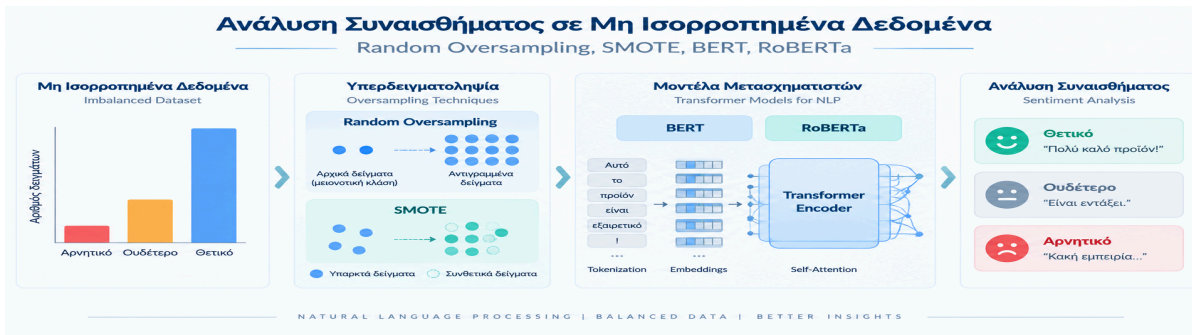


ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Εφαρμογή μεθόδων υπερδειγματοληψίας σε εφαρμογές ανάλυσης συναισθήματος με μη ισορροπημένα δεδομένα



Του φοιτητή
Γκρισπάνη Αλέξανδρου.
Αρ. Μητρώου: 185171

Επιβλέπων
Ονοματεπώνυμο Ακριτίδης
Λεωνίδας
Βαθμίδα Καθηγητής

Ημερομηνία: Σεπτέμβριος 2026

Τίτλος Δ.Ε. Εφαρμογή μεθόδων υπερδειγματοληψίας σε εφαρμογές ανάλυσης συναισθήματος με μη ισορροπημένα δεδομένα

Κωδικός Δ.Ε. 26191

Ονοματεπώνυμο φοιτητή/τών Γκρισπάνης Αλέξανδρος

Ονοματεπώνυμο εισηγητή Λεωνίδα Ακριτίδης

Ημερομηνία ανάληψης Δ.Ε. 16/3/26

Ημερομηνία περάτωσης Δ.Ε. 9/6/26

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία του φοιτητή Γκρισπάνη Αλέξανδρου που την εκπόνησε/αν. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης, ο συγγραφέας/δημιουργός εκχωρεί στο Διεθνές Πανεπιστήμιο της Ελλάδος άδεια χρήσης του δικαιώματος αναπαραγωγής, δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσης της εργασίας διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος. Η ανοικτή πρόσβαση στο πλήρες κείμενο της εργασίας, δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού, ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, πώληση, εμπορική χρήση, διανομή, έκδοση, μεταφόρτωση (downloading), ανάρτηση (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του Διεθνούς Πανεπιστημίου της Ελλάδος, δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα, εκ μέρους του Τμήματος.

Πρόλογος

Η παρούσα διπλωματική εργασία εκπονήθηκε με αντικείμενο την εφαρμογή μεθόδων υπερδειγματοληψίας σε προβλήματα ανάλυσης συναισθήματος με μη ισορροπημένα δεδομένα. Η επιλογή του θέματος προέκυψε από τη συχνή παρουσία έντονης ανισορροπίας σε πραγματικά σύνολα κριτικών, όπου ορισμένες κατηγορίες συναισθήματος ή βαθμολογίας εμφανίζονται σε πολύ μεγαλύτερο πλήθος από άλλες. Στο πλαίσιο της εργασίας μελετήθηκαν σύγχρονα μοντέλα Transformer και υλοποιήθηκαν πειραματικές ροές με BERT και RoBERTa. Η κύρια έμφαση δόθηκε στο Random Oversampling και στο SMOTE, καθώς και στον τρόπο με τον οποίο οι τεχνικές αυτές επηρεάζουν τις μειονοτικές κλάσεις και τις μετρικές αξιολόγησης. Η ανάπτυξη πραγματοποιήθηκε σε Python και Google Colab με χρήση PyTorch, Hugging Face Transformers, scikit-learn και imbalanced-learn. Ιδιαίτερη προσοχή δόθηκε στη διάκριση μεταξύ αποτελεσμάτων εκπαίδευσης και αποτελεσμάτων σε ανεξάρτητα validation δεδομένα, ώστε τα τελικά συμπεράσματα να αντανakλούν όσο το δυνατόν ακριβέστερα τις δυνατότητες και τους περιορισμούς των μεθόδων που εφαρμόστηκαν.

Περίληψη

Η ανάλυση συναισθήματος αποτελεί βασική εφαρμογή της Επεξεργασίας Φυσικής Γλώσσας, όμως η απόδοση των μοντέλων επηρεάζεται σημαντικά όταν οι διαθέσιμες κλάσεις παρουσιάζουν έντονα μη ισορροπημένη κατανομή. Σκοπός της παρούσας διπλωματικής εργασίας είναι η εφαρμογή και αξιολόγηση μεθόδων υπερδειγματοληψίας σε προβλήματα sentiment classification με χρήση μοντέλων Transformer. Μελετήθηκαν δύο κύριες προσεγγίσεις. Στην πρώτη χρησιμοποιήθηκε το BERT για ταξινόμηση εταιρικών κριτικών σε πέντε βαθμολογικές κλάσεις και εφαρμόστηκε Random Oversampling σε δύο πειραματικά σενάρια. Στο αυστηρότερο σενάριο, με ανεξάρτητο stratified validation set 12.000 δειγμάτων, το baseline BERT πέτυχε accuracy 88,12% και macro F1 0,57, ενώ μετά το Random Oversampling οι αντίστοιχες τιμές ήταν 87,23% και 0,5535. Στη δεύτερη προσέγγιση, αγγλόφωνες κριτικές μετατράπηκαν σε embeddings 768 διαστάσεων μέσω RoBERTa και το SMOTE εφαρμόστηκε αποκλειστικά στο training feature space πριν από Logistic Regression. Η baseline ταξινόμηση έδωσε accuracy 99,94% και macro F1 0,98 έναντι 99,85% και 0,96 μετά το SMOTE. Τα υψηλά αποτελέσματα του RoBERTa ερμηνεύονται με προσοχή επειδή οι στόχοι προήλθαν από pseudo-labels του ίδιου pretrained μοντέλου. Συνολικά, τα πειράματα δείχνουν ότι η εξισορρόπηση των training data δεν συνεπάγεται αυτόματα βελτίωση της γενίκευσης. Η αξιολόγηση σε μη ισορροπημένα δεδομένα πρέπει να βασίζεται σε macro F1, ανά κλάση precision/recall και confusion matrices και όχι μόνο στη συνολική accuracy.

Λέξεις-κλειδιά: Ανάλυση συναισθήματος, μη ισορροπημένα δεδομένα, υπερδειγματοληψία, Random Oversampling, SMOTE, BERT, RoBERTa.

Application of Oversampling Techniques for Sentiment Analysis on Imbalanced Datasets

«Γκρισπάνης Αλέξανδρος»

Abstract

Sentiment analysis is a major Natural Language Processing task, but model performance can be strongly affected when the available classes are highly imbalanced. The aim of this thesis is to apply and evaluate oversampling methods in sentiment classification tasks using Transformer-based representations. Two main experimental approaches were investigated. First, BERT was used to classify company reviews into five rating classes, and Random Oversampling was examined under two experimental scenarios. In the stricter scenario, which used an untouched stratified validation set of 12,000 samples, the baseline BERT model achieved 88.12% accuracy and a macro F1-score of 0.57, whereas the Random Oversampling model achieved 87.23% accuracy and a macro F1-score of 0.5535. Second, English-language reviews were represented as 768-dimensional RoBERTa embeddings and SMOTE was applied only to the training feature space before fitting a Logistic Regression classifier. The baseline classifier achieved 99.94% accuracy and a macro F1-score of 0.98, compared with 99.85% and 0.96 after SMOTE. The unusually high RoBERTa-based results are interpreted cautiously because the target sentiment labels were pseudo-labels produced by the same pretrained model family used to generate the embeddings. Overall, the experiments show that balancing the training distribution does not automatically improve generalization. For imbalanced sentiment data, evaluation should therefore focus on macro F1, per-class precision and recall, and confusion matrices rather than relying on accuracy alone.

Keywords: Sentiment analysis, imbalanced data, oversampling, Random Oversampling, SMOTE, BERT, RoBERTa.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής εργασίας για την καθοδήγηση και τις παρατηρήσεις του κατά τη διάρκεια της εκπόνησής της.

Περιεχόμενα

Πρόλογος	5
Περίληψη	6
Abstract	7
Ευχαριστίες	8
Περιεχόμενα	9
Κατάλογος Σχημάτων	15
Κατάλογος Πινάκων	15
Συντομογραφίες	15
Κεφάλαιο 1ο: Εισαγωγή	1
1.1 Εισαγωγή	1
1.2 Ανάλυση συναισθήματος	1
1.3 Το πρόβλημα της μη ισορροπημένης κατανομής δεδομένων	2
1.4 Κίνητρο της εργασίας	2
1.5 Σκοπός της διπλωματικής εργασίας	3
1.6 Στόχοι της εργασίας	3
1.7 Ερευνητικά ερωτήματα	4
1.7.1 Ερευνητικό ερώτημα 1: Επίδραση της ανισορροπίας των κλάσεων	4
1.7.2 Ερευνητικό ερώτημα 2: Επίδραση του Random Oversampling	4
1.7.3 Ερευνητικό ερώτημα 3: Επίδραση του SMOTE σε Transformer embeddings	4
1.7.4 Ερευνητικό ερώτημα 4: Κατάλληλες μετρικές αξιολόγησης	4
1.8 Μεθοδολογική προσέγγιση	4
1.9 Δομή της εργασίας	5
1.10 Επίλογος	6
Κεφάλαιο 2ο: Επεξεργασία Φυσικής Γλώσσας και Ανάλυση Συναισθήματος	7
2.1 Εισαγωγή	7
2.2 Επεξεργασία Φυσικής Γλώσσας	7
2.2.1 Ορισμός της Επεξεργασίας Φυσικής Γλώσσας	7
2.2.2 Βασικές εφαρμογές NLP	8
2.3 Ανάλυση συναισθήματος	8
2.3.1 Ορισμός	8
2.3.2 Κατηγορίες ανάλυσης συναισθήματος	9
2.3.3 Επίπεδα ανάλυσης συναισθήματος	9
2.4 Προσεγγίσεις στην ανάλυση συναισθήματος	10

2.4.1 Lexicon-based προσεγγίσεις	10
2.4.2 Παραδοσιακές τεχνικές Machine Learning	10
2.4.3 Deep Learning	11
2.4.4 Transformer-based προσεγγίσεις	11
2.5 Προεπεξεργασία κειμένου	12
2.5.1 Καθαρισμός δεδομένων	12
2.5.2 Tokenization	13
2.5.3 Padding και truncation	13
2.5.4 Ανίχνευση γλώσσας	13
2.5.5 Διαχείριση μεγάλου μήκους κειμένων	13
2.6 Σύνολα δεδομένων κριτικών	14
2.7 Προκλήσεις στην ανάλυση συναισθήματος	15
2.7.1 Σαρκασμός	15
2.7.2 Άρνηση	15
2.7.3 Πολυσημία	15
2.7.4 Ουδέτερο συναίσθημα	15
2.7.5 Ανισορροπία κλάσεων	16
2.8 Επίλογος	16
Κεφάλαιο 3ο: Μη Ισορροπημένα Δεδομένα και Μέθοδοι Υπερδειγματοληψίας	17
3.1 Εισαγωγή	17
3.2 Μη ισορροπημένα σύνολα δεδομένων	17
3.2.1 Ορισμός του class imbalance	17
3.2.2 Majority class και minority class	18
3.2.3 Βαθμός ανισορροπίας	18
3.3 Επιπτώσεις της ανισορροπίας στην εκπαίδευση	19
3.3.1 Bias προς τη majority class	19
3.3.2 Μειωμένη αναγνώριση μειονοτικών κλάσεων	19
3.3.3 Το πρόβλημα της υψηλής αλλά παραπλανητικής accuracy	20
3.4 Μέθοδοι αντιμετώπισης μη ισορροπημένων δεδομένων	20
3.4.1 Undersampling	20
3.4.2 Oversampling	21
3.4.3 Class weighting	21
3.4.4 Data augmentation	21
3.5 Random Oversampling	22
3.5.1 Βασική λειτουργία	22

3.5.2 Αλγοριθμική περιγραφή	22
3.5.3 Πλεονεκτήματα	23
3.5.4 Μειονεκτήματα	23
3.6 Synthetic Minority Over-sampling Technique (SMOTE)	24
3.6.1 Βασική ιδέα	24
3.6.2 Επιλογή κοντινών γειτόνων	24
3.6.3 Δημιουργία συνθετικών δειγμάτων	24
3.6.4 Μαθηματική διατύπωση του SMOTE	25
3.6.5 Πλεονεκτήματα	25
3.6.6 Περιορισμοί	26
3.7 SMOTE και δεδομένα κειμένου	26
3.8 SMOTE σε διανυσματικές αναπαραστάσεις	26
3.8.1 Feature vectors	27
3.8.2 Embeddings	27
3.8.3 Εφαρμογή SMOTE σε Transformer embeddings	27
3.9 Σύγκριση Random Oversampling και SMOTE	28
3.10 Επίλογος	29
Κεφάλαιο 4ο: Transformers, BERT και RoBERTa	30
4.1 Εισαγωγή	30
4.2 Η αρχιτεκτονική Transformer	30
4.2.1 Encoder και Decoder	30
4.2.2 Self-Attention	31
4.2.3 Query, Key και Value	32
4.2.4 Multi-Head Attention	32
4.2.5 Positional Encoding	33
4.3 Προεκπαιδευμένα γλωσσικά μοντέλα	33
4.4 BERT	34
4.4.1 Αρχιτεκτονική BERT	34
4.4.2 Bidirectional representation	35
4.4.3 Pre-training	35
4.4.4 Fine-tuning	35
4.4.5 BERT για text classification	36
4.5 Tokenization στο BERT	36
4.6 Embeddings του BERT	36
4.7 RoBERTa	37

4.7.1 Βασική αρχιτεκτονική	37
4.7.2 Διαφορές BERT και RoBERTa	37
4.7.3 RoBERTa για sentiment analysis	39
4.8 CardiffNLP Twitter RoBERTa	39
4.9 Transformer embeddings για κλασικούς classifiers	40
4.10 BERT και RoBERTa σε μη ισορροπημένα δεδομένα	40
4.11 Επίλογος	41
Κεφάλαιο 5ο: Μεθοδολογία και Υλοποίηση	42
5.1 Εισαγωγή	42
5.2 Ερευνητική μεθοδολογία	42
5.3 Υπολογιστικό περιβάλλον	43
5.3.1 Google Colab	43
5.3.2 GPU acceleration	44
5.3.3 Python	44
5.4 Βιβλιοθήκες λογισμικού	44
5.5 Σύνολα δεδομένων	45
5.5.1 Δεδομένα κριτικών	45
5.5.2 Δομή των συνόλων δεδομένων	45
5.5.3 Κατανομή κλάσεων	45
5.5.4 Ανίχνευση και επιλογή αγγλικών κριτικών	46
5.6 Προεπεξεργασία δεδομένων	46
5.6.1 Αντιμετώπιση κενών τιμών	46
5.6.2 Μετατροπή δεδομένων σε κείμενο	46
5.6.3 Language filtering	47
5.6.4 Tokenization	47
5.6.5 Truncation	47
5.7 Πείραμα A: BERT χωρίς oversampling	47
5.7.1 Αρχιτεκτονική	47
5.7.2 Παράμετροι εκπαίδευσης	48
5.7.3 Διαδικασία εκπαίδευσης	48
5.8 Πείραμα B1: BERT με Random Oversampling στο πλήρες training dataset	48
5.8.1 Αρχική κατανομή του πλήρους training dataset	48
5.8.2 Εφαρμογή Random Oversampling στο πλήρες training dataset	49
5.8.3 Νέα εξισορροπημένη κατανομή κλάσεων	49
5.8.4 Εκπαίδευση BERT	49

5.8.5 Ανάλυση της κατανομής προβλέψεων στο test set	49
5.9 Πείραμα Β2: BERT με Random Oversampling και Train/Validation Split	50
5.9.1 Διαχωρισμός training και validation set	50
5.9.2 Baseline BERT χωρίς ROS στο training split	50
5.9.3 Εφαρμογή ROS αποκλειστικά στο training split	50
5.9.4 Εκπαίδευση BERT στο εξισορροπημένο training set	51
5.9.5 Αξιολόγηση στο ανέγγιχτο validation set	51
5.9.6 Σύγκριση Precision, Recall, F1-score και Confusion Matrices	52
5.10 Πείραμα Γ: RoBERTa για sentiment classification	52
5.10.1 Επιλογή μοντέλου	52
5.10.2 Παραγωγή προβλέψεων	52
5.10.3 Confidence scores	52
5.10.4 Αντιμετώπιση των κλάσεων και pseudo-labeling	53
5.11 Πείραμα Δ: RoBERTa Embeddings με SMOTE	53
5.11.1 Παραγωγή embeddings	53
5.11.2 Διαχωρισμός training και validation/test δεδομένων	53
5.11.3 Αρχική ανισορροπία των embeddings	54
5.11.4 Εφαρμογή SMOTE στο training set	54
5.11.5 Παραγωγή balanced embedding dataset	54
5.11.6 Logistic Regression	54
5.11.7 Classification pipeline	55
5.12 Μέτρα αξιολόγησης	55
5.12.1 Accuracy	55
5.12.2 Precision	56
5.12.3 Recall	56
5.12.4 F1-score	56
5.12.5 Macro F1	56
5.12.6 Weighted F1	56
5.12.7 Confusion Matrix	56
5.13 Επίλογος	57
Κεφάλαιο 6ο: Πειραματικά Αποτελέσματα και Αξιολόγηση	58
6.1 Εισαγωγή	58
6.2 Ανάλυση αρχικής κατανομής δεδομένων	58
6.3 Αποτελέσματα BERT baseline	59
6.4 Αποτελέσματα πρώτης εφαρμογής Random Oversampling	59

6.5 Αποτελέσματα δεύτερης εφαρμογής Random Oversampling με Train/Validation Split	60
6.6 Συγκριτική αξιολόγηση των δύο πειραμάτων Random Oversampling	61
6.7 Αποτελέσματα RoBERTa	62
6.8 Κατανομή συναισθημάτων RoBERTa	62
6.9 Αποτελέσματα RoBERTa embeddings χωρίς SMOTE	63
6.10 Αποτελέσματα RoBERTa embeddings με SMOTE	63
6.11 Συγκριτική αξιολόγηση SMOTE	64
6.12 Συγκριτική αξιολόγηση όλων των πειραμάτων	65
6.13 Ανάλυση Confusion Matrices	66
6.14 Επίδραση της υπερδειγματοληψίας στις μειονοτικές κλάσεις	67
6.15 Συζήτηση αποτελεσμάτων	67
6.15.1 Random Oversampling	68
6.15.2 SMOTE	68
6.15.3 Πότε βοήθησε η υπερδειγματοληψία	68
6.15.4 Πότε δεν βοήθησε	68
6.15.5 Πιθανές αιτίες	68
6.16 Περιορισμοί της πειραματικής διαδικασίας	69
6.17 Επίλογος	69
Κεφάλαιο 7ο: Συμπεράσματα και Προτάσεις Βελτίωσης	71
7.1 Εισαγωγή	71
7.2 Ανακεφαλαίωση της εργασίας	71
7.3 Απάντηση στα ερευνητικά ερωτήματα	71
7.3.1 Επίδραση του class imbalance	71
7.3.2 Επίδραση του Random Oversampling	72
7.3.3 Επίδραση του SMOTE	72
7.3.4 Κατάλληλες μετρικές αξιολόγησης	72
7.4 Κύρια συμπεράσματα	73
7.5 Περιορισμοί	73
7.6 Μελλοντικές επεκτάσεις	74
7.7 Επίλογος	75
ΒΙΒΛΙΟΓΡΑΦΙΑ	76
ΠΑΡΑΡΤΗΜΑ Α: Επιλεγμένος Πηγαίος Κώδικας	80

Κατάλογος Σχημάτων

Σχήμα 2.1: Βασική ροή ενός συστήματος ανάλυσης συναισθήματος.	9
Σχήμα 2.2: Βασικές προσεγγίσεις στην ανάλυση συναισθήματος.	12
Σχήμα 2.3: Ενδεικτική ροή προεπεξεργασίας κειμένου για Transformer-based ταξινόμηση.	14
Σχήμα 3.1: Ενδεικτική απεικόνιση μη ισορροπημένης κατανομής κλάσεων.	19
Σχήμα 3.2: Βασικές κατηγορίες μεθόδων αντιμετώπισης της ανισορροπίας κλάσεων.	22
Σχήμα 3.3: Αρχή λειτουργίας του Random Oversampling.	23
Σχήμα 3.4: Δημιουργία συνθετικού δείγματος με SMOTE μεταξύ γειτονικών minority vectors.	25
Σχήμα 3.5: Εφαρμογή SMOTE στον χώρο Transformer embeddings.	28
Σχήμα 4.1: Απλοποιημένη δομή ενός Transformer encoder block.	31
Σχήμα 4.2: Βασική ροή του μηχανισμού self-attention μέσω Query, Key και Value.	32
Σχήμα 4.3: Απλοποιημένο pipeline BERT για ταξινόμηση κριτικών σε πέντε ratings.	34
Σχήμα 5.1: Πειραματική ροή BERT με Random Oversampling μόνο στο training split.	51
Σχήμα 5.2: Πειραματική ροή RoBERTa embeddings, SMOTE και Logistic Regression.	55
Σχήμα 6.1: Αρχική κατανομή των πέντε rating classes στο BERT training dataset.	59
Σχήμα 6.2: Σύγκριση F1-score ανά rating class στο κοινό validation set.	61
Σχήμα 6.3: Ακραία ανισορροπία των sentiment pseudo-labels που παράγονται από το pretrained RoBERTa.	63
Σχήμα 6.4: Σύγκριση accuracy και macro F1 της Logistic Regression με και χωρίς SMOTE.	65
Σχήμα 6.5: Confusion matrix του BERT baseline στο validation set.	66
Σχήμα 6.6: Confusion matrix του BERT μετά το Random Oversampling στο ίδιο validation set.	67

Κατάλογος Πινάκων

Πίνακας 3.1: Συγκριτικά χαρακτηριστικά Random Oversampling και SMOTE	29
Πίνακας 4.1: Βασική σύγκριση BERT-base και RoBERTa-base	39
Πίνακας 5.1: Σύνολα δεδομένων και ρόλος τους στα πειράματα	45
Πίνακας 5.2: Βασικές παράμετροι των BERT πειραμάτων	48
Πίνακας 5.3: Κατανομές κλάσεων πριν και μετά το Random Oversampling	51
Πίνακας 5.4: Ρυθμίσεις του RoBERTa embeddings + SMOTE πειράματος	54
Πίνακας 6.1: Αρχική κατανομή κλάσεων στο BERT dataset	59
Πίνακας 6.2: Training metrics πριν και μετά το ROS στο πλήρες BERT training dataset	60
Πίνακας 6.3: Validation metrics BERT πριν και μετά το Random Oversampling	61
Πίνακας 6.4: Κατανομή των pseudo-labels στο RoBERTa+SMOTE dataset	63
Πίνακας 6.5: Confusion matrices του RoBERTa embedding classifier	64
Πίνακας 6.6: Validation metrics της Logistic Regression πριν και μετά το SMOTE	65
Πίνακας 6.7: Συνοπτική εικόνα των βασικών πειραματικών αποτελεσμάτων	65

Συντομογραφίες

Δ.Ε.	Διπλωματική Εργασία
ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο της Ελλάδος

NLP	Natural Language Processing
ML	Machine Learning
DL	Deep Learning
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
ROS	Random Over-Sampling
SMOTE	Synthetic Minority Over-sampling Technique
GPU	Graphics Processing Unit

LR Logistic Regression

F1 F1-score

CUDA Compute Unified Device Architecture

API Application Programming Interface

Κεφάλαιο 1ο: Εισαγωγή

1.1 Εισαγωγή

Η ραγδαία αύξηση του όγκου κειμενικών δεδομένων που παράγονται καθημερινά σε πλατφόρμες αξιολογήσεων, κοινωνικά δίκτυα, ηλεκτρονικά καταστήματα και διαδικτυακές υπηρεσίες έχει καταστήσει την αυτόματη εξαγωγή πληροφορίας από φυσική γλώσσα ιδιαίτερα σημαντική. Ένα μεγάλο μέρος αυτών των δεδομένων δεν περιγράφει απλώς γεγονότα, αλλά εκφράζει απόψεις, κρίσεις, προτιμήσεις και συναισθηματικές στάσεις των χρηστών. Η ανάλυση συναισθήματος αποτελεί τον κλάδο της Επεξεργασίας Φυσικής Γλώσσας που μελετά υπολογιστικές μεθόδους για την αναγνώριση και κατηγοριοποίηση τέτοιων υποκειμενικών πληροφοριών [1], [46].

Η ανάπτυξη μοντέλων ανάλυσης συναισθήματος έχει εξελιχθεί σημαντικά τα τελευταία χρόνια. Οι πρώτες προσεγγίσεις βασίζονταν κυρίως σε λεξικά συναισθήματος και σε κλασικούς αλγορίθμους μηχανικής μάθησης, ενώ οι σύγχρονες εφαρμογές χρησιμοποιούν βαθιά νευρωνικά δίκτυα και, ιδιαίτερα, αρχιτεκτονικές Transformer. Η εισαγωγή της αρχιτεκτονικής Transformer [4] και στη συνέχεια προεκπαιδευμένων γλωσσικών μοντέλων όπως το BERT [5] και το RoBERTa [6] επέτρεψε τη δημιουργία ισχυρών γλωσσικών αναπαραστάσεων, οι οποίες μπορούν να προσαρμοστούν σε προβλήματα ταξινόμησης κειμένου με σχετικά περιορισμένες μεταβολές στην αρχιτεκτονική του μοντέλου.

Παρά τη βελτίωση των μοντέλων, η ποιότητα της ταξινόμησης δεν εξαρτάται αποκλειστικά από την ισχύ του αλγορίθμου. Η κατανομή των διαθέσιμων δεδομένων αποτελεί κρίσιμο παράγοντα. Σε πραγματικά σύνολα αξιολογήσεων είναι συχνό οι θετικές ή υψηλές βαθμολογίες να εμφανίζονται πολύ συχνότερα από τις αρνητικές ή τις ουδέτερες. Σε αυτή την περίπτωση προκύπτει το πρόβλημα της μη ισορροπημένης κατανομής κλάσεων, το οποίο μπορεί να οδηγήσει ένα μοντέλο σε ισχυρή προτίμηση προς την πλειονοτική κλάση [2], [35]-[39], [50]. Η παρούσα διπλωματική εργασία επικεντρώνεται στην εφαρμογή και αξιολόγηση μεθόδων υπερδειγματοληψίας σε τέτοια περιβάλλοντα ανάλυσης συναισθήματος.

1.2 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος, γνωστή διεθνώς ως sentiment analysis ή opinion mining, αποσκοπεί στην αυτόματη αναγνώριση της στάσης που εκφράζεται σε ένα κείμενο. Η στάση αυτή μπορεί να αποτυπώνεται σε δυαδικές κατηγορίες, όπως θετικό και αρνητικό συναίσθημα, σε τρεις κατηγορίες, όπως θετικό, ουδέτερο και αρνητικό, ή σε περισσότερα επίπεδα που αντιστοιχούν σε βαθμολογίες και διαφορετική ένταση συναισθήματος. Η επιλογή της μορφής του προβλήματος εξαρτάται από τα διαθέσιμα δεδομένα και από τον στόχο της εφαρμογής.

Οι διαδικτυακές κριτικές αποτελούν χαρακτηριστικό παράδειγμα δεδομένων κατάλληλων για ανάλυση συναισθήματος. Μία κριτική μπορεί να περιέχει πληροφορίες σχετικά με την ποιότητα μιας υπηρεσίας, την ικανοποίηση του χρήστη, συγκεκριμένες θετικές ή αρνητικές εμπειρίες και συνολική αξιολόγηση. Η αυτοματοποιημένη επεξεργασία μεγάλου αριθμού τέτοιων κειμένων επιτρέπει την

εξαγωγή τάσεων που θα ήταν δύσκολο να αναγνωριστούν με χειροκίνητη ανάγνωση. Για τον λόγο αυτό, το sentiment analysis χρησιμοποιείται σε πλήθος εφαρμογών, από την παρακολούθηση της ικανοποίησης πελατών έως την ανάλυση σχολίων για προϊόντα και υπηρεσίες [1], [21], [23].

Ωστόσο, η ταξινόμηση του συναισθήματος δεν είναι απλή διαδικασία. Η φυσική γλώσσα περιλαμβάνει συμφραζόμενα, άρνηση, ειρωνεία, διαφορετικούς τρόπους έκφρασης και συχνά ταυτόχρονη παρουσία θετικών και αρνητικών στοιχείων στο ίδιο κείμενο. Επιπλέον, η απόδοση του μοντέλου επηρεάζεται από τη μορφή των δεδομένων εκπαίδευσης. Ένα σύνολο που περιέχει κυρίως μία κατηγορία συναισθήματος μπορεί να οδηγήσει σε υψηλή συνολική επίδοση χωρίς αντίστοιχα καλή αναγνώριση των σπανιότερων κατηγοριών. Η συγκεκριμένη παρατήρηση αποτελεί το σημείο σύνδεσης ανάμεσα στην ανάλυση συναισθήματος και στο πρόβλημα της ανισορροπίας κλάσεων που εξετάζεται στην εργασία [1], [7], [8].

1.3 Το πρόβλημα της μη ισορροπημένης κατανομής δεδομένων

Ένα σύνολο δεδομένων χαρακτηρίζεται ως μη ισορροπημένο όταν οι διαθέσιμες παρατηρήσεις δεν κατανέμονται ομοιόμορφα μεταξύ των κλάσεων. Σε ακραίες περιπτώσεις, μία κλάση μπορεί να περιέχει χιλιάδες δείγματα ενώ μία άλλη μόνο δεκάδες ή εκατοντάδες. Η πρώτη συνήθως αναφέρεται ως πλειονοτική κλάση, ενώ η δεύτερη ως μειονοτική. Η ανισορροπία αυτή δεν αποτελεί απλώς στατιστικό χαρακτηριστικό του συνόλου δεδομένων, αλλά μπορεί να μεταβάλει ουσιαστικά τη διαδικασία μάθησης ενός ταξινομητή [2].

Οι περισσότεροι τυπικοί αλγόριθμοι βελτιστοποιούν μία συνολική συνάρτηση σφάλματος. Όταν η πλειονοτική κλάση περιλαμβάνει το μεγαλύτερο μέρος των παραδειγμάτων, η σωστή πρόβλεψή της συνεισφέρει περισσότερο στη συνολική ελαχιστοποίηση του σφάλματος. Έτσι, ένα μοντέλο μπορεί να μάθει να προβλέπει πολύ καλά την κυρίαρχη κατηγορία, αλλά να αποτυγχάνει στις περιπτώσεις που εμφανίζονται σπάνια. Στην ανάλυση συναισθήματος αυτό μπορεί, για παράδειγμα, να σημαίνει πολύ καλή αναγνώριση θετικών αξιολογήσεων αλλά χαμηλή ανάκληση για αρνητικές ή ουδέτερες αξιολογήσεις.

Το πρόβλημα γίνεται ακόμη πιο εμφανές όταν η αξιολόγηση περιορίζεται στη μετρική accuracy. Εάν το 95% των παραδειγμάτων ανήκει σε μία μόνο κλάση, ένας εξαιρετικά απλοϊκός ταξινομητής που προβλέπει συνεχώς την ίδια κλάση θα μπορούσε θεωρητικά να εμφανίζει accuracy 95%, χωρίς να παρέχει ουσιαστική δυνατότητα διάκρισης των υπόλοιπων κατηγοριών. Για τον λόγο αυτό, η βιβλιογραφία του imbalanced learning δίνει ιδιαίτερη έμφαση σε μετρικές όπως precision, recall και F1-score, καθώς και σε μακρο-μέσους όρους που αποδίδουν ίσο βάρος στις επιμέρους κλάσεις [2].

Μία κατηγορία τεχνικών αντιμετώπισης του προβλήματος είναι η αναδειγματοληψία των δεδομένων. Η υπερδειγματοληψία αυξάνει την παρουσία της μειονοτικής κλάσης στο σύνολο εκπαίδευσης. Η απλούστερη εκδοχή είναι το Random Oversampling, όπου υφιστάμενα μειονοτικά δείγματα επαναλαμβάνονται. Μία πιο σύνθετη προσέγγιση είναι το SMOTE, το οποίο δημιουργεί συνθετικά δείγματα μέσω παρεμβολής στον χώρο χαρακτηριστικών μεταξύ γειτονικών παρατηρήσεων της μειονοτικής κλάσης [3], [40]-[44], [49], [50]. Η αποτελεσματικότητα αυτών των τεχνικών δεν είναι δεδομένη σε κάθε εφαρμογή και πρέπει να εξετάζεται πειραματικά.

1.4 Κίνητρο της εργασίας

Το ενδιαφέρον για το συγκεκριμένο θέμα προέκυψε αρχικά από την παρατήρηση της κατανομής των αξιολογήσεων σε πραγματικές διαδικτυακές πλατφόρμες. Σε πλατφόρμες που αφορούν προϊόντα, καταστήματα, υπηρεσίες ή καταλύματα, όπως το Airbnb, παρατηρείται συχνά σημαντική ανισορροπία

στον αριθμό των αξιολογήσεων που αντιστοιχούν σε διαφορετικές κατηγορίες. Σε ορισμένες περιπτώσεις κυριαρχούν έντονα οι θετικές αξιολογήσεις, ενώ σε άλλες μπορεί να συγκεντρώνεται μεγαλύτερος αριθμός αρνητικών σχολίων.

Η παρατήρηση αυτή της άνισης κατανομής δημιούργησε το ενδιαφέρον να διερευνηθεί η συμπεριφορά ενός συστήματος αυτόματης ανάλυσης συναισθήματος όταν δεν διαθέτει παρόμοιο αριθμό παραδειγμάτων από όλες τις κλάσεις. Πρόκειται για ένα πρόβλημα που δεν περιορίζεται σε τεχνητά ή εργαστηριακά datasets, αλλά εμφανίζεται και σε πραγματικά δεδομένα διαδικτυακών αξιολογήσεων.

Με βάση αυτή την αφετηρία, η εργασία επικεντρώθηκε στη μελέτη της μη ισορροπημένης κατανομής των δεδομένων και στην εφαρμογή τεχνικών υπερδειγματοληψίας. Εξετάστηκαν τόσο το Random Oversampling σε συνδυασμό με BERT όσο και το SMOTE σε διανυσματικές αναπαραστάσεις που παράγονται από RoBERTa, με στόχο να διερευνηθεί κατά πόσο η αριθμητική ενίσχυση των λιγότερο συχνών κλάσεων μπορεί να οδηγήσει σε αποτελεσματικότερη αναγνώρισή τους από τα αντίστοιχα μοντέλα [46]-[50].

1.5 Σκοπός της διπλωματικής εργασίας

Σκοπός της διπλωματικής εργασίας είναι η μελέτη της επίδρασης μεθόδων υπερδειγματοληψίας σε εφαρμογές ανάλυσης συναισθήματος όταν τα δεδομένα παρουσιάζουν έντονη ανισορροπία μεταξύ των κλάσεων. Η έρευνα δεν θεωρεί εκ των προτέρων ότι η υπερδειγματοληψία θα οδηγήσει υποχρεωτικά σε υψηλότερη συνολική ακρίβεια. Αντίθετα, επιδιώκεται να εξεταστεί πειραματικά εάν και σε ποιο βαθμό οι τεχνικές αυτές επηρεάζουν την ικανότητα των μοντέλων να αναγνωρίζουν τις λιγότερο συχνές κατηγορίες.

Για την επίτευξη του σκοπού χρησιμοποιούνται σύγχρονες Transformer-based αναπαραστάσεις μέσω BERT και RoBERTa [5], [6], σε συνδυασμό με Random Oversampling και SMOTE. Η σύγκριση επικεντρώνεται τόσο στη συνολική απόδοση όσο και σε μετρικές ανά κλάση. Ιδιαίτερη σημασία αποδίδεται στο recall και στο F1-score των μειονοτικών κλάσεων, καθώς αυτές οι μετρικές μπορούν να αποκαλύψουν βελτιώσεις ή υποβαθμίσεις που δεν είναι εμφανείς από την accuracy [2], [11], [13], [35]-[37].

1.6 Στόχοι της εργασίας

Για την υλοποίηση του παραπάνω σκοπού ορίζονται οι ακόλουθοι επιμέρους στόχοι:

1. Η μελέτη του θεωρητικού υπόβαθρου της ανάλυσης συναισθήματος, των Transformer μοντέλων και του προβλήματος της μη ισορροπημένης μάθησης.
2. Η ανάλυση της κατανομής των κλάσεων στα χρησιμοποιούμενα σύνολα δεδομένων και η τεκμηρίωση του βαθμού ανισορροπίας πριν από την εφαρμογή οποιασδήποτε τεχνικής balancing.
3. Η δημιουργία baseline πειραμάτων χωρίς υπερδειγματοληψία, ώστε να υπάρχει σημείο αναφοράς για την αξιολόγηση των παρεμβάσεων στα δεδομένα.
4. Η εφαρμογή Random Oversampling σε δύο διαφορετικά πειραματικά σενάρια με BERT: αρχικά στο πλήρες σύνολο εκπαίδευσης για τη μελέτη της επίδρασης της εξισορρόπησης στην εκπαίδευση και στην κατανομή των προβλέψεων και, στη συνέχεια, αποκλειστικά στο training split ενός διαχωρισμού training/validation, ώστε να αξιολογηθεί η γενίκευση σε ανέγγιχτο validation set.

5. Η εφαρμογή SMOTE σε embeddings που παράγονται από RoBERTa και η εκπαίδευση ταξινομητή Logistic Regression πάνω στο ισορροπημένο σύνολο αναπαραστάσεων.
6. Η αξιολόγηση των μοντέλων με accuracy, precision, recall, F1-score, macro F1 και confusion matrices, με ιδιαίτερη έμφαση στις μειονοτικές κατηγορίες.
7. Η συγκριτική συζήτηση των πλεονεκτημάτων, των περιορισμών και των συνθηκών υπό τις οποίες κάθε μέθοδος υπερδειγματοληψίας μπορεί να είναι χρήσιμη σε εφαρμογές ανάλυσης συναισθήματος.

1.7 Ερευνητικά ερωτήματα

Με βάση τον σκοπό και τους επιμέρους στόχους της εργασίας, η πειραματική διερεύνηση οργανώνεται γύρω από τέσσερα ερευνητικά ερωτήματα. Τα ερωτήματα εξετάζουν την επίδραση της ανισορροπίας των κλάσεων, τη συμπεριφορά του Random Oversampling και του SMOTE, καθώς και την καταλληλότητα των μετρικών αξιολόγησης σε μη ισορροπημένα δεδομένα.

1.7.1 Ερευνητικό ερώτημα 1: Επίδραση της ανισορροπίας των κλάσεων

RQ1: Πώς επηρεάζει η ανισορροπία των κλάσεων τη συμπεριφορά και την απόδοση μοντέλων ανάλυσης συναισθήματος; Το ερώτημα αυτό εξετάζει αν η υψηλή παρουσία της πλειονοτικής κλάσης οδηγεί σε άνιση απόδοση μεταξύ των κατηγοριών και εάν το συνολικό accuracy αποτυπώνει επαρκώς αυτή τη συμπεριφορά.

1.7.2 Ερευνητικό ερώτημα 2: Επίδραση του Random Oversampling

RQ2: Σε ποιο βαθμό η εφαρμογή Random Oversampling επηρεάζει την απόδοση ενός pipeline βασισμένου σε BERT και ειδικότερα την αναγνώριση των μειονοτικών κλάσεων; Το ερώτημα εξετάζεται σε δύο πειραματικές διατάξεις: μία εφαρμογή ROS στο πλήρες training dataset και μία αυστηρότερη εφαρμογή όπου το ROS εκτελείται μόνο στο training split, ενώ η αξιολόγηση πραγματοποιείται σε ανέγγιχτο validation set. Στο πρώτο σενάριο εξετάζεται κυρίως η μεταβολή της κατανομής των προβλέψεων στο μη επισημασμένο test set, ενώ στο δεύτερο η σύγκριση βασίζεται σε accuracy, precision, recall και F1-score ανά κλάση πάνω στο κοινό validation set.

1.7.3 Ερευνητικό ερώτημα 3: Επίδραση του SMOTE σε Transformer embeddings

RQ3: Ποια είναι η επίδραση του SMOTE όταν εφαρμόζεται σε διανυσματικές αναπαραστάσεις που έχουν παραχθεί από RoBERTa; Επειδή το SMOTE λειτουργεί σε αριθμητικό χώρο χαρακτηριστικών, το ερώτημα εστιάζει στη δυνατότητα χρήσης Transformer embeddings ως κατάλληλου χώρου για τη δημιουργία συνθετικών μειονοτικών δειγμάτων.

1.7.4 Ερευνητικό ερώτημα 4: Κατάλληλες μετρικές αξιολόγησης

RQ4: Ποιες μετρικές αξιολόγησης παρέχουν την πληρέστερη εικόνα της επίδοσης σε έντονα μη ισορροπημένα δεδομένα; Εξετάζεται η σχέση μεταξύ accuracy και μετρικών ανά κλάση, καθώς και η χρησιμότητα του macro F1 όταν απαιτείται ισότιμη αξιολόγηση όλων των κατηγοριών ανεξάρτητα από το πλήθος των δειγμάτων τους.

1.8 Μεθοδολογική προσέγγιση

Η μεθοδολογική προσέγγιση της εργασίας οργανώνεται ως συγκριτική πειραματική μελέτη. Αρχικά πραγματοποιείται φόρτωση και έλεγχος των δεδομένων, καθαρισμός όπου απαιτείται και επιλογή των

κατάλληλων κειμενικών πεδίων. Σε πειράματα όπου χρησιμοποιούνται πολυγλωσσικές ή μη φιλτραρισμένες κριτικές, εφαρμόζεται ανίχνευση γλώσσας ώστε να διατηρηθούν τα αγγλικά κείμενα. Στη συνέχεια πραγματοποιείται η κατάλληλη tokenization ή η εξαγωγή embeddings, ανάλογα με το μοντέλο και το πείραμα.

Στον πρώτο βασικό πειραματικό άξονα χρησιμοποιείται BERT για ταξινόμηση κριτικών και εξετάζονται δύο περιπτώσεις Random Oversampling. Στην πρώτη, το ROS εφαρμόζεται στο πλήρες διαθέσιμο training dataset, ώστε να μελετηθεί η μεταβολή της κατανομής των κλάσεων και η συμπεριφορά των προβλέψεων μετά την εκπαίδευση σε πλήρως εξισορροπημένο σύνολο. Στη δεύτερη, προηγείται διαχωρισμός σε training και validation set και το ROS εφαρμόζεται αποκλειστικά στο training split. Με τον τρόπο αυτό το validation set παραμένει ανέγγιχτο και χρησιμοποιείται για άμεση σύγκριση του baseline BERT με το BERT που εκπαιδεύτηκε μετά την υπερδειγματοληψία.

Στον δεύτερο βασικό άξονα χρησιμοποιείται RoBERTa ως μηχανισμός παραγωγής πυκνών διανυσματικών αναπαραστάσεων. Οι επισημασμένες κριτικές διαχωρίζονται αρχικά σε training και validation δεδομένα και στη συνέχεια παράγονται ξεχωριστά τα αντίστοιχα RoBERTa embeddings. Το SMOTE εφαρμόζεται αποκλειστικά στα training embeddings, ώστε να αποφεύγεται διαρροή συνθετικής πληροφορίας προς το σύνολο αξιολόγησης. Πάνω στα αρχικά και στα εξισορροπημένα training embeddings εκπαιδεύεται Logistic Regression και συγκρίνονται οι επιδόσεις των δύο εκδοχών στο ίδιο validation set.

Η πειραματική διαδικασία υλοποιείται σε Python και Google Colab με βιβλιοθήκες όπως PyTorch, Transformers, scikit-learn και imbalanced-learn. Η επιλογή αυτών των εργαλείων επιτρέπει την αξιοποίηση GPU για τα Transformer μοντέλα και παρέχει έτοιμες, ευρέως χρησιμοποιούμενες υλοποιήσεις για τις τεχνικές ταξινόμησης και υπερδειγματοληψίας. Η λεπτομερής περιγραφή των δεδομένων, των παραμέτρων εκπαίδευσης και της ακριβούς ακολουθίας των πειραμάτων παρουσιάζεται στο Κεφάλαιο 5.

Τα δύο πειράματα Random Oversampling δεν σχεδιάστηκαν εξ αρχής ως δύο ανεξάρτητες παραλλαγές της ίδιας διαδικασίας. Στην πρώτη εφαρμογή, ο στόχος ήταν να εκπαιδευτεί το μοντέλο μετά την εξισορρόπηση του πλήρους training set και στη συνέχεια να εξεταστεί η συμπεριφορά των προβλέψεων πάνω στο διαθέσιμο test set. Κατά την επεξεργασία των δεδομένων, όμως, το test set διαπιστώθηκε ότι δεν περιείχε πραγματικές ετικέτες βαθμολογίας. Για αυτό μπορούσαν να παραχθούν predictions, αλλά δεν ήταν δυνατό να υπολογιστούν accuracy, precision, recall ή F1-score.

Για να υπάρχει δυνατότητα υπολογισμού πραγματικών μετρικών, χρησιμοποιήθηκε στη συνέχεια το επισημασμένο training dataset ως βάση για ένα δεύτερο πείραμα. Το dataset χωρίστηκε σε training και validation τμήμα και το Random Oversampling εφαρμόστηκε μόνο στο training subset. Το validation set διατηρήθηκε χωρίς αλλαγές, ώστε τόσο το baseline όσο και το oversampled μοντέλο να αξιολογηθούν πάνω στα ίδια δεδομένα με γνωστές πραγματικές ετικέτες. Η δεύτερη αυτή διάταξη αποτέλεσε επομένως την κύρια βάση για τη σύγκριση της γενίκευσης των δύο μοντέλων.

1.9 Δομή της εργασίας

Η εργασία οργανώνεται σε επτά κύρια κεφάλαια. Το παρόν Κεφάλαιο 1 εισάγει το πρόβλημα, ορίζει τον σκοπό, τους στόχους και τα ερευνητικά ερωτήματα και παρουσιάζει συνοπτικά τη μεθοδολογική προσέγγιση.

Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό υπόβαθρο της Επεξεργασίας Φυσικής Γλώσσας και της ανάλυσης συναισθήματος. Το Κεφάλαιο 3 εστιάζει στο πρόβλημα της μη ισορροπημένης μάθησης και

στις μεθόδους υπερδειγματοληψίας, με έμφαση στο Random Oversampling και στο SMOTE. Στο Κεφάλαιο 4 αναλύεται η αρχιτεκτονική Transformer και παρουσιάζονται τα μοντέλα BERT και RoBERTa, καθώς και η χρήση των embeddings τους σε προβλήματα ταξινόμησης.

Το Κεφάλαιο 5 περιγράφει αναλυτικά τη μεθοδολογία και την υλοποίηση των πειραματικών pipelines, τα χρησιμοποιούμενα datasets, το preprocessing και τις παραμέτρους εκπαίδευσης. Στο Κεφάλαιο 6 παρουσιάζονται και συζητούνται τα πειραματικά αποτελέσματα με κατάλληλες μετρικές και confusion matrices. Τέλος, στο Κεφάλαιο 7 συνοψίζονται τα κύρια συμπεράσματα, απαντώνται τα ερευνητικά ερωτήματα και καταγράφονται οι περιορισμοί και πιθανές μελλοντικές επεκτάσεις της εργασίας.

1.10 Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκε το γενικό πλαίσιο της διπλωματικής εργασίας και αναδείχθηκε η σύνδεση μεταξύ της ανάλυσης συναισθήματος και του προβλήματος της μη ισορροπημένης κατανομής δεδομένων. Η ύπαρξη ισχυρών Transformer μοντέλων δεν αναιρεί την ανάγκη προσεκτικής διαχείρισης των κλάσεων, ιδιαίτερα όταν οι σπάνιες κατηγορίες έχουν ουσιαστική σημασία για την εφαρμογή.

Με βάση το παραπάνω πλαίσιο, η εργασία εξετάζει πειραματικά δύο κύριες τεχνικές υπερδειγματοληψίας, το Random Oversampling και το SMOTE, σε συνδυασμό με BERT και RoBERTa-based αναπαραστάσεις. Στο επόμενο κεφάλαιο παρουσιάζονται αναλυτικότερα οι βασικές έννοιες της Επεξεργασίας Φυσικής Γλώσσας και της ανάλυσης συναισθήματος, οι οποίες αποτελούν το θεωρητικό υπόβαθρο για τα πειράματα που ακολουθούν.

Κεφάλαιο 2ο: Επεξεργασία Φυσικής Γλώσσας και Ανάλυση Συναισθήματος

2.1 Εισαγωγή

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP) αποτελεί κλάδο της Τεχνητής Νοημοσύνης που επικεντρώνεται στην ανάπτυξη υπολογιστικών μεθόδων για την ανάλυση, κατανόηση και παραγωγή ανθρώπινης γλώσσας. Η σημασία της έχει ενισχυθεί σημαντικά λόγω της συνεχούς αύξησης των κειμενικών δεδομένων που παράγονται σε κοινωνικά δίκτυα, ηλεκτρονικά καταστήματα, πλατφόρμες αξιολογήσεων, φόρουμ και συστήματα υποστήριξης χρηστών. Σε αντίθεση με τα δομημένα δεδομένα, το φυσικό κείμενο χαρακτηρίζεται από μεγάλη ποικιλία στη σύνταξη, στο λεξιλόγιο και στον τρόπο έκφρασης, γεγονός που καθιστά την αυτόματη επεξεργασία του ιδιαίτερα απαιτητική.

Μία από τις πιο διαδεδομένες εφαρμογές του NLP είναι η ανάλυση συναισθήματος, δηλαδή η αυτόματη εκτίμηση της στάσης ή της συναισθηματικής πολικότητας που εκφράζεται σε ένα κείμενο. Η συγκεκριμένη εργασία εστιάζει σε εφαρμογές ανάλυσης συναισθήματος πάνω σε δεδομένα κριτικών, όπου η κατανομή των κατηγοριών μπορεί να είναι έντονα μη ισορροπημένη. Για τον λόγο αυτό, πριν εξεταστούν αναλυτικά οι τεχνικές υπερδειγματοληψίας στα επόμενα κεφάλαια, είναι απαραίτητο να παρουσιαστούν οι βασικές έννοιες της Επεξεργασίας Φυσικής Γλώσσας, οι κύριες προσεγγίσεις στην ανάλυση συναισθήματος, τα συνήθη στάδια προεπεξεργασίας και οι δυσκολίες που εμφανίζονται κατά την ταξινόμηση πραγματικών κειμενικών δεδομένων.

2.2 Επεξεργασία Φυσικής Γλώσσας

Η ενότητα αυτή παρουσιάζει συνοπτικά το πεδίο της Επεξεργασίας Φυσικής Γλώσσας και τη θέση του στο ευρύτερο πλαίσιο της Τεχνητής Νοημοσύνης. Αρχικά δίνεται ο βασικός ορισμός του NLP και στη συνέχεια παρουσιάζονται χαρακτηριστικές εφαρμογές του που συνδέονται με την επεξεργασία και την ταξινόμηση κειμένου.

2.2.1 Ορισμός της Επεξεργασίας Φυσικής Γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας μελετά τον τρόπο με τον οποίο ένα υπολογιστικό σύστημα μπορεί να αναπαραστήσει και να επεξεργαστεί γλωσσικές πληροφορίες. Το πεδίο συνδυάζει στοιχεία γλωσσολογίας, στατιστικής, μηχανικής μάθησης και βαθιάς μάθησης. Παραδοσιακά, οι εφαρμογές NLP βασίζονταν σε χειροκίνητα κατασκευασμένους κανόνες, λεξικά και γραμματικές. Με την αύξηση της διαθεσιμότητας δεδομένων, οι στατιστικές και οι μηχανικές προσεγγίσεις απέκτησαν κυρίαρχο ρόλο, ενώ τα τελευταία χρόνια τα προεκπαιδευμένα νευρωνικά γλωσσικά μοντέλα έχουν μεταβάλει σημαντικά τον τρόπο επίλυσης των περισσότερων γλωσσικών προβλημάτων [9], [46].

Βασική δυσκολία του NLP είναι ότι η φυσική γλώσσα δεν αποτελεί ένα απλό σύνολο ανεξάρτητων λέξεων. Η σημασία μιας λέξης εξαρτάται από τα συμφραζόμενα, τη θέση της στην πρόταση, τις προηγούμενες και επόμενες λέξεις, τη σύνταξη και συχνά από γνώσεις που δεν αναφέρονται ρητά στο κείμενο. Η ίδια ακολουθία λέξεων μπορεί να αποκτά διαφορετική σημασία ανάλογα με το πλαίσιο, ενώ δύο διαφορετικές διατυπώσεις μπορεί να εκφράζουν ουσιαστικά την ίδια πρόθεση. Συνεπώς, ένα σύστημα NLP πρέπει να μετατρέψει ένα σύνθετο συμβολικό αντικείμενο, όπως το κείμενο, σε αναπαράσταση που μπορεί να επεξεργαστεί αριθμητικά ένας αλγόριθμος.

2.2.2 Βασικές εφαρμογές NLP

Το NLP περιλαμβάνει ένα ευρύ φάσμα εφαρμογών. Η ταξινόμηση κειμένου αποδίδει ένα έγγραφο ή μία πρόταση σε μία ή περισσότερες προκαθορισμένες κατηγορίες και αποτελεί τη γενικότερη μορφή του προβλήματος που εξετάζεται στην παρούσα εργασία. Άλλες σημαντικές εφαρμογές είναι η αναγνώριση ονοματικών οντοτήτων, όπου εντοπίζονται στοιχεία όπως πρόσωπα, οργανισμοί και τοποθεσίες, η αυτόματη περίληψη κειμένου, η μηχανική μετάφραση, η απάντηση ερωτήσεων, η εξαγωγή πληροφοριών και η παραγωγή φυσικής γλώσσας.

Ανάλογα με το πρόβλημα, η ανάλυση μπορεί να απαιτεί διαφορετικά επίπεδα γλωσσικής πληροφορίας. Η μορφολογική επεξεργασία εξετάζει τη δομή των λέξεων, η συντακτική ανάλυση τις σχέσεις μεταξύ των όρων μιας πρότασης και η σημασιολογική επεξεργασία επιχειρεί να αποδώσει νόημα στις γλωσσικές μονάδες. Σε πρακτικές εφαρμογές τα επίπεδα αυτά δεν είναι πλήρως ανεξάρτητα. Ένα σύστημα ταξινόμησης συναισθήματος, για παράδειγμα, μπορεί να χρειάζεται ταυτόχρονα πληροφορία για τη σειρά των λέξεων, την άρνηση και τη σημασία μιας έκφρασης μέσα στα συγκεκριμένα συμφραζόμενα. Η σύγχρονη τάση είναι μεγάλο μέρος αυτής της πληροφορίας να μαθαίνεται από το ίδιο το μοντέλο αντί να κωδικοποιείται αποκλειστικά με χειροποίητους κανόνες.

Η ανάλυση συναισθήματος αποτελεί ειδική περίπτωση ταξινόμησης κειμένου, στην οποία η ζητούμενη κατηγορία σχετίζεται με τη στάση του συγγραφέα. Η εφαρμογή της είναι ιδιαίτερα σημαντική σε περιβάλλοντα όπου συγκεντρώνονται μεγάλες ποσότητες σχολίων ή αξιολογήσεων. Μέσω της αυτοματοποιημένης επεξεργασίας μπορούν να εντοπιστούν τάσεις σε χιλιάδες κριτικές χωρίς να απαιτείται χειροκίνητη ανάγνωση κάθε κειμένου. Η αξία αυτή όμως εξαρτάται άμεσα από την αξιοπιστία του μοντέλου και από την ικανότητά του να αναγνωρίζει όχι μόνο την κυρίαρχη κατηγορία, αλλά και σπανιότερες μορφές αρνητικού ή ουδέτερου συναισθήματος.

2.3 Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος αποτελεί το ειδικό πρόβλημα ταξινόμησης κειμένου στο οποίο επικεντρώνεται η παρούσα εργασία. Στις υποενότητες που ακολουθούν ορίζεται η έννοια του sentiment analysis και παρουσιάζονται οι βασικές κατηγορίες και τα επίπεδα στα οποία μπορεί να πραγματοποιηθεί η ανάλυση.

2.3.1 Ορισμός

Η ανάλυση συναισθήματος, γνωστή επίσης ως opinion mining, αφορά την υπολογιστική μελέτη απόψεων, αξιολογήσεων, στάσεων και συναισθημάτων που εκφράζονται σε κείμενο [1], [7], [20]-[24]. Σε μία τυπική εφαρμογή, το σύστημα λαμβάνει ως είσοδο ένα κείμενο και αποδίδει μία ετικέτα που περιγράφει την εκφραζόμενη πολικότητα. Η απλούστερη μορφή του προβλήματος είναι η δυαδική ταξινόμηση σε θετικό και αρνητικό συναίσθημα. Σε πιο σύνθετες εφαρμογές προστίθεται ουδέτερη κατηγορία ή χρησιμοποιούνται λεπτομερέστερες κλίμακες, όπως αξιολογήσεις από ένα έως πέντε αστέρια.

Η έννοια του συναισθήματος δεν ταυτίζεται πάντοτε με ένα ρητά εκφρασμένο συναίσθημα, όπως χαρά ή θυμός. Σε συστήματα αξιολόγησης προϊόντων ή υπηρεσιών ενδιαφέρει συχνά η συνολική στάση του χρήστη απέναντι στο αντικείμενο. Για παράδειγμα, η φράση «το δωμάτιο ήταν καθαρό αλλά πολύ μικρό» περιέχει θετική και αρνητική αξιολόγηση ταυτόχρονα. Το τελικό label εξαρτάται από το επίπεδο στο οποίο πραγματοποιείται η ανάλυση και από τον τρόπο με τον οποίο έχει οριστεί το πρόβλημα ταξινόμησης.

Η βασική ροή ενός συστήματος ανάλυσης συναισθήματος αποτυπώνεται συνοπτικά στο Σχήμα 2.1.



Σχήμα 2.1: Βασική ροή ενός συστήματος ανάλυσης συναισθήματος.

Πηγή: Ιδία επεξεργασία.

2.3.2 Κατηγορίες ανάλυσης συναισθήματος

Η συνηθέστερη κατηγοριοποίηση αφορά την πολικότητα του κειμένου. Στη δυαδική περίπτωση χρησιμοποιούνται δύο κλάσεις, *positive* και *negative*, ενώ στην τριαδική προσέγγιση προστίθεται η κατηγορία *neutral*. Η τρίτη κλάση είναι χρήσιμη όταν υπάρχουν κείμενα που δεν εκφράζουν σαφή αξιολογική στάση, αλλά αυξάνει τη δυσκολία του προβλήματος, επειδή τα όρια μεταξύ ήπια θετικών, ήπια αρνητικών και πραγματικά ουδέτερων διατυπώσεων είναι συχνά ασαφή [7], [8], [24].

Μία εναλλακτική διατύπωση είναι η ταξινόμηση σε διατεταγμένη κλίμακα, όπως οι αξιολογήσεις 1 έως 5. Σε αυτή την περίπτωση οι κατηγορίες διαθέτουν φυσική σειρά, όμως ένας τυπικός classifier μπορεί να τις χειριστεί ως ανεξάρτητες κλάσεις. Το πενταβάθμιο σχήμα έχει ιδιαίτερη σχέση με τα πειράματα BERT της παρούσας εργασίας, στα οποία οι κριτικές αντιστοιχούν σε ratings από 1 έως 5. Η κατανομή αυτών των ratings δεν είναι αναγκαστικά ομοιόμορφη, με αποτέλεσμα η ταξινόμηση συναισθήματος να συνδέεται άμεσα με το πρόβλημα της μη ισορροπημένης μάθησης.

2.3.3 Επίπεδα ανάλυσης συναισθήματος

Η ανάλυση συναισθήματος μπορεί να πραγματοποιηθεί σε διαφορετικά επίπεδα λεπτομέρειας. Στο επίπεδο εγγράφου (*document-level*) θεωρείται ότι ολόκληρο το κείμενο εκφράζει μία συνολική άποψη και αποδίδεται μία ενιαία ετικέτα. Η προσέγγιση αυτή είναι κατάλληλη για πολλές διαδικτυακές κριτικές, όπου ζητούμενο είναι η συνολική αξιολόγηση του χρήστη. Στο επίπεδο πρότασης (*sentence-level*) κάθε πρόταση εξετάζεται ανεξάρτητα, επιτρέποντας την αναγνώριση διαφορετικών στάσεων μέσα στο ίδιο έγγραφο.

Σε μεγαλύτερη λεπτομέρεια, η *aspect-based sentiment analysis* επιχειρεί να εντοπίσει το αντικείμενο ή το χαρακτηριστικό στο οποίο αναφέρεται η άποψη και να αποδώσει ξεχωριστή πολικότητα σε κάθε πτυχή [7], [21]. Για μία ξενοδοχειακή κριτική, για παράδειγμα, ο χρήστης μπορεί να αξιολογεί θετικά την τοποθεσία αλλά αρνητικά την καθαριότητα. Η παρούσα εργασία δεν υλοποιεί *aspect-based* ταξινόμηση, αλλά η διάκριση των επιπέδων είναι σημαντική επειδή δείχνει ότι η έννοια του «σωστού sentiment label» εξαρτάται από τον στόχο της εφαρμογής και από τον τρόπο δημιουργίας του dataset.

2.4 Προσεγγίσεις στην ανάλυση συναισθήματος

Η ανάλυση συναισθήματος μπορεί να υλοποιηθεί με διαφορετικές μεθοδολογικές προσεγγίσεις, οι οποίες διαφέρουν ως προς τον τρόπο αναπαράστασης του κειμένου και εκμάθησης των σχετικών προτύπων. Στη συνέχεια παρουσιάζονται οι lexicon-based μέθοδοι, οι παραδοσιακές τεχνικές Machine Learning, οι προσεγγίσεις Deep Learning και τα σύγχρονα Transformer-based μοντέλα.

2.4.1 Lexicon-based προσεγγίσεις

Οι lexicon-based μέθοδοι βασίζονται σε λεξικά όπου λέξεις ή φράσεις συνδέονται με προκαθορισμένες τιμές συναισθηματικής πολικότητας. Ένα απλό σύστημα μπορεί να υπολογίζει το συνολικό score ενός κειμένου συνδυάζοντας τις θετικές και αρνητικές τιμές των λέξεων που εμφανίζονται σε αυτό. Το βασικό πλεονέκτημα είναι η ερμηνευσιμότητα και η δυνατότητα εφαρμογής χωρίς μεγάλο επισημασμένο training dataset. Επιπλέον, τέτοιες προσεγγίσεις μπορούν να αποτελέσουν γρήγορο baseline ή να χρησιμοποιηθούν όταν οι διαθέσιμοι υπολογιστικοί πόροι είναι περιορισμένοι [8], [20], [22].

Ωστόσο, τα λεξικά δυσκολεύονται να αναπαραστήσουν σύνθετα συμφραζόμενα. Η πολικότητα μιας λέξης μπορεί να αλλάξει λόγω άρνησης, ειρωνείας ή του συγκεκριμένου πεδίου εφαρμογής. Η λέξη «απρόβλεπτο», για παράδειγμα, μπορεί να είναι θετική όταν περιγράφει την πλοκή μιας ταινίας αλλά αρνητική όταν περιγράφει τη λειτουργία μιας υπηρεσίας. Για τον λόγο αυτό, οι σύγχρονες εφαρμογές υψηλής ακρίβειας βασίζονται συχνότερα σε μοντέλα που μαθαίνουν απευθείας από δεδομένα και λαμβάνουν υπόψη το γλωσσικό περιβάλλον.

2.4.2 Παραδοσιακές τεχνικές Machine Learning

Στις κλασικές προσεγγίσεις μηχανικής μάθησης, το κείμενο μετατρέπεται πρώτα σε αριθμητικό διάνυσμα χαρακτηριστικών. Συνηθισμένες αναπαραστάσεις είναι το Bag-of-Words και το TF-IDF, στις οποίες κάθε διάσταση σχετίζεται με την παρουσία ή τη σχετική σημαντικότητα μιας λέξης ή ενός n-gram. Πάνω σε αυτά τα διανύσματα μπορούν να εκπαιδευτούν αλγόριθμοι όπως Logistic Regression, Naive Bayes και Support Vector Machines. Οι μέθοδοι αυτές αποτέλεσαν για πολλά χρόνια ισχυρά baselines για την ταξινόμηση κειμένου και παραμένουν χρήσιμες όταν το μέγεθος των δεδομένων ή οι υπολογιστικοί πόροι δεν δικαιολογούν τη χρήση μεγάλων νευρωνικών μοντέλων [1], [8], [25]-[27].

Ένα σημαντικό χαρακτηριστικό των παραδοσιακών classifiers είναι ότι το στάδιο αναπαράστασης και το στάδιο ταξινόμησης είναι διακριτά. Αυτό επιτρέπει την εφαρμογή τεχνικών balancing στον αριθμητικό χώρο των χαρακτηριστικών, υπό την προϋπόθεση ότι η αναπαράσταση είναι κατάλληλη. Η συγκεκριμένη λογική σχετίζεται άμεσα με ένα από τα πειράματα της εργασίας, όπου ένα Transformer χρησιμοποιείται για παραγωγή embeddings και στη συνέχεια εφαρμόζεται SMOTE πριν από την εκπαίδευση Logistic Regression. Επομένως, οι κλασικοί classifiers δεν αντιμετωπίζονται ως ξεπερασμένη τεχνολογία, αλλά μπορούν να λειτουργήσουν συμπληρωματικά προς τις σύγχρονες νευρωνικές αναπαραστάσεις.

Οι supervised τεχνικές μηχανικής μάθησης προϋποθέτουν επίσης ένα σύνολο παραδειγμάτων με γνωστές ετικέτες. Η ποιότητα αυτών των labels είναι κρίσιμη, επειδή ο classifier μαθαίνει το όριο απόφασης από τις διαθέσιμες περιπτώσεις. Σε κειμενικά δεδομένα, οι αναπαραστάσεις Bag-of-Words και TF-IDF είναι συνήθως πολύ υψηλής διάστασης και αραιές, αφού κάθε έγγραφο περιέχει μόνο ένα μικρό μέρος του συνολικού λεξιλογίου. Παρότι τέτοιες αναπαραστάσεις μπορούν να λειτουργήσουν αποτελεσματικά με γραμμικούς ταξινομητές, δεν περιγράφουν εύκολα τη σημασιολογική σχέση

μεταξύ διαφορετικών λέξεων ή διατυπώσεων που εκφράζουν παρόμοιο νόημα.. Οι νευρωνικές αναπαραστάσεις, όπως τα embeddings, αντιμετωπίζουν διαφορετικά το πρόβλημα, μετατρέποντας το κείμενο σε πυκνότερα διανύσματα μικρότερης διάστασης. Με αυτόν τον τρόπο μπορούν να διατηρηθούν περισσότερες πληροφορίες σχετικά με το νόημα και τα συμφραζόμενα του κειμένου. Η ιδιότητα αυτή είναι ιδιαίτερα χρήσιμη στην παρούσα εργασία, καθώς επιτρέπει τη χρήση των Transformer embeddings όχι μόνο για ταξινόμηση αλλά και ως αριθμητικό χώρο πάνω στον οποίο μπορούν να εφαρμοστούν τεχνικές όπως το SMOTE.

2.4.3 Deep Learning

Η βαθιά μάθηση επέτρεψε την εκμάθηση αναπαραστάσεων απευθείας από τα δεδομένα και μείωσε την ανάγκη χειροκίνητου σχεδιασμού χαρακτηριστικών. Νευρωνικές αρχιτεκτονικές όπως τα Convolutional Neural Networks, τα Recurrent Neural Networks και ειδικότερα τα LSTM χρησιμοποιήθηκαν ευρέως σε προβλήματα ταξινόμησης κειμένου. Σε αντίθεση με το Bag-of-Words, τα μοντέλα ακολουθίας μπορούν να αξιοποιούν τη σειρά των λέξεων και να διατηρούν πληροφορία για προηγούμενα στοιχεία της πρότασης [28]-[33].

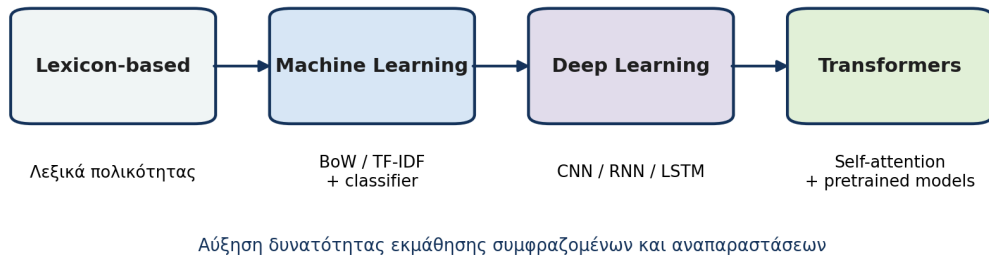
Παρότι τα RNN και LSTM αποτέλεσαν σημαντικό βήμα προς την καλύτερη μοντελοποίηση γλωσσικού περιβάλλοντος, η σειριακή φύση τους δημιουργεί περιορισμούς στην παράλληλη επεξεργασία και στην αποτύπωση πολύ μακρινών εξαρτήσεων. Οι περιορισμοί αυτοί αποτέλεσαν ένα από τα κίνητρα για την ανάπτυξη της αρχιτεκτονικής Transformer, η οποία βασίζεται στον μηχανισμό attention και μπορεί να επεξεργάζεται αποτελεσματικότερα σχέσεις μεταξύ διαφορετικών θέσεων μιας ακολουθίας [4], [32], [33].

2.4.4 Transformer-based προσεγγίσεις

Η αρχιτεκτονική Transformer εισήγαγε μία διαφορετική προσέγγιση στην επεξεργασία ακολουθιών, βασισμένη κυρίως σε μηχανισμούς self-attention αντί για αναδρομική επεξεργασία [4]. Πάνω σε αυτή την αρχιτεκτονική αναπτύχθηκαν προεκπαιδευμένα γλωσσικά μοντέλα, όπως το BERT και το RoBERTa. Το BERT εκπαιδεύεται ώστε να δημιουργεί συμφραζόμενες αναπαραστάσεις, δηλαδή η αναπαράσταση μιας λέξης εξαρτάται από το κείμενο που την περιβάλλει [5]. Το RoBERTa διατηρεί τη βασική αρχιτεκτονική του BERT, αλλά τροποποιεί τη διαδικασία προεκπαίδευσης και τις σχετικές υπερπαραμέτρους [6].

Για την ανάλυση συναισθήματος, τα μοντέλα αυτά μπορούν είτε να προσαρμοστούν μέσω fine-tuning για classification είτε να χρησιμοποιηθούν ως μηχανισμοί παραγωγής embeddings για έναν ξεχωριστό ταξινομητή. Η ικανότητα μοντελοποίησης συμφραζομένων είναι ιδιαίτερα σημαντική σε κριτικές, όπου η σημασία ενός όρου μπορεί να αλλάζει ανάλογα με τη φράση στην οποία βρίσκεται. Η παρούσα εργασία αξιοποιεί και τις δύο λογικές: BERT για απευθείας ταξινόμηση και RoBERTa για παραγωγή συναισθηματικά χρήσιμων αναπαραστάσεων. Η αρχιτεκτονική των Transformer, καθώς και οι λεπτομέρειες των BERT και RoBERTa, παρουσιάζονται αναλυτικά στο Κεφάλαιο 4.

Η εξέλιξη των κυριότερων προσεγγίσεων που παρουσιάστηκαν στην ενότητα συνοψίζεται στο Σχήμα 2.2.



Σχήμα 2.2: Βασικές προσεγγίσεις στην ανάλυση συναισθήματος.

Πηγή: Ιδία επεξεργασία.

2.5 Προεπεξεργασία κειμένου

Πριν από την εφαρμογή ενός μοντέλου ταξινόμησης, τα κειμενικά δεδομένα πρέπει να μετατραπούν σε μορφή κατάλληλη για επεξεργασία. Η ενότητα εξετάζει τα βασικά στάδια αυτής της διαδικασίας, όπως τον καθαρισμό, το tokenization, το padding και truncation, την ανίχνευση γλώσσας και τη διαχείριση κειμένων μεγάλου μήκους.

2.5.1 Καθαρισμός δεδομένων

Πριν από την εκπαίδευση ή την εφαρμογή ενός μοντέλου, τα κειμενικά δεδομένα πρέπει να ελεγχθούν ως προς την ποιότητα και τη μορφή τους. Ο καθαρισμός μπορεί να περιλαμβάνει την αντιμετώπιση κενών τιμών, τη μετατροπή μη κειμενικών αντικειμένων σε κατάλληλη μορφή, την αφαίρεση διπλότυπων όταν αυτό είναι μεθοδολογικά επιθυμητό και τον έλεγχο για προβληματικούς χαρακτήρες. Σε δεδομένα που προέρχονται από το διαδίκτυο συναντώνται συχνά URLs, HTML entities, emojis, πολλαπλά κενά ή ιδιαίτερη στίξη.

Η υπερβολική κανονικοποίηση δεν είναι πάντοτε ωφέλιμη. Σε σύγχρονα προεκπαιδευμένα μοντέλα, στοιχεία όπως η στίξη, η κεφαλαιοποίηση ή ακόμη και ορισμένα ειδικά σύμβολα μπορεί να μεταφέρουν χρήσιμη πληροφορία. Για τον λόγο αυτό, το preprocessing πρέπει να προσαρμόζεται στο μοντέλο που χρησιμοποιείται και όχι να εφαρμόζεται μηχανικά. Στην παρούσα εργασία διατηρείται όσο το δυνατόν η φυσική μορφή των reviews, ενώ γίνονται οι απαραίτητοι έλεγχοι ώστε το κείμενο να είναι έγκυρη είσοδος για τους αντίστοιχους tokenizers.

Εξίσου σημαντική είναι η συνέπεια ανάμεσα στο preprocessing της εκπαίδευσης και της αξιολόγησης. Εάν τα training δεδομένα υποστούν διαφορετικούς μετασχηματισμούς από τα validation ή test δεδομένα, μπορεί να δημιουργηθεί τεχνητή μεταβολή της κατανομής εισόδου. Για λόγους αναπαραγωγιμότητας, οι μετασχηματισμοί πρέπει να είναι σαφώς καθορισμένοι και να εφαρμόζονται με την ίδια λογική σε κάθε πειραματική εκτέλεση. Αυτό είναι ιδιαίτερα σημαντικό όταν συγκρίνονται baseline και oversampling σενάρια, επειδή η μέθοδος εξισορρόπησης πρέπει να αποτελεί την κύρια ελεγχόμενη διαφορά και όχι να συνδυάζεται με διαφορετική προεπεξεργασία.

2.5.2 Tokenization

Το tokenization είναι η διαδικασία διάσπασης του κειμένου σε μικρότερες μονάδες, τα tokens, τα οποία αποτελούν την είσοδο του γλωσσικού μοντέλου. Σε απλές προσεγγίσεις ένα token μπορεί να αντιστοιχεί σε μία ολόκληρη λέξη. Ωστόσο, τα σύγχρονα μοντέλα χρησιμοποιούν συνήθως subword tokenization, ώστε να μπορούν να χειρίζονται σπάνιες λέξεις, μορφολογικές παραλλαγές και όρους που δεν έχουν εμφανιστεί αυτούσιοι κατά την προεκπαίδευση.

Στα BERT και RoBERTa ο tokenizer είναι στενά συνδεδεμένος με το λεξιλόγιο και τον τρόπο προεκπαίδευσης του αντίστοιχου μοντέλου. Το αρχικό κείμενο μετατρέπεται σε ακολουθία token IDs, ενώ προστίθενται ειδικά tokens όπου απαιτείται. Το αποτέλεσμα δεν αποτελεί απλή αντιστοίχιση λέξεων σε ακέραιους αριθμούς, αλλά την είσοδο πάνω στην οποία θα δημιουργηθούν οι συμπραζόμενες αναπαραστάσεις. Συνεπώς, η χρήση του σωστού tokenizer για το εκάστοτε pretrained checkpoint είναι απαραίτητη προϋπόθεση για έγκυρη inference ή fine-tuning διαδικασία [5], [6].

2.5.3 Padding και truncation

Τα νευρωνικά μοντέλα εκπαιδεύονται συνήθως σε batches, δηλαδή σε ομάδες παραδειγμάτων που επεξεργάζονται ταυτόχρονα. Επειδή τα κείμενα έχουν διαφορετικό μήκος, απαιτείται μηχανισμός ώστε οι ακολουθίες ενός batch να αποκτούν συμβατές διαστάσεις. Το padding προσθέτει ειδικά tokens στις μικρότερες ακολουθίες, ενώ το attention mask ενημερώνει το μοντέλο ποια σημεία αποτελούν πραγματικό κείμενο και ποια padding.

Αντίθετα, όταν ένα κείμενο υπερβαίνει το μέγιστο επιτρεπόμενο μήκος του μοντέλου, εφαρμόζεται truncation. Η επιλογή του ορίου αποτελεί συμβιβασμό μεταξύ πληροφορίας και υπολογιστικού κόστους. Πολύ μικρό όριο μπορεί να αφαιρέσει κρίσιμο τμήμα μιας κριτικής, ενώ πολύ μεγάλο όριο αυξάνει τη χρήση μνήμης και τον χρόνο επεξεργασίας. Στα πειράματα της εργασίας οι αντίστοιχες τιμές καθορίζονται από τον tokenizer και τις ρυθμίσεις κάθε notebook και θα παρουσιαστούν αναλυτικά στο κεφάλαιο της μεθοδολογίας.

2.5.4 Ανίχνευση γλώσσας

Σε datasets που συλλέγονται από διεθνείς πλατφόρμες είναι συνηθισμένο να συνυπάρχουν κριτικές σε πολλές γλώσσες. Εάν ένα προεκπαιδευμένο μοντέλο έχει σχεδιαστεί κυρίως για αγγλικό κείμενο, η ανεξέλεγκτη εισαγωγή κειμένων διαφορετικών γλωσσών μπορεί να μειώσει την αξιοπιστία των προβλέψεων. Για αυτό μπορεί να προηγηθεί αυτόματη ανίχνευση γλώσσας και επιλογή μόνο των δειγμάτων που αντιστοιχούν στην επιθυμητή γλώσσα.

Η γλωσσική ανίχνευση δεν είναι απόλυτα ακριβής, ιδιαίτερα σε πολύ σύντομα κείμενα, ονόματα ή reviews με μεικτή χρήση γλωσσών. Επομένως, αποτελεί ένα preprocessing φίλτρο και όχι ανεξάρτητη εγγύηση ποιότητας. Στο πειραματικό pipeline της εργασίας, η επιλογή αγγλικών κριτικών είναι σημαντικό βήμα πριν από την εφαρμογή του RoBERTa sentiment model, επειδή εξασφαλίζει μεγαλύτερη συμβατότητα ανάμεσα στο περιεχόμενο του dataset και στο μοντέλο που χρησιμοποιείται.

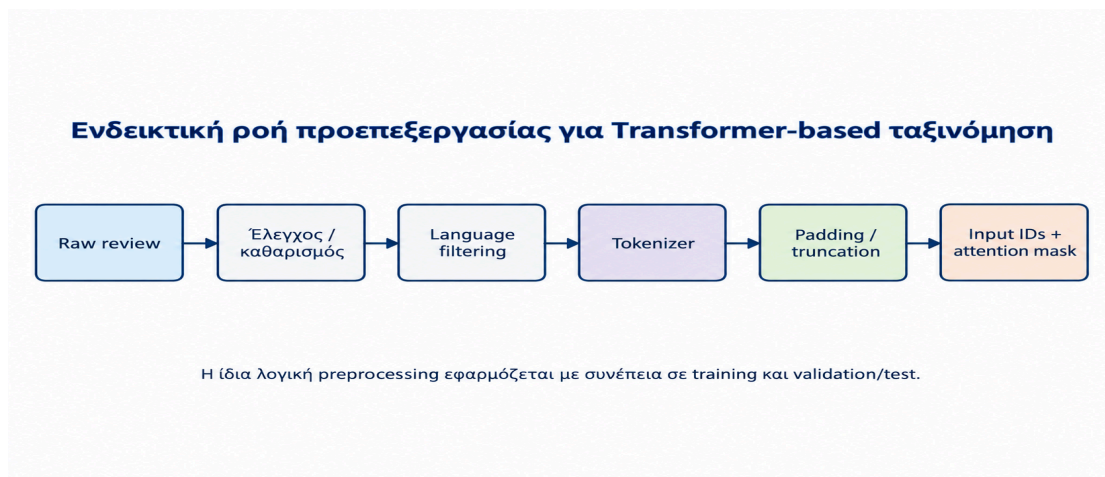
2.5.5 Διαχείριση μεγάλου μήκους κειμένων

Οι μεγάλες κριτικές δημιουργούν πρόσθετο πρόβλημα όταν ξεπερνούν το μέγιστο μήκος εισόδου ενός Transformer. Η απλούστερη λύση είναι το truncation, όμως η απόρριψη του τελευταίου μέρους του κειμένου μπορεί να αφαιρέσει την τελική αξιολόγηση του χρήστη ή μία κρίσιμη αντίθεση. Εναλλακτικά, ένα μεγάλο κείμενο μπορεί να χωριστεί σε τμήματα και οι επιμέρους προβλέψεις να

συνδυαστούν, αλλά αυτό αυξάνει την πολυπλοκότητα του pipeline και δημιουργεί το ερώτημα του τρόπου aggregation.

Για εφαρμογές document-level sentiment analysis, η επιλογή στρατηγικής εξαρτάται από την κατανομή μήκους του dataset και από τον διαθέσιμο υπολογιστικό προϋπολογισμό. Η παρούσα εργασία ακολουθεί τις ρυθμίσεις μήκους των χρησιμοποιούμενων Transformer pipelines και δεν επιχειρεί ειδικό hierarchical μοντέλο για υπερμεγέθη κείμενα. Ο περιορισμός αυτός πρέπει να λαμβάνεται υπόψη κατά την ερμηνεία των αποτελεσμάτων, καθώς ένα μέρος της πληροφορίας ενδέχεται να μην εισέρχεται στο μοντέλο όταν απαιτείται truncation.

Η συνολική ακολουθία των βασικών σταδίων προεπεξεργασίας που προηγούνται της εισόδου στο μοντέλο παρουσιάζεται στο Σχήμα 2.3.



Σχήμα 2.3: Ενδεικτική ροή προεπεξεργασίας κειμένου για Transformer-based ταξινόμηση.

Πηγή: *Ιδία επεξεργασία.*

2.6 Σύνολα δεδομένων κριτικών

Οι διαδικτυακές κριτικές αποτελούν ιδιαίτερα κατάλληλη πηγή για μελέτες ανάλυσης συναισθήματος επειδή περιέχουν φυσικά παραγόμενες αξιολογικές εκφράσεις και συχνά συνοδεύονται από αριθμητικές βαθμολογίες. Οι βαθμολογίες μπορούν να λειτουργήσουν ως labels ή ως έμμεση ένδειξη πολικότητας, ενώ το ελεύθερο κείμενο προσφέρει πλούσιο γλωσσικό περιεχόμενο. Παράλληλα, όμως, οι κριτικές περιέχουν θόρυβο, ανορθογραφίες, συντομογραφίες, διαφορετικό μήκος και διαφοροποίηση ως προς το ύφος, γεγονός που τις καθιστά ρεαλιστικό αλλά απαιτητικό πεδίο αξιολόγησης.

Ένα ακόμη χαρακτηριστικό των review datasets είναι ότι οι χρήστες δεν επιλέγουν απαραίτητα όλες τις βαθμολογίες με την ίδια συχνότητα. Σε αρκετές πλατφόρμες παρατηρείται συγκέντρωση στις υψηλές βαθμολογίες, με πολύ μικρότερο αριθμό παραδειγμάτων στις ενδιάμεσες ή αρνητικές κατηγορίες. Επομένως, η μη ισορροπημένη κατανομή δεν είναι απλώς τεχνητό πρόβλημα ενός εργαστηριακού dataset, αλλά μπορεί να αποτελεί χαρακτηριστικό του πραγματικού μηχανισμού παραγωγής δεδομένων. Αυτό καθιστά τις κριτικές κατάλληλο πεδίο για την αξιολόγηση τεχνικών όπως το Random Oversampling και το SMOTE.

Η αξιοποίηση των κριτικών απαιτεί επίσης προσεκτικό διαχωρισμό των δεδομένων. Όταν υπάρχουν πολλαπλές εκδόσεις του ίδιου δείγματος ή όταν η υπερδειγματοληψία πραγματοποιείται πριν από τον διαχωρισμό σε training και validation/test σύνολα, υπάρχει κίνδυνος παρόμοια ή ταυτόσημα

παραδείγματα να εμφανιστούν και στις δύο πλευρές της αξιολόγησης. Μία τέτοια διαρροή δεδομένων μπορεί να δημιουργήσει υπερβολικά αισιόδοξες μετρικές. Για αυτό, σε πειράματα που έχουν ως στόχο την εκτίμηση της γενίκευσης, η εξισορρόπηση πρέπει να περιορίζεται στο training set. Η αρχή αυτή θα εφαρμοστεί ρητά στη δεύτερη περίπτωση Random Oversampling της παρούσας εργασίας.

2.7 Προκλήσεις στην ανάλυση συναισθήματος

Η απόδοση ενός συστήματος sentiment analysis επηρεάζεται από γλωσσικά και δεδομενοκεντρικά φαινόμενα που δυσκολεύουν την ορθή ταξινόμηση. Οι υποενότητες που ακολουθούν εστιάζουν στον σαρκασμό, στην άρνηση, στην πολυσημία, στην ουδέτερη κατηγορία και στην ανισορροπία των κλάσεων.

2.7.1 Σαρκασμός

Ο σαρκασμός είναι μία από τις δυσκολότερες μορφές γλωσσικής έκφρασης για ένα σύστημα ανάλυσης συναισθήματος. Μία σαρκαστική πρόταση μπορεί να χρησιμοποιεί εμφανώς θετικές λέξεις για να εκφράσει αρνητική αξιολόγηση, όπως στη διατύπωση «εξαιρετική εξυπηρέτηση, περίμενα μόνο δύο ώρες». Ένα σύστημα που βασίζεται αποκλειστικά σε μεμονωμένες λέξεις μπορεί να αποδώσει λανθασμένα θετικό sentiment, ενώ η σωστή ερμηνεία απαιτεί κατανόηση της αντίθεσης μεταξύ του κυριολεκτικού και του πραγματικού νοήματος.

Τα συμφραζόμενα μοντέλα μπορούν να μειώσουν μέρος αυτού του προβλήματος, επειδή η αναπαράσταση δεν βασίζεται μόνο σε ένα μεμονωμένο token. Παρ' όλα αυτά, ο σαρκασμός μπορεί να απαιτεί πραγματολογική γνώση ή πληροφορία που βρίσκεται εκτός του συγκεκριμένου κειμένου. Συνεπώς, ακόμη και ένα ισχυρό Transformer δεν είναι εγγυημένο ότι θα αναγνωρίσει σωστά κάθε τέτοια περίπτωση.

2.7.2 Άρνηση

Η άρνηση μπορεί να αντιστρέψει ή να τροποποιήσει σημαντικά την πολικότητα μιας φράσης. Οι εκφράσεις «καλό» και «δεν ήταν καλό» περιέχουν τον ίδιο βασικό θετικό όρο, αλλά μεταφέρουν διαφορετική αξιολόγηση. Σε απλές bag-of-words ή lexicon-based προσεγγίσεις, η σχέση ανάμεσα στην άρνηση και στη λέξη που επηρεάζει μπορεί να χαθεί, ιδίως όταν μεσολαβούν αρκετές λέξεις.

Η σωστή αντιμετώπιση της άρνησης αποτελεί χαρακτηριστικό παράδειγμα του γιατί η σειρά και τα συμφραζόμενα των tokens έχουν σημασία. Τα Transformer μοντέλα έχουν τη δυνατότητα να συσχετίζουν διαφορετικές θέσεις της ακολουθίας μέσω self-attention, όμως η τελική επιτυχία εξαρτάται και από τα δεδομένα προεκπαίδευσης, το fine-tuning και το είδος των κειμένων που αναλύονται. Σε reviews με πολύπλοκες προτάσεις μπορεί επίσης να συνυπάρχουν πολλαπλές αρνήσεις ή αντιθέσεις.

2.7.3 Πολυσημία

Πολυσημία ονομάζεται η δυνατότητα μιας λέξης να έχει περισσότερες από μία σημασίες. Το φαινόμενο αυτό δυσκολεύει τις στατικές αναπαραστάσεις, επειδή η ίδια λέξη μπορεί να λαμβάνει μία ενιαία τιμή ή διάνυμα ανεξάρτητα από το πλαίσιο. Στην ανάλυση συναισθήματος το πρόβλημα είναι ακόμη εντονότερο όταν διαφορετικές σημασίες της ίδιας λέξης έχουν διαφορετική αξιολογική φόρτιση.

Οι contextual embeddings των Transformer αντιμετωπίζουν καλύτερα την πολυσημία, επειδή η αναπαράσταση ενός token προκύπτει σε συνάρτηση με τα υπόλοιπα tokens της ακολουθίας [5]. Παρόλα αυτά, σε πολύ εξειδικευμένα domains μπορεί να υπάρχουν όροι των οποίων η χρήση διαφέρει σημαντικά από το γενικό κείμενο πάνω στο οποίο έγινε η προεκπαίδευση. Για αυτό η επιλογή domain-appropriate μοντέλου ή η περαιτέρω προσαρμογή του μπορεί να επηρεάσει ουσιαστικά την απόδοση.

2.7.4 Ουδέτερο συναίσθημα

Η ουδέτερη κατηγορία είναι συχνά δυσκολότερη από όσο υποδηλώνει ο ορισμός της. Ένα κείμενο μπορεί να χαρακτηριστεί ουδέτερο επειδή δεν εκφράζει γνώμη, επειδή περιλαμβάνει ισορροπημένες θετικές και αρνητικές παρατηρήσεις ή επειδή η συνολική στάση δεν είναι αρκετά έντονη ώστε να ενταχθεί στις ακραίες κατηγορίες. Αυτές οι διαφορετικές περιπτώσεις δεν είναι απαραίτητα γλωσσικά ομοιογενείς.

Η δυσκολία ορισμού της neutral class επηρεάζει τόσο την επισήμανση των δεδομένων όσο και την αξιολόγηση του μοντέλου. Δύο annotators μπορεί να διαφωνήσουν για μία ήπια διατύπωση, ενώ ένα pretrained sentiment model μπορεί να έχει διαφορετικά decision boundaries από το dataset μιας νέας εφαρμογής. Για αυτό τα αποτελέσματα μιας τριαδικής ταξινόμησης πρέπει να ερμηνεύονται σε συνάρτηση με τον τρόπο δημιουργίας των labels και όχι μόνο ως αριθμητικές μετρικές.

2.7.5 Ανισορροπία κλάσεων

Η ανισορροπία των κλάσεων αποτελεί κεντρικό πρόβλημα της παρούσας εργασίας. Όταν μία ή περισσότερες κατηγορίες εμφανίζονται πολύ λιγότερο συχνά στο training set, το μοντέλο λαμβάνει λιγότερα παραδείγματα από τα οποία μπορεί να μάθει τα χαρακτηριστικά τους. Σε ένα review dataset όπου η πλειονότητα των χρηστών δίνει υψηλές βαθμολογίες, ένας classifier μπορεί να επιτυγχάνει υψηλή συνολική accuracy προβλέποντας σωστά κυρίως τις δημοφιλείς κατηγορίες, αλλά να αποτυγχάνει σε σπάνιες αρνητικές ή ενδιάμεσες αξιολογήσεις.

Για τον λόγο αυτό, η αξιολόγηση πρέπει να εξετάζει precision, recall και F1-score ανά κλάση, καθώς και macro averages που δίνουν ίσο βάρος στις κατηγορίες. Η ύπαρξη imbalance δεν σημαίνει αυτόματα ότι πρέπει να εφαρμοστεί συγκεκριμένη τεχνική εξισορρόπησης. Η αποτελεσματικότητα μιας μεθόδου εξαρτάται από το dataset, το μοντέλο και το πειραματικό πρωτόκολλο. Το επόμενο κεφάλαιο παρουσιάζει αναλυτικά το πρόβλημα της μη ισορροπημένης μάθησης και τις μεθόδους υπερδειγματοληψίας που εξετάζονται πειραματικά στην εργασία.

Η διάκριση μεταξύ συνολικής και ανά κλάση απόδοσης είναι ουσιώδης και για την ερμηνεία των αποτελεσμάτων. Ένα μοντέλο μπορεί να εμφανίζει μικρή μεταβολή στην accuracy, αλλά ταυτόχρονα σημαντική αλλαγή στο recall μιας μειονοτικής κατηγορίας. Αντίστροφα, μία τεχνική oversampling μπορεί να αυξήσει την προσαρμογή στα επαναλαμβανόμενα training samples χωρίς να προσφέρει αντίστοιχη βελτίωση σε ανέγγιχτα δεδομένα. Επομένως, το αν μία μέθοδος θεωρείται επιτυχημένη πρέπει να κρίνεται με βάση το ερευνητικό ερώτημα και όχι από μία μοναδική μετρική. Η προσέγγιση αυτή αποτελεί βασική αρχή της πειραματικής αξιολόγησης που ακολουθεί στα επόμενα κεφάλαια.

2.8 Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκε το βασικό θεωρητικό πλαίσιο της Επεξεργασίας Φυσικής Γλώσσας και της ανάλυσης συναισθήματος. Αναλύθηκαν οι κύριες κατηγορίες και τα επίπεδα sentiment analysis, καθώς και η εξέλιξη των προσεγγίσεων από λεξικά και παραδοσιακούς classifiers

προς τη βαθιά μάθηση και τα Transformer-based μοντέλα. Επιπλέον, παρουσιάστηκαν βασικά στάδια προεπεξεργασίας, όπως ο καθαρισμός, το tokenization, το padding, το truncation και η ανίχνευση γλώσσας.

Ιδιαίτερη έμφαση δόθηκε στις δυσκολίες που χαρακτηρίζουν πραγματικά δεδομένα κριτικών, όπως ο σαρκασμός, η άρνηση, η πολυσημία, η ουδέτερη κατηγορία και η ανισορροπία των κλάσεων. Η τελευταία αποτελεί τον συνδετικό κρίκο με το επόμενο κεφάλαιο, όπου εξετάζονται αναλυτικά οι συνέπειες της μη ισορροπημένης κατανομής και οι τεχνικές Random Oversampling και SMOTE που χρησιμοποιούνται στα πειράματα της διπλωματικής.

Κεφάλαιο 3ο: Μη Ισορροπημένα Δεδομένα και Μέθοδοι Υπερδειγματοληψίας

3.1 Εισαγωγή

Η μη ισορροπημένη κατανομή των κλάσεων αποτελεί ένα από τα σημαντικότερα πρακτικά προβλήματα στη μηχανική μάθηση και αποκτά ιδιαίτερη σημασία σε εφαρμογές ανάλυσης συναισθήματος. Σε πραγματικά σύνολα δεδομένων η συχνότητα εμφάνισης των κατηγοριών συνήθως δεν είναι ομοιόμορφη. Οι χρήστες, για παράδειγμα, μπορεί να δημοσιεύουν πολύ περισσότερες θετικές από αρνητικές κριτικές ή να επιλέγουν συχνότερα τις ακραίες βαθμολογίες από τις ενδιάμεσες. Ως αποτέλεσμα, ο αλγόριθμος εκπαίδευσης εκτίθεται σε διαφορετικό αριθμό παραδειγμάτων για κάθε κλάση και μπορεί να μάθει με μεγαλύτερη ευκολία τα χαρακτηριστικά των συχνότερων κατηγοριών. Η ύπαρξη imbalance δεν αποτελεί απλώς στατιστική ιδιότητα του dataset, αλλά παράγοντα που επηρεάζει τον τρόπο με τον οποίο πρέπει να σχεδιαστούν η εκπαίδευση, η αξιολόγηση και η ερμηνεία ενός μοντέλου [2], [10], [35]-[39], [50].

Στη βιβλιογραφία έχουν προταθεί προσεγγίσεις τόσο σε επίπεδο δεδομένων όσο και σε επίπεδο αλγόριθμου. Οι πρώτες μεταβάλλουν τη σύνθεση του training set μέσω undersampling ή oversampling, ενώ οι δεύτερες τροποποιούν τον μηχανισμό μάθησης, για παράδειγμα μέσω διαφορετικών βαρών στις κλάσεις ή διαφορετικού κόστους για τα σφάλματα. Η παρούσα διπλωματική εστιάζει κυρίως στις μεθόδους υπερδειγματοληψίας, και ειδικότερα στο Random Oversampling (ROS) και στο Synthetic Minority Over-sampling Technique (SMOTE). Το ROS εφαρμόζεται στα πειράματα BERT πάνω στα αρχικά δείγματα κειμένου, ενώ το SMOTE εφαρμόζεται σε αριθμητικές αναπαραστάσεις που έχουν εξαχθεί από RoBERTa. Η διάκριση αυτή είναι ουσιώδης, διότι οι δύο τεχνικές παρεμβαίνουν στο dataset με διαφορετικό τρόπο και δημιουργούν διαφορετικούς κινδύνους και πλεονεκτήματα [35]-[37], [40]-[44], [45], [49], [50].

3.2 Μη ισορροπημένα σύνολα δεδομένων

Η κατανόηση του class imbalance προϋποθέτει σαφή περιγραφή της κατανομής των παραδειγμάτων μεταξύ των διαθέσιμων κλάσεων. Για τον λόγο αυτό παρουσιάζονται αρχικά ο ορισμός του προβλήματος, οι έννοιες majority και minority class και ένας τρόπος αποτύπωσης του βαθμού ανισορροπίας.

3.2.1 Ορισμός του class imbalance

Ένα πρόβλημα ταξινόμησης χαρακτηρίζεται ως μη ισορροπημένο όταν οι διαθέσιμες παρατηρήσεις δεν κατανέμονται με παρόμοια συχνότητα στις κλάσεις. Στη δυαδική ταξινόμηση η περιγραφή είναι σχετικά απλή: μία κατηγορία περιλαμβάνει σημαντικά περισσότερα παραδείγματα από την άλλη. Στη multiclass ταξινόμηση, όπως στην πρόβλεψη αξιολογήσεων 1 έως 5, η εικόνα μπορεί να είναι πιο σύνθετη, καθώς μπορεί να συνυπάρχουν μία ιδιαίτερα συχνή κλάση, μία δεύτερη σχετικά συχνή και αρκετές μικρές κατηγορίες. Η έρευνα γύρω από το class imbalance έχει δείξει ότι δεν αρκεί να εξετάζεται μόνο η συνολική αναλογία των δειγμάτων. Σημασία έχουν επίσης το μέγεθος του training set, η επικάλυψη μεταξύ των κλάσεων, η παρουσία θορύβου και η δυσκολία του ίδιου του προβλήματος [10], [11], [35]-[37].

Στην ανάλυση συναισθήματος, το imbalance μπορεί να οφείλεται στην πραγματική συμπεριφορά των χρηστών και επομένως να αποτελεί χαρακτηριστικό του πληθυσμού που παρατηρείται. Ένα κατάλυμα με κυρίως ικανοποιημένους επισκέπτες είναι λογικό να συγκεντρώνει περισσότερες θετικές κριτικές. Αντίστοιχα, ένα σύνολο αξιολογήσεων προϊόντων μπορεί να περιλαμβάνει πολύ περισσότερες βαθμολογίες πέντε αστέρων από βαθμολογίες δύο ή τριών αστέρων. Η τεχνητή εξισορρόπηση του training set δεν αλλάζει αυτή την πραγματική κατανομή στον κόσμο. Αλλάζει μόνο τις πληροφορίες που παρουσιάζονται στον αλγόριθμο κατά την εκπαίδευση, με στόχο να του επιτρέψει να μάθει καλύτερα και τις λιγότερο συχνές κατηγορίες.

3.2.2 Majority class και minority class

Η κλάση με τον μεγαλύτερο αριθμό παραδειγμάτων ονομάζεται majority class, ενώ μία κατηγορία με σημαντικά μικρότερη παρουσία ονομάζεται minority class. Στα δυαδικά προβλήματα οι όροι αυτοί αντιστοιχούν συνήθως σε μία πλειονοτική και μία μειονοτική κατηγορία. Στα multiclass δεδομένα μπορούν να υπάρχουν περισσότερες από μία minority classes και οι μεταξύ τους διαφορές να είναι σημαντικές. Για παράδειγμα, σε ένα dataset πέντε ratings είναι δυνατό η κλάση 5 να αποτελεί τη majority class, η κλάση 1 να διαθέτει αρκετά παραδείγματα αλλά πολύ λιγότερα από την κλάση 5, και οι κλάσεις 2 και 3 να είναι εξαιρετικά σπάνιες.

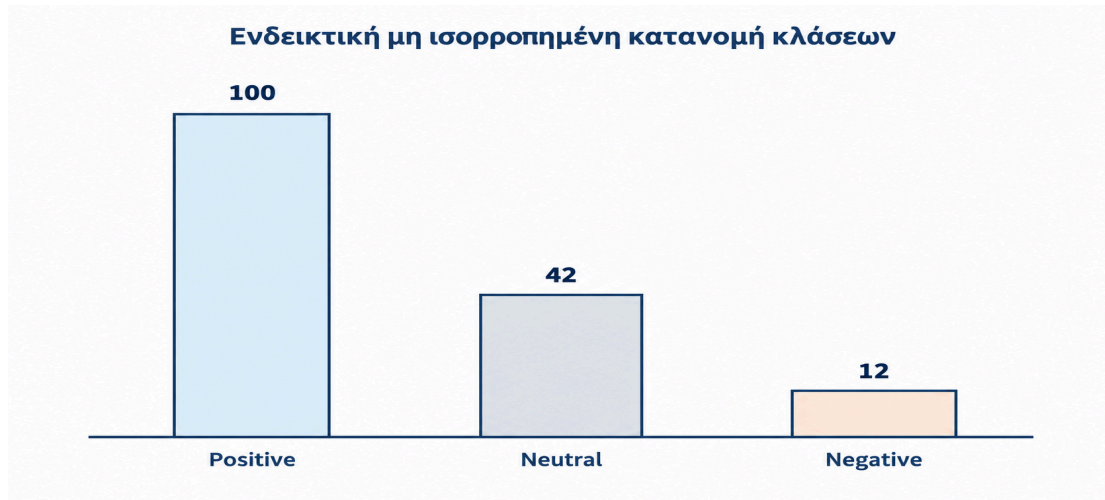
Η ορολογία majority/minority είναι χρήσιμη επειδή περιγράφει την ποσότητα των διαθέσιμων δεδομένων και όχι την επιχειρησιακή σημασία της κλάσης. Μία minority class μπορεί να είναι η σημαντικότερη κατηγορία μιας εφαρμογής, όπως ένα σπάνιο αλλά κρίσιμο αρνητικό συμβάν. Στην παρούσα εργασία, όμως, το ενδιαφέρον είναι κυρίως μεθοδολογικό: εξετάζεται αν η αύξηση της παρουσίας των σπανιότερων κατηγοριών στο training set μπορεί να μεταβάλει την ικανότητα των μοντέλων να τις αναγνωρίζουν. Η επιτυχία δεν πρέπει επομένως να κρίνεται αποκλειστικά από τη συνολική accuracy, αλλά και από την ανά κλάση συμπεριφορά.

3.2.3 Βαθμός ανισορροπίας

Ο βαθμός ανισορροπίας μπορεί να αποτυπωθεί με τον λόγο ανισορροπίας (Imbalance Ratio - IR). Σε μία απλή δυαδική περίπτωση ορίζεται ως ο λόγος του αριθμού των δειγμάτων της πλειονοτικής κλάσης προς τον αριθμό των δειγμάτων της μειονοτικής κλάσης: $IR = N_{\text{majority}} / N_{\text{minority}}$. Τιμή κοντά στο 1 υποδηλώνει περίπου ισορροπημένες κατηγορίες, ενώ μεγαλύτερες τιμές δηλώνουν αυξανόμενη απόκλιση. Στη multiclass περίπτωση μπορεί να υπολογιστεί αντίστοιχος λόγος ανά κλάση, χρησιμοποιώντας ως σημείο αναφοράς τη μεγαλύτερη κατηγορία.

Ο IR είναι χρήσιμος ως περιγραφικό μέτρο, αλλά δεν προβλέπει μόνος του τη δυσκολία ταξινόμησης. Δύο datasets με τον ίδιο λόγο ανισορροπίας μπορεί να παρουσιάζουν πολύ διαφορετική συμπεριφορά εάν στο ένα οι κλάσεις διαχωρίζονται καθαρά και στο άλλο επικαλύπτονται έντονα. Η βιβλιογραφία επισημαίνει ότι η πολυπλοκότητα του ορίου απόφασης, το μικρό απόλυτο μέγεθος της minority class και ο θόρυβος μπορούν να επιβαρύνουν περισσότερο την απόδοση από όσο δείχνει μία απλή αναλογία [10], [11], [35]-[39]. Για τον λόγο αυτό, στην πειραματική ανάλυση της εργασίας η κατανομή των κλάσεων παρουσιάζεται μαζί με precision, recall, F1-score και confusion matrices.

Ένα απλοποιημένο παράδειγμα της διαφοράς μεταξύ πλειονοτικής και μειονοτικής κλάσης παρουσιάζεται στο Σχήμα 3.1.



Σχήμα 3.1: Ενδεικτική απεικόνιση μη ισορροπημένης κατανομής κλάσεων.

Πηγή: Ιδία επεξεργασία.

3.3 Επιπτώσεις της ανισορροπίας στην εκπαίδευση

Η ανισορροπία των κλάσεων δεν επηρεάζει μόνο την αριθμητική κατανομή του dataset, αλλά μπορεί να μεταβάλει ουσιαστικά τη συμπεριφορά του ταξινομητή κατά την εκπαίδευση και την αξιολόγηση. Στη συνέχεια εξετάζονται το bias προς τη majority class, η μειωμένη αναγνώριση των μειονοτικών κατηγοριών και ο κίνδυνος μιας υψηλής αλλά παραπλανητικής accuracy.

3.3.1 Bias προς τη majority class

Κατά την εκπαίδευση ενός classifier, κάθε παράδειγμα συμβάλλει στην προσαρμογή των παραμέτρων του μοντέλου. Όταν μία κατηγορία εμφανίζεται πολύ συχνότερα, τα σφάλματα που σχετίζονται με αυτή παράγονται και επηρεάζουν την εκπαίδευση πολύ περισσότερες φορές. Έτσι, η ελαχιστοποίηση μιας συνολικής συνάρτησης απώλειας μπορεί να οδηγήσει σε λύση που ευνοεί τις συχνές κατηγορίες, ακόμη και αν η απόδοση στις σπανιότερες παραμένει χαμηλή. Το φαινόμενο έχει περιγραφεί τόσο σε παραδοσιακούς classifiers όσο και σε συστήματα βαθιάς μάθησης [2], [13], [37], [38].

Στα μοντέλα Transformer το γεγονός ότι χρησιμοποιείται ισχυρή προεκπαίδευση δεν καταργεί το πρόβλημα. Η προεκπαίδευση παρέχει πλούσιες γλωσσικές αναπαραστάσεις, αλλά το τελικό classification head και οι παράμετροι που προσαρμόζονται στο downstream task επηρεάζονται από την κατανομή των labels του συγκεκριμένου dataset. Αν ο στόχος είναι πενταταξική ταξινόμηση και οι ενδιάμεσες βαθμολογίες εμφανίζονται σπάνια, το μοντέλο μπορεί να μάθει πολύ σταθερά τις συχνές κλάσεις και να εμφανίζει αβεβαιότητα ή συστηματικές συγχύσεις στις υπόλοιπες. Αυτός είναι ένας από τους λόγους για τους οποίους στην παρούσα εργασία εξετάζεται το ROS ως παρέμβαση πριν από την εκπαίδευση BERT.

3.3.2 Μειωμένη αναγνώριση μειονοτικών κλάσεων

Η πιο άμεση συνέπεια του imbalance είναι συχνά η χαμηλή ευαισθησία στις μειονοτικές κατηγορίες. Ένας classifier μπορεί να χαρακτηρίζει μεγάλο ποσοστό των minority samples ως μέλη μιας γειτονικής ή πλειονοτικής κλάσης. Αυτό αποτυπώνεται σε χαμηλό recall, επειδή μόνο μικρό ποσοστό των πραγματικών παραδειγμάτων της συγκεκριμένης κατηγορίας αναγνωρίζεται σωστά. Σε

προβλήματα όπου η minority class είναι επιχειρησιακά κρίσιμη, η συμπεριφορά αυτή μπορεί να είναι πολύ σημαντικότερη από μία μικρή διαφορά στη συνολική accuracy.

Στα sentiment datasets το πρόβλημα μπορεί να εμφανιστεί με ιδιαίτερα ενδιαφέρον τρόπο. Οι σπάνιες κλάσεις δεν είναι απαραίτητα γλωσσικά απομονωμένες. Μία κριτική δύο αστέρων μπορεί να μοιράζεται αρνητικό λεξιλόγιο με την κλάση ενός αστεριού αλλά ταυτόχρονα να περιλαμβάνει θετικές παρατηρήσεις. Αν η κλάση 2 έχει πολύ λίγα δείγματα, το μοντέλο μπορεί να μην μάθει αρκετά παραδείγματα αυτής της ενδιάμεσης γλωσσικής συμπεριφοράς. Επομένως, η αύξηση της ποσότητας των training examples μπορεί θεωρητικά να βοηθήσει, αλλά το αποτέλεσμα εξαρτάται από το αν τα πρόσθετα δείγματα προσθέτουν πραγματική πληροφορία ή απλώς επαναλαμβάνουν τις ίδιες παρατηρήσεις.

3.3.3 Το πρόβλημα της υψηλής αλλά παραπλανητικής accuracy

Η accuracy υπολογίζει το ποσοστό των συνολικά σωστών προβλέψεων και είναι εύκολη στην ερμηνεία. Σε έντονα imbalanced δεδομένα, όμως, μπορεί να αποκρύψει σοβαρές αδυναμίες. Αν το 95% των παραδειγμάτων ανήκει σε μία κλάση, ένας υποθετικός classifier που προβλέπει πάντοτε αυτή την κατηγορία επιτυγχάνει accuracy 95%, παρότι δεν αναγνωρίζει ούτε ένα δείγμα από τις υπόλοιπες κλάσεις. Το παράδειγμα είναι ακραίο, αλλά δείχνει γιατί η υψηλή συνολική επίδοση δεν αποτελεί επαρκή ένδειξη ισόρροπης συμπεριφοράς [2], [11], [35], [41].

Για τη μελέτη της υπερδειγματοληψίας έχουν μεγαλύτερη αξία οι μετρικές ανά κλάση και οι macro averages. Το macro F1, για παράδειγμα, υπολογίζει το F1-score κάθε κλάσης και στη συνέχεια παίρνει τον απλό μέσο όρο, δίνοντας ίδιο βάρος σε κάθε κατηγορία ανεξάρτητα από τη συχνότητά της. Αντίθετα, το weighted F1 επηρεάζεται περισσότερο από τις συχνές κλάσεις. Η παράλληλη παρουσίαση accuracy, macro F1 και αναλυτικών precision/recall/F1 είναι επομένως καταλληλότερη για την αξιολόγηση των ROS και SMOTE πειραμάτων της εργασίας [34], [36].

3.4 Μέθοδοι αντιμετώπισης μη ισορροπημένων δεδομένων

Για την αντιμετώπιση της μη ισορροπημένης κατανομής έχουν αναπτυχθεί τεχνικές τόσο στο επίπεδο των δεδομένων όσο και στο επίπεδο του αλγορίθμου. Η ενότητα παρουσιάζει συνοπτικά το undersampling, το oversampling, το class weighting και το data augmentation, πριν δοθεί μεγαλύτερη έμφαση στις μεθόδους υπερδειγματοληψίας που χρησιμοποιούνται πειραματικά στην εργασία.

3.4.1 Undersampling

Το undersampling αντιμετωπίζει την ανισορροπία μειώνοντας τον αριθμό των παραδειγμάτων της majority class. Η απλούστερη μορφή είναι το Random Undersampling, όπου επιλέγεται τυχαία ένα υποσύνολο από τη συχνότερη κατηγορία. Η μέθοδος μειώνει το μέγεθος του training set και μπορεί να καταστήσει την εκπαίδευση ταχύτερη, ιδιαίτερα όταν η πλειονοτική κατηγορία είναι εξαιρετικά μεγάλη. Παράλληλα, αυξάνει τη σχετική επιρροή της minority class χωρίς να δημιουργεί τεχνητά νέα δεδομένα.

Το βασικό μειονέκτημα είναι η απώλεια πληροφορίας. Δείγματα της majority class που αφαιρούνται μπορεί να περιέχουν χρήσιμες γλωσσικές παραλλαγές, σπάνιες περιπτώσεις ή παρατηρήσεις κοντά στο decision boundary. Επομένως, παρότι το undersampling αποτελεί σημαντική κατηγορία τεχνικών, δεν επιλέχθηκε ως κεντρική παρέμβαση στην παρούσα εργασία. Ο στόχος είναι να διατηρηθούν τα διαθέσιμα παραδείγματα και να εξεταστούν μέθοδοι που αυξάνουν την παρουσία των μικρότερων κατηγοριών.

3.4.2 Oversampling

Το oversampling ακολουθεί την αντίθετη κατεύθυνση: αυξάνει τον αριθμό των training samples των μειονοτικών κλάσεων. Η αύξηση μπορεί να πραγματοποιηθεί είτε με επανάληψη ήδη υπαρχόντων παραδειγμάτων είτε με δημιουργία συνθετικών παρατηρήσεων στο feature space. Στην πρώτη κατηγορία ανήκει το Random Oversampling, ενώ στη δεύτερη ανήκει το SMOTE. Και οι δύο μέθοδοι έχουν κοινό στόχο να μειώσουν την αριθμητική κυριαρχία της majority class, αλλά δεν είναι ισοδύναμες ως προς το είδος της πληροφορίας που εισάγουν [3], [12].

Οι data-level μέθοδοι έχουν το πλεονέκτημα ότι μπορούν να χρησιμοποιηθούν με διαφορετικούς classifiers χωρίς να απαιτείται απαραίτητα αλλαγή της εσωτερικής αρχιτεκτονικής τους. Αυτό είναι ιδιαίτερα πρακτικό στα πειράματα της διπλωματικής, επειδή επιτρέπει την εφαρμογή ROS πριν από το fine-tuning του BERT και SMOTE πάνω σε RoBERTa embeddings πριν από την εκπαίδευση Logistic Regression. Ωστόσο, η υπερδειγματοληψία πρέπει να πραγματοποιείται μόνο στο training set όταν υπάρχει ανεξάρτητο validation/test set. Διαφορετικά, υπάρχει κίνδυνος leakage και υπεραισιόδοξης εκτίμησης της γενίκευσης.

3.4.3 Class weighting

Το class weighting αποτελεί algorithm-level προσέγγιση. Αντί να αλλάζει τον αριθμό των παραδειγμάτων, αλλάζει τη σχετική σημασία των σφαλμάτων που προέρχονται από διαφορετικές κλάσεις. Συνήθως οι σπανιότερες κατηγορίες λαμβάνουν μεγαλύτερο βάρος στη συνάρτηση απώλειας, ώστε μία λανθασμένη πρόβλεψη σε minority sample να επηρεάζει περισσότερο την ενημέρωση των παραμέτρων. Με αυτό τον τρόπο το dataset παραμένει αμετάβλητο, αλλά ο optimizer ενθαρρύνεται να μην αγνοεί τις μικρές κλάσεις [2], [13].

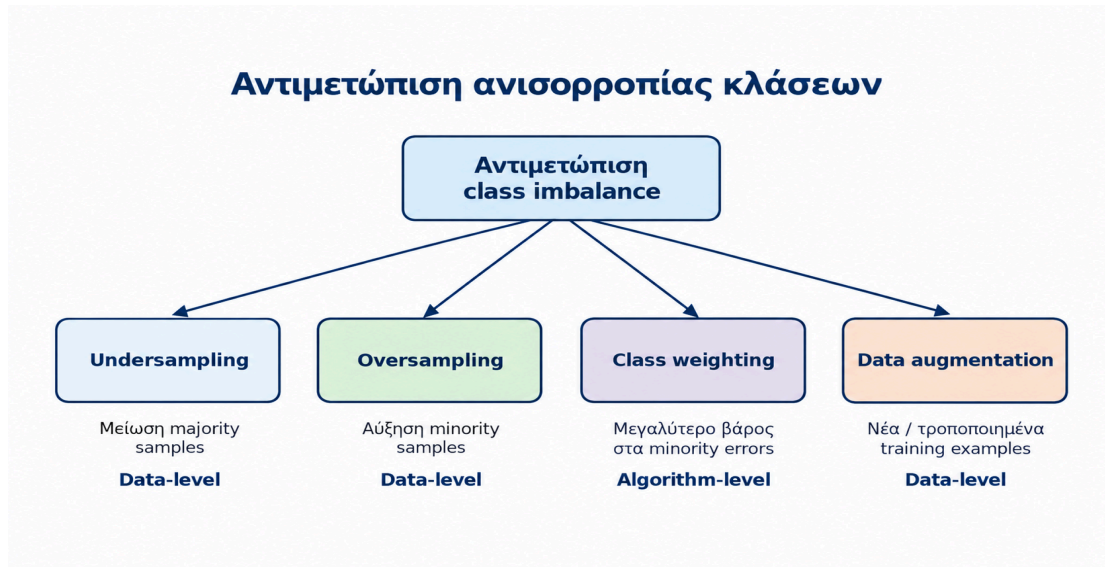
Η προσέγγιση είναι σημαντική ως εναλλακτική, αλλά δεν αποτελεί τον κύριο πειραματικό άξονα της παρούσας εργασίας, επειδή ο τίτλος και το ερευνητικό ερώτημα εστιάζουν στις μεθόδους υπερδειγματοληψίας. Παρόλα αυτά, το class weighting αποτελεί χρήσιμο σημείο αναφοράς για μελλοντική επέκταση. Θα μπορούσε να συγκριθεί με ROS και SMOTE χρησιμοποιώντας ακριβώς το ίδιο training/validation split, ώστε να εξεταστεί αν η αλλαγή της loss function προσφέρει καλύτερη ισορροπία μεταξύ majority και minority classes χωρίς επαναλήψεις ή συνθετικά features.

3.4.4 Data augmentation

Ο όρος data augmentation περιγράφει τη δημιουργία πρόσθετων training examples μέσω μετασχηματισμών που επιδιώκουν να διατηρήσουν την αρχική ετικέτα. Στην επεξεργασία εικόνας οι μετασχηματισμοί μπορούν να είναι γεωμετρικοί, ενώ στο κείμενο απαιτούν μεγαλύτερη προσοχή επειδή ακόμη και μία μικρή αλλαγή μπορεί να τροποποιήσει τη σημασία ή την πολικότητα. Έχουν προταθεί τεχνικές όπως synonym replacement, random insertion, random swap και random deletion, οι οποίες μπορούν να αυξήσουν το μέγεθος ενός text classification dataset [14].

Η data augmentation δεν ταυτίζεται με το SMOTE. Οι τεχνικές EDA δημιουργούν νέο κείμενο ή τροποποιημένες ακολουθίες tokens, ενώ το SMOTE δημιουργεί νέες αριθμητικές παρατηρήσεις μεταξύ γειτονικών minority vectors. Στην παρούσα εργασία δεν εφαρμόζεται γλωσσική augmentation στα αρχικά reviews. Η διάκριση αυτή είναι σημαντική, ώστε τα πειράματα να περιγράφονται με ακρίβεια: το ROS επαναλαμβάνει πραγματικές κριτικές, ενώ το SMOTE λειτουργεί σε embedding space και δεν παράγει νέο φυσικό κείμενο.

Οι βασικές κατηγορίες τεχνικών που εξετάστηκαν για την αντιμετώπιση του class imbalance οργανώνονται σχηματικά στο Σχήμα 3.2.



Σχήμα 3.2: Βασικές κατηγορίες μεθόδων αντιμετώπισης της ανισορροπίας κλάσεων.

Πηγή: Ιδία επεξεργασία.

3.5 Random Oversampling

Το Random Oversampling αποτελεί την απλούστερη από τις δύο κύριες τεχνικές υπερδειγματοληψίας που εξετάζονται στην παρούσα εργασία. Στις επόμενες υποενότητες παρουσιάζονται η βασική λειτουργία και η αλγοριθμική λογική της μεθόδου, καθώς και τα σημαντικότερα πλεονεκτήματα και μειονεκτήματά της.

3.5.1 Βασική λειτουργία

Το Random Oversampling είναι μία από τις απλούστερες και περισσότερο κατανοητές τεχνικές εξισορρόπησης. Για κάθε μειονοτική κλάση επιλέγονται τυχαία παραδείγματα με επανατοποθέτηση και προστίθενται στο training set μέχρι να επιτευχθεί ο επιθυμητός αριθμός δειγμάτων. Στην πλήρη εξισορρόπηση, όλες οι κλάσεις μπορούν να αυξηθούν μέχρι το μέγεθος της αρχικής majority class. Η διαδικασία δεν μεταβάλλει το περιεχόμενο των παραδειγμάτων και δεν δημιουργεί νέες γλωσσικές διατυπώσεις. Δημιουργεί επιπλέον εμφανίσεις των ίδιων training records [12].

Στην παρούσα εργασία το ROS εξετάζεται με BERT σε δύο διαφορετικά σενάρια. Στο πρώτο εφαρμόζεται στο πλήρες διαθέσιμο training dataset και χρησιμοποιείται για να μελετηθεί η εκπαίδευση σε μία πλήρως εξισορροπημένη κατανομή καθώς και η μεταβολή στην κατανομή των προβλέψεων. Στο δεύτερο και αυστηρότερο πειραματικό σενάριο προηγείται train/validation split και το ROS εφαρμόζεται αποκλειστικά στο training split. Το validation set διατηρεί την αρχική του κατανομή και χρησιμοποιείται για συγκρίσιμη αξιολόγηση πριν και μετά την υπερδειγματοληψία.

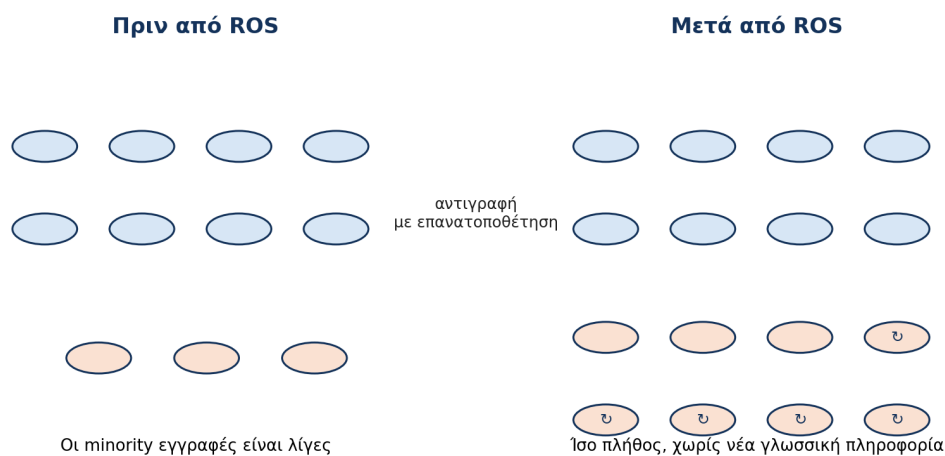
3.5.2 Αλγοριθμική περιγραφή

Έστω ότι υπάρχουν C κλάσεις και η κλάση c περιλαμβάνει N_c παραδείγματα. Αν ως στόχος εξισορρόπησης επιλεγεί N_{target} , για κάθε κλάση με $N_c < N_{\text{target}}$ απαιτούνται $N_{\text{target}} - N_c$ επιπλέον δείγματα. Το ROS επιλέγει τυχαία, με επανάθεση, παρατηρήσεις από το ήδη υπάρχον σύνολο της κλάσης c έως ότου συμπληρωθεί ο απαιτούμενος αριθμός. Η πιθανότητα επιλογής μπορεί

στην απλή μορφή να είναι ομοιόμορφη, επομένως κάθε αρχικό minority sample έχει πιθανότητα να αντιγραφεί μία ή περισσότερες φορές.

Για text classification η τεχνική είναι ιδιαίτερα απλή επειδή δεν χρειάζεται να οριστεί απόσταση μεταξύ προτάσεων ή να παραχθεί νέο κείμενο. Το oversampler χειρίζεται τις εγγραφές ως training samples και στη συνέχεια το συνηθισμένο tokenizer του BERT επεξεργάζεται τις επαναλαμβανόμενες κριτικές. Η εξισορρόπηση όμως αυξάνει το συνολικό πλήθος των training rows και επομένως αυξάνει τον χρόνο εκπαίδευσης. Στο πρώτο ROS πείραμα της εργασίας, η εξίσωση όλων των κλάσεων προς τη μεγαλύτερη κατηγορία οδηγεί σε σημαντικά μεγαλύτερο training set, γεγονός που πρέπει να λαμβάνεται υπόψη κατά την ερμηνεία του computational cost.

Η λειτουργία του Random Oversampling πριν και μετά την εξισορρόπηση μιας μειονοτικής κλάσης απεικονίζεται στο Σχήμα 3.3.



Σχήμα 3.3: Αρχή λειτουργίας του Random Oversampling.

Πηγή: Ιδία επεξεργασία.

3.5.3 Πλεονεκτήματα

Βασικά πλεονεκτήματα του Random Oversampling είναι η απλότητα εφαρμογής του και το γεγονός ότι διατηρεί όλα τα αρχικά δείγματα του συνόλου εκπαίδευσης. Η μέθοδος δεν απαιτεί δημιουργία συνθετικού κειμένου ούτε ορισμό κάποιας μετρικής απόστασης μεταξύ των κριτικών, καθώς αυξάνει την παρουσία των μειονοτικών κλάσεων μέσω επανάληψης υπαρχόντων παραδειγμάτων. Επιπλέον, μπορεί να εφαρμοστεί πριν από το tokenization και χωρίς αλλαγή της εσωτερικής αρχιτεκτονικής του ταξινομητή, γεγονός που την καθιστά εύκολη στην ενσωμάτωση σε ένα υπάρχον text-classification pipeline [12].

3.5.4 Μειονεκτήματα

Το σημαντικότερο μειονέκτημα είναι ότι το ROS δεν προσθέτει νέα πληροφορία. Αν μία minority class περιλαμβάνει λίγα ή μη αντιπροσωπευτικά δείγματα, η επανάληψή τους δεν αυξάνει τη γλωσσική ποικιλία της κατηγορίας. Αντίθετα, το μοντέλο βλέπει πολλές φορές τις ίδιες παρατηρήσεις και μπορεί να προσαρμοστεί υπερβολικά σε αυτές. Για αυτό η υπερδειγματοληψία συσχετίζεται συχνά με κίνδυνο overfitting, ιδιαίτερα όταν η αρχική minority class είναι πολύ μικρή [12].

Επιπλέον, η πλήρης εξίσωση προς την πλειονοτική κλάση μπορεί να δημιουργήσει πολύ μεγάλο training set. Η αύξηση αυτή έχει άμεσο κόστος σε χρόνο και υπολογιστικούς πόρους όταν το downstream model είναι Transformer. Τέλος, το ROS δεν αντιμετωπίζει από μόνο του προβλήματα label noise. Αν ένα λανθασμένα επισημασμένο minority sample επιλεγεί επανειλημμένα, το σφάλμα του αποκτά μεγαλύτερη παρουσία. Για τον λόγο αυτό, η τελική αξία του ROS πρέπει να κρίνεται στο ανέγγιχτο validation/test set και όχι από την απόδοση πάνω στο oversampled training set [40]-[42].

3.6 Synthetic Minority Over-sampling Technique (SMOTE)

Το SMOTE ακολουθεί διαφορετική λογική από την απλή επανάληψη μειονοτικών δειγμάτων, καθώς δημιουργεί νέες συνθετικές παρατηρήσεις στον αριθμητικό χώρο χαρακτηριστικών. Η ενότητα αναλύει τη βασική ιδέα της μεθόδου, την επιλογή κοντινών γειτόνων, τη διαδικασία δημιουργίας συνθετικών δειγμάτων, τη μαθηματική της διατύπωση και τους βασικούς περιορισμούς της.

3.6.1 Βασική ιδέα

Το SMOTE προτάθηκε από τους Chawla et al. ως μέθοδος δημιουργίας συνθετικών παραδειγμάτων της minority class αντί της απλής αντιγραφής υπαρχόντων παρατηρήσεων [3], [43]. Η βασική ιδέα είναι ότι νέα samples μπορούν να δημιουργηθούν κατά μήκος των γραμμικών τμημάτων που συνδέουν μία minority παρατήρηση με γειτονικές παρατηρήσεις της ίδιας κλάσης. Με αυτό τον τρόπο η περιοχή του feature space που καταλαμβάνει η minority class γίνεται πυκνότερη χωρίς να αποτελούν όλα τα νέα δείγματα ακριβή αντίγραφα των αρχικών.

Η μέθοδος σχεδιάστηκε για αριθμητικά χαρακτηριστικά, επειδή η έννοια της παρεμβολής προϋποθέτει ότι μπορούν να υπολογιστούν αποστάσεις και ενδιάμεσες τιμές μεταξύ των παρατηρήσεων. Αυτό δημιουργεί άμεση μεθοδολογική διαφορά από το ROS στα κείμενα. Δεν είναι λογικό να πάρουμε δύο προτάσεις και να υπολογίσουμε αριθμητικά το «ενδιάμεσο κείμενό» τους. Για αυτό, στην παρούσα εργασία το SMOTE δεν εφαρμόζεται πάνω στις λέξεις ή στα token ids αλλά πάνω σε RoBERTa embeddings, τα οποία είναι συνεχή αριθμητικά διανύσματα.

3.6.2 Επιλογή κοντινών γειτόνων

Για κάθε minority sample x_i , το SMOTE εντοπίζει k κοντινούς γείτονες που ανήκουν στην ίδια κλάση. Η επιλογή βασίζεται σε μία μετρική απόστασης στο feature space, συχνά στην Ευκλείδεια απόσταση. Στη συνέχεια επιλέγεται ένας ή περισσότεροι από αυτούς τους γείτονες ως σημεία αναφοράς για τη δημιουργία νέων παρατηρήσεων. Ο αριθμός k καθορίζει πόσο τοπική είναι η διαδικασία. Μικρό k περιορίζει τη σύνθεση σε πολύ κοντινές παρατηρήσεις, ενώ μεγαλύτερο k μπορεί να συνδέσει πιο διαφορετικά σημεία της ίδιας κλάσης [43]-[44].

Η ποιότητα της γειτονιάς έχει μεγάλη σημασία. Αν η minority class περιλαμβάνει απομονωμένα outliers ή περιοχές που βρίσκονται πολύ κοντά σε άλλη κλάση, η γραμμική παρεμβολή μπορεί να δημιουργήσει συνθετικά σημεία σε αμφίβολες περιοχές. Σε υψηλές διαστάσεις, όπως τα Transformer embeddings, οι αποστάσεις παρουσιάζουν επίσης ιδιαίτερες ιδιότητες και η γεωμετρική εγγύτητα δεν εγγυάται ότι δύο κείμενα είναι σημασιολογικά όμοια σε κάθε διάσταση. Επομένως, η εφαρμογή SMOTE σε embeddings είναι πρακτικά εφικτή αλλά χρειάζεται πειραματική αξιολόγηση αντί να θεωρείται εξ ορισμού βελτιωτική.

3.6.3 Δημιουργία συνθετικών δειγμάτων

Αφού επιλεγεί ένα αρχικό minority vector και ένας γειτονικός vector, το SMOTE δημιουργεί ένα νέο σημείο ανάμεσά τους. Το synthetic sample δεν είναι αντίγραφο κανενός από τα δύο. Οι τιμές των χαρακτηριστικών του αποτελούν παρεμβολή μεταξύ των αντίστοιχων τιμών των αρχικών διανυσμάτων. Επαναλαμβάνοντας τη διαδικασία, μπορεί να αυξηθεί η minority class μέχρι έναν προκαθορισμένο στόχο εξισορρόπησης.

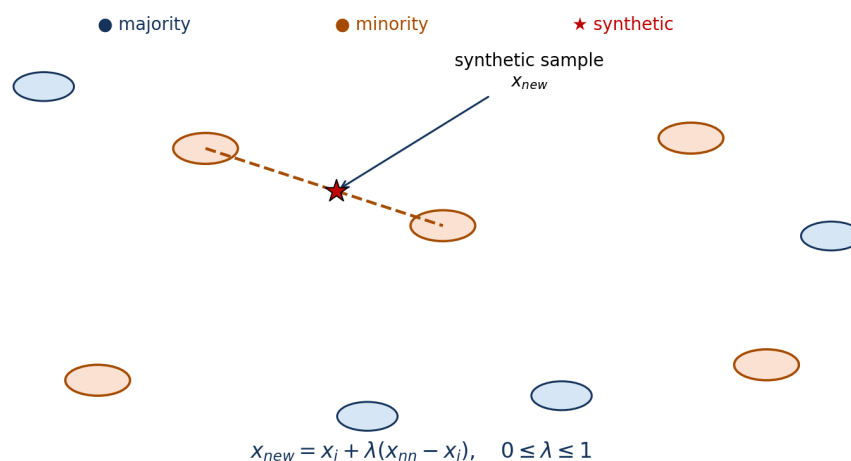
Η λογική αυτή επιδιώκει να μειώσει ένα μειονέκτημα του Random Oversampling: αντί να ενισχύει αποκλειστικά υπάρχοντα σημεία, δημιουργεί νέα σημεία στην περιοχή που ορίζεται από τα minority samples. Ωστόσο, η σύνθεση δεν αποτελεί εγγύηση ότι το νέο vector αντιστοιχεί σε πραγματικό κείμενο που θα μπορούσε να γράψει ένας χρήστης. Στην παρούσα εφαρμογή αυτό δεν αποτελεί απαίτηση, επειδή τα synthetic embeddings χρησιμοποιούνται μόνο ως training features για τη Logistic Regression. Δεν επιχειρείται αποκωδικοποίησή τους σε φυσική γλώσσα.

3.6.4 Μαθηματική διατύπωση του SMOTE

Έστω x_i ένα διάνυσμα της minority class και x_{nn} ένας από τους k κοντινότερους γείτονές του. Ένα νέο συνθετικό δείγμα μπορεί να εκφραστεί ως: $x_{new} = x_i + \lambda \cdot (x_{nn} - x_i)$, όπου λ είναι τυχαίος αριθμός στο διάστημα $[0, 1]$. Για $\lambda = 0$ το σημείο συμπίπτει με το x_i , για $\lambda = 1$ με το x_{nn} , ενώ για ενδιάμεσες τιμές το x_{new} βρίσκεται πάνω στο ευθύγραμμο τμήμα που συνδέει τα δύο vectors [3].

Η εξίσωση δείχνει και τη βασική υπόθεση της μεθόδου: οι περιοχές μεταξύ γειτονικών minority samples θεωρούνται κατάλληλες για παραγωγή πρόσθετων παρατηρήσεων της ίδιας κλάσης. Σε ένα embedding space η υπόθεση αυτή μεταφράζεται στο ότι η τοπική γεωμετρία των διανυσμάτων περιέχει χρήσιμη πληροφορία για το sentiment label. Η αποτελεσματικότητα αυτής της υπόθεσης δεν πρέπει να θεωρείται δεδομένη. Για αυτό το πείραμα RoBERTa συγκρίνει Logistic Regression πάνω στα αρχικά embeddings με Logistic Regression πάνω στο training set μετά την εφαρμογή SMOTE.

Η γεωμετρική λογική της παρεμβολής μεταξύ δύο γειτονικών minority samples παρουσιάζεται στο Σχήμα 3.4.



Σχήμα 3.4: Δημιουργία συνθετικού δείγματος με SMOTE μεταξύ γειτονικών minority vectors.

Πηγή: Ιδία επεξεργασία.

3.6.5 Πλεονεκτήματα

Το σημαντικότερο πλεονέκτημα του SMOTE έναντι της τυχαίας αντιγραφής είναι ότι δημιουργεί νέα αριθμητικά σημεία. Θεωρητικά, αυτό μπορεί να μειώσει την υπερβολική εξάρτηση από συγκεκριμένα minority samples και να βοηθήσει τον classifier να σχηματίσει ευρύτερη περιοχή απόφασης για τη μειονοτική κατηγορία. Η μέθοδος είναι επίσης γενική και μπορεί να εφαρμοστεί σε πολλά αριθμητικά feature spaces, γεγονός που την έχει καταστήσει μία από τις πλέον γνωστές τεχνικές imbalanced learning [3], [12], [43]-[44].

Στο πλαίσιο της παρούσας εργασίας, ένα ιδιαίτερο πλεονέκτημα είναι ότι το SMOTE μπορεί να συνδυαστεί με contextual representations από Transformer. Το RoBERTa χρησιμοποιείται ως feature extractor και παράγει embeddings στα οποία έχει ήδη κωδικοποιηθεί γλωσσική πληροφορία. Η υπερδειγματοληψία πραγματοποιείται μετά από αυτή την αναπαράσταση και πριν από τον τελικό classifier, επιτρέποντας να εξεταστεί μια διαφορετική στρατηγική από το ROS πάνω στο raw-text training set.

3.6.6 Περιορισμοί

Το SMOTE μπορεί να δημιουργήσει προβληματικά synthetic samples όταν οι κλάσεις επικαλύπτονται ή όταν υπάρχουν outliers. Εφόσον η μέθοδος χρησιμοποιεί μόνο την τοπική γεωμετρία της minority class, ένα νέο σημείο μπορεί να δημιουργηθεί κοντά ή μέσα σε περιοχή που στην πραγματικότητα αντιστοιχεί σε άλλη κατηγορία. Η συμπεριφορά αυτή μπορεί να αυξήσει το class overlap και να δυσκολέψει τον classifier. Επιπλέον, η επιλογή των γειτόνων και της μετρικής απόστασης επηρεάζει άμεσα το αποτέλεσμα [3], [12], [43]-[44].

Στην εφαρμογή πάνω σε Transformer embeddings υπάρχει επιπλέον ο περιορισμός της ερμηνευσιμότητας. Ένα συνθετικό διάνυσμα δεν αντιστοιχεί υποχρεωτικά σε συγκεκριμένο φυσικό κείμενο και δεν μπορεί να ελεγχθεί εύκολα από άνθρωπο ως προς το αν εκφράζει πραγματικά το label του. Επίσης, επειδή τα embeddings έχουν υψηλή διάσταση, η έννοια του κοντινού γείτονα μπορεί να μην είναι τόσο απλή όσο σε χαμηλοδιάστατα features. Επομένως, το SMOTE αντιμετωπίζεται στην εργασία ως πειραματική τεχνική της οποίας η αξία κρίνεται από τις μετρικές στο validation set και όχι ως μέθοδος που αναμένεται εκ των προτέρων να βελτιώσει την απόδοση.

3.7 SMOTE και δεδομένα κειμένου

Η άμεση εφαρμογή του SMOTE σε ακατέργαστο κείμενο δεν είναι κατάλληλη, επειδή οι λέξεις και οι προτάσεις δεν είναι συνεχή αριθμητικά χαρακτηριστικά πάνω στα οποία έχει σαφή έννοια η γραμμική παρεμβολή. Ακόμη και τα token ids ενός tokenizer είναι απλώς ακέραιοι δείκτες λεξιλογίου. Η αριθμητική απόσταση μεταξύ δύο ids δεν εκφράζει γλωσσική ή σημασιολογική απόσταση και επομένως η δημιουργία ενδιάμεσης αριθμητικής τιμής δεν θα παρήγαγε έγκυρο token.

Για να χρησιμοποιηθεί το SMOTE σε text classification απαιτείται πρώτα μετατροπή του κειμένου σε αριθμητική αναπαράσταση. Παραδοσιακά αυτή θα μπορούσε να είναι bag-of-words ή TF-IDF. Στη σύγχρονη NLP μπορούν να χρησιμοποιηθούν dense embeddings που παράγονται από νευρωνικά μοντέλα. Η επιλογή αναπαράστασης είναι κρίσιμη επειδή το SMOTE δημιουργεί synthetic samples ακριβώς σε αυτό το feature space. Η παρούσα εργασία ακολουθεί την τελευταία προσέγγιση και εφαρμόζει το SMOTE σε embeddings που εξάγονται από προεκπαιδευμένο RoBERTa [42], [47], [48].

3.8 SMOTE σε διανυσματικές αναπαραστάσεις

Η εφαρμογή του SMOTE σε κειμενικά προβλήματα απαιτεί πρώτα τη μετατροπή του κειμένου σε συνεχή αριθμητική αναπαράσταση. Για τον λόγο αυτό η ενότητα διακρίνει την έννοια των feature vectors από τα embeddings και στη συνέχεια εξετάζει ειδικά την εφαρμογή του SMOTE σε Transformer embeddings.

3.8.1 Feature vectors

Ένα feature vector είναι μία αριθμητική αναπαράσταση μιας παρατήρησης. Σε παραδοσιακή επεξεργασία κειμένου μπορεί να περιλαμβάνει συχνότητες όρων ή TF-IDF βάρη. Το βασικό πλεονέκτημα είναι ότι μετατρέπει το κείμενο σε μορφή που μπορεί να χρησιμοποιηθεί από classifiers όπως Logistic Regression, Support Vector Machines και άλλα μοντέλα. Ωστόσο, οι sparse representations συνήθως δεν αποτυπώνουν άμεσα τη σειρά των λέξεων και τη σημασία που προκύπτει από τα συμφραζόμενα.

Για oversampling, η μορφή του feature vector καθορίζει και τη γεωμετρία πάνω στην οποία θα εργαστεί το SMOTE. Αν δύο vectors θεωρούνται γείτονες, η μέθοδος υποθέτει ότι η μεταξύ τους περιοχή είναι κατάλληλη για synthetic minority samples. Επομένως, ένα feature space που οργανώνει καλύτερα τις σημασιολογικά σχετικές κριτικές μπορεί να αποτελεί καταλληλότερη βάση από μία απλή αναπαράσταση συχνοτήτων.

3.8.2 Embeddings

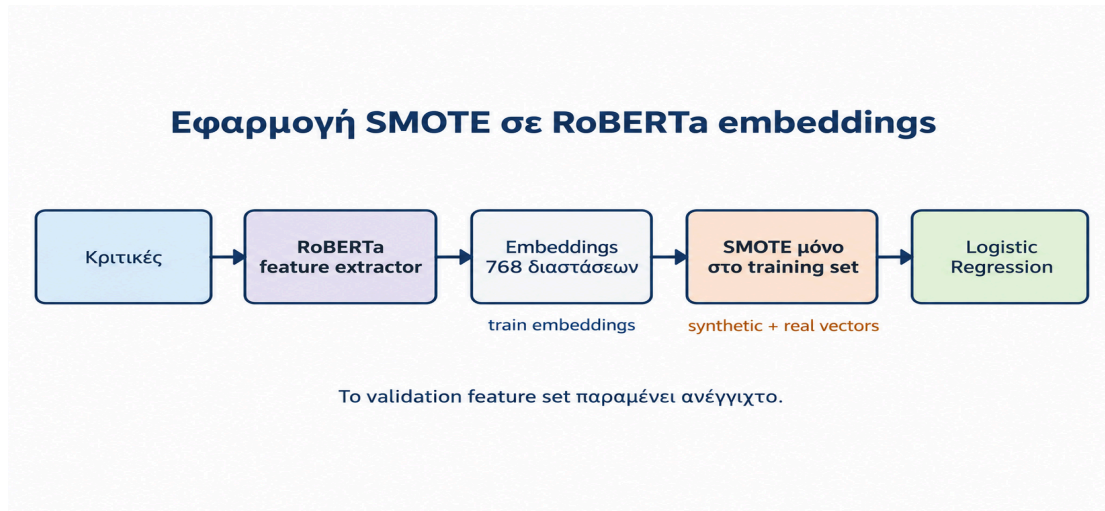
Τα embeddings είναι πυκνά αριθμητικά διανύσματα στα οποία λέξεις, προτάσεις ή έγγραφα αναπαρίστανται σε συνεχή χώρο. Στα contextual Transformer models η αναπαράσταση ενός token εξαρτάται από ολόκληρο το γλωσσικό του περιβάλλον. Αυτό επιτρέπει στα embeddings να περιέχουν πληροφορίες για τη σημασία και τη χρήση των λέξεων σε συγκεκριμένο context. Τα BERT και RoBERTa αποτελούν χαρακτηριστικά παραδείγματα μοντέλων που παράγουν τέτοιες αναπαραστάσεις [5], [6], [28]-[30].

Η χρήση embeddings ως features δημιουργεί μία υβριδική αρχιτεκτονική: ο Transformer λειτουργεί ως feature extractor και ένας κλασικός classifier πραγματοποιεί την τελική πρόβλεψη. Η προσέγγιση αυτή είναι ιδιαίτερα κατάλληλη για τη μελέτη του SMOTE, επειδή η υπερδειγματοληψία μπορεί να εφαρμοστεί στα embeddings του training set και στη συνέχεια να εκπαιδευτεί ο ίδιος classifier τόσο στα αρχικά όσο και στα balanced features. Έτσι απομονώνεται σε σημαντικό βαθμό η επίδραση της παρέμβασης στο training data.

3.8.3 Εφαρμογή SMOTE σε Transformer embeddings

Στο σχετικό πείραμα της διπλωματικής, κάθε αγγλική κριτική μετατρέπεται μέσω του RoBERTa σε διάνυσμα 768 διαστάσεων. Τα embeddings χωρίζονται σε training και validation δεδομένα και το SMOTE εφαρμόζεται αποκλειστικά στο training μέρος. Μετά την εξισορρόπηση εκπαιδεύεται Logistic Regression πάνω στα synthetic και πραγματικά training embeddings. Για τη σύγκριση εκπαιδεύεται επίσης Logistic Regression στα αρχικά, μη εξισορροπημένα embeddings.

Η εφαρμογή του SMOTE σε Transformer embeddings, αντί στο αρχικό κείμενο, συνοψίζεται στο Σχήμα 3.5.



Σχήμα 3.5: Εφαρμογή SMOTE στον χώρο Transformer embeddings.

Πηγή: *Ιδία επεξεργασία.*

Η διαδικασία μπορεί να συνοψιστεί ως Review → RoBERTa → Embedding → SMOTE → Logistic Regression. Η σειρά των βημάτων είναι ουσιώδης. Το RoBERTa δεν επανεκπαιδεύεται από τα synthetic vectors και το SMOTE δεν εφαρμόζεται στο raw text. Συνεπώς, η συγκεκριμένη προσομοίωση εξετάζει αν η εξισορρόπηση του embedding space βοηθά έναν downstream γραμμικό classifier να αναπαράγει καλύτερα τις κατηγορίες συναισθήματος. Τα αποτελέσματα θα παρουσιαστούν στο Κεφάλαιο 6 χωρίς να θεωρείται προκαταβολικά ότι η synthetic oversampling προσφέρει βελτίωση.

3.9 Σύγκριση Random Oversampling και SMOTE

Το ROS και το SMOTE ανήκουν και τα δύο στις data-level τεχνικές υπερδειγματοληψίας, αλλά διαφέρουν ουσιαστικά στον μηχανισμό τους. Το ROS αυξάνει τη συχνότητα πραγματικών παρατηρήσεων μέσω αντιγραφής, ενώ το SMOTE δημιουργεί synthetic vectors μέσω παρεμβολής. Το πρώτο μπορεί να εφαρμοστεί άμεσα σε text records πριν από το tokenization, ενώ το δεύτερο απαιτεί αριθμητική αναπαράσταση με meaningful geometry. Η απλότητα του ROS συνοδεύεται από κίνδυνο overfitting σε επαναλαμβανόμενα samples, ενώ το SMOTE μειώνει την ακριβή αντιγραφή αλλά μπορεί να εισαγάγει synthetic σημεία σε ακατάλληλες περιοχές του feature space [3], [12], [40]-[44].

Στην παρούσα εργασία οι δύο μέθοδοι δεν συγκρίνονται ως απόλυτα ισοδύναμες πάνω στο ίδιο μοντέλο. Το ROS χρησιμοποιείται με BERT πάνω στα αρχικά training records, ενώ το SMOTE χρησιμοποιείται σε RoBERTa embeddings με Logistic Regression. Η σύγκριση επομένως έχει κυρίως μεθοδολογικό χαρακτήρα: εξετάζονται δύο διαφορετικές στρατηγικές εξισορρόπησης σε δύο διαφορετικά pipelines. Για ασφαλή συμπεράσματα σχετικά με την ανωτερότητα μιας τεχνικής θα απαιτούνταν κοινό dataset, ίδια labels, ίδιο validation protocol και όσο το δυνατόν ίδιος downstream classifier. Η εργασία διατηρεί αυτή τη διάκριση ώστε τα αποτελέσματα να ερμηνεύονται με ακρίβεια.

Πίνακας 3.1: Συγκριτικά χαρακτηριστικά Random Oversampling και SMOTE

Χαρακτηριστικό	Random Oversampling (ROS)	SMOTE
Βασικός μηχανισμός	Τυχαία επανάληψη υπαρχόντων minority samples	Δημιουργία synthetic samples με παρεμβολή
Νέα πληροφορία	Δεν δημιουργεί νέα χαρακτηριστικά	Δημιουργεί νέα σημεία στο feature space
Εφαρμογή σε raw text	Ναι, σε επίπεδο εγγραφών	Όχι άμεσα
Απαίτηση αριθμητικής αναπαράστασης	Όχι	Ναι
Κύριος κίνδυνος	Overfitting σε επαναλαμβανόμενα samples	Synthetic σημεία σε αμφίβολες περιοχές/overlap
Χρήση στην παρούσα εργασία	BERT training data - δύο ROS σενάρια	RoBERTa embeddings + Logistic Regression

Ο Πίνακας 3.1 συνοψίζει τις κυριότερες διαφορές των δύο τεχνικών. Στην πειραματική αξιολόγηση, το βασικό κριτήριο δεν είναι αν το training set γίνεται αριθμητικά ισορροπημένο, αλλά αν η παρέμβαση βελτιώνει τη γενίκευση στις σπανιότερες κατηγορίες. Για αυτό τα αποτελέσματα ερμηνεύονται κυρίως με βάση το macro F1 και τις ανά κλάση τιμές precision, recall και F1-score, ενώ η accuracy χρησιμοποιείται συμπληρωματικά.

3.10 Επίλογος

Στο κεφάλαιο αυτό αναλύθηκε το πρόβλημα της μη ισορροπημένης κατανομής δεδομένων και η επίδρασή του στην εκπαίδευση και αξιολόγηση μοντέλων ταξινόμησης. Παρουσιάστηκαν οι έννοιες majority και minority class, ο λόγος ανισορροπίας και οι λόγοι για τους οποίους μία υψηλή accuracy μπορεί να είναι παραπλανητική όταν οι κλάσεις έχουν πολύ διαφορετικές συχνότητες. Εξετάστηκαν επίσης οι βασικές κατηγορίες αντιμετώπισης του imbalance, με έμφαση στις data-level μεθόδους που αποτελούν τον κεντρικό άξονα της εργασίας.

Το Random Oversampling επαναλαμβάνει πραγματικά minority samples και προσφέρει μία απλή, εύκολα ερμηνεύσιμη μέθοδο εξισορρόπησης, αλλά μπορεί να αυξήσει τον κίνδυνο overfitting. Το SMOTE, αντίθετα, δημιουργεί synthetic παρατηρήσεις μεταξύ γειτονικών minority vectors και απαιτεί κατάλληλο αριθμητικό feature space. Για κειμενικά δεδομένα αυτό οδηγεί στην εφαρμογή του πάνω σε embeddings και όχι πάνω στο raw text. Οι δύο προσεγγίσεις συνδέονται άμεσα με τα πειράματα BERT και RoBERTa της διπλωματικής. Στο επόμενο κεφάλαιο παρουσιάζονται αναλυτικότερα οι Transformer αρχιτεκτονικές και τα μοντέλα BERT και RoBERTa που χρησιμοποιούνται για την υλοποίηση αυτών των πειραματικών pipelines.

Κεφάλαιο 4ο: Transformers, BERT και RoBERTa

4.1 Εισαγωγή

Η εμφάνιση της αρχιτεκτονικής Transformer αποτέλεσε καθοριστική αλλαγή στην Επεξεργασία Φυσικής Γλώσσας, καθώς επέτρεψε την αποτελεσματική μοντελοποίηση σχέσεων μεταξύ διαφορετικών θέσεων μιας ακολουθίας χωρίς την αναδρομική επεξεργασία που χαρακτηρίζει τα RNN και LSTM. Το βασικό στοιχείο της αρχιτεκτονικής είναι ο μηχανισμός attention, μέσω του οποίου το μοντέλο υπολογίζει δυναμικά ποιες λέξεις ή υπομονάδες του κειμένου είναι περισσότερο σχετικές μεταξύ τους για τη δημιουργία της αναπαράστασης κάθε θέσης. Η προσέγγιση αυτή παρουσιάστηκε από τους Vaswani et al. [4] και αποτέλεσε τη βάση για μία μεγάλη οικογένεια προεκπαιδευμένων γλωσσικών μοντέλων.

Στην παρούσα διπλωματική ιδιαίτερη σημασία έχουν δύο Transformer μοντέλα: το BERT και το RoBERTa. Το BERT χρησιμοποιείται ως μοντέλο απευθείας ταξινόμησης κριτικών σε πέντε κατηγορίες βαθμολογίας, τόσο πριν όσο και μετά την εφαρμογή Random Oversampling. Το RoBERTa αξιοποιείται σε πειράματα ανάλυσης συναισθήματος και ως μηχανισμός παραγωγής contextual embeddings πάνω στα οποία εφαρμόζεται SMOTE και στη συνέχεια Logistic Regression. Για τον λόγο αυτό, το κεφάλαιο δεν περιορίζεται σε μία γενική παρουσίαση των Transformers, αλλά αναλύει τις ιδιότητες των BERT και RoBERTa που συνδέονται άμεσα με τα πειραματικά pipelines της εργασίας.

4.2 Η αρχιτεκτονική Transformer

Η αρχιτεκτονική Transformer βασίζεται σε ένα σύνολο μηχανισμών που επιτρέπουν την επεξεργασία μιας ακολουθίας χωρίς την αποκλειστική χρήση αναδρομικών δομών. Στις υποενότητες που ακολουθούν παρουσιάζονται ο Encoder και ο Decoder, ο μηχανισμός Self-Attention, οι έννοιες Query, Key και Value, το Multi-Head Attention και το Positional Encoding.

4.2.1 Encoder και Decoder

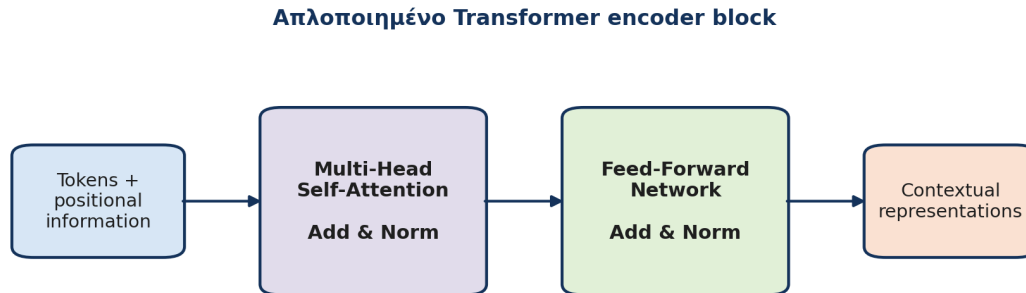
Ο αρχικός Transformer σχεδιάστηκε ως αρχιτεκτονική encoder-decoder για προβλήματα μετατροπής ακολουθιών, όπως η μηχανική μετάφραση [4]. Ο encoder λαμβάνει την ακολουθία εισόδου και δημιουργεί μία σειρά συμφραζόμενων αναπαραστάσεων. Ο decoder χρησιμοποιεί αυτές τις αναπαραστάσεις μαζί με τα tokens που έχουν ήδη παραχθεί, ώστε να υπολογίσει το επόμενο token της εξόδου. Κάθε πλευρά αποτελείται από διαδοχικά επαναλαμβανόμενα blocks, τα οποία περιλαμβάνουν attention μηχανισμούς, πλήρως συνδεδεμένα feed-forward δίκτυα, residual connections και normalization.

Για εργασίες κατανόησης κειμένου δεν είναι απαραίτητο να χρησιμοποιούνται και τα δύο μέρη. Το BERT και το RoBERTa είναι encoder-only μοντέλα: χρησιμοποιούν το τμήμα του Transformer που είναι κατάλληλο για τη δημιουργία ισχυρών αναπαραστάσεων μιας δεδομένης ακολουθίας και όχι για αυτοπαλίνδρομη παραγωγή νέου κειμένου. Η επιλογή αυτή ταιριάζει σε προβλήματα ταξινόμησης, επειδή ο στόχος είναι να εξαχθεί μία αναπαράσταση της εισόδου και στη συνέχεια να αντιστοιχιστεί σε μία ή περισσότερες ετικέτες.

Κάθε encoder block δέχεται ένα σύνολο διανυσμάτων και επιστρέφει νέα διανύσματα της ίδιας ακολουθίας, αλλά πλέον εμπλουτισμένα με πληροφορία από άλλες θέσεις. Η επανάληψη αυτής της διαδικασίας σε πολλά επίπεδα επιτρέπει στα ανώτερα layers να αποτυπώνουν πιο σύνθετες γλωσσικές

σχέσεις. Έτσι, μία λέξη δεν αναπαρίσταται μόνο με βάση την αρχική της embedding τιμή, αλλά σταδιακά ενσωματώνει πληροφορία από το περιβάλλον της.

Η βασική δομή ενός Transformer encoder block παρουσιάζεται σε απλοποιημένη μορφή στο Σχήμα 4.1.



Το BERT και το RoBERTa χρησιμοποιούν διαδοχικά encoder blocks.

Σχήμα 4.1: Απλοποιημένη δομή ενός Transformer encoder block.

Πηγή: Ιδία επεξεργασία.

4.2.2 Self-Attention

Το self-attention είναι ο βασικός μηχανισμός με τον οποίο κάθε θέση μιας ακολουθίας μπορεί να συσχετιστεί άμεσα με όλες τις υπόλοιπες θέσεις. Σε αντίθεση με ένα αναδρομικό δίκτυο, όπου η πληροφορία πρέπει να μεταφερθεί διαδοχικά από token σε token, ο Transformer μπορεί σε ένα attention layer να υπολογίσει απευθείας τη σχέση δύο απομακρυσμένων θέσεων. Αυτό είναι σημαντικό στη γλώσσα, επειδή η ερμηνεία μιας λέξης μπορεί να εξαρτάται από στοιχεία που βρίσκονται αρκετές θέσεις πριν ή μετά από αυτή.

Για κάθε token υπολογίζονται βάρη attention ως προς τα υπόλοιπα tokens. Τα βάρη αυτά λειτουργούν ως συντελεστές σχετικότητας: υψηλότερη τιμή σημαίνει ότι το περιεχόμενο μιας άλλης θέσης πρέπει να συμβάλει περισσότερο στη νέα αναπαράσταση. Η διαδικασία πραγματοποιείται για κάθε θέση της ακολουθίας, με αποτέλεσμα όλες οι αναπαραστάσεις να γίνονται context-dependent. Στην ανάλυση συναισθήματος αυτό βοηθά, για παράδειγμα, στη σύνδεση ενός επιθέτου με το αντικείμενο που περιγράφει ή μιας αρνητικής λέξης με μία προηγούμενη άρνηση.

Η βασική μορφή του scaled dot-product attention μπορεί να αποδοθεί ως εξής [4]:

$$\text{Attention}(Q, K, V) = \text{softmax}((QK^T) / \text{sqrt}(d_k)) V \quad (4.1)$$

όπου d_k είναι η διάσταση των key vectors. Η διαίρεση με $\text{sqrt}(d_k)$ λειτουργεί ως scaling factor πριν από το softmax, ώστε τα εσωτερικά γινόμενα να παραμένουν σε αριθμητικό εύρος κατάλληλο για σταθερή βελτιστοποίηση. Η εξίσωση υπολογίζεται παράλληλα για όλες τις θέσεις μέσω πράξεων πινάκων, γεγονός που συμβάλλει στην αποδοτική αξιοποίηση σύγχρονων GPUs.

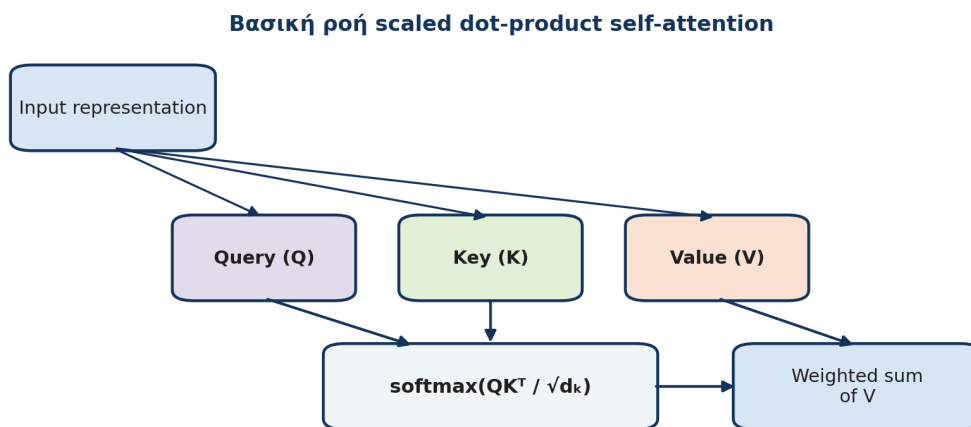
Η άμεση σύγκριση όλων των θέσεων έχει, ωστόσο, υπολογιστικό κόστος που αυξάνεται τετραγωνικά ως προς το μήκος της ακολουθίας, επειδή σχηματίζεται πίνακας attention διαστάσεων περίπου $n \times n$. Για σχετικά σύντομα κείμενα η ιδιότητα αυτή είναι διαχειρίσιμη και επιτρέπει υψηλό βαθμό

παραλληλισμού, όμως σε πολύ μεγάλες ακολουθίες αυξάνει σημαντικά τη μνήμη και τον χρόνο επεξεργασίας. Ο περιορισμός αυτός εξηγεί γιατί τα πρακτικά Transformer pipelines ορίζουν μέγιστο sequence length και γιατί το truncation που παρουσιάστηκε στο Κεφάλαιο 2 αποτελεί ουσιαστική σχεδιαστική επιλογή και όχι απλώς τεχνική ρύθμιση.

4.2.3 Query, Key και Value

Ο υπολογισμός του attention βασίζεται σε τρεις διαφορετικές γραμμικές προβολές της εισόδου, οι οποίες ονομάζονται Query (Q), Key (K) και Value (V). Διαισθητικά, το Query εκφράζει το είδος της πληροφορίας που αναζητεί μία θέση, το Key περιγράφει την πληροφορία με την οποία μπορεί να γίνει αντιστοίχιση και το Value περιέχει το περιεχόμενο που τελικά θα συνεισφέρει στη νέα αναπαράσταση. Οι τρεις πίνακες δεν αποτελούν ξεχωριστές λέξεις ή χειροκίνητα χαρακτηριστικά, αλλά μαθαίνονται από το δίκτυο μέσω διαφορετικών παραμέτρων.

Η σχέση μεταξύ Query, Key και Value στον υπολογισμό του scaled dot-product attention απεικονίζεται στο Σχήμα 4.2.



Σχήμα 4.2: Βασική ροή του μηχανισμού self-attention μέσω Query, Key και Value.

Πηγή: Ιδία επεξεργασία.

Η ομοιότητα ενός Query με τα Keys υπολογίζεται μέσω εσωτερικού γινομένου. Τα αποτελέσματα διαιρούνται με τη ρίζα της διάστασης d_k ώστε να περιορίζεται η αύξηση του μεγέθους των τιμών σε μεγάλες διαστάσεις και στη συνέχεια εφαρμόζεται softmax. Τα τελικά βάρη πολλαπλασιάζονται με τα Values. Με αυτό τον τρόπο, η νέα αναπαράσταση κάθε token είναι σταθμισμένος συνδυασμός πληροφοριών από ολόκληρη την ακολουθία. Οι πίνακες Q, K και V εκπαιδεύονται μαζί με τις υπόλοιπες παραμέτρους, επομένως το μοντέλο μαθαίνει ποιες σχέσεις είναι χρήσιμες για τον γλωσσικό στόχο.

4.2.4 Multi-Head Attention

Ένα μοναδικό attention mechanism μπορεί να επικεντρωθεί σε συγκεκριμένο είδος σχέσεων, αλλά η γλώσσα περιλαμβάνει πολλές διαφορετικές αλληλεπιδράσεις ταυτόχρονα. Για αυτό ο Transformer χρησιμοποιεί Multi-Head Attention. Η ίδια είσοδος προβάλλεται σε πολλαπλά διαφορετικά σύνολα Query, Key και Value, και κάθε head υπολογίζει ανεξάρτητα τα δικά του attention weights. Τα αποτελέσματα των heads συνενώνονται και μετασχηματίζονται ξανά σε κοινή αναπαράσταση [4].

Η ύπαρξη πολλών heads παρέχει στο μοντέλο τη δυνατότητα να αποτυπώνει παράλληλα διαφορετικούς τύπους εξαρτήσεων. Ένα head μπορεί να δίνει μεγαλύτερη έμφαση σε τοπικές συντακτικές σχέσεις, ενώ ένα άλλο μπορεί να συσχετίζει απομακρυσμένες λέξεις ή εκφράσεις. Δεν είναι απαραίτητο κάθε head να έχει μία σταθερή ανθρώπινη ερμηνεύσιμη λειτουργία, αλλά συλλογικά αυξάνουν την εκφραστική ικανότητα του μοντέλου. Στις base εκδόσεις των BERT και RoBERTa χρησιμοποιούνται 12 attention heads ανά Transformer layer [5], [6].

4.2.5 Positional Encoding

Το self-attention από μόνο του δεν περιλαμβάνει αναδρομή ή συνέλιξη και επομένως δεν ενσωματώνει εγγενώς τη σειρά των tokens. Αν δύο ακολουθίες περιείχαν ακριβώς τα ίδια tokens σε διαφορετική διάταξη, ένας καθαρός μηχανισμός attention χωρίς πληροφορία θέσης δεν θα είχε επαρκή τρόπο να διακρίνει τη σειρά τους. Για αυτό η αρχιτεκτονική Transformer προσθέτει positional information στις αναπαραστάσεις εισόδου [4].

Στον αρχικό Transformer χρησιμοποιούνται ντετερμινιστικές ημιτονοειδείς συναρτήσεις διαφορετικών συχνοτήτων. Στα BERT και RoBERTa χρησιμοποιούνται learned positional embeddings: για κάθε θέση υπάρχει ένα διάνυσμα θέσης το οποίο εκπαιδεύεται μαζί με τις υπόλοιπες παραμέτρους. Το positional embedding προστίθεται στην token representation πριν από τα Transformer layers, επιτρέποντας στο μοντέλο να διακρίνει, για παράδειγμα, τη διαφορά μεταξύ «not good» και «good, not excellent», όπου η θέση και η σύνδεση των λέξεων επηρεάζουν ουσιαστικά το συναίσθημα.

4.3 Προεκπαιδευμένα γλωσσικά μοντέλα

Η επιτυχία των σύγχρονων Transformer μοντέλων συνδέεται σε μεγάλο βαθμό με την προεκπαίδευση. Αντί ένα μοντέλο να εκπαιδεύεται αποκλειστικά πάνω σε ένα σχετικά μικρό labeled dataset για μία συγκεκριμένη εργασία, αρχικά εκπαιδεύεται σε πολύ μεγαλύτερα μη επισημασμένα σώματα κειμένου με έναν αυτοεπιβλεπόμενο γλωσσικό στόχο. Με αυτό τον τρόπο αποκτά γενικές γλωσσικές αναπαραστάσεις πριν προσαρμοστεί σε ένα downstream task. Η διαδικασία μειώνει την ανάγκη εκπαίδευσης ενός μεγάλου μοντέλου από την αρχή και επιτρέπει τη μεταφορά γνώσης από γενικά κείμενα σε ειδικότερες εφαρμογές.

Κατά την προεκπαίδευση, το δίκτυο μαθαίνει στατιστικές και σημασιολογικές κανονικότητες της γλώσσας. Δεν αποθηκεύει απλώς ένα στατικό λεξικό λέξεων, αλλά προσαρμόζει τα Transformer layers ώστε να προβλέπουν ή να ανακατασκευάζουν γλωσσικές πληροφορίες από τα συμφοραζόμενα. Μετά την ολοκλήρωση της προεκπαίδευσης, το ίδιο σύνολο βαρών μπορεί να χρησιμοποιηθεί σε classification, question answering, named entity recognition και πολλές άλλες εργασίες με περιορισμένη πρόσθετη εκπαίδευση [5].

Η λογική αυτή έχει ιδιαίτερη σημασία για τη διπλωματική, επειδή κανένα από τα μεγάλα Transformer μοντέλα δεν εκπαιδεύεται από μηδενική αρχικοποίηση. Χρησιμοποιούνται προεκπαιδευμένα checkpoints και η πειραματική διαδικασία επικεντρώνεται στην προσαρμογή ή στην αξιοποίηση των αναπαραστάσεών τους. Έτσι, το ερευνητικό ενδιαφέρον βρίσκεται κυρίως στην επίδραση της κατανομής των training labels και των oversampling τεχνικών και όχι στο κόστος εκπαίδευσης ενός γλωσσικού μοντέλου από την αρχή.

Η μεταφορά γνώσης δεν σημαίνει ότι το pretrained model είναι εξίσου κατάλληλο για κάθε domain. Η γλώσσα μιας εταιρικής κριτικής, μιας σύντομης δημοσίευσης κοινωνικού δικτύου και ενός ακαδημαϊκού κειμένου μπορεί να διαφέρει ως προς λεξιλόγιο, μήκος, ύφος και κατανομή συναισθημάτων. Το fine-tuning ή η επιλογή ενός domain-oriented checkpoint λειτουργεί ως

μηχανισμός προσαρμογής, αλλά η τελική συμπεριφορά πρέπει πάντοτε να επαληθεύεται στο dataset στόχου. Η παρατήρηση αυτή είναι ιδιαίτερα σχετική με το CardiffNLP Twitter RoBERTa, το οποίο έχει εκπαιδευτεί σε κοινωνικά δίκτυα αλλά εφαρμόζεται στην εργασία σε reviews διαφορετικού τύπου.

4.4 BERT

Το BERT αποτελεί ένα από τα βασικά Transformer μοντέλα που χρησιμοποιούνται στην πειραματική διαδικασία της εργασίας. Η ενότητα παρουσιάζει την αρχιτεκτονική του, τη bidirectional αναπαράσταση, τα στάδια pre-training και fine-tuning και τον τρόπο με τον οποίο μπορεί να χρησιμοποιηθεί για text classification.

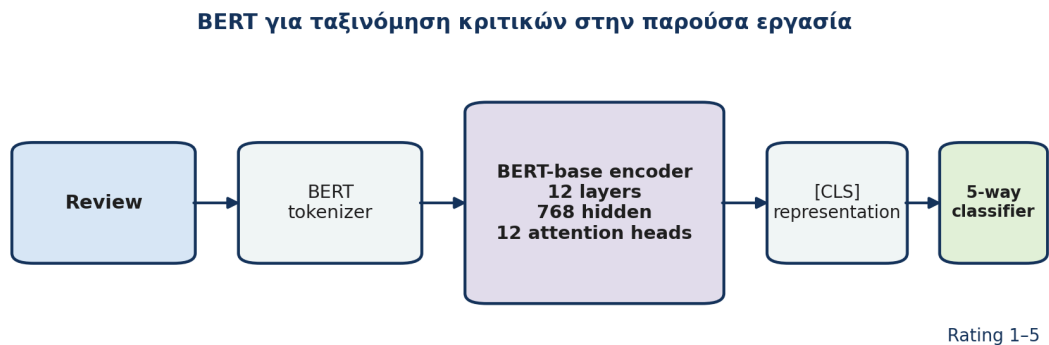
4.4.1 Αρχιτεκτονική BERT

Το BERT (Bidirectional Encoder Representations from Transformers) είναι encoder-only μοντέλο που βασίζεται στα Transformer encoder blocks και σχεδιάστηκε για τη δημιουργία βαθιά συμφραζόμενων αναπαραστάσεων [5]. Η έκδοση BERT-base αποτελείται από 12 Transformer layers, hidden size 768 και 12 attention heads, με περίπου 110 εκατομμύρια παραμέτρους. Η μεγαλύτερη έκδοση BERT-large χρησιμοποιεί περισσότερα layers και μεγαλύτερο hidden size, αλλά απαιτεί αντίστοιχα αυξημένους υπολογιστικούς πόρους.

Η είσοδος του BERT σχηματίζεται από το άθροισμα token embeddings, positional embeddings και segment embeddings. Το token embedding περιγράφει την υπομονάδα κειμένου, το positional embedding τη θέση της και το segment embedding επιτρέπει τη διάκριση δύο τμημάτων εισόδου όταν χρησιμοποιούνται ζεύγη προτάσεων. Η επεξεργασία αυτών των διανυσμάτων από τα διαδοχικά encoder layers παράγει για κάθε θέση μία τελική contextual representation 768 διαστάσεων στο BERT-base.

Στην παρούσα εργασία χρησιμοποιείται το checkpoint 'bert-base-uncased'. Η επιλογή uncased σημαίνει ότι το tokenizer και το λεξιλόγιο έχουν σχεδιαστεί χωρίς διάκριση κεφαλαίων και πεζών για το αγγλικό κείμενο. Πάνω από το BERT προστίθεται classification head πέντε εξόδων, επειδή το αντίστοιχο dataset περιλαμβάνει ratings από 1 έως 5. Με αυτό τον τρόπο το ίδιο pretrained backbone χρησιμοποιείται τόσο στο baseline όσο και στις δύο προσεγγίσεις αξιολόγησης του Random Oversampling.

Ο τρόπος με τον οποίο το BERT αξιοποιείται ως end-to-end classifier για τις πέντε κατηγορίες rating συνοψίζεται στο Σχήμα 4.3.



Σχήμα 4.3: Απλοποιημένο pipeline BERT για ταξινόμηση κριτικών σε πέντε ratings.

Πηγή: Ιδία επεξεργασία.

4.4.2 Bidirectional representation

Η βασική καινοτομία του BERT είναι ότι η αναπαράσταση ενός token διαμορφώνεται χρησιμοποιώντας συμφραζόμενα και από τις δύο κατευθύνσεις της ακολουθίας [5]. Σε ένα κλασικό αριστερό-προς-δεξιά γλωσσικό μοντέλο, η πρόβλεψη μιας θέσης βασίζεται μόνο στις προηγούμενες λέξεις. Το BERT, μέσω του Masked Language Modeling, μπορεί να χρησιμοποιεί τόσο το προηγούμενο όσο και το επόμενο κείμενο για να ανακατασκευάσει μία κρυμμένη λέξη. Αυτό επιτρέπει στις παραγόμενες αναπαραστάσεις να ενσωματώνουν πλουσιότερη πληροφορία για τη σημασία ενός όρου μέσα στην πρόταση.

Η ιδιότητα αυτή είναι σημαντική στην ανάλυση συναισθήματος. Η πολικότητα μιας λέξης μπορεί να ανατραπεί ή να περιοριστεί από λέξεις που προηγούνται ή ακολουθούν. Εκφράσεις όπως «not particularly good» ή «good at first, but disappointing overall» απαιτούν συνδυασμό πληροφορίας από διαφορετικά σημεία. Η bidirectional επεξεργασία δεν εγγυάται ότι κάθε σύνθετη περίπτωση θα ταξινομηθεί σωστά, αλλά παρέχει καταλληλότερη βάση από μία στατική ή μονοκατευθυντική αναπαράσταση.

4.4.3 Pre-training

Στο αρχικό BERT η προεκπαίδευση βασίζεται σε δύο στόχους: Masked Language Modeling (MLM) και Next Sentence Prediction (NSP) [5]. Στο MLM ένα ποσοστό των tokens επιλέγεται και το μοντέλο καλείται να προβλέψει την αρχική τους ταυτότητα χρησιμοποιώντας τα υπόλοιπα συμφραζόμενα. Η διαδικασία επιτρέπει την εκπαίδευση bidirectional representations χωρίς το μοντέλο να βλέπει απευθείας το token που πρέπει να ανακατασκευάσει.

Ο δεύτερος στόχος, NSP, παρουσιάζει στο μοντέλο δύο τμήματα κειμένου και ζητά να προβλέψει αν το δεύτερο ακολουθεί πραγματικά το πρώτο στο αρχικό corpus. Η πρόθεση ήταν να ενισχυθούν εργασίες που απαιτούν κατανόηση σχέσεων μεταξύ προτάσεων. Το RoBERTa αργότερα έδειξε ότι ο συγκεκριμένος στόχος δεν είναι απαραίτητος για ισχυρή απόδοση και τον αφαίρεσε από τη διαδικασία προεκπαίδευσης [6].

Η προεκπαίδευση του BERT πραγματοποιήθηκε σε μεγάλα αγγλικά corpora, και το αποτέλεσμα διατέθηκε ως pretrained checkpoint. Για το downstream sentiment classification δεν επαναλαμβάνεται αυτό το στάδιο. Αντίθετα, φορτώνονται τα ήδη προεκπαιδευμένα βάρη και πραγματοποιείται fine-tuning στο labeled dataset της εφαρμογής.

4.4.4 Fine-tuning

Το fine-tuning είναι το στάδιο κατά το οποίο ένα pretrained model προσαρμόζεται σε συγκεκριμένη εργασία. Συνήθως προστίθεται ένα μικρό task-specific layer πάνω από το Transformer backbone και το σύνολο ή μεγάλο μέρος των παραμέτρων ενημερώνεται μέσω backpropagation χρησιμοποιώντας το labeled training set. Σε σχέση με την προεκπαίδευση, απαιτούνται πολύ λιγότερα δεδομένα και βήματα εκπαίδευσης, επειδή το μοντέλο ξεκινά ήδη από γλωσσικά χρήσιμη κατάσταση [5].

Στο BERT experiment της εργασίας χρησιμοποιείται 'BertForSequenceClassification', δηλαδή η υλοποίηση του BERT με classification head. Το μοντέλο εκπαιδεύεται ώστε να επιστρέφει logits για πέντε κλάσεις. Η ίδια βασική αρχιτεκτονική επανεκπαιδεύεται στα διαφορετικά training scenarios, ώστε η βασική μεταβλητή ενδιαφέροντος να είναι η χρήση ή μη Random Oversampling και όχι η αλλαγή του Transformer backbone.

4.4.5 BERT για text classification

Για classification, το BERT χρησιμοποιεί μία ειδική θέση στην αρχή της ακολουθίας, το token '[CLS]'. Η τελική αναπαράσταση αυτής της θέσης τροφοδοτείται σε ένα classification layer και μετατρέπεται σε logits για τις διαθέσιμες κλάσεις. Κατά το fine-tuning, η cross-entropy loss συγκρίνει τα logits με το πραγματικό label και το gradient ενημερώνει τόσο το classification head όσο και τα βάρη του BERT. Το μοντέλο μαθαίνει έτσι να μετασχηματίζει τις γενικές γλωσσικές αναπαραστάσεις σε features χρήσιμα για τη συγκεκριμένη ταξινόμηση.

Η πενταταξική πρόβλεψη ratings είναι πιο απαιτητική από μία απλή δυαδική θετική/αρνητική διάκριση. Οι ενδιάμεσες κατηγορίες μπορεί να μοιράζονται λεξιλόγιο και συναισθηματικές εκφράσεις με γειτονικά ratings, ενώ οι ακραίες βαθμολογίες συχνά περιέχουν πιο σαφή γλωσσικά σήματα. Εάν επιπλέον οι ενδιάμεσες κλάσεις εμφανίζονται πολύ λιγότερο συχνά, το classification head λαμβάνει σαφώς λιγότερα παραδείγματα για να μάθει τα όριά τους.

Ακριβώς αυτό το σημείο συνδέει τη θεωρία του BERT με το κεντρικό θέμα της διπλωματικής. Το Random Oversampling δεν αλλάζει τη Transformer αρχιτεκτονική. Μεταβάλλει τη συχνότητα με την οποία τα minority samples εμφανίζονται κατά την εκπαίδευση. Συνεπώς, η σύγκριση baseline και ROS μοντέλων εξετάζει αν η ίδια ισχυρή γλωσσική αρχιτεκτονική γενικεύει διαφορετικά όταν εκτίθεται σε διαφορετική κατανομή training examples.

4.5 Tokenization στο BERT

Πριν το κείμενο εισαχθεί στο BERT πρέπει να μετατραπεί σε ακολουθία token ids. Το BERT χρησιμοποιεί WordPiece tokenization, όπου μία λέξη μπορεί να διατηρηθεί ως ενιαίο token εφόσον υπάρχει στο λεξιλόγιο ή να χωριστεί σε μικρότερες υπομονάδες. Η υπολεξική προσέγγιση μειώνει το πρόβλημα των άγνωστων λέξεων, επειδή ακόμη και ένας σπάνιος όρος μπορεί συχνά να αναπαρασταθεί ως συνδυασμός γνωστών subwords [5].

Στην αρχή της ακολουθίας προστίθεται '[CLS]' και στο τέλος '[SEP]'. Όταν οι ακολουθίες σε ένα batch έχουν διαφορετικό μήκος, χρησιμοποιείται padding ώστε να αποκτήσουν κοινό μήκος και

attention mask ώστε το μοντέλο να αγνοεί τις padding θέσεις. Αν μία ακολουθία ξεπερνά το επιλεγμένο μέγιστο μήκος, εφαρμόζεται truncation. Αυτές οι λειτουργίες δεν είναι απλή τεχνική λεπτομέρεια, επειδή καθορίζουν ποιο μέρος μιας μεγάλης κριτικής φτάνει πραγματικά στο Transformer.

Στα BERT πειράματα της εργασίας χρησιμοποιείται μέγιστο μήκος 128 tokens. Η ίδια ρύθμιση εφαρμόζεται στα συγκρινόμενα BERT scenarios ώστε να υπάρχει συνέπεια ως προς τη μορφή της εισόδου. Με αυτό τον τρόπο, η επίδραση του ROS δεν συγχέεται με διαφορετικό tokenization ή διαφορετικό sequence length.

4.6 Embeddings του BERT

Ο όρος embedding στο BERT μπορεί να αναφέρεται σε διαφορετικά επίπεδα αναπαράστασης. Στην είσοδο υπάρχουν token, position και segment embeddings. Μετά από κάθε Transformer layer παράγονται contextual hidden states, τα οποία ενσωματώνουν όλο και περισσότερη πληροφορία από την ακολουθία. Στο BERT-base κάθε token αντιστοιχεί τελικά σε διάνυσμα 768 διαστάσεων. Επειδή τα hidden states είναι contextual, το ίδιο subword μπορεί να έχει διαφορετική τελική αναπαράσταση σε δύο διαφορετικές προτάσεις.

Για εφαρμογές classification μπορεί να χρησιμοποιηθεί η αναπαράσταση της θέσης '[CLS]', mean pooling πάνω στα tokens ή άλλες τεχνικές συγκέντρωσης. Όταν γίνεται end-to-end fine-tuning μέσω 'BertForSequenceClassification', η επιλογή της συνολικής αναπαράστασης είναι ενσωματωμένη στην αρχιτεκτονική. Όταν ο Transformer χρησιμοποιείται ως feature extractor, αντίθετα, ο ερευνητής μπορεί να εξαγάγει ρητά τα hidden states και να τα χρησιμοποιήσει σε ανεξάρτητο classifier.

Η επιλογή του pooling επηρεάζει το τι πληροφορία διατηρείται στο τελικό vector. Η χρήση της πρώτης ειδικής θέσης προσφέρει μία απλή σταθερού μήκους αναπαράσταση, ενώ το mean pooling συνδυάζει τις αναπαραστάσεις όλων των μη padding tokens. Δεν υπάρχει μία καθολικά βέλτιστη επιλογή για κάθε dataset. Στην παρούσα υλοποίηση RoBERTa επιλέγεται η αναπαράσταση της πρώτης θέσης από το τελευταίο hidden layer, ώστε κάθε review να μετατρέπεται με συνεπή τρόπο σε ένα vector 768 διαστάσεων πριν από την εφαρμογή SMOTE. Η επιλογή διατηρεί σταθερό τον feature extractor και καθιστά συγκρίσιμα τα downstream πειράματα με και χωρίς υπερδειγματοληψία.

Η διάκριση αυτή είναι χρήσιμη για τη σύγκριση με το RoBERTa experiment. Στα βασικά BERT πειράματα η ταξινόμηση πραγματοποιείται εντός του Transformer pipeline και το μοντέλο fine-tunes πάνω στα ratings. Στο πείραμα RoBERTa+SMOTE, αντίθετα, ο Transformer χρησιμοποιείται για να παράγει σταθερά 768-dimensional embeddings και η τελική ταξινόμηση γίνεται με Logistic Regression.

4.7 RoBERTa

Το RoBERTa αποτελεί βελτιστοποιημένη παραλλαγή της λογικής προεκπαίδευσης του BERT και χρησιμοποιείται στην εργασία τόσο για sentiment analysis όσο και για παραγωγή embeddings. Στις επόμενες υποενότητες παρουσιάζονται η βασική αρχιτεκτονική του, οι κύριες διαφορές του από το BERT και η αξιοποίησή του σε προβλήματα ανάλυσης συναισθήματος.

4.7.1 Βασική αρχιτεκτονική

Το RoBERTa (Robustly Optimized BERT Pretraining Approach) διατηρεί ουσιαστικά την encoder-only αρχιτεκτονική του BERT, αλλά επανεξετάζει τον τρόπο προεκπαίδευσης [6]. Η base

έκδοση χρησιμοποιεί 12 Transformer layers, hidden size 768 και 12 attention heads, δηλαδή έχει συγκρίσιμη βασική δομή με το BERT-base. Οι διαφορές στην απόδοση δεν οφείλονται επομένως σε μία εντελώς διαφορετική attention αρχιτεκτονική, αλλά κυρίως στις επιλογές pretraining, tokenizer, δεδομένων και υπερπαραμέτρων.

Το RoBERTa χρησιμοποιεί byte-level Byte Pair Encoding (BPE), το οποίο επιτρέπει τη δημιουργία υπολεξικών μονάδων σε επίπεδο bytes και παρέχει πολύ ευρεία κάλυψη χαρακτήρων χωρίς να απαιτείται ξεχωριστό unknown token για τις περισσότερες εισόδους. Η συγκεκριμένη επιλογή είναι χρήσιμη σε θορυβώδες διαδικτυακό κείμενο, όπου εμφανίζονται ασυνήθιστες μορφές λέξεων, σύμβολα και μη τυπικές γραφές.

4.7.2 Διαφορές BERT και RoBERTa

Οι Liu et al. [6] έδειξαν ότι αρκετές επιλογές της αρχικής διαδικασίας προεκπαίδευσης του BERT μπορούσαν να βελτιστοποιηθούν. Το RoBERTa καταργεί το Next Sentence Prediction, χρησιμοποιεί dynamic masking αντί για μία στατική διαδικασία δημιουργίας masked παραδειγμάτων, εκπαιδεύεται με μεγαλύτερα batches και αξιοποιεί σημαντικά περισσότερα δεδομένα και training steps. Οι αλλαγές αυτές επιτρέπουν στο μοντέλο να εκμεταλλευθεί καλύτερα την ίδια βασική Transformer αρχιτεκτονική.

Μία ακόμη πρακτική διαφορά αφορά τον tokenizer. Το BERT-base-uncased χρησιμοποιεί WordPiece λεξιλόγιο περίπου 30 χιλιάδων tokens, ενώ το RoBERTa-base χρησιμοποιεί μεγαλύτερο byte-level BPE λεξιλόγιο. Η διαφορά επηρεάζει τον τρόπο διάσπασης των λέξεων και τον αριθμό των subwords που δημιουργούνται. Δεν σημαίνει ότι ένας tokenizer είναι καθολικά καλύτερος, αλλά ότι κάθε pretrained model πρέπει να χρησιμοποιείται με τον tokenizer που αντιστοιχεί στο checkpoint του.

Στον Πίνακα 4.1 συνοψίζονται οι βασικές διαφορές που είναι σημαντικές για την παρούσα εργασία. Και τα δύο μοντέλα χρησιμοποιούν bidirectional Transformer encoders, αλλά το BERT χρησιμοποιείται εδώ ως end-to-end classifier σε ratings, ενώ το RoBERTa αξιοποιείται τόσο ως sentiment classifier όσο και ως feature extractor.

Πίνακας 4.1: Βασική σύγκριση BERT-base και RoBERTa-base

Χαρακτηριστικό	BERT-base	RoBERTa-base
Transformer layers	12 encoder layers	12 encoder layers
Hidden size / attention heads	768 / 12	768 / 12
Tokenizer	WordPiece	Byte-level BPE
Pre-training objective	MLM + NSP	MLM χωρίς NSP
Masking	Κλασική BERT διαδικασία	Dynamic masking
Pre-training strategy	Αρχική BERT ρύθμιση	Περισσότερα δεδομένα, μεγαλύτερα batches και training
Χρήση στην εργασία	5-class rating classification + ROS	Sentiment classification και 768D embeddings + SMOTE

Ο Πίνακας 4.1 δείχνει ότι οι δύο base εκδόσεις έχουν παρόμοιο μέγεθος encoder και επομένως η σύγκρισή τους δεν πρέπει να ερμηνεύεται ως σύγκριση μίας μικρής και μίας πολύ μεγαλύτερης αρχιτεκτονικής. Οι ουσιαστικές διαφορές εντοπίζονται κυρίως στην προεκπαίδευση και στο tokenization. Στην παρούσα εργασία, επιπλέον, τα δύο μοντέλα χρησιμοποιούνται σε διαφορετικούς πειραματικούς ρόλους, γεγονός που αποκλείει την παρουσίαση των αποτελεσμάτων τους ως άμεσο head-to-head benchmark.

4.7.3 RoBERTa για sentiment analysis

Η χρήση του RoBERTa στην ανάλυση συναισθήματος ακολουθεί τη γενική λογική του Transformer classification. Η ακολουθία tokenized κειμένου επεξεργάζεται από τα encoder layers και μία συνολική αναπαράσταση τροφοδοτείται σε classification head. Με fine-tuning σε labeled sentiment data, το μοντέλο προσαρμόζει τα pretrained features ώστε να διακρίνει τις επιθυμητές κατηγορίες, όπως negative, neutral και positive.

Η πρακτική αξία ενός domain-oriented pretrained model είναι ότι έχει ήδη εκτεθεί σε γλωσσικά μοτίβα παρόμοια με το κείμενο εφαρμογής. Αυτό δεν σημαίνει ότι η μεταφορά σε άλλο είδος reviews είναι χωρίς περιορισμούς, επειδή Twitter posts και αναλυτικές κριτικές έχουν διαφορετικό ύφος και μήκος. Παρ' όλα αυτά, ένα RoBERTa μοντέλο που έχει προσαρμοστεί σε sentiment classification παρέχει ένα ισχυρό σημείο εκκίνησης τόσο για pseudo-labeling όσο και για παραγωγή sentiment-aware αναπαραστάσεων.

4.8 CardiffNLP Twitter RoBERTa

Στα RoBERTa πειράματα της εργασίας χρησιμοποιείται το checkpoint `cardiffnlp/twitter-roberta-base-sentiment`. Πρόκειται για RoBERTa-base μοντέλο που έχει υποστεί πρόσθετη προεκπαίδευση σε μεγάλο corpus αγγλικών tweets και fine-tuning για sentiment analysis στο πλαίσιο του TweetEval [15], [52]. Η σχετική εργασία TweetEval περιγράφει εκπαίδευση Twitter-specific RoBERTa σε περίπου 60 εκατομμύρια tweets, ενώ το μοντέλο sentiment χρησιμοποιεί τρεις κατηγορίες: negative, neutral και positive [15].

Η επιλογή του συγκεκριμένου checkpoint είναι σημαντική για τη μεθοδολογία, επειδή οι αρχικές sentiment labels στο RoBERTa pipeline προκύπτουν από ήδη εκπαιδευμένο classification head. Στο notebook τα scores μετατρέπονται με softmax και η κλάση με τη μεγαλύτερη πιθανότητα αντιστοιχίζεται σε negative, neutral ή positive. Σε διαφορετικό πείραμα, το classification head αφαιρείται μέσω `AutoModel` και χρησιμοποιείται το base network για την εξαγωγή embeddings.

Η χρήση pretrained sentiment μοντέλου προσφέρει πρακτική ευκολία, αλλά δημιουργεί και μεθοδολογικό περιορισμό όταν οι ίδιες προβλέψεις χρησιμοποιούνται ως pseudo-labels για downstream εκπαίδευση. Για αυτό τα αντίστοιχα αποτελέσματα της εργασίας δεν ερμηνεύονται ως

ανεξάρτητη μέτρηση ακρίβειας έναντι human ground truth. Αντίθετα, χρησιμοποιούνται για να μελετηθεί η συμπεριφορά του pipeline και, στην περίπτωση των embeddings, η επίδραση του SMOTE σε έναν downstream classifier.

4.9 Transformer embeddings για κλασικούς classifiers

Ένας Transformer δεν είναι απαραίτητο να λειτουργεί πάντοτε ως end-to-end classifier. Μπορεί να χρησιμοποιηθεί ως feature extractor: το κείμενο περνά από τα encoder layers, εξάγεται μία σταθερού μήκους διανυσματική αναπαράσταση και αυτή χρησιμοποιείται ως είσοδος σε διαφορετικό αλγόριθμο μηχανικής μάθησης. Η προσέγγιση συνδυάζει τη σημασιολογική ισχύ των contextual embeddings με την απλότητα και την ερμηνευσιμότητα ενός κλασικού classifier.

Στο RoBERTa+SMOTE πείραμα της εργασίας φορτώνεται το base RoBERTa μέσω 'AutoModel' χωρίς classification head. Από το 'last_hidden_state' επιλέγεται η αναπαράσταση της πρώτης ειδικής θέσης της ακολουθίας, δημιουργώντας για κάθε review ένα embedding 768 διαστάσεων. Το μοντέλο βρίσκεται σε evaluation mode και η εξαγωγή πραγματοποιείται χωρίς gradient calculation, επομένως τα RoBERTa weights δεν ενημερώνονται κατά τη δημιουργία των features.

Τα embeddings χωρίζονται σε training και validation sets πριν την υπερδειγματοληψία. Το SMOTE εφαρμόζεται μόνο στα training vectors και στη συνέχεια εκπαιδεύεται Logistic Regression. Για baseline εκπαιδεύεται ο ίδιος classifier στα αρχικά embeddings. Αυτή η αρχιτεκτονική επιτρέπει να εξεταστεί η επίδραση της εξισορρόπησης στον feature space χωρίς να μεταβάλλεται το Transformer backbone ή να δημιουργούνται συνθετικά raw texts.

Η χρήση ενός frozen feature extractor έχει και περιορισμούς. Επειδή το Transformer δεν προσαρμόζεται στον downstream classifier, τα embeddings παραμένουν αυτά που προκύπτουν από το pretrained checkpoint. Ένα end-to-end fine-tuned model θα μπορούσε να μετακινήσει τον representation space προς τα όρια του συγκεκριμένου dataset. Από την άλλη πλευρά, η σταθερότητα των features αποτελεί πλεονέκτημα για ένα πείραμα SMOTE, επειδή μειώνει τον αριθμό των μεταβαλλόμενων παραγόντων και επιτρέπει καθαρότερη σύγκριση πριν και μετά το oversampling.

4.10 BERT και RoBERTa σε μη ισορροπημένα δεδομένα

Η ισχυρή προεκπαίδευση των BERT και RoBERTa δεν εξαλείφει το πρόβλημα της μη ισορροπημένης εκπαίδευσης. Όταν το downstream loss υπολογίζεται πάνω σε πολύ άνισο αριθμό labels, οι συχνότερες κλάσεις εξακολουθούν να συνεισφέρουν περισσότερα training examples και μπορούν να κυριαρχήσουν στη βελτιστοποίηση. Το pretrained backbone παρέχει καλύτερες αρχικές αναπαραστάσεις, αλλά το τελικό decision boundary εξακολουθεί να εξαρτάται από τα δεδομένα που παρουσιάζονται στο μοντέλο.

Στο BERT pipeline η παρέμβαση πραγματοποιείται πριν από το tokenization και το fine-tuning: το Random Oversampling αυξάνει τον αριθμό των raw training records των minority ratings. Έτσι το ίδιο BERT architecture βλέπει συχνότερα τις σπάνιες κλάσεις. Στο RoBERTa pipeline η παρέμβαση πραγματοποιείται μετά την εξαγωγή embeddings: το SMOTE δημιουργεί synthetic training vectors και ο downstream Logistic Regression εκπαιδεύεται στον εξισορροπημένο feature space. Πρόκειται επομένως για δύο διαφορετικά σημεία παρέμβασης στο συνολικό NLP pipeline.

Η αξιολόγηση πρέπει να διατηρεί validation/test data έξω από οποιαδήποτε διαδικασία oversampling. Διαφορετικά, επαναλαμβανόμενα ή συνθετικά παραδείγματα μπορούν να διαρρεύσουν στην αξιολόγηση και να δημιουργήσουν υπερασπιδόμενες μετρικές. Για αυτό, ιδιαίτερο βάρος στην παρούσα

εργασία δίνεται στο δεύτερο BERT ROS σενάριο, όπου το validation set παραμένει ανέγγιχτο, και στο RoBERTa+SMOTE πείραμα, όπου το SMOTE εφαρμόζεται αποκλειστικά στα training embeddings.

Το βασικό συμπέρασμα σε επίπεδο αρχιτεκτονικής είναι ότι oversampling και Transformer δεν είναι ανταγωνιστικές έννοιες. Ο Transformer καθορίζει τον τρόπο αναπαράστασης και μάθησης του κειμένου, ενώ το oversampling μεταβάλλει την κατανομή των training observations ή features. Η αποτελεσματικότητα του συνδυασμού τους αποτελεί εμπειρικό ερώτημα και πρέπει να κρίνεται με macro metrics, ανά κλάση recall/F1 και confusion matrices και όχι μόνο με συνολική accuracy.

4.11 Επίλογος

Στο κεφάλαιο αυτό παρουσιάστηκε η αρχιτεκτονική Transformer και αναλύθηκαν τα βασικά στοιχεία που επιτρέπουν τη δημιουργία contextual representations: self-attention, Query-Key-Value, Multi-Head Attention και positional information. Στη συνέχεια παρουσιάστηκε η λογική της προεκπαίδευσης και της μεταφοράς γνώσης, πάνω στην οποία βασίζονται τα BERT και RoBERTa. Το BERT χρησιμοποιεί encoder-only bidirectional Transformer και μπορεί να προσαρμοστεί αποτελεσματικά σε text classification μέσω fine-tuning, ενώ το RoBERTa διατηρεί τη βασική αρχιτεκτονική αλλά τροποποιεί σημαντικές επιλογές της διαδικασίας pretraining [5], [6].

Η σύνδεση με την εργασία είναι άμεση: το `bert-base-uncased` χρησιμοποιείται στα baseline και Random Oversampling πειράματα πέντε ratings, ενώ το `cardiffnlp/twitter-roberta-base-sentiment` χρησιμοποιείται για τριταξική sentiment analysis και για εξαγωγή 768-dimensional embeddings πριν από το SMOTE και τη Logistic Regression. Το επόμενο κεφάλαιο παρουσιάζει αναλυτικά τα datasets, το υπολογιστικό περιβάλλον, τις παραμέτρους και τις πειραματικές ροές.

Κεφάλαιο 5ο: Μεθοδολογία και Υλοποίηση

5.1 Εισαγωγή

Το παρόν κεφάλαιο περιγράφει τη μεθοδολογία και την πρακτική υλοποίηση των πειραμάτων που πραγματοποιήθηκαν για τη μελέτη τεχνικών υπερδειγματοληψίας σε εφαρμογές ανάλυσης συναισθήματος με έντονα μη ισορροπημένα δεδομένα. Σε αντίθεση με τα προηγούμενα κεφάλαια, όπου παρουσιάστηκε το θεωρητικό υπόβαθρο της ανάλυσης συναισθήματος, των Transformers και των τεχνικών Random Oversampling και SMOTE, εδώ το ενδιαφέρον μεταφέρεται στην ακριβή ακολουθία επεξεργασίας των δεδομένων, στις παραμέτρους των μοντέλων και στον τρόπο με τον οποίο οργανώθηκαν οι διαφορετικές πειραματικές προσομοιώσεις.

Η εργασία περιλαμβάνει δύο κύριες οικογένειες πειραμάτων. Η πρώτη βασίζεται στο BERT και σε ένα σύνολο 60.000 επισημασμένων εταιρικών κριτικών με βαθμολογία από 1 έως 5. Στο πλαίσιο αυτό εξετάζεται η εκπαίδευση του `'bert-base-uncased'` πριν και μετά την εφαρμογή Random Oversampling. Η τεχνική ROS εξετάζεται με δύο διαφορετικούς τρόπους. Στο πρώτο σενάριο εφαρμόζεται στο πλήρες training dataset και μελετάται η συμπεριφορά του μοντέλου και η κατανομή των προβλέψεων στο μη επισημασμένο test set. Στο δεύτερο σενάριο πραγματοποιείται πρώτα stratified train/validation split και το ROS εφαρμόζεται αποκλειστικά στο training split, ώστε η αξιολόγηση να γίνεται σε ανεξάρτητο και ανέγγιχτο validation set.

Η δεύτερη οικογένεια πειραμάτων βασίζεται στο RoBERTa. Αρχικά χρησιμοποιείται το `'cardiffnlp/twitter-roberta-base-sentiment'` ως προεκπαιδευμένο μοντέλο sentiment classification για την παραγωγή sentiment labels και για ένα βοηθητικό πείραμα pseudo-labeling/fine-tuning. Στη συνέχεια, σε διαφορετικό pipeline, το classification head αφαιρείται και το RoBERTa χρησιμοποιείται ως σταθερός feature extractor. Για κάθε κριτική παράγεται contextual embedding 768 διαστάσεων, το SMOTE εφαρμόζεται μόνο στα embeddings του training set και δύο Logistic Regression classifiers συγκρίνονται πάνω στο ίδιο validation set: ένας εκπαιδευμένος στα αρχικά μη ισορροπημένα embeddings και ένας στα εξισορροπημένα embeddings.

Οι δύο οικογένειες δεν αποτελούν άμεσο benchmark BERT έναντι RoBERTa, επειδή χρησιμοποιούν διαφορετικά datasets, διαφορετικούς στόχους ταξινόμησης και διαφορετικά σημεία εφαρμογής της υπερδειγματοληψίας. Ο κοινός άξονας είναι η διερεύνηση της επίδρασης της κατανομής των κλάσεων. Για τον λόγο αυτό, η μεθοδολογία εστιάζει όχι μόνο στη συνολική accuracy, αλλά και σε precision, recall, F1-score, macro averages και confusion matrices, με ιδιαίτερη έμφαση στις μειονοτικές κλάσεις.

5.2 Ερευνητική μεθοδολογία

Η πειραματική διαδικασία οργανώθηκε ως ακολουθία σταδίων που επαναλαμβάνεται, με κατάλληλες προσαρμογές, στα διαφορετικά notebooks. Τα βασικά στάδια είναι: φόρτωση και διερεύνηση δεδομένων, προεπεξεργασία, δημιουργία baseline, εφαρμογή τεχνικής εξισορρόπησης μόνο στην περιοχή της εκπαίδευσης όπου αυτό είναι δυνατό, επανεκπαίδευση ή εκπαίδευση downstream classifier και αξιολόγηση. Η λογική αυτή επιτρέπει να διαχωριστεί η επίδραση του ίδιου του μοντέλου από την επίδραση της μεταβολής της class distribution.

Στην περίπτωση του BERT, το input παραμένει raw text μέχρι τη διαδικασία tokenization. Το RandomOverSampler της βιβλιοθήκης `'imbalanced-learn'` λειτουργεί πάνω στις εγγραφές κειμένου

και τις αντίστοιχες βαθμολογίες, αντιγράφοντας παραδείγματα από τις λιγότερο συχνές κλάσεις. Μετά το oversampling πραγματοποιείται tokenization και fine-tuning νέου 'BertForSequenceClassification'. Επομένως η παρέμβαση γίνεται πριν από την εκπαίδευση του Transformer και αλλάζει τη συχνότητα με την οποία το μοντέλο βλέπει παραδείγματα κάθε rating.

Στην περίπτωση του RoBERTa+SMOTE, η σειρά είναι διαφορετική. Το κείμενο πρώτα μετατρέπεται σε αριθμητική αναπαράσταση μέσω του pretrained RoBERTa. Τα 768-dimensional embeddings αποτελούν τον feature space πάνω στον οποίο εφαρμόζεται το SMOTE. Η τεχνική δημιουργεί νέα συνθετικά vectors για τις minority classes χωρίς να παράγει τεχνητές προτάσεις ή να παρεμβαίνει στον tokenizer. Μετά την εξισορρόπηση, η Logistic Regression χρησιμοποιείται ως downstream classifier. Έτσι είναι δυνατή η άμεση σύγκριση του ίδιου classifier με και χωρίς SMOTE, ενώ το Transformer backbone παραμένει αμετάβλητο.

Ιδιαίτερη σημασία δίνεται στον διαχωρισμό training και validation data. Το δεύτερο BERT ROS σενάριο και το RoBERTa+SMOTE πείραμα πραγματοποιούν stratified split πριν την υπερδειγματοληψία και δεν εφαρμόζουν ROS ή SMOTE στο validation set. Αυτό είναι το καταλληλότερο πρωτόκολλο για τη μέτρηση γενίκευσης. Αντίθετα, το πρώτο BERT ROS σενάριο λειτουργεί ως διερευνητική προσομοίωση της επίδρασης ενός πλήρως εξισορροπημένου training dataset και των αλλαγών στην κατανομή των test predictions. Για τον λόγο αυτό, τα αποτελέσματα των δύο ROS σεναρίων ερμηνεύονται με διαφορετικό βάρος στο Κεφάλαιο 6.

Τα notebooks περιλαμβάνουν επαναχρησιμοποίηση global μεταβλητών και επανεκτέλεση κελιών. Ειδικά στο πρώτο ROS section, ένα αποθηκευμένο output με τίτλο «before balancing» εμφανίζει ήδη ισορροπημένη κατανομή, επειδή η μεταβλητή 'train_df' είχε τροποποιηθεί σε προηγούμενο κελί. Η μεθοδολογική περιγραφή της παρούσας εργασίας βασίζεται στις αρχικές τιμές του dataset και στο σαφώς παραγόμενο 'train_df_balanced' των 173.395 εγγράφων. Η παρατήρηση αυτή λαμβάνεται επίσης υπόψη στους περιορισμούς της πειραματικής διαδικασίας.

5.3 Υπολογιστικό περιβάλλον

Η υλοποίηση των πειραμάτων απαιτεί ένα υπολογιστικό περιβάλλον που να υποστηρίζει την εκτέλεση μοντέλων Transformer και των απαραίτητων βιβλιοθηκών Machine Learning. Η ενότητα παρουσιάζει το Google Colab, τη χρήση GPU acceleration και τη γλώσσα Python, τα οποία αποτέλεσαν τη βασική υποδομή της πειραματικής εργασίας.

5.3.1 Google Colab

Η ανάπτυξη και εκτέλεση των πειραμάτων πραγματοποιήθηκε σε notebooks στο Google Colab. Η επιλογή notebook περιβάλλοντος επιτρέπει την τμηματική εκτέλεση των σταδίων, την άμεση οπτικοποίηση των κατανομών και των confusion matrices και τη χρήση Python βιβλιοθηκών για NLP, deep learning και imbalanced learning στο ίδιο περιβάλλον. Τα δεδομένα φορτώνονται από διαδρομές στο Google Drive, γεγονός που διευκολύνει τη διατήρηση των CSV αρχείων και των παραγόμενων αποτελεσμάτων μεταξύ διαφορετικών συνεδριών.

Η μορφή notebook είναι ιδιαίτερα χρήσιμη για διερευνητική πειραματική εργασία, αλλά απαιτεί προσοχή ως προς τη σειρά εκτέλεσης. Μεταβλητές που τροποποιούνται σε ένα κελί μπορούν να επηρεάσουν μεταγενέστερα outputs αν δεν γίνει επανεκκίνηση του runtime. Για αυτό, κατά την τελική ερμηνεία των πειραμάτων εξετάζονται τόσο ο κώδικας όσο και τα αποθηκευμένα outputs, ενώ οι ισχυρότερες συγκρίσεις είναι εκείνες στις οποίες η ροή ορίζει ρητά νέο dataset ή νέο model instance.

Παρότι το Google Colab αποδείχθηκε ιδιαίτερα χρήσιμο για την ανάπτυξη και την εκτέλεση των πειραμάτων, στην πράξη παρουσίασε και έναν σημαντικό περιορισμό: το runtime δεν είναι μόνιμο. Σε περίπτωση αποσύνδεσης ή λήξης της συνεδρίας, οι μεταβλητές που βρίσκονται στη μνήμη χάνονται. Αν απαιτείται μεταγενέστερα η προσθήκη ενός νέου βήματος ή η επανεξέταση ενός προηγούμενου αποτελέσματος, είναι συχνά απαραίτητη η επανεκτέλεση μεγάλου μέρους του notebook.

Το πρόβλημα γίνεται περισσότερο αισθητό σε πειράματα με Transformer μοντέλα, επειδή στάδια όπως το tokenization, η παραγωγή embeddings και κυρίως το fine-tuning του BERT είναι χρονοβόρα. Έτσι, μία απλή διακοπή της συνεδρίας μπορεί να οδηγήσει σε επανάληψη διαδικασιών που είχαν ήδη ολοκληρωθεί. Η εμπειρία αυτή ανέδειξε στην πράξη τη σημασία της καλύτερης οργάνωσης των notebooks και της αποθήκευσης ενδιάμεσων αποτελεσμάτων.

5.3.2 GPU acceleration

Τα Transformer μοντέλα μεταφέρονται σε συσκευή CUDA όταν αυτή είναι διαθέσιμη. Στα notebooks χρησιμοποιείται η έκφραση ``torch.device("cuda" if torch.cuda.is_available else "cpu")``, και τα αποθηκευμένα outputs δείχνουν ότι οι συγκεκριμένες εκτελέσεις πραγματοποιήθηκαν σε ``cuda``. Η χρήση GPU είναι ιδιαίτερα σημαντική κυρίως για το fine-tuning, το inference και την εξαγωγή embeddings των Transformer μοντέλων, τα οποία απαιτούν σημαντικά περισσότερους υπολογισμούς σε σχέση με έναν κλασικό classifier.

Τα input tensors, τα attention masks και τα labels μεταφέρονται στη συσκευή πριν την forward pass. Κατά την inference ή την εξαγωγή embeddings χρησιμοποιείται ``torch.no_grad``, ώστε να μην αποθηκεύονται gradients που δεν χρειάζονται. Στα training loops εφαρμόζεται gradient clipping με μέγιστη νόρμα 1.0, ενώ μετά από κάθε optimizer step ενημερώνεται linear learning-rate scheduler.

5.3.3 Python

Η υλοποίηση πραγματοποιήθηκε σε Python. Η γλώσσα επιλέγεται φυσικά από το οικοσύστημα των χρησιμοποιούμενων εργαλείων: PyTorch για tensors και εκπαίδευση νευρωνικών μοντέλων, Hugging Face Transformers για BERT/RoBERTa, pandas και NumPy για διαχείριση δεδομένων, scikit-learn για splitting και metrics και imbalanced-learn για ROS και SMOTE. Η χρήση κοινής γλώσσας σε όλη τη ροή μειώνει την ανάγκη μετατροπής δεδομένων μεταξύ διαφορετικών περιβαλλόντων και επιτρέπει την αναπαραγωγή του pipeline σε ένα notebook.

5.4 Βιβλιοθήκες λογισμικού

Οι βασικές βιβλιοθήκες της υλοποίησης είναι οι ``transformers``, ``torch``, ``pandas``, ``numpy``, ``scikit-learn``, ``imbalanced-learn``, ``langdetect``, ``matplotlib`` και ``seaborn``. Το Hugging Face Transformers χρησιμοποιείται για τους tokenizers και τις Transformer αρχιτεκτονικές [19], [52], το PyTorch για tensors, GPU execution και training loops [18], η scikit-learn για splitting, Logistic Regression και metrics [17], ενώ η imbalanced-learn παρέχει τις υλοποιήσεις RandomOverSampler και SMOTE [16]. Οι υπόλοιπες βιβλιοθήκες χρησιμοποιούνται για διαχείριση δεδομένων, ανίχνευση γλώσσας και οπτικοποίηση.

Οι βιβλιοθήκες οπτικοποίησης χρησιμοποιούνται για την αρχική διερεύνηση και την παρουσίαση της κατανομής των ratings ή sentiment labels, καθώς και για confusion matrices και συγκρίσεις των προβλέψεων. Η ``langdetect`` χρησιμοποιείται στο RoBERTa/SMOTE notebook για τον εντοπισμό αγγλόφωνων κριτικών. Η επιλογή αυτή είναι σημαντική επειδή το χρησιμοποιούμενο RoBERTa

checkpoint έχει εκπαιδευτεί για αγγλικό social-media text και η ανάμειξη διαφορετικών γλωσσών θα περιέπλεκε την ερμηνεία των labels.

5.5 Σύνολα δεδομένων

Τα πειράματα της εργασίας βασίζονται σε σύνολα διαδικτυακών κριτικών με διαφορετικό ρόλο στα επιμέρους pipelines. Στις υποενότητες που ακολουθούν περιγράφονται τα δεδομένα κριτικών, η δομή των αρχείων, η κατανομή των κλάσεων και η διαδικασία ανίχνευσης και επιλογής των αγγλικών κειμένων όπου αυτή απαιτείται.

5.5.1 Δεδομένα κριτικών

Για τα BERT πειράματα χρησιμοποιείται αρχείο `train.csv` με 60.000 εγγραφές και πεδία `Id`, `Review` και `Rating`. Η τιμή `Rating` λαμβάνει ακέραιες τιμές από 1 έως 5 και χρησιμοποιείται ως target για πεντατάξική ταξινόμηση. Υπάρχει επιπλέον `test.csv` με 40.000 εγγραφές που περιέχει `Id` και `Review`, χωρίς πραγματική στήλη `Rating`. Το `sample\_submission.csv` δείχνει μόνο τη μορφή της απαιτούμενης πρόβλεψης. Συνεπώς, το συγκεκριμένο test set χρησιμοποιείται για παραγωγή predictions και σύγκριση κατανομών, όχι για υπολογισμό πραγματικής test accuracy.

Για το RoBERTa+SMOTE πείραμα χρησιμοποιείται αρχείο `reviews.csv` με στήλες `comments` και `listing\_id`. Από τις πρώτες 20.000 εγγραφές πραγματοποιείται ανίχνευση γλώσσας και προκύπτουν 16.512 αγγλόφωνες κριτικές. Το pretrained RoBERTa παράγει για αυτές τρεις sentiment classes: positive, negative και neutral. Για το βοηθητικό fine-tuning RoBERTa experiment χρησιμοποιείται ξεχωριστό `english\_reviews.csv`, ενώ το τελικό αρχείο προβλέψεων περιλαμβάνει 8.688 reviews με `listing\_id`, `text` και `predicted\_sentiment` [51].

Πίνακας 5.1: Σύνολα δεδομένων και ρόλος τους στα πειράματα

Πείραμα	Αρχικό πλήθος	Labels / κλάσεις	Χρήση
BERT company reviews	60.000 train + 40.000 test	Ratings 1-5 στο train	Baseline και δύο ROS σενάρια
RoBERTa + SMOTE	20.000 reviews εξετάζονται, 16.512 αγγλικά	negative / neutral / positive από pretrained RoBERTa	768D embeddings, SMOTE, Logistic Regression
RoBERTa pseudo-label/fine-tuning	8.688 reviews στο english reviews.csv	Binary pseudo-labels μετά από 3-class RoBERTa	Βοηθητικό fine-tuning experiment

Τα datasets χρησιμοποιούνται σε διαφορετικούς ρόλους και δεν συγχωνεύονται. Η κύρια σύγκριση πραγματοποιείται εντός κάθε πειραματικού pipeline, πριν και μετά την αντίστοιχη τεχνική εξισορρόπησης.

5.5.2 Δομή των συνόλων δεδομένων

Η δομή των δύο πηγών αντικατοπτρίζει δύο διαφορετικές διατυπώσεις του sentiment problem. Στο BERT dataset η ετικέτα είναι αριθμητική βαθμολογία 1-5. Για την εκπαίδευση, οι τιμές μετατρέπονται από 1-5 σε 0-4, όπως απαιτεί συνήθως το `CrossEntropyLoss` για πέντε output logits. Στο RoBERTa dataset, αντίθετα, οι τρεις κλάσεις είναι `negative`, `neutral` και `positive`, οι οποίες αντιστοιχίζονται στους αριθμούς 0, 1 και 2 πριν τη δημιουργία embeddings και την εκπαίδευση της Logistic Regression.

Οι δύο αναπαραστάσεις δεν ενοποιούνται τεχνητά. Η πενταβάθμια κλίμακα διατηρείται στα BERT πειράματα επειδή επιτρέπει να φανεί αν οι ενδιάμεσες και σπανιότερες κλάσεις 2, 3 και 4 επηρεάζονται από την ανισορροπία. Η τριταξική προσέγγιση διατηρείται στο RoBERTa pipeline επειδή προκύπτει απευθείας από το προεκπαιδευμένο sentiment checkpoint.

5.5.3 Κατανομή κλάσεων

Το BERT training dataset είναι έντονα μη ισορροπημένο. Η Rating 5 περιλαμβάνει 34.679 δείγματα και αποτελεί τη μεγαλύτερη κλάση, ενώ η Rating 1 περιλαμβάνει 18.663. Οι ενδιάμεσες κλάσεις είναι πολύ μικρότερες: 1.629 δείγματα για Rating 2, 1.679 για Rating 3 και 3.350 για Rating 4. Η κατανομή αυτή δημιουργεί κατάλληλη περίπτωση για μελέτη ROS, επειδή η πλειονοτική κλάση είναι πάνω από είκοσι φορές μεγαλύτερη από τη μικρότερη.

Στο RoBERTa dataset η ανισορροπία είναι ακόμη εντονότερη. Από τις 16.512 αγγλόφωνες κριτικές, το αρχικό pretrained classifier αποδίδει 16.297 ως positive, 113 ως negative και 102 ως neutral. Μετά από stratified 80/20 split, το training set των 13.209 embeddings περιλαμβάνει 13.037 positive, 90 negative και 82 neutral δείγματα. Αυτή η κατανομή αποτελεί την είσοδο στο SMOTE.

5.5.4 Ανίχνευση και επιλογή αγγλικών κριτικών

Στο RoBERTa+SMOTE notebook εξετάζονται οι πρώτες 20.000 τιμές της στήλης 'comments'. Για κάθε review εφαρμόζεται η συνάρτηση 'detect' της βιβλιοθήκης 'langdetect'. Όταν επιστρέφεται η τιμή 'en', αποθηκεύονται το 'listing_id' και το κείμενο. Εξαιρέσεις κατά την ανίχνευση καταγράφονται σε ξεχωριστό DataFrame αντί να διακόπτουν τη διαδικασία. Από τα 20.000 αρχικά records, 16.512 ταξινομούνται ως αγγλικά και 25 επιπλέον εγγραφές εμφανίζονται ως 'Undetermined' λόγω σφάλματος ανίχνευσης.

Για το sentiment pipeline χρησιμοποιούνται μόνο οι εγγραφές με 'language_status = English'. Η επιλογή γίνεται πριν από την παραγωγή labels, embeddings ή το train/validation split, ώστε οι επόμενες φάσεις να λειτουργούν σε γλωσσικά ομοιογενέστερο υποσύνολο.

5.6 Προεπεξεργασία δεδομένων

Πριν από την εκπαίδευση των μοντέλων εφαρμόζεται σειρά βημάτων προεπεξεργασίας ώστε τα δεδομένα να έχουν συνεπή και κατάλληλη μορφή για κάθε πειραματικό pipeline. Η διαδικασία περιλαμβάνει αντιμετώπιση κενών τιμών, μετατροπή των πεδίων σε κείμενο, language filtering, tokenization και truncation.

5.6.1 Αντιμετώπιση κενών τιμών

Τα notebooks εφαρμόζουν κυρίως αμυντικές μετατροπές αντί εκτεταμένου linguistic cleaning. Στη RoBERTa συνάρτηση sentiment analysis, μη string τιμές μετατρέπονται σε string και η ειδική περίπτωση 'nan' επιστρέφει κενό αποτέλεσμα. Μετά την αντιστοίχιση sentiment labels σε αριθμητικά IDs, γραμμές όπου 'sentiment_label' ή 'label_id' είναι 'None' αφαιρούνται πριν από το split. Με τον τρόπο αυτό αποφεύγεται η είσοδος μη έγκυρων labels στη Logistic Regression ή στο SMOTE.

Στο BERT dataset ο κώδικας χρησιμοποιεί απευθείας τη στήλη 'Review', χωρίς ξεχωριστό στάδιο εκτεταμένης αφαίρεσης stopwords ή stemming. Αυτό είναι συμβατό με τη χρήση pretrained Transformers, καθώς ο tokenizer και το contextual model έχουν σχεδιαστεί να λειτουργούν πάνω σε φυσική γλωσσική ακολουθία και όχι σε χειροκίνητα απογυμνωμένο bag-of-words representation.

5.6.2 Μετατροπή δεδομένων σε κείμενο

Πριν από την επεξεργασία, τα reviews αντιμετωπίζονται ως συμβολοσειρές. Στο fine-tuning RoBERTa notebook χρησιμοποιείται ρητά 'df["text"].astype(str).tolist', ενώ στο sentiment function γίνεται 'str(text)' όταν η τιμή δεν είναι ήδη string. Στο BERT pipeline η 'Review' περνά απευθείας στον

tokenizer. Η συνέπεια του datatype είναι σημαντική, επειδή οι Hugging Face tokenizers αναμένουν κείμενο ή λίστα κειμένων και δεν μπορούν να επεξεργαστούν αυθαίρετες αριθμητικές ή κενές τιμές ως έγκυρες προτάσεις.

5.6.3 Language filtering

Η γλωσσική διήθηση εφαρμόζεται μόνο στο dataset κριτικών που χρησιμοποιείται με RoBERTa. Δεν υπάρχει αντίστοιχο βήμα στο company-reviews dataset του BERT notebook. Η διαφοροποίηση αυτή διατηρείται στη μεθοδολογία και δεν υποτίθεται ότι όλα τα datasets έχουν περάσει από την ίδια διαδικασία. Στο RoBERTa pipeline, το language filtering προηγείται της sentiment inference και επομένως επηρεάζει το τελικό πλήθος των 16.512 παραδειγμάτων.

5.6.4 Tokenization

Για το BERT χρησιμοποιείται `BertTokenizer.from_pretrained("bert-base-uncased")`. Κάθε review κωδικοποιείται με special tokens, attention mask, padding και truncation. Η μέγιστη ακολουθία ορίζεται στα 128 tokens. Τα `input_ids`, `attention_mask` και `token_type_ids` συνενώνονται σε tensors και χρησιμοποιούνται για τη δημιουργία `TensorDataset` και `DataLoader`. Τα ratings μετατρέπονται σε labels 0-4.

Στο RoBERTa+SMOTE pipeline χρησιμοποιείται `AutoTokenizer` του `cardiffnlp/twitter-roberta-base-sentiment` με padding σε `max_length`, truncation και μέγιστο μήκος 512. Τα tokenized δεδομένα αποθηκεύονται σε custom PyTorch Dataset και τροφοδοτούν DataLoaders με batch size 16. Στο fine-tuning pseudo-label experiment το ίδιο checkpoint χρησιμοποιείται με μέγιστο μήκος 128 για την παραγωγή pseudo-labels και την επανεκπαίδευση του binary classification head.

5.6.5 Truncation

Το truncation περιορίζει τις ακολουθίες σε ένα προκαθορισμένο μέγιστο μήκος, ενώ το padding χρησιμοποιείται για τη δημιουργία batches με συμβατές διαστάσεις. Στα BERT πειράματα χρησιμοποιείται `max_length=128`, επιλογή που μειώνει το υπολογιστικό κόστος του fine-tuning. Στο embedding experiment του RoBERTa χρησιμοποιείται `max_length=512`, επιτρέποντας μεγαλύτερο τμήμα κάθε review να συμβάλει στο embedding. Η διαφορά αυτή αποτελεί χαρακτηριστικό των δύο ανεξάρτητων pipelines και όχι παράμετρο που συγκρίνεται άμεσα.

5.7 Πείραμα A: BERT χωρίς oversampling

Το πρώτο BERT πείραμα λειτουργεί ως baseline, ώστε να αποτυπωθεί η συμπεριφορά του μοντέλου πριν από οποιαδήποτε παρέμβαση υπερδειγματοληψίας. Στις επόμενες υποενότητες περιγράφονται η αρχιτεκτονική που χρησιμοποιήθηκε, οι βασικές παράμετροι και η διαδικασία εκπαίδευσης.

5.7.1 Αρχιτεκτονική

Το BERT baseline χρησιμοποιεί `BertForSequenceClassification.from_pretrained("bert-base-uncased", num_labels=5)`. Το pretrained encoder φορτώνεται από το BERT checkpoint και το classification head προσαρμόζεται σε πέντε outputs, ένα για κάθε rating. Τα labels 1-5 μετατοπίζονται σε 0-4 πριν δοθούν στο μοντέλο. Η εκπαίδευση πραγματοποιείται με supervised cross-entropy μέσω του ενσωματωμένου loss της κλάσης `BertForSequenceClassification` [5].

Το ίδιο architecture επαναχρησιμοποιείται και στα ROS μοντέλα. Σε κάθε σημαντικό πειραματικό branch δημιουργείται νέο model instance, ώστε η εκπαίδευση μετά το oversampling να μην συνεχίζει απλώς από τα ήδη fine-tuned βάρη προηγούμενου baseline. Η επιλογή αυτή είναι αναγκαία για να αποδοθούν οι διαφορές στη νέα training distribution και όχι σε πρόσθετες epochs πάνω στο ίδιο μοντέλο.

5.7.2 Παράμετροι εκπαίδευσης

Στο αρχικό BERT training χρησιμοποιούνται batch size 32, learning rate 2×10^{-5} , AdamW optimizer [34] με epsilon 10^{-8} και τέσσερις epochs. Ο linear scheduler έχει μηδενικά warm-up steps και συνολικό αριθμό steps ίσο με τον αριθμό των batches επί τις epochs. Ο random seed ορίζεται σε 42 για Python, NumPy και PyTorch, και εφαρμόζεται gradient clipping με μέγιστη νόρμα 1.0.

Οι παράμετροι δεν είναι απολύτως ίδιες σε όλες τις προσομοιώσεις. Στο πρώτο post-balanced BERT χρησιμοποιούνται batch size 32 και δύο epochs, ενώ στο δεύτερο ROS scenario το baseline split χρησιμοποιεί batch size 32 και τέσσερις epochs, αλλά το ROS training DataLoader δημιουργείται με batch size 64 και επίσης τέσσερις epochs. Η διαφοροποίηση αυτή καταγράφεται ρητά, επειδή σημαίνει ότι η δεύτερη σύγκριση μεταβάλλει όχι μόνο την class distribution αλλά και το batch size. Συνεπώς η επίδραση του ROS πρέπει να ερμηνεύεται με αυτή την επιφύλαξη.

Πίνακας 5.2: Βασικές παράμετροι των BERT πειραμάτων

Ρύθμιση	Full baseline	Full ROS	Split baseline	Split ROS
Model	bert-base-uncased	bert-base-uncased	bert-base-uncased	bert-base-uncased
Num labels	5	5	5	5
Max length	128	128	128	128
Batch size	32	32	32	64
Epochs	4	2	4	4
Learning rate	$2e-5$	$2e-5$	$2e-5$	$2e-5$
Adam epsilon	$1e-8$	$1e-8$	$1e-8$	$1e-8$
Random seed	42	42	42	42

Ο πίνακας αναδεικνύει δύο σημαντικές διαφορές: στο πρώτο exploratory post-balanced run χρησιμοποιούνται δύο epochs, ενώ στο δεύτερο ROS scenario το batch size αυξάνεται από 32 σε 64. Οι διαφορές αυτές λαμβάνονται υπόψη κατά την ερμηνεία των αποτελεσμάτων.

5.7.3 Διαδικασία εκπαίδευσης

Σε κάθε epoch το μοντέλο τίθεται σε training mode και κάθε batch περιέχει input IDs, attention masks και labels. Πριν από κάθε forward pass μηδενίζονται τα gradients. Το loss που επιστρέφει το μοντέλο υποβάλλεται σε backward propagation, ακολουθεί gradient clipping, `optimizer.step` και `scheduler.step`. Η μέση training loss καταγράφεται ανά epoch. Κατά την αξιολόγηση χρησιμοποιείται `model.eval` και `torch.no\_grad` και τα logits μεταφέρονται στη CPU για υπολογισμό των predicted labels.

Στο αρχικό section του notebook η αξιολόγηση του πρώτου BERT πραγματοποιείται πάνω στο training DataLoader. Για αυτό οι αντίστοιχες μετρικές χρησιμοποιούνται μόνο ως ένδειξη training fit και όχι ως ανεξάρτητη εκτίμηση γενίκευσης. Η αξιολόγηση σε πραγματικό held-out σύνολο πραγματοποιείται αργότερα στο δεύτερο BERT σενάριο, όπου δημιουργείται ρητά validation set 12.000 δειγμάτων.

5.8 Πείραμα B1: BERT με Random Oversampling στο πλήρες training dataset

Το πρώτο σενάριο Random Oversampling εξετάζει την εκπαίδευση του BERT μετά την εξισορρόπηση του πλήρους διαθέσιμου training dataset. Η παρουσίαση ακολουθεί τη σειρά της πειραματικής ροής: αρχική κατανομή κλάσεων, εφαρμογή ROS, νέα εξισορροπημένη κατανομή, εκπαίδευση του BERT και ανάλυση της κατανομής των προβλέψεων στο test set.

5.8.1 Αρχική κατανομή του πλήρους training dataset

Η πρώτη εφαρμογή ROS ξεκινά από το πλήρες training dataset των 60.000 επισημασμένων reviews. Η αρχική κατανομή είναι Rating 1: 18.663, Rating 2: 1.629, Rating 3: 1.679, Rating 4: 3.350 και Rating 5: 34.679. Το έντονο χάσμα μεταξύ των ενδιάμεσων ratings και της Rating 5 είναι ο βασικός λόγος εφαρμογής υπερδειγματοληψίας.

5.8.2 Εφαρμογή Random Oversampling στο πλήρες training dataset

Οι τιμές της στήλης 'Review' διαμορφώνονται σε μονοδιάστατο feature array και οι τιμές 'Rating' αποτελούν το target. Ο 'RandomOverSampler(random_state=42)' εφαρμόζεται πάνω σε αυτά τα ζεύγη και αντιγράφει τυχαία records από κάθε minority class μέχρι να φτάσει το πλήθος της μεγαλύτερης κλάσης. Επειδή η παρέμβαση πραγματοποιείται σε raw records, κάθε επαναλαμβανόμενο review υποβάλλεται ξανά σε tokenization με τον ίδιο ακριβώς τρόπο όπως το αρχικό του αντίγραφο.

Η μέθοδος δεν δημιουργεί νέο συνθετικό κείμενο. Το πλεονέκτημά της είναι ότι διατηρεί έγκυρες προτάσεις και δεν επιχειρεί αριθμητική παρεμβολή σε token IDs. Το μειονέκτημα είναι ότι αυξάνει την επανάληψη των ίδιων μειονοτικών δειγμάτων, με πιθανό κίνδυνο overfitting.

5.8.3 Νέα εξισορροπημένη κατανομή κλάσεων

Μετά το ROS, κάθε μία από τις πέντε κλάσεις περιλαμβάνει 34.679 εγγραφές. Το συνολικό μέγεθος του 'train_df_balanced' είναι επομένως 173.395 reviews. Το tokenization με max length 128 παράγει tensor διαστάσεων 173.395×128 και ο DataLoader με batch size 32 περιλαμβάνει 5.419 training batches.

5.8.4 Εκπαίδευση BERT

Για το post-balanced dataset δημιουργείται νέο 'BertForSequenceClassification' με πέντε labels. Χρησιμοποιούνται AdamW με learning rate 2×10^{-5} και epsilon 10^{-8} , linear scheduler, random seed 42 και gradient clipping 1.0. Στο συγκεκριμένο exploratory run το μοντέλο εκπαιδεύεται για δύο epochs αντί για τις τέσσερις του αρχικού baseline. Για αυτό η σύγκριση των training metrics πριν και μετά το ROS δεν αποτελεί απολύτως ελεγχόμενο πείραμα μιας μόνο μεταβλητής.

Μετά την εκπαίδευση, το balanced model αξιολογείται επίσης πάνω στο balanced training data και εφαρμόζεται στο test dataset για την παραγωγή νέων ratings. Οι training metrics παρουσιάζονται στο επόμενο κεφάλαιο μόνο ως training performance. Η ουσιαστική χρησιμότητα του σεναρίου είναι να επιβεβαιώσει ότι το ROS μπορεί να δημιουργήσει ισόποση training distribution και να εξετάσει πώς μεταβάλλεται η κατανομή των προβλέψεων στο μη επισημασμένο test set.

5.8.5 Ανάλυση της κατανομής προβλέψεων στο test set

Το `test.csv` υποβάλλεται σε tokenization με τον ίδιο BERT tokenizer και το μοντέλο προβλέπει μία από τις πέντε κλάσεις για κάθε εγγραφή. Επειδή το αρχείο δεν περιλαμβάνει πραγματικές ετικέτες, δεν υπολογίζεται test classification report. Αντί αυτού, οι προβλέψεις του pre-balanced και post-balanced μοντέλου συγκρίνονται ως προς την κατανομή των ratings. Η προσέγγιση μπορεί να δείξει μετατόπιση της συμπεριφοράς του classifier, αλλά δεν μπορεί από μόνη της να αποδείξει βελτίωση ακρίβειας.

Η συγκεκριμένη πρώτη εφαρμογή του ROS είχε ως κύριο πρακτικό στόχο την παραγωγή και τη σύγκριση προβλέψεων πάνω στο test set. Επειδή όμως το test αρχείο δεν περιείχε πραγματικά ratings, η ανάλυση περιορίστηκε στην κατανομή των predictions και δεν ήταν δυνατό να εξαχθούν μετρικές επίδοσης. Ο περιορισμός αυτός αποτέλεσε τον βασικό λόγο δημιουργίας του δεύτερου ROS πειράματος.

5.9 Πείραμα B2: BERT με Random Oversampling και Train/Validation Split

Το δεύτερο σενάριο Random Oversampling σχεδιάστηκε ώστε η αξιολόγηση να πραγματοποιείται σε ανεξάρτητο και ανέγγιχτο validation set με γνωστές ετικέτες. Η ενότητα περιγράφει τον διαχωρισμό των δεδομένων, το baseline training, την εφαρμογή ROS αποκλειστικά στο training split και την τελική σύγκριση των μοντέλων με μετρικές ανά κλάση και confusion matrices.

5.9.1 Διαχωρισμός training και validation set

Το δεύτερο πείραμα δημιουργήθηκε ώστε να αντιμετωπιστεί η έλλειψη labels στο αρχικό test set. Αντί να χρησιμοποιηθεί το μη επισημασμένο test.csv για αξιολόγηση, το διαθέσιμο train.csv, το οποίο περιείχε τόσο τις κριτικές όσο και τις πραγματικές βαθμολογίες, χωρίστηκε σε δύο τμήματα. Με τον τρόπο αυτό δημιουργήθηκε ένα training subset για την εκπαίδευση και ένα ξεχωριστό validation subset πάνω στο οποίο μπορούσαν να υπολογιστούν οι απαιτούμενες μετρικές.

Στο δεύτερο BERT σενάριο το αρχικό `train.csv` φορτώνεται ξανά ώστε να αποφευχθούν μεταβολές από προηγούμενες προσομοιώσεις. Στη συνέχεια εφαρμόζεται `train\_test\_split` με `test\_size=0.2`, `random\_state=42` και `stratify=Rating`. Προκύπτουν 48.000 training samples και 12.000 validation samples. Η σχετική κατανομή των ratings διατηρείται σχεδόν ακριβώς ίδια στα δύο υποσύνολα, όπως αναμένεται από stratification.

Το validation set δεν συμμετέχει ούτε στο Random Oversampling ούτε στην εκπαίδευση του ROS model. Αυτό είναι βασικό στοιχείο του πειραματικού σχεδιασμού, επειδή επιτρέπει να αξιολογούνται τόσο το baseline όσο και το oversampled μοντέλο πάνω στα ίδια πραγματικά και μη επαναλαμβανόμενα δεδομένα.

5.9.2 Baseline BERT χωρίς ROS στο training split

Το training split υποβάλλεται σε tokenization με `MAX\_LEN=128`, ενώ οι labels μετατρέπονται σε τιμές 0-4. Δημιουργείται training DataLoader με RandomSampler και batch size 32 και validation DataLoader με SequentialSampler και batch size 32. Ο baseline classifier είναι νέο `bert-base-uncased` με πέντε outputs. Εκπαιδεύεται για τέσσερις epochs με AdamW, learning rate 2×10^{-5} , epsilon 10^{-8} , linear scheduler και seed 42.

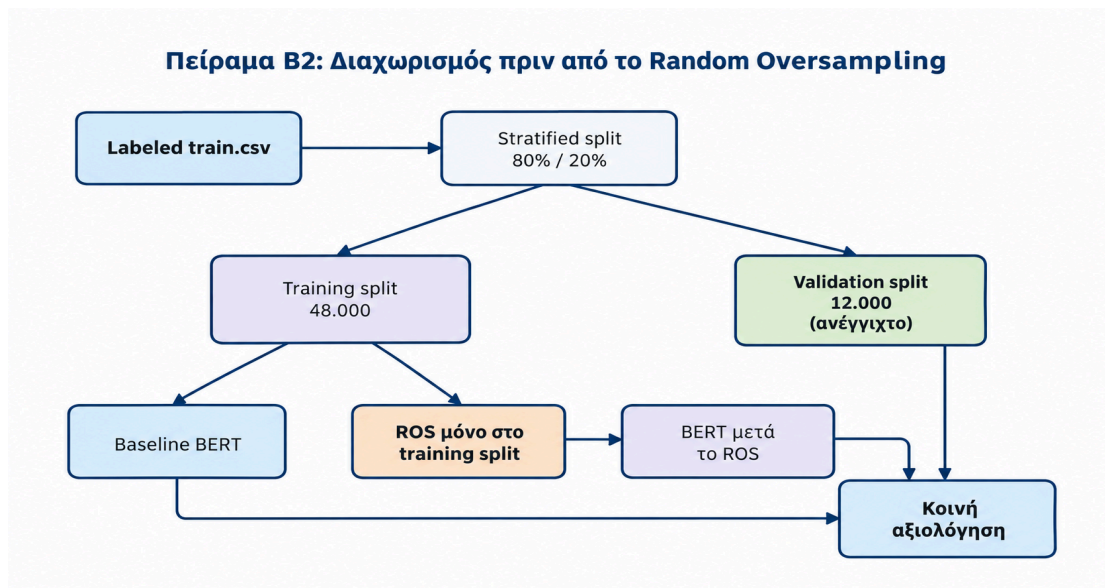
Η αξιολόγηση πραγματοποιείται στο validation DataLoader με `model.eval` και χωρίς gradients. Καταγράφονται validation loss, accuracy, classification report και confusion matrix. Αυτή η ροή αποτελεί το βασικό σημείο αναφοράς για το δεύτερο ROS πείραμα.

5.9.3 Εφαρμογή ROS αποκλειστικά στο training split

Ο RandomOverSampler εφαρμόζεται μόνο στις 48.000 training εγγραφές. Πριν το ROS, οι απόλυτοι αριθμοί είναι 14.931 για Rating 1, 1.303 για Rating 2, 1.343 για Rating 3, 2.680 για Rating 4 και 27.743 για Rating 5. Μετά την εφαρμογή, κάθε κλάση έχει 27.743 records. Το νέο training dataset περιλαμβάνει συνολικά 138.715 εγγραφές.

Το validation set των 12.000 samples παραμένει ακριβώς το ίδιο με εκείνο που χρησιμοποιήθηκε στο baseline. Με αυτό τον τρόπο αποφεύγεται oversampling leakage. Το ROS μεταβάλλει μόνο την training distribution και όχι την πραγματική κατανομή πάνω στην οποία υπολογίζονται τα τελικά validation metrics.

Η πειραματική ροή του δεύτερου και αυστηρότερου σεναρίου Random Oversampling παρουσιάζεται στο Σχήμα 5.1.



Σχήμα 5.1: Πειραματική ροή BERT με Random Oversampling μόνο στο training split.

Πηγή: Ιδία επεξεργασία.

Πίνακας 5.3: Κατανομές κλάσεων πριν και μετά το Random Oversampling

Rating	Full dataset πριν ROS	Full dataset μετά ROS	Training split πριν ROS	Training split μετά ROS
1	18.663	34.679	14.931	27.743
2	1.629	34.679	1.303	27.743
3	1.679	34.679	1.343	27.743
4	3.350	34.679	2.680	27.743
5	34.679	34.679	27.743	27.743
Σύνολο	60.000	173.395	48.000	138.715

Και στα δύο σενάρια ο ROS εξισώνει τις κλάσεις στο μέγεθος της αντίστοιχης majority class. Στο δεύτερο σενάριο η διαδικασία σταματά στο training split και δεν αγγίζει τις 12.000 validation εγγραφές.

5.9.4 Εκπαίδευση BERT στο εξισορροπημένο training set

Για το oversampled training set δημιουργείται νέος BERT classifier. Ο βοηθητικός `create_dataloader` tokenizes τα reviews με max length 128, όμως στο συγκεκριμένο ROS branch χρησιμοποιείται batch size 64. Το σύνολο των 138.715 εγγραφών δημιουργεί 2.168 batches ανά epoch. Η εκπαίδευση

πραγματοποιείται για τέσσερις epochs, με την ίδια learning rate 2×10^{-5} , epsilon 10^{-8} , linear scheduler, gradient clipping 1.0 και seed 42.

Η επιλογή batch size 64 διαφέρει από το batch size 32 του baseline split. Η διαφορά αυτή δεν αφαιρεί την αξία του πειράματος ως εφαρμογή ROS, αλλά αποτελεί συγχυτικό παράγοντα όταν επιχειρείται να αποδοθεί κάθε αλλαγή αποκλειστικά στην υπερδειγματοληψία. Η παράμετρος καταγράφεται επομένως ρητά και συζητείται στους περιορισμούς.

5.9.5 Αξιολόγηση στο ανέγγιχτο validation set

Μετά την ολοκλήρωση των τεσσάρων epochs, το ROS model αξιολογείται με την ίδια συνάρτηση `evaluate_model` στο `val_dataloader_split`. Για κάθε batch συλλέγονται logits και true labels, υπολογίζεται η μέση validation loss και τα predicted labels προκύπτουν με `argmax`. Στη συνέχεια παράγονται συνολική accuracy, classification report ανά rating και confusion matrix.

Το γεγονός ότι ακριβώς το ίδιο validation set χρησιμοποιείται πριν και μετά το ROS επιτρέπει άμεση σύγκριση των predictions. Σε αντίθεση με το πρώτο ROS σενάριο, εδώ οι μετρικές αντιπροσωπεύουν held-out validation performance και όχι fitting πάνω στο training data.

5.9.6 Σύγκριση Precision, Recall, F1-score και Confusion Matrices

Για τα δύο μοντέλα του δεύτερου σεναρίου δημιουργούνται side-by-side classification reports και confusion matrices. Η σύγκριση πραγματοποιείται ανά Rating 1 έως Rating 5, ενώ εξετάζονται επίσης macro και weighted averages. Η συγκεκριμένη μορφή είναι καταλληλότερη από μία απλή σύγκριση accuracy, επειδή οι κλάσεις 2, 3 και 4 έχουν πολύ μικρότερο support από τις κλάσεις 1 και 5.

Τα αριθμητικά αποτελέσματα δεν παρουσιάζονται στο παρόν κεφάλαιο αλλά στο Κεφάλαιο 6. Μεθοδολογικά, το κύριο κριτήριο είναι αν η υπερδειγματοληψία αυξάνει recall και F1 των minority ratings χωρίς υπερβολική επιδείνωση των υπόλοιπων κλάσεων. Η confusion matrix χρησιμοποιείται για να εντοπιστεί προς ποιες κλάσεις μεταφέρονται τα σφάλματα.

5.10 Πείραμα Γ: RoBERTa για sentiment classification

Το πείραμα αυτό εξετάζει τη χρήση ενός pretrained RoBERTa sentiment model για την παραγωγή προβλέψεων πάνω στις διαθέσιμες κριτικές. Στη συνέχεια παρουσιάζονται η επιλογή του μοντέλου, η διαδικασία παραγωγής προβλέψεων, τα confidence scores και ο τρόπος με τον οποίο αντιμετωπίζονται οι κλάσεις και τα pseudo-labels.

5.10.1 Επιλογή μοντέλου

Το βοηθητικό RoBERTa sentiment experiment χρησιμοποιεί το `cardiffnlp/twitter-roberta-base-sentiment`, το οποίο παρέχει τρεις αρχικές κατηγορίες negative, neutral και positive [15], [52]. Το checkpoint φορτώνεται μέσω `AutoTokenizer` και `AutoModelForSequenceClassification` και μεταφέρεται σε GPU όταν υπάρχει διαθέσιμη CUDA συσκευή.

Το πείραμα αυτό δεν αποτελεί το κύριο oversampling experiment της εργασίας. Χρησιμοποιείται για να εξεταστεί ένα pseudo-labeling/fine-tuning pipeline και για να παραχθούν συγκριτικές sentiment distributions. Η βασική τεχνική υπερδειγματοληψίας με RoBERTa εξετάζεται στο επόμενο πείραμα μέσω SMOTE πάνω στα embeddings.

5.10.2 Παραγωγή προβλέψεων

Οι 8.688 εγγραφές του `english\_reviews.csv` μετατρέπονται σε λίστα κειμένων και επεξεργάζονται σε batches των 32. Ο tokenizer εφαρμόζει truncation, padding και `max\_length=128`. Τα logits του pretrained μοντέλου μετατρέπονται σε πιθανότητες με softmax και η κλάση με τη μεγαλύτερη πιθανότητα θεωρείται η αρχική πρόβλεψη. Το ίδιο model χρησιμοποιείται επίσης αργότερα για baseline κατανομή δύο κλάσεων πριν από το fine-tuning.

5.10.3 Confidence scores

Για την παραγωγή pseudo-labels δεν χρησιμοποιούνται όλες οι προβλέψεις. Ως confidence ορίζεται η μέγιστη πιθανότητα μεταξύ των τριών softmax outputs και ένα review διατηρείται μόνο όταν η confidence είναι μεγαλύτερη από 0,90. Η επιλογή αυτή αποσκοπεί στη διατήρηση μόνο των προβλέψεων για τις οποίες το pretrained checkpoint εμφανίζεται ιδιαίτερα βέβαιο.

Μετά την εφαρμογή του confidence threshold και τη δυαδική χαρτογράφηση που περιγράφεται στην επόμενη ενότητα, προκύπτουν 7.936 εγγραφές στην ετικέτα 1 και μόλις 6 στην ετικέτα 0. Η ακραία αυτή κατανομή δείχνει ότι ακόμη και υψηλό confidence threshold δεν εξαλείφει την ανισορροπία του συγκεκριμένου dataset.

5.10.4 Αντιμετώπιση των κλάσεων και pseudo-labeling

Στο pseudo-labeling code η αρχική RoBERTa κλάση 0 διατηρείται ως binary label 0, ενώ οι αρχικές κλάσεις 1 και 2 συγχωνεύονται στο binary label 1. Με άλλα λόγια, οι κατηγορίες neutral και positive αντιμετωπίζονται ως μία κοινή μη αρνητική κλάση. Από τα 7.942 high-confidence pseudo-labeled samples προκύπτουν 7.936 label-1 και 6 label-0. Στο τελικό στάδιο πρόβλεψης ο κώδικας ονομάζει το label 1 `positive` και το label 0 `negative`, παρότι το label 1 έχει εκπαιδευτεί πάνω στη συγχωνευμένη κατηγορία neutral/positive.

Για τη δημιουργία balanced subset δεν εφαρμόζεται oversampling. Αντίθετα, η μεγάλη κλάση υποδειγματοληπτείται τυχαία μέχρι το μέγεθος της μικρής, με αποτέλεσμα $6+6=12$ δείγματα. Το subset χωρίζεται με stratified 80/20 split. Στη συνέχεια φορτώνεται το ίδιο RoBERTa checkpoint με `num\_labels=2` και `ignore\_mismatched\_sizes=True`, ώστε το τριταξικό classification head να αντικατασταθεί από δυαδικό head. Το fine-tuning πραγματοποιείται για τρεις epochs με train/eval batch size 16.

Το πολύ μικρό πλήθος των 12 pseudo-labeled samples σημαίνει ότι η αντίστοιχη test αξιολόγηση έχει εξαιρετικά μικρό support. Για αυτό το πείραμα αντιμετωπίζεται ως βοηθητική απόδειξη λειτουργίας του pipeline και όχι ως κύριο benchmark της διπλωματικής. Επιπλέον, επειδή τα labels προέρχονται από το ίδιο αρχικό RoBERTa, οι μετρικές δεν ισοδυναμούν με accuracy έναντι ανεξάρτητων ανθρώπινων labels.

5.11 Πείραμα Α: RoBERTa Embeddings με SMOTE

Το δεύτερο κύριο πειραματικό pipeline χρησιμοποιεί το RoBERTa ως feature extractor και εφαρμόζει SMOTE στον χώρο των embeddings πριν από την εκπαίδευση Logistic Regression. Οι υποενοότητες περιγράφουν διαδοχικά την παραγωγή των embeddings, τον διαχωρισμό training και validation δεδομένων, την αρχική ανισορροπία, την εφαρμογή SMOTE και τη σύγκριση του τελικού classification pipeline.

5.11.1 Παραγωγή embeddings

Στο κύριο RoBERTa oversampling experiment, μετά την αρχική sentiment labeling διαδικασία φορτώνεται το base model μέσω `'AutoModel.from_pretrained(MODEL)'`, δηλαδή χωρίς το classification head. Το μοντέλο τίθεται σε evaluation mode. Για κάθε tokenized batch λαμβάνεται το `'last_hidden_state'` και χρησιμοποιείται το διάνυσμα της πρώτης θέσης (`'outputs.last_hidden_state[:, 0,:]'`) ως συνολική αναπαράσταση της κριτικής. Δεν εφαρμόζεται mean pooling στα token representations. Το αποτέλεσμα έχει 768 χαρακτηριστικά ανά sample, σύμφωνα με το hidden size της base αρχιτεκτονικής [52].

Η εξαγωγή γίνεται με `'torch.no_grad'`, οπότε το RoBERTa δεν είναι fine-tuned στο συγκεκριμένο στάδιο. Δημιουργούνται πίνακες `'train_embeddings'` διαστάσεων 13.209×768 και `'val_embeddings'` διαστάσεων 3.303×768 . Οι αντίστοιχες αριθμητικές labels αποθηκεύονται ξεχωριστά.

5.11.2 Διαχωρισμός training και validation/test δεδομένων

Πριν την παραγωγή embeddings, οι 16.512 labeled αγγλόφωνες κριτικές χωρίζονται σε 80% training και 20% validation με `'train_test_split'`, stratification πάνω στο `'label_id'` και `'random_state=42'`. Προκύπτουν 13.209 training και 3.303 validation samples. Η σχετική αναλογία positive/negative/neutral διατηρείται ουσιαστικά ίδια και στα δύο subsets.

Η διάταξη split-before-SMOTE είναι κρίσιμη. Το validation set δεν συμμετέχει στον υπολογισμό των συνθετικών training vectors και παραμένει αντιπροσωπευτικό της αρχικής ακραία άνισης κατανομής. Έτσι και οι δύο Logistic Regression classifiers αξιολογούνται πάνω ακριβώς στο ίδιο unseen feature set.

5.11.3 Αρχική ανισορροπία των embeddings

Στο training embedding space η positive κλάση περιλαμβάνει 13.037 vectors, η negative 90 και η neutral 82. Το SMOTE καλείται επομένως να εξισορροπήσει δύο εξαιρετικά μικρές κλάσεις σε σχέση με την positive. Η κατάσταση αυτή αποτελεί απαιτητικό σενάριο για synthetic oversampling, επειδή οι synthetic points παράγονται από πολύ περιορισμένο αριθμό πραγματικών minority observations.

5.11.4 Εφαρμογή SMOTE στο training set

Χρησιμοποιείται `'SMOTE(random_state=42)'` της `'imbalanced-learn'` και η μέθοδος `'fit_resample(train_embeddings, train_labels)'`. Η εφαρμογή πραγματοποιείται αποκλειστικά στα 13.209 training embeddings. Το validation set δεν μεταβάλλεται. Η μέθοδος δημιουργεί νέα vectors με παρεμβολή μεταξύ κοντινών minority observations στον 768-dimensional representation space, σύμφωνα με τη λογική που παρουσιάστηκε στο Κεφάλαιο 3 [3], [16], [43]-[44].

5.11.5 Παραγωγή balanced embedding dataset

Μετά το SMOTE και οι τρεις training classes περιλαμβάνουν 13.037 samples. Το νέο dataset αποτελείται από 39.111 embeddings διαστάσεων 768. Η positive κλάση διατηρεί τα πραγματικά της δείγματα, ενώ οι negative και neutral εμπλουτίζονται με synthetic vectors μέχρι να φτάσουν το πλήθος της majority class. Η μεταβολή αυτή αφορά μόνο τον feature space και δεν παράγει νέα αναγνώσιμα reviews.

Πίνακας 5.4: Ρυθμίσεις του RoBERTa embeddings + SMOTE πειράματος

Στοιχείο	Ρύθμιση / τιμή
----------	----------------

Pretrained checkpoint	cardiffnlp/twitter-roberta-base-sentiment
Αγγλικά reviews	16.512
Train / validation	13.209 / 3.303, stratified, random state=42
Sentiment labels	negative=0, neutral=1, positive=2
Max token length	512
Embedding	πρώτη θέση του last hidden state, 768 διαστάσεις
Training counts πριν SMOTE	negative 90, neutral 82, positive 13.037
Training counts μετά SMOTE	13.037 ανά κλάση, σύνολο 39.111
Downstream classifier	LogisticRegression(max_iter=1000, random_state=42)

Η χρήση του ίδιου validation embedding set και του ίδιου classifier πριν και μετά το SMOTE αποτελεί το καθαρότερο συγκριτικό σημείο του RoBERTa pipeline.

5.11.6 Logistic Regression

Για την απομόνωση της επίδρασης του SMOTE εκπαιδεύονται δύο Logistic Regression classifiers με την ίδια βασική ρύθμιση `'max_iter=1000'` και `'random_state=42'`. Ο πρώτος (`'clf_imbalanced'`) εκπαιδεύεται στα αρχικά 13.209 training embeddings. Ο δεύτερος (`'clf_smote'`) εκπαιδεύεται στα 39.111 SMOTEd embeddings. Και οι δύο κάνουν predictions στα ίδια 3.303 validation embeddings.

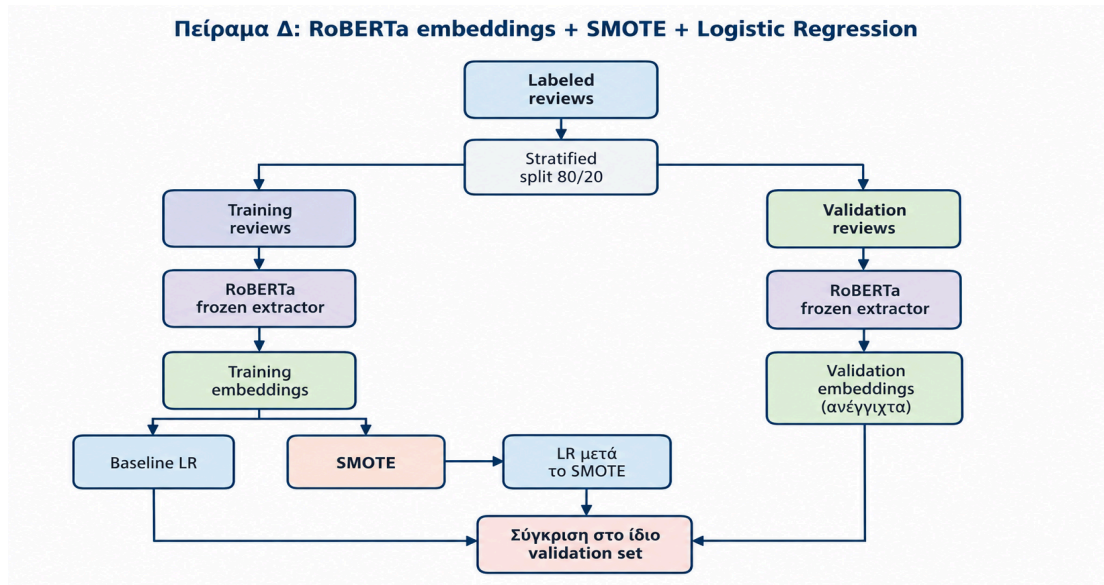
Η επιλογή ίδιου downstream classifier είναι σημαντική, επειδή η βασική διαφοροποίηση μεταξύ των δύο branches είναι η training distribution. Σε αντίθεση με το BERT fine-tuning, εδώ δεν υπάρχουν διαφορετικές Transformer epochs ή batch sizes μετά την εξαγωγή features. Το pretrained RoBERTa παραμένει frozen και η σύγκριση επικεντρώνεται στην επίδραση της υπερδειγματοληψίας στον classifier.

5.11.7 Classification pipeline

Η πλήρης ροή του πειράματος μπορεί να συνοψιστεί ως εξής: raw review → language filtering → pretrained RoBERTa sentiment label → stratified train/validation split → RoBERTa tokenization → frozen 768-dimensional embedding → SMOTE μόνο στο training branch → Logistic Regression → αξιολόγηση στο αρχικό validation set. Παράλληλα υπάρχει baseline branch όπου παραλείπεται μόνο το SMOTE και η Logistic Regression εκπαιδεύεται στα αρχικά embeddings.

Η δομή αυτή επιτρέπει σύγκριση πριν και μετά την υπερδειγματοληψία, αλλά πρέπει να ληφθεί υπόψη ότι τα sentiment labels που λειτουργούν ως target έχουν παραχθεί από το ίδιο pretrained RoBERTa checkpoint και δεν αποτελούν human ground truth. Συνεπώς οι μετρικές του πειράματος μετρούν σε μεγάλο βαθμό πόσο καλά ο downstream classifier αναπαράγει τις pseudo-label αποφάσεις του RoBERTa στον embedding space.

Η πλήρης ροή του πειράματος RoBERTa embeddings με SMOTE και Logistic Regression συνοψίζεται στο Σχήμα 5.2.



Σχήμα 5.2: Πειραματική ροή RoBERTa embeddings, SMOTE και Logistic Regression.

Πηγή: Ιδία επεξεργασία.

5.12 Μέτρα αξιολόγησης

Η αξιολόγηση σε μη ισορροπημένα δεδομένα απαιτεί περισσότερες από μία μετρικές, επειδή η συνολική accuracy δεν αποτυπώνει πάντοτε την απόδοση στις μειονοτικές κατηγορίες. Για τον λόγο αυτό χρησιμοποιούνται και παρουσιάζονται οι μετρικές accuracy, precision, recall, F1-score, macro F1, weighted F1 και confusion matrix.

5.12.1 Accuracy

Το accuracy υπολογίζεται ως ο λόγος των σωστών προβλέψεων προς το συνολικό πλήθος των δειγμάτων. Χρησιμοποιείται ως συνοπτική μέτρηση, αλλά στα συγκεκριμένα datasets δεν επαρκεί λόγω του έντονου class imbalance. Ένα μοντέλο μπορεί να εμφανίζει υψηλό accuracy προβλέποντας σωστά κυρίως τη majority class. Για αυτό το accuracy παρουσιάζεται πάντοτε μαζί με τις ανά κλάση μετρικές.

5.12.2 Precision

Το precision μιας κλάσης εκφράζει ποιο ποσοστό των samples που προβλέφθηκαν ως η συγκεκριμένη κλάση ανήκουν πραγματικά σε αυτή. Στην εργασία είναι ιδιαίτερα χρήσιμη για να διαπιστωθεί αν η υπερδειγματοληψία αυξάνει τα false positives των minority classes. Για παράδειγμα, η αύξηση recall μιας σπάνιας κλάσης δεν είναι απαραίτητα θετική αν συνοδεύεται από πολύ μεγάλη πτώση precision.

5.12.3 Recall

Το recall μετρά ποιο ποσοστό των πραγματικών samples μιας κλάσης εντοπίζεται σωστά από τον classifier. Για προβλήματα imbalance αποτελεί βασικό κριτήριο, επειδή η αποτυχία αναγνώρισης minority observations συχνά εμφανίζεται ως χαμηλό recall. Η σύγκριση πριν και μετά ROS/SMOTE εξετάζει επομένως αν η αυξημένη παρουσία των minority classes στο training data μεταφράζεται σε καλύτερη ανάκτησή τους στο ανέγγιχτο validation set.

5.12.4 F1-score

Το F1-score είναι ο αρμονικός μέσος precision και recall και επιτρέπει να αξιολογηθεί η ισορροπία μεταξύ των δύο. Στα classification reports υπολογίζεται ξεχωριστά για κάθε rating ή sentiment class. Στο BERT πείραμα ιδιαίτερο ενδιαφέρον έχουν τα F1 των Ratings 2, 3 και 4, ενώ στο RoBERTa pipeline εξετάζονται negative και neutral που έχουν πολύ μικρό support.

5.12.5 Macro F1

Το macro F1 υπολογίζει τον απλό μέσο όρο των F1-scores όλων των κλάσεων, δίνοντας σε κάθε κλάση ίσο βάρος ανεξάρτητα από το support της. Για την παρούσα εργασία αποτελεί μία από τις σημαντικότερες συνολικές μετρικές, επειδή δεν επιτρέπει στην πολύ μεγάλη majority class να κυριαρχήσει στον τελικό αριθμό. Μία πτώση του macro F1 μετά το oversampling μπορεί να δείξει ότι η εξισορρόπηση δεν βελτίωσε τελικά τη συνολική ισορροπία της ταξινόμησης.

5.12.6 Weighted F1

Το weighted F1 σταθμίζει το F1 κάθε κλάσης με βάση το πραγματικό της support. Είναι χρήσιμο ως συμπληρωματική μέτρηση της συνολικής απόδοσης πάνω στην πραγματική κατανομή, αλλά σε ακραία imbalanced δεδομένα τείνει να επηρεάζεται περισσότερο από τις majority classes. Για αυτό συγκρίνεται με το macro F1 και δεν χρησιμοποιείται μόνο του.

5.12.7 Confusion Matrix

Οι πίνακες confusion matrices καταγράφουν αναλυτικά τις πραγματικές και προβλεπόμενες κλάσεις. Στα BERT πειράματα είναι πίνακες 5×5 και επιτρέπουν να διαπιστωθεί, για παράδειγμα, αν πραγματικές Rating 4 μετακινούνται συστηματικά προς Rating 5 ή αν οι Ratings 2 και 3 συγχέονται μεταξύ τους. Στο RoBERTa experiment οι πίνακες είναι 3×3 για negative, neutral και positive.

Η ανάλυση confusion matrix είναι ιδιαίτερα σημαντική για την ερμηνεία της υπερδειγματοληψίας, επειδή δύο μοντέλα με παρόμοια accuracy μπορούν να κάνουν πολύ διαφορετικούς τύπους λαθών. Στο Κεφάλαιο 6 οι matrices εξετάζονται μαζί με τα classification reports, ώστε οι αριθμητικές μεταβολές να συνδέονται με συγκεκριμένες μετακινήσεις μεταξύ κλάσεων.

5.13 Επίλογος

Στο κεφάλαιο παρουσιάστηκε η πλήρης μεθοδολογική ροή των πειραμάτων. Για το BERT χρησιμοποιήθηκε ταξινόμηση ratings, πέντε τάξεων, με 'bert-base-uncased' και δύο διαφορετικές εφαρμογές Random Oversampling. Η πρώτη εξισορρόπησε ολόκληρο το training dataset σε 173.395 samples και χρησιμοποιήθηκε κυρίως για διερευνητική σύγκριση training και test prediction distributions. Η δεύτερη ξεκίνησε από stratified split 48.000/12.000, εφάρμοσε ROS μόνο στο training subset και αξιολόγησε baseline και oversampled μοντέλο στο ίδιο validation set.

Για το RoBERTa παρουσιάστηκαν δύο διαφορετικοί ρόλοι. Το fine-tuning/pseudo-labeling experiment χρησιμοποιεί high-confidence labels και μικρό balanced subset ως βοηθητική προσομοίωση. Το κύριο SMOTE experiment χρησιμοποιεί το pretrained RoBERTa ως frozen extractor 768-dimensional embeddings, εφαρμόζει SMOTE αποκλειστικά στο training feature space και συγκρίνει δύο Logistic Regression classifiers στο ίδιο validation set. Η μεθοδολογία αυτή επιτρέπει να εξεταστούν δύο διαφορετικές στρατηγικές υπερδειγματοληψίας: επανάληψη πραγματικών κειμένων πριν το Transformer fine-tuning και δημιουργία synthetic representations μετά το Transformer.

Το επόμενο κεφάλαιο παρουσιάζει τα πειραματικά αποτελέσματα. Η ερμηνεία θα λαμβάνει υπόψη όχι μόνο accuracy αλλά κυρίως macro F1, ανά κλάση precision/recall/F1 και confusion matrices. Παράλληλα θα διαχωρίζονται τα training-only αποτελέσματα και τα pseudo-label metrics από τις ισχυρότερες held-out validation συγκρίσεις, ώστε τα συμπεράσματα να αντανακλούν με ακρίβεια το τι υποστηρίζουν οι υλοποιημένες προσομοιώσεις.

Κεφάλαιο 6ο: Πειραματικά Αποτελέσματα και Αξιολόγηση

6.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζονται και ερμηνεύονται τα αποτελέσματα των πειραμάτων που περιγράφηκαν στο Κεφάλαιο 5. Η αξιολόγηση οργανώνεται γύρω από το βασικό ερευνητικό ερώτημα της εργασίας: κατά πόσο η μεταβολή της κατανομής των training classes μέσω Random Oversampling ή SMOTE οδηγεί σε ουσιαστικότερη αναγνώριση των λιγότερο συχνών κατηγοριών. Για τον λόγο αυτό δεν αρκεί η παράθεση μιας συνολικής accuracy. Εξετάζονται τα classification reports, το macro F1, οι ανά κλάση τιμές precision, recall και F1-score, καθώς και οι confusion matrices.

Τα αποτελέσματα δεν έχουν όλα την ίδια αποδεικτική ισχύ. Στο πρώτο BERT σενάριο υπάρχουν metrics πάνω στα ίδια training data στα οποία εκπαιδεύτηκαν τα μοντέλα, ενώ το test dataset δεν διαθέτει πραγματικές labels. Οι τιμές αυτές είναι χρήσιμες για να φανεί πόσο καλά προσαρμόζεται το μοντέλο στα δεδομένα εκπαίδευσης, αλλά δεν πρέπει να ερμηνεύονται ως test performance. Αντίθετα, το δεύτερο BERT σενάριο συγκρίνει δύο μοντέλα στο ίδιο ανέγγιχτο validation set και αποτελεί την καθαρότερη δοκιμή της επίδρασης του ROS.

Αντίστοιχα, στο RoBERTa+SMOTE pipeline το validation set παραμένει εκτός της διαδικασίας SMOTE. Ωστόσο, τα target labels έχουν παραχθεί από το ίδιο pretrained RoBERTa checkpoint που χρησιμοποιείται για την εξαγωγή embeddings. Επομένως τα πολύ υψηλά metrics μετρούν κυρίως την ικανότητα ενός γραμμικού downstream classifier να αναπαράγει τις αποφάσεις του RoBERTa στον representation space του και όχι ανεξάρτητη ακρίβεια έναντι ανθρώπινης επισημάνσης. Η διάκριση αυτή είναι κεντρική για τη σωστή ερμηνεία των αποτελεσμάτων.

6.2 Ανάλυση αρχικής κατανομής δεδομένων

Το BERT dataset περιλαμβάνει 60.000 επισημασμένες κριτικές με βαθμολογίες από 1 έως 5. Η κατανομή είναι έντονα άνιση: 34.679 εγγραφές ανήκουν στη Rating 5 και 18.663 στη Rating 1, ενώ οι Ratings 2, 3 και 4 περιλαμβάνουν μόλις 1.629, 1.679 και 3.350 samples αντίστοιχα. Συνεπώς περισσότερο από το 88% του training set συγκεντρώνεται στις δύο ακραίες βαθμολογίες 1 και 5. Η δομή αυτή δημιουργεί ένα δύσκολο πρόβλημα για τις ενδιάμεσες κατηγορίες, οι οποίες διαθέτουν πολύ λιγότερα παραδείγματα κατά το fine-tuning.

Η ανισορροπία δεν είναι απλώς αριθμητικό χαρακτηριστικό του dataset αλλά εμφανίζεται άμεσα και στην επίδοση του baseline. Όπως θα φανεί στις επόμενες ενότητες, οι ακραίες κλάσεις αποκτούν F1-score κοντά ή πάνω από 0,90 στο ανεξάρτητο validation set, ενώ οι τρεις ενδιάμεσες κλάσεις παραμένουν πολύ χαμηλότερα. Αυτό υποστηρίζει την επιλογή του συγκεκριμένου dataset ως κατάλληλο πεδίο μελέτης της υπερδειγματοληψίας.



Σχήμα 6.1: Αρχική κατανομή των πέντε rating classes στο BERT training dataset.

Πίνακας 6.1: Αρχική κατανομή κλάσεων στο BERT dataset

Rating	Πλήθος	Ποσοστό
1	18.663	31,105%
2	1.629	2,715%
3	1.679	2,798%
4	3.350	5,583%
5	34.679	57,798%
Σύνολο	60.000	100%

6.3 Αποτελέσματα BERT baseline

Στην αρχική εκτέλεση πάνω στο πλήρες training dataset, το BERT αξιολογήθηκε πάνω στα ίδια 60.000 samples μετά την εκπαίδευση. Η αποθηκευμένη training accuracy είναι 0,966217 και το macro F1 0,868901. Οι Rating 1 και Rating 5 έχουν F1 0,987554 και 0,985126, ενώ οι Rating 2, Rating 3 και Rating 4 έχουν 0,757789, 0,812575 και 0,801462 αντίστοιχα. Το αποτέλεσμα δείχνει ότι ακόμη και σε training evaluation οι μειονοτικές κλάσεις είναι δυσκολότερες από τις δύο κυρίαρχες.

Η confusion matrix της training αξιολόγησης δείχνει χαρακτηριστικές συγχύσεις μεταξύ γειτονικών βαθμολογιών. Για τη Rating 2, 285 από τα 1.629 samples ταξινομούνται ως Rating 1 και 195 ως Rating 3. Για τη Rating 4, 825 από τα 3.350 samples μεταφέρονται στη Rating 5. Οι μετακινήσεις αυτές είναι λογικές για ordinal ratings, επειδή οι γλωσσικές διαφορές ανάμεσα σε γειτονικές κατηγορίες μπορούν να είναι μικρές. Παρ' όλα αυτά, η αξιολόγηση στο training set δεν αποτελεί μέτρηση γενίκευσης και χρησιμοποιείται εδώ μόνο ως baseline συμπεριφοράς.

6.4 Αποτελέσματα πρώτης εφαρμογής Random Oversampling

Στο πρώτο ROS σενάριο το πλήρες training dataset εξισορροπήθηκε με αντιγραφή δειγμάτων μέχρι κάθε rating να περιλαμβάνει 34.679 εγγραφές. Το συνολικό training set αυξήθηκε από 60.000 σε 173.395 samples. Η αξιολόγηση του μοντέλου πάνω στο εξισορροπημένο training set έδωσε accuracy

0,993737, macro F1 0,993727 και weighted F1 0,993727. Τα F1 των Ratings 2, 3 και 4 αυξήθηκαν σε 0,997311, 0,997627 και 0,991617 αντίστοιχα.

Η εντυπωσιακή άνοδος των training metrics δεν μπορεί να θεωρηθεί από μόνη της απόδειξη ότι το ROS βελτίωσε τη γενίκευση. Το Random Oversampling επαναλαμβάνει πραγματικά minority samples, συνεπώς το μοντέλο βλέπει πολλές φορές τα ίδια παραδείγματα. Επιπλέον, το post-balanced μοντέλο αξιολογείται στο ίδιο resampled training set και όχι σε ανεξάρτητο labeled test set. Η πολύ υψηλή επίδοση είναι συνεπώς συμβατή τόσο με καλύτερη εκμάθηση των minority patterns όσο και με αυξημένη προσαρμογή ή απομνημόνευση των επαναλαμβανόμενων samples.

Το notebook εφαρμόζει το post-balanced μοντέλο και στο μη επισημασμένο test set των 40.000 κριτικών. Οι αποθηκευμένες οπτικοποιήσεις δείχνουν ότι, παρότι το training set έχει γίνει πλήρως ομοιόμορφο, οι test predictions παραμένουν έντονα συγκεντρωμένες στη Rating 5. Οι ενδιάμεσες βαθμολογίες εμφανίζονται κάπως περισσότερο σε σχέση με το pre-balanced μοντέλο, αλλά η κατανομή δεν γίνεται ισορροπημένη. Εφόσον το test set δεν περιλαμβάνει ground-truth ratings, δεν είναι δυνατό να υπολογιστεί accuracy ή F1 για αυτό το στάδιο.

Υπάρχει επίσης ένας επιπλέον παράγοντας που περιορίζει την άμεση σύγκριση: το αρχικό full-data BERT εκπαιδεύτηκε για τέσσερα epochs, ενώ το post-balanced full-data μοντέλο για δύο epochs. Επομένως η μεταβολή των training αποτελεσμάτων δεν οφείλεται αποκλειστικά στο ROS. Για αυτό το πρώτο σενάριο αντιμετωπίζεται ως διερευνητική εφαρμογή και όχι ως το κύριο πείραμα επιβεβαίωσης της αποτελεσματικότητας της υπερδειγματοληψίας.

Πίνακας 6.2: Training metrics πριν και μετά το ROS στο πλήρες BERT training dataset

Μετρική	Χωρίς ROS	Με ROS
Accuracy	0,9662	0,9937
Macro F1	0,8689	0,9937
Weighted F1	0,9646	0,9937
F1 Rating 2	0,7578	0,9973
F1 Rating 3	0,8126	0,9976
F1 Rating 4	0,8015	0,9916

6.5 Αποτελέσματα δεύτερης εφαρμογής Random Oversampling με Train/Validation Split

Στο δεύτερο ROS σενάριο το αρχικό dataset χωρίστηκε πρώτα με stratification σε 48.000 training και 12.000 validation samples. Το validation set διατήρησε την αρχική κατανομή και δεν συμμετείχε σε κανένα στάδιο υπερδειγματοληψίας. Στο baseline training split οι κλάσεις είχαν 14.931, 1.303, 1.343, 2.680 και 27.743 samples για Ratings 1 έως 5. Μετά το ROS, καθεμία από τις πέντε κλάσεις είχε 27.743 training samples.

Το baseline BERT, εκπαιδευμένο στο μη ισορροπημένο training split, πέτυχε validation loss 0,4266 και accuracy 0,8812. Το macro precision ήταν 0,59, το macro recall 0,56 και το macro F1 0,57. Σε επίπεδο κλάσεων, οι Ratings 1 και 5 εμφάνισαν F1 περίπου 0,92 και 0,95, ενώ οι Ratings 2, 3 και 4 περιορίστηκαν σε 0,22, 0,38 και 0,40. Η διαφορά αυτή αποτελεί σαφή ένδειξη ότι η συνολική accuracy αποκρύπτει πολύ διαφορετική ποιότητα ταξινόμησης μεταξύ majority και minority classes.

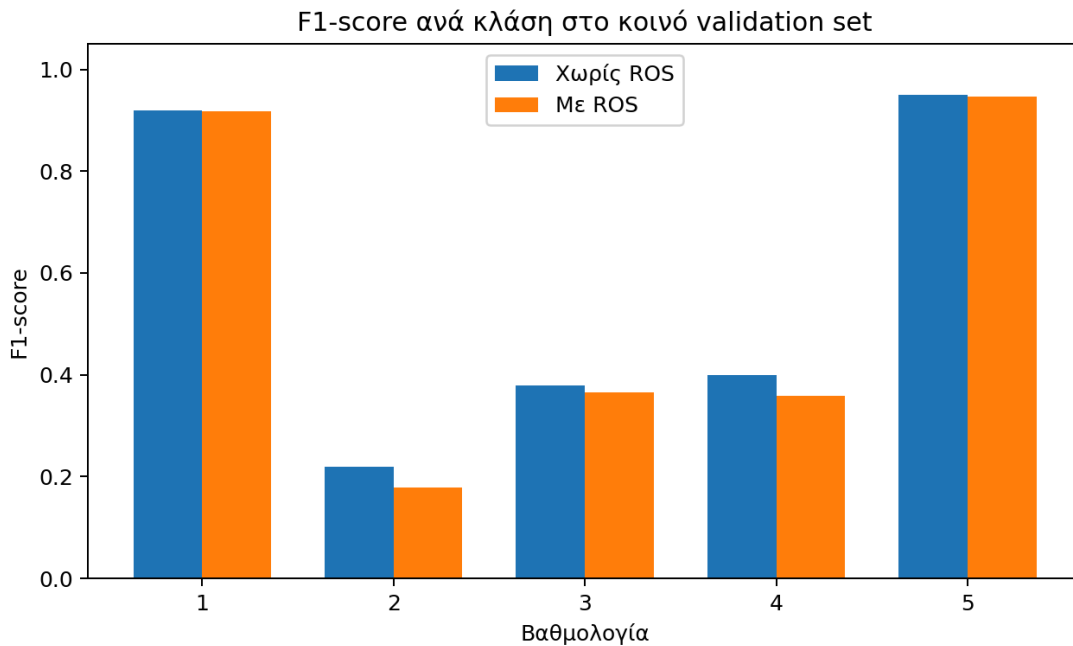
Πριν από την εκτέλεση του συγκεκριμένου πειράματος, η αρχική προσδοκία ήταν ότι το Random Oversampling θα μπορούσε να οδηγήσει σε μια μικρή βελτίωση κυρίως στις μειονοτικές κλάσεις. Ο στόχος δεν ήταν απαραίτητα να αυξηθεί το συνολικό accuracy, αλλά να δοθεί στο μοντέλο

μεγαλύτερη δυνατότητα να αναγνωρίζει τις κλάσεις Rating 2, 3 και 4, οι οποίες περιείχαν πολύ λιγότερα training samples από τις κλάσεις Rating 1 και 5.

Τα αποτελέσματα δεν επιβεβαίωσαν αυτή την προσδοκία. Παρότι το training set έγινε πλήρως ισορροπημένο, τα F1-scores των τριών ενδιάμεσων κλάσεων δεν βελτιώθηκαν στο κοινό validation set. Το αποτέλεσμα αυτό ήταν σημαντικό για την εργασία, επειδή έδειξε στην πράξη ότι η αύξηση του αριθμού των διαθέσιμων training samples μέσω επανάληψης δεν εγγυάται καλύτερη αναγνώριση νέων παραδειγμάτων.

Μετά την εφαρμογή ROS αποκλειστικά στο training split, το νέο BERT αξιολογήθηκε ακριβώς στα ίδια 12.000 validation samples. Η accuracy μειώθηκε σε 0,8723 και το macro F1 σε 0,5535. Το macro precision ήταν 0,5642 και το macro recall 0,5448. Η πτώση δεν περιορίζεται σε μία κλάση: τα F1 των Ratings 2, 3 και 4 μειώθηκαν σε 0,1792, 0,3659 και 0,3580, ενώ μικρές μειώσεις εμφανίστηκαν και στις Ratings 1 και 5.

Το αποτέλεσμα αυτό είναι ιδιαίτερα σημαντικό για τη διπλωματική, επειδή δείχνει ότι η πλήρης εξισορρόπηση του training split δεν μεταφράστηκε σε βελτίωση της επίδοσης στις μειονοτικές κλάσεις. Αντίθετα, το baseline χωρίς ROS είχε υψηλότερη συνολική accuracy και υψηλότερο macro F1. Επομένως, για το συγκεκριμένο dataset και τις συγκεκριμένες ρυθμίσεις fine-tuning, το Random Oversampling δεν παρείχε μετρήσιμο όφελος γενίκευσης.



Σχήμα 6.2: Σύγκριση F1-score ανά rating class στο κοινό validation set.

Πίνακας 6.3: Validation metrics BERT πριν και μετά το Random Oversampling

Κλάση / μέσος	Precision χωρίς ROS	Recall χωρίς ROS	F1 χωρίς ROS	Precision με ROS	Recall με ROS	F1 με ROS
Rating 1	0.92	0.93	0.92	0.9080	0.9279	0.9178
Rating 2	0.22	0.21	0.22	0.1910	0.1687	0.1792
Rating 3	0.39	0.36	0.38	0.3893	0.3452	0.3659
Rating 4	0.45	0.35	0.40	0.3936	0.3284	0.3580
Rating 5	0.94	0.96	0.95	0.9394	0.9536	0.9464
Macro avg	0.59	0.56	0.57	0.5642	0.5448	0.5535
Accuracy			0.8812			0.8723

6.6 Συγκριτική αξιολόγηση των δύο πειραμάτων Random Oversampling

Οι δύο εφαρμογές ROS οδηγούν σε φαινομενικά διαφορετική εικόνα μόνο αν αγνοηθεί το πρωτόκολλο αξιολόγησης. Στο πρώτο σενάριο, η training accuracy αυξάνεται από 96,62% σε 99,37% και το macro F1 από 0,8689 σε 0,9937. Η σύγκριση όμως γίνεται πάνω σε διαφορετικά training sets και περιλαμβάνει duplicated samples, ενώ τα δύο μοντέλα δεν εκπαιδεύονται για ίδιο αριθμό epochs. Στο δεύτερο σενάριο, όπου το validation set είναι κοινό και ανεξάρτητο, η επίδοση μειώνεται μετά το ROS.

Η αντίθεση αυτή αναδεικνύει ένα βασικό μεθοδολογικό μάθημα. Η αποτελεσματικότητα μιας τεχνικής υπερδειγματοληψίας δεν μπορεί να κριθεί από το πόσο καλά ταξινομείται το resampled training set. Ο πραγματικός στόχος είναι να αυξηθεί η αναγνώριση νέων minority observations. Εφόσον το macro F1 του held-out validation set μειώθηκε, η παρούσα εφαρμογή ROS δεν επιτυγχάνει αυτόν τον στόχο, παρότι παράγει σχεδόν τέλεια training metrics.

Ένας επιπλέον περιορισμός στο δεύτερο σενάριο είναι ότι το ROS branch χρησιμοποιεί batch size 64, ενώ το baseline split χρησιμοποιεί batch size 32. Τα epochs παραμένουν τέσσερα και το learning rate 2×10^{-5} , όμως η αλλαγή του batch size σημαίνει ότι η σύγκριση δεν απομονώνει απολύτως μία μόνο μεταβλητή. Παρ' όλα αυτά, η κατεύθυνση των αποτελεσμάτων είναι σαφής: δεν παρατηρείται βελτίωση σε καμία από τις τρεις δύσκολες ενδιάμεσες κλάσεις.

6.7 Αποτελέσματα RoBERTa

Το RoBERTa χρησιμοποιήθηκε σε δύο διαφορετικές πειραματικές ροές. Η πρώτη είναι βοηθητικό πείραμα pseudo-labeling και fine-tuning. Από το `english\_reviews.csv`, το αρχικό `cardiffnlp/twitter-roberta-base-sentiment` παρήγαγε high-confidence pseudo-labels με threshold μεγαλύτερο από 0,90. Με τη δυαδική αντιστοίχιση όπου η αρχική negative κλάση παραμένει 0 και οι neutral/positive συγχωνεύονται στην κλάση 1, προέκυψαν 7.936 samples της κλάσης 1 και μόλις 6 της κλάσης 0.

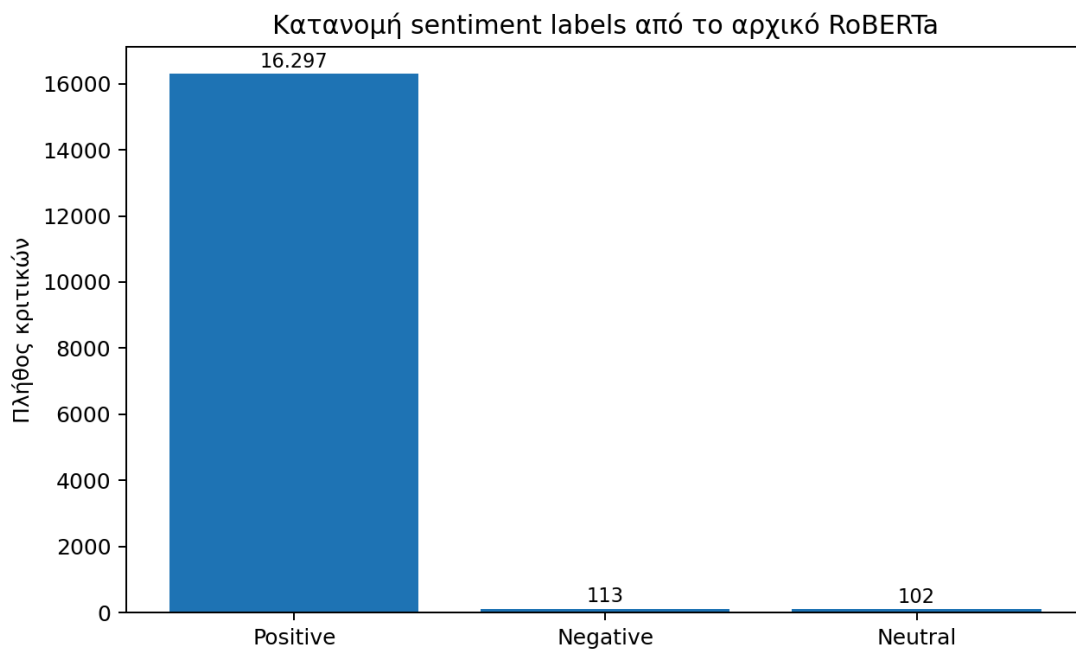
Για να δημιουργηθεί balanced subset δεν χρησιμοποιήθηκε oversampling αλλά undersampling της πλειονοτικής κλάσης: επιλέχθηκαν 6 samples από κάθε κλάση, συνολικά 12. Με split 80/20 το held-out test subset περιορίζεται σε μόλις τρία δείγματα. Η αποθηκευμένη accuracy 1,0000 σε αυτό το subset δεν αποτελεί αξιόπιστη ένδειξη γενίκευσης, τόσο λόγω του εξαιρετικά μικρού support όσο και επειδή οι labels προήλθαν από το ίδιο μοντέλο που χρησιμοποιήθηκε ως βάση για το fine-tuning.

Το συγκεκριμένο πείραμα παραμένει χρήσιμο ως ένδειξη του πόσο ακραία μπορεί να γίνει η ανισορροπία όταν χρησιμοποιείται confidence filtering. Το αρχικό RoBERTa, όταν οι τρεις κλάσεις χαρτογραφήθηκαν σε δύο, έδωσε 8.643 predictions στην κλάση neutral/positive και 45 στην negative. Μετά το fine-tuning, ο κώδικας ονομάζει την έξοδο 1 `positive` και την έξοδο 0 `negative`, με τελική κατανομή 95,626% και 4,374% αντίστοιχα. Ωστόσο, επειδή το label 1 είχε δημιουργηθεί από τη συγχώνευση των αρχικών neutral και positive προβλέψεων, η συγκεκριμένη έξοδος πρέπει να ερμηνεύεται ως η κοινή μη αρνητική δυαδική κλάση και όχι ως ανεξάρτητα επιβεβαιωμένο θετικό sentiment. Η μεταβολή της κατανομής δείχνει ότι ακόμη και ένα πολύ μικρό balanced fine-tuning subset μπορεί να μετακινήσει το decision boundary, αλλά δεν επιτρέπει ανεξάρτητη εκτίμηση accuracy.

6.8 Κατανομή συναισθημάτων RoBERTa

Στο κύριο RoBERTa+SMOTE notebook εξετάστηκαν οι πρώτες 20.000 εγγραφές του dataset και μετά το language filtering διατηρήθηκαν 16.512 αγγλόφωνες κριτικές. Η αρχική ταξινόμηση του pretrained RoBERTa παράγαγε 16.297 positive, 113 negative και 102 neutral labels. Η positive κατηγορία αντιστοιχεί περίπου στο 98,7% του συνόλου, ενώ οι δύο μειονοτικές κλάσεις μαζί αποτελούν περίπου το 1,3%.

Η ακραία αυτή κατανομή μεταφέρεται και μετά το stratified split: στο training set υπάρχουν 13.037 positive embeddings, 90 negative και 82 neutral, ενώ στο validation set τα supports είναι 3.260, 23 και 20 αντίστοιχα. Πρόκειται συνεπώς για σημαντικά δυσκολότερη αριθμητικά ανισορροπία από το BERT rating dataset. Το γεγονός αυτό καθιστά το πείραμα κατάλληλο για να εξεταστεί αν το SMOTE μπορεί να προσφέρει όφελος όταν οι minority representations είναι εξαιρετικά λίγες.



Σχήμα 6.3: Ακραία ανισορροπία των sentiment pseudo-labels που παράγονται από το pretrained RoBERTa.

Πίνακας 6.4: Κατανομή των pseudo-labels στο RoBERTa+SMOTE dataset

Sentiment	Σύνολο	Training	Validation
Positive	16.297	13.037	3.260
Negative	113	90	23
Neutral	102	82	20
Σύνολο	16.512	13.209	3.303

6.9 Αποτελέσματα RoBERTa embeddings χωρίς SMOTE

Ο πρώτος Logistic Regression classifier εκπαιδεύτηκε στα αρχικά 13.209 RoBERTa embeddings χωρίς εξισορρόπηση. Στο validation set των 3.303 samples πέτυχε accuracy 0,9994. Η negative κλάση είχε precision 1,00, recall 0,96 και F1 0,98, η neutral precision 0,91, recall 1,00 και F1 0,95, ενώ η positive εμφανίζεται στρογγυλοποιημένα με precision, recall και F1 ίσα με 1,00. Το macro F1 ήταν 0,98.

Η confusion matrix δείχνει δύο μόνο σφάλματα: από τις 23 negative κριτικές, μία ταξινομήθηκε ως neutral, και από τις 3.260 positive, μία ταξινομήθηκε ως neutral. Και οι 20 neutral ταξινομήθηκαν σωστά. Η επίδοση είναι ασυνήθιστα υψηλή για ένα τόσο imbalanced πρόβλημα και πρέπει να ερμηνευθεί σε συνδυασμό με τον τρόπο δημιουργίας των targets: τα sentiment labels είναι predictions του pretrained RoBERTa και τα features είναι embeddings του ίδιου μοντέλου. Επομένως ο Logistic Regression classifier μαθαίνει έναν representation space που έχει ήδη διαμορφωθεί για την ίδια sentiment διάκριση.

6.10 Αποτελέσματα RoBERTa embeddings με SMOTE

Με το SMOTE, οι 90 negative και 82 neutral training representations αυξήθηκαν συνθετικά ώστε όλες οι κλάσεις να έχουν 13.037 samples. Το νέο training matrix είχε 39.111 γραμμές και 768 χαρακτηριστικά. Ο δεύτερος Logistic Regression classifier εκπαιδεύτηκε στο resampled feature space και αξιολογήθηκε στο ίδιο αρχικό validation set.

Η accuracy ήταν 0,9985 και το macro F1 0,96. Η negative κλάση παρέμεινε ουσιαστικά αμετάβλητη με F1 0,98. Η neutral διατήρησε recall 1,00, αλλά η precision μειώθηκε από 0,91 σε 0,80 και το F1 από 0,95 σε 0,89. Η positive παρέμεινε στρογγυλοποιημένα στο 1,00. Η confusion matrix δείχνει πέντε συνολικά σφάλματα: μία negative και τέσσερις positive κριτικές ταξινομήθηκαν ως neutral.

Το αποτέλεσμα δείχνει ότι το SMOTE αύξησε την τάση του classifier να προβλέπει τη neutral minority class, όμως η επιπλέον ευαισθησία δεν ήταν απαραίτητη, επειδή η baseline recall της neutral ήταν ήδη 1,00. Αντί να διορθώσει false negatives, δημιούργησε περισσότερα false positives και έτσι μείωσε την precision. Με βάση το macro F1, η συνολική ισορροπία των κλάσεων έγινε ελαφρώς χειρότερη.

Πίνακας 6.5: Confusion matrices του RoBERTa embedding classifier

Πραγματική κλάση	Χωρίς SMOTE: Neg/Neu/Pos	Με SMOTE: Neg/Neu/Pos
Negative	22 / 1 / 0	22 / 1 / 0
Neutral	0 / 20 / 0	0 / 20 / 0
Positive	0 / 1 / 3259	0 / 4 / 3256

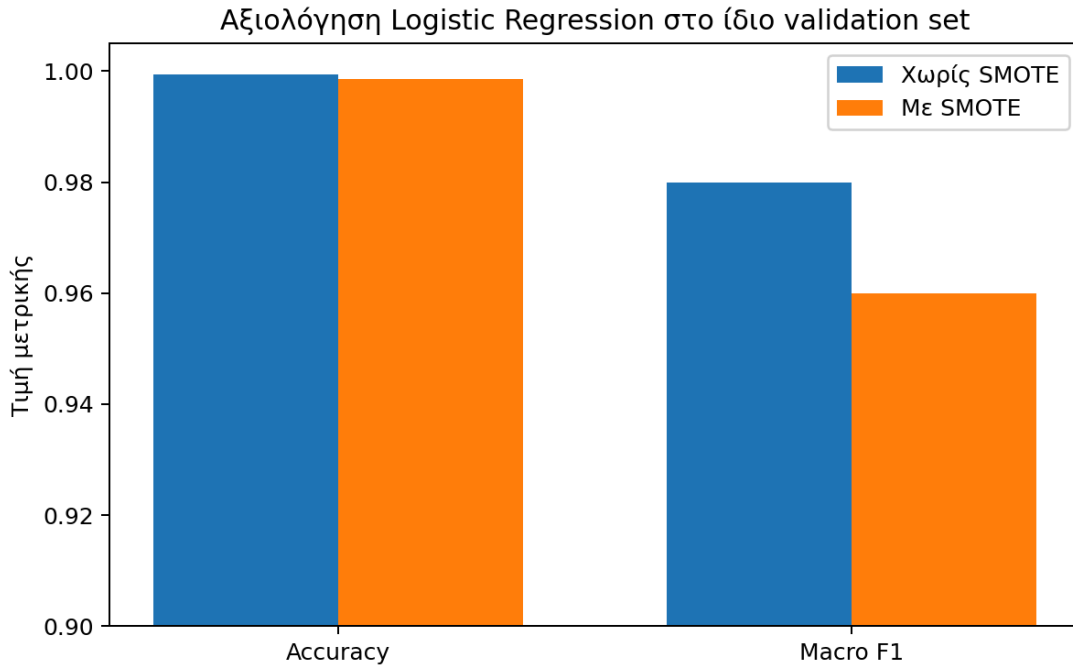
6.11 Συγκριτική αξιολόγηση SMOTE

Η άμεση σύγκριση των δύο Logistic Regression classifiers είναι μεθοδολογικά καθαρότερη από το πρώτο full-data ROS πείραμα, επειδή χρησιμοποιείται το ίδιο validation set, ο ίδιος downstream classifier και το ίδιο frozen RoBERTa feature extractor. Η μόνη κύρια διαφοροποίηση είναι αν τα training embeddings έχουν περάσει ή όχι από SMOTE. Με αυτά τα δεδομένα, το baseline υπερτερεί: accuracy 99,94% έναντι 99,85% και macro F1 0,98 έναντι 0,96.

Η απουσία βελτίωσης δεν σημαίνει ότι το SMOTE είναι γενικά ακατάλληλο για Transformer embeddings. Στο συγκεκριμένο πείραμα υπάρχουν δύο ιδιαίτερες συνθήκες. Πρώτον, το baseline είναι ήδη σχεδόν τέλει ως προς τα pseudo-label targets, οπότε υπάρχει ελάχιστος χώρος βελτίωσης. Δεύτερον, οι minority classes αποτελούνται από μόλις 90 και 82 training observations. Η παρεμβολή στον υψηλής διάστασης χώρο μπορεί να δημιουργεί σημεία που επεκτείνουν υπερβολικά την περιοχή της neutral κλάσης και αυξάνουν τα false positives.

Συνεπώς, το πρακτικό συμπέρασμα δεν είναι ότι η εξισορρόπηση αποτυγχάνει πάντοτε, αλλά ότι η χρησιμότητά της εξαρτάται από την ποιότητα και την ανεξαρτησία των labels, την γεωμετρία του feature space, το μέγεθος των minority classes και το επίπεδο απόδοσης του baseline. Όταν το baseline

έχει ήδη σχεδόν πλήρη συμφωνία με τα pseudo-labels, η τεχνητή ισορροπία μπορεί να προσθέσει θόρυβο αντί για χρήσιμη πληροφορία.



Σχήμα 6.4: Σύγκριση accuracy και macro F1 της Logistic Regression με και χωρίς SMOTE.

Πίνακας 6.6: Validation metrics της Logistic Regression πριν και μετά το SMOTE

Κλάση / μέσος	F1 χωρίς SMOTE	F1 με SMOTE
Negative	0,98	0,98
Neutral	0,95	0,89
Positive	1,00	1,00
Macro F1	0,98	0,96
Accuracy	0,9994	0,9985

6.12 Συγκριτική αξιολόγηση όλων των πειραμάτων

Η συνολική σύγκριση πρέπει να γίνεται με προσοχή, επειδή τα BERT και RoBERTa experiments δεν χρησιμοποιούν το ίδιο dataset ούτε την ίδια μορφή στόχου. Το BERT προβλέπει πέντε ανθρώπινα παρεχόμενες rating classes στο company-reviews dataset. Το RoBERTa pipeline προβλέπει τρεις sentiment pseudo-labels σε διαφορετικό corpus. Επομένως οι απόλυτες accuracy τιμές των δύο οικογενειών δεν αποτελούν μέτρο για το ποιο Transformer είναι «καλύτερο».

Αυτό που μπορεί να συγκριθεί είναι η κατεύθυνση της αλλαγής μέσα σε κάθε πείραμα. Στο held-out BERT σενάριο, η εφαρμογή ROS μείωσε accuracy και macro F1. Στο held-out RoBERTa embedding σενάριο, η εφαρμογή SMOTE επίσης μείωσε ελαφρά και τις δύο μετρικές. Έτσι, στα δύο πειράματα όπου υπάρχει κοινό validation set πριν και μετά την υπερδειγματοληψία, δεν παρατηρείται βελτίωση της γενίκευσης.

Αντίθετα, οι training-only μετρήσεις του πρώτου ROS σεναρίου αυξάνονται εντυπωσιακά. Η διαφορά αυτή αποδεικνύει γιατί είναι αναγκαίο να διαχωρίζονται τα fit metrics από τα held-out metrics. Η υπερδειγματοληψία μπορεί να κάνει το training distribution πλήρως ισορροπημένο και να αυξήσει την επίδοση πάνω σε αυτό, χωρίς να εξασφαλίζει αντίστοιχη βελτίωση σε νέα δεδομένα.

Πίνακας 6.7: Συνοπτική εικόνα των βασικών πειραματικών αποτελεσμάτων

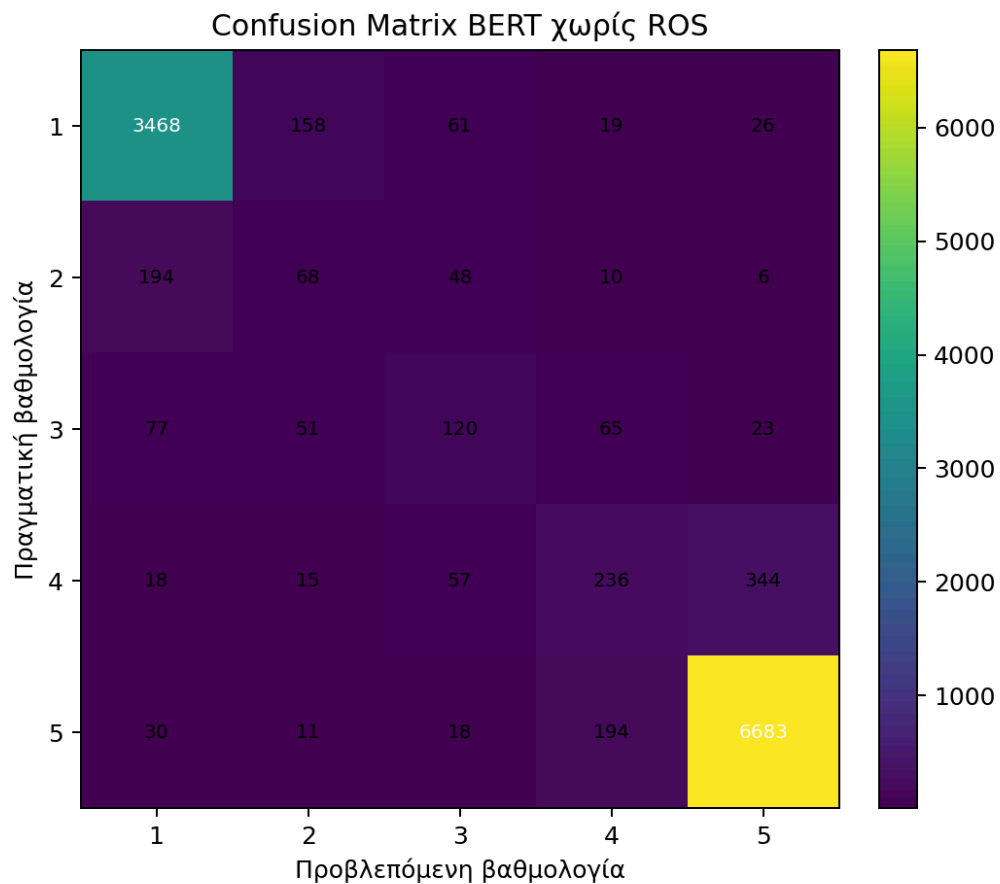
Πείραμα	Oversampling	Evaluation set	Accuracy	Macro F1	Ερμηνεία
BERT full baseline	Όχι	Training	0.9662	0.8689	Training-only
BERT full ROS	ROS	Resampled training	0.9937	0.9937	Training-only, όχι generalization
BERT split baseline	Όχι	Validation 12.000	0.8812	0.57	Held-out σύγκριση
BERT split ROS	ROS	Τόιο validation 12.000	0.8723	0.5535	Held-out σύγκριση
RoBERTa emb. + LR	Όχι	Validation 3.303	0.9994	0.98	Pseudo-label targets
RoBERTa emb. + SMOTE + LR	SMOTE	Τόιο validation 3.303	0.9985	0.96	Pseudo-label targets

6.13 Ανάλυση Confusion Matrices

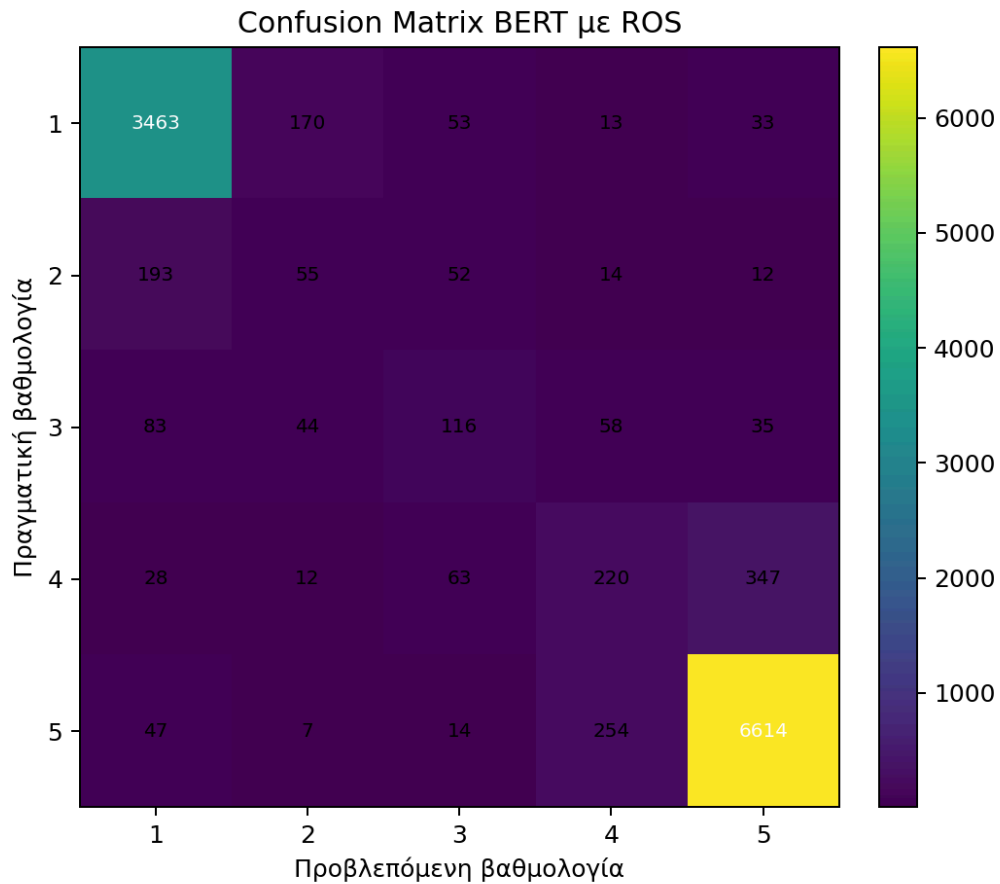
Οι confusion matrices του δεύτερου BERT πειράματος επιτρέπουν λεπτομερέστερη σύγκριση. Χωρίς ROS, 3.468 από τις 3.732 Rating 1 και 6.683 από τις 6.936 Rating 5 προβλέπονται σωστά. Οι ενδιάμεσες κλάσεις είναι πολύ δυσκολότερες: μόνο 68 από τις 326 Rating 2, 120 από τις 336 Rating 3 και 236 από τις 670 Rating 4 ταξινομούνται σωστά.

Με ROS, οι σωστές προβλέψεις γίνονται 3.463 για Rating 1, 55 για Rating 2, 116 για Rating 3, 220 για Rating 4 και 6.614 για Rating 5. Η μεγαλύτερη πρακτική μεταβολή εμφανίζεται στη Rating 4, όπου οι προβλέψεις προς Rating 5 αυξάνονται από 344 σε 347 και οι σωστές προβλέψεις μειώνονται κατά 16. Στη Rating 2 οι σωστές προβλέψεις πέφτουν από 68 σε 55. Δεν εμφανίζεται επομένως το αναμενόμενο κέρδος recall για τις minority classes.

Στο RoBERTa experiment οι matrices είναι ακόμη πιο καθαρές λόγω του μικρού αριθμού λαθών. Χωρίς SMOTE υπάρχουν δύο λάθη προς neutral. Με SMOTE τα λάθη γίνονται πέντε, όλα επίσης προς neutral. Αυτό συμφωνεί με τη μείωση της neutral precision από 0,91 σε 0,80 και δείχνει ότι η synthetic επέκταση της neutral region αυξάνει false-positive neutral predictions.



Σχήμα 6.5: Confusion matrix του BERT baseline στο validation set.



Σχήμα 6.6: Confusion matrix του BERT μετά το Random Oversampling στο ίδιο validation set.

6.14 Επίδραση της υπερδειγματοληψίας στις μειονοτικές κλάσεις

Ο βασικός στόχος της υπερδειγματοληψίας δεν είναι να εξισώσει τις προβλέψεις, αλλά να προσφέρει στο μοντέλο επαρκέστερη πληροφορία για τις σπάνιες κλάσεις. Στο BERT validation experiment, όμως, οι Ratings 2, 3 και 4 δεν εμφανίζουν βελτίωση F1 μετά το ROS. Η Rating 2 μειώνεται από περίπου 0,22 σε 0,1792, η Rating 3 από 0,38 σε 0,3659 και η Rating 4 από 0,40 σε 0,3580. Η εικόνα αυτή δείχνει ότι η απλή επανάληψη ήδη διαθέσιμων κειμένων δεν προσθέτει νέα γλωσσική ποικιλία.

Στο RoBERTa+SMOTE πείραμα, η negative κλάση παραμένει σταθερή και η neutral χειροτερεύει σε precision/F1. Το SMOTE δημιουργεί νέα σημεία στον embedding space, αλλά τα σημεία αυτά δεν αντιστοιχούν απαραίτητα σε πραγματικά νέα γλωσσικά μοτίβα. Ειδικά όταν οι αρχικές minority observations είναι ελάχιστες, η γεωμετρική παρεμβολή μπορεί να δημιουργήσει synthetic vectors κοντά σε περιοχές άλλων κλάσεων.

Τα αποτελέσματα υποστηρίζουν συνεπώς μια πιο συντηρητική θέση: η υπερδειγματοληψία είναι εργαλείο που πρέπει να αξιολογείται πειραματικά και όχι προεπιλεγμένη λύση. Η αντιμετώπιση της έντονης ανισορροπίας μπορεί να είναι σημαντική σε ορισμένα προβλήματα, αλλά η απλή αριθμητική εξισορρόπηση δεν αποτελεί από μόνη της επαρκή συνθήκη για βελτίωση της ταξινόμησης.

6.15 Συζήτηση αποτελεσμάτων

Η αρχική προσδοκία πριν από την εφαρμογή των τεχνικών υπερδειγματοληψίας ήταν σχετικά συντηρητική. Δεν θεωρήθηκε ότι η εξισορρόπηση θα οδηγούσε απαραίτητα σε μεγάλη αύξηση της συνολικής επίδοσης, αλλά αναμενόταν τουλάχιστον μία μικρή βελτίωση στην αναγνώριση των μειονοτικών κλάσεων. Η προσδοκία αυτή δεν επιβεβαιώθηκε ούτε στο αυστηρότερο πείραμα Random Oversampling ούτε στη σύγκριση των RoBERTa embeddings πριν και μετά το SMOTE.

Το αποτέλεσμα αυτό μετατόπισε το ενδιαφέρον της ανάλυσης από το αν μια τεχνική μπορεί απλώς να εξισώσει αριθμητικά το training set στο αν αυτή η εξισορρόπηση μεταφέρεται πραγματικά σε καλύτερη απόδοση πάνω σε δεδομένα που το μοντέλο δεν έχει δει προηγουμένως.

6.15.1 Random Oversampling

Το ROS έχει το πλεονέκτημα ότι δεν αλλοιώνει τα ίδια τα κείμενα. Στο BERT pipeline οι minority reviews επαναλαμβάνονται και το Transformer εκπαιδεύεται σε αυξημένη συχνότητα αυτών των παραδειγμάτων. Αυτό μπορεί να είναι χρήσιμο όταν η βασική αιτία χαμηλού recall είναι η ανεπαρκής έκθεση του μοντέλου στη minority class. Στην παρούσα εργασία, όμως, η επανάληψη δεν παρήγαγε καλύτερο validation F1. Η πιθανότερη ερμηνεία είναι ότι το μοντέλο απέκτησε μεγαλύτερη προσαρμογή στα συγκεκριμένα minority samples χωρίς να αποκτήσει νέες γλωσσικές παραλλαγές που θα βοηθούσαν σε unseen reviews.

6.15.2 SMOTE

Το SMOTE αντιμετωπίζει διαφορετικά το πρόβλημα, δημιουργώντας νέα σημεία μεταξύ κοντινών minority embeddings. Η τεχνική αποφεύγει την ακριβή αντιγραφή samples, αλλά εισάγει την υπόθεση ότι οι ενδιάμεσες θέσεις στον representation space αποτελούν έγκυρα σημεία της ίδιας κλάσης. Στα RoBERTa embeddings της συγκεκριμένης εφαρμογής, η υπόθεση αυτή δεν απέδωσε καλύτερο validation αποτέλεσμα. Η neutral recall ήταν ήδη τέλεια στο baseline, και η synthetic επέκταση αύξησε κυρίως false positives.

6.15.3 Πότε βοήθησε η υπερδειγματοληψία

Με αυστηρό κριτήριο held-out generalization, στα παρόντα πειράματα δεν τεκμηριώνεται σαφής βελτίωση από ROS ή SMOTE. Η υπερδειγματοληψία βοήθησε στο να καταστεί η training distribution ομοιόμορφη και, στο πρώτο BERT σενάριο, αύξησε σημαντικά την επίδοση πάνω στο resampled training set. Αυτό δείχνει ότι το μοντέλο μπορεί να προσαρμοστεί πολύ καλά στα επαναλαμβανόμενα minority training samples, χωρίς αυτό να συνεπάγεται καλύτερη γενίκευση σε νέα δεδομένα. Πράγματι, το όφελος δεν μεταφέρθηκε στο ανεξάρτητο validation scenario.

6.15.4 Πότε δεν βοήθησε

Το ROS δεν βοήθησε όταν η αξιολόγηση πραγματοποιήθηκε σε validation data με την πραγματική ανισόρροπη κατανομή. Το SMOTE επίσης δεν βοήθησε όταν εφαρμόστηκε σε ήδη πολύ διαχωρίσιμα RoBERTa embeddings που είχαν targets παραγόμενα από το ίδιο μοντέλο. Και στις δύο περιπτώσεις η υπερδειγματοληψία μεταβάλλει έντονα το training space, αλλά η μεταβολή δεν προσφέρει νέα πραγματική πληροφορία για τη γλωσσική ποικιλία των minority classes.

6.15.5 Πιθανές αιτίες

Πιθανές αιτίες της απουσίας βελτίωσης είναι η επανάληψη ίδιων κειμένων στο ROS, ο πολύ μικρός αριθμός πραγματικών minority observations στο SMOTE, η υψηλή διάσταση των embeddings, η φυσική αλληλοεπικάλυψη γειτονικών rating classes και η διαφορά ανάμεσα σε ισορροπημένο training prior και πραγματικό validation prior. Στο RoBERTa experiment προστίθεται η ισχυρή εξάρτηση features και pseudo-labels από το ίδιο pretrained μοντέλο, η οποία καθιστά το baseline εξαιρετικά ισχυρό και περιορίζει το περιθώριο βελτίωσης.

Επιπλέον, το oversampling δεν συνοδεύτηκε από συστηματικό hyperparameter search. Το default SMOTE χρησιμοποιεί συγκεκριμένο αριθμό γειτόνων, ενώ το ROS εφαρμόζει πλήρη εξίσωση προς τη majority class. Είναι πιθανό η μερική εξισορρόπηση, η επιλογή διαφορετικού sampling ratio ή η προσαρμογή των training hyperparameters να οδηγούσαν σε διαφορετική συμπεριφορά.

6.16 Περιορισμοί της πειραματικής διαδικασίας

Ο πρώτος περιορισμός είναι ότι τα δύο κύρια experimental families βασίζονται σε διαφορετικά datasets και διαφορετικούς στόχους. Το BERT πρόβλημα είναι πενταταξική πρόβλεψη rating, ενώ το RoBERTa πρόβλημα είναι τριταξική sentiment classification. Για αυτό δεν πραγματοποιείται άμεση σύγκριση BERT έναντι RoBERTa με βάση τις απόλυτες τιμές accuracy.

Δεύτερος περιορισμός είναι η έλλειψη labeled test set στο πρώτο BERT σενάριο. Το notebook διαθέτει 40.000 test reviews, αλλά χωρίς ground-truth ratings. Οι κατανομές predictions μπορούν να αναλυθούν, όμως δεν μπορούν να υπολογιστούν test precision, recall ή F1. Παράλληλα, η σύγκριση pre/post balancing στο full dataset βασίζεται σε training evaluation και σε διαφορετικό αριθμό epochs.

Τρίτος περιορισμός αφορά τα RoBERTa targets. Οι sentiment labels προκύπτουν από το pretrained `cardiffnlp/twitter-roberta-base-sentiment`, ενώ τα embeddings παράγονται από το backbone του ίδιου checkpoint. Η αξιολόγηση επομένως δεν είναι ανεξάρτητη από το μοντέλο που δημιούργησε τις labels. Απαιτείται human-labeled test set για ισχυρό ισχυρισμό σχετικά με πραγματική sentiment accuracy.

Τέταρτος περιορισμός είναι το μικρό support των minority classes. Στο SMOTE training split υπάρχουν μόνο 90 negative και 82 neutral πραγματικές παρατηρήσεις. Η δημιουργία πάνω από 12.900 synthetic observations ανά minority class αποτελεί πολύ μεγάλη επέκταση σε σχέση με το αρχικό δείγμα και μπορεί να ενισχύσει artifacts της τοπικής γεωμετρίας.

Πέμπτος περιορισμός είναι η εκτέλεση των πειραμάτων με ένα κύριο random seed και χωρίς repeated cross-validation. Ιδιαίτερα στις μικρές κλάσεις, διαφορετικό split μπορεί να αλλάξει αισθητά τα per-class metrics. Επιπλέον, στο δεύτερο BERT ROS branch το batch size είναι 64 έναντι 32 στο baseline, άρα η επίδραση του ROS δεν απομονώνεται απόλυτα.

Τέλος, τα notebooks δημιουργήθηκαν σε διερευνητικό περιβάλλον Google Colab και περιλαμβάνουν επανεκτελέσεις κελιών και επαναχρησιμοποίηση global μεταβλητών. Η τελική ανάλυση βασίστηκε στις ρητές αρχικές φορτώσεις δεδομένων, στα σαφή splits και στα αποθηκευμένα outputs. Για πλήρη αναπαραγωγιμότητα θα ήταν προτιμότερο κάθε τελικό experiment να βρίσκεται σε ξεχωριστό notebook ή script που εκτελείται από καθαρό runtime από την αρχή έως το τέλος.

6.17 Επίλογος

Τα αποτελέσματα δείχνουν ότι η αντιμετώπιση class imbalance δεν είναι ισοδύναμη με την απλή εξίσωση των training counts. Στο BERT, το ROS αύξησε εντυπωσιακά τα training metrics του πλήρως resampled dataset, αλλά στο πείραμα με κοινό ανέγγιχτο validation set η accuracy μειώθηκε από 88,12% σε 87,23% και το macro F1 από 0,57 σε 0,5535. Οι μειονοτικές Ratings 2, 3 και 4 δεν βελτιώθηκαν.

Στο RoBERTa embedding experiment, το baseline Logistic Regression ήδη αναπαρήγαγε σχεδόν τέλεια τα RoBERTa pseudo-labels. Μετά το SMOTE η accuracy μειώθηκε από 99,94% σε 99,85% και το macro F1 από 0,98 σε 0,96, κυρίως λόγω περισσότερων false-positive neutral predictions. Το αποτέλεσμα δεν αποτελεί γενική απόρριψη του SMOTE, αλλά δείχνει ότι η χρησιμότητα synthetic oversampling εξαρτάται από την ποιότητα του representation space και τη φύση των targets.

Συνολικά, τα δύο ισχυρότερα held-out comparisons της εργασίας συγκλίνουν στο ίδιο συμπέρασμα: οι εξεταζόμενες μορφές υπερδειγματοληψίας δεν βελτίωσαν τη γενίκευση με τις συγκεκριμένες ρυθμίσεις. Το επόμενο κεφάλαιο συνοψίζει τα συμπεράσματα, απαντά στα ερευνητικά ερωτήματα και παρουσιάζει προτάσεις για μελλοντική βελτίωση της πειραματικής διαδικασίας.

Κεφάλαιο 7ο: Συμπεράσματα και Προτάσεις Βελτίωσης

7.1 Εισαγωγή

Το τελευταίο κεφάλαιο συνοψίζει τα ευρήματα της διπλωματικής εργασίας και τα συνδέει με τα ερευνητικά ερωτήματα που τέθηκαν στην Εισαγωγή. Η εργασία είχε ως κεντρικό αντικείμενο την εφαρμογή μεθόδων υπερδειγματοληψίας σε προβλήματα ανάλυσης συναισθήματος με μη ισορροπημένα δεδομένα. Η πειραματική μελέτη δεν αντιμετώπισε το oversampling ως τεχνική που αναμένεται εκ των προτέρων να βελτιώσει την επίδοση, αλλά ως παρέμβαση της οποίας η πραγματική επίδραση πρέπει να επιβεβαιώνεται σε ανεξάρτητα δεδομένα.

Οι δύο κύριες τεχνικές που εξετάστηκαν ήταν το Random Oversampling και το SMOTE. Το ROS εφαρμόστηκε σε raw text samples πριν από το BERT fine-tuning, ενώ το SMOTE εφαρμόστηκε σε 768-dimensional RoBERTa embeddings πριν από Logistic Regression. Με αυτόν τον τρόπο η εργασία κάλυψε δύο διαφορετικά επίπεδα υπερδειγματοληψίας: επανάληψη πραγματικών παραδειγμάτων στο επίπεδο του dataset και δημιουργία synthetic examples στο επίπεδο της διανυσματικής αναπαράστασης.

7.2 Ανακεφαλαίωση της εργασίας

Αρχικά παρουσιάστηκε το θεωρητικό υπόβαθρο της Επεξεργασίας Φυσικής Γλώσσας και της ανάλυσης συναισθήματος, οι βασικές δυσκολίες της ταξινόμησης κειμένου και η μετάβαση από παραδοσιακές τεχνικές Machine Learning σε Transformer-based μοντέλα. Στη συνέχεια αναλύθηκε το πρόβλημα του class imbalance, οι συνέπειές του στις minority classes και οι βασικές στρατηγικές αντιμετώπισής του. Ιδιαίτερη έμφαση δόθηκε στο Random Oversampling και στο SMOTE, καθώς αποτελούν τις τεχνικές που χρησιμοποιήθηκαν πειραματικά.

Το θεωρητικό μέρος ολοκληρώθηκε με την παρουσίαση της αρχιτεκτονικής Transformer, του self-attention, του BERT και του RoBERTa. Η πρακτική υλοποίηση χρησιμοποίησε `'bert-base-uncased'` για πενταταξική ταξινόμηση ratings και `'cardiffnlp/twitter-roberta-base-sentiment'` για sentiment labeling και contextual embeddings. Τα πειράματα υλοποιήθηκαν σε Google Colab με PyTorch και Hugging Face Transformers, ενώ οι τεχνικές υπερδειγματοληψίας προήλθαν από την imbalanced-learn.

Για το BERT υλοποιήθηκαν δύο ROS σενάρια. Το πρώτο εξισορρόπησε το πλήρες training dataset και ανέδειξε πολύ υψηλές training επιδόσεις, χωρίς όμως labeled test set. Το δεύτερο πραγματοποίησε split πριν το ROS και αξιολόγησε baseline και balanced μοντέλο στο ίδιο validation set. Για το RoBERTa δημιουργήθηκαν embeddings σε ξεχωριστό corpus, εφαρμόστηκε SMOTE μόνο στο training subset και συγκρίθηκαν δύο Logistic Regression classifiers στο ίδιο validation set.

7.3 Απάντηση στα ερευνητικά ερωτήματα

Με βάση τα πειραματικά αποτελέσματα και τη συζήτηση που προηγήθηκε, στην ενότητα αυτή δίνονται συγκεντρωτικές απαντήσεις στα ερευνητικά ερωτήματα της εργασίας. Η παρουσίαση ακολουθεί τη σειρά των αρχικών ερωτημάτων και εστιάζει στο class imbalance, στο Random Oversampling, στο SMOTE και στις μετρικές που αποδείχθηκαν καταλληλότερες για την αξιολόγηση.

7.3.1 Επίδραση του class imbalance

Το πρώτο ερευνητικό ερώτημα αφορούσε την επίδραση της ανισορροπίας στην απόδοση των μοντέλων. Τα BERT αποτελέσματα παρέχουν σαφή απάντηση. Παρότι το baseline validation accuracy ήταν 88,12%, οι ενδιάμεσες Ratings 2, 3 και 4 είχαν F1 μόλις 0,22, 0,38 και 0,40, σε αντίθεση με τις Ratings 1 και 5 που βρίσκονταν περίπου στο 0,92 και 0,95. Η συνολική accuracy επομένως απέκρυπτε πολύ διαφορετική συμπεριφορά ανά κλάση.

Η ίδια αριθμητική ανισορροπία εμφανίζεται ακόμη εντονότερα στο RoBERTa corpus, όπου 16.297 από τις 16.512 pseudo-labeled κριτικές είναι positive. Στη συγκεκριμένη περίπτωση η baseline Logistic Regression παραμένει εξαιρετικά επιτυχής, αλλά αυτό συνδέεται με την προέλευση των features και των labels από το ίδιο RoBERTa μοντέλο. Το γενικό συμπέρασμα είναι ότι το class imbalance πρέπει να εξετάζεται μαζί με τη φύση της αναπαράστασης, την ποιότητα των labels και τις per-class μετρικές.

7.3.2 Επίδραση του Random Oversampling

Το δεύτερο ερευνητικό ερώτημα αφορούσε το κατά πόσο το Random Oversampling βελτιώνει την αναγνώριση των μειονοτικών κλάσεων. Στο συγκεκριμένο πρωτόκολλο δεν παρατηρήθηκε βελτίωση: η accuracy μειώθηκε από 0,8812 σε 0,8723, το macro F1 από 0,57 σε 0,5535 και οι Ratings 2–4 παρουσίασαν μικρότερα F1. Ωστόσο, επειδή το ROS branch χρησιμοποίησε batch size 64 αντί 32 στο baseline, οι μεταβολές δεν μπορούν να αποδοθούν αποκλειστικά στην υπερδειγματοληψία.

Η μεγάλη βελτίωση που εμφανίζεται στο πρώτο full-data ROS σενάριο αφορά training evaluation και δεν πρέπει να συγχέεται με generalization. Το γεγονός ότι το μοντέλο φτάνει macro F1 περίπου 0,994 πάνω στο resampled training set δείχνει ότι μπορεί να προσαρμοστεί στα duplicated minority examples. Το δεύτερο σενάριο, όμως, δείχνει ότι αυτή η προσαρμογή δεν μεταφέρεται απαραίτητα σε unseen samples. Επομένως το ROS δεν πρέπει να αξιολογείται από την επίδοση στο balanced training set αλλά από την επίδοση σε ανέγγιχτη κατανομή.

7.3.3 Επίδραση του SMOTE

Το τρίτο ερευνητικό ερώτημα εξέτασε την εφαρμογή SMOTE σε Transformer embeddings. Η τεχνική κατάφερε αριθμητικά να μετατρέψει το training set από 13.037 positive, 90 negative και 82 neutral embeddings σε 13.037 samples ανά κλάση. Ωστόσο η validation επίδοση δεν βελτιώθηκε. Η accuracy μειώθηκε από 0,9994 σε 0,9985 και το macro F1 από 0,98 σε 0,96.

Η σημαντικότερη αλλαγή εμφανίστηκε στη neutral κλάση: η recall παρέμεινε 1,00, αλλά η precision μειώθηκε από 0,91 σε 0,80 και το F1 από 0,95 σε 0,89. Το SMOTE δημιούργησε περισσότερες false-positive neutral predictions χωρίς να διορθώσει κάποιο πρόβλημα recall που υπήρχε στο baseline. Συνεπώς, στο συγκεκριμένο frozen RoBERTa representation space και με τα συγκεκριμένα pseudo-label targets, το synthetic oversampling δεν πρόσφερε όφελος.

7.3.4 Κατάλληλες μετρικές αξιολόγησης

Το τέταρτο ερευνητικό ερώτημα αφορούσε τις κατάλληλες μετρικές για μη ισορροπημένα δεδομένα. Τα πειράματα επιβεβαιώνουν ότι η accuracy από μόνη της είναι ανεπαρκής. Στο BERT validation set, accuracy 88,12% συνυπάρχει με F1 0,22 για Rating 2. Αν εξετάζοταν μόνο η accuracy, η αδυναμία του μοντέλου σε αυτή την κλάση θα παρέμενε σε μεγάλο βαθμό κρυμμένη.

Το macro F1 είναι ιδιαίτερα χρήσιμο επειδή δίνει ίσο βάρος σε όλες τις κλάσεις. Παράλληλα, τα per-class precision και recall είναι απαραίτητα για να διαπιστωθεί αν μια μέθοδος αυξάνει την ανίχνευση της minority class εις βάρος πολλών false positives. Οι confusion matrices συμπληρώνουν την εικόνα, επειδή δείχνουν προς ποιες κλάσεις μετακινούνται τα λάθη. Για τον λόγο αυτό η παρούσα εργασία θεωρεί τον συνδυασμό macro F1, per-class F1/recall/precision και confusion matrix καταλληλότερο από μία μόνο συνολική μετρική.

7.4 Κύρια συμπεράσματα

Το βασικό συμπέρασμα είναι ότι η υπερδειγματοληψία δεν αποτελεί αυτόματη λύση στο class imbalance. Η εξισορρόπηση της αριθμητικής κατανομής μπορεί να αλλάξει σημαντικά το training process χωρίς να προσθέσει νέα πληροφορία. Το ROS αυξάνει τη συχνότητα των ήδη διαθέσιμων minority texts, ενώ το SMOTE δημιουργεί νέες διανυσματικές θέσεις με βάση τις ήδη διαθέσιμες minority representations. Και οι δύο τεχνικές μπορούν συνεπώς να ενισχύσουν υπάρχοντα patterns χωρίς να αυξάνουν αναγκαστικά τη γλωσσική ποικιλία.

Δεύτερο βασικό συμπέρασμα είναι η σημασία του πειραματικού σχεδιασμού. Το πρώτο BERT πείραμα θα μπορούσε, αν εξεταζόταν μόνο επιφανειακά, να οδηγήσει στο συμπέρασμα ότι το ROS είναι εξαιρετικά αποτελεσματικό, επειδή το macro F1 του training set αυξάνεται από 0,8689 σε 0,9937. Η ανεξάρτητη validation σύγκριση δείχνει την αντίθετη εικόνα. Επομένως κάθε τεχνική sampling πρέπει να εφαρμόζεται μόνο στο training partition και η τελική αξιολόγηση να πραγματοποιείται σε δεδομένα που δεν συμμετείχαν στην εξισορρόπηση.

Τρίτο συμπέρασμα είναι ότι η πηγή των labels έχει καθοριστική σημασία. Στο RoBERTa pipeline τα υψηλά metrics δεν πρέπει να παρουσιαστούν ως 99,9% πραγματική accuracy sentiment analysis. Τα pseudo-labels έχουν παραχθεί από το ίδιο μοντέλο που δημιουργεί τον representation space, οπότε η downstream ταξινόμηση είναι εγγενώς ευκολότερη. Η συγκεκριμένη υλοποίηση είναι χρήσιμη για τη μελέτη του SMOTE σε embeddings, αλλά για αξιολόγηση πραγματικής ακρίβειας απαιτείται ανεξάρτητη ανθρώπινη επισήμανση.

Τέλος, η εργασία δείχνει ότι ένα αρνητικό πειραματικό αποτέλεσμα παραμένει χρήσιμο. Το γεγονός ότι ROS και SMOTE δεν βελτίωσαν τις βασικές held-out μετρικές δεν αποτελεί αποτυχία της μελέτης, αλλά εύρημα που οριοθετεί τις συνθήκες υπό τις οποίες η απλή εξισορρόπηση δεν επαρκεί. Η αξία βρίσκεται στη σύγκριση, στην καταγραφή των περιορισμών και στη δυνατότητα να σχεδιαστούν ισχυρότερα επόμενα πειράματα.

7.5 Περιορισμοί

Οι σημαντικότεροι περιορισμοί σχετίζονται με τα datasets και το evaluation protocol. Δεν υπάρχει ενιαίο human-labeled dataset πάνω στο οποίο να έχουν αξιολογηθεί BERT+ROS και RoBERTa+SMOTE με ακριβώς τις ίδιες κλάσεις. Το BERT test set του πρώτου σεναρίου είναι μη επισημασμένο, ενώ το RoBERTa sentiment dataset χρησιμοποιεί pseudo-labels. Αυτό περιορίζει τη δυνατότητα ενιαίου ranking όλων των μεθόδων.

Επιπλέον, δεν πραγματοποιήθηκαν repeated runs με πολλαπλά random seeds ή cross-validation. Οι minority classes είναι μικρές, επομένως ένα μόνο split μπορεί να επηρεάζει το αποτέλεσμα. Στο δεύτερο BERT πείραμα υπάρχει και διαφορά batch size μεταξύ baseline και ROS branch, ενώ στο πρώτο full-data σενάριο υπάρχει διαφορά στον αριθμό epochs. Αυτοί οι παράγοντες πρέπει να ελεγχθούν σε επόμενη, αυστηρότερα ελεγχόμενη πειραματική έκδοση.

Τέλος, η διερευνητική φύση των notebooks οδήγησε σε επαναχρησιμοποίηση μεταβλητών και outputs από διαφορετικές εκτελέσεις. Παρότι τα κύρια αποτελέσματα μπορούν να απομονωθούν, η αναπαραγωγικότητα θα βελτιωνόταν σημαντικά με ξεχωριστά scripts για κάθε experiment, σταθερό configuration file, αποθήκευση των seeds και αυτόματη εξαγωγή όλων των metrics σε κοινή μορφή.

Ένας επιπλέον πρακτικός περιορισμός προέκυψε από τον τρόπο οργάνωσης της πειραματικής διαδικασίας στο Google Colab. Επειδή αρκετά στάδια βασίζονταν σε μεταβλητές που διατηρούνταν μόνο στη μνήμη του ενεργού runtime, μία αποσύνδεση ή η επιστροφή στο notebook μετά από μεγάλο χρονικό διάστημα μπορούσε να απαιτήσει επανεκτέλεση μεγάλου μέρους του κώδικα. Αυτό αύξανε τον χρόνο που απαιτούνταν για μικρές αλλαγές ή συμπληρωματικές δοκιμές και δυσκόλευε την επαναληψιμότητα ορισμένων πειραμάτων.

7.6 Μελλοντικές επεκτάσεις

Η σημαντικότερη μελλοντική επέκταση είναι η δημιουργία ή χρήση ενός ενιαίου, ανεξάρτητα επισημασμένου sentiment dataset με επαρκές πλήθος παραδειγμάτων σε όλες τις κλάσεις. Πάνω στο ίδιο dataset θα μπορούσαν να συγκριθούν baseline Transformer, ROS, SMOTE-on-embeddings και άλλες imbalance-aware τεχνικές χωρίς να αλλάζει ο στόχος ταξινόμησης. Ένα untouched test set θα επέτρεπε ασφαλέστερη εξαγωγή τελικών συμπερασμάτων [47], [48].

Δεύτερη κατεύθυνση είναι η σύγκριση oversampling με cost-sensitive learning. Αντί να μεταβάλλεται το training dataset, μπορούν να χρησιμοποιηθούν class weights ή cost-sensitive objectives ώστε τα λάθη στις minority classes να έχουν μεγαλύτερη επίδραση στη βελτιστοποίηση. Έτσι αποφεύγεται η επανάληψη ίδιων κειμένων και μπορεί να διατηρηθεί η φυσική κατανομή του dataset [45].

Για το Random Oversampling θα ήταν χρήσιμη η μελέτη διαφορετικών sampling ratios αντί της πλήρους εξίσωσης όλων των κλάσεων προς τη majority class. Η υπερβολικά μεγάλη αύξηση των minority samples μπορεί να οδηγήσει σε overfitting. Μερική εξισορρόπηση, συνδυασμένη με early stopping και σταθερό batch size, θα επέτρεπε να εντοπιστεί καλύτερο σημείο ισορροπίας [40]-[42].

Για το SMOTE μπορούν να εξεταστούν διαφορετικές τιμές `k\_neighbors`, μερικό sampling και παραλλαγές όπως Borderline-SMOTE ή Safe-Level-SMOTE. Επίσης είναι ενδιαφέρον να συγκριθούν διαφορετικές αναπαραστάσεις, για παράδειγμα pooled outputs από διαφορετικά Transformer layers ή sentence-level embeddings, ώστε να εξεταστεί αν ο representation space είναι περισσότερο κατάλληλος για γεωμετρική παρεμβολή [43]-[44], [50].

Η αξιολόγηση μπορεί να ενισχυθεί με repeated stratified splits ή k-fold cross-validation και αναφορά μέσου όρου και τυπικής απόκλισης των metrics. Η εκτέλεση με πολλαπλά seeds είναι ιδιαίτερα σημαντική στις μικρές minority classes. Παράλληλα, η βελτιστοποίηση learning rate, batch size, αριθμού epochs και decision thresholds θα πρέπει να γίνεται μόνο μέσω validation data και όχι στο τελικό test set [2], [11], [13].

Σε μία νέα υλοποίηση, μία από τις πρώτες αλλαγές θα αφορούσε την οργάνωση των notebooks. Θα ήταν προτιμότερο τα διαφορετικά πειράματα και τα βασικά στάδια επεξεργασίας να είναι περισσότερο ανεξάρτητα μεταξύ τους, ώστε η εκτέλεση ενός σταδίου να μην προϋποθέτει ότι ολόκληρο το προηγούμενο notebook παραμένει ενεργό στη μνήμη.

Παράλληλα, θα μπορούσαν να πραγματοποιούνται συστηματικές ενδιάμεσες αποθηκεύσεις μετά από χρονοβόρα στάδια, όπως η προεπεξεργασία των δεδομένων, η παραγωγή embeddings ή η εκπαίδευση ενός μοντέλου. Έτσι, μετά από πιθανή αποσύνδεση του Google Colab, το πείραμα θα μπορούσε να συνεχιστεί από το τελευταίο ολοκληρωμένο στάδιο αντί να απαιτείται επανεκτέλεση μεγάλων

τμημάτων του κώδικα. Η αλλαγή αυτή θα βελτιώνει τόσο την πρακτική διαχείριση του χρόνου όσο και την επαναληψιμότητα της εκτέλεσης των πειραμάτων.

Τέλος, θα ήταν χρήσιμη η δημιουργία πλήρως αναπαραγωγίμου λογισμικού pipeline. Κάθε πείραμα μπορεί να εκτελείται από καθαρό runtime, να αποθηκεύει αυτόματα configuration, class distributions, metrics και confusion matrices και να παράγει κοινό report. Μια τέτοια δομή θα διευκόλυνε τόσο την επανάληψη της μελέτης σε άλλα datasets όσο και την προσθήκη νέων τεχνικών αντιμετώπισης της ανισορροπίας.

7.7 Επίλογος

Η διπλωματική εργασία μελέτησε στην πράξη δύο ευρέως χρησιμοποιούμενες στρατηγικές υπερδειγματοληψίας σε Transformer-based εφαρμογές ανάλυσης συναισθήματος. Τα πειράματα έδειξαν ότι η αριθμητική εξισορρόπηση του training set είναι εύκολο να επιτευχθεί, αλλά η ουσιαστική βελτίωση της γενίκευσης είναι πολύ πιο απαιτητική. Στα πειράματα με κοινό held-out validation set, ούτε το Random Oversampling ούτε το SMOTE βελτίωσαν το macro F1.

Το αποτέλεσμα υπογραμμίζει ότι η αντιμετώπιση του class imbalance πρέπει να συνδυάζεται με προσεκτικό evaluation design, ανεξάρτητες labels και μετρικές που δίνουν πραγματικό βάρος στις minority classes. Η παρούσα εργασία παρέχει ένα πρακτικό πλαίσιο εφαρμογής ROS σε raw text και SMOTE σε Transformer embeddings, καταγράφει τα όριά τους στα εξεταζόμενα δεδομένα και δημιουργεί βάση για συστηματικότερη μελλοντική διερεύνηση [46]-[50].

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval*, vol. 2, nos. 1-2, pp. 1-135, 2008, doi: 10.1561/15000000011.
- [2] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.
- [3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002, doi: 10.1613/jair.953.
- [4] A. Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998-6008.
- [5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, Minneapolis, MN, USA: Association for Computational Linguistics, 2019, pp. 4171-4186, doi: 10.18653/v1/N19-1423.
- [6] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
- [7] B. Liu, *Sentiment Analysis and Opinion Mining*, *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1. Morgan & Claypool Publishers, 2012, pp. 1-167, doi: 10.2200/S00416ED1V01Y201204HLT016.
- [8] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.
- [9] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. Upper Saddle River, NJ, USA: Pearson Prentice Hall, 2009.
- [10] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429-449, 2002, doi: 10.3233/IDA-2002-6504.
- [11] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, pp. 221-232, 2016, doi: 10.1007/s13748-016-0094-0.
- [12] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 20-29, 2004, doi: 10.1145/1007730.1007735.
- [13] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, art. no. 27, 2019, doi: 10.1186/s40537-019-0192-5.
- [14] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," in *Proceedings of EMNLP-IJCNLP*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 6382-6388, doi: 10.18653/v1/D19-1670.
- [15] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," in *Findings of EMNLP 2020*,

Online: Association for Computational Linguistics, 2020, pp. 1644-1650, doi: 10.18653/v1/2020.findings-emnlp.148.

[16] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1-5, 2017.

[17] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825-2830, 2011.

[18] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 8024-8035.

[19] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of EMNLP: System Demonstrations*, Online: Association for Computational Linguistics, 2020, pp. 38-45, doi: 10.18653/v1/2020.emnlp-demos.6.

[20] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in *Proceedings of ACL*, Philadelphia, PA, USA, 2002, pp. 417-424, doi: 10.3115/1073083.1073153.

[21] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," in *Proceedings of KDD*, Seattle, WA, USA, 2004, pp. 168-177, doi: 10.1145/1014052.1014073.

[22] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267-307, 2011, doi: 10.1162/COLI_a_00049.

[23] A. L. Maas et al., "Learning Word Vectors for Sentiment Analysis," in *Proceedings of ACL-HLT*, Portland, OR, USA, 2011, pp. 142-150.

[24] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of EMNLP*, Seattle, WA, USA, 2013, pp. 1631-1642.

[25] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," in *Machine Learning: ECML-98*, Berlin, Germany: Springer, 1998, pp. 137-142, doi: 10.1007/BFb0026683.

[26] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," in *AAAI-98 Workshop on Learning for Text Categorization*, Madison, WI, USA, 1998, pp. 41-48.

[27] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871-1874, 2008.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *arXiv preprint arXiv:1301.3781*, 2013.

[29] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Proceedings of EMNLP*, Doha, Qatar, 2014, pp. 1532-1543, doi: 10.3115/v1/D14-1162.

[30] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017, doi: 10.1162/tacl_a_00051.

- [31] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proceedings of EMNLP, Doha, Qatar, 2014, pp. 1746-1751, doi: 10.3115/v1/D14-1181.
- [32] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [33] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [34] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in Proceedings of ICLR, New Orleans, LA, USA, 2019.
- [35] P. Branco, L. Torgo, and R. P. Ribeiro, "A Survey of Predictive Modeling on Imbalanced Domains," *ACM Computing Surveys*, vol. 49, no. 2, art. no. 31, pp. 1-50, 2016, doi: 10.1145/2907070.
- [36] A. Fernandez, S. Garcia, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-319-98074-4.
- [37] H. Kaur, H. S. Pannu, and A. K. Malhi, "A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions," *ACM Computing Surveys*, vol. 52, no. 4, art. no. 79, pp. 1-36, 2019, doi: 10.1145/3343440.
- [38] M. Buda, A. Maki, and M. A. Mazurowski, "A Systematic Study of the Class Imbalance Problem in Convolutional Neural Networks," *Neural Networks*, vol. 106, pp. 249-259, 2018, doi: 10.1016/j.neunet.2018.07.011.
- [39] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 687-719, 2009, doi: 10.1142/S0218001409007326.
- [40] A. Estabrooks, T. Jo, and N. Japkowicz, "A Multiple Resampling Method for Learning from Imbalanced Data Sets," *Computational Intelligence*, vol. 20, no. 1, pp. 18-36, 2004, doi: 10.1111/j.0824-7935.2004.t01-1-00228.x.
- [41] G. M. Weiss, "Mining with Rarity: A Unifying Framework," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 7-19, 2004, doi: 10.1145/1007730.1007734.
- [42] A. Moreo, A. Esuli, and F. Sebastiani, "Distributional Random Oversampling for Imbalanced Text Classification," in Proceedings of SIGIR, Pisa, Italy, 2016, pp. 805-808, doi: 10.1145/2911451.2914722.
- [43] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing*, Berlin, Germany: Springer, 2005, pp. 878-887, doi: 10.1007/11538059_91.
- [44] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem," in *Advances in Knowledge Discovery and Data Mining*, Berlin, Germany: Springer, 2009, pp. 475-482, doi: 10.1007/978-3-642-01307-2_43.
- [45] C. Elkan, "The Foundations of Cost-Sensitive Learning," in Proceedings of IJCAI, Seattle, WA, USA, 2001, pp. 973-978.

- [46] L. Akritidis and P. Bozanis, "Machine Learning Advances and Applications on Natural Language Processing (NLP)," *Electronics*, vol. 14, no. 16, art. no. 3282, 2025, doi: 10.3390/electronics14163282.
- [47] L. Akritidis and P. Bozanis, "Low-Dimensional Text Representations for Sentiment Analysis NLP Tasks," *SN Computer Science*, vol. 4, no. 5, art. no. 474, 2023, doi: 10.1007/s42979-023-01913-y.
- [48] L. Akritidis and P. Bozanis, "How Dimensionality Reduction Affects Sentiment Analysis NLP Tasks: An Experimental Study," in *Artificial Intelligence Applications and Innovations, IFIP AICT*, vol. 647, Cham, Switzerland: Springer, 2022, pp. 301-312, doi: 10.1007/978-3-031-08337-2_25.
- [49] L. Akritidis, A. Fevgas, M. Alamaniotis, and P. Bozanis, "Conditional Data Synthesis with Deep Generative Models for Imbalanced Dataset Oversampling," in *Proceedings of IEEE ICTAI*, Atlanta, GA, USA, 2023, pp. 444-451, doi: 10.1109/ICTAI59109.2023.00071.
- [50] L. Akritidis and P. Bozanis, "A Multi-Dimensional Survey on Learning from Imbalanced Data," in *Extended Selected Papers of the 14th International Conference on Information, Intelligence, Systems, and Applications, Lecture Notes in Networks and Systems*, vol. 1093, Cham, Switzerland: Springer, 2024, pp. 13-45, doi: 10.1007/978-3-031-67426-6_2.
- [51] M. Cox, Inside Airbnb: Get the Data. Accessed: Sep. 6, 2026. [Online]. Available: <https://insideairbnb.com/get-the-data/>
- [52] CardiffNLP, cardiffnlp/twitter-roberta-base-sentiment: Hugging Face Model Card. Accessed: Sep. 6, 2026. [Online]. Available: <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>

ΠΑΡΑΡΤΗΜΑ Α: Επιλεγμένος Πηγαίος Κώδικας

Στο παράρτημα παρατίθενται επιλεγμένα αποσπάσματα από τα notebooks που αντιστοιχούν στα βασικά βήματα υπερδιδασμοληψίας και αξιολόγησης. Ο κώδικας έχει περιοριστεί στα τμήματα που είναι απαραίτητα για την κατανόηση της πειραματικής ροής. Οι πλήρεις εκτελέσεις περιλαμβάνουν επιπλέον κελιά φόρτωσης δεδομένων, οπτικοποίησης και αποθήκευσης αποτελεσμάτων.

A.1 Random Oversampling στο πλήρες BERT training dataset

Η πρώτη εφαρμογή ROS αντιγράφει minority reviews μέχρι κάθε rating να φτάσει το πλήθος της Rating 5.

```
from imblearn.over_sampling import RandomOverSampler

X = train_df['Review'].values.reshape(-1, 1)
y = train_df['Rating'].values

ros = RandomOverSampler(random_state=42)
X_resampled, y_resampled = ros.fit_resample(X, y)

train_df_balanced = pd.DataFrame({
    'Review': X_resampled.flatten(),
    'Rating': y_resampled
})

print(train_df_balanced['Rating'].value_counts())
```

A.2 Stratified split και ROS μόνο στο training subset

Το δεύτερο BERT πείραμα πραγματοποιεί πρώτα stratified split και εφαρμόζει ROS μόνο στο training branch, αφήνοντας το validation set ανέγγιχτο.

```
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import RandomOverSampler

train_df_split, val_df_split = train_test_split(
    full_original_train_df,
    test_size=0.2,
    random_state=42,
    stratify=full_original_train_df['Rating']
)

ros = RandomOverSampler(random_state=42)
X_train = train_df_split[['Review']]
y_train = train_df_split['Rating']
X_ros, y_ros = ros.fit_resample(X_train, y_train)

train_ros_df = pd.DataFrame({
    'Review': X_ros['Review'].to_numpy(),
    'Rating': y_ros.to_numpy()
})
```

A.3 Εξαγωγή RoBERTa embeddings

Στο SMOTE πείραμα το classification head δεν χρησιμοποιείται ως τελικός classifier. Το RoBERTa λειτουργεί ως frozen feature extractor και για κάθε review χρησιμοποιείται ως embedding το διάνυσμα της πρώτης θέσης του 'last_hidden_state', σύμφωνα με το πραγματικό notebook.

```
from transformers import AutoTokenizer, AutoModel

embedding_model_name = 'cardiffnlp/twitter-roberta-base-sentiment'
embedding_tokenizer = AutoTokenizer.from_pretrained(embedding_model_name)
embedding_model = AutoModel.from_pretrained(embedding_model_name)
embedding_model.to(device)

with torch.no_grad():
    outputs = embedding_model(
        input_ids=input_ids,
        attention_mask=attention_mask
    )
```

```
# First-token representation from the last hidden layer
embeddings = outputs.last_hidden_state[:, 0, :].cpu().numpy()
```

A.4 SMOTE στον χώρο embeddings

Το SMOTE εφαρμόζεται αποκλειστικά στα training embeddings. Το validation feature matrix δεν συμμετέχει στο `fit\_resample`.

```
from imblearn.over_sampling import SMOTE

sm = SMOTE(random_state=42)
X_res, y_res = sm.fit_resample(
    train_embeddings,
    train_labels
)

print(pd.Series(train_labels).value_counts())
print(pd.Series(y_res).value_counts())
print(X_res.shape) # (39111, 768)
```

A.5 Logistic Regression πριν και μετά το SMOTE

Δύο classifiers με τις ίδιες βασικές ρυθμίσεις εκπαιδεύονται στα αρχικά και στα SMOTEd embeddings και αξιολογούνται στο ίδιο validation set.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score

clf_imbalanced = LogisticRegression(max_iter=1000, random_state=42)
clf_imbalanced.fit(train_embeddings, train_labels)
val_preds_imbalanced = clf_imbalanced.predict(val_embeddings)

clf_smote = LogisticRegression(max_iter=1000, random_state=42)
clf_smote.fit(X_res, y_res)
val_preds_smote = clf_smote.predict(val_embeddings)

print(classification_report(val_labels, val_preds_imbalanced))
print(classification_report(val_labels, val_preds_smote))
print(accuracy_score(val_labels, val_preds_imbalanced))
print(accuracy_score(val_labels, val_preds_smote))
```