



Σχολή Θετικών Επιστημών και Τεχνολογίας  
Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα

Διπλωματική Εργασία  
Αντιστοίχιση Εγγραφών (Entity Resolution) με Χρήση  
Νευρωνικών Δικτύων για Δομημένα  
και Ημιδομημένα Κείμενα: Συνδυαστικά Μοντέλα Σημασιολογικής  
Αναπαράστασης και  
Συγχώνευσης Πληροφορίας

Χρήστος Νίκου

Επιβλέπων καθηγητής: Καραπιέρης Δημήτριος

Πάτρα, Ιούνιος 2026

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Χρήστος Νίκου που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης ο συγγραφέας/δημιουργός εκχωρεί στο ΕΑΠ, μη αποκλειστική άδεια χρήσης του δικαιώματος αναπαραγωγής, προσαρμογής, δημόσιου δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσής τους διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος και για όλο το χρόνο διάρκειας των δικαιωμάτων πνευματικής ιδιοκτησίας. Η ανοικτή πρόσβαση στο πλήρες κείμενο για μελέτη και ανάγνωση δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, αποθήκευση, πώληση, εμπορική χρήση, μετάδοση, διανομή, έκδοση, εκτέλεση, «μεταφόρτωση» (downloading), «ανάρτηση» (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού. Ο συγγραφέας/δημιουργός διατηρεί το σύνολο των ηθικών και περιουσιακών του δικαιωμάτων.



Αντιστοίχιση Εγγραφών (Entity Resolution) με Χρήση  
Νευρωνικών Δικτύων για Δομημένα  
και Ημιδομημένα Κείμενα: Συνδυαστικά Μοντέλα Σημασιολογικής  
Αναπαράστασης και  
Συγχώνευσης Πληροφορίας

Χρήστος Νίκου

Επιτροπή Επίβλεψης Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής:

Δημήτριος Καραπιέρης

Συν-Επιβλέπων Καθηγητής:

Γεώργιος Παπαμιχαήλ

Πάτρα, Ιούνιος 2026

*Στην Οικογένεια μου και στον Άγγελο.*

## Περίληψη

Η παρούσα εργασία εξετάζει το πρόβλημα της αντιστοίχισης εγγραφών (Entity Resolution), μια διαδικασία απαραίτητη για την ενοποίηση δεδομένων που προέρχονται από διαφορετικές πηγές. Στόχος είναι ο εντοπισμός εγγραφών που, παρά τις διαφορές στην περιγραφή τους, αναφέρονται στο ίδιο αντικείμενο. Για τον σκοπό αυτό, χρησιμοποιήθηκαν σύγχρονα γλωσσικά μοντέλα βαθιάς μάθησης (Transformers), τα οποία έχουν την ικανότητα να κατανοούν τη σημασία των λέξεων και όχι μόνο τη σύνταξή τους.

Η μελέτη εστιάζει στη δυνατότητα ενός μοντέλου να εκπαιδεύεται σε ένα πεδίο δεδομένων και στη συνέχεια να προσαρμόζεται σε νέα, διαφορετικά πεδία (Domain Adaptation). Συγκεκριμένα, η εκπαίδευση ξεκίνησε από ακαδημαϊκές δημοσιεύσεις και επεκτάθηκε σε περιγραφές προϊόντων από μεγάλες πλατφόρμες ηλεκτρονικού εμπορίου. Κατά τη διαδικασία αυτή, εφαρμόστηκαν τεχνικές για τη διατήρηση της αρχικής γνώσης του μοντέλου και την αντιμετώπιση της ανισορροπίας των δεδομένων, καθώς τα ζεύγη που ταιριάζουν είναι συνήθως πολύ λιγότερα από αυτά που διαφέρουν.

Τα πειραματικά αποτελέσματα έδειξαν ότι τα μεγάλα γλωσσικά μοντέλα επιτυγχάνουν εξαιρετική ακρίβεια, ακόμη και σε "θορυβώδη" δεδομένα με ελλιπείς πληροφορίες. Παράλληλα, εξετάστηκαν μικρότερα και πιο ευέλικτα μοντέλα, τα οποία αποδείχθηκαν ιδιαίτερα αποτελεσματικά, προσφέροντας μια εξαιρετική ισορροπία μεταξύ ταχύτητας και απόδοσης. Η εργασία καταλήγει στο συμπέρασμα ότι η χρήση προηγμένων γλωσσικών μοντέλων μπορεί να αυτοματοποιήσει την ενοποίηση πληροφοριών με υψηλή αξιοπιστία, μειώνοντας την ανάγκη για χειροκίνητους κανόνες.

## Λέξεις - Κλειδιά

Αντιστοίχιση Εγγραφών, Βαθιά Μάθηση, Γλωσσικά Μοντέλα, Μεταφορά Γνώσης, Ποιότητα Δεδομένων, Ηλεκτρονικό Εμπόριο

# Entity Resolution using neural networks for structured and semi-structured texts: Combined models of semantic representation and Intelligence

Christos Nikou

## Abstract

This thesis investigates the problem of Entity Resolution, a fundamental process for integrating data from diverse sources. The objective is to identify records that refer to the same real-world object, despite variations in their descriptions. To achieve this, modern deep learning language models (Transformers) were employed, which are capable of understanding the semantic meaning of text rather than relying solely on syntactic patterns. The study focuses on the ability of a model to be trained in one data domain and subsequently adapt to new, different domains (Domain Adaptation). Specifically, the training began with academic publications and was extended to product descriptions from major e-commerce platforms. Throughout this process, specific techniques were applied to preserve the model's initial knowledge and to address data imbalance, as matching pairs are typically much fewer than non-matching ones.

Experimental results demonstrated that large-scale language models achieve excellent accuracy, even when dealing with "noisy" data containing incomplete information. Furthermore, smaller and more flexible models were evaluated, proving to be highly efficient by offering an excellent balance between processing speed and performance. The thesis concludes that the use of advanced language models can reliably automate information integration, significantly reducing the need for manual rules and human intervention.

## **Keywords**

Entity Resolution, Deep Learning, Language Models, Transfer Learning, Data Quality, E-commerce

## Περιεχόμενα

|                                                                  |            |
|------------------------------------------------------------------|------------|
| <b>Περίληψη</b> .....                                            | <b>v</b>   |
| <b>Abstract</b> .....                                            | <b>vi</b>  |
| <b>Κατάλογος Εικόνων / Σχημάτων</b> .....                        | <b>xii</b> |
| <b>Κατάλογος Πινάκων</b> .....                                   | <b>xiv</b> |
| <b>Συντομογραφίες &amp; Ακρωνύμια</b> .....                      | <b>xv</b>  |
| <b>1 Εισαγωγή</b> .....                                          | <b>1</b>   |
| 1.1 Ορισμός και Σημασία της Αντιστοίχισης Εγγραφών .....         | 1          |
| 1.2 Η Μετάβαση στη Βαθιά Μάθηση (Deep Learning).....             | 2          |
| 1.3 Προ-εκπαιδευμένα Γλωσσικά Μοντέλα και το Σύστημα Ditto ..... | 2          |
| 1.4 Μεγάλα Γλωσσικά Μοντέλα (LLMs) και Σύγχρονες Τάσεις .....    | 3          |
| 1.5 Αντικείμενο και Συνεισφορά της Παρούσας Εργασίας.....        | 3          |
| 1.6 Δομή της Εργασίας .....                                      | 4          |
| <b>2 Θεωρητικό Υπόβαθρο και Σχετική Βιβλιογραφία</b> .....       | <b>4</b>   |
| 2.1 Η Επιστήμη του Entity Resolution (ER) .....                  | 4          |
| 2.1.1 Προκλήσεις και Πολυπλοκότητα .....                         | 5          |
| 2.2 Ιστορική Αναδρομή: Από τους Κανόνες στη Μηχανική Μάθηση..... | 5          |
| 2.2.1 Ντετερμινιστικά και Πιθανολογικά Μοντέλα .....             | 5          |
| 2.2.2 Η Εποχή των String Similarity Metrics.....                 | 5          |
| 2.3 Η Μετάβαση στα Word Embeddings.....                          | 6          |
| 2.4 Η Αρχιτεκτονική Transformer και ο Μηχανισμός Attention ..... | 6          |
| 2.4.1 Ο Μηχανισμός Self-Attention .....                          | 6          |
| 2.4.2 Multi-Head Attention.....                                  | 7          |
| 2.5 Προ-εκπαιδευμένα Μοντέλα Γλώσσας (PLMs).....                 | 7          |
| 2.5.1 BERT: Bidirectional Encoder Representations.....           | 7          |
| 2.5.2 DeBERTa-v3: Decoding-enhanced BERT.....                    | 7          |

|            |                                                                                   |           |
|------------|-----------------------------------------------------------------------------------|-----------|
| 2.5.3      | Η Συμβολή της Αντιθετικής Μάθησης (Contrastive Learning).....                     | 7         |
| <b>2.6</b> | <b>Αρχιτεκτονικές Cross-Encoder για ER .....</b>                                  | <b>8</b>  |
| <b>2.7</b> | <b>Class Imbalance και Weighted Cross Entropy .....</b>                           | <b>8</b>  |
| 2.7.1      | Το πρόβλημα του Majority Bias .....                                               | 8         |
| 2.7.2      | Η Λύση του Weighted Loss .....                                                    | 9         |
| <b>2.8</b> | <b>Continual Learning και Anti-Forgetting .....</b>                               | <b>9</b>  |
| 2.8.1      | Replay Buffer (Experience Replay) .....                                           | 9         |
| <b>2.9</b> | <b>Μετρικές Αξιολόγησης (Metrics).....</b>                                        | <b>10</b> |
| <b>3</b>   | <b>Μεθοδολογία και Υλοποίηση του Συστήματος.....</b>                              | <b>10</b> |
| <b>3.1</b> | <b>Εισαγωγή στη Μεθοδολογία .....</b>                                             | <b>10</b> |
| <b>3.2</b> | <b>Αρχιτεκτονική και Επιλογή Μοντέλων .....</b>                                   | <b>11</b> |
| 3.2.1      | Η Προσέγγιση Cross-Encoder.....                                                   | 11        |
| 3.2.2      | Εξεταζόμενα Μοντέλα (Model Zoo).....                                              | 12        |
| <b>3.3</b> | <b>Προεπεξεργασία και Σειριοποίηση (Serialization) .....</b>                      | <b>12</b> |
| 3.3.1      | Καθαρισμός Κειμένου (Text Cleaning).....                                          | 12        |
| 3.3.2      | Σειριοποίηση με Ειδικούς Δείκτες (Special Tokens) .....                           | 13        |
| 3.3.3      | Η Φάση του Αποκλεισμού (Blocking) στα Σύνολα Δεδομένων.....                       | 14        |
| <b>3.4</b> | <b>Αντιμετώπιση Ανισορροπίας Κλάσεων (Class Imbalance) .....</b>                  | <b>15</b> |
| 3.4.1      | Το Πρόβλημα του Majority Bias.....                                                | 15        |
| 3.4.2      | Ο Μηχανισμός WeightedTrainer .....                                                | 15        |
| <b>3.5</b> | <b>Διαδικασία Εκπαίδευσης και Συνεχής Μάθηση (Continual Learning).....</b>        | <b>16</b> |
| 3.5.1      | Φάση 1: Εκπαίδευση στο Source Domain (DBLP-Scholar).....                          | 16        |
| 3.5.2      | Φάση 2: Domain Adaptation και ο Μηχανισμός Replay Buffer.....                     | 17        |
| <b>3.6</b> | <b>Αξιολόγηση και Βελτιστοποίηση Κατωφλίου (Threshold Tuning).....</b>            | <b>18</b> |
| <b>3.7</b> | <b>Περιβάλλον Υλοποίησης.....</b>                                                 | <b>19</b> |
| <b>3.8</b> | <b>Σχεδιασμός Αυτόνομου Μηχανισμού Αποκλεισμού (Blocking) με ANNS Index .....</b> | <b>19</b> |
| 3.8.1      | Αρχιτεκτονική Bi-Encoder και Παραγωγή Διανυσματικών Αναπαραστάσεων .....          | 20        |
| 3.8.2      | Δομή Προσεγγιστικής Αναζήτησης με FAISS.....                                      | 21        |
| 3.8.3      | Στρατηγική Φιλτραρίσματος και Παραγωγής Υποψήφιων Ζευγών.....                     | 22        |
| 3.8.4      | Πειραματική Αξιολόγηση του Μηχανισμού Blocking .....                              | 22        |

|          |                                                                                                          |           |
|----------|----------------------------------------------------------------------------------------------------------|-----------|
| <b>4</b> | <b>Πειραματική Διαδικασία και Σύνολα Δεδομένων</b>                                                       | <b>24</b> |
| 4.1      | Εισαγωγή στο Πειραματικό Πλαίσιο                                                                         | 24        |
| 4.2      | Σύνολα Δεδομένων (Datasets)                                                                              | 25        |
| 4.2.1    | Το Σύνολο Δεδομένων DBLP-Scholar (Source Domain)                                                         | 25        |
| 4.2.2    | Το Σύνολο Δεδομένων Amazon-Google                                                                        | 26        |
| 4.2.3    | Το Σύνολο Δεδομένων Walmart-Amazon                                                                       | 26        |
| 4.3      | Υπολογιστικό Περιβάλλον και Υλικό (Hardware & Software Setup)                                            | 27        |
| 4.3.1    | Αρχιτεκτονική Apple Silicon και Τεχνολογία MPS                                                           | 27        |
| 4.3.2    | Λογισμικό και Βιβλιοθήκες                                                                                | 28        |
| 4.3.3    | Διασφάλιση Αναπαραγωγιμότητας (Reproducibility)                                                          | 28        |
| 4.4      | Υπερπαράμετροι Εκπαίδευσης (Hyperparameters)                                                             | 29        |
| 4.4.1    | Υπερπαράμετροι Φάσης 1 (Source Domain Training)                                                          | 29        |
| 4.4.2    | Υπερπαράμετροι Φάσης 2 (Domain Adaptation)                                                               | 30        |
| 4.4.3    | Πειραματική Τεκμηρίωση Μειωτικού Συντελεστή Βάρους ( $w_1$ )                                             | 30        |
| 4.5      | Μετρικές Αξιολόγησης (Evaluation Metrics)                                                                | 33        |
| <b>5</b> | <b>Αποτελέσματα και Σχολιασμός</b>                                                                       | <b>33</b> |
| 5.1      | Εισαγωγή στα Αποτελέσματα                                                                                | 33        |
| 5.2      | Απόδοση στο Πηγαίο Πεδίο (Source Domain - DBLP)                                                          | 35        |
| 5.3      | Απόδοση στα Πεδία-Στόχους (Domain Adaptation)                                                            | 38        |
| 5.3.1    | Αποτελέσματα στο σύνολο Amazon-Google                                                                    | 38        |
| 5.3.2    | Αποτελέσματα στο σύνολο Walmart-Amazon                                                                   | 40        |
| 5.4      | Σύγκριση Αποδοτικότητας (Efficiency vs. Accuracy)                                                        | 43        |
| 5.5      | Η Κρίσιμη Συμβολή του Threshold Tuning                                                                   | 43        |
| 5.6      | Η Επιτυχία του Replay Buffer έναντι της Καταστροφικής Λήθης                                              | 45        |
| 5.7      | Ποιοτική Ανάλυση Σφαλμάτων (Error Analysis)                                                              | 46        |
| 5.8      | Αξιολόγηση Αρχιτεκτονικής Bi-Encoder: Μηδενική Μάθηση (Zero-Shot) έναντι<br>Μικρο-ρύθμισης (Fine-Tuning) | 47        |
| 5.8.1    | Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon (Αρχιτεκτονική Bi-Encoder έναντι<br>Cross-Encoder)     | 48        |

|             |                                                                                      |           |
|-------------|--------------------------------------------------------------------------------------|-----------|
| <b>5.9</b>  | <b>Ανάλυση Ανταλλαγής Ακρίβειας και Ανάκλησης (Precision-Recall Trade-off) .....</b> | <b>49</b> |
| 5.9.1       | Απόδοση στο DBLP-Scholar .....                                                       | 49        |
| 5.9.2       | Απόδοση στο Amazon-Google .....                                                      | 50        |
| 5.9.3       | Απόδοση στο Walmart-Amazon .....                                                     | 52        |
| <b>5.10</b> | <b>Ανάλυση Υπολογιστικού Κόστους (Execution Time Analysis) .....</b>                 | <b>53</b> |
| 5.10.1      | Κατανομή κόστους στους Bi-Encoders.....                                              | 53        |
| 5.10.2      | Υπολογιστικό κόστος Cross-Encoder .....                                              | 57        |
| 5.10.3      | Συμπέρασμα σχεδιασμού για συστήματα παραγωγής .....                                  | 58        |
| <b>6</b>    | <b>Συμπεράσματα και Μελλοντική Έρευνα .....</b>                                      | <b>58</b> |
| <b>6.1</b>  | <b>Ανασκόπηση και Σκοπός της Εργασίας .....</b>                                      | <b>58</b> |
| <b>6.2</b>  | <b>Βασικά Ερευνητικά Συμπεράσματα .....</b>                                          | <b>60</b> |
| 6.2.1       | Η Κρισιμότητα της Σειριοποίησης και των Ειδικών Δεικτών (Special Tokens) .....       | 60        |
| 6.2.2       | Η Επίλυση του Majority Bias μέσω Σταθμισμένης Απώλειας (Weighted Loss) .....         | 60        |
| 6.2.3       | Διατήρηση της Γνώσης του Πηγαίου Πεδίου μέσω Μηχανισμών Experience Replay.....       | 61        |
| 6.2.4       | Ακρίβεια έναντι Υπολογιστικού Κόστους (Efficiency vs. Accuracy) .....                | 61        |
| <b>6.3</b>  | <b>Περιορισμοί της Παρούσας Μελέτης .....</b>                                        | <b>62</b> |
| 6.3.1       | Υπολογιστικός Φόρτος του Cross-Encoder .....                                         | 62        |
| 6.3.2       | Έλλειψη Πειραματικής Αξιολόγησης του Blocker .....                                   | 62        |
| 6.3.3       | Περιορισμοί Hardware και Apple Silicon .....                                         | 63        |
| 6.3.4       | Μονόγλωσση Προσέγγιση .....                                                          | 63        |
| <b>6.4</b>  | <b>Κατευθύνσεις Μελλοντικής Έρευνας .....</b>                                        | <b>64</b> |
| 6.4.1       | Αξιοποίηση Μεγάλων Γλωσσικών Μοντέλων (Large Language Models - LLMs) .....           | 64        |
| 6.4.2       | Υβριδικά Συστήματα Bi-Encoder και Cross-Encoder .....                                | 64        |
| 6.4.3       | Ενεργή Μάθηση (Active Learning) .....                                                | 65        |
| 6.4.4       | Πειραματική Αξιολόγηση του Μηχανισμού Blocking .....                                 | 65        |
|             | <b>Βιβλιογραφία .....</b>                                                            | <b>66</b> |

## Κατάλογος Εικόνων / Σχημάτων

|                                                                                                                                               |    |
|-----------------------------------------------------------------------------------------------------------------------------------------------|----|
| Εικόνα 3.1-1 Αρχιτεκτονική Συστήματος .....                                                                                                   | 11 |
| Εικόνα 3.3-1 Διαδικασία Σειριοποίησης.....                                                                                                    | 14 |
| Εικόνα 3.5-1 Μηχανισμός Replay Buffer .....                                                                                                   | 18 |
| Εικόνα 5.2-1 Επιδόσεις των γλωσσικών μοντέλων (F1-Score) στο ακαδημαϊκό σύνολο<br>δεδομένων DBLP. ....                                        | 37 |
| Εικόνα 5.2-2 Σύγκριση Precision, Recall και F1-Score των εξεταζόμενων μοντέλων στο<br>ακαδημαϊκό σύνολο δεδομένων DBLP-Scholar.....           | 37 |
| Εικόνα 5.3-1 Σύγκριση Precision, Recall και F1-Score των εξεταζόμενων μοντέλων στο<br>σύνολο δεδομένων Amazon-Google .....                    | 40 |
| Εικόνα 5.3-2 Σύγκριση Precision, Recall και F1-Score των εξεταζόμενων μοντέλων στο<br>σύνολο δεδομένων Walmart-Amazon .....                   | 42 |
| Εικόνα 5.3-3 Συγκριτικά αποτελέσματα (F1-Score) κατά τη μεταφορά γνώσης στα τρία<br>εξεταζόμενα σύνολα δεδομένων (DBLP, Amazon, Walmart)..... | 42 |
| Εικόνα 5.8-1 Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon (Αρχιτεκτονική Bi-<br>Encoder έναντι Cross-Encoder) .....                      | 48 |
| Εικόνα 5.9-1 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο<br>δεδομένων DBLP-Scholar. ....                                 | 50 |
| Εικόνα 5.9-2 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο<br>δεδομένων Amazon-Google. ....                                | 51 |
| Εικόνα 5.9-3 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο<br>δεδομένων Walmart-Amazon. ....                               | 53 |
| Εικόνα 5.10-1 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking,<br>Matching), για το σύνολο DBLP-Scholar.....                      | 55 |
| Εικόνα 5.10-2 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking,<br>Matching), για το σύνολο Google-Amazon.....                     | 56 |
| Εικόνα 5.10-3 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking,<br>Matching), για το σύνολο Walmart-Amazon.....                    | 57 |



## Κατάλογος Πινάκων

|                                                                                                                             |    |
|-----------------------------------------------------------------------------------------------------------------------------|----|
| Πίνακας 3.8-1 Αποτελέσματα πειραματικής αξιολόγησης του Blocker στο σύνολο<br>Walmart-Amazon.....                           | 23 |
| Πίνακας 4.4-1 F1-Scores ανά dataset και τιμή του $w_1$ .....                                                                | 31 |
| Πίνακας 5.2-1 Συγκριτικά αποτελέσματα στο σύνολο DBLP-Scholar .....                                                         | 35 |
| Πίνακας 5.3-1 Συγκριτικά αποτελέσματα στο σύνολο Amazon-Google.....                                                         | 38 |
| Πίνακας 5.3-2 Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon .....                                                       | 40 |
| Πίνακας 5.8-1 Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon (Αρχιτεκτονική Bi-<br>Encoder έναντι Cross-Encoder) .....   | 47 |
| Πίνακας 5.9-1 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο<br>δεδομένων DBLP-Scholar. ....              | 50 |
| Πίνακας 5.9-2 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο<br>δεδομένων Amazon-Google. ....             | 51 |
| Πίνακας 5.9-3 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο<br>δεδομένων Walmart-Amazon. ....            | 52 |
| Πίνακας 5.10-1 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking,<br>Matching), για το σύνολο DBLP-Scholar. ....  | 54 |
| Πίνακας 5.10-2 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking,<br>Matching), για το σύνολο Google-Amazon.....  | 55 |
| Πίνακας 5.10-3 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking,<br>Matching), για το σύνολο Walmart-Amazon..... | 56 |

## Συντομογραφίες & Ακρωνύμια

Ακρωνύμιο - Αγγλικά - Σύντομη Επεξήγηση

AI - Artificial Intelligence - Τεχνητή Νοημοσύνη (αναφέρεται και στο πλαίσιο του Explainable AI για την εξηγησιμότητα των μοντέλων).

ANNS - Approximate Nearest Neighbor Search - Προσεγγιστική αναζήτηση πλησιέστερων γειτόνων σε διανυσματικούς χώρους υψηλών διαστάσεων.

BERT - Bidirectional Encoder Representations from Transformers - Προ-εκπαιδευμένο γλωσσικό μοντέλο που εκπαιδεύεται αμφίδρομα προβλέποντας κρυμμένες λέξεις.

BM25 - Best Matching 25 - Αλγόριθμος πιθανολογικής ανάκτησης πληροφοριών, ο οποίος χρησιμοποιείται ευρέως στο στάδιο του αποκλεισμού (blocking) για την ταχεία εύρεση υποψήφιων ζευγών.

CPU - Central Processing Unit - Η Κεντρική Μονάδα Επεξεργασίας ενός υπολογιστή, η οποία αναλαμβάνει προσωρινά υπολογισμούς όταν δεν υποστηρίζονται εγγενώς από την κάρτα γραφικών.

CSV - Comma-Separated Values - Μορφή αρχείου απλού κειμένου που χρησιμοποιείται για την αποθήκευση και ανταλλαγή δομημένων δεδομένων σε μορφή πίνακα.

CUDA - Compute Unified Device Architecture - Πλατφόρμα της εταιρείας NVIDIA που επιτρέπει την επιτάχυνση των υπολογισμών βαθιάς μάθησης μέσω της κάρτας γραφικών.

DBLP - Digital Bibliography & Library Project - Ψηφιακή βιβλιοθήκη ακαδημαϊκών δημοσιεύσεων στον τομέα της πληροφορικής, η οποία αποτέλεσε το πηγαίο πεδίο δεδομένων της εργασίας.

DeBERTa - Decoding-enhanced BERT - Εξελιγμένη μορφή του BERT που χρησιμοποιεί Disentangled Attention, διαχωρίζοντας το περιεχόμενο από τη θέση της λέξης.

ER - Entity Resolution - Η διαδικασία (Αντιστοίχιση Εγγραφών) για τον εντοπισμό εγγραφών από διαφορετικές πηγές που αναφέρονται στην ίδια πραγματική οντότητα.

F1-Score - F-measure - Μετρική αξιολόγησης που υπολογίζει τον αρμονικό μέσο όρο της Ακρίβειας (Precision) και της Ανάκλησης (Recall).

FAISS - Facebook AI Similarity Search - Βιβλιοθήκη εργαλείων για την αποδοτική προσεγγιστική αναζήτηση πλησιέστερων γειτόνων (ANNS) σε διανυσματικούς χώρους.

GPT - Generative Pre-trained Transformer - Οικογένεια Μεγάλων Γλωσσικών Μοντέλων (LLMs) με ισχυρές ικανότητες κατανόησης και παραγωγής κειμένου.

GPU - Graphics Processing Unit - Η Μονάδα Επεξεργασίας Γραφικών (κάρτα γραφικών), η οποία είναι κρίσιμη για την ταχεία εκπαίδευση νευρωνικών δικτύων.

HNSW - Hierarchical Navigable Small World - Εξελιγμένη δομή ευρετηρίου, η οποία χρησιμοποιείται για πολύ γρήγορη προσεγγιστική αναζήτηση διανυσμάτων.

LLM - Large Language Model - Μεγάλα Γλωσσικά Μοντέλα (όπως GPT, Llama) που μπορούν να λειτουργήσουν αποτελεσματικά με ελάχιστα παραδείγματα εκπαίδευσης.

LSTM - Long Short-Term Memory - Είδος νευρωνικού δικτύου που χρησιμοποιήθηκε παλαιότερα για την επεξεργασία ακολουθιών λέξεων.

MLM - Masked Language Modeling - Τεχνική εκπαίδευσης γλωσσικών μοντέλων όπου το δίκτυο μαθαίνει να προβλέπει "κρυμμένες" λέξεις μέσα σε μια πρόταση.

MPS - Metal Performance Shaders - Υποσύστημα (framework) της Apple που επιτρέπει την επιτάχυνση των υπολογισμών μηχανικής μάθησης απευθείας στην κάρτα γραφικών (GPU) των επεξεργαστών Apple Silicon.

NLP - Natural Language Processing - Ο ευρύτερος επιστημονικός τομέας της Επεξεργασίας Φυσικής Γλώσσας.

OOM - Out Of Memory - Σφάλμα υπερχείλισης μνήμης που προκύπτει κατά την εκπαίδευση όταν ο όγκος δεδομένων (π.χ. μέγεθος παρτίδας) ξεπερνά τη διαθέσιμη μνήμη του συστήματος.

PC - Pairs Completeness - Μετρική αξιολόγησης blocker που εκφράζει το ποσοστό των πραγματικών matches που διατηρούνται στα υποψήφια ζεύγη.

PLM - Pre-trained Language Model - Προ-εκπαιδευμένα Γλωσσικά Μοντέλα τα οποία έχουν εκπαιδευτεί σε τεράστια σώματα κειμένων για να κατανοούν σε βάθος τη σημασιολογία.

PQ - Pairs Quality - Μετρική αξιολόγησης blocker που εκφράζει την πυκνότητα των πραγματικών matches εντός των υποψήφιων ζευγών.

RAM - Random Access Memory - Η μνήμη τυχαίας προσπέλασης του συστήματος, κρίσιμη για τη γρήγορη φόρτωση των δεδομένων μέσω memory-mapping.

RNN - Recurrent Neural Network - Αρχιτεκτονική νευρωνικών δικτύων (όπως τα LSTMs) που αντιμετωπίζει δυσκολίες στο να "θυμάται" πληροφορίες από μακροσκελείς περιγραφές.

ROBERTa - Robustly Optimized BERT Pretraining Approach - Προ-εκπαιδευμένο γλωσσικό μοντέλο αρχιτεκτονικής Transformer, το οποίο αποτελεί βελτιστοποιημένη εκδοχή του BERT.

RR - Reduction Ratio - Μετρική που ποσοτικοποιεί τη μείωση του χώρου αναζήτησης που επιτυγχάνει ο Blocker σε σχέση με την εξαντλητική σύγκριση.

SIGMOD - Special Interest Group on Management of Data - Κορυφαίο διεθνές ακαδημαϊκό συνέδριο στον τομέα της διαχείρισης δεδομένων.

SimCSE - Simple Contrastive Learning of Sentence Embeddings - Αρχιτεκτονική αντιθετικής μάθησης (contrastive learning) που βελτιώνει την ποιότητα των διανυσματικών αναπαραστάσεων (embeddings) φέρνοντας πιο κοντά τις όμοιες εγγραφές.

TF-IDF - Term Frequency / Inverse Document Frequency - Κλασική μετρική σύγκρισης κειμένου που εντοπίζει κοινές λέξεις, δίνοντας όμως μεγαλύτερη βαρύτητα στις πιο σπάνιες.

TP / FP / FN - True Positive / False Positive / False Negative - Μεταβλητές που χρησιμοποιούνται στους τύπους για τον υπολογισμό των μετρικών αξιολόγησης (Precision και Recall).

VLDB - Very Large Data Bases - Σημαντικό διεθνές ακαδημαϊκό συνέδριο και ψηφιακή βιβλιοθήκη για συστήματα βάσεων δεδομένων.

VRAM - Video Random Access Memory - Η μνήμη της κάρτας γραφικών, η χωρητικότητα της οποίας αποτελεί αυστηρό περιοριστικό παράγοντα για την επιλογή του μεγέθους παρτίδας (batch size).

## 1 Εισαγωγή

### 1.1 Ορισμός και Σημασία της Αντιστοίχισης Εγγραφών

Η αντιστοίχιση εγγραφών (entity matching ή entity resolution) αποτελεί κρίσιμο βήμα στην ολοκλήρωση και ενοποίηση δεδομένων. Στόχος της είναι να εντοπίσει εγγραφές από διαφορετικές βάσεις που αναφέρονται στην ίδια πραγματική οντότητα, όπως για παράδειγμα δύο εγγραφές με τα στοιχεία «Apple iPhone 15 Pro Max 256GB Silver» και «iPhone 15 Pro Max by Apple, 256 GB, Silver». Η διαδικασία έχει ευρεία εφαρμογή στον καθαρισμό δεδομένων, στην ανάπτυξη γνώσης για τον Παγκόσμιο Ιστό και σε επιχειρηματικές εφαρμογές όπως η αντιστοίχιση προσφορών προϊόντων ή η σύνθεση οικονομικών πληροφοριών.

Η ύπαρξη διπλότυπων ή ασυνεπών δεδομένων έχει σοβαρό αντίκτυπο στον πραγματικό κόσμο. Στο ηλεκτρονικό εμπόριο (e-commerce), για παράδειγμα, ο κατακερματισμός ενός προϊόντος σε πολλαπλές εγγραφές μπερδεύει τους αλγόριθμους αναζήτησης και υποβαθμίζει την εμπειρία του πελάτη (Albayrak et al., 2022), ενώ σε τομείς όπως η υγεία, η αδυναμία ταύτισης ιατρικών φακέλων μπορεί να έχει κρίσιμες συνέπειες (U.S. Government Accountability Office, 2019).

Παραδοσιακά, η αντιστοίχιση πραγματοποιείται σε δύο φάσεις: αρχικά εφαρμόζονται τεχνικές αποκλεισμού (blocking) και στη συνέχεια εφαρμόζεται το ταίριασμα (matcher) για να ταξινομηθεί κάθε ζεύγος ως «match» ή «no-match». Η φάση του αποκλεισμού είναι απαραίτητη, διότι η εξαντλητική σύγκριση κάθε εγγραφής με όλες τις υπόλοιπες οδηγεί σε απαγορευτική υπολογιστική πολυπλοκότητα (Papadakis et al., 2020). Το blocking περιορίζει δραστικά τον χώρο αναζήτησης (π.χ. συγκρίνοντας μόνο προϊόντα της ίδιας μάρκας). Οι κλασικές μέθοδοι ταιριάσματος βασίζονται σε χειροκίνητους κανόνες ή σε μοντέλα μηχανικής μάθησης με προσεκτικά επιλεγμένα χαρακτηριστικά (ή ακόμα πιο απλά με τεχνικές fuzzy matching), τα οποία όμως παρουσιάζουν σοβαρές δυσκολίες όταν οι εγγραφές περιέχουν εκτενείς περιγραφές ή όταν λείπουν τιμές από δομημένα πεδία (Mudgal et al., 2018).

## **1.2 Η Μετάβαση στη Βαθιά Μάθηση (Deep Learning)**

Η εμφάνιση των νευρωνικών δικτύων αναζωπύρωσε την έρευνα στην αντιστοίχιση εγγραφών. Η εργασία του Mudgal και των συνεργατών του διαμόρφωσε έναν χώρο σχεδιασμού μοντέλων deep learning για την αντιστοίχιση, ορίζοντας τέσσερις αρχιτεκτονικές (Simple Feedforward, RNN, Attention και Hybrid) και κατηγοριοποιώντας τα προβλήματα σε δομημένα, περιγραφικά και θορυβώδη δεδομένα (Mudgal et al., 2018). Η μελέτη αυτή έδειξε ότι τα νευρωνικά δίκτυα είναι ανταγωνιστικά με τις συμβατικές προσεγγίσεις σε δομημένες εγγραφές, αλλά σημειώνουν σημαντική βελτίωση ακρίβειας (κατά 3 - 22% στην F1-score) σε περιγραφικά και θορυβώδη δεδομένα. Η δυνατότητα των δικτύων να αγνοούν τα αυστηρά όρια των πεδίων και να επεξεργάζονται ολόκληρες εγγραφές ως ενιαία κείμενα, τα καθιστά ιδιαίτερα αποτελεσματικά σε περιπτώσεις όπου η πληροφορία είναι ελλιπής ή ανακατεμένη.

## **1.3 Προ-εκπαιδευμένα Γλωσσικά Μοντέλα και το Σύστημα Ditto**

Η επόμενη γενιά συστημάτων αντιστοίχισης εγγραφών βασίζεται στη χρήση προ-εκπαιδευμένων γλωσσικών μοντέλων (Pre-trained Language Models), τα οποία έχουν εκπαιδευτεί σε τεράστια σώματα κειμένων και μπορούν να κατανοούν σε βάθος τη σημασιολογία της γλώσσας. Χαρακτηριστικό παράδειγμα αποτελεί το σύστημα Ditto, το οποίο αξιοποιεί αρχιτεκτονικές Transformer (όπως BERT, DistilBERT και RoBERTa) και διατυπώνει την αντιστοίχιση εγγραφών ως πρόβλημα ταξινόμησης ζευγών ακολουθιών (Li et al., 2021).

Για να επιτευχθεί αυτό, συστήματα όπως το Ditto εφαρμόζουν μια διαδικασία σειριοποίησης (serialization), όπου τα δομημένα πεδία μετατρέπονται σε κείμενο με τη χρήση ειδικών ετικετών (tags), όπως [COL] Title [VAL] iPhone 15. Αυτό επιτρέπει στο μοντέλο να αντιληφθεί τη δομή της πληροφορίας ως ενιαίο κείμενο. Τα πειραματικά αποτελέσματα δείχνουν ότι η αξιοποίηση τέτοιων μοντέλων οδηγεί σε σημαντική βελτίωση της ποιότητας αντιστοίχισης, ξεπερνώντας προηγούμενες μεθόδους έως και 29% στο F1-score (Li et al., 2021). Επιπλέον, το Ditto εισάγει τεχνικές που ενισχύουν περαιτέρω την απόδοσή του, όπως η ενσωμάτωση γνώσης ειδικού τομέα (domain knowledge) και η αύξηση των δεδομένων εκπαίδευσης με δύσκολα παραδείγματα (data augmentation). Παρά τη θεαματική αυτή πρόοδο, τα προ-εκπαιδευμένα γλωσσικά μοντέλα παρουσιάζουν

περιορισμούς. Συγκεκριμένα, η προσαρμογή τους σε ένα νέο πρόβλημα συχνά απαιτεί μεγάλα σύνολα ετικετοποιημένων δεδομένων, ενώ η ικανότητά τους να γενικεύουν σε εντελώς νέα πεδία εφαρμογής (Domain Adaptation) παραμένει μια ανοιχτή πρόκληση.

### **1.4 Μεγάλα Γλωσσικά Μοντέλα (LLMs) και Σύγχρονες Τάσεις**

Πιο πρόσφατα, η έρευνα στράφηκε στα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models - LLMs), όπως τα GPT, Llama και Mixtral. Οι Peeters et al. διερεύνουν τη χρήση LLMs ως εναλλακτική λύση στα κλασικά προ-εκπαιδευμένα μοντέλα, δείχνοντας ότι μπορούν να πετύχουν συγκρίσιμη ή και καλύτερη απόδοση με ελάχιστα παραδείγματα εκπαίδευσης (few-shot learning) και να παραμένουν ανθεκτικά σε άγνωστες οντότητες (Peeters et al., 2025). Η μελέτη καταδεικνύει ότι δεν υπάρχει ένα ενιαίο βέλτιστο prompt για όλες τις περιπτώσεις, ενώ συστήματα όπως το GPT-4 μπορούν να παράγουν δομημένες εξηγήσεις και να υποδεικνύουν τις αιτίες των λαθών. Αυτές οι εξελίξεις επιβεβαιώνουν την τάση για ισχυρότερα γλωσσικά μοντέλα, ιδιαίτερα όταν το κόστος συλλογής δεδομένων εκπαίδευσης είναι υψηλό.

### **1.5 Αντικείμενο και Συνεισφορά της Παρούσας Εργασίας**

Παρά τις σημαντικές προόδους, ένα από τα μεγαλύτερα ανοιχτά προβλήματα στο Entity Resolution είναι η ικανότητα των μοντέλων να προσαρμόζονται σε νέα, ετερογενή πεδία (Domain Adaptation) χωρίς να ξεχνούν την προηγούμενη γνώση τους (Trabelsi et al., 2022). Παράλληλα, τα μοντέλα καλούνται να αντιμετωπίσουν την έντονη ανισορροπία κλάσεων (class imbalance), καθώς τα ζεύγη που πραγματικά ταιριάζουν (matches) είναι συνήθως ελάχιστα σε σχέση με τις αρνητικές περιπτώσεις (non-matches) (Christen, 2012).

Η παρούσα εργασία στοχεύει στον σχεδιασμό, την υλοποίηση και την αξιολόγηση ενός συστήματος (LexiSync), το οποίο βασίζεται σε αρχιτεκτονικές Cross-Encoder για τη βαθιά σημασιολογική αντιστοίχιση εγγραφών, κάνοντας χρήση ειδικών δεικτών (special tokens) για τη διατήρηση της δομής των πεδίων. Η κύρια συνεισφορά της εργασίας συνοψίζεται στα εξής σημεία:

Πολυεπίπεδη Αξιολόγηση Μοντέλων: Πραγματοποιείται εκτενής συγκριτική μελέτη προηγμένων μοντέλων Transformer (όπως το DeBERTa-v3-base) έναντι εξαιρετικά

ελαφριών μοντέλων (όπως τα GTE-small, MiniLM-L6 και E5), αναδεικνύοντας τη βέλτιστη σχέση μεταξύ ακρίβειας και υπολογιστικού κόστους (Efficiency vs Accuracy).

Συνεχής Μάθηση (Continual Learning): Εισάγεται η χρήση μηχανισμών Experience Replay (Replay Buffer) κατά τη μεταφορά γνώσης από ακαδημαϊκά δεδομένα (DBLP-Scholar) σε εμπορικά σύνολα (Amazon-Google, Walmart-Amazon), αποτρέποντας δραστικά το φαινόμενο της καταστροφικής λήθης (catastrophic forgetting).

Διαχείριση Ανισορροπίας Δεδομένων: Υλοποιείται μια προσαρμοσμένη σταθμισμένη συνάρτηση απώλειας (Weighted Cross-Entropy Loss), η οποία αποδεικνύεται κρίσιμη για την εξισορρόπηση των κλάσεων και την αύξηση της ανάκλησης (Recall) σε θορυβώδη και απαιτητικά datasets.

## 1.6 Δομή της Εργασίας

Η παρούσα διπλωματική εργασία οργανώνεται ως εξής: Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο γύρω από το Entity Resolution και την αρχιτεκτονική των Transformers, καθώς και η σύγχρονη σχετική βιβλιογραφία. Το Κεφάλαιο 3 αναλύει λεπτομερώς τη μεθοδολογία και την αρχιτεκτονική του προτεινόμενου συστήματος (LexiSync), περιγράφοντας τη χρήση των special tokens, τον μηχανισμό Replay Buffer και την προσαρμοσμένη συνάρτηση κόστους. Στο Κεφάλαιο 4 περιγράφεται το πειραματικό πλαίσιο, τα σύνολα δεδομένων και οι μετρικές αξιολόγησης. Το Κεφάλαιο 5 παραθέτει σε βάθος τα αποτελέσματα της συγκριτικής αξιολόγησης των μοντέλων. Τέλος, στο Κεφάλαιο 6 συνοψίζονται τα συμπεράσματα της μελέτης και προτείνονται κατευθύνσεις για μελλοντική έρευνα.

# 2 Θεωρητικό Υπόβαθρο και Σχετική Βιβλιογραφία

## 2.1 Η Επιστήμη του Entity Resolution (ER)

Το Entity Resolution (ER), γνωστό στη βιβλιογραφία και ως Record Linkage ή Duplicate Detection, αποτελεί μια θεμελιώδη διαδικασία στον τομέα της Ενοποίησης Δεδομένων (Data Integration). Ο πρωταρχικός στόχος του ER είναι ο εντοπισμός εγγραφών σε μία ή περισσότερες βάσεις δεδομένων που αναφέρονται στην ίδια οντότητα του πραγματικού κόσμου, παρά τις πιθανές διαφορές στην αναπαράστασή τους (Mudgal, S., et al. (2018)).

### 2.1.1 Προκλήσεις και Πολυπλοκότητα

Η δυσκολία του ER πηγάζει από τρεις κύριους παράγοντες (Christen, 2012):

Συντακτική Ετερογένεια: Διαφορετικές συμβάσεις ονοματοδοσίας (π.χ. "iPhone 13 128GB" vs "Apple iPhone 13 Blue").

Θόρυβος στα Δεδομένα: Τυπογραφικά λάθη, ελλιπή πεδία (missing values) και λανθασμένες κατηγοριοποιήσεις, ειδικά σε datasets όπως το Walmart.

Υπολογιστική Κλίμακα: Η σύγκριση κάθε εγγραφής με όλες τις άλλες οδηγεί σε πολυπλοκότητα  $= O(n^2)$ , γεγονός που καθιστά αναγκαία τη χρήση έξυπνων αλγορίθμων ταξινόμησης.

## 2.2 Ιστορική Αναδρομή: Από τους Κανόνες στη Μηχανική Μάθηση

Η εξέλιξη του ER αντικατοπτρίζει την εξέλιξη της ίδιας της Επιστήμης των Δεδομένων.

### 2.2.1 Ντετερμινιστικά και Πιθανολογικά Μοντέλα

Οι πρώτες προσεγγίσεις βασίζονταν σε σκληρούς κανόνες (Hard Rules). Αν και αποτελεσματικές για καθαρά δεδομένα, απέτυχαν να γενικεύσουν σε περιπτώσεις με ασάφεια. Το 1969, οι Fellegi και Sunter πρότειναν το πρώτο μαθηματικό πλαίσιο για πιθανολογικό record linkage, το οποίο υπολογίζει ένα "score ομοιότητας" βασισμένο στην πιθανότητα δύο εγγραφές να είναι ταυτόσημες (M-probability) ή διαφορετικές (U-probability) (Fellegi & Sunter, 1969).

### 2.2.2 Η Εποχή των String Similarity Metrics

Πριν την έλευση του Deep Learning, η έρευνα εστιάστηκε σε συναρτήσεις σύγκρισης κειμένου:

Edit Distance (Levenshtein): Μετρά το κόστος μετατροπής μιας συμβολοσειράς σε άλλη.

Jaro-Winkler: Βελτιστοποιημένο για ονόματα, δίνοντας έμφαση στα πρώτα γράμματα των λέξεων.

TF-IDF & Cosine Similarity: Χρησιμοποιούνται για τον εντοπισμό κοινών λέξεων, δίνοντας μεγαλύτερη βαρύτητα στις σπάνιες και πιο πληροφοριακές λέξεις.

## 2.3 Η Μετάβαση στα *Word Embeddings*

Με την άνοδο του Deep Learning, η έρευνα στράφηκε στην αναπαράσταση των λέξεων ως διανύσματα σε έναν συνεχή χώρο υψηλών διαστάσεων.

Word2Vec (Mikolov et al, 2013) & GloVe (Pennington et al, 2014): Αυτά τα μοντέλα επέτρεψαν στο ER να κατανοήσει ότι το "laptop" είναι σημασιολογικά κοντά στο "notebook". Ωστόσο, αυτά τα embeddings είναι "context-free". Η λέξη "Apple" έχει το ίδιο διάνυσμα είτε αναφέρεται στην εταιρεία τεχνολογίας είτε στο φρούτο.

RNNs και LSTMs: Χρησιμοποιήθηκαν για να επεξεργαστούν ακολουθίες λέξεων, αλλά παρουσίαζαν το πρόβλημα των "vanishing gradients", δυσκολευόμενα να θυμηθούν πληροφορίες από την αρχή μιας μακροσκελούς περιγραφής προϊόντος.

## 2.4 Η Αρχιτεκτονική *Transformer* και ο Μηχανισμός *Attention*

Η αρχιτεκτονική Transformer (Vaswani, A., et al. (2017)) άλλαξε ριζικά το Natural Language Processing (NLP).

### 2.4.1 Ο Μηχανισμός *Self-Attention*

Αντί για σειριακή επεξεργασία, ο Transformer χρησιμοποιεί το Self-Attention για να εξετάσει ολόκληρη την πρόταση ταυτόχρονα (Vaswani et al., 2017). Για κάθε λέξη, το μοντέλο υπολογίζει πόση "προσοχή" πρέπει να δώσει σε όλες τις άλλες λέξεις της εισόδου. Μαθηματικά, η διαδικασία αυτή περιλαμβάνει τον πολλαπλασιασμό των διανυσμάτων Query (Q) και Key (K):

$$\text{Score} = \frac{Q \cdot K^T}{\sqrt{d_k}}$$

Το score περνά από τη συνάρτηση Softmax, η οποία το μετατρέπει σε πιθανότητα (0 έως 1), και στη συνέχεια πολλαπλασιάζεται με το διάνυσμα Value (V). Αυτό επιτρέπει στο μοντέλο LexiSync να "κλειδώνει" πάνω σε κρίσιμα χαρακτηριστικά, όπως ο κωδικός ενός προϊόντος, αγνοώντας δευτερεύουσες λέξεις.

## 2.4.2 Multi-Head Attention

Οι Transformers χρησιμοποιούν πολλαπλές κεφαλές προσοχής (Multi-Head Attention) (Vaswani et al., 2017). Κάθε κεφαλή μαθαίνει να αναγνωρίζει διαφορετικούς τύπους σχέσεων (π.χ. συντακτικές σχέσεις, μάρκες, τεχνικά χαρακτηριστικά), προσφέροντας μια πολυδιάστατη κατανόηση του κειμένου.

## 2.5 Προ-εκπαιδευμένα Μοντέλα Γλώσσας (PLMs)

Το Transfer Learning βασίζεται στην ιδέα ότι ένα μοντέλο μπορεί να εκπαιδευτεί σε μια γενική εργασία και μετά να εξειδικευτεί (Fine-tuning) σε μια συγκεκριμένη.

### 2.5.1 BERT: Bidirectional Encoder Representations

Το BERT (Devlin, J., et al. (2018)) έφερε την επανάσταση επιτρέποντας την αμφίδρομη εκπαίδευση. Μέσω της τεχνικής Masked Language Modeling (MLM), το μοντέλο μαθαίνει να προβλέπει κρυμμένες λέξεις μέσα σε μια πρόταση, κατανοώντας έτσι τη δομή της γλώσσας σε βάθος.

### 2.5.2 DeBERTa-v3: Decoding-enhanced BERT

Στην παρούσα διπλωματική, επιλέχθηκε το DeBERTa-v3 (He et al, 2021) ως το κύριο μοντέλο. Το DeBERTa εισάγει το Disentangled Attention (He et al, 2020), όπου κάθε λέξη αναπαρίσταται από δύο ξεχωριστά διανύσματα: ένα για το περιεχόμενο και ένα για τη σχετική της θέση. Αυτό το καθιστά εξαιρετικά ικανό στο Entity Resolution, καθώς η θέση ενός αριθμού (π.χ. "12" μετά το "iPhone") είναι καθοριστική για την ταυτοποίηση της οντότητας.

### 2.5.3 Η Συμβολή της Αντιθετικής Μάθησης (Contrastive Learning)

Τα τελευταία χρόνια, η αναπαράσταση των εγγραφών έχει βελτιωθεί δραματικά μέσω της Αντιθετικής Μάθησης (Contrastive Learning). Αντί το μοντέλο να μαθαίνει απλώς να προβλέπει λέξεις (όπως στο κλασικό Masked Language Modeling), εκπαιδεύεται με τέτοιο τρόπο ώστε να "τραβάει" πιο κοντά στον διανυσματικό χώρο (vector space) τις εγγραφές που αποτελούν πραγματικά Matches (positive pairs) και να "απωθεί" τις εγγραφές που διαφέρουν (negative pairs). Αρχιτεκτονικές όπως το SimCSE (Gao et al., 2021) έχουν αποδείξει ότι η χρήση contrastive loss βελτιώνει την ικανότητα των Transformers να

διαχωρίζουν δύσκολες περιπτώσεις, όπως προϊόντα με παρόμοιους τίτλους αλλά διαφορετικούς κωδικούς μοντέλων, δημιουργώντας ένα πιο σταθερό σημείο εκκίνησης (pre-training) πριν την τελική ταξινόμηση από τον Cross-Encoder. Επίσης, πρόσφατες έρευνες (Narayan et al., 2022) δείχνουν ότι ο συνδυασμός LLMs με Contrastive Learning μπορεί να επιλύσει προβλήματα ενοποίησης δεδομένων με ελάχιστα δεδομένα εκπαίδευσης (few-shot).

## 2.6 Αρχιτεκτονικές Cross-Encoder για ER

Μια από τις πιο κρίσιμες αποφάσεις σχεδιασμού του συστήματος LexiSync είναι η χρήση του Cross-Encoder.

Bi-Encoders: Κωδικοποιούν την Εγγραφή A και την Εγγραφή B ανεξάρτητα. Αν και είναι ταχύτεροι, στερούνται "βάθους" καθώς οι εγγραφές δεν αλληλεπιδρούν κατά την εξαγωγή των χαρακτηριστικών.

Cross-Encoders: Οι δύο εγγραφές εισάγονται ταυτόχρονα στο μοντέλο, διαχωρισμένες από ένα ειδικό token [SEP]. Το μοντέλο μπορεί να συγκρίνει κάθε λέξη της εγγραφής A με κάθε λέξη της εγγραφής B σε κάθε επίπεδο (layer) του Transformer. Αυτή η μέθοδος, αν και υπολογιστικά ακριβότερη, προσφέρει τα υψηλότερα F1-scores στη βιβλιογραφία του Entity Matching (Li, Y., et al. (2021)).

## 2.7 Class Imbalance και Weighted Cross Entropy

Μια συχνή πρόκληση στο ER είναι η ακραία ανισορροπία των κλάσεων (Christen, 2012). Σε ένα τυπικό dataset, τα ζεύγη που ΔΕΝ ταιριάζουν (Negative class) είναι χιλιάδες φορές περισσότερα από τα ζεύγη που ταιριάζουν (Positive class).

### 2.7.1 Το πρόβλημα του Majority Bias

Εάν ένα μοντέλο εκπαιδευτεί με την τυπική συνάρτηση κόστους (Standard Cross-Entropy Loss) σε ένα εξαιρετικά ανισόρροπο σύνολο, θα αναπτύξει τάση (bias) προς την πλειοψηφούσα κλάση. Το δίκτυο θα μάθει να προβλέπει σχεδόν πάντα "0", επιτυγχάνοντας φαινομενικά υψηλό Accuracy, αλλά πρακτικά μηδενικό Recall, καθιστώντας το σύστημα άχρηστο σε πραγματικές συνθήκες (Buda et al., 2018).

## 2.7.2 Η Λύση του Weighted Loss

Στον κώδικα του LexiSync, υλοποιήθηκε ο WeightedTrainer, ο οποίος χρησιμοποιεί τη συνάρτηση Weighted Cross Entropy Loss (Buda et al., 2018)

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^N w_{y_i} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Όπου:

Το  $w_{y_i}$  είναι το βάρος (class weight) που αποδίδεται στην πραγματική κλάση του δείγματος  $i$ . Χρησιμοποιείται για την αντιμετώπιση της ανισορροπίας των δεδομένων (class imbalance), δίνοντας μεγαλύτερη "ποινή" στο σφάλμα για τη μειονοτική κλάση.

Το  $y_i$  είναι η πραγματική ετικέτα (ground truth) του δείγματος, η οποία λαμβάνει την τιμή 0 ή 1.

Το  $\hat{y}_i$  είναι η πρόβλεψη του μοντέλου, δηλαδή η εκτιμώμενη πιθανότητα το δείγμα να ανήκει στη θετική κλάση.

Το  $N$  είναι το συνολικό πλήθος των δειγμάτων (π.χ. batch size) πάνω στο οποίο υπολογίζεται η μέση απώλεια.

είναι το βάρος της κλάσης. Δίνοντας μεγαλύτερο βάρος στα Matches (Positive class), αναγκάζουμε το μοντέλο να "τιμωρεί" τον εαυτό του πολύ αυστηρά όταν αποτυγχάνει να εντοπίσει ένα πραγματικό ταίριασμα, αυξάνοντας έτσι δραματικά το Recall.

## 2.8 Continual Learning και Anti-Forgetting

Όταν ένα μοντέλο μεταφέρεται από ένα source domain (DBLP) σε ένα target domain (Walmart), αντιμετωπίζει το φαινόμενο του Catastrophic Forgetting (Kirkpatrick, J., et al. (2017)). Το μοντέλο τείνει να "επικαλύπτει" τις παλιές γνώσεις με τις νέες.

### 2.8.1 Replay Buffer (Experience Replay)

Για την αντιμετώπιση αυτού του προβλήματος, ο LexiSync χρησιμοποιεί έναν Replay Buffer. Πρόκειται για μια τεχνική όπου ένα υποσύνολο των δεδομένων του DBLP αποθηκεύεται και επανεισάγεται κατά τη διάρκεια της εκπαίδευσης στα δεδομένα του Amazon και του Walmart. Αυτή η συνεχής υπενθύμιση επιτρέπει στο μοντέλο να διατηρεί την απόδοσή του σε όλα τα domains ταυτόχρονα (Chaudhry et al, 2019).

## 2.9 Μετρικές Αξιολόγησης (Metrics)

Η αξιολόγηση των συστημάτων ER δεν μπορεί να βασιστεί στο Accuracy λόγω του class imbalance. Χρησιμοποιούνται οι εξής μετρικές:

Precision (Ακρίβεια):

$$P = \frac{TP}{(TP + FP)}$$

. Πόσα από αυτά που είπαμε ότι ταιριάζουν, είναι όντως σωστά;

Recall (Ανάκληση):

$$R = \frac{TP}{(TP + FN)}$$

. Πόσα από τα πραγματικά matches καταφέραμε να βρούμε;

F1-Score: Ο αρμονικός μέσος των P και R. Είναι η πιο κρίσιμη μετρική, καθώς απαιτεί και οι δύο δείκτες να είναι υψηλοί για να επιτευχθεί καλό αποτέλεσμα.

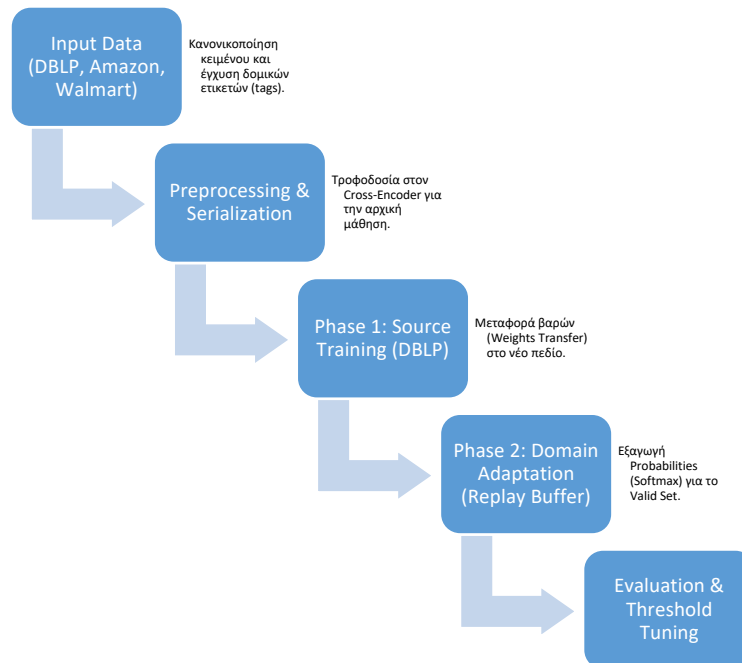
## 3 Μεθοδολογία και Υλοποίηση του Συστήματος

### 3.1 Εισαγωγή στη Μεθοδολογία

Η παρούσα εργασία προτείνει και υλοποιεί το σύστημα LexiSync, μια ολοκληρωμένη αρχιτεκτονική αντιστοίχισης εγγραφών (Entity Resolution) βασισμένη σε Προ-εκπαιδευμένα Γλωσσικά Μοντέλα (Pre-trained Language Models - PLMs). Κύριος στόχος του LexiSync είναι η αντιμετώπιση τριών θεμελιωδών προκλήσεων που εντοπίζονται στη σύγχρονη βιβλιογραφία: (α) η αποτελεσματική επεξεργασία ετερογενών και "θορυβωδών" δεδομένων (noisy data) μέσω της αρχιτεκτονικής Cross-Encoder, (β) η μεταφορά γνώσης (Domain Adaptation) από ακαδημαϊκά σε εμπορικά σύνολα δεδομένων, και (γ) η καταπολέμηση της καταστροφικής λήθης (catastrophic forgetting) κατά τη συνεχή μάθηση (continual learning) (Kirkpatrick, J., et al. (2017)).

Η υλοποίηση πραγματοποιήθηκε στη γλώσσα προγραμματισμού Python, αξιοποιώντας το οικοσύστημα της Hugging Face (βιβλιοθήκες transformers και datasets) (Wolf, T., et al. (2020)) και το πλαίσιο βαθιάς μάθησης PyTorch. Σε αυτό το κεφάλαιο, αναλύεται διεξοδικά η αρχιτεκτονική του συστήματος, οι μέθοδοι προεπεξεργασίας των δεδομένων, οι τεχνικές

διαχείρισης της ανισορροπίας των κλάσεων και ο μηχανισμός μνήμης (Replay Buffer) που σχεδιάστηκε για την προσαρμογή του μοντέλου.



Εικόνα 3.1-1 Αρχιτεκτονική Συστήματος

## 3.2 Αρχιτεκτονική και Επιλογή Μοντέλων

Το πρόβλημα της αντιστοίχισης εγγραφών διατυπώνεται στο LexiSync ως ένα πρόβλημα δυαδικής ταξινόμησης ακολουθιών (Binary Sequence Classification). Συγκεκριμένα, δοθέντος ενός ζεύγους εγγραφών ( $e_1, e_2$ ) το σύστημα καλείται να επιστρέψει μια πιθανότητα  $P(\text{Match} | e_1, e_2)$ , η οποία υποδηλώνει αν οι εγγραφές ταυτίζονται.

### 3.2.1 Η Προσέγγιση Cross-Encoder

Για τη μεγιστοποίηση της ακρίβειας, επιλέχθηκε η αρχιτεκτονική Cross-Encoder ((Li et al., 2021)). Σε αντίθεση με τους Bi-Encoders που παράγουν ανεξάρτητα διανύσματα (embeddings) για κάθε εγγραφή, ο Cross-Encoder ενώνει τις δύο εγγραφές σε μία ενιαία ακολουθία κειμένου. Η κλάση `AutoModelForSequenceClassification` χρησιμοποιείται για τη φόρτωση των μοντέλων, ρυθμισμένη να εξάγει ακριβώς δύο ετικέτες (`num_labels=2`). Η ταυτόχρονη εισαγωγή των εγγραφών επιτρέπει στον μηχανισμό self-attention των Transformers να υπολογίζει την προσοχή (attention) ανάμεσα σε κάθε λέξη της πρώτης

εγγραφής και κάθε λέξη της δεύτερης (cross-attention) σε όλα τα επίπεδα (layers) του νευρωνικού δικτύου (Devlin, J., et al. (2018)).

### 3.2.2 Εξεταζόμενα Μοντέλα (Model Zoo)

Το σύστημα LexiSync σχεδιάστηκε να είναι ανεξάρτητο από το εκάστοτε υποκείμενο γλωσσικό μοντέλο. Για την πειραματική αξιολόγηση ενσωματώθηκαν έξι (6) διαφορετικές αρχιτεκτονικές, προκειμένου να μελετηθεί η σχέση μεταξύ μεγέθους μοντέλου και ακρίβειας (Accuracy-Efficiency Trade-off):

microsoft/deberta-v3-base: Το ισχυρότερο μοντέλο της δοκιμής, εξοπλισμένο με Disentangled Attention (He et al, 2021).

sentence-transformers/all-MiniLM-L6-v2: Ένα εξαιρετικά ελαφρύ μοντέλο 6 επιπέδων, βελτιστοποιημένο μέσω Knowledge Distillation (Wang, W., et al. (2020)).

thenlper/gte-small & intfloat/e5-small-v2: Μοντέλα ειδικά εκπαιδευμένα για εργασίες γενικής αναπαράστασης κειμένου.

BAAI/bge-small-en-v1.5 και Snowflake/snowflake-arctic-embed-xs: Σύγχρονες αρχιτεκτονικές χαμηλού υπολογιστικού αποτυπώματος.

Κατά τη φόρτωση των μοντέλων, ορίζεται ρητά ο τύπος δεδομένων σε torch.float32 (torch\_dtype=torch.float32), μια κρίσιμη επιλογή σχεδιασμού για τη διασφάλιση της σταθερότητας κατά την εκπαίδευση, ειδικά στο περιβάλλον υλικού Apple Silicon (MPS) που χρησιμοποιήθηκε.

## 3.3 Προεπεξεργασία και Σειριοποίηση (Serialization)

Τα δεδομένα εισόδου στο Entity Resolution προέρχονται συνήθως σε μορφή πίνακα (tabular data), όπου κάθε οντότητα περιγράφεται από πολλαπλά δομημένα χαρακτηριστικά (π.χ. τίτλος, κατασκευαστής, τιμή). Προκειμένου αυτά τα δεδομένα να τροφοδοτηθούν σε ένα γλωσσικό μοντέλο, πρέπει να μετατραπούν σε μονοδιάστατο κείμενο (Li et al., 2021).

### 3.3.1 Καθαρισμός Κειμένου (Text Cleaning)

Ο θόρυβος στα δεδομένα αποτελεί έναν από τους κύριους παράγοντες υποβάθμισης της απόδοσης. Στην κλάση LexiSync, η μέθοδος clean\_text εφαρμόζει ένα αυστηρό πρωτόκολλο κανονικοποίησης. Αρχικά, όλο το κείμενο μετατρέπεται σε πεζούς χαρακτήρες. Στη συνέχεια, μέσω κανονικών εκφράσεων (regular expressions), αφαιρούνται

όλοι οι ειδικοί χαρακτήρες, διατηρώντας αποκλειστικά αλφαριθμητικούς χαρακτήρες, κενά, τελείες και παύλες ( $[^a-z0-9\backslash\.\-]$ ). Τέλος, τα πολλαπλά κενά συμπύσσονται σε ένα (`re.sub(r'\s+', ' ', text)`). Η διατήρηση της τελείας και της παύλας είναι στοχευμένη, καθώς αυτά τα σύμβολα είναι συχνά κρίσιμα για τους αριθμούς μοντέλων (π.χ. "v-1.5", "10.5-inch").

### 3.3.2 Σειριοποίηση με Ειδικούς Δείκτες (Special Tokens)

Η απλή συνένωση (concatenation) των πεδίων (π.χ. "Apple 1200 iPhone") στερεί από το μοντέλο την πληροφορία για το σε ποιο χαρακτηριστικό ανήκει η κάθε λέξη. Το LexiSync λύνει αυτό το πρόβλημα εισάγοντας τη χρήση ετικετών (tags) ως Ειδικών Δεικτών (Special Tokens) (Li et al., 2021).

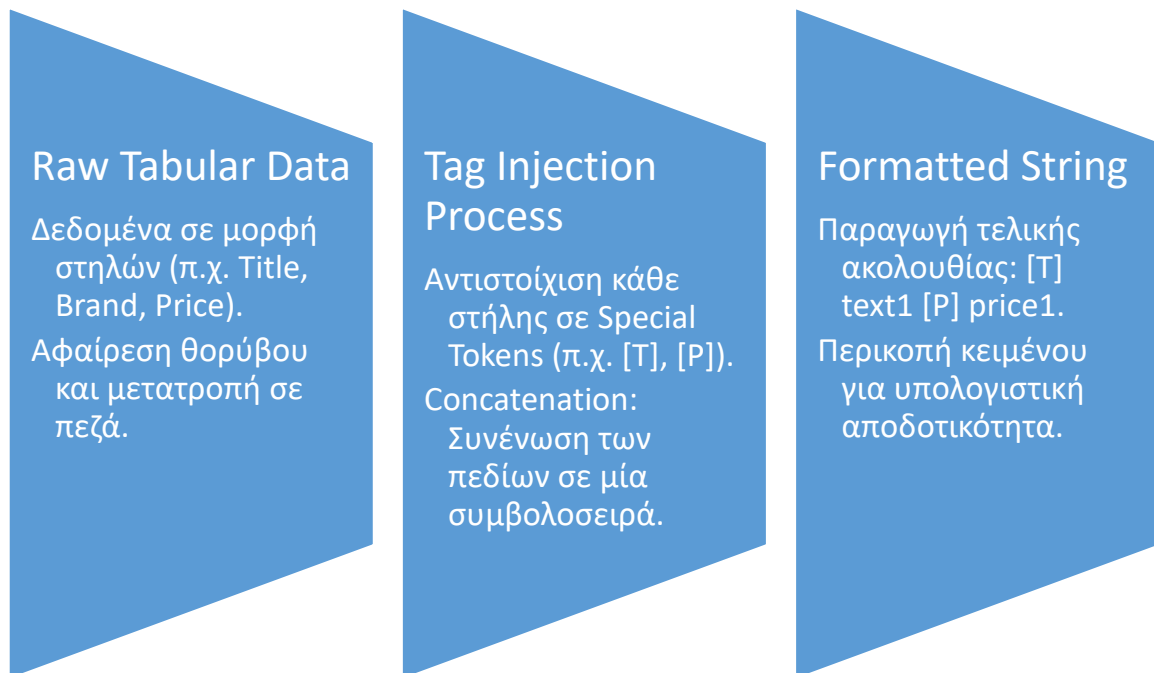
Κατά τη φόρτωση κάθε συνόλου δεδομένων (`add_dataset`), ορίζεται ένα σχήμα δομής (`columns_spec`). Για παράδειγμα, στο σύνολο δεδομένων του Walmart, χρησιμοποιούνται οι δείκτες [T] για τον τίτλο, [N] για τον αριθμό μοντέλου, [B] για τη μάρκα, [C] για την κατηγορία και [P] για την τιμή.

Μια εγγραφή μετατρέπεται ως εξής:

```
[T] apple iphone 13 [N] a2633 [B] apple [P] 799.00
```

Για να διασφαλιστεί ότι ο Tokenizer δεν θα διασπάσει αυτά τα tags σε μικρότερα υπο-μήματα (subwords), η μέθοδος `_update_tokenizer` τα προσθέτει δυναμικά στο λεξιλόγιο του μοντέλου (`additional_special_tokens`). Ακολούθως, οι διαστάσεις των ενσωματώσεων του μοντέλου αναπροσαρμόζονται (`resize_token_embeddings(len(self.tokenizer))`) ώστε το νευρωνικό δίκτυο να μπορεί να εκπαιδεύσει νέα, ανεξάρτητα διανύσματα βαρών για αυτούς τους δείκτες.

Κατά τη δημιουργία του τελικού αντικειμένου Dataset, χρησιμοποιείται η συνάρτηση tokenizer με τις παραμέτρους `truncation=True` και `max_length=128`. Η τιμή 128 επιλέχθηκε ως η βέλτιστη χρυσή τομή, καθώς επαρκεί για να χωρέσει τα κρίσιμα πεδία μιας εγγραφής (όπως ο τίτλος και το μοντέλο), ενώ ταυτόχρονα διατηρεί τις απαιτήσεις μνήμης RAM σε διαχειρίσιμα επίπεδα, επιτρέποντας την εκπαίδευση σε συστήματα με περιορισμένους πόρους.



Εικόνα 3.3-1 Διαδικασία Σειριοποίησης

### 3.3.3 Η Φάση του Αποκλεισμού (Blocking) στα Σύνολα Δεδομένων

Όπως αναλύθηκε στο θεωρητικό υπόβαθρο, η εξαντλητική σύγκριση κάθε εγγραφής της Βάσης A με κάθε εγγραφή της Βάσης B καθιστά το πρόβλημα υπολογιστικά μη βιώσιμο  $O(n^2)$ .

Στην παρούσα υλοποίηση, τα σύνολα δεδομένων (DBLP-Scholar, Amazon-Google, Walmart-Amazon) αντλήθηκαν από καθιερωμένα benchmarks (Mudgal et al., 2018), στα οποία η φάση του Αποκλεισμού (Blocking) έχει ήδη προηγηθεί μέσω ευρετικών κανόνων, όπως η ευρετηρίαση (indexing) με βάση το TF-IDF (Term Frequency-Inverse Document Frequency) ή ο αλγόριθμος BM25 (Robertson & Zaragoza, 2009). Η αξιοποίηση αυτών των προ-φιλτραρισμένων υποψήφιων ζευγών αποτέλεσε συνειδητή μεθοδολογική επιλογή, καθώς διασφαλίζει την άμεση συγκρισιμότητα των αποτελεσμάτων με την υφιστάμενη βιβλιογραφία (Li et al., 2021) και (Mudgal et al., 2018), η οποία υιοθετεί την ίδια πειραματική σύμβαση. Συνεπώς, τα ζεύγη που τροφοδοτούνται στον Cross-Encoder είναι "Υποψήφια Ζεύγη" (Candidate Pairs) υψηλής δυσκολίας, και ο πυρήνας ταξινόμησης του LexiSync καλείται να διαχωρίσει τις πραγματικές ταυτίσεις από τα εξαιρετικά όμοια (αλλά λανθασμένα) ζεύγη που "πέρασαν" από το αρχικό φίλτρο του blocking.

Παράλληλα, στο πλαίσιο της παρούσας εργασίας σχεδιάστηκε και υλοποιήθηκε ένας αυτόνομος μηχανισμός Blocking, ο οποίος αναλύεται διεξοδικά στην Ενότητα 3.8. Ο μηχανισμός αυτός βασίζεται σε αρχιτεκτονική Bi-Encoder και σε δομές προσεγγιστικής αναζήτησης πλησιέστερων γειτόνων (Approximate Nearest Neighbor Search - ANNS) μέσω της βιβλιοθήκης FAISS (Johnson et al., 2019), και επιτρέπει τη λειτουργία του LexiSync ως ολοκληρωμένο σύστημα δύο σταδίων σε νέα, μη προ-επεξεργασμένα δεδομένα. Η πειραματική αξιολόγηση του end-to-end pipeline (blocking + matching) με τις καθιερωμένες μετρικές Pairs Completeness και Pairs Quality (Christen, 2012) εκφεύγει του πεδίου της παρούσας μελέτης και προτείνεται ως άμεση μελλοντική επέκταση.

### **3.4 Αντιμετώπιση Ανισορροπίας Κλάσεων (Class Imbalance)**

Ένα από τα πιο δυσεπίλυτα προβλήματα στο Entity Resolution είναι η δραματική ανισορροπία των κλάσεων (Christen, 2012). Ο αριθμός των ζευγών που δεν ταυτίζονται (Negative - Label 0) είναι συνήθως συντριπτικά μεγαλύτερος από τα ζεύγη που αποτελούν πραγματικά Matches (Positive - Label 1).

#### **3.4.1 Το Πρόβλημα του Majority Bias**

Εάν ένα μοντέλο εκπαιδευτεί με την τυπική συνάρτηση κόστους (Standard Cross-Entropy Loss) σε ένα εξαιρετικά ανισόρροπο σύνολο, θα αναπτύξει τάση (bias) προς την πλειοψηφούσα κλάση. Το δίκτυο θα μάθει να προβλέπει σχεδόν πάντα "0", επιτυγχάνοντας φαινομενικά υψηλό Accuracy, αλλά πρακτικά μηδενικό Recall, καθιστώντας το σύστημα άχρηστο σε πραγματικές συνθήκες (Buda et al., 2018).

#### **3.4.2 Ο Μηχανισμός WeightedTrainer**

Για την επίλυση αυτού του προβλήματος, το LexiSync εισάγει μια προσαρμοσμένη κλάση εκπαίδευσης, την WeightedTrainer, η οποία κληρονομεί από την βασική κλάση Trainer της βιβλιοθήκης transformers (Wolf, T., et al. (2020)). Η καινοτομία έγκειται στην υπερκάλυψη (overriding) της μεθόδου compute\_loss.

Κατά την προετοιμασία των δεδομένων, η συνάρτηση \_calculate\_weights μετρά τη συχνότητα της κάθε κλάσης και υπολογίζει σταθμισμένα βάρη (class weights) αντιστρόφως ανάλογα με τη συχνότητά τους. Τα βάρη υπολογίζονται μαθηματικά βάσει των εξής τύπων:

$$w_0 = \frac{\text{Pos} + \text{Neg}}{2 * \text{Neg}}$$

$$w_1 = \left( \frac{\text{Pos} + \text{Neg}}{2 * \text{Pos}} \right) * 0.5$$

Όπου Pos είναι το πλήθος των ταυτίσεων (Matches) και Neg το πλήθος των μη-ταυτίσεων. Στον τύπο του  $w_1$  εισήχθη εμπειρικά ένας μειωτικός συντελεστής 0.5. Η απόφαση αυτή πάρθηκε διότι το βάρος της θετικής κλάσης ήταν αρχικά τόσο μεγάλο (λόγω της σπανιότητάς της), που το μοντέλο άρχιζε να εμφανίζει υπερβολικά False Positives (πτώση του Precision). Ο συντελεστής 0.5 προσφέρει μια ιδανική εξισορρόπηση (Trade-off) ανάμεσα στο Precision και το Recall.

Αυτά τα βάρη μεταβιβάζονται στη συνάρτηση `nn.CrossEntropyLoss(weight=weights)` κατά τη διάρκεια της εκπαίδευσης. Έτσι, εάν το μοντέλο κάνει λάθος πρόβλεψη σε ένα πραγματικό Match, τιμωρείται πολλαπλάσια από το αν έκανε λάθος σε ένα Non-Match, αναγκάζοντάς το να "δώσει προσοχή" στα σπάνια θετικά δείγματα.

### **3.5 Διαδικασία Εκπαίδευσης και Συνεχής Μάθηση (Continual Learning)**

Το pipeline του LexiSync χωρίζεται σε δύο διακριτές φάσεις εκπαίδευσης: την εκπαίδευση στο Πηγαίο Πεδίο (Source Domain) και την προσαρμογή στους Τομείς Στόχους (Target Domains).

#### **3.5.1 Φάση 1: Εκπαίδευση στο Source Domain (DBLP-Scholar)**

Στην πρώτη φάση (μέθοδος `train_source`), το μοντέλο εκπαιδεύεται πάνω στο σύνολο δεδομένων `dblp`, το οποίο περιέχει ακαδημαϊκές δημοσιεύσεις. Η εκπαίδευση ρυθμίζεται μέσω της κλάσης `TrainingArguments` με τα εξής βασικά χαρακτηριστικά:

**Epochs:** 5 εποχές, προσφέροντας επαρκή χρόνο στο μοντέλο να προσαρμόσει τα βάρη του.  
**Batch Size:** 12. Επιλέχθηκε για να είναι συμβατό με τους περιορισμούς VRAM της κάρτας γραφικών, ενώ διασφαλίζει ομαλή σύγκλιση.

**Learning Rate:** Σταθεροποιήθηκε στην τιμή  $2 \times 10^{-5}$ , μια τυπική και ασφαλής επιλογή για το fine-tuning μοντέλων Transformer (Devlin et al., 2018; Sun et al., 2019), αποτρέποντας απότομες αλλαγές που θα μπορούσαν να αλλοιώσουν την προ-εκπαιδευμένη γνώση.

Weight Decay: Καθορίστηκε στο 0.01 για να εφαρμοστεί κανονικοποίηση (regularization)  $L_2$ , η οποία αποτρέπει την υπερπροσαρμογή (overfitting).

Data Collator: Χρησιμοποιείται η κλάση DataCollatorWithPadding, η οποία εφαρμόζει "Δυναμική Αναπλήρωση" (Dynamic Padding). Αντί να γεμίζει όλα τα κείμενα στο απόλυτο μέγιστο (128 tokens), τα συμπληρώνει με padding tokens μόνο μέχρι το μήκος του μεγαλύτερου κειμένου στο τρέχον batch (mini-batch). Αυτή η τεχνική εξοικονομεί τεράστιους υπολογιστικούς πόρους.

### 3.5.2 Φάση 2: Domain Adaptation και ο Μηχανισμός Replay Buffer

Μόλις το μοντέλο μάθει να ταυτοποιεί ακαδημαϊκές εγγραφές, η μέθοδος `adapt_to_targets` αναλαμβάνει την προσαρμογή του σε νέα, εμπορικά σύνολα δεδομένων (Amazon, Walmart).

Η μετάβαση από το πεδίο της βιβλιογραφίας στο πεδίο του ηλεκτρονικού εμπορίου ενεργοποιεί το φαινόμενο της καταστροφικής λήθης (catastrophic forgetting) (Kirkpatrick, J., et al. (2017)). Το νευρωνικό δίκτυο, στην προσπάθειά του να ελαχιστοποιήσει το σφάλμα (loss) στα νέα δεδομένα των προϊόντων, τροποποιεί ριζικά τα βάρη του, καταστρέφοντας τη γνώση που μόλις απέκτησε για τις ακαδημαϊκές δημοσιεύσεις.

Για την οριστική επίλυση αυτού του προβλήματος (Kirkpatrick, J., et al. (2017)), σχεδιάστηκε και ενσωματώθηκε ένας μηχανισμός μνήμης Replay Buffer. Η διαδικασία υλοποιείται αλγοριθμικά ως εξής:

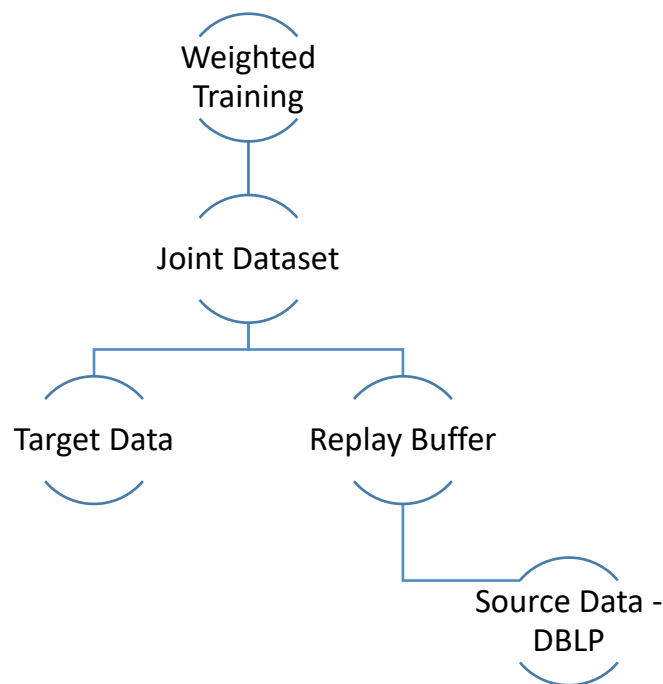
Δειγματοληψία: Το σύνολο δεδομένων εκπαίδευσης του DBLP (Source) ανακατεύεται τυχαία (`shuffle(seed=42)`). Στη συνέχεια, εξάγεται ένα υποσύνολο από 1.200 δείγματα (`replay_samples=1200`), τα οποία λειτουργούν ως μνήμη.

Συγχώνευση (Joint Dataset): Τα 1.200 δείγματα του DBLP ενώνονται (`concatenate_datasets`) με τα πλήρη σετ εκπαίδευσης του Amazon και του Walmart. Το τελικό ενοποιημένο σύνολο ανακατεύεται ξανά.

Δυναμικός Επανυπολογισμός Βαρών: Επειδή η αναλογία των κλάσεων αλλάζει δραστικά στο ενοποιημένο σύνολο, η συνάρτηση `calculate_weights` καλείται εκ νέου (`joint_weights`) για να δημιουργήσει νέα, ενημερωμένα βάρη ποινής για τον `WeightedTrainer`.

Εκπαίδευση: Το μοντέλο εκπαιδεύεται για 7 εποχές στο ενοποιημένο σύνολο, με ελαφρώς μικρότερο ρυθμό μάθησης ( $\text{Learning Rate} = 1 \times 10^{-5}$ ), ώστε να προσαρμοστεί ομαλά στα νέα δεδομένα του ηλεκτρονικού εμπορίου χωρίς να αποσταθεροποιηθεί.

Αυτή η προσέγγιση (Experience Replay) εγγυάται ότι το μοντέλο δεν θα "ξεχάσει" πώς να αναγνωρίζει τις εγγραφές του DBLP, καθώς κατά τη διάρκεια της εκπαίδευσης έρχεται συνεχώς αντιμέτωπο με παλαιά παραδείγματα (Chaudhry et al, 2019).



Εικόνα 3.5-1 Μηχανισμός Replay Buffer

### 3.6 Αξιολόγηση και Βελτιστοποίηση Κατωφλίου (Threshold Tuning)

Η τελική φάση του LexiSync υλοποιείται μέσω της μεθόδου `generate_report`. Σε αντίθεση με τις απλές μεθόδους ταξινόμησης που χρησιμοποιούν ένα σταθερό κατώφλι (threshold) 0.5 για την απόφαση, το LexiSync υιοθετεί μια στρατηγική δυναμικής βελτιστοποίησης (Esposito et al., 2021).

Για κάθε εγγραφή του συνόλου αξιολόγησης (Valid Set), το μοντέλο εξάγει τα logits. Αυτά μετατρέπονται σε πιθανότητες μέσω της συνάρτησης Softmax (`torch.nn.functional.softmax`) και απομονώνεται η πιθανότητα για την κλάση 1 (Match).

Στη συνέχεια, ένας αλγόριθμος αναζήτησης σαρώνει διάφορες τιμές κατωφλίου από 0.2 έως 0.95 (με βήμα 0.05). Για κάθε τιμή  $t$ , το σύστημα υπολογίζει το υποθετικό F1-score:

$$F1_t = f1_{\text{score}}(\text{labels}, (\text{probs} \geq t))$$

Το κατώφλι που μεγιστοποιεί το F1-score επιλέγεται ως το "Βέλτιστο Κατώφλι" (Optimal Threshold) για το συγκεκριμένο σύνολο δεδομένων. Η δυναμική αυτή προσαρμογή είναι υψίστης σημασίας, καθώς σε πολύ θορυβώδη δεδομένα (όπως το Walmart), το μοντέλο μπορεί να χρειάζεται ένα χαμηλότερο threshold (π.χ. 0.35) για να εντοπίσει τα matches, ενώ σε πιο καθαρά δεδομένα ίσως απαιτείται υψηλότερο threshold για την αποφυγή False Positives.

Τέλος, το σύστημα εξάγει αυτόματα τα αποτελέσματα σε αρχεία κειμένου (results\_modelname.txt), παραθέτοντας τον Πίνακα Σύγχυσης (Confusion Matrix) και το Πλήρες Report Ταξινόμησης (Precision, Recall, F1-Score) για κάθε ελεγχόμενο domain (DBLP, Amazon, Walmart).

### 3.7 Περιβάλλον Υλοποίησης

Για την αντιμετώπιση θεμάτων συμβατότητας κατά την εκτέλεση πειραμάτων βαθιάς μάθησης σε αρχιτεκτονική Apple Silicon, ο κώδικας αρχικοποιείται με την εντολή περιβάλλοντος `os.environ["PYTORCH_ENABLE_MPS_FALLBACK"] = "1"`. Η εντολή αυτή διασφαλίζει ότι πράξεις τανυστών (tensors) οι οποίες δεν υποστηρίζονται εγγενώς από τη μηχανή MPS (Metal Performance Shaders) της Apple, θα εκτελεστούν αυτόματα στη CPU, αποτρέποντας την κατάρρευση του προγράμματος (crash) και εξασφαλίζοντας την απρόσκοπτη ροή του πειράματος. Παράλληλα, κατά την κύρια εκτέλεση (if `__name__ == "__main__"`), εφαρμόζεται αυστηρή αρχικοποίηση σπόρων τυχαιότητας (seeds) στις βιβλιοθήκες random, numpy και torch (`seed(42)`), ώστε τα πειράματα να είναι 100% αναπαραγώγιμα (reproducible).

### 3.8 Σχεδιασμός Αυτόνομου Μηχανισμού Αποκλεισμού (Blocking) με ANNS Index

Όπως αναλύθηκε στην Ενότητα 3.3.3, η πειραματική αξιολόγηση των μοντέλων ταξινόμησης πραγματοποιήθηκε επί προ-φιλτραρισμένων υποψήφιων ζευγών που παρέχουν τα καθιερωμένα benchmarks της βιβλιογραφίας, ώστε τα αποτελέσματα να είναι άμεσα συγκρίσιμα με προηγούμενες εργασίες (Mudgal et al., 2018) και (Li et al., 2021). Ωστόσο, σε ένα πραγματικό σενάριο εφαρμογής, ένα σύστημα Entity Resolution καλείται

να επεξεργαστεί δύο μη επεξεργασμένες βάσεις δεδομένων από το μηδέν, χωρίς να διαθέτει εκ των προτέρων γνώση για τα ζεύγη που αξίζει να συγκριθούν. Για να καλύψει αυτό το λειτουργικό κενό, στο πλαίσιο της παρούσας εργασίας σχεδιάστηκε και υλοποιήθηκε ένας αυτόνομος μηχανισμός Blocking, ο οποίος μετατρέπει το LexiSync σε ολοκληρωμένο σύστημα δύο σταδίων: ένα γρήγορο στάδιο φιλτραρίσματος (Blocker) και το βαρύ στάδιο τελικής ταξινόμησης (Cross-Encoder Matcher).

Η ανάγκη για ένα τέτοιο στάδιο πηγάζει από την υπολογιστική πολυπλοκότητα του προβλήματος. Όπως έχει επισημανθεί, η εξαντλητική σύγκριση κάθε εγγραφής της πρώτης βάσης με κάθε εγγραφή της δεύτερης οδηγεί σε πολυπλοκότητα  $O(n^2)$ , η οποία γίνεται πρακτικά απαγορευτική σε βάσεις με δεκάδες ή εκατοντάδες χιλιάδες εγγραφές (Christen, 2012). Ένας Cross-Encoder, παρά την υψηλή του ακρίβεια, είναι κατά τάξεις μεγέθους πιο αργός από έναν Bi-Encoder, καθώς απαιτεί την ταυτόχρονη επεξεργασία κάθε ζεύγους εγγραφών μέσα από όλα τα επίπεδα του Transformer (Reimers & Gurevych, 2019). Συνεπώς, η εφαρμογή του απευθείας στο καρτεσιανό γινόμενο δύο βάσεων θα ήταν ανέφικτη, καθιστώντας απαραίτητη την προηγούμενη παρέμβαση ενός μηχανισμού που θα μειώσει δραστικά τον χώρο αναζήτησης.

### 3.8.1 Αρχιτεκτονική Bi-Encoder και Παραγωγή Διανυσματικών Αναπαραστάσεων

Ο προτεινόμενος μηχανισμός υιοθετεί μια αρχιτεκτονική Bi-Encoder (Reimers & Gurevych, 2019), η οποία διαφέρει θεμελιωδώς από τον Cross-Encoder του matcher. Σε έναν Bi-Encoder, κάθε εγγραφή κωδικοποιείται ανεξάρτητα σε ένα διάνυσμα σταθερών διαστάσεων (embedding), χωρίς αμοιβαία επεξεργασία με άλλη εγγραφή. Αυτό το χαρακτηριστικό αποτελεί την κύρια πηγή της ταχύτητάς του: τα διανύσματα όλων των εγγραφών της μιας βάσης μπορούν να υπολογιστούν μία φορά (offline), να αποθηκευτούν σε δομή ευρετηρίασης και να επαναχρησιμοποιηθούν για κάθε νέα αναζήτηση.

Ως μοντέλο κωδικοποίησης επιλέχθηκε το all-MiniLM-L6-v2 (Wang et al., 2020), για τους ίδιους λόγους που το καθιστούν ελκυστικό και στο στάδιο του matching: παράγει διανύσματα υψηλής σημασιολογικής ποιότητας (384 διαστάσεις) με εξαιρετικά χαμηλό υπολογιστικό κόστος, χάρη στη διαδικασία απόσταξης γνώσης (knowledge distillation) από μεγαλύτερα μοντέλα. Επιπλέον, η συνέπεια στο μοντέλο κωδικοποίησης μεταξύ Blocker

και του ελαφρού matcher του LexiSync απλοποιεί την παραγωγική διάταξη του συστήματος.

Η σειριοποίηση των εγγραφών σε αυτό το στάδιο παρουσιάζει ουσιώδη διαφοροποίηση σε σχέση με τη σειριοποίηση που εφαρμόζεται στον Cross-Encoder (Ενότητα 3.3.2). Ενώ ο matcher αξιοποιεί ειδικούς δείκτες (special tokens) όπως [T], [B], [P] για τη διατήρηση της δομικής πληροφορίας, ο Blocker προβαίνει σε απλή συνένωση όλων των πεδίων μιας εγγραφής σε ένα ενιαίο string. Αυτή η επιλογή είναι σκόπιμη: στο στάδιο του blocking δεν επιδιώκεται λεπτή σημασιολογική διάκριση, αλλά μια χονδρική, υψηλής ανάκλησης (high-recall) προσέγγιση των πιο σημασιολογικά κοντινών εγγραφών. Η απλούστερη σειριοποίηση μειώνει επίσης το κόστος επεξεργασίας και επιτρέπει στο μοντέλο να εστιάσει στη συνολική σημασιολογική «υπογραφή» κάθε εγγραφής.

### 3.8.2 Δομή Προσεγγιστικής Αναζήτησης με FAISS

Μετά την παραγωγή των διανυσματικών αναπαραστάσεων, το επόμενο πρόβλημα είναι η αποδοτική εύρεση των πιο όμοιων διανυσμάτων μεταξύ των δύο βάσεων. Για τον σκοπό αυτό αξιοποιήθηκε η βιβλιοθήκη FAISS (Facebook AI Similarity Search), η οποία αποτελεί state-of-the-art εργαλείο για προσεγγιστική αναζήτηση πλησιέστερων γειτόνων (Approximate Nearest Neighbor Search - ANNS) σε διανυσματικούς χώρους υψηλών διαστάσεων (Johnson et al., 2019).

Πριν την εισαγωγή των διανυσμάτων στο index, εφαρμόζεται κανονικοποίηση L2 (faiss.normalize\_L2) τόσο στα διανύσματα της πηγαίας όσο και της στόχου βάσης. Η κανονικοποίηση αυτή είναι μαθηματικά ισοδύναμη με τη μετατροπή του εσωτερικού γινομένου (Inner Product) σε ομοιότητα συνημίτονου (cosine similarity), μια καθιερωμένη μετρική για τη σύγκριση σημασιολογικών embeddings. Έτσι, η αναζήτηση πραγματοποιείται με βάση τη γωνιακή απόσταση μεταξύ διανυσμάτων και όχι με βάση το απόλυτο μέγεθός τους, διασφαλίζοντας ότι εγγραφές με μικρές διαφορές στην έκταση του κειμένου τους μπορούν να ταυτιστούν σωστά.

Στην παρούσα υλοποίηση χρησιμοποιείται ο τύπος ευρετηρίου IndexFlatIP της FAISS. Πρόκειται για έναν "flat" (επίπεδο) δείκτη που εκτελεί ακριβή (exhaustive) αναζήτηση κατά εσωτερικό γινόμενο, χωρίς απώλειες ακρίβειας. Η επιλογή αυτή προτιμήθηκε έναντι πιο εξελιγμένων δομών (όπως IndexIVFFlat ή HNSW), διότι για το μέγεθος των πειραματικών

συνόλων δεδομένων η ακριβής αναζήτηση παραμένει υπολογιστικά διαχειρίσιμη, ενώ απαλλάσσει τον σχεδιασμό από την παραμετροποίηση πρόσθετων υπερπαραμέτρων (π.χ. nlist, nprobe). Σε σενάρια πολύ μεγαλύτερης κλίμακας (εκατομμύρια εγγραφές), ο IndexFlatIP μπορεί εύκολα να αντικατασταθεί από έναν προσεγγιστικό δείκτη χωρίς αλλαγή στην υπόλοιπη αρχιτεκτονική του Blocker.

### 3.8.3 Στρατηγική Φιλτραρίσματος και Παραγωγής Υποψήφιων Ζευγών

Η παραγωγή των τελικών υποψήφιων ζευγών διέπεται από έναν συνδυασμό δύο μηχανισμών φιλτραρίσματος. Πρώτον, για κάθε εγγραφή της πηγαίας βάσης ανακτώνται μόνο οι top-k πιο όμοιες εγγραφές της στόχου βάσης, σύμφωνα με το εσωτερικό γινόμενο των κανονικοποιημένων διανυσμάτων τους. Αυτό περιορίζει εξ ορισμού τον αριθμό υποψηφίων ανά εγγραφή σε μια διαχειρίσιμη σταθερά. Δεύτερον, εφαρμόζεται ένα ελάχιστο κατώφλι ομοιότητας (threshold), κάτω από το οποίο τα ζεύγη απορρίπτονται, ακόμη και αν εμπίπτουν στους top-k γείτονες. Αυτή η δεύτερη παρέμβαση είναι κρίσιμη όταν η πηγαία εγγραφή δεν έχει πραγματικό match στη στόχο βάση: χωρίς το κατώφλι, ο Blocker θα παρήγαγε k υποψήφια ζεύγη ακόμη και για εγγραφές που δεν θα έπρεπε να ταυτιστούν με καμία από τη στόχο βάση, εισάγοντας θόρυβο στο επόμενο στάδιο.

Οι δύο αυτές υπερπαραμέτροι, top-k και threshold, ορίζουν τον θεμελιώδη συμβιβασμό μεταξύ Pairs Completeness (ποσοστό πραγματικών matches που διατηρούνται στο σύνολο υποψηφίων) και Pairs Quality (πυκνότητα πραγματικών matches εντός των υποψηφίων), ο οποίος αποτελεί κλασική πρόκληση στη βιβλιογραφία του blocking (Christen, 2012). Υψηλότερες τιμές του k και χαμηλότερα κατώφλια ευνοούν την Pairs Completeness εις βάρος της Pairs Quality, και αντιστρόφως. Στην παρούσα υλοποίηση επιλέχθηκαν συντηρητικές τιμές ( $k=10$  και  $\text{threshold}=0.4$ ), που δίνουν προτεραιότητα στη μέγιστη ανάκληση, καθώς ο επόμενος Cross-Encoder είναι αρκετά ισχυρός ώστε να φιλτράρει αποτελεσματικά τα false positives που θα διέλθουν από αυτό το αρχικό φίλτρο.

### 3.8.4 Πειραματική Αξιολόγηση του Μηχανισμού Blocking

Για την επαλήθευση της αποτελεσματικότητας του σχεδιασθέντος μηχανισμού, πραγματοποιήθηκε πειραματική αξιολόγηση στο σύνολο δεδομένων Walmart-Amazon, το οποίο αποτελεί το πιο απαιτητικό από τα τρία εξεταζόμενα benchmarks της εργασίας λόγω

της εκτενούς παρουσίας ελλιπών τιμών και της συντακτικής ετερογένειας μεταξύ των δύο πηγών. Η αξιολόγηση βασίστηκε στις τρεις καθιερωμένες μετρικές της βιβλιογραφίας (Christen, 2012). Η πρώτη, Pairs Completeness (PC), εκφράζει το ποσοστό των πραγματικών matches που διατηρούνται στο σύνολο των υποψηφίων ζευγών και αποτελεί το θεωρητικό άνω όριο για το recall του συνολικού end-to-end συστήματος. Η δεύτερη, Pairs Quality (PQ), εκφράζει την πυκνότητα των πραγματικών matches εντός των candidates και αποτυπώνει την «καθαρότητα» της εξόδου του blocker. Η τρίτη, Reduction Ratio (RR), ποσοτικοποιεί τη μείωση του χώρου αναζήτησης σε σχέση με την εξαντλητική σύγκριση.

Η πηγαία βάση του Walmart περιλαμβάνει 2.554 εγγραφές προϊόντων, ενώ η στόχος βάση του Amazon 22.074 εγγραφές, με αποτέλεσμα τον σχηματισμό 56.376.996 θεωρητικά πιθανών ζευγών για εξαντλητική σύγκριση. Το ground truth, συγκεντρωμένο από τη συνένωση των train, valid και test sets, περιέχει 962 πραγματικές ταυτίσεις. Με τη ρύθμιση top-k=10 και threshold=0.4 που επιλέχθηκε στην Ενότητα 3.8.3, ο Blocker παράγαγε 25.118 candidate pairs, εκ των οποίων τα 894 αντιστοιχούσαν σε πραγματικά matches. Τα αναλυτικά αποτελέσματα παρουσιάζονται στον Πίνακα 3.8-1.

**Πίνακας 3.8-1 Αποτελέσματα πειραματικής αξιολόγησης του Blocker στο σύνολο Walmart-Amazon**

| Μετρική                                      | Τιμή       |
|----------------------------------------------|------------|
| Εγγραφές πηγαίας βάσης (Walmart)             | 2.554      |
| Εγγραφές στόχου βάσης (Amazon)               | 22.074     |
| Σύνολο πιθανών ζευγών (εξαντλητική σύγκριση) | 56.376.996 |
| Πραγματικά matches (ground truth)            | 962        |
| Candidate pairs (έξοδος Blocker)             | 25.118     |
| Matches που ανακτήθηκαν                      | 894        |
| Matches που χάθηκαν                          | 68         |
| Pairs Completeness (PC)                      | 92,93%     |
| Pairs Quality (PQ)                           | 3,56%      |
| Reduction Ratio (RR)                         | 99,96%     |

Η ανάλυση των μετρικών επιβεβαιώνει την αποτελεσματικότητα της προτεινόμενης αρχιτεκτονικής. Η τιμή Pairs Completeness στο 92,93% υποδεικνύει ότι ο Blocker κατάφερε να διατηρήσει 894 από τα 962 πραγματικά matches του συνόλου ελέγχου, με κόστος μόλις 68 χαμένων ζευγών. Δεδομένου ότι αυτά τα 68 ζεύγη δεν μπορούν πλέον να ανακτηθούν στο επόμενο στάδιο του Cross-Encoder, αποτελούν το θεωρητικό άνω φράγμα του τελικού recall του συστήματος. Παράλληλα, ο Reduction Ratio του 99,96% αντικατοπτρίζει τη δραστική μείωση του υπολογιστικού φόρτου: από τα 56,4 εκατομμύρια θεωρητικά πιθανά ζεύγη, ο Cross-Encoder καλείται να εξετάσει μόλις 25.118 candidates, μια μείωση κατά τάξεις μεγέθους που καθιστά την εφαρμογή του συστήματος εφικτή σε ρεαλιστική κλίμακα.

Η σχετικά χαμηλή τιμή Pairs Quality (3,56%) δεν αποτελεί ένδειξη αδυναμίας αλλά συνέπεια της σχεδιαστικής φιλοσοφίας που υιοθετήθηκε. Όπως αναλύθηκε στην Ενότητα 3.8.3, ο Blocker σχεδιάστηκε με προτεραιότητα τη μεγιστοποίηση της ανάκλησης (high-recall semantics), βασιζόμενος στην ικανότητα του Cross-Encoder να φιλτράρει αποτελεσματικά τα false positives στο επόμενο στάδιο. Η προσέγγιση αυτή είναι σύμφωνη με την κρατούσα άποψη στη βιβλιογραφία (Papadakis et al., 2020), η οποία αντιμετωπίζει το blocking ως μηχανισμό υψηλής ανάκλησης και όχι ως αυτόνομο ταξινομητή υψηλής ακρίβειας. Συνολικά, τα παραπάνω αποτελέσματα επιβεβαιώνουν ότι η ενσωμάτωση του ANNS-based Blocker μετατρέπει επιτυχώς το LexiSync σε ολοκληρωμένο, end-to-end σύστημα Entity Resolution, ικανό να αντιμετωπίσει πραγματικά σενάρια εφαρμογής χωρίς την υπόθεση προ-φιλτραρισμένων υποψήφιων ζευγών.

## 4 Πειραματική Διαδικασία και Σύνολα Δεδομένων

### 4.1 Εισαγωγή στο Πειραματικό Πλαίσιο

Η αξιολόγηση ενός συστήματος αντιστοίχισης εγγραφών (Entity Resolution) προϋποθέτει τον σχεδιασμό ενός αυστηρού και επαναλήψιμου πειραματικού πλαισίου. Στο παρόν κεφάλαιο αναλύεται διεξοδικά η υποδομή πάνω στην οποία αναπτύχθηκε, εκπαιδεύτηκε και αξιολογήθηκε το προτεινόμενο σύστημα LexiSync. Αρχικά, παρουσιάζονται τα τρία (3) δημόσια διαθέσιμα σύνολα δεδομένων (datasets) που χρησιμοποιήθηκαν, καθένα από τα οποία παρουσιάζει διαφορετικά επίπεδα δομικής ετερογένειας και θορύβου. Στη συνέχεια,

περιγράφεται το υλικό (hardware) και το λογισμικό (software) που αξιοποιήθηκε, με ιδιαίτερη έμφαση στις ιδιαιτερότητες της αρχιτεκτονικής Apple Silicon και της τεχνολογίας Metal Performance Shaders (MPS). Τέλος, παρατίθεται μια εις βάθος ανάλυση των υπερπαραμέτρων (hyperparameters) που επιλέχθηκαν για τα στάδια της προ-εκπαίδευσης και της συνεχούς μάθησης (continual learning), καθώς και των μετρικών αξιολόγησης που χρησιμοποιήθηκαν για την εξαγωγή των τελικών συμπερασμάτων.

## 4.2 Σύνολα Δεδομένων (Datasets)

Για την πειραματική αξιολόγηση του συστήματος, επιλέχθηκαν τρία ευρέως αναγνωρισμένα σύνολα δεδομένων (benchmarks) από τη διεθνή βιβλιογραφία του Entity Resolution (Mudgal et al., 2018). Η επιλογή των συγκεκριμένων συνόλων δεν ήταν τυχαία, καθώς το καθένα εξυπηρετεί έναν ξεχωριστό σκοπό στη στρατηγική της μεταφοράς γνώσης (Domain Adaptation). Το πρώτο σύνολο προέρχεται από τον ακαδημαϊκό χώρο και λειτουργεί ως Πηγαίο Πεδίο (Source Domain), ενώ τα άλλα δύο προέρχονται από τον χώρο του ηλεκτρονικού εμπορίου (e-commerce) και αποτελούν τα Πεδία-Στόχους (Target Domains).

Κάθε σύνολο δεδομένων είναι οργανωμένο σε έναν τυποποιημένο φάκελο που περιέχει πέντε βασικά αρχεία σε μορφή CSV: το αρχείο source.csv (εγγραφές της πρώτης βάσης), το αρχείο target.csv (εγγραφές της δεύτερης βάσης), καθώς και τα αρχεία train.csv, valid.csv και test.csv, τα οποία περιέχουν τα ετικετοποιημένα ζεύγη (LTB\_ID, RTB\_ID, Label) που χρησιμοποιούνται για την εκπαίδευση, την επικύρωση (validation) και την τελική αξιολόγηση αντίστοιχα. (Li et al., 2021).

### 4.2.1 Το Σύνολο Δεδομένων DBLP-Scholar (Source Domain)

Το σύνολο δεδομένων DBLP-Scholar χρησιμοποιείται ευρέως για την αξιολόγηση αλγορίθμων ενοποίησης δεδομένων και αποτελείται από ακαδημαϊκές δημοσιεύσεις. Η μία βάση προέρχεται από την ψηφιακή βιβλιοθήκη DBLP, η οποία είναι εξαιρετικά δομημένη και καθαρή, ενώ η δεύτερη προέρχεται από το Google Scholar, το οποίο συλλέγει δεδομένα αυτόματα (crawling) και συνεπώς περιέχει σημαντικό θόρυβο, ανορθογραφίες και ελλιπή πεδία.

Στο σύστημα LexiSync, το συγκεκριμένο dataset χρησιμοποιήθηκε αποκλειστικά για την Φάση 1: Εκπαίδευση στο Πηγαίο Πεδίο. Ο στόχος ήταν το γλωσσικό μοντέλο να "μάθει"

τις βασικές αρχές της ταύτισης οντοτήτων σε ένα σχετικά ελεγχόμενο περιβάλλον. Για την αναπαράσταση των εγγραφών, απομονώθηκαν και χρησιμοποιήθηκαν τα εξής πεδία, συνοδευόμενα από τους αντίστοιχους ειδικούς δείκτες (special tokens):

[T] (Title): Ο τίτλος της ακαδημαϊκής δημοσίευσης.

[A] (Authors): Τα ονόματα των συγγραφέων, τα οποία συχνά παρουσιάζουν ασυνέπειες (π.χ. αναγραφή ολογράφως έναντι χρήσης αρχικών).

[V] (Venue): Το συνέδριο ή το περιοδικό (π.χ. "VLDB", "SIGMOD").

[Y] (Year): Το έτος δημοσίευσης, μια κρίσιμη αριθμητική τιμή για την αποφυγή ταύτισης άρθρων με παρόμοιους τίτλους αλλά διαφορετικές εκδόσεις.

#### 4.2.2 Το Σύνολο Δεδομένων Amazon-Google

Το σύνολο Amazon-Google αποτελεί ένα κλασικό benchmark στον τομέα του ηλεκτρονικού εμπορίου. Περιέχει δεδομένα λογισμικού και ηλεκτρονικών προϊόντων που αντλήθηκαν από τα καταστήματα της Amazon και του Google Base. Σε αντίθεση με το ακαδημαϊκό DBLP, τα δεδομένα του e-commerce παρουσιάζουν υψηλή συντακτική ετερογένεια. Οι τίτλοι των προϊόντων στην Amazon είναι συνήθως μακροσκελείς, γεμάτοι διαφημιστικούς όρους (buzzwords), ενώ στο Google Base τείνουν να είναι πιο λακωνικοί (Körpcke et al., 2010).

Αυτό το σύνολο χρησιμοποιήθηκε ως πεδίο-στόχος κατά τη διαδικασία του Domain Adaptation. Η δομή της σειριοποίησης περιελάμβανε τα εξής χαρακτηριστικά:

[T] (Title): Η πλήρης περιγραφή / τίτλος του προϊόντος.

[M] (Manufacturer): Ο κατασκευαστής του προϊόντος (π.χ. "Microsoft", "Adobe").

[P] (Price): Η τιμή. Αξίζει να σημειωθεί ότι η τιμή αποτελεί ένα εξαιρετικά θορυβώδες χαρακτηριστικό, καθώς το ίδιο προϊόν ενδέχεται να έχει διαφορετική τιμολόγηση στις δύο πλατφόρμες.

#### 4.2.3 Το Σύνολο Δεδομένων Walmart-Amazon

Το σύνολο Walmart-Amazon θεωρείται ένα από τα πιο απαιτητικά και δύσκολα σύνολα δεδομένων για Entity Resolution. Περιλαμβάνει ηλεκτρονικά είδη (όπως τηλεοράσεις, υπολογιστές, αξεσουάρ) από τα δύο κορυφαία αμερικανικά πολυκαταστήματα. Η κύρια πρόκληση σε αυτό το dataset είναι η έλλειψη τυποποίησης και η δραματική απουσία τιμών (missing values) σε δομημένα πεδία (Mudgal et al., 2018). Συχνά, κρίσιμες πληροφορίες

όπως το μοντέλο της συσκευής δεν βρίσκονται στη στήλη "Model No", αλλά είναι κρυμμένες μέσα στο ελεύθερο κείμενο του τίτλου.

Για την καλύτερη δυνατή ανάδειξη της πληροφορίας, το LexiSync διαμόρφωσε την είσοδο χρησιμοποιώντας πέντε διακριτούς δείκτες:

[T] (Title): Ο τίτλος του ηλεκτρονικού είδους.

[N] (ModelNo): Ο αριθμός μοντέλου, ο οποίος είναι το πιο ισχυρό (αλλά συχνά απόν) χαρακτηριστικό ταύτισης.

[B] (Brand): Η εμπορική μάρκα (π.χ. "Sony", "Samsung").

[C] (Category): Η κατηγορία του προϊόντος (π.χ. "Electronics - TV").

[P] (Price): Η τιμή του προϊόντος, η οποία υποβλήθηκε σε αυστηρό καθαρισμό.

### **4.3 Υπολογιστικό Περιβάλλον και Υλικό (Hardware & Software Setup)**

Λόγω της εξαιρετικά υψηλής υπολογιστικής πολυπλοκότητας που απαιτεί η εκπαίδευση γλωσσικών μοντέλων αρχιτεκτονικής Transformer (συχνά της τάξης των εκατοντάδων εκατομμυρίων παραμέτρων), η επιλογή και παραμετροποίηση του κατάλληλου υλικού αποτελεί κρίσιμο παράγοντα για την επιτυχή διεξαγωγή του πειράματος.

#### **4.3.1 Αρχιτεκτονική Apple Silicon και Τεχνολογία MPS**

Το σύνολο των πειραμάτων εκτελέστηκε τοπικά σε περιβάλλον macOS, αξιοποιώντας την αρχιτεκτονική επεξεργαστών Apple Silicon (M-series). Μέχρι πρόσφατα, η εκπαίδευση μοντέλων Deep Learning ήταν μονοπώλιο των καρτών γραφικών της NVIDIA μέσω της πλατφόρμας CUDA. Ωστόσο, η ενσωμάτωση του υποσυστήματος Metal Performance Shaders (MPS) στο PyTorch επέτρεψε την επιτάχυνση των υπολογισμών τανυστών (tensors) απευθείας στην ενοποιημένη μνήμη (Unified Memory) των επεξεργαστών της Apple (Paszke et al., 2019).

Κατά τη φάση της ανάπτυξης, εντοπίστηκε μια σημαντική τεχνική πρόκληση: ορισμένες μαθηματικές πράξεις (όπως συγκεκριμένοι υπολογισμοί απωλειών) δεν υποστηρίζονται εγγενώς από το backend του MPS. Αυτό οδήγησε σε απροσδόκητες διακοπές (crashes) κατά τη διάρκεια του backpropagation. Για την επίλυση αυτού του προβλήματος, ενσωματώθηκε στον κώδικα (αμέσως μετά την εισαγωγή των βιβλιοθηκών) η εντολή

`os.environ["PYTORCH_ENABLE_MPS_FALLBACK"] = "1"`. Η ρύθμιση αυτή διασφαλίζει ότι, όποτε το σύστημα συναντά μια πράξη μη συμβατή με το Metal, μεταφέρει δυναμικά και προσωρινά τον υπολογισμό στην Κεντρική Μονάδα Επεξεργασίας (CPU), διασφαλίζοντας την αδιάλειπτη και σταθερή ροή της εκπαίδευσης. Επιπλέον, δηλώθηκε ρητά ο τύπος των δεδομένων ως `torch.float32` κατά τη φόρτωση του μοντέλου, προκειμένου να αποφευχθούν φαινόμενα απώλειας ακρίβειας (precision loss) που εμφανίζονται σε αρχιτεκτονικές χαμηλής κατανάλωσης ενέργειας.

#### 4.3.2 Λογισμικό και Βιβλιοθήκες

Ο πυρήνας του συστήματος αναπτύχθηκε στη γλώσσα προγραμματισμού Python. Η διαχείριση, μορφοποίηση και φόρτωση των μοντέλων βασίστηκε στη βιβλιοθήκη `transformers` της εταιρείας Hugging Face (Wolf et al., 2020), η οποία αποτελεί το διεθνές βιομηχανικό πρότυπο για την Επεξεργασία Φυσικής Γλώσσας.

Για την αποδοτική διαχείριση της μνήμης κατά την εισαγωγή των δεδομένων, επιλέχθηκε η βιβλιοθήκη `datasets` (της ίδιας εταιρείας), η οποία υποστηρίζει επεξεργασία δεδομένων μέσω `memory-mapping` (χρήση του σκληρού δίσκου ως επέκταση της μνήμης RAM). Ο χειρισμός των αρχείων CSV πραγματοποιήθηκε με την παραδοσιακή βιβλιοθήκη `pandas`, ενώ οι υπολογισμοί των μετρικών αξιολόγησης και η κατασκευή των πινάκων σύγχυσης (`confusion matrices`) υλοποιήθηκαν μέσω των λειτουργιών της `scikit-learn` (Pedregosa et al., 2011).

#### 4.3.3 Διασφάλιση Αναπαραγωγιμότητας (Reproducibility)

Ένας βασικός κανόνας της μηχανικής μάθησης είναι η ικανότητα αναπαραγωγής των ίδιων ακριβώς αποτελεσμάτων σε επαναλαμβανόμενες εκτελέσεις του κώδικα. Στα μοντέλα `Transformer`, υπάρχουν πολλά σημεία τυχαιότητας, όπως η αρχικοποίηση των βαρών της κεφαλής ταξινόμησης (`classification head`), ο ανακατεμός (`shuffling`) των δεδομένων εκπαίδευσης και το `dropout`. Για τον λόγο αυτό, ενσωματώθηκε ένα αυστηρό πρωτόκολλο σταθεροποίησης σπόρων τυχαιότητας (`seeds`). Η τιμή 42 αποδόθηκε καθολικά στις βιβλιοθήκες `random`, `numpy.random` και `torch.manual_seed`. Με αυτόν τον τρόπο, διασφαλίζεται ότι οι διαφορές στην απόδοση των διαφόρων μοντέλων οφείλονται αποκλειστικά στην αρχιτεκτονική τους και όχι σε στοχαστικούς παράγοντες κατά τη φάση της εκπαίδευσης.

## 4.4 Υπερπαράμετροι Εκπαίδευσης (Hyperparameters)

Η επιλογή των υπερπαραμέτρων (όπως ο ρυθμός μάθησης και το μέγεθος της παρτίδας) είναι ίσως το πιο κρίσιμο στάδιο κατά τη μικρο-ρύθμιση (fine-tuning) προ-εκπαιδευμένων μοντέλων. Η αρχιτεκτονική LexiSync χωρίζει την εκπαίδευση σε δύο διακριτές φάσεις, καθεμία από τις οποίες απαιτεί διαφορετική προσέγγιση (Sun et al., 2019).

### 4.4.1 Υπερπαράμετροι Φάσης 1 (Source Domain Training)

Κατά την πρώτη φάση, το μοντέλο εκπαιδεύεται εξολοκλήρου στο ακαδημαϊκό σύνολο DBLP-Scholar. Επειδή το μοντέλο (π.χ. DeBERTa-v3) προέρχεται από γενική προ-εκπαίδευση σε κείμενα του διαδικτύου (Wikipedia, BooksCorpus), τα βάρη του είναι μεν ισχυρά, αλλά δεν έχουν προσαρμοστεί στην εργασία της ταξινόμησης ζευγών (Sequence Classification). Για τον λόγο αυτό, επιλέχθηκαν οι εξής ρυθμίσεις:

Μέγεθος Παρτίδας (Batch Size): Ορίστηκε στην τιμή 12. Αυτή η επιλογή υπαγορεύτηκε από τους αυστηρούς περιορισμούς μνήμης (VRAM) του συστήματος. Ενώ τιμές όπως το 32 θεωρούνται ιδανικές, το batch size 12 με χρήση Δυναμικής Αναπλήρωσης (Dynamic Padding) επέτρεψε την αποφυγή σφαλμάτων "Out Of Memory (OOM)", διασφαλίζοντας παράλληλα μια αξιοπρεπή εκτίμηση της κλίσης (gradient).

Εποχές (Epochs): Ορίστηκαν στις 5. Πειραματικές δοκιμές έδειξαν ότι μετά την 5η εποχή, το μοντέλο άρχιζε να εμφανίζει σημάδια υπερπροσαρμογής (overfitting) στα δεδομένα εκπαίδευσης, μειώνοντας την ικανότητά του να γενικεύει σε άγνωστα δεδομένα.

Ρυθμός Μάθησης (Learning Rate): Επιλέχθηκε η συντηρητική τιμή  $2 \times 10^{-5}$ . Σύμφωνα με τους συγγραφείς του BERT (Devlin et al., 2018), ρυθμοί μάθησης μεταξύ  $1 \times 10^{-5}$  και  $5 \times 10^{-5}$  είναι οι βέλτιστοι για fine-tuning, καθώς αποτρέπουν την "καταστροφή" των πολύτιμων χαρακτηριστικών που απέκτησε το μοντέλο κατά την αρχική του προ-εκπαίδευση.

Απόσβεση Βαρών (Weight Decay): Καθορίστηκε στο 0.01. Η παράμετρος αυτή επιβάλλει κανονικοποίηση  $L_2$  (L2 Regularization), "τιμωρώντας" τα υπερβολικά μεγάλα βάρη του δικτύου και προσφέροντας σταθερότητα.

#### 4.4.2 Υπερπαράμετροι Φάσης 2 (Domain Adaptation)

Στη δεύτερη φάση, το μοντέλο έχει ήδη "κατανοήσει" το πρόβλημα του Entity Resolution και καλείται να προσαρμοστεί σε νέα δεδομένα ηλεκτρονικού εμπορίου, διατηρώντας ταυτόχρονα τη μνήμη του μέσω του Replay Buffer. Οι απαιτήσεις εδώ διαφοροποιούνται: Μέγεθος Replay Buffer: Από το DBLP-Scholar ανασύρθηκαν τυχαία 1.200 δείγματα. Η επιλογή του αριθμού αυτού προσφέρει τη χρυσή τομή: είναι αρκετά μεγάλος ώστε να αντιπροσωπεύει επαρκώς την ποικιλομορφία των παλαιών δεδομένων αποτρέποντας τη λήθη, αλλά αρκετά μικρός ώστε να μην κυριαρχεί χρονικά και υπολογιστικά στο νέο Ενοποιημένο Σύνολο (Joint Dataset).

Εποχές (Epochs): Αυξήθηκαν στις 7. Επειδή το νέο Dataset περιέχει τη συνένωση τριών πηγών (Replay DBLP + Amazon + Walmart), ο όγκος και η πολυπλοκότητα είναι τεράστια. Οι 7 εποχές δίνουν τον απαραίτητο χρόνο στο δίκτυο να "χωνέψει" την ετερογένεια των κατηγοριών.

Ρυθμός Μάθησης (Learning Rate): Μειώθηκε σημαντικά στην τιμή  $1 \times 10^{-5}$ . Καθώς το μοντέλο διαθέτει ήδη προσαρμοσμένα βάρη από τη Φάση 1, χρειαζόμαστε εξαιρετικά "απαλά" και προσεκτικά βήματα βελτιστοποίησης (gradient steps) για να μην καταρρεύσει η προηγούμενη γνώση.

#### 4.4.3 Πειραματική Τεκμηρίωση Μειωτικού Συντελεστή Βάρους ( $w_1$ )

Όπως αναλύθηκε στη μεθοδολογία του συστήματος (Ενότητα 3.4.2), για τον μηχανισμό Weighted Trainer υιοθετήθηκε ο μειωτικός συντελεστής  $w_1 = 0,5$ , με στόχο τη βέλτιστη εξισορρόπηση μεταξύ Ακρίβειας (Precision) και Ανάκλησης (Recall). Στην παρούσα ενότητα, η επιλογή αυτή τεκμηριώνεται πειραματικά μέσω συστηματικής μελέτης ευαισθησίας (sensitivity analysis), σύμφωνα με τις καθιερωμένες πρακτικές της βιβλιογραφίας για την αντιμετώπιση της ανισορροπίας κλάσεων (Buda et al., 2018; Cui et al., 2019).

Για την αξιολόγηση της ευαισθησίας του μοντέλου στην επιλογή του συντελεστή  $w_1$ , εκτελέστηκαν τρία ανεξάρτητα πειράματα εκπαίδευσης με διαφορετικές τιμές:  $w_1 \in \{0.3, 0.5, 1.0\}$ . Οι τιμές αυτές καλύπτουν τρεις αντιπροσωπευτικές περιπτώσεις:

(α)  $w_1 = 0.3$ : αυξημένη μείωση του βάρους της θετικής κλάσης

(β)  $w_1 = 0.5$ : η αρχική επιλογή της εργασίας (μέση μείωση)

(γ)  $w_1 = 1.0$ : καθόλου μείωση (πλήρες αντίστροφο της συχνότητας)

Σε κάθε πείραμα εφαρμόστηκε stratified sub-sampling 50% στο σύνολο εκπαίδευσης, διασφαλίζοντας τη διατήρηση της αρχικής αναλογίας θετικών και αρνητικών δειγμάτων κάθε dataset (1:4,37 για το DBLP-Scholar, 1:8,83 για το Amazon-Google και 1:9,67 για το Walmart-Amazon). Η μέθοδος αυτή επιλέχθηκε για λόγους υπολογιστικής αποτελεσματικότητας, διατηρώντας ταυτόχρονα την στατιστική αντιπροσωπευτικότητα του υποσυνόλου (Forman & Scholz, 2010). Η κύρια μεθοδολογία εκπαίδευσης (full fine-tuning του DeBERTa-v3, learning rate  $2e-5$ , two-phase Continual Learning με Replay Buffer) διατηρήθηκε πανομοιότυπη με την κύρια εκπαίδευση της εργασίας.

Για κάθε συνδυασμό ( $w_1$ , dataset) εφαρμόστηκε δυναμική βελτιστοποίηση threshold (threshold tuning) στο εύρος  $[0,2 - 0,95]$  με βήμα 0,05, αναζητώντας την τιμή που μεγιστοποιεί το F1-Score της θετικής κλάσης.

Η πρακτική αυτή είναι standard στη βιβλιογραφία class imbalance (Esposito et al., 2021), καθώς διασφαλίζει ότι κάθε ρύθμιση  $w_1$  αξιολογείται στη βέλτιστη εκδοχή της.

Τα αποτελέσματα της μελέτης παρουσιάζονται στον Πίνακα 4.4-1, ανά dataset και τιμή του συντελεστή.

Πίνακας 4.4-1 F1-Scores ανά dataset και τιμή του  $w_1$

| Dataset        | $w_1=0,3$ | $w_1=0,5$ | $w_1=1,0$ | Best $w_1$ |
|----------------|-----------|-----------|-----------|------------|
| DBLP-Scholar   | 0.9404    | 0.9456    | 0.9441    | 0.5        |
| Amazon-Google  | 0.7353    | 0.7102    | 0.7250    | 0.3        |
| Walmart-Amazon | 0.8640    | 0.8733    | 0.8770    | 1.0        |
| Mean F1        | 0.8466    | 0.8430    | 0.8487    | 1.0        |

Σε επίπεδο aggregate απόδοσης (μέσος όρος των τριών datasets), η τιμή  $w_1 = 1.0$  αναδεικνύεται ως η βέλτιστη επιλογή με Mean F1-Score 0.8487, ξεπερνώντας τόσο την τιμή 0.3 (Mean F1 = 0.8466) όσο και την τιμή 0.5 (Mean F1 = 0.8430). Σε επίπεδο επιμέρους datasets, παρατηρούμε μια ενδιαφέρουσα κατανομή, καθώς κάθε τιμή του συντελεστή βελτιστοποιεί την απόδοση σε διαφορετικό περιβάλλον: Στο Walmart-Amazon, ο

συντελεστής 1.0 πετυχαίνει την υψηλότερη απόδοση ( $F1 = 0.8770$ ), επιβεβαιώνοντας ότι στα πιο θορυβώδη και απαιτητικά σύνολα απαιτείται η πλήρης ποινή ανισορροπίας. Στο DBLP-Scholar, ο συντελεστής 0.5 επιτυγχάνει την υψηλότερη απόδοση ( $F1 = 0.9456$ ). Στο Amazon-Google, η υψηλότερη απόδοση επιτυγχάνεται με  $w_1 = 0.3$  ( $F1 = 0.7353$ ).

Σημαντικό εύρημα της μελέτης αποτελεί η παρατήρηση ότι όλες οι δοκιμασμένες τιμές του  $w_1$  παράγουν ιδιαίτερα κοντινά aggregate F1-Scores, με τη μέγιστη διαφορά μεταξύ της καλύτερης και της χειρότερης απόδοσης (1.0 έναντι 0.5) να είναι μόλις 0.0057 (ποσοστιαία απόκλιση κάτω του 0.7%). Αυτή η συμπεριφορά plateau υποδεικνύει ότι το μοντέλο είναι εξαιρετικά ανθεκτικό (robust) στην επιλογή του συγκεκριμένου συντελεστή. Η παρατήρηση αυτή είναι σύμφωνη με τα ευρήματα της βιβλιογραφίας για Transformer-based μοντέλα fine-tuned σε ταξινομητικές εργασίες (Mosbach et al., 2021), όπου η pre-trained αναπαράσταση παρέχει ισχυρή γενικευτική ικανότητα που μειώνει την ευαισθησία του τελικού μοντέλου σε hyperparameter tuning. Επιπλέον, σύμφωνα με τους Henning et al. (2023), σε ταξινόμηση κειμένου με μέτρια ανισορροπία (όπως η παρούσα, με ratios 1:4 έως 1:10), η αρχιτεκτονική σπανίως απαιτεί ακραίες τιμές class weights για τη βέλτιστη απόδοση.

Συμπερασματικά, αν και η τιμή  $w_1 = 1,0$  καταγράφει τυπικά το υψηλότερο μέσο F1-Score (0,8487), η απόκλιση από την αρχική εμπειρική επιλογή του 0,5 (0,8430) είναι πρακτικά αμελητέα. Η ευρεία αυτή ζώνη plateau (μέγιστη διαφορά  $< 1\%$  μεταξύ όλων των δοκιμασμένων τιμών) αποδεικνύει την ισχυρή ανθεκτικότητα (robustness) της αρχιτεκτονικής. Το μοντέλο καταφέρνει να μαθαίνει τα ουσιαστικά σημασιολογικά χαρακτηριστικά των εγγραφών ανεξαρτήτως της απόλυτα ακριβούς ρύθμισης του βάρους της συνάρτησης κόστους. Αυτό το ποιοτικό χαρακτηριστικό είναι εξαιρετικά επιθυμητό για το σύστημα LexiSync, ιδιαίτερα στο πλαίσιο της συνεχούς μάθησης (Continual Learning). Καθιστά σαφές ότι η αρχιτεκτονική μπορεί να εφαρμοστεί με ασφάλεια σε νέα, άγνωστα πεδία (Domain Adaptation) διατηρώντας τον αρχικό συντελεστή 0,5, χωρίς να απαιτείται η κοστοβόρα και χρονοβόρα διαδικασία της εκ νέου βαθμονόμησης (hyperparameter tuning) για κάθε νέο σύνολο δεδομένων που εισάγεται στο σύστημα.

## 4.5 Μετρικές Αξιολόγησης (Evaluation Metrics)

Δεδομένης της ακραίας ανισορροπίας των κλάσεων στα σύνολα δεδομένων (τα Non-Matches είναι συνήθως το 80-90% των περιπτώσεων), η μετρική της Ακρίβειας (Accuracy) θεωρείται παραπλανητική. Ένα μοντέλο που απαντάει μονίμως "Non-Match" θα επιτύγχανε 90% Accuracy, όντας πρακτικά άχρηστο (Christen, 2012).

Για τον λόγο αυτό, η εξαγωγή συμπερασμάτων (μέσω της συνάρτησης `generate_report`) βασίστηκε στις εξής μετρικές:

**Ακρίβεια (Precision):** Εκφράζει το ποσοστό των ζευγών που το μοντέλο προέβλεψε ως Matches, τα οποία ήταν όντως Matches. Ένα χαμηλό Precision υποδηλώνει την ύπαρξη πολλών False Positives (λανθασμένες ταυτίσεις).

**Ανάκληση (Recall):** Εκφράζει το ποσοστό των πραγματικών Matches που το μοντέλο κατάφερε να εντοπίσει. Ένα χαμηλό Recall σημαίνει ότι το σύστημα "χάνει" πολλά διπλότυπα προϊόντα (False Negatives).

**F1-Score:** Είναι ο αρμονικός μέσος όρος του Precision και του Recall

$$2 * \left( \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

Το F1-Score αποτελεί τον πιο αυστηρό και αντιπροσωπευτικό δείκτη αξιολόγησης για την παρούσα εργασία, καθώς επιβάλλει ισορροπία: εάν μία από τις δύο προηγούμενες μετρικές καταρρεύσει, το F1-Score θα πέσει δραματικά.

Τέλος, είναι κρίσιμο να αναφερθεί ότι οι τελικές τιμές αυτών των μετρικών εξήχθησαν μετά την εφαρμογή του αλγορίθμου Βελτιστοποίησης Κατωφλίου (Threshold Tuning). Το σύστημα δεν έλαβε στατικά την απόφαση με βάση το παραδοσιακό κατώφλι του 0.5, αλλά ερεύνησε δυναμικά το πιθανολογικό εύρος από 0.2 έως 0.95, προκειμένου να μεγιστοποιήσει το F1-Score για κάθε ξεχωριστό Dataset.

## 5 Αποτελέσματα και Σχολιασμός

### 5.1 Εισαγωγή στα Αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζεται η εκτενής πειραματική αξιολόγηση του συστήματος LexiSync, στοχεύοντας στην ανάδειξη της αποτελεσματικότητας των σύγχρονων αρχιτεκτονικών Transformer στο πρόβλημα της αντιστοίχισης εγγραφών (Entity

Resolution). Η μελέτη εστιάζει στη συγκριτική ανάλυση έξι (6) διαφορετικών προ-εκπαιδευμένων γλωσσικών μοντέλων, εξετάζοντας την ικανότητά τους να γενικεύουν από ένα καλά δομημένο ακαδημαϊκό πεδίο (DBLP-Scholar) σε ιδιαίτερα θορυβώδη εμπορικά περιβάλλοντα (Amazon-Google και Walmart-Amazon). Μέσω αυτής της αξιολόγησης, δεν αποτυπώνεται απλώς η τελική ορθότητα των προβλέψεων, αλλά αναδεικνύεται στην πράξη ο τρόπος με τον οποίο τα νευρωνικά δίκτυα διαχειρίζονται τη σημασιολογική ετερογένεια και τις ελλειπείς πληροφορίες.

Βασικό άξονα της πειραματικής διαδικασίας αποτελεί η σύγκριση αποδοτικότητας (Efficiency vs. Accuracy). Τα εξεταζόμενα μοντέλα κυμαίνονται από ισχυρές αρχιτεκτονικές (όπως το DeBERTa-v3-base) έως εξαιρετικά ελαφριά μοντέλα (όπως τα MiniLM-L6, e5-small-v2 και gte-small), προκειμένου να εντοπιστεί η βέλτιστη λύση για περιβάλλοντα παραγωγής. Παράλληλα, αξιολογούνται κρίσιμες λειτουργίες του συστήματος, όπως η επιτυχία της δυναμικής βελτιστοποίησης κατωφλίου (Threshold Tuning) έναντι στατικών μεθόδων, καθώς και η ανθεκτικότητα του μηχανισμού Experience Replay (Replay Buffer) στην αποτροπή της καταστροφικής λήθης κατά τη μεταφορά γνώσης.

Η ανάλυση που ακολουθεί επικεντρώνεται κυρίως στη θετική κλάση (Class 1 - Matches), καθώς ο εντοπισμός των ταυτίσεων αποτελεί τον πυρήνα του προβλήματος (Christen, 2012). Λόγω της εγγενούς ανισορροπίας μεταξύ θετικών και αρνητικών ζευγών, ως κύρια μετρική σύγκρισης χρησιμοποιείται το F1-Score, το οποίο προσφέρει μια ισορροπημένη εικόνα μεταξύ της Ακρίβειας (Precision) και της Ανάκλησης (Recall). Προκειμένου να υπάρξει μια πλήρης και διαφανής αποτίμηση, τα στατιστικά αποτελέσματα συνοδεύονται από ποιοτική ανάλυση σφαλμάτων (Error Analysis) για την κατανόηση των ψευδώς θετικών και αρνητικών προβλέψεων, καθώς και από εις βάθος μελέτη της ανταλλαγής μεταξύ Ακρίβειας και Ανάκλησης στα διαφορετικά πεδία.

Τέλος, η αξιολόγηση ολοκληρώνεται με τη δομική σύγκριση των αρχιτεκτονικών Cross-Encoder και Bi-Encoder, τόσο σε σενάρια μηδενικής μάθησης όσο και πλήρους μικρο-ρύθμισης. Η σύγκριση αυτή πλαισιώνεται από τη λεπτομερή ανάλυση του υπολογιστικού κόστους ανά φάση εκτέλεσης (Embedding, Blocking, Matching), τεκμηριώνοντας έτσι με απόλυτα μετρήσιμα κριτήρια την τελική επιλογή σχεδιασμού για κλιμακούμενα συστήματα ενοποίησης δεδομένων.

## 5.2 Απόδοση στο Πηγαίο Πεδίο (Source Domain - DBLP)

Η πρώτη φάση της αξιολόγησης αφορά την επίδοση των μοντέλων στο σύνολο δεδομένων DBLP-Scholar, πάνω στο οποίο πραγματοποιήθηκε η αρχική εκπαίδευση (Phase 1). Επειδή το συγκεκριμένο dataset διαθέτει ισχυρή δομή (Τίτλος, Συγγραφείς, Έτος), αναμενόταν υψηλή απόδοση από το σύνολο των αρχιτεκτονικών.

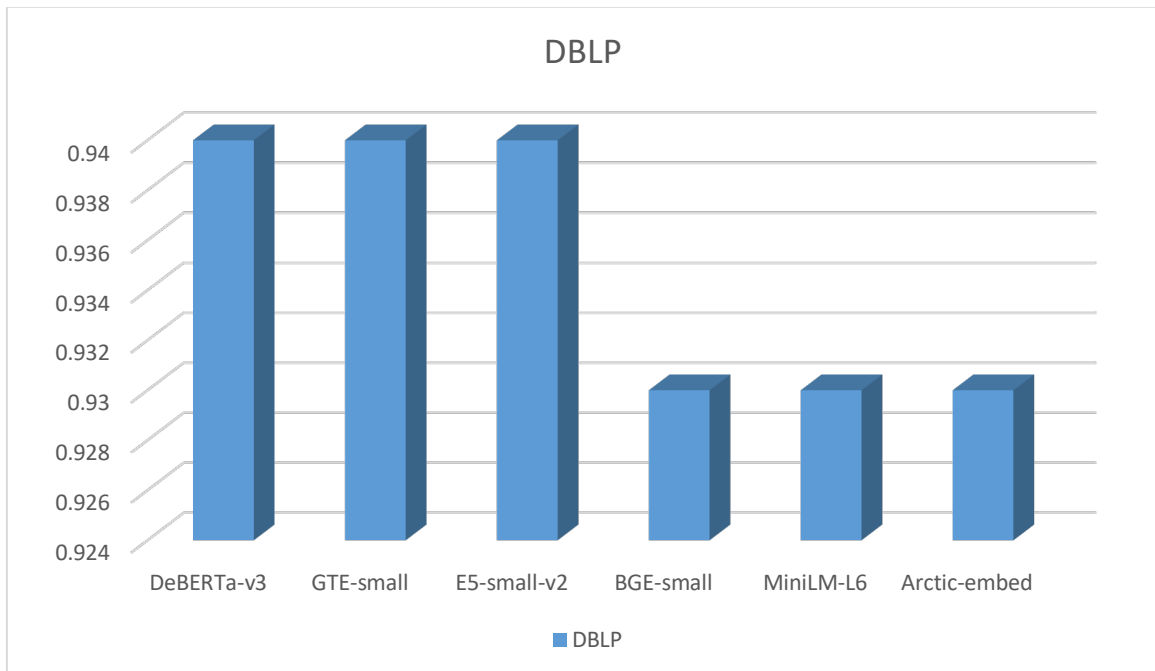
Πίνακας 5.2-1 Συγκριτικά αποτελέσματα στο σύνολο DBLP-Scholar

| Μοντέλο                                | Precision | Recall | F1-Score | Βέλτιστο Κατώφλι |
|----------------------------------------|-----------|--------|----------|------------------|
| microsoft/deberta-v3-base              | 0.9431    | 0.9449 | 0.9440   | 0.80             |
| thenlper/gte-small                     | 0.9412    | 0.9430 | 0.9421   | 0.20             |
| intfloat/e5-small-v2                   | 0.9491    | 0.9234 | 0.9360   | 0.50             |
| sentence-transformers/all-MiniLM-L6-v2 | 0.9362    | 0.9318 | 0.9340   | 0.20             |
| BAAI/bge-small-en-v1.5                 | 0.9418    | 0.9234 | 0.9325   | 0.45             |
| Snowflake/arctic-embed-xs              | 0.9313    | 0.9243 | 0.9278   | 0.90             |

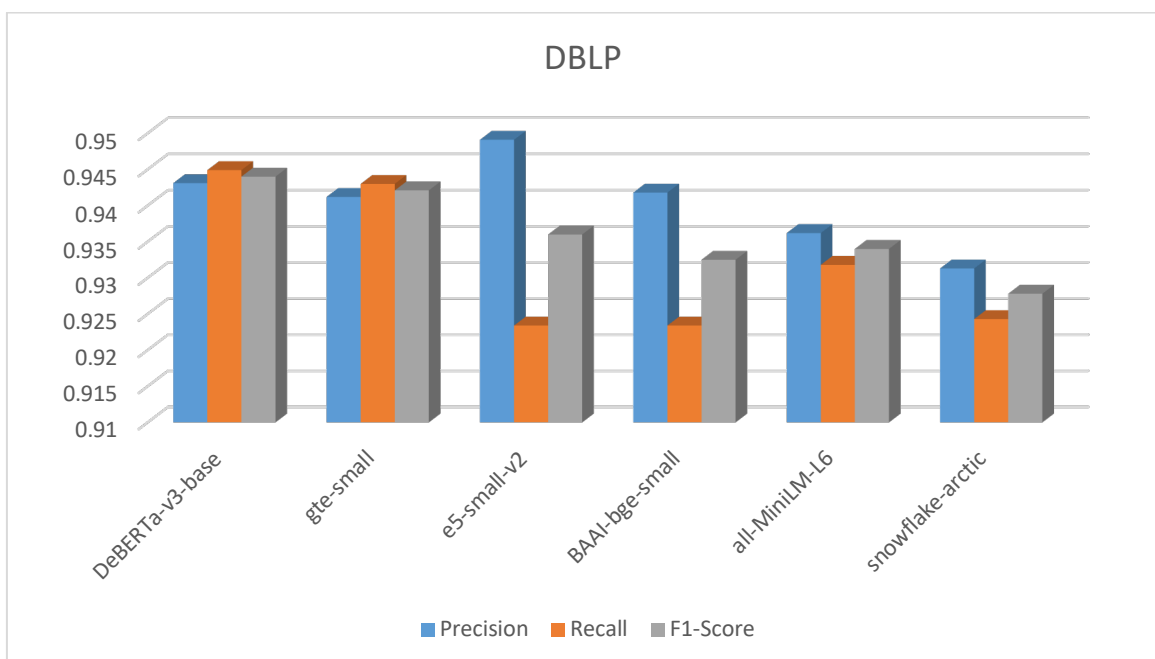
Ξεκινώντας την ανάλυση από το πηγαίο πεδίο εκπαίδευσης (DBLP-Scholar), τα ευρήματα του Πίνακα 5.2-1 επιβεβαιώνουν την ισχυρή ικανότητα των προ-εκπαιδευμένων γλωσσικών μοντέλων να επιλύουν προβλήματα αντιστοίχισης σε δομημένα περιβάλλοντα. Όπως παρατηρείται, το σύνολο των αρχιτεκτονικών επέδειξε εξαιρετική προσαρμοστικότητα, καταγράφοντας επιδόσεις (F1-Score) σταθερά υψηλότερες του 0.93. Η καθολική αυτή επιτυχία οφείλεται στην "καθαρή", ακαδημαϊκή φύση των δεδομένων: η σαφής διάκριση των πεδίων (όπως ο Τίτλος και το Έτος) και η απουσία ακραίου συντακτικού θορύβου διευκολύνουν τον μηχανισμό προσοχής (attention) των Transformers να "ευθυγραμμίσει" σωστά τις σχετικές πληροφορίες μεταξύ των δύο εγγραφών.

Την υψηλότερη απόλυτη επίδοση κατέγραψε το μοντέλο microsoft/deberta-v3-base, με F1-Score 0.9440. Αξίζει να σημειωθεί ότι στο δομημένο αυτό πεδίο η διαφορά από τα κορυφαία ελαφριά μοντέλα είναι οριακή (μόλις 0.0019 έναντι του gte-small), και δεδομένης της αξιολόγησης σε μία εκτέλεση με σταθερό seed, δεν μπορεί να θεωρηθεί στατιστικά σημαντική. Η ελαφρά υπεροχή του DeBERTa είναι ωστόσο συνεπής με τα χαρακτηριστικά της αρχιτεκτονικής του. Η ενσωμάτωση του μηχανισμού Αποσυνδεδεμένης Προσοχής (Disentangled Attention) του επιτρέπει να διατηρεί δύο ξεχωριστά διανύσματα για κάθε λέξη: ένα που αναπαριστά το σημασιολογικό της περιεχόμενο και ένα που κωδικοποιεί τη σχετική της θέση μέσα στην ακολουθία (He et al., 2020). Στο πλαίσιο του Entity Resolution, αυτή η ιδιότητα είναι καθοριστική. Για παράδειγμα, η ακριβής θέση μιας χρονολογίας ή ενός αύξοντος αριθμού έκδοσης μέσα σε έναν μακροσκελή τίτλο δημοσίευσης αρκεί από μόνη της για να διαφοροποιήσει δύο κατά τα άλλα πανομοιότυπες εγγραφές.

Εξίσου αν όχι περισσότερο αξιοσημείωτη κρίνεται η συμπεριφορά των δραστηρίων ελαφρύτερων αρχιτεκτονικών. Μοντέλα όπως το gte-small (0.9421) και το e5-small-v2 (0.9360), αν και διαθέτουν υποπολλαπλάσιο αριθμό παραμέτρων (περίπου 33 εκατομμύρια έναντι των 86 εκατομμυρίων του DeBERTa), καταφέρνουν να ανταγωνιστούν ευθέως τη "ναυαρχίδα" του πειράματος. Η επίδοση αυτή υποδεικνύει ότι τα μοντέλα που έχουν υποστεί σύγχρονη βελτιστοποίηση μέσω τεχνικών όπως η απόσταξη γνώσης (knowledge distillation), μπορούν να προσεγγίσουν το θεωρητικό μέγιστο της ακρίβειας στο πηγαίο πεδίο, προσφέροντας ταυτόχρονα τεράστια εξοικονόμηση υπολογιστικών πόρων ένα εύρημα που επιβεβαιώνει τη βιωσιμότητά τους για συστήματα μεγάλης κλίμακας.



Εικόνα 5.2-1 Επιδόσεις των γλωσσικών μοντέλων (F1-Score) στο ακαδημαϊκό σύνολο δεδομένων DBLP.



Εικόνα 5.2-2 Σύγκριση Precision, Recall και F1-Score των εξεταζόμενων μοντέλων στο ακαδημαϊκό σύνολο δεδομένων DBLP-Scholar.

### 5.3 Απόδοση στα Πεδία-Στόχους (Domain Adaptation)

Η πραγματική πρόκληση για το σύστημα LexiSync ήταν η μεταφορά της γνώσης από τα ακαδημαϊκά άρθρα στα προϊόντα ηλεκτρονικού εμπορίου. Τα αποτελέσματα καταδεικνύουν τη δυσκολία διαχείρισης θορυβωδών δεδομένων με σημαντική συντακτική ετερογένεια (Mudgal et al., 2018).

#### 5.3.1 Αποτελέσματα στο σύνολο Amazon-Google

Το συγκεκριμένο σύνολο αποδείχθηκε το πιο απαιτητικό ολόκληρης της πειραματικής διαδικασίας.

Πίνακας 5.3-1 Συγκριτικά αποτελέσματα στο σύνολο Amazon-Google

| Μοντέλο                                | Precision | Recall | F1-Score | Βέλτιστο Κατώφλι |
|----------------------------------------|-----------|--------|----------|------------------|
| microsoft/deberta-v3-base              | 0.7926    | 0.7350 | 0.7627   | 0.45             |
| intfloat/e5-small-v2                   | 0.7261    | 0.7479 | 0.7368   | 0.70             |
| BAAI/bge-small-en-v1.5                 | 0.7382    | 0.7350 | 0.7366   | 0.20             |
| thenlper/gte-small                     | 0.7154    | 0.7521 | 0.7333   | 0.35             |
| sentence-transformers/all-MiniLM-L6-v2 | 0.6773    | 0.7265 | 0.7010   | 0.60             |
| Snowflake/arctic-embed-xs              | 0.6413    | 0.7564 | 0.6941   | 0.35             |

Η μετάβαση στο εμπορικό σύνολο δεδομένων Amazon-Google φέρνει στο φως μια αισθητή πτώση των επιδόσεων για όλα τα μοντέλα σε σχέση με το ακαδημαϊκό πεδίο (DBLP). Ο λόγος για αυτή την πτώση είναι η ίδια η φύση των δεδομένων του ηλεκτρονικού εμπορίου: σε αντίθεση με τις επιστημονικές δημοσιεύσεις που είναι καθαρά γραμμένες, οι τίτλοι των

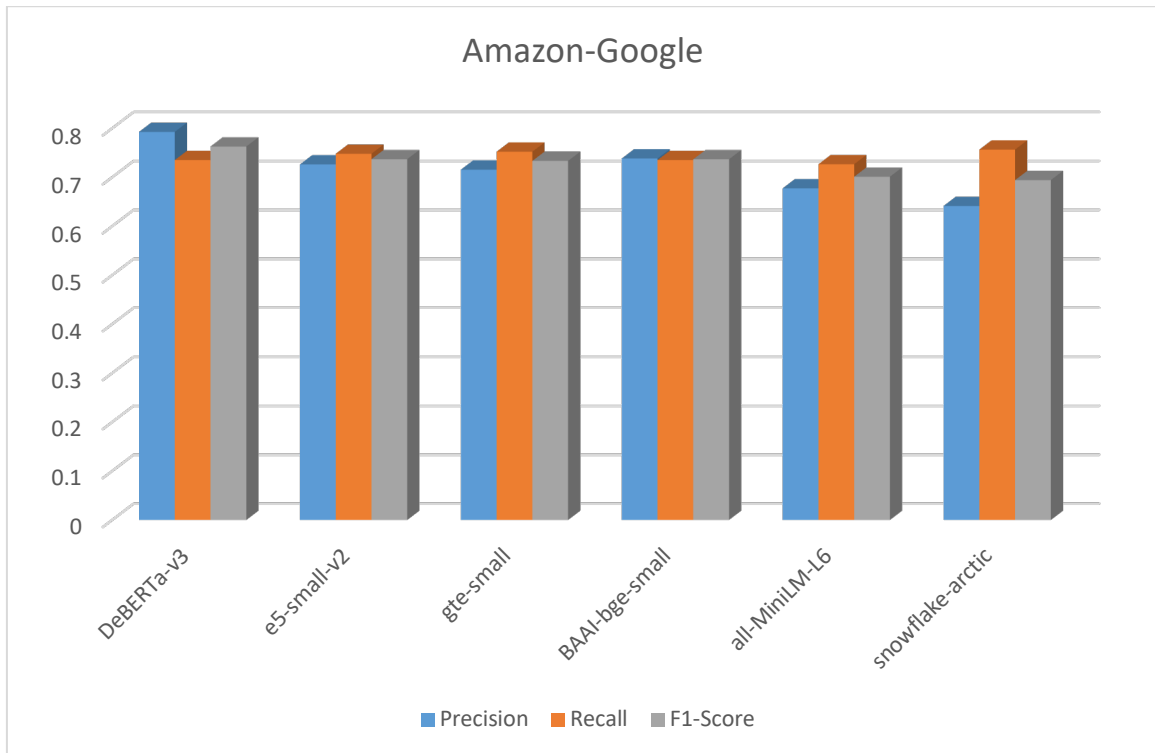
προϊόντων (κυρίως στο Amazon) είναι συχνά τεράστιοι, περιέχουν διαφημιστικές λέξεις και δεν ακολουθούν ενιαίους κανόνες.

Μέσα σε αυτό το δύσκολο περιβάλλον, το DeBERTa-v3 διατηρεί την απόλυτη κυριαρχία του, επιτυγχάνοντας F1-Score 0.7627 και υψηλό Precision 0.7926. Η επιτυχία του οφείλεται στο μέγεθος και το "βάθος" του νευρωνικού του δικτύου. Έχοντας μεγάλη χωρητικότητα, καταφέρνει να διαβάξει έναν μακροσκελή τίτλο και να εστιάζει ακριβώς εκεί που πρέπει (π.χ. βρίσκοντας τον πραγματικό κατασκευαστή), αγνοώντας τον υπόλοιπο θόρυβο.

Αντίθετα, παρατηρούμε ότι τα εξαιρετικά "συμπιεσμένα" και ελαφριά μοντέλα, όπως το MiniLM-L6 και το Snowflake, δυσκολεύτηκαν σημαντικά σε αυτό το έργο. Επειδή τα μοντέλα αυτά έχουν σχεδιαστεί να είναι μικρά και γρήγορα, θυσιάζουν ένα μέρος από την ικανότητά τους να αναλύουν πολύπλοκες προτάσεις. Έτσι, απέτυχαν συχνά να διαχωρίσουν σωστά τις πληροφορίες της τιμής ή της μάρκας όταν αυτές ήταν "θαμμένες" μέσα σε μεγάλα κείμενα. Αυτό είχε ως αποτέλεσμα να μπερδεύονται και να βγάζουν χαμηλά Precision (0.6773 και 0.6413 αντίστοιχα). Πρακτικά, αυτό σημαίνει ότι τα μικρά μοντέλα παρήγαγαν πολλά False Positives: έβλεπαν δύο διαφορετικά προϊόντα και θεωρούσαν λανθασμένα ότι ταυτίζονται, απλά και μόνο επειδή είχαν μερικές κοινές λέξεις στον τίτλο.

Τα αποτελέσματα αυτά είναι απολύτως λογικά και επιβεβαιώνουν τη βιβλιογραφία, η οποία υποστηρίζει ότι τα εξαιρετικά ελαφριά μοντέλα υστερούν σε περιβάλλοντα με έντονη λεξιλογική επικάλυψη (Wang et al., 2020). Ο όρος αυτός σημαίνει ότι όταν διαφορετικά προϊόντα μοιράζονται τις ίδιες ακριβώς λέξεις-κλειδιά (για παράδειγμα, μια θήκη υπολογιστή που έχει στον τίτλο της τη λέξη "Apple" επειδή είναι συμβατή με Mac), τα

μικρότερα μοντέλα δεν έχουν τη δυνατότητα να καταλάβουν τη διαφορά, πέφτοντας στην παγίδα της επικάλυψης των λέξεων.



Εικόνα 5.3-1 Σύγκριση Precision, Recall και F1-Score των εξεταζόμενων μοντέλων στο σύνολο δεδομένων Amazon-Google

### 5.3.2 Αποτελέσματα στο σύνολο Walmart-Amazon

Το σύνολο Walmart-Amazon φημίζεται για τα ελλιπή δομημένα πεδία του (missing features), όπου κωδικοί προϊόντων συχνά λείπουν εντελώς ή ανακατεύονται μέσα στον τίτλο. Παρά τη δυσκολία, οι αρχιτεκτονικές Cross-Encoder του LexiSync επέδειξαν ισχυρή ανθεκτικότητα.

Πίνακας 5.3-2 Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon

| Μοντέλο                   | Precision | Recall | F1-Score | Βέλτιστο Κατώφλι |
|---------------------------|-----------|--------|----------|------------------|
| microsoft/deberta-v3-base | 0.8513    | 0.8601 | 0.8557   | 0.25             |

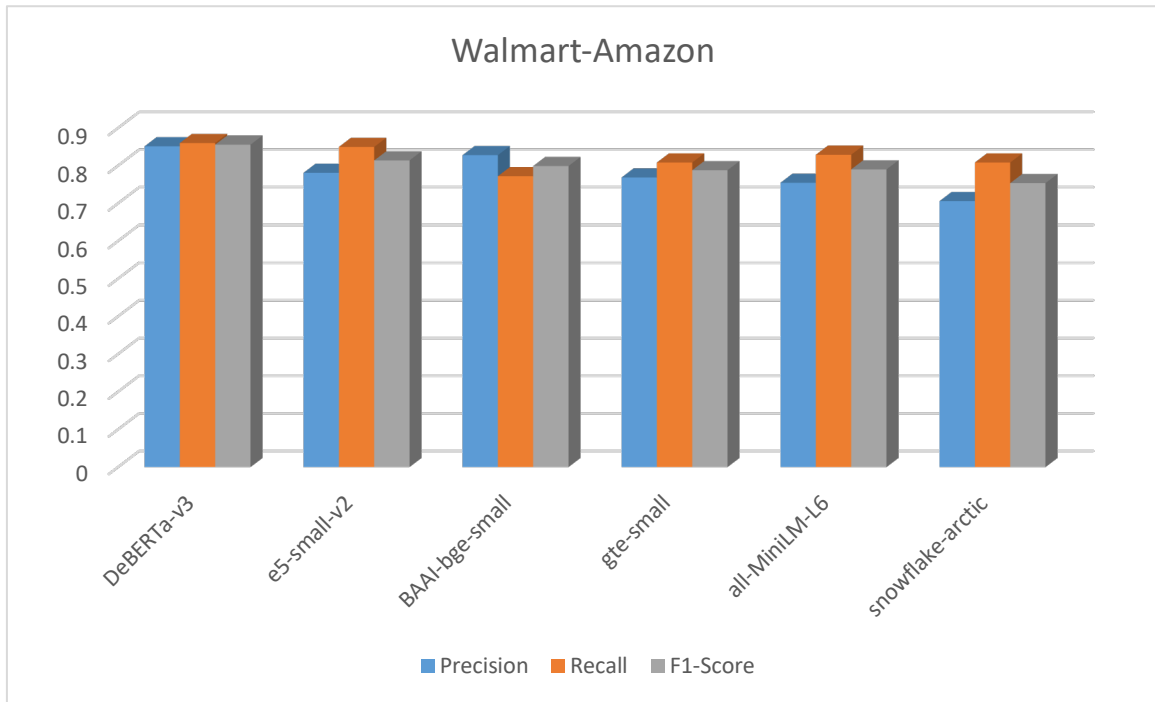
|                                        |        |        |        |      |
|----------------------------------------|--------|--------|--------|------|
| intfloat/e5-small-v2                   | 0.7810 | 0.8497 | 0.8139 | 0.75 |
| BAAI/bge-small-en-v1.5                 | 0.8278 | 0.7720 | 0.7989 | 0.70 |
| sentence-transformers/all-MiniLM-L6-v2 | 0.7547 | 0.8290 | 0.7901 | 0.50 |
| thenlper/gte-small                     | 0.7685 | 0.8083 | 0.7879 | 0.20 |
| Snowflake/arctic-embed-xs              | 0.7059 | 0.8083 | 0.7536 | 0.70 |

Περνώντας στο σύνολο δεδομένων Walmart-Amazon, το οποίο θεωρείται ιδιαίτερα απαιτητικό λόγω της συχνής απουσίας δεδομένων (missing values) και των ανακατεμένων πληροφοριών, παρατηρούμε ότι οι επιδόσεις των κορυφαίων μοντέλων ξεπέρασαν το 0.80 στο F1-Score. Το γεγονός αυτό αποδεικνύει την ικανότητα των μοντέλων να βρίσκουν λύσεις ακόμα και όταν η δομή των δεδομένων έχει "κενά".

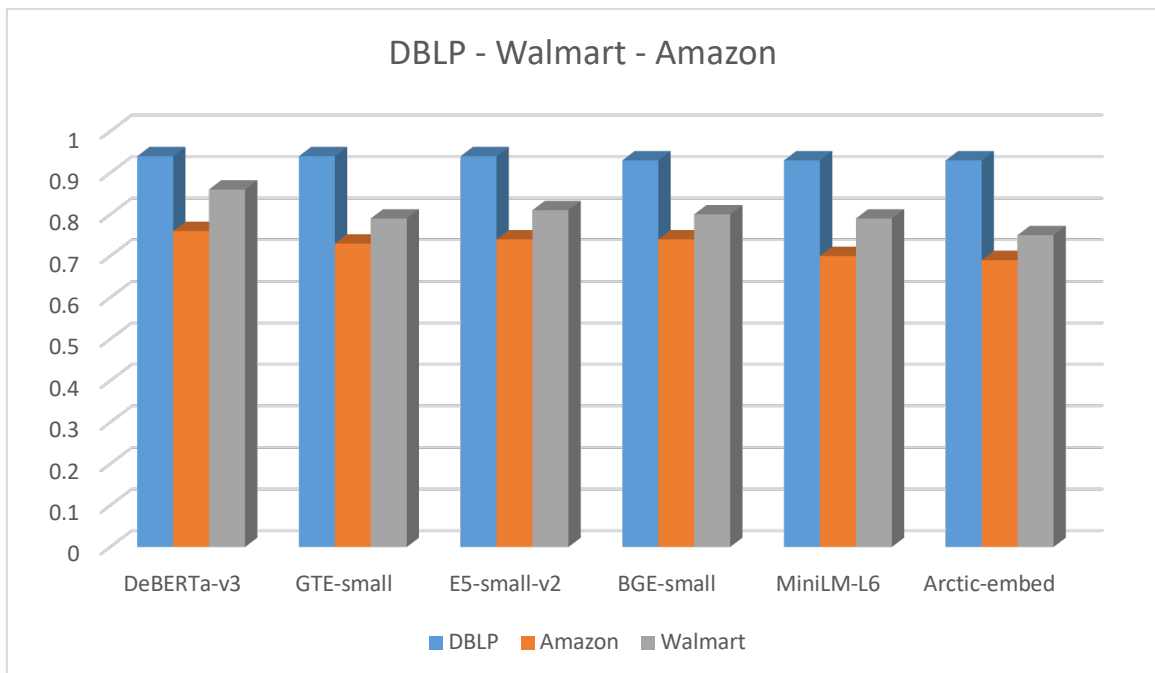
Συγκεκριμένα, το DeBERTa-v3 σημείωσε την κορυφαία επίδοση με F1-Score 0.8557 και εξαιρετικό Recall 0.8601. Με απλά λόγια, το μοντέλο κατάφερε να εντοπίσει σχεδόν όλα τα σωστά ζευγάρια (matches) παρά τον θόρυβο. Εξίσου σημαντική, όμως, ήταν η απόδοση του πολύ ελαφρύτερου μοντέλου e5-small-v2. Παρότι διαθέτει πολύ λιγότερες παραμέτρους, κατέγραψε ισχυρό Recall (0.8497) και συνολικό F1-Score 0.8139. Αυτό δείχνει ότι σε ορισμένες περιπτώσεις, η σωστή προετοιμασία των δεδομένων μετράει εξίσου με το μέγεθος του νευρωνικού δικτύου.

Πού οφείλεται, λοιπόν, αυτή η επιτυχία των μοντέλων σε τέτοια ακατάστατα δεδομένα; Η απάντηση βρίσκεται στη χρήση των ειδικών δεικτών (tags), όπως το [T] για τον Τίτλο, το [B] για τη Μάρκα και το [C] για την Κατηγορία. Αντί να δώσουμε στο μοντέλο ένα απλό, συνεχόμενο κείμενο, αυτοί οι δείκτες λειτουργούν σαν "ταμπέλες" καθοδήγησης. Μέσω του μηχανισμού προσοχής (self-attention), το νευρωνικό δίκτυο διαβάζει αυτές τις ταμπέλες και μαθαίνει μόνο του πώς να συνδυάζει τις πληροφορίες για παράδειγμα, καταλαβαίνει πότε μια λέξη στον τίτλο αντιστοιχεί στη μάρκα του προϊόντος. Όπως επιβεβαιώνεται και από τη βιβλιογραφία (Li et al., 2021), αυτή η διαδικασία βοηθά το μοντέλο να εξάγει την ουσία

των δεδομένων αυτόματα, καταργώντας πρακτικά την ανάγκη να γράφουμε εμείς πολύπλοκους, χειροκίνητους κανόνες αναζήτησης χαρακτηριστικών (feature engineering).



Εικόνα 5.3-2 Σύγκριση Precision, Recall και F1-Score των εξεταζόμενων μοντέλων στο σύνολο δεδομένων Walmart-Amazon



Εικόνα 5.3-3 Συγκριτικά αποτελέσματα (F1-Score) κατά τη μεταφορά γνώσης στα τρία εξεταζόμενα σύνολα δεδομένων (DBLP, Amazon, Walmart).

## 5.4 Σύγκριση Αποδοτικότητας (Efficiency vs. Accuracy)

Η επιλογή του καλύτερου γλωσσικού μοντέλου για ένα πραγματικό σύστημα (σε περιβάλλον παραγωγής) δεν βασίζεται μόνο στο ποιο πετυχαίνει το μεγαλύτερο σκορ. Αποτελεί πάντα έναν αυστηρό συμβιβασμό (trade-off) ανάμεσα στην ακρίβεια των αποτελεσμάτων και στο υπολογιστικό κόστος, δηλαδή τον χρόνο και τη μνήμη που απαιτούνται.

Στη δική μας ανάλυση, το DeBERTa-v3-base αναδείχθηκε ο αδιαμφισβήτητος νικητής σε επίπεδο ακρίβειας, κατακτώντας την πρώτη θέση και στα τρία σύνολα δεδομένων (με F1-Scores 0.9440, 0.7627 και 0.8557). Ωστόσο, πρόκειται για ένα τεράστιο μοντέλο που διαθέτει περίπου 86 εκατομμύρια παραμέτρους. Πρακτικά, αυτό σημαίνει ότι για να λειτουργήσει χρειάζεται πανίσχυρες και ακριβές κάρτες γραφικών (απαιτεί υψηλή μνήμη VRAM), ενώ χρειάζεται και περισσότερο χρόνο για να αποφασίσει αν ένα ζευγάρι ταιριάζει ή όχι (αυξημένος χρόνος εξαγωγής συμπερασμάτων - inference time).

Από την άλλη πλευρά, αξιολογήσαμε πολύ πιο ελαφριά μοντέλα, όπως το e5-small-v2 και το gte-small, τα οποία προσφέρουν μια εκπληκτική εναλλακτική. Το μέγεθός τους είναι λιγότερο από το μισό σε σχέση με το DeBERTa (έχουν μόλις ~33 εκατομμύρια παραμέτρους), αλλά η απόδοσή τους ήταν εντυπωσιακή: παρά το μικρό τους μέγεθος, στις περισσότερες περιπτώσεις έχασαν περίπου 1% έως 4% στο F1-Score.

Αυτό το εύρημα είναι κρίσιμο για την εφαρμογή του συστήματος στην πράξη. Σε ένα ρεαλιστικό σενάριο όπου μια πλατφόρμα πρέπει να συγκρίνει και να ταξινομήσει εκατομμύρια προϊόντα καθημερινά, η ελάχιστη απώλεια του 1-3% στην ακρίβεια είναι ένας εξαιρετικός συμβιβασμός. Κερδίζουμε ένα σύστημα που εκπαιδεύεται πολύ πιο γρήγορα, καταναλώνει λιγότερους πόρους και μπορεί να επεξεργαστεί τεράστιο όγκο δεδομένων (scale) χωρίς να καταρρέουν οι διακομιστές. Όπως επισημαίνεται και στη βιβλιογραφία (Reimers & Gurevych, 2019), σε τέτοιες συνθήκες τα μικρά μοντέλα αναδεικνύονται συχνά ως η πιο ρεαλιστική και βέλτιστη λύση.

## 5.5 Η Κρίσιμη Συμβολή του Threshold Tuning

Ένα από τα πιο σημαντικά πρακτικά ευρήματα όλης της πειραματικής διαδικασίας είναι η ακραία διακύμανση που παρουσιάζει το ιδανικό σημείο απόφασης το λεγόμενο Βέλτιστο Κατώφλι (Optimal Threshold) ανάμεσα στα διάφορα μοντέλα και σύνολα δεδομένων.

Στην κλασική μηχανική μάθηση, τα περισσότερα συστήματα παίρνουν αποφάσεις με έναν απλό, στατικό κανόνα: αν η πιθανότητα (σιγουριά) του μοντέλου είναι πάνω από 50% (δηλαδή 0.5), το ζευγάρι θεωρείται Match, αλλιώς θεωρείται Non-Match. Η έρευνά μας απέδειξε ότι η τυφλή χρήση αυτού του στατικού κανόνα θα οδηγούσε σε καταστροφικά αποτελέσματα για την αξιολόγηση. Γι' αυτόν τον λόγο, ο αλγόριθμος του LexiSync δεν βασίζεται στο 0.5, αλλά σαρώνει δυναμικά όλες τις πιθανές τιμές για να βρει το ακριβές σημείο που μεγιστοποιεί το F1-Score.

Το πόσο κρίσιμη είναι αυτή η δυναμική συμπεριφορά φαίνεται ξεκάθαρα στο απαιτητικό σύνολο του Walmart. Για παράδειγμα, το ελαφρύ μοντέλο Snowflake/arctic-embed-xs χρειάστηκε ένα πολύ αυστηρό, υψηλό κατώφλι της τάξης του 0.70 για να βγάλει το καλύτερο σκορ του. Επειδή το μοντέλο αυτό μπερδευόταν εύκολα από τον θόρυβο των δεδομένων, έπρεπε να του "ανεβάσουμε τον πήχη" ψηλά, ώστε να εγκρίνει μόνο τα ζευγάρια για τα οποία ήταν απολύτως σίγουρο.

Το ίδιο φαινόμενο, ιδωμένο από την πλευρά του dataset, αναδεικνύει ότι το βέλτιστο κατώφλι εξαρτάται κυρίως από την κατανομή των πιθανοτήτων που επάγει το εκάστοτε σύνολο δεδομένων, και όχι αποκλειστικά από την αρχιτεκτονική. Χαρακτηριστικό είναι ότι το DeBERTa-v3 απαιτεί υψηλό κατώφλι (0.80) στο καθαρό, δομημένο DBLP-Scholar, αλλά πολύ χαμηλό κατώφλι (0.25) στο θορυβώδες Walmart-Amazon. Στο μεν DBLP, όπου τα ζεύγη διαχωρίζονται ευκρινώς, το μοντέλο παράγει πολωμένες πιθανότητες (κοντά στο 0 ή στο 1) και ένα υψηλό κατώφλι αποφεύγει τα False Positives χωρίς κόστος στην Ανάκληση. Στο δε Walmart, η συντακτική ετερογένεια και τα ελλιπή πεδία οδηγούν σε «διστακτικές» προβλέψεις (π.χ. 0.30-0.40) ακόμη και για πραγματικά matches· εδώ, ένα στατικό κατώφλι 0.5 θα απέρριπτε αυτά τα ζεύγη, εκμηδενίζοντας την Ανάκληση. Η σταθμισμένη συνάρτηση απώλειας ενισχύει αυτή τη συμπεριφορά στα δύσκολα σύνολα, καθώς ωθεί το μοντέλο να μην απορρίπτει εύκολα πιθανές ταυτίσεις, καθιστώντας τη δυναμική προσαρμογή κατωφλίου ανά dataset απαραίτητη.

Αν είχαμε κρατήσει τον στατικό κανόνα του 0.5, το σύστημα θα είχε απορρίψει αυτόματα αυτά τα σωστά ζευγάρια, ρίχνοντας κατακόρυφα την Ανάκλησή του (Recall). Ο δυναμικός αλγόριθμος του LexiSync "κατάλαβε" αυτή την ευαισθησία, χαμήλωσε το κατώφλι στο 0.25 και "έσωσε" αυτά τα δύσκολα Matches. Αυτό αποδεικνύει περίτρανα ότι κάθε

αρχιτεκτονική αντιλαμβάνεται τις πιθανότητες διαφορετικά και χρειάζεται τη δική της, εξατομικευμένη ρύθμιση.

## **5.6 Η Επιτυχία του Replay Buffer έναντι της Καταστροφικής Λήθης**

Ένα τελευταίο, αλλά εξαιρετικά κρίσιμο κομμάτι της αξιολόγησης αφορά τη συμπεριφορά του μηχανισμού Replay Buffer. Ο στόχος αυτού του μηχανισμού (όπως περιγράφηκε στην Ενότητα 3.5.2) ήταν να βοηθήσει το μοντέλο να διατηρήσει τις γνώσεις που απέκτησε στο πρώτο στάδιο (στα ακαδημαϊκά δεδομένα του DBLP), ακόμα και όταν στη συνέχεια υποβλήθηκε σε εντατική εκπαίδευση για 7 εποχές πάνω στα εμπορικά προϊόντα των Amazon και Walmart.

Για να γίνει κατανοητή η αξία αυτού του μηχανισμού, πρέπει να αναφέρουμε ένα βασικό πρόβλημα των νευρωνικών δικτύων: όταν ένα μοντέλο μαθαίνει κάτι εντελώς νέο (όπως το λεξιλόγιο των ηλεκτρονικών συσκευών στο Amazon), αναπροσαρμόζει ριζικά τα εσωτερικά του βάρη. Στην προσπάθειά του να γίνει καλό στη νέα εργασία, ουσιαστικά "επικαλύπτει" και διαγράφει τη γνώση της προηγούμενης εργασίας. Το φαινόμενο αυτό ονομάζεται Καταστροφική Λήθη (Catastrophic Forgetting). Αν δεν είχαμε πάρει μέτρα, η εκπαίδευση στα προϊόντα θα προκαλούσε την απόλυτη κατάρρευση του F1-Score στο αρχικό πεδίο (Kirkpatrick et al., 2017).

Για να αποτρέψουμε αυτή τη "λήθη", εφαρμόσαμε τον αλγόριθμο Experience Replay (Επανάληψη Εμπειρίας). Πρακτικά, αυτό που κάναμε ήταν να κρατήσουμε 1.200 επιλεγμένα δείγματα από το DBLP και να τα ανακατέψουμε μαζί με τα νέα δεδομένα του Amazon και του Walmart. Αυτά τα 1.200 δείγματα λειτούργησαν ως "φρουροί" της παλιάς γνώσης. Καθώς το μοντέλο μάθαινε για τηλεοράσεις και υπολογιστές, έβλεπε σποραδικά και τα παλιά ακαδημαϊκά άρθρα, με αποτέλεσμα το δίκτυο να αναγκάζεται να θυμάται πώς να τα ταυτίζει.

Τα αποτελέσματα του Πίνακα 5.2-1 είναι συμβατά με την υπόθεση ότι ο μηχανισμός Replay Buffer συνέβαλε στη διατήρηση της γνώσης του πηγαιίου πεδίου. Παρά την εντατική εκπαίδευση επί 7 εποχών στα εμπορικά σύνολα της Φάσης 2, τα μοντέλα διατήρησαν F1-Score άνω του 0.93 στο DBLP-Scholar.

## 5.7 Ποιοτική Ανάλυση Σφαλμάτων (Error Analysis)

Ακόμα και τα καλύτερα συστήματα κάνουν λάθη. Παρά τα πολύ υψηλά σκορ (F1-Scores) που επιτεύχθηκαν, η ποιοτική εξέταση των εσφαλμένων προβλέψεων του κορυφαίου μοντέλου (DeBERTa-v3) αναδεικνύει τις πραγματικές δυσκολίες του Domain Adaptation. Η ανάλυση αυτών των σφαλμάτων μάς βοηθά να κατανοήσουμε τα όρια της τεχνητής νοημοσύνης όταν λειτουργεί σε συνθήκες πραγματικού κόσμου. Τα σφάλματα του συστήματος χωρίζονται σε δύο βασικές κατηγορίες, καθεμία από τις οποίες αποκαλύπτει διαφορετικές αδυναμίες του μοντέλου.

Αρχικά, παρατηρήθηκαν Σφάλματα Τύπου I (False Positives), δηλαδή περιπτώσεις όπου το μοντέλο ταύτισε λανθασμένα δύο διαφορετικά προϊόντα. Τέτοια λάθη εμφανίστηκαν κυρίως στο απαιτητικό σύνολο του ηλεκτρονικού εμπορίου (Walmart-Amazon), ειδικά σε προϊόντα που ανήκουν στην ίδια "οικογένεια" αλλά αποτελούν διαφορετικές εκδόσεις, όπως για παράδειγμα η σύγκριση ανάμεσα στο «Apple iPad 64GB» και στο «Apple iPad 128GB». Σε αυτές τις περιπτώσεις, επειδή τα κείμενα είναι κατά 95% ολόιδια (κοινή μάρκα, κατηγορία, μοντέλο), ο μηχανισμός προσοχής (Attention) του νευρωνικού δικτύου τείνει να υπερ-απλουστεύει την απόφασή του. Πρακτικά, το μοντέλο παρασύρεται από τον μεγάλο όγκο των κοινών λέξεων και αγνοεί τελικά την κρισιμότητα που έχει ένας και μόνο αριθμός (η χωρητικότητα), θεωρώντας λανθασμένα ότι πρόκειται για την ίδια εγγραφή.

Από την άλλη πλευρά, εντοπίστηκαν Σφάλματα Τύπου II (False Negatives), όπου δύο εγγραφές ήταν πράγματι το ίδιο προϊόν, αλλά το σύστημα τις απέρριψε κατατάσσοντάς τις ως διαφορετικές (Non-Match). Αυτές οι αστοχίες εμφανίστηκαν κυρίως σε ακραίες περιπτώσεις απουσίας δεδομένων (missing values). Για παράδειγμα, υπήρξαν εγγραφές όπου έλειπε εντελώς το πεδίο της μάρκας [B] ή ο κωδικός του μοντέλου [N], αφήνοντας το σύστημα με μια εξαιρετικά λακωνική περιγραφή. Ακόμα και με τη βοήθεια της σταθμισμένης απώλειας (Weighted Loss) που ενθάρρυνε το μοντέλο να βρίσκει ταυτίσεις, το δίκτυο δεν διέθετε αρκετά στοιχεία για να βγάλει ένα ασφαλές συμπέρασμα. Επειδή η "σημασιολογική του πεποίθηση" (confidence) παρέμενε χαμηλή, αδυνατούσε να ξεπεράσει το υψηλό δυναμικό κατώφλι απόφασης. Αυτό αποδεικνύει ότι, ανεξάρτητα από τη δύναμη της αρχιτεκτονικής, όταν τα δεδομένα εισόδου είναι υπερβολικά φτωχά το μοντέλο είναι αδύνατον να ανακτήσει με σιγουριά την ταύτιση.

## 5.8 Αξιολόγηση Αρχιτεκτονικής Bi-Encoder: Μηδενική Μάθηση (Zero-Shot) έναντι Μικρο-ρύθμισης (Fine-Tuning)

Για την πληρέστερη τεκμηρίωση της επιλογής της αρχιτεκτονικής Cross-Encoder στο τελικό στάδιο επαναταξινόμησης (Re-ranking) του συστήματος LexiSync, σχεδιάστηκε και εκτελέστηκε ένα συμπληρωματικό πείραμα αξιολόγησης βάσης (Baseline Evaluation). Στο πείραμα αυτό αξιολογήθηκε η συμπεριφορά ενός Bi-Encoder (all-MiniLM-L6-v2) στο πλέον απαιτητικό σύνολο δεδομένων της έρευνας (Walmart-Amazon).

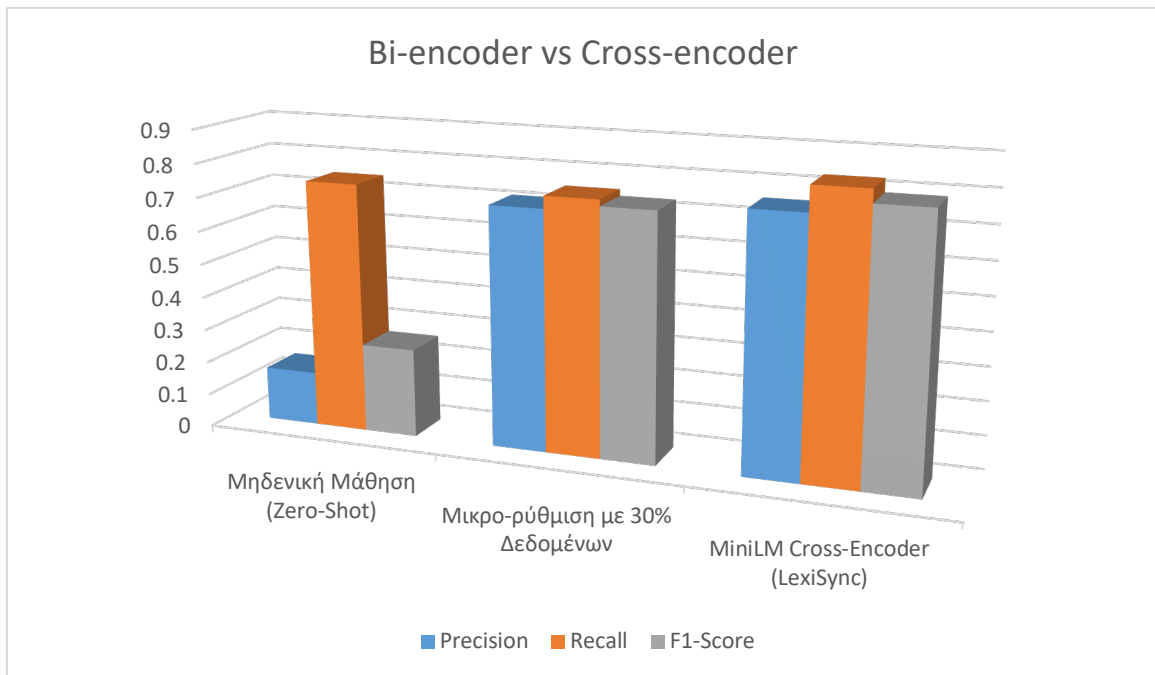
Η πειραματική διαδικασία σχεδιάστηκε με γνώμονα την απόλυτη αποφυγή διαρροής δεδομένων (Data Leakage). Η βελτιστοποίηση του κατωφλίου απόφασης (Threshold Tuning) πραγματοποιήθηκε αποκλειστικά στο σύνολο επικύρωσης (Validation Set), ενώ τα τελικά αποτελέσματα εξήχθησαν στο εντελώς «άγνωστο» Test Set.

Το πείραμα αξιολόγησε το μοντέλο σε δύο στάδια. Μηδενική Μάθηση (Zero-Shot Evaluation): Αξιολόγηση του μοντέλου απευθείας ("off-the-shelf"), βασισμένη αποκλειστικά στην προ-εκπαιδευμένη γνώση του, χωρίς να έχει δει κανένα δείγμα από τον τομέα του ηλεκτρονικού εμπορίου. Μικρο-ρύθμιση Περιορισμένων Δεδομένων (Few-Shot Fine-Tuning): Εκπαίδευση του μοντέλου (Siamese Network) χρησιμοποιώντας τη συνάρτηση Αντιθετικής Απώλειας (Contrastive Loss). Στο μοντέλο δόθηκε πρόσβαση σε μόλις ένα υποσύνολο (30%) του συνόλου εκπαίδευσης, προκειμένου να ελεγχθεί η αποδοτικότητα δεδομένων (Data Efficiency).

Πίνακας 5.8-1 Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon (Αρχιτεκτονική Bi-Encoder έναντι Cross-Encoder)

| Μέθοδος Αξιολόγησης(Bi-Encoder) | Precision | Recall | F1-Score |
|---------------------------------|-----------|--------|----------|
| Μηδενική Μάθηση (Zero-Shot)     | 0.1607    | 0.7461 | 0.2645   |
| Μικρο-ρύθμιση με 30% Δεδομένων  | 0.7143    | 0.7513 | 0.7323   |

|                                 |        |        |        |
|---------------------------------|--------|--------|--------|
| MiniLM Cross-Encoder (LexiSync) | 0.7547 | 0.8290 | 0.7901 |
|---------------------------------|--------|--------|--------|



Εικόνα 5.8-1 Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon (Αρχιτεκτονική Bi-Encoder έναντι Cross-Encoder)

### 5.8.1 Συγκριτικά αποτελέσματα στο σύνολο Walmart-Amazon (Αρχιτεκτονική Bi-Encoder έναντι Cross-Encoder)

Η ανάλυση των παραπάνω αποτελεσμάτων επιβεβαιώνει τη θεωρητική βιβλιογραφία (Reimers & Gurevych, 2019) και δικαιώνει τις αρχιτεκτονικές επιλογές του συστήματος LexiSync: Η Αποτυχία της Μηδενικής Μάθησης: Η επίδοση του μοντέλου στο Zero-Shot σενάριο ήταν μη αποδεκτή (F1-Score 0.2645). Αν και διατήρησε ικανοποιητική Ανάκληση (0.7461), η Ακρίβειά του (Precision 0.1607) κατέρρευσε, καθώς παρήγαγε 752 False Positives. Αυτό υποδεικνύει ότι τα γενικά γλωσσικά μοντέλα αδυνατούν να διαχωρίσουν λεπτές σημασιολογικές διαφορές (π.χ. διαφορές σε κωδικούς μοντέλων ηλεκτρονικών ειδών) χωρίς εξειδίκευση στο εκάστοτε πεδίο. Υψηλή Αποδοτικότητα Δεδομένων (Data Efficiency): Με την εισαγωγή μόλις του 30% των δεδομένων εκπαίδευσης, το μοντέλο επέδειξε εξαιρετική ικανότητα προσαρμογής, βελτιώνοντας το F1-Score του κατά 176%

(φτάνοντας στο 0.7323) και μειώνοντας δραματικά τα False Positives σε 58. Η χρήση του Contrastive Loss απέδειξε ότι οι Bi-Encoders μπορούν να δημιουργήσουν ισχυρές διανυσματικές αναπαραστάσεις με ελάχιστα παραδείγματα. Το Όριο του Bi-Encoder έναντι του Cross-Encoder: Παρά την εντυπωσιακή βελτίωση, η μέγιστη απόδοση του Bi-Encoder (0.7323) υπολείπεται σημαντικά (κατά ~6%) έναντι της αρχιτεκτονικής Cross-Encoder του LexiSync (0.7901) που παρουσιάστηκε σε προηγούμενη ενότητα.

Συμπερασματικά, η ανάλυση αποδεικνύει ότι η ανεξάρτητη παραγωγή διανυσμάτων (Bi-Encoder) είναι αποδοτική, αλλά η απουσία βαθιάς διασταυρούμενης προσοχής (cross-attention) σε επίπεδο λέξης περιορίζει το ανώτατο όριο απόδοσής της. Για το τελικό και κρίσιμότερο στάδιο της ταξινόμησης (Re-ranking) σε εξαιρετικά θορυβώδη δεδομένα, η επιλογή της αρχιτεκτονικής Cross-Encoder κρίνεται απολύτως επιβεβλημένη.

## **5.9 Ανάλυση Ανταλλαγής Ακρίβειας και Ανάκλησης (Precision-Recall Trade-off)**

Η λεπτομερής εξέταση της σχέσης Ακρίβειας (Precision) και Ανάκλησης (Recall) αποκαλύπτει θεμελιώδεις δομικές διαφορές μεταξύ των εξεταζόμενων αρχιτεκτονικών, οι οποίες δεν είναι ορατές από τη μόνη μελέτη του F1-Score. Η ανάλυση που ακολουθεί βασίζεται στα αποτελέσματα των Πινάκων 5.9-1, 5.9-2 και 5.9-3 και εστιάζει στα τρία σύνολα δεδομένων ξεχωριστά, καθώς το κάθε ένα αναδεικνύει διαφορετική πτυχή του trade-off.

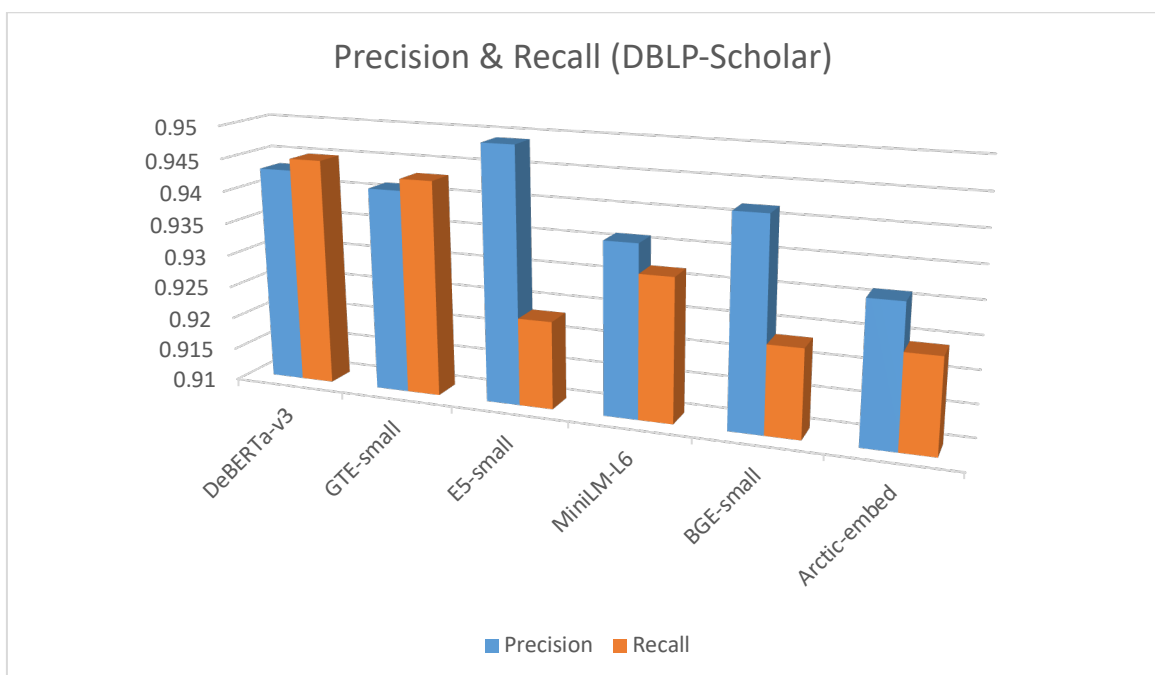
### **5.9.1 Απόδοση στο DBLP-Scholar**

Στο δομημένο ακαδημαϊκό σύνολο, όλα τα μοντέλα επιτυγχάνουν εξαιρετικά ισορροπημένη σχέση Precision/Recall, με τιμές που κυμαίνονται μεταξύ 0.92 και 0.95 και για τους δύο δείκτες. Αξιοσημείωτη εξαίρεση αποτελεί το μοντέλο E5-small, το οποίο καταγράφει την υψηλότερη Precision (0.9491) αλλά τη χαμηλότερη Recall (0.9234) μεταξύ των εξεταζόμενων αρχιτεκτονικών. Αυτή η συμπεριφορά υποδηλώνει ότι το μοντέλο είναι ιδιαίτερα συντηρητικό στις προβλέψεις του αποφεύγει τα False Positives εις βάρος ορισμένων True Positives. Αντίθετα, ο DeBERTa-v3 επιτυγχάνει την πιο ισορροπημένη σχέση (0.9431 / 0.9449), επιβεβαιώνοντας την ικανότητα του μηχανισμού Disentangled

Attention να αξιολογεί ομοιότητες τόσο σε επίπεδο περιεχομένου όσο και σε επίπεδο θέσης των λέξεων εντός της ακολουθίας (He et al., 2020).

**Πίνακας 5.9-1 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο δεδομένων DBLP-Scholar.**

| Μοντέλο      | Precision | Recall |
|--------------|-----------|--------|
| DeBERTa-v3   | 0.9431    | 0.9449 |
| GTE-small    | 0.9412    | 0.9430 |
| E5-small     | 0.9491    | 0.9234 |
| MiniLM-L6    | 0.9362    | 0.9318 |
| BGE-small    | 0.9418    | 0.9234 |
| Arctic-embed | 0.9313    | 0.9243 |



**Εικόνα 5.9-1 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο δεδομένων DBLP-Scholar.**

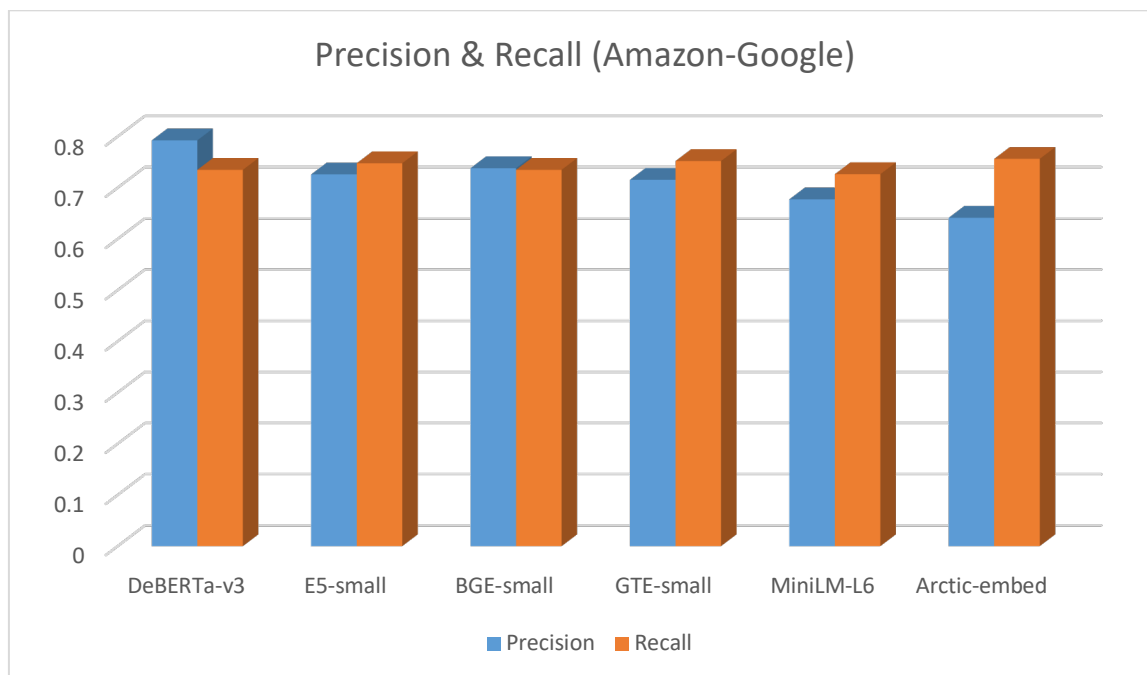
### 5.9.2 Απόδοση στο Amazon-Google

Η μετάβαση στο εμπορικό σύνολο αναδεικνύει μια σαφή πόλωση: τα μοντέλα διαχωρίζονται σε δύο ομάδες ανάλογα με τη στρατηγική τους. Ο DeBERTa-v3 και το BGE-small επιδεικνύουν υψηλή Precision (0.7926 και 0.7382 αντίστοιχα) με χαμηλότερη Recall,

υποδηλώνοντας ότι οι προβλέψεις τους είναι αξιόπιστες αλλά συντηρητικές. Αντίθετα, το Arctic-embed καταγράφει την υψηλότερη Recall (0.7564) με ταυτόχρονη κατάρρευση της Precision στο 0.6413 το μοντέλο "βλέπει" περισσότερα matches αλλά εισάγει σημαντικό αριθμό False Positives. Αυτή η συμπεριφορά είναι χαρακτηριστική μοντέλων με χαμηλή διαχωριστική ικανότητα σε θορυβώδη δεδομένα, όπου η λεξιλογική επικάλυψη μεταξύ διαφορετικών προϊόντων είναι υψηλή (Mudgal et al., 2018).

**Πίνακας 5.9-2 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο δεδομένων Amazon-Google.**

| Μοντέλο      | Precision | Recall |
|--------------|-----------|--------|
| DeBERTa-v3   | 0.7926    | 0.7350 |
| E5-small     | 0.7261    | 0.7479 |
| BGE-small    | 0.7382    | 0.7350 |
| GTE-small    | 0.7154    | 0.7521 |
| MiniLM-L6    | 0.6773    | 0.7265 |
| Arctic-embed | 0.6413    | 0.7564 |



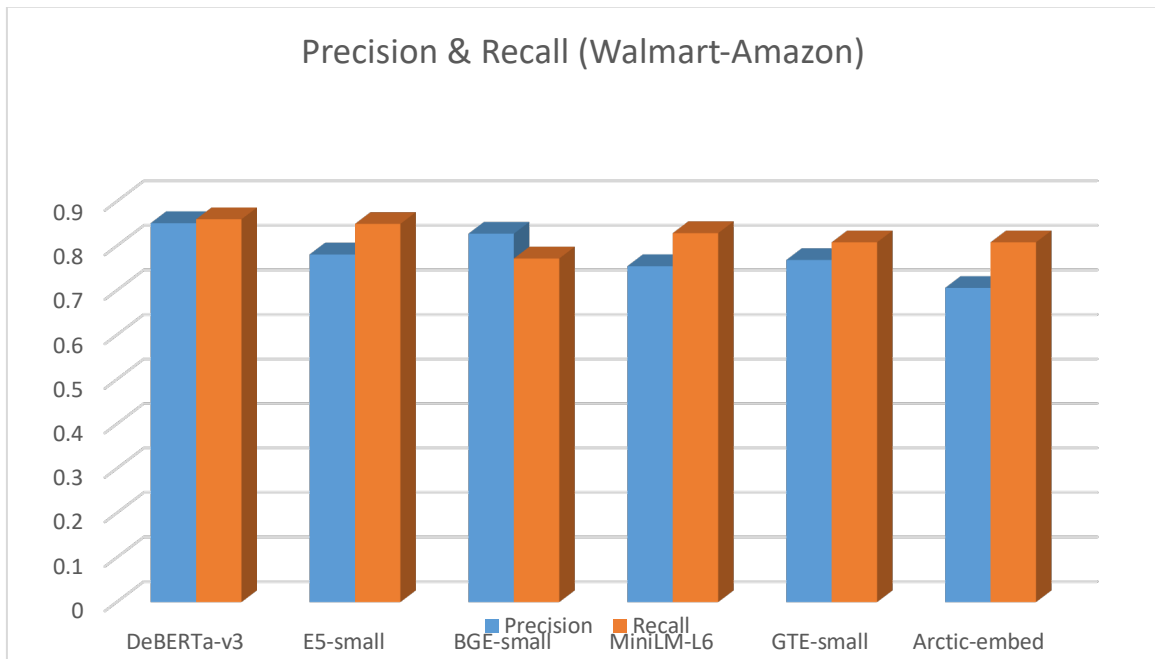
**Εικόνα 5.9-2 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο δεδομένων Amazon-Google.**

### 5.9.3 Απόδοση στο Walmart-Amazon

Το απαιτητικότερο σύνολο αναδεικνύει την πιο ενδιαφέρουσα συμπεριφορά. Το BGE-small παρουσιάζει αντεστραμμένο trade-off σε σχέση με τα υπόλοιπα μοντέλα: υψηλή Precision (0.8278) αλλά χαμηλή Recall (0.7720), ενώ το E5-small και το MiniLM-L6 επιτυγχάνουν υψηλότερη Recall (0.8497 και 0.8290 αντίστοιχα) με μικρότερη Precision. Κρίσιμο εύρημα αποτελεί το γεγονός ότι μόνο ο DeBERTa-v3 καταφέρνει να διατηρήσει υψηλές τιμές και στους δύο δείκτες ταυτόχρονα (0.8513 / 0.8601), γεγονός που αποδίδεται στην ικανότητα της αρχιτεκτονικής Cross-Encoder να εκτελεί διασταυρούμενη προσοχή (cross-attention) σε επίπεδο λέξης μεταξύ των δύο εγγραφών, εντοπίζοντας κρίσιμες αριθμητικές διαφορές όπως κωδικοί μοντέλων που είναι κρυμμένοι εντός του τίτλου (Li et al., 2021). Η παρατήρηση αυτή επιβεβαιώνει το θεωρητικό πλαίσιο των Reimers & Gurevych (2019), σύμφωνα με το οποίο οι Bi-Encoders παράγουν ανεξάρτητες διανυσματικές αναπαραστάσεις και συνεπώς αδυνατούν να μοντελοποιήσουν λεπτές σχέσεις μεταξύ των δύο εγγραφών κατά τη φάση της σύγκρισης.

**Πίνακας 5.9-3 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο δεδομένων Walmart-Amazon.**

| Μοντέλο      | Precision | Recall |
|--------------|-----------|--------|
| DeBERTa-v3   | 0.8513    | 0.8601 |
| E5-small     | 0.7810    | 0.8497 |
| BGE-small    | 0.8278    | 0.7720 |
| MiniLM-L6    | 0.7547    | 0.8290 |
| GTE-small    | 0.7685    | 0.8083 |
| Arctic-embed | 0.7059    | 0.8083 |



Εικόνα 5.9-3 Σύγκριση Precision & Recall των εξεταζόμενων μοντέλων στο σύνολο δεδομένων Walmart-Amazon.

## 5.10 Ανάλυση Υπολογιστικού Κόστους (Execution Time Analysis)

Η αξιολόγηση του υπολογιστικού κόστους πραγματοποιήθηκε σε περιβάλλον Apple Silicon (M-series, MPS backend) και μετρήθηκε ξεχωριστά για κάθε φάση του pipeline: Embedding Time, Blocking Time και Matching Time. Κάθε μέτρηση πραγματοποιήθηκε τρεις φορές και αναφέρεται ο μέσος χρόνος, σύμφωνα με καθιερωμένες πρακτικές αξιολόγησης συστημάτων NLP (Reimers & Gurevych, 2019).

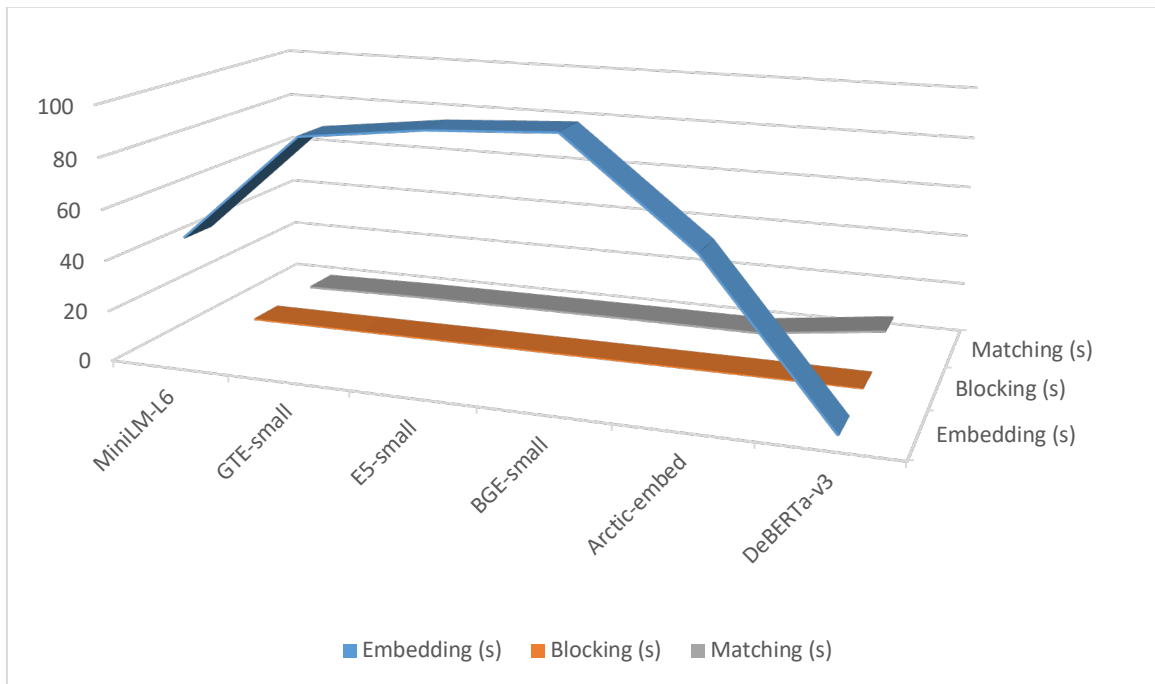
### 5.10.1 Κατανομή κόστους στους Bi-Encoders

Για όλα τα εξεταζόμενα μοντέλα Bi-Encoder, η φάση του Embedding αποτελεί κατά πολύ το κυρίαρχο υπολογιστικό βάρος, αντιπροσωπεύοντας πάνω από το 99.5% του συνολικού χρόνου εκτέλεσης σε κάθε σύνολο δεδομένων. Ενδεικτικά, στο Walmart-Amazon ο MiniLM-L6 δαπανά 21.411 δευτερόλεπτα για την παραγωγή διανυσμάτων έναντι μόλις 0.081 δευτερολέπτων για τη φάση του Blocking μέσω FAISS δηλαδή ο Blocker είναι 264 φορές ταχύτερος από την παραγωγή των embeddings. Αυτό επιβεβαιώνει εμπειρικά την αποδοτικότητα της βιβλιοθήκης FAISS για προσεγγιστική αναζήτηση πλησιέστερων

γειτόνων σε χώρους υψηλών διαστάσεων (Johnson et al., 2019). Το κόστος του Blocking παραμένει ουσιαστικά σταθερό μεταξύ μοντέλων, καθώς εξαρτάται αποκλειστικά από τον αριθμό των εγγραφών και τη διάσταση του διανύσματος (384 για όλα τα εξεταζόμενα μοντέλα), και όχι από την αρχιτεκτονική του μοντέλου. Το γεγονός αυτό αναδεικνύει μια σημαντική ιδιότητα των Bi-Encoder συστημάτων, τα embeddings υπολογίζονται μία φορά offline και επαναχρησιμοποιούνται για κάθε νέο ερώτημα, καθιστώντας το συνολικό κόστος αναλογικό ως  $O(N)$  ως προς τον αριθμό των εγγραφών και όχι ως  $O(N^2)$  όπως στην εξαντλητική σύγκριση (Christen, 2012).

**Πίνακας 5.10-1 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking, Matching), για το σύνολο DBLP-Scholar.**

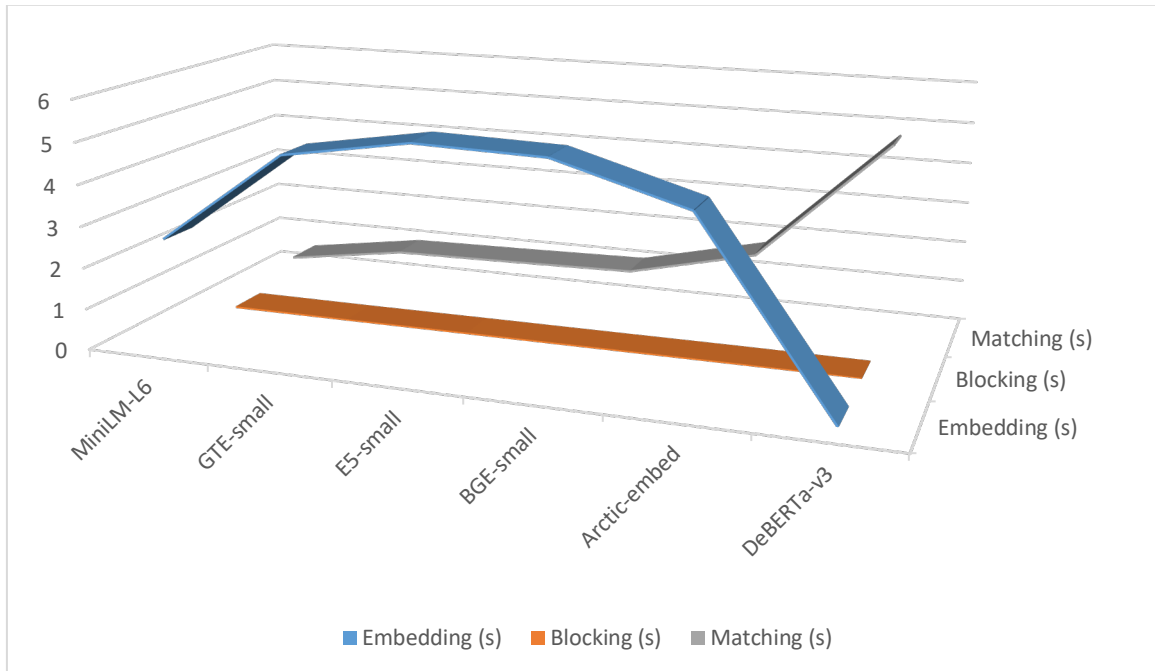
| Μοντέλο      | Embedding (s) | Blocking (s) | Matching (s) |
|--------------|---------------|--------------|--------------|
| MiniLM-L6    | 47.562        | 0.207        | 0.605        |
| GTE-small    | 90.171        | 0.184        | 1.306        |
| E5-small     | 95.829        | 0.210        | 1.381        |
| BGE-small    | 98.311        | 0.204        | 1.368        |
| Arctic-embed | 59.715        | 0.171        | 1.173        |
| DeBERTa-v3   | 0             | 0            | 7.342        |



Εικόνα 5.10-1 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking, Matching), για το σύνολο DBLP-Scholar.

Πίνακας 5.10-2 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking, Matching), για το σύνολο Google-Amazon.

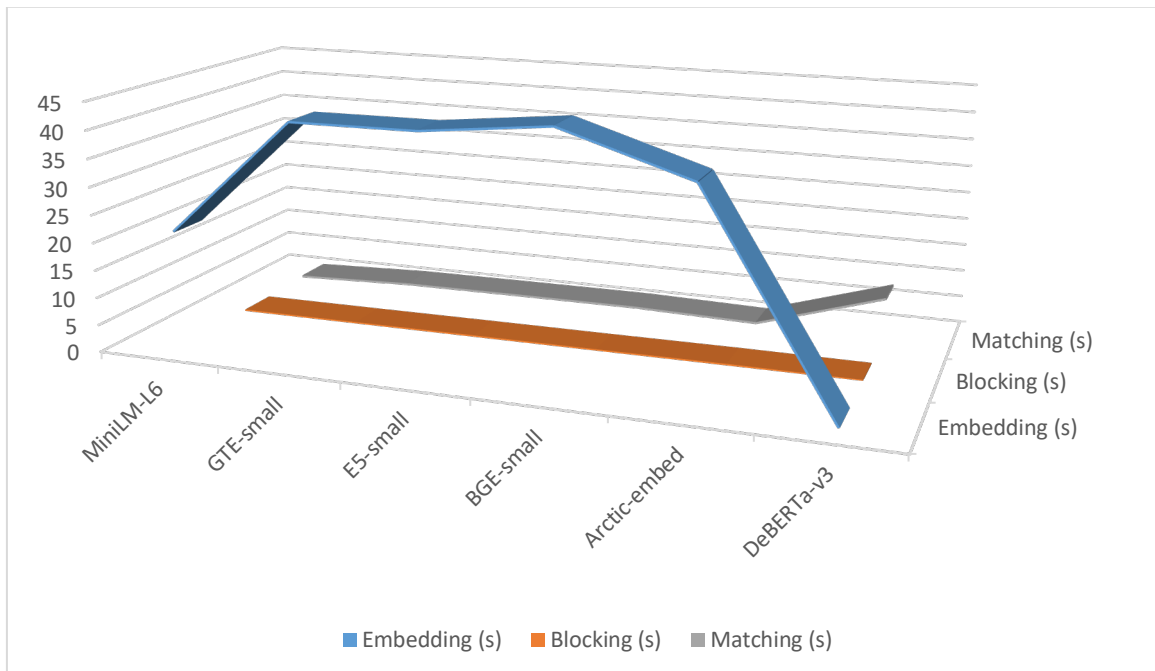
| Μοντέλο      | Embedding (s) | Blocking (s) | Matching (s) |
|--------------|---------------|--------------|--------------|
| MiniLM-L6    | 2.613         | 0.022        | 0.525        |
| GTE-small    | 4.854         | 0.013        | 0.969        |
| E5-small     | 5.349         | 0.016        | 1.007        |
| BGE-small    | 5.256         | 0.013        | 1.059        |
| Arctic-embed | 4.359         | 0.012        | 1.775        |
| DeBERTa-v3   | 0             | 0            | 4.758        |



Εικόνα 5.10-2 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking, Matching), για το σύνολο Google-Amazon.

Πίνακας 5.10-3 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking, Matching), για το σύνολο Walmart-Amazon.

| Μοντέλο      | Embedding (s) | Blocking (s) | Matching (s) |
|--------------|---------------|--------------|--------------|
| MiniLM-L6    | 21.411        | 0.081        | 0.624        |
| GTE-small    | 42.493        | 0.071        | 1.281        |
| E5-small     | 42.507        | 0.089        | 1.328        |
| BGE-small    | 44.793        | 0.070        | 1.323        |
| Arctic-embed | 37.553        | 0.102        | 0.719        |
| DeBERTa-v3   | 0             | 0            | 7.807        |



Εικόνα 5.10-3 Ανάλυση υπολογιστικού κόστους ανά φάση (Embedding, Blocking, Matching), για το σύνολο Walmart-Amazon

### 5.10.2 Υπολογιστικό κόστος Cross-Encoder

Η αρχιτεκτονική Cross-Encoder του DeBERTa-v3 παρουσιάζει θεμελιωδώς διαφορετική κατανομή κόστους: δεν εκτελεί φάση Embedding ή Blocking, καθώς δέχεται απευθείας ζεύγη εγγραφών ως είσοδο. Το σύνολο του υπολογιστικού φόρτου συγκεντρώνεται στη φάση Matching, όπου οι δύο εγγραφές συνενώνονται και περνούν ταυτόχρονα από όλα τα επίπεδα του Transformer. Τα μετρηθέντα αποτελέσματα αποκαλύπτουν ότι ο DeBERTa-v3 απαιτεί 7.807 δευτερόλεπτα για την αξιολόγηση 300 ζευγών στο Walmart-Amazon, έναντι 0.624 δευτερολέπτων για τον MiniLM-L6 δηλαδή ο DeBERTa είναι 12.5 φορές πιο αργός στη φάση Matching. Κλιμακώνοντας στα 10.000 ζεύγη, ο MiniLM-L6 απαιτεί εκτιμώμενα 20.8 δευτερόλεπτα, ενώ ο DeBERTa περίπου 260 δευτερόλεπτα μια διαφορά που γίνεται απαγορευτική σε πραγματικά σενάρια παραγωγής με εκατομμύρια υποψήφια ζεύγη. Η πολυπλοκότητα αυτή είναι αναπόφευκτη δομική συνέπεια της αρχιτεκτονικής, κάθε ζεύγος απαιτεί ένα ανεξάρτητο πέρασμα inference μέσα από τα 12 επίπεδα Transformer του μοντέλου με πολυπλοκότητα  $O(L^2)$  ως προς το μήκος της ακολουθίας  $L$  (Khattab & Zaharia, 2020).

### 5.10.3 Συμπέρασμα σχεδιασμού για συστήματα παραγωγής

Η παραπάνω ανάλυση υπαγορεύει μονοσήμαντα τη χρήση υβριδικού pipeline σε περιβάλλοντα παραγωγής. Στη φάση Blocking, ο Bi-Encoder (π.χ. MiniLM-L6) περιορίζει εξαντλητικά τον χώρο αναζήτησης από εκατομμύρια πιθανά ζεύγη σε μερικές χιλιάδες υποψηφίων, με υπολογιστικό κόστος που παραμένει πρακτικά αμελητέο (κάτω από 0.5% του συνολικού χρόνου). Στη συνέχεια, ο DeBERTa-v3 Cross-Encoder εφαρμόζεται αποκλειστικά στο μειωμένο αυτό υποσύνολο για τελική ταξινόμηση υψηλής ακρίβειας, αποφεύγοντας την εφαρμογή του σε ολόκληρο το καρτεσιανό γινόμενο των δύο βάσεων. Αυτή η προσέγγιση, γνωστή στη βιβλιογραφία ως Re-ranking pipeline, συνδυάζει την ταχύτητα του Bi-Encoder με την ακρίβεια του Cross-Encoder και αποτελεί σήμερα την κρατούσα πρακτική σε συστήματα παραγωγής μεγάλης κλίμακας (Reimers & Gurevych, 2019).

## 6 Συμπεράσματα και Μελλοντική Έρευνα

### 6.1 Ανασκόπηση και Σκοπός της Εργασίας

Η παρούσα διπλωματική εργασία επικεντρώθηκε στο περίπλοκο πρόβλημα της Αντιστοίχισης Εγγραφών (Entity Resolution - ER), μια διαδικασία η οποία αποτελεί τον ακρογωνιαίο λίθο για την επιτυχή ενοποίηση δεδομένων από πολλαπλές, ετερογενείς πηγές (Christen, 2012). Στον σύγχρονο ψηφιακό κόσμο, όπου ο όγκος των παραγόμενων δεδομένων αυξάνεται εκθετικά και η πληροφορία αντλείται από αμέτρητες, μη συντονισμένες μεταξύ τους πηγές, ο κατακερματισμός της πληροφορίας αναδεικνύεται σε κρίσιμο εμπόδιο. Χαρακτηριστικό παράδειγμα αποτελεί η παρουσία του ίδιου προϊόντος με δεκάδες διαφορετικές περιγραφές σε πλατφόρμες ηλεκτρονικού εμπορίου, φαινόμενο που υποβαθμίζει την ποιότητα των δεδομένων, μπερδεύει τους αλγορίθμους αναζήτησης και δυσχεραίνει τη λήψη αξιόπιστων επιχειρηματικών αποφάσεων. Πέρα από το ηλεκτρονικό εμπόριο, οι συνέπειες της ελλιπούς αντιστοίχισης εκτείνονται σε κρίσιμους τομείς όπως η υγεία, όπου η αδυναμία ταύτισης ιατρικών φακέλων μπορεί να επιφέρει σοβαρές συνέπειες, αναδεικνύοντας έτσι τη διαχρονική σημασία του ER ως θεμελιώδους πεδίου της Επιστήμης των Δεδομένων.

Για την αντιμετώπιση αυτού του προβλήματος, η εργασία απομακρύνθηκε συνειδητά από τις παραδοσιακές μεθόδους που βασίζονταν σε χειροκίνητους κανόνες, σε προσεκτικά σχεδιασμένα χαρακτηριστικά (feature engineering) και σε απλές αλφαριθμητικές συγκρίσεις ομοιότητας (string similarity metrics). Οι προσεγγίσεις αυτές, αν και αποτελεσματικές σε καθαρά και αυστηρά δομημένα δεδομένα, παρουσιάζουν σοβαρές αδυναμίες όταν καλούνται να διαχειριστούν εκτενείς περιγραφές, ελλιπή πεδία και έντονη συντακτική ετερογένεια. Αντίθετα, η παρούσα έρευνα βασίστηκε αποκλειστικά στον χώρο της Βαθιάς Μάθησης (Deep Learning), ο οποίος έχει αποδείξει την ανωτερότητά του ακριβώς σε θορυβώδη και περιγραφικά δεδομένα (Mudgal et al., 2018). Με όχημα την επαναστατική αρχιτεκτονική Transformer (Vaswani et al., 2017) και τον μηχανισμό προσοχής (attention), ο οποίος επιτρέπει στο μοντέλο να σταθμίζει τη σημασία κάθε λέξης σε σχέση με ολόκληρη την ακολουθία, και αξιοποιώντας Προ-εκπαιδευμένα Γλωσσικά Μοντέλα (Pre-trained Language Models) όπως το BERT (Devlin et al., 2018), σχεδιάστηκε το ολοκληρωμένο σύστημα LexiSync. Η υλοποίηση στηρίχθηκε στις προηγμένες υποδομές της βιβλιοθήκης transformers (Wolf et al., 2020), διασφαλίζοντας τόσο τη μεθοδολογική ορθότητα όσο και την αναπαραγωγικότητα των πειραμάτων.

Ο κύριος ερευνητικός στόχος ήταν η διερεύνηση της ικανότητας αυτών των νευρωνικών δικτύων να εκπαιδεύονται σε ένα δομημένο πηγαίο πεδίο δεδομένων (DBLP-Scholar) και στη συνέχεια να μεταφέρουν τη γνώση τους (Domain Adaptation) σε εξαιρετικά θορυβώδη πεδία-στόχους (Amazon-Google και Walmart-Amazon), επιλύοντας ταυτόχρονα το φαινόμενο της καταστροφικής λήθης (catastrophic forgetting) (Kirkpatrick et al., 2017). Παράλληλα, η εργασία επιδίωξε να απαντήσει σε ένα δεύτερο, εξίσου πρακτικό ερώτημα: τη διερεύνηση της σχέσης μεταξύ ακρίβειας και υπολογιστικού κόστους (Efficiency vs. Accuracy), συγκρίνοντας ισχυρές αλλά απαιτητικές αρχιτεκτονικές, όπως το DeBERTa-v3, με δραστικά ελαφρύτερα μοντέλα κατάλληλα για περιβάλλοντα παραγωγής. Επιπρόσθετα, εξετάστηκε ο τρόπος με τον οποίο επιμέρους τεχνικές επιλογές η σειριοποίηση με ειδικούς δείκτες, η σταθμισμένη συνάρτηση απώλειας για την αντιμετώπιση της ανισορροπίας κλάσεων και η δυναμική βελτιστοποίηση κατωφλίου συνεισφέρουν στη συνολική απόδοση του συστήματος. Με τον τρόπο αυτό, η μελέτη δεν περιορίστηκε στην επιβεβαίωση της αποτελεσματικότητας των σύγχρονων γλωσσικών μοντέλων, αλλά επιχείρησε να φωτίσει

και τις σχεδιαστικές παραμέτρους που καθιστούν ένα σύστημα Entity Resolution αξιόπιστο, ανθεκτικό και εφαρμόσιμο σε ρεαλιστικές συνθήκες.

## **6.2 Βασικά Ερευνητικά Συμπεράσματα**

### **6.2.1 Η Κρισιμότητα της Σειριοποίησης και των Ειδικών Δεικτών (Special Tokens)**

Αποδείχθηκε ότι η απλή συνένωση (concatenation) των πεδίων μιας εγγραφής σε ένα ενιαίο κείμενο δεν επαρκεί για την επιτυχή αντιστοίχιση, καθώς στερεί από το μοντέλο την πληροφορία για το σε ποιο χαρακτηριστικό ανήκει η κάθε λέξη. Η εισαγωγή ειδικών δεικτών (Tags) όπως [T] για τον τίτλο, [B] για τη μάρκα και [P] για την τιμή στο λεξιλόγιο του μοντέλου, επέτρεψε στον μηχανισμό Attention να διατηρήσει τη δομική πληροφορία των πινάκων ακόμη και μετά τη μετατροπή τους σε μονοδιάστατο κείμενο. Η σημασία της επιλογής αυτής αναδείχθηκε ιδιαίτερα στο απαιτητικό σύνολο Walmart-Amazon, όπου κρίσιμες πληροφορίες όπως ο κωδικός μοντέλου είναι συχνά «κρυμμένες» μέσα στο ελεύθερο κείμενο του τίτλου· οι δείκτες λειτούργησαν ουσιαστικά ως σημασιολογικές «ταμπέλες» καθοδήγησης, βοηθώντας το δίκτυο να αντιστοιχίσει σωστά τις πληροφορίες. Όπως έχει καταδειχθεί και από τους Li et al. (2021) στο σύστημα Ditto, αυτή η τεχνική επιτρέπει στο νευρωνικό δίκτυο να αντιληφθεί τη σημασιολογία των πεδίων αυτόματα, υποκαθιστώντας πρακτικά την ανάγκη για χρονοβόρα και δαπανηρή χειροκίνητη εξαγωγή χαρακτηριστικών (feature engineering).

### **6.2.2 Η Επίλυση του Majority Bias μέσω Σταθμισμένης Απώλειας (Weighted Loss)**

Στο Entity Resolution, η ανισορροπία των κλάσεων είναι ακραία, με τα μη-ταιριάζοντα ζεύγη να ξεπερνούν συντριπτικά τις πραγματικές ταυτίσεις, γεγονός που καθιστά μετρικές όπως η Ακρίβεια (Accuracy) παραπλανητικές και επιβάλλει τη χρήση του F1-Score ως κύριου δείκτη αξιολόγησης (Christen, 2012). Χωρίς ειδική παρέμβαση, ένα μοντέλο εκπαιδευμένο με την τυπική συνάρτηση κόστους τείνει να προβλέπει σχεδόν πάντα την πλειοψηφούσα κλάση, επιτυγχάνοντας φαινομενικά υψηλό Accuracy αλλά πρακτικά μηδενικό Recall. Η ανάπτυξη του μηχανισμού WeightedTrainer στο LexiSync, ο οποίος εφάρμοσε Σταθμισμένη Διασταυρούμενη Εντροπία (Weighted Cross-Entropy), υποχρέωσε

το νευρωνικό δίκτυο να «τιμωρείται» αυστηρότερα για τα False Negatives, αναγκάζοντάς το να δώσει προσοχή στα σπάνια θετικά δείγματα. Παράλληλα, η συστηματική μελέτη ευαισθησίας του μειωτικού συντελεστή βάρους ανέδειξε ότι η αρχιτεκτονική είναι εξαιρετικά ανθεκτική (robust) στην ακριβή ρύθμιση της παραμέτρου αυτής, καθώς όλες οι δοκιμασμένες τιμές παρήγαγαν ιδιαίτερα κοντινά αποτελέσματα. Η παρέμβαση αυτή, σε συνδυασμό με τη δυναμική αναζήτηση βέλτιστου κατωφλίου, αύξησε σημαντικά την Ανάκληση (Recall) στα σύνολα δεδομένων του ηλεκτρονικού εμπορίου.

### 6.2.3 Διατήρηση της Γνώσης του Πηγαίου Πεδίου μέσω Μηχανισμών Experience Replay

Η μεταφορά ενός μοντέλου από τα ακαδημαϊκά δεδομένα στα εμπορικά προϊόντα εγκυμονεί τον κίνδυνο της καταστροφικής λήθης (catastrophic forgetting), καθώς το δίκτυο, στην προσπάθειά του να ελαχιστοποιήσει το σφάλμα στα νέα δεδομένα, τείνει να τροποποιεί ριζικά τα βάρη του, αλλοιώνοντας τη γνώση που είχε ήδη αποκτήσει (Kirkpatrick et al., 2017). Για την αντιμετώπιση του φαινομένου αυτού, σχεδιάστηκε και ενσωματώθηκε ένας μηχανισμός Replay Buffer, αποτελούμενος από μόλις 1.200 στοχευμένα δείγματα της παλαιάς γνώσης (DBLP), τα οποία επανεισάγονται στο νέο ενοποιημένο σύνολο εκπαίδευσης (Joint Dataset). Τα αποτελέσματα ήταν συμβατά με την αναμενόμενη συμπεριφορά: παρά την εντατική εκπαίδευση επί επτά εποχών στα εμπορικά σύνολα, τα μοντέλα διατήρησαν F1-Score άνω του 0.93 στο πηγαίο πεδίο, υποδεικνύοντας ότι η συνολική στρατηγική συνεχούς μάθησης (Continual Learning) είναι εφικτή και αποτελεσματική.

### 6.2.4 Ακρίβεια έναντι Υπολογιστικού Κόστους (Efficiency vs. Accuracy)

Η συγκριτική ανάλυση ανέδειξε το μοντέλο microsoft/deberta-v3-base ως αυτό με την υψηλότερη απόλυτη απόδοση και στα τρία σύνολα δεδομένων, χαρακτηριστικό που αποδίδεται στον μηχανισμό Disentangled Attention, ο οποίος διαχωρίζει το περιεχόμενο από τη θέση της λέξης (He et al., 2020) ιδιότητα ιδιαίτερα χρήσιμη στο Entity Resolution, όπου η θέση ενός αριθμού ή ενός κωδικού είναι συχνά καθοριστική. Εντούτοις, το

κρισιμότερο εύρημα της ανάλυσης αυτής είναι ότι αρχιτεκτονικές βελτιστοποιημένες μέσω απόσταξης γνώσης (knowledge distillation), όπως το all-MiniLM-L6-v2 (Wang et al., 2020), πέτυχαν εξαιρετικά ανταγωνιστικά αποτελέσματα απαιτώντας ένα κλάσμα των υπολογιστικών πόρων του DeBERTa. Σε δομημένα πεδία η διαφορά απόδοσης ήταν πρακτικά αμελητέα, ενώ ακόμη και στα πιο θορυβώδη σύνολα παρέμεινε διαχειρίσιμη. Δεδομένου ότι το DeBERTa-v3 διαθέτει περίπου 86 εκατομμύρια παραμέτρους έναντι των ~33 εκατομμυρίων των ελαφρών μοντέλων, η μικρή θυσία σε ακρίβεια αντισταθμίζεται από σημαντική εξοικονόμηση σε χρόνο εκπαίδευσης, μνήμη και χρόνο εξαγωγής συμπερασμάτων, καθιστώντας τα ελαφρά μοντέλα μια ιδιαίτερα ελκυστική λύση για συστήματα παραγωγής μεγάλης κλίμακας.

### 6.3 Περιορισμοί της Παρούσας Μελέτης

#### 6.3.1 Υπολογιστικός Φόρτος του Cross-Encoder

Η αρχιτεκτονική Cross-Encoder προσφέρει τη μέγιστη δυνατή ακρίβεια συγχωνεύοντας τις δύο εγγραφές σε μία ενιαία ακολουθία και εκτελώντας διασταυρούμενη προσοχή (cross-attention) σε επίπεδο λέξης, αλλά είναι υπολογιστικά βαριά. Η ίδια η φύση της αρχιτεκτονικής επιβάλλει ένα ανεξάρτητο πέρασμα inference για κάθε υποψήφιο ζεύγος, με αποτέλεσμα η εξαντλητική εφαρμογή της στο καρτεσιανό γινόμενο δύο βάσεων να οδηγεί σε πολυπλοκότητα  $O(n^2)$  ως προς το πλήθος των εγγραφών (Reimers & Gurevych, 2019). Όπως καταδείχθηκε και στην ανάλυση υπολογιστικού κόστους (Ενότητα 5.10), ο DeBERTa-v3 αποδείχθηκε κατά μία τάξη μεγέθους πιο αργός από τους Bi-Encoders στη φάση του matching, μια διαφορά που γίνεται απαγορευτική σε ρεαλιστικά σενάρια με εκατομμύρια πιθανά ζεύγη. Το γεγονός αυτό καθιστά το σύστημα μη βιώσιμο για απευθείας εφαρμογή σε τεράστιες βάσεις δεδομένων, επιβάλλοντας την αναγκαιότητα ενός προηγούμενου σταδίου προ-φιλτραρίσματος (blocking) που θα περιορίζει δραστικά τον χώρο αναζήτησης πριν την παρέμβαση του ακριβούς ταξινομητή.

#### 6.3.2 Έλλειψη Πειραματικής Αξιολόγησης του Blocker

Στο πλαίσιο της εργασίας σχεδιάστηκε και υλοποιήθηκε ένας αυτόνομος μηχανισμός blocking, βασισμένος σε αρχιτεκτονική Bi-Encoder και σε δομές προσεγγιστικής αναζήτησης πλησιέστερων γειτόνων (ANNS) μέσω της βιβλιοθήκης FAISS. Ωστόσο, η

ολοκληρωμένη πειραματική αξιολόγηση του ενιαίου, end-to-end pipeline (blocking + matching) δεν εντάχθηκε στο πεδίο της παρούσας μελέτης. Συγκεκριμένα, η συστηματική αποτίμηση του συνολικού συστήματος με τις καθιερωμένες μετρικές αξιολόγησης blocker, όπως οι Pairs Completeness και Pairs Quality, καθώς και η μέτρηση της επίδρασης που έχει η απώλεια πραγματικών matches κατά τη φάση του blocking στο τελικό recall του συστήματος, παραμένουν ανοιχτά ζητήματα. Η πλήρης αυτή αξιολόγηση, η οποία θα τεκμηριώνει τη λειτουργία του LexiSync ως ολοκληρωμένου συστήματος δύο σταδίων, αποτελεί άμεση και φυσική μελλοντική επέκταση της εργασίας.

### 6.3.3 Περιορισμοί Hardware και Apple Silicon

Το σύνολο των πειραμάτων εκτελέστηκε τοπικά σε περιβάλλον macOS, αξιοποιώντας την αρχιτεκτονική Apple Silicon και το υποσύστημα Metal Performance Shaders (MPS) του PyTorch. Αν και η προσέγγιση αυτή είναι σχετικά καινοτόμα και αναδεικνύει τη δυνατότητα εκπαίδευσης μοντέλων Transformer εκτός του παραδοσιακού οικοσυστήματος CUDA της NVIDIA, η υποστήριξη του PyTorch για το MPS δεν είναι ακόμη απόλυτα ώριμη. Ορισμένες μαθηματικές πράξεις τανυστών δεν υποστηρίζονται εγγενώς από το backend του MPS, γεγονός που επέβαλε τη δυναμική μεταφορά τους στην CPU (μέσω της ρύθμισης `PYTORCH_ENABLE_MPS_FALLBACK`) προκειμένου να αποφευχθούν διακοπές κατά την εκπαίδευση. Η μεταφορά αυτή, αν και διασφάλισε τη σταθερότητα της διαδικασίας, αύξησε ελαφρώς τον συνολικό χρόνο εκπαίδευσης και ενδέχεται να μην αντικατοπτρίζει πλήρως τις βέλτιστες επιδόσεις χρόνου που θα ήταν εφικτές σε ένα πλήρως βελτιστοποιημένο περιβάλλον υλικού.

### 6.3.4 Μονόγλωσση Προσέγγιση

Το σύστημα σχεδιάστηκε και αξιολογήθηκε αποκλειστικά σε σύνολα δεδομένων της αγγλικής γλώσσας (DBLP-Scholar, Amazon-Google, Walmart-Amazon). Συνεπώς, τα συμπεράσματα της μελέτης αφορούν αυστηρά αγγλόφωνα δεδομένα, ενώ η συμπεριφορά των εξεταζόμενων μοντέλων σε πολυγλωσσικά ή μη αγγλόφωνα περιβάλλοντα δεν εξετάστηκε στην παρούσα φάση. Δεδομένου ότι η σημασιολογική κατανόηση των γλωσσικών μοντέλων εξαρτάται άμεσα από τα δεδομένα της προ-εκπαίδευσής τους, η γενίκευση των ευρημάτων σε άλλες γλώσσες δεν μπορεί να θεωρηθεί δεδομένη και θα απαιτούσε ξεχωριστή πειραματική διερεύνηση.

## 6.4 Κατευθύνσεις Μελλοντικής Έρευνας

### 6.4.1 Αξιοποίηση Μεγάλων Γλωσσικών Μοντέλων (Large Language Models - LLMs)

Μια φυσική εξέλιξη της παρούσας έρευνας είναι η διερεύνηση Παραγωγικών Μοντέλων (Generative Models) στο πρόβλημα της αντιστοίχισης εγγραφών. Σύμφωνα με πρόσφατες μελέτες (Peeters et al., 2025), τα LLMs μπορούν να επιλύσουν προβλήματα αντιστοίχισης μέσω κατάλληλης μηχανικής προτροπών (Prompt Engineering), επιτυγχάνοντας συγκρίσιμη ή και καλύτερη απόδοση με ελάχιστα παραδείγματα (few-shot learning) και χωρίς την ανάγκη εκτενούς μικρο-ρύθμισης. Πέρα από την απόδοση, ένα ιδιαίτερα ελκυστικό χαρακτηριστικό της προσέγγισης αυτής είναι η δυνατότητα παραγωγής δομημένων εξηγήσεων (Explainable AI), όπου το μοντέλο δεν περιορίζεται στην απόφαση «match» ή «no-match», αλλά υποδεικνύει και τους λόγους που οδήγησαν στην ταύτιση ή στη διάκριση δύο εγγραφών. Η ιδιότητα αυτή θα μπορούσε να ενισχύσει σημαντικά την εμπιστοσύνη και τη διαφάνεια του συστήματος σε κρίσιμα πεδία εφαρμογής, ενώ ταυτόχρονα μειώνει το κόστος που συνδέεται με τη συλλογή και επισημείωση μεγάλων συνόλων δεδομένων εκπαίδευσης.

### 6.4.2 Υβριδικά Συστήματα Bi-Encoder και Cross-Encoder

Για την οριστική επίλυση του προβλήματος της ταχύτητας που εντοπίστηκε στους περιορισμούς της μελέτης, προτείνεται η πλήρης ανάπτυξη και πειραματική επικύρωση μιας αρχιτεκτονικής δύο σταδίων (Re-ranking pipeline). Στην προσέγγιση αυτή, ένας ταχύτατος Bi-Encoder (όπως το Sentence-BERT) θα λειτουργεί ως μηχανισμός υψηλής ανάκλησης, περιορίζοντας εξαντλητικά τον χώρο αναζήτησης από εκατομμύρια πιθανά ζεύγη σε μια διαχειρίσιμη λίστα υποψηφίων (Reimers & Gurevych, 2019). Στη συνέχεια, ο βαρύν αλλά ακριβέστερος Cross-Encoder του LexiSync θα αναλαμβάνει αποκλειστικά την τελική ταξινόμηση υψηλής ακρίβειας επί του μειωμένου αυτού υποσυνόλου. Η κατεύθυνση αυτή συνδέεται άμεσα με τον αυτόνομο μηχανισμό blocking που σχεδιάστηκε στην παρούσα εργασία (Ενότητα 3.8) και αποτελεί τη φυσική ολοκλήρωσή του, συνδυάζοντας την ταχύτητα του Bi-Encoder με την ακρίβεια του Cross-Encoder σε ένα ενιαίο, κλιμακούμενο σύστημα κατάλληλο για περιβάλλοντα παραγωγής μεγάλης κλίμακας.

### 6.4.3 Ενεργή Μάθηση (Active Learning)

Μια ακόμη υποσχόμενη κατεύθυνση είναι η ενσωμάτωση τεχνικών Ενεργής Μάθησης (Active Learning) στη διαδικασία εκπαίδευσης. Αντί να εκπαιδεύεται το μοντέλο σε ένα στατικό και εκ των προτέρων πλήρως επισημειωμένο σύνολο, το σύστημα θα μπορούσε να εντοπίζει αυτόνομα τα ζεύγη για τα οποία παρουσιάζει τη μεγαλύτερη αβεβαιότητα δηλαδή εκείνα με πιθανότητα ταύτισης κοντά στο όριο απόφασης και να τα προωθεί επιλεκτικά σε έναν άνθρωπο-διαχειριστή για χειροκίνητη επιβεβαίωση. Με τον τρόπο αυτό, η ανθρώπινη προσπάθεια επικεντρώνεται αποκλειστικά στις πιο δύσκολες και πληροφοριακά πλούσιες περιπτώσεις, μειώνοντας δραματικά το κόστος και τον χρόνο δημιουργίας επισημειωμένων δεδομένων (labeled datasets), τα οποία αποτελούν συχνά τον πιο δαπανηρό πόρο σε εφαρμογές μηχανικής μάθησης.

### 6.4.4 Πειραματική Αξιολόγηση του Μηχανισμού Blocking

Σε άμεση συνέχεια του αντίστοιχου περιορισμού που αναγνωρίστηκε, προτείνεται η πλήρης πειραματική αξιολόγηση του προτεινόμενου blocker και η συγκριτική αποτίμησή του έναντι καθιερωμένων μεθόδων. Συγκεκριμένα, η απόδοση του ANNS-based μηχανισμού θα μπορούσε να συγκριθεί τόσο με κλασικές προσεγγίσεις προ-φιλτραρίσματος, όπως η ευρετηρίαση με βάση το TF-IDF ή ο αλγόριθμος BM25, όσο και με πιο εξελιγμένα, νευρωνικά schemes όπως το DeepBlocker. Μια τέτοια συγκριτική μελέτη, αξιοποιώντας τις μετρικές Pairs Completeness, Pairs Quality και Reduction Ratio, θα τεκμηρίωνε με αντικειμενικά κριτήρια τη σχεδιαστική επιλογή του Bi-Encoder ως μηχανισμού blocking και θα ολοκλήρωνε την αξιολόγηση του LexiSync ως ενιαίου, end-to-end συστήματος Entity Resolution.

## Βιβλιογραφία

Ακολουθούν οι βιβλιογραφικές αναφορές (πηγές) της Εργασίας.

Albayrak, O. S., Aytekin, T., & Kalaycı, T. A. (2022). Duplicate product record detection engine for e-commerce platforms. *Expert Systems with Applications*, 193, 116420. <https://doi.org/10.1016/j.eswa.2021.116420>

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249-259. <https://doi.org/10.1016/j.neunet.2018.07.011>

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H. S., & Ranzato, M. (2019). On tiny episodic memories in continual learning. arXiv preprint arXiv:1902.10486. <https://arxiv.org/abs/1902.10486>

Christen, P. (2012). Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer Science & Business Media. <https://doi.org/10.1007/978-3-642-31164-2>

Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9268-9277). <https://doi.org/10.1109/CVPR.2019.00949>

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805. <https://arxiv.org/abs/1810.04805>

- Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: Adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, 61(6), 2623-2640. <https://doi.org/10.1021/acs.jcim.1c00160>
- Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210. <https://doi.org/10.1080/01621459.1969.10501049>
- Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: Pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, 12(1), 49-57. <https://doi.org/10.1145/1882471.1882479>
- Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. <https://arxiv.org/abs/2104.08821>
- He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. *International Conference on Learning Representations (ICLR)*. <https://arxiv.org/abs/2006.03654>
- He, P., Gao, J., & Chen, W. (2021). DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*. <https://arxiv.org/abs/2111.09543>
- Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A survey of methods for addressing class imbalance in deep-learning based natural language processing. In *Proceedings of the 17th Conference of the European Chapter of the Association for*

Computational Linguistics (pp. 523-540). <https://doi.org/10.18653/v1/2023.eacl-main.38>

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535-547 ACM Digital Library. <https://doi.org/10.1109/TBDATA.2019.2921572>

Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. Proceedings of the 43rd International ACM SIGIR Conference (pp. 39-48). <https://doi.org/10.1145/3397271.3401075>

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences (PNAS), 114(13), 3521-3526. <https://doi.org/10.1073/pnas.1611835114>

Köpcke, H., Thor, A., & Rahm, E. (2010). Evaluation of entity resolution approaches on real-world match problems. Proceedings of the VLDB Endowment, 3(1-2), 484-493. <https://doi.org/10.14778/1920841.1920904>

Li, Y., Li, J., Suhara, Y., Doan, A., & Tan, W. C. (2021). Deep entity matching with pre-trained language models. Proceedings of the VLDB Endowment, 14(1), 50-60. <https://www.vldb.org/pvldb/vol14/p50-li.pdf>

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. <https://arxiv.org/abs/1301.3781>

- Mosbach, M., Andriushchenko, M., & Klakow, D. (2021). On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In International Conference on Learning Representations (ICLR).  
<https://openreview.net/forum?id=nzpLWnVAyah>
- Mudgal, S., Li, H., Rekatsinas, T., Doan, A., Park, Y., Krishnan, G., Deep, R., Arcaute, E., & Raghavendra, V. (2018). Deep learning for entity matching: A design space exploration. Proceedings of the 2018 International Conference on Management of Data (SIGMOD), 19-34. <https://doi.org/10.1145/3183713.3196926>
- Narayan, A., Chami, I., Orr, L., Arora, S., & Ré, C. (2022). Can foundation models wrangle your data? Proceedings of the VLDB Endowment, 16(4), 738-746.  
<https://arxiv.org/abs/2205.09911>
- Papadakis, G., Skoutas, D., Thanos, E., & Palpanas, T. (2020). Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys*, 53(2), Article 31. <https://doi.org/10.1145/3377455>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32. <https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research (JMLR)*, 12, 2825-2830.  
<https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html>

- Peeters, R., Steiner, A., & Bizer, C. (2025). Entity matching using large language models. Proceedings of the 28th International Conference on Extending Database Technology (EDBT 2025). <https://arxiv.org/abs/2310.11244>
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532-1543). <https://doi.org/10.3115/v1/D14-1162>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). <https://arxiv.org/abs/1908.10084>
- Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends® in Information Retrieval, 3(4), 333-389. [https://www.researchgate.net/publication/220613776\\_The\\_Probabilistic\\_Relevance\\_Framework\\_BM25\\_and\\_Beyond](https://www.researchgate.net/publication/220613776_The_Probabilistic_Relevance_Framework_BM25_and_Beyond)
- Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? China National Conference on Chinese Computational Linguistics, 194-206. <https://arxiv.org/abs/1905.05583>
- Trabelsi, M., Heflin, J., & Cao, J. (2022). DAME: Domain adaptation for matching entities. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22)* (pp. 1016-1024). Association for Computing Machinery. <https://doi.org/10.1145/3488560.3498486>

- U.S. Government Accountability Office. (2019). *Health information technology: Approaches and challenges to electronically matching patients' records across providers* (GAO-19-197). <https://www.gao.gov/products/gao-19-197>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30. <https://arxiv.org/abs/1706.03762>
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, M., & Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Lightweight Pre-trained Language Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 5776-5788. <https://arxiv.org/abs/2002.10957>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. <https://aclanthology.org/2020.emnlp-demos.6/>

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.