UNIVERSITY OF THE AEGEAN

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Α Ι Γ Α Ι Ο Υ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Ανίχνευση Εξαπάτησης από Κείμενο με Χρήση Νευρωνικών Δικτύων Βαθιάς Μάθησης

(Deception Detection from Text Using Deep Neural Networks)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Χρήστου Παπαντωνόπουλου

Α.Μ: 321/2020177

Επιβλέπων καθηγητής: ΕΥΣΤΑΘΙΟΣ ΣΤΑΜΑΤΑΤΟΣ

Εξεταστική επιτροπή: ΧΡΗΣΤΟΣ ΓΚΟΥΜΟΠΟΥΛΟΣ, ΘΕΟΔΩΡΟΣ ΚΩΣΤΟΥΛΑΣ

ΣΑΜΟΣ, [Σεπτέμβριος 2024]

Αυτή η σελίδα είναι σκόπιμα λευκή.

Πρόλογος και Ευχαριστίες

Ευχαριστώ πρώτα απ' όλα τον κύριο Ευστάθιο Σταματάτο, καθηγητή του Τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων, της Πολυτεχνικής Σχολής του Πανεπιστημίου Αιγαίου, για την επίβλεψη και την συστηματική καθοδήγησή του καθ' όλη τη διάρκεια αυτής της διπλωματικής εργασίας στο πεδίο της Τεχνητής Νοημοσύνης. Οι συμβουλές, η κριτική και η υποστήριξή του ήταν πάντοτε εύστοχες και απαραίτητες για την ολοκλήρωση αυτού του έργου.

Θέλω επίσης να εκφράσω τις ειλικρινείς μου ευχαριστίες στους φίλους και συμφοιτητές μου Αντρέα και Κυριακή, οι οποίοι με την υποστήριξή τους συνέβαλαν σημαντικά στην ολοκλήρωση αυτής της διπλωματικής. Οι συζητήσεις και οι ιδέες που αντάλλαξα μαζί τους ήταν που με ενθάρρυναν και με ενέπνευσαν να συνεχίσω και να ολοκληρώσω την εργασία αυτή.

Ακόμη, θέλω να ευχαριστήσω την οικογένειά μου για τη στήριξη και την κατανόηση κατά τη διάρκεια όλης της ακαδημαϊκής πορείας μου. Χωρίς την υποστήριξή τους, δεν θα ήταν εφικτό να φτάσω μέχρι εδώ.

Τέλος ευχαριστώ θερμά όλους όσους συνέβαλαν περισσότερο ή λιγότερο στην επιτυχή εκπόνηση αυτής της διπλωματικής εργασίας. Η εμπειρία που αποκόμισα ήταν εξαιρετικά ενδιαφέρουσα και εκπαιδευτική, και η συμβολή όσων στάθηκαν δίπλα μου είναι για μένα ανεκτίμητη.

Με εκτίμηση,

Χρήστος Παπαντωνόπουλος

2024

του

ΧΡΗΣΤΟΥ ΠΑΠΑΝΤΩΝΟΠΟΥΛΟΥ

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Αυτή η σελίδα είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Λίστα Εικόνων – Πινάκων

Λίστα Ακρωνυμίων

NLP	Natural Language Processing
DL	Deep Learning
ML	Machine Learning
RNNs	Recurrent Neural Networks
LSTM	Long Short Term Memory
CNNs	Convolutional Neural Networks
GRU	Gated Recurrent Unit
B.E.R.T	Bidirectional Encoder Representations
NER	Named Entity Recognition
AI	Artificial Intelligence
ANI	Artificial Narrow Intelligence
AGI	Artificial General Intelligence
POS	Part-of-Speech
LightGBM	Light Gradient Boosting Machine
ASI	Artificial Super Intelligence
SVM	Support Vector Machines
GPU	Graphics Processing Units
DNNs	Deep Neural Networks
MLPs	Multi-Layer Perceptrons
GPT	Generative Pre-trained Transformer
GBM	Gradient Boosting Machines
SMOTE	Synthetic Minority Over-sampling Technique
ADASYN	Adaptive Synthetic Sampling
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
RFs	Random Forests
GRNNs	Gated Recurrent Neural Networks
DRI	Deceptive Review Identification
RCNN	Recurrent Convolutional Neural Network
DFFNN	Deep Feedforward Neural Network
ReLU	Rectified Linear Unit
AP	Average Precision
TF-IDF	Term Frequency-Inverse Document Frequency
BoW	Bag of Words
CBOW	Continuous Bag of Words
GloVe	Global Vectors for Word Representation
BiRNN	Bidirectional Recurrent Neural Network
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Approach

Η σελίδα αυτή είναι σκόπιμα λευκή

Περίληψη

Στην ψηφιακή εποχή, ο πολλαπλασιασμός των ψευδών ειδήσεων παρουσιάζει σημαντικές προκλήσεις για την ακεραιότητα των πληροφοριών και την εμπιστοσύνη του κοινού. Η παρούσα διατριβή διερευνά την ανάπτυξη και εφαρμογή ενός προ-εκπαιδευμένου μοντέλου, ενός κλάδου της τεχνητής νοημοσύνης, για την ανίχνευση παραπλανητικού περιεχομένου σε άρθρα. Αξιοποιώντας τη δύναμη της Επεξεργασίας Φυσικής Γλώσσας (NLP) και της βαθιάς μάθησης, η έρευνα επικεντρώνεται στην αξιοποίηση ενός state-of-the-art προ-εκπαιδευμένου μοντέλου για τον εντοπισμό και την ταξινόμηση ψευδών ειδήσεων. Η αρχιτεκτονική του μοντέλου ενσωματώνει εξελιγμένη γλωσσική ανάλυση και ανάλυση συμφραζομένων για να διακρίνει τις λεπτές ενδείξεις εξαπάτησης που συχνά διαφεύγουν από τις παραδοσιακές μεθόδους ανίχνευσης. Εκτεταμένες εμπειρικές αξιολογήσεις καταδεικνύουν την αποτελεσματικότητα του μοντέλου όσον αφορά την επίτευξη υψηλής ακρίβειας και ευρωστίας σε διάφορα σύνολα δεδομένων. Τέλος η μελέτη αυτή επικεντρώνεται σε 5 διαφορετικά σύνολα εκπαίδευσης υλοποιώντας 7 διαφορετικά μοντέλα τα οποία με εφαρμογή κατάλληλων τεχνικών και παραμετροποίηση των παραμέτρων μας δίνουν στο τέλος τα βέλτιστα αποτελέσματα στην ανίχνευση ψευδών ειδήσεων.

Λέξεις Κλειδιά: Ταξινόμηση Ψευδών ειδήσεων, Προεπεξεργασία δεδομένων , Βαθιά Μάθηση, Τεχνητή Νοημοσύνη, Επεξεργασία Φυσικής Γλώσσας, Transformer, Προεκπαιδευμένο γλωσσικό μοντέλο.

Abstract

In the digital age, the proliferation of fake news presents significant challenges to information integrity and public trust. This thesis investigates the development and application of a pre-trained model, a branch of artificial intelligence, for detecting misleading content in articles. Leveraging the power of Natural Language Processing (NLP) and deep learning, the research focuses on utilizing a state-of-the-art pre-trained model for detecting and classifying fake news. The model architecture incorporates sophisticated linguistic and contextual analysis to discern subtle deception cues that are often missed by traditional detection methods. Extensive empirical evaluations demonstrate the effectiveness of the model in achieving high accuracy and robustness on various datasets. Finally, this study focuses on 5 different training sets implementing 7 different models which, with the application of appropriate techniques and parameterization of the parameters, give us the best results in the detection of fake news.

Keywords: False News Classification, Data preprocessing, Deep Learning, Artificial Intelligence, Natural Language Processing, Transformers, Pre-trained language model.

Δομή Διπλωματικής

Στο **1ο** κεφάλαιο γίνεται μία εισαγωγή στις εξελίξεις της πληροφορικής η οποία γίνεται κτήμα όλων των ανθρώπων φέρνοντας μαζί με τα αναμφισβήτητα οφέλη στην καθημερινή μας ζωή και προβλήματα που πρέπει να αναγνωριστούν και να αντιμετωπισθούν πάλι με τη χρήση της πληροφορικής και τους κλάδους της τεχνητής νοημοσύνης με μία επισκόπηση της επεξεργασίας φυσικής γλώσσας, που έχουν εφαρμογή στην ανίχνευση και αντιμετώπιση ψευδών ειδήσεων.

Στο **2ο** κεφάλαιο εξηγούνται πιο αναλυτικά θεωρητικά θέματα για την ταξινόμηση κειμένου και τους τύπους τους, έπειτα παρουσιάζονται τα συστήματα μέτρησης επίδοσης για τα μοντέλα που επιλέχθηκαν, τα συστήματα που ακολουθούν και αυτά που βασίζονται στη βαθιά και μηχανική μάθηση.

Το **3ο** κεφάλαιο επικεντρώνεται στην παρουσίαση αποτελεσμάτων από σχετικές ερευνητικές εργασίες που συναντώνται στην βιβλιογραφία και στην μέθοδο που ουσιαστικά αυτές ακολουθούν για να εξάγουν συμπεράσματα και αποτελέσματα σχετικά με την αντιμετώπιση των ψευδών ειδήσεων.

Στο **4ο** κεφάλαιο παρουσιάζονται η μεθοδολογία και τα σύνολα εκπαίδευσης που χρησιμοποιήθηκαν στην παρούσα εργασία.

Στο **5ο** κεφάλαιο παρουσιάζονται εκτενώς τα πειραματικά αποτελέσματα σχολιάζονται τα αποτελέσματα και η απόδοση την μοντέλων που χρησιμοποιήθηκαν και οι πιθανές μελλοντικές επεκτάσεις που θα μπορούσαν να πραγματοποιηθούν με σκοπό την ακόμη καλύτερη απόδοση των μοντέλων.

Τέλος στο **6ο** κεφάλαιο αναφέρονται τα συμπεράσματα που εξάγονται σχετικά με την προσέγγιση που ακολουθήθηκε.

Κεφάλαιο 1

Εισαγωγή

Στον 21ο αιώνα, την εποχή της πληροφορίας, τα smartphones και το διαδίκτυο έχουν γίνει αναπόσπαστο μέρος της ζωής μας. Η ψηφιοποίηση έχει αλλάξει κάθε πτυχή της καθημερινότητάς μας, επηρεάζοντας τον τρόπο με τον οποίο επικοινωνούμε και πληροφορούμαστε. Η ευκολία πρόσβασης σε πληροφορίες μέσω των κοινωνικών δικτύων έχει φέρει σημαντικά οφέλη, αλλά έχει επίσης οδηγήσει στην εξάπλωση ψευδών ειδήσεων και παραπληροφόρησης.

Από τις αρχές του 2000, η χρήση του διαδικτύου αυξήθηκε ραγδαία. Το 2000, λιγότερο από το 7% του παγκόσμιου πληθυσμού είχε πρόσβαση στο διαδίκτυο, ενώ σήμερα πάνω από το 50% είναι συνδεδεμένοι (Ritchie et al., 2023) (Dennis & Kahn, 2024). Ταυτόχρονα, η χρήση των κινητών τηλεφώνων έχει αυξηθεί από 740 εκατομμύρια συνδρομές το 2000 σε πάνω από 8 δισεκατομμύρια σήμερα, καθιστώντας τα κινητά τηλέφωνα πιο κοινά από ποτέ (Hillyer, 2020).

Η χρήση των μέσων κοινωνικής δικτύωσης έχει επίσης αυξηθεί δραματικά. Το 2004, λιγότερο από ένα εκατομμύριο άτομα χρησιμοποιούσαν το MySpace, ενώ σήμερα το Facebook έχει πάνω από 2,26 δισεκατομμύρια χρήστες (Hillyer, 2020). Παρά τα πλεονεκτήματα που προσφέρουν, τα κοινωνικά δίκτυα έχουν δημιουργήσει νέες προκλήσεις, όπως η εξάπλωση ψευδών ειδήσεων, που μπορούν να έχουν σοβαρές επιπτώσεις στην κοινωνία.

Η διάδοση ψευδών ειδήσεων αποτελεί σοβαρό πρόβλημα στον ψηφιακό κόσμο. Οι ψευδείς αναφορές, γνωστές και ως "fake news", μπορούν να δημιουργηθούν και να διαδοθούν με μεγάλη ταχύτητα, παραπλανώντας το κοινό και προκαλώντας κοινωνικές, πολιτικές και οικονομικές επιπτώσεις. Η ανίχνευση και ο εντοπισμός αυτών των ψευδών αναφορών αποτελούν σημαντική πρόκληση για τον ακαδημαϊκό χώρο και τις κυβερνήσεις παγκοσμίως.

Οι τεχνολογίες βαθιάς μάθησης και ειδικότερα οι αρχιτεκτονικές μετασχηματιστών (Transformers), όπως το γλωσσικό μοντέλο BERT, έχουν δείξει εξαιρετικά αποτελέσματα στην ταξινόμηση και ανίχνευση ψευδών ειδήσεων. Παρ' όλα αυτά, υπάρχουν ακόμη σοβαρά εμπόδια που πρέπει να αντιμετωπιστούν για την αποτελεσματική επίλυση αυτού του

προβλήματος. Αυτά περιλαμβάνουν την σαφήνεια στον ορισμό των ετικετών δεδομένων, τη γενίκευση των αποτελεσμάτων σε νέα δεδομένα, και την αντιμετώπιση των διαφόρων μορφών προκατάληψης στα δεδομένα (MCCLAIN, et al., 2021).

Τα κεντρικά σημεία:

Στη διπλωματική αυτή εργασία, θα εξετάσουμε τις τεχνικές που χρησιμοποιούνται για τον εντοπισμό ψευδών αναφορών στο διαδίκτυο, εστιάζοντας στις προκλήσεις και τις λύσεις που προσφέρονται από τις σύγχρονες τεχνολογίες τεχνητής νοημοσύνης.

1.1 Ανάλυση Φυσικής Γλώσσας – Natural Language Processing

Η Ανάλυση Φυσικής Γλώσσας (NLP) είναι ένας κλάδος της τεχνητής νοημοσύνης που επιτρέπει στους υπολογιστές να κατανοούν, να ερμηνεύουν και να παράγουν ανθρώπινη γλώσσα με τρόπους που είναι φυσικοί για τους ανθρώπους. Το NLP συνδυάζει διάφορες τεχνικές από την υπολογιστική γλωσσολογία και τη μηχανική μάθηση για να γεφυρώσει το χάσμα μεταξύ ανθρώπινης επικοινωνίας και υπολογιστικής κατανόησης. Η σημασία της NLP έγκειται στη δυνατότητά της να διευκολύνει την αλληλεπίδραση ανθρώπου-μηχανής, βελτιώνοντας την απόδοση συστημάτων και εφαρμογών σε τομείς όπως η εξυπηρέτηση πελατών, η αυτόματη μετάφραση, η ανάλυση συναισθήματος και πολλά άλλα (Holdsworth, 2024).

1.1.1 Εφαρμογές

Η NLP έχει πολλές εφαρμογές που χρησιμοποιούνται ευρέως στην καθημερινή ζωή. Μερικές από τις πιο κοινές περιλαμβάνουν:

- **Chatbots και Εικονικοί Βοηθοί**: Χρησιμοποιούνται για την αυτόματη αλληλεπίδραση με τους χρήστες, απαντώντας σε ερωτήσεις και προσφέροντας πληροφορίες. Παραδείγματα περιλαμβάνουν την Amazon Alexa και την Apple Siri, που χρησιμοποιούν την NLP για να αναγνωρίζουν και να ανταποκρίνονται σε φωνητικές εντολές (Holdsworth, 2024 ; Coursera Staff, 2024).
- **Αυτόματη Μετάφραση**: Εργαλεία όπως το Google Translate χρησιμοποιούν την NLP για να μεταφράζουν κείμενα από μια γλώσσα σε άλλη, διατηρώντας την ακρίβεια και το νόημα του πρωτότυπου κειμένου.

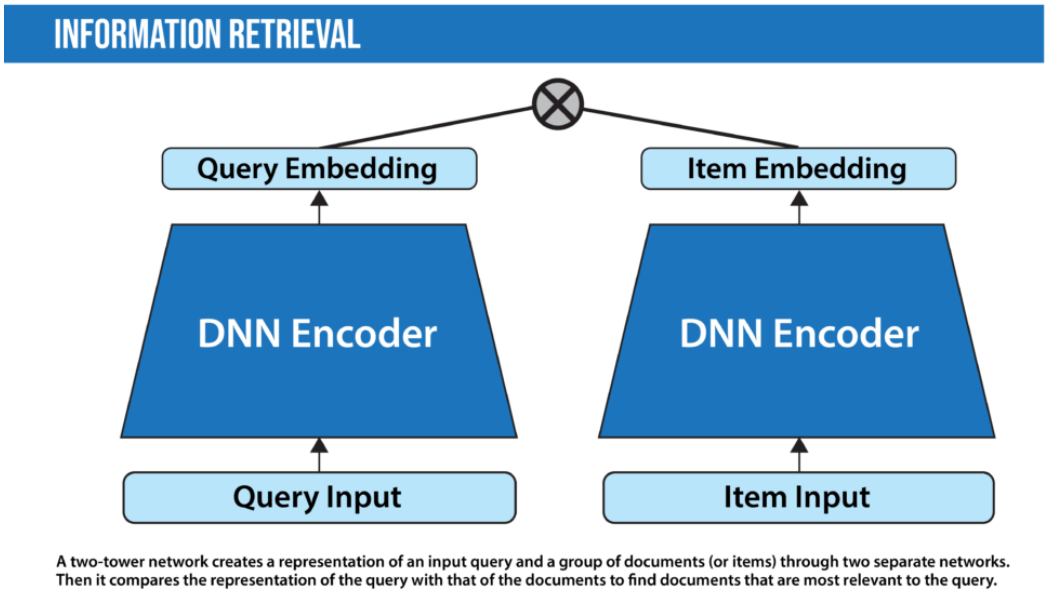
- **Ανάλυση Συναισθήματος**: Εφαρμόζεται στα μέσα κοινωνικής δικτύωσης και στις επιχειρηματικές εφαρμογές για να ανιχνεύει και να αναλύει τις συναισθηματικές αντιδράσεις των χρηστών σε πραγματικό χρόνο (Holdsworth, 2024).
- **Διόρθωση Κειμένου και Πρόβλεψη**: Η τεχνολογία που χρησιμοποιείται για την αυτόματη διόρθωση λαθών και την πρόβλεψη λέξεων σε εφαρμογές επεξεργασίας κειμένου και πληκτρολόγια κινητών τηλεφώνων (Coursera Staff, 2024).

1.1.2 Τεχνικές

Οι βασικές τεχνικές που χρησιμοποιούνται στην NLP περιλαμβάνουν:

- **Tokenization**: Η διαδικασία διαχωρισμού κειμένου σε μικρότερα κομμάτια, όπως λέξεις ή φράσεις ακόμη και χαρακτήρες που είναι γνωστά ως tokens. Είναι το πρώτο βήμα στην ανάλυση κειμένου και χρησιμοποιείται για την προετοιμασία των δεδομένων για περαιτέρω επεξεργασία (DeepLearning.AI, 2023).
- **Lemmatization και Stemming**: Μετατροπή των λέξεων στη βασική ή ριζική τους μορφή. Η lemmatization χρησιμοποιεί λεξικά για να εξάγει τη σωστή μορφή μιας λέξης, ενώ το stemming κόβει το τέλος των λέξεων για να βρει τη ρίζα τους.
- **Αναγνώριση Οντοτήτων (Named Entity Recognition - NER)**: Η αναγνώριση και κατηγοριοποίηση λέξεων ή φράσεων ως οντότητες, όπως ονόματα ατόμων, τοποθεσίες, οργανισμοί κ.λπ. Αυτό βοηθά στη δομή και την ερμηνεία του κειμένου (Holdsworth, 2024).
- **Ανάλυση Συναισθήματος**: Αποσκοπεί στην εξαγωγή υποκειμενικών πληροφοριών από το κείμενο, όπως συναισθήματα και στάσεις. Χρησιμοποιείται ευρέως στα μέσα κοινωνικής δικτύωσης και στην ανάλυση κριτικών (Holdsworth, 2024).
- **Μέρος του Λόγου (Part-of-Speech - POS) Tagging**: Η διαδικασία αναγνώρισης και κατηγοριοποίησης λέξεων ανάλογα με τον ρόλο τους στη φράση (ουσιαστικά, ρήματα, επίθετα κ.λπ.) (Holdsworth, 2024).
- **Dropout**: Το dropout είναι μια τεχνική τακτοποίησης που χρησιμοποιείται για τη μείωση της υπερεκπαίδευσης στα βαθιά νευρωνικά δίκτυα. Λειτουργεί τυχαία απενεργοποιώντας νευρώνες κατά τη διάρκεια της εκπαίδευσης, ώστε το δίκτυο να μην εξαρτάται υπερβολικά από συγκεκριμένα χαρακτηριστικά, ενισχύοντας έτσι τη γενίκευση του μοντέλου (Jason Brownlee,2019).

- **Gradient Clipping**: Το gradient clipping είναι μια τεχνική που χρησιμοποιείται για την αντιμετώπιση του προβλήματος των εκρηκτικών κλίσεων (exploding gradients) κατά την εκπαίδευση νευρωνικών δικτύων, ειδικά σε επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs). Οι εκρηκτικές κλίσεις προκαλούν πολύ μεγάλες τιμές στις παραγώγους της συνάρτησης κόστους, οδηγώντας σε αστάθεια και δυσκολία στην εκπαίδευση του μοντέλου(Jason Brownlee,2020).
- **Αρχικοποίηση Xavier**: Η Xavier Initialization, επίσης γνωστή ως Glorot Initialization, είναι μια μέθοδος αρχικοποίησης βαρών για νευρωνικά δίκτυα, προτεινόμενη από τους Xavier Glorot και Yoshua Bengio. Σκοπός της είναι να βοηθήσει στην εξισορρόπηση της διασποράς των εισόδων και εξόδων στα νευρωνικά στρώματα, ώστε να μειώσει το φαινόμενο των vanishing ή exploding gradients κατά την εκπαίδευση του μοντέλου(Jason Brownlee,2021).
- **Fine-tuning**: Το fine-tuning είναι μια τεχνική που χρησιμοποιείται στα βαθιά νευρωνικά δίκτυα για την προσαρμογή ενός προεκπαιδευμένου μοντέλου σε μια νέα, εξειδικευμένη εργασία. Αρχικά, το μοντέλο εκπαιδεύεται σε ένα μεγάλο γενικό σύνολο δεδομένων για να μάθει γενικά χαρακτηριστικά. Στη συνέχεια, το μοντέλο "λεπτορυθμίζεται" με εκπαίδευση σε ένα μικρότερο, πιο εξειδικευμένο σύνολο δεδομένων για να μάθει τα συγκεκριμένα χαρακτηριστικά της νέας εργασίας(Jason Brownlee,2020).
- **Χρήση bi-directional cells**: Σε αντίθεση με τα παραδοσιακά επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs), τα οποία χρησιμοποιούν μόνο το παρελθόν για την πρόβλεψη, τα bi-directional RNNs χρησιμοποιούν και το παρελθόν και το μέλλον, δίνοντας έτσι περισσότερες πληροφορίες στο δίκτυο για να βελτιώσει τις προβλέψεις του. Αυτό τα καθιστά ιδανικά για προβλήματα που απαιτούν την κατανόηση πλήρους συμφραζομένων της ακολουθίας(Jason Brownlee,2021).



Εικόνα 1: *Ανάκτηση Πληροφορίας*
Πηγή: https://www.deeplearning.ai/resources/natural-language-processing/

1.1.3 Προκλήσεις

Η ανάλυση φυσικής γλώσσας αντιμετωπίζει αρκετές προκλήσεις και προβλήματα:

- **Πολυπλοκότητα της Ανθρώπινης Γλώσσας**: Η ανθρώπινη γλώσσα είναι περίπλοκη και γεμάτη με ιδιωματισμούς, σαρδάμ, αμφισημίες και πολιτισμικές αποχρώσεις που δυσκολεύουν την κατανόηση από τους υπολογιστές. Η διόρθωση αυτών των προβλημάτων απαιτεί προηγμένες τεχνικές και σημαντική υπολογιστική ισχύ (Holdsworth, 2024) .
- **Διαφορετικότητα και Ποικιλομορφία**: Οι διάφορες γλώσσες και διάλεκτοι προσθέτουν επιπλέον πολυπλοκότητα. Κάθε γλώσσα έχει τις δικές της δομές, λεξιλόγιο και γραμματική, που πρέπει να ληφθούν υπόψη στην επεξεργασία φυσικής γλώσσας.
- **Διαχείριση Μεγάλων Δεδομένων**: Η επεξεργασία τεράστιων όγκων δεδομένων κειμένου απαιτεί υψηλή υπολογιστική ισχύ και αποτελεσματικούς αλγόριθμους για να διασφαλιστεί η ταχύτητα και η ακρίβεια της ανάλυσης .

- **Προκαταλήψεις και Ηθικά Ζητήματα**: Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων NLP συχνά περιέχουν προκαταλήψεις που μπορεί να οδηγήσουν σε μεροληπτικές αποφάσεις. Αυτό απαιτεί την ανάπτυξη τεχνικών για την ανίχνευση και μείωση των προκαταλήψεων στα δεδομένα και τους αλγόριθμους (Ramanathan, 2024).

Η ανάλυση φυσικής γλώσσας είναι ένας ζωτικός τομέας της τεχνητής νοημοσύνης που επιτρέπει την καλύτερη αλληλεπίδραση μεταξύ ανθρώπων και υπολογιστών. Παρά τις προκλήσεις που αντιμετωπίζει, οι εξελίξεις στις τεχνικές και την υπολογιστική ισχύ συνεχίζουν να βελτιώνουν την απόδοση και την ευχρηστία των εφαρμογών NLP, καθιστώντας τον έναν από τους πιο σημαντικούς τομείς στην τεχνολογία και την πληροφορική (Holdsworth, 2024 ; Ramanathan, 2024).

1.2 Τεχνητή Νοημοσύνη – Artificial Intelligence

Η τεχνητή νοημοσύνη (AI) ορίζεται ως η ικανότητα ενός συστήματος ή μηχανής να μιμείται τις ανθρώπινες γνωστικές λειτουργίες, όπως η μάθηση, η λήψη αποφάσεων και η επίλυση προβλημάτων. Η AI αναπτύσσεται με τη χρήση αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να εκτελούν εργασίες που παραδοσιακά απαιτούσαν ανθρώπινη νοημοσύνη. Η τεχνητή νοημοσύνη μπορεί να διακριθεί σε δύο κύριες κατηγορίες: τη στενή AI (Artificial Narrow Intelligence, ANI), που εκτελεί συγκεκριμένες εργασίες, και τη γενική AI (Artificial General Intelligence, AGI), που μπορεί να εκτελεί οποιαδήποτε νοητική εργασία που μπορεί να εκτελέσει ένας άνθρωπος.

Η ιστορία της τεχνητής νοημοσύνης ξεκινάει από τη δεκαετία του 1950 με την εργασία του Alan Turing, ο οποίος εισήγαγε την έννοια της μηχανικής σκέψης. Το 1956, το συνέδριο στο Dartmouth College θεωρείται ως το σημείο γέννησης της AI ως ξεχωριστού επιστημονικού τομέα. Κατά τις επόμενες δεκαετίες, η έρευνα στην AI πέρασε από περιόδους ενθουσιασμού και απογοήτευσης, γνωστές ως "AI winters," λόγω των περιορισμένων δυνατοτήτων του υλικού και του λογισμικού της εποχής.

Στη δεκαετία του 1980, η εμφάνιση των εξειδικευμένων συστημάτων (expert systems) έφερε νέα πνοή στον τομέα, ενώ η δεκαετία του 1990 είδε την ανάπτυξη των νευρωνικών δικτύων

και της μηχανικής μάθησης (machine learning), που επέτρεψαν την επεξεργασία μεγάλων ποσοτήτων δεδομένων με αυξανόμενη ακρίβεια. Σήμερα, η τεχνητή νοημοσύνη έχει εξελιχθεί σε ένα πολυδιάστατο πεδίο, που περιλαμβάνει υποτομείς όπως η βαθιά μάθηση (deep learning), η ανάλυση φυσικής γλώσσας (NLP), και η ρομποτική (Ahmed et al., 2021 ; Essa et al., 2023).

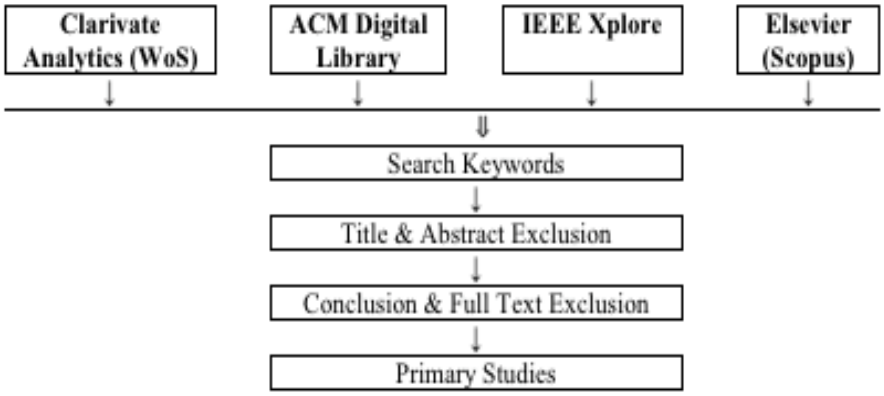
1.2.1 Σημασία στην Ανίχνευση Ψευδών Ειδήσεων

Η τεχνητή νοημοσύνη έχει αναδειχθεί ως ένα πολύτιμο εργαλείο για την ανίχνευση ψευδών ειδήσεων στο διαδίκτυο. Η αυξημένη εξάπλωση των κοινωνικών δικτύων και η ευκολία διάδοσης πληροφορίας έχουν καταστήσει την παραπληροφόρηση ένα σοβαρό πρόβλημα με κοινωνικές, πολιτικές και οικονομικές επιπτώσεις. Οι παραδοσιακές μέθοδοι επαλήθευσης πληροφοριών, όπως ο δημοσιογραφικός έλεγχος, είναι χρονοβόρες και μη ρεαλιστικές για την αντιμετώπιση του όγκου των δεδομένων που διακινούνται καθημερινά. Εδώ έρχεται η AI να προσφέρει λύσεις.

Τα συστήματα ανίχνευσης ψευδών ειδήσεων βασίζονται σε τεχνικές μηχανικής μάθησης και βαθιάς μάθησης για να αναλύσουν κείμενο και να προσδιορίσουν την εγκυρότητά του. Αυτές οι τεχνικές περιλαμβάνουν:

1. **Ανάλυση Κειμένου (Text Analysis)**: Η χρήση αλγορίθμων NLP για την ανάλυση των γλωσσικών χαρακτηριστικών ενός κειμένου. Αυτοί οι αλγόριθμοι εξετάζουν τη σύνταξη, τη γραμματική, και το ύφος του κειμένου για να εντοπίσουν ανωμαλίες που μπορεί να υποδεικνύουν ψευδείς ειδήσεις (Mishra & Sadia, 2023).
2. **Συγκριτική Ανάλυση Πηγών (Source Comparison)**: Η AI μπορεί να συγκρίνει τις πληροφορίες από διάφορες πηγές για να αξιολογήσει την ακρίβειά τους. Μηχανισμοί όπως τα αλγοριθμικά γραφήματα μπορούν να εντοπίσουν μοτίβα διασποράς πληροφορίας που σχετίζονται με ψευδείς ειδήσεις (Ahmed et al., 2021).
3. **Νευρωνικά Δίκτυα (Neural Networks)**: Η χρήση πολυστρωματικών νευρωνικών δικτύων και αρχιτεκτονικών βαθιάς μάθησης, όπως το BERT (Bidirectional Encoder Representations from Transformers), επιτρέπει την ανάλυση μεγάλων ποσοτήτων κειμένου και την ανίχνευση λεπτών γλωσσικών ενδείξεων που μπορεί να αποκαλύψουν την ψευδή φύση μιας είδησης (Essa et al., 2023).

4. **Εξόρυξη Δεδομένων (Data Mining)**: Χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων, οι αλγόριθμοι μπορούν να εντοπίσουν και να αναλύσουν μεγάλα σύνολα δεδομένων, αποκαλύπτοντας μοτίβα που σχετίζονται με την παραπληροφόρηση. Αυτή η διαδικασία επιτρέπει την αυτοματοποιημένη ανίχνευση ψευδών ειδήσεων με μεγαλύτερη ακρίβεια και ταχύτητα (Khalil et al., 2023).
5. **Μηχανική Μάθηση (Machine Learning)**: Τα μοντέλα μηχανικής μάθησης μπορούν να εκπαιδευτούν με την παροχή μεγάλων συνόλων δεδομένων που περιέχουν παραδείγματα ψευδών και αληθινών ειδήσεων. Με αυτόν τον τρόπο, τα μοντέλα αναπτύσσουν την ικανότητα να διακρίνουν μεταξύ ψευδών και αληθινών ειδήσεων με βάση χαρακτηριστικά όπως η χρήση της γλώσσας, η πηγή της πληροφορίας, και άλλα μετρικά χαρακτηριστικά (Essa et al., 2023).



Εικόνα 2: *Διάγραμμα μοντέλου διαδικασίας αναζήτησης.*
Πηγή: https://arxiv.org/pdf/2102.04458

Οι τεχνικές αυτές έχουν αποδειχθεί εξαιρετικά αποτελεσματικές σε διάφορες μελέτες. Για παράδειγμα, η χρήση του BERT σε συνδυασμό με μοντέλα LightGBM (Light Gradient Boosting Machine) έχει επιτύχει υψηλή ακρίβεια στην ανίχνευση ψευδών ειδήσεων, συνδυάζοντας την ανάλυση κειμένου με την αποδοτική εκπαίδευση και ταξινόμηση δεδομένων (Ahmed et al., 2021 ; Essa et al., 2023). Επίσης, η ενσωμάτωση μετασχηματιστικών μοντέλων όπως το RoBERTa και το ALBERT έχει ενισχύσει περαιτέρω την ικανότητα ανίχνευσης μέσω της κατανόησης πιο σύνθετων γλωσσικών δομών και περιεχομένου (Essa et al., 2023).

Επιπλέον, οι σύγχρονες τεχνικές επιτρέπουν τη συνδυαστική χρήση διαφόρων αλγορίθμων και μεθοδολογιών για την αύξηση της ακρίβειας και της αποδοτικότητας. Ένα παράδειγμα

είναι η υβριδική προσέγγιση που χρησιμοποιεί συνδυασμό CNN (Convolutional Neural Networks) και LSTM (Long Short-Term Memory) για την ανάλυση τόσο των γλωσσικών όσο και των χρονικών χαρακτηριστικών του κειμένου (Essa et al., 2023).

Η τεχνητή νοημοσύνη προσφέρει εξαιρετικά εργαλεία για την ανίχνευση ψευδών ειδήσεων, επιτρέποντας την ανάλυση μεγάλων ποσοτήτων δεδομένων με ακρίβεια και ταχύτητα που ξεπερνά τις παραδοσιακές μεθόδους. Η συνεχής εξέλιξη των αλγορίθμων και των μοντέλων AI, σε συνδυασμό με την αύξηση των διαθέσιμων δεδομένων, θα συνεχίσει να βελτιώνει την αποτελεσματικότητα της ανίχνευσης και θα συμβάλλει στην αντιμετώπιση του προβλήματος της παραπληροφόρησης στο διαδίκτυο (Mishra & Sadia, 2023 ; Ahmed et al., 2021; Essa et al., 2023).

1.2.2 Κατηγορίες

1.2.2.1 Στενή Τεχνητή Νοημοσύνη (ANI)

Η Στενή Τεχνητή Νοημοσύνη (Artificial Narrow Intelligence - ANI), γνωστή και ως Αδύναμη Τεχνητή Νοημοσύνη, αναφέρεται σε συστήματα τεχνητής νοημοσύνης που έχουν σχεδιαστεί για να εκτελούν συγκεκριμένα καθήκοντα ή ένα στενό σύνολο σχετικών καθηκόντων με υψηλό επίπεδο ακρίβειας. Αυτά τα συστήματα είναι προγραμματισμένα να λειτουργούν εντός ενός προκαθορισμένου συνόλου κανόνων και δεν μπορούν να επιδείξουν το ίδιο επίπεδο κατανόησης ή προσαρμοστικότητας με έναν άνθρωπο (Krishna, 2023).

Οι εφαρμογές της στενής τεχνητής νοημοσύνης είναι πολλές και περιλαμβάνουν:

- **Συστήματα Συστάσεων**: Υπηρεσίες όπως το Netflix και το Spotify χρησιμοποιούν στενή AI για να αναλύσουν τις προτιμήσεις των χρηστών και να παρέχουν εξατομικευμένες προτάσεις περιεχομένου.
- **Αυτόνομα Οχήματα**: Τα αυτοκινούμενα οχήματα χρησιμοποιούν AI για να ερμηνεύσουν τα δεδομένα αισθητήρων και να λαμβάνουν αποφάσεις σε πραγματικό χρόνο κατά την οδήγηση.
- **Υγεία**: Αλγόριθμοι AI βοηθούν στη διάγνωση ασθενειών, την ανάλυση ιατρικών εικόνων και την πρόβλεψη αποτελεσμάτων ασθενών (Ai—admin, 2024).

- **Οικονομικά**: Η AI χρησιμοποιείται για τον αλγοριθμικό εμπορικό συναλλαγών, την ανίχνευση απάτης και την αυτοματοποίηση της εξυπηρέτησης πελατών στον τραπεζικό τομέα (Ai—admin, 2024).
- **Κατασκευή**: Η AI βελτιστοποιεί τις διαδικασίες παραγωγής, βελτιώνει τη διαχείριση της αλυσίδας εφοδιασμού και ενισχύει τον ποιοτικό έλεγχο (Ai—admin, 2024).

1.2.2.2 Τεχνητή Γενική Νοημοσύνη (AGI)

Η Τεχνητή Γενική Νοημοσύνη (Artificial General Intelligence - AGI) αναφέρεται σε συστήματα AI που έχουν τη δυνατότητα να μάθουν, να κατανοούν και να εκτελούν οποιαδήποτε διανοητική εργασία που μπορεί να κάνει ένας άνθρωπος. Ενώ η AGI παραμένει θεωρητική και δεν έχει ακόμα επιτευχθεί, η έρευνα συνεχίζεται με στόχο τη δημιουργία συστημάτων που μπορούν να σκέφτονται και να αντιλαμβάνονται τον κόσμο με τον ίδιο τρόπο που το κάνει ένας άνθρωπος .

Οι προοπτικές της AGI είναι πολλές και περιλαμβάνουν:

- **Προσαρμοστικότητα**: Τα συστήματα AGI θα μπορούν να προσαρμόζονται σε νέες καταστάσεις χωρίς να απαιτείται επαναπρογραμματισμός.
- **Πολυλειτουργικότητα**: Η AGI θα μπορεί να εκτελεί πολλές και διαφορετικές εργασίες, από τη διάγνωση ασθενειών μέχρι τη σύνθεση μουσικής και την ανάπτυξη νέων τεχνολογιών (Awan, 2023).
- **Ανθρωποειδή Ρομπότ**: Τα ρομπότ με AGI θα μπορούν να αλληλοεπιδρούν με τους ανθρώπους με φυσικό και διαισθητικό τρόπο, προσφέροντας βοήθεια σε διάφορους τομείς της καθημερινότητας και της εργασίας.

Η AGI θα μπορούσε να φέρει σημαντικές αλλαγές στην κοινωνία, αλλά και προκλήσεις, όπως η ηθική χρήση και η υπευθυνότητα για τις αποφάσεις που λαμβάνονται από τέτοιες μηχανές (Krishna, 2023).

1.2.2.3 Τεχνητή Υπερνοημοσύνη (ASI)

Η Τεχνητή Υπερνοημοσύνη (Artificial Superintelligence - ASI) αντιπροσωπεύει την κορυφή της εξέλιξης της τεχνητής νοημοσύνης, όπου τα συστήματα όχι μόνο θα μπορούν να εκτελούν όλες τις εργασίες καλύτερα από τους ανθρώπους, αλλά και να αναπτύσσουν δικές τους

στρατηγικές σκέψης και λήψης αποφάσεων. Η ASI θα έχει ασύγκριτα ανώτερες δυνατότητες επεξεργασίας δεδομένων, μνήμης και λήψης αποφάσεων, οδηγώντας πιθανώς σε έναν κόσμο όπου η τεχνολογική ανάπτυξη θα είναι εκθετική και ανεξέλεγκτη, μια κατάσταση που συχνά αναφέρεται ως "τεχνολογική ιδιαιτερότητα" (singularity).

Οι πιθανές επιπτώσεις της ASI περιλαμβάνουν:

- **Κοινωνική Μεταμόρφωση**: Η ASI θα μπορούσε να μεταμορφώσει τη φύση της εργασίας, της οικονομίας και της κοινωνίας συνολικά, προσφέροντας λύσεις σε προβλήματα όπως η κλιματική αλλαγή και οι ανίατες ασθένειες, αλλά και δημιουργώντας νέες προκλήσεις για την ανθρώπινη απασχόληση και την οικονομική ανισότητα (Awan, 2023).
- **Ηθικά και Υπευθυνότητα**: Η ανάπτυξη της ASI θέτει σοβαρά ηθικά ζητήματα σχετικά με την ευθύνη και τον έλεγχο αυτών των συστημάτων. Η διασφάλιση ότι η ASI θα ενεργεί με τρόπους που είναι ευθυγραμμισμένοι με τις ανθρώπινες αξίες και την ηθική θα είναι ζωτικής σημασίας (Krishna, 2023).
- **Ασφάλεια και Κίνδυνοι**: Η δυνατότητα της ASI να υπερέχει σε όλες τις ανθρώπινες δραστηριότητες εγείρει ανησυχίες για την ασφάλεια, καθώς ένα υπερευφυές σύστημα θα μπορούσε να θεωρήσει την ανθρωπότητα ως απειλή και να ενεργήσει εναντίον της (Awan, 2023).

Η επίτευξη της ASI μπορεί να βρίσκεται ακόμη μακριά, αλλά η έρευνα και η προετοιμασία για τις πιθανές επιπτώσεις της είναι κρίσιμη για την κατανόηση και την αντιμετώπιση των προκλήσεων που θα φέρει.

Οι κατηγορίες της τεχνητής νοημοσύνης (ANI, AGI, ASI) αντιπροσωπεύουν διαφορετικά επίπεδα ικανοτήτων και δυνατοτήτων. Ενώ η στενή τεχνητή νοημοσύνη (ANI) είναι ήδη παρούσα και χρησιμοποιείται ευρέως σε πολλές εφαρμογές, η τεχνητή γενική νοημοσύνη (AGI) και η τεχνητή υπερνοημοσύνη (ASI) παραμένουν αντικείμενα εντατικής έρευνας και φιλοδοξίας. Καθώς προχωρά η τεχνολογική ανάπτυξη, είναι σημαντικό να προσεγγίσουμε τη δημιουργία και τη χρήση αυτών των συστημάτων με υπευθυνότητα και ηθική, λαμβάνοντας υπόψη τις ευρύτερες επιπτώσεις τους στην κοινωνία και την ανθρώπινη ζωή.

1.2.3 Εφαρμογές

Η ενσωμάτωση της τεχνητής νοημοσύνης (AI) σε διάφορους τομείς έχει επιφέρει σημαντικές αλλαγές και βελτιώσεις στη λειτουργικότητα και την αποδοτικότητα των συστημάτων. Η χρήση της AI έχει καταστεί κρίσιμη σε τομείς όπως τα ψηφιακά παιχνίδια, η επεξεργασία φυσικής γλώσσας, τα εξειδικευμένα συστήματα και τα συστήματα όρασης.

1.2.3.1 Ψηφιακά Παιχνίδια

Η τεχνητή νοημοσύνη έχει διαδραματίσει καθοριστικό ρόλο στη μεταμόρφωση του τομέα των ψηφιακών παιχνιδιών. Από την αρχική χρήση απλών συστημάτων κανόνων σε παιχνίδια όπως το Pac-Man, η AI έχει εξελιχθεί και χρησιμοποιείται πλέον για την ανάπτυξη πιο σύνθετων χαρακτήρων και συμπεριφορών στο παιχνίδι. Σήμερα, η AI επηρεάζει τη συμπεριφορά των NPCs (Non-Player Characters) και την προσαρμογή τους στις ενέργειες των παικτών, δημιουργώντας έτσι πιο δυναμικές και αλληλεπιδραστικές εμπειρίες παιχνιδιού (Takyar, 2024).

Οι προγραμματιστές παιχνιδιών χρησιμοποιούν AI για την ανάλυση δεδομένων σε πραγματικό χρόνο, επιτρέποντας την άμεση αντίδραση στις ενέργειες των παικτών και τη βελτιστοποίηση της εμπειρίας παιχνιδιού. Η ανάλυση αυτών των δεδομένων βοηθά στην προσαρμογή του επιπέδου δυσκολίας και στη δημιουργία εξατομικευμένων προκλήσεων που διατηρούν το ενδιαφέρον των παικτών. Επιπλέον, η AI συμβάλλει στην ανίχνευση και διόρθωση σφαλμάτων, βελτιώνοντας την ποιότητα των παιχνιδιών (Awan, 2023 ; Saleem, 2022).

1.2.3.2 Επεξεργασία Φυσικής Γλώσσας

Η επεξεργασία φυσικής γλώσσας (NLP) αποτελεί έναν από τους πιο σημαντικούς τομείς εφαρμογής της τεχνητής νοημοσύνης. Η NLP επιτρέπει στα μηχανήματα να κατανοούν και να επεξεργάζονται την ανθρώπινη γλώσσα, καθιστώντας δυνατή την αλληλεπίδραση μεταξύ ανθρώπων και μηχανών με φυσικό τρόπο. Χρησιμοποιείται ευρέως σε εφαρμογές όπως τα chatbots, τα συστήματα αυτόματης μετάφρασης και οι ψηφιακοί βοηθοί (Mishra & Sadia, 2023).

Η AI στην επεξεργασία φυσικής γλώσσας περιλαμβάνει τεχνικές όπως το tokenization, το lemmatization και η αναγνώριση οντοτήτων. Αυτές οι τεχνικές επιτρέπουν στα συστήματα να αναλύουν και να κατανοούν το περιεχόμενο των κειμένων, διευκολύνοντας την αυτόματη απόκριση και την παροχή ακριβών πληροφοριών στους χρήστες. Η χρήση της AI σε αυτόν τον τομέα έχει βελτιώσει σημαντικά την αποτελεσματικότητα και την ακρίβεια των συστημάτων επεξεργασίας φυσικής γλώσσας (Awan, 2023; Saleem, 2022).

1.2.3.3 Εξειδικευμένα Συστήματα

Τα εξειδικευμένα συστήματα (expert systems) αποτελούν μια κατηγορία εφαρμογών της τεχνητής νοημοσύνης που χρησιμοποιούνται για τη λήψη αποφάσεων. Αυτά τα συστήματα σχεδιάζονται να μιμούνται την ικανότητα ενός ανθρώπου ειδικού να λαμβάνει αποφάσεις βάσει δεδομένων και γνώσεων. Χρησιμοποιούνται ευρέως σε τομείς όπως η ιατρική διάγνωση, η χρηματοοικονομική ανάλυση και η διαχείριση έργων (Mishra & Sadia, 2023) (Essa et al., 2023).

Η κύρια λειτουργία των εξειδικευμένων συστημάτων είναι η λήψη τεκμηριωμένων αποφάσεων. Αυτά τα συστήματα λαμβάνουν δεδομένα ως είσοδο, τα επεξεργάζονται μέσω κανόνων και αλγορίθμων και παράγουν αποφάσεις ή συστάσεις. Η χρήση της AI σε αυτά τα συστήματα έχει βελτιώσει την ακρίβεια και την αποδοτικότητα της λήψης αποφάσεων, μειώνοντας παράλληλα τον χρόνο και τα κόστη που απαιτούνται για την ανάλυση δεδομένων (Mishra & Sadia, 2023; Essa et al., 2023).

1.2.3.4 Συστήματα Όρασης

Η τεχνητή νοημοσύνη έχει επίσης σημαντικό αντίκτυπο στα συστήματα όρασης, τα οποία χρησιμοποιούνται για την ανάλυση και την κατανόηση οπτικών πληροφοριών. Τα συστήματα όρασης βασίζονται σε τεχνικές επεξεργασίας εικόνας και βίντεο για την αναγνώριση και την ανάλυση αντικειμένων, προσώπων και σκηνών. Χρησιμοποιούνται σε τομείς όπως η ασφάλεια, η ιατρική απεικόνιση και η αυτόνομη οδήγηση (Mishra & Sadia, 2023 ; Essa et al., 2023).

Η AI στα συστήματα όρασης επιτρέπει την ανάπτυξη εφαρμογών που μπορούν να αναγνωρίζουν και να κατανοούν οπτικά δεδομένα με ακρίβεια. Αυτές οι εφαρμογές

περιλαμβάνουν την αναγνώριση προσώπων, την ανίχνευση αντικειμένων και την ανάλυση εικόνων. Η χρήση της AI έχει βελτιώσει την ικανότητα των συστημάτων όρασης να επεξεργάζονται μεγάλες ποσότητες δεδομένων και να παρέχουν ακριβείς και αξιόπιστες αναλύσεις.

Η ενσωμάτωση της τεχνητής νοημοσύνης σε διάφορους τομείς έχει επιφέρει σημαντικές βελτιώσεις στην απόδοση και την αποδοτικότητα των συστημάτων. Η χρήση της AI σε ψηφιακά παιχνίδια, επεξεργασία φυσικής γλώσσας, εξειδικευμένα συστήματα και συστήματα όρασης έχει επηρεάσει θετικά την ανάπτυξη και την καινοτομία σε αυτούς τους τομείς. Οι προοπτικές για το μέλλον της AI είναι εξαιρετικά ενθαρρυντικές, καθώς συνεχίζεται η έρευνα και η ανάπτυξη νέων τεχνολογιών και εφαρμογών (Mishra & Sadia, 2023).

1.2.4 Δυσκολίες και Προκλήσεις

Η τεχνητή νοημοσύνη (AI) προσφέρει απεριόριστες δυνατότητες και εφαρμογές, αλλά συνοδεύεται από σημαντικές προκλήσεις που πρέπει να αντιμετωπιστούν για να επιτευχθούν τα μέγιστα οφέλη. Αυτές οι προκλήσεις περιλαμβάνουν την υπολογιστική δύναμη, την ασφάλεια δεδομένων, το κόστος, τη σπανιότητα δεδομένων και την προκατάληψη στα δεδομένα.

1.2.5 Υπολογιστική Δύναμη

Η υπολογιστική δύναμη είναι ένα από τα πιο κρίσιμα ζητήματα που αντιμετωπίζει η τεχνητή νοημοσύνη. Τα συστήματα AI, ειδικά εκείνα που βασίζονται στη βαθιά μάθηση, απαιτούν τεράστιους πόρους επεξεργασίας. Οι αλγόριθμοι βαθιάς μάθησης χρειάζονται μεγάλες ποσότητες δεδομένων για την εκπαίδευση και απαιτούν ισχυρούς επεξεργαστές, όπως οι GPU (Graphics Processing Units), για να εκτελέσουν πολύπλοκους υπολογισμούς (Khalil et al ., 2023; Essa et al., 2023).

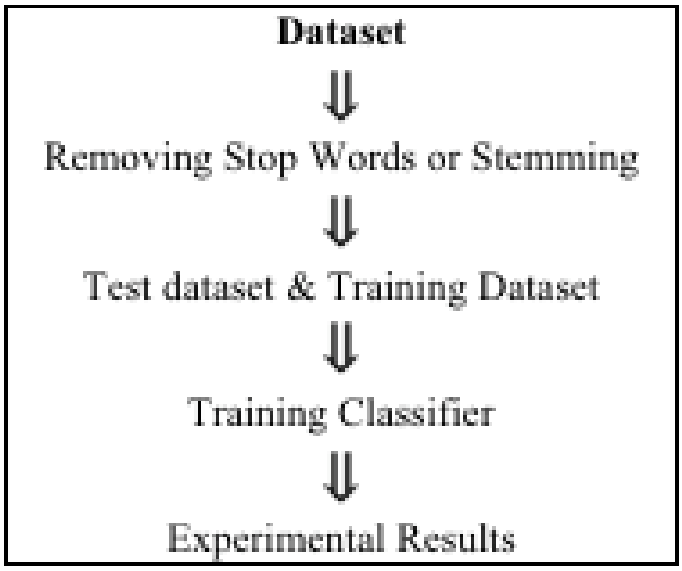
Η συνεχής αύξηση των απαιτήσεων σε υπολογιστική δύναμη σημαίνει ότι οι οργανισμοί πρέπει να επενδύουν σε ακριβά λογισμικά και υποδομές. Επιπλέον, η διαχείριση και η συντήρηση αυτών των συστημάτων μπορεί να είναι εξαιρετικά δαπανηρή και απαιτεί εξειδικευμένο προσωπικό. Καθώς οι ανάγκες για επεξεργασία αυξάνονται, η ενεργειακή

κατανάλωση των συστημάτων AI επίσης αυξάνεται, προκαλώντας ανησυχίες για την περιβαλλοντική τους επίδραση (Mishra & Sadia, 2023).

1.2.6 Απόρρητο και Ασφάλεια Δεδομένων

Η ασφάλεια και το απόρρητο των δεδομένων είναι κεντρικά ζητήματα στην εποχή της ψηφιακής πληροφορίας. Τα συστήματα AI συχνά επεξεργάζονται μεγάλες ποσότητες ευαίσθητων δεδομένων, συμπεριλαμβανομένων προσωπικών πληροφοριών και δεδομένων υγείας. Η διαχείριση αυτών των δεδομένων απαιτεί αυστηρά μέτρα ασφαλείας για την αποφυγή διαρροών και κακόβουλων επιθέσεων (Ahmed et al., 2021).

Οι παραβιάσεις δεδομένων μπορεί να έχουν σοβαρές συνέπειες, όπως οικονομικές απώλειες και ζημιά στη φήμη των οργανισμών. Επίσης, υπάρχει ο κίνδυνος της μη εξουσιοδοτημένης πρόσβασης σε προσωπικά δεδομένα, που μπορεί να οδηγήσει σε παραβιάσεις της ιδιωτικότητας και της προστασίας των προσωπικών δεδομένων. Οι κανονισμοί προστασίας δεδομένων, όπως ο GDPR στην Ευρώπη, θέτουν αυστηρούς κανόνες για την προστασία των δεδομένων, και οι οργανισμοί πρέπει να συμμορφώνονται με αυτούς για να αποφύγουν βαριά πρόστιμα και κυρώσεις (Essa et al., 2023; Ahmed et al., 2021).



Εικόνα 3: *Το σύνολο δεδομένων εκπαίδευσης εμφανίζει τα βήματα που χρησιμοποιούνται κατά την εκπαίδευση ενός ταξινομητή. Μετά την εκπαίδευση ένας ταξινομητής είναι που στη συνέχεια χρησιμοποιείται για πειράματα.* *Πηγή: https://arxiv.org/pdf/2102.04458*

1.2.7 Κόστος

Η ανάπτυξη και η υιοθέτηση της AI είναι μια δαπανηρή διαδικασία, ιδιαίτερα για μικρές και μεσαίες επιχειρήσεις. Το κόστος των απαιτούμενων υποδομών, του λογισμικού και της συντήρησης είναι υψηλό. Επιπλέον, η πρόσληψη και η εκπαίδευση εξειδικευμένου προσωπικού για τη διαχείριση των συστημάτων AI προσθέτει στο συνολικό κόστος (Ahmed et al., 2021).

Μεγάλες εταιρείες όπως η Google, η Amazon και η Microsoft έχουν τους πόρους να επενδύσουν σε τεχνολογίες AI και να αναπτύξουν πρωτοποριακές λύσεις. Ωστόσο, οι μικρότερες επιχειρήσεις συχνά δεν μπορούν να ανταγωνιστούν λόγω των υψηλών οικονομικών απαιτήσεων. Αυτό δημιουργεί ένα χάσμα στην υιοθέτηση της AI, όπου οι μεγάλες εταιρείες απολαμβάνουν τα οφέλη της τεχνολογίας, ενώ οι μικρότερες επιχειρήσεις μένουν πίσω (Mishra & Sadia, 2023).

1.2.8 Σπανιότητα Δεδομένων

Τα συστήματα AI βασίζονται σε μεγάλες ποσότητες δεδομένων για να εκπαιδευτούν και να λειτουργήσουν αποτελεσματικά. Ωστόσο, η πρόσβαση σε ποιοτικά δεδομένα είναι συχνά περιορισμένη. Η σπανιότητα δεδομένων μπορεί να οφείλεται σε διάφορους παράγοντες, όπως η ιδιωτικότητα των δεδομένων, οι περιορισμοί στην κοινή χρήση δεδομένων και η έλλειψη κατάλληλων δεδομένων για συγκεκριμένα προβλήματα (Essa et al., 2023).

Η έλλειψη δεδομένων καθιστά δύσκολη την ανάπτυξη αξιόπιστων μοντέλων AI. Επιπλέον, τα δεδομένα που είναι διαθέσιμα μπορεί να μην είναι αντιπροσωπευτικά ή να περιέχουν σφάλματα, οδηγώντας σε κακή απόδοση των μοντέλων. Η ανεύρεση και η συλλογή κατάλληλων δεδομένων είναι μια χρονοβόρα και δαπανηρή διαδικασία, που απαιτεί συνεχή επένδυση σε πόρους και τεχνολογίες (Khalil et al., 2023).

1.2.9 Προκατάληψη

Η προκατάληψη στα δεδομένα αποτελεί μια σημαντική πρόκληση για την AI. Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων μπορεί να περιέχουν σφάλματα και προκαταλήψεις που αντανακλούν τις κοινωνικές ανισότητες και τις διακρίσεις. Αυτές οι

προκαταλήψεις μπορούν να επηρεάσουν αρνητικά την απόδοση των μοντέλων και να οδηγήσουν σε αθέμιτες ή διακριτικές αποφάσεις (Mishra & Sadia, 2023; Essa et al., 2023).

Η αντιμετώπιση της προκατάληψης απαιτεί την ανάπτυξη τεχνικών για την ανίχνευση και την εξάλειψη των προκαταλήψεων στα δεδομένα. Οι ερευνητές και οι προγραμματιστές πρέπει να είναι προσεκτικοί κατά την επιλογή και την επεξεργασία των δεδομένων, εξασφαλίζοντας ότι είναι όσο το δυνατόν πιο αντιπροσωπευτικά και απαλλαγμένα από διακρίσεις. Επιπλέον, η συνεχής παρακολούθηση και αξιολόγηση των μοντέλων AI είναι απαραίτητη για την αναγνώριση και την αντιμετώπιση των προκαταλήψεων (Ahmed et al., 2021).

Η τεχνητή νοημοσύνη προσφέρει τεράστιες δυνατότητες, αλλά οι προκλήσεις που αντιμετωπίζει είναι σημαντικές. Η υπολογιστική δύναμη, η ασφάλεια δεδομένων, το κόστος, η σπανιότητα δεδομένων και η προκατάληψη αποτελούν κρίσιμους παράγοντες που πρέπει να αντιμετωπιστούν για την επίτευξη των μέγιστων οφελών από την AI. Με τη σωστή διαχείριση και την ανάπτυξη καινοτόμων λύσεων, αυτές οι προκλήσεις μπορούν να ξεπεραστούν, επιτρέποντας την πλήρη αξιοποίηση των δυνατοτήτων της τεχνητής νοημοσύνης.

1.2.10 Νοημοσύνη και Ηθική

Η χρήση της δημιουργεί μια σειρά από ηθικά διλήμματα και προκλήσεις που πρέπει να αντιμετωπιστούν προσεκτικά από την κοινωνία, τους κατασκευαστές και τους νομοθέτες. Σε αυτή την ενότητα, θα αναλύσουμε τα βασικά ηθικά ζητήματα, την υπευθυνότητα των κατασκευαστών και των προγραμματιστών AI και τη νομοθεσία που διέπει τη χρήση της AI.

1.2.11 Ηθικά Διλήμματα

Ένα από τα πιο συχνά αναφερόμενα ηθικά ζητήματα στην AI είναι η προκατάληψη και η αδικία. Τα συστήματα AI συχνά εκπαιδεύονται με δεδομένα που περιέχουν ανθρώπινες προκαταλήψεις, με αποτέλεσμα να αναπαράγουν και να ενισχύουν αυτές τις προκαταλήψεις σε εφαρμογές όπως η πρόσληψη προσωπικού, η δικαιοσύνη και η υγειονομική περίθαλψη. Για παράδειγμα, οι αλγόριθμοι πρόσληψης μπορεί να αποκλείσουν ορισμένες ομάδες ατόμων αν τα δεδομένα εκπαίδευσης περιέχουν προκαταλήψεις κατά αυτών των ομάδων (UNESCO, 2023, Boothman, 2020).

Ένα άλλο σημαντικό ηθικό ζήτημα είναι η ιδιωτικότητα των δεδομένων. Η εκπαίδευση μοντέλων AI απαιτεί τεράστια ποσά δεδομένων, πολλά από τα οποία περιέχουν προσωπικές πληροφορίες. Η συλλογή, αποθήκευση και επεξεργασία αυτών των δεδομένων δημιουργεί κινδύνους για την ιδιωτικότητα και την ασφάλεια των ατόμων, ειδικά όταν τα δεδομένα χρησιμοποιούνται χωρίς τη συγκατάθεση των ενδιαφερομένων (Green, 2020).

Η ανεργία είναι ένα ακόμη κρίσιμο ζήτημα που προκύπτει από την ευρεία χρήση της AI. Η αυτοματοποίηση και η χρήση της AI σε διάφορους τομείς μπορεί να οδηγήσει σε απώλεια θέσεων εργασίας, ιδιαίτερα σε επαγγέλματα που θεωρούνταν παραδοσιακά ασφαλή από την αυτοματοποίηση, όπως η ιατρική και η εκπαίδευση. Αυτό δημιουργεί την ανάγκη για νέες πολιτικές που θα στηρίζουν τους εργαζομένους και θα προωθούν την επανεκπαίδευσή τους (Green, 2020).

1.2.12 Υπευθυνότητα

Η υπευθυνότητα των κατασκευαστών και των προγραμματιστών AI είναι ένα από τα πιο κρίσιμα θέματα στην ανάπτυξη και τη χρήση της AI. Οι εταιρείες που αναπτύσσουν AI πρέπει να διασφαλίσουν ότι τα προϊόντα τους είναι ασφαλή, δίκαια και αξιόπιστα. Αυτό σημαίνει ότι πρέπει να υπάρχουν μηχανισμοί για τον εντοπισμό και την αντιμετώπιση των προκαταλήψεων στα δεδομένα εκπαίδευσης, καθώς και συστήματα για τη διασφάλιση της διαφάνειας στις αποφάσεις που λαμβάνονται από την AI (UNESCO, 2023).

Η υπευθυνότητα δεν περιορίζεται μόνο στους κατασκευαστές αλλά επεκτείνεται και στους χρήστες της AI. Οι οργανισμοί που χρησιμοποιούν AI πρέπει να κατανοούν τους περιορισμούς και τις δυνατότητες της τεχνολογίας και να διασφαλίζουν ότι χρησιμοποιούν τα συστήματα με τρόπο που να μην προκαλεί βλάβη. Για παράδειγμα, τα συστήματα AI που χρησιμοποιούνται στη δικαιοσύνη πρέπει να είναι διαφανή και να υπόκεινται σε αυστηρό έλεγχο για να αποφευχθούν αδικίες (Boothman, 2020).

1.2.13 Νομοθεσία

Η νομοθεσία που διέπει τη χρήση της AI είναι ακόμη σε πρώιμο στάδιο και διαφέρει από χώρα σε χώρα. Ωστόσο, υπάρχουν ορισμένα διεθνή πρότυπα και κατευθυντήριες γραμμές που προωθούνται από οργανισμούς όπως η UNESCO. Για παράδειγμα, η UNESCO έχει

εκδώσει την «Σύσταση για την Ηθική της Τεχνητής Νοημοσύνης», η οποία παρέχει κατευθυντήριες αρχές για την ανάπτυξη και τη χρήση της AI με βάση την προστασία των ανθρωπίνων δικαιωμάτων και την προώθηση της διαφάνειας και της δικαιοσύνης (UNESCO, 2023).

Η Ευρωπαϊκή Ένωση (ΕΕ) είναι επίσης πρωτοπόρος στη ρύθμιση της AI με το προτεινόμενο νομοθετικό πλαίσιο της, το οποίο στοχεύει στη διασφάλιση ότι η AI χρησιμοποιείται με τρόπο που να είναι ασφαλής και σεβαστός προς τα ανθρώπινα δικαιώματα. Το πλαίσιο αυτό περιλαμβάνει κανόνες για τη διαφάνεια, την ασφάλεια και την υπευθυνότητα των συστημάτων AI, καθώς και για την προστασία των προσωπικών δεδομένων (Boothman, 2020)
.

Στις Ηνωμένες Πολιτείες, η νομοθεσία για την AI είναι πιο κατακερματισμένη, με διάφορους νόμους και κανονισμούς να ισχύουν σε επίπεδο πολιτείας. Ωστόσο, υπάρχουν προσπάθειες για την ανάπτυξη ομοσπονδιακών κατευθυντήριων γραμμών που θα ρυθμίζουν τη χρήση της AI σε τομείς όπως η υγειονομική περίθαλψη, η χρηματοοικονομική βιομηχανία και η δικαιοσύνη (Green, 2020).

Η ηθική χρήση της τεχνητής νοημοσύνης είναι ένα σύνθετο και πολυδιάστατο ζήτημα που απαιτεί συνεργασία μεταξύ κυβερνήσεων, βιομηχανίας και κοινωνίας των πολιτών. Οι κατασκευαστές και οι προγραμματιστές AI έχουν την ευθύνη να διασφαλίσουν ότι τα συστήματα τους είναι δίκαια, ασφαλή και αξιόπιστα. Η νομοθεσία πρέπει να συνεχίσει να εξελίσσεται για να καλύπτει τα κενά και να διασφαλίζει την προστασία των ανθρωπίνων δικαιωμάτων. Μέσω της συνεργασίας και της συνεχούς αξιολόγησης, μπορούμε να επιτύχουμε την υπεύθυνη και ηθική χρήση της AI για το καλό της κοινωνίας.

1.3 Μηχανική Μάθηση – Machine Learning

Η μηχανική μάθηση (Machine Learning) είναι ένας κλάδος της τεχνητής νοημοσύνης που επικεντρώνεται στην ανάπτυξη αλγορίθμων και στατιστικών μοντέλων που επιτρέπουν στους υπολογιστές να εκτελούν συγκεκριμένες εργασίες χωρίς να είναι ρητά προγραμματισμένοι για αυτές. Η ουσία της μηχανικής μάθησης είναι η ικανότητα των συστημάτων να μαθαίνουν από δεδομένα και να βελτιώνονται με την πάροδο του χρόνου (Ahmed et al., 2021).

Η σημασία της μηχανικής μάθησης στην ανίχνευση ψευδών ειδήσεων είναι τεράστια. Οι παραδοσιακές μέθοδοι εντοπισμού ψευδών ειδήσεων βασίζονται σε ανθρώπινη παρέμβαση, η οποία είναι χρονοβόρα και αναποτελεσματική για την αντιμετώπιση του τεράστιου όγκου πληροφοριών που κυκλοφορούν καθημερινά στο διαδίκτυο. Η μηχανική μάθηση προσφέρει λύσεις που επιτρέπουν την αυτοματοποίηση της διαδικασίας ανίχνευσης, βελτιώνοντας την ακρίβεια και την ταχύτητα ανίχνευσης ψευδών ειδήσεων (Mishra & Sadia, 2023).

1.3.1 Κατηγορίες

Οι κύριες κατηγορίες της μηχανικής μάθησης περιλαμβάνουν την επιβλεπόμενη μάθηση, τη μη επιβλεπόμενη μάθηση και τη μάθηση ενίσχυσης.

Επιβλεπόμενη Μάθηση (Supervised Learning): Στην επιβλεπόμενη μάθηση, το σύστημα εκπαιδεύεται με ένα σύνολο δεδομένων εισόδου και εξόδου, γνωστό ως labeled data. Οι αλγόριθμοι μαθαίνουν να αντιστοιχούν εισόδους σε εξόδους και χρησιμοποιούνται ευρέως σε εφαρμογές όπως η ταξινόμηση κειμένων και η ανίχνευση ψευδών ειδήσεων (Ahmed et al., 2021).

Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): Η μη επιβλεπόμενη μάθηση χρησιμοποιεί δεδομένα χωρίς ετικέτες (unlabeled data) και επικεντρώνεται στην εύρεση κρυφών προτύπων ή δομών στα δεδομένα. Αυτή η προσέγγιση είναι χρήσιμη για την ανάλυση μεγάλων όγκων δεδομένων και την αναγνώριση μοτίβων που μπορούν να υποδείξουν ψευδείς ειδήσεις (Essa et al., 2023).

Μάθηση Ενίσχυσης (Reinforcement Learning): Στη μάθηση ενίσχυσης, οι αλγόριθμοι μαθαίνουν μέσω δοκιμών και σφαλμάτων, λαμβάνοντας ανταμοιβές ή ποινές για τις ενέργειές τους. Αν και η χρήση της μάθησης ενίσχυσης στην ανίχνευση ψευδών ειδήσεων δεν είναι τόσο διαδεδομένη όσο οι άλλες μέθοδοι, προσφέρει δυνατότητες για προσαρμοστικά συστήματα που βελτιώνονται με την πάροδο του χρόνου (Essa et al., 2023).

1.3.2 Αλγόριθμοι

Οι βασικοί αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται στον εντοπισμό ψευδών ειδήσεων περιλαμβάνουν:

Αλγόριθμοι Επιβλεπόμενης Μάθησης:

- **Λογιστική Παλινδρόμηση (Logistic Regression)**: Χρησιμοποιείται για τη μοντελοποίηση της πιθανότητας μιας δυαδικής μεταβλητής εξόδου. Είναι απλός και αποδοτικός για τη βασική ανίχνευση ψευδών ειδήσεων (Ahmed et al., 2021).
- **Δέντρα Απόφασης (Decision Trees)**: Αυτοί οι αλγόριθμοι διαχωρίζουν τα δεδομένα σε υποσύνολα βασισμένα σε κανόνες απόφασης. Τα δέντρα απόφασης είναι εύκολα στην κατανόηση και οπτικοποίηση και χρησιμοποιούνται συχνά σε συνδυασμό με άλλους αλγόριθμους (Essa et al., 2023).
- **Δάση Τυχαίων Δέντρων (Random Forests)**: Αποτελούν σύνολο από δέντρα απόφασης και παρέχουν καλύτερη ακρίβεια από τα μεμονωμένα δέντρα. Χρησιμοποιούνται ευρέως στην ανίχνευση ψευδών ειδήσεων για την ανάλυση μεγάλων συνόλων δεδομένων (Ahmed et al., 2021).
- **Υποστηρικτικές Διανυσματικές Μηχανές (Support Vector Machines - SVM)**: Αυτοί οι αλγόριθμοι χρησιμοποιούν υπερεπίπεδα για την ταξινόμηση δεδομένων και είναι ιδιαίτερα αποτελεσματικοί σε περιπτώσεις με υψηλή διαστασιμότητα δεδομένων (Khalil et al., 2023).

Αλγόριθμοι Μη Επιβλεπόμενης Μάθησης:

- **Αλγόριθμος K-Means**: Χρησιμοποιείται για την ομαδοποίηση δεδομένων σε k ομάδες, βρίσκοντας πρότυπα που μπορούν να υποδείξουν ψευδείς ειδήσεις χωρίς την ανάγκη ετικετών (Mishra & Sadia, 2023).
- **Ιεραρχική Ομαδοποίηση (Hierarchical Clustering)**: Δημιουργεί μια ιεραρχία από ομάδες δεδομένων, χρήσιμη για την ανακάλυψη σχέσεων μεταξύ δεδομένων που μπορεί να μην είναι άμεσα εμφανείς (Ahmed et al., 2021).

Αλγόριθμοι Μάθησης Ενίσχυσης:

- **Q-Learning**: Είναι μια μορφή μάθησης ενίσχυσης που χρησιμοποιείται για την ανάπτυξη προσαρμοστικών συστημάτων. Μπορεί να βελτιώσει τη διαδικασία ανίχνευσης ψευδών ειδήσεων μέσω συνεχούς μάθησης από τα δεδομένα (Essa et al., 2023).

1.3.3 Εφαρμογές

Η μηχανική μάθηση έχει ευρεία εφαρμογή σε διάφορους τομείς, συμπεριλαμβανομένης της ανίχνευσης ψευδών ειδήσεων:

Μέσα Κοινωνικής Δικτύωσης: Οι πλατφόρμες κοινωνικών δικτύων χρησιμοποιούν αλγορίθμους μηχανικής μάθησης για την ανίχνευση και την επισήμανση ψευδών ειδήσεων. Αυτό βοηθά στη μείωση της εξάπλωσης παραπληροφόρησης και στη διατήρηση της αξιοπιστίας των πληροφοριών (Ahmed et al., 2021).

Δημοσιογραφία και Ειδησεογραφία: Ειδησεογραφικά πρακτορεία και οργανισμοί χρησιμοποιούν εργαλεία μηχανικής μάθησης για την αξιολόγηση της ακρίβειας των ειδήσεων πριν τη δημοσίευση τους. Αυτό συμβάλλει στη διατήρηση της εμπιστοσύνης του κοινού στα μέσα ενημέρωσης (Mishra & Sadia, 2023).

Ασφάλεια και Νομοθεσία: Οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για την ανίχνευση ψευδών πληροφοριών σε νομικά έγγραφα και για την πρόληψη της απάτης. Αυτό είναι κρίσιμο για τη διασφάλιση της ακεραιότητας των νομικών συστημάτων και της ασφάλειας των δεδομένων (Essa et al., 2023).

Εκπαίδευση: Στον τομέα της εκπαίδευσης, η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την ανίχνευση πλαστών ερευνητικών εργασιών και την πρόληψη της ακαδημαϊκής απάτης. Τα συστήματα αυτά βοηθούν στη διατήρηση της ακαδημαϊκής ακεραιότητας και στην ενίσχυση της ποιότητας της εκπαίδευσης (Ahmed et al., 2021).

1.3.4 Δυσκολίες και Προκλήσεις

Η μηχανική μάθηση αντιμετωπίζει αρκετές προκλήσεις που πρέπει να ξεπεραστούν για να επιτευχθεί αποτελεσματική ανίχνευση ψευδών ειδήσεων:

Υπολογιστική Ισχύς: Οι αλγόριθμοι μηχανικής μάθησης απαιτούν σημαντική υπολογιστική ισχύ, ειδικά όταν πρόκειται για την επεξεργασία μεγάλων δεδομένων. Αυτό μπορεί να είναι ένα σημαντικό εμπόδιο για οργανισμούς με περιορισμένους πόρους (Mishra & Sadia, 2023).

1.3.5 Ποιότητα Δεδομένων

Η ποιότητα των δεδομένων είναι κρίσιμη για την επιτυχία των αλγορίθμων μηχανικής μάθησης. Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων πρέπει να είναι ακριβή, πλήρη και αντιπροσωπευτικά. Στην πράξη, τα δεδομένα μπορεί να περιέχουν θόρυβο, ελλείψεις ή προκαταλήψεις, που μπορούν να επηρεάσουν αρνητικά την απόδοση των μοντέλων. Η εξασφάλιση υψηλής ποιότητας δεδομένων απαιτεί προσεκτική προεπεξεργασία και καθαρισμό των δεδομένων, διαδικασίες που μπορεί να είναι χρονοβόρες και απαιτητικές (Ahmed et al., 2021 ; Essa et al., 2023).

Υπερεκπαίδευση και Υποεκπαίδευση: Η υπερεκπαίδευση (overfitting) συμβαίνει όταν ένα μοντέλο μαθαίνει υπερβολικά καλά τα δεδομένα εκπαίδευσης, αλλά αποτυγχάνει να γενικεύσει σε νέα δεδομένα. Αντίθετα, η υποεκπαίδευση (underfitting) συμβαίνει όταν το μοντέλο δεν καταφέρνει να μάθει επαρκώς τα δεδομένα εκπαίδευσης, οδηγώντας σε χαμηλή απόδοση. Η εξισορρόπηση μεταξύ υπερεκπαίδευσης και υποεκπαίδευσης είναι μία από τις μεγαλύτερες προκλήσεις στη μηχανική μάθηση και απαιτεί προσεκτική επιλογή αλγορίθμων και παραμέτρων (Ahmed et al., 2021 ; Essa et al., 2023).

Κλιμάκωση και Υπολογιστική Ισχύς: Οι αλγόριθμοι μηχανικής μάθησης, ειδικά αυτοί της βαθιάς μάθησης, απαιτούν σημαντικούς υπολογιστικούς πόρους για την εκπαίδευση και την εκτέλεση. Η κλιμάκωση των μοντέλων για την επεξεργασία μεγάλων συνόλων δεδομένων μπορεί να είναι δύσκολη και δαπανηρή. Επιπλέον, η εκπαίδευση αυτών των μοντέλων μπορεί να διαρκέσει από ώρες έως ημέρες, ανάλογα με την πολυπλοκότητα του αλγορίθμου και το μέγεθος των δεδομένων (Mishra & Sadia, 2023; Essa et al., 2023).

Ασφάλεια και Απόρρητο: Η διαχείριση μεγάλων ποσοτήτων δεδομένων εγείρει σοβαρά ζητήματα ασφάλειας και απορρήτου. Οι οργανισμοί πρέπει να διασφαλίσουν ότι τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων είναι ασφαλή και προστατευμένα από μη εξουσιοδοτημένη πρόσβαση. Επιπλέον, η χρήση δεδομένων που περιέχουν ευαίσθητες πληροφορίες απαιτεί συμμόρφωση με κανονισμούς προστασίας δεδομένων, όπως ο Γενικός Κανονισμός Προστασίας Δεδομένων (GDPR) στην Ευρωπαϊκή Ένωση (Ahmed et al., 2021; Essa et al., 2023).

Ερμηνευσιμότητα: Οι σύγχρονοι αλγόριθμοι μηχανικής μάθησης, ειδικά τα μοντέλα βαθιάς μάθησης, συχνά λειτουργούν ως «μαύρα κουτιά», όπου η λογική πίσω από τις αποφάσεις τους δεν είναι εύκολα κατανοητή. Η έλλειψη ερμηνευσιμότητας μπορεί να αποτελέσει σημαντικό εμπόδιο στην αποδοχή και εμπιστοσύνη των συστημάτων AI από τους χρήστες και τους ρυθμιστικούς φορείς. Η ανάπτυξη τεχνικών που καθιστούν τα μοντέλα πιο διαφανή και ερμηνεύσιμα είναι ένα ενεργό πεδίο έρευνας στη μηχανική μάθηση (Ahmed et al., 2021 ; Essa et al., 2023).

Προκαταλήψεις και Δικαιοσύνη: Τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση των μοντέλων μηχανικής μάθησης μπορεί να περιέχουν προκαταλήψεις που αντικατοπτρίζουν τις κοινωνικές ανισότητες. Αυτές οι προκαταλήψεις μπορούν να αναπαραχθούν και να ενισχυθούν από τα μοντέλα, οδηγώντας σε άδικες αποφάσεις. Η εξασφάλιση της δικαιοσύνης και της αμεροληψίας στα μοντέλα μηχανικής μάθησης απαιτεί προσεκτική ανάλυση των δεδομένων και ανάπτυξη αλγορίθμων που μπορούν να ανιχνεύσουν και να διορθώσουν τις προκαταλήψεις (Essa et al., 2023).

Η μηχανική μάθηση αποτελεί αναπόσπαστο κομμάτι της σύγχρονης τεχνολογίας και παίζει σημαντικό ρόλο στην ανίχνευση ψευδών ειδήσεων. Παρά τις προκλήσεις που αντιμετωπίζει, η συνεχής εξέλιξη των αλγορίθμων και των υπολογιστικών δυνατοτήτων, σε συνδυασμό με τη βελτίωση της ποιότητας των δεδομένων και της ερμηνευσιμότητας των μοντέλων, αναμένεται να οδηγήσει σε ακόμα πιο αποδοτικά και αξιόπιστα συστήματα στο μέλλον (Ahmed et al., 2021 ; Essa et al., 2023).

1.4 Βαθιά Μάθηση – Deep Learning

Η βαθιά μάθηση (Deep Learning) είναι ένας κλάδος της μηχανικής μάθησης που χρησιμοποιεί πολυεπίπεδα νευρωνικά δίκτυα για την ανάλυση και επεξεργασία δεδομένων. Εμπνευσμένα από τη δομή και τη λειτουργία του ανθρώπινου εγκεφάλου, τα νευρωνικά δίκτυα προσπαθούν να μιμηθούν την ικανότητα των ανθρώπων να μαθαίνουν από εμπειρίες και να αναγνωρίζουν μοτίβα. Βασικό χαρακτηριστικό της βαθιάς μάθησης είναι η χρήση πολλών στρωμάτων (layers) νευρώνων, που επιτρέπουν την εξαγωγή πολύπλοκων χαρακτηριστικών από τα δεδομένα.

Ένα βασικό στοιχείο της βαθιάς μάθησης είναι τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks, DNNs), τα οποία αποτελούνται από τρία ή περισσότερα στρώματα: το εισόδου, τα κρυφά (hidden layers), και το εξόδου. Αυτά τα δίκτυα χρησιμοποιούν αλγόριθμους βελτιστοποίησης όπως ο αλγόριθμος πίσω διάδοσης (backpropagation) για την αναπροσαρμογή των βαρών των νευρώνων, βελτιώνοντας την απόδοση του μοντέλου κατά τη διάρκεια της εκπαίδευσης.

1.4.1 Αλγόριθμοι – Αρχιτεκτονικές

Οι κύριοι αλγόριθμοι και οι αρχιτεκτονικές της βαθιάς μάθησης περιλαμβάνουν τα εξής:

1. **Πολυστρωματικοί Αντιληπτήρες (Multi-Layer Perceptrons, MLPs)**: Τα MLPs είναι απλά νευρωνικά δίκτυα που αποτελούνται από ένα στρώμα εισόδου, ένα ή περισσότερα κρυφά στρώματα και ένα στρώμα εξόδου. Χρησιμοποιούνται για την επίλυση προβλημάτων ταξινόμησης και παλινδρόμησης.
2. **Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Networks, CNNs)**: Τα CNNs είναι ειδικά σχεδιασμένα για την επεξεργασία δεδομένων με τοπική χωρική δομή, όπως εικόνες και βίντεο. Χρησιμοποιούν συνελικτικά και υποδειγματικά στρώματα για την εξαγωγή χαρακτηριστικών και τη μείωση της διάστασης των δεδομένων.
3. **Επαναληπτικά Νευρωνικά Δίκτυα (Recurrent Neural Networks, RNNs)**: Τα RNNs είναι κατάλληλα για την επεξεργασία ακολουθιακών δεδομένων, όπως το κείμενο και η ομιλία. Διαθέτουν κυκλικές συνδέσεις που τους επιτρέπουν να διατηρούν πληροφορίες από προηγούμενα χρονικά βήματα. Μια βελτίωση των RNNs είναι τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory, LSTM) που ξεπερνούν προβλήματα διασποράς και εξαφάνισης της πληροφορίας.
4. **Μετασχηματιστές (Transformers)**: Οι μετασχηματιστές έχουν φέρει επανάσταση στην επεξεργασία φυσικής γλώσσας και άλλων ακολουθιακών δεδομένων. Χρησιμοποιούν μηχανισμούς προσοχής (attention mechanisms) για να εστιάζουν σε σημαντικές πληροφορίες μέσα στις ακολουθίες δεδομένων. Διάσημα μοντέλα βασισμένα σε μετασχηματιστές είναι το BERT (Bidirectional Encoder Representations from Transformers) και το GPT (Generative Pre-trained Transformer).

1.4.2 Εφαρμογές

Η βαθιά μάθηση έχει ευρύ φάσμα εφαρμογών στον εντοπισμό ψευδών ειδήσεων, καθιστώντας την ιδιαίτερα χρήσιμη στον τομέα της ανάλυσης περιεχομένου και της αξιοπιστίας των πληροφοριών.

1. **Ανάλυση Κειμένου**: Τα μοντέλα βαθιάς μάθησης, όπως τα CNNs και τα RNNs, χρησιμοποιούνται για την ανάλυση και κατηγοριοποίηση κειμένων. Αυτά τα μοντέλα μπορούν να εντοπίζουν μοτίβα που συνδέονται με ψευδείς αναφορές, όπως υπερβολικές δηλώσεις και ανυπόστατες πληροφορίες.
2. **Ανίχνευση Στάσης Κειμένου (Stance Detection)**: Τα RNNs και οι μετασχηματιστές χρησιμοποιούνται για την ανίχνευση της στάσης που εκφράζεται σε ένα κείμενο σχετικά με ένα συγκεκριμένο θέμα. Αυτό βοηθά στον εντοπισμό αν το περιεχόμενο υποστηρίζει ή αντικρούει μια είδηση, συμβάλλοντας στην αξιολόγηση της αξιοπιστίας της.
3. **Δικτύωση και Διασπορά (Network and Propagation Analysis)**: Τα μοντέλα βαθιάς μάθησης αναλύουν τη διάδοση πληροφοριών στα κοινωνικά δίκτυα για να εντοπίσουν μοτίβα διάδοσης ψευδών ειδήσεων. Αυτή η ανάλυση βοηθά στον προσδιορισμό των πηγών και των δικτύων που συμβάλλουν στη διάδοση παραπληροφόρησης (Khalil et al., 2023).
4. **Ενσωμάτωση Πολυμέσων**: Η βαθιά μάθηση χρησιμοποιείται για την ανάλυση πολυμέσων, όπως εικόνες και βίντεο, για τον εντοπισμό αλλοιώσεων ή ψευδών στοιχείων που συνοδεύουν τις ειδήσεις. Τα CNNs είναι ιδιαίτερα χρήσιμα για την ανάλυση οπτικών δεδομένων και την ανίχνευση ψεύτικων εικόνων ή βίντεο.

1.4.3 Δυσκολίες – Προκλήσεις

Παρόλο που η βαθιά μάθηση έχει φέρει επανάσταση στην ανάλυση δεδομένων, αντιμετωπίζει αρκετές δυσκολίες και προκλήσεις:

1. **Απαιτήσεις Υπολογιστικής Ισχύος**: Η εκπαίδευση βαθιών νευρωνικών δικτύων απαιτεί σημαντικούς υπολογιστικούς πόρους, όπως ισχυρούς επεξεργαστές (CPUs) και κάρτες γραφικών (GPUs). Οι απαιτήσεις αυτές μπορεί να είναι απαγορευτικές για μικρές επιχειρήσεις ή ερευνητές με περιορισμένους πόρους (Mishra & Sadia, 2023).

2. **Δεδομένα και Προκαταλήψεις**: Η ποιότητα των δεδομένων είναι κρίσιμη για την απόδοση των μοντέλων βαθιάς μάθησης. Τα δεδομένα που περιέχουν προκαταλήψεις μπορούν να οδηγήσουν σε λανθασμένα ή μεροληπτικά αποτελέσματα. Επιπλέον, η συλλογή και ετικετοποίηση μεγάλων ποσοτήτων δεδομένων είναι χρονοβόρα και δαπανηρή διαδικασία (Ahmed et al., 2021; Essa et al., 2023).
3. **Ερμηνευσιμότητα (Interpretability)**: Τα βαθιά νευρωνικά δίκτυα συχνά χαρακτηρίζονται ως "μαύρα κουτιά" λόγω της πολυπλοκότητάς τους. Η έλλειψη διαφάνειας στον τρόπο λήψης αποφάσεων μπορεί να είναι προβληματική, ειδικά σε εφαρμογές όπου απαιτείται εξήγηση των αποτελεσμάτων, όπως η ιατρική διάγνωση ή η χρηματοοικονομική ανάλυση.
4. **Υπερεκπαίδευση (Overfitting)**: Η υπερεκπαίδευση είναι ένα συνηθισμένο πρόβλημα στα βαθιά νευρωνικά δίκτυα, όπου το μοντέλο μαθαίνει υπερβολικά καλά τα δεδομένα εκπαίδευσης αλλά αποτυγχάνει να γενικεύσει σε νέα δεδομένα. Τεχνικές όπως η κανονικοποίηση (regularization) και η χρήση dropout στρώσεων χρησιμοποιούνται για την αντιμετώπιση αυτού του προβλήματος (Ahmed et al., 2021).
5. **Ασφάλεια και Ιδιωτικότητα**: Η χρήση ευαίσθητων δεδομένων στην εκπαίδευση μοντέλων βαθιάς μάθησης εγείρει ζητήματα ασφάλειας και ιδιωτικότητας. Η προστασία των δεδομένων και η συμμόρφωση με κανονισμούς προστασίας προσωπικών δεδομένων είναι απαραίτητες για την αποφυγή παραβιάσεων και κατάχρησης των δεδομένων.

Η βαθιά μάθηση συνεχίζει να εξελίσσεται και να προσφέρει νέες δυνατότητες και εφαρμογές στην ανίχνευση ψευδών ειδήσεων και σε πολλούς άλλους τομείς της επιστήμης και της τεχνολογίας. Ωστόσο, για να συνεχίσει να προοδεύει και να προσφέρει λύσεις σε πραγματικά προβλήματα, είναι απαραίτητο να αντιμετωπιστούν οι παραπάνω προκλήσεις με καινοτόμες προσεγγίσεις και τεχνολογικές βελτιώσεις.

Αποτελεί μια από τις πιο ισχυρές τεχνολογίες στον τομέα της τεχνητής νοημοσύνης, με τεράστιες δυνατότητες εφαρμογών σε διάφορους τομείς, συμπεριλαμβανομένου του εντοπισμού ψευδών ειδήσεων. Οι βασικές έννοιες της βαθιάς μάθησης, οι κύριοι αλγόριθμοι

και οι αρχιτεκτονικές της, καθώς και οι εφαρμογές της στον εντοπισμό ψευδών ειδήσεων, παρουσιάζουν την ευρύτητα και την αποτελεσματικότητα αυτής της τεχνολογίας.

Παρά τις πολλές ευκαιρίες, η βαθιά μάθηση αντιμετωπίζει σημαντικές προκλήσεις, όπως οι απαιτήσεις υπολογιστικής ισχύος, τα ζητήματα δεδομένων και προκαταλήψεων, η ερμηνευσιμότητα των μοντέλων, η υπερεκπαίδευση, και οι ανησυχίες για την ασφάλεια και την ιδιωτικότητα. Για να αντιμετωπιστούν αυτές οι προκλήσεις, απαιτείται συνεχής έρευνα και ανάπτυξη, καθώς και η υιοθέτηση βέλτιστων πρακτικών στην ανάπτυξη και την εφαρμογή των τεχνολογιών βαθιάς μάθησης.

Η συνεχής πρόοδος στην τεχνολογία της βαθιάς μάθησης θα επιτρέψει τη δημιουργία πιο ισχυρών και αποδοτικών εργαλείων για την ανίχνευση ψευδών ειδήσεων και άλλων σκόπιμα ψευδών πληροφοριών, συμβάλλοντας στη βελτίωση της ποιότητας των πληροφοριών και της ενημέρωσης που διατίθενται στο ευρύ κοινό.

1.5 Στόχοι Διπλωματικής

Ο κύριος στόχος της διπλωματικής εργασίας είναι η χρησιμοποίηση και αξιολόγηση διαφορετικών μοντέλων τεχνητής νοημοσύνης για τον εντοπισμό ψευδών αναφορών στο διαδίκτυο, συγκεκριμένα fake news κάνοντας πειράματα με τη χρήση γνωστών ομάδων δεδομένων και γνωστών μοντέλων επεξεργασίας της φυσικής γλώσσας. Έτσι μπορεί ουσιαστικά να συμβάλει στη βελτίωση της ακρίβειας και της αποδοτικότητας των υφιστάμενων μοντέλων εντοπισμού παραπληροφόρησης. Επιμέρους στόχοι περιλαμβάνουν:

1. **Αναγνώριση και Κατηγοριοποίηση Ψευδών Αναφορών**: Ανάπτυξη μεθόδων για την αναγνώριση και κατηγοριοποίηση των ψευδών ειδήσεων σε πραγματικό χρόνο.
2. **Αξιολόγηση Απόδοσης Μοντέλων**: Συστηματική αξιολόγηση της απόδοσης διαφόρων αλγορίθμων μηχανικής μάθησης και βαθιάς μάθησης στην ανίχνευση fake news.
3. **Ενσωμάτωση Εξελιγμένων Τεχνικών NLP**: Χρήση προηγμένων τεχνικών επεξεργασίας φυσικής γλώσσας (NLP) για τη βελτίωση της ακρίβειας των προβλέψεων.
4. **Ελαχιστοποίηση Προκαταλήψεων**: Ανάπτυξη μεθόδων για την ελαχιστοποίηση των προκαταλήψεων στα δεδομένα και την εξασφάλιση δίκαιων αποτελεσμάτων.

5. **Σχεδιασμός και Υλοποίηση Πρωτοτύπου**: Δημιουργία ενός λειτουργικού πρωτοτύπου που θα εφαρμόζει τις προτεινόμενες τεχνικές σε πραγματικά δεδομένα από διάφορες πηγές ειδήσεων.

1.6 Μεθοδολογία

Η μεθοδολογία που θα ακολουθηθεί για την επίτευξη των στόχων περιλαμβάνει τα εξής στάδια:

1. **Συλλογή Δεδομένων**: Συλλογή μεγάλων συνόλων δεδομένων από διάφορες πηγές ειδήσεων, συμπεριλαμβανομένων των κοινωνικών μέσων, διαδικτυακών ειδησεογραφικών πρακτορείων και βάσεων δεδομένων με ετικετοποιημένα fake news.
2. **Προεπεξεργασία Δεδομένων**: Εφαρμογή τεχνικών προεπεξεργασίας δεδομένων για τον καθαρισμό και την κανονικοποίηση των δεδομένων. Αυτές οι τεχνικές περιλαμβάνουν την αφαίρεση θορύβου, τη μετατροπή σε μικρούς χαρακτήρες, την αφαίρεση σημείων στίξης και τη χρήση τεχνικών αποδιατύπωσης και απλοποίησης των λέξεων.
3. **Εκπαίδευση Μοντέλων**: Χρήση προηγμένων μοντέλων μηχανικής μάθησης και πιο συγκεκριμένα βαθιάς μάθησης όπως τα CNNs (Convolutional Neural Networks), RNNs *Recurrent Neural Networks), transformers (που χρησιμοποιούν attention mechanisms) και μοντέλα βασισμένα στους transformers όπως τα BERT (Bidirectional Encoder Representations from Transformers) και RoBERTa (Robustly optimized BERT approach) για την εκπαίδευση συστημάτων ανίχνευσης ψευδών ειδήσεων. Θα διερευνηθούν διαφορετικές αρχιτεκτονικές και παραμετροποιήσεις για την βελτιστοποίηση της απόδοσης.
4. **Αξιολόγηση Απόδοσης**: Χρήση μετρικών απόδοσης όπως η ακρίβεια δοκιμής, η ανάκληση, το F1 score, η ευσυνειδησία και η εξισορροπημένη ακρίβεια για την αξιολόγηση των μοντέλων. Η αξιολόγηση θα περιλαμβάνει επίσης τη σύγκριση με παραδοσιακές μεθόδους και την ανάλυση της απόδοσης σε διάφορα υποσύνολα δεδομένων.
5. **Βελτιστοποίηση Μοντέλων**: Εφαρμογή τεχνικών βελτιστοποίησης όπως η ρύθμιση υπερπαραμέτρων, η χρήση τεχνικών ενίσχυσης και η εφαρμογή μεθόδων αποφυγής

υπερπροσαρμογής (overfitting) όπως η αποκοπή (dropout) και η κανονικοποίηση (regularization).

6. **Ανάπτυξη Πρωτοτύπου**: Σχεδιασμός και υλοποίηση ενός λειτουργικού πρωτοτύπου που θα εφαρμόζει τις προτεινόμενες τεχνικές σε πραγματικά δεδομένα. Το πρωτότυπο θα αξιολογηθεί για την απόδοση του σε ζωντανά δεδομένα από διάφορες πηγές ειδήσεων.

1.7 Αναμενόμενα Αποτελέσματα

Τα αναμενόμενα αποτελέσματα από την παρούσα διπλωματική εργασία περιλαμβάνουν:

1. **Ανάπτυξη Αποτελεσματικών Μοντέλων Ανίχνευσης**: Δημιουργία μοντέλων που θα μπορούν να ανιχνεύουν ψευδείς αναφορές με υψηλή ακρίβεια και αξιοπιστία, βελτιώνοντας την απόδοση σε σχέση με υπάρχουσες μεθόδους (Ahmed et al., 2021 ; Nezafat, 2024).
2. **Βελτίωση Τεχνικών NLP**: Εφαρμογή καινοτόμων τεχνικών NLP που θα ενισχύσουν την ικανότητα των συστημάτων να κατανοούν και να επεξεργάζονται φυσική γλώσσα σε βάθος (Akhtar et al., 2023; Ashwin, 2019).
3. **Μείωση Προκαταλήψεων**: Εντοπισμός και μείωση των προκαταλήψεων στα δεδομένα και στα αποτελέσματα των μοντέλων, διασφαλίζοντας τη δικαιοσύνη και την ισότητα στις προβλέψεις (Nezafat, 2024).
4. **Συμβολή στην Έρευνα**: Παροχή νέων γνώσεων και δεδομένων στην επιστημονική κοινότητα για την ανίχνευση ψευδών ειδήσεων, συμβάλλοντας στην κατανόηση και την καταπολέμηση της παραπληροφόρησης.
5. **Πρακτική Εφαρμογή**: Δημιουργία ενός λειτουργικού συστήματος που θα μπορεί να ενσωματωθεί σε πλατφόρμες κοινωνικών μέσων και ειδησεογραφικών πρακτορείων για την αυτόματη ανίχνευση και αντιμετώπιση των ψευδών ειδήσεων σε πραγματικό χρόνο (Ahmed et al., 2021).

Με τη διπλωματική αυτή εργασία, επιδιώκεται η προώθηση της γνώσης στον τομέα της τεχνητής νοημοσύνης και της ανίχνευσης ψευδών αναφορών, προσφέροντας λύσεις που μπορούν να συμβάλουν στη βελτίωση της ποιότητας των πληροφοριών στο διαδίκτυο και στην προστασία του κοινού από την παραπληροφόρηση.

Κεφάλαιο 2

Ταξινόμηση Κειμένου για Εντοπισμό Ψευδών Αναφορών

2.1 Τύποι Ταξινόμησης

2.1.1 Δυαδική Ταξινόμηση (Binary Classification)

Η δυαδική ταξινόμηση είναι μια μορφή ταξινόμησης στην οποία ένα σύστημα καλείται να κατηγοριοποιήσει τα δεδομένα σε μία από δύο πιθανές κατηγορίες. Σε περιπτώσεις ανίχνευσης ψευδών ειδήσεων, οι κατηγορίες αυτές είναι συνήθως "ψευδείς" και "αληθείς". Το σύστημα αναλύει το κείμενο και αποφασίζει σε ποια από τις δύο κατηγορίες ανήκει το νέο κομμάτι δεδομένων.

Εφαρμογές στον εντοπισμό ψευδών ειδήσεων

Η δυαδική ταξινόμηση είναι ευρέως χρησιμοποιούμενη στον εντοπισμό ψευδών ειδήσεων. Οι αλγόριθμοι μηχανικής μάθησης και βαθιάς μάθησης εφαρμόζονται για να αναλύσουν άρθρα και αναρτήσεις στα μέσα κοινωνικής δικτύωσης, κατατάσσοντάς τα ως "αληθή" ή "ψευδή". Παραδείγματα τέτοιων εφαρμογών περιλαμβάνουν τη χρήση μοντέλων CNN-LSTM και υβριδικών νευρωνικών δικτύων που συνδυάζουν διάφορες τεχνικές για να βελτιώσουν την ακρίβεια ανίχνευσης ψευδών ειδήσεων (Akhtar et al., 2023; Khalil et al., 2023).

Πλεονεκτήματα και μειονεκτήματα

Πλεονεκτήματα:

1. **Απλότητα:** Η δυαδική ταξινόμηση είναι ευκολότερη στην υλοποίηση και απαιτεί λιγότερους υπολογιστικούς πόρους σε σύγκριση με πιο περίπλοκες μορφές ταξινόμησης, όπως η πολυκατηγορική ταξινόμηση.
2. **Υψηλή ακρίβεια σε συγκεκριμένα προβλήματα:** Σε περιπτώσεις όπου το πρόβλημα είναι σαφώς ορισμένο και τα δεδομένα είναι καλά διαχωρισμένα, οι αλγόριθμοι δυαδικής ταξινόμησης μπορούν να επιτύχουν υψηλή ακρίβεια.

3. **Ευκολία στην ερμηνεία αποτελεσμάτων:** Τα αποτελέσματα είναι εύκολα κατανοητά και επεξηγήσιμα, κάτι που είναι σημαντικό για την εφαρμογή σε τομείς όπως η δημοσιογραφία και η διακυβέρνηση.

Μειονεκτήματα:

1. **Περιορισμοί σε σύνθετα προβλήματα:** Η δυαδική ταξινόμηση μπορεί να μην είναι κατάλληλη για προβλήματα όπου τα δεδομένα δεν διαχωρίζονται εύκολα σε δύο κατηγορίες.
2. **Ευαισθησία στην ανισορροπία δεδομένων:** Όταν υπάρχει μεγάλη ανισορροπία μεταξύ των κατηγοριών (π.χ., πολύ περισσότερα αληθή από ψευδή άρθρα), η απόδοση των αλγορίθμων μπορεί να υποφέρει.
3. **Υπερεκπαίδευση:** Οι αλγόριθμοι μπορούν να υπερεκπαιδευτούν στα δεδομένα εκπαίδευσης, οδηγώντας σε μειωμένη γενίκευση και χαμηλή απόδοση σε νέα δεδομένα (Akhtar et al., 2023; Khalil et al., 2023).

Παραδείγματα αλγορίθμων

1. Logistic Regression (Λογιστική Παλινδρόμηση):

- Ένας απλός και αποδοτικός αλγόριθμος για δυαδική ταξινόμηση, ο οποίος χρησιμοποιεί μια λογιστική συνάρτηση για να μοντελοποιήσει την πιθανότητα μιας δυαδικής μεταβλητής εξόδου. Είναι ιδιαίτερα αποτελεσματικός για γραμμικά διαχωρίσιμα σύνολα δεδομένων (Khalil et al., 2023).

2. Support Vector Machines (SVM):

- Ένας αλγόριθμος που προσπαθεί να βρει το βέλτιστο διαχωριστικό υπερεπίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των δύο κατηγοριών. Τα SVMs είναι ισχυρά εργαλεία για την ανίχνευση ψευδών ειδήσεων, ειδικά όταν συνδυάζονται με τεχνικές εξαγωγής χαρακτηριστικών όπως οι λέξεις-κλειδιά και οι συνδυασμοί λέξεων (Khalil et al., 2023).

3. Naive Bayes:

- Βασίζεται στο θεώρημα του Bayes και είναι ιδιαίτερα χρήσιμος για προβλήματα ταξινόμησης κειμένου. Η απλότητά του και η ικανότητά του να λειτουργεί καλά με υψηλές διαστάσεις το καθιστούν δημοφιλές για την ανίχνευση ψευδών ειδήσεων (Akhtar et al., 2023).

4. Convolutional Neural Networks (CNNs):

- Χρησιμοποιούνται ευρέως για την επεξεργασία εικόνας, αλλά έχουν εφαρμοστεί και στην ανίχνευση ψευδών ειδήσεων με πολύ καλά αποτελέσματα. Τα CNNs μπορούν να εξάγουν χαρακτηριστικά από το κείμενο, αναγνωρίζοντας μοτίβα που υποδηλώνουν ψευδές περιεχόμενο (Akhtar et al., 2023).

5. Long Short-Term Memory Networks (LSTMs):

- Τα LSTMs είναι τύποι αναδρομικών νευρωνικών δικτύων που είναι ικανά να μάθουν μακροπρόθεσμες εξαρτήσεις. Στην ανίχνευση ψευδών ειδήσεων, χρησιμοποιούνται για την ανάλυση ακολουθιών κειμένου και την αναγνώριση σύνθετων μοτίβων ψευδούς πληροφόρησης (Khalil et al., 2023).

6. Hybrid Models (Υβριδικά Μοντέλα):

- Συνδυάζουν διάφορες τεχνικές μηχανικής και βαθιάς μάθησης για να βελτιώσουν την απόδοση. Ένα παράδειγμα είναι ο συνδυασμός CNNs με LSTMs για την ανίχνευση ψευδών ειδήσεων, που αξιοποιεί τα πλεονεκτήματα και των δύο τύπων δικτύων (Akhtar et al., 2023).

Οι παραπάνω αλγόριθμοι και μοντέλα αποτελούν τη βάση για την ανίχνευση ψευδών ειδήσεων μέσω δυαδικής ταξινόμησης, συμβάλλοντας στη βελτίωση της ακρίβειας και της αποτελεσματικότητας των συστημάτων εντοπισμού ψευδών ειδήσεων.

2.1.2 Ταξινόμηση Πολλών Κλάσεων (Multiclass Classification)

Η ταξινόμηση πολλών κλάσεων (multiclass classification) είναι μια τεχνική της μηχανικής μάθησης που χρησιμοποιείται για την κατηγοριοποίηση δειγμάτων σε περισσότερες από δύο κατηγορίες. Σε αντίθεση με την δυαδική ταξινόμηση, όπου τα δεδομένα ταξινομούνται σε

δύο κατηγορίες (π.χ., αληθές ή ψευδές), η ταξινόμηση πολλών κλάσεων επιτρέπει την κατηγοριοποίηση σε πολλαπλές, διακριτές κλάσεις. Αυτή η μορφή ταξινόμησης είναι απαραίτητη όταν τα δεδομένα παρουσιάζουν ποικιλία και δεν μπορούν να περιγραφούν επαρκώς με δύο μόνο κατηγορίες.

Για παράδειγμα, στον τομέα της αναγνώρισης εικόνας, ένα σύστημα ταξινόμησης μπορεί να κατηγοριοποιήσει εικόνες σε κατηγορίες όπως «γάτα», «σκύλος», «πουλί» κ.λπ. Αντίστοιχα, στον τομέα της ανάλυσης κειμένου, ένα σύστημα μπορεί να κατηγοριοποιήσει άρθρα ειδήσεων σε κατηγορίες όπως «πολιτική», «αθλητισμός», «ψυχαγωγία» κ.λπ. (Khalil et al., 2023; Essa et al., 2023).

Εφαρμογές σε ψευδείς αναφορές

Η ταξινόμηση πολλών κλάσεων παίζει κρίσιμο ρόλο στην ανίχνευση ψευδών ειδήσεων και άλλων μορφών παραπληροφόρησης στο διαδίκτυο. Οι ψευδείς αναφορές δεν είναι πάντα ίδιες και μπορεί να κατηγοριοποιούνται σε διάφορους τύπους με βάση τη φύση και την πρόθεσή τους. Μερικά παραδείγματα κατηγοριών περιλαμβάνουν:

1. **Προπαγάνδα**: Αναφορές που έχουν σχεδιαστεί για να επηρεάσουν την κοινή γνώμη με συγκεκριμένο πολιτικό ή κοινωνικό σκοπό.
2. **Κλικαρισμένες ειδήσεις (Clickbait)**: Τίτλοι και περιεχόμενο που έχουν σχεδιαστεί για να προσελκύσουν κλικ, συχνά με υπερβολικές ή ψευδείς δηλώσεις.
3. **Σατιρικές ειδήσεις**: Ειδήσεις που έχουν δημιουργηθεί για σατιρικούς σκοπούς και όχι για να παραπληροφορήσουν σκόπιμα.
4. **Ανακριβείς αναφορές**: Ειδήσεις που περιέχουν λάθη ή παραλείψεις χωρίς κακόβουλη πρόθεση (Mishra & Sadia, 2023; Ahmed et al., 2021).

Οι αλγόριθμοι ταξινόμησης πολλών κλάσεων μπορούν να βοηθήσουν στην αυτόματη κατηγοριοποίηση αυτών των ειδών ψευδών αναφορών, επιτρέποντας στις πλατφόρμες κοινωνικής δικτύωσης και στους ελέγχους γεγονότων να αντιμετωπίσουν πιο αποτελεσματικά την παραπληροφόρηση. Επιπλέον, βοηθούν στην αναγνώριση των μοτίβων και των χαρακτηριστικών που διακρίνουν κάθε τύπο ψευδούς αναφοράς, βελτιώνοντας έτσι την ακρίβεια και την αποτελεσματικότητα των συστημάτων ανίχνευσης.

Πλεονεκτήματα και μειονεκτήματα

Πλεονεκτήματα

1. **Ευελιξία**: Η δυνατότητα κατηγοριοποίησης σε πολλές κλάσεις προσφέρει μεγαλύτερη ευελιξία και λεπτομέρεια σε σύγκριση με τη δυαδική ταξινόμηση. Αυτό είναι ιδιαίτερα χρήσιμο σε εφαρμογές όπου τα δεδομένα παρουσιάζουν ποικιλία και πολυπλοκότητα.
2. **Αυξημένη Ακρίβεια**: Η ταξινόμηση σε περισσότερες κλάσεις μπορεί να οδηγήσει σε ακριβέστερα αποτελέσματα, καθώς επιτρέπει την καλύτερη διάκριση μεταξύ διαφορετικών τύπων δεδομένων.
3. **Βελτίωση Συστήματος Ανίχνευσης**: Στον τομέα της ανίχνευσης ψευδών ειδήσεων, η χρήση πολλαπλών κλάσεων επιτρέπει την πιο λεπτομερή κατηγοριοποίηση των ειδήσεων, βοηθώντας στην αναγνώριση των διαφορετικών μορφών παραπληροφόρησης (Ahmed et al., 2021; Essa et al., 2023).

Μειονεκτήματα

1. **Αυξημένη Πολυπλοκότητα**: Η ταξινόμηση πολλών κλάσεων είναι πιο πολύπλοκη και απαιτεί περισσότερους υπολογιστικούς πόρους σε σύγκριση με τη δυαδική ταξινόμηση. Η πολυπλοκότητα αυτή μπορεί να οδηγήσει σε μεγαλύτερους χρόνους εκπαίδευσης και πρόβλεψης.
2. **Δυσκολία στην Ετικετοποίηση Δεδομένων**: Η σωστή ετικετοποίηση δεδομένων για την εκπαίδευση ενός συστήματος ταξινόμησης πολλών κλάσεων μπορεί να είναι πιο απαιτητική και χρονοβόρα. Αυτό οφείλεται στην ανάγκη για ακριβή και συνεπή ετικετοποίηση σε πολλαπλές κλάσεις.
3. **Διαχείριση Σπανιότητας Δεδομένων**: Η ύπαρξη σπάνιων κλάσεων μπορεί να δημιουργήσει προβλήματα στη μοντελοποίηση, καθώς τα δεδομένα για αυτές τις κλάσεις μπορεί να μην είναι επαρκή για την εκπαίδευση ενός αποτελεσματικού μοντέλου (Mishra & Sadia, 2023).

Παραδείγματα αλγορίθμων

Υπάρχουν διάφοροι αλγόριθμοι που μπορούν να χρησιμοποιηθούν για την ταξινόμηση πολλών κλάσεων, ο καθένας με τα δικά του πλεονεκτήματα και μειονεκτήματα:

1. **Support Vector Machines (SVM)**: Οι SVM μπορούν να επεκταθούν για την ταξινόμηση πολλών κλάσεων χρησιμοποιώντας τεχνικές όπως το One-vs-Rest (OvR) και το One-vs-One (OvO). Αυτές οι τεχνικές δημιουργούν πολλαπλούς δυαδικούς ταξινομητές για να καλύψουν όλες τις πιθανές κλάσεις. Οι SVM είναι γνωστοί για την ακρίβειά τους και την ικανότητά τους να χειρίζονται μεγάλα σύνολα δεδομένων, αλλά μπορούν να είναι υπολογιστικά απαιτητικοί (Ahmed et al., 2021).
2. **Random Forest**: Ο αλγόριθμος Random Forest χρησιμοποιεί πολλαπλά δέντρα απόφασης για την ταξινόμηση των δεδομένων. Κάθε δέντρο ψηφίζει για την κατηγορία της εισόδου και η κατηγορία με τις περισσότερες ψήφους επιλέγεται ως η τελική πρόβλεψη. Το Random Forest είναι ανθεκτικό στον υπερπροσδιορισμό (overfitting) και μπορεί να διαχειριστεί μεγάλες ποσότητες δεδομένων και χαρακτηριστικών, αλλά η ερμηνευσιμότητα των αποτελεσμάτων μπορεί να είναι περιορισμένη (Mishra & Sadia, 2023) (Akhtar et al., 2023).
3. **Neural Networks**: Τα νευρωνικά δίκτυα, ειδικά τα βαθιά νευρωνικά δίκτυα (deep neural networks), μπορούν να χρησιμοποιηθούν για την ταξινόμηση πολλών κλάσεων. Αυτά τα δίκτυα μπορούν να μάθουν περίπλοκα μοτίβα στα δεδομένα και είναι ιδιαίτερα χρήσιμα σε εφαρμογές όπως η αναγνώριση εικόνας και η ανάλυση κειμένου. Ωστόσο, απαιτούν μεγάλο όγκο δεδομένων και υπολογιστικών πόρων για την εκπαίδευσή τους και μπορεί να είναι δύσκολο να εξηγηθούν τα αποτελέσματά τους (Ahmed et al., 2021).
4. **Gradient Boosting Machines (GBM)**: Οι GBM, όπως το XGBoost και το LightGBM, είναι αλγόριθμοι ενίσχυσης που χρησιμοποιούνται για την ταξινόμηση πολλών κλάσεων. Αυτοί οι αλγόριθμοι χτίζουν μοντέλα σε διαδοχικά στάδια, όπου κάθε νέο μοντέλο διορθώνει τα λάθη των προηγούμενων. Οι GBM είναι πολύ ισχυροί και συχνά χρησιμοποιούνται σε διαγωνισμούς μηχανικής μάθησης λόγω της υψηλής ακρίβειάς τους. Ωστόσο, η εκπαίδευσή τους μπορεί να είναι χρονοβόρα και υπολογιστικά απαιτητική (Essa et al., 2023).

2.1.3 Ταξινόμηση Πολλαπλών Ετικετών (Multi-label Classification)

Η ταξινόμηση πολλαπλών ετικετών (multi-label classification) είναι μια τεχνική μηχανικής μάθησης που επιτρέπει σε ένα μοντέλο να εκχωρήσει πολλαπλές ετικέτες ή κατηγορίες σε μια μοναδική είσοδο δεδομένων. Σε αντίθεση με την παραδοσιακή ταξινόμηση, όπου κάθε

δείγμα δεδομένων ανήκει σε μία μόνο κατηγορία, η ταξινόμηση πολλαπλών ετικετών αναγνωρίζει ότι τα δεδομένα μπορεί να ανήκουν σε περισσότερες από μία κατηγορίες ταυτόχρονα. Αυτή η προσέγγιση είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου τα δεδομένα είναι περίπλοκα και περιέχουν πολυδιάστατες πληροφορίες που δεν μπορούν να αποδοθούν επαρκώς σε μια μοναδική κατηγορία (Khalil et al., 2023; Mishra & Sadia, 2023).

Εφαρμογές στην ανίχνευση ψευδών αναφορών

Η ταξινόμηση πολλαπλών ετικετών έχει ευρεία εφαρμογή στην ανίχνευση ψευδών αναφορών στο διαδίκτυο, ειδικά σε πλατφόρμες κοινωνικής δικτύωσης και ειδησεογραφικά sites. Στην ανίχνευση ψευδών ειδήσεων, ένα άρθρο ή μια ανάρτηση μπορεί να ταξινομηθεί σε πολλαπλές κατηγορίες, όπως "ψευδές", "παραπλανητικό", "προπαγάνδα", και "πραγματικό", επιτρέποντας την πιο λεπτομερή και ακριβή ανάλυση του περιεχομένου. Η προσέγγιση αυτή επιτρέπει στους ερευνητές και τους επαγγελματίες να αναγνωρίζουν τις διάφορες διαστάσεις της παραπληροφόρησης και να αναπτύσσουν αποτελεσματικότερες στρατηγικές για την καταπολέμησή της (Mishra & Sadia, 2023; Ahmed et al., 2021).

Για παράδειγμα, κατά την ανίχνευση ψευδών ειδήσεων σχετικά με την πανδημία του COVID-19, ένα μοντέλο πολλαπλών ετικετών μπορεί να επισημάνει ένα άρθρο ως "ψευδές" και "παραπλανητικό", βοηθώντας έτσι να εντοπιστεί όχι μόνο η αναλήθεια του περιεχομένου αλλά και η πρόθεσή του να παραπλανήσει το κοινό. Αυτό προσφέρει μια πιο ολοκληρωμένη προσέγγιση στην αντιμετώπιση της παραπληροφόρησης (Akhtar et al., 2023).

Πλεονεκτήματα και μειονεκτήματα

Η ταξινόμηση πολλαπλών ετικετών προσφέρει αρκετά πλεονεκτήματα, αλλά συνοδεύεται και από ορισμένα μειονεκτήματα:

Πλεονεκτήματα

1. **Ακρίβεια και Ευελιξία**: Η δυνατότητα ταξινόμησης δεδομένων σε περισσότερες από μία κατηγορίες επιτρέπει στα μοντέλα να κατανοούν και να αποτυπώνουν τις πολυδιάστατες πληροφορίες των δεδομένων με μεγαλύτερη ακρίβεια και λεπτομέρεια.

2. **Καλύτερη Απόδοση σε Πολύπλοκα Δεδομένα**: Σε περιπτώσεις όπου τα δεδομένα είναι πολυσύνθετα και αλληλένδετα, η ταξινόμηση πολλαπλών ετικετών επιτρέπει την καλύτερη διαχείριση και ανάλυση αυτών των δεδομένων.
3. **Εφαρμογή σε Διάφορους Τομείς**: Η τεχνική αυτή είναι ευέλικτη και μπορεί να εφαρμοστεί σε μια πληθώρα τομέων, από την ανίχνευση ψευδών ειδήσεων μέχρι την κατηγοριοποίηση μουσικών ειδών και την ανάλυση συναισθημάτων στα μέσα κοινωνικής δικτύωσης (Ahmed et al., 2021).

Μειονεκτήματα

1. **Πολυπλοκότητα στην Εκπαίδευση**: Η εκπαίδευση μοντέλων πολλαπλών ετικετών μπορεί να είναι πολύπλοκη και απαιτεί μεγάλα σύνολα δεδομένων καθώς και περισσότερη υπολογιστική ισχύ.
2. **Αντιμετώπιση Συνεχών Αλλαγών**: Η συνεχής ενημέρωση και ανανέωση των δεδομένων και των ετικετών μπορεί να είναι δύσκολη και χρονοβόρα διαδικασία.
3. **Προκλήσεις στην Ερμηνεία των Αποτελεσμάτων**: Η ερμηνεία των αποτελεσμάτων ενός μοντέλου πολλαπλών ετικετών μπορεί να είναι πιο σύνθετη, καθώς απαιτεί την ανάλυση και κατανόηση πολλών διαφορετικών κατηγοριών ταυτόχρονα (Mishra & Sadia, 2023; Khalil et al., 2023).

Παραδείγματα αλγορίθμων

Για την υλοποίηση της ταξινόμησης πολλαπλών ετικετών, χρησιμοποιούνται διάφοροι αλγόριθμοι και τεχνικές. Ορισμένοι από τους πιο κοινούς αλγόριθμους περιλαμβάνουν:

1. **Binary Relevance**: Αυτή η μέθοδος διασπά το πρόβλημα πολλαπλών ετικετών σε πολλαπλά προβλήματα δυαδικής ταξινόμησης, όπου κάθε κατηγορία αντιμετωπίζεται ανεξάρτητα. Παρά το ότι είναι απλή και εύκολη στην εφαρμογή, δεν λαμβάνει υπόψη την αλληλεπίδραση μεταξύ των κατηγοριών (Khalil et al., 2023).
2. **Classifier Chains**: Αυτή η προσέγγιση δημιουργεί μια αλυσίδα ταξινομητών όπου η έξοδος του κάθε ταξινομητή χρησιμοποιείται ως πρόσθετο χαρακτηριστικό για τον επόμενο ταξινομητή στην αλυσίδα. Αυτό επιτρέπει την ανάλυση των αλληλεπιδράσεων μεταξύ των κατηγοριών, βελτιώνοντας την ακρίβεια της πρόβλεψης (Ahmed et al., 2021).

3. **Label Powerset**: Αυτή η τεχνική μετατρέπει το πρόβλημα πολλαπλών ετικετών σε ένα πρόβλημα ταξινόμησης πολλαπλών κατηγοριών, δημιουργώντας νέες κατηγορίες για κάθε μοναδικό συνδυασμό ετικετών που εμφανίζεται στο σύνολο δεδομένων. Παρόλο που μπορεί να είναι πολύ αποδοτική σε μικρά σύνολα δεδομένων, η απόδοσή της μπορεί να μειωθεί δραματικά όταν οι συνδυασμοί ετικετών αυξάνονται (Khalil et al., 2023; Mishra & Sadia, 2023).
4. **Deep Learning Approaches**: Τεχνικές βαθιάς μάθησης, όπως οι νευρωνικά δίκτυα CNNs και RNNs, χρησιμοποιούνται ευρέως για την ταξινόμηση πολλαπλών ετικετών. Αυτά τα μοντέλα μπορούν να μάθουν πολύπλοκες σχέσεις και αλληλεπιδράσεις μεταξύ των ετικετών, προσφέροντας υψηλή ακρίβεια στην πρόβλεψη. Ωστόσο, απαιτούν μεγάλη υπολογιστική ισχύ και μεγάλα σύνολα δεδομένων για την εκπαίδευσή τους (Akhtar et al., 2023; Ahmed et al., 2021).

Η ταξινόμηση πολλαπλών ετικετών αποτελεί ένα ισχυρό εργαλείο στην ανίχνευση ψευδών αναφορών, επιτρέποντας την ακριβέστερη και λεπτομερέστερη κατηγοριοποίηση των δεδομένων. Παρά τις προκλήσεις που συνοδεύουν την υλοποίησή της, τα πλεονεκτήματα που προσφέρει την καθιστούν απαραίτητη για την αποτελεσματική διαχείριση και ανάλυση της παραπληροφόρησης στο διαδίκτυο.

2.1.4 Μη ισορροπημένη ταξινόμηση (Imbalanced Classification)

Η μη ισορροπημένη ταξινόμηση αναφέρεται σε προβλήματα ταξινόμησης όπου οι κατηγορίες των δεδομένων δεν είναι ισορροπημένες, δηλαδή μία ή περισσότερες κατηγορίες περιέχουν πολύ λιγότερα δείγματα σε σχέση με άλλες. Στην περίπτωση της ανίχνευσης ψευδών ειδήσεων, αυτό το πρόβλημα εμφανίζεται συνήθως όταν τα δεδομένα περιέχουν περισσότερα αληθινά άρθρα παρά ψευδή (Akhtar et al., 2023; Mishra & Sadia, 2023).

Προκλήσεις στην ταξινόμηση ψευδών ειδήσεων

1. **Υποεκπροσώπηση Ψευδών Ειδήσεων**: Τα δεδομένα που περιέχουν περισσότερες αληθινές ειδήσεις μπορούν να οδηγήσουν σε μοντέλα που δεν καταφέρνουν να εντοπίσουν ψευδείς ειδήσεις, επειδή δεν εκπαιδεύονται επαρκώς σε δείγματα ψευδών ειδήσεων. Αυτό μειώνει την ευαισθησία (recall) των μοντέλων και τα κάνει

να αποδίδουν καλά μόνο στην πλειονότητα των δεδομένων (Akhtar et al., 2023; Mishra & Sadia, 2023).

2. **Μεροληψία Μοντέλου**: Η ανισορροπία μπορεί να οδηγήσει τα μοντέλα να γίνουν μεροληπτικά υπέρ της πλειονότητας των δεδομένων, παραμελώντας τη μειονότητα. Έτσι, τα μοντέλα μπορεί να αποτυγχάνουν να εντοπίσουν ψευδείς ειδήσεις ή να δίνουν λανθασμένα αποτελέσματα (Akhtar et al., 2023; Khalil et al., 2023).
3. **Δυσκολία στην Αξιολόγηση**: Η ανισορροπία των κατηγοριών κάνει τη χρήση παραδοσιακών μετρικών απόδοσης, όπως η ακρίβεια, μη κατάλληλη για την αξιολόγηση της απόδοσης των μοντέλων. Πρέπει να χρησιμοποιούνται μετρικές όπως η ακρίβεια, η ευαισθησία και το F1-score, που λαμβάνουν υπόψη την ανισορροπία (Shushkevich et al., 2023; Mishra & Sadia, 2023).

Μέθοδοι αντιμετώπισης

1. **Υπερδειγματοληψία (Oversampling)**: Αυτή η τεχνική περιλαμβάνει την παραγωγή συνθετικών δεδομένων για τη μειονότητα των κατηγοριών χρησιμοποιώντας μεθόδους όπως το SMOTE (Synthetic Minority Over-sampling Technique). Η τεχνική αυτή βοηθά στην εξισορρόπηση των κατηγοριών αυξάνοντας τον αριθμό των δειγμάτων της μειονότητας (Akhtar et al., 2023; Khalil et al., 2023).
2. **Υποδειγματοληψία (Undersampling)**: Αυτή η μέθοδος περιλαμβάνει τη μείωση του αριθμού των δειγμάτων της πλειονότητας των κατηγοριών για να επιτευχθεί ισορροπία. Αν και μπορεί να οδηγήσει σε απώλεια πληροφοριών, είναι χρήσιμη όταν υπάρχουν αρκετά δείγματα της πλειονότητας για να διατηρηθεί η ακρίβεια του μοντέλου (Akhtar et al., 2023; Mishra & Sadia, 2023).
3. **Δημιουργία Συνθετικών Δεδομένων (Data Augmentation)**: Χρησιμοποιώντας τεχνικές όπως η γλωσσική αναπαράσταση (language representation) και τα μοντέλα μετασχηματιστών (transformers), όπως το BERT, μπορούν να δημιουργηθούν νέα δείγματα που αυξάνουν την ποικιλομορφία και την ισορροπία των δεδομένων (Shushkevich et al., 2023).
4. **Σταθμισμένη Εκπαίδευση (Cost-Sensitive Training)**: Η προσαρμογή της λειτουργίας απώλειας του μοντέλου για να λαμβάνει υπόψη το κόστος της λάθος ταξινόμησης της μειονότητας, βοηθά τα μοντέλα να δίνουν μεγαλύτερη σημασία στις ψευδείς ειδήσεις κατά την εκπαίδευση (Akhtar et al., 2023).

Παραδείγματα αλγορίθμων

1. **SMOTE (Synthetic Minority Over-sampling Technique)**: Ένας ευρέως χρησιμοποιούμενος αλγόριθμος υπερδειγματοληψίας που δημιουργεί συνθετικά δείγματα για τη μειονότητα βασισμένα στα χαρακτηριστικά των υπαρχόντων δειγμάτων. Το SMOTE έχει αποδειχθεί αποτελεσματικό στην αντιμετώπιση της ανισορροπίας σε πολλά σενάρια ταξινόμησης (Akhtar et al., 2023).
2. **Borderline-SMOTE**: Μια παραλλαγή του SMOTE που δημιουργεί συνθετικά δείγματα μόνο κοντά στα όρια των κατηγοριών, ενισχύοντας έτσι την απόδοση του μοντέλου σε δύσκολες περιπτώσεις ταξινόμησης (Akhtar et al., 2023).
3. **ADASYN (Adaptive Synthetic Sampling)**: Ένας αλγόριθμος που προσαρμόζεται δυναμικά στις ανάγκες των δεδομένων, δημιουργώντας περισσότερα συνθετικά δείγματα για τις περιοχές που χρειάζονται περισσότερο ενίσχυση, βελτιώνοντας έτσι την ισορροπία και την απόδοση του μοντέλου (Akhtar et al., 2023).
4. **Random Oversampling and Undersampling**: Απλές τεχνικές υπερδειγματοληψίας και υποδειγματοληψίας που χρησιμοποιούνται ευρέως για την εξισορρόπηση των δεδομένων. Αν και είναι βασικές τεχνικές, μπορούν να προσφέρουν σημαντικές βελτιώσεις σε πολλά σενάρια ταξινόμησης (Akhtar et al., 2023; Mishra & Sadia, 2023).

Η ανίχνευση ψευδών ειδήσεων με χρήση τεχνικών μηχανικής μάθησης και βαθιάς μάθησης απαιτεί την αντιμετώπιση του προβλήματος της ανισορροπίας δεδομένων για την επίτευξη υψηλής ακρίβειας και αξιοπιστίας. Οι παραπάνω μέθοδοι και αλγόριθμοι παρέχουν μια ισχυρή βάση για τη βελτίωση της απόδοσης των μοντέλων ταξινόμησης σε ανισόρροπα δεδομένα, επιτρέποντας την αποτελεσματική ανίχνευση ψευδών ειδήσεων στο διαδίκτυο (Mishra & Sadia, 2023; Khalil et al., 2023; Shushkevich et al., 2023).

2.2 Μετρικά Συστήματα Επίδοσης Ταξινόμησης

Στον τομέα της μηχανικής μάθησης και της τεχνητής νοημοσύνης, η αξιολόγηση της απόδοσης των μοντέλων ταξινόμησης αποτελεί κρίσιμο βήμα για την επιλογή του καταλληλότερου μοντέλου. Υπάρχουν διάφορα μετρικά συστήματα που χρησιμοποιούνται για την εκτίμηση της απόδοσης αυτών των μοντέλων. Οι κύριες μετρικές περιλαμβάνουν την

ακρίβεια στο σετ ελέγχου (Test Accuracy), την ευσυνειδησία ή ακρίβεια (Precision), την επανακλησιμότητα (Recall), το F1 Score και την εξισορροπημένη ακρίβεια (Balanced Accuracy).

Ακρίβεια στο σετ ελέγχου (Test Accuracy)

Η ακρίβεια είναι μία από τις πιο κοινές μετρικές για την αξιολόγηση μοντέλων ταξινόμησης. Ορίζεται ως ο λόγος των σωστών προβλέψεων προς το συνολικό αριθμό προβλέψεων. Η ακρίβεια υπολογίζεται ως εξής:

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

όπου:

- TP (True Positives) είναι τα αληθώς θετικά,
- TN (True Negatives) είναι τα αληθώς αρνητικά,
- FP (False Positives) είναι τα ψευδώς θετικά,
- FN (False Negatives) είναι τα ψευδώς αρνητικά.

Η ακρίβεια είναι μια χρήσιμη μετρική όταν οι τάξεις του προβλήματος είναι ισορροπημένες. Ωστόσο, σε περιπτώσεις ανισορροπίας, όπου μία τάξη κυριαρχεί, η ακρίβεια μπορεί να δώσει παραπλανητικά αποτελέσματα, καθώς μπορεί να αντικατοπτρίζει την υπερίσχυση της επικρατούσας τάξης (Akhtar et al., 2023).

Ευσυνειδησία – Ακρίβεια (Precision)

Η ευσυνειδησία, γνωστή και ως θετική προγνωστική αξία, μετρά το ποσοστό των πραγματικών θετικών παραδειγμάτων μεταξύ των παραδειγμάτων που έχουν ταξινομηθεί ως θετικά από το μοντέλο. Υπολογίζεται ως εξής:

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

Η ευσυνειδησία είναι ιδιαίτερα χρήσιμη σε καταστάσεις όπου το κόστος των ψευδώς θετικών είναι υψηλό. Για παράδειγμα, στην ιατρική διάγνωση, ένα ψευδώς θετικό αποτέλεσμα μπορεί να οδηγήσει σε περιττές θεραπείες, γι' αυτό η υψηλή ευσυνειδησία είναι σημαντική (Ahmed et al., 2021).

Ανάκληση (Recall)

Η ανάκληση, ή ευαισθησία, μετρά την ικανότητα του μοντέλου να εντοπίζει όλα τα θετικά παραδείγματα στην πραγματικότητα. Υπολογίζεται ως εξής:

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

Η ανάκληση είναι σημαντική σε περιπτώσεις όπου το κόστος των ψευδώς αρνητικών είναι υψηλό. Για παράδειγμα, στην ανίχνευση απάτης, ένα ψευδώς αρνητικό αποτέλεσμα σημαίνει ότι μια πραγματική απάτη δεν εντοπίστηκε, κάτι που μπορεί να έχει σοβαρές συνέπειες (Shushkevich et al., 2023).

F1-Score

Το F1-Score είναι ο αρμονικός μέσος όρος της ευσυνειδησίας (precision) και της ανάκλησης (recall). Δίνει μια ισορροπημένη αποτίμηση της απόδοσης του μοντέλου και είναι χρήσιμο όταν υπάρχει ανισορροπία στις τάξεις του προβλήματος. Υπολογίζεται ως εξής:

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Το F1-Score είναι ιδιαίτερα χρήσιμο όταν χρειάζεται να βρεθεί μια ισορροπία μεταξύ της ευσυνειδησίας και της ανάκλησης, και όταν η ακρίβεια στο σετ ελέγχου δεν παρέχει αρκετά πληροφοριακή εικόνα για την απόδοση του μοντέλου (Akhtar et al., 2023; Khalil et al., 2023).

Ισορροπημένη Ακρίβεια (Balanced Accuracy)

Το **balanced accuracy** είναι μια μετρική που χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης με μη ισορροπημένα σύνολα δεδομένων. Είναι ιδιαίτερα χρήσιμη στη βαθιά μάθηση και σε άλλες τεχνικές μηχανικής μάθησης για την αξιολόγηση της απόδοσης ενός μοντέλου.

Το balanced accuracy ουσιαστικά είναι ο μέσος όρος της ευαισθησίας (sensitivity) και της ειδικότητας (specificity) ενός μοντέλου. Αυτό σημαίνει ότι λαμβάνει υπόψη τόσο τα αληθώς θετικά όσο και τα αληθώς αρνητικά αποτελέσματα, προσφέροντας μια πιο ισορροπημένη εικόνα της απόδοσης του μοντέλου.

Υπολογισμός

Ο τύπος για τον υπολογισμό του balanced accuracy είναι:

$$Balanced\ Accuracy = \frac{Sencitivity + Specificity}{2}$$

- **Sensitivity (ή Recall)**: Το ποσοστό των αληθώς θετικών αποτελεσμάτων που ανιχνεύθηκαν σωστά από το μοντέλο.
- **Specificity**: Το ποσοστό των αληθώς αρνητικών αποτελεσμάτων που ανιχνεύθηκαν σωστά από το μοντέλο.

Σε προβλήματα με μη ισορροπημένα δεδομένα, η παραδοσιακή ακρίβεια (accuracy) μπορεί να είναι παραπλανητική. Για παράδειγμα, σε ένα σύνολο δεδομένων όπου το 90% των δειγμάτων ανήκουν στην κατηγορία Α και το 10% στην κατηγορία Β, ένα μοντέλο που προβλέπει πάντα την κατηγορία Α θα έχει ακρίβεια 90%, αλλά δεν θα είναι χρήσιμο για την ανίχνευση της κατηγορίας Β.

2.3 Συστήματα Βασισμένα σε Κανόνες

Τα συστήματα βασισμένα σε κανόνες (rule-based systems) είναι μια από τις πρώτες μορφές τεχνητής νοημοσύνης και χρησιμοποιούνται για την αυτοματοποίηση της λήψης αποφάσεων σε ένα συγκεκριμένο πεδίο. Αυτά τα συστήματα λειτουργούν χρησιμοποιώντας μια συλλογή κανόνων, που συνήθως αποτελούνται από δηλώσεις "αν-τότε" (if-then statements). Οι κανόνες αυτοί κωδικοποιούν τη γνώση ειδικών και περιγράφουν τις ενέργειες που πρέπει να ληφθούν υπό συγκεκριμένες συνθήκες (Khalil et al., 2023).

Η αρχιτεκτονική των συστημάτων βασισμένων σε κανόνες συνήθως περιλαμβάνει τρία βασικά συστατικά: τη βάση γνώσεων, τον μηχανισμό συμπερασμού και το περιβάλλον χρήστη. Η βάση γνώσεων περιέχει τους κανόνες και τα δεδομένα που σχετίζονται με το

πρόβλημα. Ο μηχανισμός συμπερασμού είναι το μέρος του συστήματος που εφαρμόζει τους κανόνες στη βάση δεδομένων για να παράγει νέα γνώση ή να λάβει αποφάσεις. Το περιβάλλον χρήστη επιτρέπει στους χρήστες να εισάγουν δεδομένα και να λαμβάνουν απαντήσεις από το σύστημα (Shushkevich et al., 2023).

Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα

1. **Διαφάνεια και Εξηγησιμότητα**: Τα συστήματα βασισμένα σε κανόνες προσφέρουν διαφάνεια στις αποφάσεις που λαμβάνουν, καθώς κάθε απόφαση μπορεί να ανιχνευθεί πίσω στους κανόνες που την προκάλεσαν. Αυτό τα καθιστά εξαιρετικά εξηγήσιμα, διευκολύνοντας την κατανόηση του τρόπου με τον οποίο λειτουργούν από μη τεχνικούς χρήστες (Ahmed et al., 2021).
2. **Ευκολία Εφαρμογής και Τροποποίησης**: Οι κανόνες μπορούν εύκολα να προστεθούν, να αφαιρεθούν ή να τροποποιηθούν, καθιστώντας τα συστήματα αυτά ευέλικτα και προσαρμόσιμα στις μεταβαλλόμενες ανάγκες και απαιτήσεις. Αυτό τα καθιστά ιδανικά για πεδία όπου οι απαιτήσεις αλλάζουν συχνά (Khalil et al., 2023).
3. **Χαμηλό Κόστος Ανάπτυξης**: Η ανάπτυξη συστημάτων βασισμένων σε κανόνες μπορεί να είναι σχετικά φθηνή, ειδικά όταν οι κανόνες μπορούν να αντληθούν άμεσα από ειδικούς στον τομέα χωρίς την ανάγκη για σύνθετα δεδομένα εκπαίδευσης .

Μειονεκτήματα

1. **Περιορισμένη Κλίμακα και Ευελιξία**: Τα συστήματα βασισμένα σε κανόνες αντιμετωπίζουν δυσκολίες όταν το σύνολο των κανόνων γίνεται πολύ μεγάλο ή πολύπλοκο. Επιπλέον, δεν μπορούν να μάθουν ή να προσαρμοστούν σε νέα δεδομένα ή καταστάσεις χωρίς ανθρώπινη παρέμβαση (Khalil et al., 2023).
2. **Αναποτελεσματικότητα σε Περίπλοκα Προβλήματα**: Αυτά τα συστήματα είναι συχνά ανεπαρκή για την επίλυση περίπλοκων προβλημάτων που απαιτούν υψηλό βαθμό κατανόησης ή προσαρμογής. Για παράδειγμα, μπορεί να αποτύχουν σε καταστάσεις όπου οι κανόνες δεν καλύπτουν όλες τις πιθανές καταστάσεις .
3. **Δυσκολίες στη Συντήρηση**: Καθώς προστίθενται νέοι κανόνες και οι υπάρχοντες τροποποιούνται, μπορεί να προκύψουν συγκρούσεις ή ασυνέπειες μεταξύ των

κανόνων. Η συντήρηση της συνέπειας και της συνοχής των κανόνων μπορεί να γίνει πολύπλοκη και χρονοβόρα .

Χρήσεις στον Εντοπισμό Ψευδών Ειδήσεων

Τα συστήματα βασισμένα σε κανόνες μπορούν να εφαρμοστούν στον εντοπισμό ψευδών ειδήσεων (fake news) μέσω της χρήσης προκαθορισμένων κανόνων για τον έλεγχο της αξιοπιστίας των πληροφοριών. Οι κανόνες αυτοί μπορεί να περιλαμβάνουν διάφορα κριτήρια, όπως η πηγή των πληροφοριών, η γλώσσα που χρησιμοποιείται, και η δομή του κειμένου.

1. **Έλεγχος Πηγών**: Ένας από τους βασικούς κανόνες μπορεί να είναι ο έλεγχος της αξιοπιστίας της πηγής των ειδήσεων. Για παράδειγμα, ειδήσεις που προέρχονται από πηγές που έχουν ιστορικό διάδοσης ψευδών πληροφοριών μπορεί να χαρακτηριστούν ως ψευδείς (Ahmed et al., 2021).
2. **Ανάλυση Γλώσσας**: Οι κανόνες μπορούν να εφαρμοστούν για την ανάλυση της γλώσσας που χρησιμοποιείται στο κείμενο. Χαρακτηριστικά όπως η υπερβολή, η συναισθηματική φόρτιση και η χρήση ανώνυμων πηγών μπορεί να είναι ενδείξεις ψευδών ειδήσεων. Οι κανόνες αυτοί μπορούν να βοηθήσουν στον εντοπισμό ειδήσεων που χρησιμοποιούν παραπλανητική γλώσσα (Khalil et al., 2023).
3. **Δομή και Περιεχόμενο Κειμένου**: Οι κανόνες μπορούν επίσης να ελέγχουν τη δομή και το περιεχόμενο του κειμένου. Για παράδειγμα, ειδήσεις που περιέχουν πολλές ασυνέπειες ή αντιφάσεις μπορεί να θεωρηθούν ψευδείς. Επιπλέον, η παρουσία πολλών ορθογραφικών ή γραμματικών λαθών μπορεί να είναι ένδειξη χαμηλής ποιότητας και αξιοπιστίας του κειμένου .
4. **Συγκεκριμένα Μοτίβα και Κοινές Φράσεις**: Τα συστήματα βασισμένα σε κανόνες μπορούν να αναγνωρίζουν συγκεκριμένα μοτίβα ή φράσεις που είναι συχνές σε ψευδείς ειδήσεις. Για παράδειγμα, φράσεις όπως "αυτό πρέπει να το δείτε!" ή "είδηση σοκ" μπορεί να χρησιμοποιούνται συχνά σε clickbait τίτλους και να αποτελούν ένδειξη ψευδών ειδήσεων .
5. **Συγκριτική Ανάλυση**: Ένα άλλο σύνολο κανόνων μπορεί να περιλαμβάνει τη σύγκριση των πληροφοριών με άλλες αξιόπιστες πηγές. Αν το περιεχόμενο μιας είδησης δεν συμβαδίζει με άλλες αξιόπιστες πηγές, μπορεί να θεωρηθεί ως ψευδές. Αυτός ο

κανόνας βασίζεται στη διασταύρωση των πληροφοριών για την επιβεβαίωση της ακρίβειάς τους .

Παρά τα προαναφερθέντα πλεονεκτήματα, τα συστήματα βασισμένα σε κανόνες δεν είναι πανάκεια. Ενώ μπορούν να είναι εξαιρετικά χρήσιμα σε συγκεκριμένες εφαρμογές, έχουν περιορισμούς όσον αφορά την ευελιξία και την ικανότητα προσαρμογής τους σε νέα δεδομένα ή καταστάσεις. Για τον εντοπισμό ψευδών ειδήσεων, οι κανόνες πρέπει να ενημερώνονται συνεχώς για να ανταποκρίνονται στις νέες τεχνικές και μοτίβα που χρησιμοποιούνται για τη διάδοση παραπληροφόρησης.

Τα συστήματα βασισμένα σε κανόνες είναι ένα ισχυρό εργαλείο για την αυτοματοποίηση της λήψης αποφάσεων και τον εντοπισμό ψευδών ειδήσεων. Ωστόσο, η αποτελεσματικότητά τους εξαρτάται από την ποιότητα και την επικαιρότητα των κανόνων που χρησιμοποιούνται. Η συνεχής ενημέρωση και βελτίωση των κανόνων είναι απαραίτητη για να διατηρηθεί η αποτελεσματικότητα αυτών των συστημάτων.

2.4 Συστήματα Βασισμένα σε Μηχανική Μάθηση

Η μηχανική μάθηση (ML) είναι ένας κλάδος της τεχνητής νοημοσύνης (AI) που εστιάζει στην ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις ή αποφάσεις χωρίς να έχουν προγραμματιστεί ρητά για κάθε εργασία. Η διαδικασία αυτή περιλαμβάνει τη χρήση στατιστικών τεχνικών για την ανάλυση και τη μοντελοποίηση δεδομένων, επιτρέποντας στα συστήματα να αναγνωρίζουν μοτίβα και να προσαρμόζονται σε νέα δεδομένα.

Η μηχανική μάθηση έχει καταστεί ουσιώδης στον τομέα της ανίχνευσης ψευδών ειδήσεων, καθώς τα παραδοσιακά συστήματα κανόνων δεν μπορούν να ανταπεξέλθουν στην πολυπλοκότητα και την ποικιλία των ψευδών αναφορών που κυκλοφορούν στο διαδίκτυο. Η ικανότητα των συστημάτων ML να επεξεργάζονται μεγάλους όγκους δεδομένων και να αναγνωρίζουν σύνθετα μοτίβα τα καθιστά ιδανικά για την αντιμετώπιση αυτού του προβλήματος.

Αλγόριθμοι που Χρησιμοποιούνται στην Ανίχνευση Ψευδών Ειδήσεων

Διάφοροι αλγόριθμοι μηχανικής μάθησης έχουν αναπτυχθεί και χρησιμοποιούνται για την ανίχνευση ψευδών ειδήσεων. Οι πιο δημοφιλείς από αυτούς περιλαμβάνουν τους αλγορίθμους επιβλεπόμενης μάθησης, όπως τα Υποστηρικτικά Διανύσματα Μηχανών (SVM), τα Νευρωνικά Δίκτυα, τα Δέντρα Απόφασης, καθώς και τους αλγορίθμους μη επιβλεπόμενης μάθησης όπως οι K-Μέσοι. Ακολουθεί μια αναλυτική περιγραφή αυτών των αλγορίθμων:

1. **Υποστηρικτικά Διανύσματα Μηχανών (SVM)**

Τα SVM είναι ένας από τους πιο ευρέως χρησιμοποιούμενους αλγορίθμους για την ανίχνευση ψευδών ειδήσεων. Βασίζονται στη θεωρία των διανυσματικών χώρων και στοχεύουν να βρουν το υπερ-επίπεδο που διαχωρίζει τα δεδομένα σε διαφορετικές κατηγορίες με τον καλύτερο δυνατό τρόπο. Στην περίπτωση της ανίχνευσης ψευδών ειδήσεων, τα SVM χρησιμοποιούνται για να διαχωρίσουν τα άρθρα σε δύο κατηγορίες: αληθινά και ψευδή (Khalil et al., 2023).

2. **Νευρωνικά Δίκτυα**

Τα νευρωνικά δίκτυα, και ειδικότερα τα βαθιά νευρωνικά δίκτυα (DNN), έχουν αποδειχθεί ιδιαίτερα αποτελεσματικά στην ανίχνευση ψευδών ειδήσεων. Αυτά τα μοντέλα αποτελούνται από πολλαπλά στρώματα νευρώνων που συνεργάζονται για να αναγνωρίζουν μοτίβα και να μαθαίνουν από τα δεδομένα. Τα βαθιά νευρωνικά δίκτυα είναι ιδιαίτερα κατάλληλα για την επεξεργασία φυσικής γλώσσας (NLP), επιτρέποντας στα συστήματα να αναλύουν το περιεχόμενο των άρθρων και να ανιχνεύουν ψευδείς πληροφορίες (Khalil et al ., 2023).

3. **Δέντρα Απόφασης**

Τα δέντρα απόφασης είναι ένας άλλος δημοφιλής αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την ανίχνευση ψευδών ειδήσεων. Αυτά τα μοντέλα λειτουργούν με τη δημιουργία ενός δέντρου απόφασης όπου κάθε κόμβος αντιπροσωπεύει μια ερώτηση ή μια απόφαση, και κάθε κλάδος αντιπροσωπεύει την απάντηση σε αυτή την ερώτηση. Τα δέντρα

απόφασης μπορούν να εντοπίσουν ψευδείς ειδήσεις με βάση χαρακτηριστικά όπως η δομή του κειμένου, η συχνότητα των λέξεων και άλλα σημεία δεδομένων (Ahmed et al., 2021).

4. **K-Μέσοι (K-Means)**

Οι αλγόριθμοι K-Μέσων χρησιμοποιούνται για την ομαδοποίηση δεδομένων σε κλάσεις (clusters) με βάση την ομοιότητα των χαρακτηριστικών τους. Στην ανίχνευση ψευδών ειδήσεων, οι K-Μέσοι μπορούν να χρησιμοποιηθούν για την ομαδοποίηση άρθρων που έχουν παρόμοιες χαρακτηριστικές ιδιότητες και την αναγνώριση των κλάσεων που περιέχουν ψευδείς ειδήσεις (Shushkevich et al., 2023).

5. **Γραφικά Μοντέλα (Graphical Models)**

Τα γραφικά μοντέλα, όπως τα Bayesian Networks και τα Markov Random Fields, χρησιμοποιούνται επίσης στην ανίχνευση ψευδών ειδήσεων. Αυτά τα μοντέλα κατασκευάζουν γραφήματα που αντιπροσωπεύουν την πιθανότητα σχέσεων μεταξύ διαφόρων δεδομένων. Με βάση αυτές τις πιθανότητες, τα γραφικά μοντέλα μπορούν να ανιχνεύσουν ψευδείς ειδήσεις μέσω της ανάλυσης των σχέσεων και των συνδέσεων μεταξύ των δεδομένων (Ahmed et al., 2021).

6. **Αλγόριθμοι Βαθιάς Μάθησης**

Οι αλγόριθμοι βαθιάς μάθησης, όπως τα Συγκλίνοντα Νευρωνικά Δίκτυα (CNN) και τα Αναδρομικά Νευρωνικά Δίκτυα (RNN), έχουν χρησιμοποιηθεί ευρέως για την ανίχνευση ψευδών ειδήσεων. Τα CNN είναι ιδιαίτερα αποτελεσματικά στην ανάλυση εικόνων και κειμένων, ενώ τα RNN είναι κατάλληλα για την ανάλυση ακολουθιών δεδομένων, όπως κείμενα άρθρων ειδήσεων (Akhtar et al., 2023; Khalil et al., 2023).

Προκλήσεις και Μέλλον της Ανίχνευσης Ψευδών Ειδήσεων με Μηχανική Μάθηση

Παρόλο που οι αλγόριθμοι μηχανικής μάθησης έχουν αποδειχθεί ιδιαίτερα αποτελεσματικοί στην ανίχνευση ψευδών ειδήσεων, υπάρχουν ακόμη σημαντικές προκλήσεις που πρέπει να αντιμετωπιστούν. Μια από τις κύριες προκλήσεις είναι η διαχείριση της προκατάληψης στα δεδομένα εκπαίδευσης. Η ποιότητα των δεδομένων που χρησιμοποιούνται για την

εκπαίδευση των μοντέλων είναι κρίσιμη, καθώς τα δεδομένα με προκαταλήψεις μπορούν να οδηγήσουν σε ανακριβή και άδικα αποτελέσματα.

Επιπλέον, οι αλγόριθμοι μηχανικής μάθησης απαιτούν μεγάλους όγκους δεδομένων για την εκπαίδευσή τους, κάτι που μπορεί να είναι δύσκολο να επιτευχθεί σε περιπτώσεις που τα δεδομένα είναι περιορισμένα ή δεν είναι διαθέσιμα. Η διαχείριση της ιδιωτικότητας και της ασφάλειας των δεδομένων είναι επίσης μια σημαντική πρόκληση, καθώς η συλλογή και η αποθήκευση μεγάλων όγκων δεδομένων μπορεί να δημιουργήσει κινδύνους για την ιδιωτικότητα των χρηστών (Shushkevich et al., 2023; Khalil et al., 2023).

Η εξέλιξη της τεχνολογίας AI και ML προσφέρει πολλές υποσχέσεις για τη βελτίωση της ανίχνευσης ψευδών ειδήσεων. Οι ερευνητές συνεχίζουν να αναπτύσσουν νέες τεχνικές και αλγορίθμους για την αντιμετώπιση αυτών των προκλήσεων, εστιάζοντας στην αύξηση της ακρίβειας και της αποδοτικότητας των μοντέλων. Η χρήση συνδυαστικών προσεγγίσεων που ενσωματώνουν πολλαπλούς αλγορίθμους και τεχνικές μπορεί να βελτιώσει σημαντικά την ικανότητα των συστημάτων να εντοπίζουν ψευδείς ειδήσεις και να περιορίζουν την εξάπλωσή τους (Ahmed et al., 2021; Khalil et al., 2023).

2.5 Συστήματα Βασισμένα σε Βαθιά Μάθηση

Η βαθιά μάθηση (deep learning) αποτελεί έναν υποκλάδο της μηχανικής μάθησης που εστιάζει στη χρήση νευρωνικών δικτύων για την ανάλυση και επεξεργασία δεδομένων. Αυτά τα δίκτυα εμπνέονται από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου και αποτελούνται από πολλαπλά στρώματα (layers) νευρώνων, που επιτρέπουν την εξαγωγή πολύπλοκων χαρακτηριστικών από τα δεδομένα (Ahmed et al., 2021). Η ανάπτυξη της βαθιάς μάθησης έχει επιτρέψει τη σημαντική βελτίωση της απόδοσης των συστημάτων ανίχνευσης ψευδών ειδήσεων, καθώς μπορεί να αναγνωρίζει μοτίβα και σχέσεις σε μεγάλους όγκους δεδομένων με ακρίβεια και αποτελεσματικότητα.

Αλγόριθμοι Βαθιάς Μάθησης για Ανίχνευση Ψευδών Ειδήσεων

Η εφαρμογή της βαθιάς μάθησης στην ανίχνευση ψευδών ειδήσεων περιλαμβάνει τη χρήση διαφόρων αλγορίθμων και τεχνικών, όπως τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs), τα

Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs), και τα μοντέλα βασισμένα σε μετασχηματιστές (Transformers).

Συνελικτικά Νευρωνικά Δίκτυα (CNNs)

Τα CNNs είναι ιδιαίτερα αποτελεσματικά στην ανάλυση εικόνας και κειμένου, λόγω της ικανότητάς τους να εξάγουν τοπικά χαρακτηριστικά από τα δεδομένα μέσω των συνελικτικών στρωμάτων τους. Στην ανίχνευση ψευδών ειδήσεων, τα CNNs χρησιμοποιούνται για την ανάλυση των κειμενικών δεδομένων, αναγνωρίζοντας μοτίβα και ακολουθίες λέξεων που μπορεί να υποδηλώνουν την ύπαρξη ψευδών πληροφοριών (Akhtar et al., 2023). Τα CNNs είναι ικανά να εντοπίζουν συγκεκαλυμμένα σημάδια απάτης σε άρθρα και δημοσιεύσεις, βοηθώντας στην αυτόματη ταξινόμηση των ειδήσεων σε αληθείς και ψευδείς.

Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs)

Τα RNNs είναι σχεδιασμένα για την επεξεργασία ακολουθιακών δεδομένων, καθιστώντας τα ιδανικά για την ανάλυση κειμένων που ακολουθούν χρονολογική σειρά ή έχουν εξάρτηση από προηγούμενα στοιχεία. Μια ιδιαίτερη μορφή των RNNs, τα LSTM (Long Short-Term Memory) δίκτυα, χρησιμοποιούνται συχνά στην ανίχνευση ψευδών ειδήσεων, διότι μπορούν να διατηρούν πληροφορίες για μεγάλες ακολουθίες δεδομένων και να απομνημονεύουν σημαντικές πληροφορίες από το κείμενο (Khalil et al., 2023). Αυτό επιτρέπει στα LSTM να ανιχνεύουν αλλαγές στον τόνο, τη γλώσσα και τα μοτίβα γραφής που μπορεί να υποδεικνύουν ψευδείς ειδήσεις.

Μετασχηματιστές (Transformers)

Οι μετασχηματιστές έχουν φέρει επανάσταση στον τομέα της φυσικής γλώσσας, προσφέροντας μεγαλύτερη ακρίβεια και αποτελεσματικότητα σε σύγκριση με τα παραδοσιακά RNNs και CNNs.

Οι Transformers όπως ο GPT (Generative Pre-trained Transformer) βασίζονται στην αρχιτεκτονική της αυτοπροσοχής (self-attention), που τους επιτρέπει να επεξεργάζονται ταυτόχρονα όλα τα στοιχεία ενός κειμένου, εντοπίζοντας τις σχέσεις και τις αλληλεπιδράσεις μεταξύ των λέξεων σε διάφορα σημεία του κειμένου. Αυτή η δυνατότητα καθιστά τους μετασχηματιστές εξαιρετικά αποδοτικούς στην αναγνώριση των ψευδών ειδήσεων, διότι

μπορούν να κατανοούν το συνολικό πλαίσιο και να ανιχνεύουν τις λεπτομέρειες που μπορεί να υποδηλώνουν την ύπαρξη αναληθών πληροφοριών (Ahmed et al., 2021 ; Khalil et al., 2023).

Μοντέλα Βασισμένα σε Μετασχηματιστές (BERT, RoBERTa)

Τα μοντέλα που βασίζονται στους transformers όπως το BERT (Bidirectional Encoder Representations from Transformers) που αναπτύχθηκε από την Google το 2018 και το RoBERTa (Robustly optimized BERT approach) που αποτελεί βελτιωμένη έκδοση του BERT και αναπτύχθηκε από το Facebook AI, χρησιμοποιούνται ευρέως στην ανίχνευση ψευδών ειδήσεων. Το BERT, για παράδειγμα, εκπαιδεύεται σε τεράστια ποσά δεδομένων για να κατανοήσει το πλαίσιο κάθε λέξης σε ένα κείμενο, επιτρέποντάς του να εντοπίζει πιο σύνθετα μοτίβα ψευδών πληροφοριών (Akhtar et al., 2023; Khalil et al., 2023).

Συνδυαστικά Μοντέλα

Η χρήση συνδυαστικών μοντέλων που αξιοποιούν τις δυνατότητες διαφορετικών τύπων νευρωνικών δικτύων έχει αποδειχθεί ιδιαίτερα αποτελεσματική στην ανίχνευση ψευδών ειδήσεων. Για παράδειγμα, ένας συνδυασμός CNNs και LSTM μπορεί να χρησιμοποιηθεί για να εξάγει τοπικά χαρακτηριστικά από το κείμενο και να διατηρεί μακροχρόνιες εξαρτήσεις, βελτιώνοντας την ακρίβεια της ανίχνευσης (Akhtar et al., 2023; Ahmed et al., 2021). Επιπλέον, η ενσωμάτωση μοντέλων BERT με LightGBM έχει δείξει σημαντική βελτίωση στην απόδοση, αξιοποιώντας τις πλούσιες γλωσσικές αναπαραστάσεις του BERT και την ταχύτητα του LightGBM για την ταξινόμηση των ειδήσεων (Khalil et al., 2023).

Παρόλο που η βαθιά μάθηση έχει επιτύχει σημαντικές προόδους στην ανίχνευση ψευδών ειδήσεων, υπάρχουν ακόμα προκλήσεις που πρέπει να αντιμετωπιστούν. Η ανάγκη για μεγάλους όγκους δεδομένων υψηλής ποιότητας είναι μία από τις μεγαλύτερες προκλήσεις, καθώς τα δεδομένα αυτά μπορεί να είναι δύσκολο να συλλεχθούν και να επισημανθούν. Επιπλέον, τα συστήματα βαθιάς μάθησης είναι συχνά υποκείμενα σε προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσης, γεγονός που μπορεί να επηρεάσει την απόδοση και την αξιοπιστία τους (Shushkevich et al., 2023; Akhtar et al., 2023;).

Για να ξεπεραστούν αυτές οι προκλήσεις, οι ερευνητές εστιάζουν στη δημιουργία πιο ανθεκτικών και λιγότερο προκατειλημμένων μοντέλων. Η ενσωμάτωση τεχνικών όπως η μεταφορά μάθησης (transfer learning) και η ενίσχυση δεδομένων (data augmentation) μπορεί να βοηθήσει στη μείωση της εξάρτησης από μεγάλους όγκους δεδομένων. Επιπλέον, η ανάπτυξη μεθόδων για την αυτόματη ανίχνευση και διόρθωση προκαταλήψεων στα δεδομένα εκπαίδευσης είναι ζωτικής σημασίας για την βελτίωση της αξιοπιστίας των συστημάτων βαθιάς μάθησης (Khalil et al., 2023; Shushkevich et al., 2023).

Συμπερασματικά, η βαθιά μάθηση προσφέρει ισχυρά εργαλεία για την ανίχνευση ψευδών ειδήσεων, χρησιμοποιώντας προχωρημένες τεχνικές νευρωνικών δικτύων που μπορούν να αναγνωρίζουν πολύπλοκα μοτίβα και σχέσεις στα δεδομένα. Η συνεχής έρευνα και ανάπτυξη σε αυτόν τον τομέα αναμένεται να βελτιώσει περαιτέρω την ακρίβεια και την αποδοτικότητα των συστημάτων ανίχνευσης, συμβάλλοντας στη μείωση της διάδοσης ψευδών ειδήσεων και στην προώθηση της αξιοπιστίας της πληροφορίας στο διαδίκτυο.

2.6 Υβριδικά Συστήματα

Τα υβριδικά συστήματα στην τεχνητή νοημοσύνη (AI) αναφέρονται στη συνδυασμένη χρήση διαφόρων τεχνικών και μεθόδων από διαφορετικούς τομείς της AI για την επίλυση σύνθετων προβλημάτων. Αυτά τα συστήματα ενσωματώνουν στοιχεία από πολλαπλές τεχνολογίες, όπως η μηχανική μάθηση, τα νευρωνικά δίκτυα, οι αλγόριθμοι εξελικτικής βελτιστοποίησης και οι κανόνες βασισμένοι στη γνώση, για να δημιουργήσουν λύσεις που δεν μπορούν να επιτευχθούν μόνο με μία τεχνολογία.

Η βασική αρχή λειτουργίας των υβριδικών συστημάτων είναι η συνεργασία μεταξύ των διαφόρων αλγορίθμων για την ενίσχυση της απόδοσης και της ακρίβειας. Για παράδειγμα, ένα υβριδικό σύστημα μπορεί να χρησιμοποιεί έναν αλγόριθμο μηχανικής μάθησης για την εξαγωγή χαρακτηριστικών και έναν άλλο για την τελική ταξινόμηση. Αυτή η συνδυασμένη προσέγγιση επιτρέπει την αξιοποίηση των πλεονεκτημάτων κάθε αλγορίθμου και τη μείωση των αδυναμιών τους (Ahmed et al., 2021).

Συνδυασμός Διαφόρων Αλγορίθμων και Μεθόδων

Ο συνδυασμός διαφόρων αλγορίθμων και μεθόδων στα υβριδικά συστήματα γίνεται με στόχο τη βελτίωση της απόδοσης και την αντιμετώπιση των περιορισμών των μεμονωμένων αλγορίθμων.

Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα

1. **Βελτιωμένη Απόδοση**: Τα υβριδικά συστήματα μπορούν να συνδυάζουν τα ισχυρά σημεία διαφόρων αλγορίθμων, επιτυγχάνοντας καλύτερη απόδοση σε σύγκριση με μεμονωμένους αλγόριθμους. Για παράδειγμα, οι συνδυασμοί νευρωνικών δικτύων και εξελικτικών αλγορίθμων μπορούν να βελτιώσουν την ακρίβεια και την αποδοτικότητα (Khalil et al., 2023).
2. **Αντιμετώπιση Πολύπλοκων Προβλημάτων**: Με την ενσωμάτωση πολλαπλών τεχνικών, τα υβριδικά συστήματα μπορούν να χειριστούν πιο περίπλοκα προβλήματα που δεν θα ήταν δυνατά να επιλυθούν με μία μόνο μέθοδο. Αυτό τα καθιστά ιδιαίτερα χρήσιμα σε εφαρμογές όπως η ανίχνευση ψευδών ειδήσεων, όπου η πολυπλοκότητα και η ποικιλία των δεδομένων είναι μεγάλες (Ahmed et al., 2021).
3. **Αυξημένη Ευελιξία**: Η χρήση διαφορετικών αλγορίθμων επιτρέπει στα υβριδικά συστήματα να είναι πιο ευέλικτα και προσαρμόσιμα σε διαφορετικά είδη δεδομένων και προβλημάτων. Αυτή η ευελιξία είναι κρίσιμη για την αντιμετώπιση αλλαγών στο περιβάλλον των δεδομένων ή στις απαιτήσεις της εφαρμογής (Khalil et al., 2023).

Μειονεκτήματα

1. **Υψηλότερο Κόστος Υπολογιστικών Πόρων**: Τα υβριδικά συστήματα συχνά απαιτούν περισσότερους υπολογιστικούς πόρους λόγω της ανάγκης για την εκτέλεση πολλαπλών αλγορίθμων ταυτόχρονα. Αυτό μπορεί να αυξήσει το κόστος και την πολυπλοκότητα της υλοποίησης και της συντήρησης των συστημάτων (Ahmed et al., 2021).
2. **Πολυπλοκότητα στην Ανάπτυξη**: Η ανάπτυξη υβριδικών συστημάτων μπορεί να είναι πιο περίπλοκη και απαιτητική σε σύγκριση με τη χρήση μεμονωμένων αλγορίθμων.

Απαιτεί εξειδικευμένες γνώσεις και εμπειρία στον σχεδιασμό και την ενσωμάτωση διαφόρων τεχνολογιών (Ahmed et al., 2021).

3. **Δυσκολίες στην Επεξήγηση**: Ο συνδυασμός πολλαπλών αλγορίθμων μπορεί να κάνει τα υβριδικά συστήματα πιο δύσκολα στην επεξήγηση και την ερμηνεία. Αυτό είναι ιδιαίτερα σημαντικό σε περιπτώσεις όπου η διαφάνεια και η εξηγήσιμη AI είναι κρίσιμες, όπως στην ιατρική διάγνωση ή τη δικαιοσύνη (Khalil et al., 2023).

Παραδείγματα

Μερικά παραδείγματα συνδυασμών περιλαμβάνουν:

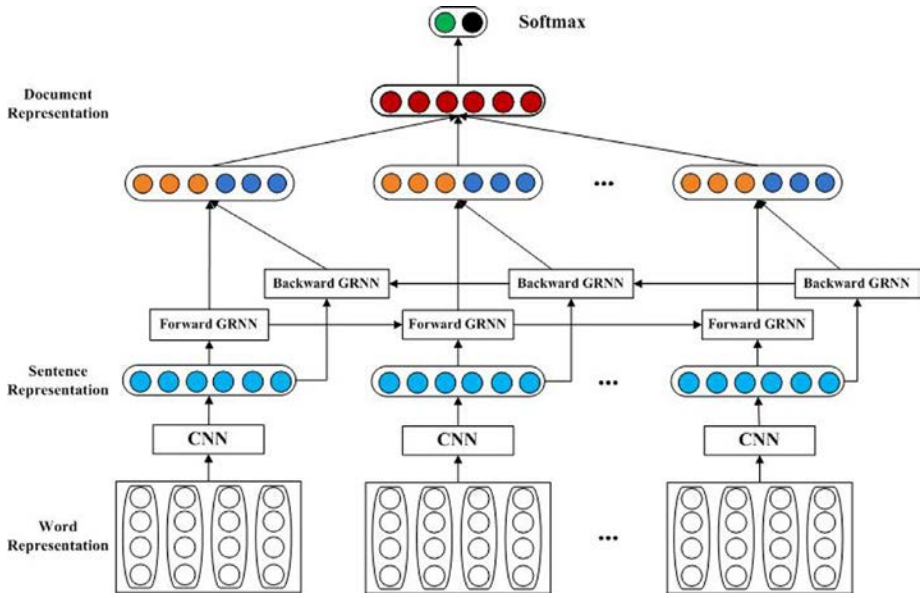
1. **Νευρωνικά Δίκτυα και Κανόνες Βασισμένοι στη Γνώση**: Ένας συνδυασμός νευρωνικών δικτύων για την εκμάθηση μοτίβων και κανόνων βασισμένων στη γνώση για τη λήψη αποφάσεων μπορεί να προσφέρει βελτιωμένη ακρίβεια και επεξηγηματικότητα (Akhtar et al., 2023).
2. **Μηχανική Μάθηση και Βαθιά Μάθηση**: Χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης για την προκαταρκτική επεξεργασία δεδομένων και νευρωνικά δίκτυα βαθιάς μάθησης για την τελική ανάλυση, μπορεί να επιτευχθεί υψηλότερη ακρίβεια και ταχύτητα (Ahmed et al., 2021).
3. **Εξελικτική Βελτιστοποίηση και Νευρωνικά Δίκτυα**: Εξελικτικοί αλγόριθμοι μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση των παραμέτρων ενός νευρωνικού δικτύου, βελτιώνοντας την απόδοσή του και μειώνοντας τον κίνδυνο υπερπροσαρμογής (Ahmed et al., 2021).

Κεφάλαιο 3

Literature Review

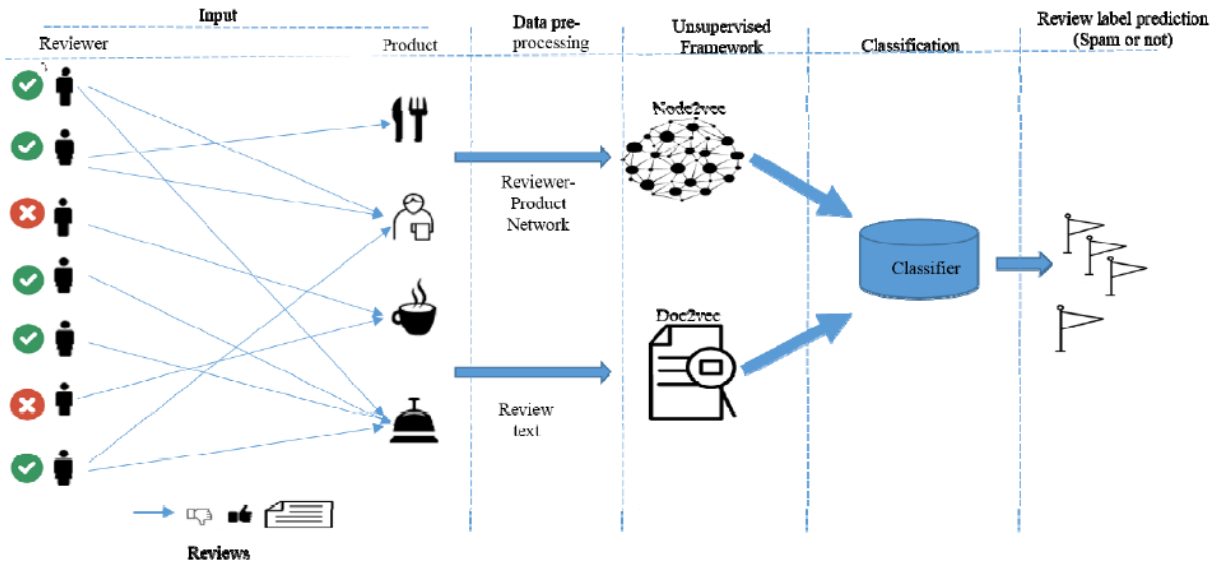
Ο **Fontanarava et al.** (2017) διεξήγαγε μια σημαντική μελέτη σχετικά με την ανίχνευση deceptive reviews, συνδυάζοντας μια ποικιλία χαρακτηριστικών που σχετίζονται τόσο με τους τους κριτικές όσο και με τους χρήστες που δημοσιεύσαν το review. Η εκμετάλλευση τόσο των χαρακτηριστικών που αφορούν την ίδια την κριτική (review-centric) όσο και των χαρακτηριστικών που αφορούν τον χρήστη (reviewer-centric) βοήθησε στο να ανιχνευτούν περίπλοκα μοτίβα τόσο στο κείμενο του review όσο και στην περαιτέρω συμπεριφορά των χρηστών. Η πρώτη κατηγορία χαρακτηριστικών, περιλαμβάνει στοιχεία κειμένου τους n-grams και εκτιμήσεις συναισθήματος, μετα-δεδομένα τους αξιολογήσεις κριτικών, αποκλίσεις αξιολογήσεων και χρονικά μοτίβα τους εκρήξεις κριτικών λόγω ξαφνικής δημοτικότητας ή επιθέσεις spam. Η δεύτερη κατηγορία, εξετάζει τους συμπεριφορές των reviewers, αναλύοντας πτυχές τους η ομοιότητα περιεχομένου μεταξύ των κριτικών που έχουν δημοσιεύσει, τα μοτίβα που ακολουθούν τους βαθμολογήσεις και τους αξιολογήσεις τους και τη συχνότητα με την οποία δραστηριοποιούνται στα πλαίσια του χρόνου, τα οποία παρέχουν βαθύτερες πληροφορίες για πιθανές δραστηριότητες spam. Οι συγγραφείς χρησιμοποίησαν έναν ταξινομητή Random Forests (RFs) για να αξιολογήσουν την αποτελεσματικότητα των προτεινόμενων χαρακτηριστικών τους για την ανίχνευση deceptive κριτικών. Χρησιμοποίησαν το dataset YelpZip, που επιλέχθηκε για την ολοκληρωμένη κάλυψη που προσφέρει και τη μεγάλη κλίμακα του, για να επικυρώσουν την προσέγγισή τους σε διάφορους τύπους επιχειρήσεων και κριτικών. Αναλύοντας συστηματικά τους επιδράσεις των επιμέρους χαρακτηριστικών μέσω συναρτήσεων αθροιστικής κατανομής και δίνοντας έμφαση στα χαρακτηριστικά έκρηξης για την ανίχνευση μεμονωμένων κριτικών, η μέθοδος αυτή επιτυγχάνει σημαντικές βελτιώσεις στην ακρίβεια, την ανάκληση, το f-score και την ακρίβεια σε σύγκριση με τους μεθοδολογίες που είχαν προταθεί μέχρι τότε στην υπάρχουσα βιβλιογραφία, επικυρώνοντας τη σημασία των διαφορετικών συνόλων χαρακτηριστικών για την ενίσχυση τους απόδοσης ταξινόμησης.

Οι **Ren and Ji et al.** (2017) διερευνήσαν την ανίχνευση παραπλανητικών κριτικών σε ένα multi-domain περιβάλλον με τη χρήση προηγμένων νευρωνικών μοντέλων. Αξιοποιώντας το σύνολο δεδομένων που προτάθηκε από τους Li et al. (2014), το οποίο περιλαμβάνει αληθείς και παραπλανητικές κριτικές που προέρχονται από πελάτες, Turkers και υπαλλήλους, η παρούσα μελέτη χρησιμοποιεί διάφορες αρχιτεκτονικές νευρωνικών δικτύων, τους CNNs, RNNs και GRNNs. Τα αποτελέσματα δείχνουν ότι το μοντέλο Bi-directional Average Gated Recurrent Neural Network (GRNN) επιτυγχάνει την υψηλότερη ακρίβεια σε ένα mixed σύνολο δεδομένων, ξεπερνώντας τα παραδοσιακά μοντέλα τους το SVM. Η έρευνα καταδεικνύει ότι η αποτελεσματικότητα των νευρωνικών μοντέλων στη σύλληψη σημασιολογικών χαρακτηριστικών είναι κρίσιμη για τη διάκριση παραπλανητικών κριτικών σε διάφορους τομείς δραστηριότητας. Επιπλέον, τα πειράματα δείχνουν ότι η ενσωμάτωση τους μηχανισμού προσοχής (attention-based mechanism), που χρησιμοποιείται για την εξέταση τους σημασίας των διαφορετικών διανυσμάτων κατάστασης, ενισχύει ακόμα περισσότερο την απόδοσή των μοντέλων. Τέλος, η ενσωμάτωση διακριτών και νευρωνικών χαρακτηριστικών αναδεικνύει τη δυνατότητα βελτίωσης τους ακρίβειας ανίχνευσης. Αυτή η έρευνα υπογραμμίζει την ευρωστία των νευρωνικών δικτύων στην αυτοματοποιημένη ανίχνευση παραπλανητικών μηνυμάτων spam, δίνοντας έμφαση στην ικανότητά τους να κωδικοποιούν αποτελεσματικά τη σημασιολογία (semantics) σε επίπεδο εγγράφου.



Εικόνα 4: *Διάγραμμα Μοντέλου νευρωνικού δικτύου για την ανίχνευση παραπλανητικών μηνυμάτων spam.* *Πηγή: https://www.sciencedirect.com/science/article/pii/S0020025517300166*

Οι Yilmaz and Durahim et al. (2018) στην ολοκληρωμένη μελέτη τους σχετικά με την ανίχνευση παραπλανητικών κριτικών, εισήγαγαν ένα πλαίσιο που ενσωματώνει πληροφορίες τόσο από τα δεδομένα κειμένου των κριτικών όσο και από δεδομένα που βασίζονται στο διασυνδεδεμένο δίκτυο κριτικών-προϊόντων. Αυτή η προσέγγιση αξιοποιεί πυκνά διανύσματα χαρακτηριστικών χαμηλής διάστασης που μαθαίνονται με μη επιβλεπόμενο τρόπο, μέσω της προσαρμογής αλγορίθμων για τη δημιουργία embeddings εγγράφων και κόμβων, δηλαδή των Doc2vec και Node2vec. Στο προτεινόμενο πλαίσιο semi-supervised μάθησης, η παραγωγή των embeddings για τα έγγραφα και τους κόμβους μπορεί να υλοποιηθεί ανεξάρτητα και παράλληλα. Στο τελικό στάδιο της διαδικασίας, αυτές οι διανυσματικές αναπαραστάσεις συνδυάζονται σε ένα ενιαίο διάνυσμα με τη συνένωση των embeddings από τις κριτικές, τους χρήστες και τα προϊόντα. Έπειτα αυτές οι αναπαραστάσεις τροφοδοτούνται σε αλγόριθμο λογιστικής παλινδρόμησης για τη δημιουργία ενός ταξινομητή με σκοπό την ανίχνευση ανεπιθύμητων κριτικών. Για την αξιολόγηση της προτεινόμενης μεθόδου, οι συγγραφείς ανέπτυξαν και συνέκριναν τρία διαφορετικά μοντέλα. Το πρώτο μοντέλο χρησιμοποιεί embeddings που προέρχονται αποκλειστικά και μόνο από κείμενα των reviews. Το δεύτερο μοντέλο βασίζεται σε embeddings που προέρχονται από το δίκτυο κριτικών-προϊόντων. Το τρίτο μοντέλο συνδυάζει με τη συνένωση των embeddings χαρακτηριστικά από τις κριτικές, τους χρήστες και τα προϊόντα. Κάθε μοντέλο αξιολογείται με τη χρήση του cross validation με βάση τη μέση ακρίβεια (AP) και της περιοχής κάτω από την καμπύλη ROC (AUC). Τα αποτελέσματα υποδεικνύουν ότι το μοντέλο που χρησιμοποιεί συνδυασμένα διανύσματα χαρακτηριστικών υπερτερεί σημαντικά έναντι τόσο των μεμονωμένων μοντέλων που βασίζονται σε κείμενο όσο και των μοντέλων που βασίζονται σε δίκτυο. Κατά συνέπεια, οι συγγραφείς καταλήγουν στο συμπέρασμα ότι η προσέγγισή τους, SPR2EP, ξεπερνά τις προηγούμενες υπάρχουσες σύγχρονες μεθόδους ανίχνευσης παραπλανητικών κριτικών.



Εικόνα 5: *Επισκόπηση του προτεινόμενου πλαισίου ανίχνευσης Spam Review.*
Πηγή: SPR2EP: A Semi-Supervised Spam Review Detection Framework

Ο Zhang et al (2018) εισάγει μια νέα προσέγγιση, που ονομάζεται DRI-RCNN (Deceptive Review Identification by Recurrent Convolutional Neural Network), η οποία αξιοποιεί πληροφορίες σχετικά με το context των λέξεων για να βελτιώσει την ανίχνευση παραπλανητικών κριτικών. Το DRI-RCNN υιοθετεί τη μέθοδο Skip-gram για να δημιουργήσει word embeddings βάση τόσο των ιδίων των λέξεων όσο και των γειτονικών τους, αναπαριστώντας κάθε λέξη με έξι συνιστώσες: παραπλανητικές και αληθινές αναπαραστάσεις μέσω word emdebbings, διανύσματα συμφραζομένων αριστερά και δεξιά από την εκάστοτε λέξη. Αυτές οι αναπαραστάσεις προκύπτουν με τη χρήση ενός αναδρομικού νευρωνικού δικτύου συνελικτικού τύπου που συλλαμβάνει τις τοπικές και διαδοχικές εξαρτήσεις στο κείμενο. Το μοντέλο χρησιμοποιεί max-pooling για την επιλογή των πιο κρίσιμων χαρακτηριστικών από τα διανύσματα επαναλαμβανόμενης συνέλιξης και στρώματα ReLU (Rectified Linear Unit) για το φιλτράρισμα και την ενίσχυση αυτών των χαρακτηριστικών, αφαιρώντας τις αρνητικές τιμές. Τέλος, εφαρμόζει ένα πλήρως συνδεδεμένο δίκτυο με αντίστροφη διάδοση (back propagation) ακολουθούμενο από μια συνάρτηση softmax για να ταξινομήσει τις κριτικές είτε ως παραπλανητικές είτε ως αληθινές. Η προσέγγιση DRI-RCNN αξιολογήθηκε σε δύο σύνολα δεδομένων αναφοράς, ένα σύνολο δεδομένων με spam περιεχόμενο και ένα σύνολο δεδομένων με κείμενα παραπλάνησης. Στα πειράματα που διεκπεραιώθηκαν αποδείχτηκε πως η παραπάνω μεθοδολογία υπερτερεί σε σύγκριση με τις υπάρχουσες σύγχρονες τεχνικές στην αναγνώριση παραπλανητικών κειμένων.

Ο Aghakhani et al. (2018) πρότεινε το FakeGAN, μια προσέγγιση που αξιοποιεί τη δύναμη των Generative Adversarial Networks (GAN) και των μεθόδων μάθησης με ημιεπίβλεψη για την ανίχνευση παραπλανητικών κειμένων.

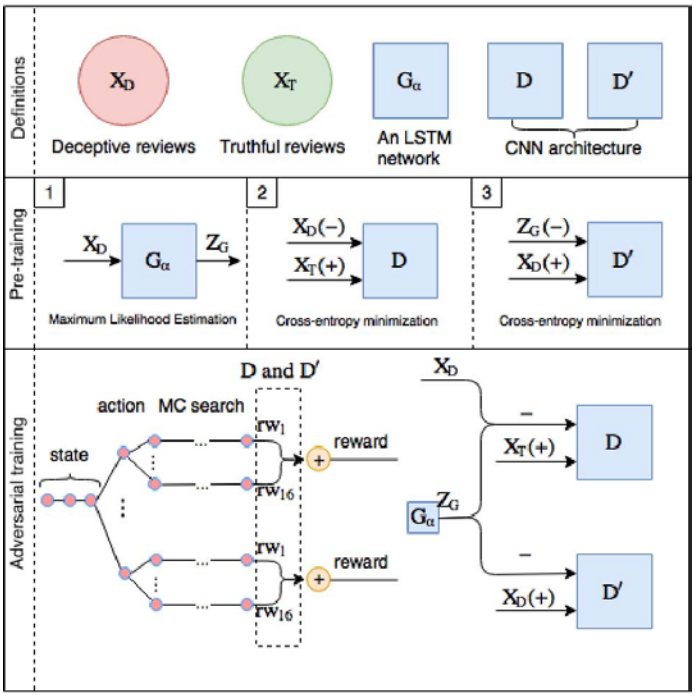
Το FakeGAN χρησιμοποιεί δύο discriminators για την ενίσχυση της σταθερότητας και της ευρωστίας του μοντέλου, παρέχοντας λύση στο πρόβλημα toy mode collapse που συχνά αντιμετωπίζεται στην εκπαίδευση των GAN και ενισχύει το σύστημα ώστε να μαθαίνει αποτελεσματικά τόσο από αληθείς όσο και από παραπλανητικές κατανομές κριτικών.

Από την άλλη πλευρά, ο generator στο FakeGAN μοντελοποιείται ως ένας στοχαστικός πράκτορας πολιτικής σε ένα περιβάλλον ενισχυτικής μάθησης, χρησιμοποιώντας την αναζήτηση Monte Carlo για την ενίσχυση της ποιότητας των παραγόμενων δειγμάτων.

Όσον αφορά την αρχιτεκτονική , ο generator κατασκευάζεται με τη χρήση αναδρομικών νευρωνικών δικτύων (RNN), συγκεκριμένα με τη χρήση μακράς βραχυπρόθεσμης μνήμης (LSTM), για να χειρίζεται ακολουθίες δεδομένων και να καταγράφει μακροπρόθεσμες εξαρτήσεις, γεγονός που βοηθά στη δημιουργία συνεκτικού κειμένου. Από την άλλη πλευρά, οι discriminators κατασκευάζονται με τη χρήση συνεπτυγμένων νευρωνικών δικτύων (CNN), τα οποία είναι αποτελεσματικά για την ταξινόμηση κειμένου.

Πιο συγκεκριμένα, το CNN χρησιμοποιεί μια μη γραμμική συνάρτηση (ReLU), τεχνικές max pooling και ένα πλήρως συνδεδεμένο στρώμα εξόδου με σιγμοειδή συνάρτηση για την παραγωγή της πιθανότητας της κατηγορίας-στόχου.

Το FakeGAN δοκιμάστηκε σε ένα σύνολο δεδομένων που αποτελείται από κριτικές ξενοδοχείων του TripAdvisor και η απόδοσή του αξιολογήθηκε με βάση τη μετρική ακρίβειας μέσω διασταυρούμενης επικύρωσης.



Εικόνα 6: *Η επισκόπηση του FakeGAN. Τα σύμβολα + και -υποδηλώνουν θετικά και αρνητικά δείγματα αντίστοιχα. Σημειώστε ότι, αυτά διαφέρουν από τις αληθινές και τις παραπλανητικές κριτικές.*

Πηγή:https://www.semanticscholar.org/reader/02fdca762652c50d46619a886cfc0754732b05c5

Η ποικιλομορφία όσον αφορά τη φύση των παραπλανητικών κειμένων σε διαφορετικούς τομείς περιπλέκει τη σύγκριση και τη γενίκευση των μοντέλων ανίχνευσης. Για τον λόγο αυτό ο **Velutharambath et al. (2023)** πρότεινε μια ολοκληρωμένη επισκόπηση των προκλήσεων στην ανίχνευση της λεκτικής εξαπάτησης σε πολλαπλούς τομείς δραστηριότητας. Για την αντιμετώπιση του παραπάνω προβλήματος, οι συγγραφείς ενοποιούν διάφορα δημόσια διαθέσιμα σύνολα δεδομένων εξαπάτησης που συναντώνται στη σχετική βιβλιογραφία σε αγγλική γλώσσα σε ένα ενιαίο σύνολο δεδομένων με την ονομασία UNIDECOR. Αυτό το σύνολο δεδομένων περιλαμβάνει κείμενα από ένα ευρύ φάσμα τομέων, όπως κριτικές στα μέσα κοινωνικής δικτύωσης, δικαστικές μαρτυρίες, δηλώσεις γνώμης και παραπλανητικούς διαλόγους από διαδικτυακά παιχνίδια στρατηγικής. Σε κάθε εγγραφή σε αυτό το ενοποιημένο σύνολο ανατίθεται μια δυαδική ετικέτα που υποδεικνύει αν είναι αληθής ή παραπλανητική. Για να μελετήσουν τη δυνατότητα γενίκευσης των γλωσσικών ενδείξεων σε διαφορετικά σύνολα δεδομένων, οι συγγραφείς διεξήγαγαν μια ανάλυση συσχέτισης χρησιμοποιώντας το μέτρο ομοιότητας των επιμέρους corpus που ορίστηκε από τους Li και Dunn (2022), με στόχο τον εντοπισμό συχνά χρησιμοποιούμενων χαρακτηριστικών που είναι γενικά και εφαρμόζονται σε όλους τους τομείς. Ωστόσο, δεν βρέθηκαν συνεπή γλωσσικά στοιχεία σε

όλους τους τομείς. Για την αξιολόγηση της αποτελεσματικότητας της ανίχνευσης εξαπάτησης, η τρέχουσα ερευνητική μελέτη χρησιμοποιεί ένα πειραματικό πλαίσιο με τη χρήση λεπτομερώς ρυθμισμένων μοντέλων RoBERTa (Liu et al., 2019). Δύο είδη πειραμάτων εκτελέστηκαν στη συγκεκριμένη εργασία. Στην αξιολόγηση within-corpus, τα μοντέλα εκπαιδεύονται και δοκιμάζονται στο ίδιο σύνολο δεδομένων μέσω διασταυρούμενης επικύρωσης. Από την άλλη πλευρά, για την αξιολόγηση cross-corpus, τα μοντέλα εκπαιδεύονται σε ένα σύνολο δεδομένων και δοκιμάζονται σε ένα άλλο, χρησιμοποιώντας μια στρατηγική διασταυρούμενης επικύρωσης σε όλα τα υποσύνολα για να διατηρηθεί η συνέπεια στην αξιολόγηση της απόδοσης. Τα αποτελέσματα αποκαλύπτουν ότι τα μοντέλα αποδίδουν γενικά καλύτερα σε περιβάλλον within-corpus από ότι σε cross-corpus. Σε γενικές γραμμές, σύνολα δεδομένων που αφορούν διαφορετικό τομέα ή έχουν διαφορές στη δομή του κειμένου τους παρουσιάζουν κακή απόδοση σε cross-corpus περιβάλλον. Ωστόσο, σύνολα δεδομένων από παρόμοιους τομείς, παρουσιάζουν σχετικά καλύτερες επιδόσεις, υποδεικνύοντας κάποια δυνατότητα χρήσης των ενδείξεων εξαπάτησης σε σύνολα δεδομένων που συσχετίζονται όσον αφορά τον τομέα.

Ο Loukas et al. (2022) προτείνει και αξιολογεί έξι μοντέλα βαθιάς μάθησης για τη βελτίωση της αυτοματοποιημένης ανίχνευσης εξαπάτησης. Τα μοντέλα αυτά περιλαμβάνουν συνδυασμούς των μεθόδων BERT (και RoBERTa), MultiHead Attention, Self-Attention και Transformers. Τα μοντέλα αυτά αξιολογούνται σε δύο σύνολα δεδομένων: το ένα προέρχεται από την μελέτη των Kleinberg and Verschuere (2021) και περιλαμβάνει statements που συλλέγονται μέσω του Prolific Academic, ενώ το δεύτερο με όνομα Open Domain Deception Dataset προέρχεται από την μελέτη Pérez-Rosas and Mihalcea (2015) με statements που προέρχονται από το Amazon Mechanical Turk. Η χρήση αυτών των συνόλων δεδομένων επιβεβαιώνει την ανθεκτικότητα των προτεινόμενων προσεγγίσεων σε διαφορετικά πλαίσια και τομείς, ενώ διασφαλίζει παράλληλα και την ολοκληρωμένη αξιολόγηση των μοντέλων ανίχνευσης εξαπάτησης. Στα πλαίσια των πειραμάτων, η μελέτη αξιολογεί τις προτεινόμενες προσεγγίσεις έναντι της ανθρώπινης κρίσης, αλλά και μοντέλων μηχανικής μάθησης που χρησιμοποιούν χαρακτηριστικά LIWC και POS. Τα αποτελέσματα δείχνουν ότι μπορούμε να βελτιώσουμε την αυτοματοποιημένη ανίχνευση εξαπάτησης χρησιμοποιώντας μοντέλα που βασίζονται σε Transformers (+2,11% στην ακρίβεια). Επιπροσθέτως, οι συγγραφείς πραγματοποίησαν επίσης μια εις βάθος ανάλυση που αποκαλύπτει τόσο τις ομοιότητες όσο

και τις διαφορές μεταξύ αληθινών και παραπλανητικών δηλώσεων, τονίζοντας τις συσχετίσεις συγκεκριμένων κατηγοριών LIWC.

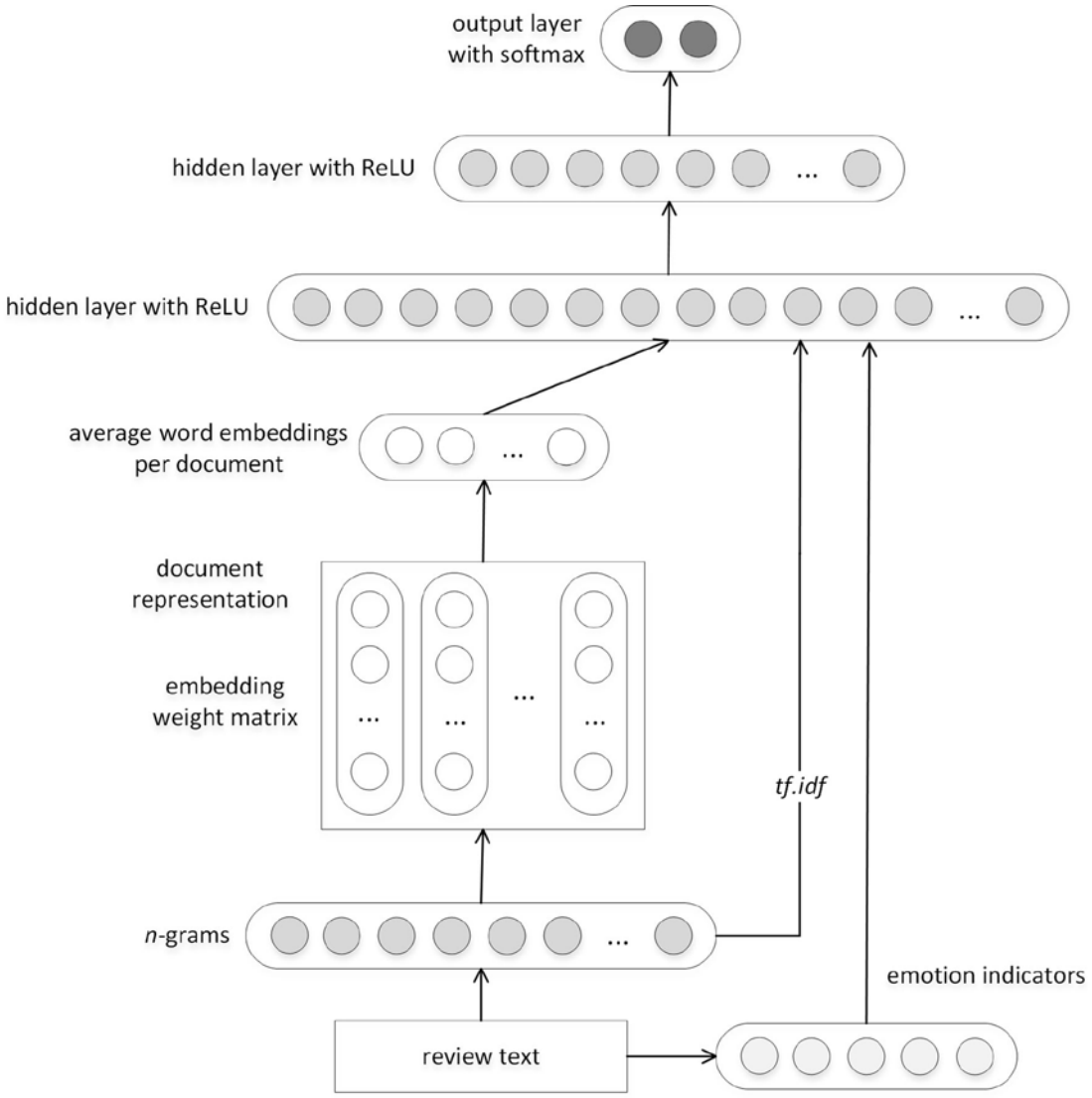
Τέλος, ενσωμάτωσαν επίσης τη μέθοδο LIME με σκοπό την ερμηνευσιμότητα των μοντέλων, το οποίο αποκάλυψε ότι συχνά συμβαίνουν λανθασμένες ταξινομήσεις λόγω των παραπλανητικών χαρακτηριστικών LIWC που εμφανίζονται σε αληθινές δηλώσεις και αντίστροφα.

O Hijek et al. (2020) αντιμετωπίζει το ζήτημα της ανίχνευσης παραπλανητικών κριτικών προτείνοντας δύο νέα μοντέλα νευρωνικών δικτύων, τα οποία χρησιμοποιούν τη λογική των BagOfWords αλλά και δείκτες συναισθήματος (emotion indicators). Τα μοντέλα αναπαριστούν τις κριτικές σε επίπεδο εγγράφου χρησιμοποιώντας n-grams, word emdeddings και δείκτες συναισθημάτων βασισμένους σε λεξικό.

Πιο συγκεκριμένα, στην παρούσα μελέτη χρησιμοποιήθηκαν τρεις τύποι δεικτών συναισθήματος, συμπεριλαμβανομένων των χαρακτηριστικών πολικότητας (polarity), ισχύος (strength) και συναισθήματος (emotion). Πιο συγκεκριμένα, οι συγγραφείς πρότειναν ένα βελτιωμένο νευρωνικό δίκτυο βαθιάς τροφοδότησης (Deep Feedforward Neural Network - DFFNN) και ένα συνελικτικό νευρωνικό δίκτυο (Convolutional Neural Network - CNN).

Το DFFNN αναπαρίσταται από ένα multi-layer perceptron με δύο hidden layers και ένα input layer, όπου τα τρία σύνολα χαρακτηριστικών εξάγονται από το ακατέργαστο κείμενο των reviews. Από την άλλη πλευρά, το CNN μοντέλο, μετατρέπει κάθε πρόταση στην v-διάστατη διανυσματική αναπαράσταση λέξεων χρησιμοποιώντας pretrained embeddings. Επιπλέον, ενσωματώνονται τα TF-IDF βάρη και οι δείκτες συναισθήματος για κάθε λέξη.

Εκτελέστηκαν πειράματα σε τέσσερα real-life σύνολα δεδομένων για να αποδειχθεί η αποτελεσματικότητα των προτεινόμενων μοντέλων. Βάσει των αποτελεσμάτων, τα προτεινόμενα μοντέλα υπερέβησαν τις τρέχουσες προσεγγίσεις και τις σύγχρονες μεθόδους ανίχνευσης παραπλανητικών κριτικών όσον αφορά την ακρίβεια, την AUC και το F-score.



Εικόνα 7: Μοντέλο DFFNN για την ανίχνευση ψεύτικων κριτικών.
Πηγή: https://www.researchgate.net/publication/338982783_Fake_consumer_review_detection_using_deep_neural_networks_integrating_word_embeddings_and_emotion_mining

Ο Peskov et al. (2020) διεξήγαγε έρευνα για την ανίχνευση εξαπάτησης σε κείμενα που προέρχονται από διαδικτυακές συνομιλίες, οι οποίες δημιουργούνται σε ένα παιχνίδι που βασίζεται σε διαπραγματεύσεις και ονομάζεται Diplomacy. Σε αυτό το παιχνίδι οι παίκτες συνάπτουν ή ακυρώνουν συμμαχίες μεταξύ τους και οι συγγραφείς στοχεύουν να ανακαλύψουν τη στρατηγική χρήση της εξαπάτησης στις μακροχρόνιες αλληλεπιδράσεις τους. Για κάθε μήνυμα που ανταλλάσσεται στο πλαίσιο του παιχνιδιού, ο αποστολέας σχολιάζει αν περιέχει ένα ACTUAL LIE, ενώ από την άλλη πλευρά ο παραλήπτης σχολιάζει αν το μήνυμα γίνεται αντιληπτό ως SUSPECTED LIE ή όχι. Ως εκ τούτου, από τα δεδομένα μπορούν να προκύψουν τέσσερις διαφορετικοί συνδυασμοί εξαπάτησης και αντίληψης. Ο αποστολέας είτε λέει ψέματα είτε λέει την αλήθεια. Ομοίως, το μήνυμα μπορεί να

ερμηνευθεί από τον παραλήπτη είτε ως παραπλανητικό είτε ως αληθινό. Στην παρούσα μελέτη, χρησιμοποιήθηκε ένα μοντέλο λογιστικής παλινδρόμησης και ένα τυπικό δίκτυο LSTM. Και τα δύο μοντέλα συγκρίθηκαν με ένα human baseline που εντόπισε σχεδόν το 90% των ψεμάτων.

Λόγω της imbalanced φύσης του συνόλου δεδομένων, καθώς μόνο το 5% των μηνυμάτων είχαν σχολιαστεί ως παραπλανητικά, οι ερευνητές υιοθέτησαν μια διαδικασία εκπαίδευσης ενσωματώνοντας ένα σταθμισμένο F1-score, προκειμένου να τιμωρηθεί περισσότερο η εσφαλμένη ταξινόμηση των περιπτώσεων της μειοψηφικής κατηγορίας (ACTUAL LIES). Δοκιμάστηκε επίσης η λεπτομερής ρύθμιση των pretrained BERT, αλλά δεν επέφερε καμία βελτίωση στην απόδοση.

Ο Capuozzo et al. (2020) παρουσίασε το DecOp (Deceptive Opinions), ένα καινοτόμο πλαίσιο για την αυτόματη ανίχνευση εξαπάτησης σε cross-domain και cross-language σενάρια. Τα δεδομένα που χρησιμοποιήθηκαν σε αυτή τη μελέτη εξήχθησαν από first person opinions, αληθινές και παραπλανητικές, σε πέντε διαφορετικούς τομείς (άμβλωση, νομιμοποίηση κάνναβης, ευθανασία, γάμος ομοφυλοφίλων και πολιτική για τους μετανάστες) και δύο γλώσσες (ΗΠΑ, ιταλικά).

Όλοι οι συμμετέχοντες, είτε από το Amazon Mechanical Turk είτε μέσω απλών Google forms, κλήθηκαν να δώσουν δύο απαντήσεις. Αρχικά, την πραγματική τους άποψη για κάποιο θέμα και στη συνέχεια μια ψεύτικη δήλωση, η οποία θα μπορούσε υποθετικά να χρησιμοποιηθεί για να πείσει κάποιον ότι αντιπροσωπεύει την πραγματική τους άποψη.

Το πρόβλημα διερευνήθηκε από τρεις διαφορετικές οπτικές γωνίες, within-topic, cross-topic and author-based. Χρησιμοποιήθηκε μια abstract έκδοση της αρχιτεκτονικής Tansformer, η οποία αποτελείται από ένα embedding layer και μια stacked ακολουθία από transformer μπλοκ, ενώ το FastText χρησιμοποιήθηκε για την κωδικοποίηση των ακολουθιών. Τα BERT, ROBERTA και ALBERT φάνηκαν ανεπαρκή στη συγκεκριμένη μελέτη.

Κεφάλαιο 4

Η Μεθοδολογία που Ακολουθήθηκε

4.1 Περιγραφή Συνόλου Δεδομένων

DeRev

Το DEREV dataset δημιουργήθηκε με σκοπό να μελετηθεί η εξαπάτηση σε κριτικές βιβλίων του Amazon, κατηγοριοποιώντας τα σε ύποπτες και αθώες ομάδες. Περιέχει 6819 κριτικές από 4811 κριτικούς για 68 βιβλία. Για τον εντοπισμό δυνητικά παραπλανητικών κριτικών χρησιμοποιήθηκαν τέσσερις ενδείξεις: η συσχέτιση της κριτικής με ύποπτα βιβλία (SB), η ομαδοποίηση των κριτικών μέσα σε σύντομο χρονικό διάστημα (Cl), η χρήση ψευδώνυμων αντί πραγματικών ονομάτων (NN) και η έλλειψη επαλήθευσης αγοράς στο Amazon (UP). Οι κριτικές ταξινομήθηκαν ευρετικά χρησιμοποιώντας αυτές τις ενδείξεις, σχηματίζοντας ένα πρότυπο πιθανής ειλικρίνειας. Για να δημιουργηθεί ένας πιο αξιόπιστος χρυσός κανόνας, 118 γνωστές ψεύτικες κριτικές από βιβλία με αγορασμένες κριτικές, αντιπαραβλήθηκαν με 118 κριτικές που θεωρήθηκαν γνήσιες, δημιουργώντας ένα σύνολο δεδομένων 236 κριτικών για επικύρωση και περαιτέρω ανάλυση.

Reference: Fornaciari, Tommaso and Massimo Poesio. “Identifying fake Amazon reviews as learning from crowds.” Conference of the European Chapter of the Association for Computational Linguistics (2014).

OP SPAM

Το Op Spam dataset, το οποίο αναπτύχθηκε από τους Ott et al. 2011, χρησιμεύει ως σύνολο δεδομένων αναφοράς για την ανίχνευση παραπλανητικών μηνυμάτων spam. Αποτελείται από κριτικές ξενοδοχείων, με επίκεντρο 20 δημοφιλή ξενοδοχεία στο Σικάγο. Οι παραπλανητικές θετικές κριτικές προήλθαν από το Amazon Mechanical Turk, όπου ελεγχόμενοι turkers με έδρα τις ΗΠΑ έγραψαν 400 κριτικές, διασφαλίζοντας την ποιότητα μέσω αυστηρών κριτηρίων. Οι αληθινές θετικές κριτικές συλλέχθηκαν από το TripAdvisor, επιλέγοντας μόνο κριτικές 5 αστέρων με περισσότερους από 150 χαρακτήρες από συγγραφείς με πολλαπλές κριτικές, με αποτέλεσμα να προκύψουν 400 αληθινές κριτικές.

Παρόμοια διαδικασία εφαρμόστηκε και για τις κριτικές αρνητικού συναισθήματος, με 400 παραπλανητικές κριτικές από το Mechanical Turk και 400 αληθινές κριτικές από διάφορες διαδικτυακές πλατφόρμες, συμπεριλαμβανομένων των Expedia και Yelp. Επιπλέον μια ομάδα εθελοντών, συγκεκριμένα μη εκπαιδευμένοι προπτυχιακοί φοιτητές, χρησιμοποιήθηκαν για να αξιολογήσουν την ανίχνευση εξαπάτησης του συνόλου δεδομένων, αποκαλύπτοντας ότι οι άνθρωποι είναι καλύτεροι στον εντοπισμό αρνητικών παραπλανητικών κριτικών (61%) από ό,τι θετικών (57%), αν και οι αυτοματοποιημένοι ταξινομητές εξακολουθούν να έχουν συνολικά καλύτερες επιδόσεις. Το σύνολο δεδομένων ενοποιήθηκε ώστε να περιλαμβάνει από 400 αληθείς και παραπλανητικές κριτικές για θετικά και αρνητικά συναισθήματα.

Reference: Ott, Myle, Yejin Choi, Claire Cardie and Jeffrey T. Hancock. “Finding Deceptive Opinion Spam by Any Stretch of the Imagination.” Annual Meeting of the Association for Computational Linguistics (2011).

Cross Cultural

Το CrossCultural dataset περιλαμβάνει διάφορα σύνολα δεδομένων σε διαφορετικές γλώσσες. Στη δική μας εφαρμογή επικεντρωθήκαμε μόνο στα αγγλικά που ομιλούνται στις Ηνωμένες Πολιτείες (EnglishUS). Αυτό το σύνολο δεδομένων αποτελείται από σύντομα δοκίμια σχετικά με τρία θέματα: απόψεις για την άμβλωση, απόψεις για τη θανατική ποινή και συναισθήματα για έναν καλύτερο φίλο. Για τη συλλογή των δεδομένων προσλήφθηκαν άτομα μέσω του Amazon Mechanical Turk, με περιορισμούς τοποθεσίας για να διασφαλιστεί ότι οι συμμετέχοντες ήταν από τις Ηνωμένες Πολιτείες. Οι συμμετέχοντες κλήθηκαν να γράψουν τόσο αληθινά όσο και παραπλανητικά δοκίμια για τα συγκεκριμένα θέματα. Κατά τη διαδικασία προετοιμασίας των δεδομένων, οι ερευνητές εντόπισαν και διόρθωσαν συστηματικά σφάλματα στίξης. Ωστόσο, άφησαν τα ορθογραφικά λάθη χωρίς διόρθωση, καθώς ήταν ομοιόμορφα κατανεμημένα τόσο στα παραπλανητικά όσο και στα ειλικρινή κείμενα, χωρίς να φανερώνουν κάποια σαφή προκατάληψη.

Reference: Perez-Rosas, V., & Mihalcea, R. (2014). Cross-cultural Deception Detection. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 440–445). Association for Computational Linguistics.

Amazon Reviews

Αυτό το σύνολο δεδομένων, το οποίο είναι διαθέσιμο στο Kaggle, αποτελείται από κριτικές προϊόντων του Amazon που έχουν χαρακτηριστεί είτε ως ψεύτικες είτε ως πραγματικές. Στο σύνολο δεδομένων, αυτές οι ετικέτες αναπαρίστανται με τη λέξη κλειδί __label1__ για τις ψεύτικες κριτικές και __label2__ για τις πραγματικές κριτικές. Κάθε καταχώρηση κριτικής περιλαμβάνει διάφορα άλλα χαρακτηριστικά, όπως η βαθμολογία που δόθηκε, αν η αγορά επαληθεύτηκε, η κατηγορία στην οποίο ανήκει το προϊόν, το μοναδικό αναγνωριστικό του προϊόντος, ο τίτλος του προϊόντος και ο τίτλος της ίδιας της κριτικής. Το σύνολο δεδομένων περιλαμβάνει συνολικά 21.000 κριτικές, ομοιόμορφα κατανεμημένες σε διάφορες κατηγορίες προϊόντων, και οι κριτικές αυτές έχουν χαρακτηριστεί ως "μη συμμορφούμενες" σύμφωνα με τις πολιτικές της Amazon. Για τη μελέτη μας, αγνοήσαμε όλα τα πρόσθετα μεταδεδομένα, εστιάζοντας μόνο στο κείμενο των κριτικών για να ακολουθήσουμε το ίδιο μοτίβο και με τα υπόλοιπα σύνολα δεδομένων.
Reference: https://www.kaggle.com/datasets/lievgarcia/amazon-reviews

Restaurants Reviews

Το RestaurantsDataset, το οποίο παρουσιάστηκε για πρώτη φορά από τον Faranak Abri [ ] σε προηγούμενη έρευνα, περιέχει κριτικές για τρία τοπικά εστιατόρια. Οι αυθεντικές κριτικές προήλθαν από πραγματικούς πελάτες μέσω διαδικτυακών πλατφορμών που δίνουν τη δυνατότητα στους πελάτες να αξιολογήσουν ένα εστιατόριο και να σχολιάσουν την εμπειρία τους. Για να δημιουργήσουν ψεύτικες κριτικές, οι συγγραφείς επιστράτευσαν τέσσερις προπτυχιακούς φοιτητές για να γράψουν φανταστικές κριτικές για τα ίδια εστιατόρια. Το σύνολο δεδομένων είναι ισορροπημένο, διαθέτοντας ίση κατανομή θετικών και αρνητικών κριτικών, καθώς και ίσο αριθμό πραγματικών και ψεύτικων κριτικών, συνολικά 110 κριτικές.
Reference: Faranak Abri, Luis Felipe Gutiérrez, Akbar Siami Namin, Keith S. Jones, and David R. W. Sears, Linguistic Features for Detecting Fake Reviews. International Conference on Machine Learning Applications (ICMLA) 2020, December 2020, Miami - Florida, US.

Κάθε σύνολο δεδομένων που παρουσιάζεται στην παραπάνω ενότητα έχει υποστεί κάποια προ επεξεργασία για να εξασφαλιστεί η ομοιομορφία μεταξύ αυτών και να εξαλειφθούν

πιθανές ασυμφωνίες στις μορφή και το περιεχόμενο, επιτρέποντας τη συνεπή ανάλυση κατά τη διάρκεια της πειραματικής φάσης στην ενότητα 5. Κατά τη διάρκεια αυτής της διαδικασίας, εξαλείφονται τα περιττά στοιχεία και τα metadata που σχετίζονται με το κείμενο της εκάστοτε κριτικής ή δοκιμίου. Οι όροι "truthful" και "deceptive" χρησιμοποιούνται για την τυποποίηση των ετικετών σε μια συνεπή ορολογία, αντικαθιστώντας δυαδικές ή ποικίλες ετικέτες όπως 0/1, Label1/Label2 ή true/false. Επιπλέον, τα σύνολα δεδομένων από κάθε πηγή βρίσκονταν σε διάφορες μορφές, για παράδειγμα: τα σύνολα δεδομένων Cross-Cultural και OP_Spam περιέχουν ξεχωριστά αρχεία txt κατανεμημένα σε πολλαπλούς υποφακέλους με βάση τη γλώσσα ή τον τομέα, ενώ το σύνολο δεδομένων DeRev αποτελείται από μεμονωμένα αρχεία XML για κάθε κριτική. Για να αντιμετωπιστεί αυτό το ζήτημα, όλα τα σύνολα δεδομένων μετατράπηκαν σε σειριακά αρχεία csv, περιλαμβάνοντας μόνο το κείμενο ως ανεξάρτητη μεταβλητή και την παρατηρούμενη κλάση (αληθής ή παραπλανητική) ως ground truth label.

4.2 Πειραματικό Πλαίσιο

Στο τρέχον κεφάλαιο περιγράφεται η μεθοδολογία που χρησιμοποιήθηκε για τη διερεύνηση της αποτελεσματικότητας διαφόρων τεχνικών Βαθιάς Μάθησης για εργασίες ταξινόμησης κειμένου σε πολλαπλά σύνολα δεδομένων με σκοπό την ανίχνευση εξαπάτησης. Η έρευνα αποσκοπεί στην αξιολόγηση των επιδόσεων αυτών των μοντέλων στην ταξινόμηση παραπλανητικών και αληθινών κειμένων, αξιοποιώντας τεχνικές βαθιάς μάθησης για την επεξεργασία φυσικής γλώσσας. Η μελέτη χρησιμοποιεί ένα συγκριτικό πειραματικό πλαίσιο για την αξιολόγηση διαφόρων αρχιτεκτονικών βαθιάς μάθησης: ξεκινώντας από ένα απλό RNN, αλλά φτάνοντας και σε προσεγγίσεις αιχμής, όπως μοντέλα Transformers με μηχανισμούς που βασίζονται στην προσοχή και τα πιο σύγχρονα BERT και RoBERTa που βασίζονται στους transformers. Τα μοντέλα αυτά επιλέχθηκαν με βάση την ικανότητά τους να συλλαμβάνουν διαδοχικές εξαρτήσεις και εξαρτήσεις μεγάλης εμβέλειας αντίστοιχα, οι οποίες είναι ζωτικής σημασίας για την κατανόηση κειμένου. Μια συλλογή διαφορετικών συνόλων δεδομένων που περιέχουν επισημασμένα παραπλανητικά και αληθινά δεδομένα κειμένου χρησιμοποιείται με σκοπό την εκπαίδευση και επικύρωσης των μοντέλων. Τα σύνολα δεδομένων προέρχονται από αξιόπιστες πηγές και καλύπτουν ένα εύρος τομέων, ώστε να διασφαλίζεται η γενίκευση των ευρημάτων. Περισσότερες λεπτομέρειες σχετικά με

τον τομέα που αναφέρεται κάθε σύνολο δεδομένων και τη διαδικασία που χρησιμοποιήθηκε για την εξαγωγή δεδομένων αναφέρθηκαν προηγουμένως.

4.3 Τεχνική Υλοποίηση

Τα πειράματα υλοποιούνται με τη χρήση της γλώσσας προγραμματισμού Python στο περιβάλλον Google Colab, αξιοποιώντας την διαθέσιμη GPU του περιβάλλοντος αυτού για αποτελεσματική και γρήγορη εκπαίδευση των μοντέλων. Βιβλιοθήκες όπως οι PyTorch, TorchText και Scikit-learn χρησιμοποιούνται για την προεπεξεργασία δεδομένων, την ανάπτυξη μοντέλων και την αξιολόγηση της απόδοσης. Τέλος χρησιμοποιείται και η Hugging Face Transformers η οποία παρέχει τα προεκπαιδευμένα μοντέλα BERT και RoBERTa μαζί με tokenizers, καθιστώντας εύκολη την αναγνωσιμότητα αυτών των μοντέλων για εργασίες ταξινόμησης ακολουθίας.

4.4 Προεπεξεργασία Δεδομένων

Χρησιμοποιώντας τον βασικό αγγλικό tokenizer (ένα εργαλείο για την επεξεργασία κειμένου που επιτρέπει τη διάσπαση μιας πρότασης ή κειμένου σε διακριτά τμήματα, όπως λέξεις ή σύμβολα) που παρέχεται από το TorchText, κάθε δείγμα κειμένου διαιρείται σε tokens και μετατρέπεται σε πεζά γράμματα για να εξασφαλιστεί η ομοιομορφία στις αναπαραστάσεις των λέξεων. Τα stopwords, κοινές λέξεις που δεν συμβάλλουν σημαντικά στο νόημα του κειμένου (π.χ. "the", "is", "and"), αφαιρούνται για να μειωθεί ο θόρυβος. Επιπλέον, τα σημεία στίξης, όπως τα κόμματα και οι τελείες, αφαιρούνται από το κείμενο για να διατηρηθεί ένα καθαρότερο σύνολο δεδομένων που επικεντρώνεται στο ουσιαστικό περιεχόμενο.
Στη συνέχεια, κατασκευάζεται ένα λεξικό από τα επεξεργασμένα tokens, ενσωματώνοντας ειδικά tokens για padding και την διαχείριση των άγνωστων λέξεων με σκοπό να διευκολυνθεί η εκπαίδευση και η αξιολόγηση του μοντέλου. Ειδικά tokens όπως <PAD> (συμπλήρωση) και <UNK> (άγνωστο) περιλαμβάνονται για να χειρίζονται ακολουθίες μεταβλητού μήκους κατά την εκπαίδευση και την αξιολόγηση του μοντέλου. Αυτά τα βήματα προεπεξεργασίας είναι απαραίτητα για την ενίσχυση της ποιότητας των δεδομένων που τροφοδοτούνται στα μοντέλα, βελτιώνοντας την ικανότητά των μοντέλων να διακρίνουν τα παραπλανητικά από τα

αληθινά κείμενα με βάση το σημασιολογικό περιεχόμενο και όχι τα ξένα γλωσσικά χαρακτηριστικά.

4.5 Διανυσματοποίηση Κειμένου

Η διαδικασία δημιουργίας λεξιλογίου περιλαμβάνει την σάρωση ενός corpus (μεγάλο και δομημένο σύνολο κειμένων) για τη συλλογή όλων των μοναδικών λέξεων, δημιουργώντας ουσιαστικά έναν πλήρη κατάλογο όρων που χρησιμοποιούνται στο σύνολο δεδομένων. Μόλις εντοπιστούν οι μοναδικές λέξεις, σε κάθε λέξη ανατίθεται ένας μοναδικός δείκτης, σχηματίζοντας μια δομή που μοιάζει με λεξικό, όπου κάθε λέξη αντιστοιχεί σε έναν ακέραιο αριθμό. Αυτό το λεξιλόγιο χρησιμεύει στη συνέχεια ως βάση για τη μετατροπή του κειμένου σε αριθμητικές αναπαραστάσεις, όπου κάθε λέξη σε ένα κείμενο αντικαθίσταται με τον αντίστοιχο δείκτη από το λεξιλόγιο. Αυτή η αναπαράσταση με δείκτες είναι απαραίτητη για την περαιτέρω επεξεργασία σε μοντέλα μηχανικής μάθησης, ιδίως σε περιβάλλοντα βαθιάς μάθησης όπου οι λέξεις συχνά ενσωματώνονται σε πυκνούς διανυσματικούς χώρους.

4.6 Διανυσματοποίηση Κειμένου Εναλλακτικές

Στην ανάλυσή μας, δοκιμάσαμε επίσης δύο διακεκριμένες τεχνικές διανυσματοποίησης κειμένου: Bag of Words (BoW) και Term Frequency-Inverse Document Frequency (TF-IDF). Η μέθοδος BoW περιλαμβάνει τη δημιουργία ενός αραιού πίνακα αναπαράστασης των δεδομένων κειμένου, όπου κάθε μοναδική λέξη στο σώμα κειμένων συσχετίζεται με μια στήλη του πίνακα. Οι εγγραφές στον πίνακα αντιπροσωπεύουν τη συχνότητα κάθε λέξης σε ένα δεδομένο έγγραφο, καταγράφοντας την παρουσία και τον αριθμό των λέξεων, αλλά αγνοώντας τη γραμματική και τη σειρά με την οποία εμφανίζονται. Από την άλλη πλευρά, ο TF-IDF ενισχύει την προσέγγιση BoW, καθώς δεν λαμβάνει υπόψη μόνο τη συχνότητα των λέξεων, αλλά σταθμίζει επίσης τη σημασία κάθε λέξης με βάση το πόσο συχνά εμφανίζεται σε όλα τα υπόλοιπα έγγραφα. Αυτή η τεχνική βοηθά στη μείωση της επιρροής των συχνά χρησιμοποιούμενων λέξεων όπως "και" ή "είναι", οι οποίες μπορεί να μην έχουν τόσο μεγάλη σημασία για τη διάκριση μεταξύ των κειμένων.

Οι Word2Vec και GloVe διαφέρουν σημαντικά από τις παραδοσιακές τεχνικές διανυσματοποίησης κειμένου, όπως τις τεχνικές Bag-of-Words (BoW) και Term Frequency-

Inverse Document Frequency (TF-IDF), τόσο ως προς τη μεθοδολογία όσο και ως προς τον μέγεθος των πληροφοριών που αποτυπώνουν. Ενώ οι BoW και TF-IDF βασίζονται στη συχνότητα των λέξεων σε ένα πλήθος κειμένων, αντιμετωπίζοντας τις λέξεις ως διακριτές μονάδες χωρίς να αποτυπώνουν τις σχέσεις τους, οι Word2Vec και GloVe βασίζονται σε κατανεμημένες αναπαραστάσεις λέξεων που ενσωματώνουν τις σημασιολογικές σχέσεις μεταξύ των λέξεων αυτών.

Το Word2Vec, για παράδειγμα, χρησιμοποιεί μια προσέγγιση βασισμένη σε νευρωνικά δίκτυα για την εκμάθηση διανυσματικών αναπαραστάσεων των λέξεων, έτσι ώστε οι λέξεις που χρησιμοποιούνται σε παρόμοια συμφραζόμενα να αντιστοιχίζονται σε κοντινά σημεία του διανυσματικού χώρου. Υπάρχουν δύο κύριοι τύποι μοντέλων Word2Vec: Continuous Bag-of-Words (CBOW) και Skip-Gram.

Το CBOW προβλέπει μια λέξη με βάση τα συμφραζόμενά της, ενώ το Skip-Gram κάνει το αντίθετο, προβλέποντας λέξεις συμφραζομένων από μια δεδομένη λέξη. Αυτή η μέθοδος συλλαμβάνει όχι μόνο συντακτικές αλλά και σημασιολογικές ομοιότητες, επιτρέποντάς της να αναγνωρίζει ότι λέξεις όπως "βασιλιάς" και "βασίλισσα" σχετίζονται με ουσιαστικό τρόπο πέρα από την απλή συνύπαρξη.

Ο GloVe (Global Vectors for Word Representation), από την άλλη πλευρά, είναι ένας αλγόριθμος μάθησης χωρίς επίβλεψη που παράγει word emdeddings με τη συγκέντρωση παγκόσμιων στατιστικών συνύπαρξης λέξεων από ένα πλήθος κειμένων. Σε αντίθεση με τον Word2Vec, ο οποίος επικεντρώνεται στο τοπικό πλαίσιο, ο GloVe εξετάζει ολόκληρο το σώμα κειμένων, παρέχοντας έτσι μια πιο ολιστική άποψη των σχέσεων των λέξεων. Δημιουργεί διανύσματα με τέτοιο τρόπο ώστε οι γραμμικές σχέσεις στο διανυσματικό χώρο να αποτυπώνουν σημαντικές γλωσσικές σχέσεις, για παράδειγμα, το "άνδρας" - "γυναίκα" είναι περίπου ίσο με το "βασιλιάς" - "βασίλισσα".

Στα πειράματα που διενεργήθηκαν στα πλαίσια της εργασίας αυτής δεν έγινε χρήση του bow,tf-idf παρά κρατήθηκε το vocabulary που είναι με την λογική των συχνών λέξεων δηλαδή αυτές τις λέξεις που δεν εμφανίζονται πολλές φορές τις πετάμε.

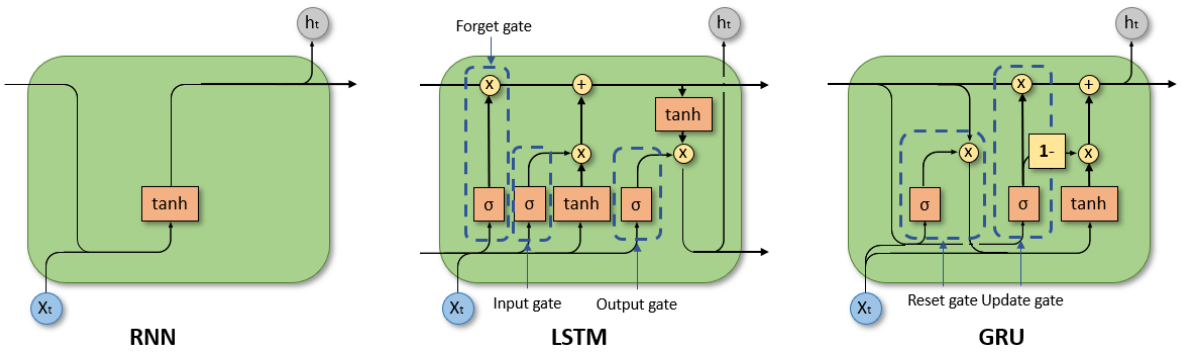
4.7 Μοντέλα Βαθιάς Μάθησης

4.7.1 Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs)

Τα αναδρομικά νευρωνικά δίκτυα (RNN), ιδίως τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM) και τα δίκτυα αναδρομής (GRU), χρησιμοποιούνται ευρέως για την επεξεργασία διαδοχικών δεδομένων, όπως το κείμενο, λόγω της ικανότητάς τους να συλλαμβάνουν πολύπλοκες εξαρτήσεις σε μεγάλες ακολουθίες. Τα LSTM είναι ιδιαίτερα αποτελεσματικά στη διαχείριση των μακροπρόθεσμων εξαρτήσεων και στην αντιμετώπιση του προβλήματος της εξαφανιζόμενης κλίσης (vanishing gradient), γεγονός που τους επιτρέπει να διατηρούν πληροφορίες σε εκτεταμένα αποσπάσματα κειμένου.

Από την άλλη πλευρά, τα GRUs προσφέρουν μια πιο απλοποιημένη αρχιτεκτονική σε σύγκριση με τα LSTM, με λιγότερες πύλες και παραμέτρους, αυξάνοντας έτσι την υπολογιστική αποδοτικότητα, διατηρώντας παράλληλα συγκρίσιμες επιδόσεις. Τόσο τα LSTM όσο και τα GRU είναι ικανά στην κατανόηση πληροφοριών που σχετίζονται με το πλαίσιο και χρησιμοποιούνται συχνά σε εργασίες όπως η ανάλυση συναισθήματος και η ανίχνευση εξαπάτησης.

Η ικανότητά τους να επεξεργάζονται και να ερμηνεύουν διαδοχικά δεδομένα τα καθιστά ιδιαίτερα κατάλληλα για την ανάλυση και τη διάκριση σύνθετων μοτίβων σε κριτικές κειμένου.



Εικόνα 8: Σύγκριση RNN, LSTM και GRU αρχιτεκτονικών τονίζοντας τους διαφορετικούς μηχανισμούς που χρησιμοποιούνται για το χειρισμό διαδοχικών δεδομένων.
Πηγή: https://medium.com/@yash9439/building-multi-layer-gru-from-scratch-305a03670fdd

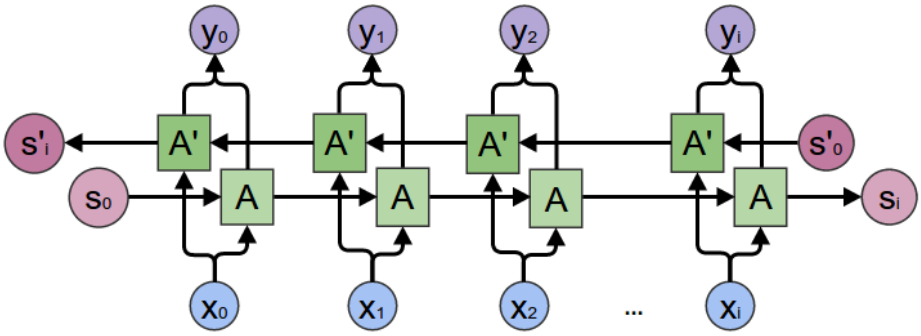
Υλοποιήθηκαν τα προαναφερθέντα μοντέλα RNN χρησιμοποιώντας τη βιβλιοθήκη PyTorch. Το τυπικό RNN αποτελείται από ένα embedding layer για τη μετατροπή των δεικτών εισόδου σε πυκνά διανύσματα, ένα RNN layer για τη σύλληψη των χρονικών εξαρτήσεων και ένα linear layer για την παραγωγή της τελικής εξόδου. Η μέθοδος forward επεξεργάζεται ένα batch ακολουθιών, χρησιμοποιώντας την τελευταία έξοδο RNN για τη δημιουργία προβλέψεων.

Η κλάση μοντέλου LSTM, όπως και το μοντέλο RNN, είναι ένα νευρωνικό δίκτυο για το χειρισμό ακολουθιών. Η κύρια διαφορά έγκειται στη χρήση ενός LSTM layer (Long Short-Term Memory) αντί ενός τυπικού RNN. Τα LSTM είναι καλύτερα στη σύλληψη μακροπρόθεσμων εξαρτήσεων και στη διαχείριση ζητημάτων εξαφανιζόμενης κλίσης (vanishing gradient) σε διαδοχικά δεδομένα. Το μοντέλο περιλαμβάνει επίσης ένα embedding layer και ένα linear layer, με τη μέθοδο forward να χρησιμοποιεί την τελική έξοδο από το LSTM για την παραγωγή προβλέψεων.

Τέλος, το μοντέλο GRU χρησιμοποιεί ένα GRU layer (Gated Recurrent Unit) αντί για ένα RNN. Οι GRU, όπως και οι LSTM, είναι αποτελεσματικές στη σύλληψη των εξαρτήσεων στις ακολουθίες και στον μετριασμό του προβλήματος της εξαφανιζόμενης κλίσης, αλλά είναι υπολογιστικά πιο αποδοτικές λόγω του ότι διαθέτουν λιγότερες πύλες. Το μοντέλο περιλαμβάνει ένα embedding layer για το μετασχηματισμό της εισόδου και ένα linear layer για τη δημιουργία της τελικής εξόδου. Η μέθοδος forward επεξεργάζεται την ακολουθία εισόδου, χρησιμοποιώντας την τελευταία έξοδο GRU για να κάνει προβλέψεις.

4.7.2 Αμφίδρομα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (bi-RNNs)

Τα αμφίδρομα αναδρομικά νευρωνικά δίκτυα (bi-RNNs) επεκτείνουν τις δυνατότητες των παραδοσιακών RNNs, επεξεργαζόμενα ακολουθίες τόσο προς τα εμπρός όσο και προς τα πίσω, ενισχύοντας έτσι την ικανότητά τους να συλλαμβάνουν πληροφορίες σχετικά με το πλαίσιο τόσο από το παρελθόν όσο και από το μέλλον μέσα σε μια ακολουθία. Σε ένα αμφίδρομο RNN, χρησιμοποιούνται δύο ξεχωριστά hidden layers: το ένα επεξεργάζεται την ακολουθία από την αρχή προς το τέλος (κατεύθυνση προς τα εμπρός) και το άλλο την επεξεργάζεται από το τέλος προς την αρχή (κατεύθυνση προς τα πίσω). Αυτή η διπλή προσέγγιση επιτρέπει στο δίκτυο να έχει πρόσβαση σε πληροφορίες από ολόκληρη την ακολουθία σε κάθε χρονικό βήμα, καθιστώντας το ιδιαίτερα αποτελεσματικό για εργασίες όπου το πλαίσιο και από τις δύο κατευθύνσεις είναι σημαντικό, όπως για παράδειγμα στην κατανόηση σύνθετων δομών προτάσεων. Τα αμφίδρομα RNN είναι αποτελεσματικά σε εφαρμογές όπως η μηχανική μετάφραση, η ανάλυση συναισθήματος και η ανίχνευση εξαπάτησης, όπου η κατανόηση του πλήρους πλαισίου μιας αναθεώρησης ή ενός αποσπάσματος κειμένου μπορεί να βελτιώσει σημαντικά την απόδοση και την ακρίβεια του μοντέλου.



Εικόνα 9: *Δομή ενός αμφίδρομου επαναλαμβανόμενου δικτύου.*
Πηγή: https://www.researchgate.net/publication/339757609_A_Spontaneous_Visible_and_Thermal_Facial_Expression_of_Human_Emotion_Database

4.7.3 Μοντέλα που Βασίζονται στην Προσοχή (Attention-based models)

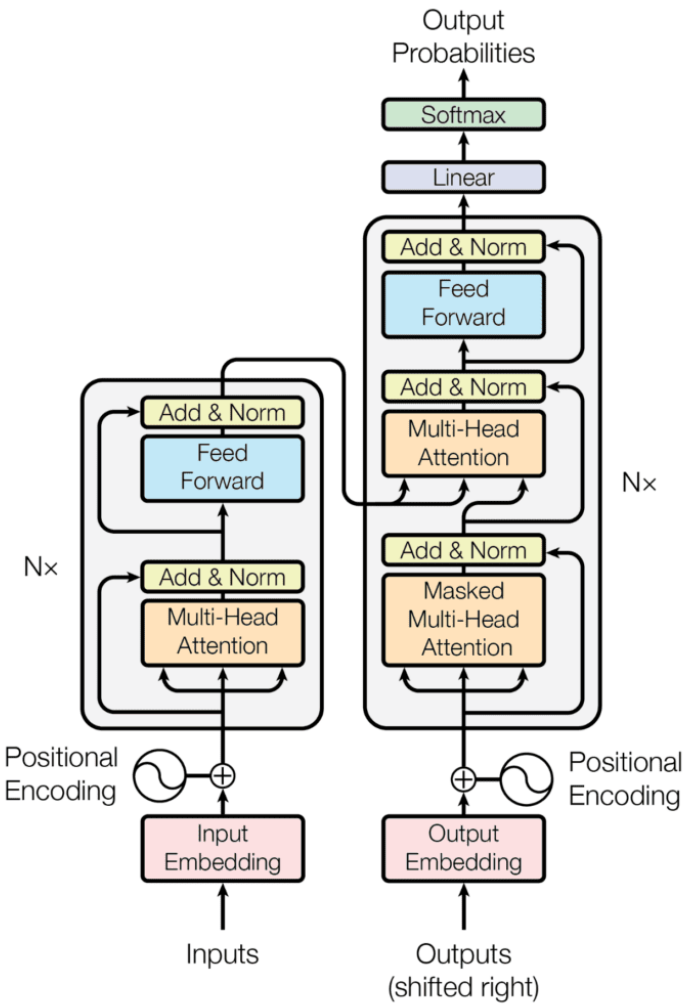
Τα μοντέλα που βασίζονται στην προσοχή (attention) είναι μια κατηγορία νευρωνικών δικτύων που έχουν αποκτήσει σημαντική θέση στους τομείς της επεξεργασίας φυσικής γλώσσας (NLP), της υπολογιστικής όρασης και άλλων περιοχών της μηχανικής μάθησης. Αυτά τα μοντέλα χρησιμοποιούν έναν μηχανισμό που ονομάζεται "attention" για να εστιάσουν σε συγκεκριμένα μέρη των δεδομένων εισόδου όταν κάνουν προβλέψεις. Αυτή η έννοια έχει ιδιαίτερα σημαντική επίδραση στη βελτίωση της απόδοσης και της ερμηνευσιμότητας των μοντέλων.

Ο μηχανισμός attention είναι σχεδιασμένος να αναδεικνύει δυναμικά τα σημαντικά μέρη των δεδομένων εισόδου. Εκχωρεί διαφορετικά βάρη (attention scores) σε διαφορετικά μέρη της εισόδου, επιτρέποντας στο μοντέλο να εστιάσει σε σχετικές πληροφορίες ενώ αγνοεί τις λιγότερο σημαντικές λεπτομέρειες. Το self-attention ή intra-attention αναφέρεται στην εφαρμογή του μηχανισμού attention μέσα σε μια μοναδική ακολουθία. Επιτρέπει στο μοντέλο να εξετάζει τις σχέσεις μεταξύ όλων των στοιχείων της ακολουθίας ταυτόχρονα. Αυτό είναι ιδιαίτερα χρήσιμο σε προβλήματα επεξεργασίας φυσικής γλώσσας, όπου η κατανόηση του πλαισίου των λέξεων σε σχέση με τις άλλες λέξεις που την περιβάλλουν είναι κρίσιμη.

Το μοντέλο Transformers, που παρουσιάστηκε από τους Vaswani et al. το 2017, αποτελεί ορόσημο στα μοντέλα που βασίζονται στην προσοχή. Βασίζεται εξ ολοκλήρου στους μηχανισμούς self-attention, παρακάμπτοντας τα παραδοσιακά αναδρομικά ή συνελικτικά

νευρωνικά δίκτυα. Η αρχιτεκτονική του Transformer αποτελείται από στρώματα multi-head self-attention και feed-forward neural networks, καθιστώντας αποδοτικό καθώς υποστηρίζει παράλληλη επεξεργασία. Η multi-head attention επιτρέπει στο μοντέλο να εστιάζει ταυτόχρονα σε πληροφορίες από διαφορετικούς υποχώρους αναπαράστασης σε διαφορετικές θέσεις. Αυτό σημαίνει ότι το μοντέλο μπορεί να συλλάβει διαφορετικούς τύπους σχέσεων μεταξύ των λέξεων ή των στοιχείων της ακολουθίας εισόδου.

Στο τομέα του NLP, τα μοντέλα που βασίζονται στην προσοχή χρησιμοποιούνται συχνά σε ένα πλαίσιο encoder-decoder. Ο encoder επεξεργάζεται την ακολουθία εισόδου και ο decoder παράγει την ακολουθία εξόδου, με τον μηχανισμό attention να βοηθά τον decoder να εστιάσει στα σχετικά μέρη της εξόδου του encoder σε κάθε βήμα.



Εικόνα 10: Απεικόνιση της αρχιτεκτονικής του μοντέλου Transformer, όπου παρουσιάζεται η δομή κωδικοποιητή-αποκωδικοποιητή με πολλαπλές κεφαλές προσοχής και στρώματα προώθησης. *Πηγή: https://medium.com/@mekarahul/what-are-self-attention-models-69fb59f6b5f8*

Στην υλοποίηση αναπτύχθηκε ο μηχανισμός αυτοπροσοχής, όπως παρουσιάστηκε στην αρχική εργασία του Transformer "Attention Is All You Need", και στη συνέχεια χρησιμοποιήθηκε αυτός ο μηχανισμός για να κατασκευαστεί ένα μοντέλο Transformer. Δημιουργήθηκαν δύο ξεχωριστές κλάσεις python: η SelfAttention η οποία ορίζεται ως ένα προσαρμοσμένο επίπεδο που ενσωματώνει τον μηχανισμό αυτοπροσοχής και το TransformerModel.

4.7.4 Μοντέλα BERT και RoBERTa (pretrained models)

BERT (Αναπαραστάσεις κωδικοποιητή διπλής κατεύθυνσης από Transformers)

Το BERT, που παρουσιάστηκε από τους Devlin et al. το 2018, είναι ένα πρωτοποριακό μοντέλο στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP). Έχει σχεδιαστεί για να προ-εκπαιδεύει βαθιές αμφίδρομες αναπαραστάσεις με κοινή κλιμάκωση τόσο του αριστερού όσο και του δεξιού πλαισίου σε όλα τα επίπεδα. Αυτό επιτρέπει στο BERT να κατανοεί το πλαίσιο μιας λέξης με βάση τις λέξεις που την περιβάλλουν, καθιστώντας το ιδιαίτερα αποτελεσματικό για διάφορες εργασίες NLP, όπως η απάντηση ερωτήσεων και η εξαγωγή συμπερασμάτων για τη γλώσσα.

Βασικά χαρακτηριστικά της BERT:

- Αμφίδρομη εκπαίδευση: Σε αντίθεση με τα προηγούμενα μοντέλα που διαβάζουν την είσοδο κειμένου διαδοχικά (από αριστερά προς τα δεξιά ή από δεξιά προς τα αριστερά), το BERT διαβάζει ολόκληρη την ακολουθία των λέξεων ταυτόχρονα, επιτρέποντάς του να κατανοεί το πλαίσιο πιο αποτελεσματικά.
- Masked Language Model (MLM): Κατά τη διάρκεια της εκπαίδευσης, κάποιο ποσοστό των σημείων εισόδου αποκρύπτονται τυχαία και το μοντέλο εκπαιδεύεται για να προβλέπει αυτά τα αποκρυπτόμενα σημεία. Αυτό βοηθά το μοντέλο να μάθει βαθιές αμφίδρομες αναπαραστάσεις.
- Πρόβλεψη επόμενης πρότασης (NSP): Το BERT εκπαιδεύεται επίσης σε μια εργασία για να προβλέψει αν μια δεδομένη πρόταση είναι η επόμενη πρόταση στο αρχικό κείμενο, η οποία βοηθά στην κατανόηση της σχέσης μεταξύ των προτάσεων.

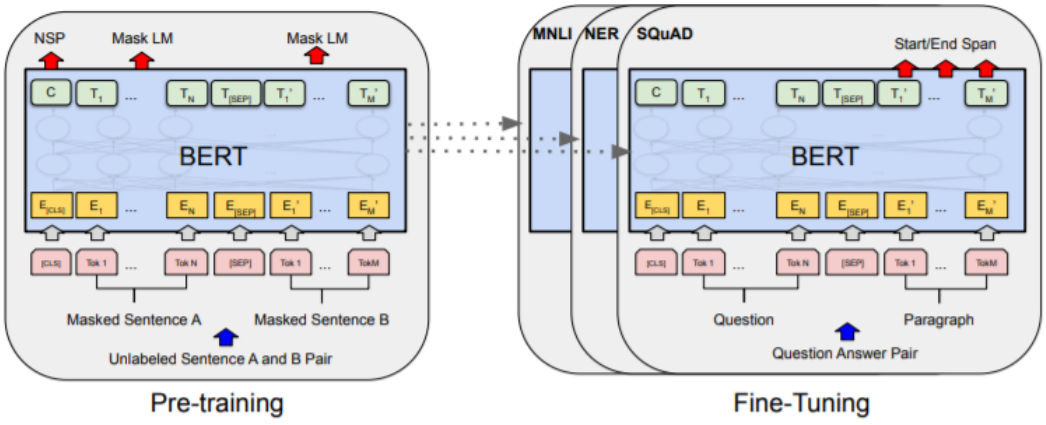
RoBERTa (Robustly Optimized BERT Pretraining Approach)

Η RoBERTa, που παρουσιάστηκε από τους Liu et al. το 2019, είναι μια βελτιστοποιημένη έκδοση της BERT. Οι συγγραφείς διαπίστωσαν ότι η BERT ήταν σημαντικά υποεκπαιδευμένη και πρότειναν διάφορες τροποποιήσεις για τη βελτίωση της απόδοσής της.

Οι βασικές βελτιώσεις της RoBERTa:

- Μεγαλύτερη εκπαίδευση: Το RoBERTa εκπαιδεύεται για μεγαλύτερο χρονικό διάστημα με μεγαλύτερες παρτίδες και περισσότερα δεδομένα, γεγονός που συμβάλλει στην επίτευξη καλύτερης απόδοσης.
- Αφαίρεση του NSP: Ο στόχος πρόβλεψης επόμενης πρότασης αφαιρείται, καθώς διαπιστώθηκε ότι είναι περιττός για τη βελτίωση της απόδοσης της επόμενης εργασίας.
- Δυναμική κάλυψη: Αντί για στατική απόκρυψη, το RoBERTa χρησιμοποιεί δυναμική απόκρυψη, όπου το μοτίβο απόκρυψης αλλάζει κατά τη διάρκεια της εκπαίδευσης, καθιστώντας το μοντέλο πιο ανθεκτικό.
- Μεγαλύτερα σύνολα δεδομένων: Το RoBERTa εκπαιδεύεται σε μεγαλύτερα σύνολα δεδομένων, συμπεριλαμβανομένου ενός νέου συνόλου δεδομένων που ονομάζεται CC-News, το οποίο είναι συγκρίσιμο σε μέγεθος με άλλα μεγάλα, ιδιωτικά χρησιμοποιούμενα σύνολα δεδομένων.

Τόσο το BERT όσο και το RoBERTa έχουν προωθήσει σημαντικά τον τομέα του NLP, με το RoBERTa να βασίζεται στα πλεονεκτήματα του BERT και να αντιμετωπίζει ορισμένους από τους περιορισμούς του για να επιτύχει ακόμη καλύτερες επιδόσεις σε διάφορα benchmarks.



Εικόνα 11: Απεικόνιση της αρχιτεκτονικής του μοντέλου BERT

Πηγή: https://medium.com/@marketing_novita.ai/introducing-roberta-base-model-a-comprehensive-overview-330338afa082

Η εικόνα 11 απεικονίζει τη ροή εργασίας του μοντέλου BERT, παρουσιάζοντας δύο βασικά στάδια: την προ-εκπαίδευση και τη λεπτομέρεια.

Στη φάση της προεκπαίδευσης, ο BERT εκπαιδεύεται σε δύο βασικές εργασίες: Masked Language Modeling (MLM), όπου οι τυχαίες λέξεις σε προτάσεις καλύπτονται και το μοντέλο μαθαίνει να τις προβλέπει, και το μοντέλο Next Sentence Prediction (NSP), όπου το μοντέλο προβλέπει εάν δύο προτάσεις διαδέχονται η μία την άλλη. Η είσοδος αποτελείται από ενσωματώσεις διακριτικών για προτάσεις, οι οποίες περιλαμβάνουν ενσωματώσεις θέσης και τμήματος.

Μετά την προεκπαίδευση, το BERT ρυθμίζεται με ακρίβεια σε συγκεκριμένες εργασίες όπως Masked natural language inference (MNLI), αναγνώριση ονομαστικής οντότητας (NER) και SQuAD (απάντηση ερωτήσεων), όπου εκπαιδεύεται περαιτέρω σε δεδομένα με ετικέτα για συγκεκριμένη εργασία. Το Fine-tuning προσαρμόζει τα προεκπαιδευμένα βάρη του BERT στις συγκεκριμένες εργασίες, προσαρμόζοντας τις εξόδους όπως διαστήματα για ζεύγη ερωτήσεων-απαντήσεων στο SQuAD.

4.8 Βέλτιστα αποτελέσματα με βάση τη βιβλιογραφία για όλα τα μοντέλα

Τα αποτελέσματα μπορεί να διαφέρουν ανάλογα με διάφορους παράγοντες, όπως η συγκεκριμένη εργασία (π.χ. ανάλυση συναισθήματος, ανίχνευση ανεπιθύμητης αλληλογραφίας), η προεπεξεργασία του συνόλου δεδομένων, ο συντονισμός των υπερπαραμέτρων και η υλοποίηση των μοντέλων. Ωστόσο, μπορώ να δώσω μια γενική επισκόπηση με βάση τη βιβλιογραφία αναφέροντας αρχικά πιο κάτω τα ποσοστά ακρίβειας που πετυχαίνονται στους μετρήσιμους δείκτες των αποτελεσμάτων για κάθε Dataset:

1. “Amazon Reviews” Dataset

- RNN: Περίπου 85-87%.
- BiRNN: Περίπου 88-90%
- LSTM: Περίπου 90-92%
- GRU: Περίπου 91-93%
- Transformer: Περίπου 94-95%
- BERT: Περίπου 95-96%
- RoBERTa: Περίπου 96-97%

2. “Op Spam” Dataset

- RNN: Περίπου 80-82%
- BiRNN: Περίπου 82-84%
- LSTM: Περίπου 84-86%
- GRU: Περίπου 85-87%
- Transformer: Περίπου 88-90%
- BERT: Περίπου 90-92%
- RoBERTa: Περίπου 92-94%

3. “De Rev” Dataset

- RNN: Περίπου 78-80%
- BiRNN: Περίπου 80-82%
- LSTM: Περίπου 82-84%
- GRU: Περίπου 83-85%
- Transformer: Περίπου 86-88%
- BERT: Περίπου 88-90%
- RoBERTa: Περίπου 90-92%

4. “Cross-Cultural” Dataset

- RNN: Περίπου 75-77%
- BiRNN: Περίπου 77-79%
- LSTM: Περίπου 79-81%
- GRU: Περίπου 80-82%
- Transformer: Περίπου 83-85%
- BERT: Περίπου 85-87%
- RoBERTa: Περίπου 87-89%

5. “Restaurant Reviews” Dataset

- RNN: 82-84% περίπου
- BiRNN: Περίπου 84-86%
- LSTM: Περίπου 86-88%
- GRU: Περίπου 87-89%
- Transformer: Περίπου 90-92%
- BERT: Περίπου 92-94%
- RoBERTa: Περίπου 94-96%

Αυτά τα ποσοστά είναι κατά προσέγγιση και μπορεί να διαφέρουν ανάλογα με τις συγκεκριμένες εκδόσεις συνόλων δεδομένων, τα βήματα προεπεξεργασίας και τις διαμορφώσεις μοντέλων που χρησιμοποιούνται σε διάφορες μελέτες.

Μερικά στοιχεία που δείχνουν την επιτυχία κάθε μοντέλου σε διάφορες περιπτώσεις έχουν ως εξής:

α. Οι RNN και οι BiRNN έχουν γενικά χαμηλότερες επιδόσεις λόγω προβλημάτων εξαφανιζόμενης κλίσης και περιορισμένης χωρητικότητας μνήμης, γεγονός που τις καθιστά λιγότερο αποτελεσματικές σε μεγάλες ακολουθίες.

β. Τα LSTM και GRU βελτιώνουν τα RNNs χειριζόμενα καλύτερα τις μακροπρόθεσμες εξαρτήσεις, οδηγώντας σε ελαφρώς υψηλότερη ακρίβεια.

γ. Οι Transformers υπερτερούν των LSTM και GRU στις περισσότερες περιπτώσεις λόγω της ικανότητάς τους να συλλαμβάνουν τις παγκόσμιες εξαρτήσεις χωρίς να βασίζονται σε διαδοχική επεξεργασία δεδομένων.

δ. Οι BERT και RoBERTa (που είναι μια βελτιστοποιημένη έκδοση του BERT) επιτυγχάνουν γενικά την υψηλότερη ακρίβεια, καθώς έχουν προ-εκπαιδευτεί σε τεράστιες ποσότητες δεδομένων και έχουν ρυθμιστεί λεπτομερώς για συγκεκριμένες εργασίες.

4.9 Ρύθμιση μεταξύ τομέων (Cross-domain Set-up)

Στο πλαίσιο της ανίχνευσης εξαπάτησης, η διασφάλιση ότι τα μοντέλα μας γενικεύονται αποτελεσματικά σε διάφορους τομείς είναι ζωτικής σημασίας για την ευρωστία και την αξιοπιστία τους. Για να το αντιμετωπίσουμε αυτό, επινοήσαμε μια νέα πειραματική πρόταση που συνδυάζει όλα τα διαθέσιμα σύνολα δεδομένων σε ένα ενιαίο, ενοποιημένο σύνολο δεδομένων. Αυτή η προσέγγιση συνδυασμένου συνόλου δεδομένων αξιοποιεί τις αρχές της διασταυρούμενης επικύρωσης για την αυστηρή αξιολόγηση της απόδοσης των μοντέλων σε διαφορετικούς τομείς. Συγκεκριμένα, εφαρμόσαμε μια στρατηγική πενταπλής διασταυρούμενης επικύρωσης όπου, σε κάθε επανάληψη, τέσσερα από τα πέντε σύνολα δεδομένων χρησιμοποιούνται για εκπαίδευση, ενώ το πέμπτο σύνολο δεδομένων προορίζεται για δοκιμή. Αυτή η προσέγγιση διασφαλίζει ότι κάθε σύνολο δεδομένων χρησιμοποιείται ως σύνολο δοκιμής ακριβώς μία φορά, παρέχοντας μια ολοκληρωμένη αξιολόγηση της ικανότητας του μοντέλου να γενικεύει σε άγνωστα δεδομένα και βελτιώνοντας την προσαρμοστικότητά του σε διαφορετικά πλαίσια ανίχνευσης εξαπάτησης. Χρησιμοποιώντας αυτό το πλαίσιο διασταυρούμενης επικύρωσης πολλαπλών συνόλων δεδομένων, ενισχύουμε την ανθεκτικότητα του μοντέλου και επικυρώνουμε την απόδοσή του με έναν πιο ολιστικό και γενικευμένο τρόπο.

Κεφάλαιο 5

Πειραματικά αποτελέσματα

5.1 Πλαίσιο Πειραμάτων

Στην παρούσα Διπλωματική εργασία, προτείνουμε ένα πειραματικό πλαίσιο σχεδιασμένο για την αξιολόγηση της απόδοσης διαφόρων μοντέλων μηχανικής μάθησης στην ανίχνευση απάτης (deception detection) σε κείμενα. Στο προτεινόμενο πειραματικό πλαίσιο ενσωματώνεται η προ-επεξεργασία δεδομένων, η ανάπτυξη αρχιτεκτονικών και εκπαίδευση των μοντέλων και η αξιολόγηση της απόδοσης, επιτρέποντας τη σύγκριση διαφορετικών μοντέλων με βάση την αποτελεσματικότητά τους στην αναγνώριση απάτης σε δεδομένα κειμένου.

Το πλαίσιο περιλαμβάνει τη χρήση διαφόρων μοντέλων, όπως Recurrent Neural Networks (RNNs) και αρχιτεκτονικές Attention-pretrained models. Συγκεκριμένα, εξετάστηκαν τα παρακάτω μοντέλα των οποίων οι δυνατότητες και οι περιορισμοί τους αναλύθηκαν στο προηγούμενο κεφάλαιο:

- Standard RNN
- Bidirectional RNN (BiRNN)
- Long Short-Term Memory (LSTM)
- Gated Recurrent Unit (GRU)
- Transformer (Self-Attention)
- Bidirectional Encoder Representations from Transformers (BERT)
- A Robustly Optimized BERT Pretraining Approach (RoBERTa)

Κάθε μοντέλο εκπαιδεύτηκε και αξιολογήθηκε σε πέντε σύνολα δεδομένων κειμένου, προσφέροντας σημαντικές γνώσεις σχετικά με την απόδοση των μοντέλων σε διαφορετικούς τομείς και διαφορετικά περιβάλλοντα ανίχνευσης απάτης.

5.2 Επεξεργασία και Προεπεξεργασία Δεδομένων

Τα σύνολα δεδομένων που χρησιμοποιήθηκαν για την ανίχνευση απάτης υποβλήθηκαν σε επεξεργασία ώστε να διασφαλιστεί η συνέπεια και η συμβατότητα με τα μοντέλα. Η προεπεξεργασία περιλάμβανε τα εξής βήματα:

- Lowercasing όλων των κειμένων.
- Αφαίρεση stopwords και σημείων στίξης, χρησιμοποιώντας τη βιβλιοθήκη NLTK.
- Tokenization των κειμένων σε tokens, χρησιμοποιώντας έναν απλό αγγλικό tokenizer.
- Padding ή Truncation των κειμένων για την διατήρηση σταθερού μήκους, διασφαλίζοντας ομοιομορφία στα input layers των μοντέλων.

5.3 Αρχιτεκτονικές Μοντέλων

- **Standard RNN**: Το Simple RNN επεξεργάζεται ακολουθίες λέξεων σειριακά και χρησιμοποιεί την τελική κρυφή κατάσταση για την ταξινόμηση.
- **Bidirectional RNN** (BiRNN): Το BiRNN λαμβάνει υπόψη τόσο το παρελθόν όσο και το μέλλον της ακολουθίας, επιτρέποντας στο μοντέλο να μάθει από το πλήρες περιεχόμενο.
- **LSTM:** Τα LSTMs επιλύουν το πρόβλημα vanishing gradients στα παραδοσιακά RNNs, διατηρώντας πληροφορίες για μακρές ακολουθίες.
- **GRU**: Τα GRUs είναι μια απλοποιημένη παραλλαγή των LSTMs με λιγότερες πύλες, διευκολύνοντας τη διαδικασία μάθησης.
- **Transformer**: Το Transformer χρησιμοποιεί Self-Attention για την κατανόηση μακροχρόνιων εξαρτήσεων. Σε αντίθεση με τα RNNs, επιτρέπει την επεξεργασία ολόκληρης της ακολουθίας παράλληλα, οδηγώντας σε πιο αποδοτική εκμάθηση.
- **BERT**: Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα μοντέλο βασισμένο στους Transformers που χρησιμοποιεί αμφίδρομη προσοχή (bidirectional attention) για να κατανοήσει το πλήρες πλαίσιο μιας πρότασης, λαμβάνοντας υπόψη τόσο τις προηγούμενες όσο και τις επόμενες λέξεις. Αυτό επιτρέπει στο BERT να επιτυγχάνει εξαιρετικές επιδόσεις σε πολλές εργασίες φυσικής γλώσσας.

- **RoBERTa**: Το RoBERTa (A Robustly Optimized BERT Pretraining Approach) αποτελεί μια βελτιστοποιημένη εκδοχή του BERT. Βασίζεται στην ίδια αρχιτεκτονική με τον BERT, αλλά υλοποιεί πιο προσεκτική εκπαίδευση με μεγαλύτερα batches, περισσότερη εισαγωγή δεδομένων και διαφορετική επεξεργασία της μάσκας. Αυτό καθιστά το RoBERTa ακόμα πιο ισχυρό σε διάφορες εργασίες επεξεργασίας φυσικής γλώσσας.

Κάθε μοντέλο υλοποιήθηκε με χρήση της βιβλιοθήκης PyTorch, προσαρμόζοντας τις παραμέτρους όπως το embedding size, τον αριθμό των hidden states και τα output layers, ώστε να ταιριάζουν στις απαιτήσεις κάθε αρχιτεκτονικής. Για τα RNN, BiRNN, LSTM και GRU χρησιμοποιήθηκαν κρυφές καταστάσεις για την αναπαράσταση των ακολουθιών, ενώ για τα Transformers, BERT και RoBERTa εφαρμόστηκαν layers self-attention και πιο εξελιγμένοι μηχανισμοί ενσωμάτωσης (embedding) για την κατανόηση μακροχρόνιων εξαρτήσεων και το πλήρες πλαίσιο της πρότασης. Οι μεταβλητές, όπως το learning rate και το batch size, ρυθμίστηκαν ανάλογα με τις απαιτήσεις κάθε μοντέλου για βέλτιστη απόδοση.

5.4 Εκπαίδευση και Αξιολόγηση

Τα μοντέλα εκπαιδεύτηκαν αρχικά στο πείραμα με preprocessing, για το μέγιστο αριθμό των 15 επαναλήψεων εκπαίδευσης (EPOCHS), χρησιμοποιώντας Cross-Entropy Loss και τον Adam optimizer. Εφαρμόστηκε η λογική του early stopping, τερματίζοντας την εκπαίδευση εάν η απόδοση στο validation set δεν βελτιωνόταν μετά από προκαθορισμένες εποχές (με patience 3 εποχών). Το 5% των δεδομένων εκπαίδευσης χρησιμοποιήθηκε για το validation των μοντέλων στα πλαίσια της εκπαίδευσης. Στη συνέχεια επαναλήφθηκε το πείραμα χωρίς preprocessing αλλάζοντας τον αριθμό επαναλήψεων εκπαίδευσης, EPOCHS=25 και χρησιμοποιώντας το 80% των δεδομένων για training set από το οποίο το 95% (δηλαδή το 76% του συνολικού συνόλου) είναι το τελικό training set και το 5% (δηλαδή το 4% του συνολικού συνόλου) χρησιμοποιείται ως validation set. Το υπόλοιπο 20% των δεδομένων αποτελεί το test set.

Για την αξιολόγηση των μοντέλων υπολογίστηκαν metrics όπως Accuracy, Balanced Accuracy, Precision, Recall και F1-score ενώ για την αναλυτική σύγκριση μεταξύ των κλάσεων (truthful, deceptive) αποτυπώνονται αναλυτικά και τα Classification Reports και Confusion Matrices. Η απόδοση των μοντέλων σε κάθε μια από τις παραπάνω μετρικές μετρήθηκε ξεχωριστά στα

πέντε διαφορετικά datasets, ενώ καταγράφηκαν επίσης ο αριθμός των παραμέτρων (βάρη που γίνονται update κατά την εκπαίδευση) και ο χρόνος εκπαίδευσης.

5.5 Οπτικοποίηση και Συγκριτική Ανάλυση

Τα αποτελέσματα οπτικοποιήθηκαν σε barcharts για να διευκολύνουν τη σύγκριση των μοντέλων. Για κάθε πείραμα, παρουσιάστηκαν διαγράμματα για την ακρίβεια, τον χρόνο εκπαίδευσης και τον αριθμό παραμέτρων προς εκπαίδευση που απαιτεί το κάθε μοντέλο.

5.6 Εφαρμογή πειραμάτων

5.6.1 Ρύθμιση Εντός Τομέα (In-domain setup -Vocabulary & Default Parameters) με Preprocessing

Στο πρώτο setup δημιουργούμε ένα vocabulary από το εκάστοτε υπό εξέταση σύνολο δεδομένων με ένα ελάχιστο όριο συχνότητας εμφάνισης λέξεων. Οι σπάνιες λέξεις αντικαταστάθηκαν με ένα "unknown" token, ενώ προστίθενται και padding tokens για τη διατήρηση σταθερού μήκους εισόδου.
Στην εκπαίδευση των μοντέλων μηχανικής μάθησης, χρησιμοποιήθηκαν οι εξής global υπερπαράμετροι:

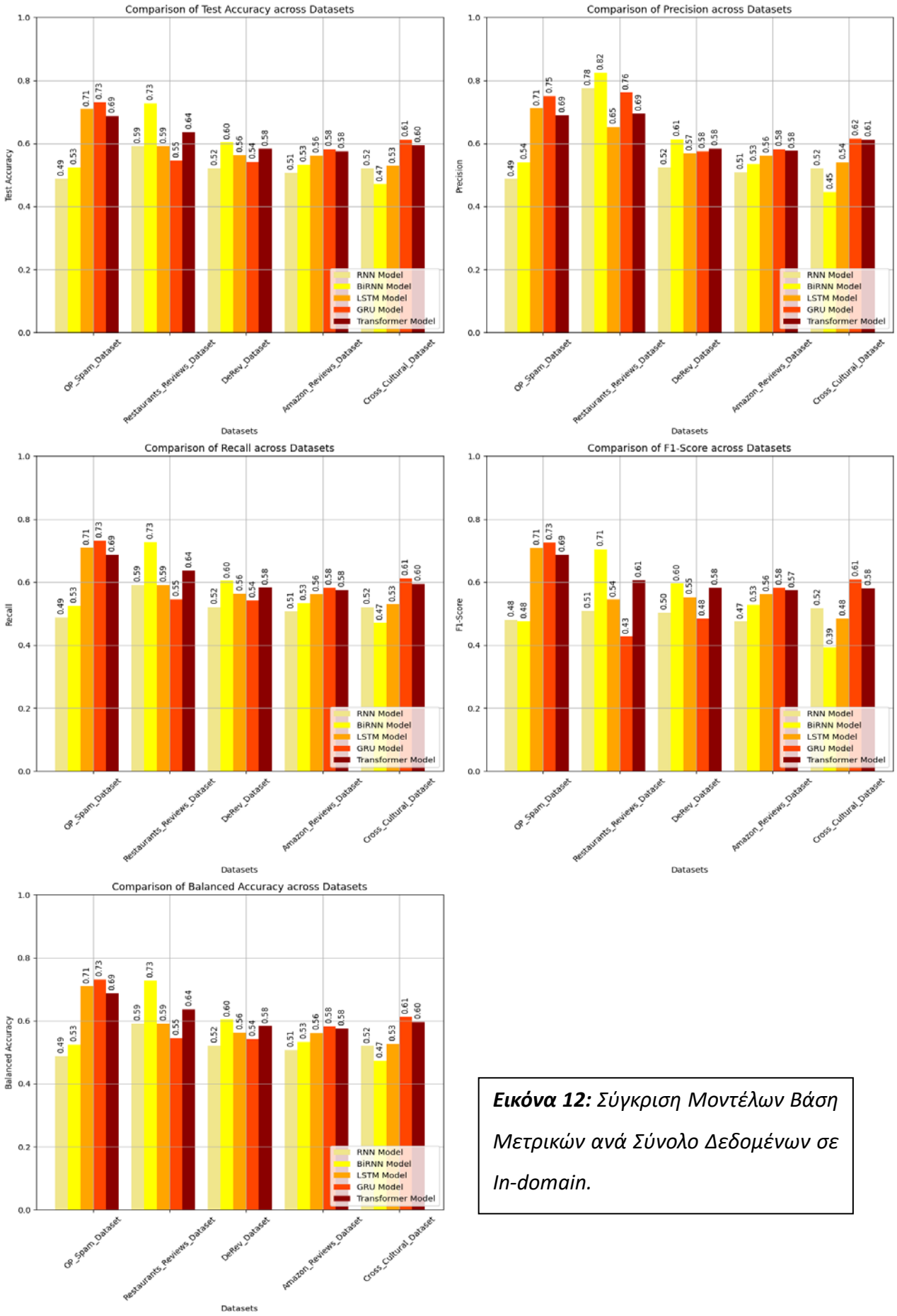
- MAX_WORDS = 25 καθορίζει το μέγιστο μήκος της ακολουθίας κειμένων, επιτρέποντας την ανάλυση έως 25 λέξεων ανά κείμενο.
- EPOCHS = 15 ορίζει τον αριθμό των επαναλήψεων εκπαίδευσης πάνω από το σύνολο δεδομένων, επιτρέποντας στο μοντέλο να βελτιωθεί σε 15 κύκλους.
- LEARNING_RATE = 1e-3 προσδιορίζει τον ρυθμό μάθησης, επηρεάζοντας τη ταχύτητα και τη σταθερότητα της εκπαίδευσης.
- BATCH_SIZE = 32 καθορίζει τον αριθμό των δειγμάτων δεδομένων που χρησιμοποιούνται σε κάθε επανάληψη για την ενημέρωση των βαρών του μοντέλου.
- EMBEDDING_DIM = 100 αναφέρεται στη διάσταση του διανύσματος αναπαράστασης λέξεων, επιτρέποντας πλουσιότερες γλωσσικές αναπαραστάσεις.
- HIDDEN_DIM = 64 καθορίζει τον αριθμό των νευρώνων στο κρυφό επίπεδο των RNNs, LSTMs ή GRUs, επηρεάζοντας την ικανότητα του μοντέλου να αποθηκεύει πληροφορίες.

- NUM_HEADS = 5 σχετίζεται με τον αριθμό των κεφαλών στον μηχανισμό προσοχής, ενισχύοντας τη δυνατότητα του μοντέλου να αποτυπώσει σύνθετες σχέσεις.

Τα αποτελέσματα ομαδοποιημένα ανά σετ δεδομένων για κάθε μοντέλο ανάλυσης φαίνονται στον παρακάτω πίνακα 1 ενώ για καλύτερη σύγκριση και πιο εποπτική παρουσίαση έχουν σχεδιαστεί και τα αντίστοιχα διαγράμματα δίνοντας τα αποτελέσματα που πετυχαίνουν τα γλωσσικά μοντέλα ανά μετρική και ανά σετ δεδομένων στην εικόνα 12

Dataset	*Model*	*Params*	*Training Time*	*Test Accuracy*	*Precision*	*Recall*	*F1*	*Balanced Accuracy*
Amazon Reviews	RNN	678254	1.97	0.51	0.51	0.51	0.47	0.51
	BiRNN	689006	2.10	0.53	0.53	0.53	0.53	0.53
	LSTM	710126	2.78	0.56	0.56	0.56	0.56	0.56
	GRU	699502	3.18	0.58	0.58	0.58	0.58	0.58
	Transforme	714494	2.88	0.58	0.58	0.58	0.57	0.58
OP Spam	RNN	172954	0.38	0.49	0.49	0.49	0.48	0.49
	BiRNN	183706	0.36	0.53	0.54	0.53	0.48	0.53
	LSTM	204826	0.36	0.71	0.71	0.71	0.71	0.71
	GRU	194202	0.24	0.73	0.75	0.73	0.73	0.73
	Transforme	209194	0.29	0.69	0.69	0.69	0.69	0.69
DeRev	RNN	38754	0.08	0.52	0.52	0.52	0.50	0.52
	BiRNN	49506	0.05	0.60	0.61	0.60	0.60	0.60
	LSTM	70626	0.06	0.56	0.57	0.56	0.55	0.56
	GRU	60002	0.05	0.54	0.58	0.54	0.48	0.54
	Transforme	74994	0.07	0.58	0.58	0.58	0.58	0.58
Cross Cultural	RNN	47554	0.07	0.52	0.52	0.52	0.52	0.52
	BiRNN	58306	0.08	0.47	0.45	0.47	0.39	0.47
	LSTM	79426	0.09	0.53	0.54	0.53	0.48	0.53
	GRU	68802	0.09	0.61	0.62	0.61	0.61	0.61
	Transforme	83794	0.11	0.60	0.61	0.60	0.58	0.60
Restaurants Reviews	RNN	19154	0.03	0.59	0.78	0.59	0.51	0.59
	BiRNN	29906	0.04	0.73	0.82	0.73	0.71	0.73
	LSTM	51026	0.02	0.59	0.65	0.59	0.54	0.59
	GRU	40402	0.03	0.55	0.76	0.55	0.43	0.55
	Transforme	55394	0.03	0.64	0.69	0.64	0.61	0.64

Πίνακας 1: *Μετρήσεις Απόδοσης ανά Dataset σε In-domain.*



Εικόνα 12: *Σύγκριση Μοντέλων Βάση Μετρικών ανά Σύνολο Δεδομένων σε In-domain.*

Για το Amazon Reviews dataset, φαίνεται πως όσο αυξάνουμε την πολυπλοκότητα του μοντέλου, αυξάνεται και η επίδοση στις εξεταζόμενες μετρικές. Συγκεκριμένα το standard RNN μοντέλο, που έχει και το μικρότερο πλήθος παράμετρών προς εκπαίδευση πετυχαίνεται το μικρότερο accuracy (0.51) και F1-score (0.47), γεγονός που αποδεικνύει ότι δυσκολεύεται να εντοπίσει λεκτικά pattern στο dataset. Από την άλλη, όσο αυξάνεται η πολυπλοκότητα του μοντέλου, όπως με τα LSTM, GRU και Tranformers, υπάρχει παράλληλα και αισθητή βελτίωση στις μετρικές, όπως για παράδειγμα στο F1 που peak-άρει στο 0.58, σχεδόν δέκα ποσοστιαίες μονάδες πάνω από τις baseline προσεγγίσεις. Συνολικά, GRU και Tranformers υπερτερούν σε ακρίβεια από τα υπόλοιπα μοντέλα. Σε γενικές γραμμές όμως η απόδοση είναι σχετικά χαμηλή γεγονός που μπορεί να επηρεάζεται από την ποικιλομορφία και του συγκεκριμένου dataset. Ίσως κάποιες τροποποιήσεις στις παραμέτρους των πιο αποδοτικών μοντέλων να μας επέτρεπε να συλλάβουμε περισσότερα Long-term dependencies και να μας έδινε ακόμα καλύτερα αποτελέσματα.

Στο OP Spam dataset, τα αποτελέσματα δείχνουν ένα σαφές πλεονέκτημα για μοντέλα όπως τα LSTM και GRU, τα οποία επιτυγχάνουν σημαντικά υψηλότερες μετρήσεις απόδοσης σε σύγκριση με τα μοντέλα RNN και BiRNN. Ειδικότερα, το μοντέλο GRU ξεχωρίζει με το υψηλότερο accuracy (0,73), precision (0,75) και F1 score (0,73), ενώ έχει και τον μικρότερο χρόνο εκπαίδευσης. Αυτό υποδηλώνει ότι το μοντέλο GRU είναι ιδιαίτερα αποτελεσματικό για αυτό το σύνολο δεδομένων, και μπορεί να εντοπίζει deceptive μοτίβα στο κείμενο. Το Transformer, αν και θεωρητικά το πιο ισχυρό μοντέλο, υστερεί ελαφρώς έναντι του GRU, υποδεικνύοντας ότι για ορισμένα μικρότερα ή πιο εξειδικευμένα σύνολα δεδομένων, απλούστερα μοντέλα όπως το GRU μπορούν να υπερτερούν έναντι των πιο σύνθετων.

Για το σύνολο δεδομένων DeRev, το μοντέλο BiRNN παρέχει την υψηλότερη τιμή σε accuracy και F1 score, ξεπερνώντας ακόμη και το μοντέλο Transformer. Αυτό προκαλεί κάποια έκπληξη, δεδομένου ότι οι Transformers συνήθως υπερέχουν σε πολλά NLP προβλήματα στη διεθνή βιβλιογραφία. Τα μοντέλα RNN και GRU, από την άλλη πλευρά, παρουσιάζουν ασθενέστερες επιδόσεις, γεγονός που μπορεί να υποδηλώνει ότι δεν είναι τόσο αποτελεσματικά στην σύλληψη λεκτικών μοτίβων σε αυτό το συγκεκριμένο σύνολο δεδομένων. Το σχετικά μικρό μέγεθος του “DeRev” Dataset ενδέχεται να ευνοεί το BiRNN, καθώς επωφελείται από την επεξεργασία του κειμένου εισόδου και προς τις δύο κατευθύνσεις, αποκτώντας έτσι καλύτερη κατανόηση του σχετικού λεκτικού πλαισίου που το περιβάλει. Η βελτίωση της ποιότητας ή η αύξηση του μεγέθους του συνόλου δεδομένων, με μεθόδους oversampling θα

μπορούσε δυνητικά να βοηθήσει τα πιο σύνθετα μοντέλα όπως το Transformers να αποδώσουν καλύτερα.

Στο σύνολο δεδομένων Cross-Cultural, το μοντέλο GRU αναδεικνύεται και πάλι ως το κορυφαίο, επιτυγχάνοντας την υψηλότερη τιμή σε accuracy (0,61) και F1 score. Ακολουθεί το μοντέλο Transformer, το οποίο επίσης έχει καλές επιδόσεις, αλλά με ελαφρώς μεγαλύτερο χρόνο εκπαίδευσης. Το BiRNN, εμφανίζει τις λιγότερο καλές επιδόσεις, ενώ τα μοντέλα RNN και LSTM προσφέρουν μέτρια αποτελέσματα, υποδεικνύοντας ότι συλλαμβάνουν κάποιες, αλλά όχι όλες τις πολυπλοκότητες του συνόλου δεδομένων. Η κύρια πρόκληση στο συγκεκριμένο σύνολο δεδομένων είναι πιθανότατα η ποικιλομορφία των γλωσσικών και πολιτισμικών προτύπων, γεγονός που καθιστά δύσκολη την καλή γενίκευση των απλούστερων μοντέλων.

Τα αποτελέσματα του συνόλου δεδομένων Restaurants Reviews παρουσιάζουν ιδιαίτερο ενδιαφέρον, με το μοντέλο BiRNN να έχει την καλύτερη απόδοση όσον αφορά το accuracy (0,73) και το F1 score (0,71). Αυτό υποδηλώνει ότι η επεξεργασία κειμένου προς τις δύο κατευθύνσεις παρέχει σημαντικό πλεονέκτημα στην κατανόηση των συναισθημάτων των πελατών. Ωστόσο, τα μοντέλα LSTM και GRU, τα οποία έχουν συνήθως καλές επιδόσεις, παρουσιάζουν χαμηλότερη Accuracy και F1-Score σε αυτό το σύνολο δεδομένων. Το μοντέλο Transformer παρέχει μια ισορροπημένη απόδοση με αξιοπρεπείς βαθμολογίες ακρίβειας και F1, αλλά δεν ξεπερνά το BiRNN. Η πρόκληση εδώ θα μπορούσε να είναι το σχετικά μικρό μέγεθος του συνόλου δεδομένων, το οποίο μπορεί να μην αξιοποιεί πλήρως τις δυνατότητες των πιο σύνθετων μοντέλων. Η βελτίωση της ποιότητας των δεδομένων ή ο συνδυασμός αυτών των μοντέλων σε ένα ensemble μοντέλο θα μπορούσε να βελτιώσει περαιτέρω τις επιδόσεις.

Σε όλα τα σύνολα δεδομένων, μια βασική πρόκληση είναι η επιλογή της σωστής αρχιτεκτονικής μοντέλου που εξισορροπεί την πολυπλοκότητα, τον χρόνο εκπαίδευσης και την απόδοση. Ενώ οι Transformers και τα GRUs έχουν γενικά καλές επιδόσεις, δεν είναι πάντα η καλύτερη επιλογή για κάθε σύνολο δεδομένων, όπως φαίνεται από την επιτυχία του BiRNN στα σύνολα δεδομένων DeRev και Restaurants Reviews. Οι κύριοι τομείς για βελτίωση, περιλαμβάνουν τον πειραματισμό με μεθόδους ensemble, την πιο προσεκτική ρύθμιση των υπερπαραμέτρων και ενδεχομένως την αύξηση των συνόλων δεδομένων ώστε να επιτρέπουν σε πιο σύνθετα μοντέλα να μάθουν πιο πολύπλοκες σχέσεις και εξαρτήσεις μεταξύ των κειμένων και να αποδίδουν με το βέλτιστο δυνατό τρόπο. Επιπλέον, η ερμηνευσιμότητα και

η επεξηγηματικότητα των μοντέλων είναι ζωτικής σημασίας, ιδίως για την κατανόηση του λόγου για τον οποίο ορισμένα μοντέλα υπερτερούν έναντι άλλων σε συγκεκριμένα σύνολα δεδομένων.

Το μοντέλο Transformer μπορεί να μην είχε την πιο καλή επίδοση σε όλα τα datasets, ωστόσο σε σχέση με τα υπόλοιπα έχει δείξει μεγαλύτερη σταθερότητα και καλύτερη γενίκευση. Οι αποκλίσεις μεταξύ των διαφορετικών datasets είναι μικρές, ενώ στα σύνολα δεδομένων με τις περισσότερες εγγραφές καταφέρνουν να εντοπίσουν καλύτερα από τα υπόλοιπα σχέσεις και εξαρτήσεις που οδηγούν σε καλύτερες προβλέψεις. Το γεγονός αυτό μας υποδεικνύει ότι πιθανόν προσαρμόζοντας τα transformers με βελτιστοποίηση της αρχιτεκτονικής και των υπερπαραμέτρων τους, θα μπορούσαμε να πετύχουμε πιο υψηλές επιδόσεις στις μετρικές.

5.6.2 Ρύθμιση μεταξύ τομέων (Cross-domain setup) με Preprocessing

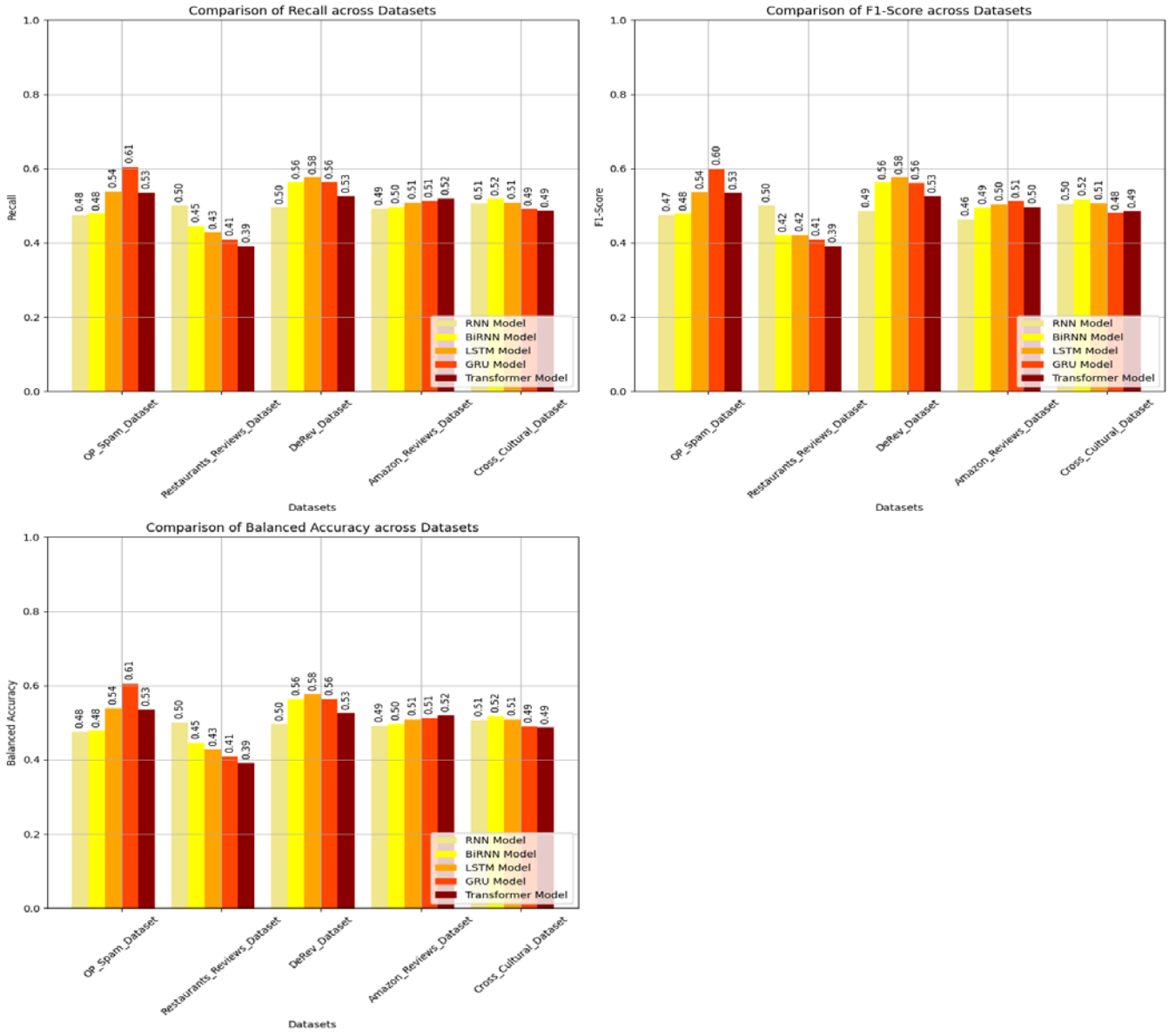
Στα πλαίσια αυτού του δεύτερου setup αξιολογούνται τα προηγούμενα μοντέλα βαθιάς μάθησης σε ένα Cross-domain περιβάλλον. Ο τρόπος που στήθηκε το τρέχον πείραμα περιγράφεται παρακάτω:

Τα πέντε διαφορετικά σύνολα δεδομένων διαβάζονται και προ-επεξεργάζονται με την ίδια λογική που περιγράφεται προηγούμενα. Η διαφοροποίηση σε αυτή τη συνθήκη είναι ότι κατά την δημιουργία των loaders που θα feed-άρουμε στα μοντέλα ακολουθείται μια διαφορετική προσέγγιση. Πιο συγκεκριμένα, αντί να δημιουργούμε train-valid-test splits ξεχωριστά για κάθε επιμέρους dataset, αυτό που κάνουμε είναι σε κάθε επανάληψη να ορίζουμε ως σύνολο εκπαίδευσης τα τέσσερα από τα πέντε datasets και σαν σύνολο ελέγχου το πέμπτο εναπόμειναν dataset. Με αυτή τη μεθοδολογία είναι σαν να εφαρμόζουμε την λογική του 5-fold cross validation, όπου σε κάθε επανάληψη ένα εκ των πέντε datasets αποτελεί το σύνολο ελέγχου. Το μοντέλο εκπαιδεύεται πάνω στα υπόλοιπα τέσσερα datasets, μαθαίνοντας μοτίβα και εξάγοντας γνώση από αυτά και μετέπειτα προσπαθεί να κάνει σωστές προβλέψεις σε ένα εντελώς άγνωστο σύνολο δεδομένων, που προέρχεται από κάποιο νέο domain, ενδεχομένως πολύ διαφορετικό από αυτά πάνω στα οποία εκπαιδεύτηκαν τα μοντέλα.

Αρχικά, εξετάζουμε την απόδοση του εκάστοτε μοντέλου στο συγκεκριμένο setting, προκειμένου να δούμε αν υπάρχει κάποιο που να υπερτερεί και να κάνει overcome την παραπάνω πρόκληση. Για το σκοπό αυτό παράγονται με το τρέξιμο του κώδικα ο πίνακας και τα διαγράμματα που παρουσιάζονται πιο κάτω:

Dataset	*Model*	*Parameters*	*Training Time*	*Test Accuracy*	*Precision*	*Recall*	*F1-Score*	*Balanced Accuracy*
Amazon_Reviews	RNN	233754	0,92	0,49	0,49	0,49	0,46	0,49
	BiRNN	244506	1,86	0,50	0,50	0,50	0,49	0,50
	LSTM	265626	1,21	0,51	0,51	0,51	0,50	0,51
	GRU	255002	2,25	0,51	0,51	0,51	0,51	0,51
	Transformer	269994	1,46	0,52	0,52	0,52	0,50	0,52
OP_Spam	RNN	713954	9,03	0,48	0,48	0,48	0,47	0,48
	BiRNN	724706	14,13	0,48	0,48	0,48	0,48	0,48
	LSTM	745826	13,32	0,54	0,54	0,54	0,54	0,54
	GRU	735202	16,91	0,61	0,61	0,61	0,60	0,61
	Transformer	750194	15,36	0,53	0,53	0,53	0,53	0,53
DeRev	RNN	757854	10,63	0,50	0,50	0,50	0,49	0,50
	BiRNN	768606	15,84	0,56	0,56	0,56	0,56	0,56
	LSTM	789726	13,55	0,58	0,58	0,58	0,58	0,58
	GRU	779102	18,81	0,56	0,56	0,56	0,56	0,56
	Transformer	794094	16,50	0,53	0,53	0,53	0,53	0,53
Cross_Cultural	RNN	762654	11,11	0,51	0,51	0,51	0,50	0,51
	BiRNN	773406	15,01	0,52	0,52	0,52	0,52	0,52
	LSTM	794526	14,05	0,51	0,51	0,51	0,51	0,51
	GRU	783902	18,82	0,49	0,49	0,49	0,48	0,49
	Transformer	798894	16,41	0,49	0,49	0,49	0,49	0,49
Restaurants_Reviews	RNN	772654	10,40	0,50	0,50	0,50	0,50	0,50
	BiRNN	783406	15,37	0,45	0,43	0,45	0,42	0,45
	LSTM	804526	14,02	0,43	0,42	0,43	0,42	0,43
	GRU	793902	18,51	0,41	0,41	0,41	0,41	0,41
	Transformer	808894	16,70	0,39	0,39	0,39	0,39	0,39

Πίνακας 2: *Μετρήσεις Απόδοσης ανά Dataset σε Cross-domain.*



Εικόνα 13: *Σύγκριση Μοντέλων Βάση Μετρικών ανά Σύνολο Δεδομένων σε In-domain.*

Συνολικά, τα αποτελέσματα δείχνουν ότι το GRU Model υπερέχει στις περισσότερες μετρήσεις και σε διάφορα σύνολα δεδομένων. Στις μετρικές όπως Test Accuracy, Recall, F1-Score και Balanced Accuracy για το Dataset “OP Spam” το GRU καταφέρνει να έχει τις υψηλότερες ή κοντά στις υψηλότερες τιμές ξεπερνώντας κάθε φορά το 60%). Ακολούθως το LSTM, αποδίδει καλύτερα στα σύνολα δεδομένων όπως το “DeRev” και το “Cross-Cultural” Datasets δίνοντας πιστότητα αποτελεσμάτων 58% και 51% αντίστοιχα σε όλους τους μετρούμενους δείκτες. Παράλληλα, τα μοντέλα BiRNN, RNN και Transformer ακολουθούν με μέτριες αποδόσεις, παρουσιάζοντας υψηλότερη ή χαμηλότερη αποτελεσματικότητα

ανάλογα με τα Datasets, καταφέρνοντας πάντως να διατηρήσουν μια αξιοπιστία που δεν πέφτει κάτω από 39%.

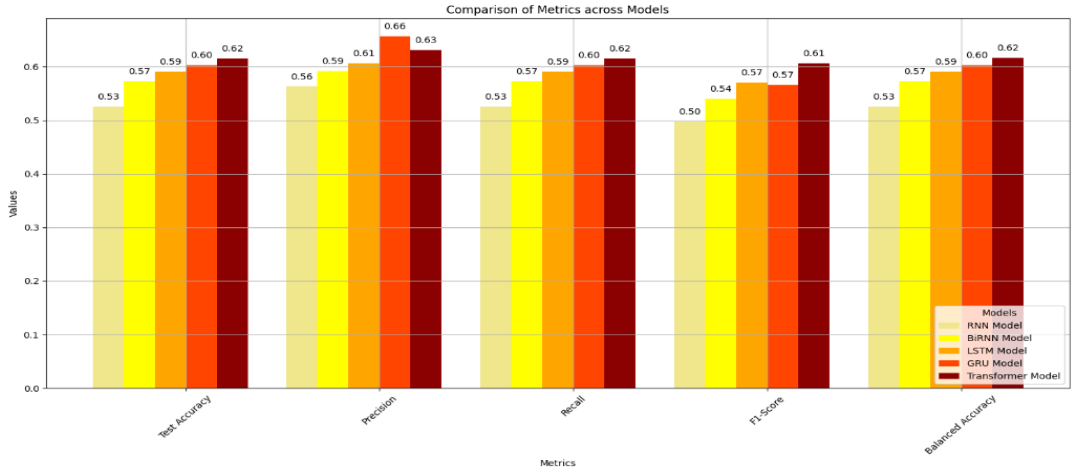
Στη συνέχεια παρουσιάζεται στους δύο πίνακες 3.1 και 3.2 η απόδοση κάθε μοντέλου συγκριτικά, σε in-domain και cross-domain περιβάλλον με preprocessing.

Dataset	*Model*	*Params*	*Training Time*	*Test Accuracy*	*Precision*	*Recall*	*F1*	*Balanced Accuracy*
Amazon Reviews	RNN	678254	1,97	0,51	0,51	0,51	0,47	0,51
OP Spam	RNN	172954	0,38	0,49	0,49	0,49	0,48	0,49
DeRev	RNN	38754	0,08	0,52	0,52	0,52	0,5	0,52
Cross Cultural	RNN	47554	0,07	0,52	0,52	0,52	0,52	0,52
Restaurants Reviews	RNN	19154	0,03	0,59	0,78	0,59	0,51	0,59
	Average			*0,53*	*0,56*	*0,53*	*0,50*	*0,53*
Amazon Reviews	BiRNN	689006	2,1	0,53	0,53	0,53	0,53	0,53
OP Spam	BiRNN	183706	0,36	0,53	0,54	0,53	0,48	0,53
DeRev	BiRNN	49506	0,05	0,6	0,61	0,6	0,6	0,6
Cross Cultural	BiRNN	58306	0,08	0,47	0,45	0,47	0,39	0,47
Restaurants Reviews	BiRNN	29906	0,04	0,73	0,82	0,73	0,71	0,73
	Average			*0,57*	*0,59*	*0,57*	*0,54*	*0,57*
Amazon Reviews	LSTM	710126	2,78	0,56	0,56	0,56	0,56	0,56
OP Spam	LSTM	204826	0,36	0,71	0,71	0,71	0,71	0,71
DeRev	LSTM	70626	0,06	0,56	0,57	0,56	0,55	0,56
Cross Cultural	LSTM	79426	0,09	0,53	0,54	0,53	0,48	0,53
Restaurants Reviews	LSTM	51026	0,02	0,59	0,65	0,59	0,54	0,59
	Average			*0,59*	*0,61*	*0,59*	*0,57*	*0,59*
Amazon Reviews	GRU	699502	3,18	0,58	0,58	0,58	0,58	0,58
OP Spam	GRU	194202	0,24	0,73	0,75	0,73	0,73	0,73
DeRev	GRU	60002	0,05	0,54	0,58	0,54	0,48	0,54
Cross Cultural	GRU	68802	0,09	0,61	0,62	0,61	0,61	0,61
Restaurants Reviews	GRU	40402	0,03	0,55	0,76	0,55	0,43	0,55
	Average			*0,60*	*0,66*	*0,60*	*0,57*	*0,60*
Amazon Reviews	Transformer	714494	2,88	0,58	0,58	0,58	0,57	0,58
OP Spam	Transformer	209194	0,29	0,69	0,69	0,69	0,69	0,69
DeRev	Transformer	74994	0,07	0,58	0,58	0,58	0,58	0,58
Cross Cultural	Transformer	83794	0,11	0,6	0,61	0,6	0,58	0,6
Restaurants Reviews	Transformer	55394	0,03	0,64	0,69	0,64	0,61	0,64
	Average			*0,62*	*0,63*	*0,62*	*0,61*	*0,62*

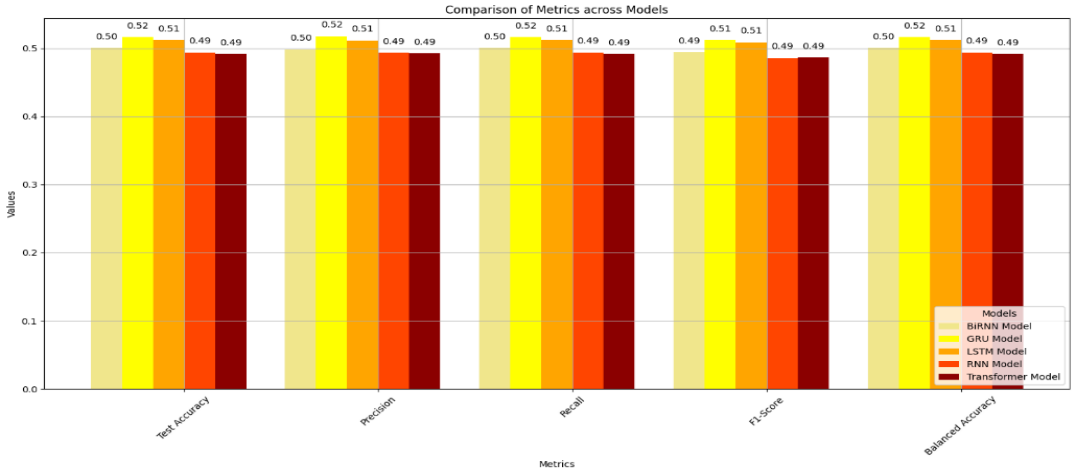
Πίνακας 3.1: *Μετρήσεις Απόδοσης ανά Μοντέλο σε In-domain.*

Dataset	Model	Parameters	Training Time	Test Accuracy	Precision	Recall	F1-Score	Balanced Accuracy
Amazon_Reviews	RNN	233754	0,92	0,49	0,49	0,49	0,46	0,49
OP_Spam	RNN	713954	9,03	0,48	0,48	0,48	0,47	0,48
DeRev	RNN	757854	10,63	0,50	0,50	0,50	0,49	0,50
Cross_Cultural	RNN	762654	11,11	0,51	0,51	0,51	0,50	0,51
Restaurants_Reviews	RNN	772654	10,40	0,50	0,50	0,50	0,50	0,50
	Average			*0,49*	*0,49*	*0,49*	*0,49*	*0,49*
Amazon_Reviews	BiRNN	244506	1,86	0,50	0,50	0,50	0,49	0,50
OP_Spam	BiRNN	724706	14,13	0,48	0,48	0,48	0,48	0,48
DeRev	BiRNN	768606	15,84	0,56	0,56	0,56	0,56	0,56
Cross_Cultural	BiRNN	773406	15,01	0,52	0,52	0,52	0,52	0,52
Restaurants_Reviews	BiRNN	783406	15,37	0,45	0,43	0,45	0,42	0,45
	Average			*0,50*	*0,50*	*0,50*	*0,49*	*0,50*
Amazon_Reviews	LSTM	265626	1,21	0,51	0,51	0,51	0,50	0,51
OP_Spam	LSTM	745826	13,32	0,54	0,54	0,54	0,54	0,54
DeRev	LSTM	789726	13,55	0,58	0,58	0,58	0,58	0,58
Cross_Cultural	LSTM	794526	14,05	0,51	0,51	0,51	0,51	0,51
Restaurants_Reviews	LSTM	804526	14,02	0,43	0,42	0,43	0,42	0,43
	Average			*0,51*	*0,51*	*0,51*	*0,51*	*0,51*
Amazon_Reviews	GRU	255002	2,25	0,51	0,51	0,51	0,51	0,51
OP_Spam	GRU	735202	16,91	0,61	0,61	0,61	0,60	0,61
DeRev	GRU	779102	18,81	0,56	0,56	0,56	0,56	0,56
Cross_Cultural	GRU	783902	18,82	0,49	0,49	0,49	0,48	0,49
Restaurants_Reviews	GRU	793902	18,51	0,41	0,41	0,41	0,41	0,41
	Average			*0,52*	*0,52*	*0,52*	*0,51*	*0,52*
OP_Spam	Transformer	750194	15,36	0,53	0,53	0,53	0,53	0,53
Restaurants_Reviews	Transformer	808894	16,70	0,39	0,39	0,39	0,39	0,39
DeRev	Transformer	794094	16,50	0,53	0,53	0,53	0,53	0,53
Amazon_Reviews	Transformer	269994	1,46	0,52	0,52	0,52	0,50	0,52
Cross_Cultural	Transformer	798894	16,41	0,49	0,49	0,49	0,49	0,49
	Average			*0,49*	*0,49*	*0,49*	*0,49*	*0,49*

Πίνακας 3.2: *Μετρήσεις Απόδοσης ανά Μοντέλο σε Cross-domain.*



Εικόνα 14: *Σύγκριση απόδοσης μοντέλων ανά μετρική σε In-domain.*



Εικόνα 15: *Σύγκριση απόδοσης μοντέλων ανά μετρική σε Cross-domain.*

Στις παραπάνω εικόνες έχουμε αποτυπώσει τη μέση απόδοση των μοντέλων και για τα πέντε datasets στα οποία δοκιμάστηκαν. Τα αποτελέσματα των πειραμάτων στο in-domain περιβάλλον δείχνουν ότι το μοντέλο Transformer υπερτερεί έναντι των υπολοίπων, επιτυγχάνοντας την υψηλότερο accuracy (0,62) και F1-Score (0,61). Τα μοντέλα GRU και LSTM έχουν επίσης καλές επιδόσεις, με το GRU να έχει το υψηλότερο precision (0,66). Τα παραδοσιακά RNNs και BiRNNs υστερούν, υποδεικνύοντας ότι τα πιο εξελιγμένα μοντέλα, όπως τα Transformers και GRUs, είναι καταλληλότερα για την ανίχνευση εξαπάτησης όταν τα δεδομένα εκπαίδευσης και ελέγχου είναι παρόμοια και προέρχονται από τον ίδιο τομέα.

Ωστόσο, στο περιβάλλον Cross-domain, όλα τα μοντέλα παρουσιάζουν σημαντική πτώση στην απόδοσή τους. Το Transformer, παρά την καλή του επίδοση στην προηγούμενη συνθήκη

(in-domain), δυσκολεύεται να γενικεύσει μεταξύ διαφορετικών τομέων μελέτης, έχοντας ίδια ή ελαφρώς χειρότερη επίδοση από απλούστερα μοντέλα όπως τα RNNs και τα BiRNNs. Τα μοντέλα GRU και LSTM επιδεικνύουν ελαφρώς καλύτερες επιδόσεις στο cross-domain περιβάλλον, γεγονός που υποδηλώνει ότι συλλαμβάνουν πιο γενικεύσιμα πρότυπα. Συνολικά, ενώ τα μοντέλα Transformer είναι ισχυρά εξατομικευμένα για κάθε dataset, τα μοντέλα GRU και LSTM προσφέρουν πιο εύρωστες επιδόσεις όταν αντιμετωπίζουν δεδομένα από διαφορετικούς τομείς, καθιστώντας τα πιο αξιόπιστα σε ποικίλα σενάρια του πραγματικού κόσμου.

5.6.3 Ρύθμιση μεταξύ τομέων (in και Cross-domain setup) χωρίς Preprocessing

Στο τρίτο και τέταρτο setup χρησιμοποιήθηκαν οι εξής global υπερπαράμετροι:

- MAX_WORDS = 256 καθορίζει το μέγιστο μήκος της ακολουθίας κειμένων, επιτρέποντας την ανάλυση έως 256 λέξεων ανά κείμενο.
- EPOCHS = 25 ορίζει τον αριθμό των επαναλήψεων εκπαίδευσης πάνω από το σύνολο δεδομένων, επιτρέποντας στο μοντέλο να βελτιωθεί σε 15 κύκλους.
- LEARNING_RATE = 2e-5 προσδιορίζει τον ρυθμό μάθησης, επηρεάζοντας τη ταχύτητα και τη σταθερότητα της εκπαίδευσης.
- BATCH_SIZE = 32 καθορίζει τον αριθμό των δειγμάτων δεδομένων που χρησιμοποιούνται σε κάθε επανάληψη για την ενημέρωση των βαρών του μοντέλου.
- EMBEDDING_DIM = 256 αναφέρεται στη διάσταση του διανύσματος αναπαράστασης λέξεων, επιτρέποντας πλουσιότερες γλωσσικές αναπαραστάσεις.
- HIDDEN_DIM = 256 καθορίζει τον αριθμό των νευρώνων στο κρυφό επίπεδο των RNNs, LSTMs ή GRUs, επηρεάζοντας την ικανότητα του μοντέλου να αποθηκεύει πληροφορίες.
- NUM_HEADS = 16 σχετίζεται με τον αριθμό των κεφαλών στον μηχανισμό προσοχής, ενισχύοντας τη δυνατότητα του μοντέλου να αποτυπώσει σύνθετες σχέσεις.

Εκτός από τις αλλαγές στις παραμέτρους, έχουν χρησιμοποιηθεί δύο ακόμη μοντέλα τα BERT και RoBERTa και έχουν εφαρμοστεί και κάποιες επιπλέον τεχνικές για την βελτιστοποίηση των αποτελεσμάτων ως ακολούθως:

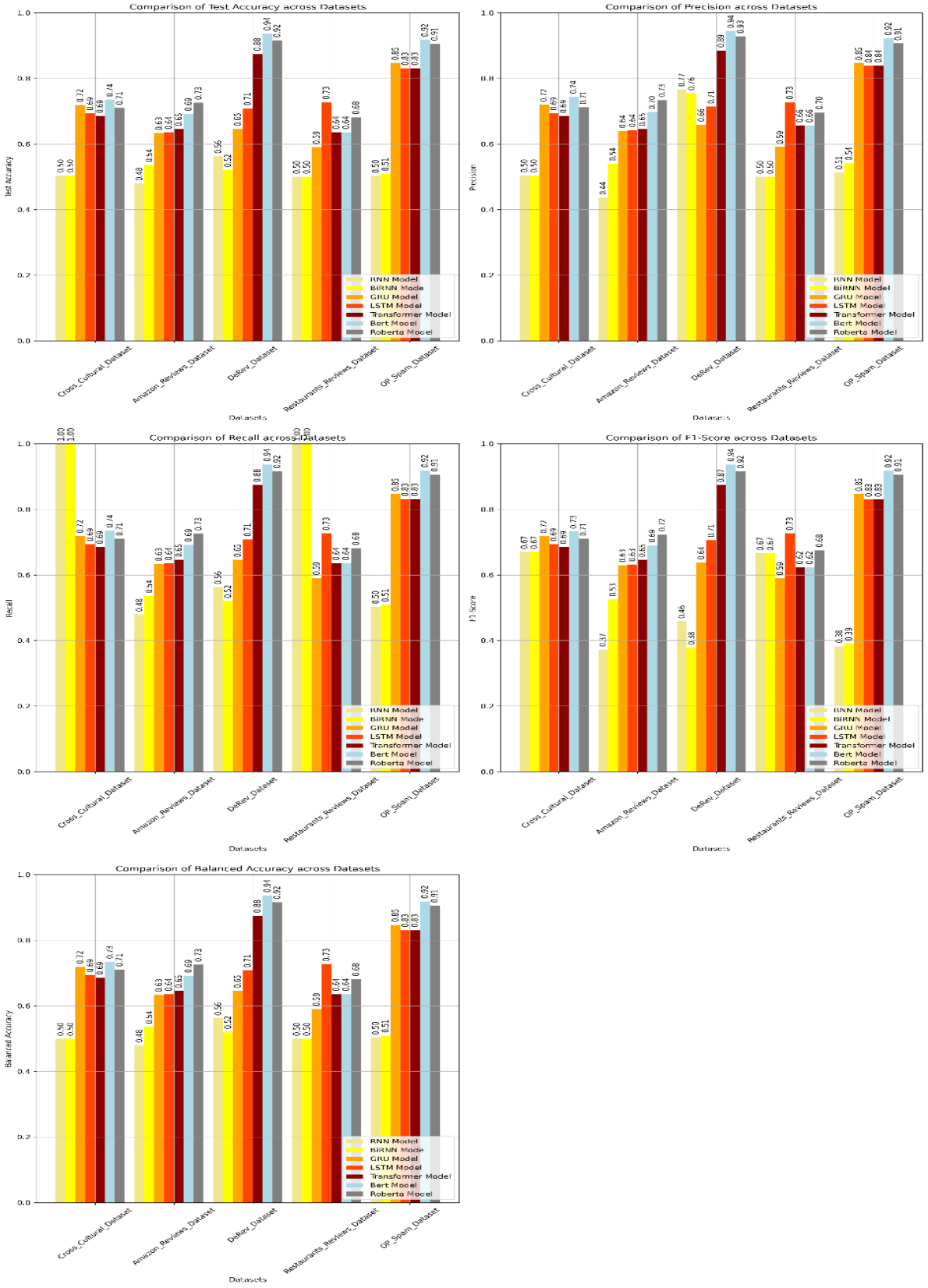
- Αρχικά έγινε προσθήκη dropout στα περισσότερα μοντέλα για καλύτερη γενίκευση που ουσιαστικά το dropout είναι μια τεχνική regularization που βοηθά στην αποφυγή υπερπροσαρμογής (overfitting) κατά την εκπαίδευση των νευρωνικών δικτύων και κατά τη διάρκεια της εκπαίδευσης, τυχαία “απενεργοποιούνται” ορισμένοι νευρώνες, μειώνοντας έτσι την πιθανότητα το μοντέλο να εξαρτηθεί υπερβολικά από συγκεκριμένα μοτίβα στα δεδομένα εκπαίδευσης.
- Έπειτα έγινε προσθήκη “gradient clipping” κατά την εκπαίδευση των μοντέλων για να εξαλειφθεί το πρόβλημα του “exploding gradient” δηλαδή είναι μια τεχνική που περιορίζει το μέγεθος των gradients κατά την εκπαίδευση, αποτρέποντας τα να γίνουν υπερβολικά μεγάλα, έτσι είναι ιδιαίτερα χρήσιμο σε RNNs (Recurrent Neural Networks), όπου τα gradients μπορούν να εκραγούν (exploding gradients), προκαλώντας αστάθεια στην εκπαίδευση.
- Στην συνέχεια έγινε αρχικοποίηση των weights με τη διαδικασία Xavier και set σε 0 για τα biases (σε όλα τα μοντέλα εκτός από Roberta/Bert) που είναι μια μέθοδος για την αρχικοποίηση των weights σε νευρωνικά δίκτυα, που βοηθά στην αποφυγή προβλημάτων με vanishing/exploding gradients, με τα biases συνήθως να αρχικοποιούνται σε 0 για να μην επηρεάζουν την αρχική εκπαίδευση.
- Ακολούθως για το Bert/Roberta, πρώτα χρησιμοποιήθηκαν τα pre-trained μοντέλα, και μετά έγινε fine-tuning με training πάνω στο εκάστοτε dataset όπου τα μοντέλα BERT και RoBERTa είναι προεκπαιδευμένα σε μεγάλα κείμενα και στη συνέχεια προσαρμόζονται (fine-tuned) σε συγκεκριμένα datasets για συγκεκριμένες εργασίες, βελτιώνοντας έτσι την απόδοσή τους σε αυτές τις εργασίες.
- Επίσης κάνουμε χρήση bi-directional cells για LSTMs και GRUs cells, αλλά και αύξηση των layers στα RNN/BiRNN/LSTM/GRU δηλαδή με τα bi-directional cells επιτρέπουν στα LSTMs και GRUs να λαμβάνουν υπόψη τους πληροφορίες τόσο από το παρελθόν όσο και από το μέλλον σε μια ακολουθία δεδομένων, όπου η αύξηση των layers μπορεί να βελτιώσει την ικανότητα του μοντέλου να μαθαίνει πιο σύνθετα μοτίβα.
- Τέλος κάνουμε χρήση του validation set για να αξιολογήσουμε και να κρατήσουμε το καλύτερο μοντέλο κατά την εκπαίδευση (με κριτήριο το validation loss) με το validation set να χρησιμοποιείται για την αξιολόγηση της απόδοσης του μοντέλου κατά την εκπαίδευση, έτσι το μοντέλο με το χαμηλότερο validation loss θεωρείται το

καλύτερο, καθώς δείχνει καλύτερη γενίκευση στα δεδομένα που δεν έχει δει κατά την εκπαίδευση.

Τα αποτελέσματα ομαδοποιημένα ανά σετ δεδομένων για κάθε μοντέλο ανάλυσης φαίνονται στον παρακάτω πίνακα 4 ενώ για καλύτερη σύγκριση και πιο εποπτική παρουσίαση έχουν σχεδιαστεί και τα αντίστοιχα διαγράμματα δίνοντας τα αποτελέσματα που πετυχαίνουν τα γλωσσικά μοντέλα ανά μετρική και ανά σετ δεδομένων στην εικόνα 16

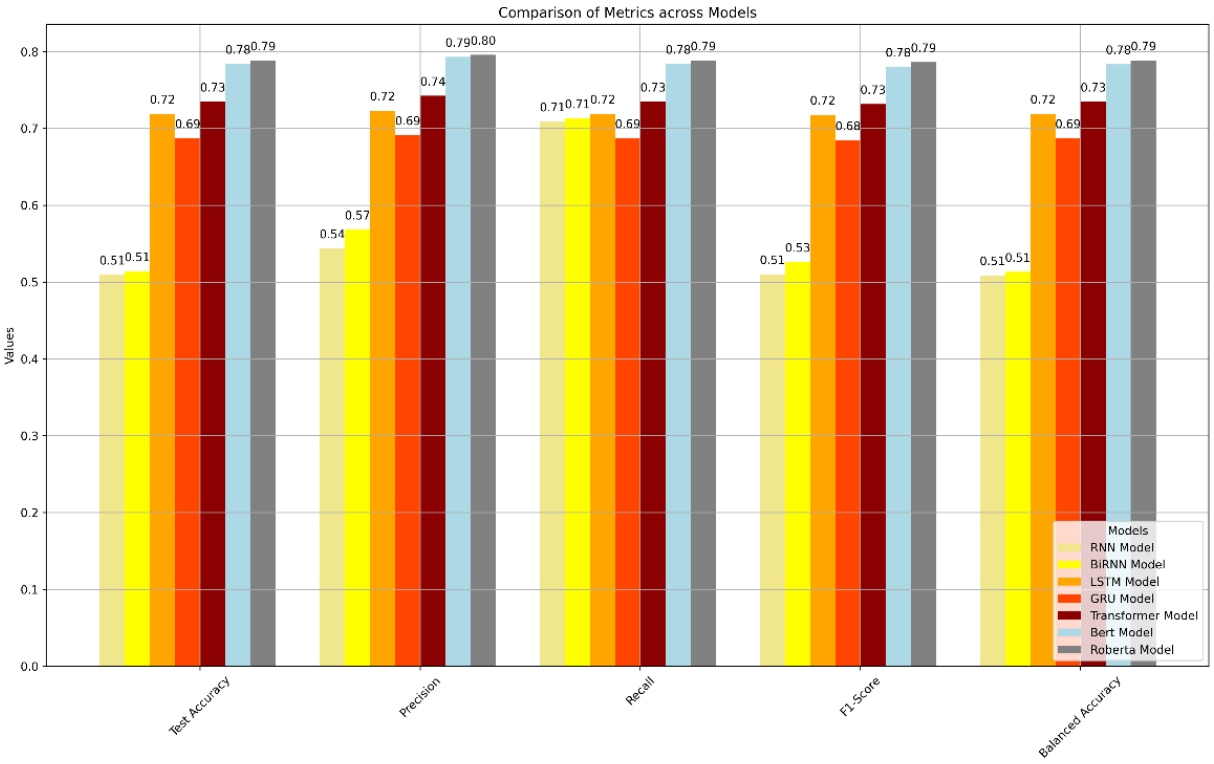
C	Dataset	Model	Parameters	Training Tim	Test Accurac	Precision	Recall	F1-Score	Balanced Acc
0	Cross_Cultural	RNN	177670	0,052	0,504	0,504	1,000	0,670	0,500
1	Amazon_Reviews	RNN	812770	1,720	0,480	0,436	0,480	0,373	0,480
2	DeRev	RNN	168870	0,023	0,563	0,767	0,563	0,459	0,563
3	Restaurants_Reviews	RNN	144970	0,010	0,500	0,500	1,000	0,667	0,500
4	OP_Spam	RNN	305870	0,157	0,503	0,514	0,503	0,383	0,503
5	Cross_Cultural	BiRNN	404742	0,078	0,504	0,504	1,000	0,670	0,500
6	Amazon_Reviews	BiRNN	1039842	2,688	0,536	0,540	0,536	0,526	0,536
7	DeRev	BiRNN	395942	0,032	0,521	0,755	0,521	0,378	0,521
8	Restaurants_Reviews	BiRNN	372042	0,015	0,500	0,500	1,000	0,667	0,500
9	OP_Spam	BiRNN	532942	0,203	0,509	0,542	0,509	0,391	0,509
10	Cross_Cultural	GRU	523014	0,076	0,719	0,720	0,719	0,719	0,719
11	Amazon_Reviews	GRU	1158114	2,607	0,634	0,640	0,634	0,630	0,634
12	DeRev	GRU	514214	0,031	0,646	0,659	0,646	0,638	0,646
13	Restaurants_Reviews	GRU	490314	0,014	0,591	0,592	0,591	0,590	0,591
14	OP_Spam	GRU	651214	0,199	0,847	0,847	0,847	0,847	0,847
15	Cross_Cultural	LSTM	1471238	0,214	0,694	0,694	0,694	0,694	0,694
16	Amazon_Reviews	LSTM	2106338	7,480	0,636	0,643	0,636	0,632	0,636
17	DeRev	LSTM	1462438	0,083	0,708	0,714	0,708	0,706	0,708
18	Restaurants_Reviews	LSTM	1438538	0,039	0,727	0,727	0,727	0,727	0,727
19	OP_Spam	LSTM	1599438	0,570	0,831	0,839	0,831	0,830	0,831
20	Cross_Cultural	Transformer	454658	0,094	0,686	0,686	0,686	0,686	0,686
21	Amazon_Reviews	Transformer	2080514	3,510	0,646	0,646	0,646	0,646	0,646
22	DeRev	Transformer	432130	0,036	0,875	0,886	0,875	0,874	0,875
23	Restaurants_Reviews	Transformer	370946	0,016	0,636	0,657	0,636	0,624	0,636
24	OP_Spam	Transformer	782850	0,244	0,831	0,839	0,831	0,830	0,831
25	Cross_Cultural	Bert	109483778	7,362	0,736	0,745	0,736	0,733	0,735
26	Amazon_Reviews	Bert	109483778	248,981	0,692	0,698	0,692	0,690	0,692
27	DeRev	Bert	109483778	2,604	0,938	0,944	0,938	0,937	0,938
28	Restaurants_Reviews	Bert	109483778	1,134	0,636	0,657	0,636	0,624	0,636
29	OP_Spam	Bert	109483778	18,772	0,919	0,923	0,919	0,919	0,919
30	Cross_Cultural	Roberta	124647170	7,666	0,711	0,712	0,711	0,710	0,711
31	Amazon_Reviews	Roberta	124647170	270,902	0,726	0,735	0,726	0,724	0,726
32	DeRev	Roberta	124647170	3,001	0,917	0,929	0,917	0,916	0,917
33	Restaurants_Reviews	Roberta	124647170	1,376	0,682	0,696	0,682	0,676	0,682
34	OP_Spam	Roberta	124647170	20,444	0,906	0,907	0,906	0,906	0,906

Πίνακας 4: *Μετρήσεις Απόδοσης ανά Μοντέλο σε In-domain.*



Εικόνα 16: *Σύγκριση Μοντέλων Βάση Μετρικών ανά Σύνολο Δεδομένων σε In-domain.*

Τελικά μπορούμε να δούμε στην παρακάτω εικόνα 17 ένα συγκριτικό διάγραμμα όπου φαίνεται η απόδοση των μοντέλων που πετυχαίνουν στις μετρικές, ανά σύνολο δεδομένων.



Εικόνα 17: *Σύγκριση απόδοσης μοντέλων ανά μετρική σε In-domain.*

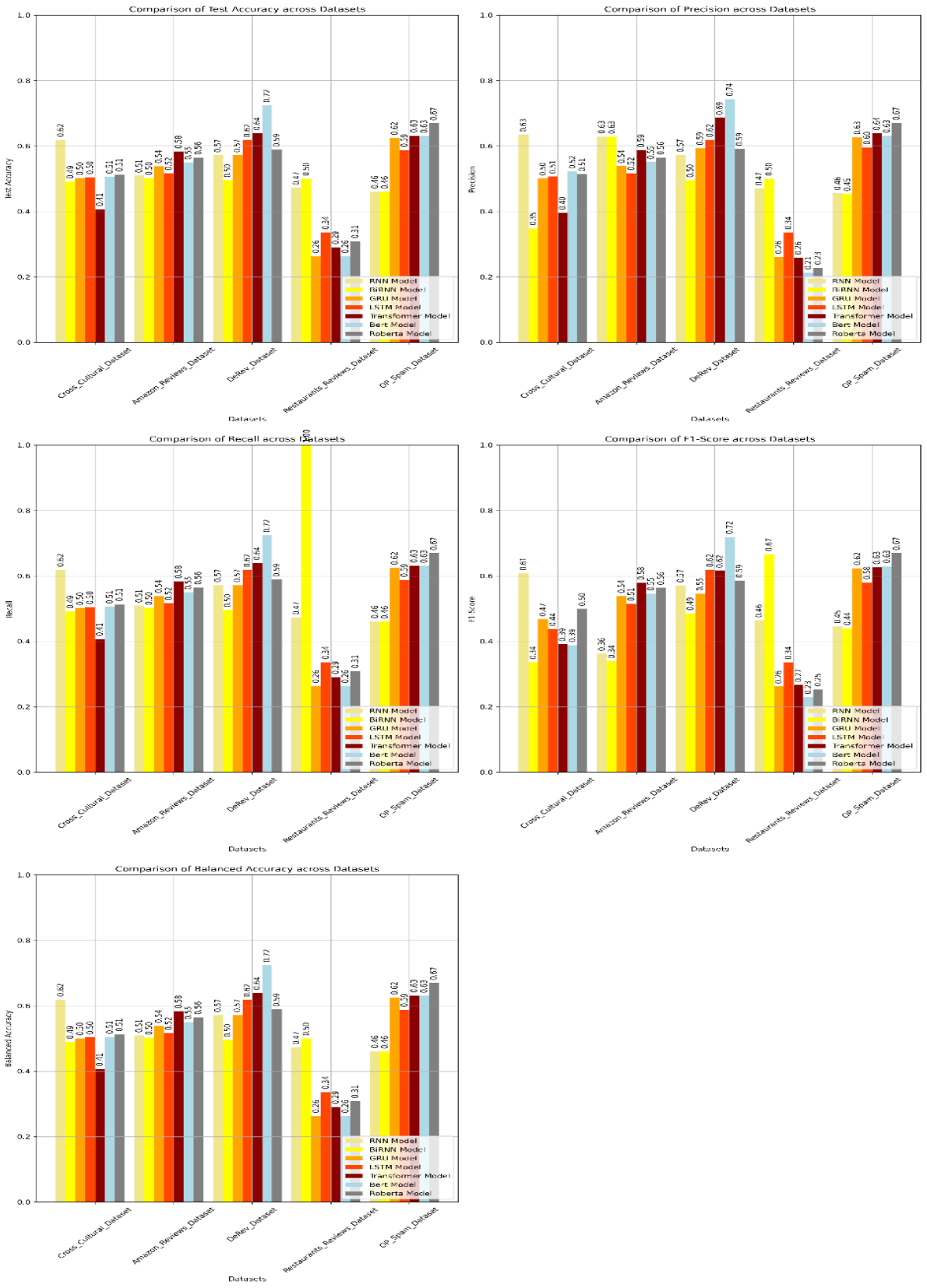
Πράγματι στο διάγραμμα αυτό φαίνεται να ξεχωρίζουν οι επιδόσεις που πετυχαίνουν τα πιο σύγχρονα μοντέλα BERT και RoBERTa αφού οι αποδόσεις τους φτάνουν στο 80% και είναι σαφώς καλύτερες από τους transformers που ακολουθούν με επιδόσεις κοντά στο 74% και το LSTM με επιδόσεις κοντά στο 72%. Παλαιότερα μοντέλα όπως τα RNN και BiRNN πετυχαίνουν επιδόσεις κοντά στο 55% που δεν είναι ικανοποιητικές για την ανίχνευση ψευδών αναφορών.

Ακολουθεί ο πίνακας 5 όπου παρουσιάζονται τα αποτελέσματα στο περιβάλλον cross-domain που περιεγράφηκε πιο πάνω ομαδοποιώντας τα αποτελέσματα που πετυχαίνει κάθε σύνολο δεδομένων με τη χρήση καθ' ενός από τα 7 μοντέλα που χρησιμοποιήθηκαν στο πείραμα.

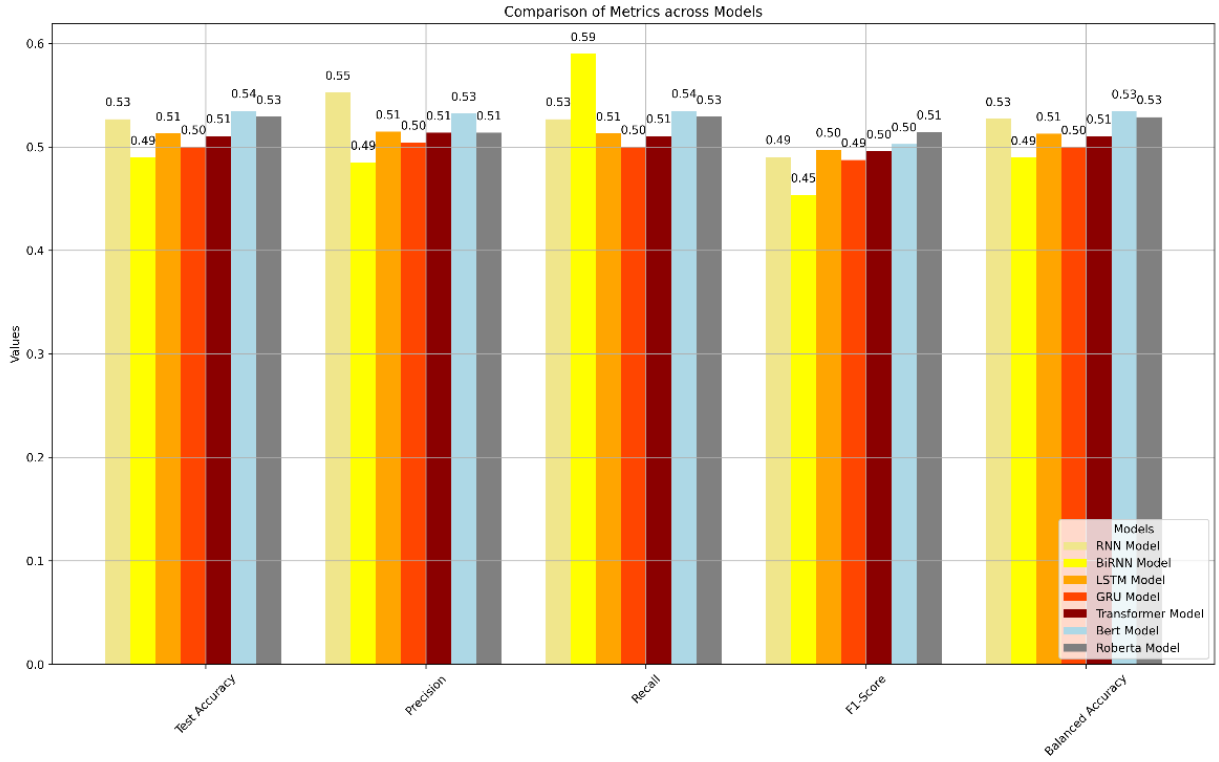
C	Dataset	Model	Paramete	Training T	Test Accu	Precision	Recall	F1-Score	Balanced
0	Cross_Cultural	RNN	909570	2,504	0,619	0,635	0,619	0,608	0,620
1	Amazon_Reviews	RNN	909570	0,290	0,510	0,629	0,510	0,362	0,510
2	DeRev	RNN	909570	2,518	0,572	0,572	0,572	0,571	0,572
3	Restaurants_Reviews	RNN	909570	2,538	0,473	0,471	0,473	0,464	0,473
4	OP_Spam	RNN	909570	2,345	0,461	0,456	0,461	0,446	0,461
5	Cross_Cultural	BiRNN	1136642	3,789	0,492	0,349	0,492	0,336	0,490
6	Amazon_Reviews	BiRNN	1136642	0,427	0,502	0,628	0,502	0,340	0,502
7	DeRev	BiRNN	1136642	3,851	0,496	0,495	0,496	0,486	0,496
8	Restaurants_Reviews	BiRNN	1136642	3,866	0,500	0,500	1,000	0,667	0,500
9	OP_Spam	BiRNN	1136642	3,626	0,460	0,453	0,460	0,440	0,460
10	Cross_Cultural	GRU	1254914	3,768	0,502	0,501	0,502	0,468	0,501
11	Amazon_Reviews	GRU	1254914	0,417	0,539	0,539	0,539	0,538	0,539
12	DeRev	GRU	1254914	3,792	0,572	0,594	0,572	0,546	0,572
13	Restaurants_Reviews	GRU	1254914	3,832	0,264	0,262	0,264	0,262	0,264
14	OP_Spam	GRU	1254914	3,569	0,625	0,627	0,625	0,624	0,625
15	Cross_Cultural	LSTM	2203138	10,507	0,505	0,507	0,505	0,438	0,504
16	Amazon_Reviews	LSTM	2203138	1,163	0,517	0,517	0,517	0,515	0,517
17	DeRev	LSTM	2203138	10,667	0,619	0,619	0,619	0,619	0,619
18	Restaurants_Reviews	LSTM	2203138	10,780	0,336	0,336	0,336	0,336	0,336
19	OP_Spam	LSTM	2203138	10,081	0,588	0,595	0,588	0,580	0,588
20	Cross_Cultural	Transformer	834386	2,597	0,407	0,396	0,407	0,392	0,407
21	Amazon_Reviews	Transformer	834386	0,293	0,583	0,587	0,583	0,579	0,583
22	DeRev	Transformer	834386	2,490	0,640	0,686	0,640	0,616	0,640
23	Restaurants_Reviews	Transformer	834386	2,683	0,291	0,259	0,291	0,267	0,291
24	OP_Spam	Transformer	834386	2,413	0,632	0,640	0,632	0,627	0,632
25	Cross_Cultural	Bert	109483778	361,467	0,507	0,522	0,507	0,389	0,505
26	Amazon_Reviews	Bert	109483778	40,308	0,550	0,551	0,550	0,546	0,550
27	DeRev	Bert	109483778	368,778	0,725	0,744	0,725	0,719	0,725
28	Restaurants_Reviews	Bert	109483778	370,633	0,264	0,214	0,264	0,230	0,264
29	OP_Spam	Bert	109483778	345,812	0,631	0,632	0,631	0,630	0,631
30	Cross_Cultural	Roberta	124647170	367,619	0,513	0,514	0,513	0,499	0,513
31	Amazon_Reviews	Roberta	124647170	40,944	0,564	0,565	0,564	0,564	0,564
32	DeRev	Roberta	124647170	374,508	0,589	0,592	0,589	0,586	0,589
33	Restaurants_Reviews	Roberta	124647170	374,957	0,309	0,228	0,309	0,254	0,309
34	OP_Spam	Roberta	124647170	351,385	0,670	0,670	0,670	0,670	0,670

Πίνακας 5: *Μετρήσεις Απόδοσης ανά Μοντέλο σε Cross-domain.*

Και αντιστοίχως παρουσιάζονται σε μορφή διαγραμμάτων στην εικόνα 18 τα ίδια αποτελέσματα για να δοθεί μια εποπτική εικόνα των μετρικών για κάθε μοντέλο και κάθε σετ δεδομένων. Ακολούθως στην εικόνα 19 παρουσιάζεται το διάγραμμα με μία συγκριτική απόδοση των μοντέλων στις 5 μετρικές που χρησιμοποιούνται για την αξιολόγηση.



Εικόνα 18: *Σύγκριση Μοντέλων Βάση Μετρικών ανά Σύνολο Δεδομένων σε Cross-domain.*



Εικόνα 19: *Σύγκριση απόδοσης μοντέλων ανά μετρική σε Cross-domain.*

Κατά την εκπαίδευση των μοντέλων με την μεθοδολογία “in_domain”, όπου κάθε dataset χρησιμοποιείται για training/testing/validation για κάθε μοντέλο, παρατηρήσαμε καλή συμπεριφορά των περισσοτέρων μοντέλων. Η μεγαλύτερη δυσκολία σε αυτό τον τρόπο εκπαίδευσης ήταν οι μεγάλες διαφορές στα μεγέθη των dataset, που μπορούν να δημιουργήσουν δυσκολίες στα πιο απλά μοντέλα. Αυτό παρατηρήθηκε κυρίως στα μοντέλα RNN και Bi-Directional RNN, τα οποία δεν μπόρεσαν να φτάσουν καλή ακρίβεια, με μέση ακρίβεια 0.51 και 0.51, αντίστοιχα. Μοντέλα όπως τα GRUs και LSTMs παρέχουν σχεδόν 40% καλύτερη ακρίβεια σε σχέση με τα προηγούμενα (με μέση ακρίβεια 0.69 και 0.71, αντίστοιχα), κυρίως λόγω της επιπλέον πολυπλοκότητας και ικανότητας να καταλαβαίνουν εξαρτήσεις σε μεγάλες ακολουθίες, όπως και στην περίπτωση των dataset που χρησιμοποιήθηκαν. Επιπλέον βελτίωση, αν και σε μικρότερο βαθμό έδωσαν απλά transformer μοντέλα με τον μηχανισμό self-attention. Την καλύτερη επίδοση είδαμε στα πιο πολύπλοκα μοντέλα μας, Bert και Roberta, που κατάφεραν να φτάσουν μέση ακρίβεια σχεδόν στα 0.8. Σημαντικό είναι να αναφέρουμε ότι η μέση ακρίβεια παρατηρήθηκε καλύτερη για τα περισσότερα μοντέλα κυρίως στα DevRev και OP_Spam datasets, ενώ στα υπόλοιπα είχαμε συγκριτικά χαμηλότερες

τιμές. Εξαίρεση αποτελούν τα Bi-RNN και RNN μοντέλα που είχαν ακριβώς την ίδια συμπεριφορά σε όλα τα datasets.

Κατά την εκπαίδευση των μοντέλων με την μεθοδολογία “cross_domain”, όπου ένα dataset χρησιμοποιείται για testing και τα υπόλοιπα για training για κάθε μοντέλο, παρατηρήσαμε κακή συμπεριφορά για όλα τα μοντέλα. Ενώ αυξάνουμε το πλήθος των δειγμάτων στο training, άρα και δίνουμε παραπάνω δυνατότητες στο μοντέλο να «μάθει», το testing δεν βρίσκεται στον χώρο εκπαίδευσης των δεδομένων. Αυτό έχει ως αποτέλεσμα την χειρότερη ακρίβεια των μοντέλων, ακόμη και με την μεγάλη εγγύτητα των προβλημάτων, δηλαδή κάθε dataset είχε κατά βάση reviews και έπρεπε να βρεθεί αν είναι αληθοφανής ή όχι. Σχεδόν όλα τα μοντέλα είχαν μέση ακρίβεια γύρω από το 0.5, που πρακτικά σημαίνει ότι ακόμη και το πιο απλό μοντέλο τυχαίας πρόβλεψης (π.χ., «κορώνα ή γράμματα»), στατιστικά θα είχε την ίδια επίδοση. Οι μόνες σοβαρές διαφορές βρέθηκαν στα DeRev και OP_spam datasets που τα Roberta/Bert μοντέλα έφτασαν ικανοποιητικά το 0.74 και 0.67. Τέλος, μεγάλα προβλήματα παρουσιάστηκαν στην εκπαίδευση με testing set το Restaurant_Reviews, που τα πιο προηγμένα μοντέλα είχαν επιδόσεις κοντά στο εύρος του 0.26-3.5, αδυνατώντας να φτάσουν ακόμη και την μέση ακρίβεια του 0.5.

5.6.4 Συμπέρασμα

Συνοψίζοντας, το προτεινόμενο πειραματικό πλαίσιο προσφέρει ένα ισχυρό εργαλείο για την αξιολόγηση μοντέλων deep learning στην ανίχνευση απάτης, ενώ προάγει την έρευνα στον τομέα, διευκολύνοντας την εξερεύνηση νέων τεχνικών και datasets.

Όσον αφορά την απόδοση των μοντέλων παρατηρήσαμε ότι τόσο τα transformers όσο και τα Transformer-based μοντέλα (BERT και RoBERTa) επέδειξαν σταθερά καλή απόδοση σε σχέση με τα RNN-based μοντέλα τα οποία πήγαν εξίσου καλά σε ορισμένες μετρικές αλλά η απόδοσή τους είχε μεγάλες διακυμάνσεις πράγμα που δείχνει ότι τα αποτελέσματά τους, χωρίς προεπεξεργασία, έχουν μεγαλύτερη ευαισθησία στην ποιότητα και μέγεθος των datasets.

Η απόδοση των μοντέλων με προεπεξεργασία σε περιβάλλον Cross-domain δίνει μια ομοιόμορφη σχετικά εικόνα με μικρή διακύμανση στην απόδοση για όλα τα μοντέλα μην επιτρέποντας ουσιαστικά να διακρίνουμε ποια μοντέλα πλεονεκτούν σε σχέση με τα

υπόλοιπα ενώ στο περιβάλλον in-domain τονίζεται η ισχύς των attention μηχανισμών στην ανίχνευση απάτης και μένουν με χαμηλή απόδοση τα RNN μοντέλα.

Ένα ακόμα συμπέρασμα στο οποίο καταλήγουμε είναι ότι είναι ιδιαίτερα δύσκολο και αποτελεί σημαντική πρόκληση να βρεθεί μια κοινώς αποδεκτή αλγοριθμική λύση που να καλύπτει όλα τα πιθανά domains με διαφορετικά vocabularies και διαφορετικά παρατηρούμενα μοτίβα στο τρόπο με τον οποίο εκφράζονται οι χρήστες στα κοινωνικά δίκτυα.

Ακόμη παρόλο που είδαμε κατά την έρευνα μας καλά αποτελέσματα στα πειράματα όπου κάθε σύνολο δεδομένων αξιολογείται ξεχωριστά, στο πειραματικό πλαίσιο με την cross-domain λογική είδαμε την απόδοση σε όλες τις μετρικές αξιολόγησης να μειώνεται σημαντικά. Αυτή η διαπίστωση έχει γίνει στο παρελθόν και σε άλλες επιστημονικές εργασίες γεγονός που καθιστά ακόμα το deception detection ένα ενδιαφέρον ζήτημα για μελέτη.

5.6.5 Μελλοντική Εργασία

Η βελτίωση της απόδοσης των μοντέλων μελλοντικά μπορεί να γίνει με την προσθήκη επιπλέον χαρακτηριστικών που αφορούν το sentiment του υπό εξέταση κειμένου όπως:

1. Χαρακτηριστικά συναισθηματικής έντασης: Προσθήκη χαρακτηριστικών που μετρούν την ένταση των συναισθημάτων (π.χ. θετικό, αρνητικό, ουδέτερο) μπορεί να βοηθήσει τα μοντέλα να κατανοήσουν καλύτερα το συναισθηματικό περιεχόμενο του κειμένου.
2. Χαρακτηριστικά συναισθηματικής πολικότητας: Αυτά τα χαρακτηριστικά μπορούν να βοηθήσουν τα μοντέλα να διακρίνουν μεταξύ θετικών και αρνητικών συναισθημάτων, βελτιώνοντας την ακρίβεια της ανάλυσης συναισθήματος.
3. Χαρακτηριστικά συναισθηματικής συνέχειας: Η χρήση χαρακτηριστικών που λαμβάνουν υπόψη τη συνέχεια των συναισθημάτων σε μεγαλύτερα κείμενα μπορεί να βελτιώσει την κατανόηση των συναισθηματικών μεταβολών.
4. Χαρακτηριστικά από εξωτερικές πηγές: Χρήση λεξικών συναισθημάτων ή άλλων εξωτερικών πηγών για την ενίσχυση των χαρακτηριστικών του μοντέλου.

Κεφάλαιο 6

Συμπεράσματα

Σύμφωνα με τα αποτελέσματα των πειραμάτων όπως προέκυψαν στην παρούσα εργασία, υπάρχει μια σύγκριση επτά διαφορετικών μοντέλων βαθιάς μάθησης - RNN, BiRNN, LSTM, GRU, Transformer, BERT και RoBERTa - με διάφορα γνωστά σύνολα δεδομένων με βάση πέντε μετρικές απόδοσης: ακρίβεια δοκιμής, ευσυνειδησία, ανάκληση, F1-Score και ισορροπημένη ακρίβεια.

6.1 Συμπεράσματα από τα Αποτελέσματα με Προεπεξεργασία

Αρχικά, όπου εφαρμόστηκε προεπεξεργασία, αξιολογήθηκαν πέντε μοντέλα (RNN, BiRNN, LSTM, GRU, και Transformer) σε πέντε σύνολα δεδομένων. Η προεπεξεργασία περιλάμβανε βήματα όπως καθαρισμό δεδομένων, κανονικοποίηση και ενδεχομένως επιλογή χαρακτηριστικών.

Χαμηλότερη Απόδοση μετά την Προεπεξεργασία:

- Τα αποτελέσματα έδειξαν ότι η απόδοση των μοντέλων με προεπεξεργασία ήταν χαμηλότερη σε σύγκριση με την εκδοχή χωρίς προεπεξεργασία, με τις μετρικές να παρουσιάζουν μειώσεις στις τιμές ακρίβειας, F1-score και άλλες. Για παράδειγμα, το Transformer, το οποίο αρχικά είχε ακρίβεια γύρω στο 60%, μετά την προεπεξεργασία σημείωσε πτώση σε ποσοστά περίπου στο 49%-51%.
- Αυτό υποδεικνύει ότι η προεπεξεργασία μπορεί να έχει αφαιρέσει κρίσιμες πληροφορίες ή να έχει αλλοιώσει τα χαρακτηριστικά των δεδομένων με τρόπο που επηρεάζει αρνητικά την απόδοση των μοντέλων.

Επίδραση της Προεπεξεργασίας στα Απλούστερα Μοντέλα:

- Τα απλούστερα μοντέλα, όπως τα RNN και BiRNN, έδειξαν μεγαλύτερη πτώση στην απόδοσή τους μετά την προεπεξεργασία. Αυτό πιθανόν να οφείλεται στο γεγονός ότι τα μοντέλα αυτά βασίζονται περισσότερο στην αρχική διαμόρφωση των δεδομένων και επηρεάζονται αρνητικά όταν αυτή αλλάζει.

Προβλήματα με τη Μέθοδο Προεπεξεργασίας:

- Η προεπεξεργασία μπορεί να περιλάμβανε διαδικασίες που αφαίρεσαν θορυβώδη δεδομένα αλλά και χρήσιμες πληροφορίες, οδηγώντας σε απώλεια κρίσιμων σημάτων για την ανίχνευση απάτης. Η υπερβολική κανονικοποίηση ή η απομάκρυνση στοιχείων που περιέχουν σημαντικά χαρακτηριστικά είναι πιθανά προβλήματα.

6.2 Συμπεράσματα από τα Αποτελέσματα χωρίς Προεπεξεργασία

Στο δεύτερο setup, χρησιμοποιήθηκαν επτά μοντέλα (RNN, BiRNN, LSTM, GRU, Transformer, BERT και RoBERTa) χωρίς καμία προεπεξεργασία στα δεδομένα.

Καλύτερη Απόδοση χωρίς Προεπεξεργασία:

- Τα μοντέλα πέτυχαν καλύτερες μετρικές απόδοσης σε σύγκριση με την έκδοση με προεπεξεργασία, με την ακρίβεια να κυμαίνεται στο 52%-62% ανάλογα με το μοντέλο. Τα πιο σύνθετα μοντέλα, όπως τα BERT και RoBERTa, επωφελήθηκαν από την απουσία προεπεξεργασίας, καταφέρνοντας να αξιοποιήσουν πλήρως τις ωμές πληροφορίες των δεδομένων.
- Η ακρίβεια των μοντέλων Transformer, BERT, και RoBERTa, που κυμαινόταν περίπου στο 60%, επιβεβαιώνει ότι τα μοντέλα αυτά είναι πιο ανθεκτικά στον θόρυβο και μπορούν να αξιοποιήσουν τα ωμά δεδομένα καλύτερα από ό,τι με προεπεξεργασία.

Πλεονεκτήματα της Χωρίς Προεπεξεργασία Προσέγγισης:

- Η διατήρηση όλων των πληροφοριών, ακόμα και αν αυτές περιέχουν θόρυβο, φαίνεται να επιτρέπει στα πιο εξελιγμένα μοντέλα να εντοπίζουν τα σημαντικά μοτίβα που οδηγούν σε καλύτερη απόδοση.
- Η απουσία προεπεξεργασίας μπορεί επίσης να έχει μειώσει το overfitting, καθώς τα μοντέλα αξιοποίησαν μεγαλύτερο όγκο δεδομένων με τη φυσική τους μορφή.

6.3 Σύγκριση των Αποτελεσμάτων με και χωρίς Προεπεξεργασία

Απώλεια Πληροφορίας λόγω Προεπεξεργασίας:

- Η προεπεξεργασία, παρά το γεγονός ότι στοχεύει στη βελτίωση της ποιότητας των δεδομένων, φαίνεται ότι σε αυτή την περίπτωση απέτυχε να διατηρήσει τα κρίσιμα χαρακτηριστικά, κάτι που είχε ως αποτέλεσμα τη χειροτέρευση της απόδοσης.
- Οι μέθοδοι που εφαρμόστηκαν ενδέχεται να αφαίρεσαν σημαντικές πληροφορίες ή να μετέβαλαν τη δομή των δεδομένων με τρόπο που επηρέασε την κατανόηση των μοντέλων.

Ανθεκτικότητα των Μοντέλων στην Ποιότητα των Δεδομένων:

- Τα πιο σύνθετα μοντέλα, όπως τα BERT και RoBERTa, έδειξαν μεγαλύτερη ανθεκτικότητα στην απουσία προεπεξεργασίας, επιτυγχάνοντας υψηλότερα ποσοστά ακρίβειας χωρίς επεξεργασμένες εισόδους. Αυτό δείχνει ότι τα πιο σύγχρονα γλωσσικά μοντέλα μπορούν να επωφεληθούν από τα ωμά δεδομένα, ενδεχομένως λόγω της ισχυρής τους ικανότητας στη μάθηση αντιπροσωπευτικών χαρακτηριστικών.

Υπεροχή των πιο προηγμένων μοντέλων:

- Οι Transformers υπερτερούν σταθερά έναντι των άλλων μοντέλων σε όλα τα σύνολα δεδομένων και τις μετρήσεις όταν τα δεδομένα δεν έχουν υποστεί preprocessing, ακολουθούμενα από τα GRU και LSTM. Στην περίπτωση που έχει γίνει preprocessing των δεδομένων τότε οι μετρήσεις δείχνουν τα GRU και LSTM να είναι πολύ κοντά με τους transformers δίνοντας εξίσου καλά αποτελέσματα. Το BiRNN παρουσιάζει αξιοπρεπείς επιδόσεις, αλλά γενικά υστερεί σε σχέση με τα LSTM και GRU. Το RNN υπολείπεται σταθερά σε όλες τις μετρικές και τα σύνολα δεδομένων. Τα αποτελέσματα υποδηλώνουν ότι ενώ τα παραδοσιακά μοντέλα όπως τα RNN και BiRNN εξακολουθούν να έχουν τη θέση τους, προηγμένες αρχιτεκτονικές όπως οι Transformers και τα BERT και RoBERTa, τα GRU και τα LSTM προσφέρουν ανώτερες επιδόσεις, ιδίως σε εργασίες που περιλαμβάνουν σύνθετα σύνολα δεδομένων.

Διαφορές σε περιβάλλον In και Cross Domain:

- Η απόδοση κατά τον πειραματισμό με το setup σε cross-domain περιβάλλον μειώνεται αισθητά κάτι που είναι αναμενόμενο. Στα αποτελέσματα που παρατέθηκαν στο προηγούμενο κεφάλαιο, φαίνεται ξεκάθαρα η μεγάλη πρόκληση της γενίκευσης μεταξύ διαφορετικών domains. Πιο συγκεκριμένα, η απόδοση σε όλες τις μετρικές πέφτει κατά 10 περίπου ποσοστιαίες μονάδες, όταν εκπαιδεύεται το μοντέλο με δεδομένα από διαφορετική πηγή

από το σύνολο δεδομένων ελέγχου. Ακόμα, τα πιο αποτελεσματικά και πολύπλοκα μοντέλα (LSTM, GRU και Transformers), στο cross-domain περιβάλλον φαίνεται ότι δεν καταφέρνουν να εκμεταλλευτούν τις δυνατότητες εκμάθησης τους, καθώς η επίδοση τους δε διαφέρει σημαντικά από τις πιο απλές μεθόδους (πχ Standard RNN).

Ο κώδικας υλοποίησης, μπορεί να βρεθεί στην εξής διεύθυνση ιστού: **Github**

Βιβλιογραφία

1. Ritchie, H., Mathieu, E., Roser, M., & Ortiz-Ospina, E. (2023). Internet. Retrieved from https://ourworldindata.org/internet
2. Ramanathan, T. (2024). Natural language processing. Retrieved from https://www.britannica.com/technology/natural-language-processing-computer-science
3. Dennis, M. A., & Kahn, R. (2024). Society and the Internet. Retrieved from https://www.britannica.com/technology/Internet/Society-and-the-Internet
4. Hillyer, M. (2020). How has technology changed - and changed us - in the past 20 years? Retrieved from https://www.weforum.org/agenda/2020/11/heres-how-technology-has-changed-and-changed-us-over-the-past-20-years/
5. McClain, C., Vogels, E. A., Perrin, A., Sechopoulos, S., & Rainie, L. (2021). How the internet and technology shaped Americans' personal experiences amid COVID-19. Retrieved from https://www.pewresearch.org/internet/2021/09/01/how-the-internet-and-technology-shaped-americans-personal-experiences-amid-covid-19/
6. Holdsworth, J. (2024). What is NLP? Retrieved from https://www.ibm.com/topics/natural-language-processing
7. Coursera Staff. (2024). What is Natural Language Processing? Definition and Examples. Retrieved from https://www.coursera.org/articles/natural-language-processing
8. DeepLearning.AI. (2023). Natural Language Processing. Retrieved from https://www.deeplearning.ai/resources/natural-language-processing/
9. Essa, E., Omar, K., & Alqahtani, A. (2023). Fake news detection based on a hybrid BERT and LightGBM models. Retrieved from https://link.springer.com/article/10.1007/s40747-023-01098-0
10. Mishra, A., & Sadia, H. (2023). A Comprehensive Analysis of Fake News Detection Models: A Systematic Literature Review and Current Challenges. Retrieved from https://www.mdpi.com/2673-4591/59/1/28
11. Khalil, A., Jarrah, M., & Aldwairi, M. (2023). Hybrid Neural Network Models for Detecting Fake News Articles. Retrieved from https://link.springer.com/article/10.1007/s44230-023-00055-x

12. Mishra, A., & Sadia, H. (2023). A Comprehensive Analysis of Fake News Detection Models: A Systematic Literature Review and Current Challenges. Retrieved from https://www.mdpi.com/2673-4591/59/1/28
13. Sri, K. (2023). What is artificial narrow intelligence (ANI)? Retrieved from https://venturebeat.com/ai/what-is-artificial-narrow-intelligence-ani/
14. Ai-admin. (2024). Narrow AI Examples – How Artificial Intelligence Applications Are Transforming Specific Industries. Retrieved from https://aiforsocialgood.ca/blog/narrow-ai-examples-how-artificial-intelligence-applications-are-transforming-specific-industries
15. Awan, A. A. (2023). What is Narrow AI? Retrieved from https://www.datacamp.com/blog/what-is-narrow-ai
16. Takyar. (2024). AI in gaming: Use cases, applications, implementation and trends. Retrieved from https://www.leewayhertz.com/ai-in-gaming/#AI-powered-innovations-in-graphics:-Transforming-gaming-realism-and-aesthetics
17. Saleem, M. (2022). How advanced gaming data and AI analytics is transforming game development? Retrieved from https://www.tekrevol.com/blogs/advanced-gaming-data-and-ai-analytics/
18. UNESCO. (2023). Artificial Intelligence: examples of ethical dilemmas. Retrieved from https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases
19. Boothman, B. (2020). Great promise but potential for peril. Retrieved from https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/
20. Green, B. P. (2020). Artificial Intelligence and Ethics: Sixteen Challenges and Opportunities. Retrieved from https://www.scu.edu/ethics/all-about-ethics/artificial-intelligence-and-ethics-sixteen-challenges-and-opportunities/
21. Nezafat, M. V. (2024). Fake news detection with retrieval augmented generative artificial intelligence. Retrieved from https://www.uwindsor.ca/science/computerscience/278956/msc-thesis-proposal-fake-news-detection-retrieval-augmented-generative-artificial
22. Ittoo, A. (2019). Fake news detection using machine learning. Retrieved from https://matheo.uliege.be/handle/2268.2/8416

23. Akhtar, P., Ghouri, A. M., Khan, H. U. R., Haq, M. A. ul, Awan, U., Zahoor, N., Khan, Z., & Ashraf, A. (2022). Detecting fake news and disinformation using artificial intelligence and machine learning to avoid supply chain disruptions. Retrieved from https://link.springer.com/article/10.1007/s10479-022-05015-5
24. Shushkevich, E., Alexandrov, M., & Cardiff, J. (2023). Improving Multiclass Classification of Fake News Using BERT-Based Models and ChatGPT-Augmented Data. Retrieved from https://www.mdpi.com/2411-5134/8/5/112/xml
25. Kian long Tan, Chin Poo Lee and Kian Ming Lim (2023) : A Hybrid Deep Learning Model for Enhanced Sentiment Analysis . Retrieved from https://www.mdpi.com/2076-3417/13 6/3915.
26. (2024) Compare the different Sequence models (RNN, LSTM, GRU, and Transformers). Retrieved from https://aiml.com/compare-the-different-sequence-models-rnn-lstm-gru-and-transformers/.
27. (2024) GitHub SalmaAmgarou/Classification Regression RNN GRU LSTM Text Generation with GPT2 Classification with BERT. Retrieved from https://github.com /OmarNouih/NLP LAB4.
28. Pablo-andres Buestan-Andrade (2024) Comparison of LSTM, GRU and Transformer Neural Network Architecture for Prediction of Wind Turbine Variables | SpringerLink. Retrieved from https://www.researchgate.net/publication/373532956 Comparison of LSTM GRU and Transformer Neural Network Architecture for Prediction of Wind Turbine Variables
29. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov (2019). RoBERTa Explained | Papers With Code. Retrieved from https://paperswithcode.com/paper/roberta-a-robustly-optimized-bert-pretraining
30. Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Retrieved from https://arxiv.org/abs/1907.11692
31. Noura A. Semary, Wesam Ahmed, Khalid Amin , Paweł Pławiak and Mohamed Hammad (2023) Frontiers | Improving sentiment classification using a RoBERTa-based hybrid model. Retrieved from https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2023.1292010/full