οποία χρησιμοποιείται ως μεταβλητή-στόχος στα μοντέλα ταξινόμησης.

Στο πλαίσιο της εργασίας υλοποιούνται και αξιολογούνται διαφορετικές προσεγγίσεις Μηχανικής Μάθησης και Βαθιάς Μάθησης. Ειδικότερα, εξετάζονται οι αλγόριθμοι Random Forest, k-Nearest Neighbors και Support Vector Machine, καθώς και τα νευρωνικά δίκτυα LSTM και Bidirectional LSTM. Μέσω της εφαρμογής διαφορετικών μοντέλων επιδιώκεται να διερευνηθεί κατά πόσο οι απλούστερες κλασικές μέθοδοι ή οι πιο σύνθετες νευρωνικές αρχιτεκτονικές μπορούν να αποτυπώσουν αποτελεσματικότερα τις σχέσεις που υπάρχουν στα δεδομένα.

Παράλληλα, εξετάζεται η αξιοποίηση διαφορετικών συνόλων χαρακτηριστικών. Αρχικά, χρησιμοποιούνται αριθμητικοί δείκτες αλληλεπίδρασης, όπως τα Likes και τα Retweets, ενώ σε επόμενες

προσεγγίσεις μπορούν να ενσωματωθούν πρόσθετες πληροφορίες, όπως η πλατφόρμα, το συναίσθημα και το κειμενικό περιεχόμενο των αναρτήσεων ή των hashtags. Με τον τρόπο αυτό διερευνάται κατά πόσο ο συνδυασμός ετερογενών χαρακτηριστικών μπορεί να βελτιώσει την απόδοση της ταξινόμησης.

Τελικός στόχος είναι η συγκριτική αξιολόγηση των μοντέλων με βάση μετρικές όπως η Accuracy, η Precision, η Recall και το F1-score, καθώς και η ανάλυση των σφαλμάτων τους μέσω πινάκων σύγχυσης. Η σύγκριση αυτή αποσκοπεί στον εντοπισμό της πιο κατάλληλης προσέγγισης για το συγκεκριμένο σύνολο δεδομένων και στην εξαγωγή συμπερασμάτων σχετικά με τους παράγοντες που επηρεάζουν την πρόβλεψη των θεματικών κατηγοριών των hashtags.

ΚΕΦΑΛΑΙΟ 2 Βιβλιογραφική Επισκόπηση

2.1 Μέσα κοινωνικής δικτύωσης

2.1.1 Ορισμός των μέσων κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης αποτελούν διαδικτυακές πλατφόρμες που επιτρέπουν στους χρήστες να δημιουργούν, να δημοσιεύουν και να ανταλλάσσουν περιεχόμενο, καθώς και να αλληλεπιδρούν μεταξύ τους σε πραγματικό χρόνο ή ασύγχρονα. Σε αντίθεση με τις παραδοσιακές ιστοσελίδες, όπου το περιεχόμενο παράγεται κυρίως από τους διαχειριστές τους, στα μέσα κοινωνικής δικτύωσης το μεγαλύτερο μέρος της πληροφορίας δημιουργείται από τους ίδιους τους χρήστες. Η δυνατότητα αυτή έχει μετατρέψει τους χρήστες από απλούς αποδέκτες πληροφοριών σε ενεργούς δημιουργούς και διαμοιραστές περιεχομένου.

Η ανάπτυξη των μέσων κοινωνικής δικτύωσης συνδέεται άμεσα με την εξέλιξη του Web 2.0, το οποίο εισήγαγε ένα πιο διαδραστικό μοντέλο λειτουργίας του διαδικτύου. Στο πλαίσιο αυτό, οι χρήστες απέκτησαν τη δυνατότητα να συμμετέχουν ενεργά στη δημιουργία και συνεχή ανανέωση του περιεχομένου μέσω ιστολογίων (blogs), διαδικτυακών κοινοτήτων, συνεργατικών εφαρμογών και πλατφορμών κοινωνικής δικτύωσης. Βασικό χαρακτηριστικό αυτής της εξέλιξης αποτελεί το περιεχόμενο που δημιουργείται από τους χρήστες (User Generated Content – UGC), το οποίο περιλαμβάνει δημοσιεύσεις, φωτογραφίες, βίντεο, σχόλια και κάθε μορφή ψηφιακού υλικού που παράγεται και κοινοποιείται από τους ίδιους.

Η ευρεία διάδοση των μέσων κοινωνικής δικτύωσης έχει αλλάξει σημαντικά τον τρόπο με τον οποίο οι άνθρωποι επικοινωνούν, ενημερώνονται και ανταλλάσσουν πληροφορίες. Παράλληλα, οι πλατφόρμες αυτές χρησιμοποιούνται πλέον όχι μόνο για προσωπική επικοινωνία αλλά και για επαγγελματική δικτύωση, εκπαίδευση, ενημέρωση, ψυχαγωγία, προώθηση προϊόντων και υπηρεσιών, καθώς και για την ανάπτυξη διαδικτυακών κοινοτήτων με κοινά ενδιαφέροντα.

2.1.2 Εξέλιξη των μέσων κοινωνικής δικτύωσης

Η εξέλιξη των μέσων κοινωνικής δικτύωσης συνδέεται στενά με τη συνεχή ανάπτυξη του διαδικτύου και των τεχνολογιών επικοινωνίας. Οι πρώτες μορφές διαδικτυακής αλληλεπίδρασης εμφανίστηκαν στα τέλη της δεκαετίας του 1970 μέσω του Usenet, ενός συστήματος ηλεκτρονικών ομάδων συζήτησης που επέτρεπε στους χρήστες να ανταλλάσσουν δημόσια μηνύματα. Στη συνέχεια, κατά τη δεκαετία του 1990, η εμφάνιση των ιστολογίων (blogs) και των πρώτων υπηρεσιών κοινωνικής δικτύωσης δημιούργησε νέες δυνατότητες για τη δημοσίευση προσωπικού περιεχομένου και την ανάπτυξη διαδικτυακών κοινοτήτων.

Η ουσιαστική ανάπτυξη των μέσων κοινωνικής δικτύωσης ξεκίνησε στις αρχές της δεκαετίας του 2000, όταν εμφανίστηκαν εξειδικευμένες πλατφόρμες όπως το Friendster, το LinkedIn και το MySpace. Οι πλατφόρμες αυτές έδωσαν στους χρήστες τη δυνατότητα να δημιουργούν προσωπικά προφίλ, να αναπτύσσουν κοινωνικά ή επαγγελματικά δίκτυα και να αλληλεπιδρούν ευκολότερα με άλλους χρήστες. Η συνεχής βελτίωση των διαδικτυακών υπηρεσιών και η αυξανόμενη πρόσβαση στο διαδίκτυο συνέβαλαν στη ραγδαία εξάπλωση αυτών των εφαρμογών.

Καθοριστικό σημείο στην εξέλιξη των μέσων κοινωνικής δικτύωσης αποτέλεσε η εμφάνιση του Facebook το 2004, το οποίο αρχικά απευθυνόταν σε φοιτητές πανεπιστημίων και στη συνέχεια επεκτάθηκε στο ευρύ κοινό. Ακολούθησαν πλατφόρμες όπως το YouTube, το Twitter και αργότερα το Instagram, οι οποίες εμπλούτισαν τις δυνατότητες δημιουργίας και διαμοιρασμού περιεχομένου μέσω εικόνων, βίντεο και σύντομων δημοσιεύσεων. Η συνεχής ανάπτυξη των κινητών συσκευών και των ευρυζωνικών δικτύων επέτρεψε την άμεση πρόσβαση στις πλατφόρμες αυτές ανά πάσα στιγμή και από οποιοδήποτε σημείο.

Σήμερα, τα μέσα κοινωνικής δικτύωσης αποτελούν αναπόσπαστο μέρος της καθημερινότητας εκατομμυρίων χρηστών παγκοσμίως. Παράλληλα με την προσωπική επικοινωνία, χρησιμοποιούνται ευρέως στην ενημέρωση, στην εκπαίδευση, στο ψηφιακό μάρκετινγκ, στην επιχειρηματικότητα και στην επιστημονική έρευνα. Η συνεχής εξέλιξη των τεχνολογιών και η εμφάνιση νέων πλατφορμών διαμορφώνουν διαρκώς νέους τρόπους αλληλεπίδρασης και παραγωγής ψηφιακού περιεχομένου.

2.1.3 Χαρακτηριστικά και σύγχρονος ρόλος των μέσων κοινωνικής δικτύωσης

Τα μέσα κοινωνικής δικτύωσης χαρακτηρίζονται από την ενεργή συμμετοχή των χρηστών στη δημιουργία και την ανταλλαγή περιεχομένου. Οι χρήστες δεν περιορίζονται μόνο στην προβολή πληροφοριών, αλλά μπορούν να δημοσιεύουν κείμενα, εικόνες και βίντεο, να σχολιάζουν αναρτήσεις, να αλληλεπιδρούν με άλλους χρήστες και να συμμετέχουν σε διαδικτυακές κοινότητες. Η συνεχής αυτή αλληλεπίδραση διαφοροποιεί τα μέσα κοινωνικής δικτύωσης από τις παραδοσιακές μορφές διαδικτυακής επικοινωνίας συμβάλλοντας στη δυναμική παραγωγή και διάδοση της πληροφορίας.

Παρόλο που όλες οι πλατφόρμες βασίζονται στην κοινωνική αλληλεπίδραση, καθεμία εξυπηρετεί διαφορετικές ανάγκες και μορφές επικοινωνίας. Ορισμένες δίνουν έμφαση στη δημιουργία προσωπικών σχέσεων και στην ανταλλαγή εμπειριών, άλλες στην επαγγελματική δικτύωση, ενώ άλλες επικεντρώνονται στη δημοσίευση φωτογραφιών, βίντεο ή σύντομων μηνυμάτων. Ως αποτέλεσμα, οι χρήστες προσαρμόζουν τον τρόπο χρήσης κάθε πλατφόρμας ανάλογα με τον σκοπό και τις δυνατότητές της.

Η συνεχής αύξηση του αριθμού των χρηστών και του παραγόμενου ψηφιακού περιεχομένου έχει καταστήσει τα μέσα κοινωνικής δικτύωσης σημαντική πηγή δεδομένων για την επιστημονική έρευνα. Οι αναρτήσεις, τα σχόλια και οι αλληλεπιδράσεις των χρηστών αξιοποιούνται σε μελέτες που αφορούν την ανθρώπινη συμπεριφορά, την επικοινωνία, την ανάλυση συναισθήματος, την εξαγωγή γνώσης και την ανάπτυξη εφαρμογών μηχανικής μάθησης. Έτσι, τα μέσα κοινωνικής δικτύωσης αποτελούν πλέον όχι μόνο εργαλεία επικοινωνίας, αλλά και πολύτιμη πηγή πληροφοριών για την ανάλυση μεγάλου όγκου δεδομένων.

Η συνεχής παραγωγή περιεχομένου στα μέσα κοινωνικής δικτύωσης συνοδεύεται από διάφορους μηχανισμούς οργάνωσης και κατηγοριοποίησης της πληροφορίας. Ένας από τους σημαντικότερους μηχανισμούς είναι τα hashtags, τα οποία χρησιμοποιούνται για την ομαδοποίηση δημοσιεύσεων, τη διευκόλυνση της αναζήτησης περιεχομένου και την ανάδειξη θεμάτων που συγκεντρώνουν αυξημένο ενδιαφέρον.

2.2 Hashtags και θεματική κατηγοριοποίηση

2.2.1 Ορισμός και λειτουργία των hashtags

Τα hashtags αποτελούν ένα χαρακτηριστικό στοιχείο της επικοινωνίας στα μέσα κοινωνικής δικτύωσης και χρησιμοποιούνται για την επισήμανση και την οργάνωση του περιεχομένου. Ένα hashtag σχηματίζεται με την προσθήκη του συμβόλου «#» πριν από μία λέξη ή μία σύντομη φράση, δημιουργώντας μια ετικέτα που μπορεί να συνοδεύει μια ανάρτηση. Η χρήση τους καθιερώθηκε αρχικά στο Twitter, με σκοπό τη συγκέντρωση αναρτήσεων που αναφέρονταν σε ένα κοινό θέμα και τη διευκόλυνση της αναζήτησής τους. Στη συνέχεια, τα hashtags υιοθετήθηκαν και από άλλες πλατφόρμες κοινωνικής δικτύωσης και ενσωματώθηκαν ευρύτερα στον τρόπο με τον οποίο οι χρήστες δημοσιεύουν και αναζητούν περιεχόμενο.

Η λειτουργία των hashtags δεν περιορίζεται μόνο στην απλή οργάνωση των αναρτήσεων. Μέσω αυτών, οι χρήστες μπορούν να συμμετέχουν σε συζητήσεις γύρω από συγκεκριμένα θέματα, να εκφράζουν απόψεις, να σχολιάζουν γεγονότα και να προωθούν προϊόντα ή εκδηλώσεις. Παράλληλα, η χρήση μιας κοινής ετικέτας μπορεί να φέρει σε επαφή χρήστες που ενδιαφέρονται για το ίδιο περιεχόμενο, συμβάλλοντας στη δημιουργία θεματικών κοινοτήτων. Επομένως, τα hashtags λειτουργούν τόσο ως ένας πρακτικός τρόπος οργάνωσης και αναζήτησης της πληροφορίας όσο και ως μέσο επικοινωνίας και αλληλεπίδρασης μεταξύ των χρηστών.

2.2.2 Hashtags ως δείκτες θεματικού περιεχομένου

Πέρα από τη γενικότερη χρήση τους στην οργάνωση του περιεχομένου, τα hashtags μπορούν να λειτουργήσουν ως ενδείξεις για το θέμα μιας ανάρτησης. Όταν πολλοί χρήστες χρησιμοποιούν το ίδιο hashtag σε αναρτήσεις που αφορούν ένα συγκεκριμένο γεγονός ή αντικείμενο, δημιουργείται ένα σύνολο περιεχομένου γύρω από μια κοινή θεματική. Ως εκ τούτου, τα θεματικά hashtags μπορούν να αξιοποιηθούν και κατά τη συλλογή δεδομένων από τα μέσα κοινωνικής δικτύωσης. Οι Bruns και Liang (2012), για παράδειγμα, χρησιμοποίησαν σχετικά hashtags για τον εντοπισμό και τη συγκέντρωση tweets που αφορούσαν συγκεκριμένες φυσικές καταστροφές και γεγονότα κρίσης.

Ωστόσο, η χρήση ενός hashtag δεν αποτελεί από μόνη της απόλυτη ένδειξη για το θέμα μιας ανάρτησης. Μία σχετική ανάρτηση μπορεί να μη διαθέτει καθόλου hashtag ή να χρησιμοποιεί διαφορετική ετικέτα από αυτή που χρησιμοποιείται ευρύτερα για το συγκεκριμένο θέμα. Επομένως, η συλλογή αναρτήσεων αποκλειστικά μέσω ενός hashtag ενδέχεται να μην περιλαμβάνει όλο το σχετικό περιεχόμενο. Παράλληλα, τα hashtags δεν χρησιμοποιούνται πάντοτε αποκλειστικά ως θεματικές ετικέτες. Η επιλογή τους μπορεί να σχετίζεται με τα προσωπικά ενδιαφέροντα των χρηστών, τη συμμετοχή τους σε ευρύτερες θεματικές κοινότητες ή άλλες μορφές διαδικτυακής συμπεριφοράς και επικοινωνίας.

Η παραπάνω ιδιαιτερότητα είναι σημαντική για τη θεματική ανάλυση του περιεχομένου των κοινωνικών δικτύων. Παρόλο που ένα hashtag μπορεί να παρέχει χρήσιμη πληροφορία σχετικά με το θέμα μιας ανάρτησης, η ερμηνεία του είναι προτιμότερο να μην βασίζεται αποκλειστικά στην ίδια την ετικέτα. Η συνεξέταση του κειμενικού περιεχομένου και άλλων διαθέσιμων χαρακτηριστικών μπορεί να προσφέρει περισσότερη πληροφορία για τη θεματική στην οποία ανήκει μια ανάρτηση. Η ανάγκη αυτή οδηγεί στη χρήση

υπολογιστικών μεθόδων για την αυτόματη θεματική κατηγοριοποίηση περιεχομένου.

2.2.3 Θεματική κατηγοριοποίηση περιεχομένου κοινωνικών δικτύων

Η μεγάλη ποσότητα περιεχομένου που παράγεται καθημερινά στα μέσα κοινωνικής δικτύωσης έχει δημιουργήσει την ανάγκη για αυτοματοποιημένες μεθόδους οργάνωσης και ταξινόμησης των αναρτήσεων σε θεματικές κατηγορίες. Η θεματική κατηγοριοποίηση επιδιώκει να αντιστοιχίσει μια ανάρτηση ή ένα σύνολο αναρτήσεων σε ένα ή περισσότερα θέματα, αξιοποιώντας πληροφορίες που προέρχονται από το κείμενο, τα hashtags και άλλα διαθέσιμα χαρακτηριστικά.

Στη σχετική βιβλιογραφία έχουν αναπτυχθεί διαφορετικές προσεγγίσεις για την ανάλυση και την πρόταση hashtags. Οι Amiri et al. (2021), μέσα από ανασκόπηση της σχετικής έρευνας, εντοπίζουν βασικές κατευθύνσεις όπως η ταξινόμηση, η θεματική μοντελοποίηση, οι μέθοδοι ομοιότητας, η μηχανική μάθηση και η βαθιά μάθηση. Οι προσεγγίσεις αυτές διαφέρουν ως προς τον τρόπο με τον οποίο αναπαριστούν το περιεχόμενο και ως προς το είδος της πληροφορίας που χρησιμοποιούν για να συνδέσουν μια ανάρτηση με ένα σχετικό θέμα ή hashtag.

Η χρήση τεχνικών μηχανικής και βαθιάς μάθησης επιτρέπει την αυτόματη εκμάθηση προτύπων από μεγάλο αριθμό παραδειγμάτων. Αντί να βασίζεται αποκλειστικά σε χειροκίνητους κανόνες, ένα μοντέλο μπορεί να εκπαιδευτεί πάνω σε ήδη κατηγοριοποιημένα δεδομένα και να χρησιμοποιήσει τα μοτίβα που εντοπίζει για να προβλέψει τη θεματική κατηγορία νέων αναρτήσεων. Παράλληλα, τεχνικές όπως το topic modeling μπορούν να αξιοποιηθούν για την αναγνώριση υποκείμενων θεμάτων χωρίς να απαιτείται πάντοτε προκαθορισμένη ετικέτα για κάθε δείγμα.

Η ύπαρξη πολλών διαφορετικών προσεγγίσεων δείχνει ότι δεν υπάρχει μία μοναδική μέθοδος που να είναι κατάλληλη για κάθε περίπτωση. Η επιλογή εξαρτάται από τη μορφή των δεδομένων, το πλήθος και την ποιότητα των διαθέσιμων δειγμάτων, τις θεματικές κατηγορίες που εξετάζονται και το αν υπάρχουν ήδη ετικέτες εκπαίδευσης. Κατά συνέπεια, η αξιολόγηση διαφορετικών μεθόδων αποτελεί σημαντικό βήμα για την επιλογή της καταλληλότερης προσέγγισης σε ένα συγκεκριμένο πρόβλημα θεματικής ταξινόμησης.

2.2.4 Αυτόματη κατηγοριοποίηση και πρόταση hashtags

Η αυτόματη κατηγοριοποίηση και πρόταση hashtags αποτελεί ένα πεδίο στο οποίο συνδυάζονται τεχνικές επεξεργασίας κειμένου, μηχανικής μάθησης και θεματικής ανάλυσης. Στόχος είναι ένα σύστημα να μπορεί να αναγνωρίσει το θέμα μιας ανάρτησης και να τη συσχετίσει με κατάλληλη θεματική κατηγορία ή με σχετικά hashtags, χωρίς να απαιτείται κάθε φορά χειροκίνητη παρέμβαση.

Ένα χαρακτηριστικό παράδειγμα αποτελεί η εργασία των Pattanayak et al. (2024), στην οποία χρησιμοποιήθηκαν 20.000 tweets καταναμεμένα σε τέσσερις θεματικές κατηγορίες: αθλητισμός, πολιτική, υγεία και τεχνολογία. Αρχικά εφαρμόστηκε καθαρισμός και tokenization του κειμένου, ενώ στη συνέχεια χρησιμοποιήθηκε το Word2Vec για την αριθμητική αναπαράσταση των λέξεων. Ακολούθησε μείωση της διάστασης των δεδομένων και ομαδοποίησή τους σε τέσσερα clusters, τα οποία συνδέθηκαν με τις αντίστοιχες θεματικές κατηγορίες μέσω ημι-επιβλεπόμενης προσέγγισης. Με

βάση τη θεματική κατηγορία που αναγνωρίζονταν για κάθε tweet, το σύστημα μπορούσε στη συνέχεια να προτείνει κατάλληλο hashtag.

Η συγκεκριμένη προσέγγιση είναι σημαντική γιατί δείχνει ότι η σχέση μεταξύ μιας ανάρτησης και ενός hashtag μπορεί να μοντελοποιηθεί υπολογιστικά μέσω του ίδιου του κειμενικού περιεχομένου. Ακόμη, η αξιολόγηση με μετρικές όπως Precision, Recall και F1-score επιτρέπει τη σύγκριση διαφορετικών μεθόδων και την εκτίμηση της ποιότητας της κατηγοριοποίησης.

Σε ευρύτερο επίπεδο, η σχετική βιβλιογραφία δείχνει ότι η αυτόματη επεξεργασία hashtags δεν περιορίζεται σε μία μόνο τεχνική. Αντίθετα, έχουν χρησιμοποιηθεί μέθοδοι ταξινόμησης, θεματικής μοντελοποίησης, ομοιότητας, μηχανικής μάθησης και βαθιάς μάθησης, ανάλογα με το είδος των δεδομένων και τον στόχο της εφαρμογής.

Η αξιοποίηση τέτοιων τεχνικών είναι ιδιαίτερα σημαντική στην περίπτωση που ο όγκος των αναρτήσεων είναι μεγάλος και η χειροκίνητη κατηγοριοποίηση είναι πρακτικά ανέφικτη. Μέσω της αυτόματης ανάλυσης είναι δυνατό να οργανωθεί το περιεχόμενο σε θεματικές ομάδες, να εντοπιστούν συσχετίσεις μεταξύ αναρτήσεων και hashtags και να δημιουργηθούν συστήματα που μπορούν να επεξεργάζονται νέες αναρτήσεις με βάση πρότυπα που έχουν μάθει από προηγούμενα δεδομένα.

2.3 Επεξεργασία Φυσικής Γλώσσας

2.3.1 Εισαγωγή στην Επεξεργασία Φυσικής Γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP) αποτελεί πεδίο της Τεχνητής Νοημοσύνης που ασχολείται με την επεξεργασία και την ανάλυση της ανθρώπινης γλώσσας από υπολογιστικά συστήματα. Στόχος της είναι η ανάπτυξη μεθόδων που επιτρέπουν στους υπολογιστές να επεξεργάζονται κειμενικά δεδομένα εξάγοντας χρήσιμη πληροφορία από αυτά. Το NLP χρησιμοποιείται σε ένα μεγάλο εύρος εφαρμογών, όπως η ανάκτηση πληροφοριών, η ανάλυση συναισθήματος και γενικότερα η ανάλυση και ταξινόμηση κειμένου.

Οι πρώτες προσεγγίσεις στην Επεξεργασία Φυσικής Γλώσσας βασίζονταν σε μεγάλο βαθμό σε προκαθορισμένους γλωσσικούς κανόνες. Η ανάπτυξη της μηχανικής μάθησης οδήγησε σταδιακά στη χρήση μεθόδων που μπορούσαν να αξιοποιήσουν δεδομένα για την εκτέλεση εργασιών επεξεργασίας και ταξινόμησης κειμένου. Στη συνέχεια, η ανάπτυξη της βαθιάς μάθησης επέτρεψε τη χρήση νευρωνικών δικτύων για την αυτόματη εκμάθηση χαρακτηριστικών από τα δεδομένα, περιορίζοντας την ανάγκη για τον χειροκίνητο καθορισμό όλων των χαρακτηριστικών που απαιτούνται για την ανάλυση του κειμένου.

Η εξέλιξη αυτή έχει ιδιαίτερη σημασία για την επεξεργασία μεγάλου όγκου κειμενικών δεδομένων, όπως αυτά που παράγονται στα μέσα κοινωνικής δικτύωσης. Οι αναρτήσεις των χρηστών μπορούν να αποτελέσουν αντικείμενο αυτόματης ανάλυσης και ταξινόμησης, με τεχνικές NLP να χρησιμοποιούνται για την προετοιμασία του κειμένου πριν από την εφαρμογή μοντέλων μηχανικής ή βαθιάς μάθησης.

2.3.2 Προεπεξεργασία κειμένου

Πριν από την εφαρμογή ενός μοντέλου μηχανικής ή βαθιάς μάθησης, τα κειμενικά δεδομένα συνήθως περνούν από ένα στάδιο προεπεξεργασίας. Στόχος αυτού του σταδίου είναι η μετατροπή του αρχικού κειμένου σε μια πιο κατάλληλη και οργανωμένη μορφή, ώστε να μπορεί να αξιοποιηθεί αποτελεσματικότερα από τα επόμενα στάδια της ανάλυσης. Η διαδικασία αυτή είναι ιδιαίτερα σημαντική στα δεδομένα που προέρχονται από μέσα κοινωνικής δικτύωσης, καθώς το περιεχόμενό τους μπορεί να περιλαμβάνει σύντομα κείμενα, hashtags και διαφορετικούς τρόπους γραφής.

Μία από τις βασικές τεχνικές προεπεξεργασίας είναι το tokenization, κατά το οποίο το κείμενο διαχωρίζεται σε μικρότερες μονάδες, τα λεγόμενα tokens. Τα tokens μπορούν να αντιστοιχούν, για παράδειγμα, σε λέξεις ή άλλα επιμέρους στοιχεία του κειμένου και να αποτελούν τη βάση για την περαιτέρω επεξεργασία του. Άλλες συνηθισμένες τεχνικές είναι η αφαίρεση συχνών λέξεων που μπορεί να προσφέρουν περιορισμένη πληροφορία για μια συγκεκριμένη εργασία (stop-word removal) και το stemming, σύμφωνα με το οποίο διαφορετικές μορφές μιας λέξης περιορίζονται σε μια κοινή βασική μορφή. Τέτοιες τεχνικές χρησιμοποιούνται για την προετοιμασία δεδομένων κοινωνικών δικτύων πριν από την εφαρμογή μοντέλων ταξινόμησης.

Ωστόσο, τα βήματα προεπεξεργασίας δεν είναι απαραίτητα ίδια σε κάθε εφαρμογή. Η επιλογή τους εξαρτάται από τη μορφή του κειμένου, τον στόχο της ανάλυσης και το μοντέλο που πρόκειται να χρησιμοποιηθεί. Επομένως, η προεπεξεργασία αποτελεί μέρος της συνολικής διαδικασίας προετοιμασίας των δεδομένων και πρέπει να προσαρμόζεται στα χαρακτηριστικά του εκάστοτε προβλήματος.

2.3.3 Αριθμητική αναπαράσταση κειμένου και embeddings

Τα μοντέλα μηχανικής και βαθιάς μάθησης δεν μπορούν να επεξεργαστούν άμεσα λέξεις και προτάσεις στη φυσική τους μορφή. Εξαιτίας αυτού, μετά την προεπεξεργασία του κειμένου είναι απαραίτητη η μετατροπή του σε αριθμητική αναπαράσταση, η οποία μπορεί να χρησιμοποιηθεί ως είσοδος από το μοντέλο. Ένας διαδεδομένος τρόπος για την πραγματοποίηση αυτής της μετατροπής είναι η χρήση των embeddings.

Τα text embeddings αποτελούν αριθμητικές αναπαραστάσεις κειμενικών στοιχείων, όπως λέξεων, φράσεων ή ακόμη και ολόκληρων κειμένων, με τη μορφή διανυσμάτων πραγματικών αριθμών. Σε αντίθεση με μια απλή αντιστοίχιση κάθε λέξης σε έναν αριθμό, τα embeddings έχουν ως στόχο να αναπαραστήσουν πληροφορίες σχετικά με το νόημα και τις σχέσεις μεταξύ των διαφορετικών στοιχείων του κειμένου. Με αυτόν τον τρόπο, κειμενικά στοιχεία που παρουσιάζουν σημασιολογική ομοιότητα μπορούν να έχουν και παρόμοιες αναπαραστάσεις στον διανυσματικό χώρο.

Η αριθμητική αναπαράσταση του κειμένου αποτελεί σημαντικό στάδιο σε προβλήματα ταξινόμησης, καθώς δημιουργεί μια μορφή δεδομένων την οποία μπορούν να επεξεργαστούν οι αλγόριθμοι μηχανικής και βαθιάς μάθησης. Τα embeddings μπορούν επομένως να λειτουργήσουν ως σύνδεση μεταξύ του αρχικού κειμενικού περιεχομένου και του μοντέλου που αναλαμβάνει την περαιτέρω ανάλυση ή ταξινόμησή του.

2.3.4 Επεξεργασία Φυσικής Γλώσσας στην ανάλυση περιεχομένου κοινωνικών δικτύων

Τα μέσα κοινωνικής δικτύωσης αποτελούν σημαντική πηγή κειμενικών δεδομένων, καθώς οι χρήστες δημοσιεύουν συνεχώς περιεχόμενο σχετικό με απόψεις, ενδιαφέροντα και γεγονότα. Η μεγάλη ποσότητα αυτού του περιεχομένου δημιουργεί την ανάγκη για αυτοματοποιημένες μεθόδους επεξεργασίας και ανάλυσης. Η Επεξεργασία Φυσικής Γλώσσας μπορεί να χρησιμοποιηθεί για την προετοιμασία και την αξιοποίηση τέτοιων δεδομένων σε εργασίες όπως η ανάλυση συναισθήματος και η ταξινόμηση κειμένου.

Χαρακτηριστικό παράδειγμα αποτελεί η ανάλυση δεδομένων από το Twitter. Οι Surekha et al. (2025) εφαρμόζουν τεχνικές NLP για την προεπεξεργασία αναρτήσεων και στη συνέχεια αξιοποιούν μοντέλο που περιλαμβάνει Bi-LSTM για την ανάλυση και ταξινόμηση του περιεχομένου τους. Η προσέγγιση αυτή δείχνει πώς η επεξεργασία φυσικής γλώσσας μπορεί να συνδυαστεί με τεχνικές βαθιάς μάθησης για την εξαγωγή πληροφορίας από μη δομημένα δεδομένα κοινωνικών δικτύων.

Η αξιοποίηση του κειμένου σε τέτοιες εφαρμογές απαιτεί επομένως μια σειρά από στάδια, τα οποία ξεκινούν από την κατάλληλη προεπεξεργασία του και συνεχίζονται με τη μετατροπή του σε αριθμητική μορφή που μπορεί να χρησιμοποιηθεί από υπολογιστικά μοντέλα. Τα embeddings αποτελούν έναν τρόπο πραγματοποίησης αυτής της αναπαράστασης, επιτρέποντας την αξιοποίηση σημασιολογικών πληροφοριών του κειμένου σε εργασίες ταξινόμησης. Με αυτόν τον τρόπο, οι τεχνικές NLP αποτελούν βασικό σύνδεσμο μεταξύ του αρχικού κειμενικού περιεχομένου και των μοντέλων μηχανικής ή βαθιάς μάθησης που χρησιμοποιούνται για την ανάλυσή του.

2.4 Μηχανική Μάθηση και Ταξινόμηση

2.4.1 Εισαγωγή στη Μηχανική Μάθηση

Η Μηχανική Μάθηση (Machine Learning – ML) αποτελεί κλάδο της Τεχνητής Νοημοσύνης που βασίζεται στην ανάπτυξη υπολογιστικών συστημάτων τα οποία μπορούν να μαθαίνουν από δεδομένα. Αντί η συμπεριφορά ενός συστήματος να καθορίζεται αποκλειστικά από ένα σύνολο προκαθορισμένων κανόνων, ένα μοντέλο μηχανικής μάθησης εκπαιδεύεται χρησιμοποιώντας διαθέσιμα δεδομένα, ώστε να αναγνωρίζει μοτίβα και να αξιοποιεί τη γνώση που αποκτά για την πραγματοποίηση προβλέψεων σε νέα δεδομένα.

Οι μέθοδοι μηχανικής μάθησης μπορούν να διακριθούν σε διαφορετικές κατηγορίες ανάλογα με τον τρόπο με τον οποίο πραγματοποιείται η εκπαίδευση. Δύο βασικές κατηγορίες είναι η επιβλεπόμενη (supervised learning) και η μη επιβλεπόμενη μάθηση (unsupervised learning). Στην επιβλεπόμενη μάθηση, το μοντέλο εκπαιδεύεται με δεδομένα για τα οποία είναι γνωστή η επιθυμητή έξοδος. Με αυτόν τον τρόπο προσπαθεί να μάθει τη σχέση μεταξύ των δεδομένων εισόδου και της αντίστοιχης εξόδου, ώστε στη συνέχεια να μπορεί να πραγματοποιεί προβλέψεις για νέα δεδομένα. Αντίθετα, στη μη επιβλεπόμενη μάθηση δεν παρέχονται προκαθορισμένες σωστές εξόδους και ο αλγόριθμος επιχειρεί να εντοπίσει δομές ή μοτίβα που υπάρχουν στα ίδια τα δεδομένα.

Η Μηχανική Μάθηση έχει ιδιαίτερη σημασία στην επεξεργασία μεγάλου όγκου κειμενικών δεδομένων. Η συνεχής παραγωγή ψηφιακού περιεχομένου,

όπως ειδήσεις, κριτικές, αναφορές και αναρτήσεις σε μέσα κοινωνικής δικτύωσης, έχει δημιουργήσει την ανάγκη για αυτοματοποιημένες μεθόδους ανάλυσης και οργάνωσης του κειμένου. Στο πλαίσιο αυτό, τεχνικές μηχανικής μάθησης χρησιμοποιούνται σε προβλήματα Επεξεργασίας Φυσικής Γλώσσας και ιδιαίτερα στην ταξινόμηση κειμένου, επιτρέποντας την επεξεργασία μεγάλων συνόλων δεδομένων που θα ήταν δύσκολο να κατηγοριοποιηθούν αποκλειστικά με χειροκίνητο τρόπο.

2.4.2 Επιβλεπόμενη μάθηση και ταξινόμηση

Η ταξινόμηση (classification) αποτελεί μία από τις βασικές εφαρμογές της επιβλεπόμενης μάθησης. Στην περίπτωση αυτή, το μοντέλο εκπαιδεύεται χρησιμοποιώντας ένα σύνολο δεδομένων στο οποίο κάθε παράδειγμα συνοδεύεται από τη σωστή κατηγορία ή ετικέτα (label). Κατά την εκπαίδευση, το μοντέλο προσπαθεί να μάθει τη σχέση μεταξύ των χαρακτηριστικών των δεδομένων εισόδου και των αντίστοιχων κατηγοριών, με στόχο να μπορεί στη συνέχεια να προβλέψει την κατηγορία στην οποία ανήκει ένα νέο παράδειγμα.

Η ύπαρξη επισημασμένων δεδομένων αποτελεί βασικό χαρακτηριστικό της επιβλεπόμενης μάθησης. Ο Bell (2022) παρουσιάζει χαρακτηριστικά την περίπτωση ταξινόμησης αναρτήσεων του Twitter, κατά την οποία κάθε ανάρτηση αποτελεί δεδομένο εισόδου και συνοδεύεται από μια γνωστή θεματική κατηγορία, όπως μουσική, ένδυση ή δεδομένα. Ένα σύνολο τέτοιων παραδειγμάτων μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός ταξινομητή, ο οποίος στη συνέχεια καλείται να προσδιορίσει την κατάλληλη κατηγορία για νέες αναρτήσεις.

Η διαδικασία αυτή επιτρέπει την αυτοματοποίηση εργασιών που διαφορετικά θα απαιτούσαν χειροκίνητη κατηγοριοποίηση μεγάλου αριθμού δεδομένων. Η αποτελεσματικότητα του μοντέλου εξαρτάται, μεταξύ άλλων, από τα δεδομένα με τα οποία εκπαιδεύεται και από το κατά πόσο μπορεί να εφαρμόσει όσα έμαθε σε δεδομένα που δεν έχει συναντήσει κατά την εκπαίδευση. Κατ'επέκταση, ο στόχος δεν είναι απλώς η σωστή κατηγοριοποίηση των παραδειγμάτων εκπαίδευσης, αλλά η δυνατότητα πραγματοποίησης αξιόπιστων προβλέψεων και σε νέα δεδομένα.

2.4.3 Ταξινόμηση κειμένου

Η ταξινόμηση κειμένου (text classification) αποτελεί διαδικασία κατά την οποία ένα κειμενικό δείγμα αντιστοιχίζεται σε μία ή περισσότερες προκαθορισμένες κατηγορίες. Η ανάγκη για αυτόματη ταξινόμηση έχει αυξηθεί σημαντικά λόγω του μεγάλου όγκου ψηφιακού περιεχομένου που παράγεται καθημερινά, όπως ειδήσεις, αναφορές, κριτικές και αναρτήσεις σε μέσα κοινωνικής δικτύωσης. Η χειροκίνητη κατηγοριοποίηση τέτοιου όγκου δεδομένων είναι ιδιαίτερα χρονοβόρα, γεγονός που έχει ενισχύσει τη χρήση αυτοματοποιημένων μεθόδων βασισμένων στη Μηχανική Μάθηση.

Οι πρώτες μέθοδοι ταξινόμησης κειμένου βασίζονταν σε προκαθορισμένους κανόνες και χειροκίνητα επιλεγμένα χαρακτηριστικά. Με την ανάπτυξη της Μηχανικής Μάθησης, η διαδικασία μετατοπίστηκε προς μοντέλα που εκπαιδεύονται απευθείας από δεδομένα και μπορούν να αναγνωρίζουν μοτίβα που σχετίζονται με τις διαφορετικές κατηγορίες. Στη βιβλιογραφία έχουν χρησιμοποιηθεί κλασικοί αλγόριθμοι, όπως ο k-Nearest Neighbors (kNN) και οι Support Vector Machines (SVM), για προβλήματα αυτόματης ταξινόμησης κειμένου.

Για να χρησιμοποιηθεί ένα κείμενο ως είσοδος σε έναν αλγόριθμο ταξινόμησης, πρέπει προηγουμένως να έχει μετατραπεί σε κατάλληλη αριθμητική αναπαράσταση. Για τον λόγο αυτό, η ταξινόμηση κειμένου συνδέεται άμεσα με τις τεχνικές Επεξεργασίας Φυσικής Γλώσσας που παρουσιάστηκαν στην προηγούμενη ενότητα, όπως η προεπεξεργασία, το tokenization και η αναπαράσταση του κειμένου σε αριθμητική μορφή. Μετά την προετοιμασία των δεδομένων, το μοντέλο μπορεί να εκπαιδευτεί ώστε να αντιστοιχίζει νέα κείμενα στις κατάλληλες κατηγορίες.

Η ταξινόμηση κειμένου βρίσκει εφαρμογή σε πολλές περιοχές, όπως η ανάλυση συναισθήματος, η κατηγοριοποίηση εγγράφων και η ανάλυση αναρτήσεων κοινωνικών δικτύων. Ειδικά στα μέσα κοινωνικής δικτύωσης, η δυνατότητα αυτόματης ταξινόμησης επιτρέπει την οργάνωση μεγάλου αριθμού αναρτήσεων σύμφωνα με το περιεχόμενο ή τη θεματική τους, γεγονός που την καθιστά ιδιαίτερα χρήσιμη σε εφαρμογές ανάλυσης ψηφιακού περιεχομένου.

2.5 Random Forest

2.5.1 Βασική αρχή και λειτουργία του Random Forest

Ο Random Forest είναι ένας αλγόριθμος Μηχανικής Μάθησης που βασίζεται στον συνδυασμό πολλών δέντρων απόφασης (decision trees). Αντί να στηρίζεται στην πρόβλεψη ενός μόνο δέντρου, δημιουργεί ένα σύνολο διαφορετικών δέντρων και συνδυάζει τα αποτελέσματά τους για να καταλήξει στην τελική πρόβλεψη. Η βασική ιδέα είναι ότι ο συνδυασμός πολλών διαφορετικών δέντρων μπορεί να οδηγήσει σε πιο αξιόπιστες προβλέψεις από τη χρήση ενός μεμονωμένου δέντρου.

Η διαφοροποίηση μεταξύ των δέντρων επιτυγχάνεται με την εισαγωγή τυχαιότητας κατά την εκπαίδευσή τους. Συγκεκριμένα, κάθε δέντρο εκπαιδεύεται χρησιμοποιώντας ένα τυχαίο δείγμα των διαθέσιμων δεδομένων εκπαίδευσης, ενώ κατά τη δημιουργία των κόμβων του εξετάζεται μόνο ένα τυχαία επιλεγμένο υποσύνολο των διαθέσιμων χαρακτηριστικών. Με αυτόν τον τρόπο, τα επιμέρους δέντρα δεν κατασκευάζονται με ακριβώς τα ίδια δεδομένα και χαρακτηριστικά και, συνεπώς, μειώνεται η μεταξύ τους συσχέτιση.

Αφού δημιουργηθούν τα επιμέρους δέντρα, οι προβλέψεις τους συνδυάζονται ώστε να προκύψει το τελικό αποτέλεσμα του Random Forest. Στα προβλήματα ταξινόμησης, κάθε δέντρο προβλέπει μία κατηγορία και η τελική απόφαση βασίζεται στην κατηγορία που συγκεντρώνει τις περισσότερες ψήφους. Αντίστοιχα, σε προβλήματα παλινδρόμησης, όπου το ζητούμενο είναι η πρόβλεψη μιας αριθμητικής τιμής, μπορούν να συνδυαστούν οι επιμέρους αριθμητικές προβλέψεις των δέντρων.

2.5.2 Random Forest στην ταξινόμηση

Στα προβλήματα ταξινόμησης, ο Random Forest χρησιμοποιεί τα επιμέρους δέντρα απόφασης ως ανεξάρτητους ταξινομητές. Κάθε δέντρο επεξεργάζεται τα χαρακτηριστικά ενός νέου δείγματος και καταλήγει σε μία πρόβλεψη για την κατηγορία στην οποία θεωρεί ότι ανήκει. Οι προβλέψεις όλων των δέντρων συνδυάζονται και η τελική κατηγορία καθορίζεται με βάση την πλειοψηφία των ψήφων τους (majority voting).

Για παράδειγμα, αν ένα Random Forest αποτελείται από 100 δέντρα και τα 70 ταξινομήσουν ένα δείγμα στην κατηγορία A, ενώ τα υπόλοιπα 30 στην κατηγορία B, η τελική πρόβλεψη του μοντέλου θα είναι η κατηγορία A. Άρα, η απόφαση δεν εξαρτάται από ένα μόνο δέντρο, αλλά από τη συλλογική απόφαση του συνόλου. Η προσέγγιση αυτή περιορίζει την επίδραση που μπορεί να έχει μια λανθασμένη πρόβλεψη από κάποιο μεμονωμένο δέντρο στο τελικό αποτέλεσμα.

Σημαντικό ρόλο στην αποτελεσματικότητα της ταξινόμησης παίζει η διαφοροποίηση μεταξύ των δέντρων. Ο Breiman (2001) συνδέει την απόδοση του Random Forest τόσο με την ικανότητα των επιμέρους δέντρων να πραγματοποιούν σωστές προβλέψεις όσο και με τον βαθμό συσχέτισης μεταξύ τους. Έτσι, στόχος είναι να δημιουργούνται δέντρα που είναι αποτελεσματικά ως ταξινομητές, χωρίς όμως να είναι υπερβολικά όμοια μεταξύ τους. Η τυχαία επιλογή δειγμάτων και χαρακτηριστικών κατά την εκπαίδευση συμβάλλει σε αυτή τη διαφοροποίηση.

2.5.3 Πλεονεκτήματα και περιορισμοί του Random Forest

Ο Random Forest παρουσιάζει αρκετά πλεονεκτήματα που τον καθιστούν κατάλληλο για προβλήματα ταξινόμησης. Ένα από τα σημαντικότερα είναι ότι συνδυάζει τις προβλέψεις πολλών δέντρων απόφασης, γεγονός που συνήθως οδηγεί σε πιο σταθερό αποτέλεσμα σε σχέση με τη χρήση ενός μόνο δέντρου. Η τυχαία επιλογή δειγμάτων και χαρακτηριστικών συμβάλλει στη δημιουργία διαφορετικών δέντρων και περιορίζει την πιθανότητα όλα τα δέντρα να συμπέσουν στα ίδια λάθη. Ο Breiman (2001) επισημαίνει επίσης ότι η αύξηση του αριθμού των δέντρων δεν οδηγεί απαραίτητα σε συνεχή υπερπροσαρμογή, καθώς το σφάλμα γενίκευσης τείνει να σταθεροποιείται.

Επιπλέον, ο Random Forest μπορεί να χρησιμοποιηθεί τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα παλινδρόμησης και μπορεί να διαχειριστεί δεδομένα με μεγάλο αριθμό χαρακτηριστικών. Παρέχει επίσης τη δυνατότητα εκτίμησης της σημαντικότητας των χαρακτηριστικών, γεγονός που μπορεί να βοηθήσει στην κατανόηση του ποια χαρακτηριστικά επηρεάζουν περισσότερο τις προβλέψεις του μοντέλου. Η δομή του επιτρέπει ακόμη την παράλληλη εκτέλεση των επιμέρους δέντρων, κάτι που μπορεί να διευκολύνει την επεξεργασία μεγαλύτερων συνόλων δεδομένων.

Παρά τα πλεονεκτήματά του, ο Random Forest παρουσιάζει και ορισμένους περιορισμούς. Η χρήση μεγάλου αριθμού δέντρων αυξάνει τις απαιτήσεις σε υπολογιστικούς πόρους και μπορεί να κάνει το μοντέλο πιο σύνθετο σε σχέση με ένα μεμονωμένο δέντρο απόφασης. Επίσης, η ερμηνεία της συνολικής λειτουργίας ενός μεγάλου συνόλου δέντρων είναι δυσκολότερη, καθώς η τελική πρόβλεψη προκύπτει από τον συνδυασμό πολλών διαφορετικών αποφάσεων. Ως εκ τούτου, αν και ο Random Forest μπορεί να προσφέρει καλή προβλεπτική απόδοση, δεν έχει την ίδια άμεση ερμηνευσιμότητα με ένα απλό δέντρο απόφασης.

2.6 k-Nearest Neighbors (kNN)

2.6.1 Βασική αρχή και λειτουργία του k-Nearest Neighbors

Ο αλγόριθμος k-Nearest Neighbors (kNN) αποτελεί μία μέθοδο Μηχανικής Μάθησης που μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης. Η λειτουργία του

βασίζεται στην υπόθεση ότι δείγματα που παρουσιάζουν παρόμοια χαρακτηριστικά βρίσκονται κοντά μεταξύ τους στον χώρο των δεδομένων. Έτσι, για την ταξινόμηση ενός νέου δείγματος, ο αλγόριθμος εξετάζει τα ήδη γνωστά δείγματα εκπαίδευσης που βρίσκονται πλησιέστερα σε αυτό.

Η παράμετρος καθορίζει τον αριθμό των πλησιέστερων γειτόνων που θα ληφθούν υπόψη για την πρόβλεψη. Αφού εντοπιστούν οι κοντινότεροι γείτονες ενός νέου δείγματος, εξετάζονται οι κατηγορίες στις οποίες ανήκουν. Στην περίπτωση της ταξινόμησης, η συνηθέστερη προσέγγιση είναι ο κανόνας της πλειοψηφίας (majority rule), σύμφωνα με τον οποίο το νέο δείγμα τοποθετείται στην κατηγορία που εμφανίζεται συχνότερα μεταξύ των επιλεγμένων γειτόνων.

Ένα ιδιαίτερο χαρακτηριστικό του kNN είναι ότι θεωρείται μέθοδος lazy learning. Σε αντίθεση με άλλους αλγορίθμους, δεν δημιουργεί κατά την εκπαίδευση ένα σύνθετο μοντέλο που στη συνέχεια χρησιμοποιείται για τις προβλέψεις. Τα δεδομένα εκπαίδευσης διατηρούνται και η αναζήτηση των κατάλληλων γειτόνων πραγματοποιείται όταν πρέπει να ταξινομηθεί ένα νέο δείγμα. Συνακόλουθα, σημαντικό μέρος της υπολογιστικής διαδικασίας μεταφέρεται στο στάδιο της πρόβλεψης.

2.6.2 Απόσταση, επιλογή του k και κανονικοποίηση των δεδομένων

Η λειτουργία του kNN βασίζεται στον υπολογισμό της απόστασης μεταξύ των δειγμάτων, ώστε να εντοπιστούν εκείνα που βρίσκονται πιο κοντά στο νέο δείγμα που θέλουμε να ταξινομήσουμε. Μία από τις συνηθέστερες μεθόδους για τον σκοπό αυτό είναι η Ευκλείδεια απόσταση (Euclidean distance), η οποία εκφράζει την απόσταση μεταξύ δύο σημείων με βάση τις τιμές των χαρακτηριστικών τους. Για δύο δείγματα και με χαρακτηριστικά, υπολογίζεται ως:

$$d(x, y) = \sqrt{\left\{ \sum_{i=1}^n (x_i - y_i)^2 \right\}} \quad (2.1)$$

Με βάση τις αποστάσεις που υπολογίζονται, επιλέγονται οι πλησιέστεροι γείτονες. Η επιλογή της τιμής του είναι σημαντική, καθώς καθορίζει πόσα δείγματα θα συμμετέχουν στην τελική απόφαση του αλγορίθμου. Δεν υπάρχει μία συγκεκριμένη τιμή του που να είναι κατάλληλη για όλα τα σύνολα δεδομένων, καθώς η κατανομή και τα χαρακτηριστικά των δεδομένων μπορεί να διαφέρουν. Για τον προσδιορισμό μιας κατάλληλης τιμής μπορεί να χρησιμοποιηθεί η διασταυρούμενη επικύρωση (cross-validation), κατά την οποία εξετάζονται διαφορετικές τιμές του και αξιολογείται η απόδοση του αλγορίθμου.

Επειδή ο kNN βασίζεται στις αποστάσεις μεταξύ των χαρακτηριστικών, σημαντικό ρόλο παίζει και η κλίμακα των δεδομένων. Όταν διαφορετικά χαρακτηριστικά έχουν πολύ διαφορετικά εύρη τιμών, εκείνα με τις μεγαλύτερες αριθμητικές τιμές μπορούν να επηρεάζουν περισσότερο τον υπολογισμό της απόστασης. Εξαιτίας αυτού, μπορεί να προηγείται κανονικοποίηση (normalization), ώστε τα χαρακτηριστικά να μεταφέρονται σε συγκρίσιμες κλίμακες. Οι Pandey και Jain (2017), για παράδειγμα, εξετάζουν τεχνικές όπως η κανονικοποίηση Min-Max και η τυποποίηση με Z-score στο πλαίσιο εφαρμογής του kNN.

2.6.3 Πλεονεκτήματα και περιορισμοί του kNN

Ο kNN είναι ένας απλός και εύκολα εφαρμόσιμος αλγόριθμος, καθώς δεν απαιτεί σύνθετη διαδικασία εκπαίδευσης και μπορεί να χρησιμοποιηθεί τόσο για ταξινόμηση όσο και για παλινδρόμηση. Η λειτουργία του βασίζεται άμεσα στα διαθέσιμα δεδομένα εκπαίδευσης, γεγονός που τον καθιστά ιδιαίτερα κατανοητό και χρήσιμο ως βασικό μοντέλο σύγκρισης.

Ωστόσο, η απόδοσή του επηρεάζεται σημαντικά από την επιλογή της τιμής k , τη συνάρτηση απόστασης και την κλίμακα των χαρακτηριστικών. Επιπρόσθετα, επειδή η αναζήτηση των πλησιέστερων γειτόνων πραγματοποιείται κατά το στάδιο της πρόβλεψης, η διαδικασία μπορεί να καταστεί υπολογιστικά απαιτητική όταν το σύνολο εκπαίδευσης είναι μεγάλο. Ένας ακόμη περιορισμός είναι ότι σε δεδομένα με ανισόρροπες κλάσεις, ο κανόνας της πλειοψηφίας μπορεί να ευνοήσει τις περισσότερες εκπροσωπούμενες κατηγορίες.

2.7 Support Vector Machine (SVM)

2.7.1 Βασική αρχή και λειτουργία του Support Vector Machine

Ο αλγόριθμος Support Vector Machine (SVM) αποτελεί μία μέθοδο επιβλεπόμενης Μηχανικής Μάθησης, η οποία χρησιμοποιείται μεταξύ άλλων για προβλήματα ταξινόμησης. Βασικός στόχος του είναι ο διαχωρισμός των δεδομένων διαφορετικών κατηγοριών μέσω ενός κατάλληλου ορίου απόφασης. Στην απλούστερη περίπτωση δύο χαρακτηριστικών, το όριο αυτό μπορεί να αναπαρασταθεί ως μία ευθεία, ενώ σε χώρους περισσότερων διαστάσεων περιγράφεται ως υπερεπίπεδο (hyperplane).

Το SVM δεν επιλέγει απλώς ένα όριο που διαχωρίζει τις κατηγορίες, αλλά αναζητά εκείνο που δημιουργεί το μεγαλύτερο δυνατό περιθώριο (margin) μεταξύ τους. Το περιθώριο αντιστοιχεί στην απόσταση μεταξύ του ορίου απόφασης και των πλησιέστερων σημείων των διαφορετικών κατηγοριών. Με τη μεγιστοποίηση αυτού του περιθωρίου επιδιώκεται η δημιουργία ενός διαχωρισμού που μπορεί να γενικεύεται καλύτερα και σε νέα δεδομένα που δεν συμμετείχαν στην εκπαίδευση του μοντέλου.

Ιδιαίτερη σημασία στη διαμόρφωση του ορίου απόφασης έχουν τα σημεία που βρίσκονται πλησιέστερα σε αυτό, τα οποία ονομάζονται διανύσματα υποστήριξης (support vectors). Τα συγκεκριμένα δείγματα επηρεάζουν άμεσα τη θέση και τον προσανατολισμό του υπερεπιπέδου και, κατά συνέπεια, τον τρόπο με τον οποίο πραγματοποιείται ο διαχωρισμός των κατηγοριών.

2.7.2 Μη γραμμικός διαχωρισμός και συναρτήσεις πυρήνα

Στην πράξη, τα δεδομένα δεν είναι πάντα δυνατό να διαχωριστούν αποτελεσματικά με ένα απλό γραμμικό όριο. Για την αντιμετώπιση τέτοιων περιπτώσεων, τα SVM μπορούν να χρησιμοποιήσουν συναρτήσεις πυρήνα (kernel functions), οι οποίες επιτρέπουν την εφαρμογή της μεθόδου και σε προβλήματα όπου η σχέση μεταξύ των κατηγοριών είναι πιο σύνθετη. Η βασική ιδέα είναι ότι τα δεδομένα μπορούν να αντιμετωπιστούν σαν να έχουν μεταφερθεί σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης, στον οποίο ο διαχωρισμός τους μπορεί να γίνει ευκολότερος.

Η διαδικασία αυτή μπορεί να πραγματοποιηθεί μέσω του λεγόμενου τεχνάσματος πυρήνα (kernel trick), χωρίς να απαιτείται ο άμεσος υπολογισμός της νέας αναπαράστασης των δεδομένων σε όλες τις επιπλέον

διαστάσεις. Με αυτόν τον τρόπο, το SVM μπορεί να δημιουργήσει μη γραμμικά όρια απόφασης, διατηρώντας παράλληλα τη βασική αρχή της εύρεσης ενός κατάλληλου διαχωρισμού μεταξύ των κατηγοριών.

Μία από τις συναρτήσεις πυρήνα που χρησιμοποιούνται στα SVM είναι η Radial Basis Function (RBF). Οι συναρτήσεις πυρήνα αυτού του τύπου επιτρέπουν στο μοντέλο να χειρίζεται δεδομένα που δεν μπορούν να διαχωριστούν ικανοποιητικά στον αρχικό χώρο χαρακτηριστικών, δημιουργώντας τη δυνατότητα για πιο σύνθετα όρια απόφασης. Ως αποτέλεσμα, η επιλογή του κατάλληλου πυρήνα αποτελεί σημαντικό μέρος της διαμόρφωσης ενός μοντέλου SVM και εξαρτάται από τη δομή και τα χαρακτηριστικά των δεδομένων που εξετάζονται.

2.7.3 Πλεονεκτήματα και περιορισμοί του SVM

Το SVM παρουσιάζει αρκετά οφέλη σε προβλήματα ταξινόμησης. Η βασική του δύναμη είναι ότι επιδιώκει να δημιουργήσει ένα όριο απόφασης με όσο το δυνατόν μεγαλύτερο περιθώριο μεταξύ των κατηγοριών, γεγονός που μπορεί να συμβάλει σε καλύτερη γενίκευση σε νέα δεδομένα. Παράλληλα, μέσω των συναρτήσεων πυρήνα μπορεί να εφαρμοστεί και σε περιπτώσεις όπου οι κατηγορίες δεν διαχωρίζονται γραμμικά στον αρχικό χώρο χαρακτηριστικών.

Ένα ακόμη θετικό στοιχείο είναι ότι η θέση του ορίου απόφασης καθορίζεται κυρίως από τα support vectors, δηλαδή από τα δείγματα που βρίσκονται κοντά στο όριο μεταξύ των κατηγοριών. Ωστόσο, η απόδοση του μοντέλου εξαρτάται από την κατάλληλη επιλογή της συνάρτησης πυρήνα και των υπερπαραμέτρων του. Εξαιτίας αυτού, η ρύθμιση του SVM συχνά απαιτεί δοκιμή διαφορετικών επιλογών και αξιολόγησή τους, για παράδειγμα μέσω cross-validation.

Παρά την αποτελεσματικότητά του στην πρόβλεψη και την ταξινόμηση, το SVM παρουσιάζει περιορισμούς ως προς την ερμηνευσιμότητα. Σε σύγκριση με απλούστερα μοντέλα, δεν είναι πάντα εύκολο να προσδιοριστεί με άμεσο τρόπο πώς κάθε χαρακτηριστικό επηρεάζει την τελική απόφαση, ιδιαίτερα όταν χρησιμοποιούνται μη γραμμικοί πυρήνες. Έτσι, το SVM είναι κατάλληλο κυρίως όταν το βασικό ζητούμενο είναι η αξιόπιστη ταξινόμηση και όχι απαραίτητα η πλήρης ερμηνεία της σχέσης μεταξύ των χαρακτηριστικών και της προβλεπόμενης κατηγορίας.

2.8 Βαθιά Μάθηση και Νευρωνικά Δίκτυα

2.8.1 Εισαγωγή στη Βαθιά Μάθηση

Η Βαθιά Μάθηση (Deep Learning) αποτελεί περιοχή της Μηχανικής Μάθησης που βασίζεται στη χρήση τεχνητών νευρωνικών δικτύων με πολλαπλά επίπεδα επεξεργασίας. Μέσω αυτών των επιπέδων, τα μοντέλα μπορούν να μαθαίνουν σύνθετες και μη γραμμικές σχέσεις μεταξύ των δεδομένων εισόδου και της επιθυμητής εξόδου. Η ανάπτυξη της Βαθιάς Μάθησης έχει οδηγήσει σε σημαντικές βελτιώσεις σε διάφορα πεδία, μεταξύ των οποίων βρίσκεται και η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP).

Ένα βασικό χαρακτηριστικό της Βαθιάς Μάθησης είναι η δυνατότητα του μοντέλου να μαθαίνει σταδιακά χρήσιμες αναπαραστάσεις των δεδομένων μέσα από διαδοχικά επίπεδα. Σε αντίθεση με πιο παραδοσιακές προσεγγίσεις

Μηχανικής Μάθησης, στις οποίες συχνά απαιτείται ο καθορισμός και η επιλογή κατάλληλων χαρακτηριστικών πριν από την εκπαίδευση, τα βαθιά νευρωνικά δίκτυα μπορούν να μαθαίνουν πιο σύνθετες αναπαραστάσεις κατά τη διαδικασία της εκπαίδευσης. Η ύπαρξη πολλαπλών κρυφών επιπέδων (hidden layers) επιτρέπει στα βαθιά νευρωνικά δίκτυα να αναπαριστούν πιο σύνθετες σχέσεις και μοτίβα στα δεδομένα.

Παρ'όλα αυτά, η αύξηση του βάθους ενός νευρωνικού δικτύου συνοδεύεται και από μεγαλύτερες απαιτήσεις. Η χρήση περισσότερων κρυφών επιπέδων αυξάνει τόσο τις υπολογιστικές ανάγκες όσο και τις απαιτήσεις σε μνήμη κατά την εκπαίδευση και χρήση του μοντέλου. Συνεπώς, η Βαθιά Μάθηση προσφέρει τη δυνατότητα μοντελοποίησης σύνθετων προβλημάτων, αλλά ταυτόχρονα απαιτεί κατάλληλη αρχιτεκτονική και επαρκείς υπολογιστικούς πόρους.

2.8.2 Τεχνητά και Βαθιά Νευρωνικά Δίκτυα

Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks – ANNs) αποτελούνται από διασυνδεδεμένες μονάδες επεξεργασίας, οι οποίες οργανώνονται σε επίπεδα. Στη βασική τους μορφή περιλαμβάνουν ένα επίπεδο εισόδου (input layer), ένα ή περισσότερα κρυφά επίπεδα (hidden layers) και ένα επίπεδο εξόδου (output layer). Η πληροφορία μεταφέρεται από το επίπεδο εισόδου προς το επίπεδο εξόδου μέσω των συνδέσεων μεταξύ των νευρώνων, στις οποίες αντιστοιχούν διαφορετικά βάρη (weights).

Κατά την εκπαίδευση, το δίκτυο προσαρμόζει τα βάρη των συνδέσεων του με στόχο τη μείωση του σφάλματος μεταξύ της πραγματικής και της επιθυμητής εξόδου. Για τον σκοπό αυτό χρησιμοποιείται η διαδικασία της οπισθοδιάδοσης του σφάλματος (backpropagation), μέσω της οποίας το σφάλμα μεταφέρεται προς τα προηγούμενα επίπεδα και τα βάρη ενημερώνονται σταδιακά. Με την επανάληψη αυτής της διαδικασίας, το δίκτυο μαθαίνει τις σχέσεις και τα μοτίβα που υπάρχουν στα δεδομένα εκπαίδευσης.

Όταν ένα νευρωνικό δίκτυο περιλαμβάνει πολλαπλά κρυφά επίπεδα, χαρακτηρίζεται ως Βαθύ Νευρωνικό Δίκτυο (Deep Neural Network – DNN). Τα διαδοχικά κρυφά επίπεδα επιτρέπουν στο δίκτυο να μαθαίνει πιο σύνθετες και αφηρημένες αναπαραστάσεις των δεδομένων. Αυτή η αυξημένη δυνατότητα μοντελοποίησης αποτελεί βασικό χαρακτηριστικό των βαθιών νευρωνικών δικτύων, συνοδεύεται όμως από μεγαλύτερες απαιτήσεις σε υπολογιστικούς πόρους και μνήμη.

2.8.3 Αναδρομικά Νευρωνικά Δίκτυα (RNN)

Τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks – RNNs) αποτελούν μια κατηγορία νευρωνικών δικτύων σχεδιασμένη για την επεξεργασία ακολουθιακών δεδομένων (sequential data). Σε αντίθεση με τα απλά νευρωνικά δίκτυα, στα RNN υπάρχουν συνδέσεις ανατροφοδότησης (feedback connections), μέσω των οποίων πληροφορία από προηγούμενα στοιχεία της ακολουθίας μπορεί να επηρεάζει την επεξεργασία των επόμενων. Με αυτόν τον τρόπο, το δίκτυο λαμβάνει υπόψη όχι μόνο την τρέχουσα είσοδο, αλλά και πληροφορία που έχει προκύψει από προηγούμενα βήματα.

Κεντρικό στοιχείο της λειτουργίας ενός RNN είναι η κρυφή κατάσταση (hidden state), η οποία μεταφέρει πληροφορία από ένα βήμα της ακολουθίας στο επόμενο. Για μια ακολουθία εισόδων, η κρυφή κατάσταση σε κάθε βήμα καθορίζεται τόσο από την τρέχουσα είσοδο όσο και από την κρυφή κατάσταση του προηγούμενου βήματος. Επιπλέον, τα ίδια βάρη του δικτύου

χρησιμοποιούνται στα διαφορετικά βήματα της ακολουθίας. Η συγκεκριμένη δομή επιτρέπει στα RNN να επεξεργάζονται ακολουθίες διαφορετικού μήκους και να αξιοποιούν τη σειρά με την οποία εμφανίζονται τα δεδομένα.

Η δυνατότητα διατήρησης πληροφορίας από προηγούμενα στοιχεία καθιστά τα RNN ιδιαίτερα χρήσιμα για δεδομένα στα οποία η σειρά έχει σημασία. Χαρακτηριστικό παράδειγμα αποτελεί το κείμενο, καθώς η σημασία μιας λέξης και γενικότερα μιας πρότασης επηρεάζεται από τις λέξεις που εμφανίζονται πριν ή μετά από αυτή. Για τον λόγο αυτό, τα RNN έχουν χρησιμοποιηθεί σε εφαρμογές Επεξεργασίας Φυσικής Γλώσσας, όπως η ταξινόμηση κειμένου, η σύνοψη κειμένου και η μηχανική μετάφραση.

Η εκπαίδευση των RNN πραγματοποιείται μέσω μιας προσαρμογής της οπισθοδιάδοσης που ονομάζεται οπισθοδιάδοση μέσω του χρόνου (Backpropagation Through Time – BPTT). Κατά τη διαδικασία αυτή, το δίκτυο εξετάζεται ως μια ακολουθία διαδοχικών βημάτων και το σφάλμα μεταδίδεται προς τα πίσω μέσα σε αυτά, ώστε να ενημερωθούν τα βάρη του. Εντούτοις, η επεξεργασία μεγάλων ακολουθιών μπορεί να δημιουργήσει δυσκολίες στην εκπαίδευση και στη διατήρηση πληροφορίας για μεγάλο αριθμό βημάτων. Η ανάγκη αποτελεσματικότερης διαχείρισης μακροχρόνιων εξαρτήσεων αποτέλεσε έναν από τους λόγους ανάπτυξης πιο εξελιγμένων αρχιτεκτονικών RNN, όπως τα Long Short-Term Memory (LSTM) δίκτυα.

2.8.4 Πλεονεκτήματα και περιορισμοί της Βαθιάς Μάθησης

Η Βαθιά Μάθηση παρουσιάζει σημαντικά προτερήματα στην επεξεργασία σύνθετων δεδομένων, καθώς τα βαθιά νευρωνικά δίκτυα μπορούν να μαθαίνουν πολύπλοκες και μη γραμμικές σχέσεις μέσα από πολλαπλά επίπεδα επεξεργασίας. Η δυνατότητα αυτή τα καθιστά ιδιαίτερα κατάλληλα για εφαρμογές όπως η Επεξεργασία Φυσικής Γλώσσας, όπου η πληροφορία δεν είναι πάντα εύκολο να περιγραφεί με απλούς, προκαθορισμένους κανόνες. Πέραν αυτού, η χρήση πολλών κρυφών επιπέδων επιτρέπει στα δίκτυα να μαθαίνουν σταδιακά πιο σύνθετες αναπαραστάσεις των δεδομένων.

Ένα ακόμη πλεονέκτημα είναι ότι τα νευρωνικά δίκτυα μπορούν να περιορίσουν την ανάγκη για εκτεταμένο χειροκίνητο καθορισμό χαρακτηριστικών, καθώς μέρος των χρήσιμων αναπαραστάσεων μπορεί να μαθαίνεται κατά την εκπαίδευση. Στην περίπτωση των ακολουθιακών δεδομένων, τα RNN μπορούν επιπλέον να αξιοποιούν πληροφορία από προηγούμενα στοιχεία της ακολουθίας, γεγονός που τα καθιστά κατάλληλα για εφαρμογές κειμενικής ανάλυσης και ταξινόμησης.

Παρά τα οφέλη αυτά, τα μοντέλα Βαθιάς Μάθησης παρουσιάζουν και σημαντικούς περιορισμούς. Η αύξηση του αριθμού των κρυφών επιπέδων και της πολυπλοκότητας του δικτύου συνοδεύεται από αυξημένες απαιτήσεις σε μνήμη και υπολογιστικούς πόρους. Η εκπαίδευση μπορεί επίσης να είναι περισσότερο απαιτητική σε σύγκριση με απλούστερες μεθόδους Μηχανικής Μάθησης, καθώς απαιτεί τη ρύθμιση μεγαλύτερου αριθμού παραμέτρων και υπερπαραμέτρων.

Στα RNN εμφανίζονται επιπλέον δυσκολίες που σχετίζονται με την επεξεργασία μεγάλων ακολουθιών και τη μεταφορά πληροφορίας για μεγάλο αριθμό διαδοχικών βημάτων. Η εκπαίδευσή τους μέσω Backpropagation Through Time είναι διαδοχική και μπορεί να είναι υπολογιστικά απαιτητική, ενώ η αποτελεσματική διατήρηση πληροφορίας από πολύ παλαιότερα σημεία της ακολουθίας δεν είναι πάντα εύκολη. Οι περιορισμοί αυτοί οδήγησαν στην ανάπτυξη πιο εξειδικευμένων αρχιτεκτονικών, όπως τα LSTM.

2.9 Long Short-Term Memory (LSTM)

2.9.1 Βασική αρχή και λειτουργία του Long Short-Term Memory (LSTM)

Τα δίκτυα Long Short-Term Memory (LSTM) αποτελούν ειδική μορφή Αναδρομικών Νευρωνικών Δικτύων (RNN), η οποία αναπτύχθηκε με στόχο την αποτελεσματικότερη διατήρηση και αξιοποίηση πληροφορίας σε μεγάλες ακολουθίες. Στα απλά RNN, κατά τη διαδικασία της εκπαίδευσης, η πληροφορία που προέρχεται από αρκετά προηγούμενα βήματα μπορεί σταδιακά να εξασθενεί, με αποτέλεσμα το δίκτυο να δυσκολεύεται να μάθει εξαρτήσεις μεταξύ στοιχείων που απέχουν πολύ μεταξύ τους μέσα σε μία ακολουθία. Το LSTM προτάθηκε για την αντιμετώπιση αυτής της δυσκολίας και ιδιαίτερα του προβλήματος της εξασθένησης της ροής του σφάλματος κατά την οπισθοδιάδοση.

Η βασική διαφοροποίηση του LSTM από ένα απλό RNN βρίσκεται στην εισαγωγή ενός πιο ελεγχόμενου μηχανισμού μνήμης. Αντί η πληροφορία να μεταφέρεται απλώς από το ένα βήμα της ακολουθίας στο επόμενο, το LSTM διαθέτει ειδικές μονάδες μνήμης και μηχανισμούς πύλης (gates), οι οποίοι ελέγχουν τη ροή της πληροφορίας μέσα στο δίκτυο. Με αυτόν τον τρόπο, το μοντέλο μπορεί να αποφασίζει ποια πληροφορία είναι χρήσιμο να διατηρηθεί και ποια δεν είναι πλέον απαραίτητη.

Η συγκεκριμένη δυνατότητα επιτρέπει στα LSTM να μαθαίνουν μακροχρόνιες εξαρτήσεις (long-term dependencies), δηλαδή σχέσεις μεταξύ στοιχείων που μπορεί να βρίσκονται σε μεγάλη απόσταση μέσα στην ίδια ακολουθία. Αυτό είναι ιδιαίτερα σημαντικό στην επεξεργασία κειμένου, όπου η ερμηνεία ενός στοιχείου μπορεί να εξαρτάται από πληροφορία που εμφανίστηκε αρκετά νωρίτερα. Κατ'επέκταση, το LSTM αποτελεί εξέλιξη της βασικής λογικής των RNN, με μεγαλύτερη δυνατότητα ελέγχου της πληροφορίας που διατηρείται κατά την επεξεργασία μιας ακολουθίας.

2.9.2 Δομή της μονάδας LSTM

Η λειτουργία ενός LSTM βασίζεται σε μια ειδική εσωτερική δομή που επιτρέπει τον ελεγχόμενο χειρισμό της πληροφορίας σε κάθε βήμα της ακολουθίας. Κεντρικό στοιχείο αυτής της δομής είναι η κατάσταση μνήμης (cell state), η οποία λειτουργεί ως ένας μηχανισμός μεταφοράς πληροφορίας από προηγούμενα βήματα προς τα επόμενα. Η πληροφορία που παραμένει στη μνήμη δεν μεταφέρεται ανεξέλεγκτα, αλλά ρυθμίζεται μέσω ειδικών μηχανισμών πύλης (gates).

Η πύλη λήθης (forget gate) καθορίζει ποιο μέρος της προηγούμενης πληροφορίας πρέπει να διατηρηθεί και ποιο μπορεί να απορριφθεί. Με αυτόν τον τρόπο, το μοντέλο αποφεύγει να κρατά πληροφορία που δεν είναι πλέον χρήσιμη για την επεξεργασία της τρέχουσας ακολουθίας. Η πύλη εισόδου (input gate) ελέγχει ποια νέα πληροφορία από το τρέχον βήμα θα προστεθεί στη μνήμη, επιτρέποντας την ενημέρωση της εσωτερικής κατάστασης με στοιχεία που θεωρούνται σημαντικά.

Τέλος, η πύλη εξόδου (output gate) καθορίζει ποιο μέρος της ενημερωμένης κατάστασης μνήμης θα χρησιμοποιηθεί για τη δημιουργία της εξόδου στο συγκεκριμένο βήμα. Με τη συνεργασία των τριών αυτών πυλών, το LSTM μπορεί να ελέγχει δυναμικά ποια πληροφορία διατηρείται, ποια ενημερώνεται και ποια απορρίπτεται. Έτσι, το δίκτυο μπορεί να διαχειρίζεται αποτελεσματικότερα πληροφορία που εκτείνεται σε μεγαλύτερο αριθμό βημάτων σε σχέση με ένα απλό RNN.

Η συγκεκριμένη δομή είναι ιδιαίτερα χρήσιμη σε δεδομένα ακολουθιών, όπως το κείμενο, όπου η σημασία ενός στοιχείου μπορεί να εξαρτάται από πληροφορίες που εμφανίστηκαν νωρίτερα. Ο ελεγχόμενος μηχανισμός μνήμης επιτρέπει στο LSTM να διατηρεί σημαντικό περιεχόμενο και να περιορίζει την επίδραση λιγότερο χρήσιμων πληροφοριών κατά την επεξεργασία της ακολουθίας.

2.9.3 Πλεονεκτήματα και περιορισμοί του LSTM

Τα LSTM παρουσιάζουν σημαντικά πλεονεκτήματα στην επεξεργασία ακολουθιακών δεδομένων, επειδή έχουν σχεδιαστεί ώστε να διατηρούν πληροφορία για μεγαλύτερα χρονικά διαστήματα σε σχέση με τα απλά RNN. Μέσω της εσωτερικής κατάστασης μνήμης και των μηχανισμών πύλης μπορούν να ελέγχουν ποια πληροφορία διατηρείται, ποια ενημερώνεται και ποια απορρίπτεται. Η δυνατότητα αυτή τα καθιστά κατάλληλα για προβλήματα στα οποία υπάρχουν μακροχρόνιες εξαρτήσεις μεταξύ των στοιχείων μιας ακολουθίας.

Ένα ακόμη θετικό στοιχείο είναι ότι τα LSTM μπορούν να χρησιμοποιηθούν αποτελεσματικά σε εργασίες επεξεργασίας κειμένου, όπου η σειρά των λέξεων και η πληροφορία από προηγούμενα σημεία της πρότασης ή της ακολουθίας είναι σημαντικές. Η αρχιτεκτονική τους επιτρέπει να αντιμετωπίζουν καλύτερα το πρόβλημα της εξασθένησης του σήματος που εμφανίζεται στα απλά RNN και να διατηρούν χρήσιμη πληροφορία σε μεγαλύτερο αριθμό διαδοχικών βημάτων.

Παρά τα προτερήματά τους, τα LSTM είναι πιο σύνθετα από τα απλά RNN, καθώς περιλαμβάνουν περισσότερες εσωτερικές λειτουργίες και παραμέτρους λόγω των μηχανισμών πύλης. Αυτό αυξάνει τις απαιτήσεις σε υπολογιστικούς πόρους και μπορεί να οδηγήσει σε μεγαλύτερο χρόνο εκπαίδευσης. Περαιτέρω, η επεξεργασία των ακολουθιών πραγματοποιείται διαδοχικά, γεγονός που περιορίζει τη δυνατότητα πλήρους παραλληλοποίησης της εκπαίδευσης.

Ως εκ τούτου, τα LSTM αποτελούν ιδιαίτερα χρήσιμη εξέλιξη των RNN όταν απαιτείται η αξιοποίηση πληροφορίας από μεγαλύτερες ακολουθίες, με το κόστος όμως μιας πιο σύνθετης αρχιτεκτονικής και αυξημένων υπολογιστικών απαιτήσεων.

2.10 Αμφίδρομα Δίκτυα LSTM (Bi-LSTM) και Υβριδικές Αρχιτεκτονικές

2.10.1 Αρχιτεκτονική και λειτουργία του Bi-LSTM

Τα αμφίδρομα δίκτυα μακράς βραχυπρόθεσμης μνήμης (Bidirectional Long Short-Term Memory – Bi-LSTM) αποτελούν επέκταση των δικτύων LSTM, με βασικό χαρακτηριστικό την επεξεργασία μιας ακολουθίας προς δύο κατευθύνσεις. Σε ένα συμβατικό LSTM, η επεξεργασία πραγματοποιείται διαδοχικά από την αρχή προς το τέλος της ακολουθίας, με αποτέλεσμα κάθε χρονικό βήμα να αξιοποιεί την πληροφορία που έχει προηγηθεί. Αντίθετα, το Bi-LSTM επιτρέπει την αξιοποίηση πληροφοριών τόσο από προηγούμενα όσο και από επόμενα στοιχεία της ακολουθίας.

Η αρχιτεκτονική του Bi-LSTM αποτελείται από δύο ανεξάρτητα LSTM δίκτυα. Το πρώτο, γνωστό ως forward LSTM, επεξεργάζεται την ακολουθία προς την κανονική της κατεύθυνση, από το πρώτο προς το τελευταίο στοιχείο.

Το δεύτερο, το backward LSTM, επεξεργάζεται την ίδια ακολουθία προς την αντίθετη κατεύθυνση, από το τελευταίο προς το πρώτο στοιχείο. Οι έξοδοι των δύο δικτύων στη συνέχεια συνδυάζονται, ώστε η τελική αναπαράσταση να περιλαμβάνει πληροφορία που έχει εξαχθεί και από τις δύο κατευθύνσεις.

Η αμφίδρομη αυτή λειτουργία επιτρέπει στο δίκτυο να σχηματίζει μια πληρέστερη εικόνα του περιβάλλοντος στο οποίο εμφανίζεται κάθε στοιχείο της ακολουθίας. Η ιδιότητα αυτή είναι ιδιαίτερα χρήσιμη όταν η ερμηνεία ενός στοιχείου δεν εξαρτάται μόνο από όσα προηγούνται, αλλά και από όσα ακολουθούν. Με αυτόν τον τρόπο, το Bi-LSTM διατηρεί τους μηχανισμούς μνήμης του LSTM, ενώ παράλληλα επεκτείνει την επεξεργασία της ακολουθίας και προς τις δύο κατευθύνσεις.

2.10.2 Bi-LSTM στην ταξινόμηση κειμένου

Τα Bi-LSTM μπορούν να χρησιμοποιηθούν αποτελεσματικά σε προβλήματα ταξινόμησης κειμένου, καθώς η αμφίδρομη λειτουργία τους επιτρέπει την επεξεργασία μιας ακολουθίας λαμβάνοντας υπόψη πληροφορίες και από τις δύο κατευθύνσεις. Η δυνατότητα αυτή είναι ιδιαίτερα χρήσιμη στην επεξεργασία φυσικής γλώσσας, όπου η σημασία μιας λέξης συχνά εξαρτάται από το περιεχόμενο που βρίσκεται τόσο πριν όσο και μετά από αυτή. Με αυτόν τον τρόπο, το μοντέλο μπορεί να σχηματίσει μια πληρέστερη αναπαράσταση των συμφραζομένων του κειμένου.

Σε προβλήματα ταξινόμησης, οι αναπαραστάσεις που παράγονται από τις δύο κατευθύνσεις μπορούν να συνδυαστούν και να αξιοποιηθούν για την τελική πρόβλεψη της κατηγορίας στην οποία ανήκει το κείμενο. Έτσι, σε αντίθεση με ένα μονοκατευθυντικό LSTM, το οποίο επεξεργάζεται την ακολουθία μόνο προς μία κατεύθυνση, το Bi-LSTM μπορεί να αξιοποιήσει ένα ευρύτερο πλαίσιο πληροφορίας κατά την αναπαράσταση της ακολουθίας.

Η συγκεκριμένη ιδιότητα καθιστά τα Bi-LSTM κατάλληλα για εργασίες επεξεργασίας φυσικής γλώσσας στις οποίες η κατανόηση των συμφραζομένων αποτελεί σημαντικό μέρος της διαδικασίας ταξινόμησης. Ωστόσο, η χρησιμότητα της αμφίδρομης επεξεργασίας εξαρτάται από τη μορφή των δεδομένων εισόδου, καθώς το βασικό πλεονέκτημα της αρχιτεκτονικής προκύπτει όταν τα δεδομένα αποτελούν πραγματική ακολουθία και υπάρχει πληροφορία που μπορεί να αξιοποιηθεί και προς τις δύο κατευθύνσεις.

2.10.3 Υβριδικές και πολυκαναλικές αρχιτεκτονικές

Εκτός από τη χρήση ενός μεμονωμένου νευρωνικού δικτύου, μπορούν να αναπτυχθούν υβριδικές αρχιτεκτονικές, στις οποίες συνδυάζονται διαφορετικοί μηχανισμοί ή επιμέρους κλάδοι επεξεργασίας. Στόχος αυτών των αρχιτεκτονικών είναι η αξιοποίηση συμπληρωματικών πληροφοριών, ώστε η τελική πρόβλεψη να βασίζεται σε μια πιο ολοκληρωμένη αναπαράσταση των δεδομένων. Ένα παράδειγμα αποτελεί ο συνδυασμός Bi-LSTM με μηχανισμό Multi-Head Attention, όπου το Bi-LSTM χρησιμοποιείται για την επεξεργασία της ακολουθίας, ενώ ο μηχανισμός προσοχής αποδίδει διαφορετική βαρύτητα στις πληροφορίες που περιέχονται σε αυτή.

Μια σχετική προσέγγιση είναι οι πολυκαναλικές αρχιτεκτονικές (multi-channel architectures), στις οποίες διαφορετικοί κλάδοι χρησιμοποιούνται για την εξαγωγή διαφορετικών χαρακτηριστικών από τα δεδομένα. Κάθε κλάδος μπορεί να επεξεργάζεται την πληροφορία με διαφορετικό τρόπο και οι αναπαραστάσεις που προκύπτουν συνενώνονται στη συνέχεια μέσω της συγχώνευσης χαρακτηριστικών (feature fusion). Η συνδυασμένη

αναπαράσταση χρησιμοποιείται τελικά για την πραγματοποίηση της ταξινόμησης.

Οι Liang et al. (2024) ακολούθησαν αυτή τη λογική σε ένα πολυκαναλικό μοντέλο ταξινόμησης κειμένου, χρησιμοποιώντας διαφορετικές μεθόδους για την εξαγωγή χαρακτηριστικών και συνδυάζοντας τις πληροφορίες που προέκυψαν από τα επιμέρους κανάλια πριν από το τελικό επίπεδο ταξινόμησης. Η προσέγγιση αυτή δείχνει πώς διαφορετικές αναπαραστάσεις μπορούν να ενσωματωθούν σε μία κοινή αρχιτεκτονική, αντί η ταξινόμηση να βασίζεται αποκλειστικά στα χαρακτηριστικά που παράγονται από έναν μόνο μηχανισμό.

2.10.4 Πλεονεκτήματα και περιορισμοί των Bi-LSTM και υβριδικών αρχιτεκτονικών

Τα Bi-LSTM παρουσιάζουν σημαντικό πλεονέκτημα στην επεξεργασία ακολουθιακών δεδομένων, καθώς μπορούν να αξιοποιούν πληροφορία και από τις δύο κατευθύνσεις της ακολουθίας. Αυτό επιτρέπει τη δημιουργία πλουσιότερων αναπαραστάσεων, ιδιαίτερα σε προβλήματα επεξεργασίας κειμένου όπου τα συμφραζόμενα πριν και μετά από ένα στοιχείο μπορούν να επηρεάζουν την ερμηνεία του.

Αντίστοιχα, οι υβριδικές και πολυκαναλικές αρχιτεκτονικές προσφέρουν τη δυνατότητα συνδυασμού διαφορετικών μορφών πληροφορίας ή διαφορετικών τρόπων εξαγωγής χαρακτηριστικών. Μέσω της χρήσης πολλαπλών κλάδων και της συγχώνευσης χαρακτηριστικών, το μοντέλο μπορεί να σχηματίσει μια πιο ολοκληρωμένη αναπαράσταση πριν από την τελική ταξινόμηση. Η λογική αυτή χρησιμοποιείται τόσο σε συνδυασμούς διαφορετικών μηχανισμών, όπως Bi-LSTM και Multi-Head Attention, όσο και σε multi-channel μοντέλα που εξάγουν χαρακτηριστικά από διαφορετικούς κλάδους.

Ωστόσο, η αυξημένη αυτή δυνατότητα συνοδεύεται και από μεγαλύτερη πολυπλοκότητα. Τα Bi-LSTM περιλαμβάνουν δύο κατευθύνσεις επεξεργασίας, ενώ οι υβριδικές αρχιτεκτονικές μπορεί να περιλαμβάνουν πολλαπλούς κλάδους, διαφορετικά επίπεδα και διαδικασίες συνένωσης. Ως αποτέλεσμα, αυξάνονται ο αριθμός των παραμέτρων και οι υπολογιστικές απαιτήσεις, ενώ απαιτείται μεγαλύτερη προσοχή στον σχεδιασμό και στη ρύθμιση της αρχιτεκτονικής.

Επιπρόσθετα, η χρήση πιο σύνθετης αρχιτεκτονικής δεν εγγυάται από μόνη της καλύτερη απόδοση. Η χρησιμότητα κάθε επιπλέον κλάδου ή μηχανισμού εξαρτάται από το αν προσφέρει πραγματικά συμπληρωματική πληροφορία για το συγκεκριμένο πρόβλημα. Κατά συνέπεια, η αποτελεσματικότητα των Bi-LSTM και των υβριδικών μοντέλων πρέπει να αξιολογείται πειραματικά σε σχέση με τα δεδομένα και τον στόχο της εφαρμογής.

2.11 Σχετικές Εργασίες

Η ανάλυση περιεχομένου που προέρχεται από μέσα κοινωνικής δικτύωσης έχει αποτελέσει αντικείμενο πολλών ερευνητικών εργασιών, με διαφορετικές προσεγγίσεις ως προς την αξιοποίηση των hashtags, την ταξινόμηση του κειμένου και τον συνδυασμό διαφορετικών τύπων πληροφορίας. Οι σχετικές μελέτες περιλαμβάνουν τόσο παραδοσιακές τεχνικές Μηχανικής Μάθησης όσο και νεότερες αρχιτεκτονικές Βαθιάς Μάθησης.

Ειδικότερα, οι Wang et al. (2016) επικεντρώθηκαν άμεσα στην αξιοποίηση των hashtags για τη θεματική ανάλυση περιεχομένου στο Twitter. Οι συγγραφείς πρότειναν το Hashtag Graph-based Topic Model (HGTM), το οποίο αξιοποιεί τις σχέσεις μεταξύ των hashtags για τον εντοπισμό σημασιολογικών συσχετίσεων μεταξύ λέξεων, ακόμη και όταν αυτές δεν εμφανίζονται μαζί στο ίδιο tweet. Η προσέγγιση σχεδιάστηκε με στόχο την αντιμετώπιση της αραιότητας και του θορύβου που χαρακτηρίζουν τα σύντομα κείμενα των κοινωνικών δικτύων και αξιολογήθηκε, μεταξύ άλλων, σε προβλήματα ομαδοποίησης tweets και ταξινόμησης hashtags. Τα αποτελέσματα της μελέτης ανέδειξαν τη χρησιμότητα των hashtags ως πηγή θεματικής και σημασιολογικής πληροφορίας.

Μια ευρύτερη εικόνα της έρευνας γύρω από τα hashtags παρουσιάζεται από τους Bansal et al. (2025), οι οποίοι πραγματοποίησαν ανασκόπηση των συστημάτων σύστασης hashtags και της εξέλιξής τους από παραδοσιακές μεθόδους έως σύγχρονες τεχνικές Βαθιάς Μάθησης. Στην ανασκόπηση εξετάζονται διαφορετικές διατυπώσεις του προβλήματος, όπως η κατάταξη, η ταξινόμηση και η παραγωγή hashtags, καθώς και unimodal, multimodal και υβριδικές προσεγγίσεις. Οι συγγραφείς επισημαίνουν επίσης τη σταδιακή μετάβαση προς μοντέλα Βαθιάς Μάθησης που αξιοποιούν περισσότερο το σημασιολογικό και συμφοραζόμενο περιεχόμενο των δεδομένων.

Στο πεδίο της ταξινόμησης αναρτήσεων κοινωνικών δικτύων, οι Yunida et al. (2023) διερεύνησαν τη χρήση LSTM και Bi-LSTM για την αναγνώριση αναφορών σχετικών με φυσικές καταστροφές. Η μελέτη πραγματοποιήθηκε σε 3.000 αναρτήσεις του Twitter, κατανεμημένες σε τρεις κατηγορίες, και συνέκρινε την απόδοση των δύο νευρωνικών αρχιτεκτονικών. Το Bi-LSTM παρουσίασε ελαφρώς υψηλότερη απόδοση, επιτυγχάνοντας Accuracy 72,17% έναντι 70,67% του LSTM. Η συγκεκριμένη εργασία αποτελεί χαρακτηριστικό παράδειγμα αξιοποίησης ακολουθιακών μοντέλων για την ταξινόμηση περιεχομένου κοινωνικών δικτύων.

Παρόμοια, οι Verma et al. (2026) πραγματοποίησαν συγκριτική αξιολόγηση κλασικών αλγορίθμων Μηχανικής Μάθησης και μοντέλου LSTM για την ανάλυση συναισθήματος σε δεδομένα κοινωνικών δικτύων. Συγκεκριμένα, εξετάστηκαν οι Naïve Bayes, Random Forest και SVM σε σύγκριση με LSTM, σε συνδυασμό με διαφορετικές τεχνικές αναπαράστασης κειμένου, και συγκεκριμένα TF-IDF, FastText και Word2Vec. Σε σύνολο 30.001 αναρτήσεων, ο συνδυασμός LSTM και Word2Vec παρουσίασε την υψηλότερη απόδοση, με Accuracy 92,65% και F1-score 94,37%. Τα αποτελέσματα της μελέτης ανέδειξαν τη σημασία όχι μόνο της επιλογής του μοντέλου, αλλά και του τρόπου με τον οποίο αναπαρίσταται η κειμενική πληροφορία.

Τέλος, οι Singh et al. (2026) εξέτασαν μια διαφορετική προσέγγιση, βασισμένη στη συνδυασμένη αξιοποίηση πολλαπλών μορφών πληροφορίας για την ανάλυση συναισθήματος σε δεδομένα κοινωνικών δικτύων. Η προτεινόμενη μεθοδολογία συνδυάζει κειμενικά και ηχητικά χαρακτηριστικά και εξετάζει διαφορετικές στρατηγικές συγχώνευσης, όπως early, late και hybrid fusion. Παράλληλα, αξιολογούνται αρχιτεκτονικές όπως LSTM, GRU και αμφίδρομα μοντέλα. Η εργασία αυτή αναδεικνύει τη δυνατότητα των υβριδικών προσεγγίσεων να συνδυάζουν διαφορετικές πηγές πληροφορίας στο πλαίσιο μιας κοινής διαδικασίας ταξινόμησης.

Συνολικά, οι παραπάνω εργασίες δείχνουν ότι η ανάλυση περιεχομένου κοινωνικών δικτύων μπορεί να προσεγγιστεί με διαφορετικούς τρόπους, από την άμεση αξιοποίηση των σχέσεων μεταξύ hashtags έως τη χρήση κλασικών

αλγορίθμων Μηχανικής Μάθησης, ακολουθιακών νευρωνικών δικτύων και υβριδικών αρχιτεκτονικών. Κοινό στοιχείο των σύγχρονων προσεγγίσεων αποτελεί η προσπάθεια αξιοποίησης πλουσιότερης πληροφορίας από το περιεχόμενο και το πλαίσιο των αναρτήσεων, με στόχο την αποτελεσματικότερη ταξινόμησή τους.

ΚΕΦΑΛΑΙΟ 3 Μεθοδολογία και Υλοποίηση

3.1 Σύνολο δεδομένων

Για την πειραματική διαδικασία της παρούσας εργασίας χρησιμοποιήθηκε το σύνολο δεδομένων Social Media Sentiments Analysis Dataset, το οποίο διατίθεται δημόσια μέσω της πλατφόρμας Kaggle. Το αρχείο που αξιοποιήθηκε έχει την ονομασία sentimentdataset.csv και περιλαμβάνει αναρτήσεις από διαφορετικές πλατφόρμες κοινωνικής δικτύωσης, συνοδευόμενες από πληροφορίες σχετικές με το περιεχόμενο, το συναίσθημα, την αλληλεπίδραση των χρηστών και τον χρόνο δημοσίευσης. Σύμφωνα με την περιγραφή του δημιουργού, το dataset έχει σχεδιαστεί για τη μελέτη συναισθημάτων, τάσεων και αλληλεπιδράσεων σε περιβάλλοντα κοινωνικής δικτύωσης.

Το αρχικό σύνολο δεδομένων αποτελείται από 732 εγγραφές, καθεμία από τις οποίες αντιστοιχεί σε μία ανάρτηση. Κατά τον αρχικό έλεγχο δεν εντοπίστηκαν ελλείψεις τιμές ή πλήρως διπλότυπες εγγραφές. Το περιορισμένο μέγεθος του dataset, σε συνδυασμό με τη μεγάλη ποικιλία των hashtags, αποτελεί σημαντικό χαρακτηριστικό της παρούσας ανάλυσης και επηρεάζει τόσο τη διαδικασία κατηγοριοποίησης όσο και την απόδοση των μοντέλων.

Κάθε εγγραφή περιλαμβάνει τα ακόλουθα βασικά χαρακτηριστικά:

Πίνακας 1 Βασικά χαρακτηριστικά του συνόλου δεδομένων

Χαρακτηριστικό	Περιγραφή
<i>Text</i>	Το κειμενικό περιεχόμενο της ανάρτησης.
<i>Sentiment</i>	Η συναισθηματική κατηγορία που έχει αποδοθεί στην ανάρτηση στο αρχικό dataset.
<i>Timestamp</i>	Η ημερομηνία και η ώρα δημοσίευσης της ανάρτησης.
<i>User</i>	Το αναγνωριστικό ή το όνομα του χρήστη που πραγματοποίησε την ανάρτηση.
<i>Platform</i>	Η πλατφόρμα κοινωνικής δικτύωσης στην οποία δημοσιεύτηκε η ανάρτηση.
<i>Hashtags</i>	Οι θεματικές ετικέτες που συνοδεύουν την ανάρτηση.
<i>Likes</i>	Ο αριθμός των επισημάνσεων «μου αρέσει» που έλαβε η ανάρτηση.
<i>Retweets</i>	Ο αριθμός των αναδημοσιεύσεων της ανάρτησης.
<i>Country</i>	Η χώρα που συνδέεται με την εγγραφή.
<i>Year</i>	Το έτος δημοσίευσης.
<i>Month</i>	Ο μήνας δημοσίευσης.
<i>Day</i>	Η ημέρα δημοσίευσης.
<i>Hour</i>	Η ώρα δημοσίευσης.

Τα διαθέσιμα χαρακτηριστικά δεν χρησιμοποιούνται όλα με τον ίδιο τρόπο σε κάθε πειραματική προσέγγιση. Στα κλασικά μοντέλα Μηχανικής Μάθησης αξιοποιήθηκαν αρχικά τα αριθμητικά χαρακτηριστικά Likes και Retweets, ενώ στις πιο σύνθετες προσεγγίσεις εξετάστηκε η ενσωμάτωση πρόσθετων πληροφοριών, όπως η πλατφόρμα, το συναίσθημα και το κειμενικό περιεχόμενο των hashtags.

Η μεταβλητή-στόχος δεν υπήρχε έτοιμη στο αρχικό dataset. Για τον λόγο αυτό δημιουργήθηκε στο πλαίσιο της εργασίας μέσω της επεξεργασίας και της θεματικής ομαδοποίησης των hashtags.

3.2 Προεπεξεργασία δεδομένων

Η προεπεξεργασία των δεδομένων αποτελεί βασικό στάδιο της παρούσας εργασίας, καθώς τα `hashtags` των αναρτήσεων εμφανίζουν μεγάλη μορφολογική ποικιλία και δεν παρέχονται εξ αρχής σε μορφή κατάλληλη για την εκπαίδευση μοντέλων ταξινόμησης. Για τον λόγο αυτό αναπτύχθηκε μία διαδοχική διαδικασία επεξεργασίας, η οποία περιλαμβάνει την εξαγωγή και κανονικοποίηση των `hashtags`, τη θεματική τους ομαδοποίηση, τον καθορισμό μίας κύριας κατηγορίας ανά ανάρτηση και τη δημιουργία των αριθμητικών και κειμενικών χαρακτηριστικών που χρησιμοποιούνται στα διαφορετικά μοντέλα.

3.2.1 Εξαγωγή και κανονικοποίηση των `hashtags`

Στο πρώτο στάδιο επεξεργάζεται η στήλη `Hashtags` του αρχικού συνόλου δεδομένων. Κάθε εγγραφή της συγκεκριμένης στήλης μπορεί να περιλαμβάνει περισσότερα από ένα `hashtags`, τα οποία αποθηκεύονται αρχικά ως μία ενιαία συμβολοσειρά.

Για την απομόνωση των επιμέρους `hashtags` υλοποιήθηκε η συνάρτηση `parse_hashtags`. Η συνάρτηση ελέγχει αρχικά αν η τιμή είναι κενή και, στη συνέχεια, χρησιμοποιεί την κανονική έκφραση:

$$\#?([A-Za-z0-9_]+) \quad (3.1)$$

Η συγκεκριμένη έκφραση εντοπίζει ακολουθίες που αποτελούνται από λατινικούς χαρακτήρες, αριθμούς και κάτω παύλες, ανεξάρτητα από το αν προηγείται το σύμβολο `#`. Μετά την εξαγωγή, όλα τα `hashtags` μετατρέπονται σε πεζούς χαρακτήρες, ώστε διαφορετικές γραφές της ίδιας λέξης να αντιμετωπίζονται με ενιαίο τρόπο.

Για παράδειγμα, η αρχική τιμή:

`#Nature #Park`

μετατρέπεται στη λίστα:

`[nature, park]`

Το αποτέλεσμα αποθηκεύεται στη νέα στήλη `hashtags_list`, η οποία περιέχει για κάθε ανάρτηση μία λίστα καθαρών `hashtags`.

3.2.2 Τμηματοποίηση σύνθετων `hashtags`

Τα `hashtags` συχνά περιλαμβάνουν περισσότερες από μία λέξεις, οι οποίες μπορεί να συνδέονται με κάτω παύλες ή να εμφανίζονται ως ενιαίες συμβολοσειρές. Για την επεξεργασία τέτοιων περιπτώσεων υλοποιήθηκαν οι συναρτήσεις `tag_to_tokens` και `split_camel_case`.

Η συνάρτηση `tag_to_tokens` αντικαθιστά τις κάτω παύλες με κενά διαστήματα και δημιουργεί ένα σύνολο λεκτικών μονάδων. Παράλληλα, η `split_camel_case` χρησιμοποιεί κανονική έκφραση για τον εντοπισμό επιμέρους λέξεων και αριθμών.

Για παράδειγμα:

`lifestyle_travel` → `[lifestyle, travel]`

Τα `tokens` που προκύπτουν χρησιμοποιούνται στη συνέχεια για την αντιστοίχιση κάθε `hashtag` με τις λέξεις-κλειδιά των θεματικών κατηγοριών.

Στην παρούσα υλοποίηση, η μετατροπή των `hashtags` σε πεζούς χαρακτήρες πραγματοποιείται πριν από την εφαρμογή της συνάρτησης διαχωρισμού. Επομένως, η αναγνώριση όρων τύπου Camel Case είναι περιορισμένη όταν δεν υπάρχουν κάτω παύλες, καθώς η πληροφορία των

κεφαλαίων έχει ήδη χαθεί. Το σημείο αυτό αποτελεί περιορισμό της συγκεκριμένης διαδικασίας προεπεξεργασίας.

3.2.3 Θεματική κατηγοριοποίηση των hashtags

Για τη θεματική ομαδοποίηση των hashtags δημιουργήθηκε ένα προκαθορισμένο λεξικό λέξεων-κλειδιών με την ονομασία CATEGORY_KEYWORDS. Το λεξικό περιλαμβάνει επτά βασικές θεματικές κατηγορίες:

- positive_emotion: θετικά συναισθήματα,
- negative_emotion: αρνητικά συναισθήματα,
- emotion_reflection: συναισθηματική κατάσταση και αναστοχασμός,
- lifestyle_travel: τρόπος ζωής, υγεία, φαγητό, φύση και ταξίδια,
- entertainment_technology: ψυχαγωγία, τέχνη και τεχνολογία,
- social_education: κοινωνικές σχέσεις, εκπαίδευση και εργασία,
- sports_politics_pets: αθλητισμός, πολιτική και κατοικίδια.

Επιπλέον, χρησιμοποιείται η κατηγορία other για τα hashtags που δεν αντιστοιχίζονται σε καμία από τις παραπάνω ομάδες. Συνεπώς, η τελική διαδικασία ταξινόμησης περιλαμβάνει συνολικά οκτώ κλάσεις.

Η αντιστοίχιση πραγματοποιείται μέσω της συνάρτησης categorize_tags, η οποία υπολογίζει ένα σκορ για κάθε πιθανή κατηγορία. Για κάθε hashtag εφαρμόζονται δύο κανόνες βαθμολόγησης:

- όταν μία λέξη-κλειδί ταυτίζεται πλήρως με ένα token του hashtag, προστίθενται τρεις βαθμοί στην αντιστοιχη κατηγορία,
- όταν μία λέξη-κλειδί με μήκος τουλάχιστον τεσσάρων χαρακτήρων περιέχεται στο hashtag ως υποσυμβολοσειρά, προστίθεται ένας βαθμός.

Η χρήση του περιορισμού των τεσσάρων χαρακτήρων μειώνει τον κίνδυνο λανθασμένων αντιστοιχίσεων από πολύ σύντομες λέξεις, όπως οι όροι ai και vr.

Μετά τον υπολογισμό των σκορ, το hashtag αποδίδεται στην κατηγορία με τη μεγαλύτερη βαθμολογία. Εάν καμία κατηγορία δεν συγκεντρώσει θετικό σκορ, το hashtag κατατάσσεται στην κατηγορία other.

3.2.4 Καθορισμός της κύριας κατηγορίας

Μία ανάρτηση μπορεί να περιλαμβάνει περισσότερα από ένα hashtags και, επομένως, να σχετίζεται με περισσότερες θεματικές κατηγορίες. Η συνάρτηση assign_category εφαρμόζει την προηγούμενη διαδικασία σε όλα τα hashtags μίας εγγραφής και συγκεντρώνει τις διαφορετικές κατηγορίες σε μία λίστα.

Όταν η κατηγορία other εμφανίζεται μαζί με μία ή περισσότερες συγκεκριμένες κατηγορίες, αφαιρείται. Με τον τρόπο αυτό, μία γενική και μη πληροφοριακή αντιστοίχιση δεν υπερισχύει μιας πιο συγκεκριμένης θεματικής κατηγορίας.

Επειδή τα μοντέλα αντιμετωπίζουν το πρόβλημα ως ταξινόμηση μίας ετικέτας ανά ανάρτηση, δημιουργήθηκε η μεταβλητή primary_category. Για την επιλογή της χρησιμοποιείται η προκαθορισμένη λίστα CATEGORY_PRIORITY, η οποία ορίζει την ακόλουθη σειρά προτεραιότητας:

positive_emotion → negative_emotion → emotion_reflection
→ lifestyle_travel → entertainment_technology → social_education
→ sports_politics_pets → other

Η συνάρτηση get_primary_category εξετάζει τις κατηγορίες που έχουν αποδοθεί στην ανάρτηση και επιλέγει την πρώτη που εμφανίζεται στη λίστα

προτεραιότητας. Η κατηγορία αυτή χρησιμοποιείται ως μεταβλητή-στόχος στα μοντέλα ταξινόμησης.

Η συγκεκριμένη μέθοδος εξασφαλίζει μία μοναδική ετικέτα για κάθε εγγραφή, αποτελεί όμως έναν ντετερμινιστικό κανόνα που έχει οριστεί στο πλαίσιο της εργασίας. Κατά συνέπεια, η σειρά προτεραιότητας μπορεί να επηρεάσει την τελική κατανομή των κλάσεων και πρέπει να αναγνωριστεί ως περιορισμός της μεθοδολογίας.

3.2.5 Κωδικοποίηση και δημιουργία αριθμητικών χαρακτηριστικών

Οι αλγόριθμοι ταξινόμησης απαιτούν η μεταβλητή-στόχος να παρέχεται σε αριθμητική μορφή. Για τον λόγο αυτό χρησιμοποιήθηκε η κλάση `LabelEncoder` της βιβλιοθήκης `scikit-learn`.

Η μεταβλητή `primary_category` μετατράπηκε στη νέα αριθμητική μεταβλητή `primary_category_encoded`. Οι οκτώ κατηγορίες αντιστοιχίστηκαν σε ακέραιες τιμές από 0 έως 7. Οι αριθμοί αυτοί χρησιμοποιούνται αποκλειστικά ως αναγνωριστικά κλάσεων και δεν εκφράζουν ιεραρχία ή αριθμητική διάταξη μεταξύ των κατηγοριών.

Με αντίστοιχο τρόπο κωδικοποιήθηκαν και οι μεταβλητές `Platform` και `Sentiment`, δημιουργώντας τις στήλες:

- `Platform_encoded`,
- `Sentiment_encoded`.

Επιπλέον, δημιουργήθηκε η μεταβλητή `hashtags_Count`, η οποία εκφράζει το πλήθος των `hashtags` που περιλαμβάνει κάθε ανάρτηση. Το χαρακτηριστικό αυτό αξιοποιείται σε ορισμένες από τις εμπλουτισμένες αρχιτεκτονικές `Bi-LSTM` μαζί με τα `Likes`, `Retweets` και `Platform_encoded`.

Στα βασικά μοντέλα `Random Forest`, `kNN` και `SVM`, τα χαρακτηριστικά εισόδου περιορίζονται στις μεταβλητές `Likes` και `Retweets`. Στις πιο σύνθετες νευρωνικές προσεγγίσεις χρησιμοποιούνται επιπλέον χαρακτηριστικά, όπως το πλήθος των `hashtags`, η πλατφόρμα και το συναίσθημα της ανάρτησης.

3.2.6 Δημιουργία κειμενικής αναπαράστασης μεταδεδομένων

Για το μοντέλο `Bi-LSTM` που συνδυάζει αριθμητικά και κειμενικά δεδομένα δημιουργήθηκε η συνθετική μεταβλητή `metadata_text`. Στην παρούσα υλοποίηση, η μεταβλητή αυτή προκύπτει από τη συνένωση των πεδίων:

```
metadata_text = Platform + " " + Hashtags + " " + Sentiment (3.2)
```

Η νέα στήλη μετατρέπεται σε ακολουθίες ακεραίων μέσω της κλάσης `Tokenizer`. Στη συνέχεια, οι ακολουθίες αποκτούν σταθερό μήκος 15 θέσεων με τη χρήση της συνάρτησης `pad_sequences`. Οι αριθμητικές ακολουθίες εισάγονται σε στρώμα `Embedding` και επεξεργάζονται από διαδοχικά αμφίδρομα `LSTM` στρώματα.

Ακόμα, τα `Likes` και `Retweets` χρησιμοποιούνται ως δεύτερη είσοδος του ίδιου μοντέλου. Οι δύο αναπαραστάσεις συνενώνονται πριν από τα τελικά πλήρως συνδεδεμένα στρώματα ταξινόμησης.

Παρ'όλα αυτά, η συγκεκριμένη υλοποίηση περιλαμβάνει τα ίδια τα `hashtags` στη μεταβλητή `metadata_text`, ενώ η μεταβλητή-στόχος έχει δημιουργηθεί από αυτά. Η επιλογή αυτή μπορεί να προκαλέσει διαρροή πληροφορίας στόχου, καθώς το μοντέλο λαμβάνει στην είσοδό του πληροφορία που σχετίζεται άμεσα με τη σωστή κατηγορία.

Για μία αυστηρότερη και πιο δίκαιη πειραματική αξιολόγηση, είναι προτιμότερο το `metadata_text` να δημιουργηθεί χωρίς το πεδίο `Hashtags`, για παράδειγμα:

$$\text{metadata_text} = \text{Platform} + " " + \text{Sentiment} + " " + \text{Text} \quad (3.3)$$

Με τον τρόπο αυτό, το μοντέλο θα επιχειρεί να προβλέψει την κατηγορία των `hashtags` αξιοποιώντας το κείμενο και τα μεταδεδομένα της ανάρτησης, χωρίς να έχει άμεση πρόσβαση στα ίδια τα `hashtags`.

3.3 Επιλογή χαρακτηριστικών εισόδου

Η επιλογή των χαρακτηριστικών που χρησιμοποιούνται ως εισοδοί στα μοντέλα αποτελεί σημαντικό στάδιο της διαδικασίας ταξινόμησης, καθώς επηρεάζει άμεσα την ικανότητα των αλγορίθμων να αναγνωρίζουν πρότυπα στα δεδομένα. Στην παρούσα εργασία δεν εφαρμόστηκε κάποια τεχνική αυτόματης επιλογής χαρακτηριστικών (`feature selection`), αλλά σχεδιάστηκαν διαφορετικά σύνολα χαρακτηριστικών, τα οποία αξιοποιήθηκαν από διαφορετικές κατηγορίες μοντέλων με σκοπό τη συγκριτική αξιολόγηση της απόδοσής τους.

Τα χαρακτηριστικά που χρησιμοποιήθηκαν μπορούν να διακριθούν στις ακόλουθες κατηγορίες.

3.3.1 Αριθμητικά χαρακτηριστικά αλληλεπίδρασης

Η πρώτη ομάδα χαρακτηριστικών περιλαμβάνει τις μεταβλητές που αποτυπώνουν τον βαθμό αλληλεπίδρασης των χρηστών με κάθε ανάρτηση.

Συγκεκριμένα χρησιμοποιήθηκαν οι μεταβλητές:

- **Likes**, που αντιστοιχούν στον αριθμό των θετικών αντιδράσεων μιας ανάρτησης.
- **Retweets**, που αντιστοιχούν στον αριθμό των αναδημοσιεύσεων της ανάρτησης.

Οι δύο αυτές μεταβλητές χρησιμοποιήθηκαν ως μοναδικά χαρακτηριστικά εισόδου στα μοντέλα `Random Forest`, `k-Nearest Neighbors (kNN)`, `Support Vector Machine (SVM)`, `LSTM` και στο βασικό μοντέλο `Bi-LSTM`.

3.3.2 Χαρακτηριστικά μεταδεδομένων

Με στόχο τον εμπλουτισμό της πληροφορίας που λαμβάνουν τα μοντέλα, δημιουργήθηκαν επιπλέον χαρακτηριστικά προερχόμενα από τα μεταδεδομένα των αναρτήσεων.

Συγκεκριμένα δημιουργήθηκαν οι μεταβλητές:

- **hashtags_Count**, η οποία εκφράζει το πλήθος των `hashtags` που περιλαμβάνει κάθε ανάρτηση.
- **Platform_encoded**, η οποία προκύπτει από την αριθμητική κωδικοποίηση της πλατφόρμας δημοσίευσης μέσω της τεχνικής `Label Encoding`.
- **Sentiment_encoded**, η οποία αποτελεί την αριθμητική αναπαράσταση της συναισθηματικής κατηγορίας της ανάρτησης.

Οι μεταβλητές αυτές χρησιμοποιήθηκαν στα πιο σύνθετα μοντέλα `Bi-LSTM`, με στόχο να εξεταστεί κατά πόσο η αξιοποίηση επιπλέον πληροφοριών μπορεί να βελτιώσει την απόδοση της ταξινόμησης.

3.3.3 Κειμενικά χαρακτηριστικά

Για το υβριδικό μοντέλο βαθιάς μάθησης δημιουργήθηκε μία επιπλέον κειμενική μεταβλητή με την ονομασία **metadata_text**.

Η μεταβλητή αυτή προκύπτει από τη συνένωση των πεδίων:

- Platform
- Hashtags
- Sentiment

σε μία ενιαία συμβολοσειρά.

Στη συνέχεια εφαρμόζεται ο Tokenizer της βιβλιοθήκης TensorFlow/Keras, ο οποίος μετατρέπει κάθε λέξη σε μοναδικό ακέραιο δείκτη. Οι ακολουθίες που προκύπτουν αποκτούν κοινό μήκος μέσω της διαδικασίας padding (μήκος 15 tokens), ώστε να μπορούν να αποτελέσουν είσοδο του στρώματος Embedding του νευρωνικού δικτύου.

3.3.4 Συνδυασμοί χαρακτηριστικών ανά μοντέλο

Η παρούσα εργασία δεν χρησιμοποιεί το ίδιο σύνολο χαρακτηριστικών σε όλα τα μοντέλα. Αντίθετα, κάθε αρχιτεκτονική αξιοποιεί διαφορετικό συνδυασμό εισόδων, ώστε να διερευνηθεί η επίδραση των διαθέσιμων πληροφοριών στην ακρίβεια ταξινόμησης.

Πίνακας 2 Συνοπτικά χαρακτηριστικά που χρησιμοποιούνται σε κάθε μοντέλο

Μοντέλο	Χαρακτηριστικά εισόδου
<i>Random Forest</i>	Likes, Retweets
<i>kNN</i>	Likes, Retweets
<i>SVM</i>	Likes, Retweets
<i>LSTM</i>	Likes, Retweets
<i>Bi-LSTM</i>	Likes, Retweets
<i>Bi-LSTM2</i>	Likes, Retweets, hashtags_Count, Platform_encoded
<i>Bi-LSTM Hybrid</i>	Likes, Retweets, hashtags_Count, Platform_encoded, Sentiment_encoded
<i>Bi-LSTM Numeric + Text</i>	Likes, Retweets + metadata_text

3.4 Υλοποίηση του μοντέλου Random Forest

Για την πρώτη προσέγγιση ταξινόμησης υλοποιήθηκε μοντέλο Random Forest μέσω της κλάσης RandomForestClassifier της βιβλιοθήκης scikit-learn. Στόχος του μοντέλου είναι η πρόβλεψη της κύριας θεματικής κατηγορίας των hashtags κάθε ανάρτησης, χρησιμοποιώντας αποκλειστικά αριθμητικά χαρακτηριστικά αλληλεπίδρασης.

3.4.1 Χαρακτηριστικά εισόδου και μεταβλητή-στόχος

Ως χαρακτηριστικά εισόδου χρησιμοποιήθηκαν οι μεταβλητές Likes και Retweets. Συνεπώς, για κάθε ανάρτηση το διάνυσμα εισόδου ορίζεται ως:

$$\mathbf{X}_i = [\mathbf{X}_{\text{Likes},i}, \mathbf{X}_{\text{Retweets},i}] \quad (3.4)$$

όπου i αντιστοιχεί σε μία εγγραφή του συνόλου δεδομένων.

Ως μεταβλητή-στόχος χρησιμοποιήθηκε η στήλη primary_category_encoded, η οποία περιέχει την αριθμητική κωδικοποίηση

της κύριας θεματικής κατηγορίας κάθε ανάρτησης. Το πρόβλημα αντιμετωπίστηκε επομένως ως πρόβλημα πολυκατηγορικής ταξινόμησης, με οκτώ πιθανές κλάσεις.

3.4.2 Διαχωρισμός των δεδομένων

Τα διαθέσιμα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης και σύνολο δοκιμής με αναλογία 80% και 20%, αντίστοιχα. Ο διαχωρισμός πραγματοποιήθηκε μέσω της συνάρτησης `train_test_split`, με τιμή `random_state=42`, ώστε ο ίδιος διαχωρισμός να μπορεί να αναπαραχθεί σε επόμενες εκτελέσεις.

Με βάση το συνολικό πλήθος των 732 εγγράφων, το σύνολο δοκιμής περιλαμβάνει 147 δείγματα, ενώ οι υπόλοιπες εγγραφές χρησιμοποιούνται για την εκπαίδευση του μοντέλου.

3.4.3 Παραμετροποίηση και εκπαίδευση

Το μοντέλο δημιουργήθηκε με τις ακόλουθες βασικές παραμέτρους:

- `n_estimators=100`: χρησιμοποιήθηκαν 100 δέντρα απόφασης.
- `random_state=42`: καθορίστηκε σταθερή κατάσταση τυχαιότητας για λόγους επαναληψιμότητας.
- `class_weight="balanced"`: τα βάρη των κλάσεων προσαρμόστηκαν αυτόματα με βάση τη συχνότητα εμφάνισής τους στο σύνολο εκπαίδευσης.

Η τελευταία επιλογή χρησιμοποιήθηκε λόγω της ανισορροπίας που παρατηρήθηκε στη μεταβλητή-στόχο. Με τον τρόπο αυτό, οι λιγότερο συχνές κατηγορίες λαμβάνουν μεγαλύτερο βάρος κατά την εκπαίδευση, ενώ οι συχνότερες κατηγορίες λαμβάνουν αναλογικά μικρότερο.

Η εκπαίδευση πραγματοποιήθηκε με τη μέθοδο `fit`, χρησιμοποιώντας τα `X_train` και `y_train`. Μετά την ολοκλήρωση της εκπαίδευσης, το μοντέλο παρήγαγε προβλέψεις για το σύνολο δοκιμής μέσω της μεθόδου `predict`.

Η βασική υλοποίηση του μοντέλου Random Forest, από τον καθορισμό των χαρακτηριστικών εισόδου έως την παραγωγή των προβλέψεων και τον υπολογισμό των μετρικών αξιολόγησης, παρουσιάζεται στην παρακάτω εικόνα.

```
def random_forest_classification(df, le):
    X = df[['Likes', 'Retweets']]
    y = df['primary_category_encoded']

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    rf_model = RandomForestClassifier(n_estimators=100, random_state=42, class_weight='balanced')
    rf_model.fit(X_train, y_train)

    y_pred = rf_model.predict(X_test)

    y_test1 = y_test.values

    accuracy = accuracy_score(y_test1, y_pred)
    print("~~~~~\n")
    print("\t-Random Forest Accuracy: ", accuracy)
    report = classification_report(y_test, y_pred, target_names=le.classes_, zero_division=0)
    print("\n\t-Random Forest Classification Report: ", report)
    print("~~~~~\n")

    cm = confusion_matrix(y_test, y_pred, normalize='true')

    disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=le.classes_)

    fig, ax = plt.subplots(figsize=(12,9))
    disp.plot(cmap=plt.cm.Blues, xticks_rotation=45, ax=ax)
    plt.title("Confusion Matrix Random Forest")
    plt.tight_layout()
    plt.show()

    return rf_model
```

Figure 1 Υλοποίηση της εκπαίδευσης και αξιολόγησης του μοντέλου Random Forest.

Όπως φαίνεται στην εικόνα, ως είσοδοι του μοντέλου χρησιμοποιούνται οι μεταβλητές Likes και Retweets, ενώ ως μεταβλητή-στόχος χρησιμοποιείται η κωδικοποιημένη κύρια κατηγορία των hashtags. Τα δεδομένα διαχωρίζονται σε σύνολο εκπαίδευσης και σύνολο δοκιμής, το μοντέλο εκπαιδεύεται με 100 δέντρα απόφασης και στη συνέχεια παράγονται οι προβλέψεις για το σύνολο δοκιμής. Η αξιολόγηση πραγματοποιείται μέσω της συνολικής ακρίβειας, του classification report και του κανονικοποιημένου πίνακα σύγχυσης.

3.4.4 Αξιολόγηση του μοντέλου

Η απόδοση του Random Forest αξιολογήθηκε με τη χρήση των ακόλουθων μετρικών:

- Accuracy,
- Precision,
- Recall,
- F1-score.

Η συνολική ακρίβεια υπολογίστηκε μέσω της συνάρτησης accuracy_score, ενώ οι επιμέρους μετρικές ανά κατηγορία παρουσιάστηκαν μέσω της classification_report.

Επιπλέον, δημιουργήθηκε κανονικοποιημένος πίνακας σύγχυσης με τη χρήση της παραμέτρου normalize="true".

Με αυτή την επιλογή, κάθε γραμμή του πίνακα παρουσιάζει την αναλογία των δειγμάτων μιας πραγματικής κατηγορίας που ταξινομήθηκαν σε κάθε προβλεπόμενη κατηγορία. Η κανονικοποιημένη μορφή διευκολύνει τη σύγκριση μεταξύ κλάσεων με διαφορετικό αριθμό δειγμάτων.

3.5 Υλοποίηση του μοντέλου k-Nearest Neighbors

Για τη δεύτερη προσέγγιση ταξινόμησης υλοποιήθηκε μοντέλο k-Nearest Neighbors μέσω της κλάσης KNeighborsClassifier της βιβλιοθήκης scikit-learn. Στόχος του μοντέλου είναι η πρόβλεψη της κύριας θεματικής κατηγορίας των hashtags κάθε ανάρτησης, χρησιμοποιώντας τα αριθμητικά χαρακτηριστικά αλληλεπίδρασης Likes και Retweets.

3.5.1 Χαρακτηριστικά εισόδου και μεταβλητή-στόχος

Ως χαρακτηριστικά εισόδου χρησιμοποιήθηκαν οι μεταβλητές Likes και Retweets. Επομένως, για κάθε ανάρτηση το διάνυσμα χαρακτηριστικών ορίζεται ως:

$$\mathbf{X}_i = [X_{\text{Likes},i}, X_{\text{Retweets},i}] \quad (3.5)$$

Πληκτρολογήστε την εξίσωση εδώ.

Ως μεταβλητή-στόχος χρησιμοποιήθηκε η στήλη primary_category_encoded, η οποία περιλαμβάνει την αριθμητική αναπαράσταση της κύριας θεματικής κατηγορίας κάθε εγγραφής. Το μοντέλο καλείται συνεπώς να επιλέξει μία από τις οκτώ διαθέσιμες κλάσεις.

3.5.2 Διαχωρισμός δεδομένων

Ο διαχωρισμός των δεδομένων πραγματοποιήθηκε με τη συνάρτηση train_test_split. Το 80% των εγγραφών χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου και το υπόλοιπο 20% για την αξιολόγησή του. Η παράμετρος random_state=42 διατηρήθηκε σταθερή, ώστε ο διαχωρισμός να είναι επαναλήψιμος και συγκρίσιμος με εκείνον των υπόλοιπων μοντέλων.

3.5.3 Παραμετροποίηση και εκπαίδευση

Το μοντέλο δημιουργήθηκε με την παράμετρο: `n_neighbors=5`. Άρα, για την ταξινόμηση κάθε νέου δείγματος λαμβάνονται υπόψη τα πέντε πλησιέστερα δείγματα του συνόλου εκπαίδευσης. Η τελική πρόβλεψη προκύπτει από την κατηγορία που εμφανίζεται συχνότερα μεταξύ των πέντε αυτών γειτόνων.

Στην παρούσα υλοποίηση δεν ορίστηκε διαφορετική μετρική απόστασης, συνεπώς χρησιμοποιήθηκαν οι προεπιλεγμένες ρυθμίσεις της `KNeighborsClassifier`. Για τη συγκεκριμένη παραμετροποίηση, η απόσταση αντιστοιχεί στην Ευκλείδεια απόσταση:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{j=1}^n (p_j - q_j)^2} \quad (3.6)$$

όπου $\mathbf{p}$ και $\mathbf{q}$ είναι δύο δείγματα στον χώρο των χαρακτηριστικών.

Η εκπαίδευση πραγματοποιήθηκε με τη μέθοδο `fit`, χρησιμοποιώντας τα `X_train` και `y_train`, ενώ οι προβλέψεις για το σύνολο δοκιμής παράχθηκαν με τη μέθοδο `predict`.

Η βασική υλοποίηση του μοντέλου `kNN`, από τον καθορισμό των χαρακτηριστικών εισόδου έως την παραγωγή των προβλέψεων και τον υπολογισμό των μετρικών αξιολόγησης, παρουσιάζεται στην παρακάτω εικόνα.

```
def knn_classification(df, le):
    X = df[['Likes', 'Retweets']]
    y = df['primary_category_encoded']

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    knn_model = KNeighborsClassifier(n_neighbors=5)
    knn_model.fit(X_train, y_train)

    y_pred = knn_model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    print("~~~~~\n")
    print("\t-KNN Accuracy: ", accuracy)
    report = classification_report(y_test, y_pred, target_names= le.classes_, zero_division=0)
    print("\n\t-KNN Classification Report: ", report)
    print("~~~~~\n")

    cm = confusion_matrix(y_test, y_pred, normalize='true')

    disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=le.classes_)

    fig, ax = plt.subplots(figsize=(12,9))
    disp.plot(cmap=plt.cm.pink, xticks_rotation=45, ax=ax)
    plt.title("Confusion Matrix kNN")
    plt.tight_layout()
    plt.show()

    return knn_model
```

Figure 2 Υλοποίηση της εκπαίδευσης και αξιολόγησης του μοντέλου `k-Nearest Neighbors`.

Όπως φαίνεται παραπάνω, ως εισοδοί χρησιμοποιούνται οι μεταβλητές `Likes` και `Retweets`, ενώ η παράμετρος `n_neighbors` ορίστηκε ίση με 5. Μετά την εκπαίδευση, το μοντέλο παράγει προβλέψεις για το σύνολο δοκιμής και αξιολογείται μέσω της συνολικής ακρίβειας, του `classification report` και του κανονικοποιημένου πίνακα σύγχυσης.

3.5.4 Αξιολόγηση του μοντέλου

Η απόδοση του `kNN` αξιολογήθηκε με τις ίδιες μετρικές που χρησιμοποιήθηκαν και στο μοντέλο `Random Forest`:

- Accuracy,

- Precision,
- Recall,
- F1-score.

Η συνολική ακρίβεια υπολογίστηκε με τη συνάρτηση `accuracy_score`, ενώ οι επιμέρους μετρικές ανά κατηγορία παρουσιάστηκαν μέσω της `classification_report`.

Πέραν αυτού, δημιουργήθηκε κανονικοποιημένος πίνακας σύγχυσης με την παράμετρο: `normalize="true"`.

Με αυτή τη μορφή, οι τιμές του πίνακα εκφράζουν την αναλογία των δειγμάτων κάθε πραγματικής κατηγορίας που ταξινομήθηκαν σε κάθε προβλεπόμενη κατηγορία.

3.6 Υλοποίηση του μοντέλου Support Vector Machine

Για την τρίτη προσέγγιση ταξινόμησης υλοποιήθηκε μοντέλο Support Vector Machine μέσω της κλάσης SVC της βιβλιοθήκης `scikit-learn`. Στόχος του μοντέλου είναι η πρόβλεψη της κύριας θεματικής κατηγορίας των `hashtags` κάθε ανάρτησης, χρησιμοποιώντας τα αριθμητικά χαρακτηριστικά Likes και Retweets.

3.6.1 Χαρακτηριστικά εισόδου και μεταβλητή-στόχος

Ως χαρακτηριστικά εισόδου χρησιμοποιήθηκαν οι μεταβλητές Likes και Retweets. Για κάθε εγγραφή, το διάνυσμα εισόδου ορίζεται ως:

$$\mathbf{x}_i = [X_{\text{Likes},i}, X_{\text{Retweets},i}] \quad (3.7)$$

Ως μεταβλητή-στόχος χρησιμοποιήθηκε η στήλη `primary_category_encoded`, η οποία περιέχει την αριθμητική κωδικοποίηση της κύριας θεματικής κατηγορίας κάθε ανάρτησης. Το μοντέλο αντιμετωπίζει επομένως το πρόβλημα ως πολυκατηγορική ταξινόμηση σε οκτώ κλάσεις.

3.6.2 Διαχωρισμός των δεδομένων

Τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης και σύνολο δοκιμής με αναλογία 80% και 20%, αντιστοίχα, μέσω της συνάρτησης `train_test_split`.

Η παράμετρος `random_state=42` χρησιμοποιήθηκε ώστε ο διαχωρισμός να παραμένει σταθερός και να είναι δυνατή η επανάληψη του πειράματος με τα ίδια δεδομένα εκπαίδευσης και δοκιμής.

3.6.3 Παραμετροποίηση και εκπαίδευση

Για την υλοποίηση χρησιμοποιήθηκε η κλάση SVC με τις εξής βασικές παραμέτρους:

- `kernel="rbf"`: επιλέχθηκε πυρήνας Radial Basis Function, ώστε το μοντέλο να μπορεί να διαχωρίσει κλάσεις που δεν είναι γραμμικά διαχωρίσιμες στον αρχικό χώρο των χαρακτηριστικών.
- `class_weight="balanced"`: εφαρμόστηκε αυτόματη προσαρμογή των βαρών των κλάσεων με βάση τη συχνότητά τους στο σύνολο εκπαίδευσης.
- `random_state=42`: χρησιμοποιήθηκε σταθερή τιμή τυχειότητας για λόγους επαναληψιμότητας.

Η χρήση του RBF `kernel` επιτρέπει στο μοντέλο να δημιουργεί μη γραμμικά όρια απόφασης, γεγονός που είναι χρήσιμο όταν η σχέση μεταξύ

των χαρακτηριστικών Likes, Retweets και των κατηγοριών δεν μπορεί να περιγραφεί με ένα απλό γραμμικό υπερεπίπεδο.

Η εκπαίδευση πραγματοποιήθηκε με τη μέθοδο fit, χρησιμοποιώντας τα X_train και y_train, ενώ οι προβλέψεις για το σύνολο δοκιμής παράχθηκαν με τη μέθοδο predict.

Η βασική υλοποίηση του μοντέλου SVM, από τον καθορισμό των χαρακτηριστικών εισόδου έως την παραγωγή των προβλέψεων και τον υπολογισμό των μετρικών αξιολόγησης, παρουσιάζεται στην παρακάτω εικόνα.

```
def svm_classification(df, le):
    X = df[['Likes', 'Retweets']]
    y = df['primary_category_encoded']

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    svm_model = SVC(kernel='rbf', class_weight='balanced', random_state=42)
    svm_model.fit(X_train, y_train)

    y_pred = svm_model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    print("~~~~~\n")
    print("\t-SVM Accuracy: ", accuracy)
    report = classification_report(y_test, y_pred, target_names=le.classes_, zero_division=0)
    print("\n\t-SVM Classification Report: ", report)
    print("~~~~~\n")

    cm = confusion_matrix(y_test, y_pred, normalize='true')

    disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=le.classes_)

    fig, ax = plt.subplots(figsize=(12,9))
    disp.plot(cmap=plt.cm.Greens, xticks_rotation=45, ax=ax)
    plt.title("Confusion Matrix SVM")
    plt.tight_layout()
    plt.show()

    return svm_model
```

Figure 3 Υλοποίηση της εκπαίδευσης και αξιολόγησης του μοντέλου Support Vector Machine.

Όπως φαίνεται παραπάνω, ως εισοδοί χρησιμοποιούνται οι μεταβλητές Likes και Retweets, ενώ το μοντέλο υλοποιείται με πυρήνα RBF και ισορροπημένα βάρη κλάσεων. Μετά την εκπαίδευση, παράγονται οι προβλέψεις για το σύνολο δοκιμής και υπολογίζονται οι αντίστοιχες μετρικές αξιολόγησης.

3.6.4 Αξιολόγηση του μοντέλου

Η απόδοση του SVM αξιολογήθηκε με τις ίδιες μετρικές που χρησιμοποιήθηκαν και στα προηγούμενα μοντέλα, ώστε να είναι δυνατή η άμεση σύγκριση των αποτελεσμάτων:

- Accuracy,
- Precision,
- Recall,
- F1-score.

Η συνολική ακρίβεια υπολογίστηκε μέσω της συνάρτησης accuracy_score, ενώ οι επιμέρους μετρικές ανά κατηγορία παρουσιάστηκαν με τη χρήση της classification_report.

Παράλληλα, δημιουργήθηκε κανονικοποιημένος πίνακας σύγχυσης με την επιλογή: normalize="true"

Η κανονικοποίηση επιτρέπει την παρουσίαση των αποτελεσμάτων ως αναλογίες ανά πραγματική κατηγορία, διευκολύνοντας την ερμηνεία των σφαλμάτων και τη σύγκριση μεταξύ κλάσεων με διαφορετικό αριθμό δειγμάτων.

3.7 Υλοποίηση του μοντέλου LSTM

Για τη διερεύνηση μιας προσέγγισης βαθιάς μάθησης υλοποιήθηκε μοντέλο Long Short-Term Memory μέσω της βιβλιοθήκης TensorFlow/Keras. Το μοντέλο χρησιμοποιήθηκε για την πρόβλεψη της κύριας θεματικής κατηγορίας των hashtags, με βάση τα αριθμητικά χαρακτηριστικά Likes και Retweets.

3.7.1 Χαρακτηριστικά εισόδου και μεταβλητή-στόχος

Ως χαρακτηριστικά εισόδου χρησιμοποιήθηκαν οι μεταβλητές Likes και Retweets, ενώ ως μεταβλητή-στόχος χρησιμοποιήθηκε η κωδικοποιημένη στήλη `primary_category_encoded`.

Το διάνυσμα χαρακτηριστικών κάθε ανάρτησης ορίζεται ως:

$$\mathbf{X}_i = [X_{\text{Likes},i}, X_{\text{Retweets},i}] \quad (3.8)$$

Τα δεδομένα διαχωρίστηκαν σε σύνολο εκπαίδευσης και σύνολο δοκιμής με αναλογία 80% και 20%, αντίστοιχα, μέσω της συνάρτησης `train_test_split`. Για λόγους επαναληψιμότητας χρησιμοποιήθηκε η τιμή `random_state=42`.

3.7.2 Αναδιαμόρφωση των δεδομένων εισόδου

Τα επίπεδα LSTM απαιτούν είσοδο τριών διαστάσεων, με τη γενική μορφή:

$$(\text{samples, time steps, features}). \quad (3.9)$$

Για τον λόγο αυτό, τα αρχικά αριθμητικά δεδομένα αναδιαμορφώθηκαν από διοδιάστατους πίνακες σε τριοδιάστατους πίνακες με σχήμα:

$$(\text{samples}, 1, 2) \quad (3.10)$$

Η πρώτη διάσταση αντιστοιχεί στο πλήθος των εγγραφών, η δεύτερη σε ένα χρονικό βήμα και η τρίτη στα δύο χαρακτηριστικά, δηλαδή τα Likes και Retweets.

Στην παρούσα υλοποίηση χρησιμοποιείται μόνο ένα χρονικό βήμα. Συνεπώς, το μοντέλο δεν αξιοποιεί πλήρεις χρονικές ακολουθίες, αλλά εφαρμόζει την αρχιτεκτονική LSTM σε αριθμητικά χαρακτηριστικά σταθερού μήκους. Η επιλογή αυτή πραγματοποιήθηκε ώστε να συγκριθεί η απόδοσή του με τα υπόλοιπα μοντέλα στο ίδιο σύνολο εισόδων.

3.7.3 Αρχιτεκτονική του δικτύου

Το μοντέλο υλοποιήθηκε ως διαδοχική αρχιτεκτονική `Sequential` και περιλαμβάνει τα ακόλουθα επίπεδα:

- **LSTM επίπεδο:** περιλαμβάνει 64 μονάδες και χρησιμοποιεί συνάρτηση ενεργοποίησης `ReLU`.
- **Dropout επίπεδο:** εφαρμόζει ποσοστό απόρριψης 20%, με στόχο τον περιορισμό της υπερπροσαρμογής.
- **Πλήρως συνδεδεμένο επίπεδο:** περιλαμβάνει 32 νευρώνες με συνάρτηση ενεργοποίησης `ReLU`.
- **Επίπεδο εξόδου:** περιλαμβάνει αριθμό νευρώνων ίσο με το πλήθος των τελικών κλάσεων και χρησιμοποιεί συνάρτηση ενεργοποίησης `Softmax`.

Η συνάρτηση `Softmax` μετατρέπει τις εξόδους του τελευταίου επιπέδου σε κατανομή πιθανοτήτων:

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (3.11)$$

όπου K είναι το συνολικό πλήθος των κατηγοριών. Η κλάση με τη μεγαλύτερη πιθανότητα επιλέγεται ως τελική πρόβλεψη.

3.7.4 Μεταγλώττιση και εκπαίδευση

Το μοντέλο μεταγλωττίστηκε με:

- βελτιστοποιητή Adam,
- συνάρτηση απώλειας `sparse_categorical_crossentropy`,
- μετρική παρακολούθησης την `accuracy`.

Η εκπαίδευση πραγματοποιήθηκε για 25 εποχές, με μέγεθος παρτίδας 16: `epochs=25`

`batch_size=16`

Η επιλογή της `sparse_categorical_crossentropy` είναι κατάλληλη, καθώς η μεταβλητή-στόχος περιέχει ακέραιες κωδικοποιημένες κλάσεις και όχι one-hot διανύσματα.

Μετά την εκπαίδευση, το μοντέλο παρήγαγε για κάθε δείγμα του συνόλου δοκιμής ένα διάνυσμα πιθανοτήτων. Η τελική κατηγορία προέκυψε μέσω της συνάρτησης `argmax`, η οποία επέλεξε τη θέση με τη μεγαλύτερη πιθανότητα.

Η υλοποίηση της προετοιμασίας των δεδομένων, της αρχιτεκτονικής και της εκπαίδευσης του μοντέλου LSTM παρουσιάζεται παρακάτω.

```
def lstm_classification(df, le):
    X = df[['Likes', 'Retweets']].values
    y = df['primary_category_encoded'].values

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    X_train_3d = X_train.reshape((X_train.shape[0], 1, X_train.shape[1]))
    X_test_3d = X_test.reshape((X_test.shape[0], 1, X_test.shape[1]))

    lstm_model = Sequential([
        LSTM(64, input_shape=(1, 2), activation='relu'),
        Dropout(0.2),
        Dense(32, activation='relu'),
        Dense(len(le.classes_), activation='softmax')
    ])

    lstm_model.compile(optimizer='adam',
                      loss='sparse_categorical_crossentropy',
                      metrics=['accuracy'])

    history = lstm_model.fit(X_train_3d, y_train, epochs=25, batch_size=16, verbose=1)

    y_pred_probs = lstm_model.predict(X_test_3d)
    y_pred = np.argmax(y_pred_probs, axis=1)

    accuracy = accuracy_score(y_test, y_pred)
    print("~~~~~\n")
    print("t-LSTM Accuracy: ", accuracy)
    report = classification_report(y_test, y_pred, target_names=le.classes_, zero_division=0)
    print("\n\t-LSTM Classification Report: ", report)
    print("~~~~~\n")

    plt.figure(figsize=(12, 5))
    plt.subplot(1, 2, 1)
    plt.plot(history.history['accuracy'], label='Train Accuracy', color='lightgreen', marker='o')
    plt.title('LSTM Model Accuracy progression')
    plt.xlabel('Epoch')
    plt.ylabel('Accuracy')
    plt.legend()
    plt.grid(True, linestyle='--', alpha=0.6)
```

Figure 4 Υλοποίηση της αρχιτεκτονικής και της διαδικασίας εκπαίδευσης του μοντέλου LSTM.

Όπως φαίνεται στην παραπάνω εικόνα, τα δύο αριθμητικά χαρακτηριστικά αναδιαμορφώνονται σε τρισδιάστατη μορφή πριν εισαχθούν στο επίπεδο LSTM. Η αρχιτεκτονική περιλαμβάνει ένα LSTM επίπεδο 64 μονάδων, Dropout 20%, ένα πλήρως συνδεδεμένο επίπεδο 32 νευρώνων και τελικό επίπεδο Softmax για την πολυκατηγορική ταξινόμηση.

3.7.5 Αξιολόγηση του μοντέλου

Η αξιολόγηση πραγματοποιήθηκε με τις ίδιες μετρικές που χρησιμοποιήθηκαν στα προηγούμενα μοντέλα:

- Accuracy,

Η αξιολόγηση πραγματοποιήθηκε με τις ίδιες μετρικές που χρησιμοποιήθηκαν στα προηγούμενα μοντέλα:

- Accuracy,
- Precision,
- Recall,
- F1-score.

Συν τοις άλλοις, καταγράφηκαν η ακρίβεια και η απώλεια κατά τη διάρκεια των εποχών εκπαίδευσης, ώστε να παρακολουθείται η εξέλιξη της διαδικασίας μάθησης. Δημιουργήθηκε επίσης κανονικοποιημένος πίνακας σύγχυσης για την ανάλυση των σωστών και λανθασμένων προβλέψεων ανά κατηγορία.

3.8 Υλοποίηση μοντέλων Bidirectional LSTM

Στο πλαίσιο της παρούσας εργασίας αναπτύχθηκαν τέσσερις διαφορετικές εκδοχές μοντέλων Bidirectional Long Short-Term Memory. Στόχος των διαδοχικών υλοποιήσεων ήταν να εξεταστεί κατά πόσο ο εμπλουτισμός των αριθμητικών χαρακτηριστικών, ο διαχωρισμός των εισόδων σε ανεξάρτητους κλάδους και η ενσωμάτωση κειμενικής πληροφορίας μπορούν να βελτιώσουν την πρόβλεψη της κύριας θεματικής κατηγορίας των hashtags.

Τα μοντέλα υλοποιήθηκαν με τη βιβλιοθήκη TensorFlow/Keras και αξιολογήθηκαν με βάση τις ίδιες μετρικές που χρησιμοποιήθηκαν στις προηγούμενες προσεγγίσεις. Για κάθε μοντέλο, τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης 80% και σύνολο δοκιμής 20%, με σταθερή τιμή `random_state=42`.

3.8.1 Βασικό μοντέλο Bi-LSTM

Η πρώτη υλοποίηση αποτελεί τη βασική εκδοχή του μοντέλου Bi-LSTM και χρησιμοποιεί αποκλειστικά τις αριθμητικές μεταβλητές Likes και Retweets. Ως μεταβλητή-στόχος χρησιμοποιήθηκε η `primary_category_encoded`.

Τα δεδομένα εισόδου αναδιαμορφώθηκαν σε τρισδιάστατους πίνακες της μορφής:

$$(\text{samples}, 1, 2) \quad (3.12)$$

όπου το 1 αντιστοιχεί σε ένα χρονικό βήμα και το 2 στα δύο χαρακτηριστικά εισόδου.

Η αρχιτεκτονική του δικτύου περιλαμβάνει:

- ένα αμφίδρομο επίπεδο Bidirectional(LSTM(64)) με συνάρτηση ενεργοποίησης ReLU,
- ένα επίπεδο Dropout(0.2),
- ένα πλήρως συνδεδεμένο επίπεδο 32 νευρώνων με ενεργοποίηση ReLU,
- ένα τελικό επίπεδο Softmax με αριθμό νευρώνων ίσο με το πλήθος των κλάσεων.

Το μοντέλο μεταγλωττίστηκε με βελτιστοποιητή Adam και συνάρτηση απώλειας `sparse_categorical_crossentropy`. Η εκπαίδευση πραγματοποιήθηκε για 25 εποχές, με μέγεθος παρτίδας 16.

Επειδή η αριθμητική είσοδος περιλαμβάνει μόνο ένα χρονικό βήμα, η αμφίδρομη επεξεργασία δεν αξιοποιεί πραγματικές χρονικές εξαρτήσεις. Η συγκεκριμένη υλοποίηση χρησιμοποιήθηκε κυρίως ως βάση σύγκρισης με τις επόμενες, εμπλουτισμένες αρχιτεκτονικές.

```
def bilstm_classification(df, le):
    X = df[['Likes', 'Retweets']].values
    y = df['primary_category_encoded'].values

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    X_train_3d = X_train.reshape((X_train.shape[0], 1, X_train.shape[1]))
    X_test_3d = X_test.reshape((X_test.shape[0], 1, X_test.shape[1]))

    bilstm_model = Sequential([
        Bidirectional(LSTM(64, input_shape=(1, 2), activation='relu')),
        Dropout(0.2),
        Dense(32, activation='relu'),
        Dense(len(le.classes_), activation='softmax')
    ])

    bilstm_model.compile(optimizer='adam',
                        loss='sparse_categorical_crossentropy',
                        metrics=['accuracy'])

    history = bilstm_model.fit(X_train_3d, y_train, epochs=25, batch_size=16, verbose=1)

    y_pred_probs = bilstm_model.predict(X_test_3d)
    y_pred = np.argmax(y_pred_probs, axis=1)

    accuracy = accuracy_score(y_test, y_pred)
    print("~~~~~\n")
    print("\tBi-LSTM Accuracy: ", accuracy)
    report = classification_report(y_test, y_pred, target_names=le.classes_, zero_division=0)
    print("\n\tBi-LSTM Classification Report: ", report)
    print("~~~~~\n")

    plt.figure(figsize=(12, 5))
    plt.subplot(1, 2, 1)
    plt.plot(history.history['accuracy'], label='Train Accuracy', color='lightblue', marker='o')
    plt.title('Bi-LSTM Model Accuracy progression')
    plt.xlabel('Epoch')
    plt.ylabel('Accuracy')
    plt.legend()
    plt.grid(True, linestyle='--', alpha=0.6)
```

Figure 5 Υλοποίηση του βασικού μοντέλου Bi-LSTM με χαρακτηριστικά εισόδου τα Likes και Retweets.

Όπως φαίνεται στην παραπάνω εικόνα, τα δύο αριθμητικά χαρακτηριστικά αναδιαμορφώνονται σε τριοδιάστατη μορφή και εισάγονται σε ένα αμφίδρομο επίπεδο LSTM 64 μονάδων. Η έξοδος του επιπέδου οδηγείται σε Dropout, σε πλήρως συνδεδεμένο επίπεδο και τελικά σε επίπεδο Softmax για την πολυκατηγορική ταξινόμηση.

3.8.2 Εμπλουτισμένο μοντέλο Bi-LSTM

Η δεύτερη υλοποίηση, με την ονομασία Bi-LSTM2, επεκτείνει το βασικό μοντέλο προσθέτοντας δύο ακόμη χαρακτηριστικά εισόδου:

- `hashtags_Count`,
- `Platform_encoded`.

Επομένως, το σύνολο χαρακτηριστικών περιλαμβάνει:

$$\mathbf{X}_i = [X_{\text{Likes},i}, X_{\text{Retweets},i}, X_{\text{hashtags_Count},i}, X_{\text{Platform_encoded},i}] \quad (3.13)$$

Τα δεδομένα αναδιαμορφώθηκαν σε πίνακες σχήματος:

$$(\text{samples}, 1, 4) \quad (3.14)$$

Η αρχιτεκτονική διατηρεί την ίδια βασική δομή με το προηγούμενο μοντέλο:

- Bidirectional(LSTM(64)),
- Dropout(0.2),
- Dense(32, activation="relu"),
- τελικό επίπεδο Softmax.

Η εκπαίδευση πραγματοποιήθηκε για 30 εποχές με μέγεθος παρτίδας 16. Η προσθήκη του πλήθους των hashtags και της πλατφόρμας είχε ως στόχο να εξεταστεί αν τα χαρακτηριστικά δομής και προέλευσης της ανάρτησης προσφέρουν επιπλέον πληροφορία για τη θεματική της κατηγορία.

Και σε αυτή την περίπτωση χρησιμοποιείται μόνο ένα χρονικό βήμα. Ως αποτέλεσμα, η αρχιτεκτονική λειτουργεί περισσότερο ως μη γραμμικός ταξινομητής αριθμητικών χαρακτηριστικών παρά ως μοντέλο πραγματικής χρονικής ακολουθίας.

```
def bilstm_classification2(df, le):
    X = df[['Likes', 'Retweets', 'hashtags_Count', 'Platform_encoded']].values
    y = df['primary_category_encoded'].values

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

    X_train_3d = X_train.reshape((X_train.shape[0], 1, 4))
    X_test_3d = X_test.reshape((X_test.shape[0], 1, 4))

    bilstm_model2 = Sequential([
        Input(shape=(1,4)),
        Bidirectional(LSTM(64, activation='relu')),
        Dropout(0.2),
        Dense(32, activation='relu'),
        Dense(len(le.classes_), activation='softmax')
    ])

    bilstm_model2.compile(optimizer='adam',
                          loss='sparse_categorical_crossentropy',
                          metrics=['accuracy'])

    history = bilstm_model2.fit(X_train_3d, y_train, epochs=30, batch_size=16, verbose=1)

    y_pred_probs = bilstm_model2.predict(X_test_3d)
    y_pred = np.argmax(y_pred_probs, axis=1)

    accuracy = accuracy_score(y_test, y_pred)
    print("~~~~~\n")
    print("\t-Bi-LSTM2 Accuracy: ", accuracy)
    report = classification_report(y_test, y_pred, target_names=le.classes_, zero_division=0)
    print("\n\t-Bi-LSTM2 Classification Report: ", report)
    print("~~~~~\n")

    plt.figure(figsize=(12, 5))
    plt.subplot(1, 2, 1)
    plt.plot(history.history['accuracy'], label='Train Accuracy', color='Lightgrey', marker='o')
    plt.title('Bi-LSTM2 Model Accuracy progression')
    plt.xlabel('Epoch')
    plt.ylabel('Accuracy')
    plt.legend()
```

Figure 6 Υλοποίηση του μοντέλου Bi-LSTM2 με τέσσερα αριθμητικά χαρακτηριστικά εισόδου.

Παραπάνω φαίνεται η προσθήκη των χαρακτηριστικών hashtags_Count και Platform_encoded, καθώς και η αναδιαμόρφωση των δεδομένων σε σχήμα (samples, 1, 4).

3.8.3 Υβριδικό μοντέλο Bi-LSTM πολλαπλών εισόδων

Η τρίτη υλοποίηση σχεδιάστηκε ως υβριδικό μοντέλο δύο παράλληλων εισόδων μέσω του Keras Functional API. Τα διαθέσιμα χαρακτηριστικά χωρίστηκαν σε δύο ομάδες:

$$\mathbf{X}^{(1)} = [\text{Likes}, \text{Retweets}] \quad (3.15)$$

και

$$\mathbf{X}^{(2)} = [\text{hashtags_Count}, \text{Platform_encoded}, \text{Sentiment_encoded}] \quad (3.16)$$

Η πρώτη είσοδος περιλαμβάνει τους δείκτες αλληλεπίδρασης, ενώ η δεύτερη περιλαμβάνει χαρακτηριστικά δομής και μεταδεδομένων.

Κάθε ομάδα εισόδων επεξεργάζεται από ανεξάρτητο κλάδο:

- ο πρώτος κλάδος δέχεται είσοδο σχήματος (1,2),
- ο δεύτερος κλάδος δέχεται είσοδο σχήματος (1,3).

Κάθε κλάδος περιλαμβάνει:

- Bidirectional(LSTM(64)),
- Dropout(0.2).

Οι έξοδοι των δύο κλάδων συνενώνονται μέσω της συνάρτησης concatenate. Η ενιαία αναπαράσταση εισάγεται σε πλήρως συνδεδεμένο επίπεδο 32 νευρώνων με ReLU και στη συνέχεια στο τελικό επίπεδο Softmax.

Το μοντέλο εκπαιδεύτηκε για 30 εποχές με μέγεθος παρτίδας 16. Ο διαχωρισμός των χαρακτηριστικών σε δύο κλάδους επιτρέπει στο δίκτυο να επεξεργάζεται ανεξάρτητα τις διαφορετικές ομάδες πληροφορίας πριν από τη σύνθεσή τους.

Όμως, και οι δύο εισοδοί έχουν μόνο ένα χρονικό βήμα, συνεπώς δεν αξιοποιούνται ουσιαστικά αμφίδρομες χρονικές εξαρτήσεις.

```
def bilstm_hybrid(df, le):
    X1 = df[['Likes', 'Retweets']].values
    X2 = df[['hashtags_Count', 'Platform_encoded', 'Sentiment_encoded']].values
    y = df['primary_category_encoded'].values

    X1_train, X1_test, X2_train, X2_test, y_train, y_test = train_test_split(X1, X2, y, test_size=0.2, random_state=42)

    X1_train_3d = X1_train.reshape((X1_train.shape[0], 1, 2))
    X1_test_3d = X1_test.reshape((X1_test.shape[0], 1, 2))
    X2_train_3d = X2_train.reshape((X2_train.shape[0], 1, 3))
    X2_test_3d = X2_test.reshape((X2_test.shape[0], 1, 3))

    input1 = Input(shape=(1, 2))
    lstm1 = Bidirectional(LSTM(64, activation='relu'))(input1)
    drop1 = Dropout(0.2)(lstm1)

    input2 = Input(shape=(1, 3))
    lstm2 = Bidirectional(LSTM(64, activation='relu'))(input2)
    drop2 = Dropout(0.2)(lstm2)

    merged = concatenate([drop1, drop2])

    dense1 = Dense(32, activation='relu')(merged)
    output = Dense(len(le.classes_), activation='softmax')(dense1)

    model = Model(inputs=[input1, input2], outputs=output)

    model.compile(optimizer='adam',
                  loss='sparse_categorical_crossentropy',
                  metrics=['accuracy'])

    history = model.fit([X1_train_3d, X2_train_3d], y_train, epochs=30, batch_size=16, verbose=1)

    y_pred_probs = model.predict([X1_test_3d, X2_test_3d])
    y_pred = np.argmax(y_pred_probs, axis=1)

    accuracy = accuracy_score(y_test, y_pred)
    print("~~~~~\n")
    print("\t-Bi-LSTM-hybrid Accuracy: ", accuracy)
```

Figure 7 Υλοποίηση του υβριδικού Bi-LSTM μοντέλου με δύο παράλληλους κλάδους αριθμητικών εισόδων.

Όπως παρουσιάζεται στην παραπάνω εικόνα, οι δύο ομάδες χαρακτηριστικών επεξεργάζονται από ανεξάρτητα αμφίδρομα LSTM επίπεδα. Οι παραγόμενες αναπαραστάσεις συνενώνονται πριν από το τελικό πλήρως συνδεδεμένο επίπεδο ταξινόμησης.

3.8.4 Υβριδικό μοντέλο αριθμητικών και κειμενικών δεδομένων

Η τέταρτη και πιο σύνθετη υλοποίηση συνδυάζει αριθμητικά χαρακτηριστικά με κειμενική πληροφορία. Το μοντέλο δέχεται δύο διαφορετικές εισόδους:

$$\mathbf{X}^{(\text{num})} = [\text{Likes}, \text{Retweets}] \quad (3.17)$$

και την κειμενική μεταβλητή metadata_text.

Η κειμενική είσοδος μετατράπηκε αρχικά σε ακολουθίες ακεραιών μέσω του Tokenizer. Στη συνέχεια εφαρμόστηκε padding ώστε όλες οι ακολουθίες να έχουν σταθερό μήκος:

$$L = 15.$$

A) Αριθμητικός κλάδος

Ο αριθμητικός κλάδος δέχεται είσοδο σχήματος (1,2) και περιλαμβάνει:

- Bidirectional(LSTM(128)),
- Dropout(0.3).

B) Κειμενικός κλάδος

Ο κειμενικός κλάδος δέχεται ακολουθίες μήκους 15 και περιλαμβάνει:

- επίπεδο Embedding με διάσταση εξόδου 128,
- SpatialDropout1D(0.2),
- πρώτο Bidirectional(LSTM(64, return_sequences=True)),
- δεύτερο Bidirectional(LSTM(64)),
- Dropout(0.3).

Στον συγκεκριμένο κλάδο αξιοποιείται πραγματική ακολουθιακή πληροφορία, καθώς το μοντέλο επεξεργάζεται έως 15 διαδοχικά tokens. Το πρώτο Bi-LSTM επιστρέφει ολόκληρη την ακολουθία εξόδων, η οποία εισάγεται στο δεύτερο Bi-LSTM για περαιτέρω επεξεργασία.

Γ) Συνένωση και τελική ταξινόμηση

Οι εξοδοί των δύο κλάδων συνενώνονται και ακολουθούν:

- BatchNormalization,
- πλήρως συνδεδεμένο επίπεδο 128 νευρώνων με ReLU,
- Dropout(0.3),
- τελικό επίπεδο Softmax.

Το μοντέλο μεταγλωττίστηκε με βελτιστοποιητή Adam και συνάρτηση απώλειας `sparse_categorical_crossentropy`. Η εκπαίδευση πραγματοποιήθηκε για 40 εποχές, με μέγεθος παρτίδας 32.

```
def bilstm_numeric_text(df, le):
    X1 = df[['Likes', 'Retweets']].values
    X2 = df['metadata_text']
    y = df['primary_category_encoded'].values

    tokenizer = Tokenizer()
    tokenizer.fit_on_texts(X2)
    sequences = tokenizer.texts_to_sequences(X2)
    X2_padded = pad_sequences(sequences, maxlen=15)

    vocab_size = len(tokenizer.word_index) + 1

    X1_train, X1_test, X2_train, X2_test, y_train, y_test = train_test_split(X1, X2_padded, y, test_size=0.2, random_state=42)

    X1_train_3d = X1_train.reshape((X1_train.shape[0], 1, 2))
    X1_test_3d = X1_test.reshape((X1_test.shape[0], 1, 2))

    input_num = Input(shape=(1, 2))
    lstm_num = Bidirectional(LSTM(128))(input_num)
    drop_num = Dropout(0.3)(lstm_num)

    input_text = Input(shape=(15, ))
    embed_text = Embedding(input_dim=vocab_size, output_dim=128)(input_text)
    spatial_drop = SpatialDropout1D(0.2)(embed_text)

    lstm_text_1 = Bidirectional(LSTM(64, return_sequences=True))(spatial_drop)
    lstm_text_2 = Bidirectional(LSTM(64))(lstm_text_1)
    drop_text = Dropout(0.3)(lstm_text_2)

    merged = concatenate([drop_num, drop_text])

    dense = BatchNormalization()(merged)
    dense = Dense(128, activation='relu')(dense)
    dense = Dropout(0.3)(dense)

    output = Dense(len(le.classes_), activation='softmax')(dense)

    model = Model(inputs=[input_num, input_text], outputs=output)

    model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

Figure 8 Υλοποίηση του υβριδικού μοντέλου Bi-LSTM Numeric + Text.

Όπως φαίνεται παραπάνω, το μοντέλο διαθέτει δύο διακριτούς κλάδους. Ο πρώτος επεξεργάζεται τα αριθμητικά χαρακτηριστικά αλληλεπίδρασης, ενώ ο δεύτερος επεξεργάζεται τις κειμενικές ακολουθίες μέσω Embedding και δύο διαδοχικών επιπέδων Bi-LSTM. Οι δύο αναπαραστάσεις συνενώνονται πριν από το τελικό στάδιο ταξινόμησης.

3.8.5 Αξιολόγηση των μοντέλων Bi-LSTM

Για κάθε μία από τις τέσσερις υλοποιήσεις υπολογίστηκαν οι μετρικές Accuracy, Precision, Recall και F1-score. Επιπλέον, δημιουργήθηκαν καμπύλες ακρίβειας και απώλειας κατά την εκπαίδευση, καθώς και κανονικοποιημένοι πίνακες σύγχυσης.

ΚΕΦΑΛΑΙΟ 4 Πειράματα και Αποτελέσματα

4.1 Πειραματική διαδικασία

Στο πλαίσιο της πειραματικής διαδικασίας εκπαιδεύτηκαν και αξιολογήθηκαν διαφορετικά μοντέλα Μηχανικής και Βαθιάς Μάθησης, με στόχο την πρόβλεψη της κύριας θεματικής κατηγορίας των hashtags κάθε ανάρτησης. Η μεταβλητή-στόχος ήταν η `primary_category_encoded`, η οποία προέκυψε από την αριθμητική κωδικοποίηση των οκτώ τελικών κατηγοριών.

Τα μοντέλα που εξετάστηκαν ήταν τα ακόλουθα:

- Random Forest,
- k-Nearest Neighbors,
- Support Vector Machine,
- LSTM,
- βασικό Bidirectional LSTM,
- εμπλουτισμένο Bidirectional LSTM,
- υβριδικό Bidirectional LSTM πολλών αριθμητικών εισόδων,
- υβριδικό Bidirectional LSTM αριθμητικών και κειμενικών δεδομένων.

Η χρήση πολλών διαφορετικών μοντέλων επέτρεψε τη σύγκριση κλασικών αλγορίθμων Μηχανικής Μάθησης με νευρωνικές αρχιτεκτονικές διαφορετικής πολυπλοκότητας.

4.1.1 Διαχωρισμός του συνόλου δεδομένων

Για κάθε μοντέλο πραγματοποιήθηκε διαχωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης και σύνολο δοκιμής μέσω της συνάρτησης `train_test_split`. Το 80% των εγγραφών χρησιμοποιήθηκε για την εκπαίδευση και το υπόλοιπο 20% για την τελική αξιολόγηση:

$$\text{Train set} = 80\%, \text{ Test set} = 20\% \quad (4.1)$$

Η παράμετρος `random_state` ορίστηκε ίση με 42, ώστε ο διαχωρισμός των δεδομένων να είναι επαναλήψιμος. Δεδομένου ότι το αρχικό `dataset` περιλαμβάνει 732 εγγραφές, το σύνολο δοκιμής αποτελείται από 147 δείγματα, ενώ τα υπόλοιπα 585 χρησιμοποιούνται για την εκπαίδευση.

Στον τελικό κώδικα δεν χρησιμοποιήθηκε ξεχωριστό σύνολο επικύρωσης. Επομένως, η εκπαίδευση πραγματοποιήθηκε στο `train set` και η τελική αξιολόγηση στο `test set`.

4.1.2 Σύνολα χαρακτηριστικών

Δεν χρησιμοποιήθηκε το ίδιο σύνολο χαρακτηριστικών σε όλες τις πειραματικές προσεγγίσεις. Τα βασικά μοντέλα Random Forest, kNN, SVM, LSTM και Bi-LSTM χρησιμοποιούν αποκλειστικά τις μεταβλητές:

$$X = [\text{Likes}, \text{Retweets}] \quad (4.2)$$

Το μοντέλο Bi-LSTM2 χρησιμοποιεί τέσσερα χαρακτηριστικά:

$$X = [\text{Likes}, \text{Retweets}, \text{hashtags_Count}, \text{Platform_encoded}] \quad (4.3)$$

Το υβριδικό Bi-LSTM διαχωρίζει τα χαρακτηριστικά σε δύο εισόδους. Η πρώτη περιλαμβάνει τα Likes και Retweets, ενώ η δεύτερη περιλαμβάνει τα `hashtags_Count`, `Platform_encoded` και `Sentiment_encoded`.

Τέλος, το μοντέλο Bi-LSTM Numeric + Text συνδυάζει τα αριθμητικά χαρακτηριστικά Likes και Retweets με τη μεταβλητή `metadata_text`, η οποία μετατρέπεται σε ακολουθίες ακεραίων και συμπληρώνεται σε μήκος 15 tokens.

4.1.3 Ρυθμίσεις εκπαίδευσης

Τα κλασικά μοντέλα εκπαιδεύτηκαν με τις ακόλουθες βασικές ρυθμίσεις:

Πίνακας 3 Βασικές ρυθμίσεις κλασικών μοντέλων

Μοντέλο	Κύριες παράμετροι
Random Forest	100 δέντρα, <code>class_weight="balanced"</code>
kNN	$k = 5$ γείτονες
SVM	RBF kernel, <code>class_weight="balanced"</code>

Για τα νευρωνικά δίκτυα χρησιμοποιήθηκε ο βελτιστοποιητής Adam και η συνάρτηση απώλειας `sparse_categorical_crossentropy`.

Πίνακας 4 Βασικές ρυθμίσεις νευρωνικών δικτύων

Μοντέλο	Εποχές	Μέγεθος παρτίδας
LSTM	25	16
Bi-LSTM	25	16
Bi-LSTM2	30	16
Bi-LSTM Hybrid	30	16
Bi-LSTM Numeric + Text	40	32

Η έξοδος όλων των νευρωνικών δικτύων παράγεται από πλήρως συνδεδεμένο επίπεδο με ενεργοποίηση Softmax και αριθμό νευρώνων ίσο με το πλήθος των κατηγοριών.

4.1.4 Μετρικές αξιολόγησης

Η αξιολόγηση όλων των μοντέλων πραγματοποιήθηκε με κοινές μετρικές, ώστε να είναι δυνατή η συγκριτική ανάλυση των αποτελεσμάτων. Συγκεκριμένα χρησιμοποιήθηκαν:

- Accuracy, η οποία εκφράζει το ποσοστό των σωστών προβλέψεων στο σύνολο δοκιμής.
- Precision, η οποία δείχνει πόσες από τις προβλέψεις μιας κατηγορίας ήταν σωστές.
- Recall, η οποία δείχνει πόσα από τα πραγματικά δείγματα μιας κατηγορίας αναγνωρίστηκαν σωστά.
- F1-score, το οποίο αποτελεί τον αρμονικό μέσο των Precision και Recall.

Η συνολική ακρίβεια υπολογίστηκε μέσω της `accuracy_score`, ενώ οι μετρικές ανά κλάση παρήχθησαν μέσω της `classification_report`.

Για κάθε μοντέλο δημιουργήθηκε επίσης κανονικοποιημένος πίνακας σύγχυσης:

```
confusion_matrix(y_test, y_pred, normalize="true")
```

Η κανονικοποίηση ως προς τις πραγματικές κατηγορίες επιτρέπει την παρουσίαση των αποτελεσμάτων ως αναλογίες και διευκολύνει την ερμηνεία της συμπεριφοράς του μοντέλου σε κλάσεις με διαφορετικό αριθμό δειγμάτων.

Για τα μοντέλα LSTM και Bi-LSTM καταγράφηκαν επιπλέον η ακρίβεια και η απώλεια κατά τη διάρκεια της εκπαίδευσης. Οι αντίστοιχες καμπύλες

χρησιμοποιούνται για την παρακολούθηση της εξέλιξης της μάθησης ανά εποχή.

4.1.5 Διαδικασία σύγκρισης

Η σύγκριση πραγματοποιήθηκε σε δύο επίπεδα. Αρχικά εξετάστηκε η συνολική απόδοση κάθε μοντέλου μέσω της Accuracy και των μέσων τιμών Precision, Recall και F1-score. Στη συνέχεια αναλύθηκε η απόδοση ανά κατηγορία με βάση το classification report και τον κανονικοποιημένο πίνακα σύγχυσης.

Για τα νευρωνικά μοντέλα εξετάστηκε επιπλέον η εξέλιξη της ακρίβειας και της απώλειας κατά την εκπαίδευση. Στόχος ήταν να διερευνηθεί κατά πόσο η αυξημένη πολυπλοκότητα των αρχιτεκτονικών και η προσθήκη περισσότερων χαρακτηριστικών οδηγούν σε ουσιαστική βελτίωση της ταξινόμησης.

4.2 Αποτελέσματα Random Forest

4.2.1 Συνολική απόδοση

Το μοντέλο Random Forest αξιολογήθηκε στο σύνολο δοκιμής, το οποίο περιλάμβανε 147 εγγραφές. Ως χαρακτηριστικά εισόδου χρησιμοποιήθηκαν οι μεταβλητές Likes και Retweets, ενώ ως μεταβλητή-στόχος χρησιμοποιήθηκε η κωδικοποιημένη κύρια κατηγορία των hashtags.

Η συνολική ακρίβεια του μοντέλου ήταν:

$$\text{Accuracy} = 0,1905 \quad (4.4)$$

ή, ισοδύναμα:

$$\text{Accuracy} = 19,05\% \quad (4.5)$$

Το αποτέλεσμα αυτό δείχνει ότι το μοντέλο ταξινόμησε σωστά περίπου το ένα πέμπτο των δειγμάτων του συνόλου δοκιμής.

4.2.2 Πίνακας σύγχυσης

Η συμπεριφορά του μοντέλου ανά κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα σύγχυσης.

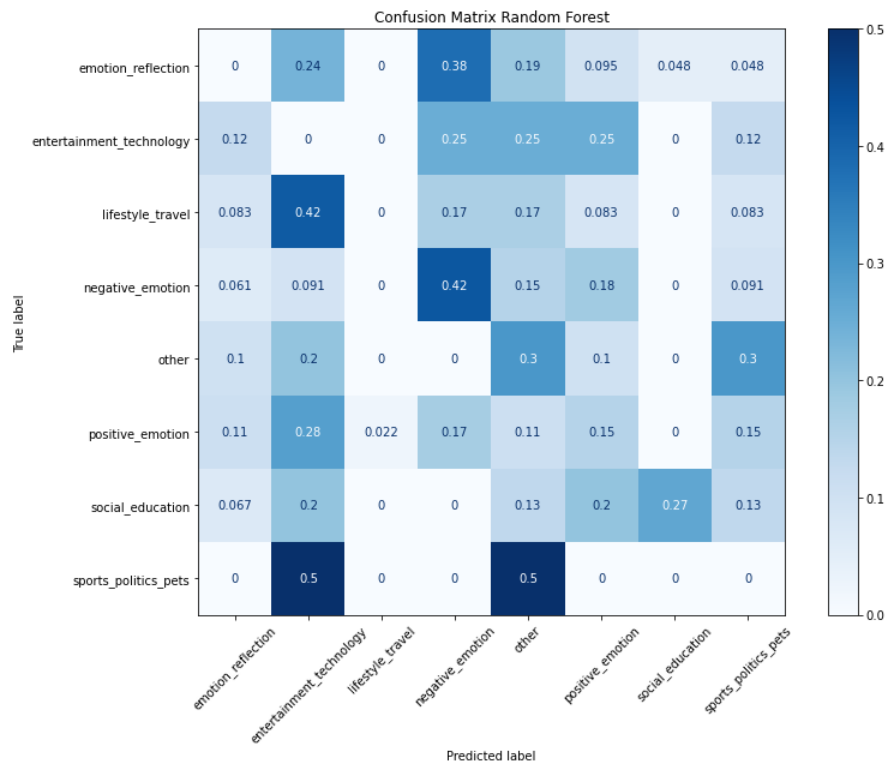


Figure 9 Κανονικοποιημένος πίνακας σύγχυσης του μοντέλου Random Forest.

Οι τιμές της κύριας διαγωνίου αντιστοιχούν στο ποσοστό των δειγμάτων κάθε πραγματικής κατηγορίας που ταξινομήθηκαν σωστά. Παρατηρείται ότι η μεγαλύτερη τιμή της διαγωνίου εμφανίζεται στην κατηγορία *negative_emotion*, με Recall 0,42. Αντίθετα, οι κατηγορίες *emotion_reflection*, *entertainment_technology*, *lifestyle_travel* και *sports_politics_pets* δεν παρουσιάζουν σωστές προβλέψεις στο συγκεκριμένο σύνολο δοκιμής.

Το μοντέλο εμφανίζει επίσης σημαντική σύγχυση μεταξύ διαφορετικών κατηγοριών. Για παράδειγμα, το 38% των δειγμάτων της κατηγορίας *emotion_reflection* προβλέφθηκε ως *negative_emotion*, ενώ το 42% των δειγμάτων της κατηγορίας *lifestyle_travel* ταξινομήθηκε ως *entertainment_technology*. Επιπλέον, τα δύο δείγματα της κατηγορίας *sports_politics_pets* προβλέφθηκαν εξ ολοκλήρου ως *entertainment_technology* ή *other*.

Η συγκέντρωση αρκετών υψηλών τιμών εκτός της κύριας διαγωνίου δείχνει ότι τα χαρακτηριστικά Likes και Retweets δεν επιτρέπουν στο μοντέλο να σχηματίσει σαφή όρια μεταξύ των θεματικών κατηγοριών.

4.2.3 Αναλυτικές μετρικές ταξινόμησης

Οι τιμές Precision, Recall και F1-score για κάθε κατηγορία παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 5 Αναλυτικές μετρικές του μοντέλου Random Forest

Κατηγορία	Precision	Recall	F1-score	Support
n	l			rt
emotion_reflection	0,00	0,00	0,00	21
entertainment_technology	0,00	0,00	0,00	8
lifestyle_travel	0,00	0,00	0,00	12

negative_emotion	0,41	0,42	0,42	33
other	0,12	0,30	0,18	10
positive_emotion	0,32	0,15	0,21	46
social_education	0,80	0,27	0,40	15
sports_politics_pets	0,00	0,00	0,00	2
Macro average	0,21	0,14	0,15	147
Weighted average	0,28	0,19	0,21	147

Η κατηγορία `social_education` παρουσιάζει την υψηλότερη τιμή Precision, ίση με 0,80. Ωστόσο, το Recall της είναι μόλις 0,27, γεγονός που δείχνει ότι το μοντέλο έκανε λίγες προβλέψεις για αυτή την κατηγορία, οι οποίες ήταν σχετικά ακριβείς, αλλά απέτυχε να αναγνωρίσει το μεγαλύτερο μέρος των πραγματικών δειγμάτων της.

Η κατηγορία `negative_emotion` εμφανίζει την καλύτερη συνολική ισορροπία μεταξύ Precision και Recall, με F1-score 0,42. Αντίθετα, τέσσερις κατηγορίες παρουσιάζουν μηδενικό F1-score, καθώς δεν αναγνωρίστηκε σωστά κανένα δείγμα τους.

Η τιμή του macro F1-score είναι 0,15, σημαντικά χαμηλότερη από το weighted F1-score 0,21. Η διαφορά αυτή οφείλεται στην ανισορροπία των κλάσεων, καθώς οι συχνότερες κατηγορίες επηρεάζουν περισσότερο τον σταθμισμένο μέσο όρο.

4.2.4 Συμπέρασμα για το μοντέλο

Τα αποτελέσματα δείχνουν ότι το Random Forest δεν επιτυγχάνει ικανοποιητική διάκριση των θεματικών κατηγοριών όταν χρησιμοποιούνται αποκλειστικά τα χαρακτηριστικά Likes και Retweets. Η συνολική ακρίβεια 19,05%, οι χαμηλές τιμές macro Recall και macro F1-score, καθώς και η έντονη σύγχυση μεταξύ των κλάσεων υποδηλώνουν ότι οι συγκεκριμένοι δείκτες αλληλεπίδρασης δεν περιέχουν επαρκή θεματική πληροφορία.

Το μοντέλο χρησιμοποιείται ως βασικό σημείο αναφοράς για τη σύγκριση με τους υπόλοιπους αλγόριθμους. Στις επόμενες ενότητες εξετάζεται κατά πόσο διαφορετικές μέθοδοι ταξινόμησης και εμπλουτισμένα σύνολα χαρακτηριστικών μπορούν να επιτύχουν καλύτερη απόδοση.

4.3 Αποτελέσματα k-Nearest Neighbors

4.3.1 Συνολική απόδοση

Το μοντέλο k-Nearest Neighbors αξιολογήθηκε στο ίδιο σύνολο δοκιμής των 147 εγγράφων, χρησιμοποιώντας ως χαρακτηριστικά εισόδου τις μεταβλητές Likes και Retweets. Ο αριθμός των πλησιέστερων γειτόνων είχε οριστεί ίσος με 5.

Η συνολική ακρίβεια του μοντέλου ήταν:

$$\text{Accuracy} = 0,2109 \quad (4.6)$$

ή:

$$\text{Accuracy} = 21,09\% \quad (4.7)$$

Η επίδοση αυτή είναι ελαφρώς υψηλότερη από την ακρίβεια του Random Forest, η οποία ήταν 19,05%. Παρόλα αυτά, το συνολικό ποσοστό σωστών προβλέψεων παραμένει χαμηλό.

4.3.2 Πίνακας σύγχυσης

Η συμπεριφορά του μοντέλου ανά κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα σύγχυσης.

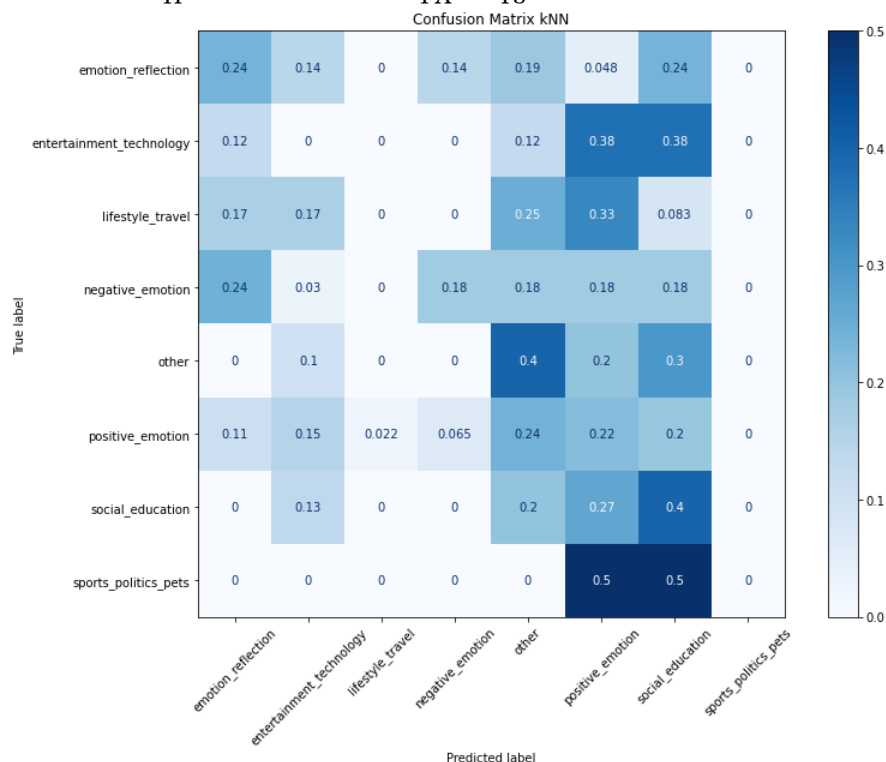


Figure 10 Κανονικοποιημένος πίνακας σύγχυσης του μοντέλου k-Nearest Neighbors.

Από τον πίνακα σύγχυσης παρατηρείται ότι το μοντέλο k-Nearest Neighbors παρουσίασε περιορισμένη ικανότητα διαχωρισμού των θεματικών κατηγοριών. Η κατηγορία other και η κατηγορία social_education εμφάνισαν το υψηλότερο Recall, με τιμή 0,40, γεγονός που σημαίνει ότι το μοντέλο αναγνώρισε σωστά το 40% των πραγματικών δειγμάτων των συγκεκριμένων κατηγοριών. Παράλληλα, η κατηγορία emotion_reflection παρουσίασε Recall 0,24, ενώ η κατηγορία positive_emotion Recall 0,22.

Αντίθετα, οι κατηγορίες entertainment_technology, lifestyle_travel και sports_politics_pets δεν παρουσίασαν καμία σωστή πρόβλεψη. Ειδικότερα, για την κατηγορία sports_politics_pets, κανένα από τα δύο διαθέσιμα δείγματα δεν ταξινομήθηκε σωστά, καθώς τα δείγματα προβλέφθηκαν ως positive_emotion και social_education.

Σημαντική σύγχυση παρατηρείται επίσης στην κατηγορία entertainment_technology, όπου το 38% των δειγμάτων ταξινομήθηκε ως positive_emotion και άλλο 38% ως social_education. Αντίστοιχα, το 33% των δειγμάτων της κατηγορίας lifestyle_travel προβλέφθηκε ως positive_emotion.

Η παρουσία αρκετών υψηλών τιμών εκτός της κύριας διαγωνίου δείχνει ότι η χρήση μόνο των χαρακτηριστικών Likes και Retweets δεν παρέχει επαρκή πληροφορία για τον σαφή διαχωρισμό των θεματικών κατηγοριών. Οι διαφορετικές κατηγορίες εμφανίζουν παρόμοια επίπεδα αλληλεπίδρασης, με αποτέλεσμα η απόσταση μεταξύ των δειγμάτων να μην αποτελεί επαρκές κριτήριο για την αξιόπιστη ταξινόμησή τους.

4.3.3 Αναλυτικές μετρικές ταξινόμησης

Πίνακας 6 Αναλυτικές μετρικές του μοντέλου k-Nearest Neighbors

Κατηγορία	Precision	Recall	F1-score	Support
n	l			rt
emotion_reflection	0,24	0,24	0,24	21
entertainment_technology	0,00	0,00	0,00	8
lifestyle_travel	0,00	0,00	0,00	12
negative_emotion	0,50	0,18	0,27	33
other	0,12	0,40	0,19	10
positive_emotion	0,32	0,22	0,26	46
social_education	0,18	0,40	0,24	15
sports_politics_pets	0,00	0,00	0,00	2
Macro average	0,17	0,18	0,15	147
Weighted average	0,27	0,21	0,21	147

Η υψηλότερη τιμή Precision εμφανίζεται στην κατηγορία negative_emotion, με τιμή 0,50. Ωστόσο, το Recall της ίδιας κατηγορίας είναι μόλις 0,18, γεγονός που δείχνει ότι το μοντέλο αναγνώρισε σωστά μικρό μέρος των πραγματικών δειγμάτων της.

Οι κατηγορίες other και social_education εμφανίζουν Recall 0,40, αλλά χαμηλές τιμές Precision, ίσες με 0,12 και 0,18 αντίστοιχα. Αυτό σημαίνει ότι το μοντέλο προβλέπει συχνά τις συγκεκριμένες κατηγορίες, αλλά σημαντικό μέρος αυτών των προβλέψεων είναι λανθασμένο.

Το macro F1-score παρέμεινε στο 0,15, ενώ το weighted F1-score ήταν 0,21. Οι τιμές αυτές δείχνουν ότι η μικρή βελτίωση της συνολικής ακρίβειας δεν συνοδεύτηκε από ουσιαστική βελτίωση της ισορροπημένης απόδοσης σε όλες τις κλάσεις.

4.3.4 Συμπέρασμα για το μοντέλο

Το kNN πέτυχε ακρίβεια 21,09%, παρουσιάζοντας μικρή βελτίωση σε σχέση με το Random Forest. Ωστόσο, η απόδοση παραμένει χαμηλή και αρκετές κατηγορίες δεν αναγνωρίστηκαν καθόλου.

Η συμπεριφορά αυτή υποδηλώνει ότι οι αναρτήσεις διαφορετικών θεματικών κατηγοριών εμφανίζουν παρόμοιες τιμές Likes και Retweets, με αποτέλεσμα να βρίσκονται κοντά μεταξύ τους στον διδιάστατο χώρο χαρακτηριστικών. Επομένως, η απόσταση μεταξύ των δειγμάτων δεν αποτελεί επαρκές κριτήριο για την αξιόπιστη πρόβλεψη της θεματικής κατηγορίας.

Επιπλέον, επειδή στην παρούσα υλοποίηση δεν εφαρμόστηκε κανονικοποίηση ή τυποποίηση των χαρακτηριστικών πριν από την εκπαίδευση, η απόδοση του kNN ενδέχεται να επηρεάζεται από τη διαφορετική αριθμητική κλίμακα των Likes και Retweets.

4.4 Αποτελέσματα Support Vector Machine

4.4.1 Συνολική απόδοση

Το μοντέλο Support Vector Machine αξιολογήθηκε στο ίδιο σύνολο δοκιμής των 147 εγγράφων, χρησιμοποιώντας ως χαρακτηριστικά εισόδου τις μεταβλητές Likes και Retweets. Για την υλοποίηση χρησιμοποιήθηκε πυρήνας Radial Basis Function και η παράμετρος class_weight="balanced" για την προσαρμογή των βαρών των κλάσεων.

Η συνολική ακρίβεια του μοντέλου ήταν:

$$\text{Accuracy} = 0,1565 \quad (4.8)$$

ή:

$$\text{Accuracy} = 15,65\% \quad (4.9)$$

Η συγκεκριμένη επίδοση είναι χαμηλότερη τόσο από την αντίστοιχη του Random Forest, η οποία ήταν 19,05%, όσο και από εκείνη του kNN, η οποία ήταν 21,09%.

4.4.2 Πίνακας σύγχυσης

Η συμπεριφορά του μοντέλου ανά κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα.

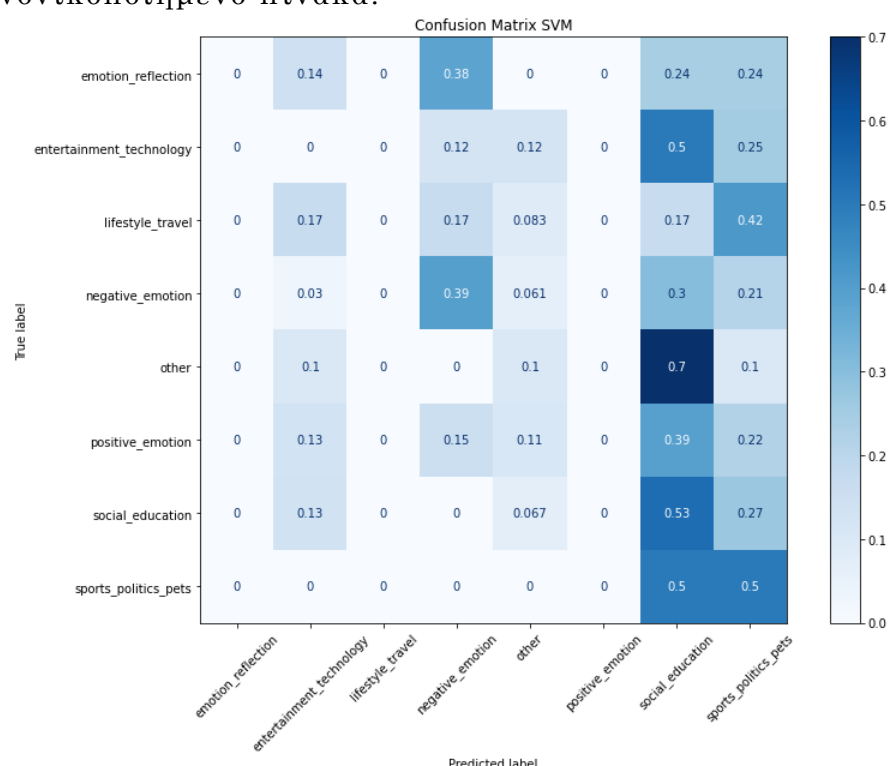


Figure 11 Κανονικοποιημένος πίνακας σύγχυσης του μοντέλου Support Vector Machine.

Ο πίνακας σύγχυσης δείχνει ότι το μοντέλο συγκέντρωσε μεγάλο μέρος των προβλέψεών του σε περιορισμένο αριθμό κατηγοριών. Η κατηγορία *social_education* παρουσίασε Recall ίσο με 0,53, ενώ η *sports_politics_pets* εμφάνισε Recall 0,50. Ωστόσο, η δεύτερη κατηγορία εκπροσωπείται από μόλις δύο δείγματα στο σύνολο δοκιμής, άρα η συγκεκριμένη τιμή πρέπει να ερμηνευτεί με ιδιαίτερη προσοχή.

Η κατηγορία *negative_emotion* παρουσίασε Recall 0,39 και αποτέλεσε την κατηγορία με την καλύτερη ισορροπία μεταξύ Precision και Recall. Αντίθετα, οι κατηγορίες *emotion_reflection*, *entertainment_technology*, *lifestyle_travel* και *positive_emotion* δεν αναγνωρίστηκαν σωστά σε κανένα δείγμα.

Ιδιαίτερα έντονη είναι η τάση του μοντέλου να ταξινομεί δείγματα ως *social_education*. Συγκεκριμένα, το 70% των δειγμάτων της κατηγορίας *other*, το 50% των δειγμάτων της *entertainment_technology*, το 39% των

δειγμάτων της `positive_emotion` και το 30% των δειγμάτων της `negative_emotion` προβλέφθηκαν ως `social_education`.

Παρατηρείται επίσης ότι το 38% των δειγμάτων της κατηγορίας `emotion_reflection` προβλέφθηκε ως `negative_emotion`, ενώ το 42% των δειγμάτων της `lifestyle_travel` ταξινομήθηκε ως `sports_politics_pets`. Η συγκέντρωση υψηλών τιμών εκτός της κύριας διαγωνίου επιβεβαιώνει τη δυσκολία του μοντέλου να διαχωρίσει αποτελεσματικά τις οκτώ θεματικές κατηγορίες.

4.4.3 Αναλυτικές μετρικές ταξινόμησης

Οι τιμές Precision, Recall και F1-score για κάθε κατηγορία παρουσιάζονται παρακάτω.

Πίνακας 7 Αναλυτικές μετρικές του μοντέλου Support Vector Machine

Κατηγορία	Precisio n	Recal l	F1- score	Suppo rt
<code>emotion_reflection</code>	0,00	0,00	0,00	21
<code>entertainment_technology</code>	0,00	0,00	0,00	8
<code>lifestyle_travel</code>	0,00	0,00	0,00	12
<code>negative_emotion</code>	0,42	0,39	0,41	33
<code>other</code>	0,09	0,10	0,10	10
<code>positive_emotion</code>	0,00	0,00	0,00	46
<code>social_education</code>	0,15	0,53	0,23	15
<code>sports_politics_pets</code>	0,03	0,50	0,05	2
Macro average	0,09	0,19	0,10	147
Weighted average	0,12	0,16	0,12	147

Η κατηγορία `negative_emotion` παρουσίασε την καλύτερη συνολική επίδοση, με Precision 0,42, Recall 0,39 και F1-score 0,41. Το αποτέλεσμα αυτό δείχνει ότι το μοντέλο αναγνώρισε ένα μέρος των δειγμάτων της συγκεκριμένης κατηγορίας με σχετικά περιορισμένο αριθμό λανθασμένων θετικών προβλέψεων.

Η κατηγορία `social_education` εμφάνισε υψηλότερο Recall, ίσο με 0,53, αλλά χαμηλό Precision, ίσο με 0,15. Αυτό σημαίνει ότι το μοντέλο εντόπισε περισσότερα από τα μισά πραγματικά δείγματα της κατηγορίας, αλλά παράλληλα απέδωσε λανθασμένα την ίδια ετικέτα σε μεγάλο αριθμό δειγμάτων άλλων κατηγοριών.

Αντίστοιχα, το Recall 0,50 της κατηγορίας `sports_politics_pets` συνοδεύεται από Precision μόλις 0,03. Συνεπώς, παρότι αναγνωρίστηκε σωστά ένα από τα δύο πραγματικά δείγματα, σχεδόν όλες οι προβλέψεις του μοντέλου για αυτή την κατηγορία ήταν λανθασμένες.

Το macro F1-score ήταν 0,10 και το weighted F1-score 0,12. Οι χαμηλές αυτές τιμές δείχνουν ότι το μοντέλο δεν παρουσίασε σταθερή απόδοση στο σύνολο των κατηγοριών. Το μηδενικό F1-score σε τέσσερις από τις οκτώ κλάσεις επιβεβαιώνει ότι η συνολική ικανότητα διάκρισης ήταν περιορισμένη.

4.4.4 Συμπέρασμα για το μοντέλο

Το μοντέλο SVM πέτυχε συνολική ακρίβεια 15,65%, η οποία αποτελεί τη χαμηλότερη επίδοση μεταξύ των τριών κλασικών αλγορίθμων που έχουν εξεταστεί μέχρι αυτό το σημείο.

Η απόδοσή του δείχνει ότι τα χαρακτηριστικά Likes και Retweets δεν παρέχουν επαρκή πληροφορία για τον σαφή διαχωρισμό των θεματικών κατηγοριών. Το μοντέλο παρουσίασε έντονη τάση πρόβλεψης συγκεκριμένων κλάσεων, κυρίως των `social_education`, `negative_emotion` και `sports_politics_pets`, ενώ απέτυχε να αναγνωρίσει τέσσερις κατηγορίες.

Επιπρόσθετα, πριν από την εφαρμογή του SVM δεν πραγματοποιήθηκε τυποποίηση των αριθμητικών χαρακτηριστικών. Καθώς οι Μηχανές Διανυσμάτων Υποστήριξης επηρεάζονται από την κλίμακα των μεταβλητών εισόδου, η απουσία `scaling` ενδέχεται να έχει επηρεάσει αρνητικά την απόδοση.

4.5 Αποτελέσματα LSTM

4.5.1 Συνολική απόδοση

Το μοντέλο LSTM αξιολογήθηκε στο ίδιο σύνολο δοκιμής των 147 εγγράφων. Ως είσοδος χρησιμοποιήθηκαν τα αριθμητικά χαρακτηριστικά Likes και Retweets, τα οποία αναδιαμορφώθηκαν σε τρισδιάστατο πίνακα ώστε να αποτελέσουν είσοδο του αναδρομικού νευρωνικού δικτύου.

Η εκπαίδευση πραγματοποιήθηκε για 25 εποχές (epochs), με βελτιστοποιητή Adam, συνάρτηση απώλειας Sparse Categorical Crossentropy και μέγεθος παρτίδας (batch size) ίσο με 16.

Η συνολική ακρίβεια του μοντέλου ήταν:

$$\text{Accuracy} = 0,3605\% \quad (4.10)$$

ή ισοδύναμα:

$$\text{Accuracy} = 36,05\% \quad (4.11)$$

Η επίδοση αυτή αποτελεί βελτίωση σε σχέση με τους κλασικούς αλγόριθμους που εξετάστηκαν προηγουμένως, παρουσιάζοντας υψηλότερη ακρίβεια από τα μοντέλα Random Forest, kNN και SVM. Ωστόσο, η πραγματική απόδοση του μοντέλου πρέπει να εξεταστεί σε συνδυασμό με τις επιμέρους μετρικές και τον πίνακα σύγχυσης, καθώς η συνολική ακρίβεια επηρεάζεται σημαντικά από την κατανομή των προβλέψεων στις διαφορετικές κατηγορίες.

4.5.2 Εξέλιξη της εκπαίδευσης

Η πορεία της ακρίβειας και της συνάρτησης απώλειας κατά τη διάρκεια της εκπαίδευσης παρουσιάζεται στο γράφημα.

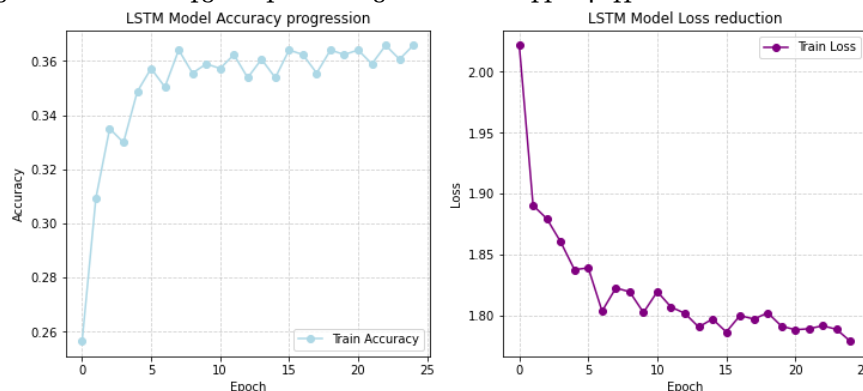


Figure 12 Εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά την εκπαίδευση του μοντέλου LSTM.

Από το διάγραμμα παρατηρείται ότι η ακρίβεια του μοντέλου αυξάνεται σημαντικά κατά τις πρώτες εποχές εκπαίδευσης, από περίπου 25% σε τιμές κοντά στο 36%. Μετά τις πρώτες εποχές η καμπύλη σταθεροποιείται, παρουσιάζοντας μικρές διακυμάνσεις γύρω από την τελική τιμή της.

Παράλληλα, η συνάρτηση απώλειας μειώνεται σταδιακά από περίπου 2,02 σε περίπου 1,78, γεγονός που δείχνει ότι το μοντέλο συγκλίνει ομαλά κατά τη διάρκεια της εκπαίδευσης χωρίς να εμφανίζει έντονες αστάθειες.

Η περιορισμένη βελτίωση μετά τις πρώτες εποχές υποδηλώνει ότι το μοντέλο έχει ήδη μάθει τα βασικά πρότυπα που υπάρχουν στα διαθέσιμα χαρακτηριστικά εισόδου και οι επιπλέον εποχές δεν οδηγούν σε σημαντική αύξηση της απόδοσης.

4.5.3 Πίνακας σύγχυσης

Η συμπεριφορά του μοντέλου ανά κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα σύγχυσης.

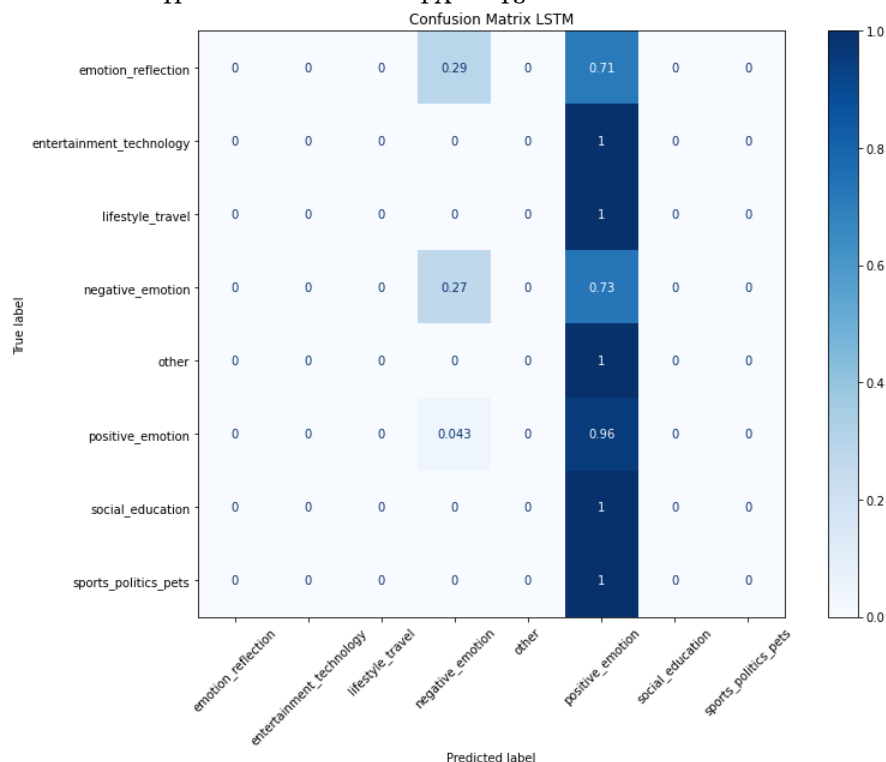


Figure 13 Κανονικοποιημένος πίνακας σύγχυσης του μοντέλου LSTM.

Ο πίνακας σύγχυσης δείχνει ότι το μοντέλο συγκεντρώνει μεγάλο μέρος των προβλέψεών του στην κατηγορία `positive_emotion`.

Συγκεκριμένα, όλα τα δείγματα των κατηγοριών `entertainment_technology`, `lifestyle_travel`, `other`, `social_education` και `sports_politics_pets` ταξινομήθηκαν ως `positive_emotion`. Παρόμοια συμπεριφορά εμφανίζεται και στις κατηγορίες `emotion_reflection` και `negative_emotion`, όπου το μεγαλύτερο μέρος των δειγμάτων μεταφέρεται επίσης στην ίδια κατηγορία.

Η κατηγορία `positive_emotion` είναι η μόνη που αναγνωρίζεται σε μεγάλο βαθμό, με το 96% των πραγματικών δειγμάτων της να ταξινομείται σωστά. Αντίστοιχα, η κατηγορία `negative_emotion` παρουσιάζει μικρότερη αλλά

υπαρκτή ικανότητα αναγνώρισης, με περίπου 27% των δειγμάτων της να προβλέπονται σωστά.

Παρότι το μοντέλο παρουσιάζει υψηλότερη συνολική ακρίβεια σε σχέση με τα προηγούμενα μοντέλα, ο πίνακας σύγχυσης αποκαλύπτει έντονη τάση προς την κυρίαρχη κατηγορία *positive_emotion*, περιορίζοντας την ικανότητα γενίκευσης στις υπόλοιπες κλάσεις.

4.5.4 Αναλυτικές μετρικές ταξινόμησης

Οι τιμές Precision, Recall και F1-score παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 8 Αναλυτικές μετρικές του μοντέλου LSTM

Κατηγορία	Precisio n	Recal l	F1- score	Suppo rt
emotion_reflection	0,00	0,00	0,00	21
entertainment_technology	0,00	0,00	0,00	8
lifestyle_travel	0,00	0,00	0,00	12
negative_emotion	0,53	0,27	0,36	33
other	0,00	0,00	0,00	10
positive_emotion	0,34	0,96	0,50	46
social_education	0,00	0,00	0,00	15
sports_politics_pets	0,00	0,00	0,00	2
Macro average	0,11	0,15	0,11	147
Weighted average	0,22	0,36	0,24	147

Η κατηγορία *positive_emotion* εμφανίζει Recall ίσο με 0,96, γεγονός που σημαίνει ότι η πλειονότητα των πραγματικών δειγμάτων της συγκεκριμένης κατηγορίας αναγνωρίστηκε σωστά. Ωστόσο, η τιμή Precision είναι μόλις 0,34, καθώς σημαντικός αριθμός δειγμάτων από άλλες κατηγορίες ταξινομήθηκε επίσης ως *positive_emotion*.

Η κατηγορία *negative_emotion* παρουσιάζει Precision 0,53 και F1-score 0,36, αποτελώντας τη δεύτερη καλύτερη κατηγορία του μοντέλου. Παρόλα αυτά, το Recall της παραμένει χαμηλό (0,27), γεγονός που δείχνει ότι μόνο ένα μικρό μέρος των πραγματικών δειγμάτων της αναγνωρίστηκε σωστά.

Αντίθετα, οι υπόλοιπες έξι κατηγορίες εμφανίζουν μηδενικές τιμές Precision, Recall και F1-score, καθώς το μοντέλο δεν κατάφερε να αναγνωρίσει σωστά κανένα δείγμα τους.

Αν και η συνολική Accuracy έφτασε το 36,05%, οι τιμές του macro F1-score (0,11) και του macro Recall (0,15) παραμένουν χαμηλές. Το γεγονός αυτό δείχνει ότι η απόδοση του μοντέλου δεν είναι ισορροπημένη μεταξύ των διαφορετικών κλάσεων και επηρεάζεται κυρίως από την επιτυχή πρόβλεψη της κατηγορίας *positive_emotion*.

4.5.5 Συμπέρασμα για το μοντέλο

Το LSTM πέτυχε Accuracy 36,05%, παρουσιάζοντας την υψηλότερη απόδοση μέχρι στιγμής σε σχέση με τα κλασικά μοντέλα Random Forest, kNN και SVM. Ωστόσο, η αναλυτική αξιολόγηση μέσω του πίνακα σύγχυσης και των επιμέρους μετρικών δείχνει ότι το μοντέλο παρουσιάζει έντονη προτίμηση προς την κατηγορία *positive_emotion*.

Η συμπεριφορά αυτή πιθανότατα σχετίζεται με την ανισορροπία του συνόλου δεδομένων, αλλά και με το γεγονός ότι η είσοδος του μοντέλου αποτελείται αποκλειστικά από τα χαρακτηριστικά Likes και Retweets, τα

οποία δεν περιέχουν επαρκή σημασιολογική πληροφορία για τον διαχωρισμό των θεματικών κατηγοριών.

Παρόλα αυτά, η βελτίωση της συνολικής ακρίβειας σε σχέση με τα προηγούμενα μοντέλα δείχνει ότι τα αναδρομικά νευρωνικά δίκτυα μπορούν να αξιοποιήσουν αποτελεσματικότερα τις διαθέσιμες σχέσεις μεταξύ των χαρακτηριστικών. Τα αποτελέσματα αυτά αποτέλεσαν τη βάση για την επόμενη αξιολόγηση πιο σύνθετων αρχιτεκτονικών Bi-LSTM και υβριδικών μοντέλων με εμπλουτισμένα χαρακτηριστικά.

4.6 Αποτελέσματα μοντέλων Bi-LSTM

4.6.1 Αποτελέσματα βασικού Bi-LSTM

4.6.1.1 Συνολική απόδοση

Το βασικό μοντέλο Bidirectional Long Short-Term Memory αξιολογήθηκε στο ίδιο σύνολο δοκιμής των 147 εγγράφων. Ως χαρακτηριστικά εισόδου χρησιμοποιήθηκαν αποκλειστικά οι αριθμητικές μεταβλητές Likes και Retweets, οι οποίες αναδιαμορφώθηκαν σε τρισδιάστατη μορφή πριν από την εισαγωγή τους στο αμφίδρομο αναδρομικό νευρωνικό δίκτυο.

Η εκπαίδευση πραγματοποιήθηκε για 25 εποχές, με βελτιστοποιητή Adam, συνάρτηση απώλειας `sparse_categorical_crossentropy` και μέγεθος παρτίδας ίσο με 16.

Η συνολική ακρίβεια του μοντέλου ήταν:

$$\text{Accuracy} = 0,3605\% \quad (4.12)$$

ή, ισοδύναμα:

$$\text{Accuracy} = 36,05\% \quad (4.13)$$

Η απόδοση αυτή είναι αντίστοιχη με αυτήν του απλού μοντέλου LSTM, το οποίο επίσης πέτυχε ακρίβεια 36,05%. Εντούτοις, παραμένει υψηλότερη από τις επιδόσεις των κλασικών μοντέλων Random Forest, kNN και SVM.

4.6.1.2 Εξέλιξη της εκπαίδευσης

Η εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά τη διάρκεια των 25 εποχών παρουσιάζεται στο παρακάτω γράφημα.

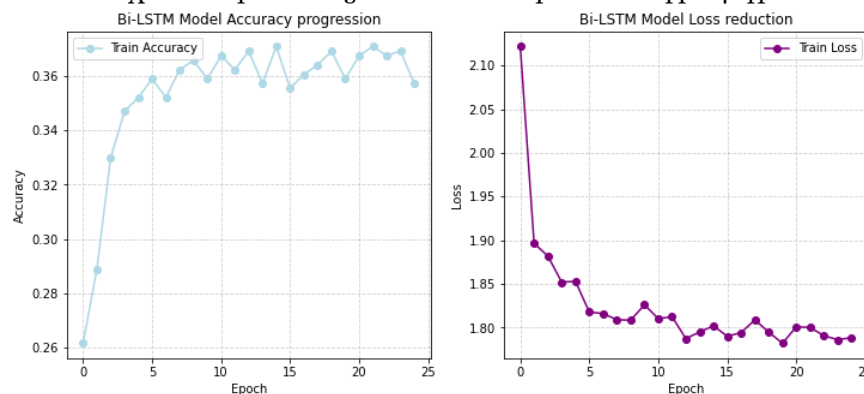


Figure 14 Εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά την εκπαίδευση του βασικού μοντέλου Bi-LSTM.

Η ακρίβεια εκπαίδευσης αυξάνεται σημαντικά κατά τις πρώτες εποχές, από περίπου 26% στην πρώτη εποχή σε τιμές κοντά στο 35% ήδη από τις πρώτες επαναλήψεις. Η υψηλότερη καταγεγραμμένη τιμή εμφανίζεται περίπου στη δέκατη πέμπτη εποχή και προσεγγίζει το 37%.

Μετά το αρχικό στάδιο βελτίωσης, η ακρίβεια παρουσιάζει μικρές διακυμάνσεις και σταθεροποιείται κυρίως μεταξύ 35% και 37%. Η τελική τιμή ακρίβειας στο σύνολο εκπαίδευσης διαμορφώθηκε περίπου στο 36%.

Παράλληλα, η συνάρτηση απώλειας μειώνεται από περίπου 2,12 στην πρώτη εποχή σε τιμές κοντά στο 1,80 μετά τις πρώτες εποχές. Στη συνέχεια, η καμπύλη παραμένει σχετικά σταθερή, με μικρές διακυμάνσεις γύρω από την τελική τιμή.

Η συγκεκριμένη συμπεριφορά δείχνει ότι το μοντέλο συγκλίνει σχετικά γρήγορα και ότι οι επιπλέον εποχές μετά το αρχικό στάδιο δεν οδηγούν σε σημαντική περαιτέρω βελτίωση. Όμως, επειδή στον κώδικα δεν καταγράφονται ξεχωριστές καμπύλες επικύρωσης, δεν είναι δυνατή η ασφαλής αξιολόγηση της γενίκευσης ή της πιθανής υπερπροσαρμογής μόνο από τις συγκεκριμένες καμπύλες εκπαίδευσης.

4.6.1.3 Πίνακας σύγχυσης

Η συμπεριφορά του βασικού μοντέλου Bi-LSTM ανά θεματική κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα σύγχυσης.

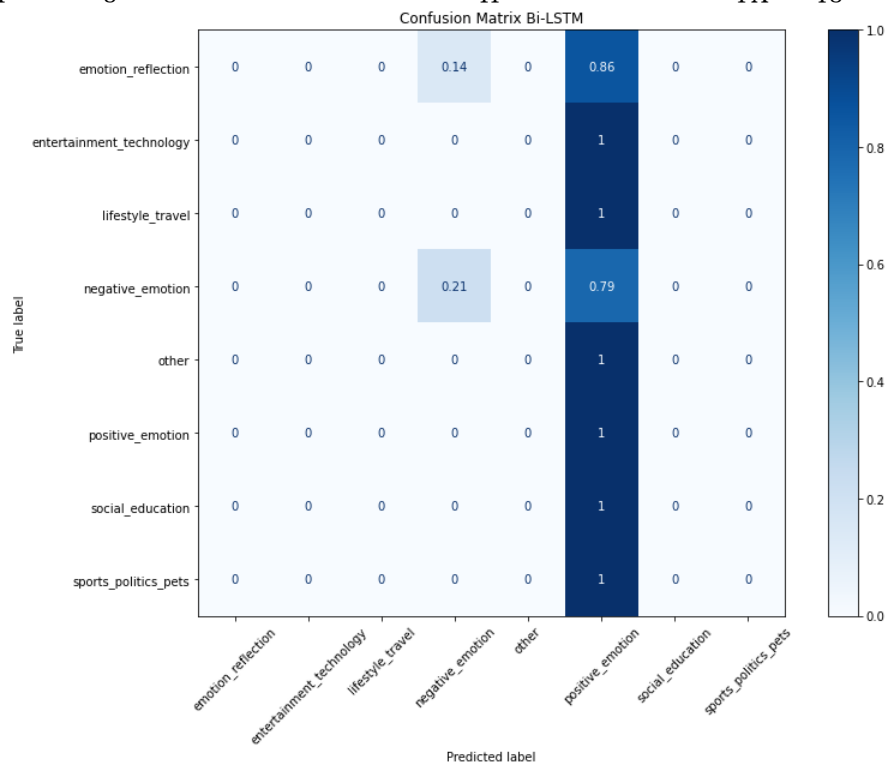


Figure 15 Κανονικοποιημένος πίνακας σύγχυσης του βασικού μοντέλου Bi-LSTM.

Ο πίνακας σύγχυσης δείχνει ότι το μοντέλο συγκεντρώνει σχεδόν όλες τις προβλέψεις του στην κατηγορία `positive_emotion`.

Συγκεκριμένα:

- το 86% των δειγμάτων της κατηγορίας `emotion_reflection`,
- το σύνολο των δειγμάτων των κατηγοριών `entertainment_technology`, `lifestyle_travel`, `other`, `positive_emotion`, `social_education`, και `sports_politics_pets` προβλέφθηκαν ως `positive_emotion`.

Η μοναδική κατηγορία εκτός της `positive_emotion` που παρουσιάζει σωστές προβλέψεις είναι η `negative_emotion`, όπου το 21% των πραγματικών δειγμάτων ταξινομήθηκε σωστά.

Παράλληλα, το 79% των δειγμάτων της *negative_emotion* ταξινομήθηκε λανθασμένα ως *positive_emotion*, ενώ το 14% των δειγμάτων της *emotion_reflection* ταξινομήθηκε ως *negative_emotion*.

Η εικόνα αυτή δείχνει ότι το μοντέλο δεν κατορθώνει να σχηματίσει διακριτά όρια μεταξύ των θεματικών κατηγοριών και παρουσιάζει έντονη προτίμηση προς την κατηγορία *positive_emotion*.

4.6.1.4 Αναλυτικές μετρικές ταξινόμησης

Οι τιμές Precision, Recall και F1-score του βασικού μοντέλου Bi-LSTM παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 9 Αναλυτικές μετρικές του βασικού μοντέλου Bi-LSTM

Κατηγορία	Precision	Recall	F1-score	Support
emotion_reflection	0,00	0,00	0,00	21
entertainment_technology	0,00	0,00	0,00	8
lifestyle_travel	0,00	0,00	0,00	12
negative_emotion	0,70	0,21	0,33	33
other	0,00	0,00	0,00	10
positive_emotion	0,34	1,00	0,50	46
social_education	0,00	0,00	0,00	15
sports_politics_pets	0,00	0,00	0,00	2
Macro average	0,13	0,15	0,10	147
Weighted average	0,26	0,36	0,23	147

Η κατηγορία *positive_emotion* παρουσιάζει το υψηλότερο Recall, ίσο με 1,00, γεγονός που σημαίνει ότι όλα τα πραγματικά δείγματα της συγκεκριμένης κατηγορίας αναγνωρίστηκαν σωστά. Ωστόσο, η τιμή Precision είναι μόλις 0,34, καθώς πολλά δείγματα άλλων κατηγοριών προβλέφθηκαν λανθασμένα ως θετικά συναισθήματα.

Η κατηγορία *negative_emotion* εμφανίζει Precision 0,70, Recall 0,21 και F1-score 0,33. Παρότι η Precision είναι η υψηλότερη μεταξύ όλων των κατηγοριών, το χαμηλό Recall δείχνει ότι το μοντέλο αναγνωρίζει σωστά μόνο μικρό μέρος των πραγματικών δειγμάτων της.

Οι υπόλοιπες έξι κατηγορίες παρουσιάζουν μηδενικές τιμές Precision, Recall και F1-score, καθώς δεν αναγνωρίστηκε σωστά κανένα δείγμα τους.

Το macro F1-score διαμορφώθηκε στο 0,10, ενώ το weighted F1-score στο 0,23. Η διαφορά μεταξύ των δύο τιμών δείχνει ότι η συνολική απόδοση επηρεάζεται έντονα από τις πολυπληθέστερες κατηγορίες. Το χαμηλό macro F1-score επιβεβαιώνει ότι το μοντέλο δεν επιτυγχάνει ισορροπημένη απόδοση στο σύνολο των οκτώ κλάσεων.

4.6.1.5 Συμπέρασμα για το μοντέλο

Το βασικό μοντέλο Bi-LSTM πέτυχε συνολική ακρίβεια 36,05%, εμφανίζοντας παρόμοια απόδοση με το απλό μοντέλο LSTM και υψηλότερη απόδοση σε σχέση με τα κλασικά μοντέλα Μηχανικής Μάθησης.

Η ανάλυση των αποτελεσμάτων δείχνει ότι η χρήση αμφίδρομης επεξεργασίας δεν οδήγησε σε ουσιαστική βελτίωση σε σχέση με το LSTM. Το γεγονός αυτό μπορεί να συνδέεται με τη μορφή των δεδομένων εισόδου, καθώς τα Likes και Retweets αναδιαμορφώθηκαν σε ακολουθία ενός μόνο χρονικού βήματος. Επομένως, το μοντέλο δεν είχε τη δυνατότητα να αξιοποιήσει πραγματικές χρονικές εξαρτήσεις.

Παρά την υψηλότερη Accuracy σε σχέση με τους κλασικούς αλγορίθμους, το μοντέλο εμφάνισε έντονη συγκέντρωση προβλέψεων στην κατηγορία `positive_emotion`. Η μηδενική αναγνώριση έξι από τις οκτώ κατηγορίες και το χαμηλό macro F1-score δείχνουν ότι η συνολική του απόδοση δεν είναι ισορροπημένη.

Τα αποτελέσματα αυτά επιβεβαιώνουν την ανάγκη προσθήκης περισσότερων χαρακτηριστικών και διαφορετικών τύπων πληροφορίας, όπως κειμενικά χαρακτηριστικά και μεταδεδομένα, ώστε το μοντέλο να μπορεί να αξιοποιήσει πλουσιότερες αναπαραστάσεις των αναρτήσεων.

4.6.2 Αποτελέσματα εμπλουτισμένου Bi-LSTM2

4.6.2.1 Συνολική απόδοση

Το εμπλουτισμένο μοντέλο Bi-LSTM2 αξιολογήθηκε στο ίδιο σύνολο δοκιμής των 147 εγγραφών. Σε αντίθεση με το βασικό Bi-LSTM, το μοντέλο χρησιμοποίησε τέσσερα χαρακτηριστικά εισόδου:

- Likes
- Retweets
- `hashtags_Count`
- `Platform_encoded`

Τα δεδομένα αναδιαμορφώθηκαν σε τρισδιάστατη μορφή, ώστε να εισαχθούν στο αμφίδρομο αναδρομικό δίκτυο. Η εκπαίδευση πραγματοποιήθηκε για 30 εποχές, με βελτιστοποιητή Adam, συνάρτηση απώλειας `sparse_categorical_crossentropy` και μέγεθος παρτίδας 16.

Η συνολική ακρίβεια του μοντέλου ήταν:

$$\text{Accuracy} = 0,3537\% \quad (4.14)$$

ή:

$$\text{Accuracy} = 35,37\% \quad (4.15)$$

Η επίδοση αυτή είναι ελαφρώς χαμηλότερη από την αντίστοιχη του βασικού Bi-LSTM, το οποίο πέτυχε ακρίβεια 36,05%, καθώς και από το απλό LSTM. Συνεπώς, η προσθήκη των χαρακτηριστικών `hashtags_Count` και `Platform_encoded` δεν οδήγησε σε βελτίωση της συνολικής απόδοσης.

4.6.2.2 Εξέλιξη της εκπαίδευσης

Η εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά τη διάρκεια των 30 εποχών παρουσιάζεται στα ακόλουθα γραφήματα.

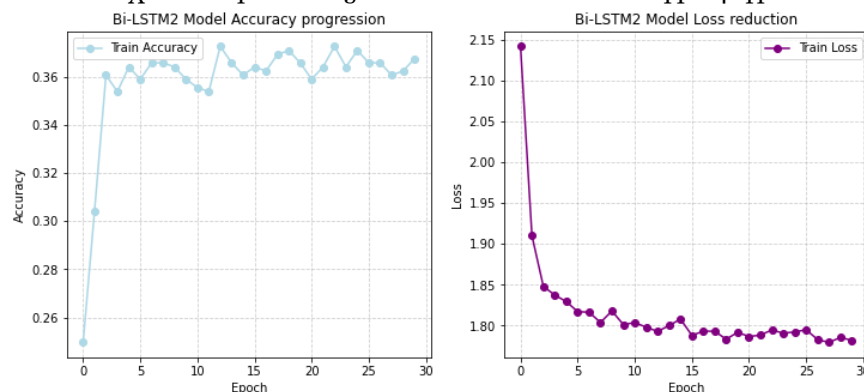


Figure 16 Εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά την εκπαίδευση του μοντέλου Bi-LSTM2.

Η ακρίβεια εκπαίδευσης αυξάνεται γρήγορα κατά τις πρώτες εποχές, από περίπου 25% στην πρώτη εποχή σε τιμές άνω του 35% μέχρι την τρίτη εποχή. Η υψηλότερη τιμή εμφανίζεται περίπου στη δέκατη τρίτη εποχή και φτάνει περίπου το 37,26%.

Μετά το αρχικό στάδιο βελτίωσης, η ακρίβεια σταθεροποιείται κυρίως μεταξύ 35% και 37%, χωρίς σαφή ανοδική τάση προς το τέλος της εκπαίδευσης. Η τελική τιμή ακρίβειας στο σύνολο εκπαίδευσης διαμορφώθηκε περίπου στο 36,75%.

Η συνάρτηση απώλειας μειώνεται από περίπου 2,14 στην πρώτη εποχή σε τιμές κοντά στο 1,78 στο τέλος της εκπαίδευσης. Μετά τις πρώτες εποχές παρουσιάζει μικρές διακυμάνσεις γύρω από την τελική τιμή.

Η συμπεριφορά αυτή δείχνει ότι το μοντέλο συγκλίνει σχετικά γρήγορα και ότι η αύξηση των εποχών δεν οδηγεί σε σημαντική περαιτέρω βελτίωση. Επειδή δεν καταγράφεται ξεχωριστή καμπύλη επικύρωσης, δεν είναι δυνατή η ασφαλής αξιολόγηση πιθανής υπερπροσαρμογής μόνο από τις καμπύλες εκπαίδευσης.

4.6.2.3 Πίνακας σύγχυσης

Η συμπεριφορά του μοντέλου ανά κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα σύγχυσης.

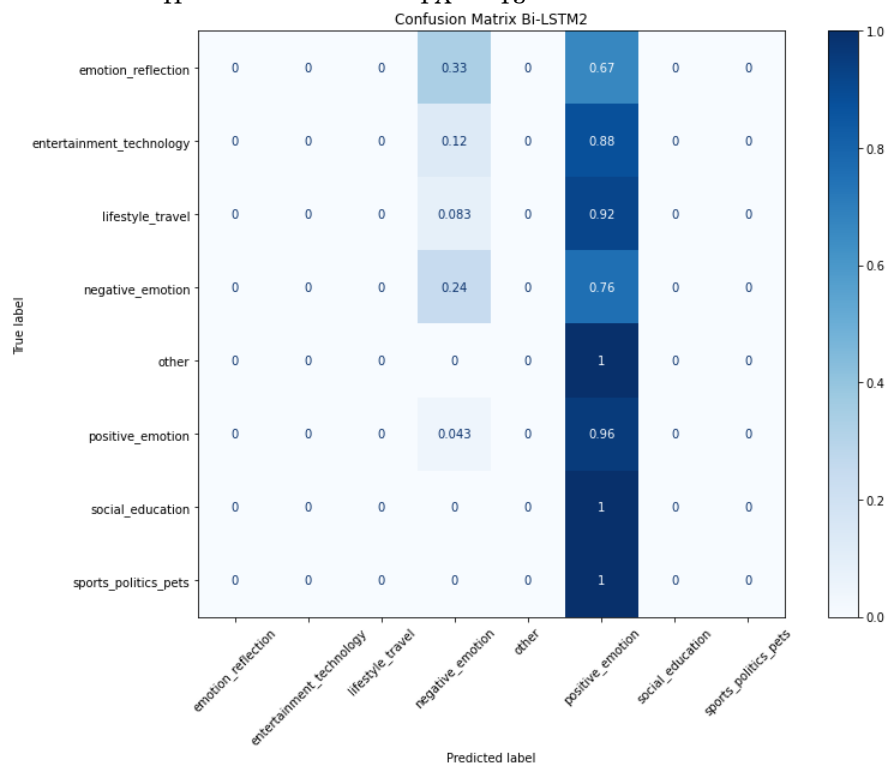


Figure 17 Κανονικοποιημένος πίνακας σύγχυσης του μοντέλου Bi-LSTM2.

Ο πίνακας σύγχυσης δείχνει ότι το μοντέλο συγκεντρώνει σχεδόν όλες τις προβλέψεις του στην κατηγορία `positive_emotion`.

Συγκεκριμένα:

- το 67% των δειγμάτων της `emotion_reflection`,
- το 88% της `entertainment_technology`,
- το 92% της `lifestyle_travel`,
- το 76% της `negative_emotion`,

- και το σύνολο των δειγμάτων των κατηγοριών other, social_education και sports_politics_pets προβλέφθηκαν ως positive_emotion.

Η κατηγορία positive_emotion παρουσιάζει υψηλή ικανότητα αναγνώρισης, καθώς το 96% των πραγματικών δειγμάτων της ταξινομήθηκε σωστά.

Παράλληλα, η κατηγορία negative_emotion αποτελεί τη δεύτερη κατηγορία στην οποία εμφανίζονται σωστές προβλέψεις, με Recall 0,24.

Η έντονη συγκέντρωση των προβλέψεων στην positive_emotion δείχνει ότι η προσθήκη των χαρακτηριστικών hashtags_Count και Platform_encoded δεν ήταν αρκετή για να βελτιώσει ουσιαστικά τον διαχωρισμό μεταξύ των θεματικών κατηγοριών.

4.6.2.4 Αναλυτικές μετρικές ταξινόμησης

Πίνακας 10 Αναλυτικές μετρικές του μοντέλου Bi-LSTM2

Κατηγορία	Precisio n	Recal l	F1- score	Suppo rt
emotion_reflection	0,00	0,00	0,00	21
entertainment_technology	0,00	0,00	0,00	8
lifestyle_travel	0,00	0,00	0,00	12
negative_emotion	0,42	0,24	0,31	33
other	0,00	0,00	0,00	10
positive_emotion	0,34	0,96	0,51	46
social_education	0,00	0,00	0,00	15
sports_politics_pets	0,00	0,00	0,00	2
Macro average	0,10	0,15	0,10	147
Weighted average	0,20	0,35	0,23	147

Η κατηγορία positive_emotion παρουσιάζει το υψηλότερο Recall, ίσο με 0,96, γεγονός που σημαίνει ότι η πλειονότητα των πραγματικών δειγμάτων της συγκεκριμένης κατηγορίας αναγνωρίστηκε σωστά. Από την άλλη πλευρά, η Precision είναι μόλις 0,34, επειδή αρκετά δείγματα άλλων κατηγοριών ταξινομήθηκαν επίσης ως positive_emotion.

Η κατηγορία negative_emotion εμφανίζει Precision 0,42, Recall 0,24 και F1-score 0,31. Οι τιμές αυτές δείχνουν ότι το μοντέλο μπορεί να αναγνωρίσει περιορισμένο αριθμό δειγμάτων της συγκεκριμένης κατηγορίας, αλλά η συνολική του απόδοση παραμένει χαμηλή.

Οι υπόλοιπες έξι κατηγορίες παρουσιάζουν μηδενικές τιμές Precision, Recall και F1-score, καθώς δεν αναγνωρίστηκε σωστά κανένα δείγμα τους.

Το macro F1-score διαμορφώθηκε στο 0,10 και το weighted F1-score στο 0,23. Οι τιμές αυτές δείχνουν ότι η απόδοση του μοντέλου επηρεάζεται σημαντικά από τις κυρίαρχες κατηγορίες και δεν παρουσιάζει ισορροπημένη συμπεριφορά μεταξύ όλων των κλάσεων.

4.6.2.5 Συμπέρασμα για το μοντέλο

Το μοντέλο Bi-LSTM2 πέτυχε συνολική ακρίβεια 35,37%, παρουσιάζοντας ελαφρώς χαμηλότερη απόδοση σε σχέση με το βασικό Bi-LSTM και το απλό LSTM.

Παρόλο που χρησιμοποιήθηκαν δύο πρόσθετα χαρακτηριστικά, το hashtags_Count και το Platform_encoded, η συμπεριφορά του μοντέλου παρέμεινε έντονα προσανατολισμένη προς την κατηγορία positive_emotion.

Η μηδενική αναγνώριση έξι από τις οκτώ κατηγορίες και το χαμηλό macro F1-score δείχνουν ότι η επιπλέον αριθμητική πληροφορία δεν ήταν επαρκής για τη δημιουργία σαφών ορίων μεταξύ των θεματικών κλάσεων.

Το αποτέλεσμα αυτό υποδηλώνει ότι τα αριθμητικά χαρακτηριστικά αλληλεπίδρασης και τα μεταδεδομένα της ανάρτησης δεν παρέχουν από μόνα τους αρκετή σημασιολογική πληροφορία. Για τον λόγο αυτό, το επόμενο μοντέλο εξετάζει μια υβριδική αρχιτεκτονική πολλαπλών εισόδων, όπου διαφορετικές ομάδες χαρακτηριστικών επεξεργάζονται ξεχωριστά πριν από τη συνένωσή τους.

4.6.3 Αποτελέσματα υβριδικού μοντέλου Bi-LSTM

4.6.3.1 Συνολική απόδοση

Το υβριδικό μοντέλο Bi-LSTM αξιολογήθηκε στο ίδιο σύνολο δοκιμής των 147 εγγράφων. Η αρχιτεκτονική του βασίστηκε σε δύο παράλληλους κλάδους εισόδου. Ο πρώτος κλάδος επεξεργάστηκε τα χαρακτηριστικά Likes και Retweets, ενώ ο δεύτερος τα χαρακτηριστικά hashtags_Count, Platform_encoded και Sentiment_encoded.

Οι δύο ομάδες χαρακτηριστικών επεξεργάστηκαν αρχικά ανεξάρτητα και στη συνέχεια οι παραγόμενες αναπαραστάσεις συνενώθηκαν πριν από το τελικό στάδιο ταξινόμησης.

Η εκπαίδευση πραγματοποιήθηκε για 30 εποχές, με βελτιστοποιητή Adam, συνάρτηση απώλειας sparse_categorical_crossentropy και μέγεθος παρτίδας ίσο με 16.

Η συνολική ακρίβεια του μοντέλου ήταν:

$$\text{Accuracy} = 0,38095\% \quad (4.16)$$

ή, ισοδύναμα:

$$\text{Accuracy} = 38,10\% \quad (4.17)$$

Η επίδοση αυτή αποτελεί την υψηλότερη που επιτεύχθηκε μεταξύ όλων των προηγούμενων μοντέλων. Συγκεκριμένα, παρουσιάζει βελτίωση σε σχέση με το απλό LSTM, το βασικό Bi-LSTM και το Bi-LSTM2, γεγονός που δείχνει ότι η συνδυαστική επεξεργασία διαφορετικών ομάδων χαρακτηριστικών μπορεί να οδηγήσει σε αποτελεσματικότερη αξιοποίηση της διαθέσιμης πληροφορίας.

4.6.3.2 Εξέλιξη της εκπαίδευσης

Η εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά τις 30 εποχές παρουσιάζεται παρακάτω.

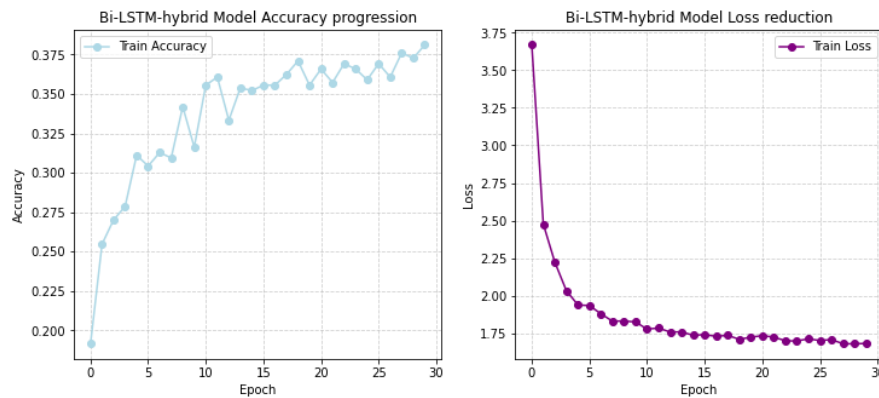


Figure 18 Εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά την εκπαίδευση του υβριδικού μοντέλου Bi-LSTM.

Η ακρίβεια εκπαίδευσης αυξάνεται από περίπου 19% στην πρώτη εποχή σε τιμές κοντά στο 31% μέχρι την πέμπτη εποχή. Στη συνέχεια παρουσιάζονται μικρές διακυμάνσεις, ενώ η γενική τάση παραμένει ανοδική.

Από τη δέκατη περίπου εποχή και μετά, η ακρίβεια σταθεροποιείται κυρίως μεταξύ 35% και 38%. Η υψηλότερη τιμή ακρίβειας εμφανίζεται στην τελευταία εποχή και φτάνει περίπου το 38,12%.

Η συνάρτηση απώλειας μειώνεται έντονα κατά τις πρώτες εποχές, από 3,67 στην πρώτη εποχή σε τιμές κάτω από 2,00 περίπου μετά την πέμπτη εποχή. Στη συνέχεια συνεχίζει να μειώνεται σταδιακά, καταλήγοντας περίπου στο 1,68 στην τελευταία εποχή.

Η συγκεκριμένη συμπεριφορά δείχνει ότι το μοντέλο συνεχίζει να βελτιώνεται καθ' όλη τη διάρκεια της εκπαίδευσης. Ωστόσο, καθώς δεν καταγράφονται ξεχωριστές μετρικές επικύρωσης, δεν μπορεί να εξαχθεί ασφαλές συμπέρασμα σχετικά με πιθανή υπερπροσαρμογή.

4.6.3.3 Πίνακας σύγκυσης

Η συμπεριφορά του μοντέλου ανά κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα σύγκυσης.

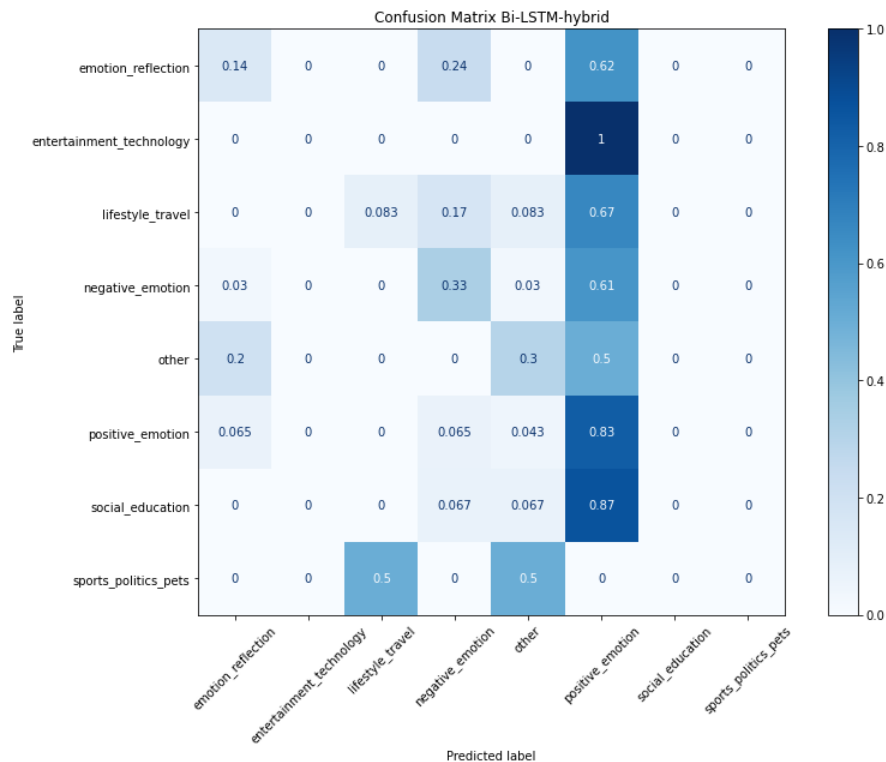


Figure 19 Κανονικοποιημένος πίνακας σύγχυσης του υβριδικού μοντέλου Bi-LSTM.

Ο πίνακας σύγχυσης δείχνει ότι το μοντέλο εξακολουθεί να παρουσιάζει τάση πρόβλεψης προς την κατηγορία *positive_emotion*. Ωστόσο, σε αντίθεση με τα προηγούμενα μοντέλα, εμφανίζει σωστές προβλέψεις σε περισσότερες κατηγορίες.

Η κατηγορία *positive_emotion* παρουσιάζει Recall ίσο με 0,83.

Παράλληλα, το μοντέλο αναγνωρίζει σωστά:

- το 14% των δειγμάτων της *emotion_reflection*,
- το 8% της *lifestyle_travel*,
- το 33% της *negative_emotion*,
- και το 30% της *other*.

Παρόλα αυτά, σημαντικό ποσοστό δειγμάτων εξακολουθεί να ταξινομείται ως *positive_emotion*. Συγκεκριμένα:

- το 62% των δειγμάτων της *emotion_reflection*,
- το 100% της *entertainment_technology*,
- το 67% της *lifestyle_travel*,
- το 61% της *negative_emotion*,
- το 50% της *other*,
- και το 87% της *social_education*

προβλέφθηκαν ως *positive_emotion*.

Για την κατηγορία *sports_politics_pets*, τα δύο διαθέσιμα δείγματα ταξινομήθηκαν ισομερώς ως *lifestyle_travel* και *other*.

Η εικόνα αυτή δείχνει ότι η υβριδική αρχιτεκτονική βελτίωσε τη διασπορά των προβλέψεων σε σχέση με τα προηγούμενα μοντέλα, χωρίς όμως να εξαλείψει πλήρως την επίδραση της κυρίαρχης κατηγορίας.

4.6.3.4 Αναλυτικές μετρικές ταξινόμησης

Οι αναλυτικές μετρικές του μοντέλου παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 11 Αναλυτικές μετρικές του υβριδικού μοντέλου Bi-LSTM

Κατηγορία	Precisio n	Recal l	F1- score	Suppo rt
emotion_reflection	0,33	0,14	0,20	21
entertainment_technology	0,00	0,00	0,00	8
lifestyle_travel	0,50	0,08	0,14	12
negative_emotion	0,50	0,33	0,40	33
other	0,33	0,30	0,32	10
positive_emotion	0,36	0,83	0,50	46
social_education	0,00	0,00	0,00	15
sports_politics_pets	0,00	0,00	0,00	2
Macro average	0,25	0,21	0,20	147
Weighted average	0,34	0,38	0,31	147

Η κατηγορία positive_emotion παρουσιάζει το υψηλότερο Recall, ίσο με 0,83, γεγονός που σημαίνει ότι το μοντέλο αναγνώρισε σωστά το μεγαλύτερο μέρος των πραγματικών δειγμάτων της συγκεκριμένης κατηγορίας. Όμως, η Precision παραμένει χαμηλή (0,36), καθώς αρκετά δείγματα άλλων κατηγοριών ταξινομήθηκαν επίσης ως positive_emotion.

Η κατηγορία negative_emotion εμφανίζει Precision 0,50, Recall 0,33 και F1-score 0,40. Η επίδοση αυτή είναι πιο ισορροπημένη σε σχέση με την positive_emotion, αν και το Recall παραμένει περιορισμένο.

Η κατηγορία lifestyle_travel εμφανίζει σχετικά υψηλή Precision, ίση με 0,50, αλλά Recall μόλις 0,08. Αυτό δείχνει ότι οι λίγες προβλέψεις του μοντέλου για τη συγκεκριμένη κατηγορία ήταν σχετικά ακριβείς, αλλά το μεγαλύτερο μέρος των πραγματικών δειγμάτων δεν αναγνωρίστηκε.

Οι κατηγορίες entertainment_technology, social_education και sports_politics_pets δεν παρουσίασαν σωστές προβλέψεις.

Το macro F1-score διαμορφώθηκε στο 0,20, ενώ το weighted F1-score στο 0,31. Οι τιμές αυτές είναι υψηλότερες από τις αντίστοιχες των προηγούμενων μοντέλων LSTM και Bi-LSTM, γεγονός που δείχνει ότι η υβριδική αρχιτεκτονική βελτίωσε όχι μόνο τη συνολική ακρίβεια αλλά και την ισορροπία των προβλέψεων μεταξύ των διαφορετικών κατηγοριών.

4.6.3.5 Συμπέρασμα για το μοντέλο

Το υβριδικό μοντέλο Bi-LSTM πέτυχε συνολική ακρίβεια 38,10%, η οποία αποτελεί την υψηλότερη επίδοση μεταξύ των μοντέλων που χρησιμοποιούν αριθμητικά και κωδικοποιημένα χαρακτηριστικά.

Η διαχείριση των χαρακτηριστικών μέσω δύο ανεξάρτητων κλάδων εισόδου επέτρεψε στο μοντέλο να αξιοποιήσει αποτελεσματικότερα διαφορετικές ομάδες πληροφορίας. Σε σχέση με τα προηγούμενα μοντέλα, αυξήθηκαν τόσο το macro F1-score όσο και το weighted F1-score, ενώ περισσότερες κατηγορίες παρουσίασαν μη μηδενικές τιμές στις επιμέρους μετρικές.

Παρόλα αυτά, το μοντέλο εξακολουθεί να εμφανίζει τάση προς την κατηγορία positive_emotion, ενώ οι κατηγορίες entertainment_technology, social_education και sports_politics_pets δεν αναγνωρίστηκαν σωστά.

Συνολικά, τα αποτελέσματα δείχνουν ότι η υβριδική αρχιτεκτονική αποτελεί βελτίωση σε σχέση με τις απλούστερες προσεγγίσεις, αλλά η προσθήκη σημασιολογικής πληροφορίας μέσω κειμενικών χαρακτηριστικών παραμένει απαραίτητη για την αποτελεσματικότερη διάκριση των θεματικών κατηγοριών.

4.6.4 Αποτελέσματα Bi-LSTM Numeric + Text

4.6.4.1 Συνολική απόδοση

Το μοντέλο Bi-LSTM Numeric + Text αποτελεί την πιο σύνθετη αρχιτεκτονική που υλοποιήθηκε στο πλαίσιο της εργασίας. Σε αντίθεση με τα προηγούμενα μοντέλα, συνδυάζει δύο διαφορετικές μορφές πληροφορίας.

Ο αριθμητικός κλάδος χρησιμοποιεί τα χαρακτηριστικά Likes και Retweets, ενώ ο κειμενικός κλάδος επεξεργάζεται τη μεταβλητή `metadata_text`. Η κειμενική πληροφορία μετατράπηκε σε αριθμητικές ακολουθίες μέσω Tokenizer και εφαρμόστηκε padding, ώστε όλες οι ακολουθίες να έχουν σταθερό μήκος 15 tokens.

Η εκπαίδευση του μοντέλου πραγματοποιήθηκε για 40 εποχές, με βελτιστοποιητή Adam, συνάρτηση απώλειας `sparse_categorical_crossentropy` και μέγεθος παρτίδας ίσο με 32.

Η συνολική ακρίβεια στο σύνολο δοκιμής ήταν:

$$\text{Accuracy} = 0,6394 \quad (3.18)$$

ή, ισοδύναμα:

$$\text{Accuracy} = 63,94\% \quad (3.19)$$

Η συγκεκριμένη επίδοση είναι σημαντικά υψηλότερη από όλες τις προηγούμενες προσεγγίσεις. Το αμέσως προηγούμενο καλύτερο αποτέλεσμα είχε επιτευχθεί από το υβριδικό Bi-LSTM, με ακρίβεια 38,10%.

4.6.4.2 Εξέλιξη της εκπαίδευσης

Η εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά τη διάρκεια των 40 εποχών παρουσιάζεται στο ακόλουθο διάγραμμα.

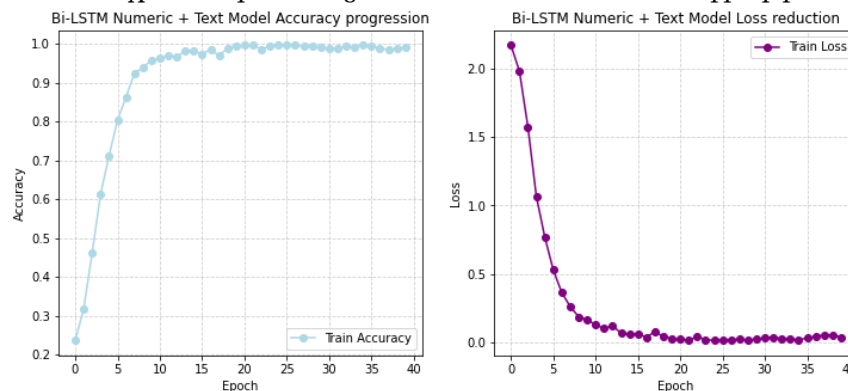


Figure 20 Εξέλιξη της ακρίβειας και της συνάρτησης απώλειας κατά την εκπαίδευση του μοντέλου Bi-LSTM Numeric + Text.

Η ακρίβεια εκπαίδευσης αυξάνεται γρήγορα κατά τις πρώτες εποχές. Από περίπου 23,59% στην πρώτη εποχή, ξεπερνά το 60% στην τέταρτη εποχή και το 90% περίπου στην ένατη εποχή. Από τη δωδέκατη εποχή και μετά, η ακρίβεια διατηρείται σε ιδιαίτερα υψηλά επίπεδα, συνήθως άνω του 97%.

Κατά τις τελευταίες εποχές, η ακρίβεια εκπαίδευσης προσεγγίζει το 100%, φτάνοντας την τιμή 98,80% στην τριακοστή ένατη και 99,15% στην τεσσαρακοστή εποχή.

Παράλληλα, η συνάρτηση απώλειας παρουσιάζει έντονη πτώση. Από την αρχική τιμή 2,17 μειώνεται κάτω από το 1,00 μέσα στις πρώτες πέντε εποχές και προσεγγίζει το μηδέν προς το τέλος της εκπαίδευσης.

Η συγκεκριμένη πορεία δείχνει ότι το μοντέλο προσαρμόζεται σε πολύ υψηλό βαθμό στα δεδομένα εκπαίδευσης. Ωστόσο, η μεγάλη διαφορά μεταξύ της ακρίβειας εκπαίδευσης, η οποία προσεγγίζει το 100%, και της ακρίβειας στο σύνολο δοκιμής, η οποία είναι 63,95%, αποτελεί ένδειξη πιθανής υπερπροσαρμογής.

Η αξιολόγηση της υπερπροσαρμογής δεν μπορεί να γίνει οριστικά μόνο από τις διαθέσιμες καμπύλες, επειδή στον κώδικα δεν καταγράφονται ξεχωριστές τιμές ακρίβειας και απώλειας για σύνολο επικύρωσης. Παρόλα αυτά, η διαφορά μεταξύ της επίδοσης εκπαίδευσης και δοκιμής πρέπει να ληφθεί υπόψη κατά την ερμηνεία των αποτελεσμάτων.

4.6.4.3 Πίνακας σύγχυσης

Η απόδοση του μοντέλου ανά θεματική κατηγορία παρουσιάζεται στον κανονικοποιημένο πίνακα.



Figure 21 Κανονικοποιημένος πίνακας σύγχυσης του μοντέλου Bi-LSTM Numeric + Text.

Σε αντίθεση με τα προηγούμενα νευρωνικά μοντέλα, το Bi-LSTM Numeric + Text επιτυγχάνει σωστές προβλέψεις και για τις οκτώ κατηγορίες.

Οι υψηλότερες τιμές Recall εμφανίζονται στις κατηγορίες:

- negative_emotion, με Recall 0,94,
- emotion_reflection, με Recall 0,76,
- positive_emotion, με Recall 0,76,
- sports_politics_pets, με Recall 0,50.

Η κατηγορία negative_emotion αναγνωρίζεται σωστά στο 94% των περιπτώσεων. Μικρό μέρος των δειγμάτων της ταξινομείται ως positive_emotion.

Η κατηγορία *emotion_reflection* εμφανίζει επίσης υψηλή αναγνώριση, καθώς το 76% των δειγμάτων ταξινομείται σωστά. Οι λανθασμένες προβλέψεις κατανέμονται κυρίως στις κατηγορίες *social_education*, *negative_emotion* και *positive_emotion*.

Η *positive_emotion* εμφανίζει Recall 0,76, καθώς το 76% των δειγμάτων της αναγνωρίζεται σωστά. Οι υπόλοιπες λανθασμένες προβλέψεις κατανέμονται κυρίως στις κατηγορίες *entertainment_technology*, *lifestyle_travel*, *negative_emotion* και *social_education*.

Χαμηλότερη επίδοση παρουσιάζεται στην κατηγορία *entertainment_technology*, με Recall 0,25. Τα δείγματά της κατανέμονται σε πολλές διαφορετικές προβλεπόμενες κατηγορίες, κυρίως στις *emotion_reflection*, *negative_emotion* και *social_education*.

Η κατηγορία *sports_politics_pets* παρουσιάζει Recall 0,50. Όμως, επειδή περιλαμβάνει μόνο δύο δείγματα στο σύνολο δοκιμής, η συγκεκριμένη τιμή πρέπει να ερμηνευτεί με προσοχή.

Η συνολική εικόνα του πίνακα σύγχυσης είναι σαφώς πιο ισορροπημένη σε σχέση με τα προηγούμενα μοντέλα. Οι υψηλότερες τιμές συγκεντρώνονται στην κύρια διαγώνιο και παρατηρούνται πλέον σωστές προβλέψεις και στις οκτώ κατηγορίες.

4.6.4.4 Αναλυτικές μετρικές ταξινόμησης

Οι αναλυτικές μετρικές του μοντέλου παρουσιάζονται στον ακόλουθο πίνακα.

Πίνακας 12 Αναλυτικές μετρικές του μοντέλου Bi-LSTM Numeric + Text

Κατηγορία	Precisio n	Recal l	F1- score	Suppo rt
emotion_reflection	0,94	0,76	0,84	21
entertainment_technology	0,20	0,25	0,22	8
lifestyle_travel	0,50	0,33	0,40	12
negative_emotion	0,63	0,94	0,76	33
other	0,50	0,20	0,29	10
positive_emotion	0,76	0,76	0,76	46
social_education	0,27	0,20	0,23	15
sports_politics_pets	0,50	0,50	0,50	2
Macro average	0,54	0,49	0,50	147
Weighted average	0,63	0,64	0,62	147

Η κατηγορία *negative_emotion* παρουσιάζει την υψηλότερη συνολική επίδοση, με Recall 0,94, Precision 0,63 και F1-score 0,76. Πολύ υψηλή είναι και η απόδοση της κατηγορίας *emotion_reflection*, με F1-score 0,84.

Η κατηγορία *positive_emotion* εμφανίζει Precision 0,76, Recall 0,76 και F1-score 0,76. Σε αντίθεση με τα προηγούμενα μοντέλα, η υψηλή αναγνώριση της συγκεκριμένης κατηγορίας δεν συνοδεύεται από τόσο έντονη συγκέντρωση λανθασμένων προβλέψεων.

Η *lifestyle_travel* παρουσιάζει Recall, ίσο με 0,33, και Precision ίση με 0,50. Αυτό σημαίνει ότι το μοντέλο αναγνωρίζει μόνο μέρος των πραγματικών δειγμάτων της κατηγορίας, ενώ οι προβλέψεις που πραγματοποιεί για τη συγκεκριμένη κατηγορία είναι σχετικά ακριβείς.

Η χαμηλότερη επίδοση παρατηρείται στην κατηγορία *entertainment_technology*, όπου οι τρεις μετρικές ισούνται με 0,20, 0,25 και

0,22 αντίστοιχα. Το αποτέλεσμα δείχνει ότι η συγκεκριμένη κατηγορία παραμένει δύσκολο να διαχωριστεί από τις υπόλοιπες.

Το macro F1-score είναι 0,50, ενώ το weighted F1-score είναι 0,62. Η σημαντική αύξηση του macro F1-score σε σχέση με όλα τα προηγούμενα μοντέλα δείχνει ότι η βελτίωση δεν περιορίζεται μόνο στις συχνότερες κατηγορίες, αλλά αφορά σε μεγαλύτερο βαθμό το σύνολο των κλάσεων.

4.6.4.5 Συμπέρασμα για το μοντέλο

Το Bi-LSTM Numeric + Text πέτυχε συνολική ακρίβεια 63,95%, αποτελώντας το μοντέλο με την υψηλότερη απόδοση στην παρούσα πειραματική διαδικασία.

Η προσθήκη κειμενικής πληροφορίας οδήγησε σε σημαντική αύξηση τόσο της Accuracy όσο και των μέσων τιμών Precision, Recall και F1-score. Σε αντίθεση με τα προηγούμενα μοντέλα, το συγκεκριμένο μοντέλο αναγνώρισε σωστά δείγματα και από τις οκτώ θεματικές κατηγορίες.

Παρόλα αυτά, η σχεδόν πλήρης προσαρμογή στο σύνολο εκπαίδευσης αποτελεί ένδειξη πιθανής υπερπροσαρμογής. Επιπλέον, η μεταβλητή `metadata_text` περιλαμβάνει τα `hashtags`, από τα οποία προέκυψε η μεταβλητή-στόχος. Συνεπώς, υπάρχει άμεση σχέση μεταξύ μέρους της εισόδου και της κατηγορίας που καλείται να προβλέψει το μοντέλο.

Το συγκεκριμένο σημείο πρέπει να αναγνωριστεί ως περιορισμός της πειραματικής διαδικασίας. Επομένως, το αποτέλεσμα δείχνει ότι η κειμενική πληροφορία είναι ιδιαίτερα χρήσιμη για την ταξινόμηση, αλλά η ακριβής δυνατότητα γενίκευσης του μοντέλου θα πρέπει να επιβεβαιωθεί με αυστηρότερη πειραματική διαδικασία, όπως χρήση ανεξάρτητου συνόλου επικύρωσης και διαφορετική διαμόρφωση της κειμενικής εισόδου.

4.7 Συγκριτική αξιολόγηση των μοντέλων

Για τη συνολική αξιολόγηση των πειραμάτων συγκρίναμε τις επιδόσεις όλων των μοντέλων που υλοποιήθηκαν. Η σύγκριση βασίστηκε στην Accuracy, καθώς και στις σταθμισμένες τιμές Precision, Recall και F1-score. Οι σταθμισμένες μετρικές λαμβάνουν υπόψη το διαφορετικό πλήθος δειγμάτων κάθε κατηγορίας και, επομένως, είναι χρήσιμες για το συγκεκριμένο σύνολο δεδομένων, στο οποίο οι κατηγορίες δεν εκπροσωπούνται ισοποσα.

Τα συγκεντρωτικά αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα.

Πίνακας 13 Συγκριτικά αποτελέσματα των μοντέλων ταξινόμησης

Μοντέλο	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score
Random Forest	19,05%	0,28	0,19	0,21
kNN	21,09%	0,27	0,21	0,21
SVM	15,65%	0,12	0,16	0,12
LSTM	36,05%	0,22	0,36	0,24
Bi-LSTM	36,05%	0,26	0,36	0,23
Bi-LSTM2	35,37%	0,20	0,35	0,23
Bi-LSTM Hybrid	38,10%	0,34	0,38	0,31

Bi-LSTM Numeric + Text	63,95%	0,63	0,64	0,62
---------------------------------------	--------	------	------	------

Από τα αποτελέσματα παρατηρούμε ότι οι τρεις κλασικοί αλγόριθμοι Μηχανικής Μάθησης παρουσίασαν τις χαμηλότερες επιδόσεις. Το kNN είχε την καλύτερη Accuracy μεταξύ αυτών, με 21,09%, ενώ το SVM εμφάνισε τη χαμηλότερη τιμή, ίση με 15,65%. Το Random Forest βρέθηκε ενδιάμεσα, με Accuracy 19,05%.

Η περιορισμένη απόδοση των συγκεκριμένων μοντέλων συνδέεται με το γεγονός ότι χρησιμοποιούν αποκλειστικά τα χαρακτηριστικά Likes και Retweets. Τα χαρακτηριστικά αυτά περιγράφουν την αλληλεπίδραση των χρηστών με μία ανάρτηση, αλλά δεν αποτυπώνουν άμεσα το θεματικό της περιεχόμενο. Για τον λόγο αυτό, αναρτήσεις διαφορετικών κατηγοριών μπορεί να έχουν παρόμοιο αριθμό Likes και Retweets και να είναι δύσκολο να διαχωριστούν.

Η μετάβαση στα νευρωνικά μοντέλα οδήγησε σε αύξηση της συνολικής Accuracy. Το LSTM πέτυχε Accuracy 36,05%, ενώ το βασικό Bi-LSTM παρουσίασε την ίδια τιμή, ίση με 36,05%. Επομένως, η αμφίδρομη επεξεργασία δεν προσέφερε βελτίωση όταν η είσοδος αποτελούνταν μόνο από τα Likes και Retweets.

Παρόμοια συμπεριφορά παρουσίασε και το Bi-LSTM2. Παρά την προσθήκη των χαρακτηριστικών `hashtags_Count` και `Platform_encoded`, η Accuracy διαμορφώθηκε στο 35,37%, ενώ το weighted F1-score ήταν 0,23. Το αποτέλεσμα αυτό δείχνει ότι τα δύο επιπλέον χαρακτηριστικά δεν παρείχαν αρκετή πληροφορία για τη σαφέστερη διάκριση των κατηγοριών.

Το Bi-LSTM Hybrid παρουσίασε καλύτερη συνολική απόδοση, με Accuracy 38,10% και weighted F1-score 0,31. Η χρήση δύο ξεχωριστών κλάδων εισόδου επέτρεψε στο μοντέλο να επεξεργαστεί ανεξάρτητα τα χαρακτηριστικά αλληλεπίδρασης και τα πρόσθετα μεταδεδομένα πριν από τη συνένωσή τους. Παρότι η βελτίωση ήταν περιορισμένη, το μοντέλο εμφάνισε πιο ισορροπημένες προβλέψεις σε σχέση με τα προηγούμενα LSTM και Bi-LSTM.

Η υψηλότερη επίδοση επιτεύχθηκε από το μοντέλο Bi-LSTM Numeric + Text. Το συγκεκριμένο μοντέλο πέτυχε Accuracy 63,95%, weighted Precision 0,63, weighted Recall 0,64 και weighted F1-score 0,62. Η διαφορά από το δεύτερο καλύτερο μοντέλο, το Bi-LSTM Hybrid, ήταν ιδιαίτερα μεγάλη, καθώς η Accuracy αυξήθηκε κατά 25,85 ποσοστιαίες μονάδες.

Η βελτίωση αυτή δείχνει ότι η κειμενική πληροφορία της μεταβλητής `metadata_text` συνδέεται πιο άμεσα με τη θεματική κατηγορία της ανάρτησης από ό,τι οι απλοί αριθμητικοί δείκτες. Το μοντέλο μπόρεσε να αξιοποιήσει τις λεκτικές πληροφορίες και να αναγνωρίσει σωστά δείγματα και από τις οκτώ κατηγορίες.

Σύγκριση με βάση το macro F1-score

Επειδή το σύνολο δεδομένων παρουσιάζει ανισορροπία μεταξύ των κατηγοριών, εξετάσαμε επιπλέον το macro F1-score. Η συγκεκριμένη μετρική υπολογίζει το F1-score κάθε κατηγορίας ξεχωριστά και στη συνέχεια λαμβάνει τον μέσο όρο, δίνοντας ίσο βάρος σε όλες τις κατηγορίες.

Πίνακας 14 Σύγκριση macro και weighted F1-score

Μοντέλο	Macro F1-score	Weighted F1-score
Random Forest	0,15	0,21

kNN	0,15	0,21
SVM	0,10	0,12
LSTM	0,10	0,23
Bi-LSTM	0,10	0,23
Bi-LSTM2	0,10	0,23
Bi-LSTM Hybrid	0,20	0,31
Bi-LSTM Numeric + Text	0,50	0,62

Οι τιμές του macro F1-score δείχνουν ότι τα μοντέλα LSTM, Bi-LSTM και Bi-LSTM2 δεν είχαν ισορροπημένη απόδοση στις οκτώ κατηγορίες, παρόλο που η Accuracy τους ήταν υψηλότερη από εκείνη των κλασικών μοντέλων. Η αύξηση της Accuracy οφειλόταν κυρίως στην καλύτερη αναγνώριση των συχνότερων κατηγοριών, ιδιαίτερα της *positive_emotion*.

Το Bi-LSTM Hybrid βελτίωσε το macro F1-score στο 0,20, γεγονός που δείχνει ότι άρχισε να αναγνωρίζει περισσότερες κατηγορίες. Ωστόσο, η πιο σημαντική βελτίωση παρατηρήθηκε στο Bi-LSTM Numeric + Text, το οποίο πέτυχε macro F1-score 0,50. Το αποτέλεσμα αυτό επιβεβαιώνει ότι η καλύτερη επίδοση του τελικού μοντέλου δεν περιορίζεται μόνο στις πολυπληθέστερες κατηγορίες, αλλά επεκτείνεται σε μεγαλύτερο μέρος των κλάσεων.

Συγκριτικό διάγραμμα

Η σύγκριση της συνολικής ακρίβειας των μοντέλων παρουσιάζεται στο ακόλουθο διάγραμμα.

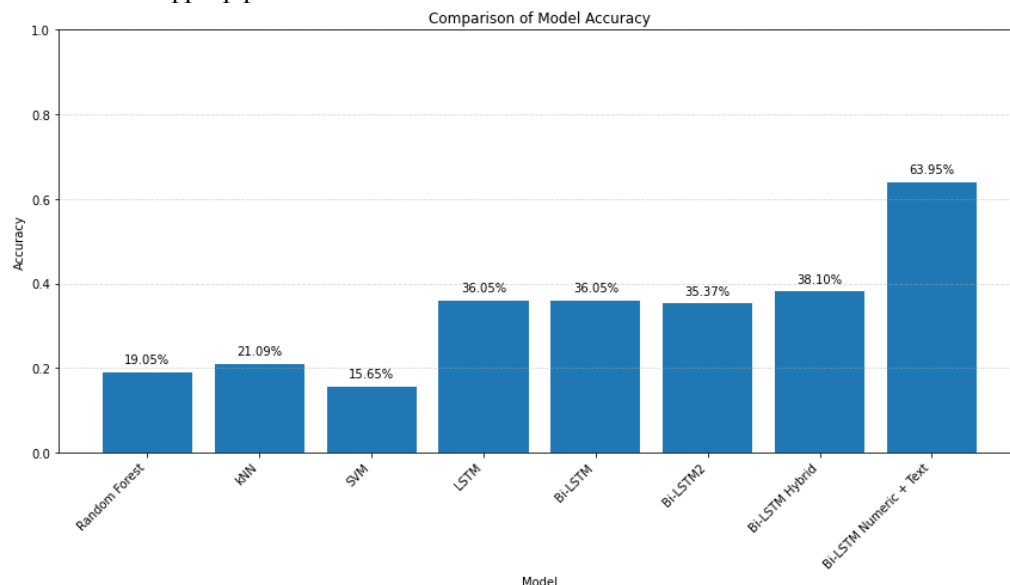


Figure 22 Συγκριτική παρουσίαση της Accuracy των μοντέλων ταξινόμησης.

Από το ραβδόγραμμα γίνεται άμεσα εμφανής η διαφορά μεταξύ του μοντέλου Bi-LSTM Numeric + Text και των υπόλοιπων προσεγγίσεων. Τα μοντέλα που χρησιμοποιούν μόνο αριθμητικά χαρακτηριστικά παραμένουν κάτω από το 40%, ενώ η προσθήκη κειμενικής πληροφορίας αυξάνει την Accuracy στο 63,95%.

Συνολικά, το Bi-LSTM Numeric + Text παρουσίασε την καλύτερη επίδοση σύμφωνα με όλες τις βασικές μετρικές αξιολόγησης. Εντούτοις, η υψηλή ακρίβεια εκπαίδευσης και η χρήση των *hashtags* μέσα στη μεταβλητή *metadata_text* αποτελούν σημεία που πρέπει να ληφθούν υπόψη κατά την τελική ερμηνεία των αποτελεσμάτων.

4.8 Συζήτηση αποτελεσμάτων

Στην παρούσα εργασία υλοποιήθηκαν και αξιολογήθηκαν οκτώ διαφορετικά μοντέλα ταξινόμησης, με στόχο την αυτόματη κατηγοριοποίηση των hashtags. Τα αποτελέσματα έδειξαν ότι η απόδοση των μοντέλων εξαρτάται σε μεγάλο βαθμό από το είδος και την ποσότητα της πληροφορίας που χρησιμοποιείται ως είσοδος.

Οι κλασικοί αλγόριθμοι μηχανικής μάθησης (Random Forest, kNN και SVM) παρουσίασαν σχετικά χαμηλή απόδοση. Το γεγονός αυτό ήταν αναμενόμενο, καθώς βασίστηκαν αποκλειστικά στα αριθμητικά χαρακτηριστικά Likes και Retweets, τα οποία δεν περιέχουν επαρκή πληροφορία για την περιγραφή του θεματικού περιεχομένου ενός hashtag.

Η χρήση μοντέλων βαθιάς μάθησης οδήγησε σε καλύτερα αποτελέσματα. Τα μοντέλα LSTM και Bi-LSTM κατάφεραν να αναγνωρίσουν πιο σύνθετα πρότυπα στα δεδομένα, παρουσιάζοντας σημαντική βελτίωση σε σχέση με τους κλασικούς αλγόριθμους. Η περαιτέρω ενσωμάτωση επιπλέον χαρακτηριστικών στο Bi-LSTM Hybrid βελτίωσε ακόμη περισσότερο την απόδοση του μοντέλου.

Η υψηλότερη ακρίβεια επιτεύχθηκε από το προτεινόμενο μοντέλο Bi-LSTM Numeric + Text. Το αποτέλεσμα αυτό δείχνει ότι η συνδυασμένη αξιοποίηση αριθμητικών χαρακτηριστικών και κειμενικής πληροφορίας επιτρέπει στο μοντέλο να κατανοήσει καλύτερα το περιεχόμενο κάθε ανάρτησης και να πραγματοποιήσει ακριβέστερη ταξινόμηση των hashtags.

Παρότι το προτεινόμενο μοντέλο παρουσίασε αισθητά καλύτερη απόδοση, εξακολουθούν να υπάρχουν κατηγορίες με χαμηλότερες επιδόσεις, όπως η Entertainment & Technology και η Sports, Politics & Pets. Το γεγονός αυτό πιθανότατα οφείλεται στον μικρό αριθμό διαθέσιμων δειγμάτων για τις συγκεκριμένες κατηγορίες, καθώς και στην ανομοιόμορφη κατανομή των δεδομένων του συνόλου εκπαίδευσης.

Συνολικά, τα αποτελέσματα της εργασίας δείχνουν ότι η ενσωμάτωση διαφορετικών πηγών πληροφορίας και η αξιοποίηση αρχιτεκτονικών βαθιάς μάθησης μπορούν να βελτιώσουν σημαντικά την απόδοση ενός συστήματος αυτόματης κατηγοριοποίησης hashtags.

4.9 Σύγκριση με προηγούμενες μελέτες

Η απόδοση των μοντέλων που αναπτύχθηκαν στην παρούσα εργασία συγκρίθηκε με αντίστοιχες προσεγγίσεις της βιβλιογραφίας, με στόχο την αξιολόγηση της θέσης της προτεινόμενης μεθοδολογίας σε σχέση με προηγούμενες εργασίες που αφορούν την ταξινόμηση περιεχομένου από δεδομένα κοινωνικών δικτύων. Η άμεση αριθμητική σύγκριση των αποτελεσμάτων δεν είναι απολύτως αντιπροσωπευτική, καθώς οι διαφορετικές μελέτες χρησιμοποιούν διαφορετικά σύνολα δεδομένων, διαφορετικό αριθμό κατηγοριών και διαφορετικές διαδικασίες προεπεξεργασίας. Ως εκ τούτου, η σύγκριση πραγματοποιείται κυρίως σε επίπεδο μεθοδολογίας και σχεδιασμού των μοντέλων.

Οι Yunida et al. (2023) μελέτησαν την ταξινόμηση αναρτήσεων από το Twitter σχετικά με φυσικές καταστροφές, συγκρίνοντας μοντέλα LSTM και Bi-LSTM. Η εργασία τους αποτελεί μία από τις πιο σχετικές προσεγγίσεις με την παρούσα μελέτη, καθώς αξιοποιεί επίσης δεδομένα κοινωνικών δικτύων και ακολουθιακά νευρωνικά δίκτυα για προβλήματα πολυκατηγορικής ταξινόμησης. Το Bi-LSTM παρουσίασε ελαφρώς καλύτερη απόδοση σε σχέση

με το απλό LSTM, γεγονός που υποδεικνύει τη χρησιμότητα της αμφίδρομης επεξεργασίας ακολουθιών. Παρόμοια, στην παρούσα εργασία εξετάστηκε η χρήση Bi-LSTM, αρχικά με αριθμητικά χαρακτηριστικά και στη συνέχεια μέσω εμπλουτισμένων υβριδικών αρχιτεκτονικών.

Από την άλλη πλευρά, υπάρχει σημαντική διαφοροποίηση μεταξύ των δύο προσεγγίσεων. Η εργασία των Yunida et al. βασίζεται αποκλειστικά σε κειμενικές αναπαραστάσεις μέσω Word2Vec, ενώ η παρούσα εργασία εξετάζει τον συνδυασμό διαφορετικών τύπων χαρακτηριστικών. Συγκεκριμένα, το τελικό μοντέλο Bi-LSTM Numeric + Text ενσωματώνει τόσο αριθμητικά χαρακτηριστικά αλληλεπίδρασης των αναρτήσεων όσο και την κειμενική πληροφορία, επιτρέποντας την αξιοποίηση διαφορετικών πηγών πληροφορίας.

Παρόμοια προσέγγιση σύγκρισης μεταξύ παραδοσιακών αλγορίθμων Μηχανικής Μάθησης και μοντέλων Βαθιάς Μάθησης παρουσιάστηκε από τους Verma et al. (2026). Οι συγγραφείς αξιολόγησαν μοντέλα όπως Naïve Bayes, Random Forest, SVM και LSTM σε δεδομένα κοινωνικών δικτύων, εξετάζοντας παράλληλα διαφορετικές τεχνικές αναπαράστασης κειμένου. Τα αποτελέσματα της μελέτης έδειξαν ότι η χρήση βαθύτερων νευρωνικών αρχιτεκτονικών σε συνδυασμό με κατάλληλες αναπαραστάσεις κειμένου μπορεί να βελτιώσει την ικανότητα ταξινόμησης.

Αντίστοιχα, στην παρούσα εργασία πραγματοποιήθηκε σύγκριση μεταξύ κλασικών μοντέλων Μηχανικής Μάθησης (Random Forest, kNN και SVM) και νευρωνικών αρχιτεκτονικών LSTM/Bi-LSTM. Τα αποτελέσματα έδειξαν ότι τα απλά μοντέλα βασισμένα αποκλειστικά σε αριθμητικά χαρακτηριστικά είχαν περιορισμένη απόδοση, ενώ η ενσωμάτωση κειμενικής πληροφορίας οδήγησε σε σημαντική βελτίωση. Συγκεκριμένα, το υβριδικό μοντέλο Bi-LSTM Numeric + Text πέτυχε την υψηλότερη απόδοση μεταξύ όλων των εξεταζόμενων προσεγγίσεων, υποδεικνύοντας ότι ο συνδυασμός δομημένων και μη δομημένων χαρακτηριστικών μπορεί να προσφέρει πιο ολοκληρωμένη αναπαράσταση του περιεχομένου των κοινωνικών δικτύων.

Συνολικά, η παρούσα εργασία διαφοροποιείται από προηγούμενες προσεγγίσεις καθώς δεν περιορίζεται αποκλειστικά στην επεξεργασία κειμένου ή στην αξιοποίηση μεμονωμένων χαρακτηριστικών, αλλά εξετάζει μία υβριδική αρχιτεκτονική που συνδυάζει αριθμητικά και κειμενικά δεδομένα. Η προσέγγιση αυτή επιτρέπει την αξιοποίηση τόσο της συμπεριφορικής πληροφορίας των αναρτήσεων όσο και του σημασιολογικού περιεχομένου τους, αποτελώντας ένα πιο ολοκληρωμένο πλαίσιο για τη θεματική ταξινόμηση δεδομένων κοινωνικών δικτύων.

4.10 Συμπεράσματα κεφαλαίου

Στο παρόν κεφάλαιο παρουσιάστηκαν τα πειραματικά αποτελέσματα των μοντέλων που αξιολογήθηκαν για τη θεματική ταξινόμηση δεδομένων κοινωνικών δικτύων. Αρχικά, εξετάστηκαν κλασικές μέθοδοι Μηχανικής Μάθησης, όπως Random Forest, k-Nearest Neighbors και Support Vector Machine, οι οποίες παρουσίασαν περιορισμένη απόδοση όταν χρησιμοποιήθηκαν αποκλειστικά αριθμητικά χαρακτηριστικά.

Στη συνέχεια, αξιολογήθηκαν νευρωνικές αρχιτεκτονικές βασισμένες σε LSTM και Bi-LSTM. Τα αποτελέσματα έδειξαν ότι η χρήση βαθύτερων ακολουθιακών μοντέλων μπορεί να βελτιώσει την ικανότητα ταξινόμησης, ιδιαίτερα όταν αξιοποιούνται πιο πλούσιες αναπαραστάσεις δεδομένων. Η σημαντικότερη βελτίωση παρατηρήθηκε στο υβριδικό μοντέλο Bi-LSTM

Numeric + Text, το οποίο συνδύασε αριθμητικά χαρακτηριστικά αλληλεπίδρασης με την κειμενική πληροφορία των αναρτήσεων.

Η συνολική αξιολόγηση έδειξε ότι η ενσωμάτωση διαφορετικών τύπων χαρακτηριστικών αποτελεί σημαντικό παράγοντα για την αποτελεσματικότερη κατανόηση περιεχομένου κοινωνικών δικτύων. Παράλληλα, η σύγκριση με προηγούμενες μελέτες ανέδειξε ότι η επιλογή αρχιτεκτονικής και η ποιότητα της πληροφορίας εισόδου επηρεάζουν σημαντικά την απόδοση των μοντέλων.

ΚΕΦΑΛΑΙΟ 5 Συμπεράσματα

5.1 Σύνοψη της εργασίας

Η παρούσα πτυχιακή εργασία είχε ως αντικείμενο τη μελέτη και την ανάπτυξη μοντέλων μηχανικής μάθησης και βαθιάς μάθησης για την αυτόματη ταξινόμηση hashtags σε θεματικές κατηγορίες. Η ανάγκη για την ανάπτυξη τέτοιων συστημάτων προκύπτει από τη συνεχή αύξηση του όγκου των δεδομένων που παράγονται καθημερινά στα μέσα κοινωνικής δικτύωσης, όπου τα hashtags χρησιμοποιούνται για την οργάνωση και την περιγραφή του περιεχομένου των αναρτήσεων.

Για την υλοποίηση της εργασίας χρησιμοποιήθηκε το σύνολο δεδομένων Social Media Sentiments Analysis Dataset, το οποίο περιλαμβάνει αναρτήσεις από διαφορετικές πλατφόρμες κοινωνικής δικτύωσης μαζί με πληροφορίες όπως τα hashtags, το συναίσθημα της ανάρτησης, την πλατφόρμα δημοσίευσης, καθώς και δείκτες αλληλεπίδρασης, όπως τα Likes και τα Retweets. Πρώτα πραγματοποιήθηκε προεπεξεργασία των δεδομένων, η οποία περιλάμβανε τον καθαρισμό των hashtags, τον διαχωρισμό σύνθετων λέξεων, την κατηγοριοποίησή τους με βάση ένα σύνολο λέξεων-κλειδιών και την ανάθεση μίας κύριας θεματικής κατηγορίας σε κάθε ανάρτηση. Έπειτα πραγματοποιήθηκε η κατάλληλη κωδικοποίηση των κατηγορικών μεταβλητών και η προετοιμασία των χαρακτηριστικών που χρησιμοποιήθηκαν ως είσοδος στα μοντέλα.

Στο πειραματικό μέρος της εργασίας υλοποιήθηκαν τρεις κλασικοί αλγόριθμοι μηχανικής μάθησης (Random Forest, k-Nearest Neighbors και Support Vector Machine) καθώς και μία σειρά μοντέλων βαθιάς μάθησης βασισμένων σε δίκτυα LSTM και Bidirectional LSTM. Εκτός από τις βασικές αρχιτεκτονικές, αναπτύχθηκαν επίσης υβριδικά μοντέλα τα οποία αξιοποίησαν συνδυαστικά αριθμητικά χαρακτηριστικά και πληροφορία κειμένου, με στόχο τη βελτίωση της απόδοσης της ταξινόμησης.

Η αξιολόγηση όλων των μοντέλων πραγματοποιήθηκε χρησιμοποιώντας κοινό σύνολο δοκιμής και βασικές μετρικές αξιολόγησης ταξινόμησης, όπως η Accuracy, η Precision, η Recall και ο F1-score. Η σύγκριση των αποτελεσμάτων επέτρεψε την αξιολόγηση τόσο των παραδοσιακών αλγορίθμων μηχανικής μάθησης όσο και των προτεινόμενων αρχιτεκτονικών βαθιάς μάθησης, αναδεικνύοντας τα πλεονεκτήματα και τους περιορισμούς κάθε προσέγγισης.

5.2 Συμπεράσματα

Η συγκριτική αξιολόγηση των μοντέλων έδειξε ότι η απόδοσή τους επηρεάζεται σημαντικά από τα χαρακτηριστικά που χρησιμοποιούνται ως είσοδος και από την αρχιτεκτονική κάθε αλγορίθμου. Τα αποτελέσματα ανέδειξαν σαφή διαφορά μεταξύ των κλασικών αλγορίθμων μηχανικής μάθησης και των μοντέλων βαθιάς μάθησης, ιδιαίτερα όταν αξιοποιήθηκε και η πληροφορία κειμένου.

Οι αλγόριθμοι Random Forest, k-Nearest Neighbors (kNN) και Support Vector Machine (SVM) παρουσίασαν σχετικά χαμηλή απόδοση, με τιμές ακρίβειας που κυμάνθηκαν μεταξύ 15,65% και 21,09%. Η περιορισμένη αυτή επίδοση μπορεί να αποδοθεί στο γεγονός ότι τα συγκεκριμένα μοντέλα εκπαιδεύτηκαν αποκλειστικά στα αριθμητικά χαρακτηριστικά Likes και

Retweets, τα οποία δεν περιγράφουν άμεσα το θεματικό περιεχόμενο των hashtags.

Η εφαρμογή των μοντέλων βαθιάς μάθησης οδήγησε σε αισθητά καλύτερα αποτελέσματα. Τα μοντέλα LSTM, Bi-LSTM και Bi-LSTM2 πέτυχαν υψηλότερη ή παρόμοια ακρίβεια σε σχέση με τους κλασικούς αλγορίθμους, με τιμές 36,05%, 36,05% και 35,37% αντίστοιχα. Τα αποτελέσματα δείχνουν ότι οι συγκεκριμένες αρχιτεκτονικές μπορούν να αξιοποιήσουν αποτελεσματικότερα τις σχέσεις μεταξύ των διαθέσιμων χαρακτηριστικών. Παρόλα αυτά, η συνολική τους απόδοση παρέμεινε περιορισμένη, καθώς εξακολουθούσαν να βασίζονται κυρίως σε αριθμητικά δεδομένα και παρουσίασαν χαμηλές τιμές macro F1-score.

Η εισαγωγή επιπλέον πληροφοριών μέσω του μοντέλου Bi-LSTM Hybrid, το οποίο συνδυάζει διαφορετικές ομάδες αριθμητικών χαρακτηριστικών, οδήγησε σε περαιτέρω βελτίωση της ακρίβειας ταξινόμησης. Το μοντέλο πέτυχε Accuracy 38,10% και weighted F1-score 0,31, παρουσιάζοντας καλύτερη απόδοση από τα προηγούμενα μοντέλα. Το αποτέλεσμα αυτό υποδηλώνει ότι η αξιοποίηση περισσότερων πληροφοριών σχετικά με κάθε ανάρτηση συμβάλλει στην αποτελεσματικότερη εκμάθηση των προτύπων που σχετίζονται με τις θεματικές κατηγορίες.

Η υψηλότερη απόδοση επιτεύχθηκε από το μοντέλο Bi-LSTM Numeric + Text, το οποίο συνδυάζει αριθμητικά χαρακτηριστικά με πληροφορία κειμένου μέσω δύο παράλληλων κλάδων επεξεργασίας. Το μοντέλο αυτό πέτυχε ακρίβεια 63,95%, weighted Precision 0,63, weighted Recall 0,64 και weighted F1-score 0,62, υπερβαίνοντας σημαντικά όλα τα υπόλοιπα μοντέλα που αξιολογήθηκαν. Η διαφορά από το δεύτερο καλύτερο μοντέλο, το Bi-LSTM Hybrid, ανέρχεται σε 25,85 ποσοστιαίες μονάδες. Παράλληλα, το macro F1-score διαμορφώθηκε στο 0,50, γεγονός που δείχνει ότι η βελτίωση δεν περιορίστηκε μόνο στις συχνότερες κατηγορίες.

Το αποτέλεσμα αυτό δείχνει ότι η ενσωμάτωση της κειμενικής πληροφορίας μπορεί να βελτιώσει ουσιαστικά την απόδοση ενός συστήματος αυτόματης ταξινόμησης hashtags, καθώς επιτρέπει στο μοντέλο να αξιοποιήσει όχι μόνο στοιχεία αλληλεπίδρασης αλλά και πληροφορίες σχετικές με το περιεχόμενο της ανάρτησης.

Ωστόσο, κατά την ερμηνεία της επίδοσης του Bi-LSTM Numeric + Text πρέπει να ληφθεί υπόψη ότι η μεταβλητή metadata_text περιλαμβάνει τα hashtags, από τα οποία προέκυψε η μεταβλητή-στόχος. Επομένως, υπάρχει άμεση σχέση μεταξύ μέρους της εισόδου και της κατηγορίας που καλείται να προβλέψει το μοντέλο. Επιπλέον, η πολύ υψηλή ακρίβεια εκπαίδευσης σε σχέση με την ακρίβεια στο σύνολο δοκιμής αποτελεί ένδειξη πιθανής υπερπροσαρμογής. Για τον λόγο αυτό, η υψηλή επίδοση του μοντέλου θα πρέπει να ερμηνευθεί με προσοχή και να επιβεβαιωθεί σε αυστηρότερη πειραματική διαδικασία.

Συνολικά, τα αποτελέσματα της εργασίας δείχνουν ότι η χρήση μοντέλων βαθιάς μάθησης σε συνδυασμό με κατάλληλη αναπαράσταση των διαθέσιμων δεδομένων μπορεί να οδηγήσει σε σημαντικά καλύτερη απόδοση σε σχέση με τις κλασικές τεχνικές μηχανικής μάθησης για το συγκεκριμένο πρόβλημα ταξινόμησης. Επιπρόσθετα, τα αποτελέσματα δείχνουν ότι η προσθήκη κειμενικής πληροφορίας είχε τη μεγαλύτερη επίδραση στην απόδοση, αν και η συγκεκριμένη προσέγγιση παρουσιάζει περιορισμούς που πρέπει να ληφθούν υπόψη κατά την αξιολόγηση της δυνατότητας γενίκευσης του μοντέλου.

5.3 Περιορισμοί της εργασίας

Παρά το γεγονός ότι τα αποτελέσματα της παρούσας εργασίας ήταν ιδιαίτερα ικανοποιητικά, υπάρχουν ορισμένοι περιορισμοί που πρέπει να ληφθούν υπόψη κατά την ερμηνεία τους.

Αρχικά, το σύνολο δεδομένων που χρησιμοποιήθηκε αποτελείται από 732 αναρτήσεις, αριθμός ο οποίος θεωρείται σχετικά μικρός για την εκπαίδευση μοντέλων βαθιάς μάθησης. Η περιορισμένη διαθεσιμότητα δεδομένων ενδέχεται να επηρεάζει την ικανότητα των μοντέλων να γενικεύουν αποτελεσματικά σε νέα, άγνωστα δεδομένα.

Επιπλέον, το σύνολο δεδομένων παρουσιάζει ανισορροπία μεταξύ των θεματικών κατηγοριών, καθώς ορισμένες κατηγορίες περιλαμβάνουν σημαντικά περισσότερα δείγματα από άλλες. Η ανισορροπία αυτή δυσκολεύει τη διαδικασία εκπαίδευσης και επηρεάζει την απόδοση των μοντέλων, ιδιαίτερα στις κατηγορίες με μικρό αριθμό παραδειγμάτων.

Ένας ακόμη περιορισμός αφορά τη διαδικασία δημιουργίας των ετικετών στόχου (target labels). Η κατηγοριοποίηση των hashtags δεν βασίστηκε σε χειροκίνητη επισημάνση από ειδικούς, αλλά πραγματοποιήθηκε μέσω ενός συστήματος αντιστοίχισης λέξεων-κλειδιών που αναπτύχθηκε στο πλαίσιο της εργασίας. Παρότι η προσέγγιση αυτή επέτρεψε την αυτόματη δημιουργία του συνόλου εκπαίδευσης, είναι πιθανό ορισμένα hashtags να έχουν ταξινομηθεί σε κατηγορίες που δεν αποδίδουν πλήρως το πραγματικό θεματικό τους περιεχόμενο.

Τέλος, παρατηρήθηκε ότι το προτεινόμενο μοντέλο Bi-LSTM Numeric + Text παρουσίασε πολύ υψηλή ακρίβεια κατά την εκπαίδευση, ενώ η απόδοσή του στο σύνολο δοκιμής ήταν αισθητά χαμηλότερη. Η διαφορά αυτή αποτελεί ένδειξη πιθανής υπερπροσαρμογής (overfitting), δηλαδή το μοντέλο φαίνεται να προσαρμόστηκε ιδιαίτερα καλά στα δεδομένα εκπαίδευσης χωρίς να επιτυγχάνει την ίδια απόδοση σε νέα δεδομένα. Κατά συνέπεια, τα αποτελέσματα θα πρέπει να ερμηνεύονται με προσοχή και να επιβεβαιωθούν σε μεγαλύτερα και περισσότερο αντιπροσωπευτικά σύνολα δεδομένων.

Πέραν αυτού, στο προτεινόμενο μοντέλο Bi-LSTM Numeric + Text, η κειμενική είσοδος (metadata_text) δημιουργήθηκε με τη συνένωση της πλατφόρμας δημοσίευσης, των hashtags και του συναισθήματος της ανάρτησης. Δεδομένου ότι οι θεματικές κατηγορίες της εργασίας προέκυψαν από την επεξεργασία των ίδιων των hashtags, η κειμενική είσοδος περιλαμβάνει πληροφορία που σχετίζεται άμεσα με τη διαδικασία δημιουργίας της μεταβλητής-στόχου. Το γεγονός αυτό ενδέχεται να συνέβαλε στη βελτίωση της απόδοσης του συγκεκριμένου μοντέλου και αποτελεί έναν παράγοντα που θα πρέπει να λαμβάνεται υπόψη κατά την ερμηνεία και τη σύγκριση των αποτελεσμάτων.

5.4 Μελλοντική εργασία

Η παρούσα εργασία ανέδειξε ότι η αξιοποίηση τεχνικών βαθιάς μάθησης και ο συνδυασμός διαφορετικών τύπων πληροφορίας μπορούν να βελτιώσουν σημαντικά την απόδοση ενός συστήματος αυτόματης ταξινόμησης hashtags. Ωστόσο, υπάρχουν αρκετές δυνατότητες περαιτέρω επέκτασης και βελτίωσης της προτεινόμενης μεθοδολογίας.

Αρχικά, σημαντική προοπτική αποτελεί η αξιοποίηση μεγαλύτερων και περισσότερο ισορροπημένων συνόλων δεδομένων, τα οποία θα περιλαμβάνουν μεγαλύτερο αριθμό αναρτήσεων και πιο ομοιόμορφη κατανομή των θεματικών

κατηγοριών. Με τον τρόπο αυτό θα μπορούσε να βελτιωθεί η γενίκευση των μοντέλων και η απόδοσή τους σε κατηγορίες που σήμερα εκπροσωπούνται από περιορισμένο αριθμό δειγμάτων.

Περαιτέρω, θα μπορούσε να εξεταστεί η χρήση προεκπαιδευμένων γλωσσικών μοντέλων, όπως τα BERT, RoBERTa ή άλλα μοντέλα βασισμένα στην αρχιτεκτονική Transformer, τα οποία έχουν αποδείξει ιδιαίτερα υψηλή απόδοση σε προβλήματα επεξεργασίας φυσικής γλώσσας. Η αξιοποίηση τέτοιων μοντέλων θα μπορούσε να επιτρέψει την καλύτερη κατανόηση του νοήματος των hashtags και του συνοδευτικού κειμένου, οδηγώντας πιθανώς σε ακόμη υψηλότερη ακρίβεια ταξινόμησης.

Μία ακόμη ενδιαφέρουσα κατεύθυνση αφορά την αξιοποίηση πραγματικών δεδομένων από πλατφόρμες κοινωνικής δικτύωσης μέσω των αντίστοιχων API, ώστε το σύστημα να αξιολογηθεί σε σύγχρονα δεδομένα και σε πραγματικές συνθήκες χρήσης. Με τον τρόπο αυτό θα μπορούσε να διερευνηθεί η δυνατότητα εφαρμογής του σε περιβάλλοντα όπου οι θεματικές τάσεις μεταβάλλονται διαρκώς.

Τέλος, θα είχε ιδιαίτερο ενδιαφέρον η διερεύνηση εναλλακτικών τρόπων δημιουργίας των κατηγοριών στόχου, όπως η χειροκίνητη επισήμανση από ειδικούς ή η αξιοποίηση ανεξάρτητων πηγών σχολιασμού. Συν τοις άλλοις, θα μπορούσε να εξεταστεί η χρήση διαφορετικών μορφών κειμενικής εισόδου ή ανεξάρτητων συνόλων δεδομένων, ώστε να αξιολογηθεί κατά πόσο η απόδοση του προτεινόμενου μοντέλου διατηρείται και σε διαφορετικά σενάρια εφαρμογής.

Συνολικά, η εργασία αυτή αποτελεί μια πρώτη προσέγγιση στην αυτόματη ταξινόμηση hashtags με χρήση τεχνικών μηχανικής μάθησης και βαθιάς μάθησης. Τα αποτελέσματά της δείχνουν ότι η αξιοποίηση συνδυαστικών πληροφοριών μπορεί να βελτιώσει σημαντικά την απόδοση των μοντέλων, ενώ παράλληλα δημιουργεί τις βάσεις για περαιτέρω έρευνα και ανάπτυξη πιο προηγμένων συστημάτων ταξινόμησης.

BIBΛΙΟΓΡΑΦΙΑ

- [1] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons*, 53(1), 59-68.
- [2] Dhingra, M., & Mudgal, R. K. (2019, March). Historical evolution of social media: An overview. In *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019*, Uttaranchal University, Dehradun, India.
- [3] Rhee, L., Bayer, J. B., Lee, D. S., & Kuru, O. (2021). Social by definition: How users define social platforms and why it matters. *Telematics and informatics*, 59, 101538.
- [4] Fan, W., & Gordon, M. D. (2014). The power of social media analytics. *Communications of the ACM*, 57(6), 74-81.
- [5] Batrinca, B., & Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1), 89-116.
- [6] Caleffi, P. M. (2015). The 'hashtag': A new word or a new rule?. *SKASE journal of theoretical linguistics*, 12(2), 46.
- [7] Bruns, A., & Liang, Y. E. (2012). Tools and methods for capturing Twitter data during natural disasters. *First Monday*, 17(4), 1-8.
- [8] Hoang, T. A., & Lim, E. P. (2017). Modeling topics and behavior of microbloggers: An integrated approach. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8(3), 1-37.
- [9] Amiri, B., Karimianghadim, R., Yazdanjue, N., & Hossain, L. (2021). Research topics and trends of the hashtag recommendation domain. *Scientometrics*, 126, 2689 - 2735. <https://doi.org/10.1007/s11192-021-03874-6>
- [10] Pattanayak, P. K., Tripathy, R. M., & Padhy, S. (2024). A Semi-supervised Approach of Cluster-Based Topic Modeling for Effective Tweet Hashtag Recommendation. *SN Computer Science*, 5. <https://doi.org/10.1007/s42979-024-03299-x>
- [11] Li, Z. (2026). Application Analysis of NLP Technology in Understanding Psychological Intervention for College Students. In *ITM Web of Conferences* (Vol. 88, p. 01030). EDP Sciences.
- [12] Surekha, N., Teja, A., Raja, S., Jakeer, S., Dheeraj, A., Bansal, S., Prakash, K., Singh, S., Rashed, M., Faruque, I., & Al-Mugren, K. (2025). Leveraging hybrid model for accurate sentiment analysis of Twitter data. *Scientific Reports*, 15. <https://doi.org/10.1038/s41598-025-09794-2>
- [13] Wei, F., Neary, R., Qin, H., Mao, Q., & Zhang, J. (2025). Empirical Evaluation of Embedding Models in the Context of Text Classification in Document Review in Construction Delay Disputes. *ArXiv*, abs/2501.09859. <https://doi.org/10.48550/arxiv.2501.09859>
- [14] Bell, J. (2022). What is machine learning?. *Machine learning and the city: applications in architecture and urban design*, 207-216.
- [15] Dogra, V., Verma, S., K., Chatterjee, P., Shafi, J., Choi, J., & Ijaz, M. F. (2022). A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience*, 2022. <https://doi.org/10.1155/2022/1883698>

- [16] Allam, H., Makubvure, L., Gyamfi, B., Graham, K. N., & Akinwolere, K. (2025). Text classification: How machine learning is revolutionizing text categorization. *Information*, 16(2), 130.
- [17] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [18] Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024, 69-79.
- [19] Pandey, A., & Jain, A. (2017). Comparative analysis of KNN algorithm using various normalization techniques. *International Journal of Computer Network and Information Security*, 10(11), 36.
- [20] Zhang, S. (2022). Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering*, 34(10), 4663–4675.
- [21] Valkenborg, D., Rousseau, A. J., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. *American journal of orthodontics and dentofacial orthopedics*, 164(5), 754-757.
- [22] Du, K.-L., Jiang, B., Lu, J., Hua, J., & Swamy, M. (2024). Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions. *Mathematics*. <https://doi.org/10.3390/math12243935>
- [23] Fan, J., Ma, C., & Zhong, Y. (2020). A selective overview of deep learning. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(2), 264.
- [24] Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., ... & Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. **Diagnostics**, 13*(15), 2582.
- [25] Das, S., Tariq, A., Santos, T., Kantareddy, S. S., & Banerjee, I. (2023). Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research. *Machine learning for Brain disorders*, 117-138.
- [26] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- [27] Staudemeyer, R. C., & Morris, E. R. (2019). Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. *arXiv preprint arXiv:1909.09586*.
- [28] Wang, Y., Xiao, Z., Fang, S., Li, W., Wang, J., & Zhao, X. (2022). BI-Directional long short-term memory for automatic detection of sleep apnea events based on single channel EEG signal. *Computers in biology and medicine*, 142, 105211.
- [29] Long, F., Zhou, K., & Ou, W. (2019). Sentiment Analysis of Text Based on Bidirectional LSTM With Multi-Head Attention. *IEEE Access*, 7, 141960-141969. <https://doi.org/10.1109/access.2019.2942614>
- [30] Liang, K., Wang, B., & Yuan, F. (2024). Research on Multi channel Text Classification Method Based on Neural Network. 2024 14th International Conference on Information Technology in Medicine and Education (ITME), 850-855. <https://doi.org/10.1109/itme63426.2024.00171>
- [31] Wang, Y., Liu, J., Huang, Y., & Feng, X. (2016). Using Hashtag Graph-Based Topic Model to Connect Semantically-Related Words Without

- Co-Occurrence in Microblogs. IEEE Transactions on Knowledge and Data Engineering, 28, 1919-1933. <https://doi.org/10.1109/tkde.2016.2531661>
- [32] Bansal, S., Gowda, K., AnupamaSureshababu, K., Kothari, C., & Kumar, N. (2025). A Comprehensive Review on Hashtag Recommendation: From Traditional to Deep Learning and Beyond. ArXiv, abs/2503.18669. <https://doi.org/10.48550/arxiv.2503.18669>
- [33] Yunida, R., Faisal, M. R., Indriani, F., Abadi, F., Budiman, I., & Prastya, S. E. (2023). LSTM and Bi-LSTM models for identifying natural disasters reports from social media. Journal of Electronics, Electromedical Engineering, and Medical Informatics, 5(4), 241-249.
- [34] Verma, M., Jain, R., & Monga, S. (2026). Comparative analysis of machine and deep learning models with text embeddings for sentiment analysis. Review of Computer Engineering Research. <https://doi.org/10.18488/76.v13i1.4709>
- [35] Singh, S., Kumar, K., & Kumar, B. (2026). A fusion of social media context for efficient multimodal sentiment analysis using deep learning techniques. Discov. Artif. Intell., 6, 329. <https://doi.org/10.1007/s44163-026-00986-x>