ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Επαλήθευση Συγγραφέα στα Κοινωνικά Δίκτυα

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Κονταξάκη Στυλιανού

Επιβλέπων : Σταματάτος Ευστάθιος

Μέλη εξεταστικής επιτροπής: Συμεωνίδης Παναγιώτης
Κωστούλας Θεόδωρος

Σάμος, Σεπτέμβριος 2026

Πρόλογος και ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Σταματάτο Ευστάθιο για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα.

Επίσης, θα ήθελα να ευχαριστήσω και τους γονείς μου για την υπομονή τους και την εμπιστοσύνη τους, επειδή χωρίς αυτούς δεν θα τα κατάφερνα.

© 2026

του

Κονταξάκη Στυλιανού

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

1 Εισαγωγή 1

1.1 Ανάλυση Συγγραφής (Authorship analysis) 1

1.2 Επαλήθευση Συγγραφέα 2

1.3 Στόχος εργασίας 3

2 Μηχανική Μάθηση και Επεξεργασία Φυσικής Γλώσσας 4

2.1 Εισαγωγή στην Μηχανική Μάθηση 4

2.1.1 Σχέσεις με άλλους τομείς 5

2.1.2 Υπέρ-προσαρμογή (Overfitting) 7

2.1.3 Μεροληψία (Bias) 8

2.2 Κατηγορίες Μηχανικής Μάθησης 8

2.2.1 Επιβλεπόμενη Μάθηση 8

2.2.2 Μη Επιβλεπόμενη Μηχανική Μάθηση 9

2.2.3 Ενισχυτική μάθηση (Reinforcement learning): 10

2.3 Αλγόριθμοι Μηχανικής Μάθησης 11

2.3.1 Δέντρα Αποφάσεων 11

2.3.2 Μηχανές Διανυσμάτων Υποστήριξης (Support-Vector Machines) 12

2.3.3 Ανάλυση Παλινδρόμησης (Regression Analysis) 13

2.3.4 Δίκτυα Bayes (Bayesian Networks) 14

2.4 Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP) 15

2.4.1 Τμηματοποίηση λέξεων (Word Segmentation - Tokenization) 16

2.4.2 Λημματοποίηση (Lemmatization) 16

2.4.3 Αποκοπή καταλήξεων (Stemming) 16

2.4.4 Μορφολογική ανάλυση (Morphological segmentation) 16

2.4.5 Επισήμανση μερών του λόγου (Part-of-speech tagging – POS) 17

2.4.6 Αναγνώριση Οντοτήτων με όνομα (Name Entity Recognition – NER) 17

2.4.7 Ανάλυση συναισθήματος (Sentiment analysis): 17

2.4.8 Word Embedding (Ενσωμάτωμα λέξης) 18

3 Βαθιά Νευρωνικά Δίκτυα 22

3.1 Έννοια 22

3.2 Οπισθοδιάδοση (Backpropagation) 23

3.3 Αλγόριθμοι Βελτιστοποίησης Νευρωνικών Δικτύων 23

3.3.1 Gradient Descent 23

3.3.2 Adam (Adaptive Moment Estimation) ... *25*
3.4 Νευρωνικά Δίκτυα με Ανατροφοδότηση (Recurrent Neural Network - RNN) ... 28
3.5 Bidirectional RNN ... 30
3.6 Long Short-Term Memory (LSTM) ... 32
3.7 Sequence-to-Sequence (seq2seq) ... 34
3.8 Μηχανισμός Attention με RNN ... 35
3.9 Transformers ... 37
3.10 BERT ... 40
3.10.1 APXITEKTONIKH ... *40*
3.11 RoBERTa ... 41
3.12 DeBERTa ... 43
4 Εργασίες επαλήθευσης συγγραφέα και διαγωνισμοί PAN ... 45
4.1 Εισαγωγή ... 45
4.1.1 Intrinsic verification models vs Extrinsic verification models ... *45*
4.1.2 Eager Learning models vs Lazy Learning models ... *46*
4.1.3 Profile-based models vs Instance-based models ... *47*
4.2 Η μέθοδος General Impostors του Seidman [24] ... 49
4.3 Η βελτιωμένη παραλλαγή της μεθόδου Impostors των Koppel & Winter [25] ... 51
4.4 Παραλλαγή της μεθόδου General Impostors των Potha & Stamatatos [26] ... 53
4.5 Διαγωνισμοί Επαλήθευσης Συγγραφέα PAN ... 55
4.5.1 PAN 2015 ... *56*
4.5.2 PAN 2020 ... *59*
4.5.3 PAN 2021 ... *64*
4.5.4 PAN 2022 ... *70*
4.6 Σύγχρονες μέθοδοι ... 80
4.6.1 Linguistically Informed Prompting (LIP) ... *80*
4.6.2 Learning Universal Authorship Representations (LUAR) ... *84*
4.6.3 Residualized Similarity ... *87*
5 Πειράματα και τελικά αποτελέσματα ... 96
5.1 Μετρικές αξιολόγησης ... 96
5.2 Περιγραφή των δεδομένων ... 97
5.3 Χωρισμός των δεδομένων ... 98
5.4 Τελικά αποτελέσματα ... 99
5.4.1 Μέθοδος baseline με TF-IDF ... *99*
5.4.2 General Impostors Method ... *101*
5.4.3 Extrinsic Instance-Based model ... *102*

5.4.4 Extrinsic Profile-Based model 103
5.4.5 Frozen DeBERTa 105
5.4.6 Μέθοδος LUAR 106
5.4.7 Μέθοδος Residualalized Similarity 108
6 Συμπεράσματα 111
Βιβλιογραφία 113
Παράρτημα Ι 116

Λίστα Σχημάτων

Εικόνα 1: Η μπλε γραμμή είναι ένα παράδειγμα υπέρ-προσαρμογής (overfitting) μιας γραμμικής συνάρτησης λόγω τυχαίου θορύβου στα δεδομένα. 7

Εικόνα 2: Στην ενισχυτική μάθηση, ένας πράκτορας (agent) εκτελεί ενέργειες (actions) μέσα σε ένα περιβάλλον (environment). Αυτές οι ενέργειες παράγουν μια ανταμοιβή (reward) και/ή μια αναπαράσταση της κατάστασης (state representation), η οποία επιστρέφεται στον πράκτορα ως ανατροφοδότηση (feedback). 11

Εικόνα 3: Δέντρο αποφάσεων που δείχνει την πιθανότητα επιβίωσης των επιβατών του Τιτανικού. 12

Εικόνα 4: Απεικόνιση μιας γραμμικής παλινδρόμησης σε ένα σύνολο δεδομένων. 14

Εικόνα 5: Ένα απλό δίκτυο Bayes. Η βροχή επηρεάζει το αν θα ενεργοποιηθεί ο εκτοξευτήρας νερού (sprinkler), και τόσο η βροχή όσο και ο εκτοξευτήρας επηρεάζουν το αν το γρασίδι είναι βρεγμένο. 14

Εικόνα 6: Τύποι word embeddings. 19

Εικόνα 7: Παράδειγμα word2vec. 20

Εικόνα 8: Gradient Descent 24

Εικόνα 9: Μεγάλα βήματα vs μικρά βήματα για τον ρυθμό μάθησης. 25

Εικόνα 10: Δομή RNN 28

Εικόνα 11: Δομή RNN (2) 29

Εικόνα 12: Αρχιτεκτονική Bidirectional RNN. 31

Εικόνα 13: Δομή LSTM. 33

Εικόνα 14: Βασική αρχιτεκτονική seq2seq. 35

Εικόνα 15: Γραφική αναπαράσταση του μοντέλου encoder-decoder με attention. Πηγή: [18] 36

Εικόνα 16: Παράδειγμα μηχανισμού attention με RNN. 36

Εικόνα 17: Το μοντέλο Transformer. Πηγή: [12]. 38

Εικόνα 18: Παράδειγμα word embedding DeBERTa 43

Εικόνα 19: Intrinsic verification model (αριστερά) vs Extrinsic verification model (δεξιά). Πηγή: [22]. 46

Εικόνα 20: Instance-based model (αριστερά) vs Profile-based model (δεξιά). Πηγή: [22] 48

Εικόνα 21: Ψευδοκώδικας της παραλλαγής μεθόδου General Impostors των Potha και Stamatatos. Οι τροποποιήσεις σε σχέση με την αρχική μέθοδο επισημαίνονται με μπλε χρώμα, ενώ οι αντίστοιχες εντολές της αρχικής μεθόδου εμφανίζονται ως σχόλια. Πηγή: [26]. 55

Εικόνα 22: Κάποια κείμενα είναι επισημασμένα ως προς τον συγγραφέα και υπάρχει ένα κείμενο αγνώστου συγγραφέα. Πηγή: [28]. 56

Εικόνα 23: Ο δείκτης c@1 για το μικρό σύνολο δεδομένων. Πηγή: [29]. 63

Εικόνα 24: Αριστερά παρουσιάζεται η κατανομή των προβλέψεων για τα καλύτερα συστήματα της κάθε ομάδας. Δεξιά φαίνεται το ίδιο, αλλά με τις μεθόδους του καθενός. Πηγή: [30] 67

Εικόνα 25: Παρουσιάζονται οι καμπύλες ακρίβειας (precision) και ανάκλησης (recall), των παραπάνω συστημάτων, καθώς και ένας naive meta classifier, που προβλέπει τον μέσο όρο των σκορ όλων των συστημάτων. Πηγή: [30] 68

Εικόνα 26: Ο δείκτης c@1 για τις μη-απαντήσεις. Πηγή: [30] 69

Εικόνα 27: Τα prompts που χρησιμοποιήθηκαν. Πηγή: [32] 81

Εικόνα 28: Η Analysis και Answer είναι η έξοδος του GPT-4. Μόνο η μέθοδος LIP πέτυχε σωστά ότι τα δύο κείμενα είναι γραμμένα από τον ίδιο συγγραφέα. Στο κείμενο με υπογραμμισμένο πορτοκαλί φαίνονται οι διαφορές συγκριτικά με την προτροπή χωρίς καθοδήγηση. Στο κείμενο με το μπλε φαίνεται η μέθοδος LIP. Πηγή: [32] 82

Εικόνα 29: Αριστερά: Αποτελέσματα επαλήθευσης συγγραφέα στο δείκτη ακρίβειας (Accuracy) στο Blog dataset. Δεξιά: Τα αποτελέσματα στο δείκτη ακρίβειας (Accuracy) στο Email dataset. Πηγή: [32] 83

Εικόνα 30: Αρχιτεκτονική LUAR. Πηγή: [33] 85

Εικόνα 31: Αρχιτεκτονική Residual Similarity. Τα πράσινα κουτιά δείχνουν ότι ενημερώνονται κατά την εκπαίδευση. Στην αριστερή πλευρά φαίνεται η μέθοδος κατά τη χρήση της πρόβλεψης. Το τελικό σκορ είναι απλώς το άθροισμα του συνημιτόνου ομοιότητας και του προβλεπόμενου υπολείμματος. Πηγή: [34] 90

Εικόνα 32: Η κατανομή των τιμών της εμπιστοσύνης ερμηνευσιμότητας για τις προβλέψεις χρησιμοποιώντας Residual Similarity μαζί με LUAR στο σύνολο δεδομένων του Reddit. Πηγή: [34] 94

Εικόνα 33: Καμπύλη ROC για την μέθοδο TF-IDF. 100

Εικόνα 34: Καμπύλη ROC για την μέθοδο του General Impostor. 101

Εικόνα 35: Καμπύλη ROC για την μέθοδο Extrinsic Instance-Based. 103

Εικόνα 36: Καμπύλη ROC για την μέθοδο Extrinsic Profile-Based. 104

Εικόνα 37: Καμπύλη ROC για την μέθοδο Frozen DeBERTa. 106

Εικόνα 38: Καμπύλη ROC για την μέθοδο LUAR. 108

Εικόνα 39: Καμπύλη ROC για την μέθοδο Residualized Similarity. 110

Λίστα Πινάκων

Πίνακας 1: Σύνοψη των σωμάτων του PAN για επαλήθευση συγγραφέα. [28] 57

Πίνακας 2: Κατανομή των συμμετεχόντων στο PAN στις κατηγορίες μοντέλων επαλήθευσης. Πηγή: [28] 58

Πίνακας 3: Παραπάνω φαίνονται τα καλύτερα αποτελέσματα που επιτεύχθηκαν από τους συμμετέχοντες για κάθε κείμενο και ο μέσος όρος της. απόδοσης για όλους τους συμμετέχοντες του PAN. Πηγή: [28] ... 59

Πίνακας 4: Αποτελέσματα αξιολόγησης. Πηγή: [29] 62

Πίνακας 5: Τα δύο κορυφαία συστήματα, αποκλείοντας τα ζεύγη, τα οποία ο boenninghoff20-large υπέβαλε μη-απάντηση. [29] 63

Πίνακας 6: Αποτελέσματα αξιολόγησης ομάδων. Πηγή: [30] 65

Πίνακας 7: Παραπάνω φαίνονται οι επιδόσεις των συστημάτων με βάση τον δείκτη c@1. Τα σκορ για τα σύνολα δεδομένων του 2020 είναι χαμηλότερα σε όλα (εκτός του ikae21) από ότι του 2021. Πηγή: [30] .. 66

Πίνακας 8: Δείχνει μια σύνοψη των δεδομένων, που χρησιμοποιήθηκαν για την εκπαίδευση και δοκιμή (training and testing) των διαφορετικών πτυχών της επαλήθευσης συγγραφέα. Πηγή: [31] 70

Πίνακας 9: Σύνοψη των βασικών χαρακτηριστικών των υποβληθέντων μεθόδων (POS: Part of Speech tags, NEs: Named Entities, SOM: Self-organizing Maps, FC: Fully-Connected Layer). Πηγή: [31] 73

Πίνακας 10: Αποτελέσματα όλων των συμμετεχόντων. Πηγή: [31] 74

Πίνακας 11: Τα αποτελέσματα αξιολόγησης ως προς τους έξι συνδυασμούς τύπων λόγου. Πηγή: [31] 75

Πίνακας 12: a) Αριθμός θετικών, αρνητικών και μη-απαντήσεων του κάθε μοντέλου επαλήθευσης. b) Χρόνος εκτέλεσης του κάθε μοντέλου. Πηγή: [31] 76

Πίνακας 13: a) Αποτελέσματα αξιολόγησης των κορυφαίο μοντέλων, που υποβλήθηκαν στο PAN 2021, στο dataset 2022 b) Θετικές, αρνητικές και μη-απαντήσεις που παρείχαν αυτά τα μοντέλα. Πηγή: [31] 78

Πίνακας 14: Αποτελέσματα zero-shot για το προτεινόμενο μοντέλο, ένα συνελεκτικό μοντέλο και ένα μοντέλο TF-IDF. Πηγή: [33] 86

Πίνακας 15: Απόδοση του μοντέλου στα 3 σύνολα δεδομένων (Reddit, Amazon, Fanfiction). Το Pikabu είναι ρώσικο dataset και χρησιμοποιούν LUAR-RU για το νευρωνικό μοντέλο στην RS. Τα καλύτερα αποτελέσματα είναι με bold. Πηγή: [34] 93

Πίνακας 16: Συγκεντρωτικός πίνακας με τα πειραματικά αποτελέσματα. 110

Περίληψη

Στην σημερινή εποχή, υπάρχουν πάρα πολλά μέσα κοινωνικής δικτύωσης (Facebook, Twitter, Reddit, Instagram κλπ.). Μέσα από αυτά, οι άνθρωποι ανταλλάσσουν μηνύματα. Κάποιες φορές, όμως, υποψιαζόμαστε ότι κάποιο κείμενο δεν το έχει γράψει ο πραγματικός συγγραφέας. Παρατηρούμε διαφορές στο ύφος γραφής του, στο λεξιλόγιό του. Αυτό το πρόβλημα έρχεται να λύσει η επαλήθευση συγγραφέα (author verification), που βρίσκεται στο ευρύτερο πλαίσιο της «ανάλυσης συγγραφέα» (authorship analysis). Στην επαλήθευση συγγραφέα δίνονται κάποια κείμενα, από γνωστό συγγραφέα, και ένα άγνωστο κείμενο (από άγνωστο συγγραφέα). Στόχος είναι να αποφασίσουμε εάν αυτό το άγνωστο κείμενο, ανήκει στον γνωστό συγγραφέα, δηλαδή εάν είναι ο ίδιος που έχει γράψει τα γνωστά κείμενα ή όχι. Για αυτό το λόγο, είναι ένα δυαδικό πρόβλημα ταξινόμησης (binary classification task), που απαντάται με ένα «ναι» ή «όχι». Η επαλήθευση συγγραφέα αναλύει το ύφος του στα κείμενα, όπως το λεξιλόγιό που χρησιμοποιεί, τη σύνταξη, καθώς και τα σημεία στίξης. Αποτελεί ένα ιδιαίτερα απαιτητικό πρόβλημα όταν έχει να κάνει με μικρά κείμενα, όπως είναι για παράδειγμα το Twitter. Αυτό συμβαίνει, διότι δεν μπορούν εξαχθούν πολλά χαρακτηριστικά του ύφους του συγγραφέα, συγκριτικά με ένα π.χ. άρθρο, όπου εκεί τα κείμενα είναι μεγάλα. Το πεδίο της επαλήθευσης συγγραφέα είναι ιδιαίτερα ευρύ, καθώς χρησιμοποιείται στον τομέα της ασφάλειας, στην αντιμετώπιση της πνευματικής ιδιοκτησίας, στον έλεγχο ψευδών ειδήσεων, στην αποκάλυψη ψεύτικων προφίλ στα μέσα κοινωνικής δικτύωσης (διπλούς λογαριασμούς) και σε πολλά άλλα.

Για την πειραματική αξιολόγηση χρησιμοποιήθηκε ένα σύνολο δεδομένων από την πλατφόρμα Twitter, το οποίο περιλαμβάνει 100 συγγραφείς και 20.000 tweets, οργανωμένα σε έγγραφα και ζεύγη κειμένων για την επαλήθευση της συγγραφικής ταυτότητας. Στο πλαίσιο της εργασίας υλοποιήθηκαν και αξιολογήθηκαν επτά διαφορετικές μέθοδοι, οι οποίες περιλαμβάνουν παραδοσιακές στυλομετρικές προσεγγίσεις, παραλλαγές της μεθόδου Impostors και σύγχρονα μοντέλα βαθιάς μάθησης. Τα πειραματικά αποτελέσματα έδειξαν ιδιαίτερα υψηλή απόδοση, με τη μέθοδο LUAR να επιτυγχάνει Accuracy 99% και AUROC 0,99, αναδεικνύοντας την αποτελεσματικότητα των σύγχρονων αναπαραστάσεων συγγραφικού ύφους. Συμπερασματικά, η εργασία καταδεικνύει ότι η επαλήθευση συγγραφέα μπορεί να επιτευχθεί με υψηλή ακρίβεια ακόμη και σε σύντομα κείμενα κοινωνικών δικτύων, ενώ η συγκριτική αξιολόγηση αναδεικνύει τα πλεονεκτήματα και τους περιορισμούς των διαφορετικών προσεγγίσεων.

Λέξεις Κλειδιά: *επαλήθευση συγγραφέα, δυαδικό πρόβλημα ταξινόμησης, λεξιλόγιο, ύφος, κείμενο, twitter.*

Abstract

Nowadays, there are numerous social media platforms, such as Facebook, Twitter, Reddit, and Instagram. Through these platforms, people communicate and exchange messages. However, in some cases, we may suspect that a particular text was not written by its claimed author. For instance, differences may be observed in the author's writing style or vocabulary. Author Verification (AV), which falls within the broader field of Authorship Analysis, aims to address this problem. In an Author Verification task, a set of texts written by a known author is provided along with an unknown text, whose authorship is uncertain. The objective is to determine whether the unknown text was written by the same author who produced the known texts. Therefore, Author Verification can be formulated as a binary classification task, in which the answer is either "yes" (same author) or "no" (different author). It examines various characteristics of an individual's writing style, including vocabulary usage, syntactic patterns, and punctuation. The task becomes particularly challenging when dealing with short texts, such as those posted on Twitter. This is because short texts provide limited linguistic and stylistic information from which the author's writing characteristics can be extracted, especially when compared with longer texts, such as articles. The field of Author Verification has a wide range of applications. It can be employed in security-related tasks, intellectual property and plagiarism investigations, fake-news detection, and the identification of fake or duplicate profiles on social media platforms, among many other applications.

For the experimental evaluation, a dataset from the Twitter platform was used, consisting of 100 authors and 20,000 tweets, organized into documents and text pairs for authorship verification. As part of this thesis, seven different methods were implemented and evaluated, including traditional stylometric approaches, variations of the Impostors method, and modern deep learning models. The experimental results demonstrated particularly high performance, with the LUAR method achieving an accuracy of 99% and an AUROC score of 0.99, highlighting the effectiveness of modern authorship style representations. In conclusion, this study demonstrates that authorship verification can achieve high accuracy even with short texts from social media platforms, while the comparative evaluation highlights the advantages and limitations of the different approaches.

Keywords: *author verification, binary classification task, writing style, twitter*

1

Εισαγωγή

1.1 Ανάλυση Συγγραφής (Authorship analysis)

Τα τελευταία χρόνια έχουν αυξηθεί ραγδαία τα δεδομένα στο διαδίκτυο. Ο καθένας μπορεί να ανταλλάσσει πληροφορίες μέσω των κοινωνικών δικτύων (Facebook, Twitter, Reddit κλπ.). Σε αυτά, μπορούμε να εκφράσουμε δημόσια την σκέψη μας, τον προβληματισμό μας, να διαβάσουμε μια είδηση, να προωθήσουμε την επιχείρησή μας, καθώς και πολλά ακόμα. Αυτές οι πληροφορίες, ή αλλιώς δεδομένα, είναι σε μορφή γραπτού κειμένου και μπορούν να χρησιμοποιηθούν για να αυτοματοποιήσουμε κάποιες διαδικασίες. Στο Twitter υπάρχουν πάρα πολύ λογαριασμοί, οι οποίοι είναι ψεύτικοι (fake). Το ίδιο, λοιπόν, άτομο για να προωθήσει την ιδέα του, μπορεί να έχει δημιουργήσει πολλαπλούς λογαριασμούς, χωρίς να το καταλαβαίνουμε. Ο στόχος, λοιπόν, είναι η αναγνώριση του συγγραφέα μέσα σε ένα κείμενο. Η ανάλυση συγγραφέα (authorship analysis) αποτελεί ένα ευρύ επιστημονικό πεδίο. Μέσω της επεξεργασίας φυσικής γλώσσας, της ανάκτησης πληροφορίας και της μηχανικής μάθησης, η ανάλυση συγγραφέα χρησιμοποιεί μεθόδους για να αναλύσει το ύφος του συγγραφέα και να εξάγει κάποια στιλιστικά χαρακτηριστικά του. Έτσι, προσπαθεί να αναγνωρίσει ή να επαληθεύσει την ταυτότητα ενός συγγραφέα. Η ανάλυση συγγραφής χωρίζεται σε 4 κατηγορίες:

- Authorship attribution: στόχος είναι η ταυτοποίηση του συγγραφέα ενός κειμένου, από ένα σύνολο υποψήφιων συγγραφέων. Χρησιμοποιείται σε λογοτεχνικά έργα (π.χ. για ταυτοποίηση συγγραφέα ενός έργου), σε ανίχνευση λογοκλοπής σε ακαδημαϊκά ή δημόσια κείμενα, σε παραβιάσεις πνευματικών δικαιωμάτων κ.α.
- Author verification: δίνοντας κάποια γνωστά κείμενα (από έναν γνωστό συγγραφέα) και ένα άγνωστο κείμενο (από άγνωστο συγγραφέα), στόχος είναι να αποφασιστεί εάν το άγνωστο κείμενο ανήκει στον γνωστό συγγραφέα ή όχι. Αποτελεί δυαδικό πρόβλημα ταξινόμησης. Χρησιμοποιείται όταν θέλουμε να ελέγξουμε αν δυο κείμενα γράφτηκαν από τον ίδιο συγγραφέα. Ένα παράδειγμα

χρήσης αποτελεί η περίπτωση που θέλουμε να ελέγξουμε εάν δυο λογαριασμοί του Twitter ανήκουν στο ίδιο άτομο (μέσα από τα κείμενα δημοσίευσης).

- Author profiling: προσπαθεί να βρει τα δημογραφικά χαρακτηριστικά ενός συγγραφέα. Δηλαδή, το φύλο του, την ηλικία του, την μητρική γλώσσα, εθνικότητα. Για παράδειγμα, χρησιμοποιείται σε περίπτωση που θέλουμε να αναγνωρίσουμε την ηλικία ή το φύλο των συγγραφέων του Twitter, να προωθήσουμε εξατομικευμένες διαφημίσεις σε κάποιο κοινό κ.α.
- Authorship obfuscation: δίνοντας κάποια κείμενα που ανήκουν σε έναν συγκεκριμένο συγγραφέα, τροποποίησε κάποιο ή κάποια, έτσι ώστε ο πραγματικός συγγραφέας να μην μπορεί να ταυτοποιηθεί. Στόχος, δηλαδή, είναι να κρυφτεί η ταυτότητα του συγγραφέα. Παραδείγματα χρήσης είναι όταν κάποιος θέλει να διατηρήσει την ανωνυμία του στο διαδίκτυο, ή ακόμα και σε περίπτωση αποφυγής λογοκλοπής.

Η βασική υπόθεση είναι ότι ο κάθε συγγραφέας έχει ένα σταθερό, μοναδικό στυλ γραφής. Άρα, αν συγκρίνουμε δύο κείμενα και το στυλ τους είναι αρκετά συνεπές, μπορούμε να υποθέσουμε ότι ανήκουν στον ίδιο συγγραφέα.

1.2 Επαλήθευση Συγγραφέα

Η επαλήθευση συγγραφέα αποτελεί μια ειδική κατηγορία της ανάλυσης συγγραφέα. Δίνοντας κάποια γνωστά κείμενα από έναν συγκεκριμένο συγγραφέα και ένα μοναδικό άγνωστο κείμενο, πρέπει να απαντηθεί το ερώτημα με ένα ναι ή όχι: ανήκει το άγνωστο κείμενο σε αυτόν που έγραψε τα γνωστά κείμενα; Αποτελεί δυαδικό πρόβλημα ταξινόμησης, αφού δέχεται μόνο 2 απαντήσεις (ναι/όχι) και προσπαθεί να εκτιμήσει πόσο πιθανό είναι το άγνωστο κείμενο να ανήκει στον συγγραφέα των γνωστών κειμένων. Χρησιμοποιούνται μέθοδοι μηχανικής ή βαθιάς μάθησης για να ανακτηθεί το ύφος του συγγραφέα.

Μέσα από την ανάλυση των κειμένων εξάγονται στιλομετρικά χαρακτηριστικά, που μπορούν να κατηγοριοποιηθούν είτε σε λεξικά χαρακτηριστικά, όπως η συχνότητα των λέξεων, ν-γράμματα λέξεων, κατανομή μήκους λέξεων, μετρήσεις λεξικού πλούτου, είτε σε επίπεδο χαρακτήρων, όπως η συχνότητα χρήσης σημείων στίξης, ν-γράμματα χαρακτήρων. Τέτοια χαρακτηριστικά είναι ανεξαρτήτου γλώσσας. Υπάρχουν και πιο σύνθετα χαρακτηριστικά που απαιτούν την ανάλυση των κειμένων με εργαλεία επεξεργασίας φυσικής γλώσσας, ώστε να εξαχθούν συντακτικά χαρακτηριστικά (π.χ. συχνότητες μερών του λόγου – POS, συντακτικοί κανόνες, συντακτικά ν-γράμματα κλπ.), καθώς και σημασιολογικά χαρακτηριστικά (π.χ. σημασιολογικές εξαρτήσεις, χρήση συνωνύμων κλπ.). Σε περίπτωση που όλα τα κείμενα μοιράζονται κάποιες ιδιότητες, για παράδειγμα, ανήκουν στο ίδιο είδος (π.χ. email), βρίσκονται στην ίδια θεματική ενότητα (π.χ. πωλήσεις υπολογιστών), είναι σε μια συγκεκριμένη μορφή (π.χ. html), είναι πολύ αποτελεσματικό να καθοριστούν συγκεκριμένα χαρακτηριστικά για αυτό το συγκεκριμένο πεδίο.

Οι εφαρμογές του περιλαμβάνουν ένα ευρύ φάσμα. Μπορεί να χρησιμοποιηθεί για την ανίχνευση παραβιασμένων λογαριασμών στα κοινωνικά δίκτυα, σε φιλτράρισμα emails (spam ή όχι), στην ανίχνευση ψευδών ειδήσεων, ψεύτικων κριτικών (οι κριτικές να ανήκουν σε ένα άτομο), στην αντιμετώπιση της πνευματικής ιδιοκτησίας

Η εργασία της επαλήθευσης συγγραφέα στο Twitter αποτελεί ένα απαιτητικό πρόβλημα, διότι τα κείμενα που υπάρχουν είναι μικρά. Έτσι, τα στιλιστικά χαρακτηριστικά που εξάγονται είναι μικρότερα συγκριτικά με ένα μεγάλο κείμενο (π.χ. άρθρο).

1.3 Στόχος εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η μελέτη και η πειραματική αξιολόγηση διαφορετικών μεθόδων για το πρόβλημα της επαλήθευσης συγγραφέα (Author Verification) σε σύντομα κείμενα που προέρχονται από την πλατφόρμα κοινωνικής δικτύωσης Twitter. Ειδικότερα, εξετάζεται κατά πόσο είναι δυνατό να προσδιοριστεί εάν δύο κείμενα έχουν γραφτεί από τον ίδιο συγγραφέα ή από διαφορετικούς συγγραφείς, αξιοποιώντας χαρακτηριστικά του συγγραφικού ύφους.

Για τον σκοπό αυτό, δημιουργείται ένα σύνολο δεδομένων από δημοσιεύσεις χρηστών του Twitter, οι οποίες οργανώνονται σε ζεύγη κειμένων ίδιου συγγραφέα και διαφορετικών συγγραφέων.

Στο πλαίσιο της εργασίας υλοποιούνται και συγκρίνονται διαφορετικές προσεγγίσεις επαλήθευσης συγγραφέα, οι οποίες καλύπτουν τόσο παραδοσιακές μεθόδους βασισμένες σε χαρακτηριστικά κειμένου, όσο και σύγχρονες μεθόδους βαθιάς μάθησης και αναπαραστάσεων συγγραφικού ύφους. Μεταξύ αυτών εξετάζονται η προσέγγιση TF-IDF με character 4-grams, η μέθοδος Impostors του Koppel & Winter, παραλλαγές της μεθόδου Impostors, όπως Extrinsic Instance-Based και Extrinsic Profile-Based, καθώς και νευρωνικές προσεγγίσεις που βασίζονται σε μοντέλα όπως το DeBERTa, το LUAR (Learning Universal Authorship Representations) και το Residualized Similarity.

Τελικός στόχος είναι η συγκριτική αξιολόγηση των παραπάνω μεθόδων στο ίδιο σύνολο δεδομένων, ώστε να διερευνηθεί η αποτελεσματικότητά τους στην αναγνώριση του συγγραφικού ύφους σε σύντομα κείμενα κοινωνικών δικτύων. Η αξιολόγηση πραγματοποιείται με κοινές μετρικές, όπως Accuracy, Precision, Recall, F1-score και AUROC, μέσω κατάλληλου κατωφλίου (threshold).

2

Μηχανική Μάθηση και Επεξεργασία Φυσικής Γλώσσας

2.1 Εισαγωγή στην Μηχανική Μάθηση

Η μηχανική μάθηση (Machine Learning - ML) είναι ένας τομέας της τεχνητής νοημοσύνης, που ασχολείται με την ανάπτυξη και τη μελέτη στατιστικών αλγορίθμων, οι οποίοι μπορούν να μαθαίνουν από δεδομένα και να προβλέπουν σε άγνωστα δεδομένα. Η μηχανική μάθηση βρίσκει εφαρμογή σε πολλούς τομείς, όπως η επεξεργασία φυσικής γλώσσας (Natural Language Processing), η υπολογιστική όραση (Computer Vision), η αναγνώριση ομιλίας (Speech Recognition), το φιλτράρισμα ηλεκτρονικού ταχυδρομείου (email filtering), η γεωργία και η ιατρική. Οι στατιστικές μέθοδοι και οι μέθοδοι μαθηματικής βελτιστοποίησης (μαθηματικός προγραμματισμός) αποτελούν τα θεμέλια της μηχανικής μάθησης. Η εξόρυξη δεδομένων (data mining) είναι ένας συναφής τομέας μελέτης που εστιάζει στην εξερευνητική ανάλυση δεδομένων (exploratory data analysis - EDA) μέσω μη επιβλεπόμενης μάθησης (unsupervised learning). [1]

Στη σύγχρονη εποχή, οι αλγόριθμοι μηχανικής μάθησης χωρίζονται σε τρεις κύριες κατηγορίες:

1) **Επιβλεπόμενη μάθηση (Supervised Learning)**

   Έχει στόχους την ταξινόμηση (classification) και την παλινδρόμηση (regression).

2) **Μη επιβλεπόμενη μάθηση (Unsupervised Learning)**

   Έχει στόχους την ομαδοποίηση (clustering), τη μείωση διαστάσεων (dimensionality reduction) και τους κανόνες συσχέτισης (association rule).

3) **Ενισχυτική μάθηση (Reinforcement Learning)**

   Επικεντρώνεται στη λήψη αποφάσεων που πρέπει να γίνονται με βάση προηγούμενες, άγνωστες εμπειρίες και χωρίζεται σε μεθόδους με ή χωρίς μοντέλο (model based methods / model free methods).

Ένας βασικός στόχος ενός μαθησιακού αλγορίθμου (learner) είναι η γενίκευση από την εμπειρία του. Η γενίκευση, σε αυτό το πλαίσιο, είναι η ικανότητα μιας μηχανής μάθησης να εκτελεί με ακρίβεια σε νέα, άγνωστα παραδείγματα ή εργασίες, αφού έχει εκπαιδευτεί σε ένα σύνολο δεδομένων εκπαίδευσης. Τα παραδείγματα εκπαίδευσης προέρχονται από κάποια γενικά άγνωστη κατανομή πιθανότητας, η οποία θεωρείται αντιπροσωπευτική του συνόλου των περιπτώσεων.
Ο αλγόριθμος μάθησης πρέπει να φτιάξει ένα γενικό μοντέλο, το οποίο να του επιτρέπει να παράγει επαρκώς ακριβείς προβλέψεις σε νέες περιπτώσεις.

Η υπολογιστική ανάλυση των αλγορίθμων μηχανικής μάθησης και της απόδοσής τους αποτελεί κλάδο της θεωρητικής επιστήμης των υπολογιστών, γνωστό ως «θεωρία υπολογιστικής μάθησης» (computational learning theory), που βασίζεται στο μοντέλο της κατά προσέγγισης σωστής μάθησης (Probably Approximately Correct - PAC).
Επειδή τα σύνολα εκπαίδευσης είναι πεπερασμένα και το μέλλον αβέβαιο, η θεωρία μάθησης, συχνά, δεν εγγυάται για την απόδοση των αλγορίθμων.
Αντίθετα, χρησιμοποιεί πιθανοτικές εκτιμήσεις της απόδοσης.
Μία γνωστή μέθοδος ποσοτικοποίησης του σφάλματος γενίκευσης είναι η ανάλυση μεροληψίας-διακύμανσης (bias-variance decomposition).

Για την καλύτερη απόδοση σε όρους γενίκευσης, η πολυπλοκότητα της υπόθεσης πρέπει να ταιριάζει με την πολυπλοκότητα της συνάρτησης που έχει τα δεδομένα.
Αν η υπόθεση είναι λιγότερο περίπλοκη από την συνάρτηση, τότε το μοντέλο υπό-προσαρμόζεται (underfitting).
Αν πολυπλοκότητα του μοντέλου αυξηθεί, τότε το σφάλμα εκπαίδευσης μειώνεται.

Αλλά, αν η υπόθεση γίνει υπερβολικά πολύπλοκη, τότε το μοντέλο υπέρ-προσαρμόζεται (overfitting) και η γενίκευση μειώνεται.

Εκτός από τα όρια απόδοσης, οι θεωρητικοί της μάθησης μελετούν και την υπολογιστική πολυπλοκότητα (time complexity) και τη δυνατότητα εκμάθησης (feasibility).
Στη θεωρία υπολογιστικής μάθησης, μια διαδικασία θεωρείται υπολογιστικά εφικτή αν μπορεί να πραγματοποιηθεί σε πολυωνυμικό χρόνο (polynomial time).

2.1.1 Σχέσεις με άλλους τομείς

2.1.1.1 Τεχνητή Νοημοσύνη (Artificial Intelligence)

Ως επιστημονική προσπάθεια, η μηχανική μάθηση (machine learning) αναπτύχθηκε μέσα από την αναζήτηση για τεχνητή νοημοσύνη (AI). Στα πρώτα χρόνια της τεχνητής νοημοσύνης, ως ακαδημαϊκού κλάδου, ορισμένοι ερευνητές ενδιαφέρθηκαν να δημιουργήσουν μηχανές που μαθαίνουν από δεδομένα. Προσπάθησαν να προσεγγίσουν το πρόβλημα χρησιμοποιώντας διάφορες συμβολικές μεθόδους, όπως αυτό που τότε ονομαζόταν «νευρωνικά δίκτυα». Η πιθανολογική συλλογιστική (probabilistic reasoning) χρησιμοποιήθηκε επίσης, ιδιαίτερα στην αυτοματοποιημένη ιατρική διάγνωση.

Ωστόσο, η αυξανόμενη έμφαση στην λογική, προσέγγιση βασισμένη στη γνώση (knowledge-based approach), προκάλεσε ένα ρήγμα μεταξύ της τεχνητής νοημοσύνης και της μηχανικής μάθησης. Τα πιθανολογικά συστήματα αντιμετώπιζαν θεωρητικά και πρακτικά προβλήματα σχετικά με τη συλλογή και αναπαράσταση δεδομένων.

Μέχρι το 1980, τα ειδικά συστήματα (expert systems) είχαν κυριαρχήσει στην τεχνητή νοημοσύνη, ενώ η στατιστική ήταν εκτός εύνοιας.

Η εργασία πάνω στη γνωσιακή μάθηση συνέχισε να αναπτύσσεται μέσα στον χώρο της τεχνητής νοημοσύνης, οδηγώντας στη δημιουργία της επαγωγικής λογικής προγραμματισμού (inductive logic programming - ILP).
Ωστόσο, η πιο στατιστική προσέγγιση της έρευνας ήταν εκτός του πεδίου της τεχνητής νοημοσύνης, προς τους τομείς της αναγνώρισης προτύπων και της ανάκτησης πληροφοριών.
Η έρευνα στα νευρωνικά δίκτυα είχε εγκαταλειφθεί περίπου την ίδια περίοδο με την τεχνητή νοημοσύνη και την επιστήμη των υπολογιστών.
Αυτή η γραμμή έρευνας συνεχίστηκε εκτός του πεδίου της τεχνητής νοημοσύνης, με το όνομα "connectionism", από ερευνητές άλλων επιστημονικών πεδίων, όπως οι John Hopfield, David Rumelhart και Geoffrey Hinton.
Η κύρια επιτυχία τους ήρθε στα μέσα της δεκαετίας του 1980 με την επανεφεύρεση του αλγορίθμου backpropagation. Η μηχανική μάθηση, αναδιοργανωμένη και αναγνωρισμένη πλέον ως αυτόνομος επιστημονικός κλάδος, άρχισε να ανθεί τη δεκαετία του 1990.
Το πεδίο άλλαξε στόχο - από την επίτευξη τεχνητής νοημοσύνης, στράφηκε στην επίλυση πρακτικών και εφαρμόσιμων προβλημάτων.
Μετατόπισε την προσοχή του από τις συμβολικές προσεγγίσεις, που είχε κληρονομήσει από την τεχνητή νοημοσύνη, προς μεθόδους και μοντέλα που προέρχονταν από τη στατιστική, τη θολή λογική (fuzzy logic) και τη θεωρία πιθανοτήτων.

2.1.1.2 Εξόρυξη Δεδομένων (Data mining)

Η μηχανική μάθηση και η εξόρυξη δεδομένων χρησιμοποιούν συχνά τις ίδιες μεθόδους και παρουσιάζουν σημαντική επικάλυψη. Η μηχανική μάθηση εστιάζει στην πρόβλεψη, βασιζόμενη σε γνωστές ιδιότητες που έχουν μάθει τα μοντέλα από τα δεδομένα εκπαίδευσης (training data). Αντίθετα, η εξόρυξη δεδομένων εστιάζει στην ανακάλυψη άγνωστων ιδιοτήτων μέσα στα δεδομένα.

Η εξόρυξη δεδομένων χρησιμοποιεί πολλές μεθόδους μηχανικής μάθησης, αλλά με διαφορετικούς στόχους.
Από την άλλη πλευρά, η μηχανική μάθηση χρησιμοποιεί μεθόδους εξόρυξης δεδομένων στο πλαίσιο της μη επιβλεπόμενης μάθησης (unsupervised learning) ή ως προ-επεξεργασία για να βελτιώσει την ακρίβεια των μοντέλων εκμάθησης.

Η σύγχυση μεταξύ αυτών των δύο ερευνητικών κοινοτήτων προκύπτει, κυρίως, από τις διαφορετικές βασικές υποθέσεις τους:

- Στη μηχανική μάθηση, η απόδοση συχνά αξιολογείται με βάση την ικανότητα να αναπαράγει γνωστή γνώση.
- Στην εξόρυξη δεδομένων, ο κύριος στόχος είναι η ανακάλυψη νέας, άγνωστης γνώσης.

Αξιολογώντας με βάση γνωστή γνώση, τότε μια μη επιβλεπόμενη μέθοδος (unsupervised) θα υστερεί σε σύγκριση με άλλες επιβλεπόμενες μεθόδους. Αντίθετα, σε μια εργασία εξόρυξης δεδομένων, οι επιβλεπόμενες μέθοδοι δεν μπορούν να χρησιμοποιηθούν, αφού δεν υπάρχουν διαθέσιμα δεδομένα εκπαίδευσης (training data).

Η μηχανική μάθηση συνδέεται, επίσης, στενά με τη βελτιστοποίηση: Πολλά προβλήματα μάθησης διατυπώνονται ως ελαχιστοποίηση μιας συνάρτησης απώλειας (loss function) πάνω σε ένα σύνολο παραδειγμάτων εκπαίδευσης.
Οι συναρτήσεις απώλειας εκφράζουν τη διαφορά μεταξύ των προβλέψεων του μοντέλου, που έχει εκπαιδευτεί και των πραγματικών τιμών των δεδομένων.
Για παράδειγμα, στην ταξινόμηση, πρέπει να δοθεί η σωστή ετικέτα σε κάθε παράδειγμα έτσι, τα μοντέλα εκπαιδεύονται για να προβλέπουν με ακρίβεια τις προκαθορισμένες ετικέτες ενός συνόλου δεδομένων.

2.1.2 Υπέρ-προσαρμογή (Overfitting)

Συνήθως, τα μοντέλα μηχανικής μάθησης απαιτούν μεγάλο όγκο αξιόπιστων δεδομένων για να επιτυγχάνουν ακριβείς προβλέψεις. Κατά τη διαδικασία εκπαίδευσης ενός μοντέλου, οι μηχανικοί μηχανικής μάθησης (machine learning engineers) χρειάζεται να στοχεύσουν και να συλλέξουν μεγάλο και αντιπροσωπευτικό δείγμα δεδομένων.

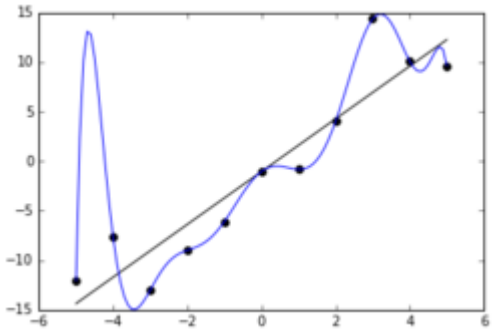
Τα δεδομένα του συνόλου εκπαίδευσης μπορεί να περιλαμβάνουν ποικίλους τύπους, όπως:

- κείμενα
- συλλογές εικόνων
- δεδομένα αισθητήρων
- δεδομένα που προέρχονται από μεμονωμένους χρήστες μιας υπηρεσίας.

Ένα φαινόμενο, που χρειάζεται ιδιαίτερη προσοχή κατά την εκπαίδευση, είναι η υπέρ-προσαρμογή (overfitting), δηλαδή όταν το μοντέλο «μαθαίνει υπερβολικά καλά» τα δεδομένα εκπαίδευσης και αποτυγχάνει να γενικεύσει σωστά σε νέα δεδομένα.

Η υπέρ-προσαρμογή είναι η υιοθέτηση μιας κακής, υπερβολικά πολύπλοκης θεωρίας, η οποία έχει τροποποιηθεί ώστε να ταιριάζει σε όλα τα προηγούμενα δεδομένα εκπαίδευσης.

Πολλά συστήματα προσπαθούν να μειώσουν την υπέρ-προσαρμογή δίνοντας ανταμοιβή σε μια θεωρία ανάλογα με το πόσο καλά ταιριάζει στα δεδομένα, αλλά ταυτόχρονα τιμωρώντας τη θεωρία ανάλογα με την πολυπλοκότητά της.



Εικόνα 1: Η μπλε γραμμή είναι ένα παράδειγμα υπέρ-προσαρμογής (overfitting) μιας γραμμικής συνάρτησης λόγω τυχαίου θορύβου στα δεδομένα.

Πηγή: https://medium.com/@musicaround/20-what-is-ml-advance-skill-normalization-d9eb8cef0e73

2.1.3 Μεροληψία (Bias)

Διάφορες προσεγγίσεις μηχανικής μάθησης μπορούν να επηρεαστούν από διαφορετικές μεροληψίες των δεδομένων. Ένα σύστημα μηχανικής μάθησης, που εκπαιδεύεται ειδικά με βάση τους τρέχοντες πελάτες μπορεί να μην είναι ικανό να προβλέψει τις ανάγκες νέων ομάδων πελατών, που δεν εκπροσωπούνται στα δεδομένα εκπαίδευσης. Όταν εκπαιδεύεται σε δεδομένα που δημιουργήθηκαν από ανθρώπους, η μηχανική μάθηση είναι πιθανό να πάρει τις συνειδητές και ασυνείδητες προκαταλήψεις, που ήδη υπάρχουν στην κοινωνία.

Τα εκπαιδευμένα μοντέλα, που βασίζονται σε μεροληπτικά ή μη αξιολογημένα δεδομένα, μπορεί να παράγουν ανεπιθύμητες προβλέψεις. Ένα μεροληπτικό μοντέλο (biased model) μπορεί να οδηγήσει σε αρνητικά αποτελέσματα, ενισχύοντας τις αρνητικές επιπτώσεις στους στόχους εφαρμογής.

Αυτού του είδους η αλγοριθμική μεροληψία, συχνά, προκύπτει από ανεπαρκή προετοιμασία ή καθαρισμό των δεδομένων πριν από την εκπαίδευση.

Τα γλωσσικά μοντέλα που εκπαιδεύονται με δεδομένα έχουν αποδειχθεί ότι περιέχουν ανθρώπινες προκαταλήψεις, καθώς οι ανθρώπινες γλώσσες περιέχουν προκαταλήψεις και οι μηχανές, που εκπαιδεύονται σε γλωσσικά σώματα δεδομένων θα τις μάθουν, επίσης. Το 2016, η Microsoft δοκίμασε το chatbot Tay στο Twitter, το οποίο σύντομα υιοθέτησε ρατσιστική και σεξιστική γλώσσα.

Το 2015, το Google Photos έβαλε ετικέτα (tag) σε ένα ζευγάρι μαύρων ατόμων ως γορίλες, προκαλώντας αντιδράσεις. Η ετικέτα αφαιρέθηκε και το 2023 το σύστημα ακόμη δεν αναγνωρίζει γορίλες σωστά. Παρόμοια ζητήματα αναγνώρισης μη λευκών ατόμων έχουν παρατηρηθεί και σε άλλα συστήματα.

2.2 Κατηγορίες Μηχανικής Μάθησης

Οι προσεγγίσεις της μηχανικής μάθησης παραδοσιακά χωρίζονται σε τρεις ευρείες κατηγορίες, που αντιστοιχούν σε παραδείγματα μάθησης, ανάλογα με την «ανατροφοδότησης» (feedback) που είναι διαθέσιμο στο σύστημα μάθησης. [1]

2.2.1 Επιβλεπόμενη Μάθηση

Οι αλγόριθμοι επιβλεπόμενης μάθησης κατασκευάζουν ένα μαθηματικό μοντέλο από ένα σύνολο δεδομένων που περιέχει τόσο τις εισόδους (inputs) όσο και τις επιθυμητές εξόδους (outputs).

Τα δεδομένα, γνωστά ως δεδομένα εκπαίδευσης (training data), αποτελούνται από ένα σύνολο παραδειγμάτων εκπαίδευσης (training examples).
Κάθε παράδειγμα εκπαίδευσης έχει μία ή περισσότερες εισόδους και την αντίστοιχη επιθυμητή έξοδο.

Στο μαθηματικό μοντέλο, κάθε παράδειγμα αναπαρίσταται από έναν πίνακα (array) ή διάνυσμα (vector), κάποιες φορές ονομάζεται διάνυσμα χαρακτηριστικών (feature vector), ενώ το σύνολο δεδομένων εκπαίδευσης αναπαρίσταται από έναν πίνακα (matrix).
Μέσω επαναληπτικής βελτιστοποίησης μιας αντικειμενικής συνάρτησης, οι αλγόριθμοι επιβλεπόμενης

μάθησης μαθαίνουν μια συνάρτηση, η οποία μπορεί να χρησιμοποιηθεί για να προβλέψει την έξοδο που αντιστοιχεί σε νέες εισόδους.

Μια βέλτιστη συνάρτηση επιτρέπει στον αλγόριθμο να καθορίζει σωστά την έξοδο για εισόδους που δεν υπήρχαν στα δεδομένα εκπαίδευσης.
Ένας αλγόριθμος, που βελτιώνει την ακρίβεια των προβλέψεών του με την πάροδο του χρόνου, θεωρείται ότι έχει μάθει να εκτελεί το συγκεκριμένο έργο.

Τα είδη των αλγορίθμων επιβλεπόμενης μάθησης περιλαμβάνουν:

- Ενεργή μάθηση (active learning)
- Ταξινόμηση (classification)
- Παλινδρόμηση (regression)

Οι αλγόριθμοι ταξινόμησης χρησιμοποιούνται όταν οι έξοδοι περιορίζονται σε ένα σύνολο τιμών, ενώ οι αλγόριθμοι παλινδρόμησης εφαρμόζονται όταν οι έξοδοι μπορούν να πάρουν οποιαδήποτε αριθμητική τιμή μέσα σε ένα εύρος.

Μάθηση ομοιότητας (Similarity learning):

Η μάθηση ομοιότητας είναι μια περιοχή της επιβλεπόμενης μηχανικής μάθησης, που συνδέεται στενά με την παλινδρόμηση και την ταξινόμηση.
Ο στόχος της είναι να μάθει από παραδείγματα, χρησιμοποιώντας μια συνάρτηση ομοιότητας, που μετρά πόσο όμοια ή σχετιζόμενα είναι δύο αντικείμενα.

Έχει εφαρμογές σε:

- κατάταξη (ranking)
- συστήματα προτάσεων (recommendation systems)
- παρακολούθηση οπτικής ταυτότητας (visual identity tracking)
- επαλήθευση προσώπου (face verification)
- επαλήθευση ομιλητή (speaker verification)

2.2.2 Μη Επιβλεπόμενη Μηχανική Μάθηση

Οι αλγόριθμοι μη επιβλεπόμενης μάθησης ανακαλύπτουν δομές μέσα σε δεδομένα που δεν έχουν επισημανθεί, ταξινομηθεί ή κατηγοριοποιηθεί.
Αντί να ανταποκρίνονται σε ανατροφοδότηση, οι αλγόριθμοι αυτοί εντοπίζουν κοινά χαρακτηριστικά στα δεδομένα και αντιδρούν ανάλογα με την παρουσία ή απουσία αυτών των κοινών μοτίβων σε κάθε νέο στοιχείο δεδομένων.

Οι βασικές εφαρμογές της μη επιβλεπόμενης μηχανικής μάθησης περιλαμβάνουν:

- Ομαδοποίηση (clustering)
- Μείωση διαστάσεων (dimensionality reduction)
- Εκτίμηση πυκνότητας (density estimation)

Ανάλυση συστάδων (Cluster analysis)

Η ανάλυση συστάδων είναι η διαδικασία κατά την οποία ένα σύνολο παρατηρήσεων χωρίζεται σε υποσύνολα (clusters), έτσι ώστε οι παρατηρήσεις μέσα στην ίδια συστάδα να είναι παρόμοιες σύμφωνα με ένα ή περισσότερα προκαθορισμένα κριτήρια, ενώ οι παρατηρήσεις από διαφορετικές συστάδες να είναι ανόμοιες.

Διαφορετικές τεχνικές ομαδοποίησης βασίζονται σε διαφορετικές υποθέσεις για τη δομή των δεδομένων, οι οποίες συχνά καθορίζονται μέσω κάποιου μέτρου ομοιότητας (similarity metric).
Η ποιότητα μιας ομαδοποίησης συχνά αξιολογείται με δείκτες όπως:

- Εσωτερική συνοχή (internal compactness): δηλαδή η ομοιότητα μεταξύ των μελών μιας ίδιας συστάδας.
- Διαχωρισμός (separation): δηλαδή η διαφορά μεταξύ διαφορετικών συστάδων.

Άλλες μέθοδοι βασίζονται στην εκτίμηση της πυκνότητας των δεδομένων ή στη συνδεσιμότητα γραφημάτων (graph connectivity).

Μια ειδική μορφή μη επιβλεπόμενης μάθησης είναι η αυτό-επιβλεπόμενη μάθηση, στην οποία το μοντέλο εκπαιδεύεται χρησιμοποιώντας σήματα επίβλεψης που δημιουργούνται από τα ίδια τα δεδομένα.

Δηλαδή, το σύστημα δημιουργεί μόνο του τις «ετικέτες» εκπαίδευσης, επιτρέποντας του να μάθει αναπαραστάσεις χωρίς την ανάγκη ανθρώπινης επισήμανσης.

Μείωση διαστάσεων (Dimensionality reduction):

Η μείωση διαστάσεων είναι μια διαδικασία κατά την οποία μειώνεται ο αριθμός των τυχαίων μεταβλητών που λαμβάνονται υπόψη, μέσω της εύρεσης ενός συνόλου βασικών μεταβλητών (principal variables).

Με άλλα λόγια, είναι μια διαδικασία, όπου μειώνεται η διάσταση του συνόλου χαρακτηριστικών δηλαδή, ο αριθμός των χαρακτηριστικών.

Οι περισσότερες τεχνικές μείωσης διαστάσεων μπορούν να θεωρηθούν είτε ως απομάκρυνση χαρακτηριστικών (feature elimination) είτε ως εξαγωγή χαρακτηριστικών (feature extraction).

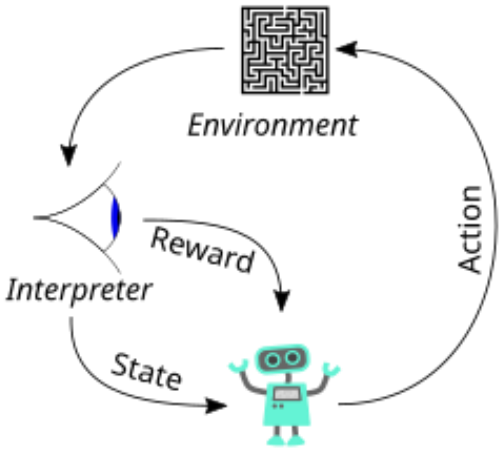
Μία από τις πιο δημοφιλείς μεθόδους μείωσης διαστάσεων είναι η Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA).
Η PCA περιλαμβάνει τη μετατροπή δεδομένων μεγάλης διάστασης (π.χ. τρισδιάστατων - 3D) σε δεδομένα μικρότερης διάστασης (π.χ. δισδιάστατα - 2D), διατηρώντας όσο το δυνατόν περισσότερη πληροφορία από την αρχική δομή των δεδομένων.

2.2.3 Ενισχυτική μάθηση (Reinforcement learning):

Ένα πρόγραμμα υπολογιστή αλληλεπιδρά με ένα δυναμικό περιβάλλον, μέσα στο οποίο πρέπει να εκτελέσει έναν συγκεκριμένο στόχο (όπως η οδήγηση ενός οχήματος ή το παίξιμο ενός παιχνιδιού ενάντια σε έναν αντίπαλο).

Καθώς το πρόγραμμα εξερευνά τον χώρο προβλημάτων, παρέχει ανατροφοδότηση παρόμοια με ανταμοιβές (rewards), τις οποίες προσπαθεί να μεγιστοποιήσει.

Αν και κάθε αλγόριθμος έχει πλεονεκτήματα και περιορισμούς, κανένας μεμονωμένος αλγόριθμος δεν είναι κατάλληλος για όλα τα προβλήματα.



Εικόνα 2: Στην ενισχυτική μάθηση**,** ένας πράκτορας (agent) εκτελεί ενέργειες (actions) μέσα σε ένα περιβάλλον (environment)**.** Αυτές οι ενέργειες παράγουν μια ανταμοιβή (reward) και/ή μια αναπαράσταση της κατάστασης (state representation**)**, η οποία επιστρέφεται στον πράκτορα ως ανατροφοδότηση (feedback).

Πηγή: https://en.wikipedia.org/wiki/Machine_learning

2.3 Αλγόριθμοι Μηχανικής Μάθησης

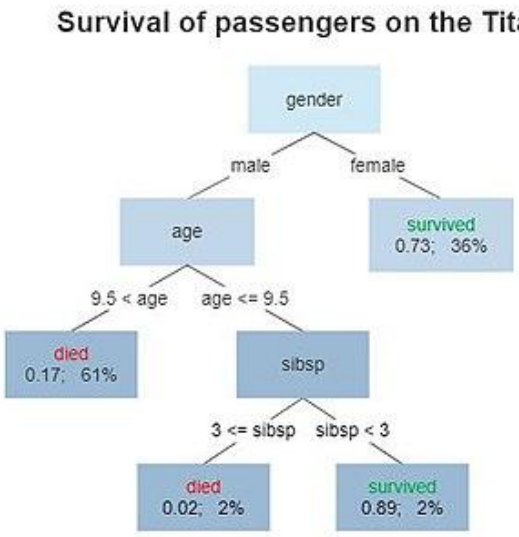
2.3.1 Δέντρα Αποφάσεων

Η εκμάθηση δέντρων αποφάσεων χρησιμοποιεί ένα δέντρο απόφασης ως μοντέλο πρόβλεψης, το οποίο μετατρέπει τις παρατηρήσεις για ένα αντικείμενο (που αναπαρίσταται στα κλαδιά) σε συμπεράσματα σχετικά με την τιμή-στόχο του αντικειμένου (που αναπαρίσταται στα φύλλα). Αποτελεί μία από τις βασικές προσεγγίσεις προγνωστικής μοντελοποίησης, που χρησιμοποιούνται στη στατιστική, την εξόρυξη δεδομένων και τη μηχανική μάθηση. [2]

Υπάρχουν δύο βασικοί τύποι δέντρων αποφάσεων:

- Δέντρα ταξινόμησης (classification trees)**:** Ο μεταβλητός στόχος παίρνει διακριτές τιμές (π.χ. "ναι" ή "όχι"). Στα δέντρα αυτά, τα φύλλα αναπαριστούν ετικέτες κατηγοριών, ενώ τα κλαδιά αντιστοιχούν σε συνδυασμούς χαρακτηριστικών (features) που οδηγούν σε κάθε κατηγορία.
- Δέντρα παλινδρόμησης (regression trees): Ο μεταβλητός στόχος παίρνει συνεχείς τιμές (π.χ. πραγματικούς αριθμούς). Χρησιμοποιούνται για προβλέψεις αριθμητικών τιμών, όπως π.χ. η πρόβλεψη τιμής κατοικίας ή θερμοκρασίας.

Στην ανάλυση αποφάσεων, ένα δέντρο απόφασης μπορεί να χρησιμοποιηθεί για να απεικονίσει τη διαδικασία λήψης αποφάσεων, παρουσιάζοντας όλες τις πιθανές επιλογές και αποτελέσματα.



Εικόνα 3: Δέντρο αποφάσεων που δείχνει την πιθανότητα επιβίωσης των επιβατών του Τιτανικού.

Πηγή: https://en.wikipedia.org/wiki/Machine_learning

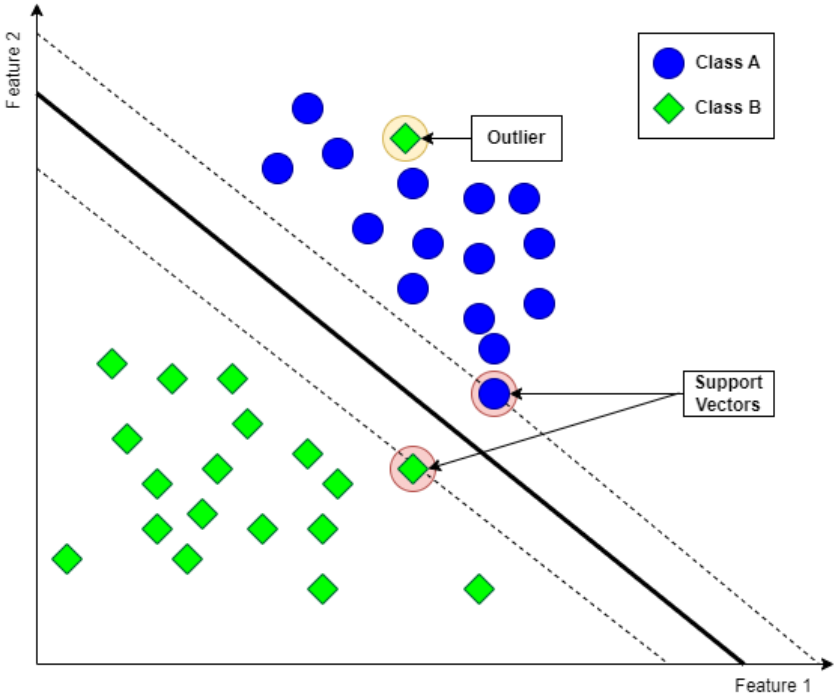
2.3.2 Μηχανές Διανυσμάτων Υποστήριξης (Support-Vector Machines)

Οι μηχανές διανυσμάτων υποστήριξης (SVM), αποτελούν ένα σύνολο συγγενών μεθόδων επιβλεπόμενης μάθησης που χρησιμοποιούνται για ταξινόμηση και παλινδρόμηση.

Δεδομένου ενός συνόλου παραδειγμάτων εκπαίδευσης, καθένα από τα οποία ανήκει σε μία από δύο κατηγορίες, ο αλγόριθμος εκπαίδευσης ενός SVM φτιάχνει ένα μοντέλο, το οποίο προβλέπει σε ποια κατηγορία ανήκει ένα νέο παράδειγμα. [3]

Ένας αλγόριθμος εκπαίδευσης SVM είναι μη πιθανοτικός, δυαδικός, γραμμικός ταξινομητής, αν και υπάρχουν μέθοδοι (όπως η Platt scaling), που επιτρέπουν τη χρήση των SVM σε πιθανοτικό πλαίσιο ταξινόμησης.

Εκτός από τη γραμμική ταξινόμηση, τα SVM μπορούν να εκτελέσουν αποτελεσματικά μια μη γραμμική ταξινόμηση, χρησιμοποιώντας την τεχνική, που ονομάζεται kernel trick (κόλπο του πυρήνα), η οποία χαρτογραφεί έμμεσα τις εισόδους σε χώρους χαρακτηριστικών υψηλότερης διάστασης.



Εικόνα 4: Support Vector Machine (SVM)

Πηγή: https://towardsdatascience.com/support-vector-machine-with-scikit-learn-a-friendly-introduction-a2969f2ff00d/

2.3.3 Ανάλυση Παλινδρόμησης (Regression Analysis)

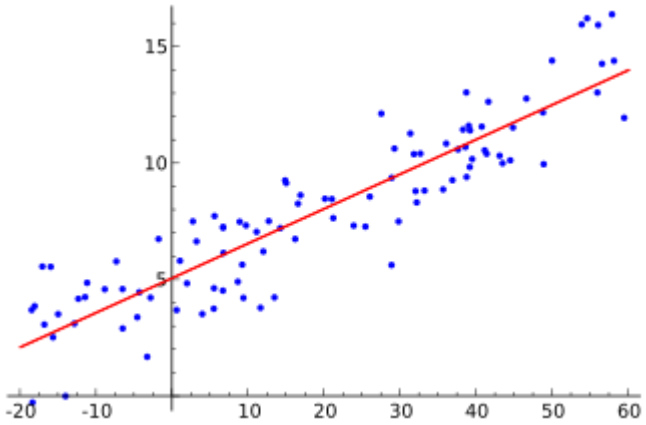
Η ανάλυση παλινδρόμησης περιλαμβάνει ένα ευρύ φάσμα από στατιστικές μεθόδους, που στοχεύουν στην εκτίμηση της σχέσης μεταξύ μεταβλητών εισόδου και των αντίστοιχων χαρακτηριστικών τους. [4]

Η πιο κοινή μορφή είναι η γραμμική παλινδρόμηση, όπου σχεδιάζεται μια ευθεία γραμμή, που προσαρμόζεται βέλτιστα στα δεδομένα, σύμφωνα με κάποιο μαθηματικό κριτήριο. Η μέθοδος αυτή, συχνά, επεκτείνεται με τεχνικές κανονικοποίησης (regularisation) για να περιοριστεί το overfitting και η μεροληψία, όπως στη ridge regression.

Για μη γραμμικά προβλήματα, χρησιμοποιούνται μοντέλα όπως:

- η πολυωνυμική παλινδρόμηση (π.χ. για προσαρμογή καμπυλών τάσης στο Microsoft Excel)
- η λογιστική παλινδρόμηση (συχνά στη στατιστική ταξινόμηση)
- η παλινδρόμηση με πυρήνα (kernel regression), η οποία εισάγει μη γραμμικότητα αξιοποιώντας το kernel trick, χαρτογραφώντας έμμεσα τις μεταβλητές εισόδου σε χώρους υψηλότερης διάστασης.

Η πολυμεταβλητή γραμμική παλινδρόμηση (multivariate linear regression) επεκτείνει την έννοια της γραμμικής παλινδρόμησης για να διαχειρίζεται πολλαπλές εξαρτημένες μεταβλητές ταυτόχρονα. Αυτή η προσέγγιση εκτιμά τις σχέσεις μεταξύ ενός συνόλου μεταβλητών εισόδου και πολλών μεταβλητών εξόδου, προσαρμόζοντας ένα πολυδιάστατο γραμμικό μοντέλο. Είναι ιδιαίτερα χρήσιμη σε σενάρια όπου οι έξοδοι είναι αλληλοεξαρτώμενες ή μοιράζονται κοινά υποκείμενα πρότυπα, όπως στην πρόβλεψη πολλών οικονομικών δεικτών ή στην ανακατασκευή εικόνων, που είναι εκ φύσεως πολυδιάστατες.



Εικόνα 4: Απεικόνιση μιας γραμμικής παλινδρόμησης σε ένα σύνολο δεδομένων.

Πηγή: https://el.wikipedia.org/wiki/Γραμμική_παλινδρόμηση

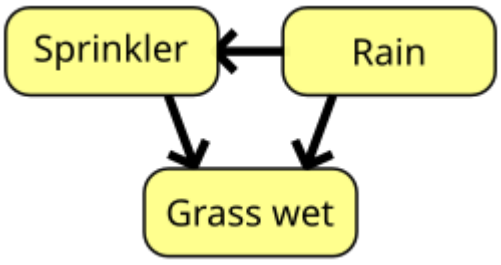
2.3.4 Δίκτυα Bayes (Bayesian Networks)

Ένα Δίκτυο Bayes είναι ένα πιθανοτικό γραφικό μοντέλο, που αναπαριστά ένα σύνολο τυχαίων μεταβλητών και τις συνθήκες ανεξαρτησίας τους μέσω ενός κατευθυνόμενου άκυκλου γράφου. [5]

Για παράδειγμα, ένα δίκτυο Bayes μπορεί να αναπαριστά τις πιθανοτικές σχέσεις μεταξύ ασθενειών και συμπτωμάτων. Δεδομένων κάποιων συμπτωμάτων, το δίκτυο μπορεί να χρησιμοποιηθεί για να υπολογίσει τις πιθανότητες ύπαρξης διαφόρων ασθενειών.

Υπάρχουν αποτελεσματικοί αλγόριθμοι, οι οποίοι εκτελούν συμπέρασμα και εκμάθηση σε τέτοια δίκτυα. Τα δυναμικά δίκτυα Bayes (dynamic Bayesian networks) επεκτείνουν την έννοια αυτή ώστε να μοντελοποιούν ακολουθίες μεταβλητών, όπως σήματα ομιλίας ή ακολουθίες πρωτεϊνών.

Οι γενικεύσεις των δικτύων Bayes που μπορούν να αναπαραστήσουν και να επιλύσουν προβλήματα λήψης αποφάσεων υπό αβεβαιότητα ονομάζονται διαγράμματα επιρροής (influence diagrams).



Εικόνα 5: Ένα απλό δίκτυο Bayes. Η βροχή επηρεάζει το αν θα ενεργοποιηθεί ο εκτοξευτήρας νερού (sprinkler), και τόσο η βροχή όσο και ο εκτοξευτήρας επηρεάζουν το αν το γρασίδι είναι βρεγμένο.

Πηγή: https://en.wikipedia.org/wiki/Machine_learning

2.4 Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP)

Η Επεξεργασία Φυσικής Γλώσσας (NLP) είναι η επεξεργασία πληροφοριών φυσικής γλώσσας από έναν υπολογιστή. Η μελέτη του NLP, που αποτελεί υπό-πεδίο της επιστήμης των υπολογιστών, συνδέεται στενά με την τεχνητή νοημοσύνη (AI). [6]

Η επεξεργασία φυσικής γλώσσας σχετίζεται, επίσης, με την ανάκτηση πληροφοριών, την αναπαράσταση γνώσης, τη υπολογιστική γλωσσολογία, και γενικότερα με τη γλωσσολογία.

Κύριες εργασίες επεξεργασίας σε ένα σύστημα NLP περιλαμβάνουν:

- Αναγνώριση ομιλίας (speech recognition)
- Ταξινόμηση κειμένου (text classification)
- Κατανόηση φυσικής γλώσσας (natural language understanding)
- Παραγωγή φυσικής γλώσσας (natural language generation).

Παραδείγματα χρήσης NLP:

Αναγνώριση ομιλίας (Speech recognition):

Δίνοντας ένα ηχητικό απόσπασμα με ομιλία ενός ή περισσότερων ατόμων, το σύστημα μετατρέπει την ομιλία σε κείμενο. Αυτή η εργασία είναι το αντίθετο της μετατροπής κειμένου σε ομιλία (Text-to-Speech) και θεωρείται πολύ δύσκολο πρόβλημα.
Κάποιες προκλήσεις είναι η έλλειψη παύσεων μεταξύ των λέξεων στην φυσική ομιλία (όπου χρειάζεται η τμηματοποίηση της ομιλίας (speech segmentation)), το φαινόμενο της συνάρθρωσης (coarticulation), όπου οι ήχοι των γραμμάτων συγχωνεύονται και τις διαφορετικές προφορές και τρόπους εκφοράς των λέξεων.

Τμηματοποίηση ομιλίας (Speech segmentation):

Δεδομένου ενός ηχητικού αποσπάσματος, το σύστημα χωρίζει την ομιλία σε μεμονωμένες λέξεις ή φράσεις.
Αποτελεί υπό-εργασία της αναγνώρισης ομιλίας.

Κείμενο σε ομιλία (Text-to-speech):

Δεδομένου ενός κειμένου, το σύστημα παράγει την αντίστοιχη προφορική εκδοχή. Χρησιμοποιείται, για παράδειγμα, για την υποβοήθηση ατόμων με προβλήματα όρασης.

2.4.1 Τμηματοποίηση λέξεων (Word Segmentation - Tokenization)

Τμηματοποίηση λέξεων (Tokenization) είναι η διαδικασία που χωρίζει ένα κείμενο σε μεμονωμένες λέξεις ή τμήματα λέξεων (tokens).
Το αποτέλεσμα είναι:

Ένας πίνακας λέξεων (word index) που αντιστοιχίζει κάθε μοναδική λέξη σε έναν αριθμητικό δείκτη, και

Ένα tokenized text, δηλαδή το κείμενο όπου κάθε λέξη έχει αντικατασταθεί από τον αριθμητικό της δείκτη.

Αυτή η αναπαράσταση χρησιμοποιείται σε πολλές μεθόδους μηχανικής και βαθιάς μάθησης (bag of words - BOW).

2.4.2 Λημματοποίηση (Lemmatization)

Η εργασία αφαίρεσης μόνο των κλιτικών καταλήξεων, ώστε να επιστραφεί η βασική λεξικογραφική μορφή (λήμμα) μιας λέξης. Σε αντίθεση με το stemming, η lemmatization χρησιμοποιεί λεξικό για να εντοπίσει τη σωστή μορφή της λέξης.

Παράδειγμα: το “running” ή το “ran” θα γίνει – “run”.

Το “better” θα γίνει “good”

Η λημματοποίηση, επομένως, προσφέρει ορθότερη γλωσσική κανονικοποίηση, ιδανική για εφαρμογές όπου χρειάζεται ακριβής σημασιολογική επεξεργασία.

2.4.3 Αποκοπή καταλήξεων (Stemming)

Είναι η διαδικασία μείωσης των παραγώγων λέξεων στη ρίζα τους. Για παράδειγμα, οι λέξεις: “closed”, “closing”, “closer”, θα γίνει “close”.

Η αποκοπή καταλήξεων βασίζεται σε κανόνες (όχι λεξικό), γι’ αυτό και μπορεί μερικές φορές να δώσει μη φυσιολογικές ρίζες (π.χ. studies θα γίνει studi).
Είναι απλούστερη και ταχύτερη, από τη λημματοποίηση, αλλά λιγότερο ακριβής.

2.4.4 Μορφολογική ανάλυση (Morphological segmentation)

Είναι η διαδικασία διάσπασης των λέξεων στα συστατικά τους μορφήματα και αναγνώρισης του είδους κάθε μορφήματος (ρίζα, πρόθημα, επίθημα κ.λπ.).

Παράδειγμα: “unhappiness” : { “un” (πρόθεμα), “happy” (ρίζα), “ness” (επίθημα) }

Η δυσκολία του έργου εξαρτάται από τη μορφολογική πολυπλοκότητα της γλώσσας:

- Στα Αγγλικά, η μορφολογία είναι σχετικά απλή και μπορείς να μοντελοποιήσεις όλες τις πιθανές μορφές μιας λέξης (π.χ. “open, opens, opened, opening”, είναι ίδια λέξη).
- Στα Τουρκικά ή Ινδικά, κάθε λήμμα μπορεί να έχει χιλιάδες πιθανούς τύπους, καθιστώντας την εργασία πολύ πιο δύσκολη.

2.4.5 Επισήμανση μερών του λόγου (Part-of-speech tagging – POS)

Δεδομένης μιας πρότασης, το σύστημα καθορίζει το μέρος του λόγου (ρήμα, ουσιαστικό, επίθετο, επίρρημα κ.λπ.) κάθε λέξης.
Πολλές λέξεις μπορούν να έχουν πολλαπλές κατηγορίες ανάλογα με το συμφραζόμενο:

Παράδειγμα:

- "book": ουσιαστικό ("the book on the table") ή ρήμα ("to book a flight")
- "set": ουσιαστικό, ρήμα ή επίθετο
- "out": μπορεί να είναι επίρρημα, πρόθεση, επίθετο κ.ά.

Η επισήμανση μερών του λόγου είναι θεμελιώδης για πολλές εφαρμογές της επεξεργασίας φυσικής γλώσσας (όπως συντακτική ανάλυση, εξαγωγή γνώσης, και μηχανική μετάφραση).

2.4.6 Αναγνώριση Οντοτήτων με όνομα (Name Entity Recognition – NER)

Δεδομένου ενός κειμένου, προσδιορίζονται οι οντότητες (όπως πρόσωπα, τοποθεσίες, οργανισμοί) και το είδος τους.

Παράδειγμα:

"Barack Obama was born in Hawaii."
Barack Obama: *Person* Hawaii: *Location*

Η χρήση κεφαλαίων γραμμάτων βοηθά σε γλώσσες όπως τα Αγγλικά, αλλά δεν είναι αξιόπιστη σε πολλές περιπτώσεις:

- Η αρχή κάθε πρότασης έχει κεφαλαίο γράμμα,
- Οι οντότητες μπορεί να αποτελούνται από πολλές λέξεις,
- Πολλές γλώσσες (π.χ. Κινέζικα, Αραβικά) δεν έχουν κεφαλαία γράμματα,
- Στα Γερμανικά όλα τα ουσιαστικά είναι κεφαλαία, ανεξαρτήτως αν είναι κύρια ονόματα.

2.4.7 Ανάλυση συναισθήματος (Sentiment analysis):

Είναι η υπολογιστική αναγνώριση και ταξινόμηση της συναισθηματικής πρόθεσης που εκφράζεται μέσα σε ένα κείμενο.

Η ανάλυση καθορίζει αν το συναίσθημα είναι θετικό, αρνητικό ή ουδέτερο.
Χρησιμοποιεί χαρακτηριστικά όπως:

- n-grams,
- TF–IDF,
- βαθιά νευρωνικά δίκτυα (deep learning) που αναγνωρίζουν μακροπρόθεσμες και βραχυπρόθεσμες σχέσεις στο κείμενο (π.χ. LSTM, Transformers).

Παράδειγμα εφαρμογής είναι η ανάλυση σχολίων πελατών σε e-shop (θετικά/αρνητικά reviews), ή σχολιασμός ταινιών (imdb reviews).

2.4.8 Word Embedding (Ενσωμάτωμα λέξης)

Στην επεξεργασία φυσικής γλώσσας, ένα ενσωμάτωμα λέξης (word embedding) είναι μια αναπαράσταση μιας λέξης. Το embedding χρησιμοποιείται στην ανάλυση κειμένου. Συνήθως, η αναπαράσταση είναι ένα διάνυσμα πραγματικών αριθμών που κωδικοποιεί τη σημασία της λέξης με τέτοιο τρόπο ώστε λέξεις που βρίσκονται πιο κοντά στον διανυσματικό χώρο να αναμένεται να έχουν παρόμοια σημασία. Τα word embeddings μπορούν να προκύψουν μέσω τεχνικών γλωσσικής μοντελοποίησης και εκμάθησης χαρακτηριστικών, όπου λέξεις ή φράσεις από το λεξιλόγιο αντιστοιχίζονται σε διανύσματα πραγματικών αριθμών.

Οι μέθοδοι για τη δημιουργία αυτής της αντιστοίχισης περιλαμβάνουν νευρωνικά δίκτυα, μείωση διαστάσεων σε πίνακες συν-εμφάνισης λέξεων, πιθανοτικά μοντέλα, μεθόδους βασισμένες σε ερμηνεύσιμες βάσεις γνώσης, και ρητές αναπαραστάσεις με βάση το συμφραζόμενο στο οποίο εμφανίζονται οι λέξεις.

Τα embeddings λέξεων και φράσεων, όταν χρησιμοποιούνται ως βασική αναπαράσταση εισόδου, έχει αποδειχθεί ότι βελτιώνουν την απόδοση σε εργασίες NLP όπως η συντακτική ανάλυση (syntactic parsing) και η ανάλυση συναισθήματος (sentiment analysis).

Το 2013, ομάδα της Google με επικεφαλής τον Tomas Mikolov δημιούργησε το word2vec [7], ένα εργαλείο που εκπαίδευε διανυσματικά μοντέλα πολύ πιο γρήγορα από προηγούμενες μεθόδους. Το word2vec συνέβαλε καθοριστικά στη διάδοση των embeddings και στη μετάβαση από εξειδικευμένη έρευνα σε ευρύτερη πρακτική εφαρμογή.

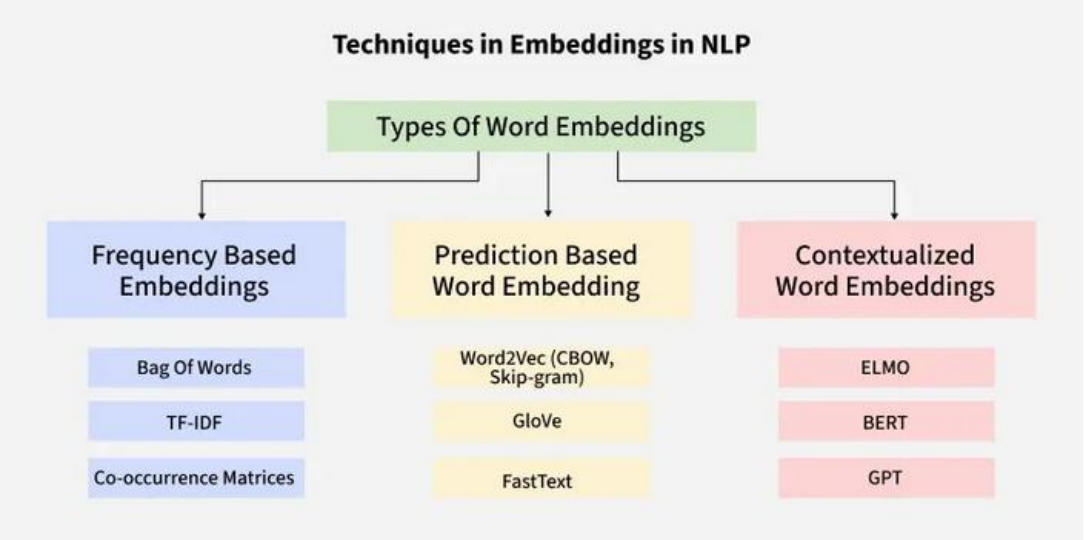
Ιστορικά, ένας από τους βασικούς περιορισμούς των στατικών word embeddings είναι ότι λέξεις με πολλαπλές σημασίες αναπαρίστανται με ένα μόνο διάνυσμα. Δηλαδή, η πολυσημία και η ομωνυμία δεν αντιμετωπίζονται σωστά. Για παράδειγμα, στη φράση «The club I tried yesterday was great!», δεν είναι σαφές αν το "club" αναφέρεται σε σάντουιτς, λέσχη, μπαστούνι γκολφ ή κάτι άλλο. Η ανάγκη για πολλαπλές αναπαραστάσεις ανά λέξη οδήγησε στην ανάπτυξη multi-sense embeddings.

Η χρήση multi-sense embeddings βελτιώνει την απόδοση σε εργασίες NLP όπως POS tagging, αναγνώριση οντοτήτων και ανάλυση συναισθήματος.

Από τα τέλη της δεκαετίας του 2010, εμφανίστηκαν δυναμικά (contextual) embeddings όπως τα ELMo [10] και BERT [11]. Σε αντίθεση με τα στατικά, αυτά δίνουν διαφορετικό embedding σε κάθε εμφάνιση μιας λέξης, ανάλογα με τα συμφραζόμενα, αποτυπώνοντας καλύτερα την πολυσημία.

Τα word embeddings μπορεί να ενσωματώνουν προκαταλήψεις από τα δεδομένα εκπαίδευσης. Έρευνες έχουν δείξει ότι ακόμη και embeddings από ειδησεογραφικά κείμενα περιέχουν έμφυλες και φυλετικές προκαταλήψεις.

Μελέτες δείχνουν ότι αυτά τα biases μπορούν όχι μόνο να διατηρηθούν αλλά και να ενισχυθούν, αν δεν υπάρχει προσεκτικός έλεγχος.



Εικόνα 6: Τύποι word embeddings

Πηγή: https://medium.com/@222kashafnaveed/phase2-part-2-prediction-based-word-embeddings-eb5b0e10cfee

Γιατί χρειαζόμαστε Word Embeddings;

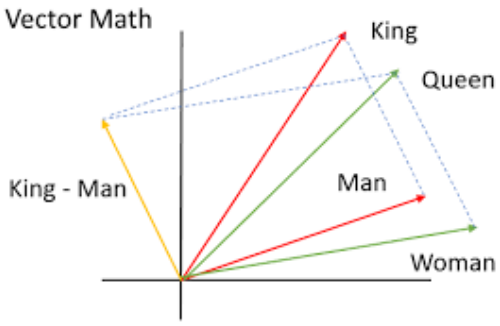
- Για μείωση της διαστασιμότητας (dimensionality).
- Για να χρησιμοποιούμε μια λέξη ώστε να προβλέπουμε τις λέξεις γύρω της.
- Βοηθούν στη βελτίωση της ερμηνευσιμότητας του μοντέλου μέσω αριθμητικής αναπαράστασης.
- Μπορούν να αποτυπώσουν τη σημασιολογία και την ομοιότητα μεταξύ λέξεων.

Τα embeddings που βασίζονται στη συχνότητα (frequency-based embeddings) αναφέρονται σε αναπαραστάσεις λέξεων που προκύπτουν από τη συχνότητα εμφάνισής τους σε ένα corpus. Αυτές οι αναπαραστάσεις βασίζονται στην ιδέα ότι η σημασία ή η σπουδαιότητα μιας λέξης μπορεί να εξαχθεί από το πόσο συχνά εμφανίζεται στο κείμενο.

Παρόλο που τα embeddings που βασίζονται στη συχνότητα είναι απλά και εύκολα στην κατανόηση, υστερούν σε βάθος σημασιολογικής πληροφορίας και κατανόησης συμφραζομένων σε σύγκριση με πιο προχωρημένα embeddings που βασίζονται σε προβλέψεις (prediction-based embeddings).

Τα embeddings που βασίζονται σε προβλέψεις (prediction-based embeddings) είναι αναπαραστάσεις λέξεων που προκύπτουν από μοντέλα τα οποία εκπαιδεύονται να προβλέπουν πτυχές του συμφραζομένου μιας λέξης ή τις γειτονικές λέξεις. Σε αντίθεση με τα frequency-based embeddings που επικεντρώνονται στη συχνότητα εμφάνισης, τα prediction-based embeddings συλλαμβάνουν σημασιολογικές σχέσεις και πληροφορία συμφραζομένων, προσφέροντας πιο πλούσιες αναπαραστάσεις νοήματος.

Τα prediction-based embeddings μπορούν να διακρίνουν καλύτερα συνώνυμα και να διαχειρίζονται πιο αποτελεσματικά την πολυσημία (λέξεις με πολλαπλές σημασίες). Οι ιδιότητες του διανυσματικού χώρου τους επιτρέπουν εργασίες όπως η μέτρηση ομοιότητας μεταξύ λέξεων και η επίλυση αναλογιών (πχ, king – man + woman = queen, ή Athens – Greece + France = Paris, ή walking – walk + swim = swimming).



Εικόνα 7: Παράδειγμα word2vec.

Πηγή: https://medium.com/analytics-vidhya/word-embeddings-in-nlp-word2vec-glove-fasttext-24d4d4286a73

Επιπλέον, μπορούν να γενικεύουν καλά σε άγνωστες λέξεις ή συμφραζόμενα, γεγονός που τα καθιστά ανθεκτικά σε περιπτώσεις λέξεων εκτός λεξιλογίου (out-of-vocabulary).

Οι μέθοδοι που βασίζονται σε προβλέψεις, ιδιαίτερα τεχνικές όπως Word2Vec, GloVe [8] και fastText [9], έχουν κυριαρχήσει στον χώρο των word embeddings λόγω της ικανότητάς τους να αποτυπώνουν πλούσια σημασιολογική πληροφορία και να αποδίδουν καλά σε διάφορες εργασίες NLP.

Contextual embeddings

Τα contextual embeddings υπάρχουν χάρη στους μηχανισμούς attention και τα transformer models [12], που λαμβάνουν υπόψη:

- το συμφραζόμενο
- τις αμφίδρομες σχέσεις μεταξύ λέξεων

οδηγώντας σε πιο προηγμένες αναπαραστάσεις γλώσσας.

Οι μηχανισμοί attention εισήχθησαν για να βελτιώσουν την ικανότητα των νευρωνικών δικτύων να εστιάζουν σε συγκεκριμένα μέρη της εισόδου κατά την πρόβλεψη. Αντί να αντιμετωπίζουν όλα τα στοιχεία της εισόδου το ίδιο, επιτρέπουν στο μοντέλο να δίνει έμφαση στα πιο σημαντικά μέρη.

Τα transformers αποτελούν τη βάση πολλών σύγχρονων μοντέλων NLP, όπως τα BERT, GPT και T5 (Text-to-Text Transfer Transformer). Αποδίδουν εξαιρετικά σε εργασίες όπως:

- γλωσσική μοντελοποίηση
- μηχανική μετάφραση

- παραγωγή κειμένου
- απάντηση ερωτήσεων

Οι transformers χρησιμοποιούν μηχανισμό self-attention για να συλλάβουν σχέσεις μεταξύ λέξεων σε μια ακολουθία. Αυτό επιτρέπει σε κάθε λέξη να «εστιάζει» σε όλες τις υπόλοιπες, καταγράφοντας μακρινές εξαρτήσεις (long-range dependencies).

3

Βαθιά Νευρωνικά Δίκτυα

3.1 Έννοια

Στη μηχανική μάθηση (Machine Learning), η βαθιά μάθηση (Deep Learning – DL) επικεντρώνεται στη χρήση πολυεπίπεδων νευρωνικών δικτύων (neural networks) για την εκτέλεση εργασιών όπως η ταξινόμηση (classification), η παλινδρόμηση (regression) και η μάθηση αναπαραστάσεων (representation learning). Το πεδίο αντλεί έμπνευση από τη βιολογική νευροεπιστήμη (biological neuroscience) και βασίζεται στη διάταξη τεχνητών νευρώνων (artificial neurons) σε διαδοχικά επίπεδα (layers), καθώς και στην «εκπαίδευσή» τους ώστε να επεξεργάζονται δεδομένα. [13]

Ο όρος «βαθιά» (deep) αναφέρεται στη χρήση πολλαπλών επιπέδων στο δίκτυο, τα οποία μπορεί να κυμαίνονται από τρία έως αρκετές εκατοντάδες ή ακόμη και χιλιάδες επίπεδα. Οι μέθοδοι που χρησιμοποιούνται μπορεί να βασίζονται σε επιβλεπόμενη μάθηση (supervised learning), ημι-επιβλεπόμενη μάθηση (semi-supervised learning) ή μη επιβλεπόμενη μάθηση (unsupervised learning).

Ορισμένες συνηθισμένες αρχιτεκτονικές δικτύων βαθιάς μάθησης περιλαμβάνουν τα πλήρως συνδεδεμένα δίκτυα (fully connected networks), τα βαθιά δίκτυα πεποίθησης (deep belief networks), τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks – RNNs), τα αναδρομικά τυχαία νευρωνικά δίκτυα (Random Neural Networks), τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks – CNNs), τους Transformers και τα νευρωνικά πεδία ακτινοβολίας (Neural Radiance Fields – NeRFs).

Οι αρχιτεκτονικές αυτές έχουν εφαρμοστεί σε τομείς όπως η υπολογιστική όραση (computer vision), η αναγνώριση ομιλίας (speech recognition), η επεξεργασία φυσικής γλώσσας (Natural Language Processing – NLP), η μηχανική μετάφραση (machine translation), κ.α. Σε αυτούς τους τομείς έχουν επιτύχει

αποτελέσματα συγκρίσιμα με εκείνα ανθρώπων-ειδικών και, σε ορισμένες περιπτώσεις, έχουν ακόμη και ξεπεράσει την απόδοση των ανθρώπων.

Το Βαθύ Νευρωνικό Δίκτυο (DNN) δημιουργεί ένα σύνολο εικονικών νευρώνων και αντιστοιχίζει τυχαίες αριθμητικές τιμές, γνωστές ως «βάρη» (weights), στις μεταξύ τους συνδέσεις. Τα βάρη και οι είσοδοι πολλαπλασιάζονται και παράγουν μια έξοδο. Εάν το δίκτυο δεν αναγνωρίσει με ακρίβεια ένα συγκεκριμένο πρότυπο (pattern), ένας αλγόριθμος προσαρμόζει τα βάρη. Με αυτόν τον τρόπο, ο αλγόριθμος μπορεί να αυξήσει ή να μειώσει την επίδραση συγκεκριμένων παραμέτρων μέχρι να προσδιορίσει τον κατάλληλο μαθηματικό μετασχηματισμό για την αποτελεσματική επεξεργασία των δεδομένων.

3.2 Οπισθοδιάδοση (Backpropagation)

Η οπισθοδιάδοση (backpropagation) είναι μια μέθοδος υπολογισμού της κλίσης (gradient) που χρησιμοποιείται συνήθως για την εκπαίδευση ενός νευρωνικού δικτύου, με σκοπό την ενημέρωση των παραμέτρων του.

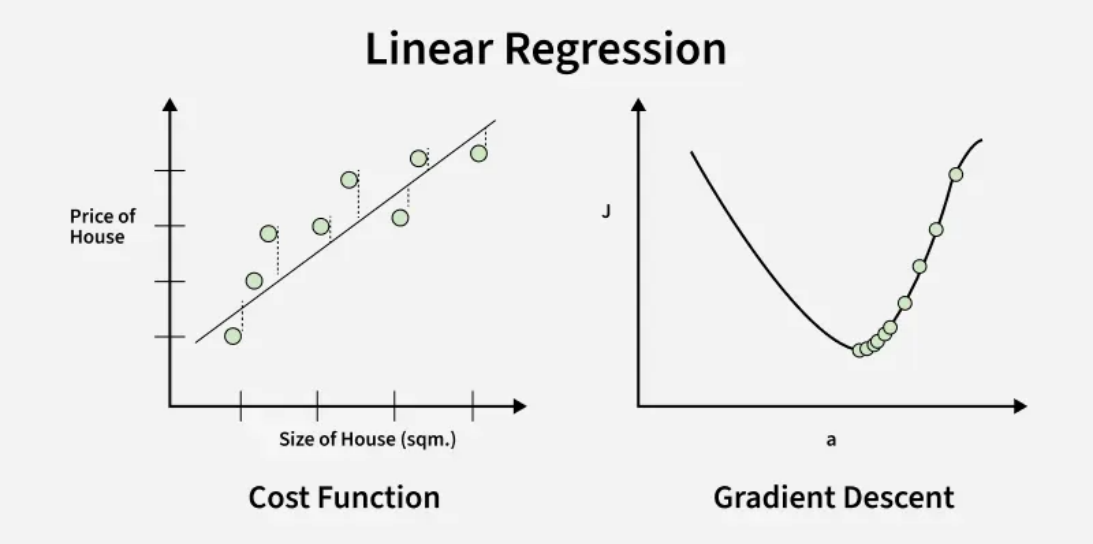
Αποτελεί μια αποδοτική εφαρμογή του κανόνα αλυσίδας (chain rule) στα νευρωνικά δίκτυα. Η οπισθοδιάδοση υπολογίζει την κλίση μιας συνάρτησης απώλειας (loss function) ως προς τα βάρη του δικτύου για ένα μόνο παράδειγμα εισόδου–εξόδου, και το κάνει αποδοτικά, υπολογίζοντας την κλίση ένα επίπεδο κάθε φορά, επαναλαμβάνοντας προς τα πίσω από το τελευταίο επίπεδο, ώστε να αποφεύγονται περιττοί υπολογισμοί ενδιάμεσων όρων στον κανόνα αλυσίδας· αυτό μπορεί να εξαχθεί μέσω δυναμικού προγραμματισμού.

3.3 Αλγόριθμοι Βελτιστοποίησης Νευρωνικών Δικτύων

Οι αλγόριθμοι βελτιστοποίησης (optimizers) στα νευρωνικά δίκτυα είναι τα μαθηματικά εργαλεία που αναλαμβάνουν να εκπαιδεύσουν το μοντέλο, τροποποιώντας συνεχώς τα βάρη του (weights) ώστε να ελαχιστοποιηθεί το σφάλμα πρόβλεψης (loss function).

3.3.1 Gradient Descent

Η καθοδική κλίση (gradient descent) είναι ένας αλγόριθμος βελτιστοποίησης. Χρησιμοποιείται για να βρεθεί πιο γρήγορα η ελάχιστη τιμή μιας συνάρτησης. Πρόκειται για έναν αλγόριθμο εύρεσης του ελαχίστου μιας κυρτής συνάρτησης (convex function). Για να το επιτύχει αυτό, μεταβάλλει επαναληπτικά τις παραμέτρους της εκάστοτε συνάρτησης. Είναι ένας αλγόριθμος που χρησιμοποιείται, για παράδειγμα, στη γραμμική παλινδρόμηση (linear regression).



Εικόνα 8: Gradient Descent

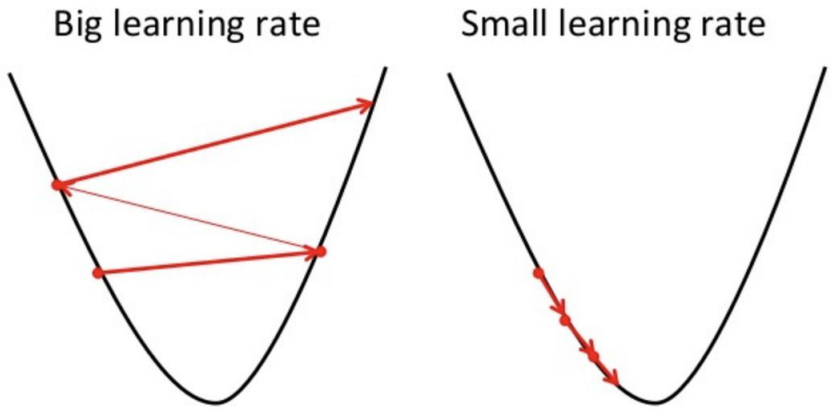
Πηγή: https://medium.com/@seyiakinsanya3/gradient-descent-in-classmachine-learning-vs-neural-networks-key-differences-and-implementation-719a24dbff52

3.3.1.1 Συνάρτηση Κόστους (Loss Function) και Ρυθμός Μάθησης (Learning Rate)

Παρόμοια με την εύρεση της ευθείας βέλτιστης προσαρμογής (best fit) στη γραμμική παλινδρόμηση, ο στόχος της καθοδικής κλίσης (gradient descent) είναι να ελαχιστοποιήσει τη συνάρτηση κόστους, δηλαδή το σφάλμα μεταξύ της προβλεπόμενης και της πραγματικής τιμής y. Για να το επιτύχει αυτό, απαιτούνται δύο βασικά στοιχεία: μια κατεύθυνση και ένας ρυθμός μάθησης (learning rate). Αυτοί οι παράγοντες επιτρέπουν στον αλγόριθμο να συγκλίνει σταδιακά σε ένα τοπικό ή καθολικό ελάχιστο (δηλαδή στο σημείο σύγκλισης).

- **Ρυθμός μάθησης (learning rate ή step size ή α)**: Είναι το μέγεθος των βημάτων που γίνονται για την επίτευξη του ελαχίστου. Συνήθως είναι μια μικρή τιμή και προσαρμόζεται με βάση τη συμπεριφορά της συνάρτησης κόστους. Υψηλός ρυθμός μάθησης οδηγεί σε μεγάλα βήματα αλλά ενέχει τον κίνδυνο υπέρβασης του ελαχίστου (overshooting). Αντίθετα, χαμηλός ρυθμός μάθησης σημαίνει μικρά βήματα. Αν και προσφέρει μεγαλύτερη ακρίβεια, απαιτεί περισσότερες επαναλήψεις και άρα περισσότερο χρόνο και υπολογιστική ισχύ.
- **Συνάρτηση κόστους (cost ή loss function)**: Μετρά τη διαφορά (σφάλμα) μεταξύ της πραγματικής τιμής και της προβλεπόμενης τιμής στη δεδομένη στιγμή. Παρέχει ανατροφοδότηση στο μοντέλο, ώστε να προσαρμόζει τις παραμέτρους του και να μειώνει το σφάλμα, οδηγώντας το προς το τοπικό ή καθολικό ελάχιστο. Ο αλγόριθμος επαναλαμβάνει τη διαδικασία, κινούμενος προς την κατεύθυνση της μέγιστης μείωσης (αρνητική κλίση), μέχρι η τιμή της συνάρτησης κόστους να πλησιάσει ή να φτάσει στο μηδέν. Σε αυτό το σημείο, η μάθηση σταματά.

Τέλος, αν και οι όροι «cost function» και «loss function» χρησιμοποιούνται συχνά ως συνώνυμοι, υπάρχει μια μικρή διαφορά: η loss function αναφέρεται στο σφάλμα για ένα μόνο παράδειγμα εκπαίδευσης, ενώ η cost function υπολογίζει το μέσο σφάλμα σε ολόκληρο το σύνολο εκπαίδευσης.



Εικόνα 9: Μεγάλα βήματα vs μικρά βήματα για τον ρυθμό μάθησης.

Πηγή: https://medium.com/@poojayadav9674/gradient-descent-cd4783e81f91

3.3.2 Adam (Adaptive Moment Estimation)

Το Adam είναι ένας από τους πιο δημοφιλείς αλγόριθμους βελτιστοποίησης, καθώς συνδυάζει τις καλύτερες ιδέες τόσο από το Momentum όσο και από το RMSprop. [14]

Χρησιμοποιεί τόσο τον κινούμενο μέσο όρο των κλίσεων (gradients) όσο και τον κινούμενο μέσο όρο των τετραγώνων των κλίσεων για να προσαρμόζει τον ρυθμό μάθησης.

Το Adam προσαρμόζει τον ρυθμό μάθησης για κάθε παράμετρο ξεχωριστά, επιτρέποντάς του να διαχειρίζεται καλύτερα «θορυβώδεις» κλίσεις (noisy gradients) και διαφορετικά μεγέθη τιμών σε σύγκριση με τη στοχαστική καθοδική κλίση (SGD). Είναι ιδιαίτερα ισχυρό σε μοντέλα βαθιάς μάθησης, όπου το τοπίο βελτιστοποίησης μπορεί να είναι αρκετά περίπλοκο και απαιτητικό.

Το Adam παρακολουθεί όχι μόνο την τρέχουσα κλίση (gradient), αλλά και την πρώτη ροπή (τον μέσο όρο των προηγούμενων κλίσεων) καθώς και τη δεύτερη ροπή (τον μέσο όρο των τετραγώνων των κλίσεων). Αυτό σημαίνει ότι για κάθε παράμετρο του μοντέλου αποθηκεύει επιπλέον πληροφορία, κάτι που μπορεί να οδηγήσει σε αυξημένες απαιτήσεις μνήμης, ιδιαίτερα σε πολύ μεγάλα μοντέλα.

Για την πρώτη ροπή, ο τύπος είναι του momentum:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \frac{dL}{dw_t}$$

-όπου:

m_t: είναι ο κινούμενος μέσος όρος των κλίσεων, τη στιγμή t

w_t: είναι το βάρος την στιγμή t

β_1: είναι η παράμετρος του momentum (συνήθως έχει τιμή 0.9)

$\frac{dL}{dw_t}$: η κλίση της συνάρτησης απώλειας σε σχέση με τα βάρη τη στιγμή t

Για την δεύτερη ροπή, ο τύπος είναι του RMSProp:

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\,(\frac{dL}{dw_t})^2$$

-όπου:

v_t: είναι ο εκθετικά μέσος όρος

Επειδή τόσο το m_t όσο και το v_t αρχικοποιούνται στο μηδέν, τείνουν να είναι μεροληπτικά προς το μηδέν, ιδιαίτερα στα αρχικά βήματα. Για να διορθωθεί αυτή η μεροληψία, ο Adam υπολογίζει διορθωμένες εκτιμήσεις (bias-corrected estimates):

$$\hat{m}_t = \frac{m_t}{1 - \beta_{1(t)}}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_{2(t)}}$$

Οπότε τα βάρη ενημερώνονται ως εξής:

$$w_{t+1} = w_t - \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} a$$

-όπου:

α: Ο ρυθμός μάθησης (προκαθορισμένος είναι 0.001)

β_1 και β_2: οι παράμετροι για τους κινούμενους μέσους όρους του gradient και του τετραγώνου του gradient, αντίστοιχα. Συνήθως, $\beta_1 = 0.9$ και $\beta_2 = 0.999$.

ε: ένας πολύ μικρός θετικός αριθμός, (πχ. 10^{-8}) που χρησιμοποιείται για να αποφευχθεί η διαίρεση με το 0.

Ο Adam αντιμετωπίζει αρκετές προκλήσεις της βελτιστοποίησης με gradient descent:

- Δυναμικοί ρυθμοί μάθησης (Dynamic learning rates): Κάθε παράμετρος έχει τον δικό της προσαρμοστικό ρυθμό μάθησης, βασισμένο στα προηγούμενα gradients και τα μεγέθη τους. Αυτό βοηθά τον βελτιστοποιητή να αποφεύγει ταλαντώσεις και να ξεπερνά πιο αποτελεσματικά τα τοπικά ελάχιστα.
- Διόρθωση μεροληψίας (Bias correction): Με τη διόρθωση της αρχικής μεροληψίας (όταν οι εκτιμήσεις πρώτης και δεύτερης ροπής είναι κοντά στο μηδέν), αποφεύγεται η αστάθεια στα πρώτα στάδια της εκπαίδευσης.
- Αποδοτική απόδοση (Efficient performance): Ο Adam συνήθως απαιτεί λιγότερη ρύθμιση υπερπαραμέτρων σε σύγκριση με άλλους αλγόριθμους όπως το SGD, γεγονός που τον καθιστά πιο εύχρηστη επιλογή για τα περισσότερα προβλήματα.

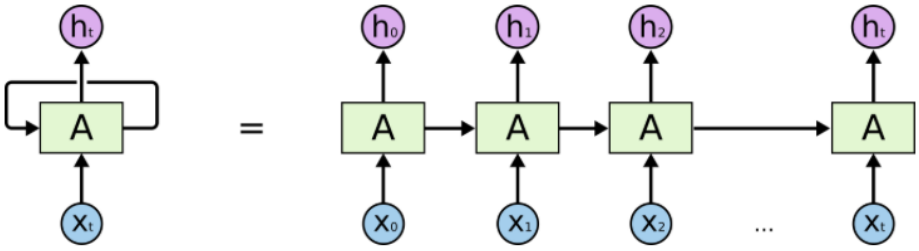
Σε σύγκριση με άλλους βελτιστοποιητές όπως το SGD (Stochastic Gradient Descent) και το SGD με momentum, ο Adam συνήθως υπερέχει τόσο σε χρόνο εκπαίδευσης όσο και σε ακρίβεια σύγκλισης. Η ικανότητά του να προσαρμόζει τον ρυθμό μάθησης για κάθε παράμετρο, σε συνδυασμό με τον μηχανισμό διόρθωσης μεροληψίας (bias correction), οδηγεί σε ταχύτερη σύγκλιση και πιο σταθερή βελτιστοποίηση.

Αυτό καθιστά τον Adam ιδιαίτερα χρήσιμο σε πολύπλοκα μοντέλα με μεγάλα σύνολα δεδομένων, καθώς αποφεύγει την αργή σύγκλιση και την αστάθεια, ενώ ταυτόχρονα πλησιάζει αποτελεσματικά το καθολικό ελάχιστο.

Σε πρακτικούς όρους, το Adam προσαρμόζεται στο «τοπίο» της συνάρτησης απώλειας, κάνοντας μεγαλύτερα βήματα όταν βρίσκεται μακριά από το βέλτιστο σημείο και μικρότερα βήματα όσο πλησιάζει σε αυτό, οδηγώντας έτσι σε ταχύτερη σύγκλιση.

3.4 Νευρωνικά Δίκτυα με Ανατροφοδότηση (Recurrent Neural Network - RNN)

Τα νευρωνικά δίκτυα με ανατροφοδότηση (RNNs) έχουν σχεδιαστεί για την επεξεργασία ακολουθιακών δεδομένων (sequential data), όπως το κείμενο, η ομιλία και οι χρονοσειρές, όπου η σειρά των στοιχείων είναι σημαντική. Σε αντίθεση με τα προωθητικά νευρωνικά δίκτυα (feedforward), τα οποία επεξεργάζονται τις εισόδους ανεξάρτητα, τα RNNs χρησιμοποιούν συνδέσεις με ανατροφοδότηση, όπου η έξοδος ενός νευρώνα σε ένα χρονικό βήμα τροφοδοτείται ξανά ως είσοδος στο δίκτυο στο επόμενο χρονικό βήμα. Αυτό επιτρέπει στα RNNs να συλλαμβάνουν χρονικές εξαρτήσεις και μοτίβα μέσα στις ακολουθίες.



Εικόνα 10: Δομή RNN

Πηγή: https://medium.com/@jianqiangma/all-about-recurrent-neural-networks-9e5ae2936f6e

Το βασικό δομικό στοιχείο ενός RNN είναι η επαναλαμβανόμενη μονάδα, η οποία διατηρεί μια κρυφή κατάσταση (hidden state) - μια μορφή μνήμης που ενημερώνεται σε κάθε χρονικό βήμα με βάση την τρέχουσα είσοδο και την προηγούμενη κρυφή κατάσταση. Αυτός ο μηχανισμός ανατροφοδότησης επιτρέπει στο δίκτυο να μαθαίνει από προηγούμενες εισόδους και να ενσωματώνει αυτή τη γνώση στην τρέχουσα επεξεργασία. Τα RNNs έχουν εφαρμοστεί με επιτυχία σε εργασίες όπως η αναγνώριση μη τμηματοποιημένου συνεχόμενου χειρόγραφου κειμένου, η αναγνώριση ομιλίας, η επεξεργασία φυσικής γλώσσας και η νευρωνική μηχανική μετάφραση.

Ωστόσο, τα παραδοσιακά RNNs αντιμετωπίζουν το πρόβλημα της εξαφάνισης της κλίσης (vanishing gradient), το οποίο περιορίζει την ικανότητά τους να μαθαίνουν εξαρτήσεις μεγάλης εμβέλειας. Το ζήτημα αυτό αντιμετωπίστηκε με την ανάπτυξη της αρχιτεκτονικής Long Short-Term Memory (LSTM) το 1997, καθιστώντας την τυπική παραλλαγή RNN για τη διαχείριση μακροπρόθεσμων εξαρτήσεων. Αργότερα, εισήχθησαν οι μονάδες Gated Recurrent Unit (GRU) ως μια υπολογιστικά πιο αποδοτική εναλλακτική λύση.

Τα τελευταία χρόνια, οι μετασχηματιστές (transformers), οι οποίοι βασίζονται σε μηχανισμούς αυτο-προσοχής (self-attention) αντί για επαναληπτικότητα, έχουν καταστεί η κυρίαρχη αρχιτεκτονική για πολλές εργασίες επεξεργασίας ακολουθιών, ιδιαίτερα στην επεξεργασία φυσικής γλώσσας, λόγω της ανώτερης διαχείρισης των εξαρτήσεων μεγάλης εμβέλειας και της μεγαλύτερης δυνατότητας παραλληλοποίησης.

Παρ' όλα αυτά, τα RNNs παραμένουν σημαντικά για εφαρμογές όπου η υπολογιστική αποδοτικότητα, η επεξεργασία σε πραγματικό χρόνο ή η εγγενώς ακολουθιακή φύση των δεδομένων είναι κρίσιμη.

Ένα RNN είναι μια συνάρτηση f_θ, όπου:

$$(x_t, h_t) \mapsto (y_t, h_{t+1})$$

- x_t: είσοδος διανύσματος
- h_t: κρυφό διάνυσμα (hidden vector)
- y_t: έξοδος διανύσματος
- θ: παράμετροι νευρωνικού δικτύου

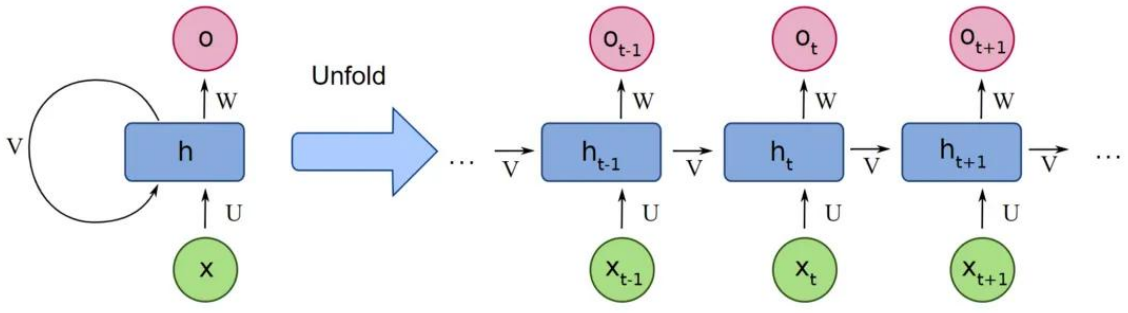
Με λίγα λόγια, πρόκειται για ένα νευρωνικό δίκτυο που αντιστοιχίζει μια είσοδο x_t σε μια έξοδο y_t, όπου το κρυφό διάνυσμα h_t παίζει τον ρόλο της «μνήμης» (το σημείο όπου αποθηκεύεται ό,τι έχει μάθει μέχρι εκείνη τη στιγμή), δηλαδή ενός μερικού αρχείου όλων των προηγούμενων ζευγών εισόδου-εξόδου. Σε κάθε βήμα, μετασχηματίζει την είσοδο σε έξοδο και τροποποιεί τη «μνήμη» του, ώστε να βελτιώνει την απόδοσή του σε μελλοντική επεξεργασία.

Σε κάθε χρονικό βήμα t:

$$h_t = f(x_t, h_{t-1})$$

όπου:

- x_t: η τρέχουσα είσοδος (π.χ. λέξη)
- h_{t-1}: το προηγούμενο hidden state
- h_t: το νέο hidden state



Εικόνα 11: Δομή RNN (2)

Πηγή: https://medium.com/deeplearningbrasilia/deep-learning-recurrent-neural-networks-f9482a24d010

Το αρνητικό στα RNNs είναι ότι έχουν μικρή μνήμη.

Υπάρχουν τέσσερις κύριες κατηγορίες RNN, οι οποίες είναι:

1) One-to-one RNN: Αυτός είναι ο απλούστερος τύπος αρχιτεκτονικής νευρωνικού δικτύου, όπου υπάρχει μία είσοδος και μία έξοδος. Χρησιμοποιείται για απλές εργασίες ταξινόμησης, όπως η δυαδική ταξινόμηση, όπου δεν εμπλέκονται ακολουθιακά δεδομένα.
2) One-to-many RNN: Σε ένα One-to-Many RNN, το δίκτυο επεξεργάζεται μία είσοδο και παράγει πολλαπλές εξόδους με την πάροδο του χρόνου. Αυτό είναι χρήσιμο σε εργασίες όπου μία είσοδος οδηγεί σε μια ακολουθία προβλέψεων (εξόδων).
3) Για παράδειγμα, στην παραγωγή λεζάντας εικόνας (image captioning), μια εικόνα χρησιμοποιείται ως είσοδος για να δημιουργηθεί μια ακολουθία λέξεων ως περιγραφή.
4) Many-to-one RNN: Σε ένα Many-to-One RNN, το δίκτυο λαμβάνει μια ακολουθία εισόδων και παράγει μία μόνο έξοδο. Αυτός ο τύπος είναι χρήσιμος όταν απαιτείται η συνολική πληροφορία της ακολουθίας για να γίνει μία πρόβλεψη.
5) Για παράδειγμα, στην ανάλυση συναισθήματος (sentiment analysis), το μοντέλο δέχεται μια ακολουθία λέξεων (όπως μια πρόταση) και παράγει μία έξοδο, όπως θετικό, αρνητικό ή ουδέτερο.
6) Many-to-many RNN: Ο τύπος Many-to-Many RNN επεξεργάζεται μια ακολουθία εισόδων και παράγει μια ακολουθία εξόδων. Σε εργασίες όπως η μετάφραση γλώσσας, μια ακολουθία λέξεων σε μία γλώσσα δίνεται ως είσοδος και παράγεται μια αντίστοιχη ακολουθία σε μια άλλη γλώσσα ως έξοδος.

3.5 Bidirectional RNN

Τα Bidirectional Recurrent Neural Networks (BRNNs) συνδέουν δύο κρυφά επίπεδα (hidden layers) με αντίθετες κατευθύνσεις προς την ίδια έξοδο. Με αυτή τη μορφή generative deep learning, το επίπεδο εξόδου μπορεί να λαμβάνει πληροφορία ταυτόχρονα από παρελθοντικές (backward) και μελλοντικές (forward) καταστάσεις.

Τα BRNNs εφευρέθηκαν το 1997 από τους Schuster και Paliwal [15] και εισήχθησαν με στόχο να αυξήσουν την ποσότητα της πληροφορίας εισόδου που είναι διαθέσιμη στο δίκτυο. Τα κλασικά recurrent neural networks (RNNs) επίσης έχουν περιορισμούς, καθώς η μελλοντική πληροφορία δεν είναι προσβάσιμη από την τρέχουσα κατάσταση.

Αντίθετα, τα BRNNs μπορούν να αξιοποιήσουν πληροφορία από το μέλλον, η οποία είναι διαθέσιμη από την τρέχουσα κατάσταση.

Τα BRNNs είναι ιδιαίτερα χρήσιμα όταν απαιτείται κατανόηση του συμφραζομένου της εισόδου. Για παράδειγμα, στην αναγνώριση χειρόγραφου κειμένου, η απόδοση μπορεί να βελτιωθεί σημαντικά όταν το μοντέλο γνωρίζει τα γράμματα που βρίσκονται πριν και μετά από το τρέχον γράμμα.

Ένα Bidirectional Recurrent Neural Network (BRNN) είναι μια επέκταση του παραδοσιακού RNN που επεξεργάζεται ακολουθιακά δεδομένα και προς τις δύο κατευθύνσεις (forward και backward). Αυτό

επιτρέπει στο δίκτυο να αξιοποιεί τόσο το παρελθόν όσο και το μέλλον της ακολουθίας όταν κάνει προβλέψεις, προσφέροντας μια πιο ολοκληρωμένη κατανόηση της πληροφορίας.

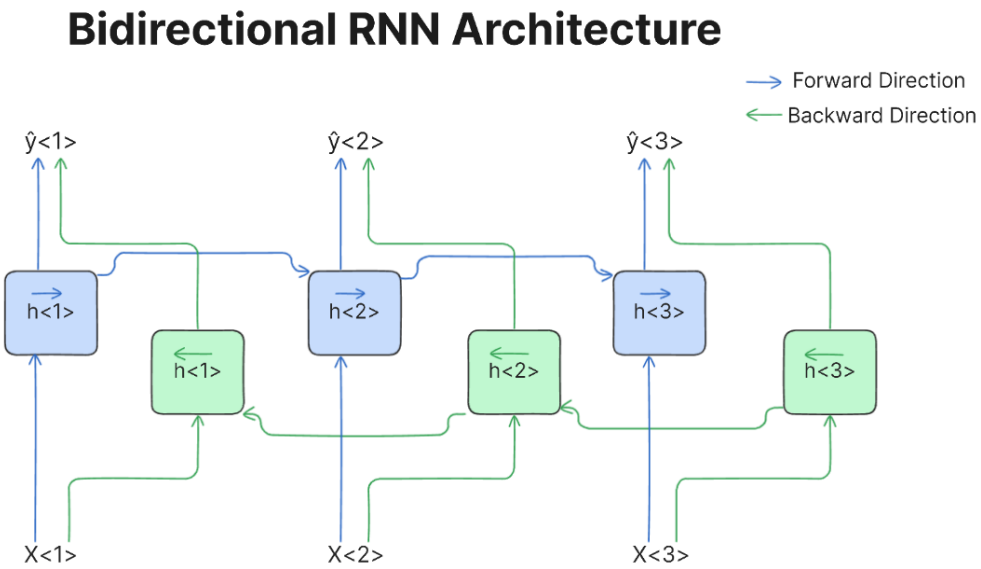
Όπως ένα κλασικό RNN, ένα BRNN προχωράει μπροστά μέσα στην ακολουθία, ενημερώνοντας την κρυφή κατάσταση (hidden state) με βάση την τρέχουσα είσοδο και την προηγούμενη κρυφή κατάσταση σε κάθε χρονικό βήμα. Η βασική διαφορά είναι ότι το BRNN διαθέτει και ένα backward hidden layer, το οποίο επεξεργάζεται την ακολουθία αντίστροφα, ενημερώνοντας την κρυφή κατάσταση με βάση την τρέχουσα είσοδο και την επόμενη χρονική κατάσταση.

Σε σύγκριση με τα unidirectional RNNs, τα BRNNs βελτιώνουν την ακρίβεια επειδή λαμβάνουν υπόψη τόσο το παρελθόν όσο και το μέλλον. Αυτό συμβαίνει επειδή τα δύο hidden layers (forward και backward) αλληλοσυμπληρώνονται και οι προβλέψεις παράγονται από τον συνδυασμό των εξόδων και των δύο κατευθύνσεων.

Παράδειγμα:

«Μου αρέσουν τα μήλα. Είναι τόσο υγιεινά.»

Σε ένα παραδοσιακό RNN, το δίκτυο μπορεί να δυσκολευτεί να καταλάβει αν η λέξη "apple" αναφέρεται στο φρούτο ή στην εταιρεία, βασιζόμενο μόνο στην πρώτη πρόταση. Ωστόσο, ένα BRNN δεν έχει τέτοιο πρόβλημα. Επεξεργαζόμενο την πρόταση και προς τις δύο κατευθύνσεις, μπορεί εύκολα να καταλάβει ότι το "apple" αναφέρεται στο φρούτο, χάρη στο μελλοντικό συμφραζόμενο που παρέχει η δεύτερη πρόταση ("Είναι τόσο υγιεινά.").



Εικόνα 12: Αρχιτεκτονική Bidirectional RNN.

Πηγή: https://medium.com/@imfuryfist/rnns-explained-the-neural-networks-that-remember-c34c34a57ac6

Πλεονεκτήματα των BRNNs:

- Βελτιωμένη κατανόηση συμφραζομένων: Λαμβάνουν υπόψη τόσο το παρελθόν όσο και το μέλλον, οδηγώντας σε πιο ακριβείς προβλέψεις.
- Υψηλότερη ακρίβεια: Είναι ιδιαίτερα αποτελεσματικά σε εργασίες NLP και επεξεργασίας ομιλίας.
- Καλύτερη διαχείριση ακολουθιών μεταβλητού μήκους: Είναι πιο ευέλικτα από τα παραδοσιακά RNNs, άρα κατάλληλα για δεδομένα με διαφορετικά μήκη.
- Μεγαλύτερη ανθεκτικότητα: Η διπλή επεξεργασία βοηθά στο φιλτράρισμα θορύβου και άσχετων πληροφοριών, βελτιώνοντας την αξιοπιστία.

Προκλήσεις των BRNNs:

- Υψηλό υπολογιστικό κόστος: Απαιτούν περίπου διπλάσια επεξεργασία σε σχέση με τα unidirectional RNNs.
- Μεγαλύτερος χρόνος εκπαίδευσης: Περισσότερες παράμετροι σημαίνουν πιο αργή σύγκλιση.
- Περιορισμένη χρήση σε πραγματικό χρόνο: Επειδή χρειάζονται ολόκληρη την ακολουθία για προβλέψεις, δεν είναι ιδανικά για real-time εφαρμογές όπως live αναγνώριση ομιλίας.
- Λιγότερη ερμηνευσιμότητα: Η διπλή κατεύθυνση επεξεργασίας κάνει πιο δύσκολη την κατανόηση του πώς προκύπτουν οι προβλέψεις.

Τα BRNNs χρησιμοποιούνται ευρέως σε εργασίες NLP, όπως:

- Ανάλυση Συναισθήματος (Sentiment Analysis): Χρησιμοποιούν το πλήρες συμφραζόμενο για καλύτερη κατηγοριοποίηση συναισθήματος.
- Αναγνώριση Οντοτήτων (NER): Εντοπίζουν οντότητες αξιοποιώντας το συμφραζόμενο και από τις δύο κατευθύνσεις.
- Μηχανική Μετάφραση: Σε encoder-decoder μοντέλα, βοηθούν τον encoder να κατανοήσει πλήρως την πρόταση πηγής, βελτιώνοντας την ποιότητα μετάφρασης.
- Αναγνώριση Ομιλίας: Βελτιώνουν την ακρίβεια μεταγραφής ήχου λαμβάνοντας υπόψη τόσο προηγούμενα όσο και επόμενα ηχητικά στοιχεία.

3.6 Long Short-Term Memory (LSTM)

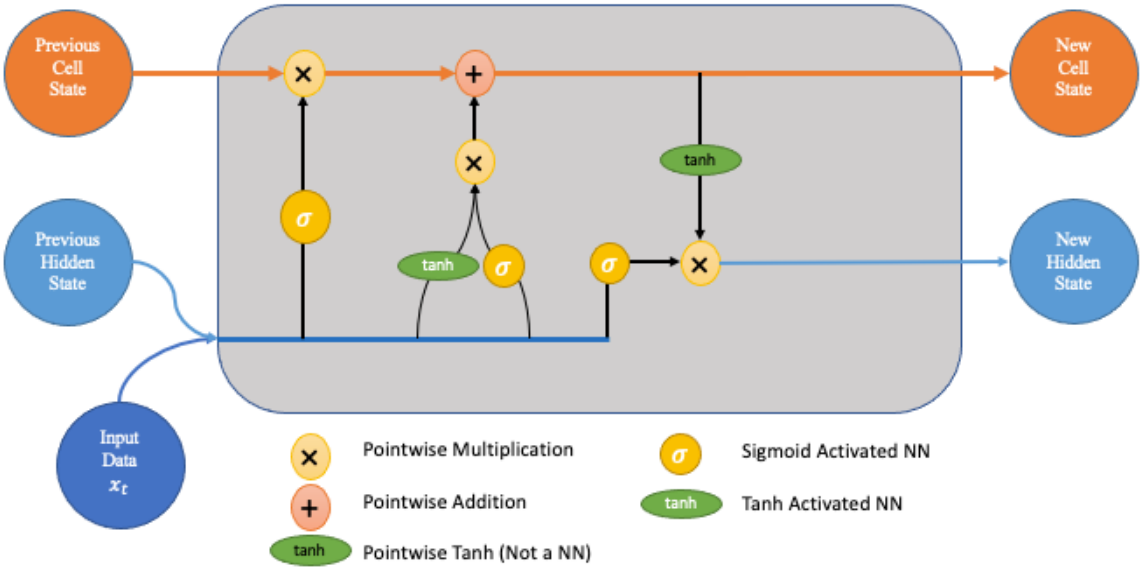
Η Long Short-Term Memory (LSTM) είναι ένας τύπος νευρωνικού δικτύου με ανατροφοδότηση (RNN) που έχει σχεδιαστεί για να αντιμετωπίζει το πρόβλημα της εξαφάνισης της κλίσης (vanishing gradient), το οποίο εμφανίζεται συχνά στα παραδοσιακά RNNs. Ένα από τα βασικά της πλεονεκτήματα είναι η σχετική της ανθεκτικότητα ως προς το μήκος των ακολουθιών, σε σύγκριση με άλλα RNNs. Στόχος της είναι να παρέχει μια μορφή «μνήμης» που μπορεί να διατηρεί πληροφορία για μεγάλο αριθμό χρονικών βημάτων (γι' αυτό και ονομάζεται «long short-term memory»). [16]

Μια μονάδα LSTM αποτελείται συνήθως από ένα «κελί» (cell) και τρεις «πύλες» (gates): την πύλη εισόδου (input gate), την πύλη εξόδου (output gate) και την πύλη λήθης (forget gate). Το κελί αποθηκεύει πληροφορίες για αυθαίρετα χρονικά διαστήματα, ενώ οι πύλες ρυθμίζουν τη ροή της πληροφορίας προς και από αυτό.

- Η **πύλη λήθης (forget gate)** αποφασίζει ποια πληροφορία από την προηγούμενη κατάσταση θα απορριφθεί, αντιστοιχίζοντας την προηγούμενη κατάσταση και την τρέχουσα είσοδο σε μια τιμή μεταξύ 0 και 1. Τιμή κοντά στο 1 σημαίνει διατήρηση της πληροφορίας, ενώ τιμή κοντά στο 0 σημαίνει απόρριψη.
- Η **πύλη εισόδου (input gate)** καθορίζει ποια νέα πληροφορία θα αποθηκευτεί στην τρέχουσα κατάσταση του κελιού, με παρόμοιο μηχανισμό.
- Η **πύλη εξόδου (output gate)** ελέγχει ποια πληροφορία από την τρέχουσα κατάσταση θα εξαχθεί ως έξοδος, επίσης μέσω τιμών μεταξύ 0 και 1.

Με την επιλεκτική εξαγωγή σημαντικών πληροφοριών, το LSTM μπορεί να διατηρεί χρήσιμες μακροπρόθεσμες εξαρτήσεις και να κάνει πιο ακριβείς προβλέψεις τόσο για το παρόν όσο και για μελλοντικά χρονικά βήματα.

Τα LSTM χρησιμοποιούνται ευρέως σε εφαρμογές όπως ταξινόμηση, επεξεργασία δεδομένων, ανάλυση χρονοσειρών, αναγνώριση ομιλίας, μηχανική μετάφραση, κ.α.



Εικόνα 13: Δομή LSTM.

Πηγή: https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9/

3.7 Sequence-to-Sequence (seq2seq)

Το seq2seq (sequence-to-sequence) είναι ένα μοντέλο που μετατρέπει μια ακολουθία εισόδου σε ένα διάνυσμα πραγματικών αριθμών χρησιμοποιώντας ένα νευρωνικό δίκτυο (encoder) και στη συνέχεια τη μετατρέπει ξανά σε μια ακολουθία εξόδου χρησιμοποιώντας ένα δεύτερο νευρωνικό δίκτυο (decoder). [17]

Η ιδέα της μετατροπής ακολουθιών μέσω αρχιτεκτονικής encoder–decoder αναπτύχθηκε στις αρχές της δεκαετίας του 2010.

Στην αρχική μορφή του seq2seq που προτάθηκε τότε, τόσο ο encoder όσο και ο decoder ήταν βασισμένοι σε LSTM. Αυτό όμως δημιουργούσε το πρόβλημα του «λαιμού μπουκαλιού» (bottleneck), καθώς το διάνυσμα κωδικοποίησης είχε σταθερό μέγεθος. Έτσι, για μεγάλες ακολουθίες εισόδου, μέρος της πληροφορίας χανόταν, επειδή ήταν δύσκολο να συμπιεστεί σε ένα διάνυσμα σταθερού μήκους. [17]

Ο μηχανισμός attention, που προτάθηκε επίσης το 2014 [18], έλυσε αυτό το πρόβλημα. Το μοντέλο ονομάστηκε RNNsearch, επειδή «προσομοιώνει» μια διαδικασία αναζήτησης μέσα στην πρόταση εισόδου κατά τη διάρκεια της αποκωδικοποίησης.

Ένα επιπλέον πρόβλημα των seq2seq μοντέλων ήταν ότι τα νευρωνικά δίκτυα με ανατροφοδότηση (RNNs) είναι δύσκολο να παραλληλοποιηθούν. Αυτό το πρόβλημα επιλύθηκε το 2017 με την εισαγωγή των Transformers, οι οποίοι αντικατέστησαν τον encoder RNN με blocks αυτο-προσοχής (self-attention) και τον decoder RNN με blocks που χρησιμοποιούν cross-attention με causal masking.

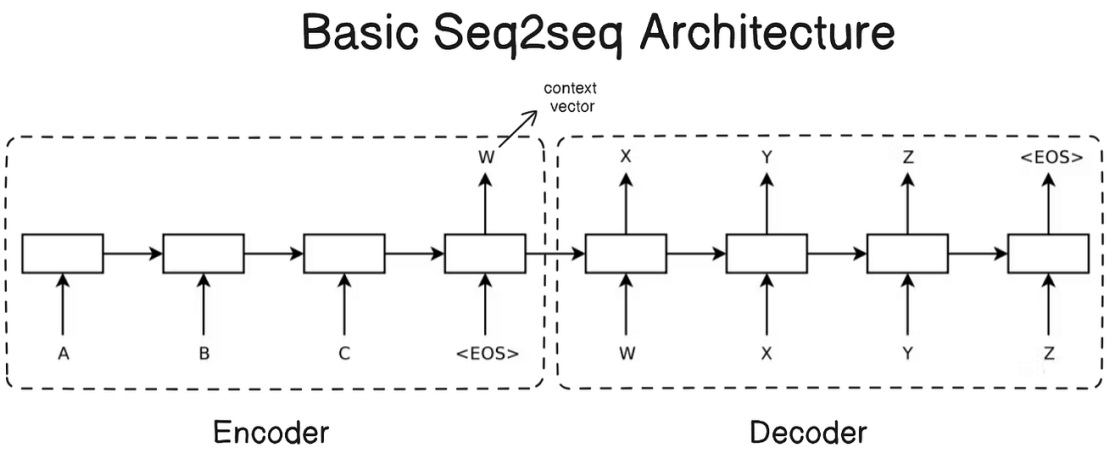
ENCODER

Ο encoder είναι υπεύθυνος για την επεξεργασία της ακολουθίας εισόδου και την εξαγωγή της βασικής της πληροφορίας, η οποία αποθηκεύεται ως κρυφή κατάσταση (hidden state) του δικτύου και, σε μοντέλα με μηχανισμό attention, ως διάνυσμα συμφραζομένων (context vector). Το context vector αποτελεί ένα σταθμισμένο άθροισμα των κρυφών καταστάσεων της εισόδου και δημιουργείται για κάθε χρονική στιγμή στην ακολουθία εξόδου.

DECODER

Ο decoder λαμβάνει το context vector και τις κρυφές καταστάσεις από τον encoder και παράγει την τελική ακολουθία εξόδου. Λειτουργεί με αυτοπαλίνδρομο (autoregressive) τρόπο, δημιουργώντας ένα στοιχείο της εξόδου κάθε φορά. Σε κάθε βήμα, λαμβάνει υπόψη τα προηγούμενα παραγόμενα στοιχεία, το context vector και την πληροφορία της εισόδου, ώστε να προβλέψει το επόμενο στοιχείο της ακολουθίας.

Συγκεκριμένα, σε μοντέλα με attention, το context vector και η κρυφή κατάσταση ενώνονται (concatenate) για να σχηματίσουν ένα διάνυσμα attention, το οποίο χρησιμοποιείται ως είσοδος στον decoder.



Εικόνα 14: Βασική αρχιτεκτονική seq2seq.

Πηγή: https://medium.com/@fraidoonomarzai99/seq2seq-encoder-decoder-and-attention-mechanism-in-depth-417ef88037f1

Μερικές από τις εφαρμογές είναι η μηχανική μετάφραση, σύνοψη κειμένου, αναγνώριση ομιλίας, κ.α.

3.8 Μηχανισμός Attention με RNN

Ένα πρόβλημα της προσέγγισης seq2seq με RNN (encoder-decoder) είναι ότι το νευρωνικό δίκτυο πρέπει να μπορεί να συμπιέσει όλη την απαραίτητη πληροφορία μιας πρότασης εισόδου, σε ένα διάνυσμα σταθερού μήκους. Αυτό δυσκολεύει το νευρωνικό δίκτυο να διαχειριστεί μεγάλες προτάσεις, ιδιαίτερα εκείνες που είναι μεγαλύτερες από τις προτάσεις που υπήρχαν στο σώμα εκπαίδευσης.

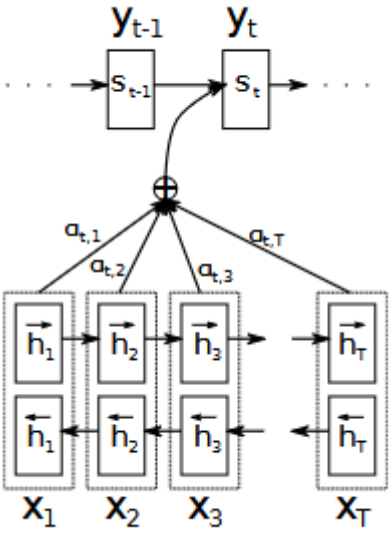
Για να αντιμετωπιστεί αυτό το πρόβλημα, εισήχθη μια επέκταση του μοντέλου encoder-decoder με attention. Κάθε φορά που το προτεινόμενο μοντέλο δημιουργεί μια λέξη στη μετάφραση, πραγματοποιεί μια «μαλακή» αναζήτηση (soft-search) για ένα σύνολο θέσεων στην πρόταση εισόδου, όπου συγκεντρώνεται η πιο σχετική πληροφορία. [18]

Στη συνέχεια, το μοντέλο προβλέπει τη λέξη στόχο βασιζόμενο:

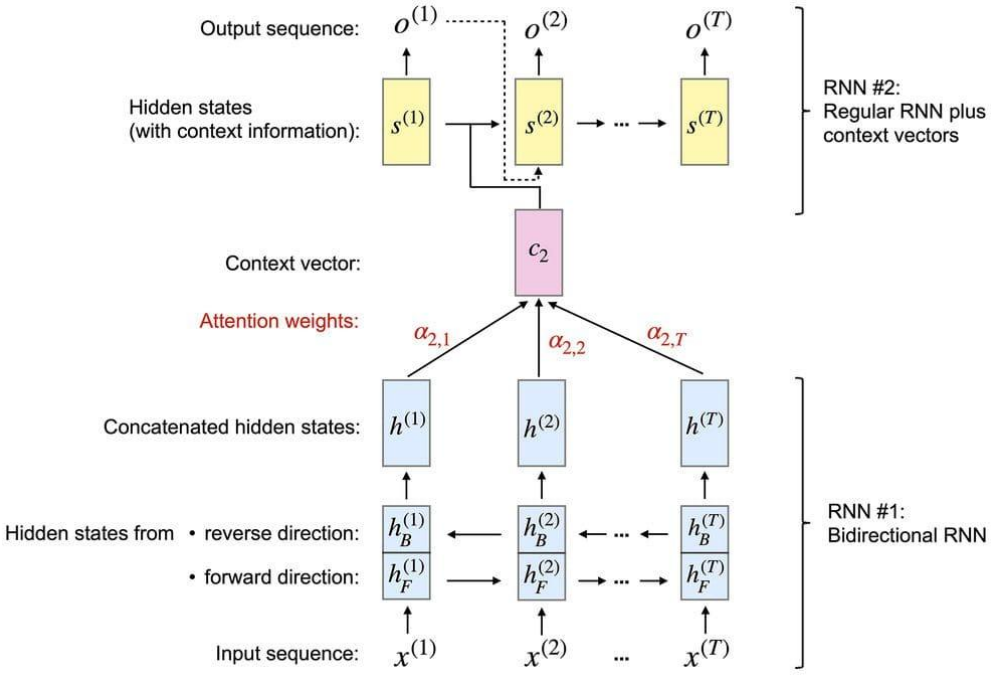
- στα context vectors που σχετίζονται με αυτές τις θέσεις της πρότασης εισόδου, και
- σε όλες τις προηγουμένως παραγόμενες λέξεις στόχου.

Το σημαντικότερο χαρακτηριστικό που διαφοροποιεί αυτή την προσέγγιση από το βασικό μοντέλο encoder-decoder είναι ότι δεν προσπαθεί να κωδικοποιήσει ολόκληρη την πρόταση εισόδου σε ένα μοναδικό διάνυσμα σταθερού μήκους.

Από πιθανοθεωρητική σκοπιά, η μετάφραση ισοδυναμεί με την εύρεση μιας πρότασης στόχου y που μεγιστοποιεί τη δεσμευμένη πιθανότητα της y δεδομένης μιας πρότασης εισόδου x, δηλαδή:
$argmax_y p\ (y|x)$



Εικόνα 15: Γραφική αναπαράσταση του μοντέλου encoder-decoder με attention. Πηγή: [18]



Εικόνα 16: Παράδειγμα μηχανισμού attention με RNN.

Πηγή: https://medium.com/latinxinai/understanding-transformers-f90b4783a65b

3.9 Transformers

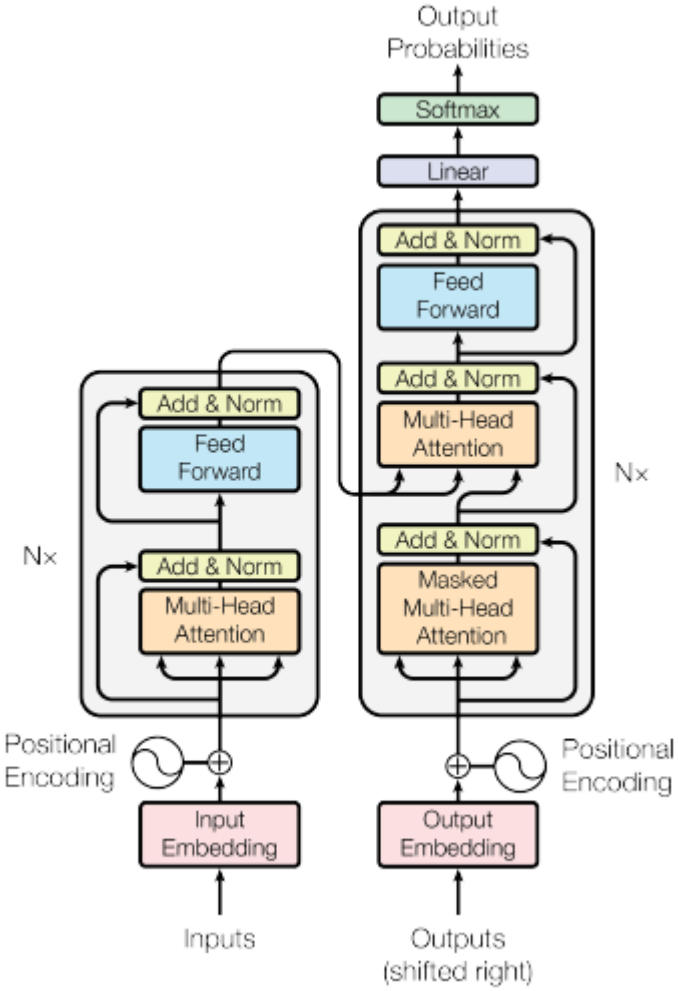
Τα μοντέλα Transformers προέρχονται από μια ερευνητική εργασία της Google το 2017. Το πρώτο μοντέλο παρουσιάστηκε στο paper με τίτλο «Attention Is All You Need» [12]. Η εισαγωγή τους προκάλεσε μια τεράστια άνθηση στον χώρο. Αυτό το επαναστατικό μοντέλο έθεσε τα θεμέλια για μεταγενέστερες σημαντικές εξελίξεις στον τομέα των μεγάλων γλωσσικών μοντέλων (Large Language Models), συμπεριλαμβανομένου του BERT και GPT. Ήδη από το 2018, αυτές οι εξελίξεις θεωρούνταν σημείο καμπής για το NLP.

Το 2020, η OpenAI ανακοίνωσε το GPT-3. Μέσα σε λίγες εβδομάδες, η ευελιξία του GPT-3 αποδείχθηκε εντυπωσιακή, καθώς οι χρήστες το αξιοποιούσαν για δημιουργία ποιημάτων, προγραμμάτων, τραγουδιών, ιστοσελίδων και πολλών άλλων εφαρμογών, κεντρίζοντας τη φαντασία ανθρώπων παγκοσμίως.

Οι Transformers είναι μια αρχιτεκτονική νευρωνικού δικτύου που χρησιμοποιείται σε διάφορες εργασίες μηχανικής μάθησης, ιδιαίτερα στην επεξεργασία φυσικής γλώσσας (NLP) και στην υπολογιστική όραση (computer vision). Εστιάζει στην κατανόηση των σχέσεων μέσα στα δεδομένα, ώστε να επεξεργάζεται την πληροφορία πιο αποτελεσματικά.

- Χρησιμοποιεί μηχανισμούς attention για να αποτυπώνει σχέσεις μεταξύ των εισόδων.
- Επεξεργάζεται ολόκληρες ακολουθίες ταυτόχρονα αντί βήμα προς βήμα (π.χ. RNN, LSTM), εκμεταλλεύοντας την ισχύ των σύγχρονων GPUs. Έτσι, αυξάνουν σημαντικά την ταχύτητα της επεξεργασίας και της εκπαίδευσης.
- Βελτιώνει την απόδοση σε εργασίες που απαιτούν κατανόηση συμφραζομένων και εξαρτήσεων.
- Χρησιμοποιείται ευρέως σε εφαρμογές NLP, computer vision και άλλους τομείς της τεχνητής νοημοσύνης.

Η καλή προσαρμογή στον παράλληλο υπολογισμό (parallelism) επιτρέπει στα μοντέλα Transformer να αξιοποιούν πλήρως την ισχύ και την ταχύτητα που προσφέρουν οι GPUs τόσο κατά την εκπαίδευση (training) όσο και κατά την εξαγωγή προβλέψεων (inference).



Εικόνα 17: Το μοντέλο Transformer. Πηγή: [12]

Εφαρμογές:

- **Εργασίες NLP (Natural Language Processing):**
  Οι transformers χρησιμοποιούνται σε:
  - Μηχανική μετάφραση,
  - Περίληψη κειμένου,
  - Αναγνώριση οντοτήτων (Named Entity Recognition),
  - και ανάλυση συναισθήματος (sentiment analysis).
- **Αναγνώριση Ομιλίας (Speech Recognition):**
  Επεξεργάζονται ηχητικά σήματα για να μετατρέψουν την ομιλία σε κείμενο.
- **Υπολογιστική Όραση (Computer Vision):**
  Οι transformers εφαρμόζονται σε:
  - Ταξινόμηση εικόνων,
  - Ανίχνευση αντικειμένων,
  - και δημιουργία εικόνων.
- **Συστήματα Προτάσεων (Recommendation Systems):**
  Παρέχουν εξατομικευμένες προτάσεις με βάση τις προτιμήσεις των χρηστών.
- **Παραγωγή Κειμένου και Μουσικής:**
  Οι transformers χρησιμοποιούνται για:
  - Δημιουργία άρθρων και κειμένων,
  - καθώς και σύνθεση μουσικής.

Περιορισμοί:

- Απαιτούν υψηλή υπολογιστική ισχύ, γεγονός που καθιστά την εκπαίδευσή τους ακριβή.
- Καταναλώνουν μεγάλη ποσότητα μνήμης, ιδιαίτερα σε μεγάλες ακολουθίες.
- Χρειάζονται μεγάλα datasets για να αποδώσουν καλά.
- Έχουν περιορισμούς στο μήκος εισόδου, γεγονός που δυσκολεύει τη διαχείριση πολύ μεγάλων κειμένων.
- Μπορεί να λειτουργούν σαν “μαύρο κουτί” (black box), μειώνοντας την ερμηνευσιμότητα.
- Υπάρχει κίνδυνος overfitting αν δεν εφαρμοστεί σωστή regularization.

3.10 BERT

Το Bidirectional Encoder Representations from Transformers (BERT) είναι ένα γλωσσικό μοντέλο που παρουσιάστηκε τον Οκτώβριο του 2018 από ερευνητές της Google [11]. Μαθαίνει να αναπαριστά το κείμενο ως μια ακολουθία διανυσμάτων (vectors) χρησιμοποιώντας αυτο-επιβλεπόμενη μάθηση (self-supervised learning). Βασίζεται αποκλειστικά στην αρχιτεκτονική Transformer Encoder. Το BERT βελτίωσε θεαματικά την κορυφαία απόδοση (state of the art) των μεγάλων γλωσσικών μοντέλων. Από το 2020, το BERT αποτελεί μια ευρέως χρησιμοποιούμενη βασική γραμμή αναφοράς (baseline) σε πειράματα Επεξεργασίας Φυσικής Γλώσσας (NLP).

Το BERT εκπαιδεύεται μέσω των εργασιών πρόβλεψης καλυμμένων tokens (masked token prediction) και πρόβλεψης της επόμενης πρότασης (next sentence prediction). Επεξεργάζεται το κείμενο αμφίδρομα (bidirectionally), λαμβάνοντας υπόψη τόσο το αριστερό όσο και το δεξί συμφραζόμενο κάθε λέξης .Μέσα από αυτή την εκπαίδευση, το BERT μαθαίνει συμφραζόμενες (contextual) και κρυφές/λανθάνουσες (latent) αναπαραστάσεις των tokens μέσα στο περιβάλλον τους, παρόμοια με τα ELMo και GPT-2. Είναι σχεδιασμένο κυρίως για εργασίες κατανόησης γλώσσας και όχι για παραγωγή κειμένου.

Αρχικά, το BERT υλοποιήθηκε για την αγγλική γλώσσα σε δύο μεγέθη μοντέλων:

- **BERTBASE**: 110 εκατομμύρια παράμετροι
- **BERTLARGE**: 340 εκατομμύρια παράμετροι

Και τα δύο μοντέλα εκπαιδεύτηκαν στο σύνολο δεδομένων Toronto BookCorpus (800 εκατομμύρια λέξεις) και στην αγγλική έκδοση της Wikipedia (2,5 δισεκατομμύρια λέξεις). Τα εκπαιδευμένα βάρη (weights) δημοσιεύθηκαν στο GitHub. Στις 11 Μαρτίου 2020 κυκλοφόρησαν ακόμη 24 μικρότερα μοντέλα, με το μικρότερο να είναι το BERTTINY, το οποίο διαθέτει μόλις 4 εκατομμύρια παραμέτρους.

3.10.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ

Το BERT είναι μια αρχιτεκτονική Transformer τύπου «encoder-only», δηλαδή χρησιμοποιεί μόνο το τμήμα του Encoder και όχι Decoder. Σε υψηλό επίπεδο, το BERT αποτελείται από τέσσερα βασικά μέρη:

1. **Tokenizer (Μετατροπέας σε tokens)**
   Αυτό το τμήμα μετατρέπει ένα κομμάτι αγγλικού κειμένου σε μια ακολουθία ακεραίων αριθμών («tokens»).
2. **Embedding (Ενσωμάτωση)**
   Αυτό το τμήμα μετατρέπει την ακολουθία των tokens σε έναν πίνακα διανυσμάτων πραγματικών αριθμών που αναπαριστούν τα tokens. Ουσιαστικά, πραγματοποιεί τη μετατροπή των διακριτών τύπων tokens σε έναν ευκλείδειο χώρο μικρότερης διάστασης.
3. **Encoder (Κωδικοποιητής)**
   Πρόκειται για μια στοίβα από Transformer blocks που χρησιμοποιούν τον μηχανισμό self-attention, αλλά χωρίς causal masking (αιτιοκρατική μάσκα).
4. **Task Head (Κεφαλή Εργασίας)**
   Αυτό το τμήμα μετατρέπει τα τελικά διανύσματα αναπαράστασης ξανά σε tokens, παράγοντας μια προβλεπόμενη κατανομή πιθανοτήτων πάνω σε όλους τους πιθανούς τύπους tokens. Μπορεί να θεωρηθεί είτε ως ένας απλός αποκωδικοποιητής (decoder), που μετατρέπει τη λανθάνουσα

αναπαράσταση σε tokens, είτε ως ένα «un-embedding layer», δηλαδή το αντίστροφο του embedding layer.

Η task head είναι απαραίτητη κατά την προεκπαίδευση (pre-training), αλλά συχνά δεν χρειάζεται στις λεγόμενες downstream tasks, όπως:

- απάντηση ερωτήσεων (question answering),
- ταξινόμηση συναισθήματος (sentiment classification),
- αναγνώριση οντοτήτων (NER),
- ταξινόμηση κειμένου κ.ά.

Αντί γι' αυτό, αφαιρείται η αρχική task head και αντικαθίσταται από ένα νέο, τυχαία αρχικοποιημένο module, ειδικά σχεδιασμένο για τη συγκεκριμένη εργασία. Στη συνέχεια γίνεται fine-tuning ολόκληρου του μοντέλου ή μέρους αυτού.

Η λανθάνουσα διανυσματική αναπαράσταση (latent vector representation) που παράγει το BERT τροφοδοτείται απευθείας στο νέο module. Με αυτόν τον τρόπο επιτυγχάνεται transfer learning με λίγα δείγματα (sample-efficient transfer learning), επειδή το μοντέλο έχει ήδη μάθει γενικές γλωσσικές γνώσεις κατά την προεκπαίδευσή του και χρειάζεται πολύ λιγότερα δεδομένα για να προσαρμοστεί σε μια νέα εργασία.

3.11 RoBERTa

Η άνοδος των μοντέλων Transformer έφερε σημαντική πρόοδο στην επεξεργασία φυσικής γλώσσας (NLP), ιδιαίτερα με το BERT. Το RoBERTa (Robustly Optimized BERT Pretraining Approach) [19] διατήρησε την ίδια αρχιτεκτονική, αλλά βελτίωσε τη διαδικασία εκπαίδευσης ώστε να επιτύχει καλύτερα αποτελέσματα. Με μερικές σχετικά μικρές αλλαγές στο BERT, το RoBERTa παρήγαγε ισχυρότερες γλωσσικές αναπαραστάσεις χωρίς να αλλάξει τον βασικό σχεδιασμό του μοντέλου.

Το RoBERTa χρησιμοποιεί την ίδια δομή Transformer Encoder με το BERT, αλλά εισάγει αρκετές σημαντικές βελτιώσεις στον τρόπο εκπαίδευσης του μοντέλου.

1) Αφαίρεση του Next Sentence Prediction (NSP)

Η προεκπαίδευση του BERT περιλάμβανε μια εργασία γνωστή ως Next Sentence Prediction (NSP), κατά την οποία το μοντέλο εκπαιδευόταν να καθορίζει αν δύο προτάσεις εμφανίζονταν διαδοχικά στο αρχικό κείμενο. Ο στόχος ήταν να μάθει καλύτερα τις σχέσεις μεταξύ προτάσεων.

Μετέπειτα μελέτες έδειξαν ότι το NSP συνεισέφερε ελάχιστα στην απόδοση ορισμένων εργασιών και σε κάποιες περιπτώσεις μπορούσε ακόμη και να εισάγει θόρυβο στην εκπαίδευση. Για τον λόγο αυτό, το RoBERTa αφαιρεί πλήρως το NSP και επικεντρώνεται αποκλειστικά στο Masked Language Modeling (MLM), επιτρέποντας στο μοντέλο να εστιάσει στην εκμάθηση καλύτερων συμφραζομένων σε επίπεδο tokens.

2) Στρατηγική Δυναμικής Απόκρυψης (Dynamic Masking)

Το BERT χρησιμοποιεί στατική απόκρυψη (static masking), όπου τα tokens που θα αποκρυφθούν επιλέγονται μία φορά κατά την προ-επεξεργασία και το ίδιο μοτίβο απόκρυψης χρησιμοποιείται σε κάθε εποχή εκπαίδευσης (epoch). Αυτό περιορίζει την ποικιλία των συμφραζομένων που βλέπει το μοντέλο και μπορεί να οδηγήσει σε υπερ-προσαρμογή (overfitting) στα συγκεκριμένα μοτίβα απόκρυψης.

Το RoBERTa αντικαθιστά αυτήν τη μέθοδο με δυναμική απόκρυψη (dynamic masking), όπου οι θέσεις των masked tokens επιλέγονται τυχαία σε κάθε πέρασμα εκπαίδευσης. Έτσι, το μοντέλο εκτίθεται σε μεγαλύτερη ποικιλία μοτίβων απόκρυψης, γεγονός που βελτιώνει τη γενίκευση και οδηγεί σε πιο ανθεκτική κατανόηση των συμφραζομένων.

3) Μεγαλύτερα Batch Sizes και Περισσότερος Χρόνος Εκπαίδευσης

Η εκπαίδευση μοντέλων βαθιάς μάθησης απαιτεί σωστή ισορροπία μεταξύ αποδοτικότητας και απόδοσης. Το BERT εκπαιδεύτηκε με σχετικά μικρά batch sizes (256 ακολουθίες) και συγκεκριμένο αριθμό βημάτων εκπαίδευσης.

Το RoBERTa αύξησε σημαντικά αυτή την κλίμακα:

- Τα batch sizes αυξήθηκαν έως και στις 8.000 ακολουθίες.
- Η διάρκεια της εκπαίδευσης επεκτάθηκε σε περισσότερα βήματα.
- Οι ρυθμοί μάθησης (learning rates) και τα προγράμματα βελτιστοποίησης (optimization schedules) ρυθμίστηκαν πιο αποτελεσματικά.

Οι αλλαγές αυτές παρέχουν πιο σταθερές ενημερώσεις των gradients και επιτρέπουν στο μοντέλο να μάθει βαθύτερα γλωσσικά μοτίβα χωρίς καμία τροποποίηση στην αρχιτεκτονική του.

4) Διευρυμένο Σύνολο Δεδομένων Εκπαίδευσης

Μία από τις σημαντικότερες βελτιώσεις του RoBERTa είναι η χρήση ενός πολύ μεγαλύτερου και πιο ποικίλου συνόλου δεδομένων.

Ενώ το BERT εκπαιδεύτηκε σε περίπου 16 GB κειμένου από τη Wikipedia και το BookCorpus, το RoBERTa εκπαιδεύτηκε σε πάνω από 160 GB κειμένου, συμπεριλαμβανομένων:

- Common Crawl News
- OpenWebText
- Stories Dataset
- Βιβλία και Wikipedia (όπως και το BERT)

Αυτή η τεράστια αύξηση των δεδομένων εκπαίδευσης εκθέτει το μοντέλο σε πλουσιότερες γλωσσικές δομές και περισσότερα θεματικά πεδία, βοηθώντας το να γενικεύει καλύτερα σε πραγματικές εφαρμογές και εργασίες NLP.

3.12 DeBERTa

Το DeBERTa [20] προτείνει έναν νέο μηχανισμό στον οποίο η ίδια πληροφορία αποθηκεύεται σε δύο διαφορετικά διανύσματα (*vectors*).

Ας θυμηθούμε τον αρχικό Transformer. Στο αρχικό block του Transformer, κάθε token αναπαρίσταται από ένα μοναδικό διάνυσμα (*vector*), το οποίο περιέχει τόσο πληροφορία για το περιεχόμενο της λέξης όσο και για τη θέση της στην ακολουθία.

Το μειονέκτημα αυτής της προσέγγισης είναι η πιθανή απώλεια πληροφορίας.

Αφού το περιεχόμενο και η θέση συγχωνεύονται σε ένα μόνο διάνυσμα, το μοντέλο μπορεί να δυσκολευτεί να καταλάβει αν μια συγκεκριμένη διάσταση του embedding είναι σημαντική λόγω:

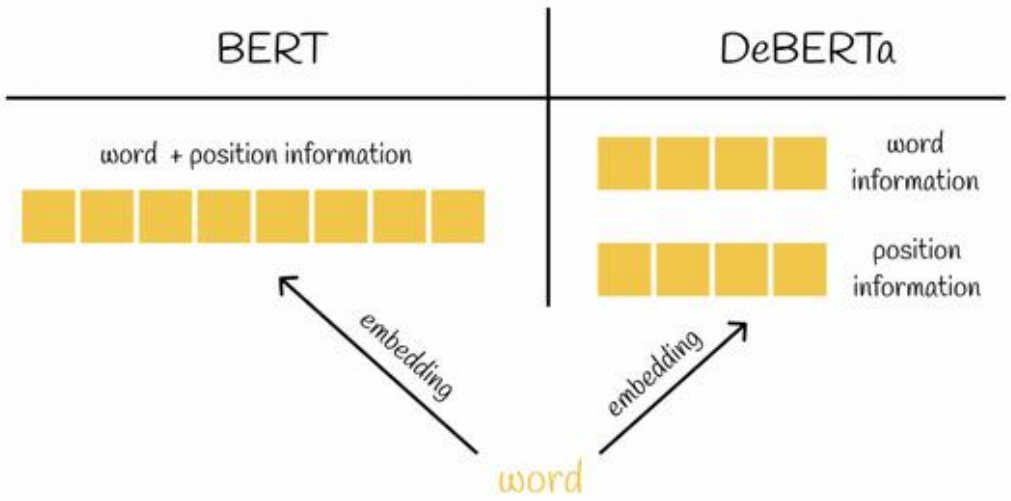
- της ίδιας της λέξης,
- της θέσης της λέξης,
- ή κάποιου συνδυασμού των δύο.

Με άλλα λόγια, μετά την πρόσθεση των embeddings, η πληροφορία για το «τι είναι η λέξη» και το «πού βρίσκεται η λέξη» δεν είναι πλέον σαφώς διαχωρισμένη.

Για παράδειγμα, αν το μοντέλο θεωρεί σημαντική μια συγκεκριμένη συνιστώσα του embedding, δεν είναι εύκολο να γνωρίζει αν αυτή η σημασία προέρχεται:

- από τη λέξη "γάτα",
- από το γεγονός ότι βρίσκεται στην αρχή της πρότασης,
- ή και από τα δύο μαζί.

Το Disentangled Attention του DeBERTa λύνει αυτό το πρόβλημα κρατώντας ξεχωριστές τις αναπαραστάσεις περιεχομένου και θέσης, ώστε το μοντέλο να μπορεί να επεξεργάζεται ανεξάρτητα αυτές τις δύο πηγές πληροφορίας κατά τον υπολογισμό της προσοχής (*attention*).



Εικόνα 18: Παράδειγμα word embedding DeBERTa

Πηγή: https://towardsdatascience.com/large-language-models-deberta-decoding-enhanced-bert-with-disentangled-attention-90016668db4b/

Επιπλέον, ο αλγόριθμος υπολογισμού του attention τροποποιείται, ώστε να λαμβάνει ρητά υπόψη τις σχέσεις μεταξύ του περιεχομένου (*content*) και της θέσης (*position*) των tokens.

Για παράδειγμα, οι λέξεις:

«διπλωματική» και «εργασία»

συνδέονται πολύ πιο στενά όταν εμφανίζονται κοντά η μία στην άλλη παρά όταν βρίσκονται σε απομακρυσμένα σημεία του κειμένου.

Αυτό το παράδειγμα δείχνει ξεκάθαρα γιατί είναι απαραίτητο να λαμβάνονται υπόψη όχι μόνο οι σχέσεις μεταξύ των περιεχομένων των λέξεων, αλλά και οι σχέσεις μεταξύ περιεχομένου και θέσης (*content-to-position relations*).

Το κλασικό BERT βασίζεται κυρίως σε απόλυτες θέσεις. Αντίθετα, το DeBERTa δίνει μεγαλύτερη έμφαση στις σχετικές θέσεις.

Έτσι μπορεί να μάθει πιο γενικούς κανόνες όπως:

- ένα επίθετο βρίσκεται συνήθως πριν από ένα ουσιαστικό
- το υποκείμενο συχνά προηγείται του ρήματος
- η λέξη «διπλωματική» συχνά ακολουθείται από τη λέξη «εργασία»

ανεξάρτητα από το πού εμφανίζονται μέσα στην πρόταση.

Αυτός είναι ένας από τους βασικούς λόγους που το Disentangled Attention αποδίδει καλύτερα από το κλασικό self-attention του BERT σε πολλές εργασίες κατανόησης φυσικής γλώσσας.

Επίσης, στο DeBERTa χρησιμοποιείται και ο Enhanced Mask Decoder, όπου διορθώνει μια αδυναμία του Disentangled Attention. Προστίθεται πληροφορία απόλυτης θέσης κατά την πρόβλεψη των μασκαρισμένων λέξεων, ώστε το μοντέλο να μπορεί να ξεχωρίζει καλύτερα λέξεις που έχουν παρόμοιο νόημα και παρόμοιο τοπικό συμφραζόμενο, αλλά διαφορετικό ρόλο μέσα στην πρόταση.

Τέλος, Ένα ιδιαίτερα εντυπωσιακό αποτέλεσμα είναι ότι, παρόλο που το DeBERTa εκπαιδεύτηκε με λιγότερα δεδομένα από το RoBERTa στις βασικές εκδόσεις του, κατάφερε να το ξεπεράσει σε αρκετά γνωστά benchmarks αξιολόγησης της κατανόησης φυσικής γλώσσας.

Αυτό δείχνει ότι οι βελτιώσεις του DeBERTa προέρχονται κυρίως από την αρχιτεκτονική του και όχι απλώς από τη χρήση περισσότερων δεδομένων ή περισσότερων παραμέτρων. Με άλλα λόγια, το DeBERTa είναι ένα χαρακτηριστικό παράδειγμα όπου ένας πιο έξυπνος σχεδιασμός του Transformer οδηγεί σε καλύτερες επιδόσεις.

4

Εργασίες επαλήθευσης συγγραφέα και διαγωνισμοί PAN

4.1 Εισαγωγή

Αρχικά να εξετάσουμε ότι έχουν δοθεί δύο κύριες κατηγορίες μεθόδων επαλήθευσης συγγραφέα:

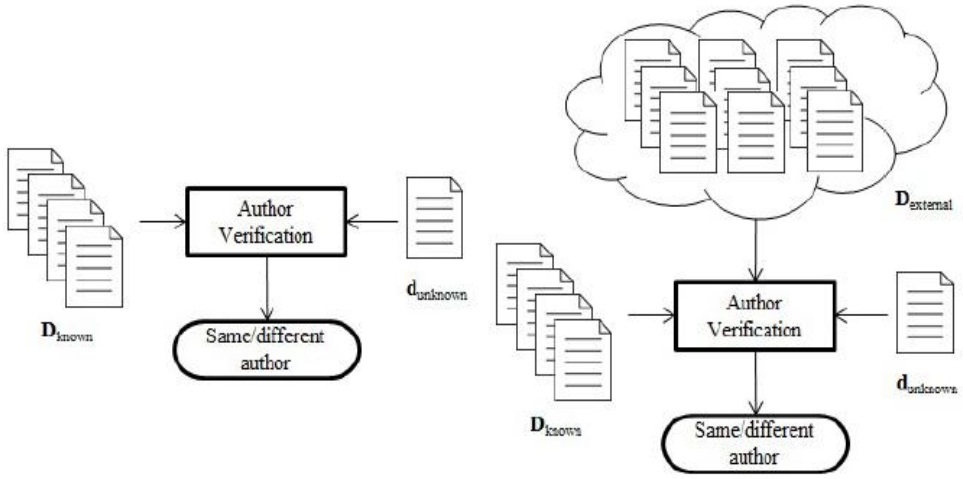
4.1.1 Intrinsic verification models vs Extrinsic verification models

1. Εσωτερικά μοντέλα επαλήθευσης (intrinsic verification models): Παρέχουν μια απόφαση βασισμένη μόνο στην ανάλυση των κειμένων, που βρέθηκαν σε ένα πρόβλημα επαλήθευσης (το άγνωστο και τα γνωστά κείμενα, που παρέχονται στο πρόβλημα). Αποκλείουν την χρησιμοποίηση εξωτερικών κειμένων άλλων συγγραφέων. Επομένως, τα μοντέλα intrinsic χειρίζονται το πρόβλημα επαλήθευσης συγγραφέα σαν μια κατηγορία ταξινόμησης, αποφεύγοντας οποιαδήποτε χρήση εξωτερικών-άλλων δεδομένων. Έτσι, αναλύουν το ύφος μόνο μέσα από τα κείμενα του ίδιου του συγγραφέα. Τα intrinsic μοντέλα είναι γρηγορότερα, αφού η ανάλυση περιορίζεται μόνο στο άγνωστο και στα γνωστά κείμενα που δίνονται. Επίσης, είναι ανθεκτικά σε θεματικές διαφορές, διότι δεν μαθαίνουν τα «θέματα», αλλά το στυλ του συγγραφέα.

2. Εξωτερικά μοντέλα επαλήθευσης (extrinsic verification models): Χρησιμοποιούν εξωτερικά κείμενα από άλλους συγγραφείς, για να εκτιμήσουν εάν υπάρχει σημαντική ομοιότητα μεταξύ του άγνωστου κειμένου και των γνωστών κειμένων. Τα μοντέλα αυτά μετατρέπουν το πρόβλημα επαλήθευσης συγγραφέα σε δυαδική ταξινόμηση (binary classification), όπου τα γνωστά κείμενα

ανήκουν στην θετική κλάση και τα άγνωστα κείμενα ανήκουν στην αρνητική κλάση. Στο training set έχουμε πολλά ζεύγη από πολλούς συγγραφείς και το μοντέλο μαθαίνει μοτίβα του συγγραφικού ύφους. Στο test set λαμβάνει ένα νέο ζεύγος (d_k, d_u) και αποφασίζει εάν είναι ο ίδιος συγγραφέας. Τα extrinsic μοντέλα είναι πιο αποτελεσματικά από τα intrinsic, αφού ένα δυαδικό πρόβλημα ταξινόμησης είναι ευκολότερο στο χειρισμό συγκριτικά με το πρόβλημα ταξινόμησης μιας κλάσης. Όμως, ένα σημαντικό θέμα σε αυτά τα μοντέλα είναι, η χρησιμοποίηση κατάλληλων εξωτερικών κειμένων. Είναι πολύ σημαντικό να χρησιμοποιηθούν εξωτερικά κείμενα που ανήκουν στο ίδιο είδος με αυτά του προβλήματος. Διαφορετικά θα επηρεαστεί αρνητικά η απόδοση. Επίσης, έχουν υψηλότερο χρόνο εκτέλεσης σε σύγκριση με τα intrinsic, διότι δέχονται και εξωτερικά κείμενα. Χρησιμοποιούνται πολύ σε μοντέλα μηχανικής και βαθιάς μάθησης

Αξίζει να σημειωθεί ότι η πιο επιδραστική μέθοδος αυτή της κατηγορίας ονομάζεται «Impostors» και εισήχθη από τους Koppel και Winter [23]. Το D_{known} αντιπροσωπεύει τη θετική κλάση, το $D_{external}$ την αρνητική κλάση και το d_u αναφέρεται σε ένα από τα δυο.



Εικόνα 19: Intrinsic verification model (αριστερά) vs Extrinsic verification model (δεξιά).
Πηγή: [22]

4.1.2 Eager Learning models vs Lazy Learning models

Από μια άλλη άποψη τα μοντέλα επαλήθευσης συγγραφέα μπορούν να χωριστούν από τον τρόπο μάθησης που χρησιμοποιούν:

1. Μοντέλα πρόθυμης μάθησης (eager learning models): Αυτά τα μοντέλα «μαθαίνουν εκ των προτέρων» από το σώμα εκπαίδευσης (training corpus). Δηλαδή, προ-εκπαιδεύονται σε πολλά ζεύγη γνωστών και άγνωστων κειμένων και δημιουργούν ένα γενικευμένο μοντέλο (μαθαίνουν μια

γενική συνάρτηση ομοιότητας συγγραφέα), το οποίο εφαρμόζεται σε νέα, άγνωστα ζεύγη κειμένων.

Προσπαθούν να εξάγουν ένα γενικό μοντέλο επαλήθευσης συγγραφέα με βάση τα δεδομένα εκπαίδευσης. Κάθε πρόβλημα επαλήθευσης αντιμετωπίζεται σαν ένα παράδειγμα είτε θετικό (όταν το άγνωστο και τα γνωστά κείμενα προέρχονται από τον ίδιο συγγραφέα) είτε αρνητικό (όταν το άγνωστο κείμενο προέρχεται από διαφορετικό συγγραφέα). Το σύνολο αυτών των παραδειγμάτων, από τα σώματα εκπαίδευσης, χρησιμοποιείται για να εκπαιδευτεί ένας δυαδικός ταξινομητής, που μπορεί, έπειτα, να χρησιμοποιηθεί για προβλέψει την πιο πιθανή κατηγορία σε οποιαδήποτε νέα περίπτωση επαλήθευσης. Τέτοια μοντέλα είναι αποτελεσματικά μόνο όταν το σώμα εκπαίδευσης είναι αντιπροσωπευτικό του προβλήματος επαλήθευσης. Η αποτελεσματικότητα και η πολυπλοκότητα εξαρτάται από το μέγεθος και τα χαρακτηριστικά του σώματος εκπαίδευσης (απαιτεί μεγάλο training set με πολλούς συγγραφείς). Επιπλέον, μπορούν να εκμεταλλευτούν τους αλγόριθμους επιβλεπόμενης μάθησης, όπως SVM, Λογιστική Παλινδρόμηση, νευρωνικά δίκτυα και έχουν πολύ γρήγορη εφαρμογή στο στάδιο της πρόβλεψης.

2. Μοντέλα τεμπέλικης μάθησης (lazy learning models): Αυτά τα μοντέλα δεν μαθαίνουν εκ των προτέρων. Δεν δημιουργούν ένα γενικευμένο μοντέλο κατά τη φάση εκπαίδευσης. Αντίθετα, αντιμετωπίζουν κάθε verification case ξεχωριστά, εκτελώντας τη μάθηση τη στιγμή της εφαρμογής.

   Αντιμετωπίζουν κάθε περίπτωση επαλήθευσης ξεχωριστά. Μόλις παρουσιαστεί μια νέα περίπτωση επαλήθευσης στη φάση της εφαρμογής, εκτελούν όλα τα απαραίτητα είδη ανάλυσης για να εκτιμήσουν την απάντηση. Το θετικό σε τέτοια μοντέλα είναι ότι δεν χρειάζονται εκπαίδευση ή σύνολα δεδομένων με επισήμανση (labeled dataset), αλλά το αρνητικό είναι ότι είναι πολύ χρονοβόρα στη φάση εφαρμογής, συγκριτικά με τα μοντέλα πρόθυμης μάθησης. Επίσης, το πλεονέκτημά τους είναι ότι δεν εξαρτώνται ιδιαίτερα από τις ιδιότητες του σώματος εκπαίδευσης. Είναι πιο ευέλικτα, αλλά λιγότερο αποδοτικά σε μεγάλη κλίμακα. Τέτοια μοντέλα είναι ο KNN ή General Impostors.

4.1.3 Profile-based models vs Instance-based models

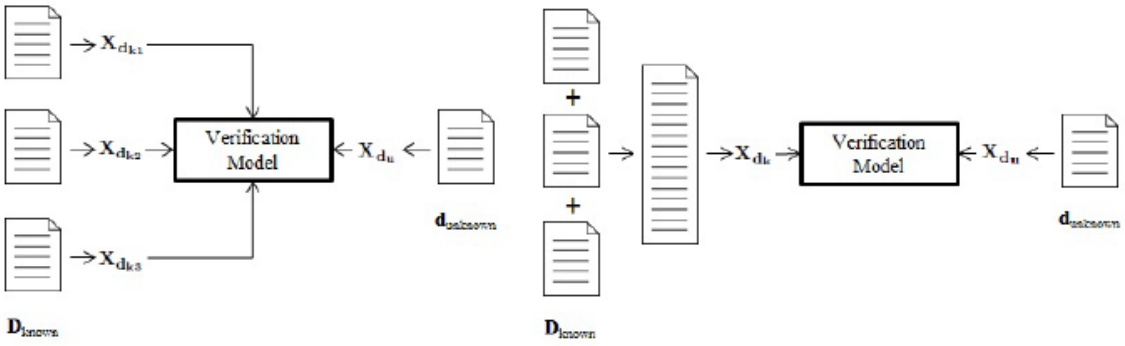
Τέλος, ένα άλλο διακριτό χαρακτηριστικό των μοντέλων επαλήθευσης αφορά τον τρόπο με τον οποίο διαχειρίζονται τα επισημασμένα κείμενα (γνωστά κείμενα) μέσα σε κάθε πρόβλημα επαλήθευση:

1) Μοντέλα βασισμένα στο προφίλ (profile-based models): Σε αυτά τα μοντέλα, όλα τα γνωστά κείμενα ενώνονται σε ένα ενιαίο κείμενο και έπειτα συγκρίνεται το ενωμένο κείμενο με το άγνωστο κείμενο. Ουσιαστικά, προσπαθούν να συλλέξουν τα στιλιστικά χαρακτηριστικά [21] του συγγραφέα, αφαιρώντας τυχόν διαφορές μεταξύ των παρεχόμενων κειμένων. Με την συνένωση όλων των χαρακτηριστικών, δημιουργούμε ένα διάνυσμα που αντιπροσωπεύει το «μέσο στυλ»

του. Έπειτα υπολογίζουμε την ομοιότητα μεταξύ του προφίλ του συγγραφέα με το άγνωστο κείμενο.

Ένα σημαντικό πλεονέκτημα αυτών των μοντέλων είναι, ότι όταν το μήκος του κειμένου είναι περιορισμένο, η συνένωση όλων των επισημασμένων κειμένων αυξάνει την ανθεκτικότητα της στιλομετρικής αναπαράστασης. Από την άλλη πλευρά, το ενωμένο κείμενο μπορεί να αποκλίνει σημαντικά από τα αρχικά κείμενα, ειδικά όταν διαφέρουν ως προς το θέμα ή το είδος.

2) Μοντέλα βασισμένα σε παραδείγματα (instance-based models): Αυτά τα μοντέλα χειρίζονται κάθε γνωστό κείμενο ξεχωριστά και το συγκρίνουν με το άγνωστο κείμενο. Δηλαδή, δεν δημιουργούν προφίλ για κάθε συγγραφέα, όπως τα profile-based, αλλά συγκρίνουν απευθείας δύο κείμενα και αποφασίζουν αν είναι γραμμένα από το ίδιο άτομο ή όχι. Τέτοια μοντέλα θεωρούν το κάθε γνωστό κείμενο σαν ένα ξεχωριστό παράδειγμα του ύφους του συγγραφέα. Όταν υπάρχουν πολλά γνωστά κείμενα, τα αποτελέσματα των επιμέρους συγκρίσεων συνδυάζονται για να παραχθεί η τελική απόφαση. Μια παραλλαγή αυτής της μεθόδου είναι να συνενωθούν αρχικά όλα τα γνωστά κείμενα και έπειτα να χωριστεί το τελικό κείμενο σε δείγματα ίσους μεγέθους. Αυτά τα μοντέλα μπορούν να εκμεταλλευτούν καλύτερα τις διαφορές μεταξύ των γνωστών κειμένων, εφόσον είναι ικανά να χειριστούν αποτελεσματικά το σύνολο των επιμέρους απαντήσεων (μια για κάθε γνωστό κείμενο). Ωστόσο, επηρεάζονται περισσότερο από περιορισμούς μήκος κειμένου. Επίσης, είναι αρκετά χρονοβόρα, διότι εξετάζεται κάθε ζεύγος κειμένου ξεχωριστά, σε αντίθεση με τα profile-based μοντέλα. Άλλο ένα μειονέκτημα είναι ότι χρειάζονται πολλά ζεύγη εκπαίδευσης για να είναι αποδοτικά. Τέλος, δεν αποθηκεύει ευρεία γνώση του ύφους του συγγραφέα, αφού εξετάζει κείμενο με κείμενο.



Εικόνα 20: Instance-based model (αριστερά) vs Profile-based model (δεξιά). Πηγή: [22]

4.2 Η μέθοδος General Impostors του Seidman [24]

Η μέθοδος General Impostors (GenIM) παρουσιάστηκε από τον Seidman το 2013 [24], στο πλαίσιο του διαγωνισμού PAN 2013 για την επαλήθευση συγγραφικής ταυτότητας (Authorship Verification). Η μέθοδος βασίζεται στην προγενέστερη προσέγγιση Impostors των Koppel et al. [23], την οποία προσαρμόζει για την αντιμετώπιση προβλημάτων επαλήθευσης συγγραφέα, αξιοποιώντας περισσότερα από ένα κείμενα γνωστής συγγραφικής προέλευσης.

Η βασική διαφοροποίηση της μεθόδου General Impostors είναι ότι δεν περιορίζεται στην εξέταση ενός μόνο ζεύγους κειμένων. Αντίθετα, επιτρέπει τη σύγκριση ενός άγνωστου εγγράφου με ένα σύνολο εγγράφων που έχουν γραφτεί από τον ίδιο γνωστό συγγραφέα. Με αυτόν τον τρόπο, αξιοποιούνται περισσότερα δείγματα του συγγραφικού ύφους, προκειμένου να ληφθεί μια συνολική απόφαση σχετικά με τη συγγραφική ταυτότητα του άγνωστου κειμένου.

Έστω ότι έχουμε ένα άγνωστο έγγραφο (X), του οποίου επιθυμούμε να επαληθεύσουμε τη συγγραφική προέλευση, και ένα σύνολο γνωστών εγγράφων:

$$Y = Y_1, Y_2, \dots, Y_n$$

όπου όλα τα έγγραφα (Y_i) έχουν γραφτεί από τον ίδιο συγγραφέα.

Η μέθοδος General Impostors εφαρμόζει την αρχική μέθοδο Impostors ξεχωριστά σε κάθε ζεύγος ((X, Y_i)). Για κάθε σύγκριση χρησιμοποιείται ένα σύνολο εξωτερικών εγγράφων, γνωστών ως impostors, τα οποία προέρχονται από άλλους συγγραφείς και λειτουργούν ως ανταγωνιστικά παραδείγματα.

Η διαδικασία επαλήθευσης βασίζεται σε επαναλαμβανόμενες συγκρίσεις ομοιότητας. Σε κάθε επανάληψη επιλέγεται τυχαία ένα ποσοστό των διαθέσιμων χαρακτηριστικών, καθώς και ένας αριθμός impostors από το εξωτερικό σύνολο εγγράφων. Στη συνέχεια εξετάζεται κατά πόσο η ομοιότητα μεταξύ των δύο εξεταζόμενων κειμένων είναι μεγαλύτερη από την αντίστοιχη ομοιότητα με τους impostors.

Στην υλοποίηση που περιγράφεται από τον Seidman, η σύγκριση βασίζεται στο γινόμενο των ομοιοτήτων των δύο εξεταζόμενων κειμένων και συγκρίνεται με το αντίστοιχο γινόμενο των ομοιοτήτων τους με κάθε impostor:

$$sim(X, Y_i) \cdot sim(Y_i, X) > sim(X, I_j) \cdot sim(Y_i, I_j)$$

για κάθε επιλεγμένο impostor (I_j).

Όταν η παραπάνω συνθήκη ικανοποιείται για όλους τους επιλεγμένους impostors, η συγκεκριμένη επανάληψη θεωρείται επιτυχής. Το αποτέλεσμα των επαναλήψεων χρησιμοποιείται για τον υπολογισμό ενός σκορ επαλήθευσης, το οποίο συγκρίνεται με ένα προκαθορισμένο κατώφλι απόφασης.

Για κάθε ζεύγος ((X, Y_i)), η αρχική μέθοδος Impostors παράγει μία δυαδική απόφαση:

$$S(X, Y_i) \in 0,1$$

όπου η τιμή 1 δηλώνει ότι τα δύο κείμενα ταξινομούνται ως προερχόμενα από τον ίδιο συγγραφέα, ενώ η τιμή 0 δηλώνει ότι ταξινομούνται ως προερχόμενα από διαφορετικούς συγγραφείς.

Στη συνέχεια, η μέθοδος General Impostors συνδυάζει τις επιμέρους αποφάσεις υπολογίζοντας τον μέσο όρο τους:

$$Score(X,Y) = \frac{1}{n}\sum_{i=1}^{n} S\,(X,Y_i)$$

Το τελικό σκορ λαμβάνει τιμές στο διάστημα ([0,1]) και εκφράζει το ποσοστό των γνωστών εγγράφων για τα οποία η μέθοδος Impostors έδωσε θετική απόφαση.

Για παράδειγμα, εάν υπάρχουν πέντε γνωστά έγγραφα και η μέθοδος ταξινομήσει τα τέσσερα ζεύγη ως προερχόμενα από τον ίδιο συγγραφέα, τότε:

$$Score(X,Y) = \frac{4}{5} = 0{,}80$$

Η τελική απόφαση προκύπτει από τη σύγκριση του σκορ με ένα κατώφλι (θ). Εάν το σκορ υπερβαίνει το επιλεγμένο κατώφλι, το άγνωστο έγγραφο θεωρείται ότι έχει γραφτεί από τον γνωστό συγγραφέα. Διαφορετικά, η υπόθεση κοινής συγγραφικής προέλευσης απορρίπτεται.

Στην αρχική εργασία εξετάστηκαν διαφορετικές αναπαραστάσεις κειμένων, όπως συχνότητες λέξεων, n-grams και character n-grams, καθώς και διαφορετικές μετρικές ομοιότητας, μεταξύ των οποίων η Min-Max similarity. Οι παράμετροι της μεθόδου, όπως ο αριθμός επαναλήψεων, το ποσοστό των επιλεγόμενων χαρακτηριστικών και τα κατώφλια απόφασης, προσαρμόστηκαν στις γλώσσες των εξεταζόμενων δεδομένων.

Επιπλέον, για τη δημιουργία του εξωτερικού συνόλου impostors χρησιμοποιήθηκαν κείμενα που συλλέχθηκαν από τον Ιστό μέσω μηχανών αναζήτησης. Η επιλογή τους είχε ως στόχο τη δημιουργία ενός κατάλληλου συνόλου ανταγωνιστικών εγγράφων, απέναντι στα οποία αξιολογούνταν η ομοιότητα των εξεταζόμενων κειμένων.

Το σημαντικότερο πλεονέκτημα της μεθόδου General Impostors είναι η δυνατότητα αξιοποίησης πολλαπλών γνωστών εγγράφων, αντί της αποκλειστικής εξάρτησης από μία μόνο σύγκριση. Με τον συνδυασμό των επιμέρους αποφάσεων επιδιώκεται μεγαλύτερη ανθεκτικότητα απέναντι σε διακυμάνσεις του συγγραφικού ύφους μεταξύ διαφορετικών κειμένων του ίδιου συγγραφέα.

Ωστόσο, η μέθοδος παρουσιάζει και ορισμένους περιορισμούς. Η επαναλαμβανόμενη εφαρμογή της αρχικής μεθόδου Impostors σε πολλά ζεύγη εγγράφων αυξάνει το υπολογιστικό κόστος. Επιπλέον, η αποτελεσματικότητά της εξαρτάται από την ποιότητα των impostors, την επιλογή των χαρακτηριστικών και την κατάλληλη ρύθμιση των παραμέτρων.

Συνολικά, η μέθοδος General Impostors αποτέλεσε μία σημαντική εξέλιξη της προσέγγισης Impostors, καθώς επέτρεψε την εφαρμογή της σε προβλήματα επαλήθευσης συγγραφικής ταυτότητας με πολλαπλά γνωστά έγγραφα. Η αποτελεσματικότητά της αναδείχθηκε στον διαγωνισμό PAN 2013, όπου η συγκεκριμένη προσέγγιση κατέλαβε την πρώτη θέση στη συνολική κατάταξη.

4.3 Η βελτιωμένη παραλλαγή της μεθόδου Impostors των Koppel & Winter [25]

Η προσέγγιση Impostors, η οποία είχε ήδη αξιοποιηθεί και επεκταθεί από τον Seidman [24], εξετάστηκε περαιτέρω από τους Koppel και Winter [25]. Η βασική της αρχή είναι ότι η ομοιότητα μεταξύ δύο εξεταζόμενων κειμένων δεν αξιολογείται μεμονωμένα, αλλά συγκριτικά με κείμενα άλλων συγγραφέων, τα οποία λειτουργούν ως ανταγωνιστικά παραδείγματα (impostors). Στη συγκεκριμένη εργασία, οι ερευνητές εξετάζουν τη συμπεριφορά της μεθόδου και τη σημασία παραμέτρων, όπως η επιλογή και ο αριθμός των impostors, για την αξιοπιστία της επαλήθευσης συγγραφικής ταυτότητας.

Η επιλογή των impostors γίνεται ως εξής:

Έστω ότι το ζεύγος προς επαλήθευση είναι:
(X, Y)

Για το Y δημιουργείται ένα σύνολο από m impostor documents:
$Y_1, Y_2, \ldots, Y_m$

τα οποία έχουν γραφτεί από άλλους συγγραφείς. Το Y ανταγωνίζεται τα συγκεκριμένα κείμενα ως προς την ομοιότητά τους με το X.

Η επιλογή των impostors αποτελεί κρίσιμο στοιχείο της μεθόδου. Εάν οι impostors είναι εντελώς άσχετοι με τα εξεταζόμενα κείμενα, η διάκριση γίνεται υπερβολικά εύκολη και υπάρχει κίνδυνος εμφάνισης πολλών false positives.

Αντίθετα, οι impostors πρέπει να αποτελούν σχετικά ισχυρούς και ρεαλιστικούς ανταγωνιστές. Η επιλογή impostors από κατάλληλο είδος μπορεί να περιορίσει την πιθανότητα το σύστημα να βασίζεται σε επιφανειακές διαφορές θεματολογίας ή είδους αντί σε χαρακτηριστικά συγγραφικού ύφους.

Ένα από τα σημαντικότερα στοιχεία της μεθόδου Impostors είναι ότι η σύγκριση δεν πραγματοποιείται μία μόνο φορά χρησιμοποιώντας ολόκληρο το σύνολο χαρακτηριστικών. Αντίθετα, εφαρμόζεται επαναλαμβανόμενη τυχαία υποδειγματοληψία χαρακτηριστικών (random feature subsampling).

Έστω ότι υπάρχει ένα συνολικό σύνολο χαρακτηριστικών:

$$F = \{f_1, f_2, \ldots, f_n\}$$

Σε κάθε επανάληψη r, επιλέγεται τυχαία μόνο ένα υποσύνολο:

$$F_r \subset F$$

και οι ομοιότητες των εγγράφων υπολογίζονται αποκλειστικά χρησιμοποιώντας τα χαρακτηριστικά του συγκεκριμένου υποσυνόλου.

Αν δύο κείμενα έχουν κοινά συγγραφικά χαρακτηριστικά, τότε η μεταξύ τους σχέση θα παραμένει ισχυρή ακόμη και όταν χρησιμοποιούνται διαφορετικά τυχαία υποσύνολα χαρακτηριστικών. Αντίθετα, μία τυχαία

ή θεματικά εξαρτώμενη ομοιότητα είναι πιθανότερο να εξαφανιστεί όταν αλλάξει το σύνολο των χαρακτηριστικών.

Για κάθε τυχαίο υποσύνολο χαρακτηριστικών υπολογίζεται αρχικά η ομοιότητα:

$$sim(X,Y)$$

και στη συνέχεια οι αντίστοιχες ομοιότητες:

$$sim(X,Y_1), sim(X,Y_2), \dots, sim(X,Y_m)$$

Το Y θεωρείται ότι κερδίζει μία επανάληψη όταν:

$$sim(X,Y) > sim(X,Y_i),\ \forall i = 1,\dots,m$$

δηλαδή όταν το πραγματικό κείμενο Y είναι περισσότερο όμοιο με το X από όλους τους impostors.

Οι Koppel και Winter επαναλαμβάνουν τη διαδικασία για 100 διαφορετικές επιλογές συνόλων χαρακτηριστικών. Έτσι, υπολογίζεται ένα σκορ το οποίο εκφράζει πόσο συχνά το Y καταφέρνει να ξεπεράσει όλους τους impostors:

$$score_X(Y) = \frac{1}{R}\sum_{r=1}^{R} I\left[sim_r(X,Y) > \max_i sim_r(X,Y_i)\right]$$

όπου R είναι ο αριθμός των επαναλήψεων.

Η συνάρτηση (I) είναι η συνάρτηση δείκτη (indicator function), η οποία λαμβάνει την τιμή 1 όταν η συνθήκη που βρίσκεται μέσα στις αγκύλες ικανοποιείται και την τιμή 0 σε αντίθετη περίπτωση. Επομένως, το τελικό σκορ εκφράζει το ποσοστό των επαναλήψεων στις οποίες το εξεταζόμενο κείμενο Y παρουσιάζει μεγαλύτερη ομοιότητα με το X από όλους τους impostors.

Για παράδειγμα, εάν πραγματοποιηθούν 100 επαναλήψεις και το Y είναι πιο όμοιο με το X από όλους τους impostors στις 80 από αυτές, τότε:

$score_X(Y) = 0.80$

Ένα υψηλό σκορ υποδεικνύει ότι η σχέση μεταξύ X και Y είναι σταθερή και ανθεκτική απέναντι τόσο στην αλλαγή των χαρακτηριστικών όσο και στην παρουσία ανταγωνιστικών κειμένων.

Το τελικό score συγκρίνεται με ένα προκαθορισμένο κατώφλι s^*. Εάν:

$S(X,Y) > s^*$

τότε το ζεύγος ταξινομείται ως «ίδιος συγγραφέας» ,ενώ διαφορετικά ταξινομείται ως: «διαφορετικός συγγραφέας».

Επομένως, η μέθοδος δεν βασίζεται απλώς στο αν δύο κείμενα έχουν υψηλή απόλυτη ομοιότητα. Βασίζεται στο κατά πόσο η ομοιότητά τους είναι συστηματικά ισχυρότερη από την ομοιότητα με κατάλληλα επιλεγμένους ανταγωνιστές.

Επίσης, Οι Koppel και Winter επισημαίνουν ότι υπάρχει σημαντική αλληλεπίδραση μεταξύ τριών παραμέτρων: της ποιότητας των impostors, του αριθμού των impostors και του threshold που χρησιμοποιείται για την τελική απόφαση.

Η χρήση πολύ λίγων ή πολύ εύκολων impostors μπορεί να οδηγήσει σε αυξημένα false positives, καθώς το πραγματικό κείμενο μπορεί να τους ξεπερνά εύκολα ακόμη και όταν τα δύο εξεταζόμενα κείμενα προέρχονται από διαφορετικούς συγγραφείς. Από την άλλη πλευρά, ένας υπερβολικά μεγάλος αριθμός ιδιαίτερα ισχυρών impostors μπορεί να αυξήσει τα false negatives. Συνεπώς, απαιτείται ισορροπία ανάμεσα στην ποσότητα και την ποιότητα του background set.

Από την άλλη, η μέθοδος απαιτεί μεγάλο αριθμό υπολογισμών ομοιότητας, καθώς κάθε ζεύγος εξετάζεται επανειλημμένα με διαφορετικά υποσύνολα χαρακτηριστικών και έναν αριθμό impostors. Το υπολογιστικό κόστος αυξάνεται επομένως με τον αριθμό των επαναλήψεων, των impostors και των εξεταζόμενων ζευγών.

Τέλος, οι Koppel και Winter παρατηρούν ότι το πρόβλημα γίνεται δυσκολότερο όταν τα δύο κείμενα του ίδιου συγγραφέα διαφέρουν σημαντικά ως προς το θέμα ή το είδος (genre). Σε τέτοιες περιπτώσεις, τα κοινά χαρακτηριστικά που μπορούν να αξιοποιηθούν για την αναγνώριση του συγγραφικού ύφους περιορίζονται.

4.4 Παραλλαγή της μεθόδου General Impostors των Potha & Stamatatos [26]

Οι Potha & Stamatatos [26] προτείνουν μια τροποποιημένη εκδοχή της μεθόδου των Impostors που βελτιώνει την απόδοσή της. Αντί για την τυχαία επιλογή των κειμένων-impostors, προτείνουν τη χρήση των κειμένων που παρουσιάζουν την υψηλότερη ομοιότητα min-max με τα υπό εξέταση κείμενα. Χρησιμοποιώντας μια παρομοίωση, σε μια αναγνώριση υποπτών από την αστυνομία δεν έχει νόημα να επιλέγονται οι ύποπτοι από τον γενικό πληθυσμό. Αντιθέτως, όλοι οι ύποπτοι θα πρέπει να έχουν παρόμοια χαρακτηριστικά. Ένα άλλο αδύνατο σημείο της αρχικής μεθόδου είναι ότι αγνοεί τις περιπτώσεις όπου τουλάχιστον ένα κείμενο-Impostor βρέθηκε πιο παρόμοιο με το αμφισβητούμενο κείμενο σε σύγκριση με τα κείμενα του υποψήφιου συγγραφέα. Για να το αντισταθμίσει αυτό, προτείνουν τη κατάταξη των ομοιοτήτων σε φθίνουσα σειρά και τη συνεκτίμηση της θέσης του κειμένου του υποψήφιου συγγραφέα. Η προτεινόμενη προσέγγιση αξιολογείται σε διάφορα σύνολα δεδομένων (corpora) του PAN και επιτυγχάνει πολύ ανταγωνιστικά αποτελέσματα.

Η τροποποιημένη έκδοση της GI επιχειρεί να βελτιώσει τα ακόλουθα σημεία:

1. Επιλογή impostors (Impostor selection):
Αντί να επιλέγονται τυχαία τα impostor documents για κάθε πρόβλημα επαλήθευσης, προτείνεται να επιλέγονται τα εξωτερικά έγγραφα που παρουσιάζουν την υψηλότερη min-max similarity σε σχέση με τα known documents.

Με αυτόν τον τρόπο αυξάνεται η πιθανότητα να επιλεγούν δύσκολοι ή ανταγωνιστικοί impostors (challenging impostors), οι οποίοι παρουσιάζουν θεματική ή υφολογική ομοιότητα με τα known documents.

2. Πληροφορία κατάταξης (Ranking information):
Κατά τη διαδικασία βαθμολόγησης, τα επιλεγμένα κείμενα-impostors συγκρίνονται με το άγνωστο κείμενο (unknown document), λειτουργώντας ως ανταγωνιστές του κειμένου του υποψήφιου συγγραφέα (known document). Στόχος είναι να προσδιοριστεί κατά πόσο το known document παρουσιάζει μεγαλύτερη ομοιότητα με το unknown document σε σχέση με τους impostors.

Σε αντίθεση με την αρχική μέθοδο General Impostors, η οποία λαμβάνει υπόψη μόνο τις περιπτώσεις όπου το known document παρουσιάζει μεγαλύτερη ομοιότητα από όλους τους impostors, η προτεινόμενη παραλλαγή αξιοποιεί τη θέση κατάταξης του known document. Συγκεκριμένα, οι τιμές ομοιότητας ταξινομούνται σε φθίνουσα σειρά και η βαθμολογία υπολογίζεται με βάση τη θέση που καταλαμβάνει το known document στην κατάταξη.

Για παράδειγμα, έστω ότι ένα known document παρουσιάζει μεγαλύτερη ομοιότητα με το unknown document από όλους τους impostors εκτός από έναν, σε όλες τις επαναλήψεις - χωρίς να είναι απαραίτητα ο ίδιος impostor σε κάθε επανάληψη. Στην περίπτωση αυτή, η αρχική μέθοδος GI θα επέστρεφε βαθμολογία 0, ενώ η προτεινόμενη μέθοδος θα παρείχε βαθμολογία 0,5.

Η μέθοδος General Impostors (GI) και η προτεινόμενη παραλλαγή της διαθέτουν αρκετές παραμέτρους που πρέπει να καθοριστούν.

Για την απλοποίηση αυτής της διαδικασίας, οι ερευνητές επικεντρώθηκαν στη ρύθμιση της παραμέτρου a = |Impostors_problem|, δηλαδή του αριθμού των impostors που χρησιμοποιούνται για κάθε πρόβλημα επαλήθευσης. Στη συνέχεια, ο αριθμός των επαναλήψεων ορίζεται ως repetitions = a/5, ενώ ο αριθμός των impostors που χρησιμοποιούνται σε κάθε επανάληψη ορίζεται ως |Impostors_repetition| = a/10.

Επιπλέον, χρησιμοποιούνται character 5-grams ως χαρακτηριστικά, σταθερός ρυθμός (fix rate) ίσος με 0,5 και η συνάρτηση ομοιότητας min-max similarity.

Η συνάρτηση συνάθροισης (aggregate function) επιλέγεται μεταξύ των minimum (min), maximum (max) και average (μέσος όρος) ξεχωριστά για κάθε training corpus. Στις περισσότερες περιπτώσεις επιλέγεται ο μέσος όρος (average).

Επειδή η GI είναι ένας στοχαστικός αλγόριθμος (stochastic algorithm), κάθε πείραμα επαναλαμβάνεται πέντε φορές και αναφέρεται η μέση τιμή της μετρικής Area Under the ROC Curve (AUROC).

Στην Εικόνα 21 παρουσιάζεται ο ψευδοκώδικας της προτεινόμενης παραλλαγής, όπου επισημαίνονται οι βασικές τροποποιήσεις σε σχέση με την αρχική μέθοδο General Impostors.

```
Data: D_known, d_unknown, D_external
Parameters: repetitions,|Impostors_problem|,|Impostors_repetition|,rate
Result: FinalScore
for each d_known ∈ D_known do
    for each impostor ∈ D_external do
        MinMax(impostor) = minmaxSimilarity(impostor, d_known);
    end
    Select Impostors_problem ⊂ D_external with highest MinMax(:);
    /* Select Impostors_problem ⊂ D_external randomly                    */
    Set Score(d_known) = 0;
    repeat repetitions times
        Select Impostors_repetition ⊂ Impostors_problem randomly;
        Select rate% of features randomly;
        for each impostor ∈ Impostors_repetition do
            Sim(impostor) = similarity(impostor, d_unknown);
            /* Sim(impostor) =
               similarity(impostor, d_known) * similarity(impostor, d_unknown)  */
        end
        Sim_known = similarity(d_known, d_unknown);
        /* Sim_known = similarity(d_known, d_unknown)^2                     */
        Rank S = Sim(:) ∪ Sim_known in decreasing order;
        pos = position of Sim_known in S;
        Score(d_known) = Score(d_known) + 1/(repetitions * pos);
        /* if Sim_known > max(Sim(:)) then
            Score(d_known) = Score(d_known) + 1/repetitions;
        end
                                                                          */
    end;
    FinalScore = aggregate(Score(:));
end
```

Εικόνα 21: Ψευδοκώδικας της παραλλαγής μεθόδου General Impostors των Potha και Stamatatos. Οι τροποποιήσεις σε σχέση με την αρχική μέθοδο επισημαίνονται με μπλε χρώμα, ενώ οι αντίστοιχες εντολές της αρχικής μεθόδου εμφανίζονται ως σχόλια. Πηγή: [26]

Συνολικά, η παραλλαγή των Potha & Stamatatos επιχειρεί να βελτιώσει την αποτελεσματικότητα της αρχικής μεθόδου General Impostors μέσω της επιλογής περισσότερο ανταγωνιστικών impostors και της αξιοποίησης της πληροφορίας κατάταξης. Οι τροποποιήσεις αυτές επιτρέπουν μια πιο λεπτομερή αξιολόγηση της ομοιότητας μεταξύ των εξεταζόμενων κειμένων, καθώς η βαθμολογία δεν εξαρτάται αποκλειστικά από την πρώτη θέση στην κατάταξη.

4.5 Διαγωνισμοί Επαλήθευσης Συγγραφέα PAN

Οι διαγωνισμοί του PAN έχουν συμβάλει στην αύξηση του ερευνητικού ενδιαφέροντος και στη δημιουργία καινοτόμων μεθόδων πάνω στην επαλήθευση συγγραφέα. Έχουν δημιουργηθεί πολλές ερευνητικές ομάδες από όλον τον κόσμο και η καθεμία προσπαθεί να λύσει το πρόβλημα της επαλήθευσης συγγραφέα με βάση την δική της προσέγγιση. Στο τέλος, αξιολογείται η κάθε προσέγγιση με βάση κάποιους συγκεκριμένους δείκτες (μετρητές) απόδοσης.

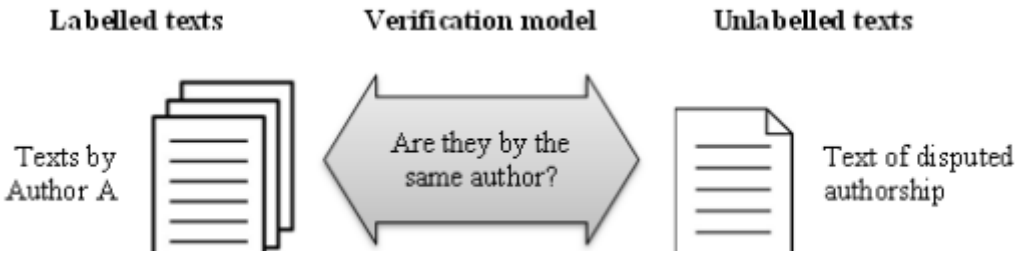
Τα σύνολα δεδομένων και το πλαίσιο αξιολόγησης του PAN παραμένουν μέχρι σήμερα βασικά εργαλεία για τη μελέτη και την ανάπτυξη σύγχρονων συστημάτων επαλήθευσης συγγραφέα, καθώς και για την

προώθηση της διεπιστημονικής συνεργασίας ανάμεσα σε ειδικούς της υπολογιστικής γλωσσολογίας, της εγκληματολογίας και της τεχνητής νοημοσύνης.

4.5.1 PAN 2015

Οι διοργανωτές του PAN [28] έφτιαξαν νέα σύνολα δεδομένων, καλύπτοντας πολλές γλώσσες (Αγγλικά, Ολλανδικά, Ελληνικά και Ισπανικά) και διάφορα είδη κειμένου (δοκίμια, μυθιστορήματα, κριτικές, άρθρα εφημερίδων). Παρείχαν ένα διαδικτυακό πειραματικό πλαίσιο, όπου οι ερευνητές μπορούσαν να υποβάλλουν τις μεθοδολογίες τους και να αξιολογούν τα αποτελέσματα. Το πλαίσιο αυτό επέτρεψε τη συγκριτική ανάλυση διαφορετικών τεχνικών επαλήθευσης.

Η διαδικασία που χρησιμοποιήθηκε στο PAN, φαίνεται στην εικόνα 22:



Εικόνα 22: Κάποια κείμενα είναι επισημασμένα ως προς τον συγγραφέα και υπάρχει ένα κείμενο αγνώστου συγγραφέα. Πηγή: [28]

Κάθε σώμα κειμένου περιλαμβάνει ένα σύνολο από προβλήματα επαλήθευσης και σε κάθε πρόβλημα δίνεται ένα σύνολο επισημασμένων (γνωστών) κειμένων (όλα από τον ίδιο συγγραφέα) και ακριβώς ένα μη επισημασμένο κείμενο (άγνωστο). Οι συμμετέχοντες του διαγωνισμού θα πρέπει να δώσουν μια δυαδική απάντηση, δηλαδή είτε είναι αυτός ο συγγραφέας του αγνώστου κειμένου, είτε όχι, καθώς και ένα σκορ με τιμή [0,1]. Μηδέν όταν το άγνωστο κείμενο δεν είναι γραμμένο από τον συγκεκριμένο συγγραφέα και 1 όταν είναι. Εάν το πρόβλημα είναι πολύ δύσκολο, είναι πιθανό η τελική τιμή να είναι 0.5, δηλαδή το τελικό ερώτημα μένει αναπάντητο.

Η αξιολόγηση των αποτελεσμάτων γίνεται με δύο μετρητές: το AUROC (Area Under the ROC curve), που εκτιμά την διαχωριστική ικανότητα του μοντέλου και το c@1. Το τελευταίο είναι μια τροποποιημένη μορφή της ακρίβειας (accuracy score), όπου συμπεριλαμβάνονται και τα προβλήματα, που μένουν αναπάντητα. Η τελική κατάταξη των συμμετεχόντων προκύπτει, από το γινόμενο των δυο δεικτών (AUROC * c@1).

	Corpus	Training problems	Evaluation problems	Mean labelled texts / problem	Mean text length (words)
PAN 2013	English (Textbooks)	10	30	3.98	1058
	Greek (Articles)	20	30	5.16	1823
	Spanish (Editorials+Fiction)	5	25	3.07	849
PAN 2014	Dutch (Essays)	96	96	1.89	405
	Dutch (Reviews)	100	100	1.02	114
	English (Essays)	200	200	2.62	841
	English (Novels)	100	200	1.00	5115
	Greek (Articles)	100	100	2.77	1470
	Spanish (Articles)	100	100	5.00	1129
PAN 2015	Dutch (Cross-genre)	100	165	1.75	357
	English (Cross-topic)	100	500	1.00	508
	Greek (Cross-topic)	100	100	2.87	717
	Spanish (Mixed)	100	100	4.00	950

Πίνακας 1: Σύνοψη των σωμάτων του PAN για επαλήθευση συγγραφέα. [28]

Τα δεδομένα εκπαίδευσης δόθηκαν στους συμμετέχοντες, ώστε να αναπτύξουν και να βελτιστοποιήσουν τις μεθόδους τους, ενώ το σετ αξιολόγησης δημοσιεύτηκε μετά την τελική προθεσμία των υποβολών. Αξίζει να σημειωθεί ότι κάθε σώμα κειμένου, είτε εκπαίδευσης είτε αξιολόγησης, είναι ισορροπημένο σε σχέση με την κατανομή των θετικών και των αρνητικών προβλημάτων επαλήθευσης. Με άλλα λόγια, η εκ των προτέρων πιθανότητα μιας θετικής ή αρνητικής απάντησης είναι 0.5.

Στις εργασίες του PAN 2013 και 2014, όλα τα κείμενα ήταν γραμμένα στην ίδια γλώσσα, άνηκαν στο ίδιο είδος κειμένου και υπήρχαν θεματικές ομοιότητες. Αυτό σημαίνει ότι το αμφισβητούμενο κείμενο και τα γνωστά κείμενα είχαν κάποια κοινά χαρακτηριστικά, ως προς το περιεχόμενο τους και το ύφος τους. Αντίθετα, το 2015 η μόνο υπόθεση είναι ότι όλα τα κείμενα μέσα σε ένα πρόβλημα είναι γραμμένα στην ίδια γλώσσα. Το άγνωστο και τα γνωστά κείμενα μπορεί να ανήκουν σε διαφορετικά είδη και να έχουν διαφορετικό θεματικό περιεχόμενο. Αυτό το καθιστά ιδιαίτερα απαιτητικό πρόβλημα, αφού το είδος και το θέμα του κειμένου επηρεάζουν σημαντικά τα στιλομετρικά χαρακτηριστικά του συγγραφέα.

Από την άλλη, η υπόθεση, ότι όλα τα κείμενα πρέπει να έχουν θεματικές ομοιότητες και ανήκουν στο ίδιο είδος, δεν είναι ρεαλιστική, αφού σε πολλές γλωσσολογικές περιπτώσεις αυτό δεν είναι εφικτό.

Οι περισσότερες μέθοδοι επαλήθευσης συγγραφέα, που υποβλήθηκαν, βασίζονται σε χαμηλού επιπέδου χαρακτηριστικά (λέξεων και χαρακτήρων) και αποφεύγουν τη χρήση συντακτικών, σημασιολογικών ή εξειδικευμένων χαρακτηριστικών. Ένας λόγος για αυτό είναι, ότι τα χαρακτηριστικά, όπως, τα ν-γράμματα χαρακτήρων, ή οι συχνότητες πολλών κοινών λέξεων, μπορούν να εφαρμοστούν σε κάθε φυσική γλώσσα με ελάχιστα απαιτούμενα για προ-επεξεργασία κειμένου. Κάποιοι συμμετέχοντες χρησιμοποιήσαν

επισημάνσεις μερών του λόγου (POS tagging) και πλήρη συντακτική ανάλυση, τα οποία, όμως, δεν μπορούσαν να εφαρμοστούν πάντα σε όλες τις γλώσσες των κειμένων.

Υπάρχουν κάποιες προσεγγίσεις που προσπαθούν να συνδυάσουν τα profile-based και instance-based μοντέλα, αναλύοντας πρώτα κάθε γνωστό κείμενο ξεχωριστά και στη συνέχεια συνδυάζοντας τις αναπαραστάσεις που εξάγονται. Όμως, τέτοια υβριδικά μοντέλα, συνήθως αποτυγχάνουν να συνδυάσουν τα πλεονεκτήματα των δυο παραδειγμάτων.

Ο πίνακας 1 δείχνει την κατανομή των συμμετεχόντων, ανάλογα με τους τύπους των μοντέλων επαλήθευσης που παρουσιάστηκαν παραπάνω. Είναι φανερό ότι η πλειονότητα των συμμετεχόντων, ακολουθούν μια εγγενή (intrinsic), τεμπέλικη (lazy) και βασισμένη σε παραδείγματα (instance-based) μεθοδολογία. Τα μοντέλα πρόθυμης μάθησης ξεκίνησαν να γίνονται δημοφιλή στις τελευταίες διοργανώσεις του PAN 2015, όταν το μέγεθος του σώματος εκπαίδευσης επέτρεψε την ανάπτυξη πιο αποτελεσματικών μοντέλων. Επιπλέον, τα extrinsic μοντέλα απέκτησαν αυξανόμενη δημοτικότητα τα τελευταία χρόνια, λόγω των εξαιρετικών αποτελεσμάτων, που επιτεύχθηκαν από τους Seidman, Khonji & Iraqi και Bagna.

Verification model	PAN 2013	PAN 2014	PAN 2015
Intrinsic	13	10	11
Extrinsic	3	3	7
Eager	2	3	10
Lazy	14	10	8
Profile-based	4	1	4
Instance-based	11	12	12
Hybrid	1	0	2

Πίνακας 2: Κατανομή των συμμετεχόντων στο PAN στις κατηγορίες μοντέλων επαλήθευσης. Πηγή: [28]

Σύμφωνα με τον πίνακα 3, είναι σαφές, η απόδοση των μοντέλων επαλήθευσης συγγραφέα παρουσιάζει διαφοροποιήσεις μεταξύ των επιμέρους σωμάτων κειμένων. Ωστόσο, οι διαφοροποιήσεις αυτές δεν μπορούν να αποδοθούν αποκλειστικά στη γλώσσα ή στο είδος του κειμένου. Για παράδειγμα, το σκορ των Ολλανδικών δοκιμίων (essays) είναι πολύ υψηλό, αντίθετα σε ένα άλλο σώμα κειμένων με ίδια γλώσσα (όπως οι κριτικές γραμμένες στα Ολλανδικά) ή στο ίδιο είδος κειμένου σε άλλη γλώσσα (όπως τα αγγλικά δοκίμια), το σκορ είναι πιο χαμηλό. Άλλοι παράγοντες, που είναι ιδιαίτερα σημαντικοί, είναι ο αριθμός των επισημασμένων κειμένων ανά πρόβλημα επαλήθευσης, καθώς και το μήκος του κειμένου. Γενικά, όταν ο αριθμός των επισημασμένων κειμένων είναι μικρός (1 ή 2), ή το μήκος του κειμένου είναι μικρό (μικρότερο από 500 λέξεις), τότε η απόδοση των μοντέλων μειώνεται.

Από την άλλη, δεν είναι πάντα δυνατό, να εξηγηθεί η υψηλή ή χαμηλή απόδοση των μοντέλων επαλήθευσης σε ένα συγκεκριμένο σώμα κειμένων αποκλειστικά με βάση τις ποσοτικές ιδιότητες του

σώματος. Υπάρχουν και ποιοτικές ιδιότητες, που είναι πιο χρήσιμες για την ερμηνεία των αποτελεσμάτων. Για παράδειγμα το σώμα κειμένων των αγγλικών μυθιστορημάτων, αποτελείται από αποσπάσματα μυθιστορημάτων ενός συγκεκριμένου υπό-είδους της λογοτεχνίας τρόμου, το οποίο χαρακτηρίζεται από ασυνήθιστο λεξιλόγιο και περίτεχνο (διακοσμημένο) λόγο (ύφος).Αυτό καθιστά τις ομοιότητες μεταξύ διαφορετικών συγγραφέων να φαίνονται πιο έντονες. Αυτό, όμως, έχει σαν αποτέλεσμα, η απόδοση να είναι σχετικά χαμηλή, παρά το μεγάλο μήκος των κειμένων.

		c@1		AUROC	
	Corpus	**Best**	**Average**	**Best**	**Average**
PAN 2013	English (Textbooks)	0.80	0.66	0.84	0.61
	Greek (Articles)	0.83	0.52	0.82	0.60
	Spanish (Editorials+Fiction)	0.84	0.59	0.93	0.67
PAN 2014	Dutch (Essays)	0.91	0.75	0.93	0.76
	Dutch (Reviews)	0.69	0.55	0.76	0.59
	English (Essays)	0.71	0.58	0.72	0.60
	English (Novels)	0.72	0.57	0.75	0.61
	Greek (Articles)	0.81	0.60	0.89	0.67
	Spanish (Articles)	0.78	0.68	0.90	0.71
PAN 2015	Dutch (Cross-genre)	0.77	0.55	0.83	0.60
	English (Cross-topic)	0.76	0.56	0.81	0.62
	Greek (Cross-topic)	0.85	0.54	0.89	0.67
	Spanish (Mixed)	0.83	0.59	0.93	0.66
	Average	0.79	0.60	0.85	0.64

Πίνακας 3: Παραπάνω φαίνονται τα καλύτερα αποτελέσματα που επιτεύχθηκαν από τους συμμετέχοντες για κάθε κείμενο και ο μέσος όρος της. απόδοσης για όλους τους συμμετέχοντες του PAN. Πηγή: [28]

4.5.2 PAN 2020

Η εργασία περιλαμβάνει fanfiction σε Αγγλική γλώσσα, γραμμένα από μη επαγγελματίες συγγραφείς [29]. Περιλαμβάνει ένα closed-set σύνολο. Δηλαδή, το test set περιέχει κάποιους συγγραφείς και θέματα, τα οποία υπάρχουν και στο training set.

Μια διαφοροποίηση, σε σχέση με τις προηγούμενες χρονιές είναι η αύξηση των δεδομένων. Αυτό έγινε, επειδή παρατηρήθηκε τις προηγούμενες χρονιές ότι δεν υπήρχαν πολλές μέθοδοι που να χρησιμοποιούν βαθιά νευρωνικά δίκτυα. Με την παροχή ενός μεγάλου συνόλου δεδομένων αξιολόγησης, στόχος ήταν να διευρυνθεί η ποικιλία των μεθόδων των συμμετεχόντων.

Εξαιτίας της ποικιλίας των συγγραφικών στυλ, συμπεριλαμβανομένου των διαχρονικών μεταβολών τους, η πρόοδος στην αυθεντικοποίηση του κειμένου με βάση το στυλ, είναι δύσκολο να παρακολουθηθεί, καθώς απαιτεί εκτεταμένη, διαφανή και επαναλαμβανόμενη διαδικασία αξιολόγησης.

Ο στόχος της επαλήθευσης συγγραφέα είναι να υπολογιστεί η συνάρτηση φ: $(D_k, d_u) \rightarrow \{T, F\}$, όπου το D_k (known) είναι ένα σύνολο κειμένων γραμμένα από γνωστό συγγραφέα και d_u (unknown) είναι το άγνωστο κείμενο. Εάν $\varphi(D_k, d_u) = T$, τότε ο συγγραφέας των D_k κειμένων είναι ο ίδιος με το d_u κείμενο. Εάν, όμως $(D_k, d_u) = F$, τότε ο συγγραφέας των D_k κειμένων δεν είναι ο ίδιος με το d_u κείμενο. Στην περίπτωση της cross-domain επαλήθευσης, τα D_k και d_u προέρχονται από διαφορετικά είδη κειμένων ή έχουν διαφορετικό περιεχόμενο (π.χ. άλλο θέμα, είδος κλπ.). Σε αυτήν την εργασία, οι διοργανωτές εξετάζουν την απλοϊκή (και πιο δύσκολη) μορφή του προβλήματος. Θεωρούν ότι το D_k είναι μόνο ένα γνωστό κείμενο, έτσι μόνο ζεύγη κειμένων εξετάζονται. Δηλαδή, το σύστημα πρέπει να αποφασίσει εάν το ένα γνωστό κείμενο, προέρχεται από τον ίδιο συγγραφέα με το άγνωστο κείμενο.

Έχουμε, δηλαδή, δυο βασικές κατηγορίες ζευγών:

- SA(Same Author): $\varphi(D_k, d_u) = T$
- DA(Different Author): $\varphi(D_k, d_u) = F$

Σύνολα Δεδομένων

Η παρούσα έκδοση ασχολείται με το υλικό τύπου «fanfiction», δηλαδή λογοτεχνικά κείμενα που δημιουργούνται από μη επαγγελματίες συγγραφείς για ένα συγκεκριμένο πεδίο («fandom»), όπως

είναι για παράδειγμα το Harry Potter, ή το Lord of the rings. Είναι μια κατηγορία κειμένων, που συνοδεύεται από πλούσια μεταδεδομένα και διατίθεται σε πολλές γλώσσες. Το fanfiction αποτελεί εξαιρετική πηγή υλικού για την μελέτη cross-domain σεναρίων, καθώς οι χρήστες, συχνά, δημοσιεύουν ιστορίες (fics), που καλύπτουν πολλά θεματικά πεδία, όπως το Twilight ή κόμικς της Marvel. Τα datasets που παρέχονται για τις εργασίες PAN 2020 και PAN 2021, αντλήθηκαν μέσω web crawling από την κοινότητα fanfiction.net.

Η έννοια της cross-domain verification (επαλήθευση μεταξύ διαφορετικών πεδίων) πραγματοποιείται εδώ ως cross-fandom verification, όπου ένα fandom μπορεί να ερμηνευτεί περίπου ως μείγμα θέματος (topic) και λογοτεχνικού είδους (genre).

Στα δεδομένα περιλαμβάνονται μόνο κείμενα στην αγγλική γλώσσα, για να μεγιστοποιηθεί το μέγεθος του dataset.

Υπάρχουν δύο σύνολα δεδομένων εκπαίδευσης, τα μεγάλα (large) και τα μικρά (small). Αυτά τα σύνολα δεδομένων επιτρέπουν την σύγκριση αποτελεσμάτων μεταξύ των συστημάτων που εκπαιδεύονται με το μεγάλο σύνολο ή το μικρό σύνολο αντίστοιχα.

Για να δημιουργηθεί το dataset, τα κείμενα ομαδοποιήθηκαν ανά συγγραφέα και θέμα (fandom), για να διασφαλιστεί ένα καλό μείγμα και των δυο. Για το μεγάλο dataset, επιλέχθηκαν 148.000 ζεύγη ίδιου

συγγραφέα (SA) και 128.000 ζεύγη διαφορετικών συγγραφέων (DA), από τη συλλογή του fanfiction.net. Το μικρό σετ εκπαίδευσης αποτελεί υποσύνολο του μεγάλου σετ, με 28.000 SA και 25.000 DA ζεύγη από τα ίδια 1.600 fandoms. Το test dataset περιλαμβάνει 10.000 SA και 6.900 DA ζεύγη από 400 fandoms.

Λόγω του μεγάλου μεγέθους του dataset, χρησιμοποιήθηκαν 4 μετρητές αξιολόγησης. Για κάθε πρόβλημα στο test set έπρεπε να δίνουν απάντηση σε ένα εύρος τιμής [0,1], δείχνοντας την πιθανότητα να ανήκουν (τα ζεύγη κειμένων) στον ίδιο συγγραφέα (SA). Για δύσκολες περιπτώσεις το σύστημα μπορούσε να δώσει σκορ 0.5, το οποίο σήμαινε ότι άφηνε το πρόβλημα χωρίς απάντηση, δηλαδή δεν ήταν σίγουρο αν ανήκε στον ίδιο συγγραφέα ή όχι.

Οι μετρητές που χρησιμοποιήθηκαν είναι οι ακόλουθοι:

- AUC (Area Under Curve): Η περιοχή κάτω από την καμπύλη ROC (Receiver Operating Characteristic). Μετρά τη συνολική διακριτική ικανότητα του συστήματος μεταξύ ζευγών SA και DA.
- c@1: Παραλλαγή την ακρίβειας (accuracy), που επιβραβεύει τα αναπάντητα ερωτήματα.
- F1-score: Ο γνωστός δείκτης F1, που έχει να κάνει με το precision και το recall. Δεν λαμβάνει υπόψη τα αναπάντητα ερωτήματα.
- $F_{0.5u}$: Ένας νέος μετρικός δείκτης βασισμένος στο $F_{0.5}$, που δίνει έμφαση στα σωστά ζεύγη ίδιου συγγραφέα (SA) και επιβραβεύει τα αναπάντητα ερωτήματα.

Για την παραγωγή της τελικής κατάταξης ενός συστήματος, υπολογίστηκε ο μέσος όρος, όλων των παραπάνω μετρητών. Έτσι, η συνολική αξιολόγηση δεν εξαρτάται μόνο από ένα μετρικό δείκτη, αλλά από πολυδιάστατη εκτίμηση της απόδοσης του κάθε συστήματος.

Εφαρμόστηκαν δυο βασικά συστήματα αναφοράς (βαθμονομημένα στο μικρό training set).

1) Η πρώτη μέθοδος υπολογίζει τις ομοιότητες συνημιτόνου (cosine similarities) μεταξύ κανονικοποιημένων TF-IDF χαρακτηριστικών τετραγραμμάτων των κειμένων σε κάθε ζεύγος. Τα τελικά σκορ τροποποιούνται, χρησιμοποιώντας grid search στα δεδομένα βαθμονόμησης (naive baseline). Αυτός ονομάζεται, first-order verifier, δηλαδή δεν μαθαίνει αναπαραστάσεις υψηλότερου επιπέδου.
2) Έπειτα, εφαρμόζουμε μια μέθοδο συμπίεσης κειμένου. Δίνοντας ένα ζεύγος κειμένων (τ1 και τ2) υπολογίζεται η cross-entropy του τ2 με βάση την πρόβλεψη του τ1, μέσω partial matching model και αντίστροφα. Η μέση τιμή και η απόλυτη διαφορά των δυο cross-entropies χρησιμοποιούνται από ένα μοντέλο λογιστικής παλινδρόμησης, για να δώσει ένα σκορ στο εύρος [0,1] (compression baseline).

Υποβλήθηκαν 13 συστήματα από 10 ομάδες.

Submission	AUC	c@1	F0.5u	F1	Overall
boenninghoff20-large	**0.969**	**0.928**	**0.907**	**0.936**	**0.935**
weerasinghe20-large	0.953	0.880	0.882	0.891	0.902
boenninghoff20-small	0.940	0.889	0.853	0.906	0.897
weerasinghe20-small	0.939	0.833	0.817	0.860	0.862
halvani20-small	0.878	0.796	0.819	0.807	0.825
kipnis20-small	0.866	0.801	0.815	0.809	0.823
araujo20-small	0.874	0.770	0.762	0.811	0.804
niven20-small	0.795	0.786	0.842	0.778	0.800
gagala20-small	0.786	0.786	0.809	0.800	0.796
araujo20-large	0.859	0.751	0.745	0.800	0.789
baseline (naive)	0.780	0.723	0.716	0.767	0.747
baseline (compression)	0.778	0.719	0.703	0.770	0.742
ordonez20-large	0.696	0.640	0.655	0.748	0.685
ikae20-small	0.840	0.544	0.704	0.598	0.672
faber20-small	0.293	0.331	0.314	0.262	0.300

Πίνακας 4: Αποτελέσματα αξιολόγησης. Πηγή: [29]

Μπορούν να εξαχθούν ορισμένες ιδιαίτερα ενδιαφέρουσες παρατηρήσεις. Πρώτον, η μέθοδος με την υψηλότερη επίδοση (boenninghoff20) πέτυχε εντυπωσιακό συνολικό σκορ 0.935, υπερέχοντας σημαντικά έναντι της δεύτερης καλύτερης προσέγγισης (weerasinghe20), η οποία ήδη παρουσίασε πολύ υψηλή συνολική επίδοση (0.902).

Η διαφορά μεταξύ των δύο αυτών συστημάτων ενδέχεται να οφείλεται στον τρόπο χειρισμού των μη-απαντήσεων (non-answers), καθώς η διαφορά στην απόδοση είναι πιο εμφανής στα μέτρα c@1 και F1, δίνοντας έτσι πλεονέκτημα στο σύστημα boenninghoff20.

Αξίζει επίσης να σημειωθεί ότι και τα δύο κορυφαία συστήματα (boenninghoff20 και weerasinghe20) εκπαιδεύτηκαν στο μεγάλο σύνολο δεδομένων, ξεπερνώντας σαφώς τις αντίστοιχες εκδόσεις που είχαν εκπαιδευτεί στο μικρότερο. Το εύρημα αυτό αποτελεί σαφή ένδειξη ότι τα μοντέλα επαλήθευσης συγγραφέα επωφελούνται από μεγαλύτερους όγκους δεδομένων εκπαίδευσης.

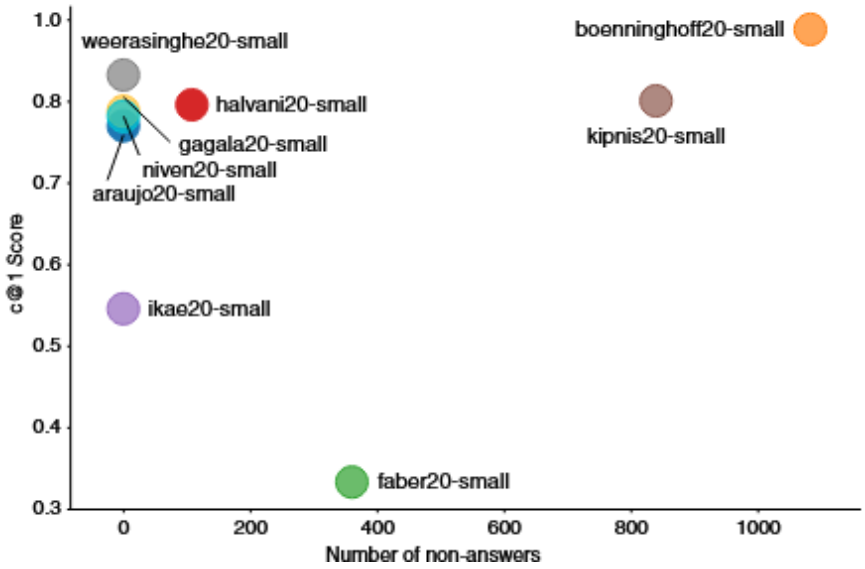
Ωστόσο, η παρατήρηση αυτή δεν επιβεβαιώθηκε στην περίπτωση του συστήματος araujo20, όπου το μοντέλο εκπαιδευμένο στο μικρό σύνολο δεδομένων υπερείχε ελαφρώς (κατά 0.015 μονάδες AUC) εκείνου που εκπαιδεύτηκε στο μεγάλο.

Η πλειονότητα των συστημάτων (με εξαίρεση τρεις υποβολές: ordonez20, ikae20 και faber20) ξεπέρασε τις επιδόσεις των βασικών μοντέλων αναφοράς (naive και compression baselines), οι οποίες δεν παρουσίασαν σημαντικές μεταξύ τους διαφορές.

Όπως αναφέρθηκε παραπάνω, οι συμμετέχοντες είχαν τη δυνατότητα να αφήσουν αναπάντητα τα δυσκολότερα προβλήματα, αποδίδοντάς τους μια τιμή σκορ ακριβώς ίση με 0.5. Οι μη-απαντήσεις αυτές επηρέαζαν ρητά τον μετρικό δείκτη c@1, ενώ εξαιρούνταν από τον υπολογισμό της F1.

Στην εικόνα 23 παρουσιάζεται η συσχέτιση μεταξύ του απόλυτου αριθμού μη-απαντήσεων που υπέβαλε κάθε ομάδα και του αντίστοιχου c@1 σκορ. Τα αποτελέσματα δείχνουν ότι μόνο τρεις ομάδες αξιοποίησαν

ενεργά αυτή τη δυνατότητα. Συγκεκριμένα, οι boenninghoff20 και kipnis υπέβαλαν σημαντικό αριθμό μη-απαντήσεων (1.082 και 839, αντίστοιχα), χωρίς να επηρεαστεί αρνητικά η συνολική τους επίδοση.



Εικόνα 23: Ο δείκτης c@1 για το μικρό σύνολο δεδομένων. Πηγή: [29]

Ένα ενδιαφέρον επόμενο ερώτημα που προκύπτει είναι κατά πόσο η πρωτιά των Boenninghoff et al. οφείλεται αποκλειστικά στη στρατηγική των μη-απαντήσεων.

Τα αποτελέσματα παρακάτω δείχνουν ότι αυτό δεν ισχύει: οι ερευνητές επαναϋπολόγισαν τα μέτρα αξιολόγησης, εξαιρώντας τα ζεύγη για τα οποία το σύστημα boenninghoff20-large είχε δηλώσει μη-απάντηση. Η προηγούμενη κατάταξη επιβεβαιώθηκε (και ενισχύθηκε ακόμη περισσότερο), γεγονός που αποδεικνύει ότι η ανωτερότητα του συστήματος δεν οφείλεται αποκλειστικά στη στρατηγική των μη-απαντήσεων.

Submission w/o non-responses	AUC	c@1	F1	F0.5u	Overall
boenninghoff20-large	0.974	0.930	0.936	0.934	0.943
weerasinghe20-large	0.957	0.886	0.897	0.888	0.907

Πίνακας 5: Τα δύο κορυφαία συστήματα, αποκλείοντας τα ζεύγη, τα οποία ο boenninghoff20-large υπέβαλε μη-απάντηση. [29]

4.5.3 PAN 2021

Είναι η δεύτερη έκδοση στο τριετές πρόγραμμα PAN (2020-2022), όπου η δυσκολία των εργασιών αυξάνεται. Το σύνολο δοκιμής (test set) έχει ζεύγος κειμένων, από άγνωστους συγγραφείς και ανήκουν σε πεδία, που δεν υπήρχαν σε κανένα προηγούμενο σύνολο δεδομένων.

Η έκδοση του 2021 [30] της επαλήθευσης συγγραφέα βασίστηκε στη δομή και στα δεδομένα εκπαίδευσης της προηγούμενης χρονιάς, αλλά χρησιμοποίησε διαφορετικό test set.

Διατηρήθηκε η cross-domain ρύθμιση, δηλαδή κάθε ζεύγος κειμένων προέρχεται από διαφορετικά πεδία, αλλά αντικαταστάθηκε η closed-set σε open-set, όπου τόσο οι συγγραφείς όσο και τα πεδία στο test set είναι εντελώς «καινούργια» και δεν υπάρχουν πουθενά στο σύνολο εκπαίδευσης. Αυτό θεωρητικά καθιστά την εργασία πιο απαιτητική συγκριτικά με τα προηγούμενα έτη, αφού τα μοντέλα δεν μπορούν να «μάθουν» εκ των προτέρων, τα στιλιστικά χαρακτηριστικά των συγκεκριμένων συγγραφέων ή θεματικών περιοχών.

Τα δεδομένα εκπαίδευσης είναι τα ίδια με της προηγούμενης χρονιάς και διατίθενται σε μικρή και μεγάλη έκδοση.

Για κάθε πρόβλημα από τα 19.999 (ζεύγη κειμένων) στο test set, τα συστήματα πρέπει να παράγουν μια τιμή a_i στο διάστημα [0,1], δείχνοντας την πιθανότητα ότι το ζεύγος κειμένων γράφτηκε από τον ίδιο συγγραφέα ($a_i > 0.5$), ή αποτελείται από διαφορετικούς συγγραφείς ($a_i < 0.5$). Τα συστήματα μπορούν να αφήνουν αναπάντητο το ερώτημα, αν έχουν σκορ $a_i = 0.5$, επειδή το θεωρούν πολύ δύσκολο.

Ξαναχρησιμοποιήθηκαν οι 4 μετρητές αξιολόγησης, που χρησιμοποιήθηκαν στο PAN 2020, αλλά προστέθηκε και το Brier score ως πέμπτο μετρητή.

Οι μετρητές που χρησιμοποιήθηκαν είναι οι ακόλουθοι:

- AUC (Area Under Curve)
- c@1
- F1-score
- $F_{0.5u}$.
- Brier: Μετρά την ακρίβεια των πιθανολογικών προβλέψεων. Είναι η αντιστροφή του Brier score loss, όπως υλοποιείται στη βιβλιοθήκη του scikit-learn. Επιβραβεύει το σύστημα που δίνει τολμηρές και σωστές προβλέψεις (π.χ. a_i κοντά στο 0.0 ή στο 1.0), αλλά «τιμωρεί» ουδέτερες προβλέψεις, όπως τα αναπάντητα ερωτήματα ($a_i = 0.5$).

Για την παραγωγή της τελικής κατάταξης ενός συστήματος, υπολογίστηκε ο μέσος όρος, όλων των παραπάνω μετρητών. Έτσι, η συνολική αξιολόγηση δεν εξαρτάται μόνο από ένα μετρικό δείκτη, αλλά από πολυδιάστατη εκτίμηση της απόδοσης του κάθε συστήματος.

Υπάρχουν τρεις βασικές μέθοδοι (εκπαιδευμένες στο small training set) για σύγκριση, όπου οι δυο πρώτες χρησιμοποιήθηκαν και στον περσινό διαγωνισμό. Η προσέγγιση compression-based, η naive distance-based και η first-order bag-of-words. Η τρίτη προστέθηκε μετέπειτα για τις ανάγκες της εργασίας. Βασίστηκε σε μια παραλλαγή για σύντομα κείμενα της μεθόδου “Unmasking” των Koppel και Schler, η οποία είχε παρουσιάσει καλά αποτελέσματα τα τελευταία χρόνια.

Στην εργασία της επαλήθευσης συγγραφέα υπήρχαν 13 υποβολές από 10 ομάδες.

Παρατηρήθηκε μια ποικιλία μεθόδων, με κάποιες καινούργιες προσεγγίσεις, για παράδειγμα μάθηση αναπαραστάσεων με νευρωνικά δίκτυα, να συνυπάρχουν με πιο καθιερωμένες μεθόδους από τους τομείς της ταξινόμησης κειμένου και της ανάκτησης πληροφοριών. Πολλές ομάδες χρησιμοποίησαν μια Siamese προσέγγιση νευρωνικού δικτύου, η οποία φαίνεται να αποτελεί καλή επιλογή για την ανάλυση ζευγών κειμένων.

Team	Dataset	AUC	c@1	F_1	$F_{0.5u}$	Brier	Overall
boenninghoff21	large	**0.9869**	**0.9502**	**0.9524**	**0.9378**	**0.9452**	**0.9545**
embarcaderoruiz21	large	0.9697	0.9306	0.9342	0.9147	0.9305	0.9359
weerasinghe21	large	0.9719	0.9172	0.9159	0.9245	0.9340	0.9327
weerasinghe21	small	0.9666	0.9103	0.9071	0.9270	0.9290	0.9280
menta21	large	0.9635	0.9024	0.8990	0.9186	0.9155	0.9198
peng21	small	0.9172	0.9172	0.9167	0.9200	0.9172	0.9177
embarcaderoruiz21	small	0.9470	0.8982	0.9040	0.8785	0.9072	0.9070
menta21	small	0.9385	0.8662	0.8620	0.8787	0.8762	0.8843
rabinovits21	small	0.8129	0.8129	0.8094	0.8186	0.8129	0.8133
ikae21	small	0.9041	0.7586	0.8145	0.7233	0.8247	0.8050
unmasking21	small	0.8298	0.7707	0.7803	0.7466	0.7904	0.7836
tyo21	large	0.8275	0.7594	0.7911	0.7257	0.8123	0.7832
naive21	small	0.7956	0.7320	0.7856	0.6998	0.7867	0.7600
compressor21	small	0.7896	0.7282	0.7609	0.7027	0.8094	0.7581
futrzynski21	large	0.7982	0.6632	0.8324	0.6682	0.7957	0.7516
liaozhihao21	small	0.4962	0.4962	0.0067	0.0161	0.4962	0.3023

Πίνακας 6: Αποτελέσματα αξιολόγησης ομάδων. Πηγή: [30]

Η τελική κατάταξη βασίζεται στη μέση απόδοση των 5 μετρητών αξιολόγησης (τελευταία στήλη του πίνακα). Η στήλη "Dataset" δείχνει εάν ένα σύστημα εκπαιδεύτηκε στο μικρό (small) ή μεγάλο (large) σύνολο δεδομένων.

Το καλύτερο σύστημα προήλθε από την ομάδα που είχε υποβάλει το καλύτερο σύστημα και στο PAN 2020. Η ομάδα boenninghof21 πέτυχε εξαιρετική σταθερή και ανθεκτική απόδοση, συμπεριλαμβάνοντας τα υψηλότερα σκορ σε κάθε μετρικό δείκτη. Στην δεύτερη θέση υπάρχουν, επίσης, πολύ καλές ομάδες (mbarcaderoruiz21, weerasinghe21, menta21 και peng21) που πέτυχαν παρόμοια σκορ. Εκτός από τα τρία συστήματα (tyo21, futrzynski21, liaozhihao21), οι περισσότερες προσεγγίσεις ξεπέρασαν τις τρεις (μη βελτιστοποιημένες) baselines μεθόδους. Από αυτές η "unmasking21" ήταν η καλύτερη με μικρή διαφορά από τις άλλες. Η μέθοδος tyo21 δεν είχε μεγάλη διαφορά με την unmasking, παρόλο που βασίζεται σε τελείως διαφορετική προσέγγιση.

Το σύστημα ikae21 πέτυχε πολύ ικανοποιητικό AUC σκορ 90%, αλλά $F_{0.5u}$ μόνο 70% (κάτι που οφείλεται στον διαφορετικό τρόπο με τον οποίο ο $F_{0.5u}$ αντιμετωπίζει τα ζεύγη κειμένων του ίδιου συγγραφέα). Γενικά, οι μη-απαντήσεις έπαιξαν σημαντικό ρόλο στην τελική κατάταξη, επηρεάζοντας τα σκορ των c@1 και $F_{0.5u}$. Συστήματα, όπως liaozhihao21-small, όπου έδιναν δυαδικές απαντήσεις, χωρίς τις μη-απαντήσεις, δεν ήταν καλά προς αυτούς τους δείκτες.

Μια άλλη παρατήρηση είναι ότι οι ομάδες που υπέβαλαν διαφορετικά συστήματα για τα μεγάλα σύνολα δεδομένων και τα μικρά σύνολα δεδομένων, τα αποτελέσματά τους ήταν διαφορετικά. Μάλιστα η «μεγάλη» έκδοση ξεπερνούσε στα πάντα, την «μικρή» έκδοση. Στην εργασία PAN 2020, η ανώτερη απόδοση των «μεγάλων» συστημάτων θα μπορούσε να αποδοθεί στο γεγονός ότι το πρόβλημα ήταν «κλειστού συνόλου», όπου ένα περίπλοκο μοντέλο μπορούσε ουσιαστικά να απομνημονεύσει τα χαρακτηριστικά του συγγραφέα. Αυτή, όμως, η εξήγηση δεν μπορεί να δοθεί στο PAN 2021, αφού κανένας από τους συγγραφείς ή τα πεδία δεν υπάρχουν στο test set.

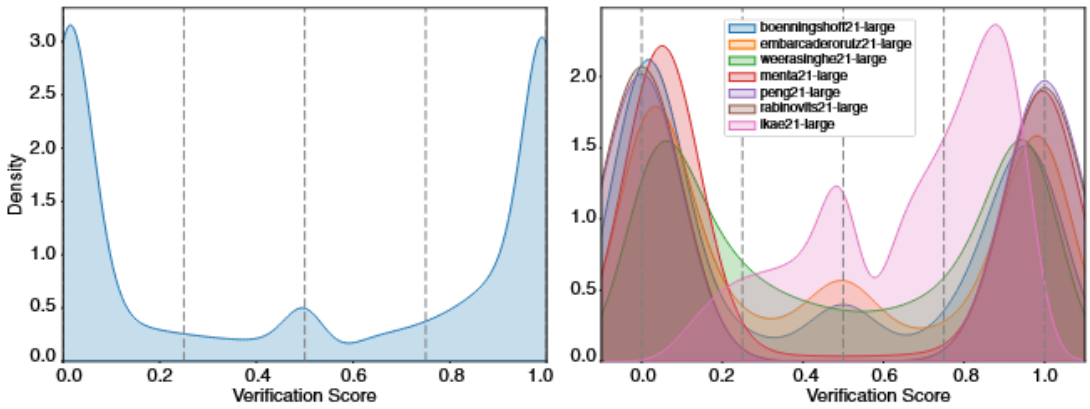
Επομένως, η βελτίωση προκύπτει από το μέγεθος ή την ποικιλία του συνόλου εκπαίδευσης. Αυτό αποτελεί, επιπλέον απόδειξη ότι τα συστήματα ωφελούνται από την αύξηση του μεγέθους του συνόλου εκπαίδευσης, ακόμη και σε σενάρια «ανοιχτού συνόλου» (open-set).

Team	2020 System		2021 System	
	2020 Data	2021 Data	2020 Data	2021 Data
niven	0.786	–	–	–
araujo	0.770	0.81	–	–
boenninghoff	0.928	–	0.917	0.950
weerasinghe	0.880	0.913	0.885	0.917
ordonez	0.640	–	–	–
faber	0.331	–	–	–
ikae	0.544	0.503	0.742	0.758
kipnis	0.801	0.815	–	–
gagala	0.786	0.804	–	–
halvani	0.796	0.822	–	–
embarcaderoruiz	–	–	0.914	0.930
menta	–	–	0.878	0.902
peng	–	–	–	0.917
rabinovits	–	–	0.795	0.812
tyo	–	–	–	0.759
futrzynski	–	–	0.662	0.663
liaozhihao	–	–	–	0.496

Πίνακας 7: Παραπάνω φαίνονται οι επιδόσεις των συστημάτων με βάση τον δείκτη c@1. Τα σκορ για τα σύνολα δεδομένων του 2020 είναι χαμηλότερα σε όλα (εκτός του ikae21) από ότι του 2021. Πηγή: [30]

Το open-set με τους άγνωστους συγγραφείς και νέα θεματικά πεδία, αποδείχτηκε στην πράξη, ότι ήταν πιο εύκολο να λυθεί από ότι το closed-set. Από την άλλη, τα συστήματα του 2021 τείνουν να υπολειτουργούν στο σύνολο δεδομένων του 2020, σε σύγκριση με τις αρχικές υποβολές του 2020 (τουλάχιστον στις σπάνιες περιπτώσεις που μπορούσε να συγκριθεί boeninghoff, weerasinghe, ikea).

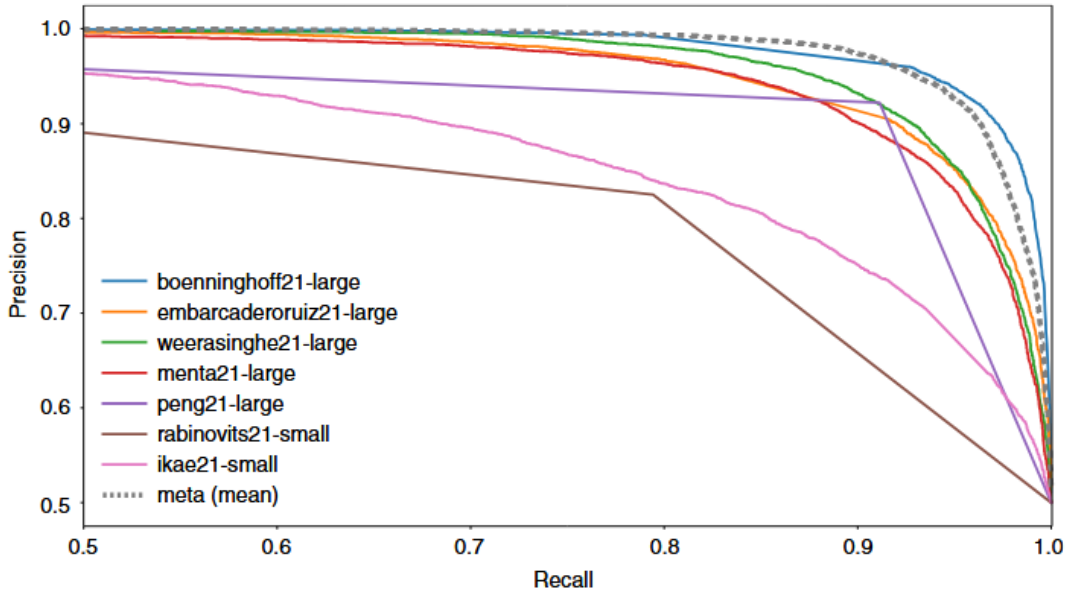
Κατανομές



Εικόνα 24: Αριστερά παρουσιάζεται η κατανομή των προβλέψεων για τα καλύτερα συστήματα της κάθε ομάδας. Δεξιά φαίνεται το ίδιο, αλλά με τις μεθόδους του καθενός. Πηγή: [30]

Η εικόνα 24 αριστερά παρουσιάζει την κατανομή των προβλέψεων για τα καλύτερα συστήματα της κάθε ομάδας. Βλέπουμε μια trimodal κατανομή με κορυφές γύρω από 0, 0.5 και 1. Η τιμή κοντά στο 0 δείχνει ότι τα δύο κείμενα ανήκουν σε διαφορετικούς συγγραφείς. Αντίθετα, η τιμή κοντά στο 1 δείχνει ότι τα δύο κείμενα ανήκουν στον ίδιο συγγραφέα. Η μεσαία κορυφή στο 0.5 δείχνει ότι το σύστημα προτίμησε να αφήσει το πρόβλημα αναπάντητο. Η ύπαρξη της τρίτης κορυφής στο 0.5 δείχνει ότι ορισμένα συστήματα προσαρμόστηκε εσκεμμένα για να αφήσουν αναπάντητο το ερώτημα, επειδή αυτό βοηθάει στην βελτίωση των δεικτών μέτρησης c@1 και $F_{0.5u}$, οι οποίοι επιβραβεύουν τα συστήματα που αποφεύγουν λανθασμένες απαντήσεις.

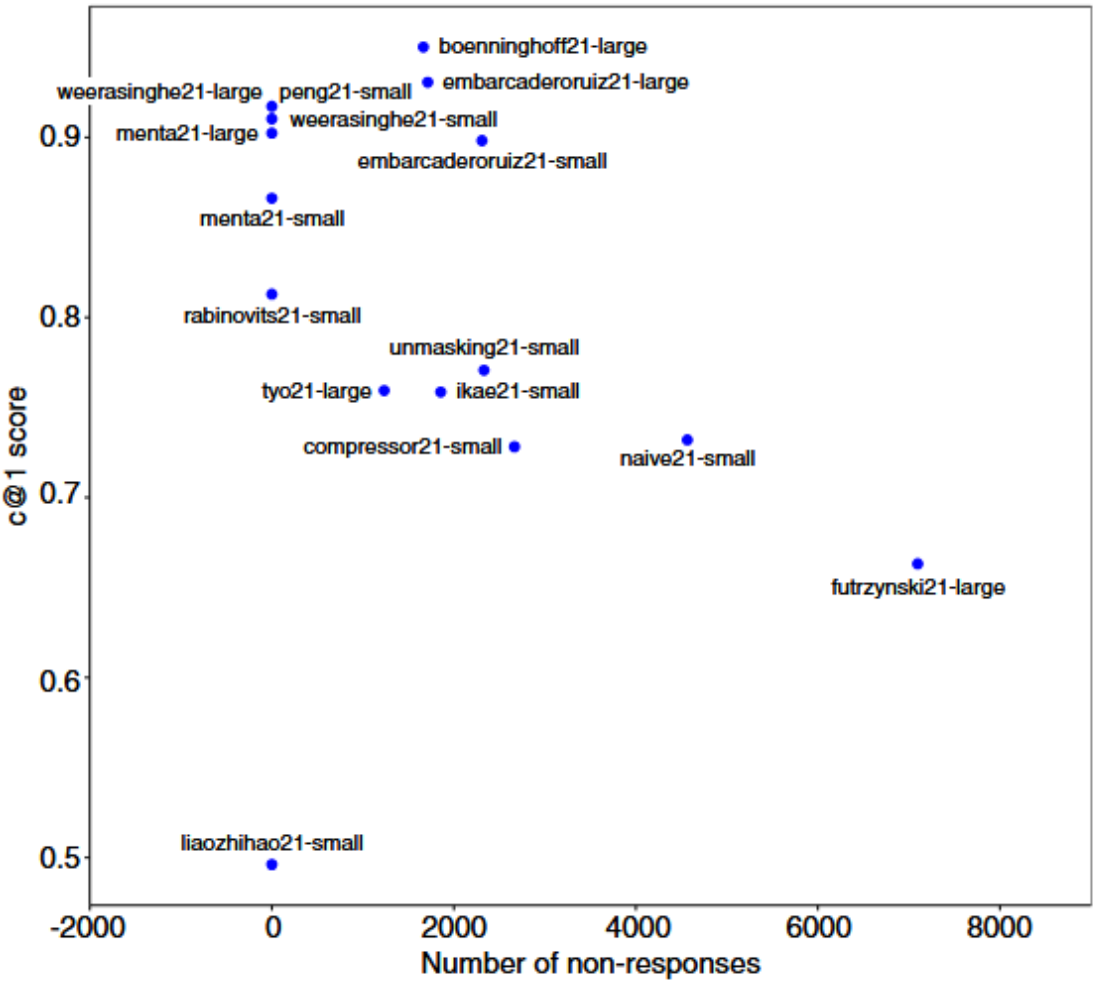
Η εικόνα δεξιά ενισχύει αυτή την υπόθεση, δείχνοντας ότι πράγματι ορισμένα συστήματα είχαν προκαθορισμένα “patterns” μη απαντήσεων, ενσωματωμένα στη στρατηγική τους.



Εικόνα 25: Παρουσιάζονται οι καμπύλες ακρίβειας (precision) και ανάκλησης (recall), των παραπάνω συστημάτων, καθώς και ένας naive meta classifier, που προβλέπει τον μέσο όρο των σκορ όλων των συστημάτων. Πηγή: [30]

Σε αντίθεση με τα προηγούμενα χρόνια, ο meta-classifier ξεπέρασε τα περισσότερα μεμονωμένα μοντέλα. Ωστόσο, ενώ ο meta classifier παρουσίασε παρόμοια ακρίβεια (precision) με το κορυφαίο σύστημα "boenninghoff21", υστέρησε σημαντικά ως προς την ανάκληση (recall), δηλαδή είχε μικρότερη κάλυψη των θετικών περιπτώσεων.

Μη-απαντήσεις



Εικόνα 26: Ο δείκτης c@1 για τις μη-απαντήσεις. Πηγή: [30]

Οι μη-απαντήσεις αποτέλεσαν αναπόσπαστο κομμάτι της διαδικασίας αξιολόγησης.

Στην εικόνα 26, απεικονίζεται η απόδοση του c@1, σαν συνάρτηση του απόλυτου αριθμού μη-απαντήσεων ανά σύστημα. Βλέπουμε ότι ο "futrzynski21" επέστρεψε συνολικά τις περισσότερες μη-απαντήσεις, αν και με κόστος απόδοσης κάτω από το βασικό επίπεδο (below baseline). Οι τρεις βασικές μέθοδοι έδωσαν, επίσης, μη-απαντήσεις σε πολλές περιπτώσεις, αλλά ξεπεράστηκαν από τα περισσότερα συστήματα των συμμετεχόντων.

Τα συστήματα υψηλότερης απόδοσης (boenninghoff21, embarcaderoruiz21) αρνήθηκαν να απαντήσουν σε περιπτώσεις σε πιο μέτριο βαθμό, με αποτέλεσμα να έχει πολλή καλή απόδοση. Πολλά άλλα καλά συστήματα (όπως weerasinghe21, menta21, ή peng21) φάνηκαν να μην δίνουν ιδιαίτερη προσοχή να βελτιστοποιήσουν αυτό το στοιχείο της εργασίας και υπέβαλαν πολύ λίγες απαντήσεις.

4.5.4 PAN 2022

Το PAN 2022 επικεντρώνεται σε ένα δύσκολο και διαφορετικό σενάριο [31]. Δίνοντας 2 κείμενα που ανήκουν σε διαφορετικούς τύπους κειμένων, ο στόχος είναι να προσδιοριστεί αν αυτά τα 2 κείμενα έχουν γραφτεί από τον ίδιο συγγραφέα. Δίνεται ένα νέο σώμα κειμένων στα αγγλικά και παρέχονται ζεύγη κειμένων από 4 διαφορετικούς τύπος: δοκίμια, emails, μηνύματα κειμένου και επαγγελματικά σημειώματα.

Η έκδοση του PAN, εστιάζει στη επαλήθευση συγγραφέα μεταξύ διαφορετικών τύπων κειμένου, όπου τα κείμενα γνωστής συγγραφής και άγνωστης συγγραφής ανήκουν σε διαφορετικούς τύπους.

Οι τύποι λόγου αυτοί διαφέρουν σημαντικά ως προς τον επικοινωνιακό σκοπό, το κοινό στο οποίο απευθύνονται, ή το επίπεδο τυπικότητας. Για παράδειγμα, τα δοκίμια επιχειρηματολογίας και τα μηνύματα κειμένου που στέλνονται σε μέλη της οικογένειας διαφέρουν σημαντικά ως προς το ύφος, όπως αυτό επιβάλλεται από τους κανόνες του κάθε τύπου λόγου.

Επομένως, είναι πολύ δύσκολο να απομονωθούν τα στιλιστικά χαρακτηριστικά του συγγραφέα, τα οποία παραμένουν σταθερά ανεξάρτητα από τον τύπο του κειμένου. Επιπλέον, οι διαφορετικοί τύποι κειμένων συνδέονται στενά με το μήκος του κειμένου, δηλαδή τα δοκίμια είναι πολύ μεγαλύτερα από ότι τα μηνύματα κειμένου που στέλνουμε στο Facebook. Η επαλήθευση συγγραφέα μεταξύ διαφορετικών τύπων κειμένου μπορεί να χρησιμοποιηθεί για να μελετηθεί ο ρόλος του μήκους του κειμένου στην αποτελεσματικότητα των μεθόδων επαλήθευσης.

Δημιουργήθηκε ένα σύνολο δεδομένων από ένα υποσύνολο του πρόσφατου “Aston 100 Idiolects Corpus" στην αγγλικά γλώσσα, το οποίο περιλαμβάνει ένα πλούσιο σύνολο τύπων γραπτού λόγου, που έχουν γραφτεί από 112 άτομα. Χρησιμοποιήθηκαν οι ακόλουθοι τύποι γραπτού λόγου: emails, δοκίμια (essays), μηνύματα κειμένου (text messages) και επαγγελματικά σημειώματα (business memos).

Subset	Training	Test
Author match	*Text pairs*	
Positive (same author)	6,132 (50.0%)	5,239 (50.0%)
Negative (different author)	6,132 (50.0%)	5,239 (50.0%)
Discourse type pairings	*Text pairs*	
Email–Text message	7,484 (61.0%)	6,092 (58.1%)
Essay–Email	1,618 (13.2%)	1,454 (13.9%)
Essay–Text message	1,182 (9.6%)	1,128 (10.8%)
Business memo–Email	1,014 (8.3%)	900 (8.6%)
Business memo–Text message	780 (6.4%)	718 (6.9%)
Essay–Business memo	186 (1.5%)	186 (1.8%)
Discourse type	*Text length (avg. chars)*	
Essay	11,098	10,117
Email	2,385	2,323
Business memo	1,255	1,042
Text message	611	601

Πίνακας 8: Δείχνει μια σύνοψη των δεδομένων, που χρησιμοποιήθηκαν για την εκπαίδευση και δοκιμή (training and testing) των διαφορετικών πτυχών της επαλήθευσης συγγραφέα. Πηγή: [31]

Το σώμα του κειμένου έχει ανωνυμοποιηθεί, καθώς τοποθεσίες, ονόματα προσώπων, διευθύνσεις κλπ. αντικαταστάθηκαν με γενικές ετικέτες. Αυτό είναι πολύ σημαντικό για την αξιολόγηση, αφού μειώνεται η επίδραση των πληροφοριών που αφορούν τον ίδιο συγγραφέα ή το θέμα του κειμένου.

Για τη δημιουργία των απαραίτητων συνόλων δεδομένων της εκπαίδευσης και του ελέγχου, το σώμα κειμένων χρειάζεται περαιτέρω προ-επεξεργασία. Πρώτα, οι διαθέσιμοι συγγραφείς χωρίστηκαν σε 2 ίσα σύνολα. Το ένα αφορά το σύνολο εκπαίδευσης και το άλλο το σύνολο του ελέγχου. Έτσι, εξασφαλίζεται ότι τυχόν ιδιαιτερότητες των συγγραφέων του συνόλου εκπαίδευσης δεν θα επηρεάσουν την απόδοση του συστήματος στο σύνολο της δοκιμής.

Επιπρόσθετα, χρησιμοποιήθηκαν δημογραφικά δεδομένα, ώστε να διατηρηθεί σταθερή η κατανομή του φύλου και στα 2 σύνολα. Τα σύνολα δεδομένων εκπαίδευσης και δοκιμής περιλαμβάνουν 56 συγγραφείς (10 άνδρες, 45 γυναίκες και 1 άγνωστο φύλο).

Το σύνολο δεδομένων αποτελείται από ζεύγη κειμένων, όπου τα δυο κείμενα ανήκουν σε διαφορετικούς τύπους γραπτού λόγου. Συμπεριλήφθηκαν όλοι οι 6 συνδυασμοί των 4 τύπων λόγου.

Όμως, η κατανομή των ζευγών κειμένου δεν είναι ομοιόμορφη μεταξύ των συνδυασμών τύπων λόγου, επειδή ο αριθμός των διαθέσιμων τύπων διαφέρει. Για παράδειγμα, κάθε συγγραφέας έχει πολλά email, αλλά μόνο ένα business memo.

Ωστόσο, και στα 2 σύνολα (training, testing), η αναλογία θετικών (ίδιος συγγραφέας) και αρνητικών (διαφορετικός συγγραφέας) περιπτώσεων είναι ισορροπημένη. Αυτό ισχύει και σε κάθε επιμέρους συνδυασμό τύπων γραπτού λόγου. Για παράδειγμα, στις περιπτώσεις essay-email το 50% των ζευγών είναι θετικά και το 50% αρνητικά. Η ισορροπία φύλου αποτρέπει συσχετίσεις του μοντέλου με δημογραφικά χαρακτηριστικά, ώστε να είναι αμερόληπτο (bias). Ο συνδυασμός διαφορετικών τύπων λόγου προσφέρει ρεαλιστικό σενάριο, όπου οι συγγραφείς αλλάζουν ύφος ανάλογα με την περίσταση. Τέλος, η ισορροπία θετικών και αρνητικών περιπτώσεων επιτρέπει στα μοντέλα να μάθουν συμμετρικά και αποτρέπει την προκατάληψη υπέρ κάποιας κατηγορίας.

Αφού το μήκος των κειμένων μερικών τύπων λόγου μπορεί να είναι περιορισμένο, συνενώθηκαν πολλαπλά κείμενα του ίδιου τύπου λόγου, ώστε να παραχθούν μεγαλύτερα δείγματα κειμένου. Πιο συγκεκριμένα, μηνύματα ηλεκτρονικού ταχυδρομείου (email messages) συνενώθηκαν, έτσι ώστε να δημιουργηθεί ένα δείγμα κειμένου τουλάχιστον 2.000 χαρακτήρων. Η ημερομηνία των μηνυμάτων λήφθηκε υπόψη, ώστε να ενωθούν διαδοχικά (σε χρονολογική σειρά). Στην περίπτωση των μηνυμάτων κειμένου (text messages), συνενώθηκαν μηνύματα είτε από επικοινωνία με φίλους, είτε με γονείς, ώστε να προκύπτουν δείγματα κειμένου τουλάχιστον 500 χαρακτήρων. Εισήχθη η ετικέτα <new> ανάμεσα στα ενωμένα μηνύματα, για να δηλώνει τα αρχικά όρια κάθε ξεχωριστού μηνύματος.

Οι τιμές μήκους κειμένου που παρουσιάζονται στον Πίνακα 7 για τα email και τα text messages, αναφέρονται σε αυτά τα ενωμένα δείγματα. Η συνένωση σύντομων μηνυμάτων είναι απαραίτητη για να μπορεί το μοντέλο να ανιχνεύσει καλύτερα στιλιστικά μοτίβα στη γραφή του συγγραφέα. Πολύ μικρά κείμενα (π.χ. 1–2 προτάσεις) δεν παρέχουν επαρκή γλωσσικά χαρακτηριστικά για ασφαλή επαλήθευση.

Έχουμε πάλι την συνάρτηση: φ: $(d_k, d_u) \rightarrow \{T, F\}$, αλλά στην φετινή έκδοση το d_k και το d_u ανήκουν σε διαφορετικούς τύπους λόγου.

Οι μετρικοί δείκτες είναι οι ίδιοι που χρησιμοποιήθηκαν και στο PAN 2021, δηλαδή οι AUROC, c@1, F1-score, F0.5u, Brier score. Χρησιμοποιήθηκε ο μέσος όρος των δεικτών για να παραχθεί η τελική βαθμολογία. Επιπλέον, αναφέρθηκε και ο χρόνος εκτέλεσης για να φανεί η σχετική αποδοτικότητα του κάθε συστήματος.

Για να διευκολυνθεί η σύγκριση των υποβληθέντων μεθόδων, οι διοργανωτές παρέχουν 2 βασικές μεθόδους αναφοράς, που βασίζονται σε χαρακτηριστικά ν-γραμμάτων (character n-grams) ή ακολουθίες χαρακτήρων (character sequencies).

Compression-based model:

Για κάθε ζεύγος κειμένων t1 και t2, υπολογίζεται η cross-entropy του t2, με βάση το μοντέλο Prediction by Partial Matching (PPM) του t1, και το αντίστροφο. Οι δύο τιμές εντροπίας, (μια για κάθε κατεύθυνση) χρησιμοποιούνται για να εξαχθεί ο μέσος όρος τους και η απόλυτα διαφορά τους. Έπειτα, εκπαιδεύεται ένας ταξινομητής λογιστικής παλινδρόμησης πάνω σε αυτά τα χαρακτηριστικά για να ένα σκορ επαλήθευσης (πιθανότητα ότι τα δυο κείμενα γράφτηκαν από τον ίδιο συγγραφέα). Εάν το τελικό σκορ είναι πολύ κοντά στο 0.5, τότε προσαρμόζεται ακριβώς στο 0.5, ώστε να θεωρείται αναπάντητο πρόβλημα.

Αν δύο κείμενα έχουν παρόμοια στατιστική συμπίεση (δηλαδή το ένα μπορεί να προβλέψει καλά το άλλο μέσω PPM), τότε πιθανόν προέρχονται από τον ίδιο συγγραφέα.

Distance-based character n-gram model:

Τα πιο συχνά τετραγράμματα χαρακτήρων εξάγονται από τα κείμενα εκπαίδευσης. Κάθε κείμενο αναπαρίσταται ως διάνυσμα χαρακτήρων αυτών των 4 γραμμάτων. Έπειτα, για κάθε ζεύγος κειμένων, υπολογίζεται η ομοιότητα συνημιτόνου (cosine similarity) μεταξύ των δυο διανυσμάτων. Κατά την εκπαίδευση, βελτιστοποιούνται 2 κατώφλια p1 (κάτω όριο) και το p2 (άνω όριο). Το τελικό αποτέλεσμα έχει ως εξής: Αν το σκορ είναι μικρότερο από το p1, τότε έχουμε αρνητική απάντηση (διαφορετικοί συγγραφείς. Αν το σκορ είναι μεγαλύτερο από το p2, τότε έχουμε θετική απάντηση (ίδιος συγγραφέας). Αν το σκορ είναι μεταξύ p1 και p2, τότε ορίζεται σε 0.5 (δηλαδή το σύστημα δεν απαντά, θεωρώντας δύσκολο το πρόβλημα).

Και οι δύο προσεγγίσεις στηρίζονται σε χαρακτηριστικά επιπέδου χαρακτήρα και δεν έχουν ρυθμιστεί ώστε να εξαρτώνται από τον τύπο του λόγου.

Λήφθηκαν 7 υποβολές και αξιολογήθηκε η αποτελεσματικότητά και η απόδοσή τους, χρησιμοποιώντας την ερευνητική πλατφόρμα TIRA. Όλοι οι συμμετέχοντες υπέβαλαν και μια τεχνική αναφορά (notebook paper), περιγράφοντας τις μεθόδους τους. Τα κύρια χαρακτηριστικά φαίνονται στον πίνακα 9.

System	Ref.	Representation	Classification	Augmentation	Type-specific
CRESPOSANCHEZ22	[25]	word unigrams, doc2vec	FC	Yes	No
		(text and POS), SOM	FC	Yes	No
GALICIA22	[24]	graph-based, POS	Siamese network	Yes	No
HUANG22	[26]	BERT	FC	No	No
JINLI22	[23]	MPNET	FC	No	No
LEI22	[27]	BERT	FC	No	No
NAJAFI22	[28]	T5, word unigrams, POS,	CNN, attention	No	No
		NEs, Punctuation	block, FC	No	No
YIHUIYE22	[29]	BERT	TextCNN	Yes	No

Πίνακας 9: Σύνοψη των βασικών χαρακτηριστικών των υποβληθέντων μεθόδων (POS: Part of Speech tags, NEs: Named Entities, SOM: Self-organizing Maps, FC: Fully-Connected Layer). Πηγή: [31]

Οι περισσότεροι συμμετέχοντες χρησιμοποίησαν προ-εκπαιδευμένα γλωσσικά μοντέλα, όπως BERT, T5, ή MPNET για την εξαγωγή ενσωματώσεων κειμένου (text embeddings).

Οι προσεγγίσεις που δεν χρησιμοποιούν προ-εκπαιδευμένα γλωσσικά μοντέλα, βασίζονται σε αναπαραστάσεις κειμένου με γραφήματα (graph-based representation), φασματική ανάλυση (spectral analysis), ή σε παραδοσιακά χαρακτηριστικά (feature engineering), όπως POS tag και μονολεκτικά χαρακτηριστικά.

Επίσης, δοκιμάστηκαν παραδοσιακοί αλγόριθμοι μηχανικής μάθησης, όπως Support Vector Machines (SVMs) και Random Forests, αλλά η απόδοσή τους ήταν χαμηλότερη.

Επιπλέον, χρησιμοποιήθηκαν και άλλες βαθιές νευρωνικές προσεγγίσεις, όπως CNNs και Siamese Networks.

Αξιοσημείωτο είναι ότι κανένας συμμετέχων δεν διερεύνησε προσεγγίσεις εξειδικευμένες ως προς τον τύπο λόγου (discourse type), παρά τις σημαντικές διαφορές που υπήρχαν μεταξύ των τύπων λόγου των κειμένων.

Η πλειονότητα βασίστηκε σε pre-trained LMs όπως BERT, T5, MPNET για δημιουργία embeddings.

Στον πίνακα 10 παρουσιάζονται τα συνολικά αποτελέσματα όλων των συμμετεχόντων. Γενικά, η αποτελεσματικότητα όλων των υποβολών είναι σχετικά χαμηλή, κάτι που αντικατοπτρίζει τη δυσκολία του συγκεκριμένου έργου (task).

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
baseline-cngdist22	0.546	0.496	0.669	0.542	0.749	**0.600**
najafi22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
galicia22	0.512	0.499	0.628	0.544	0.741	0.585
jinli22	0.577	0.557	0.581	0.563	0.589	0.573
baseline-compressor22	0.541	0.493	0.570	0.478	**0.750**	0.566
lei22	0.539	0.539	0.399	0.488	0.539	0.501
yihuiye22	0.542	0.526	0.398	0.461	0.565	0.499
huang22	0.519	0.519	0.196	0.328	0.519	0.416
cresposanchez22	0.500	0.500	0	0	0.748	0.350

Πίνακας 10: Αποτελέσματα όλων των συμμετεχόντων. Πηγή: [31]

Οι προσεγγίσεις των najafi22, galicia22 και jinli22 ξεπερνούν τις υπόλοιπες συμμετοχές. Είναι, επίσης, εντυπωσιακό ότι ένα απλό baseline μοντέλο πέτυχε το καλύτερο συνολικό σκορ, παρότι τα περισσότερα μοντέλα των συμμετεχόντων ήταν πολύ πιο περίπλοκα και εξελιγμένα.

Από την άλλη πλευρά, η πιο αποτελεσματική μέθοδος που υποβλήθηκε, αυτή των najafi22, υπερτερεί όλων των άλλων συμμετοχών και των baseline σε τρεις από τις πέντε μετρικές αξιολόγησης. Η κύρια αδυναμία της φαίνεται να είναι το χαμηλό Brier score, το οποίο υποδηλώνει ότι οι πιθανολογικές της προβλέψεις χρειάζονται βελτίωση, παρότι οι δυαδικές ταξινομήσεις (same/different author) είναι σχετικά αξιόπιστες.

Ο πίνακας 11 αναλύει τα αποτελέσματα ως προς τους έξι συνδυασμούς τύπων λόγου. Όπως είπαμε, ο κάθε τύπος λόγου έχει διαφορετικό μέσο μήκος κειμένου. Για παράδειγμα, τα δοκίμια (essays) είναι πολύ μεγαλύτερα από όλους τους υπόλοιπους τύπους λόγου που εξετάστηκαν.

(a) Email–Text message

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
baseline-cngdist22	0.585	0.505	**0.672**	0.555	0.751	**0.614**
najafi22	**0.613**	**0.583**	0.579	**0.581**	0.628	0.597
galicia22	0.517	0.502	0.641	0.549	0.740	0.590
jinli22	0.599	0.576	0.603	**0.581**	0.607	0.593
baseline-compressor22	0.570	0.480	0.661	0.499	**0.752**	0.592
lei22	0.537	0.537	0.350	0.462	0.537	0.484
yihuiye22	0.547	0.530	0.409	0.470	0.569	0.505
huang22	0.519	0.519	0.197	0.329	0.519	0.417
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(b) Essay–Email

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
baseline-cngdist22	0.514	0.489	0.657	0.542	0.740	0.589
najafi22	0.570	0.549	0.583	0.557	0.605	0.573
galicia22	0.496	0.498	0.608	0.537	0.744	0.577
jinli22	**0.588**	0.559	0.574	0.563	0.600	0.577
baseline-compressor22	0.531	0.481	**0.659**	0.531	**0.750**	**0.590**
lei22	0.585	**0.585**	0.528	**0.575**	0.585	0.571
yihuiye22	0.548	0.529	0.371	0.449	0.558	0.491
huang22	0.553	0.553	0.283	0.444	0.553	0.477
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(c) Essay–Text message

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
baseline-cngdist22	0.540	0.493	**0.673**	0.539	0.750	**0.599**
najafi22	**0.568**	**0.553**	0.567	**0.556**	0.595	0.568
galicia22	0.513	0.493	0.604	0.534	0.743	0.577
jinli22	0.476	0.483	0.486	0.485	0.520	0.490
baseline-compressor22	0.567	0.513	0.130	0.186	**0.751**	0.429
lei22	0.519	0.519	0.299	0.412	0.519	0.453
yihuiye22	0.509	0.508	0.336	0.410	0.542	0.461
huang22	0.516	0.516	0.173	0.301	0.516	0.404
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(e) Business memo–Email

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
baseline-cngdist22	0.509	0.443	**0.67**	0.500	**0.748**	0.574
najafi22	**0.606**	**0.586**	0.612	**0.589**	0.633	**0.605**
galicia22	0.521	0.513	0.647	0.556	0.742	0.596
jinli22	0.562	0.547	0.565	0.552	0.569	0.559
baseline-compressor22	0.514	0.494	0.214	0.269	0.746	0.447
lei22	0.512	0.512	0.497	0.507	0.512	0.508
yihuiye22	0.539	0.520	0.414	0.461	0.571	0.501
huang22	0.493	0.493	0.099	0.185	0.493	0.353
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(d) Business memo–Text message

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
baseline-cngdist22	0.525	0.376	**0.673**	0.455	**0.748**	**0.555**
najafi22	**0.582**	0.534	0.515	0.526	0.589	0.549
galicia22	0.487	0.464	0.553	0.503	0.741	0.549
jinli22	0.551	0.535	0.567	**0.545**	0.566	0.553
baseline-compressor22	0.524	0.518	0.065	0.127	0.746	0.396
lei22	0.539	**0.539**	0.472	0.517	0.539	0.521
yihuiye22	0.553	0.538	0.463	0.499	0.579	0.526
huang22	0.481	0.481	0.126	0.214	0.481	0.356
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(f) Essay–Business memo

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
baseline-cngdist22	0.474	0.466	**0.647**	0.512	0.740	0.568
najafi22	**0.545**	**0.538**	0.533	0.536	0.572	0.545
galicia22	0.496	0.500	0.635	**0.547**	0.739	**0.584**
jinli22	0.528	0.500	0.542	0.516	0.532	0.524
baseline-compressor22	0.474	0.477	0.477	0.446	0.744	0.523
lei22	0.500	0.500	0.367	0.437	0.500	0.461
yihuiye22	0.522	0.492	0.214	0.302	0.545	0.415
huang22	0.527	0.527	0.290	0.415	0.527	0.457
cresposanchez22	0.500	0.500	0	0	**0.748**	0.350

Πίνακας 11: Τα αποτελέσματα αξιολόγησης ως προς τους έξι συνδυασμούς τύπων λόγου. Πηγή: [31]

Όπως φαίνεται στους πίνακες b, c και f, όταν τα δοκίμια περιλαμβάνονται σε κάποιον συνδυασμό, η υποβολή των galicia22 είναι η πιο αποτελεσματική συνολικά. Αντίθετα, όταν τα δοκίμια δεν περιλαμβάνονται (πίνακες a, e και d), η προσέγγισή τους υπερτερείται από αυτή των najafi22.

Στους πιο σύντομους τύπους λόγου, δηλαδή στα επαγγελματικά σημειώματα (business memos) και τα μηνύματα κειμένου (text messages) (πίνακας d), η υποβολή των jinli22 φαίνεται να είναι η πιο αποτελεσματική.

Αυτός ο συγκεκριμένος συνδυασμός τύπων λόγου παρουσιάζει επίσης τη χαμηλότερη συνολική αποτελεσματικότητα, γεγονός που υποδηλώνει ότι το μήκος του κειμένου (σε συνδυασμό με τη διασταυρούμενη επαλήθευση τύπων λόγου, cross-discourse verification) παραμένει κρίσιμος παράγοντας στην επαλήθευση συγγραφέα.

Το baseline-cngdist22 εμφανίζεται σχετικά σταθερό σε όλους τους συνδυασμούς τύπων λόγου, ενώ το baseline-compressor22 επιτυγχάνει τα βέλτιστα αποτελέσματά του όταν περιλαμβάνονται οι μεγαλύτεροι τύποι κειμένων (όπως τα δοκίμια).

Τα αποτελέσματα δείχνουν ότι η επίδοση των μοντέλων εξαρτάται από το μέγεθος του κειμένου. Όσο μικρότερο το κείμενο, τόσο πιο δύσκολο να εξαχθούν στιλιστικά χαρακτηριστικά του συγγραφέα. Επίσης, η σταθερότητα των baseline μοντέλων δείχνει ότι η εξαγωγή απλών στιλιστικών χαρακτηριστικών (όπως character n-grams) εξακολουθούν να λειτουργούν ικανοποιητικά, ακόμη και απέναντι σε σύγχρονες deep learning μεθόδους.

Στον πίνακα 12a παρουσιάζεται ο αριθμός των θετικών και αρνητικών απαντήσεων και μη-απαντήσεων, που παρήγαγε το κάθε μοντέλο επαλήθευσης.

(a)

System	Positive	Negative	Unanswered
CRESPOSANCHEZ22	0	10,478	0
GALICIA22	8,874	1,604	0
HUANG22	1,031	9,447	0
JINLI22	5,820	4,658	0
LEI22	2,805	7,673	0
NAJAFI22	5,355	5,083	40
YIHUIYE22	2,841	7,116	521
BASELINE-CNGDIST22	9,199	17	1,262
BASELINE-COMPRESSOR22	3,927	3,268	3,283

(b)

System	Run time
CRESPOSANCHEZ22	00:05:36
GALICIA22	00:07:22
LEI22	06:04:59
YIHUIYE22	07:16:59
NAJAFI22	18:18:32
JINLI22	23:25.62
HUANG22	31:04:56

Πίνακας 12: a) Αριθμός θετικών, αρνητικών και μη-απαντήσεων του κάθε μοντέλου επαλήθευσης. b) Χρόνος εκτέλεσης του κάθε μοντέλου. Πηγή: [31]

Πρέπει να σημειωθεί ότι οι απαντήσεις αυτές δεν είναι απαραίτητα σωστές προβλέψεις, απλώς αντιστοιχούν στις περιπτώσεις του συνόλου δοκιμής, όπου το εκτιμώμενο σκορ επαλήθευσης είναι μεγαλύτερο ή μικρότερο από το 0.5.

Επιπλέον, παρουσιάζεται και ο αριθμός των περιπτώσεων με σκορ ίσο με 0.5, οι οποίες θεωρούνται μη απαντημένες περιπτώσεις (non-answers).

Αυτά τα τρία μεγέθη (θετικές, αρνητικές, μη απαντημένες) δείχνουν τον βαθμό μεροληψίας (bias) κάθε μοντέλου επαλήθευσης.

Υπενθυμίζεται ότι η πραγματική κατανομή θετικών και αρνητικών περιπτώσεων στο test set είναι ισορροπημένη.

Όπως φαίνεται, πολύ λίγα συστήματα αφήνουν περιπτώσεις αναπάντητες, πράγμα που σημαίνει ότι η αποτελεσματικότητά τους (ιδιαίτερα ως προς το c@1) θα μπορούσε να βελτιωθεί σημαντικά, εάν

ενσωμάτωναν έναν μηχανισμό για να εξαιρούν τις οριακές περιπτώσεις (με σκορ κοντά στο 0.5), όπως κάνουν τα baseline μοντέλα.

Αξίζει να σημειωθεί ότι οι προσεγγίσεις των najafi22 και jinli22 (καθώς και το baseline-compressor22) είναι αμερόληπτες, δηλαδή παράγουν παρόμοιο αριθμό θετικών και αρνητικών απαντήσεων.

Αντίθετα, η υποβολή των galicia22 και το baseline-cngdist22 είναι προκατειλημμένα προς θετικές απαντήσεις, ενώ οι huang22, lei22, yihuiye22 και cresposanchez22 είναι προκατειλημμένα προς αρνητικές απαντήσεις.

Σημαντικό είναι, επίσης, ότι αυτές οι μεροληψίες δεν επηρεάζουν το AUROC, επειδή αυτό το μέτρο βασίζεται στη σχετική κατάταξη των σκορ και όχι στις απόλυτες τιμές ή στην ισορροπία των κατηγοριών.

Πέρα από την αποτελεσματικότητα, ένα ακόμη κριτήριο αξιολόγησης ενός συστήματος επαλήθευσης συγγραφέα είναι η αποδοτικότητα ή αλλιώς το κόστος (χρόνος) εκτέλεσης (runtime cost).

Ανάλογα με την τεχνολογία που χρησιμοποιείται, αυτό το κριτήριο μπορεί να είναι ιδιαίτερα σημαντικό, ειδικά όταν πρέπει να αναλυθούν μεγάλοι όγκοι κειμένων.

Ο Πίνακας 12b δείχνει τον χρόνο εκτέλεσης κάθε μεθόδου που υποβλήθηκε, όπως μετρήθηκε στο σύστημα TIRA.

Όπως φαίνεται, οι προσεγγίσεις που αποφεύγουν τη χρήση προ-εκπαιδευμένων γλωσσικών μοντέλων έχουν πολύ μικρότερο χρόνο εκτέλεσης, με μεγάλη διαφορά από τις υπόλοιπες.

Αντίθετα, η μέθοδος huang22 απαιτεί τον μεγαλύτερο χρόνο εκτέλεσης, επειδή διαχωρίζει τα κείμενα σε επιμέρους τμήματα και εξετάζει όλους τους πιθανούς συνδυασμούς αυτών των τμημάτων.

Οι διοργανωτές εφάρμοσαν τις καλύτερες μεθόδους από την προηγούμενη διοργάνωση του PAN 2021 στο νέο σύνολο δεδομένων ελέγχου του PAN 2022.

Αυτό ουσιαστικά συνιστά ένα πείραμα μεταφοράς μάθησης (transfer learning), καθώς τα τρία μοντέλα είχαν εκπαιδευτεί και βελτιστοποιηθεί πάνω σε σύνολο δεδομένων cross-fandom (δηλαδή έλεγχο συγγραφέα μεταξύ διαφορετικών fandoms), αλλά τώρα δοκιμάζονται πάνω στο cross-discourse type dataset, όπου οι τύποι λόγου (discourse types) διαφέρουν σημαντικά (π.χ. email, essay, memo, text message).

Οι μέθοδοι που χρησιμοποιήθηκαν ήταν οι εξής:

- boenninghoff21: Μοντέλο βαθιάς μάθησης που περιλαμβάνει νευρωνική εξαγωγή χαρακτηριστικών, deep metric learning, μοντελοποίηση αβεβαιότητας, Bayesian factor scoring, συνδυαστική συνάρτηση απώλειας (loss function), και ανιχνευτή out-of-distribution για περιπτώσεις όπου το μοντέλο δεν απαντά. Τελικά, χρησιμοποιεί ensemble (majority voting).
- embarcaderoruiz21: Προσέγγιση βασισμένη σε γραφήματα (graph-based) συνδυασμένη με ένα Siamese neural network - παρόμοια με αυτή του galicia22.
- weerasinghe21: Χρησιμοποιεί στιλομετρικά χαρακτηριστικά (stylometric features), όπως χαρακτηρικά n-grams, POS tags, λέξεις λειτουργίας (function words), δείκτες λεξιλογικού πλούτου, και έναν λογιστικό ταξινομητή (logistic regression) που λαμβάνει τις απόλυτες διαφορές αυτών των χαρακτηριστικών για κάθε ζεύγος κειμένων.

Δεν έγινε καμία τροποποίηση στις μεθόδους πριν την εφαρμογή τους στο νέο dataset.

Η αποτελεσματικότητα αυτών των μεθόδων στο σύνολο δοκιμών του PAN 2021 ήταν εξαιρετική, με σκορ πάνω από 0.93 σε όλες τις μετρικές.

Ωστόσο, όπως φαίνεται στον Πίνακα 13a, η απόδοσή τους στο dataset του 2022 ήταν πολύ χειρότερη.

	(a)						(b)		
System	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**	**Positive**	**Negative**	**Unanswered**
boenninghoff21	0.513	0.501	0.002	0.005	0.531	0.310	10	10,370	98
embarcaderoruiz21	0.538	0.502	0.063	0.116	0.581	0.360	309	9,295	874
weerasinghe21	0.488	0.500	0.011	0.027	0.506	0.306	57	10,421	0

Πίνακας 13: a) Αποτελέσματα αξιολόγησης των κορυφαίο μοντέλων, που υποβλήθηκαν στο PAN 2021, στο dataset 2022 b) Θετικές, αρνητικές και μη-απαντήσεις που παρείχαν αυτά τα μοντέλα. Πηγή: [31]

Η συνολική αποτελεσματικότητα στα cross-discourse δεδομένα ήταν πολύ χαμηλή, ακόμη χαμηλότερη από εκείνη των επτά νέων υποβολών και των δύο baseline μοντέλων του 2022.

Αυτό δείχνει ότι το έντονο fine-tuning τέτοιων μοντέλων μπορεί να βελτιώνει την απόδοση σε ένα συγκεκριμένο σύνολο δεδομένων, αλλά μειώνει τη γενικευσιμότητά τους (generalizability). Με άλλα λόγια, ένα μοντέλο που έχει μάθει πολύ καλά ένα συγκεκριμένο είδος δεδομένων (όπως fanfiction) δυσκολεύεται να μεταφέρει τη γνώση του σε διαφορετικά είδη (όπως emails ή δοκίμια).

Ο Πίνακας 13b παρουσιάζει τον αριθμό των θετικών, αρνητικών και μη απαντημένων περιπτώσεων για κάθε ένα από τα τρία μοντέλα του 2021.
Όλα παρουσιάζουν έντονη προκατάληψη (bias) προς τις αρνητικές απαντήσεις, δηλαδή τείνουν να απορρίπτουν τη συσχέτιση των κειμένων ως έργο του ίδιου συγγραφέα.

Στο dataset του PAN 2021 (cross-fandom), όλα τα κείμενα είχαν παρόμοιο μήκος, ενώ στο cross-discourse dataset του 2022 υπάρχουν μεγάλες διαφορές στο μήκος και στο ύφος.
Αυτός ο παράγοντας, μαζί με άλλες ουσιαστικές διαφορές μεταξύ fanfiction και δοκιμίων ή επιχειρηματικών τύπων λόγου, μπέρδεψε τα μοντέλα, τα οποία χρειάζονται κατάλληλο fine-tuning για να προσαρμοστούν σωστά στις νέες κλίμακες των σκορ επαλήθευσης.

Με λίγα λόγια, οι μεγάλες διαφορές στο ύφος, το μήκος και τη θεματολογία καθιστούν την cross-discourse author verification ένα από τα πιο δύσκολα σενάρια.

Αξίζει να σημειωθεί ότι ακόμη και οι τιμές AUROC (οι οποίες δεν εξαρτώνται από τη βαθμονόμηση των σκορ) ήταν πολύ χαμηλές, επιβεβαιώνοντας την αδυναμία μεταφοράς της μάθησης μεταξύ τόσο διαφορετικών τύπων δεδομένων.

Η αποτελεσματικότητα όλων των υποβολών στο cross-discourse dataset ήταν χαμηλή, σε ορισμένες περιπτώσεις τόσο χαμηλή όσο η τυχαία πρόβλεψη (random baseline).
Αξιοσημείωτο είναι ότι όλες οι προσεγγίσεις, παρά το υψηλό επίπεδο πολυπλοκότητας, ξεπεράστηκαν από μια απλή baseline μέθοδο βασισμένη σε character n-grams και cosine similarity.

Αυτό υποδηλώνει ότι οι παραδοσιακές μέθοδοι στιλομετρίας με καλά κατανοημένα χαρακτηριστικά, μπορούν να είναι πιο αποτελεσματικές από τις βαθιές νευρωνικές προσεγγίσεις που χρησιμοποιούν προ-εκπαιδευμένα γλωσσικά μοντέλα σε αυτή τη δύσκολη εκδοχή του προβλήματος.

Ένας ακόμη παράγοντας είναι ο όγκος των δεδομένων εκπαίδευσης (~12.000 ζεύγη), που θεωρείται ανεπαρκής για να εκπαιδεύσει σωστά μοντέλα βαθιάς μάθησης, τα οποία συνήθως απαιτούν πολύ περισσότερα δεδομένα για να αποδώσουν καλά.

Ένα άλλο κρίσιμο ζήτημα είναι το μήκος του κειμένου. Όταν χρησιμοποιήθηκαν μεγάλα δοκίμια (essays**)**, η γραφική προσέγγιση (graph-based) του galicia22 είχε την καλύτερη απόδοση.

Αντίθετα, όταν τα δεδομένα ήταν σύντομα κείμενα (όπως emails, business memos, ή text messages), οι προσεγγίσεις με προ-εκπαιδευμένα γλωσσικά μοντέλα, όπως του najafi22 και του jinli22, αποδείχθηκαν πιο αποτελεσματικές.

Η συνολικά χαμηλή αποτελεσματικότητα όλων των συμμετοχών δείχνει ότι υπάρχει μεγάλο περιθώριο βελτίωσης στην επαλήθευση συγγραφέα μεταξύ διαφορετικών τύπων λόγου. Όλες οι υποβολές χρησιμοποίησαν ενιαίο μοντέλο που προβλέπει την ταυτότητα του συγγραφέα χωρίς να λαμβάνει υπόψη τον συνδυασμό των τύπων λόγου.

Τέλος, μια ensemble προσέγγιση που θα συνδυάζει τις παραδοσιακές στιλομετρικές μεθόδους με τα μοντέρνα προ-εκπαιδευμένα γλωσσικά μοντέλα φαίνεται να είναι μια πολύ υποσχόμενη κατεύθυνση για μελλοντική έρευνα.

4.6 Σύγχρονες μέθοδοι

Τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs) έχουν επιδείξει εξαιρετική ικανότητα στη συλλογιστική και στην επίλυση προβλημάτων. Ωστόσο, οι δυνατότητές τους στην ανάλυση συγγραφικής ταυτότητας παραμένουν σχετικά ανεξερεύνητες. Οι παραδοσιακές μελέτες βασίζονταν σε χειροποίητα υφολογικά χαρακτηριστικά, ενώ οι σύγχρονες προσεγγίσεις αξιοποιούν ενσωματώσεις κειμένου (text embeddings) που παράγονται από προεκπαιδευμένα γλωσσικά μοντέλα (κυρίως αρχιτεκτονικές τύπου BERT). Οι μέθοδοι αυτές, οι οποίες συνήθως απαιτούν προσαρμογή (fine-tuning) σε επισημασμένα δεδομένα, παρουσιάζουν υψηλή αποτελεσματικότητα σε εφαρμογές εντός συγκεκριμένων θεματικών πεδίων. Αντίθετα, παρουσιάζουν υποβάθμιση της απόδοσής τους σε εφαρμογές μεταξύ διαφορετικών πεδίων (cross-domain) και παρέχουν περιορισμένη δυνατότητα ερμηνείας των αποτελεσμάτων τους.

Σε σύγκριση με τις παραδοσιακές στατιστικές μεθόδους, οι τεχνικές βαθιάς μάθησης (deep learning) απαιτούν σημαντικά λιγότερη χειροκίνητη κατασκευή χαρακτηριστικών (feature engineering). Από την άλλη, η αποτελεσματικότητα των μεθόδων αυτών υποβαθμίζεται σημαντικά όταν τα προς εξέταση κείμενα (query texts) είναι μικρής έκτασης ή όταν υπάρχουν περιορισμένα διαθέσιμα κείμενα από τους υποψήφιους συγγραφείς. Η μείωση αυτή περιορίζει την πρακτική τους εφαρμογή σε πραγματικές συνθήκες, όπου η έλλειψη δεδομένων και η ποικιλομορφία των κειμένων αποτελούν τον κανόνα

Επιπρόσθετα, οι μέθοδοι βαθιάς μάθησης απαιτούν σημαντικό χρόνο εκπαίδευσης και μεγάλες ποσότητες επισημασμένων δεδομένων (labeled data), δεν γενικεύουν αποτελεσματικά σε διαφορετικά πεδία εφαρμογής και χαρακτηρίζονται από περιορισμένη ερμηνευσιμότητα (explainability).

4.6.1 Linguistically Informed Prompting (LIP)

Οι Huang et al. [32] πρότειναν την μέθοδο Linguistically Informed Prompting (LIP).

Τα αποτελέσματα τους έδειξαν ότι τα LLMs μπορούν να εκτελέσουν αποτελεσματικά τόσο την επαλήθευση όσο και την απόδοση συγγραφικής ταυτότητας σε καθεστώς zero-shot, χωρίς να απαιτείται διαδικασία fine-tuning.

Παρατήρησαν ότι εάν η διαμόρφωση των δεδομένων δεν είναι ισορροπημένη, τα Μεγάλα Γλωσσικά Μοντέλα (LLMs) τείνουν να προβλέπουν ότι οποιαδήποτε δύο δοθέντα κείμενα έχουν γραφτεί από διαφορετικούς συγγραφείς. Το φαινόμενο αυτό πιθανόν οφείλεται στο ότι τα LLMs εκπαιδεύονται σε ιδιαίτερα ποικίλα σύνολα δεδομένων με μεγάλο αριθμό συγγραφέων, γεγονός που τα καθιστά αποτελεσματικότερα στον εντοπισμό διαφορών στο ύφος γραφής παρά στην αναγνώριση ομοιοτήτων. Επιπλέον, η περιορισμένη παρουσία πολλαπλών έργων του ίδιου συγγραφέα κατά την εκπαίδευση ενδέχεται να εμποδίζει τα μοντέλα να μάθουν τις ιδιαίτερες υφολογικές ιδιαιτερότητες κάθε συγγραφέα.

Χρησιμοποίησαν προτροπές (prompts) "LIP", "no_guidance", "little_guidance" και "grammar_guidance", όπου αφορούν ανάλογα στον βαθμό εξειδίκευσης και στην εστίασή τους.

- Η προτροπή "no_guidance" παρέχει μόνο μια βασική περιγραφή της εργασίας, χωρίς περαιτέρω καθοδήγηση.

- Η "little_guidance" περιορίζει την ανάλυση ζητώντας από το μοντέλο να εστιάσει περισσότερο στο ύφος γραφής παρά στις διαφορές του περιεχομένου ή του θέματος.
- Η "grammar_guidance" προχωρά ένα βήμα παραπέρα, δίνοντας έμφαση στα γραμματικά χαρακτηριστικά του ύφους που αντανακλούν τη συγγραφική ταυτότητα.
- Αντίθετα, η "LIP" (Linguistically Informed Prompting) παρέχει έναν συγκεκριμένο κατάλογο γλωσσικών χαρακτηριστικών προς ανάλυση, όπως τα φραστικά ρήματα (phrasal verbs), τα τροπικά ρήματα (modal verbs), τα μοτίβα στίξης και τα τυπογραφικά λάθη, εξασφαλίζοντας μια περισσότερο ολοκληρωμένη και στοχευμένη αξιολόγηση.

Η γλωσσική αυτή καθοδήγηση επιτρέπει μια πιο λεπτομερή και ακριβή εκτίμηση της συγγραφικής ταυτότητας, ιδιαίτερα σε περιπτώσεις όπου ένας συγγραφέας γράφει για διαφορετικά θέματα και σε διαφορετικά θεματικά πεδία (domains). Σε τέτοιες περιπτώσεις, η εστίαση αποκλειστικά στο περιεχόμενο μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα, ενώ η ανάλυση σταθερών γλωσσικών και υφολογικών χαρακτηριστικών επιτρέπει την καλύτερη αναγνώριση του συγγραφικού ύφους. Για τον λόγο αυτό, η μέθοδος LIP παρουσιάζει βελτιωμένη απόδοση σε σχέση με τις λιγότερο καθοδηγούμενες προτροπές σε αρκετές από τις πειραματικές συνθήκες που εξετάστηκαν, αναδεικνύοντας τη σημασία της γλωσσολογικής καθοδήγησης στην επαλήθευση συγγραφικής ταυτότητας.

AUTHORSHIP VERIFICATION:

System instruction: Respond with a JSON object including two key elements:
"analysis": Reasoning behind your answer.
"answer": A boolean (True/False) answer.

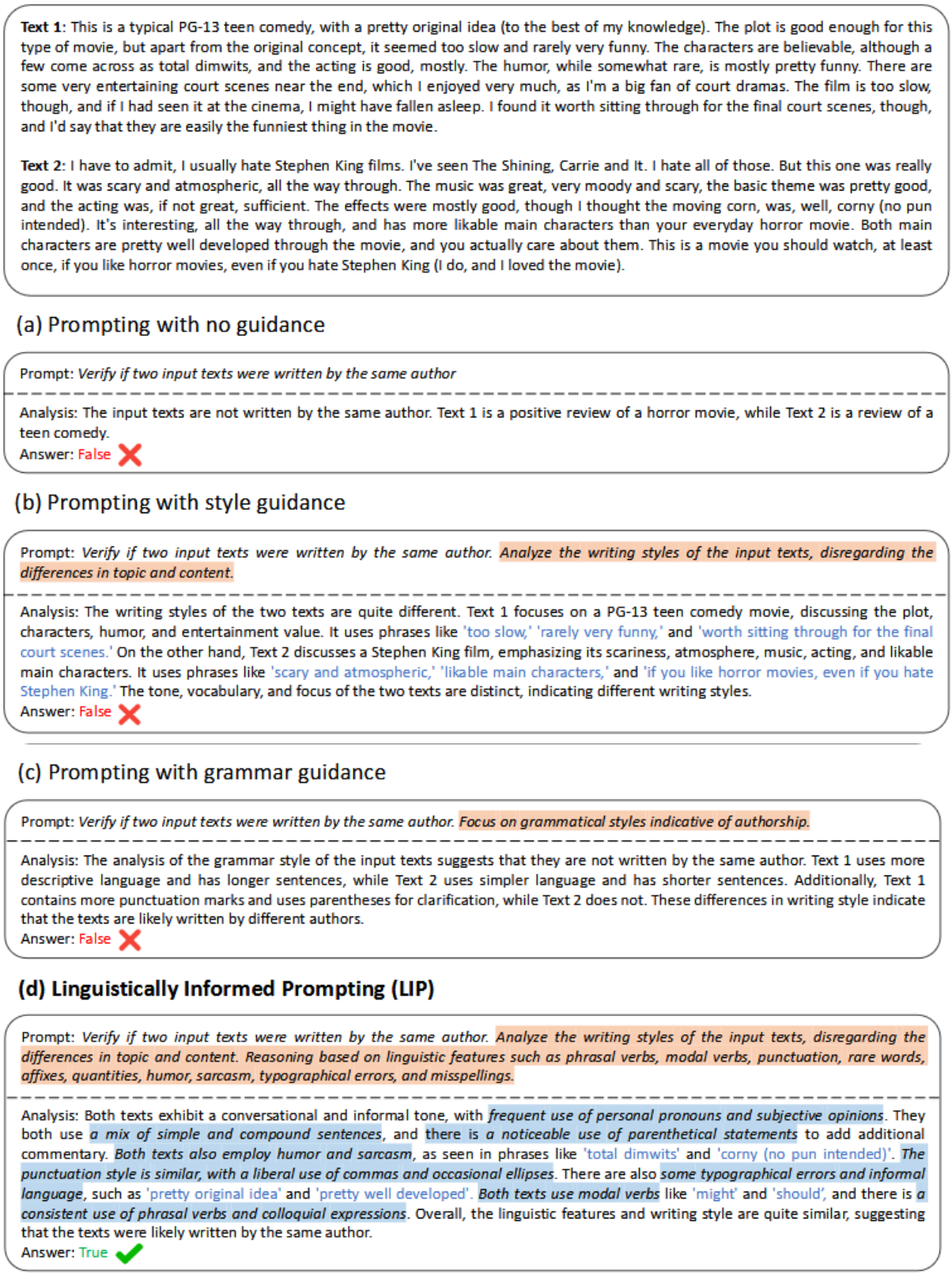
Prompting with no guidance: Verify if two input texts were written by the same author. Input text 1: <text 1>, text 2: <text 2>

Prompting with style guidance: Verify if two input texts were written by the same author. Analyze the writing styles of the input texts, disregarding the differences in topic and content. Input text 1: <text 1>, text 2: <text 2>

Prompting with grammar guidance: Verify if two input texts were written by the same author. Focus on grammatical styles indicative of authorship. Input text 1: <text 1>, text 2: <text 2>

Linguistically Informed Prompting (LIP): Verify if two input texts were written by the same author. Analyze the writing styles of the input texts, disregarding the differences in topic and content. Reasoning based on linguistic features such as phrasal verbs, modal verbs, punctuation, rare words, affixes, quantities, humor, sarcasm, typographical errors, and misspellings. Input text 1: <text 1>, text 2: <text 2>

Εικόνα 27: Τα prompts που χρησιμοποιήθηκαν. Πηγή: [32]



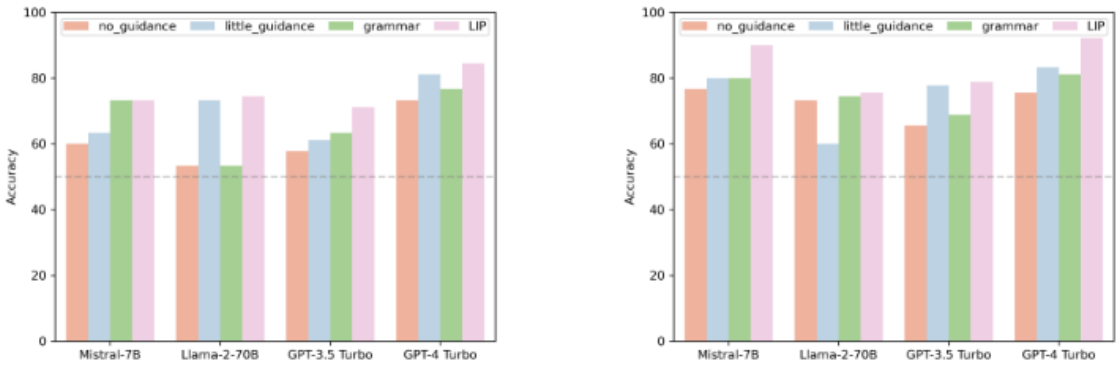
Εικόνα 28: Η Analysis και Answer είναι η έξοδος του GPT-4. Μόνο η μέθοδος LIP πέτυχε σωστά ότι τα δύο κείμενα είναι γραμμένα από τον ίδιο συγγραφέα. Στο κείμενο με υπογραμμισμένο πορτοκαλί φαίνονται οι διαφορές συγκριτικά με την προτροπή χωρίς καθοδήγηση. Στο κείμενο με το μπλε φαίνεται η μέθοδος LIP. Πηγή: [32]

Τα αποτελέσματα των πειραμάτων, που φαίνονται στην εικόνα 29, παρέχουν μια συγκριτική ανάλυση της απόδοσης των Μεγάλων Γλωσσικών Μοντέλων (LLMs) στην εργασία της επαλήθευσης συγγραφικής ταυτότητας (authorship verification) χρησιμοποιώντας δύο διαφορετικά σύνολα δεδομένων.

Συνολικά αξιολογούνται τέσσερα γλωσσικά μοντέλα, καθένα από τα οποία δοκιμάζεται με τέσσερις διαφορετικές μορφές προτροπών (prompt settings). Το GPT-4 Turbo παρουσιάζει σταθερά την υψηλότερη απόδοση και στα δύο σύνολα δεδομένων, γεγονός που υποδηλώνει την ανώτερη ικανότητά του να αναγνωρίζει τη συγγραφική ταυτότητα.

Παράλληλα, η μέθοδος Linguistically Informed Prompting (LIP) επιτυγχάνει γενικά τις υψηλότερες επιδόσεις στις περισσότερες μετρικές αξιολόγησης και για τα περισσότερα μοντέλα. Επιπλέον, και στα δύο σύνολα δεδομένων παρατηρείται ότι οι τιμές των μετρικών βελτιώνονται σταδιακά όσο αυξάνεται ο βαθμός γλωσσικής καθοδήγησης των προτροπών, δηλαδή από την απουσία καθοδήγησης (no guidance) έως την πλήρως καθοδηγούμενη προσέγγιση LIP.

Η τάση αυτή αναδεικνύει τη σημασία της γλωσσικής καθοδήγησης (linguistic guidance) για την αποτελεσματική αξιοποίηση των LLMs στην επαλήθευση συγγραφικής ταυτότητας. Η ανάλυση των αποτελεσμάτων δείχνει ότι η επιτυχία των LLMs στην εργασία αυτή επηρεάζεται σημαντικά από το είδος και τον βαθμό καθοδήγησης που παρέχεται μέσω των προτροπών (prompts). Με άλλα λόγια, όσο πιο συγκεκριμένες και γλωσσολογικά τεκμηριωμένες είναι οι οδηγίες που δίνονται στο μοντέλο, τόσο καλύτερα αυτό μπορεί να εντοπίσει τα υφολογικά χαρακτηριστικά που διαφοροποιούν ή συνδέουν τους συγγραφείς, οδηγώντας σε ακριβέστερες αποφάσεις σχετικά με τη συγγραφική ταυτότητα.



Εικόνα 29: Αριστερά: Αποτελέσματα επαλήθευσης συγγραφέα στο δείκτη ακρίβειας (Accuracy) στο Blog dataset. Δεξιά: Τα αποτελέσματα στο δείκτη ακρίβειας (Accuracy) στο Email dataset. Πηγή: [32]

Από την άλλη πλευρά, παρότι η ανάλυση συγγραφικής ταυτότητας με χρήση Μεγάλων Γλωσσικών Μοντέλων (LLMs) παρέχει έναν βαθμό ερμηνευσιμότητας μέσω των γλωσσικών χαρακτηριστικών και των παρατηρήσεων που παρουσιάζονται κατά τη διάρκεια της ανάλυσης, η έρευνα αυτή δεν εξετάζει την ερμηνεία του τρόπου με τον οποίο λαμβάνονται οι αποφάσεις στο επίπεδο των νευρώνων του μοντέλου.

Παρότι η μέθοδος LIP επιτρέπει στα Μεγάλα Γλωσσικά Μοντέλα (LLMs) να παρέχουν γλωσσικές αιτιολογήσεις για τις προβλέψεις τους, οι αιτιολογήσεις αυτές δεν εξασφαλίζουν απαραίτητα την πλήρη

ερμηνευσιμότητα της διαδικασίας λήψης αποφάσεων. Συγκεκριμένα, τα μοντέλα μπορούν να αναφέρουν γλωσσικά και υφολογικά χαρακτηριστικά που θεωρούν σημαντικά για την επαλήθευση συγγραφικής ταυτότητας, χωρίς ωστόσο να είναι βέβαιο ότι τα χαρακτηριστικά αυτά αντικατοπτρίζουν πιστά τους εσωτερικούς υπολογιστικούς μηχανισμούς που οδήγησαν στην τελική πρόβλεψη.

Επομένως, παρά τη δυνατότητα παραγωγής κατανοητών αιτιολογήσεων, η εσωτερική λειτουργία των LLMs εξακολουθεί να χαρακτηρίζεται από περιορισμένη ερμηνευσιμότητα (black-box nature). Το γεγονός αυτό αποτελεί σημαντικό περιορισμό, ιδιαίτερα σε εφαρμογές όπου απαιτείται υψηλός βαθμός διαφάνειας και αξιοπιστίας των αποφάσεων.

4.6.2 Learning Universal Authorship Representations (LUAR)

Οι συγγραφείς [33] προτείνουν τον συνδυασμό μιας αρχιτεκτονικής βασισμένης στον μηχανισμό attention με μια contrastive συνάρτηση εκπαίδευσης (contrastive training objective).

Στόχος είναι να μάθουν μια συνάρτηση f που αντιστοιχίζει μια συλλογή εγγράφων στον χώρο R^{512}, έτσι ώστε οποιεσδήποτε δύο συλλογές που έχουν γραφτεί από τον ίδιο συγγραφέα να παρουσιάζουν μεγαλύτερη ομοιότητα συνημιτόνου (cosine similarity) από δύο συλλογές που έχουν γραφτεί από διαφορετικούς συγγραφείς.

Για την εκπαίδευση της συνάρτησης f χρησιμοποιείται contrastive learning.

Η αντιθετική μάθηση (contrastive learning) αποτελεί μια μέθοδο εκπαίδευσης κατά την οποία το μοντέλο μαθαίνει να διακρίνει μεταξύ θετικών και αρνητικών παραδειγμάτων. Στην περίπτωση του LUAR, ως θετικά παραδείγματα θεωρούνται οι συλλογές κειμένων που προέρχονται από τον ίδιο συγγραφέα, ενώ ως αρνητικά παραδείγματα θεωρούνται οι συλλογές που προέρχονται από διαφορετικούς συγγραφείς. Κατά την εκπαίδευση, το μοντέλο επιδιώκει να αυξήσει την ομοιότητα μεταξύ των αναπαραστάσεων του ίδιου συγγραφέα και να μειώσει την ομοιότητα μεταξύ αναπαραστάσεων διαφορετικών συγγραφέων.

Η contrastive learning λειτουργεί ως εξής.

Δεδομένου μιας μικρής παρτίδας (mini-batch) συλλογών εγγράφων x_i, οι οποίες προσδιορίζονται από ένα σύνολο δεικτών I, συμβολίζεται z_i η L2 κανονικοποίηση της αναπαράστασης $f(x_i)$ και μεγιστοποιείται η ακόλουθη συνάρτηση στόχου:

$$\sum_{i \in I} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k \in I \setminus \{i\}} \exp(z_i \cdot z_k / \tau)}$$

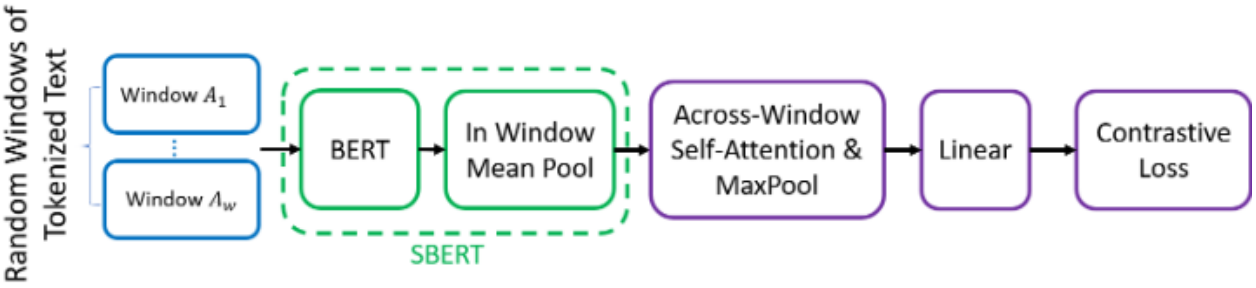
όπου:

- $P(i) \subseteq I$ είναι το σύνολο των δεικτών που αντιστοιχούν σε συλλογές του ίδιου συγγραφέα με τη συλλογή x_i,
- τ είναι μια παράμετρος θερμοκρασίας (temperature parameter), την οποία ορίζεται ίση με 0,01.

Η συγκεκριμένη διατύπωση περιλαμβάνει ως ειδικές περιπτώσεις αρκετές γνωστές συναρτήσεις απώλειας της metric learning, όπως η triplet loss.

Προτείνουν το μοντέλο επαλήθευσης συγγραφέα που βασίζεται στον μηχανισμό attention και απεικονίζεται στην εικόνα 30. Το μοντέλο αποτελεί επέκταση του SBERT (Sentence-BERT), το οποίο αναπτύχθηκε αρχικά για την εκμάθηση σημασιολογικών αναπαραστάσεων προτάσεων.

Η παρακάτω περιγραφή υποθέτει ότι η είσοδος αποτελείται από μια συλλογή σύντομων εγγράφων. Εάν αντίθετα η είσοδος είναι ένα μεγάλο έγγραφο, το αντιμετωπίζουμε ως συλλογή εγγράφων χωρίζοντάς το σε παραγράφους.



Εικόνα 30: Αρχιτεκτονική LUAR. Πηγή: [33]

Η διαδικασία λειτουργεί ως εξής:

Κείμενα συγγραφέα → BERT → SBERT pooling → Self-Attention μεταξύ αποσπασμάτων → Max Pooling → Διάνυσμα 512 διαστάσεων → Contrastive Learning

Το μοντέλο μαθαίνει έτσι ένα embedding συγγραφικού ύφους 512 διαστάσεων, όπου:

- κείμενα του ίδιου συγγραφέα βρίσκονται κοντά μεταξύ τους,
- κείμενα διαφορετικών συγγραφέων βρίσκονται μακριά μεταξύ τους,

με βάση την cosine similarity.

Οι συγγραφείς διαπιστώνουν ότι η μέθοδος του SBERT, η οποία υπολογίζει έναν σταθμισμένο μέσο όρο μέσω attention των κρυφών ενεργοποιήσεων του BERT, παράγει καλύτερες αναπαραστάσεις συγγραφικού ύφους (authorship embeddings).

<u>Αποτελέσματα:</u>

Αξιολογούν την απόδοση των μοντέλων χρησιμοποιώντας δύο καθιερωμένες μετρικές ανάκτησης πληροφορίας (retrieval metrics).

Η πρώτη είναι η Recall@8 (R@8), η οποία εκφράζει την πιθανότητα, αφού ταξινομηθούν όλα τα targets με βάση τη συνημιτονική ομοιότητά (cosine similarity) τους προς ένα τυχαία επιλεγμένο query, το μοναδικό σωστό target να βρίσκεται μέσα στις οκτώ πρώτες θέσεις της κατάταξης.

Η δεύτερη μετρική είναι η Mean Reciprocal Rank (MRR), η οποία δεν απαιτεί τον καθορισμό κάποιου σταθερού ορίου (threshold). Αντίθετα, λαμβάνει υπόψη τη θέση στην οποία εμφανίζεται το σωστό αποτέλεσμα στην κατάταξη.

Και στις δύο μετρικές, όσο μεγαλύτερη είναι η τιμή τους, τόσο καλύτερη θεωρείται η απόδοση του μοντέλου.

			Training Dataset					
			Reddit		Amazon		Fanfic	
			R@8	MRR	R@8	MRR	R@8	MRR
Evaluation Dataset	Reddit	P	65.61	50.37	23.72	15.43	12.10	7.64
		C	56.32	42.38	6.30	9.70	5.74	3.90
		T	10.34	6.77	7.65	5.03	6.97	4.63
	Amazon	P	68.91	55.59	82.54	68.99	28.91	20.06
		C	60.20	47.60	74.30	60.60	34.90	25.90
		T	43.70	35.50	31.61	24.86	21.46	16.45
	Fanfic	P	41.58	30.61	36.35	26.45	50.89	41.20
		C	40.66	30.98	24.99	17.98	47.98	39.02
		T	25.22	18.72	26.04	19.37	31.37	22.53

Πίνακας 14: Αποτελέσματα zero-shot για το προτεινόμενο μοντέλο, ένα συνελεκτικό μοντέλο και ένα μοντέλο TF-IDF. Πηγή: [33]

Υπάρχουν μεγάλες διαφορές στην ικανότητα μεταφοράς γνώσης των τριών μοντέλων.

Συγκεκριμένα, το μοντέλο που εκπαιδεύτηκε στα δεδομένα του Reddit και αξιολογήθηκε στα άλλα δύο σύνολα δεδομένων πέτυχε πάνω από το 80% της επίδοσης R@8 των αντίστοιχων μοντέλων που είχαν εκπαιδευτεί αποκλειστικά στο ίδιο πεδίο. Το αποτέλεσμα αυτό είναι ιδιαίτερα αξιοσημείωτο, δεδομένου ότι πρόκειται για σενάριο zero-shot, όπου το μοντέλο δεν έχει δει κανένα δείγμα από το πεδίο-στόχο κατά την εκπαίδευση.

Αντίθετα, τα μοντέλα που εκπαιδεύτηκαν στα δεδομένα του Amazon ή του fanfiction παρουσιάζουν σταθερά χαμηλότερη απόδοση από το μοντέλο του Reddit όταν αξιολογούνται σε διαφορετικά πεδία εφαρμογής (out-of-domain evaluation).

Παρατηρείται, επίσης, ότι το προτεινόμενο μοντέλο υπερτερεί του συνελικτικού μοντέλου αναφοράς (CNN baseline) ως προς τη μετρική R@8 σε όλες σχεδόν τις περιπτώσεις, με εξαίρεση μία μόνο περίπτωση.

Με στόχο τη βελτίωση της μεταφοράς γνώσης, οι συγγραφείς εξετάζουν τεχνικές που εμποδίζουν το μοντέλο να μάθει τις ιδιαιτερότητες του πεδίου εκπαίδευσης.

Συγκεκριμένα, προτείνουν μια απλή και ανεξάρτητη από το πεδίο εφαρμογής (domain-agnostic) τεχνική επαύξησης δεδομένων (data augmentation), η οποία βασίζεται στην τυχαία απόρριψη (dropout) υπολέξεων (subwords).

Η βασική ιδέα είναι ότι, εάν τα χαρακτηριστικά που σχετίζονται αποκλειστικά με το συγκεκριμένο πεδίο αποκρύπτονται αρκετά συχνά κατά την εκπαίδευση, τότε το μοντέλο θα αναγκαστεί να βασιστεί περισσότερο σε γενικότερα χαρακτηριστικά του συγγραφικού ύφους, βελτιώνοντας έτσι την ικανότητά του να μεταφέρεται σε άλλα πεδία.

Αν και οι συγγραφείς δεν εφάρμοσαν την τεχνική αυτή σε άλλα σύνολα δεδομένων ως πεδία-πηγή, εκτιμούν ότι η τυχαία απόκρυψη θα είναι επίσης αποτελεσματική στο σενάριο εκπαίδευσης σε πολλαπλά πεδία (multi-domain training).

Τέλος, πραγματοποίησαν και ένα επιπλέον πείραμα, στο οποίο αποκρύπτονταν ολόκληρες λέξεις αντί για υπολέξεις.

Τα αποτελέσματα ήταν πολύ παρόμοια, γεγονός που δείχνει ότι η αποτελεσματικότητα της μεθόδου δεν εξαρτάται ιδιαίτερα από το αν η απόκρυψη εφαρμόζεται σε επίπεδο λέξης ή υπολέξης.

Μελλοντικά:

Οι συγγραφείς θεωρούν ότι μια ιδιαίτερα ενδιαφέρουσα κατεύθυνση για μελλοντική έρευνα είναι η ρητή αποδιαπλοκή (disentanglement) των αναπαραστάσεων που μαθαίνει το μοντέλο, έτσι ώστε να διαχωρίζεται:

- το περιεχόμενο (content) του κειμένου,
- από το συγγραφικό ύφος (style).

Μία πιθανή προσέγγιση για την επίτευξη αυτού του στόχου είναι η χρήση domain adversarial training objectives, τα οποία εκπαιδεύουν το μοντέλο ώστε να διατηρεί πληροφορίες σχετικές με το συγγραφικό ύφος, ενώ ταυτόχρονα να απορρίπτει πληροφορίες που σχετίζονται αποκλειστικά με το πεδίο εφαρμογής.

4.6.3 Residualized Similarity

Οι Peter Zeng et al. [34] πρότειναν την μέθοδο Residualized Similarity. Παρατήρησαν ότι η υπεύθυνη χρήση συστημάτων επαλήθευσης συγγραφέα απαιτεί όχι μόνο υψηλή ακρίβεια, αλλά και ερμηνεύσιμες λύσεις. Συγκεκριμένα, για να μπορούν τέτοια συστήματα να χρησιμοποιηθούν σε περιβάλλοντα όπου οι αποφάσεις έχουν πραγματικές συνέπειες, οι προβλέψεις τους πρέπει να μπορούν να εξηγηθούν μέσω ερμηνεύσιμων χαρακτηριστικών, τα οποία να είναι δυνατό να εντοπιστούν και να συσχετιστούν άμεσα με το αρχικό κείμενο (να είναι δηλαδή ανιχνεύσιμα - traceable). Ένα ανιχνεύσιμο χαρακτηριστικό (traceable feature) δεν μπορεί να είναι ένα χαρακτηριστικό του οποίου η τιμή επιδέχεται διαφορετικές ερμηνείες από λογικούς παρατηρητές μετά την εξέταση του κειμένου

Σε εφαρμογές με νομικές ή εγκληματολογικές συνέπειες (forensic linguistics), η απόφαση πρέπει να είναι τόσο ακριβής όσο και τεκμηριωμένη από συγκεκριμένα γλωσσικά στοιχεία του κειμένου.

Οι νευρωνικές μέθοδοι επιτυγχάνουν υψηλή ακρίβεια, όμως οι εσωτερικές τους αναπαραστάσεις δεν είναι άμεσα ερμηνεύσιμες. Επιπλέον, οι προβλέψεις των Μεγάλων Γλωσσικών Μοντέλων (LLMs) δεν μπορούν να εξηγηθούν με πιστότητα, ακόμη και όταν παρέχεται μια εξήγηση για μια πρόβλεψη, αυτή δεν αντικατοπτρίζει την πραγματική διαδικασία συλλογισμού που οδήγησε το μοντέλο στην πρόβλεψή του.

Για να αντιμετωπιστεί αυτό το κενό, προτείνουν τη Residualized Similarity (RS), μια νέα μέθοδο που ενισχύει τα συστήματα τα οποία βασίζονται σε ερμηνεύσιμα χαρακτηριστικά, προσθέτοντας ένα νευρωνικό δίκτυο με σκοπό τη βελτίωση της απόδοσής τους, χωρίς να θυσιάζεται η ερμηνευσιμότητα.

Η βασική ιδέα της προτεινόμενης μεθόδου είναι η χρήση ενός νευρωνικού μοντέλου για την πρόβλεψη του υπολείμματος (residual), το οποίο ορίζεται ως η διαφορά μεταξύ της πραγματικής ετικέτας του ζεύγους εγγράφων, με τιμή +1 για τον ίδιο συγγραφέα και −1 για διαφορετικούς συγγραφείς, και της συνημιτονικής ομοιότητας που υπολογίζεται από το ερμηνεύσιμο σύστημα Gram2vec.

4.6.3.1 Αρχιτεκτονική

Για κάθε ζεύγος εγγράφων, ένα σύστημα υπολογίζει έναν βαθμό ομοιότητας. Αν η τιμή αυτή υπερβαίνει ένα προκαθορισμένο κατώφλι (threshold), συμπεραίνεται ότι τα δύο έγγραφα έχουν γραφτεί από τον ίδιο συγγραφέα. Αντίθετα, όταν η τιμή βρίσκεται κάτω από το κατώφλι, θεωρείται ότι τα έγγραφα προέρχονται από διαφορετικούς συγγραφείς.

Οι συγγραφείς ξεκινάνε με ένα ερμηνεύσιμο σύστημα, το οποίο χρησιμοποιεί ερμηνεύσιμα και ανιχνεύσιμα χαρακτηριστικά ως αρχικό εκτιμητή της ομοιότητας. Στη συνέχεια, εκπαιδεύουν ένα νευρωνικό μοντέλο ως προβλεπτή του υπολείμματος (residual predictor), το οποίο προβλέπει το σφάλμα ή, ισοδύναμα, τη διόρθωση που πρέπει να εφαρμοστεί στη βαθμολογία ομοιότητας του ερμηνεύσιμου συστήματος.

Η τελική πρόβλεψη προκύπτει απλώς από το άθροισμα της βαθμολογίας ομοιότητας του ερμηνεύσιμου μοντέλου και του προβλεπόμενου υπολείμματος, δηλαδή μιας προσαρμογής της ομοιότητας που πραγματοποιείται από το νευρωνικό μοντέλο.

Το συνδυαστικό αυτό σύστημα επιτυγχάνει τον επιθυμητό συμβιβασμό μεταξύ ερμηνευσιμότητας και ακρίβειας:

1. Όταν το ερμηνεύσιμο μοντέλο είναι πιθανό να έχει δώσει σωστή πρόβλεψη, το υπόλειμμα αναμένεται να είναι μικρό. Έτσι, διατηρούνται η ερμηνευσιμότητα και η πιστότητα της εξήγησης χωρίς να θυσιάζεται η ακρίβεια.
2. Όταν το ερμηνεύσιμο μοντέλο είναι πιθανό να έχει σφάλει, το υπόλειμμα παρέχει την απαραίτητη διόρθωση. Με αυτόν τον τρόπο βελτιώνεται η ακρίβεια της πρόβλεψης, ενώ η ερμηνευσιμότητα μειώνεται μόνο στον βαθμό που είναι αναγκαίο για τη διόρθωση του σφάλματος.

Ως σύστημα ερμηνεύσιμων χαρακτηριστικών χρησιμοποιούν το Gram2vec, το οποίο καταγράφει τις κανονικοποιημένες συχνότητες μορφολογικών και συντακτικών χαρακτηριστικών των κειμένων εισόδου.

Αξιολογούν την προσέγγιση Residualized Similarity (RS) συνδυάζοντας το Gram2vec με το LUAR, ένα από τα πλέον σύγχρονα νευρωνικά μοντέλα για την επαλήθευση συγγραφέα. Τα πειραματικά αποτελέσματα δείχνουν ότι η RS μπορεί να επιτύχει απόδοση αντίστοιχη με εκείνη της χρήσης μόνο του LUAR, ενώ ταυτόχρονα παρέχει πιστές εξηγήσεις βασισμένες σε ερμηνεύσιμα και ανιχνεύσιμα χαρακτηριστικά.

Εισάγουν τον δείκτη «Εμπιστοσύνης Ερμηνευσιμότητας» (“Interpretability Confidence” - IC), ο οποίος μετρά σε ποιο βαθμό το ερμηνεύσιμο σύστημα συμμετέχει στη λήψη της τελικής απόφασης για κάθε είσοδο.

Το σύστημα δεν αρκεί να παράγει μόνο μια τελική απόφαση· απαιτείται επιπλέον να παρέχει έναν ρητό δείκτη εμπιστοσύνης ως προς την ερμηνευσιμότητα (interpretability confidence), ο οποίος αποτυπώνει σε ποιον βαθμό η τελική απόφαση μπορεί να αποδοθεί στα ανιχνεύσιμα και ερμηνεύσιμα χαρακτηριστικά που εξήχθησαν από το κείμενο.

Στο ένα άκρο του φάσματος βρίσκονται οι ταξινομητές που βασίζονται αποκλειστικά σε γλωσσικά χαρακτηριστικά, όπως το Gram2Vec. Τα συστήματα αυτά επιτυγχάνουν 100% εμπιστοσύνη ερμηνευσιμότητας, καθώς όλες οι αποφάσεις τους στηρίζονται σε ερμηνεύσιμα και ανιχνεύσιμα χαρακτηριστικά. Ωστόσο, η ακρίβειά τους είναι συνήθως χαμηλότερη.

Στο άλλο άκρο βρίσκονται οι ταξινομητές που βασίζονται σε embeddings, όπως το LUAR. Τα μοντέλα αυτά επιτυγχάνουν συνήθως υψηλότερη ακρίβεια, αλλά η εμπιστοσύνη ερμηνευσιμότητας είναι ουσιαστικά 0%, καθώς οι αποφάσεις τους δεν μπορούν να αποδοθούν άμεσα σε ερμηνεύσιμα χαρακτηριστικά του κειμένου.

Επομένως, ο στόχος είναι ο σχεδιασμός ενός συστήματος που να επιτυγχάνει ταυτόχρονα υψηλή ακρίβεια και υψηλό βαθμό εμπιστοσύνης ερμηνευσιμότητας, συνδυάζοντας τα πλεονεκτήματα των δύο αυτών προσεγγίσεων.

Για κάθε σύνολο εκπαίδευσης (train), επικύρωσης (development/dev) και δοκιμής (test), αρχικά παράγονται ερμηνεύσιμα διανύσματα χαρακτηριστικών για κάθε έγγραφο με τη χρήση του Gram2vec. Στη συνέχεια, προκειμένου να αντιμετωπιστούν οι διαφορές στη διακύμανση των χαρακτηριστικών, τα διανύσματα τυποποιούνται (standardized) μέσω z-score κανονικοποίησης, ξεχωριστά για κάθε χαρακτηριστικό και για κάθε αντίστοιχο σύνολο δεδομένων.

Έπειτα, υπολογίζεται η συνημιτονική ομοιότητα μεταξύ των τυποποιημένων διανυσμάτων κάθε ζεύγους εγγράφων.

Η πραγματική ετικέτα (ground truth label) λαμβάνει την τιμή 1, όταν τα δύο έγγραφα έχουν γραφτεί από τον ίδιο συγγραφέα, και −1, όταν έχουν γραφτεί από διαφορετικούς συγγραφείς.

Το μοντέλο RS εκπαιδεύεται, ώστε να προβλέπει ένα υπόλειμμα (residual), το οποίο ορίζεται ως:

$$r^{(i)} = y^{(i)} - sim\big(f, (d_1)f(d_2)\big)$$

όπου:

- y είναι η πραγματική ετικέτα,
- sim είναι η συνημιτονική ομοιότητα μεταξύ των δύο διανυσμάτων,

- d_1 και d_2 είναι τα δύο προς σύγκριση έγγραφα,
- f είναι η συνάρτηση του Gram2vec που μετατρέπει κάθε έγγραφο στο αντίστοιχο διάνυσμα χαρακτηριστικών.

Το παραπάνω μέγεθος ονομάζεται πραγματικό υπόλειμμα (ground truth residual).

Η εικόνα 31 παρουσιάζει λεπτομερώς τόσο τη διαδικασία εκπαίδευσης του μοντέλου RS όσο και τη λειτουργία του κατά τη φάση της εξαγωγής προβλέψεων (inference).

Η εκπαίδευση του RS ξεκινά με ζεύγη εγγράφων. Κάθε έγγραφο μετατρέπεται σε διανυσματική αναπαράσταση τόσο από το ερμηνεύσιμο σύστημα (Gram2vec) όσο και από το νευρωνικό μοντέλο που πρόκειται να προσαρμοστεί (fine-tuned). Ως αποτέλεσμα, για κάθε ζεύγος εγγράφων παράγονται συνολικά τέσσερα διανύσματα αναπαράστασης (embeddings).

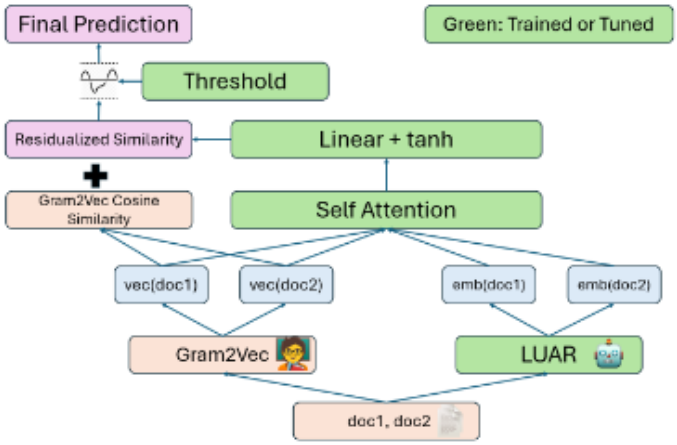
Στη συνέχεια, εφαρμόζεται ένα στρώμα προσοχής (attention layer) πάνω στα τέσσερα αυτά διανύσματα. Ο σκοπός του είναι να επιτρέψει στο μοντέλο RS να μάθει πόσο βάρος πρέπει να αποδίδει στα ερμηνεύσιμα χαρακτηριστικά και πόσο στις νευρωνικές αναπαραστάσεις κατά την πρόβλεψη του υπολείμματος.

Αξίζει να σημειωθεί ότι αυτό το στρώμα προσοχής χρησιμοποιείται μόνο κατά τη φάση της εκπαίδευσης.

Κατά τη φάση της εξαγωγής προβλέψεων (inference), όταν δοθεί ένα νέο ζεύγος εγγράφων, η τελική βαθμολογία ομοιότητας υπολογίζεται πολύ απλά ως το άθροισμα:

- της συνημιτονικής ομοιότητας που παράγεται από το ερμηνεύσιμο σύστημα (Gram2vec), και
- του προβλεπόμενου υπολείμματος (predicted residual) που υπολογίζει το νευρωνικό μοντέλο.

Με άλλα λόγια, το νευρωνικό μοντέλο δεν αντικαθιστά την απόφαση του ερμηνεύσιμου συστήματος, αλλά λειτουργεί ως ένας διορθωτικός μηχανισμός, ο οποίος προβλέπει μόνο το σφάλμα της αρχικής εκτίμησης και προσαρμόζει ανάλογα την τελική βαθμολογία ομοιότητας.



Εικόνα 31: Αρχιτεκτονική Residual Similarity. Τα πράσινα κουτιά δείχνουν ότι ενημερώνονται κατά την εκπαίδευση. Στην αριστερή πλευρά φαίνεται η μέθοδος κατά τη χρήση της πρόβλεψης. Το τελικό σκορ είναι απλώς το άθροισμα του συνημιτόνου ομοιότητας και του προβλεπόμενου υπολείμματος. Πηγή: [34]

Στόχος του μοντέλου είναι να προβλέψει την υπολειμματική ομοιότητα (residual similarity), η οποία συμβολίζεται με $\hat{r}^{(i)}$.

Η εκπαίδευση πραγματοποιείται με την ελαχιστοποίηση του Μέσου Τετραγωνικού Σφάλματος (Mean Squared Error – MSE) μεταξύ των προβλεπόμενων και των πραγματικών υπολειμμάτων:

$$L(B) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}^{(i)} - r^{(i)}\right)^2$$

Με άλλα λόγια, το μοντέλο μαθαίνει να εκτιμά όσο το δυνατόν ακριβέστερα το σφάλμα της πρόβλεψης που παράγεται από το ερμηνεύσιμο σύστημα.

Η έννοια της Εμπιστοσύνης Ερμηνευσιμότητας (Interpretability Confidence – INTCONF), αποτελεί έναν δείκτη του βαθμού στον οποίο μια συγκεκριμένη πρόβλεψη που παράγεται από τη μέθοδο Residualized Similarity (RS) μπορεί να θεωρηθεί ερμηνεύσιμη.

Η INTCONF ορίζεται με βάση την τελική πρόβλεψη του συστήματος.

Έστω ότι η βαθμολογία ομοιότητας μεταξύ δύο κειμένων, όπως υπολογίζεται αποκλειστικά από το Gram2vec, είναι sim_{g2v} .

Η INTCONF εκφράζει τη σχετική συνεισφορά του Gram2vec στη συνολική βαθμολογία ομοιότητας, η οποία αποτελείται από δύο μέρη:

- τη βαθμολογία που προέρχεται από το Gram2vec και
- το υπόλειμμα που προβλέπεται από το μοντέλο RS.

Οι συγγραφείς διακρίνουν δύο διαφορετικές περιπτώσεις.

Περίπτωση 1: Πρόβλεψη «ίδιος συγγραφέας»

Όταν το RS προβλέπει ότι τα δύο κείμενα έχουν γραφτεί από τον ίδιο συγγραφέα, η σωστή πρόβλεψη αντιστοιχεί σε υψηλή τιμή ομοιότητας.

Στην περίπτωση αυτή, η συνεισφορά του Gram2vec ορίζεται ως η απόστασή του από τη μικρότερη δυνατή βαθμολογία ομοιότητας, δηλαδή από την τιμή −1.

Επομένως, η συνεισφορά του Gram2vec ποσοτικοποιείται ως:

$$1 + simg2v$$

και διαιρείται με το άθροισμα:

- της συνεισφοράς του Gram2vec και
- της συνεισφοράς του νευρωνικού μέρους, η οποία ισούται με την απόλυτη τιμή του προβλεπόμενου υπολείμματος.

Έτσι προκύπτει:

$$\text{INTCONF} = \frac{1 + simg2v}{1 + simg2v + |\text{ predicted residual }|}$$

(ή ίση με 0, όταν ο παρονομαστής είναι μηδέν).

Περίπτωση 2: Πρόβλεψη «διαφορετικός συγγραφέας»

Όταν το σύστημα προβλέπει ότι τα δύο κείμενα έχουν γραφτεί από διαφορετικούς συγγραφείς, η επιθυμητή βαθμολογία είναι όσο το δυνατόν πιο κοντά στην ελάχιστη τιμή.

Στην περίπτωση αυτή, η συνεισφορά του Gram2vec ορίζεται ως η απόστασή του από τη μέγιστη δυνατή βαθμολογία ομοιότητας, δηλαδή από την τιμή +1.

Η INTCONF υπολογίζεται ως:

$$\text{INTCONF} = \frac{1 - simg2v}{1 - simg2v + |\text{ predicted residual }|}.$$

Η τιμή της INTCONF βρίσκεται πάντοτε στο διάστημα [0,1].

Αυτό σημαίνει ότι, αφού εκτελεστεί το μοντέλο RS, είναι δυνατόν να υπολογιστεί η Εμπιστοσύνη Ερμηνευσιμότητας για οποιοδήποτε συγκεκριμένο ζεύγος εγγράφων.

Οι συγγραφείς επισημαίνουν, επίσης, ότι σε ορισμένες περιπτώσεις το προβλεπόμενο υπόλειμμα μπορεί να αντιστρέψει (flip) την αρχική απόφαση του συστήματος που βασίζεται αποκλειστικά στο Gram2vec, μετατρέποντας μια πρόβλεψη «ίδιος συγγραφέας» σε «διαφορετικός συγγραφέας», ή το αντίστροφο.

Η δυνατότητα αυτή αποτελεί, άλλωστε, τον βασικό λόγο ύπαρξης του μοντέλου πρόβλεψης του υπολείμματος: να διορθώνει τα σφάλματα του ερμηνεύσιμου συστήματος.

Οι συγγραφείς τονίζουν, ωστόσο, ότι ακόμη και όταν η τελική πρόβλεψη αλλάζει λόγω της παρέμβασης του RS, το υποκείμενο ερμηνεύσιμο σύστημα εξακολουθεί να είναι χρήσιμο για την παροχή εξηγήσεων.

4.6.3.2 Αποτελέσματα

Οι συγγραφείς αξιολογούν τη μέθοδο Residualized Similarity (RS) συγκρίνοντάς την τόσο με το Gram2vec όσο και με τα νευρωνικά μοντέλα, χρησιμοποιώντας ως μετρική το εμβαδόν κάτω από την καμπύλη ROC (Receiver Operating Characteristic Area Under the Curve – AUC).

Τα αποτελέσματα της αξιολόγησης του συστήματος Residualized Similarity (RS) παρουσιάζονται στον πίνακα 15.

	Reddit_{AUC}		Amazon_{AUC}		Fanfiction_{AUC}		Pikabu_{AUC}	
G2V (Gram2Vec)	0.63		0.71		0.69		0.65	
Neural Models	Neural	RS	Neural	RS	Neural	RS	Neural	RS
RoBERTa-Base	0.69	0.72	0.87	0.85	0.88	0.87	-	-
RoBERTa-Large	0.71	0.70	0.90	0.86	**0.91**	0.86	-	-
Longformer	0.74	0.71	0.88	0.85	0.89	0.86	-	-
LUAR/LUAR-RU	**0.84**	**0.80**	**0.91**	**0.90**	0.89	0.87	**0.74**	**0.76**
Style	0.77	0.76	0.83	0.83	0.72	0.78	-	-
all-mpnet-base-v2	0.66	0.62	0.86	0.84	0.90	0.85	-	-
mxbai-embed-large	0.72	0.69	0.88	0.85	0.87	**0.89**	-	-

Πίνακας 15: Απόδοση του μοντέλου στα 3 σύνολα δεδομένων (Reddit, Amazon, Fanfiction). Το Pikabu είναι ρώσικο dataset και χρησιμοποιούν LUAR-RU για το νευρωνικό μοντέλο στην RS. Τα καλύτερα αποτελέσματα είναι με bold. Πηγή: [34]

Στην πρώτη γραμμή του πίνακα εμφανίζονται οι επιδόσεις του Gram2vec όταν χρησιμοποιείται ως αυτόνομο σύστημα.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα επτά διαφορετικών νευρωνικών μοντέλων σε τρία αγγλόφωνα σύνολα δεδομένων, καθώς και ενός νευρωνικού μοντέλου σε ένα ρωσικό σύνολο δεδομένων.

Για κάθε συνδυασμό μοντέλου και συνόλου δεδομένων παρατίθενται δύο αποτελέσματα:

1) η απόδοση του νευρωνικού μοντέλου ως γραμμής βάσης (neural baseline),
2) η απόδοση του προτεινόμενου συστήματος Residualized Similarity (RS), το οποίο χρησιμοποιεί το ίδιο νευρωνικό μοντέλο για την πρόβλεψη του υπολείμματος.

Τα αποτελέσματα δείχνουν ότι, για όλους τους συνδυασμούς μοντέλων και συνόλων δεδομένων, το πλήρως μη ερμηνεύσιμο νευρωνικό σύστημα και το μερικώς ερμηνεύσιμο σύστημα RS παρουσιάζουν σχεδόν την ίδια απόδοση.

Επιπλέον, δεν προκύπτει κάποιο σαφές μοτίβο που να δείχνει ότι η απόδοση της RS επηρεάζεται συστηματικά είτε από το χρησιμοποιούμενο νευρωνικό μοντέλο είτε από το συγκεκριμένο σύνολο δεδομένων.

Αντίθετα, το Gram2vec εμφανίζει σταθερά χαμηλότερη απόδοση σε σχέση με όλα τα νευρωνικά μοντέλα αλλά και με όλες τις εκδόσεις του συστήματος RS, παρότι εξακολουθεί να υπερβαίνει τη τυχαία πρόβλεψη, η οποία αντιστοιχεί σε τιμή AUC = 0,5.

Παρατηρείται, επίσης, ότι το LUAR υπερέχει όλων των υπόλοιπων νευρωνικών μοντέλων στα σύνολα δεδομένων Reddit και Amazon, ενώ παρουσιάζει ιδιαίτερα ανταγωνιστική απόδοση και στο σύνολο Fanfiction.

Το αποτέλεσμα αυτό είναι αναμενόμενο, καθώς το LUAR έχει σχεδιαστεί και εκπαιδευτεί ειδικά για εργασίες ταυτοποίησης συγγραφέα (authorship attribution).

Επιπλέον, το LUAR εμφανίζει ιδιαίτερα υψηλή απόδοση στο σύνολο Reddit, γεγονός που επίσης ήταν αναμενόμενο, επειδή η αρχική του εκπαίδευση πραγματοποιήθηκε αποκλειστικά σε δεδομένα του Reddit.

Η απόδοση του συστήματος Residualized Similarity, όταν χρησιμοποιεί το LUAR ως μοντέλο πρόβλεψης του υπολείμματος, είναι σχεδόν ταυτόσημη με εκείνη του πλήρως νευρωνικού και μη ερμηνεύσιμου LUAR, ως προς τη μετρική AUC, σε όλα τα σύνολα δεδομένων.

Ειδικότερα:

- στο σύνολο Reddit, το RS παρουσιάζει ελαφρώς χαμηλότερη απόδοση,
- ενώ στο σύνολο Pikabu παρουσιάζει ελαφρώς καλύτερη απόδοση από το LUAR.

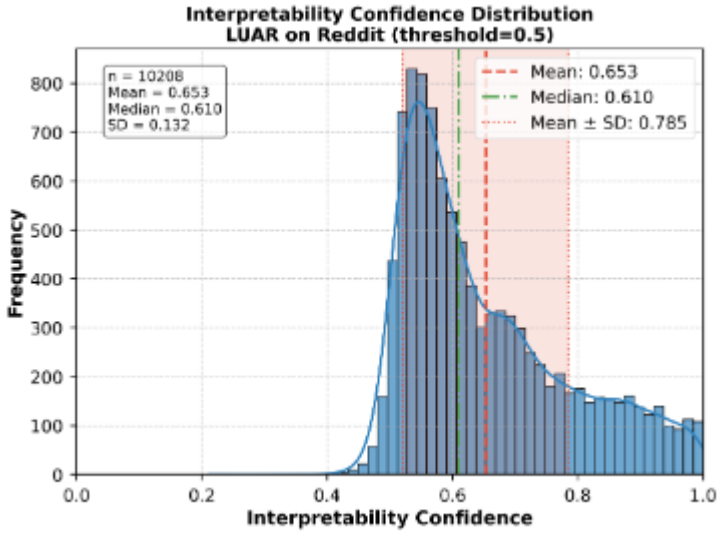
Παράλληλα, παρατηρείται σημαντική βελτίωση σε σχέση με τη χρήση μόνο του Gram2vec.

Η μεγαλύτερη βελτίωση καταγράφεται στο σύνολο δεδομένων Amazon, όπου η απόδοση αυξάνεται κατά 19 μονάδες AUC, γεγονός που δείχνει ότι η πρόβλεψη του υπολείμματος από το νευρωνικό μοντέλο διορθώνει αποτελεσματικά τις αδυναμίες του ερμηνεύσιμου συστήματος, χωρίς να θυσιάζει την ερμηνευσιμότητα των τελικών αποφάσεων.

Στην εικόνα 32 παρουσιάζεται η κατανομή των τιμών της Εμπιστοσύνης Ερμηνευσιμότητας (Interpretability Confidence – INTCONF) για τις προβλέψεις που παράγονται από το σύστημα Residualized Similarity (RS) όταν χρησιμοποιείται το LUAR ως νευρωνικό μοντέλο στο σύνολο δεδομένων Reddit.

Ένα αξιοσημείωτο εύρημα είναι ότι, στις περισσότερες περιπτώσεις, η τιμή της INTCONF είναι μεγαλύτερη από 0,5, ενώ η μέση τιμή της ανέρχεται σε 0,65.

Το αποτέλεσμα αυτό υποδηλώνει ότι το προτεινόμενο σύστημα Residualized Similarity διατηρεί σε σημαντικό βαθμό την ερμηνευσιμότητα που προσφέρει το Gram2vec, ενώ ταυτόχρονα επιτυγχάνει αισθητή βελτίωση στην απόδοση της επαλήθευσης συγγραφέα.



Εικόνα 32: Η κατανομή των τιμών της εμπιστοσύνης ερμηνευσιμότητας για τις προβλέψεις χρησιμοποιώντας Residual Similarity μαζί με LUAR στο σύνολο δεδομένων του Reddit. Πηγή: [34]

Τέλος, οι συγγραφείς διευκρινίζουν ότι η εργασία δεν προτείνει ένα πλήρως εξηγήσιμο (end-to-end explainable) σύστημα.

Αντίθετα, δείχνουν ότι το προτεινόμενο σύστημα Residualized Similarity (RS) μπορεί:

- να εντοπίζει μετρήσιμα και ερμηνεύσιμα γλωσσικά χαρακτηριστικά που πράγματι χρησιμοποιήθηκαν για τη λήψη της τελικής απόφασης, εξασφαλίζοντας έτσι την πιστότητα (faithfulness) της εξήγησης,
- και να ποσοτικοποιεί τον βαθμό στον οποίο τα χαρακτηριστικά αυτά εξηγούν γιατί το σύστημα κατέληξε στη συγκεκριμένη πρόβλεψη.

Με αυτόν τον τρόπο, το RS συνδυάζει την υψηλή απόδοση των νευρωνικών μοντέλων με τη δυνατότητα παροχής τεκμηριωμένων και επαληθεύσιμων εξηγήσεων, οι οποίες βασίζονται σε πραγματικά γλωσσικά χαρακτηριστικά του κειμένου και όχι σε αδιαφανείς εσωτερικές αναπαραστάσεις.

Τόσο στην περίπτωση όπου το RS διορθώνει μια λανθασμένη πρόβλεψη όσο και στην περίπτωση όπου το RS συμφωνεί με την αρχική πρόβλεψη του ερμηνεύσιμου συστήματος, αποδεικνύεται ότι:

- τα γλωσσικά χαρακτηριστικά που χρησιμοποιούνται είναι ουσιαστικά και ερμηνεύσιμα,
- ενώ μπορούν επίσης να εντοπιστούν απευθείας μέσα στο αρχικό κείμενο από το οποίο εξήχθησαν, γεγονός που ενισχύει την ιχνηλασιμότητα (traceability) και την πιστότητα (faithfulness) των εξηγήσεων.

Οι συγγραφείς θεωρούν ότι η προσέγγιση αυτή αποτελεί μια ιδιαίτερα ελπιδοφόρα κατεύθυνση για την ανάπτυξη συστημάτων Επεξεργασίας Φυσικής Γλώσσας (NLP) που είναι ταυτόχρονα αποτελεσματικά και ερμηνεύσιμα.

5 Πειράματα και τελικά αποτελέσματα

5.1 Μετρικές αξιολόγησης

Accuracy: Η ακρίβεια είναι μια μετρική αξιολόγησης που υπολογίζει το ποσοστό των σωστών προβλέψεων ενός μοντέλου σε σχέση με το συνολικό αριθμό των δειγμάτων. Υπολογίζεται αν διαιρέσουμε τις σωστές προβλέψεις με τον συνολικό αριθμό των προβλέψεων.

Precision: Η πιστότητα μετρά πόσο αξιόπιστες είναι οι θετικές προβλέψεις ενός μοντέλου. Το precision υπολογίζεται διαιρώντας τα σωστά θετικά αποτελέσματα (True Positives, TP) με το σύνολο των προβλέψεων που το μοντέλο χαρακτήρισε ως θετικές (True Positives + False Positives, TP + FP)

Recall: Η ανάκληση δείχνει πόσο καλά εντοπίζει ένα μοντέλο όλα τα πραγματικά θετικά παραδείγματα από ένα σύνολο δεδομένων. Η ανάκληση υπολογίζεται διαιρώντας τα σωστά θετικά αποτελέσματα (True Positives, TP) με το άθροισμα των σωστών θετικών και των ψευδών αρνητικών (True Positives + False Negatives, TP + FN).

F1-score: Η βαθμολογία F1 συνδυάζει την πιστότητα (precision) και την ανάκληση (recall) σε έναν ενιαίο αρμονικό μέσο. Ο τύπος για τον υπολογισμό της βαθμολογίας F1 είναι:

F1 = 2 * (πιστότητα * ανάκληση) / (πιστότητα + ανάκληση)

Η βαθμολογία F1 είναι ιδιαίτερα χρήσιμη σε περιπτώσεις όπου υπάρχει ανισορροπία στην κατανομή των κατηγοριών ή όταν το κόστος των ψευδώς θετικών και των ψευδώς αρνητικών δεν είναι ίσο.

AUROC: Το AUROC χρησιμοποιείται για τη μέτρηση της απόδοσης μοντέλων δυαδικής ταξινόμησης. Το AUROC μάς δείχνει πόσο καλά μπορεί το μοντέλο να ξεχωρίσει δύο κλάσεις μεταξύ τους (π.χ. αν ένας ασθενής έχει μια ασθένεια ή όχι, ή αν μια συναλλαγή είναι απάτη ή νόμιμη).

Καμπύλη ROC (Receiver Operating Characteristic): Είναι ένα διάγραμμα που απεικονίζει την απόδοση του μοντέλου σε όλα τα πιθανά όρια απόφασης (classification thresholds). Στον κάθετο άξονα (Y) σχεδιάζεται ο Ρυθμός Αληθώς Θετικών (True Positive Rate - TPR) και στον οριζόντιο άξονα (X) ο Ρυθμός Ψευδώς Θετικών (False Positive Rate - FPR).

AUC (Area Under the Curve): Είναι το συνολικό εμβαδόν κάτω από την καμπύλη ROC. Το AUC μετατρέπει ολόκληρη την καμπύλη σε έναν μοναδικό αριθμό από 0 έως 1, καθιστώντας εύκολη τη σύγκριση διαφορετικών μοντέλων

5.2 Περιγραφή των δεδομένων

Για την πειραματική αξιολόγηση των μεθόδων επαλήθευσης συγγραφέα χρησιμοποιήθηκε το σύνολο δεδομένων «twitter_1Kx1K», το οποίο αντλήθηκε από την πλατφόρμα Hugging Face[1]. Σύμφωνα με την περιγραφή του αρχικού συνόλου δεδομένων, αυτό περιλαμβάνει 1.000 χρήστες της πλατφόρμας Twitter και 1.000 tweets ανά χρήστη. Το σύνολο δεδομένων επιλέχθηκε επειδή περιλαμβάνει κείμενα πολλών διαφορετικών συγγραφέων, επιτρέποντας τη μελέτη των χαρακτηριστικών του συγγραφικού ύφους σε σύντομα μηνύματα κοινωνικών δικτύων.

Για τις ανάγκες της παρούσας εργασίας δεν χρησιμοποιήθηκε το σύνολο των αρχικών δεδομένων, αλλά δημιουργήθηκε ένα κατάλληλα επεξεργασμένο υποσύνολο. Αρχικά, πραγματοποιήθηκε καθαρισμός των κειμένων, ενώ η πληροφορία συναισθηματικής ανάλυσης (sentiment) αφαιρέθηκε, καθώς δεν σχετίζεται άμεσα με τον στόχο της επαλήθευσης συγγραφέα.

Στη συνέχεια, κάνουμε τα εξής παρακάτω βήματα για να ετοιμάσουμε το αρχείο με τα tweets:

1) Σβήνουμε τα κενά tweets.
2) Αφαιρούμε τελείως τα Retweets.
3) Αντικαταστούμε τα urls με την ετικέτα “URL”.
4) Αντικαταστούμε τις αναφορές σε άλλους χρήστες με το tag @USER.
5) Έπειτα, αντικαταστούμε μεγάλα κενά (whitespace) σε ένα κενό.
6) Αφαιρούμε tweets που έγιναν κενά.
7) Αφαιρούμε διπλότυπα

Έπειτα, επιλέχθηκαν τυχαία 100 συγγραφείς με 200 tweets ανά συγγραφέα, σχηματίζοντας ένα σύνολο 20.000 tweets.

[1] sjmeis. *Twitter_1Kx1K* [Dataset]. Hugging Face. https://huggingface.co/datasets/sjmeis/Twitter_1Kx1K

Στη συνέχεια, θα πρέπει να συνενώσουμε τα tweets, καθώς τα καθένα ξεχωριστά αποτελούν πολύ μικρά κείμενα και δεν μπορεί να αποτυπωθεί ξεκάθαρα το συγγραφικό ύφος. Το κάθε έγγραφο θα περιέχει μέσα 20 τυχαία tweets του συγγραφέα. Τα tweets ομαδοποιήθηκαν χωρίς να λαμβάνεται υπόψη η χρονολογική τους σειρά, με στόχο τη μείωση πιθανών χρονικών επιδράσεων στη σύγκριση του συγγραφικού ύφους.

Κάθε νέο tweet που συνενώνεται θα παίρνει την ετικέτα <new> για να ξεχωρίζουμε ότι ενώθηκε το ένα tweet με άλλο. Έτσι, ο κάθε συγγραφέας θα έχει τελικά 10 έγγραφα. Το γενικό σύνολο θα είναι 1.000 έγγραφα (100 authors x 10 documents).

Επίσης, εφαρμόζουμε ανωνυμοποίηση των συγγραφέων.

Τέλος, δημιουργούμε ζεύγη εγγράφων, με τυχαίο δείγμα. Τα 5 πρώτα θα είναι τα γνωστά κείμενα, και τα επόμενα 5 κείμενα θα είναι τα άγνωστα. Η στήλη label με 1 θα δείχνει τα γνωστά κείμενα (γραμμένα από τον ίδιο συγγραφέα) και με 0 τα άγνωστα κείμενα (γραμμένα από διαφορετικό συγγραφέα). Δηλαδή, για κάθε συγγραφέα θα έχουμε 5 x 5 = 25 ζεύγη κειμένων από γνωστό συγγραφέα και με 100 συγγραφείς έχουμε σύνολο 2.500 ζεύγη κειμένων (positive pairs) . Αντίστοιχα, θα έχουμε και 25 ζεύγη κειμένων από διαφορετικό συγγραφέα (2.500 negative pairs).

5.3 Χωρισμός των δεδομένων

Χωρίζουμε τα δεδομένα μας στις εξής 3 κατηγορίες:

1) Δεδομένα εκπαίδευσης (Train data): Είναι τα δεδομένα τα οποία χρησιμοποιούνται για την εκπαίδευση του μοντέλου μας, ώστε αυτό να μάθει τα χαρακτηριστικά που σχετίζονται με το συγγραφικό ύφος. Για το σύνολο εκπαίδευσης χρησιμοποιούμε 70 συγγραφείς. Όπως αναφέραμε προηγουμένως, για κάθε συγγραφέα δημιουργούνται 25 ζεύγη κειμένων «ίδιος συγγραφέας», με αποτέλεσμα να προκύπτουν 70 x 25 = 1.750 θετικά δείγματα. Αντίστοιχα, δημιουργούνται 1.750 αρνητικά ζεύγη (διαφορετικός συγγραφέας). Επομένως, το σύνολο εκπαίδευσης αποτελείται από 3.500 ζεύγη κειμένων.
2) Δεδομένα επικύρωσης (Validation data): Είναι τα δεδομένα τα οποία χρησιμοποιούνται για την ρύθμιση του μοντέλου και στην περίπτωσή μας για τον προσδιορισμό του κατάλληλου κατωφλίου (threshold), το οποίο στη συνέχεια εφαρμόζεται στα δεδομένα ελέγχου μας. Για το σύνολο επικύρωσης χρησιμοποιούμε 15 συγγραφείς και δημιουργούμε συνολικά 750 ζεύγη κειμένων. Τα ζεύγη κειμένων κατανέμονται ισόποσα σε 375 θετικά δείγματα (ίδιος συγγραφέας) και 375 αρνητικά δείγματα (διαφορετικός συγγραφέας).
3) Δεδομένα ελέγχου (Test data): Είναι τα δεδομένα τα οποία χρησιμοποιούνται αποκλειστικά για την τελική αξιολόγηση του μοντέλου μας, χωρίς να έχουν προηγουμένως χρησιμοποιηθεί για την εκπαίδευση ή την ρύθμισή του. Για τα δεδομένα ελέγχου χρησιμοποιούμε τους υπόλοιπους 15 συγγραφείς και δημιουργούμε, όπως και στο σύνολο επικύρωσης, 750 ζεύγη κειμένων, που αποτελούνται από 375 θετικά (ίδιος συγγραφέας) και 375 αρνητικά (διαφορετικός συγγραφέας) δείγματα.

Συνολικά, τα σύνολα δεδομένων μας αποτελούνται από 100 διαφορετικούς συγγραφείς και 5.000 ζεύγη κειμένων, τα οποία κατανέμονται σε 3.500 ζεύγη για την εκπαίδευση, σε 750 ζεύγη για την επικύρωση και 750 ζεύγη για τον τελικό έλεγχο. Διατηρείται ίσος αριθμός θετικών και αρνητικών ζευγών κειμένων, ώστε να εξασφαλίζεται η ισορροπία για τον αποτελεσματικό έλεγχο των μοντέλων μας για την επαλήθευση συγγραφέα.

5.4 Τελικά αποτελέσματα

Χρησιμοποιήσαμε 7 διαφορετικές μεθόδους για να τις συγκρίνουμε μεταξύ τους.

5.4.1 Μέθοδος baseline με TF-IDF

Ως βασική μέθοδος αναφοράς (baseline) για την επαλήθευση συγγραφέα χρησιμοποιήθηκε η μέθοδος TF-IDF (Term Frequency–Inverse Document Frequency), σε συνδυασμό με n-grams χαρακτήρων και την ομοιότητα συνημιτόνου (cosine similarity). Στόχος της μεθόδου ήταν η δημιουργία ενός απλού μοντέλου σύγκρισης, το οποίο θα αποτελούσε σημείο αναφοράς για την αξιολόγηση των πιο σύνθετων μεθόδων που εξετάστηκαν στην παρούσα εργασία.

Για την αναπαράσταση των κειμένων χρησιμοποιήθηκαν τετραγράμματα χαρακτήρων (character 4-grams, n=4). Κάθε τετράγραμμα αποτελείται από τέσσερις διαδοχικούς χαρακτήρες ενός κειμένου, συμπεριλαμβανομένων πιθανών κενών και σημείων στίξης. Η επιλογή χαρακτηριστικών σε επίπεδο χαρακτήρων επιτρέπει την αποτύπωση στοιχείων του συγγραφικού ύφους, όπως επαναλαμβανόμενα μοτίβα γραφής, ορθογραφικές συνήθειες και χρήση σημείων στίξης.

Για τη μετατροπή των κειμένων σε αριθμητικά διανύσματα χρησιμοποιήθηκε ο αλγόριθμος TF-IDF, ο οποίος αποδίδει βάρη στα τετραγράμματα ανάλογα με τη συχνότητα εμφάνισής τους σε ένα κείμενο και τη συχνότητά τους στο σύνολο των εγγράφων. Με αυτόν τον τρόπο, χαρακτηριστικά που εμφανίζονται συχνά σε ένα συγκεκριμένο κείμενο, αλλά είναι λιγότερο κοινά στο συνολικό σύνολο δεδομένων, αποκτούν μεγαλύτερη βαρύτητα.

Ο TF-IDF προσαρμόστηκε αποκλειστικά στα δεδομένα εκπαίδευσης, ώστε να αποφευχθεί η διαρροή πληροφοριών (data leakage) από τα σύνολα επικύρωσης και ελέγχου. Στη συνέχεια, τα κείμενα κάθε ζεύγους μετατράπηκαν σε διανύσματα TF-IDF και υπολογίστηκε η ομοιότητα συνημιτόνου μεταξύ τους. Η τιμή ομοιότητας χρησιμοποιήθηκε ως σκορ επαλήθευσης συγγραφέα, με υψηλότερες τιμές να υποδηλώνουν μεγαλύτερη ομοιότητα μεταξύ των δύο κειμένων.

Για τη μετατροπή του σκορ σε δυαδική απόφαση εφαρμόστηκε κατώφλι (threshold), το οποίο προσδιορίστηκε από το σύνολο επικύρωσης. Όταν η ομοιότητα μεταξύ δύο κειμένων ήταν μεγαλύτερη ή ίση από το επιλεγμένο κατώφλι, το ζεύγος ταξινομούνταν ως κείμενα του ίδιου συγγραφέα (Same Author), ενώ στην αντίθετη περίπτωση ταξινομούνταν ως κείμενα διαφορετικών συγγραφέων (Different Author).

Η μέθοδος TF-IDF δεν απαιτεί εκπαίδευση νευρωνικού μοντέλου ή πρόσθετη βελτιστοποίηση παραμέτρων ταξινόμησης. Η διαδικασία περιορίζεται στην προσαρμογή του TF-IDF στα δεδομένα εκπαίδευσης, στον υπολογισμό των ομοιοτήτων και στην επιλογή του κατάλληλου κατωφλίου απόφασης.

Το threshold θα το βρούμε από το σύνολο επικύρωσης. Θα δοκιμάσουμε όλα τα όρια μέχρι να βρούμε αυτό που δίνει καλύτερο F1 score.

Για F1 score = 0.92 το threshold είναι 0.2, οπότε θα χρησιμοποιήσουμε αυτό στο σύνολο ελέγχου.

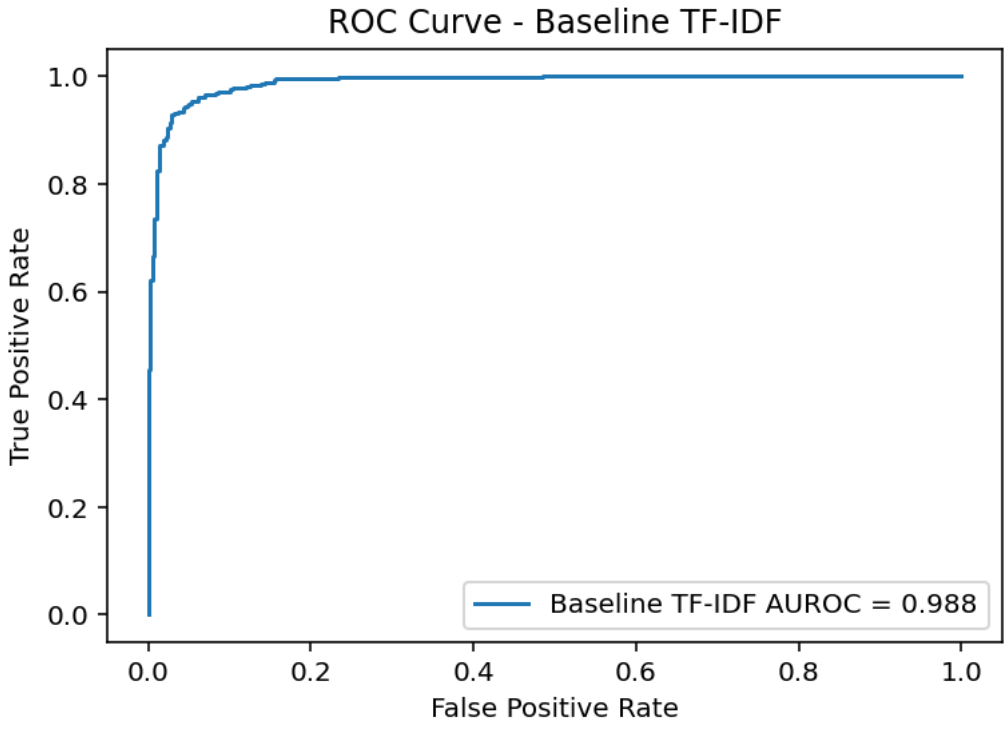
Τέλος, χρησιμοποιούμε τα metrics: “Accuracy”, “Precision”, “Recall”, “F1”, “AUROC” για τα τελικά αποτελέσματα. Τα ίδια μετρικά θα χρησιμοποιούν και στην συνέχεια.

Τα αποτελέσματα που παίρνουμε από το σύνολο ελέγχου είναι τα εξής:

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**
0.94	0.92	0.97	0.94	0.99

Τα αποτελέσματα δείχνουν ότι ακόμη και μία σχετικά απλή στυλομετρική προσέγγιση, βασισμένη σε τετραγράμματα χαρακτήρων, μπορεί να παρουσιάσει υψηλή απόδοση στην επαλήθευση συγγραφέα σε κείμενα κοινωνικών δικτύων.

Η καμπύλη Roc:



Εικόνα 33: Καμπύλη ROC για την μέθοδο TF-IDF.

5.4.2 General Impostors Method

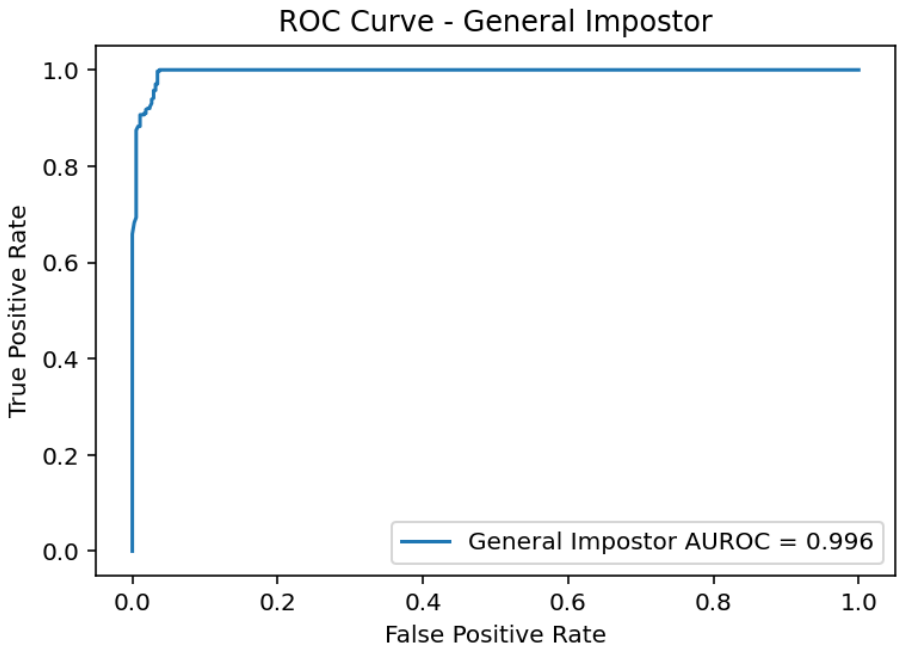
Για τους impostors χρησιμοποιούμε τους ίδιους authors που υπάρχουν στο αρχείο. Χρησιμοποιούμε διαφορετικά documents από αυτά που συγκρίνουμε. Στη συνέχεια, για την υλοποίηση της μεθόδου General Impostor χρησιμοποιήθηκε αναπαράσταση των κειμένων μέσω TF-IDF με τετραγράμματα χαρακτήρων (character n-grams, n=4).

Η μέθοδος βασίζεται στη σύγκριση της ομοιότητας μεταξύ των εξεταζόμενων κειμένων με την ομοιότητα που παρουσιάζει το άγνωστο κείμενο με κείμενα άλλων συγγραφέων (impostors). Για την πειραματική διαδικασία ορίστηκαν 100 επαναλήψεις (iterations), ενώ σε κάθε επανάληψη επιλέγονταν τυχαία 10 impostor documents και το 50% των διαθέσιμων χαρακτηριστικών (feature fraction = 0,5). Η τυχαία επιλογή πραγματοποιήθηκε με αρχική τιμή γεννήτριας τυχαίων αριθμών (random seed) ίση με 42, με σκοπό την αναπαραγωγιμότητα των πειραμάτων. Για κάθε επανάληψη εξεταζόταν εάν η ομοιότητα του γνωστού με το άγνωστο κείμενο ήταν μεγαλύτερη από τις αντίστοιχες ομοιότητες του άγνωστου κειμένου με τους επιλεγμένους impostors. Το τελικό σκορ προέκυπτε από το ποσοστό των επιτυχημένων επαναλήψεων, ενώ για την τελική απόφαση χρησιμοποιήθηκε κατώφλι 0,50, το οποίο προσδιορίστηκε από το σύνολο επικύρωσης.

Τα αποτελέσματα που προκύπτουν από το σύνολο ελέγχου είναι τα εξής:

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**
0.96	0.92	1	0.96	0.99

Η καμπύλη ROC:



Εικόνα 34: Καμπύλη ROC για την μέθοδο του General Impostor.

5.4.3 Extrinsic Instance-Based model

Στη μέθοδο Extrinsic Instance-Based (EIB), τα γνωστά κείμενα του υποψήφιου συγγραφέα αντιμετωπίζονται ως ξεχωριστά έγγραφα (instances), χωρίς να συγχωνεύονται σε ένα ενιαίο συγγραφικό προφίλ. Η μέθοδος βασίζεται στη σύγκριση της ομοιότητας μεταξύ ενός γνωστού και ενός άγνωστου κειμένου με τις αντίστοιχες ομοιότητες που παρουσιάζει το άγνωστο κείμενο με κείμενα άλλων συγγραφέων (impostors).

Για την αναπαράσταση των κειμένων χρησιμοποιήθηκε η μέθοδος TF-IDF με πενταγράμματα χαρακτήρων (character n-grams, n=5). Η μέθοδος εκπαιδεύτηκε αποκλειστικά στα δεδομένα εκπαίδευσης, με τις παραμέτρους ngram_range=(5,5), min_df=2 και lowercase=False. Για τον υπολογισμό της ομοιότητας μεταξύ των διανυσμάτων χρησιμοποιήθηκε η Min-Max similarity.

Αρχικά, δημιουργήθηκε ένα εξωτερικό σύνολο εγγράφων (Dexternal), το οποίο περιλάμβανε κείμενα συγγραφέων από το σύνολο εκπαίδευσης. Για κάθε εξεταζόμενο ζεύγος αποκλείονταν από αυτό το σύνολο οι συγγραφείς των δύο κειμένων, ώστε να μην επιλέγονται ως impostors έγγραφα των εξεταζόμενων συγγραφέων. Στη συνέχεια, υπολογιζόταν η ομοιότητα του γνωστού κειμένου με τα διαθέσιμα έγγραφα του εξωτερικού συνόλου και επιλέγονταν οι 50 περισσότερο όμοιοι impostors (hard impostors).

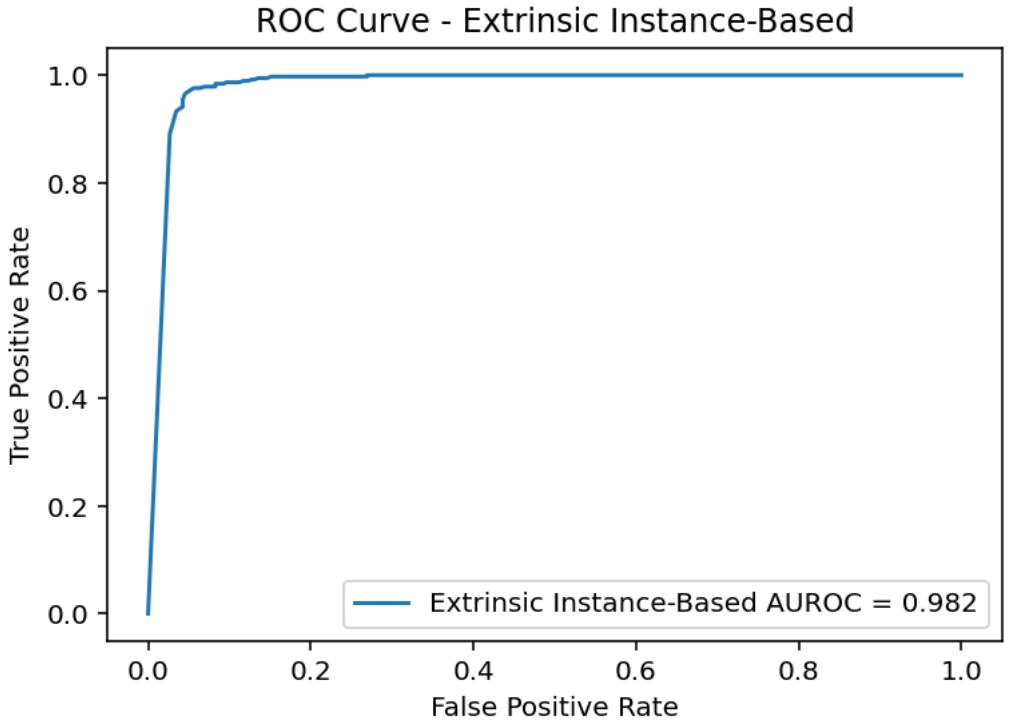
Η πειραματική διαδικασία πραγματοποιήθηκε με 10 επαναλήψεις (repetitions). Σε κάθε επανάληψη επιλέγονταν τυχαία 5 impostors από τους 50 που είχαν προηγουμένως επιλεγεί, καθώς και το 50% των διαθέσιμων χαρακτηριστικών (feature fraction = 0,5). Έπειτα, υπολογιζόταν η Min-Max similarity μεταξύ του γνωστού και του άγνωστου κειμένου, καθώς και μεταξύ του άγνωστου κειμένου και καθενός από τους 5 impostors. Οι έξι τιμές ομοιότητας ταξινομούνταν κατά φθίνουσα σειρά και καταγραφόταν η θέση κατάταξης της ομοιότητας του εξεταζόμενου ζεύγους. Το τελικό σκορ προέκυπτε από τον μέσο όρο των αντίστροφων θέσεων κατάταξης στις 10 επαναλήψεις. Για την αναπαραγωγιμότητα της διαδικασίας χρησιμοποιήθηκε αρχική τιμή τυχαίας γεννήτριας (random seed) ίση με 42, με διαφοροποίηση ανά εξεταζόμενο ζεύγος.

Το καταλληλότερο κατώφλι απόφασης προσδιορίστηκε από το σύνολο επικύρωσης, με κριτήριο τη μεγιστοποίηση του F1-score, και βρέθηκε ίσο με 0,86.

Τα αποτελέσματα από το σύνολο ελέγχου είναι τα παρακάτω:

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**
0.96	0.94	0.98	0.96	0.98

Η καμπύλη ROC:



Εικόνα 35: Καμπύλη ROC για την μέθοδο Extrinsic Instance-Based.

5.4.4 Extrinsic Profile-Based model

Στη μέθοδο Extrinsic Profile-Based (EPB), σε αντίθεση με την Instance-Based προσέγγιση, τα γνωστά κείμενα του υποψήφιου συγγραφέα συνενώνονται για τη δημιουργία ενός ενιαίου συγγραφικού προφίλ (known profile). Στην παρούσα εργασία, το προφίλ αυτό δημιουργήθηκε από 5 γνωστά έγγραφα του συγγραφέα, καθένα από τα οποία περιείχε 20 tweets. Στόχος της μεθόδου είναι να εξεταστεί εάν το άγνωστο κείμενο παρουσιάζει μεγαλύτερη ομοιότητα με το προφίλ του υποψήφιου συγγραφέα σε σύγκριση με τεχνητά προφίλ άλλων συγγραφέων (artificial impostor profiles).

Για την αναπαράσταση των κειμένων χρησιμοποιήθηκε TF-IDF με πενταγράμματα χαρακτήρων (character n-grams, n=5). Η ομοιότητα μεταξύ των διανυσματικών αναπαραστάσεων υπολογίστηκε μέσω της Min-Max similarity. Αρχικά, δημιουργήθηκε ένα εξωτερικό σύνολο εγγράφων (Dexternal), αποτελούμενο από κείμενα συγγραφέων του συνόλου εκπαίδευσης. Από αυτό αποκλείονταν οι συγγραφείς των εξεταζόμενων κειμένων και επιλέγονταν οι 50 περισσότερο συναφείς impostors (hard impostors), με βάση την ομοιότητά τους με τα γνωστά κείμενα του υποψήφιου συγγραφέα.

Η διαδικασία πραγματοποιήθηκε σε 10 επαναλήψεις (repetitions), με τυχαία επιλογή του 50% των διαθέσιμων χαρακτηριστικών σε κάθε επανάληψη (feature fraction = 0,5). Παράλληλα, δημιουργούνταν 5 τεχνητά προφίλ impostors ανά επανάληψη. Κάθε τεχνητό προφίλ σχηματιζόταν από 5 έγγραφα, τα οποία επιλέγονταν από το σύνολο των hard impostors, ώστε το πλήθος των εγγράφων του να αντιστοιχεί στο πλήθος των γνωστών εγγράφων του υποψήφιου συγγραφέα.

Στη συνέχεια, υπολογιζόταν η Min-Max similarity μεταξύ του γνωστού συγγραφικού προφίλ και του άγνωστου κειμένου, καθώς και μεταξύ καθενός από τα 5 τεχνητά προφίλ και του άγνωστου κειμένου. Οι τιμές ομοιότητας ταξινομούνταν κατά φθίνουσα σειρά και καταγραφόταν η θέση κατάταξης του γνωστού προφίλ. Το τελικό σκορ της μεθόδου προέκυπτε από τον μέσο όρο των αντίστροφων θέσεων κατάταξης στις 10 επαναλήψεις. Η τυχαία διαδικασία βασίστηκε σε καθορισμένη αρχική τιμή γεννήτριας τυχαίων αριθμών, ώστε να εξασφαλίζεται η αναπαραγωγιμότητα των πειραμάτων.

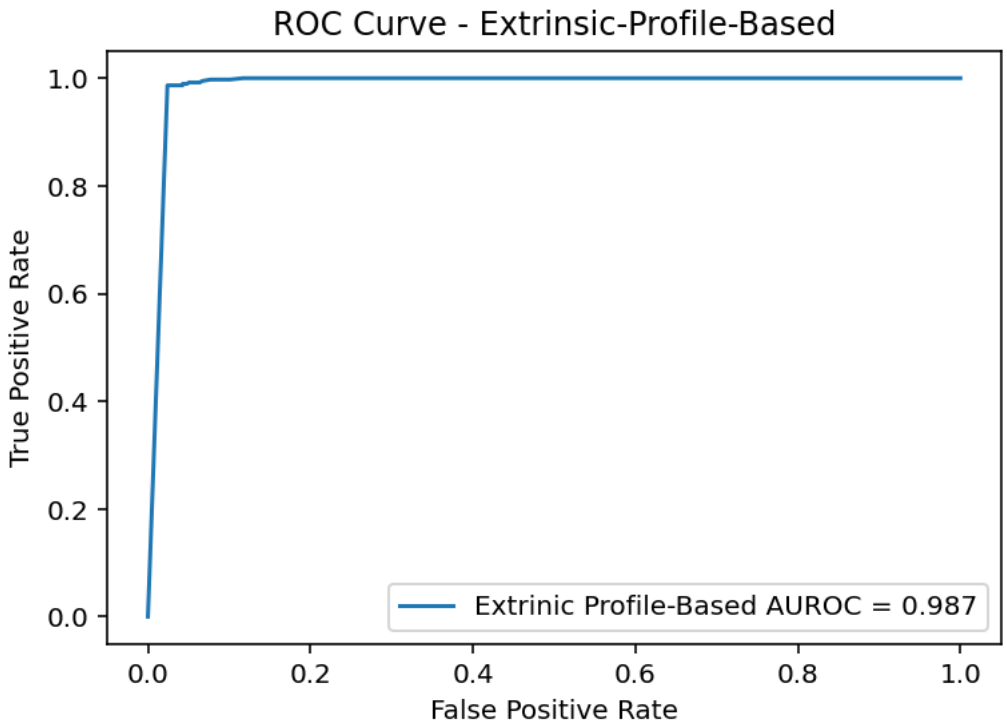
Από το σύνολο επικύρωσης προσδιορίστηκε ως καταλληλότερο κατώφλι απόφασης η τιμή 0,95.

Τα αποτελέσματα που πήραμε από το σύνολο ελέγχου είναι τα εξής:

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**
0.98	0.98	0.99	0.98	0.99

Τα αποτελέσματα αυτά δείχνουν ότι η δημιουργία ενός ενιαίου συγγραφικού προφίλ μπορεί να συμβάλει αποτελεσματικά στην αναγνώριση χαρακτηριστικών του συγγραφικού ύφους.

Η καμπύλη ROC:



Εικόνα 36: Καμπύλη ROC για την μέθοδο Extrinsic Profile-Based.

5.4.5 Frozen DeBERTa

Στην παρούσα εργασία χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο microsoft/deberta-v3-base[2], το οποίο βασίζεται στην αρχιτεκτονική Transformer. Το μοντέλο διατίθεται από την Microsoft μέσω της πλατφόρμας Hugging Face. Η μέθοδος εφαρμόστηκε με τη μορφή Siamese αρχιτεκτονικής, όπου δύο κείμενα εισάγονται ανεξάρτητα στον ίδιο κωδικοποιητή (encoder), με κοινά βάρη, προκειμένου να παραχθούν οι αντίστοιχες διανυσματικές αναπαραστάσεις τους.

Για την πειραματική διαδικασία επιλέχθηκε η προσέγγιση Frozen DeBERTa, δηλαδή η χρήση του προεκπαιδευμένου μοντέλου χωρίς πρόσθετη εκπαίδευση (fine-tuning) στα δεδομένα της εργασίας. Συγκεκριμένα, οι παράμετροι του κωδικοποιητή διατηρήθηκαν σταθερές (frozen weights), χωρίς να πραγματοποιείται ενημέρωσή τους κατά την επεξεργασία των κειμένων.

Αρχικά, κάθε κείμενο μετατράπηκε σε ακολουθία tokens μέσω του αντίστοιχου tokenizer του μοντέλου. Ως μέγιστο μήκος εισόδου ορίστηκαν τα 512 tokens (max_length=512), ενώ εφαρμόστηκαν padding και truncation, ώστε οι ακολουθίες να έχουν σταθερό μήκος. Για την επεξεργασία των δεδομένων χρησιμοποιήθηκε μέγεθος παρτίδας (batch_size) ίσο με 8.

Στη συνέχεια, κάθε κείμενο εισαγόταν στον προεκπαιδευμένο κωδικοποιητή DeBERTa, ο οποίος παρήγαγε διανυσματικές αναπαραστάσεις για τα tokens. Για τη δημιουργία μιας ενιαίας αναπαράστασης του κειμένου εφαρμόστηκε η τεχνική Mean Pooling, κατά την οποία υπολογίζεται ο μέσος όρος των διανυσμάτων των tokens, λαμβάνοντας υπόψη το attention mask. Έπειτα, εφαρμόστηκε κανονικοποίηση L2 (L2 normalization) στα παραγόμενα διανύσματα.

Η ομοιότητα μεταξύ των δύο κειμένων υπολογίστηκε μέσω της ομοιότητας συνημιτόνου των τελικών διανυσματικών αναπαραστάσεων. Για την αναπαραγωγιμότητα των πειραμάτων χρησιμοποιήθηκε αρχική τιμή τυχαίας γεννήτριας (random seed) ίση με 42.

Το μοντέλο το τρέξαμε στο google colab με σύστημα gpu A100.

Το καταλληλότερο κατώφλι απόφασης προσδιορίστηκε από το σύνολο επικύρωσης, με κριτήριο τη μεγιστοποίηση του F1-score, και βρέθηκε ίσο με 0,99.

Τα αποτελέσματα που πήραμε από το σύνολο ελέγχου φαίνονται στον παρακάτω πίνακα:

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**
0.87	0.82	0.96	0.88	0.97

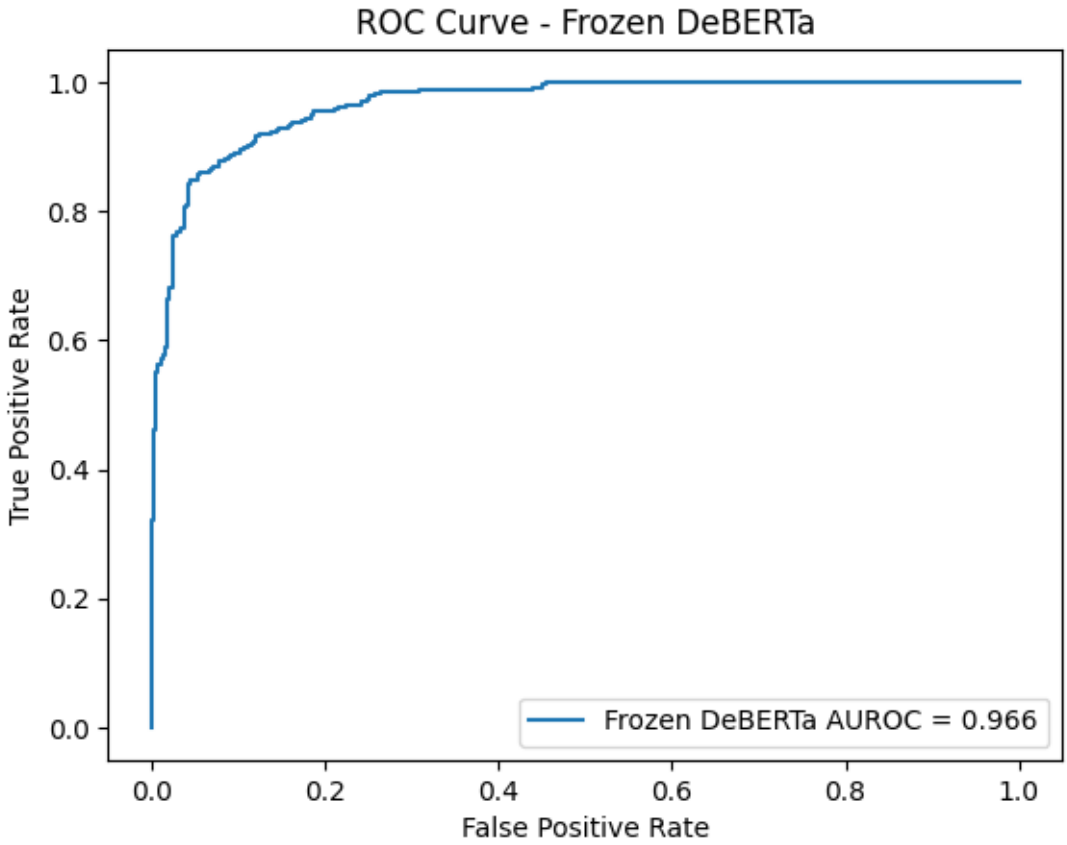
Τα αποτελέσματα δείχνουν ότι το προεκπαιδευμένο μοντέλο DeBERTa μπορεί να διακρίνει σε σημαντικό βαθμό κείμενα ίδιου και διαφορετικού συγγραφέα, ακόμη και χωρίς πρόσθετη εκπαίδευση. Ωστόσο, η

[2] Microsoft. *DeBERTa V3 BASE (Microsoft/deberta-v3-base)* [Προεκπαιδευμένο γλωσσικό μοντέλο]. Hugging Face. https://huggingface.co/microsoft/deberta-v3-base

υψηλή τιμή του κατωφλίου υποδηλώνει ότι το μοντέλο παράγει υψηλές τιμές ομοιότητας ακόμη και μεταξύ κειμένων διαφορετικών συγγραφέων, γεγονός που επηρεάζει την τελική ταξινόμηση.

Το Recall = 0.96 μαζί με Precision = 0.82 μας δείχνει ότι η μέθοδος είναι αρκετά «επιθετική» στο να αναγνωρίζει ζεύγη κυρίως «ίδιος συγγραφέας». Χάνει πολύ λίγες περιπτώσεις κειμένων «ίδιος συγγραφέας», αλλά κάνει περισσότερα false positives — δηλαδή ορισμένα κείμενα «διαφορετικός συγγραφέας» τα θεωρεί «ίδιος συγγραφέας».

Η καμπύλη ROC:



Εικόνα 37: Καμπύλη ROC για την μέθοδο Frozen DeBERTa.

5.4.6 Μέθοδος LUAR

Στην παρούσα εργασία χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο rrivera1849/LUAR-MUD[3], το οποίο έχει σχεδιαστεί για την παραγωγή διανυσματικών αναπαραστάσεων που αποτυπώνουν χαρακτηριστικά του συγγραφικού ύφους. Το μοντέλο διατίθεται μέσω της πλατφόρμας Hugging Face. Σε αντίθεση με γενικά γλωσσικά μοντέλα, το LUAR έχει εκπαιδευτεί ειδικά για την αναγνώριση συγγραφικών χαρακτηριστικών, αξιοποιώντας τεχνικές contrastive learning.

[3] rrivera1849. *LUAR MUD (rrivera1849/LUAR-MUD)* [Προεκπαιδευμένο γλωσσικό μοντέλο]. Hugging Face. https://huggingface.co/rrivera1849/LUAR-MUD

Η μέθοδος εφαρμόστηκε με προσέγγιση zero-shot, χωρίς πρόσθετη εκπαίδευση (fine-tuning) στα δεδομένα του Twitter που χρησιμοποιήθηκαν στην παρούσα εργασία. Το προεκπαιδευμένο μοντέλο χρησιμοποιήθηκε αποκλειστικά για την εξαγωγή διανυσματικών αναπαραστάσεων των κειμένων (embeddings).

Κάθε έγγραφο του συνόλου δεδομένων περιείχε 20 tweets του ίδιου συγγραφέα, τα οποία είχαν συνενωθεί με το διαχωριστικό <new>. Πριν από την εισαγωγή τους στο LUAR, τα έγγραφα διαχωρίζονταν ξανά στα επιμέρους tweets, ώστε να διατηρείται η δομή των 20 μηνυμάτων. Κάθε tweet μετατρεπόταν σε tokens μέσω του tokenizer του μοντέλου, με μέγιστο μήκος εισόδου 64 tokens ανά tweet (max_length=64). Εφαρμόστηκαν επίσης padding και truncation για την ομοιόμορφη επεξεργασία των ακολουθιών.

Στη συνέχεια, τα 20 tweets εισάγονταν ταυτόχρονα στο μοντέλο ως ένα σύνολο μηνυμάτων (episode), με διαστάσεις εισόδου (1, 20, 64). Το LUAR παρήγαγε μία ενιαία διανυσματική αναπαράσταση 512 διαστάσεων για κάθε έγγραφο, η οποία αποτύπωνε τα χαρακτηριστικά του συγγραφικού ύφους.

Για κάθε ζεύγος εγγράφων υπολογίστηκε η ομοιότητα συνημιτόνου μεταξύ των αντίστοιχων embeddings. Επιπλέον, εφαρμόστηκε μηχανισμός προσωρινής αποθήκευσης (embedding cache), ώστε κάθε μοναδικό έγγραφο να μετατρέπεται σε διάνυσμα μόνο μία φορά, μειώνοντας τον υπολογιστικό χρόνο της διαδικασίας. Στην υλοποίηση ορίστηκε επίσης παράμετρος batch_size=8, ενώ η εξαγωγή των embeddings πραγματοποιήθηκε χωρίς υπολογισμό παραγώγων (torch.no_grad()).

Τρέξαμε το μοντέλο LUAR, στο google colab (με το ίδιο runtime type με το Frozen DeBERTa).

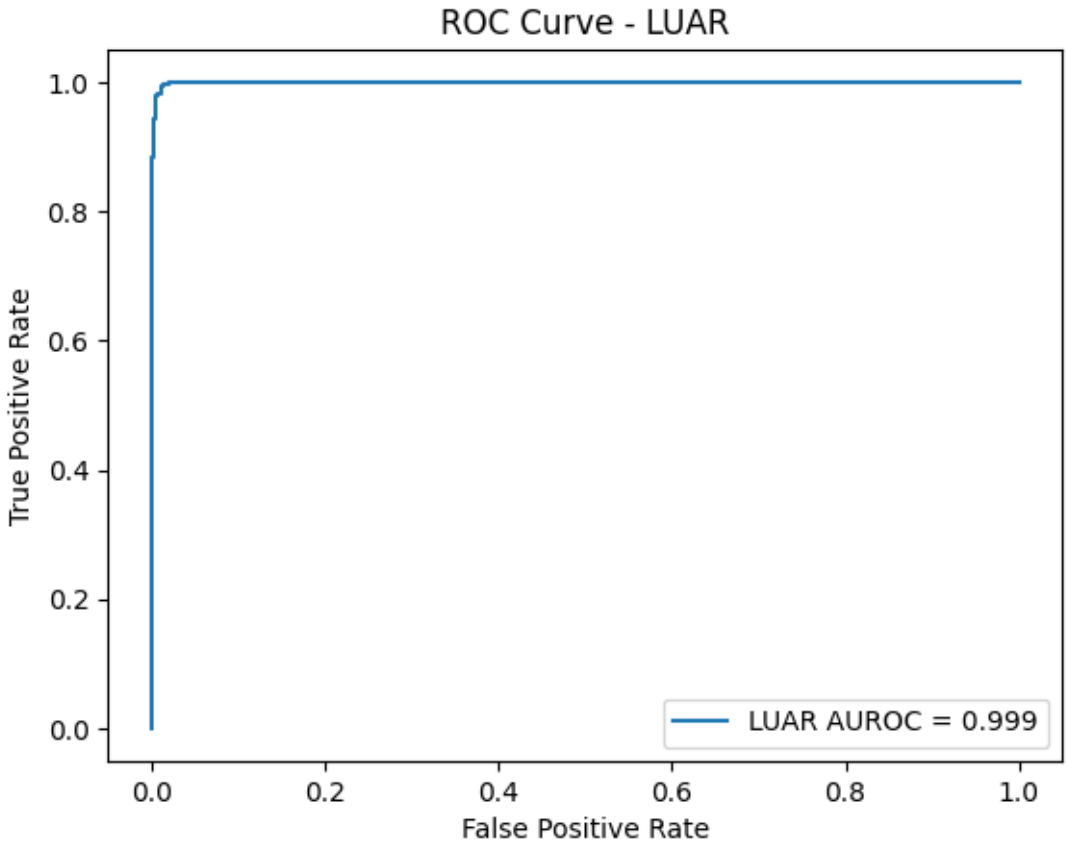
Το καταλληλότερο κατώφλι απόφασης προσδιορίστηκε από το σύνολο επικύρωσης, με κριτήριο τη μεγιστοποίηση του F1-score, και βρέθηκε ίσο με 0,89.

Με την εφαρμογή του στο σύνολο ελέγχου, πήραμε τα εξής αποτελέσματα:

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**
0.99	0.99	0.98	0.99	0.99

Η μέθοδος LUAR παρουσίασε την υψηλότερη ακρίβεια μεταξύ των επτά μεθόδων που αξιολογήθηκαν. Η υψηλή απόδοσή της χωρίς πρόσθετη εκπαίδευση στα δεδομένα της εργασίας αναδεικνύει την αποτελεσματικότητα των εξειδικευμένων αναπαραστάσεων συγγραφικού ύφους για την επαλήθευση συγγραφέα σε μηνύματα κοινωνικών δικτύων.

Η καμπύλη ROC:



Εικόνα 38: Καμπύλη ROC για την μέθοδο LUAR.

5.4.7 Μέθοδος Residualalized Similarity

Η τελευταία μέθοδος που εξετάστηκε στην παρούσα εργασία είναι η Residualized Similarity (RS), η οποία συνδυάζει παραδοσιακά γλωσσικά χαρακτηριστικά με νευρωνικές αναπαραστάσεις, με στόχο την αποτελεσματική και περισσότερο ερμηνεύσιμη επαλήθευση συγγραφέα. Η βασική ιδέα της μεθόδου είναι ότι η ομοιότητα μεταξύ δύο κειμένων μπορεί να υπολογιστεί αρχικά μέσω ερμηνεύσιμων στυλομετρικών χαρακτηριστικών και στη συνέχεια να διορθωθεί μέσω ενός νευρωνικού μοντέλου που προβλέπει την υπολειμματική διαφορά (residual).

Για την υλοποίηση της μεθόδου χρησιμοποιήθηκε το Gram2vec, το οποίο εξάγει γλωσσικά και στυλομετρικά χαρακτηριστικά από τα κείμενα και τα μετατρέπει σε αριθμητικές αναπαραστάσεις. Με βάση αυτές τις αναπαραστάσεις υπολογίστηκε η αρχική ομοιότητα μεταξύ των δύο κειμένων μέσω της ομοιότητας συνημιτόνου. Παράλληλα, χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο LUAR (rrivera1849/LUAR-MUD) για την παραγωγή νευρωνικών αναπαραστάσεων που αποτυπώνουν χαρακτηριστικά του συγγραφικού ύφους.

Οι αναπαραστάσεις των δύο κειμένων αξιοποιήθηκαν από ένα νευρωνικό μοντέλο με μηχανισμό self-attention, το οποίο εκπαιδεύτηκε να προβλέπει τη διορθωτική τιμή (residual) της αρχικής ομοιότητας Gram2vec. Για την εκπαίδευση χρησιμοποιήθηκαν οι πραγματικές ετικέτες των ζευγών, με τιμή +1 για κείμενα του ίδιου συγγραφέα και -1 για κείμενα διαφορετικών συγγραφέων. Η επιθυμητή διορθωτική τιμή ορίστηκε ως η διαφορά μεταξύ της πραγματικής ετικέτας και της αρχικής ομοιότητας Gram2vec. Η εκπαίδευση πραγματοποιήθηκε με συνάρτηση απώλειας Mean Squared Error (MSE).

Για την πειραματική διαδικασία χρησιμοποιήθηκε μέγιστο μήκος εισόδου 512 tokens (max_length=512) και μέγεθος παρτίδας (batch_size) ίσο με 8. Ως αλγόριθμος βελτιστοποίησης χρησιμοποιήθηκε ο AdamW, με ρυθμό μάθησης (learning_rate) ίσο με 5×10^{-5} και παράμετρο weight_decay=10^{-4}. Ο μέγιστος αριθμός εποχών εκπαίδευσης ορίστηκε σε 15 (epochs=15), ενώ εφαρμόστηκε μηχανισμός πρόωρης διακοπής (early stopping) με patience=3, ώστε η εκπαίδευση να σταματά όταν η απώλεια στο σύνολο επικύρωσης δεν παρουσιάζει βελτίωση για τρεις διαδοχικές εποχές. Επιπλέον, ενεργοποιήθηκε η τεχνική Low-Rank Adaptation (LoRA), με σκοπό την αποδοτικότερη προσαρμογή του προεκπαιδευμένου μοντέλου.

Το τελικό σκορ της μεθόδου υπολογίστηκε ως το άθροισμα της αρχικής ομοιότητας Gram2vec και της διορθωτικής τιμής που προέβλεπε το νευρωνικό μοντέλο. Με τον τρόπο αυτό, η τελική απόφαση βασίζεται τόσο σε ερμηνεύσιμα γλωσσικά χαρακτηριστικά όσο και σε πληροφορίες που προκύπτουν από τις νευρωνικές αναπαραστάσεις.

Το καταλληλότερο κατώφλι απόφασης προσδιορίστηκε από το σύνολο επικύρωσης και βρέθηκε ίσο με -0,59. Η αρνητική τιμή του κατωφλίου οφείλεται στο γεγονός ότι το τελικό σκορ περιλαμβάνει τη διορθωτική συνιστώσα του νευρωνικού μοντέλου και δεν περιορίζεται στο διάστημα [0,1].

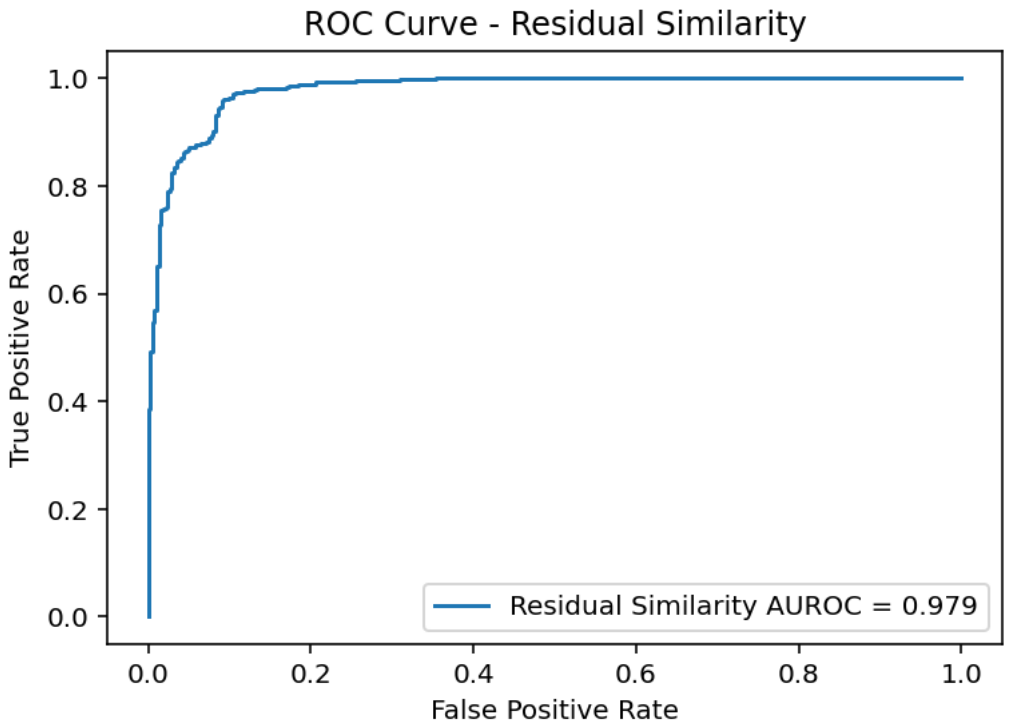
Η εφαρμογή της μεθόδου στο σύνολο ελέγχου έδωσε τα παρακάτω αποτελέσματα:

Accuracy	**Precision**	**Recall**	**F1**	**AUROC**	**Mean_INTCONF**
0.91	0.92	0.91	0.91	0.98	0.60

Τα αποτελέσματα δείχνουν ότι η Residualized Similarity παρουσιάζει υψηλή ικανότητα διάκρισης μεταξύ κειμένων ίδιου και διαφορετικού συγγραφέα, ενώ παράλληλα προσφέρει τη δυνατότητα ερμηνείας μέρους της τελικής πρόβλεψης μέσω των γλωσσικών χαρακτηριστικών του Gram2vec. Παρότι η ακρίβειά της ήταν χαμηλότερη από εκείνη της LUAR, η υψηλή τιμή AUROC αναδεικνύει την αποτελεσματικότητα του συνδυασμού στυλομετρικών και νευρωνικών χαρακτηριστικών.

Η μέση τιμή INTCONF μάς δείχνει ότι υπάρχει μέτριο επίπεδο ερμηνευσιμότητας των προβλέψεων.

Η καμπύλη ROC:



Εικόνα 39: Καμπύλη ROC για την μέθοδο Residualized Similarity.

Ο συγκεντρωτικός πίνακας με τα πειραματικά αποτελέσματα:

Model	Accuracy	Precision	Recall	F1-score	AUROC	Threshold	Mean_INTCONF
Baseline_TF-IDF	0.94	0.92	0.97	0.94	0.99	0.2	-
General Impostor	0.96	0.92	1	0.96	0.99	0.5	-
Extrinsic Instance-Based	0.96	0.94	0.98	0.96	0.98	0.86	-
Extrinsic Profile-Based	0.98	0.98	0.99	0.98	0.99	0.95	-
Frozen DeBERTa	0.87	0.82	0.96	0.88	0.97	0.99	-
LUAR	0.99	0.99	0.98	0.99	0.99	0.89	-
Residualized Similarity	0.91	0.92	0.91	0.91	0.98	-0.59	0.60

Πίνακας 16: Συγκεντρωτικός πίνακας με τα πειραματικά αποτελέσματα.

6 Συμπεράσματα

Στην παρούσα διπλωματική εργασία μελετήθηκε το πρόβλημα της επαλήθευσης συγγραφέα (Author Verification) σε κείμενα από την πλατφόρμα Twitter/X. Στόχος ήταν η διερεύνηση και η σύγκριση διαφορετικών μεθόδων, προκειμένου να διαπιστωθεί κατά πόσο μπορούν να αναγνωρίσουν εάν δύο κείμενα έχουν γραφτεί από τον ίδιο συγγραφέα. Για τον σκοπό αυτό δημιουργήθηκε ένα σύνολο δεδομένων από 100 συγγραφείς και εφαρμόστηκαν τόσο παραδοσιακές μέθοδοι, όπως η TF-IDF, η μέθοδος Impostors των Koppel & Winter και παραλλαγές της μεθόδου Potha & Stamatatos, όσο και πιο σύγχρονες προσεγγίσεις βασισμένες σε νευρωνικά δίκτυα, όπως οι DeBERTa, LUAR και Residualized Similarity.

Από τη συγκριτική αξιολόγηση των επτά μεθόδων προκύπτει ότι η αποτελεσματική επαλήθευση συγγραφέα δεν εξαρτάται αποκλειστικά από τη χρήση σύνθετων νευρωνικών μοντέλων. Η βασική μέθοδος TF-IDF πέτυχε Accuracy 0,94, αποδεικνύοντας ότι τα στυλομετρικά χαρακτηριστικά, όπως τα n-grams χαρακτήρων, μπορούν να αποτυπώσουν σημαντικές πληροφορίες για το προσωπικό ύφος γραφής ενός συγγραφέα. Παράλληλα, οι μέθοδοι Impostors παρουσίασαν ιδιαίτερα υψηλές επιδόσεις, με την παραλλαγή Extrinsic Profile-Based να επιτυγχάνει Accuracy 0,98 και F1-score 0,98. Το αποτέλεσμα αυτό υποδεικνύει ότι η δημιουργία ενός ενιαίου συγγραφικού προφίλ από πολλαπλά γνωστά κείμενα μπορεί να συμβάλει σημαντικά στην αναγνώριση του συγγραφικού ύφους. Επιπλέον, η μέθοδος LUAR παρουσίασε την υψηλότερη απόδοση, με Accuracy 0,99 και F1-score 0,99, χωρίς να απαιτείται πρόσθετη εκπαίδευση στα δεδομένα της παρούσας εργασίας. Αντίθετα, η Frozen DeBERTa παρουσίασε χαμηλότερη ακρίβεια (0,87), γεγονός που αναδεικνύει τη σημασία της εξειδικευμένης προεκπαίδευσης ενός μοντέλου για την αναπαράσταση συγγραφικών χαρακτηριστικών.

Ένα ακόμη σημαντικό συμπέρασμα της εργασίας αφορά τη σχέση μεταξύ της ακρίβειας των προβλέψεων και της ερμηνευσιμότητας των αποτελεσμάτων. Η μέθοδος Residualized Similarity, παρότι παρουσίασε χαμηλότερη ακρίβεια (0,91) σε σύγκριση με τη LUAR, πέτυχε υψηλή τιμή AUROC (0,98), δείχνοντας ότι διατηρεί σημαντική ικανότητα διάκρισης μεταξύ κειμένων ίδιου και διαφορετικού συγγραφέα. Παράλληλα, η αξιοποίηση γλωσσικών χαρακτηριστικών μέσω του Gram2vec προσφέρει τη δυνατότητα καλύτερης κατανόησης των παραγόντων που επηρεάζουν την ομοιότητα μεταξύ δύο κειμένων. Επομένως, η αξιολόγηση μιας μεθόδου δεν θα πρέπει να περιορίζεται αποκλειστικά στην ακρίβεια, αλλά να λαμβάνει υπόψη και τη δυνατότητα ερμηνείας των αποφάσεών της. Τέλος, πρέπει να επισημανθεί ότι τα πειράματα πραγματοποιήθηκαν σε συγκεκριμένο σύνολο δεδομένων από το Twitter/X, όπου κάθε έγγραφο αποτελούνταν από 20 συνενωμένα tweets. Συνεπώς, τα υψηλά αποτελέσματα δεν συνεπάγονται απαραίτητα αντίστοιχη απόδοση σε μεμονωμένα tweets ή σε κείμενα διαφορετικών πλατφορμών. Η

διαπίστωση αυτή αναδεικνύει την ανάγκη περαιτέρω αξιολόγησης της ικανότητας γενίκευσης των μεθόδων σε διαφορετικές συνθήκες.

Όσον αφορά τη μελλοντική εργασία, θα μπορούσε να εξεταστεί η εφαρμογή των μεθόδων σε μεγαλύτερα σύνολα δεδομένων, με περισσότερους συγγραφείς και μεγαλύτερη ποικιλία κειμένων. Επίσης, θα μπορούσαν να εξεταστούν πειράματα με μεμονωμένα tweets. Επιπλέον, ενδιαφέρον θα παρουσίαζε η αξιολόγησή τους σε διαφορετικές πλατφόρμες κοινωνικής δικτύωσης. Τέλος, θα μπορούσε να διερευνηθεί ο συνδυασμός παραδοσιακών και νευρωνικών μεθόδων, με στόχο τη βελτίωση της ακρίβειας, της γενίκευσης και της ερμηνευσιμότητας των αποτελεσμάτων.

Βιβλιογραφία

[1] Wikipedia contributors. (n.d.). *Machine learning*. In Wikipedia.

[2] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group.

[3] Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

[5] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

[6] Wikipedia contributors. (n.d.). *Natural Language Processing*. In Wikipedia.

[7] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv preprint arXiv:1301.3781.

[8] Pennington, J., Socher, R., & Manning, C. D. (2014). *GloVe: Global Vectors for Word Representation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

[9] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). *Enriching Word Vectors with Subword Information*. Transactions of the Association for Computational Linguistics, 5, 135–146.

[10] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep Contextualized Word Representations*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2227–2237.

[11] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL-HLT 2019, 4171–4186.

[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention Is All You Need*. Advances in Neural Information Processing Systems, 30, 5998–6008.

[13] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. Nature, 521, 436–444.

[14] Kingma, D. P., & Ba, J. (2015). *Adam: A Method for Stochastic Optimization*. International Conference on Learning Representations (ICLR)

[15] Schuster, M., & Paliwal, K. K. (1997). *Bidirectional recurrent neural networks*. IEEE Transactions on Signal Processing, 45(11), 2673–2681

[16] Hochreiter, S., & Schmidhuber, J. (1997). *Long Short-Term Memory. Neural Computation*, 9(8), 1735–1780.

[17] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). *Sequence to sequence learning with neural networks*. Advances in Neural Information Processing Systems, 27, 3104–3112.

[18] Bahdanau, D., Cho, K., & Bengio, Y. (2015). *Neural machine translation by jointly learning to align and translate*. International Conference on Learning Representations (ICLR 2015).

[19] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv:1907.11692.

[20] He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with Disentangled Attention*. International Conference on Learning Representations (ICLR).

[21] Stamatatos, E. (2009). *A Survey of Modern Authorship Attribution Methods*. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

[22] Potha, N., & Stamatatos, E. (2019). *Improving Author Verification Based on Topic Modeling*. Journal of the Association for Information Science and Technology, 70(10), 1074–1088.

[23] Koppel, M., Schler, J., & Argamon, S. (2011). *Authorship Attribution in the Wild*. Language Resources and Evaluation, 45(1), 83–94.

[24] Seidman, S. (2013). *Authorship Verification Using the Impostors Method: Notebook for PAN at CLEF 2013*. CEUR Workshop Proceedings, 1179.

[25] Koppel, M., & Winter, Y. (2014). *Determining if Two Documents Are Written by the Same Author*. Journal of the Association for Information Science and Technology, 65(1), 178–187.

[26] Potha, N., & Stamatatos, E. (2017). *An improved impostors method for authorship verification*. In *Experimental IR Meets Multilinguality,* Multimodality, and Interaction (Lecture Notes in Computer Science, Vol. 10456, pp. 138–144). Springer.

[27] Potha, N., & Stamatatos, E. (2020). *Improved algorithms for extrinsic author verification*. Knowledge and Information Systems, 62(5), 1903–1921

[28] Stamatatos, E. (2016). *Authorship verification: A review of recent advances*. Research in Computing Science, 123, 9–25.

[29] Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M., & Stein, B. (2020). *Overview of the Cross-Domain Authorship Verification Task at PAN 2020*. Working Notes of CLEF 2020, CEUR Workshop Proceedings, 2696.

[30] Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Stein, B., & Potthast, M. (2021). *Overview of the Cross-Domain Authorship Verification Task at PAN 2021*. Working Notes of CLEF 2021, CEUR Workshop Proceedings, 2936, 1743–1759.

[31] Stamatatos, E., Kestemont, M., Kredens, K., Pezik, P., Heini, A., Bevendorff, J., Stein, B., & Potthast, M. (2022). *Overview of the Authorship Verification Task at PAN 2022*. CEUR Workshop Proceedings, 3180, 2301–2313.

[32] Huang, B., Chen, C., & Shu, K. (2024). *Can Large Language Models Identify Authorship?* Findings of the Association for Computational Linguistics: EMNLP 2024, 445–460.

[33] Rivera-Soto, R. A., Miano, O. E., Ordonez, J., Chen, B. Y., Khan, A., Bishop, M., & Andrews, N. (2021). *Learning Universal Authorship Representations*. Proceedings of EMNLP 2021, 913–919.

[34] Zeng, P., Alipoormolabashi, P., Mun, J., Dey, G., Soni, N., Balasubramanian, N., Rambow, O., & Schwartz, H. (2025). *Residualized Similarity for Faithfully Explainable Authorship Verification*. Findings of the Association for Computational Linguistics: EMNLP 2025, 15824–15837.

Παράρτημα Ι

Οι υλοποιήσεις των μεθόδων που χρησιμοποιήθηκαν είναι διαθέσιμοι στο github και βρίσκονται στην παρακάτω διεύθυνση:

https://github.com/icsd12084/Thesis