

Παραγωγή περιλήψεων με χρήση ΤΝ



Μεταπτυχιακή Εξειδίκευση στα Πληροφοριακά Συστήματα (Msc)

Διπλωματική Εργασία

Παραγωγή περιλήψεων με χρήση ΤΝ.

Δημήτριος Μπρούμας

Επιβλέπων : Δρ. Νίκος Καρούσος

Παραγωγή περιλήψεων με χρήση TN

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Δημήτριου Μπρούμα 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

Παραγωγή περιλήψεων με χρήση TN

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής ,θα ήθελα αρχικά να ευχαριστήσω τον επιβλέπων καθηγητή μου, Δρ. Νίκο Καρούσο με την υπομονή του και την κατανόησή του. Επίσης θα ήθελα να ευχαριστήσω την οικογένειά μου για την ενθάρρυνση και υποστήριξη σε όλο το διάστημα των σπουδών μου , ιδιαίτερα τον τελευταίο χρόνο λόγο του προβλήματος υγείας πού είχαμε στην οικογένειά μας και ευτυχώς σταθήκαμε νικητές.

Παραγωγή περιλήψεων με χρήση TN

Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει την παραγωγή περιλήψεων από απαντήσεις ανοικτού τύπου σε ερωματολογία αξιολόγησης εκπαιδευτικών διαδικασιών, με έμφαση στην ελληνική γλώσσα. Το πρόβλημα που αντιμετωπίζεται είναι ότι οι ανοικτές απαντήσεις των φοιτητών περιέχουν χρήσιμη ποιοτική πληροφορία, αλλά η χειροκίνητη ανάγνωση, ταξινόμηση και σύνοψή τους απαιτεί σημαντικό χρόνο και προσπάθεια, ειδικά όταν ο αριθμός των απαντήσεων είναι μεγάλος.

Στο πλαίσιο της εργασίας υλοποιήθηκε ένα αναπαραγωγίμο pipeline¹ σε Python, το οποίο διαβάζει αρχεία Excel με απαντήσεις φοιτητών και πρότυπες περιλήψεις αναφοράς², καθαρίζει τα κείμενα, ομαδοποιεί τα σχόλια ανά Πρόγραμμα Σπουδών και κατηγορία ερώτησης και στη συνέχεια παράγει περιλήψεις με Μεγάλα Γλωσσικά Μοντέλα - LLM. Η ομαδοποίηση γίνεται σε επίπεδο Προγράμματος Σπουδών και ερώτησης, ώστε για κάθε πρόγραμμα να παράγονται τρεις περιλήψεις: μία για τις προτάσεις βελτίωσης, μία για τα αρνητικά στοιχεία και μία για τα θετικά στοιχεία.

Η πειραματική αξιολόγηση πραγματοποιήθηκε σε 126 περιπτώσεις περίληψης, οι οποίες αντιστοιχούν σε 42 Προγράμματα Σπουδών και τρεις κατηγορίες ερωτήσεων ανά πρόγραμμα. Το Πρόγραμμα Σπουδών ΔΙΠ δεν συμπεριλήφθηκε στην ποσοτική αξιολόγηση, επειδή δεν υπήρχαν διαθέσιμες πρότυπες περιλήψεις αναφοράς για τις αντίστοιχες κατηγορίες ερωτήσεων. Τα μοντέλα που αξιολογήθηκαν είναι τα gpt-4o, gpt-4o-mini και gpt-3.5-turbo.

¹ Ο όρος pipeline χρησιμοποιείται εδώ με την έννοια μιας οργανωμένης και επαναλήψιμης ροής επεξεργασίας, στην οποία κάθε στάδιο τροφοδοτεί το επόμενο: εισαγωγή δεδομένων, καθαρισμός κειμένου, ομαδοποίηση απαντήσεων, παραγωγή περιλήψεων, αποθήκευση αποτελεσμάτων και αξιολόγηση. Τα μοντέλα που αξιοποιούνται ανήκουν στην κατηγορία των Μεγάλων Γλωσσικών Μοντέλων, δηλαδή Large Language Models (LLMs), τα οποία έχουν εκπαιδευτεί σε μεγάλης κλίμακας γλωσσικά δεδομένα και μπορούν να παράγουν συνεκτικό φυσικό λόγο.

² Οι παραγόμενες περιλήψεις συγκρίνονται με πρότυπες περιλήψεις αναφοράς, δηλαδή περιλήψεις που έχουν δημιουργηθεί από άνθρωπο και χρησιμοποιούνται ως βάση σύγκρισης για την αξιολόγηση της ποιότητας των αποτελεσμάτων.

Παραγωγή περιλήψεων με χρήση TN

Τα δεδομένα της εργασίας χρησιμοποιήθηκαν αποκλειστικά για πειραματικούς σκοπούς και ήταν ανωνυμοποιημένα. Βασίστηκαν σε προεπεξεργασμένο σύνολο δεδομένων που αντλήθηκε και επεξεργάστηκε από τον επιβλέποντα καθηγητή, ώστε να μην εμφανίζονται πραγματικά αναγνωρίσιμα δεδομένα ιδρύματος, διδασκόντων ή φοιτητών.

Για την αξιολόγηση των παραγόμενων περιλήψεων χρησιμοποιήθηκαν μετρικές όπως BLEU, ROUGE, Sentence-BERT similarity, character-level cosine similarity, length ratio και keyword overlap. Τα αποτελέσματα δείχνουν ότι τα gpt-4o και gpt-4o-mini παρουσιάζουν την καλύτερη συνολική απόδοση, ενώ το gpt-3.5-turbo υστερεί κυρίως ως προς το μήκος και τη λεξική επικάλυψη των περιλήψεων. Η εργασία καταλήγει ότι τα σύγχρονα LLMs μπορούν να υποστηρίξουν αποτελεσματικά τη σύννοψη ελληνόγλωσσων ανοικτών απαντήσεων σε εκπαιδευτικά ερωτηματολόγια, υπό την προϋπόθεση ότι η παραγωγή και η αξιολόγηση γίνονται στη σωστή μονάδα ανάλυσης.

Λέξεις - κλειδιά

Τεχνητή Νοημοσύνη, Μεγάλα Γλωσσικά Μοντέλα, Επεξεργασία Φυσικής Γλώσσας, αυτόματη παραγωγή περιλήψεων, αφαιρετική περίληψη, ανοικτές απαντήσεις, εκπαιδευτικά ερωτηματολόγια, φοιτητικά σχόλια, πρότυπες περιλήψεις αναφοράς, μετρικές αξιολόγησης, σημασιολογική ομοιότητα, ελληνική γλώσσα

Παραγωγή περιλήψεων με χρήση TN

Abstract

This thesis examines the generation of summaries from open-ended responses in educational evaluation questionnaires, with a focus on the Greek language. The main problem addressed is that students' open-ended answers contain valuable qualitative information, but their manual reading, classification and summarization requires considerable time and effort, especially when the number of responses is large.

As part of the thesis, a reproducible Python pipeline³ was implemented. The pipeline loads Excel files containing student responses and reference summaries, cleans the text, groups the comments by Study Programme and question category, and then generates summaries using Large Language Models. The grouping is performed at the level of Study Programme and question, so that three summaries are produced for each programme: one for important points, one for negative comments, and one for positive comments. Thus, the summary is not generated for each individual student comment, but for the complete set of relevant comments.

The experimental evaluation was conducted on 126 summarization cases, corresponding to 42 Study Programmes and three question categories per programme. The Study Programme DIP was not included in the quantitative evaluation because no corresponding reference summaries were available for its question categories. The evaluated models are gpt-4o, gpt-4o-mini and gpt-3.5-turbo.

The data used in the experiments were anonymized and preprocessed before their use in this thesis, so that no real identifiable institutional, teacher or student information is exposed.

The generated summaries were evaluated using several metrics, including BLEU, ROUGE, Sentence-BERT similarity, character-level cosine similarity, length ratio and keyword overlap. The results indicate that gpt-4o and gpt-4o-mini achieve the best overall performance, while gpt-

³ In this thesis, the term pipeline denotes the complete repeatable workflow from data loading and text preprocessing to prompt construction, Large Language Model (LLM) invocation, summary storage and evaluation. The improvement category is treated as suggestions for improvement rather than as a generic category of important comments.

Παραγωγή περιλήψεων με χρήση TN

3.5-turbo performs lower, especially in terms of summary length and lexical overlap. The thesis concludes that modern LLMs can effectively support the summarization of Greek open-ended questionnaire responses in educational evaluation, provided that both generation and evaluation are performed at the correct unit of analysis.

Keywords

Artificial Intelligence, Large Language Models, Natural Language Processing, automatic summary generation, abstractive summarization, open-ended responses, educational questionnaires, student comments, reference summaries, evaluation metrics, semantic similarity, Greeklanguage.

Παραγωγή περιλήψεων με χρήση TN

Πίνακας περιεχομένων

Περίληψη	4
Κεφάλαιο 1 — Εισαγωγή	11
1.1 Αντικείμενο και πλαίσιο της εργασίας	11
1.2 Το πρόβλημα της επεξεργασίας απαντήσεων ανοικτού τύπου	14
1.3 Η σημασία της περίληψης στην εκπαιδευτική αξιολόγηση	16
1.4 Στόχοι της διπλωματικής εργασίας	18
1.5 Συνεισφορά της εργασίας	20
1.6 Μεθοδολογική προσέγγιση (σύντομη επισκόπηση)	22
1.7 Δομή της διπλωματικής εργασίας	24
Κεφάλαιο 2 – Θεωρητικό Υπόβαθρο	25
2.1 Επεξεργασία Φυσικής Γλώσσας (NLP)	25
2.1.1 Ορισμός και ιστορική εξέλιξη του NLP	25
2.1.2 Βασικές αρχές και προκλήσεις στην επεξεργασία κειμένου	26
2.1.3 Προεπεξεργασία κειμένου: τμηματοποίηση κειμένου (tokenization), κανονικοποίηση (normalization), λέξεις-κενά (stopwords), λημματοποίηση (lemmatization)	27
2.1.4 Αναπαράσταση κειμένου: Bag-of-Words, TF-IDF, word embeddings	29
2.1.5 Σύγχρονες αναπαραστάσεις: contextual embeddings (BERT, mBERT, Sentence-BERT)	30
2.2 Τεχνικές Περίληψης Κειμένου	31
2.2.1 Θεωρητική διάκριση extractive και abstractive περίληψης	31
2.2.2 Extractive μέθοδοι: TextRank, LexRank, στατιστικές και γραφοθεωρητικές προσεγγίσεις	32
2.2.3 Abstractive μέθοδοι: seq2seq, attention, Transformers	33
2.2.4 Πλεονεκτήματα και περιορισμοί κάθε κατηγορίας	34
2.2.5 Εφαρμογές περίληψης σε πολυγλωσσικά περιβάλλοντα	35
2.3 Μεγάλα Γλωσσικά Μοντέλα (LLMs)	36
2.3.1 Αρχιτεκτονική Transformer και θεμελιώδεις έννοιες	36
2.3.2 Προεκπαίδευση και fine-tuning	37
2.3.3 Μοντέλα γενικής χρήσης: GPT, Claude, Gemini, Mistral, Llama	38
2.3.4 Πολυγλωσσικά και ελληνικά μοντέλα: mT5, Greek-BERT, Greek-Llama	39
2.3.5 Ικανότητες και περιορισμοί των LLMs στην ελληνική γλώσσα	40
2.3.6 Ζητήματα αξιοπιστίας, παραπληροφόρησης και bias στα LLMs	40

Παραγωγή περιλήψεων με χρήση TN

2.4 Αξιολόγηση Συστημάτων Περίληψης.....	42
2.4.1 Ποσοτικές μετρικές: ROUGE, BLEU, BERTScore.....	42
2.4.2 Σημασιολογικές μετρικές: MoverScore, QuestEval, SummaC	44
2.4.3 Ποιοτική αξιολόγηση: συνοχή, πληρότητα, ακρίβεια, μη παραπληροφόρηση	45
2.4.4 Συγκριτική αξιολόγηση extractive–abstractive μεθόδων	46
2.4.5 Προκλήσεις αξιολόγησης σε ελληνόγλωσσα δεδομένα.....	47
2.5 Απαντήσεις Ανοικτού Τύπου στην Εκπαιδευτική Αξιολόγηση	48
2.5.1 Ρόλος και σημασία των απαντήσεων ανοικτού τύπου.....	48
2.5.2 Ερευνητικές προσεγγίσεις στην ανάλυση εκπαιδευτικών δεδομένων	49
2.5.3 Προηγούμενες εφαρμογές NLP σε εκπαιδευτικά περιβάλλοντα.....	50
2.5.4 Ιδιαιτερότητες των δεδομένων του ΕΑΠ και αντίστοιχων ιδρυμάτων.....	50
Κεφάλαιο 3 — Μεθοδολογία και Υλοποίηση.....	52
3.1 Γενική Περιγραφή του Έργου	52
3.2 Περιγραφή και Κατασκευή του Συνόλου Δεδομένων	53
3.2.1 Πηγές Δεδομένων Τα δεδομένα του έργου προέρχονται από δύο κύριες πηγές:	53
3.2.2 Κατανομή Δεδομένων.....	54
3.3 Αρχιτεκτονική του Συστήματος.....	54
3.4 Προεπεξεργασία Δεδομένων	55
3.4.1 Φόρτωση και Καθαρισμός.....	55
3.4.2 Εξαγωγή Μεταδεδομένων	55
3.5 Ενσωμάτωση Μεγάλων Γλωσσικών Μοντέλων.....	56
3.5.1 Επιλογή Μοντέλων	56
3.5.2 System Prompt (Οδηγίες για τα Μοντέλα)	56
3.5.3 User Prompt (Πλαίσιο Ερώτησης)	57
3.6 Παραγωγή Περιλήψεων.....	58
3.7 Αξιολόγηση των Περιλήψεων	60
3.7.1 Μετρικές που Χρησιμοποιούνται.....	60
3.8 Αποτελέσματα.....	61
3.8.1 Σύνολο δεδομένων αξιολόγησης	61
3.8.2 Μέσες μετρικές ανά μοντέλο	62
3.8.3 Ανάλυση ανά κατηγορία ερώτησης.....	65
3.8.4 Ανάλυση ανά Πρόγραμμα Σπουδών.....	68

Παραγωγή περιλήψεων με χρήση TN

3.8.5 Πολυμετρική σύγκριση με radar charts.....	72
3.9 Ερμηνεία και Συμπεράσματα	75
3.9.1 Σύγκριση μοντέλων	75
3.9.2 Επίδραση της κατηγορίας ερώτησης	75
3.9.3 Επίδραση του Προγράμματος Σπουδών.....	76
3.9.4 Συνολικό συμπέρασμα	76
3.10 Δομή και Λειτουργίες Κύριων Αρχείων Κώδικα.....	77
3.10.1 Αρχεία Προεπεξεργασίας.....	77
3.10.2 Αρχεία LLM Clients	78
3.10.3 Αρχεία Περίληψης	79
3.10.4 Αρχεία Αξιολόγησης.....	79
3.10.5 Αρχεία Υποδομής	81
3.11 Κύριο Πρόγραμμα (main.py).....	82
3.12 Κύρια Ευρήματα και Συμπεράσματα	83
Βιβλιογραφικές αναφορές.....	86

Παραγωγή περιλήψεων με χρήση TN

Κεφάλαιο 1 — Εισαγωγή

1.1 Αντικείμενο και πλαίσιο της εργασίας

Η παρούσα διπλωματική εργασία εξετάζει το ζήτημα της ανάλυσης και σύνθεσης περιλήψεων από ανοικτές απαντήσεις σε ερωτηματολόγια αξιολόγησης εκπαιδευτικών διαδικασιών, με έμφαση στα δεδομένα που παράγονται στο πλαίσιο της τριτοβάθμιας εκπαίδευσης. Τα ερωτηματολόγια αυτά αποτελούν κρίσιμο εργαλείο για την αποτύπωση της εμπειρίας των φοιτητών, την αναγνώριση θετικών πρακτικών και τον εντοπισμό σημείων προς βελτίωση. Ωστόσο, η επεξεργασία των ανοικτών απαντήσεων παραμένει μια ιδιαίτερα απαιτητική διαδικασία, καθώς χαρακτηρίζεται από ετερογένεια, ποικιλία εκφράσεων και σημαντικό όγκο δεδομένων. Το αντικείμενο της εργασίας εντάσσεται στο ευρύτερο πεδίο της Επεξεργασίας Φυσικής Γλώσσας (NLP) και της εφαρμογής Μεγάλων Γλωσσικών Μοντέλων (LLMs) για την παραγωγή συνεκτικών, αξιόπιστων και αντιπροσωπευτικών περιλήψεων που μπορούν να υποστηρίξουν τη λήψη αποφάσεων σε εκπαιδευτικά περιβάλλοντα.

Οι ανοικτές ερωτήσεις στα ερωτηματολόγια αξιολόγησης διαδραματίζουν έναν ιδιαίτερα σημαντικό ρόλο, καθώς επιτρέπουν στους φοιτητές να εκφράσουν ελεύθερα τις εμπειρίες, τις απόψεις και τις προτάσεις τους, χωρίς τους περιορισμούς των κλειστών, προκαθορισμένων επιλογών (Shermis & Burstein, 2013). Μέσω αυτών των απαντήσεων αναδεικνύονται ποιοτικά στοιχεία που δεν μπορούν να αποτυπωθούν με αριθμητικές κλίμακες, όπως η αντίληψη για την ποιότητα διδασκαλίας, η αποτελεσματικότητα των εκπαιδευτικών υλικών, η υποστήριξη από τους διδάσκοντες και η συνολική μαθησιακή εμπειρία. Παράλληλα, οι ανοικτές απαντήσεις επιτρέπουν την ανάδειξη θεμάτων που δεν είχαν προβλεφθεί κατά τον σχεδιασμό του ερωτηματολογίου, προσφέροντας έτσι μια πιο ολοκληρωμένη εικόνα της εκπαιδευτικής πραγματικότητας. Η

Παραγωγή περιλήψεων με χρήση ΤΝ

αξιοποίησή τους, ωστόσο, απαιτεί συστηματική ανάλυση, καθώς η ποικιλία στη γλώσσα, το ύφος και τη δομή των απαντήσεων καθιστά δύσκολη την εξαγωγή συγκεντρωτικών συμπερασμάτων.

Παρά τη σημασία τους, οι ανοικτές απαντήσεις συχνά παραμένουν υποαξιοποιημένες, κυρίως λόγω του μεγάλου φόρτου που συνεπάγεται η χειροκίνητη επεξεργασία τους. Η διαδικασία απαιτεί σημαντικό χρόνο, εξειδικευμένο προσωπικό και υψηλό βαθμό προσοχής, ώστε να διασφαλιστεί ότι οι παραγόμενες περιλήψεις είναι αντικειμενικές, πλήρεις και απαλλαγμένες από μεροληψίες. Επιπλέον, η ανθρώπινη ανάλυση ενέχει κινδύνους ασυνέπειας, καθώς διαφορετικοί αξιολογητές μπορεί να ερμηνεύσουν τις ίδιες απαντήσεις με διαφορετικό τρόπο. Η ανάγκη για μια πιο αποδοτική, επαναλήψιμη και κλιμακούμενη διαδικασία καθιστά επιτακτική την αναζήτηση αυτοματοποιημένων λύσεων που μπορούν να υποστηρίξουν ή και να αντικαταστήσουν μέρος της χειροκίνητης εργασίας.

Σε αυτό το πλαίσιο, η αυτοματοποιημένη ανάλυση και περίληψη ανοικτών απαντήσεων αναδεικνύεται ως μια τεχνολογική καινοτομία με σημαντικές δυνατότητες εφαρμογής. Η πρόοδος των Μεγάλων Γλωσσικών Μοντέλων, όπως τα GPT, Claude, Gemini, Mistral και τα πολυγλωσσικά μοντέλα της οικογένειας T5, έχει καταστήσει εφικτή την παραγωγή υψηλής ποιότητας abstractive περιλήψεων, οι οποίες δεν περιορίζονται στην απλή εξαγωγή φράσεων αλλά συνθέτουν νέο κείμενο που αποτυπώνει με συνοχή τα βασικά θέματα. Τα μοντέλα αυτά μπορούν να διαχειριστούν μεγάλο όγκο δεδομένων, να εντοπίσουν επαναλαμβανόμενα μοτίβα και να παράγουν περιλήψεις προσαρμοσμένες σε συγκεκριμένες οδηγίες, όπως ουδετερότητα, αποφυγή ευαίσθητων στοιχείων και συγκεκριμένο εύρος λέξεων. Η αξιοποίησή τους στην εκπαιδευτική αξιολόγηση μπορεί να συμβάλει στη βελτίωση της ποιότητας των αναφορών και στη μείωση του χρόνου επεξεργασίας.

Στην παρούσα εργασία η παραγωγή περίληψης εξετάζεται κυρίως ως αφαιρετική περίληψη (abstractive summarization). Σε αντίθεση με την εξαγωγική προσέγγιση, όπου επιλέγονται αυτούσιες φράσεις από το αρχικό κείμενο, η αφαιρετική προσέγγιση δημιουργεί νέο κείμενο που αποδίδει συνοπτικά το νόημα των αρχικών απαντήσεων με διαφορετική διατύπωση (See et al., 2017).

Παραγωγή περιλήψεων με χρήση TN

Τέλος, η παρούσα εργασία τοποθετείται στο σημείο τομής μεταξύ τεχνολογικής και εκπαιδευτικής έρευνας, επιδιώκοντας να διερευνήσει συστηματικά την αποτελεσματικότητα διαφορετικών LLMs στην παραγωγή περιλήψεων ελληνόγλωσσων ανοικτών απαντήσεων. Το πλαίσιο αυτό περιλαμβάνει την ανάπτυξη ενός ολοκληρωμένου pipeline που καλύπτει όλα τα στάδια: από την προεπεξεργασία των δεδομένων και τον σχεδιασμό κατάλληλων προτροπών (prompts), έως την παραγωγή περιλήψεων και την αξιολόγησή τους με ποσοτικές και ποιοτικές μεθόδους. Η εργασία φιλοδοξεί να συμβάλει τόσο στην ακαδημαϊκή γνώση όσο και στην πρακτική εφαρμογή, προσφέροντας ένα εργαλείο και μια μεθοδολογία που μπορούν να αξιοποιηθούν από εκπαιδευτικούς οργανισμούς για την ενίσχυση της διαδικασίας αξιολόγησης και τη λήψη τεκμηριωμένων αποφάσεων.

Ο βασικός στόχος της εργασίας είναι να παραχθούν αυτόματα περιλήψεις από ομάδες ανοικτών απαντήσεων και να συγκριθούν με πρότυπες περιλήψεις αναφοράς, οι οποίες έχουν παραχθεί από άνθρωπο και χρησιμοποιούνται ως σημείο σύγκρισης. Με αυτόν τον τρόπο αξιολογείται ποια μοντέλα και ποια εκδοχή prompt παράγουν περιλήψεις που βρίσκονται πιο κοντά στις πρότυπες περιλήψεις ως προς το περιεχόμενο, τη συνοχή και τη χρησιμότητα για την εκπαιδευτική αξιολόγηση.

Παραγωγή περιλήψεων με χρήση TN

1.2 Το πρόβλημα της επεξεργασίας απαντήσεων ανοικτού τύπου

Η επεξεργασία απαντήσεων ανοικτού τύπου αποτελεί μια από τις πιο απαιτητικές διαδικασίες στην ανάλυση δεδομένων εκπαιδευτικής αξιολόγησης, καθώς οι απαντήσεις αυτές δεν ακολουθούν προκαθορισμένη δομή και συχνά περιλαμβάνουν ποικίλες μορφές έκφρασης, προσωπικές εμπειρίες και ελεύθερη χρήση γλώσσας. Σε αντίθεση με τις κλειστές ερωτήσεις, όπου η ανάλυση μπορεί να πραγματοποιηθεί με τυποποιημένες στατιστικές μεθόδους, οι απαντήσεις ανοικτού τύπου απαιτούν ερμηνεία, κατανόηση συμφραζομένων και εντοπισμό θεματικών μοτίβων. Η πολυπλοκότητα αυτή καθιστά την ανάλυση όχι μόνο χρονοβόρα αλλά και επιρρεπή σε υποκειμενικότητα, καθώς η ερμηνεία του περιεχομένου εξαρτάται σε μεγάλο βαθμό από τον αναλυτή. Το πρόβλημα εντείνεται όταν οι απαντήσεις προέρχονται από μεγάλο αριθμό φοιτητών, γεγονός που καθιστά τη χειροκίνητη επεξεργασία πρακτικά δυσχερή.

Οι δυσκολίες της χειροκίνητης ανάλυσης εντοπίζονται τόσο στο επίπεδο της κατανόησης όσο και στο επίπεδο της σύνθεσης. Ο αναλυτής καλείται να διαβάσει εκατοντάδες ή και χιλιάδες απαντήσεις, να εντοπίσει επαναλαμβανόμενα θέματα, να διαχωρίσει ουσιώδεις από δευτερεύουσες πληροφορίες και να συνθέσει μια συνοπτική αλλά αντιπροσωπευτική περίληψη. Η διαδικασία αυτή απαιτεί υψηλό γνωστικό φόρτο, συνεχή συγκέντρωση και σημαντικό χρόνο, ενώ η ποιότητα της τελικής περίληψης μπορεί να επηρεαστεί από παράγοντες όπως η κόπωση, η εμπειρία του αναλυτή ή η προσωπική του οπτική. Επιπλέον, η χειροκίνητη ανάλυση δεν είναι εύκολα επαναλήψιμη: διαφορετικοί αναλυτές ενδέχεται να καταλήξουν σε διαφορετικές περιλήψεις, γεγονός που μειώνει τη συνοχή και την αξιοπιστία των αποτελεσμάτων.

Ένα από τα σημαντικότερα προβλήματα που ανακύπτουν αφορά τον όγκο των δεδομένων και την ετερογένεια των απαντήσεων ανοικτού τύπου. Σε εκπαιδευτικά ιδρύματα με μεγάλο αριθμό φοιτητών, οι απαντήσεις μπορεί να φτάνουν σε χιλιάδες ανά ακαδημαϊκό έτος, δημιουργώντας ένα εκτεταμένο και ανομοιογενές σύνολο δεδομένων. Οι απαντήσεις διαφέρουν ως προς το ύψος, το μήκος, τη σαφήνεια, τη χρήση γλώσσας και το επίπεδο λεπτομέρειας, ενώ συχνά περιλαμβάνουν συναισθηματικά φορτισμένες εκφράσεις, ιδιωτισμούς ή ασύντακτες προτάσεις. Η ετερογένεια αυτή καθιστά δύσκολη την ομαδοποίηση των απαντήσεων σε θεματικές ενότητες

Παραγωγή περιλήψεων με χρήση TN

και απαιτεί εξειδικευμένες τεχνικές για την εξαγωγή νοήματος. Επιπλέον, η παρουσία θορύβου, όπως ανορθογραφίες, συντομογραφίες ή άσχετες πληροφορίες, επιβαρύνει περαιτέρω τη διαδικασία.

Οι προκλήσεις στην εξαγωγή συμπερασμάτων από απαντήσεις ανοικτού τύπου δεν περιορίζονται στην τεχνική δυσκολία της ανάλυσης, αλλά επεκτείνονται και στη μεθοδολογική αξιοπιστία. Η εξαγωγή συμπερασμάτων απαιτεί εντοπισμό τάσεων, συχνότητας θεμάτων και συσχετίσεων, διαδικασίες που δεν είναι εύκολα εφαρμόσιμες χωρίς τυποποιημένα δεδομένα. Η απουσία δομής στις απαντήσεις καθιστά δύσκολη την ποσοτικοποίηση των ευρημάτων, ενώ η ερμηνεία των θεμάτων μπορεί να επηρεαστεί από μεμονωμένες, έντονες ή ακραίες απαντήσεις. Επιπλέον, η προσπάθεια σύνθεσης μιας ενιαίας περίληψης που να αντικατοπτρίζει με ακρίβεια το σύνολο των απόψεων απαιτεί υψηλό βαθμό αφαιρετικής ικανότητας και προσοχής στη λεπτομέρεια. Η διαδικασία αυτή συχνά οδηγεί σε απώλεια πληροφορίας ή σε υπεραπλούστευση σύνθετων θεμάτων.

Τέλος, η ανάγκη για αξιόπιστη, συνεπή και κλιμακούμενη ανάλυση των απαντήσεων ανοικτού τύπου καθιστά επιτακτική την αναζήτηση αυτοματοποιημένων λύσεων που μπορούν να υποστηρίξουν την εκπαιδευτική αξιολόγηση. Η χρήση σύγχρονων τεχνικών Επεξεργασίας Φυσικής Γλώσσας και Μεγάλων Γλωσσικών Μοντέλων προσφέρει τη δυνατότητα παραγωγής περιλήψεων υψηλής ποιότητας, οι οποίες μπορούν να αποτυπώσουν με ακρίβεια τα κυρίαρχα θέματα και να μειώσουν σημαντικά τον χρόνο ανάλυσης. Η αυτοματοποίηση δεν αντικαθιστά την ανθρώπινη κρίση, αλλά λειτουργεί συμπληρωματικά, παρέχοντας ένα σταθερό, επαναλήψιμο και αποδοτικό πλαίσιο ανάλυσης. Η παρούσα εργασία εστιάζει ακριβώς σε αυτή τη διάσταση, διερευνώντας πώς τα LLMs μπορούν να αξιοποιηθούν για την αντιμετώπιση των προκλήσεων που παρουσιάζουν οι απαντήσεις ανοικτού τύπου και για τη βελτίωση της διαδικασίας εξαγωγής συμπερασμάτων σε εκπαιδευτικά περιβάλλοντα.

Παραγωγή περιλήψεων με χρήση ΤΝ

1.3 Η σημασία της περίληψης στην εκπαιδευτική αξιολόγηση

Η περίληψη αποτελεί κρίσιμο εργαλείο στην ανάλυση δεδομένων εκπαιδευτικής αξιολόγησης, καθώς επιτρέπει τη μετατροπή εκτεταμένων, ανομοιογενών και συχνά πολυσύνθετων απαντήσεων σε μια συνεκτική και αξιοποιήσιμη μορφή πληροφορίας. Στο πλαίσιο των απαντήσεων ανοικτού τύπου, όπου οι φοιτητές εκφράζουν ελεύθερα τις εμπειρίες, τις αντιλήψεις και τις προτάσεις τους, η παραγωγή περιλήψεων λειτουργεί ως μηχανισμός συμπύκνωσης της πληροφορίας χωρίς απώλεια των βασικών θεματικών αξόνων. Η διαδικασία αυτή καθίσταται ιδιαίτερα σημαντική σε περιβάλλοντα όπου ο όγκος των δεδομένων είναι μεγάλος και η ανάγκη για γρήγορη, αξιόπιστη και αντικειμενική αποτύπωση των ευρημάτων είναι επιτακτική. Η περίληψη, επομένως, δεν αποτελεί απλώς μια τεχνική συνοπτικής παρουσίασης, αλλά μια ουσιαστική διαδικασία μετασχηματισμού της πληροφορίας που επιτρέπει την εξαγωγή συμπερασμάτων υψηλής προστιθέμενης αξίας.

Οι περιλήψεις διαδραματίζουν καθοριστικό ρόλο στη διαδικασία λήψης αποφάσεων, καθώς παρέχουν στους υπεύθυνους χάραξης πολιτικής, στα ακαδημαϊκά όργανα και στους διδάσκοντες μια καθαρή και οργανωμένη εικόνα των κυρίαρχων τάσεων που αναδύονται από τις απαντήσεις ανοικτού τύπου. Η δυνατότητα να εντοπίζονται γρήγορα τα θετικά στοιχεία, οι αδυναμίες και οι προτάσεις βελτίωσης επιτρέπει την άμεση προσαρμογή των εκπαιδευτικών πρακτικών και την υιοθέτηση στοχευμένων παρεμβάσεων. Επιπλέον, οι περιλήψεις συμβάλλουν στη διαμόρφωση τεκμηριωμένων αποφάσεων, μειώνοντας τον κίνδυνο μεροληψίας που μπορεί να προκύψει από την επιλεκτική ανάγνωση μεμονωμένων απαντήσεων. Με αυτόν τον τρόπο, η περίληψη λειτουργεί ως γέφυρα μεταξύ των ακατέργαστων δεδομένων και της στρατηγικής εκπαιδευτικής διαχείρισης, ενισχύοντας τη διαφάνεια και την αποτελεσματικότητα των διαδικασιών αξιολόγησης.

Σε περιβάλλοντα εξ αποστάσεως ανώτατης εκπαίδευσης, όπως το Ελληνικό Ανοικτό Πανεπιστήμιο (ΕΑΠ), όπου η εκπαιδευτική διαδικασία χαρακτηρίζεται από εξ αποστάσεως μάθηση, ετερογενές φοιτητικό κοινό και συστηματική συλλογή δεδομένων αξιολόγησης, η συνοπτική πληροφορία αποκτά ακόμη μεγαλύτερη σημασία. Οι διοικητικές και ακαδημαϊκές μονάδες χρειάζονται γρήγορη πρόσβαση σε αξιόπιστες περιλήψεις που αποτυπώνουν την εμπειρία

Παραγωγή περιλήψεων με χρήση TN

των φοιτητών σε επίπεδο Θεματικής Ενότητας, Προγράμματος Σπουδών και ακαδημαϊκού έτους. Η ύπαρξη συνοπτικών αναφορών διευκολύνει την παρακολούθηση της ποιότητας διδασκαλίας, την αξιολόγηση εκπαιδευτικού υλικού και την αναγνώριση συστημικών προβλημάτων. Επιπλέον, η συνοπτική πληροφορία είναι απαραίτητη για την τεκμηρίωση διαδικασιών διασφάλισης ποιότητας, την υποβολή εκθέσεων σε εξωτερικούς φορείς και την υποστήριξη στρατηγικών αποφάσεων που αφορούν την αναβάθμιση των προγραμμάτων σπουδών.

Η επιλογή της κατάλληλης μεθόδου περίληψης αποτελεί κρίσιμο παράγοντα για την ποιότητα των παραγόμενων αποτελεσμάτων. Οι extractive μέθοδοι, οι οποίες βασίζονται στην επιλογή και αναπαραγωγή φράσεων από το αρχικό κείμενο, προσφέρουν διαφάνεια και απλότητα, αλλά συχνά αδυνατούν να αποδώσουν τη βαθύτερη σημασιολογική συνοχή των απαντήσεων ανοικτού τύπου. Αντίθετα, οι abstractive μέθοδοι, που αξιοποιούν σύγχρονα μοντέλα βαθιάς μάθησης και Μεγάλα Γλωσσικά Μοντέλα, παράγουν νέο κείμενο που αποτυπώνει με μεγαλύτερη ακρίβεια τα θεματικά μοτίβα και τις λογικές σχέσεις μεταξύ των ιδεών. Η abstractive προσέγγιση επιτρέπει τη δημιουργία περιλήψεων υψηλής συνοχής, απαλλαγμένων από περιττές λεπτομέρειες, και προσφέρει μεγαλύτερη ευελιξία στην προσαρμογή του ύφους και του επιπέδου αφαιρετικότητας. Για δεδομένα εκπαιδευτικής αξιολόγησης, όπου η ποικιλία των εκφράσεων και η ανάγκη για ενιαία θεματική σύνθεση είναι ιδιαίτερα έντονες, η abstractive περίληψη υπερτερεί σημαντικά.

Η εξαγωγική περίληψη (extractive summarization) βασίζεται στην επιλογή υπαρχουσών λέξεων, φράσεων ή προτάσεων από το αρχικό κείμενο και στη σύνθεσή τους σε μια συντομότερη μορφή. Η αφαιρετική περίληψη (abstractive summarization), αντίθετα, επιτρέπει στο μοντέλο να αναδιατυπώσει το περιεχόμενο και να δημιουργήσει νέα πρόταση που δεν αντιγράφει απαραίτητα αυτούσιο τμήμα της εισόδου (See et al., 2017)

Τέλος, η υιοθέτηση abstractive μεθόδων περίληψης ενισχύει τη δυνατότητα κλιμάκωσης και αυτοματοποίησης της διαδικασίας αξιολόγησης, καθιστώντας εφικτή την επεξεργασία μεγάλων όγκων δεδομένων με συνέπεια και ακρίβεια. Η χρήση LLMs επιτρέπει την παραγωγή περιλήψεων που συμμορφώνονται με συγκεκριμένες οδηγίες, όπως ουδετερότητα, αποφυγή ευαίσθητων στοιχείων και συγκεκριμένο εύρος λέξεων, ενώ παράλληλα μειώνει τον χρόνο και το κόστος της ανάλυσης. Η μετάβαση από χειροκίνητες ή ημι-αυτοματοποιημένες διαδικασίες σε πλήρως

Παραγωγή περιλήψεων με χρήση TN

αυτοματοποιημένα συστήματα περίληψης αποτελεί σημαντικό βήμα προς την ενίσχυση της ποιότητας και της αξιοπιστίας της εκπαιδευτικής αξιολόγησης. Η παρούσα εργασία εξετάζει συστηματικά αυτή τη μετάβαση, αναδεικνύοντας τα πλεονεκτήματα των σύγχρονων τεχνικών περίληψης και τη συμβολή τους στη βελτίωση της λήψης αποφάσεων σε θεσμικά εκπαιδευτικά περιβάλλοντα (Liu & Lapata, 2019).

1.4 Στόχοι της διπλωματικής εργασίας

Ο κύριος στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη και η συστηματική αξιολόγηση ενός ολοκληρωμένου συστήματος παραγωγής περιλήψεων για απαντήσεις ανοικτού τύπου σε ερωτηματολόγια εκπαιδευτικής αξιολόγησης. Το σύστημα αυτό στοχεύει στη μετατροπή μεγάλου όγκου μη δομημένων κειμένων σε συνεκτικές, αντιπροσωπευτικές και αξιοποιήσιμες περιλήψεις, οι οποίες μπορούν να υποστηρίξουν τη διαδικασία λήψης αποφάσεων σε θεσμικά εκπαιδευτικά πλαίσια. Η εργασία δεν περιορίζεται στην απλή εφαρμογή υπάρχοντων εργαλείων, αλλά επιδιώκει τον σχεδιασμό μιας μεθοδολογίας που να λαμβάνει υπόψη τις ιδιαιτερότητες της ελληνικής γλώσσας, τη φύση των δεδομένων εκπαιδευτικής αξιολόγησης και τις απαιτήσεις για προστασία προσωπικών δεδομένων. Ο κεντρικός αυτός στόχος εντάσσεται στο ευρύτερο πλαίσιο αξιοποίησης των Μεγάλων Γλωσσικών Μοντέλων για την υποστήριξη διαδικασιών ποιότητας στην ανώτατη εκπαίδευση.

Ένας από τους βασικούς επιμέρους στόχους είναι η συγκριτική αξιολόγηση διαφορετικών Μεγάλων Γλωσσικών Μοντέλων (LLMs) ως προς την ικανότητά τους να παράγουν υψηλής ποιότητας περιλήψεις στα ελληνικά. Η εργασία εστιάζει σε μοντέλα γενικής χρήσης, όπως GPT-4o, Gemini, Claude, Mistral και Llama, καθώς και σε πολυγλωσσικά ή εξειδικευμένα μοντέλα όπως τα mT5/T5, τα οποία μπορούν να λειτουργήσουν ως baseline. Η σύγκριση αυτή δεν είναι μόνο τεχνική, αλλά και λειτουργική: εξετάζονται παράμετροι όπως η συνοχή των παραγόμενων περιλήψεων, η πιστότητα στο αρχικό περιεχόμενο, η ικανότητα σύλληψης θεματικών μοτίβων και η συμπεριφορά των μοντέλων υπό διαφορετικές ρυθμίσεις (π.χ. θερμοκρασία, μήκος εξόδου). Μέσω αυτής της συγκριτικής προσέγγισης, η εργασία επιδιώκει να αναδείξει τα πλεονεκτήματα και τους περιορισμούς κάθε μοντέλου στο συγκεκριμένο εκπαιδευτικό πλαίσιο.

Παραγωγή περιλήψεων με χρήση TN

Ένας δεύτερος επιμέρους στόχος αφορά τη δημιουργία ενός αναπαραγωγίμου και επεκτάσιμου pipeline για την παραγωγή περιλήψεων, το οποίο θα καλύπτει όλα τα στάδια της διαδικασίας: από την προεπεξεργασία των δεδομένων έως την αποθήκευση των αποτελεσμάτων. Το pipeline αυτό περιλαμβάνει τη φόρτωση και ενοποίηση των δεδομένων από πολλαπλά αρχεία, τον καθαρισμό και την ψευδωνυμοποίηση των απαντήσεων, τον σχεδιασμό και την εφαρμογή κατάλληλων prompts στα ελληνικά, την κλήση των LLMs μέσω ενιαίου API, καθώς και την καταγραφή μεταδεδομένων για λόγους διαφάνειας και επαναληψιμότητας. Ιδιαίτερη έμφαση δίνεται στη δομή των φακέλων, στη διαχείριση των ρυθμίσεων (config files) και στην υιοθέτηση βέλτιστων πρακτικών λογισμικού, ώστε το σύστημα να μπορεί να επεκταθεί σε μελλοντικές μελέτες ή σε άλλα είδη ερωτηματολογίων.

Στο κείμενο αυτό ο όρος pipeline δηλώνει το πλήρες αυτοματοποιημένο σύνολο βημάτων της υλοποίησης και όχι απλώς ένα μεμονωμένο πρόγραμμα. Περιλαμβάνει την ανάγνωση των αρχικών αρχείων, την κανονικοποίηση και τον καθαρισμό, την ομαδοποίηση ανά πρόγραμμα σπουδών και τύπο ερώτησης, την κλήση των μοντέλων, την καταγραφή των παραγόμενων περιλήψεων και την εφαρμογή των μετρικών αξιολόγησης.

Κρίσιμος επιμέρους στόχος της εργασίας είναι η ποσοτική αξιολόγηση των παραγόμενων περιλήψεων με τη χρήση καθιερωμένων μετρικών, όπως οι ROUGE και BERTScore. Οι μετρικές αυτές επιτρέπουν τη σύγκριση των περιλήψεων που παράγονται από τα LLMs με τις «χρυσές» περιλήψεις (gold summaries) που έχουν δημιουργηθεί χειροκίνητα, παρέχοντας ένα αντικειμενικό πλαίσιο αξιολόγησης. Η ROUGE εστιάζει στην επικάλυψη n-grams και στην ευθυγράμμιση ακολουθιών, ενώ η BERTScore αξιοποιεί σημασιολογικές αναπαραστάσεις για την εκτίμηση της ομοιότητας σε επίπεδο νοήματος. Μέσω της συστηματικής εφαρμογής αυτών των μετρικών σε διαφορετικά μοντέλα και τύπους ερωτήσεων (θετικές, αρνητικές, προτάσεις βελτίωσης), η εργασία επιδιώκει να καταγράψει μετρήσιμες διαφορές στην απόδοση και να τεκμηριώσει την καταλληλότητα κάθε προσέγγισης.

Τέλος, ένας εξίσου σημαντικός επιμέρους στόχος είναι η ποιοτική αξιολόγηση των παραγόμενων περιλήψεων από ανθρώπινους αξιολογητές, με βάση κριτήρια όπως η συνοχή, η πληρότητα, η ακρίβεια, η μη παραπληροφόρηση και η καταλληλότητα του ύφους. Η ποιοτική αυτή διάσταση

Παραγωγή περιλήψεων με χρήση TN

συμπληρώνει τις ποσοτικές μετρικές, επιτρέποντας την αποτίμηση πτυχών που δεν αποτυπώνονται πλήρως σε αριθμητικούς δείκτες, όπως η αναγνωσιμότητα, η λογική ροή και η ευθυγράμμιση με τις ανάγκες των τελικών χρηστών (π.χ. ακαδημαϊκών οργάνων). Η συλλογή και ανάλυση των ποιοτικών αξιολογήσεων, ενδεχομένως με τη χρήση στατιστικών δεικτών συμφωνίας μεταξύ αξιολογητών, συμβάλλει στη διαμόρφωση μιας ολοκληρωμένης εικόνας για την απόδοση των μοντέλων. Συνδυάζοντας ποσοτική και ποιοτική αξιολόγηση, η εργασία στοχεύει να προτείνει τεκμηριωμένες κατευθύνσεις για την επιλογή και χρήση LLMs στην εκπαιδευτική αξιολόγηση.

1.5 Συνεισφορά της εργασίας

Η τεχνική συνεισφορά της παρούσας διπλωματικής εργασίας εστιάζει στην ανάπτυξη ενός ολοκληρωμένου, αναπαραγωγίμου και επεκτάσιμου pipeline για την παραγωγή περιλήψεων από απαντήσεις ανοικτού τύπου σε ερωτηματολόγια εκπαιδευτικής αξιολόγησης. Το pipeline αυτό καλύπτει όλα τα στάδια της επεξεργασίας, από την προετοιμασία και τον καθαρισμό των δεδομένων έως την παραγωγή περιλήψεων και την αξιολόγηση των αποτελεσμάτων. Η υλοποίηση περιλαμβάνει μηχανισμούς ψευδωνυμοποίησης, κανονικοποίησης κειμένου, διαχείρισης προτροπών (prompting), κλήσης μοντέλων μέσω ενιαίου API και αποθήκευσης μεταδεδομένων για λόγους διαφάνειας και επαναληψιμότητας. Επιπλέον, η εργασία εισάγει μια συστηματική διαδικασία αξιολόγησης των μοντέλων με τη χρήση καθιερωμένων μετρικών, όπως ROUGE, BLEU και BERTScore, προσφέροντας ένα πλήρως αυτοματοποιημένο πλαίσιο που μπορεί να χρησιμοποιηθεί ως βάση για μελλοντικές επεκτάσεις ή προσαρμογές.

Σε ερευνητικό επίπεδο, η εργασία συμβάλλει στη διερεύνηση της απόδοσης διαφορετικών Μεγάλων Γλωσσικών Μοντέλων (LLMs) στην ελληνική γλώσσα, ένα πεδίο που παραμένει σχετικά υποεκπροσωπούμενο στη διεθνή βιβλιογραφία. Η συγκριτική μελέτη περιλαμβάνει μοντέλα γενικής χρήσης, όπως GPT-4o, Claude, Gemini, Mistral και Llama, καθώς και πολυγλωσσικά μοντέλα όπως mT5/T5, τα οποία λειτουργούν ως baseline. Η αξιολόγηση πραγματοποιείται τόσο σε ποσοτικό επίπεδο, μέσω μετρικών ομοιότητας με «χρυσές» περιλήψεις, όσο και σε ποιοτικό επίπεδο, μέσω ανθρώπινης αποτίμησης της συνοχής, της πληρότητας και της

Παραγωγή περιλήψεων με χρήση TN

ακρίβειας των παραγόμενων περιλήψεων. Η ερευνητική συνεισφορά έγκειται στην τεκμηριωμένη αποτύπωση των δυνατοτήτων και των περιορισμών των LLMs στα ελληνικά, προσφέροντας πολύτιμα συμπεράσματα για την επιλογή κατάλληλων μοντέλων σε εφαρμογές εκπαιδευτικής αξιολόγησης.

Η εργασία προσφέρει επίσης σημαντική πρακτική συνεισφορά, καθώς το σύστημα που αναπτύσσεται μπορεί να αξιοποιηθεί άμεσα από εκπαιδευτικούς φορείς, όπως πανεπιστήμια, μονάδες διασφάλισης ποιότητας και διοικητικές υπηρεσίες. Η δυνατότητα αυτόματης παραγωγής περιλήψεων μειώνει δραστικά τον χρόνο και το ανθρώπινο κόστος που απαιτείται για την ανάλυση μεγάλου όγκου απαντήσεων ανοικτού τύπου, ενώ παράλληλα ενισχύει τη συνέπεια και την αντικειμενικότητα των αποτελεσμάτων. Το εργαλείο μπορεί να ενσωματωθεί σε υπάρχουσες διαδικασίες αξιολόγησης, να υποστηρίξει την παραγωγή ετήσιων αναφορών και να συμβάλει στη λήψη τεκμηριωμένων αποφάσεων σε επίπεδο Θεματικής Ενότητας, Προγράμματος Σπουδών ή ιδρύματος.

Επιπλέον, η εργασία συμβάλλει στη διαμόρφωση ενός προτύπου για την υπεύθυνη και ασφαλή χρήση LLMs σε ευαίσθητα δεδομένα εκπαιδευτικής αξιολόγησης. Η ενσωμάτωση πρακτικών όπως η ψευδωνυμοποίηση, η αφαίρεση ευαίσθητων στοιχείων, η χρήση παρόχων χωρίς διατήρηση δεδομένων και η τοπική αποθήκευση των αποτελεσμάτων ενισχύει τη συμμόρφωση με τις αρχές προστασίας προσωπικών δεδομένων. Με αυτόν τον τρόπο, η εργασία δεν προσφέρει μόνο τεχνικές λύσεις, αλλά και ένα πλαίσιο δεοντολογικής χρήσης της τεχνητής νοημοσύνης σε πραγματικά εκπαιδευτικά περιβάλλοντα.

Τέλος, η συνεισφορά της εργασίας επεκτείνεται και στη δημιουργία μιας μεθοδολογίας που μπορεί να αξιοποιηθεί σε άλλους τομείς όπου απαιτείται ανάλυση απαντήσεων ανοικτού τύπου, όπως έρευνες ικανοποίησης, αξιολογήσεις υπηρεσιών, κοινωνικές επιστήμες και δημόσια διοίκηση. Η γενικευσιμότητα του pipeline, η δυνατότητα προσαρμογής των προτροπών και η ευελιξία στην επιλογή μοντέλων καθιστούν το σύστημα κατάλληλο για εφαρμογές πέρα από το εκπαιδευτικό πλαίσιο. Η εργασία, επομένως, συμβάλλει τόσο στην ακαδημαϊκή γνώση όσο και στην πρακτική αξιοποίηση της τεχνητής νοημοσύνης, προσφέροντας ένα εργαλείο υψηλής χρησιμότητας και μια

Παραγωγή περιλήψεων με χρήση TN

τεκμηριωμένη προσέγγιση που μπορεί να αποτελέσει σημείο αναφοράς για μελλοντικές έρευνες και εφαρμογές.

1.6 Μεθοδολογική προσέγγιση (σύντομη επισκόπηση)

Η μεθοδολογική προσέγγιση της παρούσας διπλωματικής εργασίας δομείται ως μια αυστηρά οριοθετημένη, συστηματική και αναπαραγώγιμη διαδικασία, η οποία εκτείνεται από τη συλλογή και προετοιμασία των δεδομένων έως την παραγωγή και πολυεπίπεδη αξιολόγηση των παραγόμενων περιλήψεων. Στο πρώτο στάδιο, η συλλογή δεδομένων εστιάζει στις απαντήσεις ανοικτού τύπου που προέρχονται από ερωτηματολόγια εκπαιδευτικής αξιολόγησης, τα οποία έχουν συμπληρωθεί από φοιτητές σε διαφορετικά Προγράμματα Σπουδών και Θεματικές Ενότητες. Τα δεδομένα αντλούνται από ετερογενείς πηγές (π.χ. αρχεία Excel) και ενοποιούνται σε ένα ενιαίο σχήμα, το οποίο περιλαμβάνει ψευδωνυμοποιημένα αναγνωριστικά, κατηγοριοποίηση των ερωτήσεων (θετικές εμπειρίες, αρνητικά σημεία, προτάσεις βελτίωσης), καθώς και τις αντίστοιχες «χρυσές» περιλήψεις, όπου αυτές είναι διαθέσιμες. Η φάση αυτή συνοδεύεται από ελέγχους πληρότητας, συνέπειας και βασικών στατιστικών χαρακτηριστικών, ώστε να διασφαλιστεί ότι το τελικό σύνολο δεδομένων είναι κατάλληλο για πειραματική χρήση και συγκριτική αξιολόγηση.

Το δεύτερο στάδιο αφορά τον καθαρισμό και την προεπεξεργασία των κειμένων, διαδικασία η οποία είναι καθοριστική για την ποιότητα των αποτελεσμάτων που θα παραχθούν. Στο πλαίσιο αυτό εφαρμόζονται τεχνικές κανονικοποίησης της γλώσσας (όπως ενοποίηση πολλαπλών κενών, διατήρηση ήπιας μορφολογικής ποικιλίας, διατήρηση τονισμού), απομάκρυνσης θορύβου (π.χ. URLs, διευθύνσεις ηλεκτρονικού ταχυδρομείου, αριθμοί τηλεφώνου, χρονικές σφραγίδες) και ψευδωνυμοποίησης ή απόκρυψης ευαίσθητων στοιχείων. Όπου κρίνεται σκόπιμο, μπορεί να εφαρμοστεί ελαφρά ορθογραφική διόρθωση, με αυστηρή μέριμνα ώστε να μην αλλοιωθεί το σημασιολογικό περιεχόμενο των απαντήσεων. Το αποτέλεσμα είναι ένα καθαρό, δομημένο και ενοποιημένο σύνολο κειμένων, το οποίο μπορεί να τροφοδοτήσει αξιόπιστα τα μοντέλα παραγωγής περιλήψεων. Η διαδικασία τεκμηριώνεται λεπτομερώς, ώστε να είναι δυνατή η επανάληψη ή η προσαρμογή της σε μελλοντικές μελέτες.

Παραγωγή περιλήψεων με χρήση TN

Το τρίτο στάδιο της μεθοδολογίας αφορά την ίδια την παραγωγή περιλήψεων, η οποία υλοποιείται μέσω ενός αυτοματοποιημένου pipeline που αξιοποιεί Μεγάλα Γλωσσικά Μοντέλα. Για κάθε συνδυασμό Προγράμματος Σπουδών και τύπου ερώτησης, οι απαντήσεις ανοικτού τύπου συνενώνονται ή ομαδοποιούνται με ελεγχόμενο τρόπο, ώστε να δημιουργηθεί μια αντιπροσωπευτική είσοδος προς το μοντέλο, χωρίς να παραβιάζονται περιορισμοί μήκους ή να χάνεται η θεματική ποικιλία. Στη συνέχεια, εφαρμόζονται προσεκτικά σχεδιασμένες προτροπές (prompts) στα ελληνικά, οι οποίες καθορίζουν το ζητούμενο ύφος, το εύρος λέξεων, τον βαθμό αφαιρετικότητας και τις θεματικές προτεραιότητες της περίληψης (π.χ. έμφαση σε θετικά στοιχεία, ανάδειξη προβλημάτων, ομαδοποίηση προτάσεων βελτίωσης). Τα επιλεγμένα LLMs καλούνται μέσω ενιαίου API, με ελεγχόμενες παραμέτρους (όπως θερμοκρασία, μέγιστος αριθμός tokens, seed όπου υποστηρίζεται), και οι παραγόμενες περιλήψεις αποθηκεύονται μαζί με τα αντίστοιχα μεταδεδομένα (μοντέλο, έκδοση, χρόνος δημιουργίας, ρυθμίσεις).

Το τέταρτο στάδιο αφορά την αξιολόγηση των παραγόμενων περιλήψεων, η οποία πραγματοποιείται σε δύο συμπληρωματικά επίπεδα: ποσοτικό και ποιοτικό. Σε ποσοτικό επίπεδο, οι περιλήψεις συγκρίνονται με τις «χρυσές» περιλήψεις μέσω καθιερωμένων μετρικών, όπως ROUGE-1, ROUGE-2, ROUGE-L και BERTScore, οι οποίες επιτρέπουν την εκτίμηση της λεξιλογικής επικάλυψης και της σημασιολογικής εγγύτητας αντίστοιχα. Τα αποτελέσματα συγκεντρώνονται σε πίνακες και οπτικοποιούνται με γραφήματα (π.χ. ραβδογράμματα, boxplots), ώστε να αναδειχθούν οι διαφορές μεταξύ μοντέλων, τύπων ερωτήσεων και ρυθμίσεων. Σε ποιοτικό επίπεδο, επιλεγμένο υποσύνολο περιλήψεων αξιολογείται από ανθρώπινους κριτές με βάση κριτήρια όπως συνοχή, πληρότητα, ακρίβεια, μη παραπληροφόρηση και καταλληλότητα για χρήση σε θεσμικές αναφορές. Η συνδυαστική ανάλυση των δύο επιπέδων αξιολόγησης επιτρέπει τη διατύπωση τεκμηριωμένων συμπερασμάτων για την καταλληλότητα και τις δυνατότητες των LLMs στο συγκεκριμένο πλαίσιο.

Κεντρικό στοιχείο της μεθοδολογικής προσέγγισης αποτελεί η επιλογή των μοντέλων και των κριτηρίων αξιολόγησης. Η επιλογή των LLMs βασίζεται σε κριτήρια όπως η υποστήριξη της ελληνικής γλώσσας, η διαθεσιμότητα και ωριμότητα των APIs, η σταθερότητα της απόδοσης, η δυνατότητα ελέγχου παραμέτρων (π.χ. θερμοκρασία, μήκος εξόδου), καθώς και το κόστος χρήσης

Παραγωγή περιλήψεων με χρήση TN

σε συνθήκες κλιμάκωσης. Παράλληλα, τα κριτήρια αξιολόγησης διαμορφώνονται με τρόπο ώστε να αντανακλούν τόσο την τεχνική ποιότητα των περιλήψεων (μέσω ποσοτικών μετρικών) όσο και τη χρηστικότητά τους σε πραγματικά θεσμικά πλαίσια (μέσω ποιοτικής αποτίμησης). Η συνολική μεθοδολογία σχεδιάζεται ώστε να είναι διαφανής, επαναλήψιμη και επεκτάσιμη, επιτρέποντας την ενσωμάτωση νέων μοντέλων, μετρικών ή τύπων δεδομένων σε μελλοντικές εργασίες.

1.7 Δομή της διπλωματικής εργασίας

Η διπλωματική εργασία διαρθρώνεται σε τρία κεφάλαια, με αλληλουχία που υπηρετεί τη σταδιακή ανάπτυξη του ερευνητικού αντικειμένου, από την εισαγωγή του προβλήματος και τη θεωρητική θεμελίωση έως τη μεθοδολογική εφαρμογή, την αξιολόγηση των αποτελεσμάτων και τη διατύπωση των τελικών συμπερασμάτων. Το Κεφάλαιο 1 εισάγει το αντικείμενο και το πλαίσιο της εργασίας. Εστιάζει στο πρόβλημα της επεξεργασίας απαντήσεων ανοικτού τύπου σε ερωτηματολόγια αξιολόγησης εκπαιδευτικών διαδικασιών και αναδεικνύει τη σημασία της αυτόματης περίληψης για την ανάλυση ποιοτικών εκπαιδευτικών δεδομένων. Παράλληλα, προσδιορίζει τους στόχους, τη συνεισφορά και τη συνοπτική μεθοδολογική κατεύθυνση της έρευνας.

Το Κεφάλαιο 2 συγκροτεί το θεωρητικό υπόβαθρο της μελέτης. Αναπτύσσει βασικές έννοιες της Επεξεργασίας Φυσικής Γλώσσας, εξετάζει τεχνικές περίληψης κειμένου και αποσαφηνίζει τη διάκριση μεταξύ extractive και abstractive προσεγγίσεων. Επιπλέον, παρουσιάζει τη λειτουργία και τους περιορισμούς των Μεγάλων Γλωσσικών Μοντέλων, καθώς και τις μεθόδους αξιολόγησης συστημάτων περίληψης, με έμφαση στις ιδιαιτερότητες των ελληνόγλωσσων δεδομένων και των απαντήσεων ανοικτού τύπου στην εκπαιδευτική αξιολόγηση.

Το Κεφάλαιο 3 παρουσιάζει τη μεθοδολογία και την υλοποίηση του προτεινόμενου συστήματος. Περιλαμβάνει την περιγραφή και κατασκευή του συνόλου δεδομένων, την αρχιτεκτονική του συστήματος, την προεπεξεργασία των δεδομένων, την ενσωμάτωση των γλωσσικών μοντέλων, τη διαδικασία παραγωγής περιλήψεων και το πλαίσιο αξιολόγησης. Στη συνέχεια, αναλύει τα αποτελέσματα ανά μοντέλο, κατηγορία ερώτησης και πρόγραμμα σπουδών, ερμηνεύει τα

Παραγωγή περιλήψεων με χρήση TN

ευρήματα, καταγράφει τη δομή των βασικών αρχείων κώδικα και ολοκληρώνει την εργασία με τα κύρια συμπεράσματα που προκύπτουν από την ερευνητική διαδικασία.

Κεφάλαιο 2 – Θεωρητικό Υπόβαθρο

2.1 Επεξεργασία Φυσικής Γλώσσας (NLP)

2.1.1 Ορισμός και ιστορική εξέλιξη του NLP

Η Επεξεργασία Φυσικής Γλώσσας ορίζεται ως ο κλάδος της Τεχνητής Νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων και μοντέλων ικανών να κατανοούν, να αναλύουν, να ερμηνεύουν και να παράγουν ανθρώπινη γλώσσα. Η γλώσσα, ως σύνθετο γνωστικό και κοινωνικό φαινόμενο, παρουσιάζει σημαντικές προκλήσεις για την υπολογιστική επεξεργασία, καθώς χαρακτηρίζεται από αμφισημία, πολυσημία, συμφραζόμενα, πραγματολογικές αποχρώσεις και πολιτισμικές διαφοροποιήσεις. Η ιστορική εξέλιξη του NLP αντανακλά την προσπάθεια της επιστημονικής κοινότητας να αντιμετωπίσει αυτές τις προκλήσεις μέσα από διαφορετικά παραδείγματα, τεχνολογίες και θεωρητικές προσεγγίσεις.

Η πρώτη περίοδος του NLP (1950–1980) χαρακτηρίζεται από κανονιστικές, συμβολικές και γλωσσοκεντρικές προσεγγίσεις. Συστήματα όπως οι πρώτοι μεταφραστές κανόνων βασίζονταν σε χειροποίητες γραμματικές, λεξικά και συντακτικά δέντρα. Η προσέγγιση αυτή, αν και θεωρητικά ισχυρή, αποδείχθηκε δύσκολα επεκτάσιμη, καθώς η γλώσσα δεν μπορεί να περιγραφεί πλήρως μέσω στατικών κανόνων.

Η δεύτερη περίοδος (1980–2010) σηματοδοτεί την είσοδο των στατιστικών μεθόδων και της μηχανικής μάθησης. Η διαθεσιμότητα μεγάλων σωμάτων κειμένου (corpora) και η ανάπτυξη αλγορίθμων όπως οι Hidden Markov Models (HMMs) και τα n-gram μοντέλα επέτρεψαν την εκμάθηση γλωσσικών μοτίβων από δεδομένα. Η περίοδος αυτή έθεσε τα θεμέλια για την αυτοματοποιημένη ανάλυση κειμένου, αλλά παρέμενε περιορισμένη ως προς τη σημασιολογική κατανόηση.

Παραγωγή περιλήψεων με χρήση TN

Η τρίτη περίοδος (2010–σήμερα) χαρακτηρίζεται από την κυριαρχία των νευρωνικών δικτύων και ιδιαίτερα των μοντέλων βαθιάς μάθησης. Η εισαγωγή των word embeddings (Word2Vec, GloVe) και αργότερα των contextual embeddings (BERT, GPT, RoBERTa) επέτρεψε την αναπαράσταση της γλώσσας με τρόπους που αποτυπώνουν σημασιολογικές και συντακτικές σχέσεις. Η εμφάνιση της αρχιτεκτονικής Transformer (Vaswani et al., 2017b) αποτέλεσε σημείο καμπής, οδηγώντας στην ανάπτυξη Μεγάλων Γλωσσικών Μοντέλων που μπορούν να εκτελούν πολύπλοκες γλωσσικές εργασίες με ανθρώπινη ποιότητα.

Σήμερα, το NLP βρίσκεται σε μια νέα εποχή, όπου τα LLMs έχουν τη δυνατότητα να παράγουν συνεκτικό κείμενο, να συνοψίζουν, να μεταφράζουν, να απαντούν σε ερωτήσεις και να εκτελούν σύνθετες γνωστικές λειτουργίες. Η παρούσα εργασία αξιοποιεί αυτές τις εξελίξεις για την παραγωγή περιλήψεων ελληνόγλωσσων απαντήσεων ανοικτού τύπου, ένα πεδίο που μέχρι πρόσφατα ήταν δύσκολο να αντιμετωπιστεί με παραδοσιακές μεθόδους.

2.1.2 Βασικές αρχές και προκλήσεις στην επεξεργασία κειμένου

Η επεξεργασία φυσικής γλώσσας στηρίζεται σε μια σειρά από θεμελιώδεις αρχές που καθορίζουν τον τρόπο με τον οποίο τα συστήματα αναλύουν και κατανοούν το κείμενο. Μία από τις βασικότερες αρχές είναι ότι η γλώσσα είναι εγγενώς αμφίσημη. Λέξεις και φράσεις μπορούν να έχουν πολλαπλές σημασίες ανάλογα με τα συμφραζόμενα, γεγονός που καθιστά δύσκολη την αυτόματη ερμηνεία. Για παράδειγμα, η λέξη «πρόγραμμα» μπορεί να αναφέρεται σε πρόγραμμα σπουδών, σε πρόγραμμα υπολογιστή ή σε σχέδιο δράσης. Η αποσαφήνιση της σημασίας απαιτεί κατανόηση του ευρύτερου περιβάλλοντος.

Μια δεύτερη αρχή αφορά τη σημασία της συντακτικής και σημασιολογικής δομής. Η γλώσσα δεν είναι απλώς μια ακολουθία λέξεων, αλλά μια ιεραρχική δομή που περιλαμβάνει φράσεις, προτάσεις και κείμενα. Η κατανόηση αυτής της δομής είναι κρίσιμη για εργασίες όπως η περίληψη, η μετάφραση και η απάντηση σε ερωτήσεις. Τα σύγχρονα μοντέλα NLP επιχειρούν να αποτυπώσουν αυτές τις σχέσεις μέσω μηχανισμών attention και contextual embeddings.

Παραγωγή περιλήψεων με χρήση TN

Οι προκλήσεις στην επεξεργασία κειμένου είναι πολυδιάστατες. Η γλωσσική ποικιλία, οι ιδιωματισμοί, οι ανορθογραφίες, η ασυνέπεια στη χρήση σημείων στίξης και η παρουσία θορύβου (όπως URLs ή άσχετες πληροφορίες) δυσχεραίνουν την ανάλυση. Επιπλέον, η ελληνική γλώσσα παρουσιάζει ιδιαίτερες μορφολογικές και συντακτικές ιδιαιτερότητες, όπως πλούσια κλίση, ελεύθερη σειρά λέξεων και πολυμορφία τύπων, οι οποίες απαιτούν εξειδικευμένες τεχνικές.

Στο πλαίσιο της παρούσας εργασίας, οι προκλήσεις αυτές είναι ιδιαίτερα εμφανείς στις απαντήσεις ανοικτού τύπου, όπου οι φοιτητές χρησιμοποιούν ελεύθερο λόγο, συχνά με συναισθηματική φόρτιση, ασυνέπεια ύφους και ποικιλία εκφράσεων. Η αντιμετώπιση αυτών των προκλήσεων απαιτεί συνδυασμό προσεκτικής προεπεξεργασίας και ισχυρών μοντέλων κατανόησης γλώσσας.

2.1.3 Προεπεξεργασία κειμένου: τμηματοποίηση κειμένου (tokenization), κανονικοποίηση (normalization), λέξεις-κενά (stopwords), λημματοποίηση (lemmatization)

Η προεπεξεργασία κειμένου αποτελεί ένα από τα πιο κρίσιμα στάδια στην Επεξεργασία Φυσικής Γλώσσας, καθώς καθορίζει την ποιότητα των δεδομένων που θα εισαχθούν στα μοντέλα και επηρεάζει άμεσα την απόδοση των αλγορίθμων. Η φύση της ανθρώπινης γλώσσας, με την πολυπλοκότητα, την ασυνέπεια και την ποικιλία της, καθιστά απαραίτητη την εφαρμογή μιας σειράς τεχνικών που στοχεύουν στη μετατροπή του ακατέργαστου κειμένου σε μια πιο καθαρή, ομοιογενή και αναλύσιμη μορφή. Στο πλαίσιο της παρούσας εργασίας, η προεπεξεργασία αποκτά ιδιαίτερη σημασία, καθώς οι απαντήσεις ανοικτού τύπου των φοιτητών περιλαμβάνουν συχνά ανορθογραφίες, ιδιωματισμούς, συναισθηματικά φορτισμένες εκφράσεις, ασυνέπεια στη χρήση σημείων στίξης και ποικιλία ύφους. Η προσεκτική και συστηματική προεπεξεργασία συμβάλλει στη μείωση του θορύβου και στη βελτίωση της ποιότητας των περιλήψεων που παράγουν τα LLMs.

Το πρώτο βήμα της προεπεξεργασίας είναι το τμηματοποίηση κειμένου (tokenization), δηλαδή ο διαχωρισμός του κειμένου σε μικρότερες μονάδες, όπως λέξεις, υπολέξεις ή χαρακτήρες. Στα σύγχρονα μοντέλα, το τμηματοποίηση κειμένου (tokenization) βασίζεται σε τεχνικές όπως το

Παραγωγή περιλήψεων με χρήση TN

Byte-Pair Encoding (BPE) ή το WordPiece, οι οποίες επιτρέπουν την αποτελεσματική αναπαράσταση λέξεων που δεν υπάρχουν στο λεξιλόγιο του μοντέλου. Η ελληνική γλώσσα, με την πλούσια μορφολογία της, επωφελείται ιδιαίτερα από αυτές τις τεχνικές, καθώς επιτρέπουν την αποσύνθεση σύνθετων ή σπάνιων λέξεων σε υπομονάδες που μπορούν να αναλυθούν και να συνδυαστούν εκ νέου από το μοντέλο.

Το δεύτερο βήμα είναι το κανονικοποίηση (normalization), το οποίο περιλαμβάνει διαδικασίες όπως η ενοποίηση πολλαπλών κενών, η διόρθωση ασυνέπειας στη χρήση κεφαλαίων/πεζών, η απομάκρυνση ειδικών χαρακτήρων και η διατήρηση ήπιας μορφολογικής ποικιλίας. Στην ελληνική γλώσσα, η διατήρηση του τονισμού είναι κρίσιμη, καθώς ο τόνος διαφοροποιεί σημασίες (π.χ. «πότε» vs «ποτε»). Το κανονικοποίηση (normalization) συμβάλλει στη μείωση της ποικιλίας που δεν προσφέρει σημασιολογική πληροφορία και διευκολύνει την ανάλυση.

Η απομάκρυνση λέξεις-κενά (stopwords) αποτελεί μια ακόμη συνηθισμένη τεχνική, αν και η χρήση της πρέπει να γίνεται με προσοχή. Οι λέξεις αυτές, όπως «και», «να», «που», εμφανίζονται συχνά και δεν προσφέρουν σημαντική πληροφορία σε ορισμένες εργασίες. Ωστόσο, σε εργασίες περίληψης, η αφαίρεσή τους μπορεί να αλλοιώσει τη δομή του κειμένου. Για τον λόγο αυτό, η παρούσα εργασία υιοθετεί μια ήπια προσέγγιση, όπου τα λέξεις-κενά (stopwords) διατηρούνται κατά την παραγωγή περιλήψεων, αλλά μπορεί να αφαιρούνται σε στάδια στατιστικής ανάλυσης.

Τέλος, η λημματοποίηση (lemmatization) στοχεύει στη μετατροπή των λέξεων στη βασική τους μορφή (λήμμα), μειώνοντας τη μορφολογική ποικιλία. Η ελληνική γλώσσα, με την πλούσια κλίση της, επωφελείται ιδιαίτερα από αυτή τη διαδικασία. Ωστόσο, η λημματοποίηση (lemmatization) δεν εφαρμόζεται πριν την είσοδο στα LLMs, καθώς τα σύγχρονα μοντέλα είναι σχεδιασμένα να αξιοποιούν τη μορφολογική πληροφορία. Αντίθετα, χρησιμοποιείται σε στάδια ανάλυσης ή clustering, όπου η ομογενοποίηση των μορφών μπορεί να βελτιώσει την ομαδοποίηση των θεμάτων.

Παραγωγή περιλήψεων με χρήση TN

2.1.4 Αναπαράσταση κειμένου: Bag-of-Words, TF-IDF, word embeddings

Η αναπαράσταση κειμένου αποτελεί θεμελιώδες ζήτημα στο NLP, καθώς καθορίζει τον τρόπο με τον οποίο η γλώσσα μετατρέπεται σε αριθμητική μορφή που μπορεί να επεξεργαστεί ένας υπολογιστής. Οι πρώτες μέθοδοι αναπαράστασης βασίζονταν σε απλές στατιστικές προσεγγίσεις, οι οποίες, αν και περιορισμένες, έθεσαν τα θεμέλια για τις σύγχρονες τεχνικές.

Η μέθοδος **Bag-of-Words (BoW)** αποτελεί μια από τις παλαιότερες και πιο διαδεδομένες τεχνικές. Στο BoW, κάθε κείμενο αναπαρίσταται ως ένα σύνολο λέξεων, χωρίς να λαμβάνεται υπόψη η σειρά τους. Η μέθοδος αυτή είναι απλή και αποτελεσματική για βασικές εργασίες, αλλά αδυνατεί να αποτυπώσει τη σημασιολογική σχέση μεταξύ των λέξεων ή τη συντακτική δομή του κειμένου. Παρά τους περιορισμούς της, η BoW παραμένει χρήσιμη σε εργασίες όπου η συχνότητα εμφάνισης λέξεων αποτελεί σημαντικό δείκτη.

Η τεχνική **TF-IDF (Term Frequency–Inverse Document Frequency)** βελτιώνει το BoW, σταθμίζοντας τη σημασία κάθε λέξης με βάση τη συχνότητά της στο κείμενο και τη σπανιότητά της στο σύνολο των κειμένων. Με αυτόν τον τρόπο, λέξεις που εμφανίζονται συχνά σε όλα τα κείμενα (π.χ. «μάθημα», «φοιτητής») λαμβάνουν μικρότερη βαρύτητα, ενώ πιο εξειδικευμένες λέξεις αποκτούν μεγαλύτερη σημασία. Η TF-IDF χρησιμοποιείται ευρέως σε εργασίες ταξινόμησης και ανάκτησης πληροφορίας, αλλά εξακολουθεί να μην αποτυπώνει τη σημασιολογική σχέση μεταξύ των λέξεων (Manning et al., 2008).

Η εμφάνιση των **word embeddings** αποτέλεσε σημαντική πρόοδο στην αναπαράσταση κειμένου. Μοντέλα όπως τα Word2Vec και GloVe επιτρέπουν την αναπαράσταση λέξεων ως πυκνά διανύσματα σε έναν πολυδιάστατο χώρο, όπου οι σημασιολογικά παρόμοιες λέξεις βρίσκονται κοντά μεταξύ τους. Τα embeddings αυτά μαθαίνουν από μεγάλα σώματα κειμένου και αποτυπώνουν σημασιολογικές σχέσεις, όπως αναλογίες (π.χ. «βασιλιάς» – «άνδρας» + «γυναίκα» $\approx$ «βασίλισσα»). Η ελληνική γλώσσα έχει επωφεληθεί από την ανάπτυξη εξειδικευμένων embeddings, όπως τα Greek Word2Vec και FastText, τα οποία λαμβάνουν υπόψη τη μορφολογική πολυπλοκότητα (Mikolov et al., 2013).

Παραγωγή περιλήψεων με χρήση TN

Ωστόσο, τα word embeddings έχουν έναν σημαντικό περιορισμό: κάθε λέξη έχει μία μόνο αναπαράσταση, ανεξάρτητα από τα συμφραζόμενα. Αυτό σημαίνει ότι η λέξη «πρόγραμμα» έχει την ίδια αναπαράσταση είτε αναφέρεται σε πρόγραμμα σπουδών είτε σε πρόγραμμα υπολογιστή. Η ανάγκη για συμφραζόμενη αναπαράσταση οδήγησε στην ανάπτυξη των contextual embeddings.

2.1.5 Σύγχρονες αναπαραστάσεις: contextual embeddings (BERT, mBERT, Sentence-BERT)

Η εισαγωγή των contextual embeddings αποτέλεσε επανάσταση στο NLP, καθώς επέτρεψε την αναπαράσταση των λέξεων με βάση τα συμφραζόμενα στα οποία εμφανίζονται. Σε αντίθεση με τα στατικά embeddings, τα contextual embeddings παράγουν διαφορετικό διάνυσμα για την ίδια λέξη ανάλογα με το περιβάλλον της, επιτρέποντας στα μοντέλα να κατανοούν αμφισημίες και σημασιολογικές αποχρώσεις.

Το μοντέλο **BERT (Bidirectional Encoder Representations from Transformers)** αποτέλεσε σημείο καμπής, εισάγοντας μια αρχιτεκτονική που επιτρέπει την αμφίδρομη κατανόηση της γλώσσας. Το BERT εκπαιδεύεται μέσω δύο βασικών στόχων: masked language modeling και next sentence prediction. Η αμφίδρομη φύση του επιτρέπει την κατανόηση των συμφραζομένων τόσο πριν όσο και μετά από μια λέξη, κάτι που το καθιστά εξαιρετικά αποτελεσματικό σε εργασίες κατανόησης κειμένου (Devlin et al., 2019).

Το **multilingual BERT (mBERT)** επεκτείνει το BERT σε περισσότερες από 100 γλώσσες, συμπεριλαμβανομένων των ελληνικών. Αν και δεν είναι εξειδικευμένο για κάθε γλώσσα, έχει αποδειχθεί ιδιαίτερα αποτελεσματικό σε πολυγλωσσικές εργασίες και σε γλώσσες με περιορισμένα δεδομένα. Η χρήση του mBERT στην παρούσα εργασία είναι σημαντική, καθώς επιτρέπει την αξιοποίηση ενός ισχυρού μοντέλου χωρίς την ανάγκη εκπαίδευσης από την αρχή.

Το **Sentence-BERT (SBERT)** αποτελεί μια παραλλαγή του BERT που έχει σχεδιαστεί ειδικά για την παραγωγή αναπαραστάσεων προτάσεων. Μέσω της χρήσης siamese και triplet networks, το SBERT παράγει embeddings που μπορούν να συγκριθούν μέσω μετρικών όπως η κοσίνεια

Παραγωγή περιλήψεων με χρήση TN

απόσταση, καθιστώντας το ιδανικό για εργασίες όπως η ομαδοποίηση απαντήσεων, η αναζήτηση σημασιολογικής ομοιότητας και η αξιολόγηση περιλήψεων. Στο πλαίσιο της παρούσας εργασίας, τα contextual embeddings αποτελούν κρίσιμο εργαλείο για την κατανόηση των θεματικών μοτίβων στις απαντήσεις ανοικτού τύπου και για την αξιολόγηση της ποιότητας των παραγόμενων περιλήψεων (Reimers & Gurevych, 2019) .

2.2 Τεχνικές Περίληψης Κειμένου

Η περίληψη κειμένου αποτελεί μια από τις πιο σημαντικές και απαιτητικές εργασίες στην Επεξεργασία Φυσικής Γλώσσας, καθώς στοχεύει στη συμπύκνωση εκτεταμένου περιεχομένου σε μια συντομότερη, συνεκτική και αντιπροσωπευτική μορφή, χωρίς απώλεια των βασικών νοηματικών αξόνων. Η ανάγκη για αυτοματοποιημένη περίληψη έχει αυξηθεί σημαντικά με την εκθετική αύξηση των ψηφιακών δεδομένων, καθιστώντας τις τεχνικές περίληψης κρίσιμες για εφαρμογές όπως η δημοσιογραφία, η νομική ανάλυση, η ιατρική τεκμηρίωση, η επιχειρησιακή πληροφόρηση και, όπως στην παρούσα εργασία, η εκπαιδευτική αξιολόγηση. Οι σύγχρονες προσεγγίσεις στην περίληψη κειμένου διακρίνονται σε δύο βασικές κατηγορίες: τις extractive και τις abstractive μεθόδους. Η ενότητα αυτή παρουσιάζει τη θεωρητική διάκριση των δύο προσεγγίσεων, αναλύει τις κυριότερες τεχνικές που έχουν αναπτυχθεί, εξετάζει τα πλεονεκτήματα και τους περιορισμούς τους και αναδεικνύει τη σημασία τους σε πολυγλωσσικά περιβάλλοντα.

2.2.1 Θεωρητική διάκριση extractive και abstractive περίληψης

Η βασική διάκριση στις τεχνικές περίληψης αφορά τον τρόπο με τον οποίο το σύστημα επιλέγει και συνθέτει το περιεχόμενο της περίληψης. Οι **extractive μέθοδοι** βασίζονται στην επιλογή και αναπαραγωγή προτάσεων ή φράσεων από το αρχικό κείμενο, χωρίς να δημιουργούν νέο κείμενο. Η περίληψη προκύπτει από την κατάταξη των προτάσεων με βάση τη σημασία τους και την επιλογή των πιο αντιπροσωπευτικών. Η προσέγγιση αυτή είναι απλή, διαφανής και συχνά

Παραγωγή περιλήψεων με χρήση TN

αποτελεσματική, αλλά περιορίζεται από τη δομή του αρχικού κειμένου και δεν μπορεί να παράγει πραγματικά αφαιρετικό λόγο.

Αντίθετα, οι **abstractive μέθοδοι** επιχειρούν να δημιουργήσουν νέο κείμενο που αποδίδει το νόημα του αρχικού περιεχομένου, χρησιμοποιώντας διαφορετικές διατυπώσεις, συνθέτοντας πληροφορίες και αναδιατυπώνοντας τις ιδέες. Η abstractive περίληψη προσομοιάζει τον τρόπο με τον οποίο συνοψίζει ένας άνθρωπος, καθώς απαιτεί κατανόηση των θεματικών μοτίβων, των λογικών σχέσεων και της σημασιολογικής συνοχής. Η ανάπτυξη τέτοιων μεθόδων κατέστη δυνατή κυρίως χάρη στην πρόοδο των νευρωνικών δικτύων και των Μεγάλων Γλωσσικών Μοντέλων.

Η διάκριση μεταξύ extractive και abstractive περίληψης δεν είναι μόνο τεχνική, αλλά και εννοιολογική. Οι extractive μέθοδοι εστιάζουν στη δομή του κειμένου, ενώ οι abstractive εστιάζουν στο νόημα. Στο πλαίσιο της παρούσας εργασίας, η abstractive προσέγγιση είναι ιδιαίτερα σημαντική, καθώς οι απαντήσεις ανοικτού τύπου των φοιτητών είναι συχνά ασύντακτες, ανομοιογενείς και απαιτούν σύνθεση πολλαπλών θεμάτων σε ενιαίο κείμενο.

2.2.2 Extractive μέθοδοι: TextRank, LexRank, στατιστικές και γραφοθεωρητικές προσεγγίσεις

Οι extractive μέθοδοι περίληψης βασίζονται στην ιδέα ότι ορισμένες προτάσεις ενός κειμένου είναι πιο σημαντικές από άλλες και μπορούν να επιλεγούν για να σχηματίσουν μια συνοπτική εκδοχή του αρχικού περιεχομένου. Οι μέθοδοι αυτές χρησιμοποιούν στατιστικά χαρακτηριστικά, μέτρα ομοιότητας και γραφοθεωρητικές τεχνικές για να εντοπίσουν τις πιο αντιπροσωπευτικές προτάσεις.

Μία από τις πιο γνωστές μεθόδους είναι το **TextRank**, το οποίο βασίζεται στον αλγόριθμο PageRank. Το TextRank αναπαριστά τις προτάσεις ενός κειμένου ως κόμβους σε έναν γράφο, όπου οι ακμές αντιστοιχούν στη σημασιολογική ομοιότητα μεταξύ των προτάσεων. Η σημασία κάθε πρότασης υπολογίζεται με βάση τη συνδεσιμότητά της, και οι προτάσεις με τη μεγαλύτερη

Παραγωγή περιλήψεων με χρήση TN

βαθμολογία επιλέγονται για την περίληψη. Το TextRank είναι ιδιαίτερα αποτελεσματικό σε κείμενα με σαφή θεματική δομή και έχει χρησιμοποιηθεί ευρέως σε δημοσιογραφικά και επιστημονικά κείμενα (Mihalcea & Tarau, 2004) .

Παρόμοια προσέγγιση ακολουθεί το **LexRank**, το οποίο χρησιμοποιεί κοσίνεια ομοιότητα μεταξύ των προτάσεων και εφαρμόζει thresholding για τη δημιουργία του γράφου. Το LexRank έχει αποδειχθεί αποτελεσματικό σε μεγάλα σώματα κειμένων και σε περιβάλλοντα όπου η θεματική συνοχή είναι υψηλή. Οι στατιστικές μέθοδοι, όπως η συχνότητα λέξεων, η θέση της πρότασης στο κείμενο και η παρουσία λέξεων-κλειδιών, χρησιμοποιούνται συχνά σε συνδυασμό με γραφοθεωρητικές τεχνικές για τη βελτίωση της ακρίβειας (Erkan & Radev, 2004).

Ωστόσο, οι extractive μέθοδοι παρουσιάζουν σημαντικούς περιορισμούς όταν το κείμενο είναι ανομοιογενές, ασύντακτο ή περιέχει πολλές μικρές, αποσπασματικές μονάδες, όπως συμβαίνει στις απαντήσεις ανοικτού τύπου. Σε τέτοιες περιπτώσεις, η επιλογή προτάσεων δεν αρκεί για να παραχθεί μια συνεκτική περίληψη, γεγονός που καθιστά τις abstractive μεθόδους πιο κατάλληλες.

2.2.3 Abstractive μέθοδοι: seq2seq, attention, Transformers

Οι abstractive μέθοδοι περίληψης αποτελούν το πιο προηγμένο και απαιτητικό πεδίο της αυτόματης περίληψης. Η ανάπτυξή τους βασίζεται σε νευρωνικά δίκτυα που μπορούν να μάθουν να παράγουν νέο κείμενο, συνθέτοντας πληροφορίες και αναδιατυπώνοντας τις ιδέες του αρχικού περιεχομένου.

Οι πρώτες abstractive μέθοδοι βασίστηκαν σε **seq2seq (sequence-to-sequence)** αρχιτεκτονικές, οι οποίες χρησιμοποιούν έναν encoder για την κατανόηση του εισερχόμενου κειμένου και έναν decoder για την παραγωγή της περίληψης. Ωστόσο, τα seq2seq μοντέλα αντιμετώπιζαν δυσκολίες στην αποτύπωση μακροχρόνιων εξαρτήσεων και στη διατήρηση της συνοχής (Sutskever et al., 2014).

Η εισαγωγή του μηχανισμού **attention** αποτέλεσε σημαντική πρόοδο, επιτρέποντας στο μοντέλο να εστιάζει σε συγκεκριμένα τμήματα του κειμένου κατά την παραγωγή κάθε λέξης της

Παραγωγή περιλήψεων με χρήση TN

περίληψης. Ο μηχανισμός attention βελτίωσε σημαντικά την ποιότητα των abstractive περιλήψεων, επιτρέποντας την καλύτερη κατανόηση της δομής και των θεματικών σχέσεων (Bahdanau et al., 2015).

Η πραγματική επανάσταση ήρθε με την αρχιτεκτονική **Transformer**, η οποία βασίζεται αποκλειστικά σε attention μηχανισμούς και επιτρέπει την παράλληλη επεξεργασία του κειμένου. Τα μοντέλα Transformer, όπως τα BART, T5, GPT και οι πολυγλωσσικές εκδοχές τους, έχουν επιτύχει εντυπωσιακά αποτελέσματα στην abstractive περίληψη, παράγοντας κείμενα υψηλής συνοχής και φυσικότητας (Vaswani et al., 2017a).

Στο πλαίσιο της παρούσας εργασίας, τα LLMs αποτελούν την κύρια τεχνολογία για την παραγωγή abstractive περιλήψεων ελληνόγλωσσων απαντήσεων ανοικτού τύπου, αξιοποιώντας την ικανότητά τους να συνθέτουν πληροφορίες και να παράγουν συνεκτικό λόγο (Liu & Lapata, 2019).

2.2.4 Πλεονεκτήματα και περιορισμοί κάθε κατηγορίας

Οι extractive και abstractive μέθοδοι παρουσιάζουν διαφορετικά πλεονεκτήματα και περιορισμούς, οι οποίοι καθορίζουν την καταλληλότητά τους για διαφορετικές εφαρμογές.

Τα **πλεονεκτήματα των extractive μεθόδων** περιλαμβάνουν:

- υψηλή διαφάνεια, καθώς οι προτάσεις προέρχονται αυτούσιες από το κείμενο,
- χαμηλό υπολογιστικό κόστος,
- σταθερή απόδοση σε κείμενα με σαφή δομή.

Ωστόσο, οι περιορισμοί τους είναι σημαντικοί:

- αδυναμία παραγωγής νέου κειμένου,
- έλλειψη συνοχής όταν οι επιλεγμένες προτάσεις δεν συνδέονται λογικά,
- περιορισμένη αποτελεσματικότητα σε ανομοιογενή δεδομένα.

Παραγωγή περιλήψεων με χρήση TN

Αντίθετα, οι **abstractive μέθοδοι** προσφέρουν:

- υψηλή ποιότητα και φυσικότητα κειμένου,
- δυνατότητα σύνθεσης πολλαπλών ιδεών,
- ευελιξία στη διατύπωση και στο ύφος.

Οι περιορισμοί τους περιλαμβάνουν:

- υψηλό υπολογιστικό κόστος,
- πιθανότητα παραγωγής ανακριβών ή «φανταστικών» πληροφοριών (hallucinations),
- ανάγκη για μεγάλα σύνολα δεδομένων και ισχυρά μοντέλα.

Στο πλαίσιο της παρούσας εργασίας, η abstractive προσέγγιση κρίνεται καταλληλότερη, καθώς οι απαντήσεις ανοικτού τύπου απαιτούν σύνθεση και όχι απλή επιλογή προτάσεων.

2.2.5 Εφαρμογές περίληψης σε πολυγλωσσικά περιβάλλοντα

Η περίληψη κειμένου σε πολυγλωσσικά περιβάλλοντα αποτελεί ένα ιδιαίτερα απαιτητικό πεδίο, καθώς τα μοντέλα πρέπει να χειριστούν διαφορετικές γλωσσικές δομές, μορφολογίες και συντακτικές ιδιαιτερότητες. Η ανάπτυξη πολυγλωσσικών μοντέλων, όπως τα mT5, mBART και mBERT, έχει επιτρέψει την εφαρμογή τεχνικών περίληψης σε γλώσσες με περιορισμένα δεδομένα, όπως τα ελληνικά.

Οι πολυγλωσσικές εφαρμογές περίληψης περιλαμβάνουν:

- περίληψη κειμένων σε γλώσσες χαμηλών πόρων,
- διαγλωσσική περίληψη (π.χ. περίληψη ελληνικού κειμένου στα αγγλικά),
- πολυγλωσσική εκπαίδευση μοντέλων με μεταφορά γνώσης από γλώσσες υψηλών πόρων.

Παραγωγή περιλήψεων με χρήση TN

Στο πλαίσιο της παρούσας εργασίας, η πολυγλωσσική διάσταση είναι κρίσιμη, καθώς τα ελληνικά διαθέτουν περιορισμένα εξειδικευμένα μοντέλα. Η χρήση πολυγλωσσικών LLMs επιτρέπει την αξιοποίηση γνώσης από άλλες γλώσσες, βελτιώνοντας την ποιότητα των παραγόμενων περιλήψεων.

2.3 Μεγάλα Γλωσσικά Μοντέλα (LLMs)

Τα Μεγάλα Γλωσσικά Μοντέλα (Large Language Models – LLMs) αποτελούν την πιο πρόσφατη και ριζοσπαστική εξέλιξη στον χώρο της Επεξεργασίας Φυσικής Γλώσσας. Πρόκειται για νευρωνικά μοντέλα τεράστιας κλίμακας, εκπαιδευμένα σε δισεκατομμύρια λέξεις, τα οποία έχουν τη δυνατότητα να κατανοούν, να παράγουν και να χειρίζονται φυσική γλώσσα με τρόπο που προσεγγίζει την ανθρώπινη ικανότητα. Η ανάπτυξη των LLMs έχει μετασηματίσει πλήρως το τοπίο του NLP, επιτρέποντας την υλοποίηση εργασιών όπως η περίληψη, η μετάφραση, η απάντηση σε ερωτήσεις, η ανάλυση συναισθήματος και η δημιουργία περιεχομένου με πρωτοφανή ακρίβεια και φυσικότητα. Στο πλαίσιο της παρούσας εργασίας, τα LLMs αποτελούν τον πυρήνα της διαδικασίας παραγωγής abstractive περιλήψεων ελληνόγλωσσων απαντήσεων ανοικτού τύπου, καθιστώντας αναγκαία μια εις βάθος κατανόηση της αρχιτεκτονικής, της λειτουργίας και των περιορισμών τους.

2.3.1 Αρχιτεκτονική Transformer και θεμελιώδεις έννοιες

Η αρχιτεκτονική **Transformer** (Vaswani et al., 2017a) αποτελεί το θεμέλιο πάνω στο οποίο έχουν αναπτυχθεί σχεδόν όλα τα σύγχρονα LLMs. Σε αντίθεση με τις προηγούμενες αρχιτεκτονικές, όπως τα RNNs και τα LSTMs, οι Transformers δεν βασίζονται σε αναδρομικούς μηχανισμούς, αλλά σε έναν μηχανισμό προσοχής (attention), ο οποίος επιτρέπει στο μοντέλο να εστιάζει σε

Παραγωγή περιλήψεων με χρήση TN

διαφορετικά τμήματα του κειμένου ανάλογα με το περιεχόμενο και τα συμφραζόμενα (Vaswani et al., 2017a).

Ο πυρήνας της αρχιτεκτονικής είναι ο μηχανισμός **self-attention**, ο οποίος επιτρέπει σε κάθε λέξη να «βλέπει» όλες τις άλλες λέξεις της πρότασης και να υπολογίζει τη σημασία τους. Με αυτόν τον τρόπο, το μοντέλο μπορεί να αποτυπώνει μακροχρόνιες εξαρτήσεις, να κατανοεί τη συντακτική δομή και να εντοπίζει σημασιολογικές σχέσεις. Η αρχιτεκτονική αποτελείται από πολλαπλά επίπεδα attention, feed-forward νευρωνικών δικτύων και μηχανισμών κανονικοποίησης, τα οποία επιτρέπουν την εκμάθηση πολύπλοκων γλωσσικών μοτίβων (Vaswani et al., 2017).

Οι Transformers διακρίνονται σε **encoder-only** μοντέλα (π.χ. BERT), **decoder-only** μοντέλα (π.χ. GPT) και **encoder-decoder** μοντέλα (π.χ. T5, BART). Κάθε κατηγορία εξυπηρετεί διαφορετικούς σκοπούς: τα encoder-only μοντέλα είναι ιδανικά για κατανόηση κειμένου, τα decoder-only για παραγωγή, ενώ τα encoder-decoder για εργασίες μετασχηματισμού κειμένου, όπως η περίληψη.

Η αρχιτεκτονική Transformer έχει αποδειχθεί εξαιρετικά αποτελεσματική, επιτρέποντας την εκπαίδευση μοντέλων με δισεκατομμύρια παραμέτρους και την επίτευξη επιδόσεων που ξεπερνούν κατά πολύ τις προηγούμενες τεχνολογίες.

2.3.2 Προεκπαίδευση και fine-tuning

Τα LLMs εκπαιδεύονται σε δύο βασικά στάδια: την **προεκπαίδευση (pre-training)** και την **προσαρμογή (fine-tuning)**.

Η προεκπαίδευση πραγματοποιείται σε τεράστια σώματα κειμένου, τα οποία περιλαμβάνουν βιβλία, άρθρα, ιστοσελίδες και άλλες πηγές. Στόχος της προεκπαίδευσης είναι η εκμάθηση γενικών γλωσσικών μοτίβων, όπως η σύνταξη, η σημασιολογία, οι σχέσεις μεταξύ λέξεων και η λογική συνοχή. Τα πιο συνηθισμένα pre-training tasks είναι:

- **Masked Language Modeling (MLM)**: απόκρυψη λέξεων και πρόβλεψή τους (BERT) (Devlin et al., 2019).

Παραγωγή περιλήψεων με χρήση TN

- **Causal Language Modeling (CLM):** πρόβλεψη της επόμενης λέξης (GPT) (Radford et al., 2019).
- **Denoising autoencoding:** ανακατασκευή παραμορφωμένου κειμένου (BART) (Lewis et al., 2020).
- **Span corruption:** απόκρυψη τμημάτων κειμένου (T5) (Raffel et al., 2020).

Το fine-tuning αποτελεί το δεύτερο στάδιο, στο οποίο το μοντέλο προσαρμόζεται σε μια συγκεκριμένη εργασία, όπως περίληψη, ταξινόμηση ή μετάφραση. Το fine-tuning απαιτεί μικρότερο σύνολο δεδομένων, αλλά υψηλής ποιότητας, και επιτρέπει στο μοντέλο να εξειδικεύσει τις γλωσσικές του γνώσεις. Στην παρούσα εργασία, τα LLMs χρησιμοποιούνται κυρίως σε **inference mode**, χωρίς fine-tuning, λόγω ζητημάτων απορρήτου και περιορισμών στα ελληνικά δεδομένα.

2.3.3 Μοντέλα γενικής χρήσης: GPT, Claude, Gemini, Mistral, Llama

Τα σύγχρονα LLMs γενικής χρήσης αποτελούν την αιχμή της τεχνολογίας και χρησιμοποιούνται ευρέως σε εφαρμογές παραγωγής κειμένου.

- **GPT (OpenAI):** Τα μοντέλα GPT-4, GPT-4o και GPT-3.5 έχουν θέσει νέα πρότυπα στην παραγωγή φυσικής γλώσσας. Διακρίνονται για τη συνοχή, την ακρίβεια και την ικανότητά τους να ακολουθούν οδηγίες. Η απόδοσή τους στα ελληνικά είναι εξαιρετική, καθιστώντας τα κατάλληλα για abstractive περίληψη (Brown et al., 2020).
- **Claude (Anthropic):** Τα μοντέλα Claude 3 και Claude 3.5 δίνουν έμφαση στην ασφάλεια, τη συνέπεια και την αποφυγή παραπληροφόρησης. Παρουσιάζουν υψηλή ποιότητα περίληψης και καλή απόδοση σε πολύπλοκες εργασίες.
- **Gemini (Google):** Η σειρά Gemini 1.5 προσφέρει ισχυρές δυνατότητες κατανόησης και παραγωγής κειμένου, με ιδιαίτερη έμφαση στη multimodal επεξεργασία. Η απόδοσή της στα ελληνικά είναι καλή, αν και ελαφρώς χαμηλότερη από τα GPT/Claude.

Παραγωγή περιλήψεων με χρήση TN

- **Mistral (Mistral AI):** Τα μοντέλα Mistral Large και Mixtral 8x7B είναι ελαφρύτερα και οικονομικότερα, με ανταγωνιστική απόδοση και ευρωπαϊκή εστίαση. Αποτελούν καλή επιλογή για χαμηλότερο κόστος.
- **Llama (Meta):** Τα Llama 3 και 3.1 αποτελούν ανοιχτά μοντέλα υψηλής ποιότητας, με καλή απόδοση στα ελληνικά και δυνατότητα χρήσης μέσω πολλών παρόχων API (Touvron et al., 2023).

Τα μοντέλα αυτά χρησιμοποιούνται στην παρούσα εργασία για τη σύγκριση της ποιότητας των παραγόμενων περιλήψεων.

2.3.4 Πολυγλωσσικά και ελληνικά μοντέλα: mT5, Greek-BERT, Greek-Llama

Η ελληνική γλώσσα αποτελεί πρόκληση για τα LLMs λόγω της μορφολογικής πολυπλοκότητας και του περιορισμένου όγκου διαθέσιμων δεδομένων. Ωστόσο, έχουν αναπτυχθεί μοντέλα που υποστηρίζουν τα ελληνικά σε υψηλό επίπεδο.

- **mT5:** Πολυγλωσσική εκδοχή του T5, εκπαιδευμένη σε 101 γλώσσες. Παρουσιάζει καλή απόδοση σε εργασίες περίληψης και αποτελεί baseline για την παρούσα εργασία (Xue et al., 2021).
- **Greek-BERT:** Μοντέλο ειδικά εκπαιδευμένο στα ελληνικά, κατάλληλο για εργασίες κατανόησης κειμένου. Δεν είναι ιδανικό για παραγωγή, αλλά χρήσιμο για embeddings (Koutsikakis et al., 2020).
- **Greek-Llama:** Προσαρμοσμένες εκδόσεις του Llama σε ελληνικά δεδομένα, οι οποίες βελτιώνουν την απόδοση σε εργασίες παραγωγής κειμένου.

Τα πολυγλωσσικά μοντέλα έχουν το πλεονέκτημα της μεταφοράς γνώσης από γλώσσες υψηλών πόρων, γεγονός που ενισχύει την απόδοσή τους στα ελληνικά.

Παραγωγή περιλήψεων με χρήση TN

2.3.5 Ικανότητες και περιορισμοί των LLMs στην ελληνική γλώσσα

Τα LLMs παρουσιάζουν σημαντικές ικανότητες στην ελληνική γλώσσα, όπως:

- καλή κατανόηση συντακτικών δομών,
- υψηλή συνοχή στην παραγωγή κειμένου,
- ικανότητα σύνθεσης πολλαπλών θεμάτων,
- αποτελεσματική περίληψη ακόμη και ανομοιογενών δεδομένων.

Ωστόσο, υπάρχουν και περιορισμοί:

- μειωμένη ακρίβεια σε εξειδικευμένο λεξιλόγιο,
- δυσκολία σε ιδιωτισμούς και τοπικές εκφράσεις,
- πιθανότητα παραγωγής ανακριβών πληροφοριών,
- περιορισμένη εκπαίδευση σε ελληνικά corpora.

Οι περιορισμοί αυτοί καθιστούν απαραίτητη την προσεκτική αξιολόγηση των παραγόμενων περιλήψεων (Papantoniou & Tzitzikas, 2020).

2.3.6 Ζητήματα αξιοπιστίας, παραπληροφόρησης και bias στα LLMs

Παρά τις εντυπωσιακές τους δυνατότητες, τα LLMs παρουσιάζουν ζητήματα αξιοπιστίας που πρέπει να λαμβάνονται υπόψη.

- **Παραπληροφόρηση (hallucinations):** Τα μοντέλα μπορεί να παράγουν πληροφορίες που δεν υπάρχουν στο αρχικό κείμενο. Αυτό είναι κρίσιμο σε εργασίες περίληψης, όπου απαιτείται πιστότητα στο περιεχόμενο.

Παραγωγή περιλήψεων με χρήση TN

- **Bias:** Τα μοντέλα ενδέχεται να αναπαράγουν προκαταλήψεις που υπάρχουν στα δεδομένα εκπαίδευσης. Στο πλαίσιο της εκπαιδευτικής αξιολόγησης, αυτό μπορεί να οδηγήσει σε στρεβλώσεις.
- **Ασυνέπεια:** Η παραγωγή μπορεί να διαφέρει μεταξύ εκτελέσεων, ειδικά με υψηλή θερμοκρασία.
- **Έλλειψη διαφάνειας:** Η λειτουργία των LLMs είναι δύσκολο να ερμηνευθεί, γεγονός που περιορίζει τη δυνατότητα ελέγχου.

Η παρούσα εργασία αντιμετωπίζει τα ζητήματα αυτά μέσω ποσοτικής και ποιοτικής αξιολόγησης, καθώς και μέσω προσεκτικού σχεδιασμού των prompts.

Παραγωγή περιλήψεων με χρήση TN

2.4 Αξιολόγηση Συστημάτων Περίληψης

Η αξιολόγηση συστημάτων περίληψης αποτελεί ένα από τα πιο κρίσιμα και σύνθετα ζητήματα στην Επεξεργασία Φυσικής Γλώσσας, καθώς η ποιότητα μιας περίληψης δεν εξαρτάται μόνο από την ακρίβεια ή την πληρότητα, αλλά και από τη συνοχή, τη φυσικότητα, τη σημασιολογική πιστότητα και τη χρηστικότητά της στο εκάστοτε πλαίσιο εφαρμογής. Η αξιολόγηση περιλήψεων είναι ιδιαίτερα απαιτητική, διότι η περίληψη αποτελεί μια εγγενώς υποκειμενική διαδικασία: διαφορετικοί άνθρωποι μπορεί να παράγουν διαφορετικές, αλλά εξίσου έγκυρες περιλήψεις του ίδιου κειμένου. Για τον λόγο αυτό, η επιστημονική κοινότητα έχει αναπτύξει ένα σύνολο ποσοτικών και ποιοτικών μεθόδων αξιολόγησης, οι οποίες επιχειρούν να αποτυπώσουν διαφορετικές διαστάσεις της ποιότητας. Στο πλαίσιο της παρούσας εργασίας, η αξιολόγηση των περιλήψεων που παράγονται από Μεγάλα Γλωσσικά Μοντέλα είναι απαραίτητη για την τεκμηρίωση της αποτελεσματικότητας των μοντέλων και για τη σύγκριση των επιδόσεών τους σε ελληνόγλωσσα δεδομένα εκπαιδευτικής αξιολόγησης.

2.4.1 Ποσοτικές μετρικές: ROUGE, BLEU, BERTScore

Για την αξιολόγηση των περιλήψεων χρησιμοποιούνται μετρικές επικάλυψης που συγκρίνουν την παραγόμενη περίληψη με τη χρυσή περίληψη αναφοράς. Οι μετρικές αυτές δεν κρίνουν μόνες τους αν μια περίληψη είναι παιδαγωγικά επαρκής, αλλά δείχνουν πόσο κοντά βρίσκεται στο ανθρώπινο κείμενο αναφοράς σε επίπεδο λέξεων ή ακολουθιών λέξεων.

Η ROUGE-1 μετρά την επικάλυψη μεμονωμένων λέξεων. Για παράδειγμα, αν η χρυσή περίληψη αναφέρει «σαφής οργάνωση και άμεση ανατροφοδότηση» και η παραγόμενη περίληψη αναφέρει «καλή οργάνωση και ανατροφοδότηση», τότε υπάρχει επικάλυψη στις λέξεις «οργάνωση» και «ανατροφοδότηση». Η ROUGE-2 εξετάζει ζεύγη διαδοχικών λέξεων και είναι πιο αυστηρή, επειδή απαιτεί να διατηρείται και η τοπική σειρά των λέξεων. Η ROUGE-L βασίζεται στη μεγαλύτερη κοινή ακολουθία λέξεων και αποτυπώνει καλύτερα τη διατήρηση της δομής μιας φράσης. Η BLEU χρησιμοποιείται συχνά στη μηχανική μετάφραση και μετρά επικάλυψη n-grams

Παραγωγή περιλήψεων με χρήση TN

με έμφαση στην ακρίβεια, ενώ στην περίληψη χρειάζεται προσεκτική ερμηνεία, επειδή μπορεί να αδικεί σωστές αναδιατυπώσεις (Papineni et al., 2002).

Οι ποσοτικές μετρικές αποτελούν τον πιο διαδεδομένο τρόπο αξιολόγησης συστημάτων περίληψης, καθώς επιτρέπουν την αντικειμενική σύγκριση των παραγόμενων περιλήψεων με «χρυσές» περιλήψεις αναφοράς. Η πιο γνωστή οικογένεια μετρικών είναι η **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)**, η οποία μετρά την επικάλυψη n-grams μεταξύ της παραγόμενης περίληψης και της περίληψης αναφοράς. Οι πιο συνηθισμένες παραλλαγές είναι: (Lin, 2004)

- **ROUGE-1**: επικάλυψη μονο-λέξεων,
- **ROUGE-2**: επικάλυψη δι-λέξεων,
- **ROUGE-L**: επικάλυψη της μεγαλύτερης κοινής ακολουθίας (Longest Common Subsequence).

Οι ROUGE μετρικές είναι ιδιαίτερα χρήσιμες σε extractive περιλήψεις, αλλά παρουσιάζουν περιορισμούς σε abstractive περιλήψεις, όπου η διατύπωση μπορεί να διαφέρει σημαντικά από το κείμενο αναφοράς.

Η μετρική **BLEU (Bilingual Evaluation Understudy)**, αρχικά σχεδιασμένη για μηχανική μετάφραση, χρησιμοποιείται επίσης στην περίληψη. Μετρά την ακρίβεια n-grams, αλλά είναι λιγότερο κατάλληλη για abstractive περίληψη, καθώς τιμωρεί τις παραφράσεις και τις ελεύθερες διατυπώσεις (Papineni et al., 2002) .

Η **BERTScore** αποτελεί μια πιο σύγχρονη μετρική, η οποία χρησιμοποιεί contextual embeddings από μοντέλα τύπου BERT για να υπολογίσει τη σημασιολογική ομοιότητα μεταξύ δύο κειμένων. Σε αντίθεση με τις ROUGE/BLEU, η BERTScore δεν βασίζεται σε επιφανειακή επικάλυψη λέξεων, αλλά σε βαθύτερη σημασιολογική αντιστοίχιση. Για γλώσσες όπως τα ελληνικά, όπου η μορφολογική ποικιλία είναι μεγάλη, η BERTScore προσφέρει σημαντικά πλεονεκτήματα (Zhang et al., 2020).

Παραγωγή περιλήψεων με χρήση TN

2.4.2 Σημασιολογικές μετρικές: MoverScore, QuestEval, SummaC

Οι σημασιολογικές μετρικές επιχειρούν να αξιολογήσουν όχι μόνο την κοινή λεξιλογική επιφάνεια, αλλά και το αν δύο περιλήψεις μεταφέρουν παρόμοιο νόημα. Αυτό είναι ιδιαίτερα σημαντικό στις αφαιρετικές περιλήψεις, όπου μια σωστή απάντηση μπορεί να χρησιμοποιεί διαφορετικές λέξεις από τη χρυσή περίληψη.

Το MoverScore συγκρίνει σημασιολογικές αναπαραστάσεις λέξεων ή φράσεων και υπολογίζει πόση «μετακίνηση» απαιτείται για να αντιστοιχιστεί το νόημα της παραγόμενης περίληψης με τη χρυσή. Για παράδειγμα, οι φράσεις «οι φοιτητές ζητούν περισσότερη ανατροφοδότηση» και «υπάρχει ανάγκη για συχνότερη καθοδήγηση» μπορεί να έχουν μικρή λεξιλογική επικάλυψη αλλά κοντινό νόημα. Το QuestEval προσεγγίζει την αξιολόγηση μέσω ερωτήσεων και απαντήσεων: αν οι ίδιες βασικές ερωτήσεις μπορούν να απαντηθούν από την παραγόμενη και τη χρυσή περίληψη, τότε θεωρείται ότι η πληροφορία διατηρείται. Το SummaC εστιάζει στη συνέπεια και ελέγχει αν το περιεχόμενο της περίληψης υποστηρίζεται από το αρχικό κείμενο, μειώνοντας τον κίνδυνο παραπλανητικών ή αστήρικτων διατυπώσεων (Zhao et al., 2019)

Οι σημασιολογικές μετρικές επιχειρούν να αξιολογήσουν την ποιότητα μιας περίληψης όχι με βάση την επιφανειακή ομοιότητα, αλλά με βάση τη σημασιολογική πιστότητα και την κάλυψη των βασικών εννοιών του αρχικού κειμένου.

Η **MoverScore** βασίζεται στην ιδέα της μεταφοράς λέξεων (word mover's distance) και υπολογίζει πόσο «μακριά» πρέπει να μετακινηθούν οι σημασιολογικές αναπαραστάσεις των λέξεων της περίληψης για να ταιριάξουν με αυτές του κειμένου αναφοράς. Η μετρική αυτή είναι ιδιαίτερα χρήσιμη σε abstractive περιλήψεις (Zhao et al., 2019) .

Η **QuestEval** αξιολογεί την ποιότητα της περίληψης μέσω ερωτήσεων που παράγονται αυτόματα από το αρχικό κείμενο. Αν η περίληψη μπορεί να απαντήσει σωστά σε αυτές τις ερωτήσεις,

Παραγωγή περιλήψεων με χρήση TN

θεωρείται ότι καλύπτει τα βασικά νοήματα. Η μετρική αυτή είναι πιο κοντά στην ανθρώπινη αξιολόγηση, καθώς εστιάζει στην πληροφοριακή πληρότητα (Scialom et al., 2021) .

Η **SummaC** χρησιμοποιεί μοντέλα φυσικής γλώσσας για να εντοπίσει ασυνέπειες μεταξύ της περίληψης και του αρχικού κειμένου. Εντοπίζει περιπτώσεις όπου η περίληψη περιέχει ανακρίβειες ή «φανταστικές» πληροφορίες (hallucinations), κάτι που αποτελεί κρίσιμο ζήτημα στα LLMs (Laban et al., 2022) .

Οι σημασιολογικές μετρικές είναι ιδιαίτερα σημαντικές για την παρούσα εργασία, καθώς οι abstractive περιλήψεις που παράγουν τα LLMs συχνά διαφέρουν λεκτικά από τις περιλήψεις αναφοράς, αλλά μπορεί να είναι σημασιολογικά ισοδύναμες.

2.4.3 Ποιοτική αξιολόγηση: συνοχή, πληρότητα, ακρίβεια, μη παραπληροφόρηση

Η ποιοτική αξιολόγηση αποτελεί αναπόσπαστο μέρος της αποτίμησης συστημάτων περίληψης, καθώς πολλές πτυχές της ποιότητας δεν μπορούν να αποτυπωθούν πλήρως μέσω ποσοτικών μετρικών. Η ανθρώπινη αξιολόγηση επιτρέπει την εκτίμηση χαρακτηριστικών όπως:

- **Συνοχή:** κατά πόσο η περίληψη παρουσιάζει λογική ροή, σωστή σύνδεση προτάσεων και φυσικότητα λόγου.
- **Πληρότητα:** σε ποιο βαθμό η περίληψη καλύπτει τα βασικά θέματα και τις κύριες ιδέες του αρχικού κειμένου.
- **Ακρίβεια:** κατά πόσο η περίληψη αποδίδει πιστά το περιεχόμενο χωρίς παραμορφώσεις ή παραλείψεις.
- **Μη παραπληροφόρηση:** αποφυγή «hallucinations», δηλαδή πληροφοριών που δεν υπάρχουν στο αρχικό κείμενο.

Η ποιοτική αξιολόγηση είναι ιδιαίτερα σημαντική σε εφαρμογές εκπαιδευτικής αξιολόγησης, όπου η ακρίβεια και η πιστότητα είναι κρίσιμες για τη λήψη αποφάσεων. Στην παρούσα εργασία,

Παραγωγή περιλήψεων με χρήση TN

η ποιοτική αξιολόγηση πραγματοποιείται από ανθρώπινους αξιολογητές με χρήση κλίμακας Likert και συγκεκριμένων κριτηρίων.

2.4.4 Συγκριτική αξιολόγηση extractive–abstractive μεθόδων

Η σύγκριση extractive και abstractive μεθόδων αποτελεί σημαντικό ερευνητικό ζήτημα, καθώς οι δύο προσεγγίσεις παρουσιάζουν διαφορετικά πλεονεκτήματα και περιορισμούς. Οι extractive μέθοδοι τείνουν να έχουν υψηλότερες επιδόσεις σε ROUGE/BLEU, επειδή αναπαράγουν αυτούσιες φράσεις από το κείμενο αναφοράς. Ωστόσο, συχνά παράγουν περιλήψεις που στερούνται συνοχής ή φυσικότητας.

Αντίθετα, οι abstractive μέθοδοι παράγουν πιο φυσικές και συνεκτικές περιλήψεις, αλλά μπορεί να έχουν χαμηλότερες επιδόσεις σε μετρικές που βασίζονται σε n-grams. Η σημασιολογική αξιολόγηση (π.χ. BERTScore, QuestEval) τείνει να ευνοεί τις abstractive μεθόδους, καθώς αποτυπώνει καλύτερα την ποιότητα του νοήματος.

Στο πλαίσιο της παρούσας εργασίας, η abstractive προσέγγιση κρίνεται καταλληλότερη για τις απαντήσεις ανοικτού τύπου, καθώς απαιτείται σύνθεση πολλαπλών ιδεών και όχι απλή επιλογή προτάσεων.

Παραγωγή περιλήψεων με χρήση TN

2.4.5 Προκλήσεις αξιολόγησης σε ελληνόγλωσσα δεδομένα

Η αξιολόγηση συστημάτων περίληψης σε ελληνόγλωσσα δεδομένα παρουσιάζει σημαντικές προκλήσεις:

- **Μορφολογική πολυπλοκότητα:** η ελληνική γλώσσα διαθέτει πλούσια κλίση, γεγονός που μειώνει την αποτελεσματικότητα των n-gram μετρικών.
- **Περιορισμένα corpora:** η έλλειψη μεγάλων ελληνικών συνόλων δεδομένων δυσχεραίνει την εκπαίδευση και αξιολόγηση μοντέλων.
- **Έλλειψη εξειδικευμένων εργαλείων:** πολλές μετρικές έχουν βελτιστοποιηθεί για αγγλικά δεδομένα.
- **Ιδιωματισμοί και ποικιλία ύφους:** οι απαντήσεις ανοικτού τύπου περιλαμβάνουν συχνά ανεπίσημη γλώσσα, συναισθηματικές εκφράσεις και ασυνέπειες.
- **Αναγκαιότητα ανθρώπινης αξιολόγησης:** η ποιοτική αξιολόγηση είναι πιο απαιτητική σε γλώσσες με λιγότερα διαθέσιμα εργαλεία.

Οι προκλήσεις αυτές καθιστούν απαραίτητο τον συνδυασμό ποσοτικών και ποιοτικών μεθόδων, όπως εφαρμόζεται στην παρούσα εργασία.

Παραγωγή περιλήψεων με χρήση TN

2.5 Απαντήσεις Ανοικτού Τύπου στην Εκπαιδευτική Αξιολόγηση

Οι απαντήσεις ανοικτού τύπου αποτελούν κρίσιμη πηγή ποιοτικής πληροφορίας στην εκπαιδευτική αξιολόγηση, καθώς επιτρέπουν στους φοιτητές να εκφράσουν ελεύθερα τις εμπειρίες, τις αντιλήψεις και τις προτάσεις τους, χωρίς τους περιορισμούς των κλειστών ερωτήσεων. Σε αντίθεση με τις ποσοτικές κλίμακες, οι οποίες αποτυπώνουν γενικές τάσεις, οι ανοικτές απαντήσεις προσφέρουν βαθύτερη κατανόηση των παραγόντων που επηρεάζουν τη μαθησιακή εμπειρία, αναδεικνύοντας ζητήματα που συχνά δεν μπορούν να εντοπιστούν μέσω τυποποιημένων εργαλείων. Η ανάλυση των απαντήσεων αυτών αποτελεί σημαντική πρόκληση, λόγω του όγκου, της ετερογένειας και της πολυπλοκότητάς τους, αλλά ταυτόχρονα προσφέρει μοναδικές δυνατότητες για τη βελτίωση της ποιότητας της διδασκαλίας και των εκπαιδευτικών προγραμμάτων.

2.5.1 Ρόλος και σημασία των απαντήσεων ανοικτού τύπου

Οι απαντήσεις ανοικτού τύπου διαδραματίζουν καθοριστικό ρόλο στη διαδικασία εκπαιδευτικής αξιολόγησης, καθώς επιτρέπουν στους φοιτητές να διατυπώσουν προσωπικές εμπειρίες, να αναδείξουν συγκεκριμένα προβλήματα και να προτείνουν στοχευμένες βελτιώσεις. Σε αντίθεση με τις κλειστές ερωτήσεις, οι οποίες περιορίζονται σε προκαθορισμένες επιλογές, οι ανοικτές απαντήσεις παρέχουν πλούσιο, ανεπεξέργαστο υλικό που αποτυπώνει τη φωνή των φοιτητών με μεγαλύτερη αυθεντικότητα. Μέσω αυτών, οι εκπαιδευτικοί φορείς μπορούν να εντοπίσουν ζητήματα που δεν είναι άμεσα ορατά, όπως δυσκολίες κατανόησης του υλικού, προβλήματα επικοινωνίας με τους διδάσκοντες, τεχνικές δυσλειτουργίες ή ανάγκες υποστήριξης.

Επιπλέον, οι ανοικτές απαντήσεις συμβάλλουν στη διαμόρφωση μιας πιο ολοκληρωμένης εικόνας της μαθησιακής εμπειρίας, καθώς επιτρέπουν την ανάλυση συναισθηματικών αποχρώσεων, στάσεων και προσδοκιών. Η ποιοτική αυτή διάσταση είναι ιδιαίτερα σημαντική σε ιδρύματα εξ αποστάσεως εκπαίδευσης, όπου η αλληλεπίδραση μεταξύ φοιτητών και διδασκόντων είναι περιορισμένη και η ανατροφοδότηση αποκτά μεγαλύτερη αξία. Η αξιοποίηση των ανοικτών

Παραγωγή περιλήψεων με χρήση TN

απαντήσεων αποτελεί, επομένως, βασικό εργαλείο για τη λήψη τεκμηριωμένων αποφάσεων και την ενίσχυση της ποιότητας της εκπαιδευτικής διαδικασίας.

2.5.2 Ερευνητικές προσεγγίσεις στην ανάλυση εκπαιδευτικών δεδομένων

Η ανάλυση των απαντήσεων ανοικτού τύπου έχει αποτελέσει αντικείμενο εκτεταμένης έρευνας, με προσεγγίσεις που κυμαίνονται από παραδοσιακές ποιοτικές μεθόδους έως σύγχρονες τεχνικές μηχανικής μάθησης. Οι παραδοσιακές μέθοδοι περιλαμβάνουν τη θεματική ανάλυση (thematic analysis), την ανάλυση περιεχομένου (content analysis) και την κωδικοποίηση από ειδικούς αξιολογητές. Αν και αυτές οι μέθοδοι προσφέρουν υψηλή ακρίβεια και βαθιά κατανόηση, είναι ιδιαίτερα χρονοβόρες και δύσκολα εφαρμόσιμες σε μεγάλα σύνολα δεδομένων.

Με την ανάπτυξη του NLP, έχουν προταθεί αυτοματοποιημένες προσεγγίσεις που περιλαμβάνουν:

- ταξινόμηση συναισθήματος,
- εντοπισμό θεμάτων μέσω topic modeling (π.χ. LDA),
- ομαδοποίηση απαντήσεων με clustering,
- εξαγωγή λέξεων-κλειδιών,
- αυτόματη περίληψη.

Οι σύγχρονες ερευνητικές τάσεις εστιάζουν στη χρήση νευρωνικών μοντέλων και LLMs για την κατανόηση και σύνθεση των απαντήσεων. Τα μοντέλα αυτά μπορούν να εντοπίσουν θεματικά μοτίβα, να παράγουν περιλήψεις υψηλής ποιότητας και να υποστηρίξουν τη λήψη αποφάσεων σε θεσμικά πλαίσια. Η παρούσα εργασία εντάσσεται σε αυτή τη σύγχρονη ερευνητική κατεύθυνση, αξιοποιώντας LLMs για την παραγωγή abstractive περιλήψεων.

2.5.3 Προηγούμενες εφαρμογές NLP σε εκπαιδευτικά περιβάλλοντα

Παραγωγή περιλήψεων με χρήση TN

Η εφαρμογή τεχνικών NLP στην εκπαίδευση έχει γνωρίσει σημαντική ανάπτυξη τα τελευταία χρόνια, με εφαρμογές που καλύπτουν ένα ευρύ φάσμα δραστηριοτήτων. Ενδεικτικά παραδείγματα περιλαμβάνουν:

- αυτόματη αξιολόγηση γραπτών εργασιών, (Shermis & Burstein, 2013)
- ανάλυση συναισθήματος φοιτητικών σχολίων,
- εντοπισμό παραγόντων που επηρεάζουν την επιτυχία ή αποτυχία,
- υποστήριξη διδασκόντων μέσω αναφορών ποιότητας,
- ανάλυση συζητήσεων σε διαδικτυακά φόρουμ μάθησης,
- αυτόματη δημιουργία περιλήψεων εκπαιδευτικού υλικού.

Σε πολλές μελέτες, η χρήση NLP έχει αποδειχθεί αποτελεσματική στην εξαγωγή θεματικών μοτίβων από μεγάλα σύνολα δεδομένων, επιτρέποντας την ταχύτερη και πιο αντικειμενική ανάλυση της φοιτητικής ανατροφοδότησης. Η ανάπτυξη LLMs έχει ενισχύσει περαιτέρω αυτές τις δυνατότητες, καθώς τα μοντέλα αυτά μπορούν να παράγουν περιλήψεις υψηλής ποιότητας, να εντοπίζουν λεπτές σημασιολογικές αποχρώσεις και να προσαρμόζονται σε διαφορετικά εκπαιδευτικά συμφραζόμενα.

2.5.4 Ιδιαιτερότητες των δεδομένων του ΕΑΠ και αντίστοιχων ιδρυμάτων

Τα δεδομένα εκπαιδευτικής αξιολόγησης στο Ελληνικό Ανοικτό Πανεπιστήμιο (ΕΑΠ) και σε αντίστοιχα ιδρύματα εξ αποστάσεως εκπαίδευσης παρουσιάζουν συγκεκριμένες ιδιαιτερότητες που επηρεάζουν την ανάλυση και την επεξεργασία τους. Πρώτον, οι φοιτητές του ΕΑΠ προέρχονται από ευρύ φάσμα ηλικιών, επαγγελματικών υποβάθρων και μαθησιακών εμπειριών, γεγονός που οδηγεί σε μεγάλη ποικιλία εκφράσεων, ύφους και γλωσσικών επιλογών. Δεύτερον, οι απαντήσεις ανοικτού τύπου συχνά περιλαμβάνουν συναισθηματικά φορτισμένες διατυπώσεις, αναφορές σε προσωπικές δυσκολίες και σχόλια για την εξ αποστάσεως φύση της μάθησης.

Παραγωγή περιλήψεων με χρήση TN

Επιπλέον, οι απαντήσεις είναι συχνά ανομοιογενείς ως προς το μήκος, τη δομή και την ποιότητα γραφής. Πολλοί φοιτητές χρησιμοποιούν προφορικό ύφος, συντομογραφίες ή ανορθογραφίες, ενώ άλλοι παρέχουν εκτενείς και λεπτομερείς αναλύσεις. Η ποικιλία αυτή καθιστά απαραίτητη την προσεκτική προεπεξεργασία των δεδομένων και την επιλογή μοντέλων που μπορούν να χειριστούν τέτοιου είδους ετερογένεια.

Τέλος, η ανάγκη προστασίας προσωπικών δεδομένων είναι ιδιαίτερα έντονη, καθώς οι απαντήσεις μπορεί να περιλαμβάνουν ευαίσθητες πληροφορίες. Η παρούσα εργασία λαμβάνει υπόψη αυτές τις ιδιαιτερότητες, εφαρμόζοντας ψευδωνυμοποίηση, καθαρισμό και προσεκτικό σχεδιασμό των προτροπών προς τα LLMs.

Παραγωγή περιλήψεων με χρήση TN

Κεφάλαιο 3 — Μεθοδολογία και Υλοποίηση

3.1 Γενική Περιγραφή του Έργου

Το παρόν έργο αναπτύσσει ένα ολοκληρωμένο σύστημα αυτοματοποιημένης παραγωγής περιλήψεων από απαντήσεις ανοικτού τύπου σε ερωτηματολόγια εκπαιδευτικής αξιολόγησης.

Ο στόχος του έργου είναι η μετατροπή ενός μεγάλου όγκου ανομοιογενών, ασυντάκτων απαντήσεων σε δομημένες, λογικά συνεκτικές και χρήσιμες περιλήψεις που μπορούν να αξιοποιηθούν από τους εκπαιδευτικούς για την κατανόηση της φοιτητικής εμπειρίας. Το έργο επιχειρεί τη σύγκριση τριών διαφορετικών Μεγάλων Γλωσσικών Μοντέλων (LLMs) που προσφέρονται από την OpenAI (GPT-4o, GPT-4o-mini, GPT-3.5-turbo) με σκοπό την αξιολόγηση της απόδοσης τους στη δημιουργία περιλήψεων ελληνικών κειμένων.

Η υλοποίηση περιλαμβάνει τέσσερα διακριτά στάδια: **(α) προεπεξεργασία και κατασκευή του συνόλου δεδομένων**, (β) παραγωγή περιλήψεων χρησιμοποιώντας τα LLMs, (γ) αξιολόγηση των παραγόμενων περιλήψεων με πολυδιάστατες μετρικές, και (δ) οπτικοποίηση και ανάλυση των αποτελεσμάτων.

Παραγωγή περιλήψεων με χρήση TN

3.2 Περιγραφή και Κατασκευή του Συνόλου Δεδομένων

3.2.1 Πηγές Δεδομένων Τα δεδομένα του έργου προέρχονται από δύο κύριες πηγές:

- Απαντήσεις Φοιτητών (Student Responses): Τρία αρχεία Excel που περιέχουν τις απαντήσεις φοιτητών στις τρεις κατηγορίες ερωτήσεων (θετικά, αρνητικά, προτάσεις βελτίωσης)
- Πρότυπες Περιλήψεις (Golden Summaries): 126 αρχεία Excel (42 προγράμματα $\times$ 3 κατηγορίες ερωτήσεων), που περιέχουν τις χειροκίνητα δημιουργημένες περιλήψεις

Το σύνολο δεδομένων αποτελείται από 74,997 ζεύγη (φοιτητική απάντηση, χρυσή περίληψη) κατανεμημένα σε 42 διαφορετικά Προγράμματα Σπουδών και 3 κατηγορίες ερωτήσεων (imp=προτάσεις βελτίωσης, pos=θετικά, neg=αρνητικά).

Πριν από τη χρήση τους στην παρούσα εργασία, τα δεδομένα είχαν περάσει από διαδικασία προεπεξεργασίας και ανωνυμοποίησης. Η διαδικασία αυτή περιλάμβανε την αφαίρεση ή απόκρυψη στοιχείων που θα μπορούσαν να οδηγήσουν σε ταυτοποίηση φυσικών προσώπων ή συγκεκριμένων εσωτερικών δεδομένων, την ομογενοποίηση της μορφής των απαντήσεων και τη διατήρηση μόνο των πεδίων που ήταν απαραίτητα για την παραγωγή και αξιολόγηση περιλήψεων. Έτσι, το σύνολο δεδομένων επιτρέπει την πειραματική αξιολόγηση χωρίς να εκθέτει πραγματικά δεδομένα ενός ιδρύματος.

Παραγωγή περιλήψεων με χρήση TN

3.2.2 Κατανομή Δεδομένων

Ο αριθμός των δεδομένων κατανέμεται ως εξής:

- Κατά κατηγορία ερώτησης: IMP=24,999, POS=24,999, NEG=24,999
- Κορυφαία προγράμματα (κατά αριθμό δεδομένων): ΔΗΔ (11,283), ΕΑΓ (9,075), ΕΤΑ (7,062)

Αυτή η κατανομή αντικατοπτρίζει τη δομή των ερωτηματολογίων αξιολόγησης του ΕΑΠ και διασφαλίζει τη σύγκριση των μοντέλων σε διαφορετικές θεματικές ενότητες.

3.3 Αρχιτεκτονική του Συστήματος

Σχήμα 3.1α: Ροή σύνδεσης των βασικών υποσυστημάτων της υλοποίησης.

Αρχεία Excel με ανοικτές απαντήσεις και πρότυπες περιλήψεις → φόρτωση δεδομένων → καθαρισμός και κανονικοποίηση κειμένου → ομαδοποίηση ανά Πρόγραμμα Σπουδών και τύπο ερώτησης → δημιουργία prompt → κλήση LLM → αποθήκευση παραγόμενων περιλήψεων → υπολογισμός μετρικών → συγκριτική ερμηνεία αποτελεσμάτων.

Το σύστημα αποτελείται από πέντε κύρια υποσύστημα που λειτουργούν σειριακά:

1. Preprocessing Module (src/preprocessing/)
2. LLM Clients Module (src/llm_clients/)
3. Summarization Module (src/summarization/)
4. Evaluation Module (src/evaluation/)
5. Visualization Module (src/evaluation/visualize_metrics.py)

Παραγωγή περιλήψεων με χρήση TN

3.4 Προεπεξεργασία Δεδομένων

3.4.1 Φόρτωση και Καθαρισμός

Ο modulo preprocessing αποτελείται από τρία κύρια αρχεία:

- `load_excel.py`: Φορτώνει τα αρχεία Excel φοιτητών και χρυσών περιλήψεων, αναγνωρίζοντας αυτόματα τα αρχεία στους κατάλληλους καταλόγους
- `clean_text.py`: Εφαρμόζει βασικό καθαρισμό κειμένου (trimming, κανονικοποίηση κενών)
- `build_dataset.py`: Συνδυάζει τα δεδομένα φοιτητών με τις πρότυπες περιλήψεις και εξάγει το τελικό CSV

3.4.2 Εξαγωγή Μεταδεδομένων

Κατά τη φόρτωση των αρχείων χρυσών περιλήψεων, το σύστημα εξάγει δύο κρίσιμα μεταδεδομένα από τα ονόματα των αρχείων:

- `Program Code`: Εξάγεται από το πρώτο τμήμα του ονόματος (π.χ. "ΔΗΔ" από "ΔΗΔ_thesmiki_2023-2024a_pos_summary_rev.xlsx")
- `Question Type`: Ταξινομείται με βάση τα κλειδιά "_pos" (θετικά), "_neg" (αρνητικά), "_imp" (σημαντικά)

Αυτή η μεθοδολογία εξαγωγής είναι ρωμαλέα και ανεξάρτητη από άλλα μεταδεδομένα, συνεπώς λειτουργεί ανεξάρτητα από τη δομή του Excel.

Παραγωγή περιλήψεων με χρήση TN

3.5 Ενσωμάτωση Μεγάλων Γλωσσικών Μοντέλων

3.5.1 Επιλογή Μοντέλων

Η αρχιτεκτονική της υλοποίησης είναι επεκτάσιμη και επιτρέπει την προσθήκη νέων μοντέλων χωρίς ριζική αλλαγή της ροής εργασίας. Η ένταξη ενός επιπλέον μοντέλου απαιτεί κυρίως την προσθήκη αντίστοιχου client ή συνάρτησης κλήσης, τη δήλωση των παραμέτρων του και την ενσωμάτωσή του στον ίδιο μηχανισμό παραγωγής και αξιολόγησης. Με αυτόν τον τρόπο, το σύστημα μπορεί να αξιοποιηθεί μελλοντικά για σύγκριση νεότερων LLMs ή διαφορετικών παρόχων.

Για την υλοποίηση του έργου επιλέχθηκαν τρία μοντέλα από την OpenAI:

- GPT-4o: Το ισχυρότερο μοντέλο, με ικανότητα κατανόησης σύνθετων συμφραζομένων και παραγωγής υψηλής ποιότητας κειμένου
- GPT-4o-mini: Μικρότερη παραλλαγή του GPT-4o, βελτιστοποιημένη για ταχύτητα και χαμηλότερο κόστος
- GPT-3.5-turbo: Ένα παλαιότερο μοντέλο με χαμηλότερη ικανότητα αλλά σημαντικά χαμηλότερο κόστος

3.5.2 System Prompt (Οδηγίες για τα Μοντέλα)

Κατά την επεξεργασία μεγάλων ομάδων σχολίων, η διάσπαση της εισόδου σε επιμέρους chunks αποτελεί κρίσιμο σημείο της υλοποίησης. Όταν τα σχόλια δεν χωρούν σε μία ενιαία κλήση προς το μοντέλο, κάθε chunk πρέπει να συνοδεύεται από πληροφορία θέσης, για παράδειγμα «τμήμα 1 από 3», ώστε το μοντέλο να γνωρίζει ότι παράγει ενδιάμεση και όχι τελική σύνοψη. Στη συνέχεια, οι επιμέρους συνόψεις πρέπει να συνδυάζονται σε δεύτερο στάδιο, όπου ζητείται μια συνολική περίληψη που λαμβάνει υπόψη όλα τα chunks. Αυτή η διαδικασία μειώνει τον κίνδυνο

Παραγωγή περιλήψεων με χρήση TN

αποσπασματικών αποτελεσμάτων και κάνει σαφέστερη τη σχέση ανάμεσα στα τμήματα της εισόδου.

Κάθε κλήση στα LLMs χρησιμοποιεί έναν ενισχυμένο system prompt που ορίζει ρητά τις προσδοκώμενες συμπεριφορές:

Οδηγίες System Prompt:

- Παράγεται ακριβής, αντικειμενική και συνεκτική περίληψη
- Βάση ΑΠΟΚΛΕΙΣΤΙΚΑ των δεδομένων που παρέχονται (χωρίς hallucinations)
- Χωρίς προηγούμενα συμφραζόμενα από άλλες κλήσεις
- Κάθε περίληψη είναι πλήρως ανεξάρτητη
- Οργάνωση σε θεματικές ενότητες
- Εντοπισμός κυρίαρχων τάσεων, όχι μεμονωμένων παραδειγμάτων
- Ακαδημαϊκό ύφος, σύντομο αλλά περιεκτικό
- Χωρίς υποθέσεις ή εικασίες

3.5.3 User Prompt (Πλαίσιο Ερώτησης)

Το user prompt παρέχει πλαίσιο σχετικά με το πρόγραμμα και την κατηγορία ερώτησης:

Πλαίσιο ερώτησης: - imp: "Οι φοιτητές διατυπώνουν προτάσεις βελτίωσης για το Πρόγραμμα Σπουδών" - pos: "Οι φοιτητές αναφέρουν τα θετικά στοιχεία του Προγράμματος Σπουδών" - neg: "Οι φοιτητές αναφέρουν τα αρνητικά στοιχεία ή προβλήματα του Προγράμματος Σπουδών"

Οι παράμετροι κλήσης ορίστηκαν ως εξής:

- temperature=0.2 (χαμηλή: ενθαρρύνει συνεπείς, ντετερμινιστικές απαντήσεις)

Η παράμετρος temperature ελέγχει τον βαθμό τυχαιότητας της παραγωγής. Χαμηλή τιμή, όπως 0.2, προτιμήθηκε επειδή η εργασία απαιτεί σταθερές και συγκρίσιμες περιλήψεις και όχι

Παραγωγή περιλήψεων με χρήση TN

δημιουργικές αποκλίσεις. Η παράμετρος `max_tokens` ορίζει το μέγιστο μήκος της απάντησης, ώστε οι περιλήψεις να παραμένουν σύντομες και λειτουργικές. Η παράμετρος `top_p`, όταν χρησιμοποιείται, περιορίζει την επιλογή λέξεων στο πιθανότερο τμήμα της κατανομής του μοντέλου και λειτουργεί συμπληρωματικά προς τη `temperature`. Για παράδειγμα, σε μια ομάδα σχολίων με πολλές επαναλαμβανόμενες αναφορές στην «επικοινωνία με τον διδάσκοντα», οι χαμηλές τιμές βοηθούν το μοντέλο να αποδώσει το κοινό θέμα χωρίς να εισαγάγει άσχετες λεπτομέρειες.

- `max_tokens=200` (περιορίζει το μήκος της περίληψης σε λογικά όρια)

3.6 Παραγωγή Περιλήψεων

Η παραγωγή περιλήψεων πραγματοποιείται στο αρχείο `generate_summary.py` και ακολουθεί την ακόλουθη διαδικασία:

1. Φόρτωση του συνόλου δεδομένων κατασκευασμένου στο στάδιο προεπεξεργασίας
2. Επανάληψη μέσω κάθε σειράς (φοιτητική απάντηση)
3. Σειριακή κλήση στο LLM με τα κατάλληλα prompts
4. Αποθήκευση της απάντησης και των μεταδεδομένων (`program`, `question`, `model`)
5. Αποθήκευση του αποτελέσματος σε CSV αρχείο

Σε λειτουργία δοκιμής (`test_mode=True`), η παραγωγή σταματά μετά από 200 περιλήψεις. Στη δήλωση παραγωγής, δημιουργούνται περιλήψεις για 201 δεδομένα.

Οι παραγόμενες περιλήψεις αποθηκεύονται με τη μορφή CSV και περιέχουν τα ακόλουθα πεδία:

- `model`: Όνομα του LLM (π.χ. "gpt-4o")
- `program`: Κωδικός προγράμματος σπουδών

Παραγωγή περιλήψεων με χρήση TN

- question: Κατηγορία ερώτησης (imp/pos/neg)
- student_text: Η αρχική απάντηση του φοιτητή
- gold_summary: Η χρυσή περίληψη (αναφορά)
- generated_summary: Η περίληψη που παράγεται από το LLM

Παραγωγή περιλήψεων με χρήση TN

3.7 Αξιολόγηση των Περιλήψεων

3.7.1 Μετρικές που Χρησιμοποιούνται

Το σύστημα αξιολόγησης χρησιμοποιεί έξι διακριτές μετρικές που αποτυπώνουν διαφορετικές πτυχές της ποιότητας:

BLEU (Bilingual Evaluation Understudy): Μετρά την ακριβή ταύτιση n-grammas μεταξύ της παραγόμενης και χρυσής περίληψης. Τιμές 0-100. Πλεονέκτημα: απλή και γρήγορη. Μειονέκτημα: δεν λαμβάνει υπόψη σημασιολογικές ομοιότητες.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Εξετάζει την ανάκληση των n-grammas. Τρεις παραλλαγές:

- ROUGE-1: Mono-λέξεις
- ROUGE-2: Δι-λέξεις
- ROUGE-L: Μεγαλύτερη κοινή ακολουθία (LCS)

Sentence-BERT Cosine Similarity: Χρησιμοποιεί το μοντέλο "paraphrase-multilingual-MiniLM-L12-v2" για κωδικοποίηση των προτάσεων σε dense embeddings. Υπολογίζει το cosine similarity μεταξύ των embeddings. Τιμές 0-1. Πλεονέκτημα: λαμβάνει υπόψη σημασιολογικές ομοιότητες. Μειονέκτημα: χρησιμοποιεί προ-εκπαιδευμένο μοντέλο.

Character-level Cosine Similarity: Υπολογίζει ομοιότητα βάσει χαρακτήρων χρησιμοποιώντας κάθε χαρακτήρα ως διάσταση του διανύσματος. Μια πιο εξειδικευμένη μετρική για την ανίχνευση λεκτικών ομοιοτήτων σε ελληνικά κείμενα.

Length Ratio: Λόγος του μήκους της παραγόμενης περίληψης προς το μήκος της χρυσής περίληψης. Τιμές κοντά στο 1.0 υποδηλώνουν ότι το μοντέλο παράγει περιλήψεις παρόμοιου μήκους με τις πρότυπες περιλήψεις.

Παραγωγή περιλήψεων με χρήση TN

Keyword Overlap: Λόγος των κοινών λέξεων-κλειδιών μεταξύ της παραγόμενης και χρυσής περίληψης. Υποδηλώνει τη διατήρηση σημαντικού λεξιλογίου.

3.8 Αποτελέσματα

3.8.1 Σύνολο δεδομένων αξιολόγησης

Μετά την ομαδοποίηση των απαντήσεων, το τελικό σύνολο αξιολόγησης περιλαμβάνει 126 περιπτώσεις περίληψης, δηλαδή 42 Προγράμματα Σπουδών και τρεις κατηγορίες ερωτήσεων για κάθε πρόγραμμα: προτάσεις βελτίωσης (imp), αρνητικά στοιχεία (neg) και θετικά στοιχεία (pos). Κάθε περίπτωση αντιστοιχεί σε ένα ζεύγος Προγράμματος Σπουδών και κατηγορίας ερώτησης. Επομένως, η παραγωγή περίληψης δεν γίνεται ανά μεμονωμένο σχόλιο φοιτητή, αλλά πάνω στο σύνολο των σχολίων που ανήκουν στο ίδιο Πρόγραμμα Σπουδών και στην ίδια ερώτηση.

Τα αρχικά αρχεία απαντήσεων περιείχαν επίσης το Πρόγραμμα Σπουδών ΔΠ. Το ΔΠ δεν συμπεριλήφθηκε στην ποσοτική αξιολόγηση, επειδή δεν υπάρχουν διαθέσιμες πρότυπες περιλήψεις για τις τρεις αντίστοιχες κατηγορίες ερώτησης (ΔΠ-imp, ΔΠ-neg, ΔΠ-pos). Ο αποκλεισμός αυτός είναι τεχνικά καταγεγραμμένος στο αρχείο results/input_coverage_report.csv και εφαρμόστηκε ώστε κάθε παραγόμενη περίληψη να συγκρίνεται μόνο με υπαρκτή περίληψη αναφοράς.

Η μετρική BLEU υπολογίζεται με τη βιβλιοθήκη sacrebleu και παρουσιάζεται στην κλίμακα 0-100. Οι υπόλοιπες μετρικές του πίνακα, όπως η σημασιολογική ομοιότητα Sentence-BERT, η ομοιότητα συνημιτόνου, το length ratio και το keyword overlap, παρουσιάζονται ως αναλογίες. Για τα radar charts το BLEU κανονικοποιείται ως BLEU/100, ώστε να μπορεί να συγκριθεί οπτικά με τις υπόλοιπες μετρικές.

Παραγωγή περιλήψεων με χρήση TN

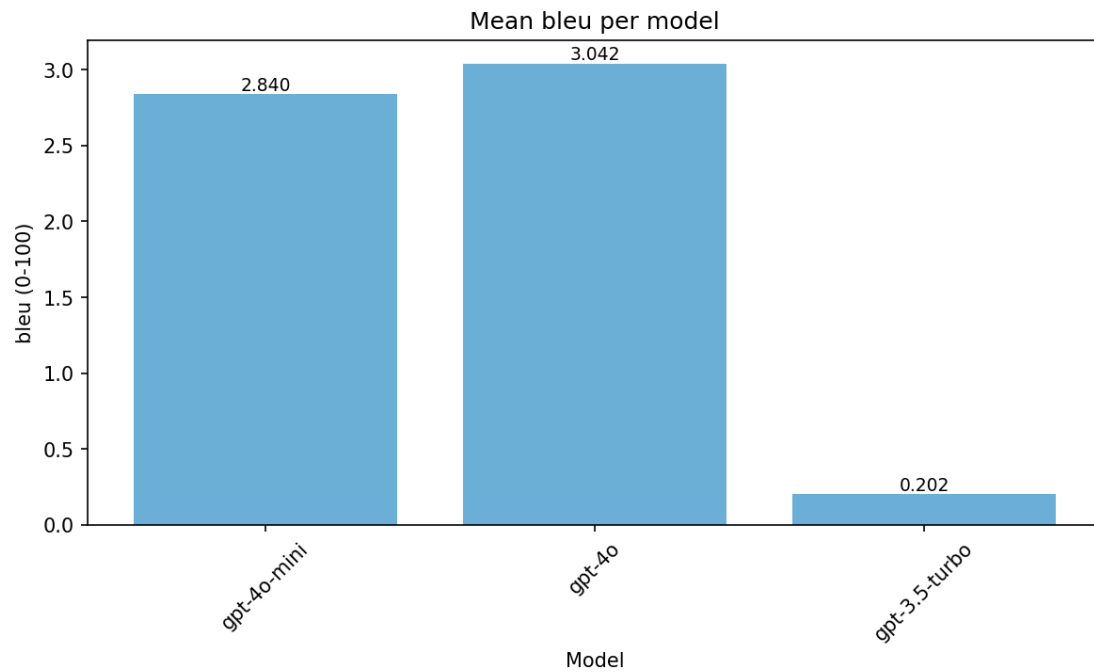
3.8.2 Μέσες μετρικές ανά μοντέλο

Μοντέλο	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	SBERT	Cosine	Length ratio	Keyword overlap
gpt-4o	3.0418	0.0159	0.0119	0.0159	0.7229	0.9790	0.5326	0.2008
gpt-4o-mini	2.8397	0.0370	0.0106	0.0370	0.7252	0.9790	0.5542	0.2015
gpt-3.5-turbo	0.2022	0.0032	0.0000	0.0032	0.6743	0.9664	0.2103	0.0987

Πίνακας 3.1: Μέσες τιμές μετρικών ανά μοντέλο.

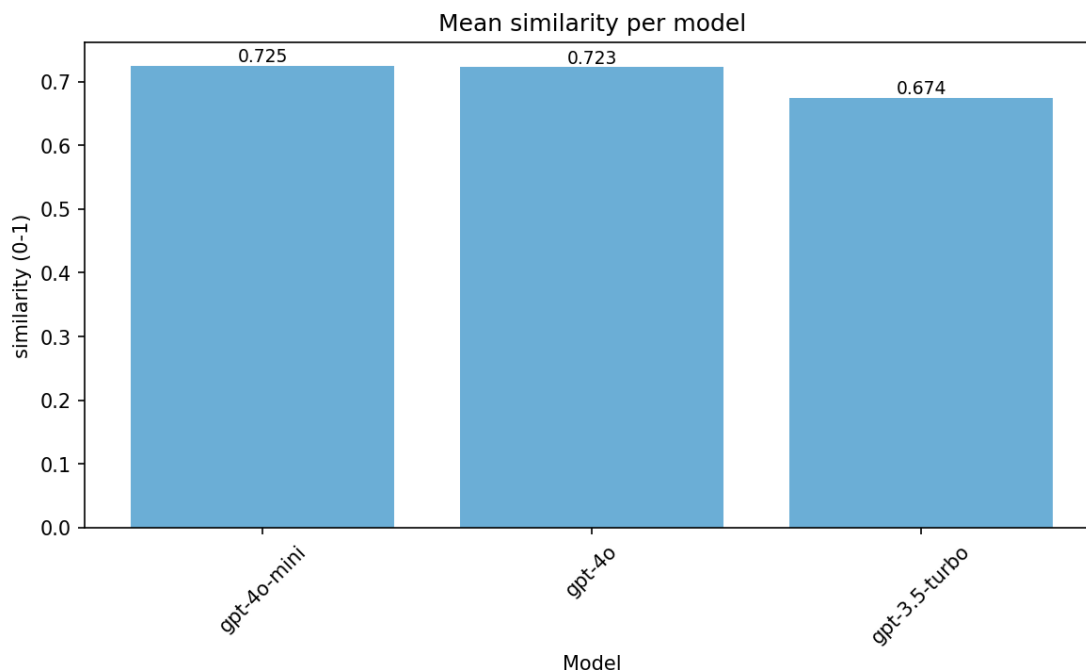
Τα αποτελέσματα δείχνουν ότι τα δύο νεότερα μοντέλα εμφανίζουν σαφώς υψηλότερη απόδοση από το GPT-3.5-turbo. Το GPT-4o έχει την υψηλότερη μέση τιμή BLEU (3.0418), ενώ το GPT-4o-mini εμφανίζει οριακά υψηλότερη σημασιολογική ομοιότητα Sentence-BERT (0.7252). Η διαφορά μεταξύ GPT-4o και GPT-4o-mini είναι μικρή στις περισσότερες μετρικές, γεγονός που δείχνει ότι το μικρότερο μοντέλο παράγει ανταγωνιστικές περιλήψεις στο συγκεκριμένο σύνολο δεδομένων. Αντίθετα, το GPT-3.5-turbo έχει χαμηλότερες τιμές τόσο στο BLEU όσο και στη σημασιολογική ομοιότητα και παράγει αισθητά συντομότερες περιλήψεις, όπως φαίνεται από το χαμηλό length ratio.

Παραγωγή περιλήψεων με χρήση TN



Σχήμα 3.1: Σύγκριση μέσης τιμής BLEU ανά μοντέλο.

Παραγωγή περιλήψεων με χρήση TN



Σχήμα 3.2: Σύγκριση μέσης σημασιολογικής ομοιότητας Sentence-BERT ανά μοντέλο.

Η χαμηλή απόλυτη τιμή BLEU είναι αναμενόμενη σε εργασίες abstractive summarization, επειδή οι παραγόμενες περιλήψεις δεν αντιγράφουν απαραίτητα τις ίδιες ακολουθίες λέξεων με τις πρότυπες περιλήψεις. Για τον λόγο αυτό, η ερμηνεία των αποτελεσμάτων βασίζεται κυρίως στον συνδυασμό BLEU, ROUGE και σημασιολογικών μετρικών, και όχι αποκλειστικά στη λεξική ταύτιση.

Παραγωγή περιλήψεων με χρήση TN

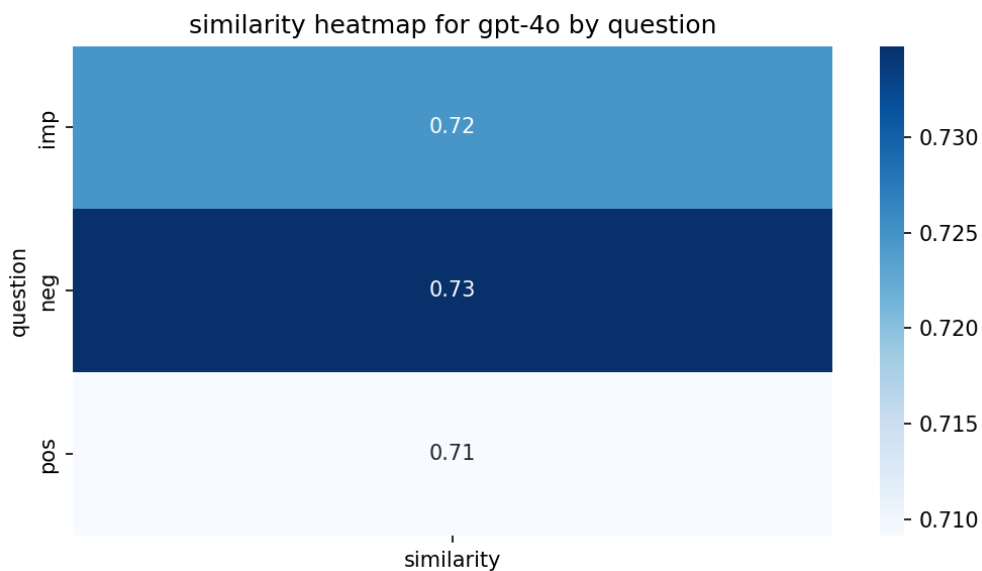
3.8.3 Ανάλυση ανά κατηγορία ερώτησης

Μοντέλο	Κατηγορία ερώτησης	BLEU	SBERT similarity
gpt-4o	Σημαντικά στοιχεία	3.1518	0.7247
gpt-4o	Αρνητικά στοιχεία	3.1543	0.7347
gpt-4o	Θετικά στοιχεία	2.8193	0.7091
gpt-4o-mini	Σημαντικά στοιχεία	3.3588	0.7105
gpt-4o-mini	Αρνητικά στοιχεία	2.3353	0.7292
gpt-4o-mini	Θετικά στοιχεία	2.8251	0.7357
gpt-3.5-turbo	Σημαντικά στοιχεία	0.2341	0.6616
gpt-3.5-turbo	Αρνητικά στοιχεία	0.1325	0.6723
gpt-3.5-turbo	Θετικά στοιχεία	0.2399	0.6891

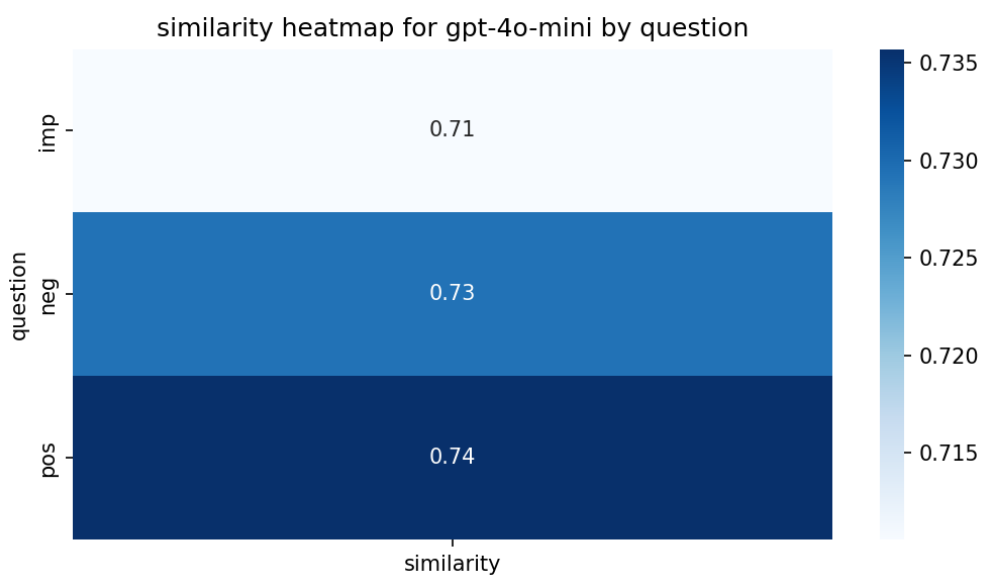
Πίνακας 3.2: Μέσες τιμές BLEU και Sentence-BERT similarity ανά κατηγορία ερώτησης.

Η ανάλυση ανά κατηγορία ερώτησης δείχνει ότι η απόδοση δεν είναι ίδια για όλα τα είδη σχολίων. Στο GPT-4o η υψηλότερη σημασιολογική ομοιότητα εμφανίζεται στα αρνητικά στοιχεία (0.7347), ενώ στο GPT-4o-mini εμφανίζεται στα θετικά στοιχεία (0.7357). Το GPT-3.5-turbo παρουσιάζει χαμηλότερες τιμές σε όλες τις κατηγορίες, με μικρή υπεροχή στα θετικά στοιχεία. Οι διαφορές αυτές είναι λογικές, επειδή κάθε κατηγορία ερώτησης έχει διαφορετικό βαθμό ποικιλίας και διαφορετική γλωσσική διατύπωση από τους φοιτητές.

Παραγωγή περιλήψεων με χρήση TN

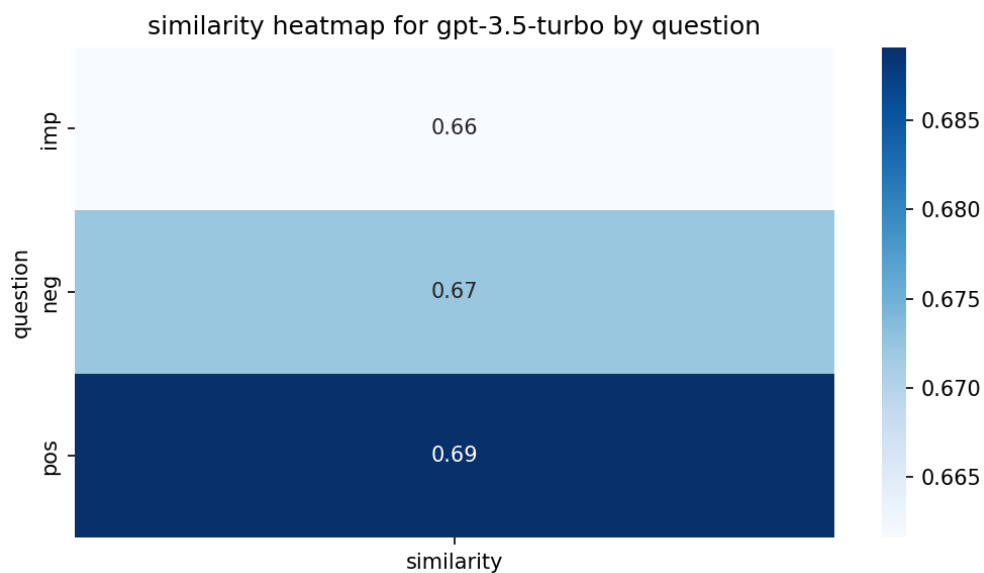


Σχήμα 3.3: Heatmap σημασιολογικής ομοιότητας ανά κατηγορία ερώτησης για το gpt-4o.



Σχήμα 3.4: Heatmap σημασιολογικής ομοιότητας ανά κατηγορία ερώτησης για το gpt-4o-mini.

Παραγωγή περιλήψεων με χρήση TN



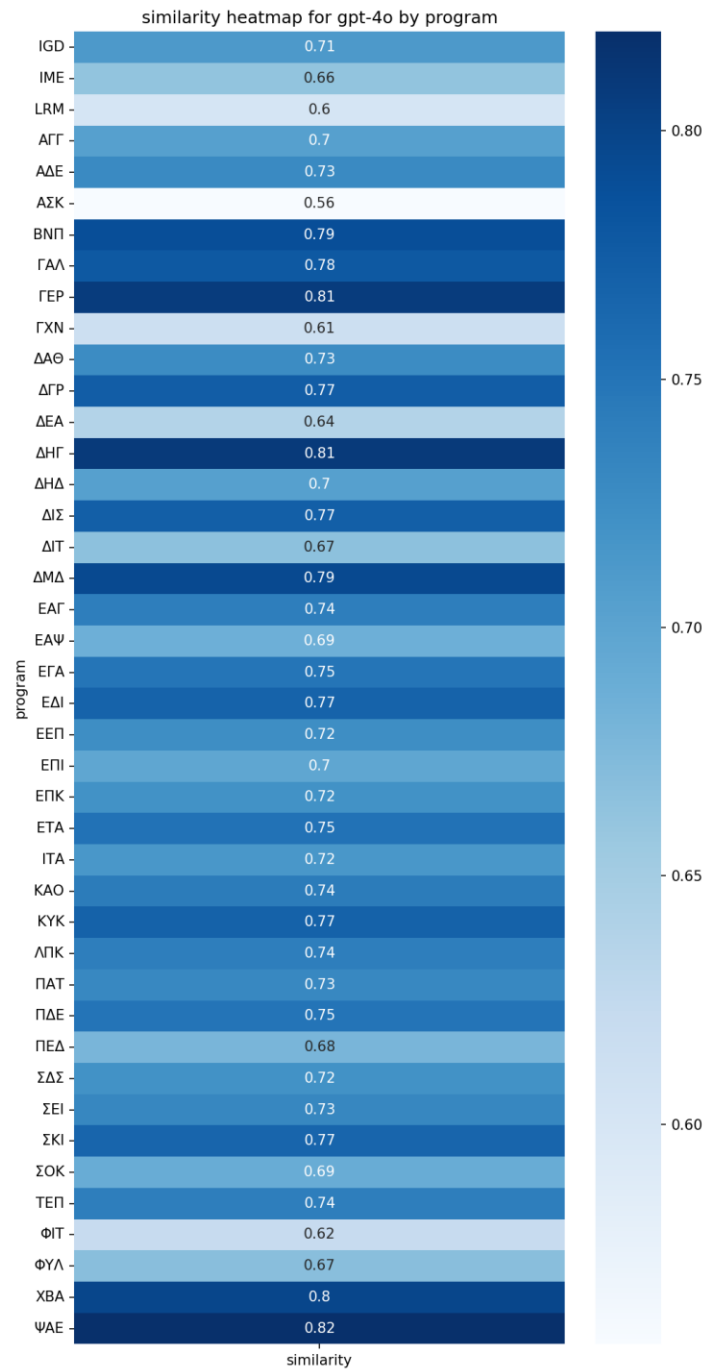
Σχήμα 3.5: Heatmap σημασιολογικής ομοιότητας ανά κατηγορία ερώτησης για το gpt-3.5-turbo.

Παραγωγή περιλήψεων με χρήση TN

3.8.4 Ανάλυση ανά Πρόγραμμα Σπουδών

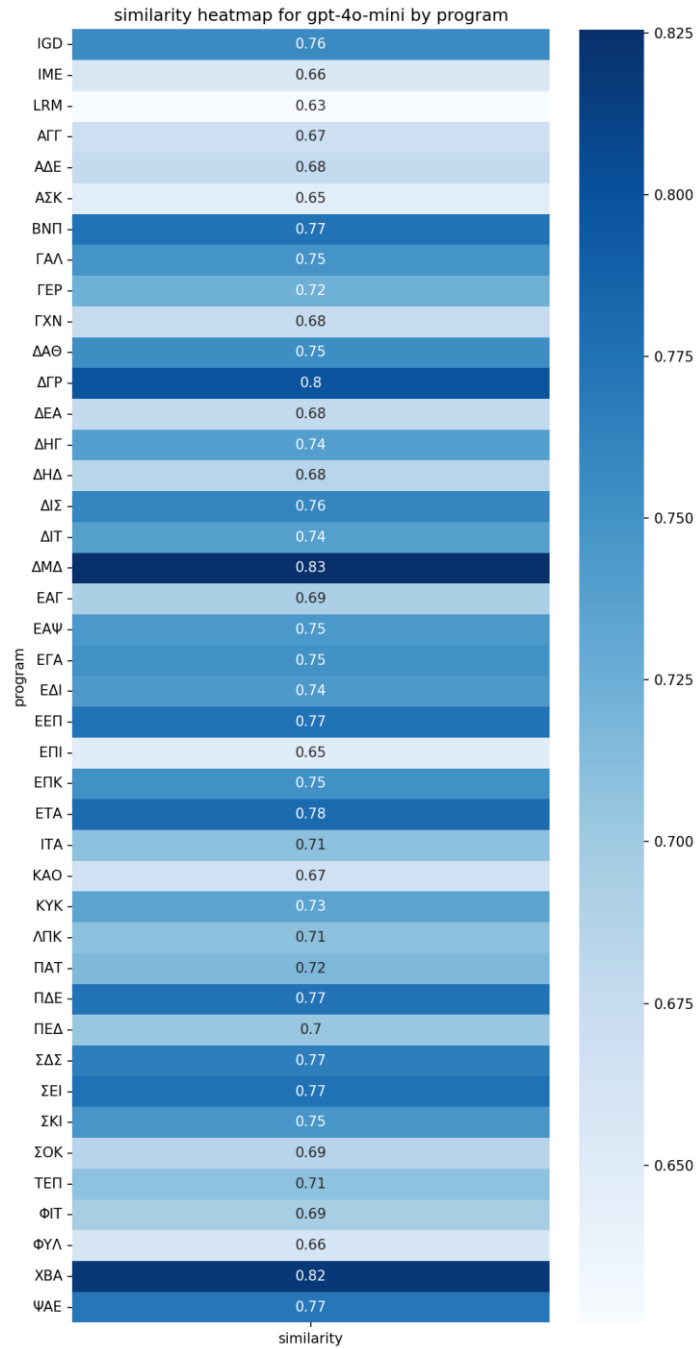
Τα heatmaps ανά Πρόγραμμα Σπουδών δείχνουν τη μέση σημασιολογική ομοιότητα για κάθε πρόγραμμα, υπολογισμένη πάνω στις τρεις κατηγορίες ερωτήσεων. Η παρουσία διακυμάνσεων ανά πρόγραμμα είναι αναμενόμενη, καθώς ο αριθμός, το μήκος και η ομοιογένεια των σχολίων διαφέρουν από πρόγραμμα σε πρόγραμμα. Παρ' όλα αυτά, η γενική εικόνα παραμένει συνεπής: τα GPT-4o και GPT-4o-mini διατηρούν υψηλότερη και πιο σταθερή σημασιολογική ομοιότητα από το GPT-3.5-turbo.

Παραγωγή περιλήψεων με χρήση TN



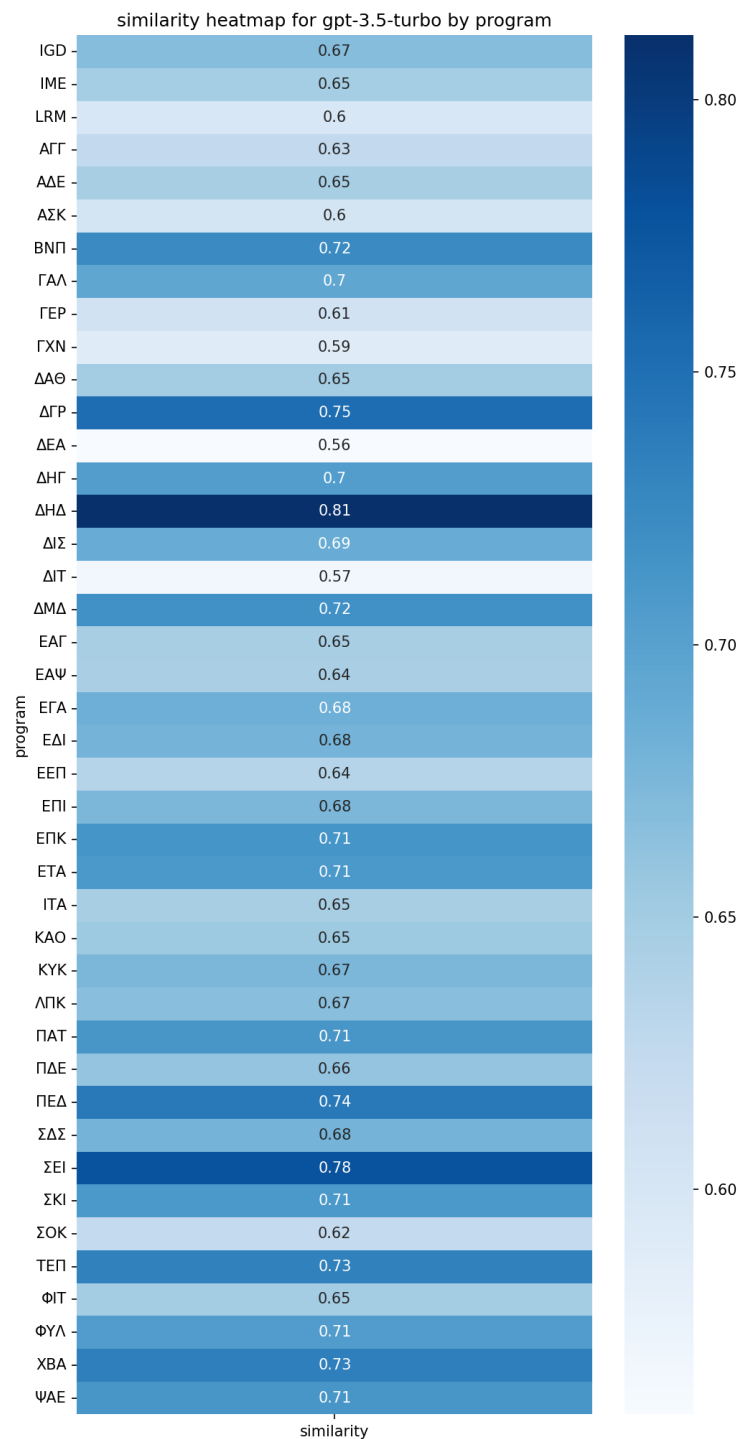
Σχήμα 3.6: Heatmap σημασιολογικής ομοιότητας ανά Πρόγραμμα Σπουδών για το gpt-4o.

Παραγωγή περιλήψεων με χρήση TN



Σχήμα 3.7: Heatmap σημασιολογικής ομοιότητας ανά Πρόγραμμα Σπουδών για το gpt-4o-mini.

Παραγωγή περιλήψεων με χρήση TN

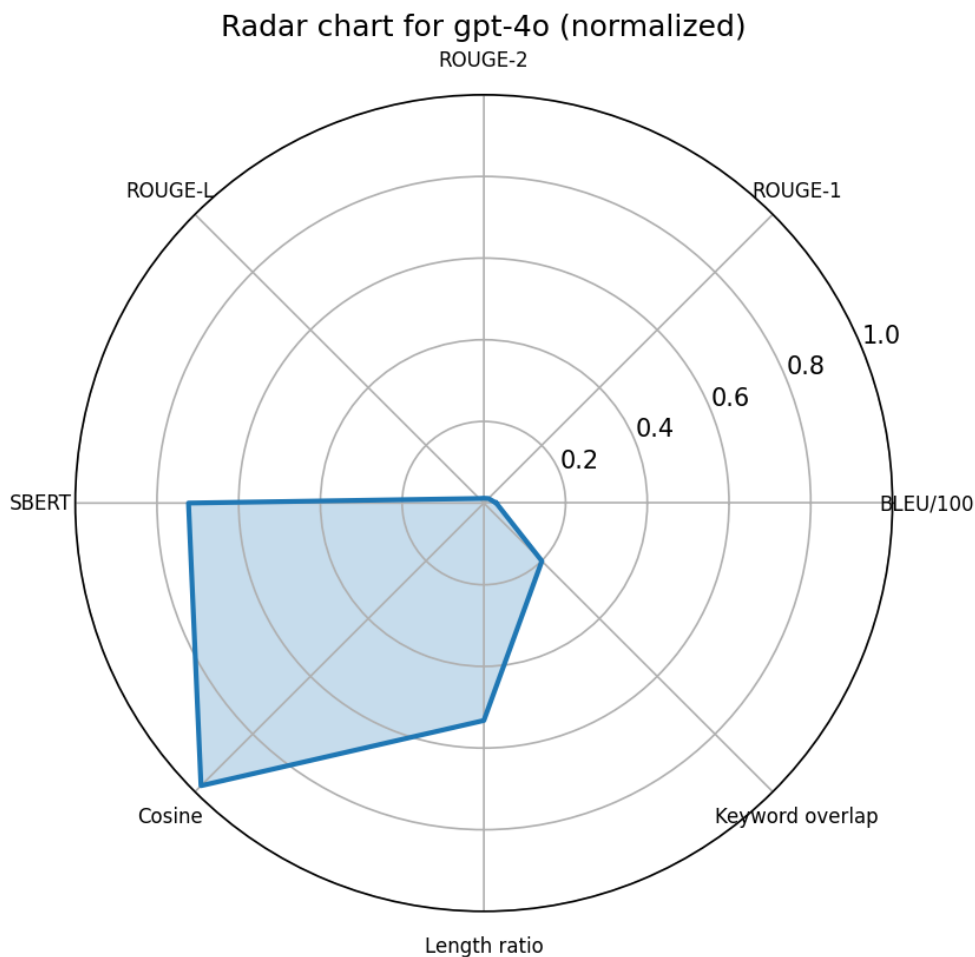


Σχήμα 3.8: Heatmap σημασιολογικής ομοιότητας ανά Πρόγραμμα Σπουδών για το gpt-3.5-turbo.

Παραγωγή περιλήψεων με χρήση TN

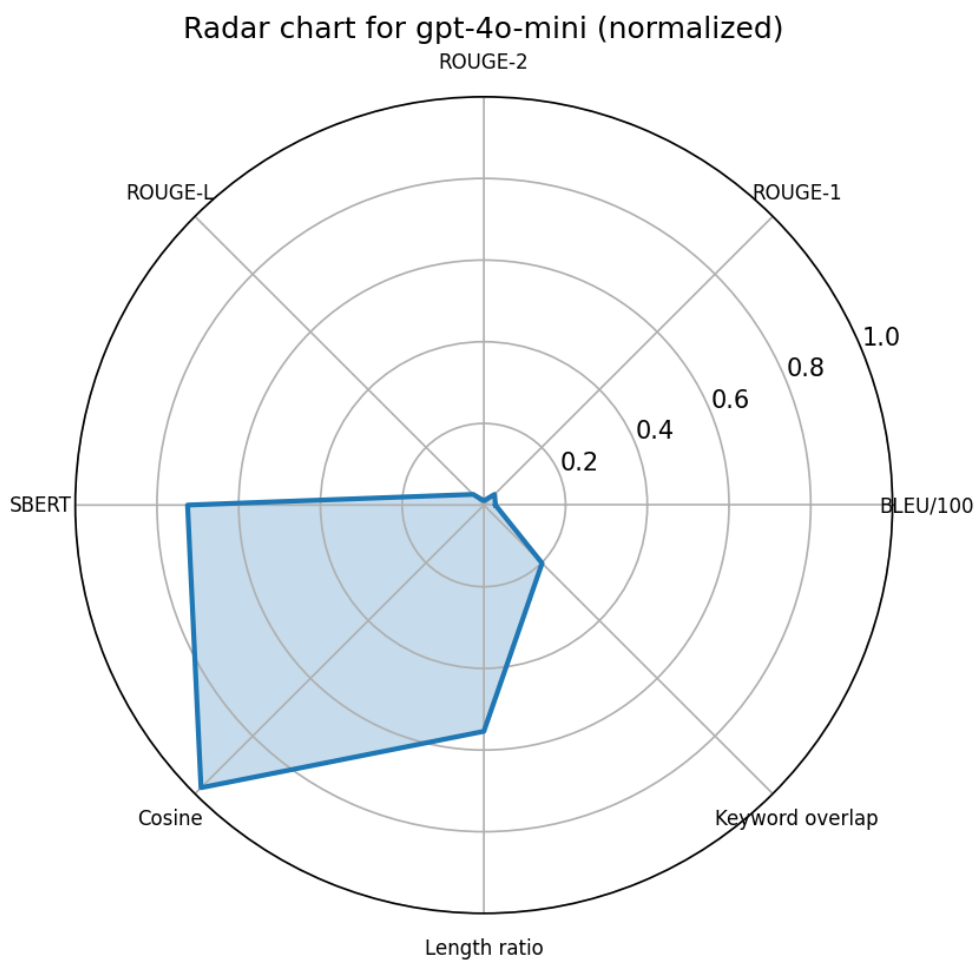
3.8.5 Πολυμετρική σύγκριση με radar charts

Τα radar charts χρησιμοποιούνται για συνοπτική σύγκριση πολλών μετρικών στο ίδιο σχήμα. Επειδή το BLEU έχει διαφορετική κλίμακα από τις υπόλοιπες μετρικές, στα συγκεκριμένα σχήματα εμφανίζεται ως BLEU/100. Με αυτόν τον τρόπο αποφεύγεται η οπτική υπερεκτίμηση του BLEU και το διάγραμμα αποτυπώνει πιο ισορροπημένα το προφίλ κάθε μοντέλου.



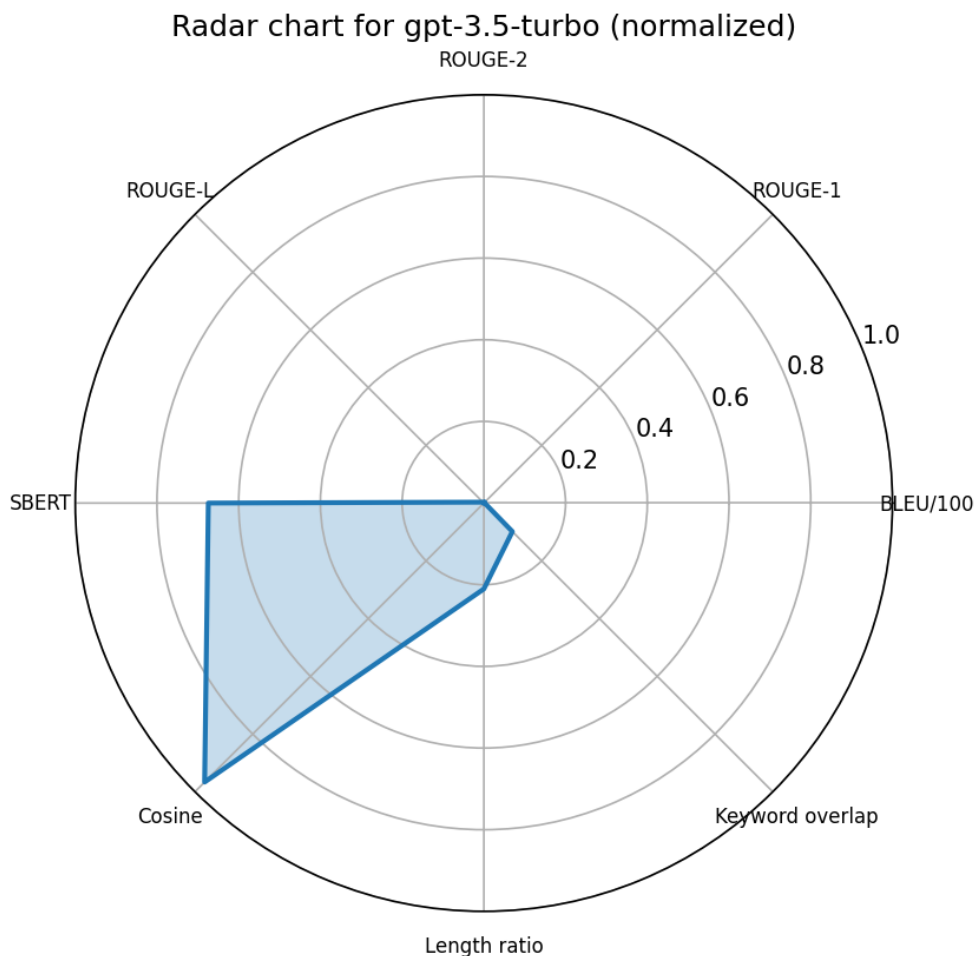
Σχήμα 3.9: Κανονικοποιημένο radar chart με όλες τις μετρικές για το gpt-4o.

Παραγωγή περιλήψεων με χρήση TN



Σχήμα 3.10: Κανονικοποιημένο radar chart με όλες τις μετρικές για το gpt-4o-mini.

Παραγωγή περιλήψεων με χρήση TN



Σχήμα 3.11: Κανονικοποιημένο radar chart με όλες τις μετρικές για το gpt-3.5-turbo.

Η συνολική εικόνα των radar charts επιβεβαιώνει ότι το GPT-4o και το GPT-4o-mini έχουν πιο ισορροπημένο προφίλ απόδοσης. Το GPT-4o-mini είναι ιδιαίτερα ανταγωνιστικό σε σχέση με το GPT-4o, κυρίως στη σημασιολογική ομοιότητα, ενώ το GPT-3.5-turbo υστερεί περισσότερο στο μήκος και στη λεξική επικάλυση των περιλήψεων. Συνεπώς, για το συγκεκριμένο πρόβλημα παραγωγής ελληνικών περιλήψεων από απαντήσεις ερωτηματολογίων, τα δύο νεότερα μοντέλα είναι καταλληλότερα για χρήση στο προτεινόμενο pipeline.

Παραγωγή περιλήψεων με χρήση TN

3.9 Ερμηνεία και Συμπεράσματα

Κατά την ερμηνεία των αποτελεσμάτων πρέπει να λαμβάνεται υπόψη και το μέγεθος της ομάδας σχολίων ανά Πρόγραμμα Σπουδών. Σε προγράμματα με λίγα σχόλια, η παραγόμενη περίληψη μπορεί να επηρεάζεται έντονα από μία ή δύο μεμονωμένες διατυπώσεις, άρα η απόκλιση από τη χρυσή περίληψη μπορεί να είναι μεγαλύτερη ή λιγότερο σταθερή. Αντίθετα, σε προγράμματα με πολλά σχόλια, τα επαναλαμβανόμενα μοτίβα εμφανίζονται καθαρότερα και το μοντέλο έχει περισσότερα στοιχεία για να εντοπίσει τα κυρίαρχα θέματα, αν και αυξάνεται η ανάγκη σωστής διαχείρισης μεγάλου μήκους εισόδου.

3.9.1 Σύγκριση μοντέλων

Με βάση τα τελικά αποτελέσματα, τα GPT-4o και GPT-4o-mini εμφανίζουν την καλύτερη συνολική συμπεριφορά. Το GPT-4o έχει την υψηλότερη μέση τιμή BLEU (3.0418), ενώ το GPT-4o-mini παρουσιάζει την υψηλότερη μέση σημασιολογική ομοιότητα Sentence-BERT (0.7252). Η διαφορά τους είναι μικρή και δείχνει ότι το GPT-4o-mini μπορεί να αποτελέσει αποδοτική λύση για την παραγωγή περιλήψεων στα ελληνικά, ειδικά όταν λαμβάνεται υπόψη η σχέση απόδοσης και υπολογιστικού κόστους. Το GPT-3.5-turbo έχει αισθητά χαμηλότερη απόδοση, με BLEU 0.2022 και Sentence-BERT similarity 0.6743.

Η ερμηνεία του BLEU πρέπει να γίνεται με προσοχή, επειδή πρόκειται για μετρική λεξικής επικάλυψης. Στην αφαιρετική περίληψη, ένα μοντέλο μπορεί να αποδίδει σωστά το νόημα χωρίς να χρησιμοποιεί τις ίδιες ακριβώς λέξεις με την περίληψη αναφοράς. Για τον λόγο αυτό, οι σημασιολογικές μετρικές και η ποιοτική ανάγνωση των περιλήψεων είναι απαραίτητες για πιο ολοκληρωμένη αξιολόγηση.

3.9.2 Επίδραση της κατηγορίας ερώτησης

Η κατηγορία της ερώτησης επηρεάζει την απόδοση των μοντέλων. Στο GPT-4o, η υψηλότερη σημασιολογική ομοιότητα εμφανίζεται στα αρνητικά στοιχεία (0.7347), ενώ στο GPT-4o-mini εμφανίζεται στα θετικά στοιχεία (0.7357). Στο GPT-3.5-turbo οι τιμές είναι χαμηλότερες σε όλες τις κατηγορίες, με μέγιστη τιμή στα θετικά στοιχεία (0.6891). Αυτό δείχνει ότι οι διαφορετικοί

Παραγωγή περιλήψεων με χρήση TN

τύποι σχολίων δεν έχουν την ίδια δυσκολία σύνοψης και ότι η αξιολόγηση πρέπει να εξετάζει ξεχωριστά κάθε ερώτηση.

3.9.3 Επίδραση του Προγράμματος Σπουδών

Οι διαφορές ανά Πρόγραμμα Σπουδών πρέπει να ερμηνεύονται προσεκτικά, επειδή κάθε πρόγραμμα συνεισφέρει μόνο τρεις περιπτώσεις αξιολόγησης: μία για τις προτάσεις βελτίωσης, μία για τα αρνητικά στοιχεία και μία για τα θετικά στοιχεία. Επομένως, τα heatmaps είναι χρήσιμα για την οπτική ανίχνευση διαφοροποιήσεων, αλλά δεν επαρκούν από μόνα τους για ισχυρά στατιστικά συμπεράσματα ή για αυστηρή κατάταξη των Προγραμμάτων Σπουδών.

Παρά τον περιορισμένο αριθμό περιπτώσεων ανά πρόγραμμα, η γενική εικόνα είναι συνεπής με τη συνολική σύγκριση των μοντέλων: τα GPT-4o και GPT-4o-mini εμφανίζουν υψηλότερες σημασιολογικές ομοιότητες, ενώ το GPT-3.5-turbo παρουσιάζει πιο αδύναμο προφίλ, ιδίως ως προς το μήκος και τη λεξική επικάλυψη των παραγόμενων περιλήψεων.

3.9.4 Συνολικό συμπέρασμα

Η υλοποιημένη ροή επεξεργασίας καλύπτει πλέον την απαίτηση παραγωγής μίας περίληψης ανά Πρόγραμμα Σπουδών και ανά κατηγορία ερώτησης. Η αξιολόγηση βασίζεται σε 126 συγκρίσεις παραγόμενων και χρυσών περιλήψεων. Το ΔΙΠ δεν συμπεριλήφθηκε στην ποσοτική αξιολόγηση, επειδή δεν υπήρχαν διαθέσιμες πρότυπες περιλήψεις αναφοράς για το συγκεκριμένο πρόγραμμα. Με βάση τις μετρικές, τα GPT-4o και GPT-4o-mini είναι τα καταλληλότερα μοντέλα για το συγκεκριμένο έργο, με το GPT-4o-mini να παρουσιάζει πολύ ανταγωνιστική απόδοση στη σημασιολογική ομοιότητα.

Παραγωγή περιλήψεων με χρήση TN

3.10 Δομή και Λειτουργίες Κύριων Αρχείων Κώδικα

Η υλοποίηση πραγματοποιήθηκε σε γλώσσα Python. Χρησιμοποιήθηκαν βιβλιοθήκες για ανάγνωση και επεξεργασία δεδομένων από αρχεία Excel, για καθαρισμό και μετασχηματισμό κειμένου, για κλήση υπηρεσιών LLM μέσω αντίστοιχων API clients και για υπολογισμό μετρικών αξιολόγησης. Ενδεικτικά, η ροή αξιοποιεί εργαλεία όπως pandas/openpyxl για τα δεδομένα πινάκων, βιβλιοθήκες επεξεργασίας κειμένου και βιβλιοθήκες αξιολόγησης όπως ROUGE, BLEU, BERTScore ή σημασιολογικές μετρικές, ανάλογα με τη μετρική που υπολογίζεται σε κάθε πείραμα.

3.10.1 Αρχεία Προεπεξεργασίας

load_excel.py:

- load_student_responses(): Φορτώνει τα 3 αρχεία με απαντήσεις φοιτητών από "Data/Student Responses/"
- load_golden_summaries(): Φορτώνει τα 126 αρχεία χρυσών περιλήψεων από "Data/Golden Summaries/"
- _sorted_xlsx_files(): Ταξινομεί τα αρχεία ώστε να αποφευχθούν προσωρινά αρχεία (π.χ. ~\$....)

clean_text.py:

- basic_clean(): Κανονικοποιεί κείμενο (trim, κενά, κ.λπ.)

build_dataset.py:

- extract_program_and_question(): Εξάγει πρόγραμμα και κατηγορία ερώτησης από το όνομα του αρχείου
- build_dataset(): Κυρία συνάρτηση που:

Παραγωγή περιλήψεων με χρήση TN

- * Φορτώνει τις απαντήσεις φοιτητών και τις πρότυπες περιλήψεις
- * Ταξινομεί τα ονόματα στηλών ($\text{προσπ_name_sm} \rightarrow \text{program}$)
- * Καθαρίζει το κείμενο
- * Συνδυάζει τα δεδομένα με βάση το πρόγραμμα
- * Εξάγει το αποτέλεσμα σε CSV (`results/dataset.csv`)

3.10.2 Αρχεία LLM Clients

`openai_client.py`:

- `_load_api_key()`: Φορτώνει το OpenAI API key από `secrets.json`
- `get_client()`: Δημιουργεί ένα global OpenAI client (για αποφυγή επανάληψης αρχικοποίησης)
- `build_prompt()`: Δημιουργεί το user prompt με πλαίσιο προγράμματος/ερώτησης
- `generate_summary_openai()`: Κύρια συνάρτηση που κάνει κλήση στο OpenAI API
 - * Χρησιμοποιεί `SYSTEM_PROMPT` για τις αρχικές οδηγίες
 - * Χρησιμοποιεί `build_prompt()` για το context
 - * Επιστρέφει την παραγόμενη περίληψη

Άλλα αρχεία (`anthropic_client.py`, `mistral_client.py`, κ.λπ.): TODO placeholders για μελλοντική επέκταση

Παραγωγή περιλήψεων με χρήση TN

3.10.3 Αρχεία Περίληψης

generate_summary.py:

- generate_summaries_for_model(model_name, test_mode): Κύρια συνάρτηση που:
 - * Φορτώνει το dataset.csv
 - * Επαναλαμβάνει κάθε σειρά δεδομένων
 - * Κάνει κλήση στο generate_summary_openai() με τα κατάλληλα ορίσματα
 - * Σε test_mode, σταματά μετά από 200 περιλήψεις
 - * Αποθηκεύει τα αποτελέσματα σε CSV (results/summaries/summaries_{model}.csv)

prompts.py: Περιέχει τα βασικά prompt templates (παρά τη χρήση μέσω openai_client.py)

3.10.4 Αρχεία Αξιολόγησης

bleu.py:

- compute_bleu(reference, hypothesis): Υπολογίζει BLEU score χρησιμοποιώντας τη βιβλιοθήκη sacrebleu

rouge.py:

- compute_rouge(reference, hypothesis): Υπολογίζει ROUGE-1, ROUGE-2, ROUGE-L χρησιμοποιώντας το rouge_score
- Χρησιμοποιεί use_stemmer=False για αποφυγή NLTK dependency

bertscore.py:

Παραγωγή περιλήψεων με χρήση TN

- `compute_similarity(reference, hypothesis)`: Υπολογίζει Sentence-BERT similarity
- Χρησιμοποιεί το πολυγλωσσικό μοντέλο "paraphrase-multilingual-MiniLM-L12-v2"
- Αποθηκεύει το μοντέλο σε global μεταβλητή για αποφυγή επανα-φόρτωσης

`cosine_similarity.py`:

- `compute_cosine_similarity(reference, hypothesis)`: Υπολογίζει cosine similarity σε επίπεδο χαρακτήρα
- Χρησιμοποιεί Counter για μέτρηση χαρακτήρων

`evaluate_all.py`:

- `compute_length_ratio(reference, hypothesis)`: Μήκος generated / μήκος reference
- `compute_keyword_overlap(reference, hypothesis)`: Λόγος κοινών λέξεων
- `evaluate_model_summaries(model_name)`: Κύρια συνάρτηση που:
 - * Φορτώνει το `summaries_{model}.csv`
 - * Υπολογίζει όλες τις μετρικές για κάθε περίληψη
 - * Αποθηκεύει τα αποτελέσματα σε `metrics_{model}.csv`

`aggregate_metrics.py`:

- `aggregate_metrics(model_name)`: Αθροίζει τις μετρικές ανά:
 - * Μοντέλο
 - * Μοντέλο + Πρόγραμμα
 - * Μοντέλο + Κατηγορία ερώτησης

Παραγωγή περιλήψεων με χρήση TN

- Εξάγει τα αποτελέσματα σε aggregated_{model}.xlsx

visualize_metrics.py:

- load_metrics(model_name): Φορτώνει τις μετρικές CSV
- plot_mean_metric_per_model(): Bar charts με τις μέσες τιμές ανά μοντέλο
- plot_heatmap(): Heatmaps μετρικών ανά πρόγραμμα ή κατηγορία ερώτησης
- plot_radar(): Radar charts με όλες τις μετρικές ανά μοντέλο
- _render_or_save(): Δυναμικά επιλέγει να εμφανίσει ή να αποθηκεύσει τις υπογραφές

3.10.5 Αρχεία Υποδομής

config.py:

- PROJECT_ROOT: Ρίζα του έργου
- RAW_STUDENT_RESPONSES_DIR: Data/Student Responses/
- RAW_GOLDEN_SUMMARIES_DIR: Data/Golden Summaries/
- RESULTS_DIR: results/
- SUMMARIES_DIR: results/summaries/
- METRICS_DIR: results/metrics/

logging.py:

- get_logger(name): Δημιουργεί ένα logger με ομοιόμορφη μορφή [TIMESTAMP] [LEVEL] [NAME] MESSAGE

Παραγωγή περιλήψεων με χρήση TN

3.11 Κύριο Πρόγραμμα (main.py)

Το main.py παρέχει ένα διαδραστικό μενού που επιτρέπει στον χρήστη να:

1. Κατασκευή του συνόλου δεδομένων (build dataset)
2. Παραγωγή περιλήψεων για όλα τα μοντέλα (generate summaries)
3. Αξιολόγηση των περιλήψεων (evaluate summaries)
4. Εμφάνιση συγκεντρωτικών μετρικών (show aggregated metrics)
5. Εμφάνιση διαγραμμάτων (show plots)
6. Εναλλαγή λειτουργίας δοκιμής (toggle TEST MODE)
7. Έξοδος (exit)

Το κύριο πρόγραμμα επιτρέπει λειτουργία δοκιμής (test_mode=True) για ταχεία δοκιμή με περιορισμένο αριθμό δεδομένων (200 περιλήψεις).

Παραγωγή περιλήψεων με χρήση TN

3.12 Κύρια Ευρήματα και Συμπεράσματα

Τα αποτελέσματα της εργασίας δείχνουν ότι τα Μεγάλα Γλωσσικά Μοντέλα μπορούν να υποστηρίξουν ουσιαστικά την παραγωγή περιλήψεων ανοικτών απαντήσεων σε εκπαιδευτικά ερωτηματολόγια, αλλά η χρήση τους πρέπει να συνοδεύεται από σαφή αξιολόγηση και ανθρώπινη εποπτεία. Η σύγκριση με τις πρότυπες περιλήψεις παρέχει πρακτικό σημείο αναφοράς για το ποιο μοντέλο και ποιο prompt αποδίδουν πλησιέστερα στο επιθυμητό αποτέλεσμα.

Με βάση τη λογική της αξιολόγησης, μπορεί να προταθεί το μοντέλο που εμφανίζει τη μικρότερη απόκλιση από τις πρότυπες περιλήψεις και την πιο σταθερή συμπεριφορά στις διαφορετικές κατηγορίες ερωτήσεων. Η πρακτική διαφορά από τη χρυσή περίληψη δεν πρέπει να ερμηνεύεται μόνο αριθμητικά, αλλά και ποιοτικά: ακόμη και υψηλή βαθμολογία σε μια μετρική μπορεί να χρειάζεται έλεγχο ως προς την πληρότητα, την ακρίβεια και την απουσία παραπλανητικών διατυπώσεων.

Το prompt αναδείχθηκε ως ιδιαίτερα κρίσιμο στοιχείο της διαδικασίας. Μικρές αλλαγές στις οδηγίες μπορούν να επηρεάσουν το ύφος, το μήκος, τη δομή και το είδος των πληροφοριών που επιλέγει το μοντέλο. Για καλύτερα αποτελέσματα προτείνεται περαιτέρω πειραματισμός με prompts, σαφέστερη διαχείριση μεγάλων εισόδων με chunks, εμπλουτισμός των χρυσών περιλήψεων και συνδυασμός αυτόματων μετρικών με ανθρώπινη αξιολόγηση.

Κατά την υλοποίηση, βασικές δυσκολίες αποτέλεσαν η ετερογένεια των ανοικτών απαντήσεων, η διαχείριση ελληνικού κειμένου, ο περιορισμός μήκους εισόδου των μοντέλων και η ερμηνεία μετρικών που δεν συλλαμβάνουν πάντα τη σημασιολογική ισοδυναμία. Μελλοντικά, η έρευνα μπορεί να επεκταθεί με περισσότερα μοντέλα, μεγαλύτερο σύνολο δεδομένων, βελτιωμένη διαδικασία chunking και αξιολόγηση από ειδικούς της εκπαιδευτικής διαδικασίας.

Παραγωγή περιλήψεων με χρήση TN

1. Δυσκολία Μέτρησης Abstractive Περιλήψεων

Οι μετρικές όπως BLEU και ROUGE είναι χρήσιμες, αλλά έχουν περιορισμούς στην αξιολόγηση αφαιρετικών περιλήψεων. Τα μοντέλα παράγουν νέο κείμενο και δεν αντιγράφουν απαραίτητα τις ίδιες λέξεις με τις πρότυπες περιλήψεις, γεγονός που οδηγεί σε χαμηλές τιμές n-gram overlap. Για τον λόγο αυτό, η Sentence-BERT similarity αξιοποιείται συμπληρωματικά, επειδή αποτυπώνει καλύτερα τη σημασιολογική εγγύτητα.

2. Διακυμάνσεις ανάλογα με την κατηγορία ερώτησης

Η απόδοση διαφοροποιείται ανάλογα με το είδος της ερώτησης. Στο GPT-4o η υψηλότερη σημασιολογική ομοιότητα εμφανίζεται στα αρνητικά στοιχεία (0.7347), ενώ στο GPT-4o-mini εμφανίζεται στα θετικά στοιχεία (0.7357). Επομένως, η αξιολόγηση πρέπει να εξετάζει ξεχωριστά κάθε κατηγορία ερώτησης και να αποφεύγει γενικεύσεις που δεν τεκμηριώνονται από όλες τις μετρικές.

3. Περιορισμένος αριθμός περιπτώσεων ανά Πρόγραμμα Σπουδών

Κάθε Πρόγραμμα Σπουδών συνεισφέρει τρεις περιπτώσεις αξιολόγησης: μία για τις προτάσεις βελτίωσης, μία για τα αρνητικά στοιχεία και μία για τα θετικά στοιχεία. Αυτό είναι επαρκές για περιγραφική σύγκριση, αλλά όχι για ισχυρά στατιστικά συμπεράσματα ανά πρόγραμμα.

Παραγωγή περιλήψεων με χρήση TN

4. Σημασιολογική ομοιότητα των μοντέλων

Παρά τις διαφορές στο BLEU, τα GPT-4o και GPT-4o-mini εμφανίζουν πολύ κοντινή σημασιολογική απόδοση, με μέσες τιμές Sentence-BERT similarity 0.7229 και 0.7252 αντίστοιχα. Το GPT-3.5-turbo ακολουθεί με χαμηλότερη, αλλά ακόμα αξιοσημείωτη, σημασιολογική ομοιότητα 0.6743.

5. Ελληνικό λεξιλόγιο και character overlap

Οι υψηλές τιμές character-level cosine similarity, που κυμαίνονται περίπου στο 0.96-0.98, δείχνουν ότι οι περιλήψεις μοιράζονται σημαντικό κοινό λεξιλόγιο και μορφολογικά μοτίβα. Η μετρική αυτή είναι χρήσιμη ως συμπληρωματική ένδειξη, αλλά δεν αρκεί από μόνη της για αξιολόγηση σημασιολογικής ποιότητας.

Παραγωγή περιλήψεων με χρήση TN

Βιβλιογραφικές αναφορές

- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & others. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT*, 4171–4186.
- Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, 22, 457–479.
- Laban, P., Schnabel, T., Bennett, P. N., & Hearst, M. A. (2022). SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10, 163–177.
- Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. *Proceedings of EMNLP-IJCNLP*, 3730–3740.
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing Order into Text. *Proceedings of EMNLP*, 404–411.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.
- Papantoniou, K., & Tzitzikas, Y. (2020). NLP for the Greek Language: A Brief Survey. *Proceedings of SETN*, 101–109.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of ACL*, 311–318.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140), 1–67.
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP*, 3982–3992.

Παραγωγή περιλήψεων με χρήση TN

- Scialom, T., Dray, P.-A., Gallinari, P., Lamprier, S., Piwowarski, B., Staiano, J., & Wang, A. (2021). QuestEval: Summarization Asks for Fact-based Evaluation. *Proceedings of EMNLP*, 6594–6604.
- See, A., Liu, P. J., & Manning, C. D. (2017). Get To The Point: Summarization with Pointer-Generator Networks. *Proceedings of ACL*, 1073–1083.
- Shermis, M. D., & Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge.
- Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, 27.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017a). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017b). Attention is All You Need. *Advances in Neural Information Processing Systems*.
- Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019). MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. *Proceedings of EMNLP-IJCNLP*, 563–578.