ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Ανίχνευση αλλαγών ύφους σε κείμενα πολλαπλών συγγραφέων
(Style change detection in multi-author documents)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Καραγιάννης Νικόλαος

Επιβλέπων : Δρ. Σταματάτος Ευστάθιος

Μέλη εξεταστικής επιτροπής:

Σάμος, Σεπτέμβριος 2025

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πρόλογος και ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του Προγράμματος Σπουδών του Τμήματος Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων του Πανεπιστημίου Αιγαίου. Το αντικείμενό της αφορά την ανάλυση και ανίχνευση αλλαγών συγγραφικού ύφους σε κείμενα πολλαπλών συγγραφέων, με χρήση σύγχρονων μεθοδολογιών βαθιάς μάθησης και συγκεκριμένα μοντέλων βασισμένων σε γλωσσικές αναπαραστάσεις (LUAR, SBERT).

Η μελέτη του συγγραφικού ύφους αποτελεί ένα σύνθετο πρόβλημα με εφαρμογές στην υπολογιστική γλωσσολογία, την ανίχνευση λογοκλοπής, καθώς και την ανάλυση περιεχομένου σε διαδικτυακά περιβάλλοντα. Η εργασία αυτή επιδιώκει να συμβάλει στη βιβλιογραφία διερευνώντας πειραματικά διαφορετικές παραλλαγές μοντέλων (παγωμένα/προσαρμοσμένα, με ομοιότητες cosine ή concatenated embeddings + classification head) και αξιολογώντας τα σε δεδομένα του διεθνούς διαγωνισμού PAN-CLEF.

Η εργασία οργανώνεται σε διακριτές ενότητες: αρχικά παρουσιάζεται το θεωρητικό υπόβαθρο και η σχετική βιβλιογραφία, στη συνέχεια αναλύεται η μεθοδολογία και η διαδικασία προεπεξεργασίας δεδομένων, έπειτα περιγράφεται η πειραματική υλοποίηση και τέλος παρατίθενται τα αποτελέσματα, τα συμπεράσματα και οι μελλοντικές κατευθύνσεις.

Αρχικά, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή μου, Δρ. Σταματάτος Ευστάθιος, για την πολύτιμη καθοδήγηση, τις χρήσιμες παρατηρήσεις και την υποστήριξή του σε όλη τη διάρκεια της εκπόνησης της εργασίας. Η συμβολή του υπήρξε καθοριστική τόσο στην επιστημονική κατεύθυνση όσο και στη διαμόρφωση του τελικού αποτελέσματος.

Ευχαριστώ επίσης τα μέλη ΔΕΠ και το προσωπικό του Τμήματος για τις γνώσεις και τα εφόδια που μου προσέφεραν κατά τη διάρκεια των σπουδών μου.

Ιδιαίτερες ευχαριστίες οφείλω στους συμφοιτητές και φίλους μου για τη συντροφιά, τις συζητήσεις και την αλληλοϋποστήριξη σε αυτή την πορεία.

Τέλος, θα ήθελα να ευχαριστήσω από καρδιάς την οικογένειά μου για την αδιάκοπη αγάπη, στήριξη και υπομονή τους όλα αυτά τα χρόνια. Χωρίς αυτούς, η ολοκλήρωση των σπουδών μου δεν θα ήταν εφικτή.

του/της

Καραγιάννης Νικόλαος

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Η σελίδα αυτή είναι σκόπιμα λευκή.

Πίνακας περιεχομένων

Λίστα Σχημάτων

Λίστα Πινάκων

Ακρωνύμια

LUAR	Learning Universal Authorship Representations
SBERT	Sentence Bidirectional Encoder Representations from Transformers
RNN	Recurrent Neural Networks
CNN	Convolutional Neural Networks
LSTM	Long-Short Term Memory
MLP	Multi-Layer Perceptron
CRF	Conditional Random Fields

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με την ανάλυση και ανίχνευση αλλαγών συγγραφικού ύφους σε κείμενα πολλών συγγραφέων στο επίπεδο της παραγράφου. Το πρόβλημα αυτό έχει ιδιαίτερο ενδιαφέρον στον χώρο της Υπολογιστικής Γλωσσολογίας και της Επεξεργασίας Φυσικής Γλώσσας, καθώς σχετίζεται με εφαρμογές όπως η ανίχνευση λογοκλοπής, η αναγνώριση ταυτότητας συγγραφέα και η ασφάλεια πληροφοριών.

Για την υλοποίηση της εργασίας αξιοποιήθηκαν σύγχρονα μοντέλα βαθιάς μάθησης βασισμένα σε γλωσσικές αναπαραστάσεις, όπως το **SBERT** και το **LUAR**. Εξετάστηκαν διαφορετικές παραλλαγές (παγωμένα και προσαρμοσμένα μοντέλα, μέθοδοι βασισμένες σε ομοιότητα cosine ή σε concatenated embeddings + classification head), ενώ τα πειράματα πραγματοποιήθηκαν σε δεδομένα του διεθνούς διαγωνισμού **PAN-CLEF**.

Τα αποτελέσματα δείχνουν ότι οι μέθοδοι που αξιοποιούν αναπαραστάσεις συγγραφικού ύφους σε επίπεδο παραγράφου μπορούν να αποδώσουν με ακρίβεια στην ανίχνευση αλλαγών συγγραφέα, αναδεικνύοντας τη σημασία των μοντέλων γλωσσικών αναπαραστάσεων σε αυτό το πεδίο.

Λέξεις Κλειδιά: ανάλυση συγγραφικού ύφους, αλλαγή συγγραφέα, SBERT, LUAR, PAN-CLEF, επεξεργασία φυσικής γλώσσας

Abstract

This thesis focuses on the analysis and detection of writing style changes in multi-author texts on paragraph level. This problem is of particular interest in the fields of Computational Linguistics and Natural Language Processing, as it is closely related to applications such as plagiarism detection, authorship identification, and information security.

For the implementation of this work, modern deep learning models based on language representations were utilized, such as **SBERT** and **LUAR**. Different variations were examined (frozen and fine-tuned models, methods based on cosine similarity, or concatenated embeddings combined with a classification head), while the experiments were conducted on datasets from the international **PAN-CLEF** competition.

The results indicate that methods leveraging paragraph-level authorial style representations can achieve accurate detection of author changes, highlighting the importance of language representation models in this domain.

Keywords: writing style analysis, author change detection, SBERT, LUAR, PAN-CLEF, natural language processing

1

Εισαγωγή

1.1 Επεξεργασία Φυσικής Γλώσσας και Υπολογιστική Γλωσσολογία

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP) αποτελεί έναν από τους ταχύτερα αναπτυσσόμενους κλάδους της Τεχνητής Νοημοσύνης [1], καθώς επικεντρώνεται στη δημιουργία μεθόδων και εργαλείων που επιτρέπουν στους υπολογιστές να κατανοούν, να επεξεργάζονται και να παράγουν ανθρώπινη γλώσσα. Στο πλαίσιο αυτό, η Υπολογιστική Γλωσσολογία επιχειρεί να μοντελοποιήσει τη γλωσσική γνώση με τρόπο που να είναι κατανοητός και αξιοποιήσιμος από τα υπολογιστικά συστήματα.

Η ανάπτυξη του διαδικτύου και η ευρεία χρήση των ψηφιακών μέσων οδήγησαν σε μια εκθετική αύξηση του όγκου των κειμένων που παράγονται καθημερινά. Ιδιαίτερα οι πλατφόρμες κοινωνικής δικτύωσης, τα ιστολόγια και τα διαδικτυακά φόρουμ δημιούργησαν ένα περιβάλλον όπου η γλώσσα χρησιμοποιείται με νέους, δυναμικούς και πολλές φορές ανορθόδοξους τρόπους. Αυτή η ποικιλομορφία εγείρει σημαντικές ερευνητικές προκλήσεις, καθώς τα κείμενα διαφέρουν ως προς το ύφος, την πολυπλοκότητα, το επίπεδο τυπικότητας και τον σκοπό επικοινωνίας.

Παράλληλα, η ανάγκη για αυτοματοποιημένη ανάλυση της γλώσσας έχει αυξηθεί, καθώς σχετίζεται με κρίσιμες εφαρμογές, όπως η ανίχνευση λογοκλοπής, η κατηγοριοποίηση κειμένων, η αναγνώριση συναισθήματος, η εξαγωγή πληροφορίας και η αναγνώριση ταυτότητας συγγραφέα. Ωστόσο, η μελέτη αυτών των φαινομένων δεν είναι απλή, καθώς η ανθρώπινη γλώσσα χαρακτηρίζεται από αμφισημία, υποκειμενικότητα και συνεχή εξέλιξη.

Ένα από τα πλέον σύνθετα ζητήματα στον χώρο είναι η ανάλυση του συγγραφικού ύφους [2], δηλαδή των χαρακτηριστικών που διαφοροποιούν τη γραφή ενός ατόμου από ενός άλλου. Η μελέτη αυτή απαιτεί μεθοδολογίες ικανές να συλλάβουν λεπτές διαφοροποιήσεις, να ανιχνεύσουν μοτίβα και να αντιμετωπίσουν την ετερογένεια που εμφανίζεται σε πολυσυγγραφικά κείμενα. Παρά την πρόοδο που έχει σημειωθεί με τη χρήση σύγχρονων μοντέλων βαθιάς μάθησης και γλωσσικών αναπαραστάσεων, παραμένουν σημαντικές προκλήσεις που καθιστούν τον χώρο αυτό εξαιρετικά ενδιαφέρον και ανοιχτό σε περαιτέρω έρευνα.

1.2 Αντικείμενο διπλωματικής

Η παρούσα διπλωματική εργασία επικεντρώνεται στην ανάλυση και ανίχνευση αλλαγών συγγραφικού ύφους σε πολυσυγγραφικά κείμενα. Το βασικό πρόβλημα που επιχειρεί να αντιμετωπίσει είναι η αυτόματη διάκριση σημείων στα οποία μεταβάλλεται η ταυτότητα του συγγραφέα μέσα σε ένα κείμενο στο επίπεδο των παραγράφων. Το ζήτημα αυτό είναι ιδιαίτερα απαιτητικό, καθώς οι διαφορές στο ύφος μεταξύ συγγραφέων μπορεί να είναι λεπτές, ενώ ταυτόχρονα το περιεχόμενο και τα συμφραζόμενα μπορούν να επηρεάσουν σημαντικά τη γλωσσική έκφραση.

Πιο συγκεκριμένα, η εργασία στοχεύει:

- Στην εφαρμογή σύγχρονων μεθόδων βαθιάς μάθησης για την ανάλυση του συγγραφικού ύφους.
- Στην αξιολόγηση και σύγκριση διαφορετικών προσεγγίσεων (παγωμένα μοντέλα, προσαρμοσμένα μοντέλα, μεθόδους βασισμένες σε ομοιότητα και σε ταξινομητές).
- Στην πειραματική επαλήθευση των μεθόδων αυτών σε δεδομένα διεθνούς κύρους, και συγκεκριμένα στο πλαίσιο του διαγωνισμού **PAN-CLEF**.

Η υλοποίηση παρουσιάζει σημαντικές τεχνικές δυσκολίες. Αφενός, η ανάλυση ύφους σε επίπεδο παραγράφου απαιτεί την εξαγωγή αναπαραστάσεων υψηλής ποιότητας, οι οποίες να αποτυπώνουν με ακρίβεια τις διαφορές μεταξύ συγγραφέων. Αφετέρου, η εκπαίδευση ή προσαρμογή βαθιών νευρωνικών δικτύων προϋποθέτει σωστή επιλογή αρχιτεκτονικής, μεθόδων βελτιστοποίησης και στρατηγικών τακτοποίησης (regularization), ώστε να αποφευχθεί η υπερπροσαρμογή (overfitting).

Η διπλωματική εργασία αξιοποιεί μοντέλα γλωσσικών αναπαραστάσεων όπως τα **SBERT** και **LUAR**, τα οποία προσφέρουν δυνατότητες κατανόησης της σημασιολογικής και υφολογικής πληροφορίας. Δοκιμάζονται διαφορετικές παραλλαγές, όπως:

- **Χρήση παγωμένων embeddings** και μέτρων ομοιότητας (π.χ. cosine similarity).
- **Συνδυασμός embeddings με classification head**, ώστε να επιτευχθεί άμεση πρόβλεψη αλλαγής συγγραφέα.
- **Fine-tuning** των μοντέλων σε ειδικά προσαρμοσμένα δεδομένα, με στόχο την καλύτερη απόδοση σε πολυσυγγραφικά κείμενα.

Με αυτόν τον τρόπο, η εργασία αποσκοπεί να διερευνήσει σε ποιο βαθμό τα σύγχρονα μοντέλα βαθιάς μάθησης μπορούν να ανταποκριθούν στο δύσκολο έργο της ανίχνευσης αλλαγών συγγραφικού ύφους και να συμβάλει στην κατανόηση των ορίων και των δυνατοτήτων τους στον συγκεκριμένο ερευνητικό χώρο.

1.2.1 Ερευνητικά ερωτήματα και στόχοι

Η διπλωματική εργασία επικεντρώνεται σε τρία βασικά ερευνητικά ερωτήματα:
(α) σε ποιο βαθμό τα σύγχρονα μοντέλα βαθιάς μάθησης μπορούν να ανιχνεύσουν με ακρίβεια τις αλλαγές συγγραφέα σε πολυσυγγραφικά κείμενα,
(β) ποιες μεθοδολογικές επιλογές (π.χ. frozen embeddings, fine-tuning, μέθοδοι ταξινόμησης, σύγκριση ομοιότητας) προσφέρουν καλύτερη απόδοση στο συγκεκριμένο έργο, και
Βασικός στόχος είναι η ανάπτυξη και αξιολόγηση μεθόδων που να συνδυάζουν την υπολογιστική

αποδοτικότητα με υψηλή ακρίβεια, προσφέροντας ταυτόχρονα νέες γνώσεις σχετικά με την υφολογική διαφοροποίηση σε κείμενα.

1.2.2 Τεχνικές προκλήσεις

Η ανάλυση συγγραφικού ύφους σε επίπεδο παραγράφου εγείρει πλήθος τεχνικών προκλήσεων. Οι διαφορές μεταξύ συγγραφέων συχνά εκδηλώνονται σε λεπτά υφολογικά χαρακτηριστικά, τα οποία είναι δύσκολο να αποτυπωθούν με συμβατικές μεθόδους. Επιπλέον, η γλωσσική έκφραση επηρεάζεται από το περιεχόμενο και τα συμφραζόμενα, γεγονός που καθιστά δυσχερή τη διάκριση του ύφους από τη θεματική πληροφορία. Παράλληλα, η εκπαίδευση νευρωνικών μοντέλων σε περιορισμένα δεδομένα ενέχει τον κίνδυνο υπερπροσαρμογής (overfitting), απαιτώντας ιδιαίτερη προσοχή στην επιλογή αρχιτεκτονικών, στρατηγικών regularization και διαδικασιών αξιολόγησης.

1.2.3 Μεθοδολογική προσέγγιση

Η μεθοδολογική προσέγγιση της παρούσας εργασίας βασίζεται στη χρήση σύγχρονων μοντέλων γλωσσικών αναπαραστάσεων. Συγκεκριμένα, αξιοποιούνται τα SBERT και LUAR, τα οποία επιτρέπουν την αποτύπωση τόσο της σημασιολογικής όσο και της υφολογικής πληροφορίας. Εξετάζονται διαφορετικές παραλλαγές, όπως η χρήση παγωμένων embeddings σε συνδυασμό με μέτρα ομοιότητας (π.χ. cosine similarity), ο συνδυασμός embeddings με classification head για άμεση πρόβλεψη αλλαγών συγγραφέα, καθώς και το fine-tuning των μοντέλων. Η πειραματική διαδικασία στηρίζεται σε δεδομένα του διαγωνισμού PAN-CLEF, τα οποία παρέχουν υψηλής ποιότητας και διεθνώς αναγνωρισμένο υλικό αξιολόγησης.

1.2.4 Αναμενόμενη συνεισφορά

Η εργασία φιλοδοξεί να συμβάλει στη βαθύτερη κατανόηση της δυνατότητας εφαρμογής μοντέλων βαθιάς μάθησης στην ανίχνευση αλλαγών συγγραφικού ύφους. Μέσω της συγκριτικής ανάλυσης διαφορετικών μεθόδων, επιδιώκεται να αναδειχθούν οι ισχυρές και αδύναμες πλευρές κάθε προσέγγισης, παρέχοντας χρήσιμα συμπεράσματα για μελλοντικές εφαρμογές στην υπολογιστική γλωσσολογία, την ανάλυση κειμένων και την ασφάλεια πληροφοριών.

1.3 Δομή της διπλωματικής

Η δομή της παρούσας διπλωματικής εργασίας έχει ως εξής:

Στο **Κεφάλαιο 2** παρουσιάζονται οι βασικές έννοιες και οι σχετικές ερευνητικές εργασίες που συνδέονται με την ανάλυση συγγραφικού ύφους και την ανίχνευση αλλαγής συγγραφέα. Το **Κεφάλαιο 3**, η σχετική έρευνα που αφορά αποτελέσματα των διαγωνισμών. Στο **Κεφάλαιο 4** περιγράφεται αναλυτικά η μοντελοποίηση των πειραμάτων, τα εργαλεία που χρησιμοποιήθηκαν και τα χαρακτηριστικά των δεδομένων που αξιοποιήθηκαν. Το **Κεφάλαιο 5** παρουσιάζει τα πειράματα και τα αποτελέσματα που προέκυψαν από την εφαρμογή των μεθόδων, καθώς και συγκριτική ανάλυση με εναλλακτικές προσεγγίσεις. Τέλος, το **Κεφάλαιο 6** συνοψίζει τα βασικά συμπεράσματα της εργασίας και προτείνει κατευθύνσεις για μελλοντική έρευνα.

2

Θεωρητικό Υπόβαθρο

2.1 Υφομετρία και Ανάλυση Συγγραφικού Ύφους

Η υφομετρία αποτελεί τον επιστημονικό κλάδο που ασχολείται με τη μελέτη του συγγραφικού ύφους, αξιοποιώντας γλωσσολογικά και στατιστικά χαρακτηριστικά για την ανάλυση κειμένων. Οι πρώτες προσεγγίσεις στηρίχθηκαν σε απλές μετρικές, όπως η συχνότητα εμφάνισης λέξεων, το μήκος προτάσεων ή η χρήση σημείων στίξης. Παρά την απλότητά τους, οι μέθοδοι αυτές παρείχαν χρήσιμα αποτελέσματα σε εφαρμογές όπως η αναγνώριση ταυτότητας συγγραφέα.

Με την εξέλιξη της Υπολογιστικής Γλωσσολογίας και της Επεξεργασίας Φυσικής Γλώσσας [3], η ανάλυση συγγραφικού ύφους [2]απέκτησε πιο σύνθετα εργαλεία, βασισμένα σε γραμματικά, συντακτικά και σημασιολογικά χαρακτηριστικά. Σταδιακά, το ενδιαφέρον της ερευνητικής κοινότητας στράφηκε προς τη χρήση μηχανικής μάθησης, επιτρέποντας την αυτόματη ανίχνευση μοτίβων που διακρίνουν τους συγγραφείς με μεγαλύτερη ακρίβεια.

2.2 Ανίχνευση Αλλαγών Συγγραφέα (Authorship Change Detection)

Η ανίχνευση αλλαγών συγγραφέα αποτελεί εξειδικευμένο υποπρόβλημα της υφομετρίας [3] και εστιάζει στον εντοπισμό σημείων μέσα σε ένα κείμενο όπου μεταβάλλεται η ταυτότητα του συγγραφέα. Σε αντίθεση με την κλασική απόδοση συγγραφέα (authorship attribution), όπου εξετάζεται ένα σύνολο κειμένων γνωστής ή άγνωστης προέλευσης, εδώ το ζητούμενο είναι η ενδοκειμενική ανάλυση, δηλαδή ο εντοπισμός των ορίων μεταξύ τμημάτων που ανήκουν σε διαφορετικούς συγγραφείς. Το πρόβλημα είναι ιδιαίτερα απαιτητικό, καθώς το στυλ του συγγραφέα συνυπάρχει με θεματικές και πραγματολογικές διαφοροποιήσεις, οι οποίες μπορεί να «θολώνουν» τα υφολογικά ίχνη.

Οι πρώτες μελέτες βασίστηκαν σε απλές μεθόδους, όπως η τεχνική των sliding window [4]s, όπου το κείμενο χωριζόταν σε επικαλυπτόμενα τμήματα στα οποία υπολογίζονταν στατιστικά χαρακτηριστικά (π.χ. συχνότητες λέξεων, μήκος προτάσεων, χρήση σημείων στίξης). Η απόφαση για το αν αλλάζει ο συγγραφέας βασιζόταν σε συγκρίσεις μεταξύ των γειτονικών τμημάτων. Παρά την απλότητά τους, οι μέθοδοι αυτές παρείχαν τα πρώτα ενθαρρυντικά αποτελέσματα σε περιορισμένα σύνολα δεδομένων.

Με την εξέλιξη της Υπολογιστικής Γλωσσολογίας και της Μηχανικής Μάθησης, η έρευνα στράφηκε σε πιο σύνθετες μεθοδολογίες. Ιδιαίτερη σημασία απέκτησε η επιλογή και εξαγωγή χαρακτηριστικών (lexical, character, syntactic, semantic), τα οποία μπορούν να αποτυπώσουν με μεγαλύτερη ακρίβεια το συγγραφικό ύφος. Για παράδειγμα, τα **χαρακτηριστικά χαρακτήρων** (character n-grams) αποδείχθηκαν ιδιαίτερα αποτελεσματικά σε θορυβώδη κείμενα, ενώ τα

συντακτικά και σημασιολογικά χαρακτηριστικά προσφέρουν πιο βαθιά ανάλυση αλλά απαιτούν ισχυρά γλωσσικά εργαλεία. Σε ορισμένες περιπτώσεις, χρησιμοποιήθηκαν επίσης **εφαρμογο-ειδικά χαρακτηριστικά**, όπως οι χαιρετισμοί ή η δομή σε διαδικτυακά φόρουμ, τα οποία ενισχύουν την ανάλυση σε συγκεκριμένα πεδία.

Τα τελευταία χρόνια, η πρόοδος των **νευρωνικών δικτύων** και ειδικότερα των **μετασχηματιστών (transformers)** [5]άλλαξε ριζικά το τοπίο. Μοντέλα όπως τα BERT, RoBERTa και Sentence-BERT επέτρεψαν την αναπαράσταση κειμένου σε εννοιολογικά πλούσιους χώρους (embeddings), διευκολύνοντας την ανίχνευση αλλαγών ύφους ακόμη και σε πολυσυγγραφικά κείμενα. Οι προσεγγίσεις αυτές συχνά ενσωματώνουν τεχνικές ομαδοποίησης (clustering), μέτρα ομοιότητας (cosine similarity), καθώς και ταξινομητές που μαθαίνουν να προβλέπουν μεταβάσεις συγγραφέα.

Σημαντικό ρόλο στην ανάπτυξη της περιοχής διαδραματίζουν οι διεθνείς διαγωνισμοί **PAN-CLEF**, οι οποίοι από το 2010 και έπειτα παρέχουν τυποποιημένα σύνολα δεδομένων και πρωτόκολλα αξιολόγησης για την ανίχνευση αλλαγής συγγραφέα. Οι εργασίες που παρουσιάστηκαν στο πλαίσιο αυτό υιοθέτησαν ευρύ φάσμα μεθόδων, από RNNs και CNNs έως state-of-the-art transformer μοντέλα. Η ύπαρξη κοινών benchmarks επέτρεψε την αντικειμενική σύγκριση αλγορίθμων και την πρόοδο προς πιο γενικεύσιμες λύσεις.

Συνοψίζοντας, η ανίχνευση αλλαγών συγγραφέα αποτελεί ένα από τα πιο απαιτητικά προβλήματα της υφομετρίας. Ο συνδυασμός χαρακτηριστικών χαμηλού επιπέδου (lexical, character-based) με μοντέλα βαθιάς μάθησης φαίνεται να προσφέρει τις καλύτερες επιδόσεις. Ωστόσο, παραμένουν προκλήσεις όπως η διάκριση μεταξύ θεματικών και υφολογικών διαφορών, η ανάγκη για μεγάλα και ισορροπημένα σύνολα δεδομένων, καθώς και η προσαρμογή των μεθόδων σε πολυγλωσσικά ή υβριδικά περιβάλλοντα.

2.3 Προσεγγίσεις με Βαθιά Μάθηση

Στη σύγχρονη υφομετρία, η χρήση νευρωνικών δικτύων έχει επιτρέψει την ανάλυση των συγγραφικών χαρακτηριστικών με μεγαλύτερη ακρίβεια και ευελιξία. Τα RNNs και τα LSTM δίκτυα αποτέλεσαν τα πρώτα εργαλεία για την επεξεργασία ακολουθιακών δεδομένων [5], [6], αξιοποιώντας τη δυνατότητα να διατηρούν πληροφορία για προηγούμενες λέξεις ή παραγράφους και να εντοπίζουν εξαρτήσεις μακράς διάρκειας μέσα στο κείμενο. Παράλληλα, τα CNNs εισήχθησαν για την αναγνώριση τοπικών μοτίβων, όπως n-grams, μέσω συνελικτικών φίλτρων που εντοπίζουν χαρακτηριστικές αλληλουχίες λέξεων ή χαρακτήρων.

Η επανάσταση όμως ήρθε με τα transformer-based μοντέλα, όπως BERT, RoBERTa, SBERT και LUAR, τα οποία μέσω των μηχανισμών self-attention μπορούν να συλλαμβάνουν τις σχέσεις μεταξύ όλων των λέξεων ενός κειμένου, παρέχοντας πλούσιες και δυναμικές αναπαραστάσεις. Στη υφομετρία, αυτά τα embeddings σε επίπεδο πρότασης ή παραγράφου έχουν αποδειχθεί ιδιαίτερα χρήσιμα για την ανίχνευση αλλαγών συγγραφέα, καθώς ενσωματώνουν τόσο στυλιστικές όσο και σημασιολογικές πληροφορίες.

Η εκπαίδευση αυτών των μοντέλων μπορεί να γίνει είτε μέσω fine-tuning, προσαρμόζοντας όλα τα βάρη στις ανάγκες της συγκεκριμένης εργασίας, είτε με frozen embeddings, όπου τα προεκπαιδευμένα embeddings παραμένουν σταθερά και τροφοδοτούν ένα νέο classification head. Επιπλέον, η ανίχνευση αλλαγών συγγραφέα μπορεί να πραγματοποιηθεί είτε με μετρικές ομοιότητας, όπως η cosine similarity, είτε μέσω επιβλεπόμενων μοντέλων ταξινόμησης που

προβλέπουν άμεσα την παρουσία αλλαγής. Η επιλογή της κατάλληλης μεθόδου εξαρτάται τόσο από τα χαρακτηριστικά των δεδομένων όσο και από τους στόχους της ανάλυσης.

2.4 Αναπαραστάσεις Κειμένου και Γλωσσικά Embeddings

Η αναπαράσταση του κειμένου αποτελεί θεμέλιο για την επιτυχή εφαρμογή νευρωνικών δικτύων στη υφομετρία. Οι πρώτες τεχνικές βασίζονταν σε Bag-of-Words και n-grams, οι οποίες μέτρησαν απλώς τη συχνότητα εμφάνισης λέξεων ή ακολουθιών, χωρίς να λαμβάνουν υπόψη το συμφραζόμενο ή τη σημασιολογία. Η εισαγωγή των word embeddings, όπως το Word2Vec και το GloVe, επέτρεψε την αναπαράσταση των λέξεων σε συνεχή διανυσματικούς χώρους, συλλαμβάνοντας σχέσεις σημασίας μεταξύ τους.

Τα στατικά embeddings όμως δεν διαφοροποιούν τη σημασία μιας λέξης ανάλογα με το περιβάλλον της, κάτι που καλύπτεται από τα δυναμικά embeddings, όπως το BERT, τα οποία δημιουργούν διαφορετικές αναπαραστάσεις για κάθε λέξη ανάλογα με το συμφραζόμενο. Επιπλέον, η μετάβαση σε sentence-level embeddings, όπως στο SBERT [5], προσφέρει συνεκτικές αναπαραστάσεις ολόκληρων προτάσεων, διευκολύνοντας tasks όπως η ταξινόμηση και η ανίχνευση αλλαγών συγγραφέα σε μεγαλύτερες ενότητες κειμένου.

Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα ειδικά embeddings για υφομετρία, όπως αυτά που παρέχει το LUAR, τα οποία ενσωματώνουν πληροφορίες σε επίπεδο παραγράφου, συνδυάζοντας στυλιστικά και σημασιολογικά χαρακτηριστικά. Η μετάβαση από τα στατικά embeddings σε δυναμικές και συμφραζόμενες αναπαραστάσεις, όπως αυτές που εισήγαγαν τα μοντέλα BERT και SBERT, άνοιξε τον δρόμο για μια νέα γενιά αρχιτεκτονικών: τα Μεγάλα Γλωσσικά Μοντέλα (LLMs). Τα μοντέλα αυτά όχι μόνο επεκτείνουν την έννοια της γλωσσικής αναπαράστασης, αλλά και θέτουν τις βάσεις για πιο σύνθετες προσεγγίσεις στη υφομετρία, όπως αναλύεται στην επόμενη ενότητα.

2.5 Μεγάλα Γλωσσικά Μοντέλα και Εξελίξεις στις Αναπαραστάσεις

Η πρόσφατη πρόοδος στην Επεξεργασία Φυσικής Γλώσσας έχει συνδεθεί άμεσα με την ανάπτυξη των **Μεγάλων Γλωσσικών Μοντέλων (Large Language Models – LLMs)** [5], τα οποία βασίζονται σε αρχιτεκτονικές μετασχηματιστών (Transformers). Τα LLMs εκπαιδεύονται σε τεράστιους όγκους δεδομένων και επιτυγχάνουν να συλλάβουν πολύπλοκες σχέσεις μεταξύ λέξεων, προτάσεων και παραγράφων. Η ικανότητά τους να κατανοούν συμφραζόμενα και να παράγουν συνεκτικό λόγο έχει μετατρέψει τα μοντέλα αυτά σε θεμέλιο της σύγχρονης Υπολογιστικής Γλωσσολογίας.

Ένα από τα σημαντικότερα χαρακτηριστικά των LLMs είναι η χρήση **contextual embeddings**, δηλαδή διανυσματικών αναπαραστάσεων που προσαρμόζονται δυναμικά στο περιβάλλον εμφάνισης κάθε λέξης. Σε αντίθεση με τα στατικά word embeddings (Word2Vec, GloVe), τα οποία αποδίδουν σε κάθε λέξη ένα μοναδικό διάνυσμα ανεξαρτήτως συμφραζομένων, τα μοντέλα τύπου BERT, RoBERTa και GPT παράγουν διαφορετική αναπαράσταση για κάθε χρήση της λέξης, ενσωματώνοντας πληροφορία σημασιολογική και συντακτική.

Η εξέλιξη αυτή επέτρεψε τη μετάβαση σε **sentence-level και paragraph-level embeddings** [7], μέσω μοντέλων όπως το SBERT, τα οποία συμπυκνώνουν ολόκληρες προτάσεις ή παραγράφους σε συνεκτικούς διανυσματικούς χώρους. Οι αναπαραστάσεις αυτές χρησιμοποιούνται ευρέως σε

εργασίες υφομετρίας, καθώς προσφέρουν τη δυνατότητα μέτρησης ομοιότητας ύφους και σημασίας με υψηλή ακρίβεια.

Περαιτέρω εξειδίκευση παρατηρείται σε μοντέλα όπως το LUAR, τα οποία σχεδιάστηκαν ειδικά για υφομετρικές εφαρμογές. Συνδυάζοντας στυλιστικά και σημασιολογικά χαρακτηριστικά σε επίπεδο παραγράφου, προσφέρουν μια πιο στοχευμένη αναπαράσταση για την ανάλυση αλλαγών συγγραφέα.

Συνολικά, τα LLMs και οι εξελίξεις στα embeddings διαμορφώνουν το τρέχον πλαίσιο ερευνητικής δραστηριότητας στη υφομετρία.

3

Σχετική Έρευνα

3.1 Εισαγωγή

Σε αυτή την ενότητα παρουσιάζονται οι κυριότερες προηγούμενες προσεγγίσεις για το πρόβλημα της ανίχνευσης αλλαγών συγγραφέα (style change detection). Έμφαση δίνεται στις κορυφαίες συμμετοχές στους διαγωνισμούς PAN-CLEF, καθώς και σε επιλεγμένες ερευνητικές εργασίες που διαμόρφωσαν την πρόοδο του πεδίου. Η ανάλυση επικεντρώνεται τόσο στις τεχνικές που εφαρμόστηκαν όσο και στα πλεονεκτήματα και μειονεκτήματά τους.

3.2 Προσεγγίσεις στα Datasets

3.2.1 PAN21

Μορφή διαγωνισμού

Την χρονια εκείνη ο διαγωνσιμός είχε 3 tasks:

1. Να βρεθει αν το κειμενο είναι γραμμένο από έναν συγγραφέα ή πολλούς.
2. Να βρεθούν τα σημεία αλλαγής συγγραφέα στο επίπεδο παραγράφου.
3. Να αποδοθούν οι παραγράφοι σε συγγραφείς.

Το dataset αποτελείται με posts από το StackExchange και αποτελείτε από 16.000 αρχεία. Υπάρχει ίσος αριθμός αρχείων με 1 συγγραφέα – 2 συγγραφείς – 3 συγγραφείς και 4 συγγραφείς. Με split train-validation-test στην μορφή 70%-15%-15%. [8]

Κορυφαίες συμμετοχές:

Αυτό το έτος υπήρχαν 5 υποβολές:

1. **Deibel and Löfflad** → multi-layer perceptrons + bidirectional LSTMs, βασιζόμενοι σε χαρακτηριστικά όπως μέσο μήκος πρότασης, vocab richness και fastText embeddings
2. **Nath** → Siamese networks για να υπολογίζει ομοιότητα ανάμεσα σε παραγράφους, με GloVe embeddings
3. **Singh et al** → ακολούθησαν προσέγγιση verification (υπολόγιζαν διαφορά μεταξύ χαρακτηριστικών των παραγράφων και χρησιμοποιούσαν logistic regression)
4. **Strøm** → BERT embeddings + stylistic features
5. **Zang et al** → (Single BERT) χρησιμοποίησαν ένα BERT fine-tuned μοντέλο για όλα τα tasks — χρησιμοποιούσαν τις προβλέψεις για Task 2 και Task 3 και συνέθεταν το Task 1 απ' αυτές.

Πίνακας 1: Αποτελέσματα διαγωνισμού (PAN21) καταταγμένα κατά την απόδοση τους ανά task.

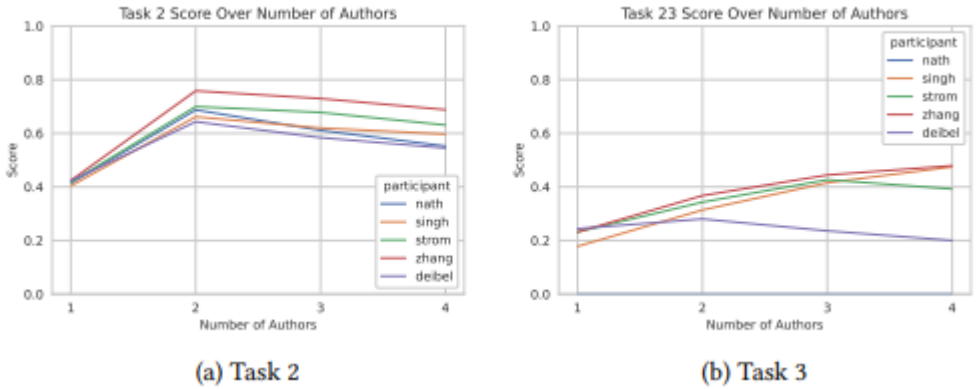
Participant	Task1 F1	Task2 F1	Task3 F1
Zhang et al.	0.753	**0.751**	**0.501**
Stø m	**0.795**	0.707	0.424
Singh et al.	0.634	0.657	0.432
Deibel et al.	0.621	0.669	0.263
Nath	0.704	0.647	–
Baseline	0.457	0.470	0.329

Το καλύτερο αποτέλεσμα για κάθε επίπεδο δυσκολίας επισημαίνεται με έντονη γραφή, καθώς προήλθε από διαφορετικές προσεγγίσεις.

Η baseline βασίστηκε σε τυχαία κατανομή προβλέψεων.

Όπως φαίνεται από τα αποτελέσματα, όλες οι συμμετοχές υπερέβησαν σημαντικά τη baseline σε όλα τα tasks, με μοναδική εξαίρεση την προσέγγιση των **Deibel et al.** για το Task 3, η οποία εμφάνισε χαμηλότερη επίδοση.
Η καλύτερη επίδοση στο Task 1 (αναγνώριση εάν ένα έγγραφο είναι μονοσυγγραφικό ή πολυσυγγραφικό) επιτεύχθηκε από τον **Strøm**, ενώ η υψηλότερη απόδοση στις εργασίες πραγματικής ανίχνευσης αλλαγών ύφους (Tasks 2 και 3) ανήκει στους **Zhang et al.** Σε όλες τις περιπτώσεις, η κορυφαία προσέγγιση παρουσίασε ουσιαστικά ανώτερα αποτελέσματα από όλες τις υπόλοιπες συμμετοχές.

έλος, εξετάστηκε πώς η απόδοση των συστημάτων μεταβάλλεται ανάλογα με τον πραγματικό αριθμό συγγραφέων ανά έγγραφο.
Για το **Task 2**, η απόδοση κορυφώνεται στα **δύο άτομα**, γεγονός που δείχνει ότι οι μέθοδοι ανιχνεύουν με μεγαλύτερη ακρίβεια αλλαγές μεταξύ δύο διαφορετικών συγγραφών. Αντίθετα, στο **Task 3**, δύο από τις υποβληθείσες προσεγγίσεις (**Zhang et al.** και **Singh et al.**) βελτίωναν σταδιακά την απόδοσή τους όσο αυξανόταν ο αριθμός των συγγραφέων, φτάνοντας στο μέγιστο όταν το έγγραφο είχε **τέσσερις συγγραφείς**.



(a) Task 2 (b) Task 3

Εικόνα 1: F1 scores ανάλογα με τον αριθμό συγγραφεών (PAN21)

Το εύρημα αυτό υποδεικνύει ότι η παρουσία περισσότερων συγγραφών μπορεί να δημιουργεί **πιο έντονες διαφοροποιήσεις ύφους**, οι οποίες διευκολύνουν τη διαδικασία ανίχνευσης. Οι

διοργανωτές προτείνουν, με βάση το αποτέλεσμα αυτό, την εξέταση μεγαλύτερου αριθμού συγγραφέων σε μελλοντικές εκδόσεις του διαγωνισμού.

3.2.2 PAN22

Μορφή διαγωνισμού

Την χρονια εκείνη ο διαγωνσιμός είχε 3 tasks:

1. Να βρεθεί η θέση αλλαγής συγγραφέα στο επίπεδο της παραγράφου (μια αλλαγή – 2 συγγραφείς).
2. Να βρεθεί η θέση αλλαγής συγγραφέα στο επίπεδο της παραγράφου (εως 5 συγγραφείς).
3. Να βρεθεί η θέση αλλαγής συγγραφέα στο επίπεδο της πρότασης.

Το dataset αποτελείται με posts από το StackExchange και αποτελείτε από 2.000 αρχεία για το task1, 10.000 για το task2 και 10.000 για το task3 (για το task3 τα dataset αφορούν προτάσεις αντί για παραγράφους). Με split train-validation-test στην μορφή 70%-15%-15%.[9]

Κορυφαίες συμμετοχές

Αυτό το έτος υπήρχαν 9 υποβολές (8 intrinsic, 1 extrinsic):

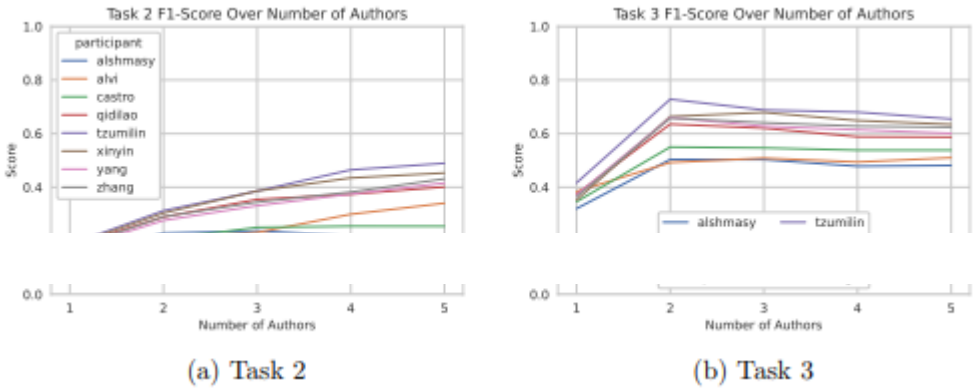
1. **Alshamasi and Menai** → Χρησιμοποίησαν λεξικά και συντακτικά χαρακτηριστικά σε επίπεδο πρότασης και παραγράφου. Εφάρμοσαν k-means clustering για την ομαδοποίηση παραγράφων με βάση την ομοιότητα ύφους (με χρήση cosine distance). Για Subtask 1 έθεσαν k=2, ενώ για Subtasks 2–3 τα clusters χρησιμοποιήθηκαν ως προβλέψεις συγγραφέα.
2. **Lao et al** → Βασίστηκαν σε προεκπαιδευμένο BERT, το οποίο fine-tuned πάνω στα δεδομένα του διαγωνισμού. Το output του BERT περνάει από 1D convolution και max pooling, καταλήγοντας σε binary classification layer για ανίχνευση αλλαγής ύφους. Για Subtask 2 χρησιμοποίησαν pairwise cosine similarities μεταξύ παραγράφων.
3. **Jiang et al** → Εφάρμοσαν transformer-based neural networks βασισμένα στο ELECTRA, που θεωρείται πιο αποδοτικό στην εκπαίδευση από το BERT. Χρησιμοποίησαν τρεις διαφορετικές εκδόσεις του ELECTRA, μία για κάθε υποεργασία, ανάλογα με την πολυπλοκότητα και τον όγκο δεδομένων.
4. **Zi et al** → Χρησιμοποίησαν BERT embeddings με masked language modeling, τα οποία ενίσχυσαν με Bi-LSTM για καλύτερη αποτύπωση συμφραζομένων. Στη συνέχεια, εφάρμοσαν convolution + max pooling, και τέλος ένα fully connected layer για τις τελικές προβλέψεις.
5. **Rodríguez-Losada and Castro** → Συνδύασαν transformer embeddings, συχνότητα σημείων στίξης και συχνότητα discourse markers. Για Subtask 1, υπολόγισαν similarity μεταξύ διαδοχικών παραγράφων και προέβλεψαν αλλαγή εκεί όπου η ομοιότητα ήταν η χαμηλότερη. Για Subtasks 2–3, όρισαν thresholds για κάθε κατηγορία χαρακτηριστικών και προέβλεψαν αλλαγή όταν οι περισσότερες τιμές ήταν κάτω από αυτά.
6. **Zhang et al** → Εισήγαγαν prompt-based μοντέλο με BERT, όπου το δίκτυο «συμπληρώνει το κενό» σε προτάσεις τύπου "They are the [blank] writing style: {Paragraph1} and {Paragraph2}", με λέξεις όπως same, different, unlike κ.ά. Χρησιμοποίησαν αυτές τις προβλέψεις για όλα τα subtasks.

7. **Lin et al** → Ανέπτυξαν ensemble μοντέλο βασισμένο σε τρεις προεκπαιδευμένους language models (BERT, RoBERTa, ALBERT). Κάθε μοντέλο χρησιμοποιήθηκε για binary ταξινόμηση (ίδιος/διαφορετικός συγγραφέας), και τα αποτελέσματα συνδυάστηκαν με majority voting.
8. **Alvi et al** → Εφάρμοσαν χειροκίνητα χαρακτηριστικά (handcrafted features) με βάση discourse markers. Για Subtask 1 εντόπισαν συνομιλιακά μοτίβα (conversational patterns) για τον εντοπισμό αλλαγής. Για Subtasks 2–3 χρησιμοποίησαν random forest για τον υπολογισμό του αριθμού συγγραφέων και k-means clustering για την ομαδοποίηση παραγράφων.

Πίνακας 2: Αποτελέσματα διαγωνισμού (PAN22) καταταγμένα κατά την απόδοση τους ανά task.

Participant	Task1	Task2	Task3
tzumilin22	**0.754**	**0.51**	**0.7156**
xinyin22	0.7346	0.4687	0.672
qidilao22	0.7471	0.417	0.641
zhang22	0.7162	0.4174	0.6581
yang22	0.669	0.4011	0.6483
alvi22	0.7052	0.3213	0.5636
castro22a	0.5661	0.2735	0.5565
alshmasy22	0.5272	0.2207	0.4995
Baseline	0.3222	0.2651	0.4809

Τα υψηλότερα συνολικά αποτελέσματα σημειώθηκαν από την ομάδα **Lin et al.**, της οποίας η προσέγγιση πέτυχε τις καλύτερες επιδόσεις σε όλα τα μετρικά που εξετάστηκαν. Παρόμοια, αν και ελαφρώς χαμηλότερη απόδοση, παρουσίασαν και οι **Jiang et al.** και **Lao et al.**, γεγονός που δείχνει μια γενικότερη σύγκλιση των συστημάτων σε ό,τι αφορά την ακρίβεια εντοπισμού αλλαγών συγγραφέα. Η ανάλυση έδειξε ότι στα **Tasks 2 και 3**, όπου τα έγγραφα περιλαμβάνουν πολλαπλούς συγγραφείς, η απόδοση των μοντέλων εξαρτάται έντονα από τον πραγματικό αριθμό των συγγραφέων. Στα έγγραφα με δύο συγγραφείς, οι περισσότερες προσεγγίσεις εμφάνισαν σταθερά υψηλές τιμές F1, ωστόσο καθώς αυξανόταν ο αριθμός των συγγραφέων, η απόδοση παρουσίαζε πτώση.



(a) Task 2 (b) Task 3

Εικόνα 2: F1 scores ανάλογα με τον αριθμό συγγραφεών (PAN22)

3.2.3 PAN23

Μορφή διαγωνισμού

Μετανομάστηκε το πρόβλημα σε *Multi-Author Writing Style Analysis* και εισήγαγε τρεις βαθμίδες δυσκολίας (easy, medium, hard). Για πρώτη φορά, έγινε έλεγχος της θεματικής ποικιλίας (topic control), ώστε να μην αποτελεί "υποκατάστατο" της υφολογικής διαφοράς. Τα δεδομένα προήλθαν από το Reddit, με συγγραφείς από διαφορετικά subreddits (r/worldnews r/politics r/askhistorians r/legaladvice).

Easy – Τα κείμενα καλύπτουν ποικιλία θεμάτων, επιτρέποντας στις προσεγγίσεις να αξιοποιήσουν θεματικές μεταβολές μεταξύ παραγράφων ως ενδείξεις αλλαγής συγγραφέα.

Medium – Τα κείμενα παρουσιάζουν περιορισμένη θεματική ποικιλία, γεγονός που απαιτεί μεγαλύτερη έμφαση σε υφολογικά παρά σε θεματικά χαρακτηριστικά για την ανίχνευση των αλλαγών.

Hard – Όλες οι παράγραφοι ενός κειμένου αφορούν το ίδιο θέμα. Σε αυτή την περίπτωση, οι θεματικές πληροφορίες παύουν να είναι χρήσιμες, και η ανίχνευση αλλαγών συγγραφέα βασίζεται αποκλειστικά σε λεπτές υφολογικές διαφοροποιήσεις.

Κάθε dataset αποτελείτε από 6.000 αρχεία και το split παραμένει 70:15:15.[10]

Κορυφαίες συμμετοχές:

Αυτό το έτος υπήρχαν 6 υποβολές:

1. **Ye et al**. → Χρησιμοποίησαν το προεκπαιδευμένο DeBERTaV3 LLM σε συνδυασμό με contrastive learning για την ανίχνευση αλλαγών ύφους. Εφάρμοσαν dense neural network με ειδική contrastive loss και υλοποίησαν prompt-based προσέγγιση, όπου το σύστημα μαθαίνει αυτόματα πώς να ρωτά αν δύο παράγραφοι γράφτηκαν από τον ίδιο συγγραφέα ή όχι.
2. **Chen et al**. → Εφάρμοσαν επίσης contrastive learning πάνω στο DeBERTa, δημιουργώντας πρόσθετα ζεύγη παραγράφων για εκπαίδευση — όχι μόνο από διαδοχικές αλλά και μη γειτονικές παραγράφους, ώστε να αυξηθεί η ποικιλία των παραδειγμάτων και να βελτιωθεί η γενίκευση του μοντέλου.
3. **Kucukkaya et al**. → Χρησιμοποίησαν DeBERTaV3 και μοντελοποίησαν το πρόβλημα ως Natural Language Inference (NLI) task. Για κάθε δύο διαδοχικές παραγράφους δημιούργησαν είσοδο με 256 tokens ανά παράγραφο, διαχωρισμένα με [SEP] token, και ένα [CLS] token για δυαδική ταξινόμηση (ίδιος ή διαφορετικός συγγραφέας). Πρότειναν δύο μεθόδους περικοπής για μεγάλα κείμενα:
   – *First-256 truncation* (για το hard dataset), και
   – *Transition-focused truncation* (για easy/medium), που κρατά tokens γύρω από το πιθανό σημείο αλλαγής ύφους.
4. **Huang et al**. → Βασίστηκαν στο πολυγλωσσικό mT0-xl language model, εφαρμόζοντας knowledge distillation ώστε να εκπαιδεύσουν ένα μικρότερο student model με χαμηλότερο υπολογιστικό κόστος. Ήταν η μόνη ομάδα που ενσωμάτωσε επιπλέον δεδομένα εκπαίδευσης, χρησιμοποιώντας το PAN2020 authorship verification dataset εκτός του επίσημου συνόλου.
5. **Jacobo et al.** → Αντιμετώπισαν το πρόβλημα ως authorship verification και εφάρμοσαν παραδοσιακές μεθόδους Μηχανικής Μάθησης. Χρησιμοποίησαν ως χαρακτηριστικά term-

document matrices και Prediction by Partial Matching (PPM), τα οποία εισήγαγαν σε Support Vector Machines (SVM) και logistic regression. Για τα easy/medium datasets χρησιμοποίησαν PPM + logistic regression, ενώ για το hard set term-document matrix + SVM.

6. **Hashemi et al.** → Συνδύασαν τα προεκπαιδευμένα BERT, RoBERTa και ELECTRA μοντέλα με ένα binary classification layer για την πρόβλεψη αλλαγής συγγραφέα. Εφάρμοσαν data augmentation δημιουργώντας ζεύγη μη γειτονικών παραγράφων και mirrored pairs για εμπλουτισμό των δεδομένων εκπαίδευσης. Επιπλέον, διερεύνησαν ensemble μοντέλα εκπαιδευμένα ξεχωριστά στα τρία datasets του διαγωνισμού.

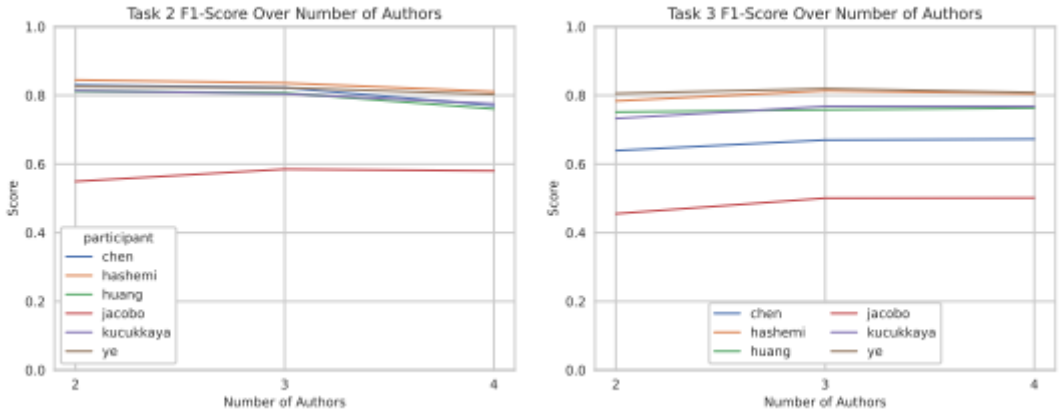
Πίνακας 3: Αποτελέσματα διαγωνισμού (PAN23) καταταγμένα κατά την απόδοση τους ανά task.

Participant	Easy F1	Medium F1	Hard F1
Chen et al.	0.914	0.82	0.676
Hashemi et al.	**0.984**	**0.843**	0.812
Huang et al.	0.968	0.806	0.769
Jacobo et al.	0.793	0.591	0.498
Kucukkaya et al.	0.982	0.81	0.772
Ye et al.	0.983	0.83	**0.821**

Το καλύτερο αποτέλεσμα για κάθε σύνολο δεδομένων επισημαίνεται με έντονη γραφή. Για τα **easy** και **medium** σύνολα δεδομένων, την υψηλότερη απόδοση πέτυχαν οι **Hashemi et al.** , ενώ για το **hard** σύνολο, τα καλύτερα αποτελέσματα προήλθαν από τους **Ye et al.**

Σε γενικές γραμμές, παρατηρείται σαφής διαφοροποίηση στην απόδοση των μοντέλων μεταξύ των τριών βαθμίδων δυσκολίας. Οι περισσότερες προσεγγίσεις επιτυγχάνουν υψηλές τιμές F1 στο **easy** dataset, ενώ η απόδοση μειώνεται σημαντικά στο **medium** και παρουσιάζει ακόμη μία, μικρότερη πτώση στο **hard**. Το γεγονός αυτό επιβεβαιώνει την επιτυχία του στόχου δημιουργίας dataset με διαβαθμισμένη δυσκολία, όπου η αύξηση της θεματικής ομοιογένειας καθιστά δυσκολότερη την ανίχνευση υφολογικών αλλαγών.

Επιπλέον, διερευνήθηκε η συσχέτιση της απόδοσης των συστημάτων με τον αριθμό των συγγραφέων ανά έγγραφο. Επειδή οι επιδόσεις στο **easy** dataset ήταν ήδη υψηλές, η ανάλυση επικεντρώθηκε στα **medium** και **hard** σύνολα. Παρατηρήθηκε ότι η απόδοση των περισσότερων μοντέλων —ιδίως στο hard dataset— βελτιώνεται σε έγγραφα με τρεις συγγραφείς σε σχέση με εκείνα που περιέχουν μόνο δύο, ενώ στη συνέχεια μειώνεται ελαφρώς για έγγραφα τεσσάρων συγγραφέων.



Εικόνα 3: F1 scores ανάλογα με τον αριθμό συγγραφεών (PAN23)

Το εύρημα αυτό υποδεικνύει ότι οι χρησιμοποιούμενες μέθοδοι ενδέχεται να παρουσιάζουν ελαφρά προκατάληψη υπέρ κειμένων με περισσότερους από δύο συγγραφείς, κάτι που μπορεί να σχετίζεται με την ποικιλία υφολογικών στοιχείων που διευκολύνουν τη διάκριση.

3.2.4 PAN24

Μορφή διαγωνισμού

Ο διαγωνισμός PAN24 ακολούθησε τη γενική δομή του PAN23, με το ίδιο πειραματικό πλαίσιο και τις ίδιες αρχές αξιολόγησης, επεκτείνοντας ωστόσο το πρόβλημα υπό τον τίτλο Multi-Author Writing Style Analysis.

Και σε αυτή την έκδοση χρησιμοποιήθηκαν τρεις βαθμίδες δυσκολίας (easy, medium, hard), με ελεγχόμενη θεματική ποικιλία (topic control), ώστε η αναγνώριση αλλαγών συγγραφέα να βασίζεται αποκλειστικά σε υφολογικά χαρακτηριστικά και όχι σε θεματικές διαφοροποιήσεις.

Τα δεδομένα προέρχονται εκ νέου από το Reddit, με δείγματα από διάφορα subreddits (π.χ. r/worldnews, r/politics, r/askhistorians, r/legaladvice).

Η διάκριση των βαθμίδων δυσκολίας παραμένει η ίδια:

Τα easy κείμενα επιτρέπουν θεματικές μεταβολές, τα medium περιορίζουν τη θεματική ποικιλία, ενώ τα hard είναι πλήρως θεματικά ομοιογενή.

Κάθε dataset περιλαμβάνει 6.000 αρχεία, με το ίδιο split εκπαίδευσης/επικύρωσης/ελέγχου (70:15:15) όπως και στην προηγούμενη διοργάνωση. [11]

Κορυφαίες συμμετοχές:

Αυτό το έτος υπήρχαν 16 υποβολές:

1. **Lv et al.** → αξιοποίησαν τον decoder του LLaMA-3 για την εξαγωγή διανυσματικών αναπαραστάσεων παραγράφων, οι οποίες τροφοδοτήθηκαν σε ένα feed-forward δίκτυο για δυαδική ταξινόμηση· η εκπαίδευση επιταχύνθηκε μέσω low-rank adaptation (LoRA).

2. **Lin et al.** → εφάρμοσαν ensemble μοντέλο με ReBERTa, DeBERTa και ERNIE, ενώ πρόσθεσαν μετα-επεξεργασία με βάση τη σημασιολογική ομοιότητα διαδοχικών παραγράφων στα *easy* και *medium* datasets για βελτίωση ακρίβειας.
3. **Ye et al.** → βασίστηκαν στη continual learning προσέγγιση με learned progress prompts για μεταφορά γνώσης μεταξύ διαφορετικών βαθμίδων δυσκολίας.
4. **Huang και Kong** → χρησιμοποίησαν DeBERTa-v3 με regularized dropout και early stopping ώστε να αποφευχθεί το overfitting κατά το fine-tuning.
5. **Huang και Kong** → δοκίμασαν πολλαπλά BERT-based μοντέλα, επιλέγοντας DeBERTa για τα *easy* και *hard* datasets και RoBERTa για το *medium*.
6. **Wu, Kong, και Ye** → αξιοποίησαν RoBERTa με contrastive learning, στοχεύοντας στη μείωση της cosine απόστασης για θετικά pairs και στην αύξησή της για αρνητικά.
7. **Liu et al.** → εφάρμοσαν επίσης contrastive learning με RoBERTa encoder, δημιουργώντας feature matrix από embeddings και απόλυτες διαφορές, το οποίο τροφοδότησαν σε fully connected layer για πρόβλεψη.
8. **Księżniak et al**. → χρησιμοποίησαν RoBERTa και DeBERTa, ενσωματώνοντας tags με υφομετρικά χαρακτηριστικά (stylometric features) στα κείμενα για εμπλουτισμό πληροφορίας.
9. **Chen, Hand, και Yi** → βασίστηκαν σε RoBERTa και χρησιμοποίησαν R-Drop regularization για να μειώσουν το overfitting και να διασφαλίσουν συνεπή αποτελέσματα μεταξύ όμοιων εισόδων.
10. **Wu et al.** → συνέκριναν BERT, RoBERTa και DistilBERT, καταλήγοντας στη χρήση RoBERTa· τα contextual features τροφοδοτήθηκαν σε Virtual Softmax layer για τριταξική ταξινόμηση (three-class classification) ώστε να ενισχυθεί η διάκριση ορίων μεταξύ κλάσεων.
11. **Khan et al.** → εφάρμοσαν weighted sampling για ισορροπία κλάσεων, συνέκριναν RoBERTa, Electra, DeBERTa και Squeeze-BERT, επιλέγοντας το πρώτο· ενίσχυσαν το training με data augmentation μέσω αντιστροφής παραγράφων χωρίς αλλαγή συγγραφέα.
12. **Sheykhlan et al.** → χρησιμοποίησαν fine-tuned BERT, RoBERTa και ELECTRA· για το *easy* dataset κράτησαν μόνο το RoBERTa, ενώ για *medium* και *hard* εφάρμοσαν ensemble όλων των μοντέλων.
13. Sanjesh και Mang**ai** → χρησιμοποίησαν υφομετρικά χαρακτηριστικά (π.χ. TF-IDF χαρακτήρων, συχνότητα stopwords, μήκη λέξεων/χαρακτήρων) για εξαγωγή embeddings, τα οποία εισήχθησαν σε CNN + Bi-LSTM αρχιτεκτονική.
14. **Liang και Lei** → χρησιμοποίησαν GPT-3.5 ως teacher model για data generation με ερωτήσεις τύπου similarity (topic, style, vocabulary) και T5-small ως student για fine-tuning στο task.
15. **Liu, Chen, και Lv** → εφάρμοσαν τη μέθοδο Entropy-based Stability-Plasticity (ESP) με BERT encoder, ρυθμίζοντας το learning rate ανά layer βάσει εντροπίας για ισορροπία μεταξύ σταθερότητας και προσαρμοστικότητας.

Πίνακας 4: Αποτελέσματα διαγωνισμού (PAN24) καταταγμένα κατά την απόδοση τους ανά task.

Participant	Easy F1	Medium F1	Hard F1
fosu-stu	0.987	**0.887**	0.834
nycu-nlp	0.964	0.857	**0.863**
no-999	**0.991**	0.83	0.832
huangzhijian	0.985	0.815	0.826
text-understanding-and-analysi	**0.991**	0.815	0.818
bingezzzleep	0.985	0.818	0.807
openfact	0.981	0.821	0.805
chen	0.968	0.822	0.807
baker	0.976	0.816	0.77
gladiators	0.956	0.809	0.783
khaldi-abderrahmane	0.905	0.806	0.641
karami-sh	0.972	0.664	0.642
riyahsanjesh	0.825	0.712	0.599
liuc0757	0.696	0.717	0.503
lxflcl66666	0.606	0.455	0.484
foshan-university-of-guangdong	0.517	0.394	0.352
Baseline	0.414	0.506	0.495

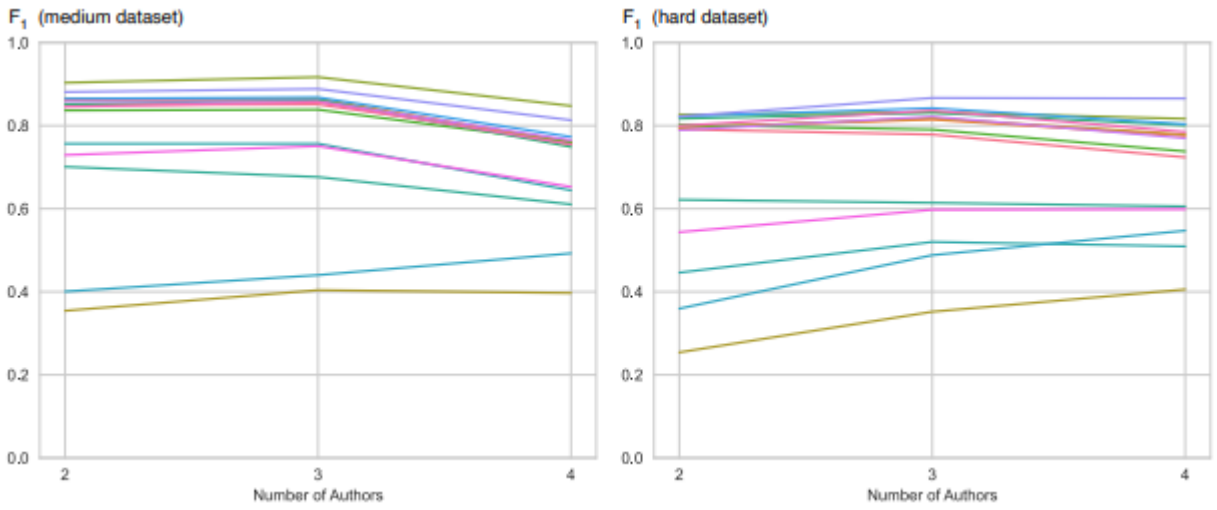
Το καλύτερο αποτέλεσμα για κάθε επίπεδο δυσκολίας επισημαίνεται με έντονη γραφή, καθώς προήλθε από διαφορετικές προσεγγίσεις.

Στο **easy dataset**, οι **Ye et al.** και **Huang & Kong** κατέλαβαν από κοινού την πρώτη θέση. Στο **medium dataset**, την υψηλότερη επίδοση πέτυχαν οι **Lv et al.** ,ενώ στο **hard dataset** η καλύτερη επίδοση προήλθε από τους **Lin et al.**

Παρατηρείται ότι, αν και εξακολουθεί να υπάρχει σαφής διαφορά στην απόδοση των μοντέλων μεταξύ των τριών βαθμίδων δυσκολίας, τα αποτελέσματα έχουν συγκλίνει σημαντικά σε σχέση με τα προηγούμενα έτη. Οι επιδόσεις στα **medium** και **hard** datasets είναι αισθητά βελτιωμένες, ενώ στο **easy dataset** τα μοντέλα προσεγγίζουν πλέον σχεδόν τέλεια αποτελέσματα.

Επιπλέον, διερευνήθηκε η συσχέτιση της απόδοσης των συστημάτων με τον αριθμό των συγγραφέων ανά έγγραφο. Και ακολουθεί το ίδιο μοτίβο με τα δύο προηγούμενα έτη, δηλαδή ότι η απόδοση των περισσότερων μοντέλων —ιδίως στο hard dataset— βελτιώνεται σε έγγραφα με τρεις συγγραφείς σε σχέση με εκείνα που περιέχουν μόνο δύο, ενώ στη συνέχεια μειώνεται ελαφρώς για έγγραφα τεσσάρων συγγραφέων.

Το εύρημα αυτό υποδεικνύει ότι οι χρησιμοποιούμενες μέθοδοι ενδέχεται να παρουσιάζουν ελαφρά προκατάληψη υπέρ κειμένων με περισσότερους από δύο συγγραφείς, κάτι που μπορεί να σχετίζεται με την ποικιλία υφολογικών στοιχείων που διευκολύνουν τη διάκριση.



Εικόνα 4: F1 scores ανάλογα με τον αριθμό συγγραφεών (PAN24)

3.3 Συγκριτική Αξιολόγηση

Η ενότητα αυτή παρουσιάζει συγκριτικά τα αποτελέσματα των κορυφαίων προσεγγίσεων που συμμετείχαν στα τελευταία τέσσερα έτη του διαγωνισμού PAN-CLEF για την ανίχνευση αλλαγών συγγραφικού ύφους. Για κάθε χρονιά, συνοψίζονται οι καλύτερες επιδόσεις ανά task ή επίπεδο δυσκολίας, καθώς και τα βασικά χαρακτηριστικά των αντίστοιχων μεθόδων. Στόχος είναι να αποτυπωθεί η εξέλιξη των τεχνικών, από τα πρώτα μοντέλα βασισμένα σε κλασικά χαρακτηριστικά έως τις σύγχρονες προσεγγίσεις με μεγάλα γλωσσικά μοντέλα (LLMs).

Πίνακας 5: Κορυφαία προσέγγιση ανά task

Διαγωνισμός	**Καλύτερη Προσέγγιση**	**Task**	**F1**	**Μέθοδος**
PAN21	Støm [12]	1	0.795	BERT embeddings + stylistic features
PAN21	Zhang et al. [13]	2	0.751	(Single BERT) χρησιμοποίησαν ένα BERT fine-tuned μοντέλο για όλα τα tasks — χρησιμοποιούσαν τις προβλέψεις για Task 2 και Task 3 και συνέθεταν το Task 1 απ' αυτές.
PAN21	Zhang et al.	3	0.501	>>
PAN22	Lin et al. [14]	1	0.754	Ανέπτυξαν ensemble μοντέλο βασισμένο σε τρεις προεκπαιδευμένους language models (BERT, RoBERTa, ALBERT). Κάθε μοντέλο χρησιμοποιήθηκε για binary ταξινόμηση (ίδιος/διαφορετικός συγγραφέας), και τα αποτελέσματα συνδυάστηκαν με majority voting.
PAN22	Lin et al.	2	0.510	>>

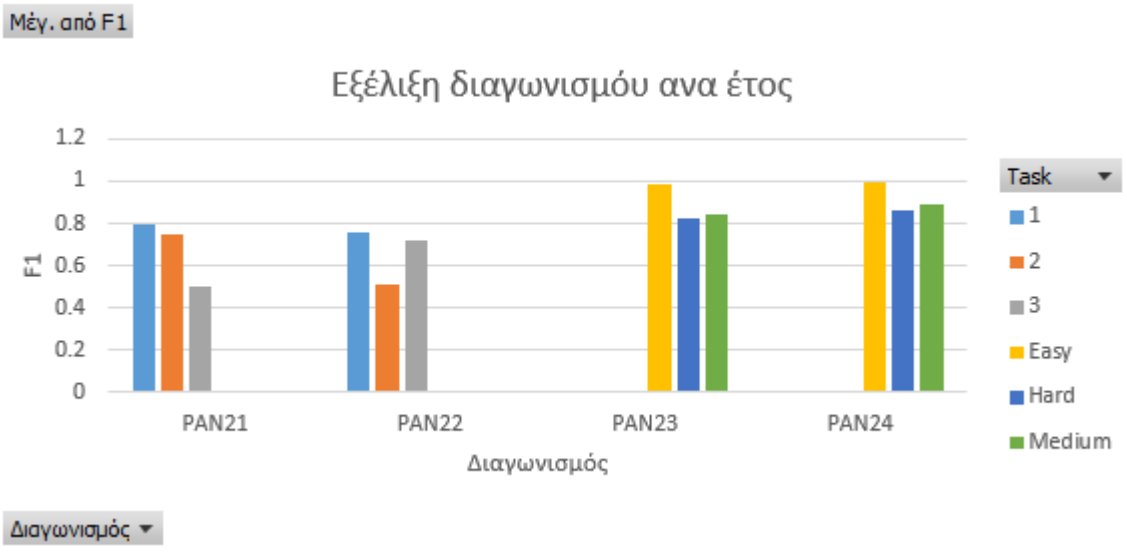
PAN22	Lin et al.	3	0.7156	>>
PAN23	Hashemi et al. [15]	Easy	0.984	Συνδύασαν τα προεκπαιδευμένα BERT, RoBERTa και ELECTRA μοντέλα με ένα binary classification layer για την πρόβλεψη αλλαγής συγγραφέα. Εφάρμοσαν data augmentation δημιουργώντας ζεύγη μη γειτονικών παραγράφων και mirrored pairs για εμπλουτισμό των δεδομένων εκπαίδευσης. Επιπλέον, διερεύνησαν ensemble μοντέλα εκπαιδευμένα ξεχωριστά στα τρία datasets του διαγωνισμού.
PAN23	Hashemi et al.	Medium	0.843	>>
PAN23	Ye et al. [16]	Hard	0.821	Χρησιμοποίησαν το προεκπαιδευμένο DeBERTaV3 LLM σε συνδυασμό με contrastive learning για την ανίχνευση αλλαγών ύφους. Εφάρμοσαν dense neural network με ειδική contrastive loss και υλοποίησαν prompt-based προσέγγιση, όπου το σύστημα μαθαίνει αυτόματα πώς να ρωτά αν δύο παράγραφοι γράφτηκαν από τον ίδιο συγγραφέα ή όχι.
PAN24	Ye et al [17]	Easy	0.991	Βασίστηκαν στη continual learning προσέγγιση με learned progress prompts για μεταφορά γνώσης μεταξύ διαφορετικών βαθμίδων δυσκολίας.
PAN24	Huang & Kong [18]	Easy	0.991	Χρησιμοποίησαν DeBERTa-v3 με regularized dropout και early stopping ώστε να αποφευχθεί το overfitting κατά το fine-tuning
PAN24	Lv et al. [19]	Medium	0.887	Αξιοποίησαν τον decoder του LLaMA-3 για την εξαγωγή διανυσματικών αναπαραστάσεων παραγράφων, οι οποίες τροφοδοτήθηκαν σε ένα feed-forward δίκτυο για δυαδική ταξινόμηση· η εκπαίδευση επιταχύνθηκε μέσω low-rank adaptation (LoRA).
PAN24	Lin et al. [20]	Hard	0.863	Εφάρμοσαν ensemble μοντέλο με ReBERTa, DeBERTa και ERNIE, ενώ πρόσθεσαν μετα-επεξεργασία με βάση τη σημασιολογική ομοιότητα διαδοχικών παραγράφων στα easy και medium datasets για βελτίωση ακρίβειας.

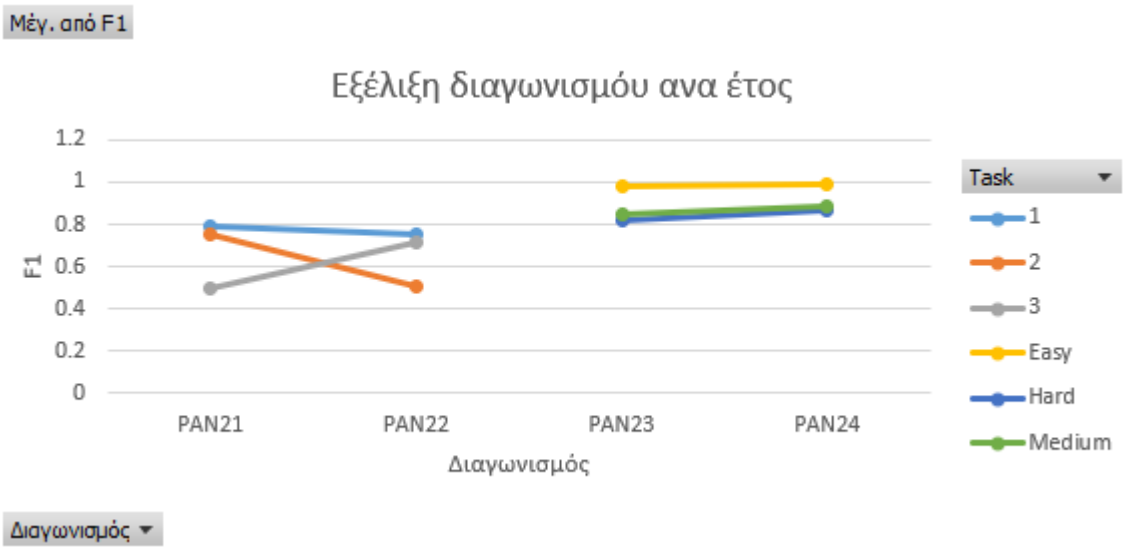
Παρατηρείται σαφής μετατόπιση της ερευνητικής τάσης από κλασικές μεθόδους υφομετρικής ανάλυσης προς πλήρως νευρωνικές προσεγγίσεις, κυρίως μετά το 2020.

- Στα **PAN21–PAN22**, οι ομάδες αξιοποιούσαν **προεκπαιδευμένα transformer μοντέλα** (BERT, RoBERTa, ALBERT), συνδυασμένα συχνά με **ensemble ή rule-based συστήματα**. Η επίδοση παρέμενε σχετικά χαμηλή (F1 < 0.8), ειδικά στα tasks υψηλής πολυπλοκότητας, γεγονός που ανέδειξε τα όρια της προσέγγισης χωρίς fine-tuning σε εξειδικευμένα δεδομένα.
- Από το **PAN23** και έπειτα, οι ομάδες στράφηκαν σε **LLMs** (BERT, DeBERTa, ELECTRA, mT0-xl), ενσωματώνοντας τεχνικές **contrastive learning, prompt learning** και **knowledge distillation**. Αυτές οι τεχνικές αύξησαν σημαντικά την απόδοση, φτάνοντας σε F1 πάνω από 0.98 για τα easy datasets.
- Στο **PAN24**, η έμφαση δόθηκε στη **βελτιστοποίηση της μάθησης** (continual learning, low-rank adaptation, regularized dropout) και στη **σημασιολογική συνέπεια μεταξύ παραγράφων** (semantic post-processing), δείχνοντας ότι το πρόβλημα έχει πλέον μετατοπιστεί από τη βασική κατανόηση ύφους στην αποτελεσματική προσαρμογή μεγάλων μοντέλων σε περιορισμένα δεδομένα.

Συνολικά, τα αποτελέσματα δείχνουν ότι η απόδοση των συστημάτων βελτιώνεται σταθερά χρόνο με τον χρόνο, κυρίως λόγω της αυξανόμενης **χρήσης προεκπαιδευμένων LLMs και τεχνικών fine-tuning**. Παράλληλα, οι συμμετέχοντες έχουν αρχίσει να αντιμετωπίζουν τη δυσκολία των θεματικά ελεγχόμενων συνόλων (topic-controlled corpora) με μεθόδους ενίσχυσης της γενίκευσης (π.χ. data augmentation, ensemble learning, continual learning). Εντούτοις, παραμένει εμφανής η πτώση επίδοσης στα **medium και hard datasets**, γεγονός που επιβεβαιώνει ότι η θεματική ποικιλία εξακολουθεί να επηρεάζει τη δυνατότητα απομόνωσης του καθαρά συγγραφικού ύφους.

Η συνολική πορεία του διαγωνισμού αναδεικνύει τη μετάβαση από την **υφομετρία με ρητά χαρακτηριστικά** προς την **υφομετρία βασισμένη σε καθολικές αναπαραστάσεις**, με τα contrastive και prompt-based μοντέλα να αποτελούν το νέο στάνταρ για το πρόβλημα.





Εικόνα 5: Σύγκριση των κορυφαίων αποτελεσμάτων του διαγωνισμού ανά χρονιά

Συνοψίζοντας, παρατηρούμε ότι η πρόοδος στο πρόβλημα της ανίχνευσης αλλαγών συγγραφικού ύφους ακολουθεί στενά την εξέλιξη των μεγάλων γλωσσικών μοντέλων. Οι πρώτες προσεγγίσεις επικεντρώνονταν σε ρητά χαρακτηριστικά και απλές αρχιτεκτονικές, ενώ οι πιο πρόσφατες αξιοποιούν καθολικές αναπαραστάσεις ύφους μέσω fine-tuned LLMs και contrastive learning. Η σταθερή αύξηση των επιδόσεων αποδεικνύει ότι το πρόβλημα έχει μετατραπεί από στατιστικό σε καθαρά αναπαραστατικό (representation learning task), όπου η επιτυχία εξαρτάται από την ικανότητα του μοντέλου να διαχωρίζει σημασιολογικά και υφολογικά σήματα.

4

Μοντελοποίηση

4.1 Εισαγωγή

Στο παρόν κεφάλαιο περιγράφεται αναλυτικά η μεθοδολογία που ακολουθήθηκε για την ανίχνευση αλλαγών συγγραφικού ύφους. Αρχικά παρουσιάζονται τα μοντέλα βάσης που αξιοποιήθηκαν, ενώ στη συνέχεια αναλύονται οι δύο κύριες μεθοδολογικές προσεγγίσεις εκπαίδευσης και πρόβλεψης — η προσέγγιση με συνημιτονική ομοιότητα (cosine similarity) και η προσέγγιση με συνένωση αναπαραστάσεων και ταξινομητή (concatenation + classification head). Τέλος, εξετάζεται ο ρόλος των embeddings και ο τρόπος με τον οποίο το «πάγωμα» ή η προσαρμογή τους (fine-tuning) επηρεάζει τη μάθηση και την απόδοση του συστήματος.

4.2 Μοντέλα

Στο πλαίσιο της παρούσας διπλωματικής εργασίας μελετήθηκαν και συγκρίθηκαν διαφορετικές παραλλαγές μοντέλων βαθιάς μάθησης για την ανίχνευση αλλαγών συγγραφικού ύφους. Συγκεκριμένα, αξιοποιήθηκαν δύο αρχιτεκτονικές: το **SBERT (Sentence-BERT)** και το **LUAR (Learning Universal Authorship Representations)** [21], οι οποίες προσεγγίζουν το πρόβλημα από διαφορετική σκοπιά. Παράλληλα, διερευνήθηκαν διαφορετικοί τρόποι εκπαίδευσης και αξιοποίησης των αναπαραστάσεων:

- **Frozen embeddings**: χρήση των προεκπαιδευμένων αναπαραστάσεων χωρίς περαιτέρω προσαρμογή.
- **Fine-tuned embeddings**: προσαρμογή των παραμέτρων των μοντέλων στα δεδομένα της συγκεκριμένης εργασίας.

Επιπλέον, εξετάστηκαν δύο διαφορετικές μεθοδολογικές προσεγγίσεις για την τελική πρόβλεψη:

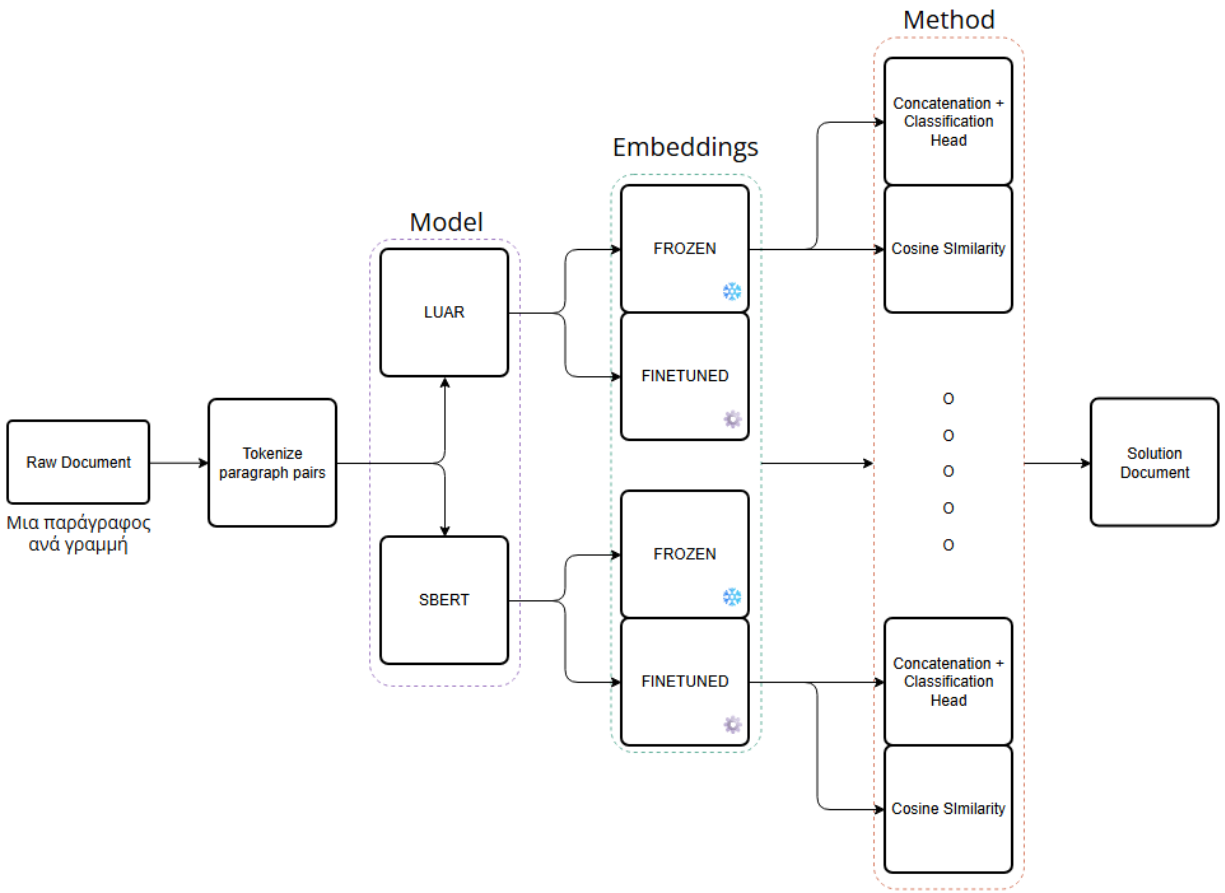
- **Μέτρα ομοιότητας** (cosine similarity) μεταξύ των παραγράφων.
- **Συνδυασμός αναπαραστάσεων με ταξινομητή** (concatenation of embeddings + classification head).

Ο στόχος των συγκρίσεων ήταν να αξιολογηθεί η επίδραση τόσο της επιλογής μεταξύ **frozen και fine-tuned embeddings**, όσο και της χρήσης **ομοιότητας έναντι ταξινομητή**, στην τελική απόδοση των μοντέλων.

Και τα δύο μοντέλα χρησιμοποιούν ως backbone έναν προεκπαιδευμένο μετασχηματιστή (*paraphrase-distilroberta-base-v1*), αλλά διαφοροποιούνται ως προς τον στόχο εκμάθησης: το SBERT στοχεύει σε σημασιολογική εγγύτητα, ενώ το LUAR σε συγγραφική ομοιότητα.

Συνολικά, στην Εικόνα 6 παρουσιάζεται η αρχιτεκτονική των πειραμάτων. Παρόλο που τα embeddings δεν αποτελούν ξεχωριστό στάδιο επεξεργασίας, αλλά ρύθμιση του εκάστοτε μοντέλου, η διάκριση αυτή αποτυπώνεται σχηματικά ώστε να αναδειχθεί ο διαφορετικός τρόπος εκπαίδευσης (frozen ή fine-tuned) και η επίδρασή του στα αποτελέσματα.

Η απεικόνιση αποσκοπεί στη σύνοψη όλων των παραλλαγών που υλοποιήθηκαν, δηλαδή των οκτώ συνδυασμών μοντέλου (SBERT ή LUAR), στρατηγικής εκπαίδευσης (frozen ή fine-tuned) και μεθόδου σύγκρισης (cosine similarity ή concatenation με classification head). Με αυτόν τον τρόπο γίνεται φανερό πώς κάθε στάδιο του pipeline —από τη δημιουργία των paragraph pairs έως την τελική απόφαση αλλαγής συγγραφέα— επηρεάζεται από τις διαφορετικές σχεδιαστικές επιλογές.



Εικόνα 6: Αρχιτεκτονική Πειραμάτων

4.2.1 SBERT

Το **Sentence-BERT (SBERT)** αποτελεί επέκταση του BERT, σχεδιασμένη για την παραγωγή σημασιολογικά πλούσιων αναπαραστάσεων προτάσεων και παραγράφων. Μέσω μηχανισμών self-attention και ενός επιπλέον pooling layer, το SBERT αποδίδει συνεκτικά embeddings που χρησιμοποιούνται εκτενώς σε εφαρμογές σύγκρισης κειμένων, αναζήτησης πληροφορίας και

αναγνώρισης ομοιότητας. Η αρχιτεκτονική αυτή χρησιμοποιήθηκε ως σημείο αναφοράς (benchmark) για τη σύγκριση με πιο εξειδικευμένα μοντέλα, όπως το LUAR.

4.2.2 LUAR

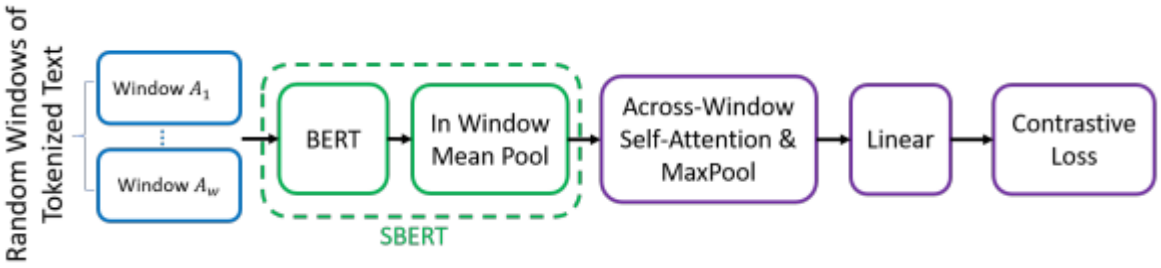
Το **Learning Universal Authorship Representations (LUAR)** επεκτείνει τη λογική του SBERT, δίνοντας έμφαση στην αποτύπωση του συγγραφικού ύφους. [21]

Το LUAR αποτελεί ένα νευρωνικό μοντέλο που έχει σχεδιαστεί ειδικά για την ανάπτυξη αναπαραστάσεων συγγραφικού ύφους, με στόχο τη διάκριση μεταξύ κειμένων διαφορετικών συγγραφέων. Σε αντίθεση με τα κλασικά μοντέλα sentence embeddings, όπως το SBERT, τα οποία βελτιστοποιούνται κυρίως ως προς τη σημασιολογική εγγύτητα προτάσεων, το LUAR επιδιώκει τη δημιουργία στιλιστικά διακριτών αναπαραστάσεων που παραμένουν σταθερές ακόμη και όταν μεταβάλλεται το περιεχόμενο, το θέμα ή το είδος του κειμένου. Βασικός σκοπός του μοντέλου είναι η εκμάθηση μιας απεικόνισης που αντιστοιχίζει συλλογές κειμένων του ίδιου συγγραφέα σε κοντινές περιοχές ενός ενιαίου διανυσματικού χώρου, ενώ αντιθέτως απομακρύνει τις αναπαραστάσεις κειμένων που προέρχονται από διαφορετικούς συγγραφείς. Η διαδικασία αυτή υλοποιείται μέσω supervised contrastive learning, το οποίο ενισχύει τη συνοχή εντός συγγραφέα και μεγιστοποιεί την απόσταση μεταξύ διαφορετικών συγγραφικών στυλ.

Αρχιτεκτονικά, το LUAR αξιοποιεί τον SBERT ως αρχικό μηχανισμό εξαγωγής σημασιολογικών αναπαραστάσεων, παράγοντας ενδιάμεσα διανύσματα 768 διαστάσεων. Στη συνέχεια, η πληροφορία επεξεργάζεται από έναν μηχανισμό αυτο-προσοχής (self-attention), ο οποίος αποσκοπεί στον εντοπισμό παραγράφων με εντονότερα υφολογικά χαρακτηριστικά. Η συσσωρευμένη πληροφορία υποβάλλεται σε max-pooling, ώστε να αναδειχθούν τα κυρίαρχα στιλιστικά μοτίβα, και τελικά προβάλλεται σε έναν χώρο 512 διαστάσεων, ο οποίος αντιστοιχεί στο τελικό authorship embedding του κειμένου. Με τον τρόπο αυτό, το LUAR δεν παράγει απλώς μια αναπαράσταση πρότασης, αλλά ένα σύνθετο υφολογικό αποτύπωμα που προκύπτει από πολλαπλά αποσπάσματα και συσχετίσεις μεταξύ τους.

Κατά την εκπαίδευση, το μοντέλο τροφοδοτείται με μικρά και τυχαία δείγματα κειμένου, ώστε να αποφευχθεί η υπερπροσαρμογή στο θεματικό περιεχόμενο και να ενισχυθεί η γενίκευση ως προς το συγγραφικό ύφος. Στο στάδιο της αξιολόγησης, χρησιμοποιούνται εκτενέστερα αποσπάσματα, με αποτέλεσμα η τελική αναπαράσταση να ενσωματώνει πλουσιότερη στυλιστική πληροφορία. Η συγκεκριμένη διαδικασία καθιστά το LUAR ανθεκτικό σε αλλαγές θεματικής και λεξιλογίου, επιδιώκοντας να απομονώσει σταθερά συγγραφικά χαρακτηριστικά ανεξάρτητα από το περιεχόμενο.

Συνολικά, το LUAR αξιοποιεί τη σημασιολογική ισχύ του SBERT, αλλά την επεκτείνει με πρόσθετους μηχανισμούς που στοχεύουν αποκλειστικά στην αποτύπωση υφολογικών χαρακτηριστικών. Ως εκ τούτου, το μοντέλο καθίσταται ιδιαιτέρως κατάλληλο για εργασίες συγγραφικής ταυτοποίησης, όπως author verification, authorship clustering και style change detection, όπου η στιλιστική διαφοροποίηση υπερέχει σε σημασία έναντι της καθαρά σημασιολογικής ομοιότητας.



Εικόνα 7: η αρχιτεκτονική του LUAR μοντέλου για την προ εκπαίδευση του.

4.3 Μέθοδοι Εκπαίδευσης

Στο πλαίσιο της εργασίας υλοποιήθηκαν και συγκρίθηκαν δύο διαφορετικοί μηχανισμοί εκπαίδευσης και λήψης απόφασης. Οι δύο προσεγγίσεις διαφέρουν ως προς το πώς χρησιμοποιούν τα παραγόμενα embeddings για να αποφανθούν εάν δύο παράγραφοι ανήκουν στον ίδιο ή διαφορετικό συγγραφέα.

4.3.1 Cosine Similarity

Η πρώτη μεθοδολογία βασίζεται στον υπολογισμό της συνημιτονικής ομοιότητας μεταξύ των διανυσματικών αναπαραστάσεων δύο παραγράφων.

Για κάθε ζεύγος παραγράφων (p_1, p_2), υπολογίζεται η ομοιότητα:

$$\cos(p_1, p_2) = \frac{p_1 \cdot p_2}{\| p_1 \| \| p_2 \|}$$

Όπου $p_1, p_2 \in \mathbb{R}^D$ είναι τα κανονικοποιημένα embeddings των παραγράφων.
Η εκπαίδευση πραγματοποιείται με τη συνάρτηση **CosineEmbeddingLoss**, η οποία επιβραβεύει υψηλή ομοιότητα για παραγράφους του ίδιου συγγραφέα και χαμηλή για διαφορετικούς:

$$L = \begin{cases} 1 - \cos(p1, p2), & \text{ανήκουν στον ίδιο συγγραφέα} \\ max(0, cos(p1, p2) - margin), & \text{ανήκουν σε διαφορετικούς συγγραφείς} \end{cases}$$

όπου το margin ελέγχει την απόσταση ανάμεσα στις δύο κατηγορίες.

Κατά την αξιολόγηση, η τελική απόφαση βασίζεται σε ένα προσαρμοζόμενο κατώφλι (threshold) t, το οποίο προσδιορίζεται στο σύνολο επικύρωσης (validation) έτσι ώστε να μεγιστοποιεί το macro F1-score.
Η διαδικασία αυτή επιτρέπει στο μοντέλο να διατηρεί ευελιξία, προσαρμόζοντας το όριο απόφασης στις στατιστικές ιδιότητες κάθε συνόλου δεδομένων.

4.3.2 Concatenation + Classification Head

Η δεύτερη μεθοδολογία αξιοποιεί έναν επιβλεπόμενο ταξινομητή για την πρόβλεψη αλλαγής συγγραφέα.
Αντί να βασίζεται αποκλειστικά στη μέτρηση ομοιότητας, το μοντέλο μαθαίνει απευθείας τη συνάρτηση που διαχωρίζει τις δύο κατηγορίες (“ίδιος συγγραφέας” / “διαφορετικός συγγραφέας”).

Η διαδικασία έχει ως εξής:

1. **Υπολογισμός paragraph embeddings:**
   Κάθε παράγραφος προβάλλεται μέσω του encoder σε ένα διάνυσμα $e_i \in \mathbb{R}^D$.
2. **Συνένωση (Concatenation):**
   Για κάθε διαδοχικό ζεύγος παραγράφων (p_i, p_{i+1}), σχηματίζεται το διάνυσμα
   $$h_i = [e(p_i); e(p_{i+1})] \in \mathbb{R}^{2D}$$

το οποίο συνδυάζει πληροφορία και από τις δύο παραγράφους.

3. **Classification Head:**
   Το διάνυσμα h_iεισάγεται σε ένα μικρό νευρωνικό δίκτυο δύο επιπέδων (fully-connected layers με ενεργοποίηση ReLU και έξοδο Softmax).
   Το δίκτυο επιστρέφει πιθανότητες για τις δύο κλάσεις:
   - 0 → Ίδιος συγγραφέας
   - 1 → Διαφορετικός συγγραφέας
4. **Εκπαίδευση:**
   Η εκπαίδευση βασίζεται στη συνάρτηση **CrossEntropyLoss**, ελαχιστοποιώντας τη διαφορά μεταξύ των προβλεπόμενων πιθανοτήτων και των πραγματικών ετικετών.

Σε αντίθεση με τη μέθοδο cosine similarity, εδώ δεν απαιτείται αναζήτηση κατωφλίου, καθώς το μοντέλο μαθαίνει απευθείας το διαχωριστικό όριο μέσω των παραμέτρων του ταξινομητή. Η μέθοδος αυτή είναι πιο εκφραστική, αλλά απαιτεί επαρκή ποικιλία δεδομένων ώστε να αποφευχθεί η υπερπροσαρμογή (overfitting).

4.4 Embeddings

Στο στάδιο της εξαγωγής αναπαραστάσεων εφαρμόστηκαν δύο διαφορετικές στρατηγικές ως προς τη χρήση του encoder:

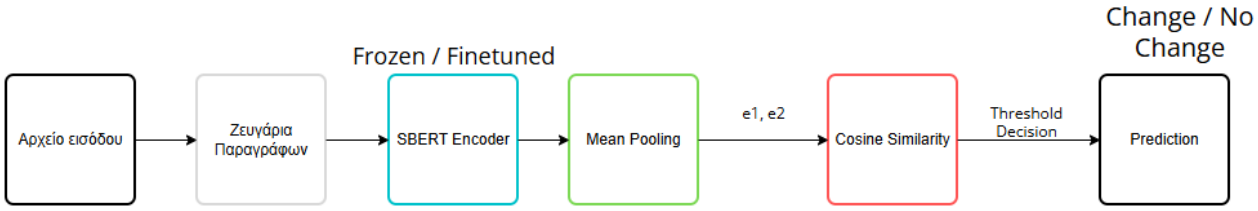
Frozen Embeddings:

Ο encoder παραμένει παγωμένος, λειτουργώντας αποκλειστικά ως feature extractor. Η προσέγγιση αυτή μειώνει το υπολογιστικό κόστος και αποτρέπει την υπερπροσαρμογή, αλλά περιορίζει την ικανότητα του μοντέλου να προσαρμόζεται στο συγκεκριμένο corpus.

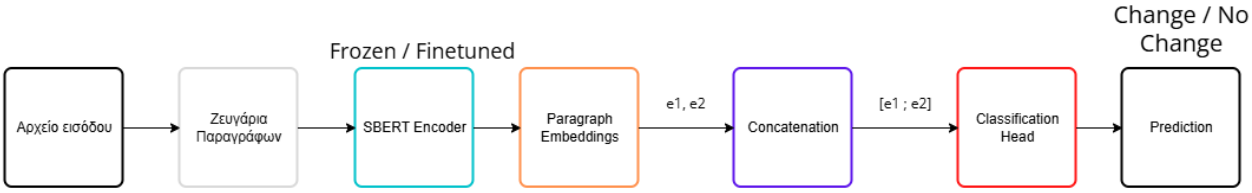
Fine-tuned Embeddings:

Ο encoder εκπαιδεύεται περαιτέρω πάνω στα δεδομένα της εργασίας. Το fine-tuning επιτρέπει τη βελτίωση της αναπαράστασης στυλομετρικών χαρακτηριστικών, αλλά απαιτεί προσεκτική ρύθμιση υπερπαραμέτρων (learning rate, batch size) για να αποφευχθεί η αλλοίωση της προεκπαιδευμένης γνώσης.

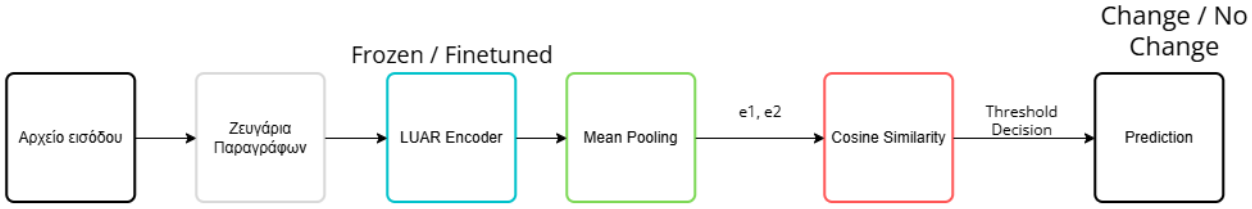
4.5 Αρχιτεκτονική των Πειραμάτων



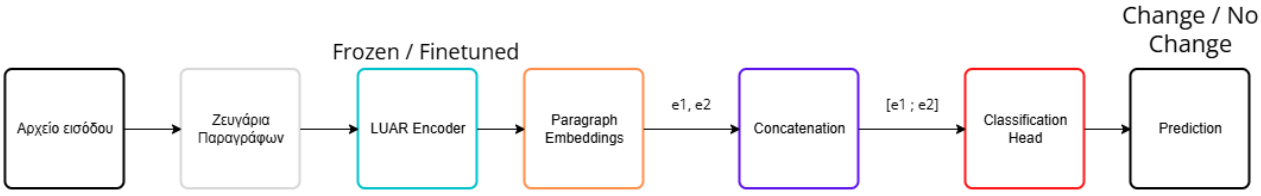
Εικόνα 8: Ροή Δεδομένων για SBERT - Cosine Similarity



Εικόνα 9: Ροή Δεδομένων για SBERT – Concatenation + Classification head



Εικόνα 10: Ροή Δεδομένων για LUAR - Cosine Similarity



Εικόνα 11: Ροή Δεδομένων για LUAR - Concatenation + Classification Head

Οι Εικόνες 8–11 απεικονίζουν τη ροή δεδομένων των τεσσάρων υλοποιημένων παραλλαγών, οι οποίες προκύπτουν από τον συνδυασμό των δύο μοντέλων (**SBERT** και **LUAR**) με τις δύο μεθοδολογίες πρόβλεψης (**Cosine Similarity** και **Concatenation + Classification Head**).

Σε όλες τις περιπτώσεις, η διαδικασία ξεκινά με τη **δημιουργία ζευγών διαδοχικών παραγράφων** από το κείμενο εισόδου. Οι παράγραφοι μετατρέπονται σε **διανυσματικές αναπαραστάσεις (embeddings)** μέσω του εκάστοτε encoder, ο οποίος λειτουργεί είτε σε **frozen** είτε σε **fine-tuned** μορφή, ανάλογα με τη στρατηγική εκπαίδευσης. Τα embeddings αυτά αποτελούν το βασικό χαρακτηριστικό εισόδου στις επόμενες φάσεις, οι οποίες διαφοροποιούνται ανάλογα με τη μέθοδο σύγκρισης.

Εικόνες 8 & 10 – Cosine Similarity

Στις εκδόσεις με **cosine similarity**, υπολογίζεται η **συνημιτονική ομοιότητα** μεταξύ των διανυσμάτων των δύο παραγράφων (e_1, e_2). Η τελική απόφαση ("Change" ή "No Change") προκύπτει μέσω ενός **προσαρμοζόμενου κατωφλίου (threshold)**, το οποίο επιλέγεται στο σύνολο επικύρωσης για τη μεγιστοποίηση του macro F1-score. Η προσέγγιση αυτή είναι **μη παραμετρική**, χωρίς επιπλέον εκπαιδεύσιμα βάρη πέρα από τον encoder, και προσφέρει σταθερότητα και αποδοτικότητα σε σενάρια με περιορισμένα δεδομένα. Στην περίπτωση του **LUAR (Εικόνα 10)**, η διαδικασία ενισχύεται από επιπρόσθετους μηχανισμούς **self-attention και mean pooling**, που επιτρέπουν τη σύλληψη χαρακτηριστικών συγγραφικού ύφους πέρα από τη σημασιολογία του κειμένου.

Εικόνες 9 & 11 – Concatenation + Classification Head

Στις εκδόσεις με **classification head**, τα embeddings των δύο παραγράφων **συνενώνονται (concatenation)** σε ένα ενιαίο διάνυσμα [e_1; e_2], το οποίο εισάγεται σε έναν **επιβλεπόμενο ταξινομητή** δύο επιπέδων (fully-connected layers με ReLU και Softmax). Ο ταξινομητής εκπαιδεύεται με **CrossEntropyLoss** ώστε να προβλέπει την πιθανότητα αλλαγής συγγραφέα. Η μέθοδος αυτή είναι **παραμετρική**, επιτρέποντας στο δίκτυο να μάθει **μη γραμμικά διαχωριστικά όρια** στον χώρο των αναπαραστάσεων.
Η εκδοχή με **LUAR (Εικόνα 11)** συνδυάζει τη στυλομετρική εξειδίκευση του μοντέλου με την εκφραστικότητα του ταξινομητή, επιτυγχάνοντας μεγαλύτερη προσαρμοστικότητα στις διαφοροποιήσεις ύφους.

4.6 Συμπεράσματα

Το κεφάλαιο αυτό παρουσίασε τη συνολική αρχιτεκτονική και τις μεθοδολογικές επιλογές που υλοποιήθηκαν στο πλαίσιο της εργασίας. Οι δύο κύριες προσεγγίσεις (Cosine Similarity και Concatenation + Classification Head) αντιπροσωπεύουν δύο διαφορετικές φιλοσοφίες εκμάθησης. Η πρώτη επικεντρώνεται στη μέτρηση ομοιότητας σε έναν προκαθορισμένο χώρο αναπαράστασης, ενώ η δεύτερη μαθαίνει απευθείας τη συνάρτηση διάκρισης μεταξύ κατηγοριών.
Η παράμετρος του fine-tuning προσθέτει έναν ακόμη άξονα διαφοροποίησης, επιτρέποντας την αξιολόγηση της ικανότητας των μοντέλων να προσαρμόζονται στα ιδιαίτερα χαρακτηριστικά των δεδομένων PAN. Η σύγκριση όλων των παραλλαγών πραγματοποιείται στο επόμενο κεφάλαιο, όπου αναλύεται η πειραματική διαδικασία και η αποδοτικότητα κάθε μεθόδου.

4.7 Περιορισμοί και Πλεονεκτήματα της Μεθοδολογίας

Η προτεινόμενη μεθοδολογική προσέγγιση βασίζεται σε γλωσσικές αναπαραστάσεις υψηλής σημασιολογικής πυκνότητας (embeddings) και σε μοντέλα όπως το LUAR, τα οποία εστιάζουν όχι μόνο στη σημασιολογία αλλά και στο συγγραφικό ύφος. Η αξιοποίηση μηχανισμών self-attention και contrastive learning ενισχύει την ικανότητα των μοντέλων να αποτυπώνουν λεπτές διαφοροποιήσεις, επιτρέποντας πιο αξιόπιστη ανίχνευση αλλαγών συγγραφέα σε πολυσυγγραφικά κείμενα.

Ωστόσο, η μεθοδολογία παρουσιάζει ορισμένους περιορισμούς. Ο σχετικά μικρός αριθμός δεδομένων ανά dataset, σε συνδυασμό με την ανομοιογένεια θεματικών πεδίων (π.χ. τεχνικά κείμενα, συζητήσεις, λογοτεχνία), μπορεί να περιορίσει τη γενικευσιμότητα των μοντέλων σε νέα ή ετερογενή δεδομένα. Επιπλέον, η εξάρτηση από προεκπαιδευμένα embeddings σημαίνει ότι τα μοντέλα κληρονομούν προκαταλήψεις ή αδυναμίες της αρχικής εκπαίδευσης.

Παρά τα παραπάνω, η μεθοδολογία συνδυάζει ακρίβεια και ευελιξία, παρέχοντας ένα πλαίσιο που μπορεί να προσαρμοστεί σε διαφορετικά σενάρια. Η ισορροπία ανάμεσα στη χρήση ενός benchmark μοντέλου (SBERT) και ενός εξειδικευμένου (LUAR) προσφέρει ολοκληρωμένη εικόνα της επίδοσης και επιτρέπει την εξαγωγή ασφαλέστερων συμπερασμάτων για τις σύγχρονες τάσεις στη υφομετρία.

5

Αποτελέσματα και Ανάλυση

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν με βάση τη μεθοδολογία που περιγράφηκε στο Κεφάλαιο 4. Εξετάζονται οι διαφορετικές παραλλαγές των μοντέλων SBERT και LUAR, με συνδυασμούς frozen/finetuned embeddings και cosine similarity/classification head, και αξιολογείται η επίδοσή τους, στην ανίχνευση αλλαγών συγγραφικού ύφους σε επίπεδο παραγράφου. Η αξιολόγηση έγινε με χρήση των μετρικών F1-score (macro), Accuracy, Precision και Recall, δίνοντας μεγαλύτερο βάρος στην F1. Τα αποτελέσματα του διαγωνισμου βρίσκονται εδώ (Zangerle et al., 2021, 2022, 2023, 2024).

5.2 Δεδομένα

5.2.1 Διαγωνισμός PAN-CLEF και Task Ανίχνευσης Αλλαγής Ύφους

Το αντικείμενο του task “style change detection” αφορά την ανίχνευση των σημείων σε ένα πολυσυγγραφικό κείμενο όπου ο συγγραφέας αλλάζει [8], [9], [10], [11]. Βασικό ερώτημα που τίθεται είναι το εξής: Εάν ένα κείμενο έχει γραφεί από πολλούς συγγραφείς, μπορούμε να εντοπίσουμε στοιχεία που να το αποδεικνύουν; Υπάρχει τρόπος να ανιχνεύσουμε μεταβολές στο ύφος γραφής; Η απάντηση σε αυτό το ερώτημα αποτελεί μία από τις πιο δύσκολες και ενδιαφέρουσες προκλήσεις στον τομέα της αναγνώρισης συγγραφέα: η ανίχνευση αλλαγής ύφους αποτελεί τον μοναδικό τρόπο να εντοπιστεί λογοκλοπή σε ένα έγγραφο όταν δεν υπάρχουν άλλα συγκριτικά κείμενα. Παράλληλα, η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για την αποκάλυψη “gift authorships”, την επαλήθευση ισχυρισμένης συγγραφής ή την ανάπτυξη νέων τεχνολογιών υποστήριξης γραφής.

Σε προηγούμενες εκδόσεις του task ανάλυσης πολυσυγγραφικού ύφους στόχος ήταν, για παράδειγμα, να εντοπιστεί αν ένα κείμενο είναι μονοσυγγραφικό ή πολυσυγγραφικό (2018), ο ακριβής αριθμός συγγραφέων σε ένα έγγραφο (2019), η ύπαρξη αλλαγής ύφους μεταξύ διαδοχικών παραγράφων (2020–2022) και η ακριβής θέση αυτών των αλλαγών (2021–2022). Το 2022, η ανίχνευση αλλαγών επεκτάθηκε και σε επίπεδο πρότασης. Τα datasets των προηγούμενων ετών παρουσίαζαν μεγάλη θεματική ποικιλία, γεγονός που επέτρεπε στους συμμετέχοντες να χρησιμοποιούν την πληροφορία του θέματος ως ένδειξη αλλαγής συγγραφέα.

Στους συμμετέχοντες ζητείται να λύσουν το πρόβλημα της ανίχνευσης αλλαγής ύφους: για ένα δοσμένο κείμενο, να εντοπιστούν όλα τα σημεία αλλαγής ύφους σε επίπεδο παραγράφου, δηλαδή για κάθε ζεύγος διαδοχικών παραγράφων να εκτιμηθεί αν υπάρχει αλλαγή συγγραφικού ύφους. Η ταυτόχρονη αλλαγή συγγραφέα και θέματος ελέγχεται προσεκτικά και παρέχονται datasets τριών επιπέδων δυσκολίας:

- **Easy:** Οι παράγραφοι ενός κειμένου καλύπτουν διάφορα θέματα, επιτρέποντας στις μεθόδους να αξιοποιήσουν πληροφορία θέματος για τον εντοπισμό αλλαγής συγγραφέα.
- **Medium:** Η θεματική ποικιλία είναι περιορισμένη (αν και παρούσα), αναγκάζοντας τις μεθόδους να εστιάσουν περισσότερο στο συγγραφικό ύφος για την επίλυση του task.
- **Hard:** Όλες οι παράγραφοι καλύπτουν το ίδιο θέμα.

Με αυτόν τον τρόπο, το PAN-CLEF παρέχει datasets σταδιακής δυσκολίας, που επιτρέπουν την αποτίμηση της ικανότητας των μοντέλων να απομονώνουν υφολογικά χαρακτηριστικά ανεξάρτητα από τη θεματική πληροφορία.

Όλα τα έγγραφα παρέχονται στα Αγγλικά και μπορεί να περιέχουν αυθαίρετο αριθμό αλλαγών ύφους. Ωστόσο, οι αλλαγές ύφους μπορεί να εμφανίζονται μόνο μεταξύ παραγράφων, καθώς κάθε παράγραφος ανήκει σε έναν μόνο συγγραφέα και δεν περιέχει εσωτερικές αλλαγές ύφους. Αντίστοιχα με τα easy, medium, hard ανά dataset υπάρχουν και τα Dataset1, Dataset2, Dataset3.

5.2.2 Μορφή δεδομένων

Η πειραματική διαδικασία βασίστηκε σε δεδομένα που προέρχονται από τον διεθνή διαγωνισμό **PAN-CLEF**, ο οποίος παρέχει τυποποιημένα corpora για προβλήματα υφομετρίας. Τα δεδομένα προέρχονται από το αποθετήριο του διαγωνισμού (https://pan.webis.de/data.html). Για την ολοκλήρωση του πειραματικού συνόλου, χρησιμοποιήθηκαν επιπλέον test sets κατόπιν επικοινωνίας με την Dr. Eva Zangerle (PC Chair του διαγωνισμού).

- **Πηγή Δεδομένων:** Χρησιμοποιήθηκαν τα παρακάτω δεδομένα από τον διαγωνισμό:
  1. PAN21 & PAN22: Style Change Detection
  2. PAN23 & PAN24: Multi-Author Writing Style Analysis
- **Μορφή Δεδομένων:** Τα δεδομένα παρέχονται σε μορφή ενός προβλήματος (txt) μαζί με το ground-truth (json) ως ζεύγη. Problem-X.txt περιέχει το κείμενο (μία παράγραφο ανά γραμμή) και truth-problem-X.json με την ακόλουθη μορφή:

PAN23 & 24

```
{
"authors": NUMBER_OF_AUTHORS,
"changes": RESULT_ARRAY_TASK
}
```

PAN21 & 22

```
{
"authors": NUMBER_OF_AUTHORS,
"site": SOURCE_SITE,
"changes": RESULT_ARRAY_TASK1 or RESULT_ARRAY_TASK3,
"paragraph-authors": RESULT_ARRAY_TASK2
}
```

Για την αξιολόγηση των λύσεων στα επιμέρους tasks, τα αποτελέσματα πρέπει να αποθηκεύονται σε ένα ξεχωριστό αρχείο για κάθε έγγραφο εισόδου και για κάθε dataset. Σημειώνεται ότι απαιτείται η δημιουργία ενός αρχείου λύσης για κάθε πρόβλημα εισόδου σε κάθε dataset. Η δομή των δεδομένων κατά τη φάση της αξιολόγησης είναι αντίστοιχη με αυτήν της φάσης εκπαίδευσης, με τη διαφορά ότι απουσιάζουν τα αρχεία ground truth.

Για κάθε πρόβλημα **problem-X.txt**, θα πρέπει να παράγει το αντίστοιχο αρχείο λύσης **solution-problem-X.json**, το οποίο περιέχει ένα JSON αντικείμενο με τη λύση του εκάστοτε task. Η λύση

είναι ένας πίνακας (array) που περιλαμβάνει δυαδικές τιμές για κάθε ζεύγος διαδοχικών παραγράφων.

Ένα ενδεικτικό παράδειγμα αρχείου λύσης είναι το παρακάτω:

```
{
"changes": [0, 0, 1, 0, 0, ...]
}
```

Οι υποβολές αξιολογούνται με βάση το μέτρο **F1-score (macro)**, το οποίο υπολογίζεται για όλα τα ζεύγη παραγράφων. Οι λύσεις για κάθε dataset αξιολογούνται ανεξάρτητα, με βάση τα σκορ που προκύπτουν από την αξιολόγηση.

Παρέχεται επίσης script για τον υπολογισμό του F1-score με βάση τα παραγόμενα αρχεία εξόδου, το οποία χρησιμοποιούμε για την αξιολόγηση.

5.3 Μέτρα αξιολόγησης

Η αξιολόγηση έγινε με βάση μετρικές που αντικατοπτρίζουν την απόδοση στην ανίχνευση αλλαγής συγγραφέα, και προσθέτω στο **F1** (το οποίο ήταν το μόνο που υπολογιζόταν κατά τον διαγωνισμό) τις παρακάτω μετρικές για μια πιο ολοκληρωμένη εικόνα του θέματος:

- **Accuracy / Precision / Recall**

Η επιλογή αυτών των μετρικών έγινε με γνώμονα τα χαρακτηριστικά του προβλήματος και τους στόχους της εργασίας. Η **Accuracy** παρέχει μια γενική εικόνα του πόσο συχνά το μοντέλο κάνει σωστές προβλέψεις, ωστόσο σε προβλήματα με ανισοκατανομή κλάσεων (π.χ. λίγες αλλαγές συγγραφέα σε σχέση με τον συνολικό αριθμό παραγράφων) μπορεί να δίνει παραπλανητικά υψηλές τιμές. Γι' αυτό, κεντρικό ρόλο στην αξιολόγηση έχει το **F1-score**, το οποίο συνδυάζει το **Precision** (την ακρίβεια στις θετικές προβλέψεις) και το **Recall** (την ικανότητα εντοπισμού όλων των πραγματικών αλλαγών). Το F1-score είναι ιδιαίτερα κατάλληλο στο συγκεκριμένο πρόβλημα, καθώς εξισορροπεί την ανάγκη να εντοπιστούν όλες οι αλλαγές συγγραφικού ύφους χωρίς να αυξάνονται οι ψευδώς θετικές προβλέψεις. Επιπλέον, οι μετρικές Precision και Recall παρέχουν επιπλέον πληροφορίες για τη συμπεριφορά του μοντέλου σε διαφορετικά σενάρια, π.χ. για κείμενα με λίγες ή πολλές αλλαγές συγγραφέα.

5.4 Ρυθμίσεις πειραμάτων

Η εκπαίδευση των μοντέλων πραγματοποιήθηκε με τη χρήση του πλαισίου **PyTorch Lightning**, το οποίο παρέχει ευκολία στη διαχείριση πειραμάτων και αυτοματοποιεί βασικές λειτουργίες (logging, early stopping, checkpointing).

- **Απώλεια (Loss function):** Για τις παραλλαγές που βασίζονται σε μέτρα ομοιότητας εφαρμόστηκε η **CosineEmbeddingLoss**, ενώ για τις παραλλαγές με classification head χρησιμοποιήθηκε η **CrossEntropyLoss** για το **SBERT** μοντέλο. Χρησιμοποιήθηκε

CosineEmbeddingLoss ενώ για τις παραλλαγές με classification head χρησιμοποιήθηκε η **CrossEntropyLoss** σε συνδιασμό με την για το LUAR μοντέλο.

Cross Entropy	**Cosine Similarity**	**Contrasive**
Η cross-entropy loss χρησιμοποιείται σε **μοντέλα βασισμένα σε ταξινόμηση**, όπου ο στόχος είναι η πρόβλεψη του αν υπάρχει αλλαγή συγγραφέα μεταξύ διαδοχικών παραγράφων. Μετρά τη διαφορά μεταξύ της **προβλεπόμενης κατανομής πιθανοτήτων** του μοντέλου (π.χ. "αλλαγή συγγραφέα" vs "ίδιος συγγραφέας") και της πραγματικής ετικέτας.	Η cosine similarity loss χρησιμοποιείται σε μοντέλα που εξάγουν **διανυσματικές εκπροσωπήσεις (embeddings)** αντί για πιθανότητες κλάσεων. Ενθαρρύνει τα embeddings **παραγράφων του ίδιου συγγραφέα** να είναι κοντά στον χώρο των διανυσμάτων, ενώ τα embeddings **διαφορετικών συγγραφέων** να απομακρύνονται.	Η contrastive loss εφαρμόζεται σε **Siamese ή μοντέλα με metric learning**, όπου λαμβάνονται υπόψη **θετικά (ίδιος συγγραφέας)** και **αρνητικά (διαφορετικός συγγραφέας)** ζεύγη παραγράφων. Ενθαρρύνει τα embeddings **θετικών ζευγών** να συγκλίνουν και τα embeddings **αρνητικών ζευγών** να διαχωρίζονται κατά τουλάχιστον ένα **όριο (margin)**

- **Batch size:** Ορίστηκε σε 16 (με ένα πρόβλημα ανά batch), λαμβάνοντας υπόψη τους περιορισμούς μνήμης GPU.
- **Epochs:** Η εκπαίδευση ολοκληρωνόταν με βάση κριτήρια early stopping στο validation set, με μέγιστο αριθμό 5 epochs.
- **Υλοποίηση labels:** Για κάθε κείμενο, δημιουργήθηκαν ετικέτες binary αλλαγής/μη αλλαγής συγγραφέα ανά μετάβαση παραγράφου, οι οποίες αποτέλεσαν τη βάση για τον υπολογισμό του F1-score.

Με τον τρόπο αυτό διατηρείται συνέπεια στην εκπαίδευση όλων των παραλλαγών, διευκολύνοντας την άμεση σύγκριση των αποτελεσμάτων.

5.5 Αποτελέσματα Κατά Μοντέλο

5.5.1 SBERT

- Η εκδοχή **concat (concatenation+classification head)** είχε γενικά καλύτερες επιδόσεις από την cosine, με υψηλά F1 σε PAN24 easy (fn: 0.8272, frozen: 0.7859) και PAN23 Dataset1 (fn: 0.7808, frozen: 0.7743).

- Η εκδοχή **cosine similarity** έδωσε ασταθή αποτελέσματα: σε ορισμένα datasets (π.χ. PAN23 Dataset2) εμφάνισε σχετικά καλά F1 (0.6441), αλλά αλλού (π.χ. PAN22 Dataset2 frozen) είχε χαμηλά F1 (0.3811).
- Το SBERT φάνηκε πιο σταθερό από το LUAR στις concat εκδοχές, αλλά λιγότερο ισχυρό στις cosine.

Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan21	fn	concat	0.5904	0.5901	0.5911	0.5916
pan21	fn	cosine	0.5419	0.3704	0.5774	0.5050
pan21	frozen	concat	0.6122	**0.6031**	0.6091	0.6042
pan21	frozen	cosine	0.5390	0.3538	0.5864	0.5010

Πίνακας 6: Αναλυτικά Αποτελέσματα στο Dataset PAN21 με χρήση μοντέλου SBERT

Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan22 Dataset 1	frozen	concat	0.8823	0.6839	0.7844	0.6472
pan22 Dataset1	fn	concat	0.8720	**0.7190**	0.7342	0.7067
pan22 Dataset1	fn	cosine	0.8260	0.4848	0.4965	0.4987
pan22 Dataset1	frozen	cosine	0.4142	0.3793	0.5023	0.5044
pan22 Dataset 2	fn	concat	0.6224	0.5894	0.5942	0.5883
pan22 Dataset 2	frozen	concat	0.6450	**0.5996**	0.6167	0.5992
pan22 Dataset2	fn	cosine	0.6160	0.3926	0.5667	0.5026
pan22 Dataset2	frozen	cosine	0.6158	0.3811	0.3078	0.5000
pan22 Dataset 3	fn	concat	0.5562	**0.5559**	0.5582	0.5586
pan22 Dataset 3	frozen	concat	0.5865	0.5412	0.5859	0.5619
pan22 Dataset3	fn	cosine	0.5497	0.3731	0.5646	0.5042
pan22 Dataset3	frozen	cosine	0.5497	0.3701	0.5732	0.5039

Πίνακας 7: Αναλυτικά Αποτελέσματα στο Dataset PAN22 με χρήση μοντέλου SBERT

Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan23 Dataset1	fn	concat	0.8986	**0.7808**	0.7734	0.7890
pan23 Dataset1	fn	cosine	0.8743	0.5664	0.7229	0.5541
pan23 Dataset1	frozen	concat	0.8993	0.7743	0.7766	0.7721
pan23 Dataset1	frozen	cosine	0.8764	0.6056	0.7284	0.5819
pan23 Dataset2	fn	concat	0.6895	0.6728	0.6840	0.6707
pan23 Dataset2	fn	cosine	0.6458	0.6441	0.6825	0.6710
pan23 Dataset2	frozen	concat	0.7029	**0.6933**	0.6959	0.6919
pan23 Dataset2	frozen	cosine	0.6346	0.6308	0.6836	0.6644
pan23 Dataset3	fn	concat	0.5809	0.5796	0.5864	0.5845
pan23 Dataset3	fn	cosine	0.5995	0.5809	0.6401	0.6100

pan23 Dataset3	frozen	concat	0.6121	**0.6101**	0.6110	0.6101
pan23 Dataset3	frozen	cosine	0.5682	0.5207	0.6526	0.5834

Πίνακας 8: Αναλυτικά Αποτελέσματα στο Dataset PAN23 με χρήση μοντέλου SBERT

Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan24 easy	fn	concat	0.9304	**0.8272**	0.8617	0.8007
pan24 easy	fn	cosine	0.8760	0.4795	0.9379	0.5064
pan24 easy	frozen	concat	0.9228	0.7859	0.8772	0.7377
pan24 easy	frozen	cosine	0.8744	0.4665	0.4372	0.5000
pan24 hard	fn	concat	0.6031	0.6031	0.6058	0.6056
pan24 hard	fn	cosine	0.6361	**0.6287**	0.6652	0.6465
pan24 hard	frozen	concat	0.6184	0.6184	0.6199	0.6201
pan24 hard	frozen	cosine	0.6157	0.5975	0.6694	0.6302
pan24 medium	fn	cosine	0.6514	0.6035	0.7363	0.6369
pan24 medium	frozen	concat	0.7296	**0.7291**	0.7291	0.7293
pan24 medium	frozen	cosine	0.6533	0.5979	0.7669	0.6377
pan24 medium	fn	concat	0.7011	0.6995	0.7007	0.6993

Πίνακας 9: Αναλυτικά Αποτελέσματα στο Dataset PAN24 με χρήση μοντέλου SBERT

5.5.2 LUAR

- Στη ρύθμιση **cosine similarity**, το LUAR πέτυχε τις υψηλότερες τιμές F1 στα PAN22 (π.χ. Dataset1 frozen: 0.9297, fine-tuned: 0.8446).
- Η εκδοχή **concat (concatenation+classification head)** εμφάνισε μέτρια αποτελέσματα, με σχετικά σταθερότητα αλλά χαμηλότερες κορυφές σε σχέση με το cosine.
- Ειδικά στο **PAN24 (hard)**, η εκδοχή LUAR-cosine fine-tuned φτάνει F1=0.7346, υπερέχοντας της concat εκδοχής (F1=0.6233).

Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan21	fn	concat	0.5857	0.5711	0.6280	0.6033
pan21	frozen	concat	0.5871	0.5738	0.6268	0.6041
pan21	fn	cosine	0.6061	0.6048	0.6176	0.6140
pan21	frozen	cosine	0.6062	**0.6049**	0.6178	0.6142

Πίνακας 10: Αναλυτικά Αποτελέσματα στο Dataset PAN21 με χρήση μοντέλου LUAR

Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan22 Dataset1	fn	concat	0.8508	0.4628	0.4542	0.4965
pan22 Dataset1	frozen	concat	0.8504	0.4596	0.4290	0.4948
pan22 Dataset1	fn	cosine	0.9128	0.8446	0.8082	0.9032
pan22 Dataset1	frozen	cosine	0.9634	**0.9297**	0.9028	0.9634

pan22 Dataset2	fn	concat	0.6352	0.5511	0.6034	0.5653
pan22 Dataset2	frozen	concat	0.6356	0.5565	0.6036	0.5683
pan22 Dataset2	fn	cosine	0.6100	**0.6059**	0.6135	0.6198
pan22 Dataset2	frozen	cosine	0.6100	**0.6059**	0.6135	0.6198
pan22 Dataset3	fn	concat	0.5611	0.5536	0.5548	0.5538
pan22 Dataset3	frozen	concat	0.5613	0.5521	0.5543	0.5527
pan22 Dataset3	fn	cosine	0.5805	**0.5793**	0.5939	0.5909
pan22 Dataset3	frozen	cosine	0.5800	0.5787	0.5933	0.5903

Πίνακας 11: Αναλυτικά Αποτελέσματα στο Dataset PAN22 με χρήση μοντέλου LUAR

Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan23 Dataset1	fn	concat	0.8030	**0.5726**	0.5710	0.5745
pan23 Dataset1	frozen	concat	0.7907	0.4920	0.4920	0.4934
pan23 Dataset1	fn	cosine	0.5339	0.4798	0.5704	0.6548
pan23 Dataset1	frozen	cosine	0.5241	0.4750	0.5736	0.6608
pan23 Dataset2	fn	concat	0.6628	**0.6557**	0.6555	0.6559
pan23 Dataset2	frozen	concat	0.6605	0.6546	0.6541	0.6555
pan23 Dataset2	fn	cosine	0.6606	0.6196	0.6701	0.6270
pan23 Dataset2	frozen	cosine	0.6587	0.6256	0.6603	0.6295
pan23 Dataset3	fn	concat	0.6225	0.6205	0.6215	0.6205
pan23 Dataset3	frozen	concat	0.5333	0.3896	0.5899	0.5128
pan23 Dataset3	fn	cosine	0.7473	0.7257	0.7473	0.7363
pan23 Dataset3	frozen	cosine	0.7491	**0.7273**	0.8258	0.7380

Πίνακας 12: Αναλυτικά Αποτελέσματα στο Dataset PAN23 με χρήση μοντέλου LUAR

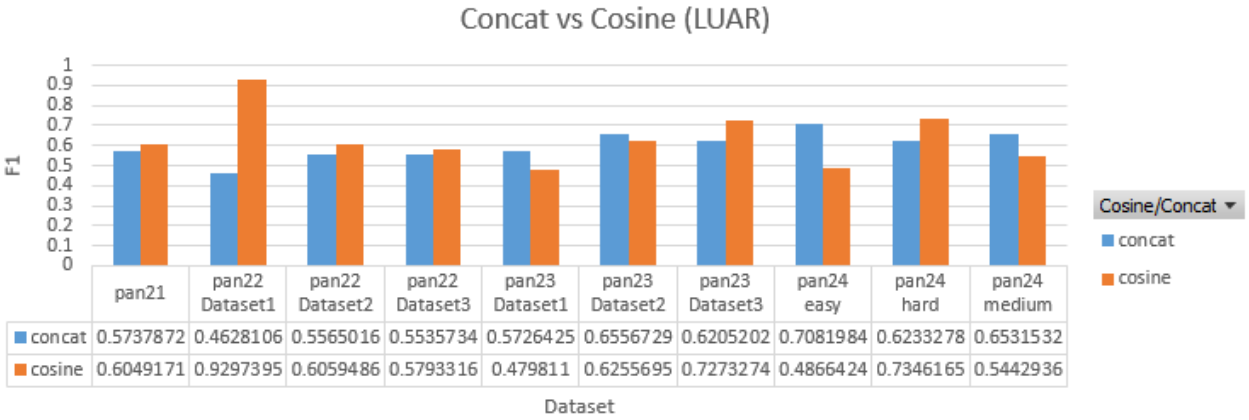
Dataset Name	Model	cosine/concat	accuracy	f1_score	precision	recall
pan24 easy	fn	concat	0.8620	**0.7082**	0.6953	0.7247
pan24 easy	frozen	concat	0.8140	0.4672	0.4612	0.4777
pan24 easy	fn	cosine	0.5412	0.4866	0.5782	0.6763
pan24 easy	frozen	cosine	0.5428	0.4857	0.5742	0.6677
pan24 medium	fn	concat	0.6566	**0.6532**	0.6567	0.6534
pan24 medium	frozen	concat	0.6072	0.5871	0.6164	0.5984
pan24 medium	fn	cosine	0.5581	0.5343	0.5925	0.5697
pan24 medium	frozen	cosine	0.5653	0.5443	0.5985	0.5764
pan24 hard	fn	concat	0.6253	0.6233	0.6237	0.6232
pan24 hard	frozen	concat	0.6201	0.6072	0.6225	0.6116
pan24 hard	fn	cosine	0.7571	**0.7346**	0.8331	0.7426
pan24 hard	frozen	cosine	0.7541	0.7312	0.8299	0.7396

Πίνακας 13: Αναλυτικά Αποτελέσματα στο Dataset PAN22 με χρήση μοντέλου LUAR

5.6 Συγκριτική Ανάλυση

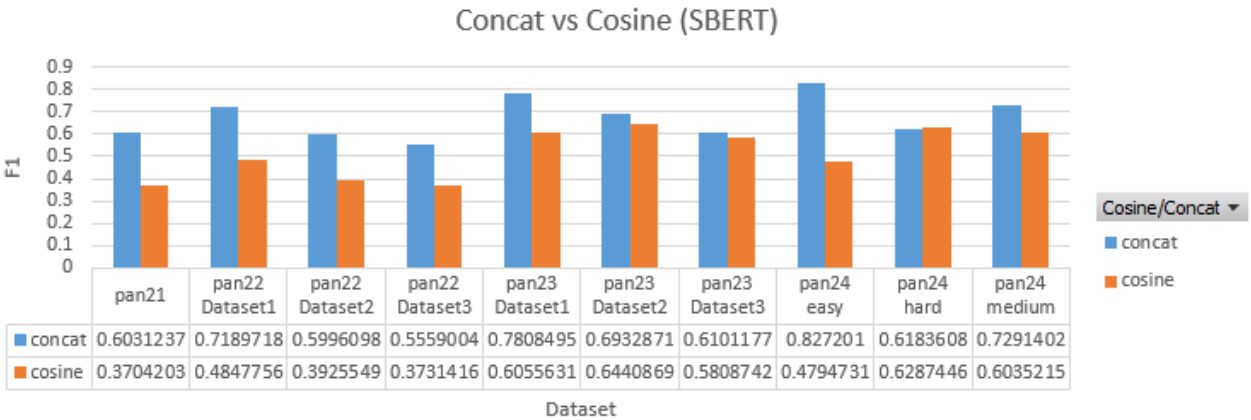
Η σύγκριση LUAR και SBERT δείχνει διαφορετικά προφίλ:

- **LUAR (cosine)** → καλύτερο στις περισσότερες περιπτώσεις, ιδιαίτερα σε PAN22 και PAN24 (hard).
- **SBERT (concat)** → αποδίδει πιο σταθερά, με υψηλές τιμές σε PAN23 και PAN24 easy.
- **Frozen vs Fine-tuned:** Το fine-tuning βελτίωσε την απόδοση σε ορισμένα datasets (π.χ. LUAR cosine σε PAN22), αλλά σε άλλα δεν προσέφερε σημαντικό όφελος ή έφερε πτώση (πιθανό overfitting).
- **Cosine vs Concat:** Για το LUAR η cosine εκδοχή υπερέχει, ενώ για το SBERT καλύτερα αποτελέσματα είχε η concat εκδοχή.



	pan21	pan22 Dataset1	pan22 Dataset2	pan22 Dataset3	pan23 Dataset1	pan23 Dataset2	pan23 Dataset3	pan24 easy	pan24 hard	pan24 medium
concat	0.5737872	0.4628106	0.5565016	0.5535734	0.5726425	0.6556729	0.6205202	0.7081984	0.6233278	0.6531532
cosine	0.6049171	0.9297395	0.6059486	0.5793316	0.479811	0.6255695	0.7273274	0.4866424	0.7346165	0.5442936

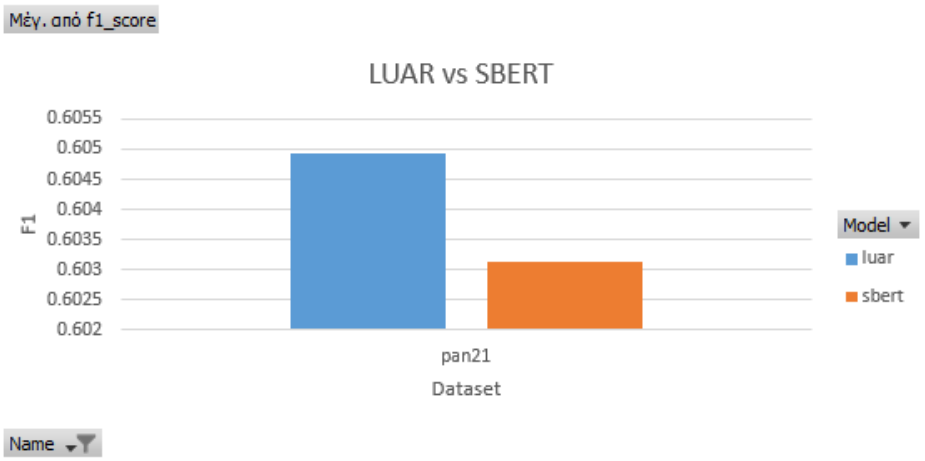
Εικόνα 12: Σύγκριση των Μεθόδων Cosine/Concatenation+Classification Head για το μοντέλο LUAR



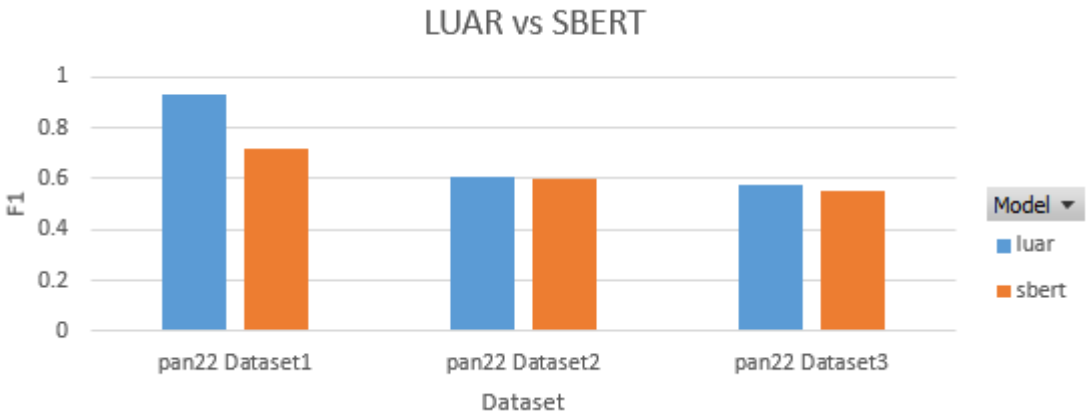
	pan21	pan22 Dataset1	pan22 Dataset2	pan22 Dataset3	pan23 Dataset1	pan23 Dataset2	pan23 Dataset3	pan24 easy	pan24 hard	pan24 medium
concat	0.6031237	0.7189718	0.5996098	0.5559004	0.7808495	0.6932871	0.6101177	0.827201	0.6183608	0.7291402
cosine	0.3704203	0.4847756	0.3925549	0.3731416	0.6055631	0.6440869	0.5808742	0.4794731	0.6287446	0.6035215

Εικόνα 13: Σύγκριση των Μεθόδων Cosine/Concatenation+Classification Head για το μοντέλο SBERT

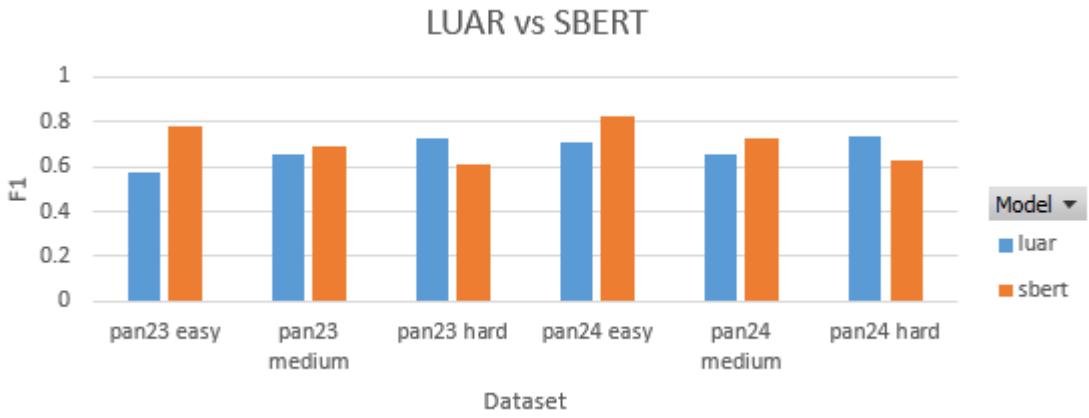
Η σύγκριση μεταξύ SBERT και LUAR δείχνει ότι το LUAR υπερτερεί στα PAN21, PAN23 – Dataset 3 και PAN24 Hard, ενώ υστερεί του SBERT στα PAN23-Dataset , PAN24-Easy/Medium και παρατηρούμε παρόμοιες αποδόσεις στα υπόλοιπα Dataset.



Εικόνα 14: Σύγκριση LUAR/SBERT στις καλύτερες F1 ανά Dataset (PAN21)



Εικόνα 15: Σύγκριση LUAR/SBERT στις καλύτερες F1 ανά Dataset (PAN22)



Εικόνα 16: Σύγκριση LUAR/SBERT στις καλύτερες F1 ανά Dataset (PAN23 & 24)

5.7 Περιορισμοί και Παρατηρήσεις

Η αξιολόγηση στα datasets **PAN21–PAN24** ανέδειξε ότι η απόδοση των μοντέλων διαφοροποιείται αισθητά ανάλογα με το σύνολο δεδομένων και τη μεθοδολογική παραλλαγή. Συνολικά, οι παραλλαγές **SBERT με concatenation** κατέγραψαν τις υψηλότερες επιδόσεις σε αρκετά σύνολα (π.χ. PAN23 Dataset1, PAN24 easy), ενώ οι παραλλαγές **LUAR με cosine similarity** αποδείχθηκαν

ιδιαίτερα αποτελεσματικές σε άλλα (π.χ. PAN22 Dataset1, PAN24 hard). Αυτό υποδεικνύει ότι καμία μεμονωμένη προσέγγιση δεν υπερέχει συστηματικά σε όλα τα σενάρια.

Παράλληλα, παρατηρούνται **μεγάλες διακυμάνσεις μεταξύ υποσυνόλων**. Ενδεικτικά, στο **PAN23 Dataset3** οι παραλλαγές με cosine similarity υπερτερούν σημαντικά έναντι εκείνων με classification head, ενώ στο **PAN22 Dataset1** επιτυγχάνονται εξαιρετικά υψηλές τιμές F1 (≈0.93) που αντικατοπτρίζουν ευκολότερες συνθήκες διαχωρισμού συγγραφέων. Αντίθετα, σε πιο «δύσκολα» σύνολα, όπως το PAN24 medium, οι αποδόσεις παραμένουν μέτριες και πιο ισορροπημένες.

Οι κύριοι περιορισμοί που αναδείχθηκαν σχετίζονται με:

- την **ετερογένεια των datasets** και των domains, που δυσχεραίνει τη γενίκευση·
- την **ευαισθησία στη μεθοδολογική επιλογή** (cosine vs concat, frozen vs fine-tuned)·
- την πιθανή επίδραση παραγόντων όπως **class imbalance** και μικρό μέγεθος δειγμάτων.

Συνολικά, τα αποτελέσματα καταδεικνύουν ότι η επιλογή μεθοδολογίας πρέπει να προσαρμόζεται στα ιδιαίτερα χαρακτηριστικά κάθε dataset, ενώ απαιτείται περαιτέρω διερεύνηση (π.χ. error analysis, cross-domain δοκιμές) για να επιβεβαιωθεί η ικανότητα γενίκευσης των μοντέλων.

5.8 Συνοπτικά Συμπεράσματα Αποτελεσμάτων

Η ανάλυση των αποτελεσμάτων ανέδειξε ορισμένες σαφείς τάσεις. Το μοντέλο **LUAR**, όταν συνδυάζεται με μετρικές ομοιότητας και ειδικότερα με το **cosine similarity**, πέτυχε τις υψηλότερες επιδόσεις, με χαρακτηριστικά παραδείγματα το **PAN22** και το σενάριο **hard του PAN24**. Αντίθετα, το **SBERT** σε συνδυασμό με **concatenation των embeddings και ταξινομητή** παρουσίασε μεγαλύτερη σταθερότητα, αποδίδοντας ιδιαίτερα καλά στα datasets του **PAN23** και στο **easy σενάριο του PAN24**.

Αξιοσημείωτο είναι ότι ο παράγοντας **frozen έναντι fine-tuned embeddings** δεν είχε πάντοτε θετική επίδραση· στα μικρότερα datasets μάλιστα παρατηρήθηκαν ενδείξεις υπερπροσαρμογής, γεγονός που περιόρισε την ικανότητα γενίκευσης. Τέλος, η επιλογή της μεθοδολογικής προσέγγισης για την τελική πρόβλεψη αποδείχθηκε καθοριστική: το **LUAR ευνοείται ξεκάθαρα από τη χρήση cosine similarity**, ενώ το **SBERT εμφανίζεται πιο αποδοτικό με τη μέθοδο concatenation και ταξινομητή**.

5.9 Συγκριτικά με τα αποτελέσματα του διαγωνισμού

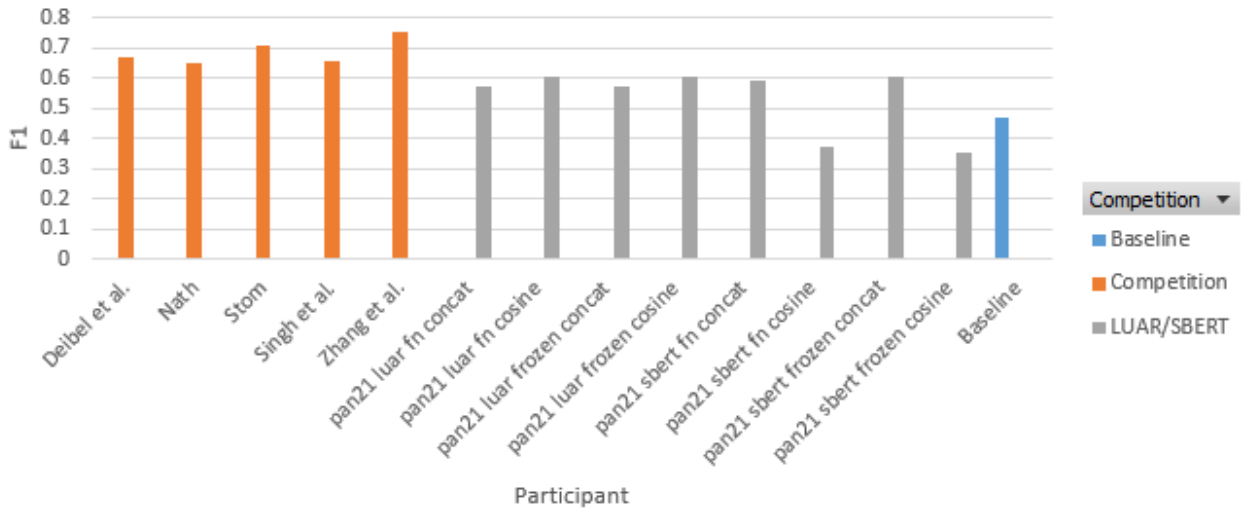
PAN21

Participant	Results (F1)
Zhang et al.	**0.751**
Stom	0.707
Singh et al.	0.657
Deibel et al.	0.669
Nath	0.647
pan21 luar fn concat	0.5711
pan21 luar frozen concat	0.5738

pan21 sbert fn concat	0.5901
pan21 sbert frozen concat	0.6031
pan21 luar fn cosine	0.6048
pan21 luar frozen cosine	0.6049
pan21 sbert fn cosine	0.3704
pan21 sbert frozen cosine	0.3538
Baseline	0.47

Πίνακας 14: Αποτελέσματα Διαγωνσιμου και μεθόδων της εργασίας για PAN21

Στον διαγωνισμό του 2021, οι καλύτερες συμμετοχές (π.χ. Zhang et al., Stom) πέτυχαν F1-score στην περιοχή **0.65–0.75**. Τα δικά μας μοντέλα κατέγραψαν αποτελέσματα από **0.35 έως 0.60**, με τις παραλλαγές **LUAR (cosine)** και **SBERT (concat)** να επιτυγχάνουν τις καλύτερες επιδόσεις (~0.60), ξεπερνώντας το baseline (0.47) αλλά υπολειπόμενες των κορυφαίων προσεγγίσεων.



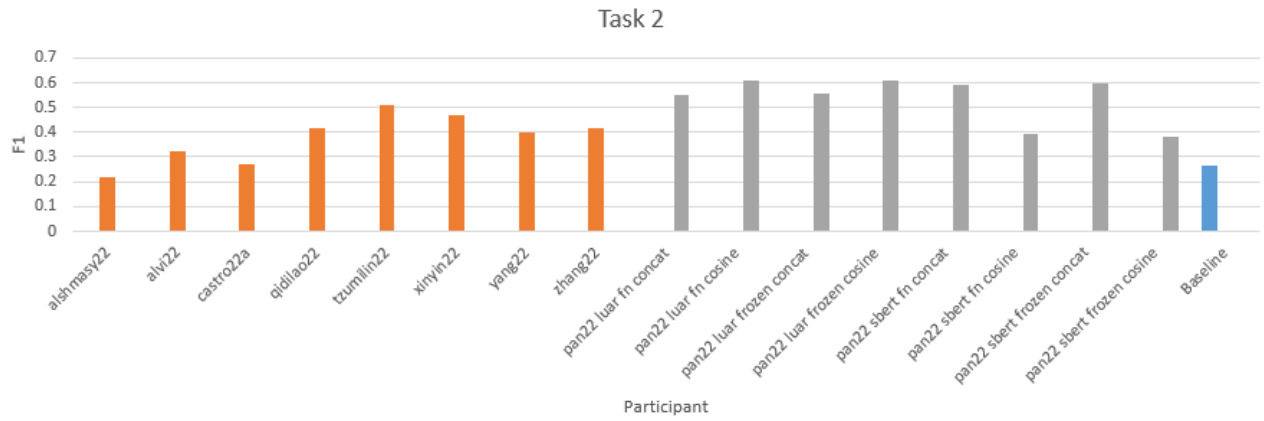
Εικόνα 17: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN21

Πίνακας 15: Αποτελέσματα Διαγωνσιμου και μεθόδων της εργασίας για PAN22

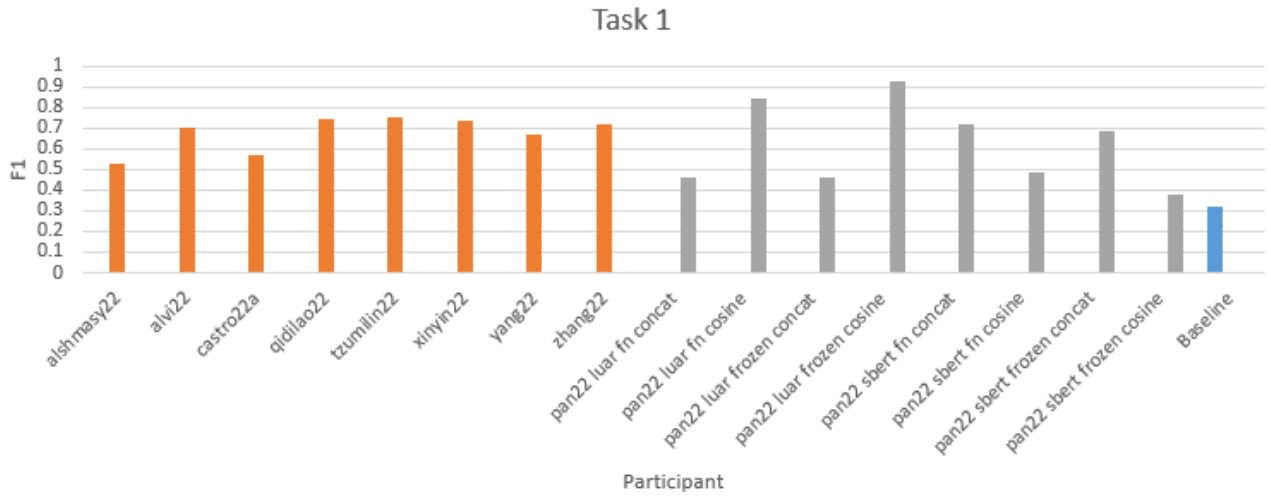
Participant	**Task1**	**Task2**	**Task3**
tzumilin22	**0.754**	**0.51**	**0.7156**
xinyin22	0.7346	0.4687	0.672
qidilao22	0.7471	0.417	0.641
zhang22	0.7162	0.4174	0.6581
yang22	0.669	0.4011	0.6483
alvi22	0.7052	0.3213	0.5636
castro22a	0.5661	0.2735	0.5565
alshmasy22	0.5272	0.2207	0.4995
pan22 luar fn concat	0.4628	0.5511	0.5536
pan22 luar frozen concat	0.4596	0.5565	0.5521
pan22 sbert fn concat	0.7190	0.5894	0.5559
pan22 sbert frozen concat	0.6839	0.5996	0.5412
pan22 luar fn cosine	0.8446	**0.6059**	0.5793

pan22 luar frozen cosine	**0.9297**	**0.6059**	0.5787
pan22 sbert fn cosine	0.4848	0.3926	0.3731
pan22 sbert frozen cosine	0.3793	0.3811	0.3701
Baseline	0.3222	0.2651	0.4809

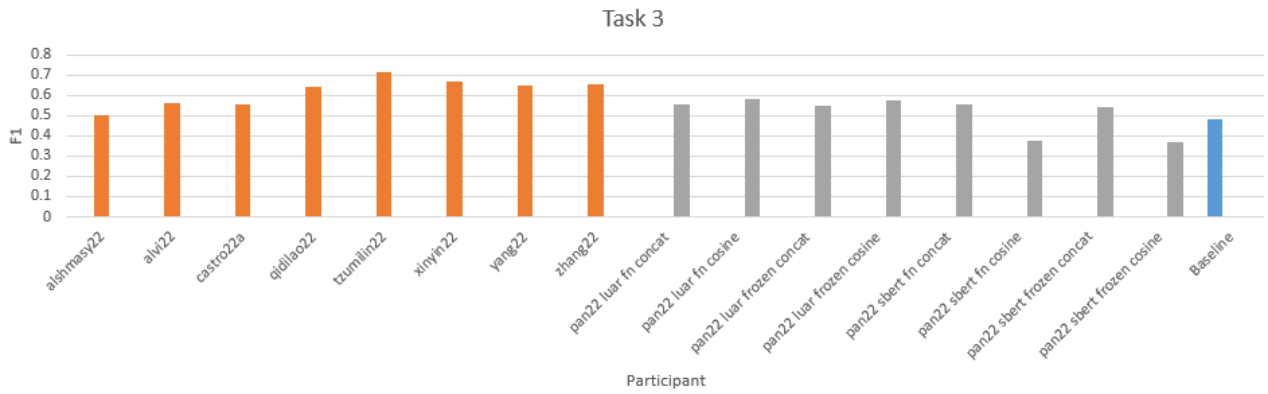
Στο διαγωνιστικό πλαίσιο του 2022, οι κορυφαίοι συμμετέχοντες κατέγραψαν τιμές F1 έως και **0.75** (Task 1), ~**0.51** (Task 2) και ~**0.71** (Task 3). Οι παραλλαγές μας με **LUAR + cosine similarity** πέτυχαν ιδιαίτερα υψηλά αποτελέσματα (έως **0.93** στο Task 1), συγκρίσιμα ή και ανώτερα από τις κορυφαίες συμμετοχές. Αντίθετα, οι παραλλαγές με **SBERT cosine** και **LUAR concat** παρουσίασαν χαμηλότερες επιδόσεις. Το εύρημα αυτό δείχνει ότι το LUAR, σε συνδυασμό με μέτρα ομοιότητας, είναι εξαιρετικά αποδοτικό στο συγκεκριμένο dataset.



Εικόνα 18: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN22 Dataset 1



Εικόνα 19: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN22 Dataset 2

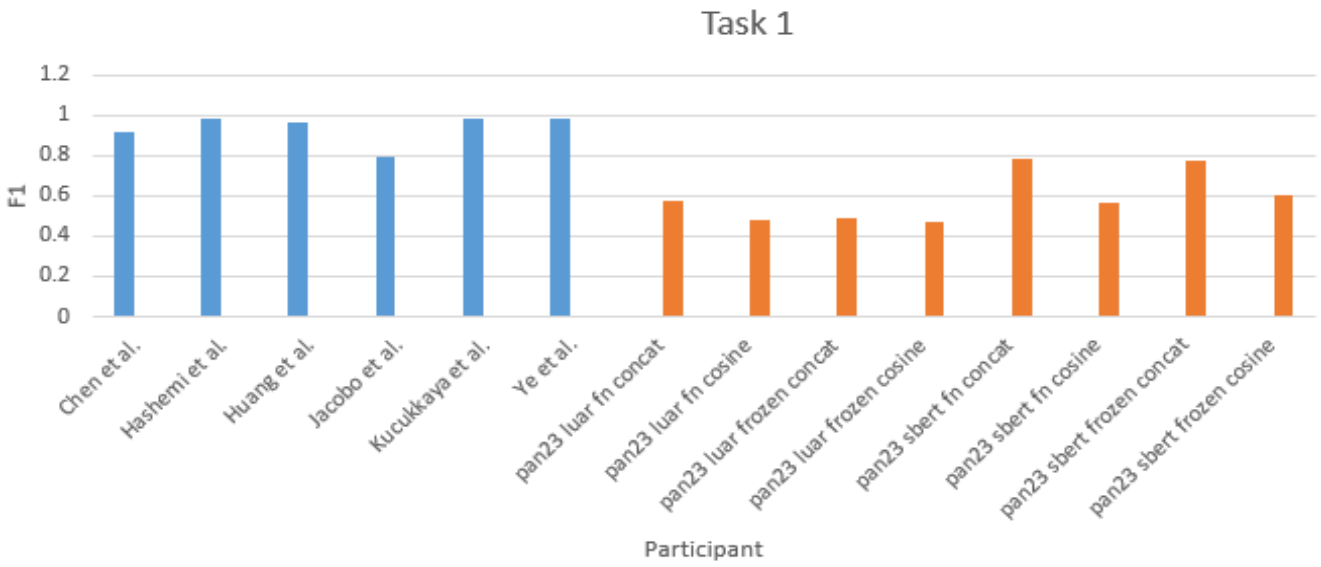


Εικόνα 20: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN22 Dataset 3

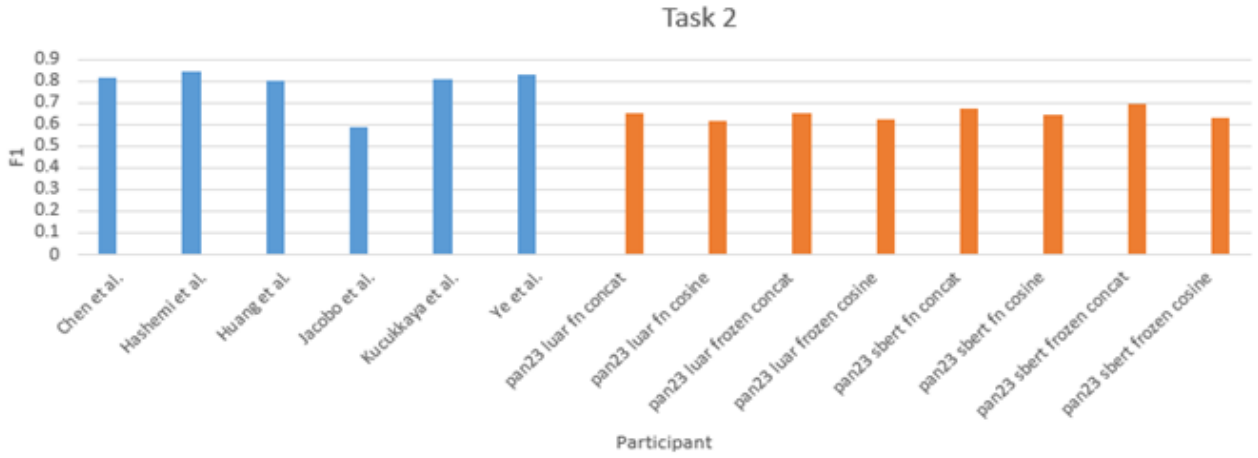
Πίνακας 16: Αποτελέσματα Διαγωνσιμου και μεθόδων της εργασίας για PAN23

Participant	Easy	Medium	Hard
Chen et al.	0.914	0.82	0.676
Hashemi et al.	**0.984**	**0.843**	0.812
Huang et al.	0.968	0.806	0.769
Jacobo et al.	0.793	0.591	0.498
Kucukkaya et al.	0.982	0.81	0.772
Ye et al.	0.983	0.83	**0.821**
pan23 luar fn concat	0.5726	0.6557	0.6205
pan23 luar frozen concat	0.4920	0.6546	0.3896
pan23 sbert fn concat	0.7808	0.6728	0.5796
pan23 sbert frozen concat	0.7743	0.6933	0.6101
pan23 luar fn cosine	0.4798	0.6196	0.7257
pan23 luar frozen cosine	0.4750	0.6256	0.7273
pan23 sbert fn cosine	0.5664	0.6441	0.5809
pan23 sbert frozen cosine	0.6056	0.6308	0.5207

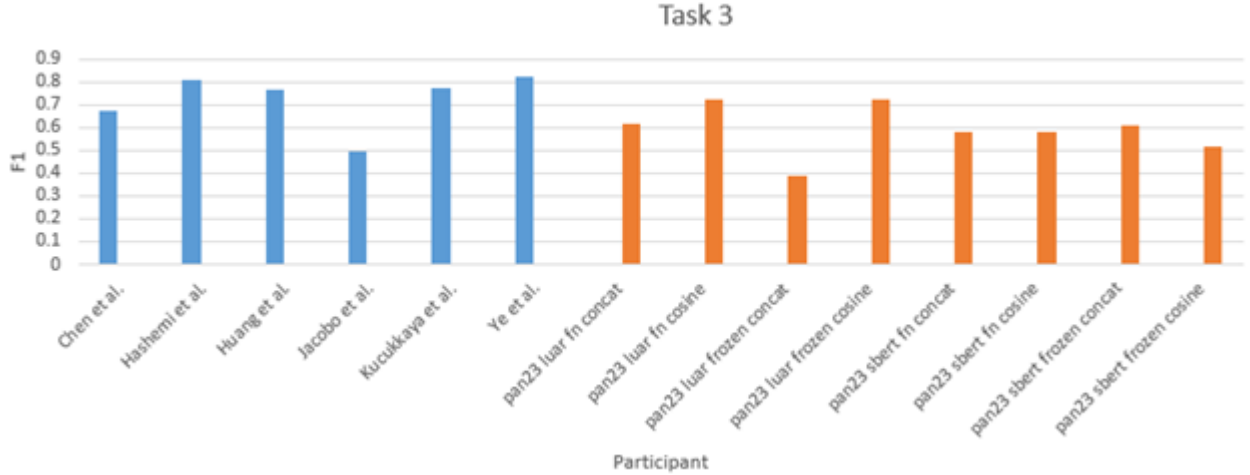
Για το 2023, οι καλύτερες συμμετοχές πέτυχαν επιδόσεις άνω του **0.90** (easy) και >**0.80** (medium). Τα δικά μας αποτελέσματα κυμάνθηκαν σε χαμηλότερα επίπεδα: **0.57–0.78** για τις παραλλαγές SBERT (concat), ενώ οι LUAR παραλλαγές κινήθηκαν γενικά χαμηλότερα, ειδικά στο easy σενάριο. Εντούτοις, σε σενάρια αυξημένης δυσκολίας (hard), κάποιες εκδοχές LUAR (cosine) κατέγραψαν ανταγωνιστικές επιδόσεις (0.72–0.73), γεγονός που δείχνει ότι η προσέγγιση μπορεί να αποδίδει καλύτερα σε πιο απαιτητικές περιπτώσεις.



Εικόνα 21: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN22 Dataset 1



Εικόνα 22: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN22 Dataset 2



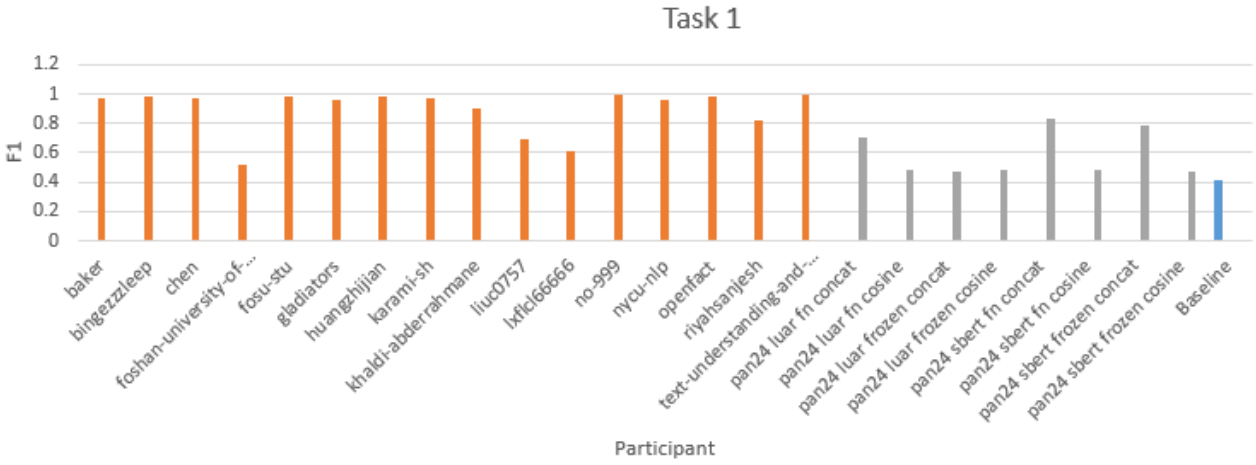
Εικόνα 23: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN22 Dataset 3

Πίνακας 17: Αποτελέσματα Διαγωνσιμου και μεθόδων της εργασίας για PAN24

Participant	**Easy**	**Medium**	**Hard**
fosu-stu	0.987	**0.887**	0.834
nycu-nlp	0.964	0.857	**0.863**
no-999	**0.991**	0.83	0.832
huangzhijian	0.985	0.815	0.826
text-understanding-and-analysi	**0.991**	0.815	0.818
bingezzzleep	0.985	0.818	0.807
openfact	0.981	0.821	0.805
chen	0.968	0.822	0.807
baker	0.976	0.816	0.77
gladiators	0.956	0.809	0.783
khaldi-abderrahmane	0.905	0.806	0.641
karami-sh	0.972	0.664	0.642
riyahsanjesh	0.825	0.712	0.599
liuc0757	0.696	0.717	0.503
lxflcl66666	0.606	0.455	0.484
foshan-university-of-guangdong	0.517	0.394	0.352
pan24 luar fn concat	0.7082	0.6233	0.6532
pan24 luar frozen concat	0.4672	0.6072	0.5871
pan24 sbert fn concat	0.8272	0.6031	0.6996
pan24 sbert frozen concat	0.7859	0.6184	0.7291
pan24 luar fn cosine	0.4866	0.7346	0.5343
pan24 luar frozen cosine	0.4857	0.7312	0.5443
pan24 sbert fn cosine	0.4795	0.6287	0.6035
pan24 sbert frozen cosine	0.4665	0.5975	0.5979
Baseline	0.414	0.506	0.495

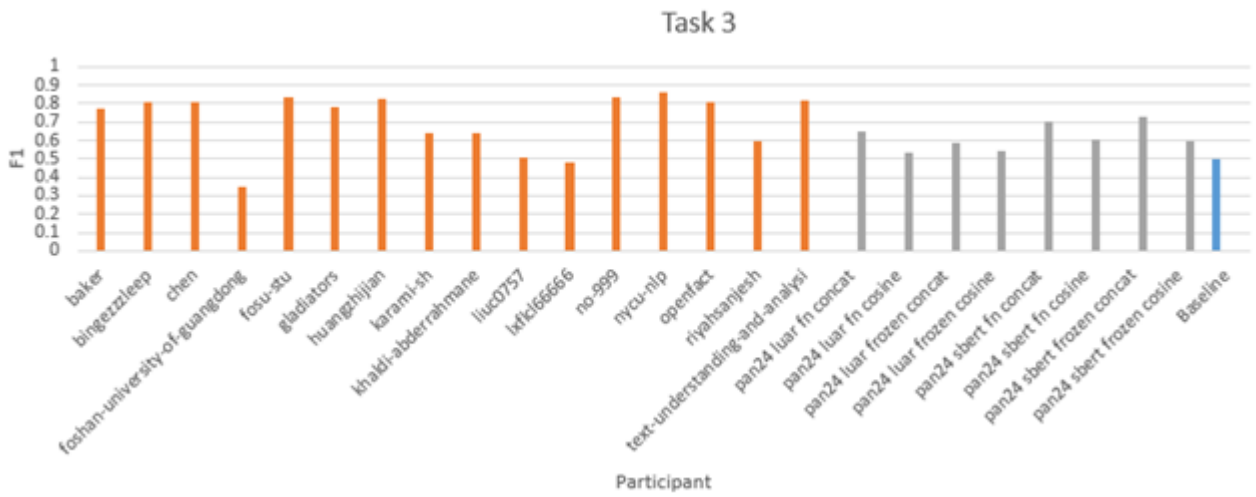
Στον πιο πρόσφατο διαγωνισμό (2024), οι κορυφαίοι συμμετέχοντες κατέγραψαν εντυπωσιακές επιδόσεις (**F1 > 0.95** στο easy). Τα δικά μας αποτελέσματα, αν και σαφώς χαμηλότερα, δείχνουν συνεπή βελτίωση σε σχέση με το baseline (0.41–0.50). Οι παραλλαγές **SBERT fn concat** ξεχώρισαν με επίδοση ~**0.82** στο easy σενάριο, πλησιάζοντας τις μεσαίες επιδόσεις των κορυφαίων

ομάδων. Σε πιο δύσκολες ρυθμίσεις (medium/hard), οι αποδόσεις κυμάνθηκαν στο **0.60–0.73**, αναδεικνύοντας τις προκλήσεις που παραμένουν ανοιχτές.



Εικόνα 24: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN24 Easy

Εικόνα 25: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN24 Medium



Εικόνα 26: Αποτελέσματα Διαγωνισμού και των μεθόδων της εργασίας για PAN24 Hard

Συνολικά, τα αποτελέσματα δείχνουν ότι τα πειραματικά μας μοντέλα:

- **Υπερτερούν σαφώς του baseline** σε σχεδόν όλα τα datasets.
- **Είναι ανταγωνιστικά** σε ορισμένα σενάρια (ιδίως PAN22 Task 1).
- **Υπολείπονται των κορυφαίων προσεγγίσεων** σε πολυπλοκότερα datasets (PAN23, PAN24), γεγονός που υπογραμμίζει την ανάγκη για περαιτέρω βελτιώσεις (π.χ. καλύτερη αξιοποίηση fine-tuning).

6

Συμπεράσματα

6.1 Επισκόπηση

Η παρούσα εργασία είχε ως στόχο τη διερεύνηση της ανίχνευσης αλλαγών συγγραφικού ύφους σε επίπεδο παραγράφου, αξιοποιώντας σύγχρονες αναπαραστάσεις κειμένου βασισμένες σε προεκπαιδευμένα γλωσσικά μοντέλα. Για τον σκοπό αυτό, υλοποιήθηκαν και αξιολογήθηκαν οκτώ διαφορετικές παραλλαγές που συνδύαζαν δύο μοντέλα (SBERT και LUAR) με διαφορετικές μεθοδολογίες υπολογισμού ομοιότητας ή ταξινόμησης (cosine similarity και concatenation με classification head), καθώς και με δύο ρυθμίσεις εκπαίδευσης (frozen και fine-tuned embeddings). Η αξιολόγηση πραγματοποιήθηκε σε τέσσερα σύνολα δεδομένων από τους διαγωνισμούς PAN21 έως PAN24, προσφέροντας μια αντιπροσωπευτική εικόνα της προόδου στο πεδίο.

6.2 Συμπεράσματα από τα Πειράματα

Τα πειραματικά αποτελέσματα ανέδειξαν σαφείς τάσεις και διαφοροποιήσεις μεταξύ των προσεγγίσεων. Το **LUAR**, όταν συνδυάστηκε με **cosine similarity**, εμφάνισε σταθερά υψηλές επιδόσεις στα περισσότερα datasets, ιδίως στα πιο απαιτητικά σενάρια όπως το *PAN22 Dataset 1* και το *PAN24 hard*. Η αρχιτεκτονική του, η οποία δίνει έμφαση στη μάθηση αναπαραστάσεων με βάση τη σημασιολογική συνοχή σε επίπεδο συγγραφέα, αποδείχθηκε αποτελεσματική στην απομόνωση του ύφους ανεξάρτητα από θεματικές μεταβολές.

Αντίθετα, το **SBERT** υπερείχε στις εκδοχές με **concatenation και classification head**, προσφέροντας μεγαλύτερη σταθερότητα σε σύνολα με πιο ομοιογενή θεματολογία, όπως το *PAN23 Dataset 1* και το *PAN24 easy*. Η ταξινομητική προσέγγιση φαίνεται να ωφελεί την απόδοση σε περιπτώσεις όπου οι διαφορές ύφους είναι πιο έντονες και μπορούν να αποτυπωθούν μέσω εκμάθησης διακριτικών ορίων.

Όσον αφορά τον τρόπο εκπαίδευσης, η **fine-tuning** προσέγγιση προσέφερε βελτιωμένα αποτελέσματα σε ορισμένα σενάρια (π.χ. LUAR–cosine στο *PAN22*), αλλά σε άλλα οδήγησε σε υπερπροσαρμογή, ιδιαίτερα σε μικρότερα ή λιγότερο ισορροπημένα datasets. Αυτό υποδεικνύει ότι το fine-tuning είναι ωφέλιμο μόνο όταν υπάρχει επαρκής ποικιλία δεδομένων, διαφορετικά η χρήση frozen embeddings αποδεικνύεται πιο αξιόπιστη.

Συνολικά, η σύγκριση μεταξύ LUAR και SBERT καταδεικνύει ότι **καμία προσέγγιση δεν υπερέχει καθολικά**. Το LUAR εμφανίζεται πιο αποτελεσματικό σε περιπτώσεις όπου η διαφοροποίηση ύφους είναι λεπτή και απαιτεί υψηλή σημασιολογική διάκριση, ενώ το SBERT αποδίδει καλύτερα σε περιβάλλοντα με πιο ευδιάκριτες αλλαγές ύφους ή θεματική μετάβαση.

Τέλος, από την ανάλυση των αποτελεσμάτων προκύπτει ότι το LUAR διαθέτει σημαντικά περιθώρια περαιτέρω βελτίωσης. Με πιο εκτεταμένο fine-tuning, προσεκτική παραμετροποίηση των υπερπαραμέτρων και χρήση τεχνικών όπως το layer-wise unfreezing ή το domain-adaptive pretraining, το μοντέλο θα μπορούσε να ξεπεράσει τις επιδόσεις των κορυφαίων συμμετοχών στους διαγωνισμούς PAN. Η σημασιολογική του δομή το καθιστά ιδιαίτερα κατάλληλο για την εκμάθηση σταθερών αναπαραστάσεων ύφους, εφόσον αξιοποιηθεί σωστά.

6.3 Περιορισμοί και Παρατηρήσεις

Η ανάλυση ανέδειξε ορισμένους περιορισμούς που επηρέασαν την τελική απόδοση των μοντέλων. Πρώτον, η **ετερογένεια των συνόλων δεδομένων** των PAN διαγωνισμών καθιστά δύσκολη τη σύγκριση και τη γενίκευση, καθώς κάθε έτος διαφοροποιείται ως προς τη θεματική ποικιλία, τον αριθμό συγγραφέων και το επίπεδο επιμέλειας των κειμένων. Δεύτερον, η **ανισορροπία κλάσεων** (περισσότερες μη-αλλαγές ύφους από αλλαγές) επηρέασε τη σταθερότητα των ταξινομητικών μοντέλων. Επιπλέον, οι **παραλλαγές με fine-tuning** παρουσίασαν αυξημένη ευαισθησία στις υπερπαραμέτρους, με ενδείξεις υπερπροσαρμογής σε ορισμένα σύνολα (ιδίως στα easy datasets).

Τέλος, οι υλοποιήσεις βασίστηκαν σε paragraph-level ανάλυση, γεγονός που περιορίζει την κατανόηση μακροδομικών αλλαγών ύφους ή συγγραφικής συνέπειας σε εκτενέστερα κείμενα. Μια προσέγγιση που θα συνδύαζε paragraph- και document-level αναπαραστάσεις πιθανότατα θα παρείχε πιο ολοκληρωμένα αποτελέσματα.

6.4 Μελλοντικές Κατευθύνσεις

Η παρούσα εργασία ανοίγει διάφορες κατευθύνσεις για περαιτέρω έρευνα. Ένα πρώτο βήμα θα ήταν η **διεύρυνση της μελέτης** με την ενσωμάτωση και άλλων LLMs (όπως LLaMA, DeBERTa ή Mistral) σε συνδυασμό με τεχνικές low-rank adaptation (LoRA) ή parameter-efficient fine-tuning (PEFT), ώστε να βελτιωθεί η ισορροπία μεταξύ απόδοσης και υπολογιστικού κόστους. Επίσης, η **διερεύνηση cross-domain σεναρίων** (εκπαίδευση σε ένα corpus και δοκιμή σε άλλο) θα μπορούσε να αποκαλύψει σε ποιο βαθμό τα μοντέλα μαθαίνουν καθολικά χαρακτηριστικά ύφους ή εξαρτώνται από θεματικά συμφραζόμενα.

Επιπλέον, η **συνδυαστική αξιοποίηση contrastive learning και continual learning** φαίνεται ιδιαίτερα υποσχόμενη, όπως υποδεικνύεται από τις πρόσφατες επιτυχημένες προσεγγίσεις των διαγωνισμών PAN24. Τέλος, ένα ενδιαφέρον μελλοντικό βήμα θα ήταν η **ερμηνευσιμότητα των μοντέλων**, δηλαδή η ανάλυση των χαρακτηριστικών που συμβάλλουν περισσότερο στη διάκριση συγγραφέων, με στόχο μια πιο διαφανή και ερμηνεύσιμη υφομετρική ανάλυση.

Βιβλιογραφία

[1] J. Daniel and J. H. Martin, “Speech and Language Processing,” 2024.

[2] E. Stamatatos, “A Survey of Modern Authorship Attribution Methods.”

[3] V. A. Oloo, C. Otieno, and L. A. Wanzare, “A Literature Survey on Writing Style Change Detection Based on Machine Learning: State-Of-The-Art - Review,” *International Journal of Computer Trends and Technology*, vol. 70, no. 5, pp. 15–32, May 2022, doi: 10.14445/22312803/ijctt-v70i5p103.

[4] G. Rios-Toledo, J. P. F. Posadas-Duran, G. Sidorov, and N. A. Castro-Sanchez, “Detection of changes in literary writing style using N-grams as style markers and supervised machine learning,” *PLoS One*, vol. 17, no. 7 July, Jul. 2022, doi: 10.1371/journal.pone.0267590.

[5] Andriy. Burkov, *The hundred-page language models book*. True Positive Inc., 2025.

[6] J. Daniel and J. H. Martin, “Speech and Language Processing,” 2024.

[7] J. Daniel and J. H. Martin, “Speech and Language Processing,” 2024.

[8] E. Zangerle, M. Mayerl, M. Potthast, and B. Stein, “Overview of the Style Change Detection Task at PAN 2021,” 2021. [Online]. Available: http://pan.webis.de

[9] E. Zangerle, M. Mayerl, M. Potthast, and B. Stein, “Overview of the Style Change Detection Task at PAN 2022,” 2022. [Online]. Available: http://ceur-ws.org

[10] E. Zangerle, M. Mayerl, M. Potthast, and B. Stein, “Overview of the Multi-Author Writing Style Analysis Task at PAN 2023,” 2023. [Online]. Available: https://pan.webis.de

[11] E. Zangerle, M. Mayerl, M. Potthast, and B. Stein, “Overview of the Multi-Author Writing Style Analysis Task at PAN 2024,” 2024. [Online]. Available: https://pan.webis.de

[12] E. Strøm, “Multi-label Style Change Detection by Solving a Binary Classification Problem Notebook for PAN at CLEF 2021,” 2021. [Online]. Available: https://stackexchange.com/

[13] Z. Zhang *et al.*, “Style Change Detection Based On Writing Style Similarity Notebook for PAN at CLEF 2021,” 2021.

[14] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, and L.-H. Lee, “Ensemble Pre-trained Transformer Models for Writing Style Change Detection,” 2022. [Online]. Available: http://ceur-ws.org

[15] A. Hashemi and W. Shi, “Enhancing Writing Style Change Detection using Transformer-based Models and Data Augmentation Notebook for PAN at CLEF 2023,” 2023. [Online]. Available: http://ceur-ws.org

[16] Z. Ye, C. Zhong, H. Qi, and Y. Han, “Supervised Contrastive Learning for Multi-Author Writing Style Analysis.”

[17] Z. Ye, Y. Zhong, C. Huang, and L. Kong, “Continual Transfer Learning With Progress Prompt for Multi-Author Writing Style Analysis,” 2024.

[18] Z. Huang and L. Kong, “DeBERTa-v3 with R-Drop regularization for Multi-Author Writing Style Analysis,” 2024.

[19] J. Lv, Y. Yi, and H. Qi, “Team fosu-stu at PAN: Supervised Fine-Tuning of Large Language Models for Multi Author Writing Style Analysis,” 2024.

[20] T.-M. Lin, Y.-H. Wu, and L.-H. Lee, “Team NYCU-NLP at PAN 2024: Integrating Transformers with Similarity Adjustments for Multi-Author Writing Style Analysis,” 2024. [Online]. Available: https://huggingface.co/roberta-base

[21] R. A. Rivera-Soto *et al.*, “Learning Universal Authorship Representations,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 913–919. doi: 10.18653/v1/2021.emnlp-main.70.