ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

Τίτλος Πτυχιακής Εργασίας Thesis Title	Συγκριτική Μελέτη Αλγορίθμων Βαθιάς Μάθησης Comparative Study Of Deep Learning Algorithms
Ονοματεπώνυμο Φοιτητή	ΘΕΡΜΙΩΤΗΣ ΣΤΥΛΙΑΝΟΣ
Πατρώνυμο	ΠΑΝΑΓΙΩΤΗΣ
Αριθμός Μητρώου	Π/18244
Επιβλέπων	ΔΙΟΝΥΣΙΟΣ ΣΩΤΗΡΟΠΟΥΛΟΣ,ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ

Ημερομηνία Παράδοσης **Ιανουάριος 2025**

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Διονύσιο Σωτηρόπουλο για την εμπιστοσύνη που μου έδειξε δίνοντας μου την ευκαιρία να εκπονήσω αυτήν την πτυχιακή εργασία και την υποστήριξή του κατά τη διάρκεια της εκπόνησης της εργασίας. Επίσης θα ήθελα να ευχαριστήσω τους γονείς μου τόσο για την ψυχική όσο και για την οικονομική υποστήριξη.

Περίληψη

Η εργασία εξετάζει την ταξινόμηση συναισθημάτων σε κείμενα με τη χρήση προηγμένων μοντέλων μηχανικής μάθησης, όπως το BERT, το Ensemble of BERT Models και ένα Linear Layer. Χρησιμοποιήθηκαν τρία σύνολα δεδομένων σε μορφή CSV, περιέχοντας πολλαπλές ετικέτες συναισθημάτων. Τα δεδομένα επεξεργάστηκαν μέσω καθαρισμού, tokenization, padding και διαχωρισμού σε σύνολα εκπαίδευσης και αξιολόγησης.
Η μεθοδολογία περιλάμβανε:

- Εκπαίδευση πολλαπλών μοντέλων BERT με στόχο τη βελτίωση της ακρίβειας.
- Υπολογισμό μετρικών, όπως AUROC και F1-score, για την αξιολόγηση των μοντέλων.
- Χρήση τεχνικής Ensemble για τη συνδυαστική ενίσχυση της απόδοσης των μοντέλων.

Τα αποτελέσματα δείχνουν ακρίβεια που αγγίζει το **0.93 AUROC** για συναισθήματα όπως "admiration", "gratitude", και "love", με σημαντική βελτίωση στην ταξινόμηση έναντι παραδοσιακών μεθόδων. Παρόλο που υπήρξαν περιορισμοί σε συναισθήματα όπως "pride" και "relief", η χρήση της προσέγγισης Ensemble μείωσε τις επιδόσεις των χαμηλότερων κλάσεων.
.

Λέξεις Κλειδιά:

- Ανάλυση Συναισθημάτων
- BERT
- Ensemble Models
- Μηχανική Μάθηση
- Ταξινόμηση Κειμένων
- AUROC
- F1-Score
- Προεπεξεργασία Δεδομένων
- Tokenization
- Padding

Abstract

This study examines sentiment classification in text using advanced machine learning models such as BERT, Ensemble of BERT Models, and a Linear Layer. Three CSV datasets containing multi-label sentiment annotations were utilized. The data were processed through cleaning, tokenization, padding, and splitting into training and evaluation sets.
The methodology included:

- Training multiple BERT models to improve accuracy.
- Computing metrics such as AUROC and F1-score for model evaluation.
- Applying an Ensemble technique to boost model performance collectively.

The results demonstrate an accuracy reaching **0.93 AUROC** for emotions like "admiration," "gratitude," and "love," significantly outperforming traditional methods. While limitations were noted for emotions such as "pride" and "relief," the Ensemble approach mitigated the lower performance in these categories.

Key Words:

- *Sentiment Analysis*
- *BERT*
- *Ensemble Models*
- *Machine Learning*
- *Text Classification*
- *AUROC*
- *F1-Score*
- *Data Preprocessing*
- *Tokenization*
- *Padding*

Αφιερώσεις

Αφιερωμένο στους γονείς μου Παναγιώτη και Γεωργία
και στα αδέρφια μου Στράτο και Μαρία!

Πίνακας Περιεχομένων

Ευχαριστίες 3

Περίληψη (Abstract) 4

Αφιερώσεις 6

Πίνακας Περιεχομένων 7

Κατάλογος Εικόνων 9

Κατάλογος Διαγραμμάτων 10

Εισαγωγή 11

1. Σχετικά Έργα.....12
2. Θεωρητικό Υπόβαθρο.....13
   - 2.1 Τι είναι η ανάλυση συναισθημάτων.....13
   - 2.2 Machine Learning vs Deep Learning 13
   - 2.3 Τεχνολογίες που εφαρμόστηκαν.....14
   - 2.4 Μοντέλο Bert.....19
3. Μεθοδολογία.....21
   - 3.1 Περιγραφή δεδομένων.....21
   - 3.2 Επεξεργασία και καθαρισμός.....22
   - 3.3 Μοντέλα και αρχιτεκτονικές.....23
   - 3.4 Υπολογιστική υποδομή.....24
4. Τεχνολογία.....25
   - 4.1 Python.....25
   - 4.2 Google Colab.....25
   - 4.3 TensorFlow.....25
   - 4.4 PyTorch.....25
5. Πειραματική Διαδικασία.....26
   - 5.1 Εφαρμογή των μοντέλων.....26
   - 5.2 Εκπαίδευση και αξιολόγηση.....28
     - 5.2.1 Εκπαίδευση των μοντέλων.....28
     - 5.2.2 Σημασία των Epochs στην εκπαίδευση των λοντέλων.....28
     - 5.2.3 Epochs σε μοντέλα Bert και Transformer.....31
     - 5.2.4 Visualization of Training Over Epochs (Οπτικοποίηση της εκπαίδευσης κατά τη διάρκεια των εποχών).....32
     - 5.2.5 Αξιολόγηση των μοντέλων.....32
   - 5.3 Συγκριση με αλλες μεθόδους.....34
6. Κώδικας Εκπαίδευσης Μοντέλου.....35
   - 6.1 Βιβλιοθήκες.....35
   - 6.2 Data Preprocessing.....36
     - 6.2.1 Data Loading.....36
     - 6.2.2 Tokenization.....36
     - 6.2.3 Padding/Truncation.....36
     - 6.2.4 Splitting.....37

6.2.5 Υπολογισμός συναισθηματικής ισορροπίας……..37
6.2.6 Ταξινόμηση δεδομένων.....38
6.2.7 Αφαίρεση και διόρθωση δεδομένων.....38
6.2.8 Επιστροφή επεξεργασμένων δεδομένων.....38
6.2.9 Data Preparation with Custom DataModule.....38
6.3 Training.....39
6.3.1 Διαμόρφωση εκπαίδευσης.....39
6.3.2 Αρχικοποιηση μοντέλου.....39
6.3.3 Ελεγχος μοντέλου και έγκαιρη διακοπή.....40
6.3.4 TensorBoard Logger.....40
6.3.5 Αρχικοποίηση εκπαίδευσης.....40
6.3.6 Εκπαίδευση μοντέλου.....41
7. Αποτελέσματα.....42
7.1 Auroc……...42
7.2 F1-Score……...43
7.3 Σύγκριση των μοντέλων με βάση τα μοντέλα.....44
8. Συμπεράσματα.....45
9. Βιβλιογραφία.....46

Κατάλογος Εικόνων

Εικόνα 1. Machine Learning vs Deep Learning……14
Εικονα 2. Self Attention Mechanism.....16
Εικόνα 3. Encoder-Decoder Architecture.....18
Εικονα 3. Bert (Biderrectional Encoder Representations from Transformers) ……23
Εικόνα 4. Data Preparation Stages…….26

Κατάλογος Διαγραμμάτων

Διάγραμμα 1. Model Train Stages.....27
Διάγραμμα 2. Training and Validation Loss Fold 1.....29
Διάγραμμα 3. Training and Validation Loss Fold 2.....29
Διάγραμμα 4. Training and Validation Loss Fold 3.....30
Διάγραμμα 5. Training and Validation Loss Fold 4.....30
Διάγραμμα 6. Training and Validation Loss Fold 5.....31
Διάγραμμα 7. Loss Convergence Chart (Διάγραμμα Συγκλισης Απωλειών) 33
Διάγραμμα 8. Evaluation Metrics Chart (Διάγραμμα Μετρήσεων Αξιόλογησης) 33
Διάγραμμα 2. Αποτελέσματα Auroc.....42
Διάγραμμα 2. Αποτελέσματα F1-Score.....43

Εισαγωγή

Η ανάλυση συναισθήματος είναι ένα κρίσιμο εργαλείο στην επεξεργασία φυσικής γλώσσας (NLP), που επιτρέπει την κατανόηση και ταξινόμηση των ανθρώπινων συναισθημάτων που εκφράζονται σε κείμενο. Οι εφαρμογές της εκτείνονται από την ανάλυση των ανατροφοδοτήσεων των χρηστών και τη βελτίωση της εμπειρίας των πελατών έως την υποστήριξη διαδικασιών λήψης αποφάσεων σε διάφορους τομείς. Σκοπός της παρούσας έρευνας είναι ο εντοπισμός και η κατανόηση των συναισθημάτων σε κείμενο που προέρχεται από σενάρια του πραγματικού κόσμου, δίνοντας έμφαση στον αντίκτυπό τους στις αλληλεπιδράσεις των χρηστών και στη συνολική διαχείριση των επιδόσεων. Αξιοποιώντας τις πλέον σύγχρονες μεθοδολογίες NLP, ιδίως το μοντέλο BERT, καθώς και τεχνικές μηχανικής μάθησης συνόλου, η παρούσα μελέτη αποσκοπεί στην ενίσχυση της ακρίβειας και της αξιοπιστίας των συστημάτων ταξινόμησης συναισθημάτων. Η εργασία αυτή διερευνά τις εφαρμογές της ανάλυσης συναισθήματος σε διάφορους τομείς, όπως η εξυπηρέτηση πελατών, η παρακολούθηση των μέσων κοινωνικής δικτύωσης και η επιχειρηματική ευφυΐα. Η ενσωμάτωση προηγμένων αλγορίθμων όπως ο BERT και τεχνικών ensemble επιτρέπει βελτιωμένη πρόβλεψη συναισθήματος και συναισθηματική κατανόηση σε περιβάλλοντα όπου οι παραδοσιακές μέθοδοι δεν επαρκούν. Η συμβολή της παρούσας έρευνας έγκειται στην ικανότητά της να γεφυρώνει τα κενά στην ανάλυση συναισθήματος, επιτυγχάνοντας υψηλότερα ποσοστά ακρίβειας και αντιμετωπίζοντας τις προκλήσεις που σχετίζονται με τα ανισόρροπα σύνολα δεδομένων και τις λεπτές συναισθηματικές αποχρώσεις. Με τον τρόπο αυτό, όχι μόνο προάγει τις δυνατότητες της ταξινόμησης συναισθήματος, αλλά παρέχει επίσης πρακτικές ιδέες για την αξιοποίηση της μηχανικής μάθησης και του NLP σε εφαρμογές του πραγματικού κόσμου.

‘

1. Σχετικά Έργα

Πολυάριθμες μελέτες έχουν διερευνήσει τον τομέα της ανάλυσης συναισθήματος, εστιάζοντας στην ταξινόμηση των συναισθημάτων και την εξαγωγή νοήματος από δεδομένα κειμένου. Οι πρώτες προσεγγίσεις βασίστηκαν σε μεγάλο βαθμό σε συστήματα βασισμένα σε κανόνες και στατιστικές μεθόδους, όπως οι Naive Bayes και Support Vector Machines, οι οποίες παρείχαν μια θεμελιώδη κατανόηση της ταξινόμησης συναισθήματος, αλλά δυσκολεύονταν με την επεκτασιμότητα και τις αποχρώσεις του πλαισίου.Η έλευση της βαθιάς μάθησης έφερε επανάσταση στον τομέα, εισάγοντας αρχιτεκτονικές νευρωνικών δικτύων ικανές να μαθαίνουν ιεραρχικές αναπαραστάσεις κειμένου. Τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) και τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM) επέδειξαν βελτιωμένη απόδοση στην καταγραφή των διαδοχικών εξαρτήσεων, οι οποίες είναι απαραίτητες για την κατανόηση του συναισθήματος σε σύνθετες προτάσεις.Πιο πρόσφατα, τα μοντέλα που βασίζονται σε μετασχηματιστές, όπως το BERT (Bidirectional Encoder Representations from Transformers), έθεσαν ένα νέο σημείο αναφοράς στο NLP. Η αμφίδρομη εκπαίδευση του BERT του δίνει τη δυνατότητα να κατανοεί τα συμφραζόμενα πιο αποτελεσματικά από τα προηγούμενα μονόδρομα μοντέλα, με αποτέλεσμα σημαντικές βελτιώσεις σε εργασίες ανάλυσης συναισθήματος. Μελέτες έχουν δείξει ότι η λεπτομερής ρύθμιση του BERT σε συγκεκριμένα σύνολα δεδομένων μπορεί να αποδώσει κορυφαίες επιδόσεις στην ταξινόμηση συναισθήματος. Συνδυάζοντας προβλέψεις από πολλαπλά μοντέλα, οι μέθοδοι ensemble μετριάζουν τις αδυναμίες των μεμονωμένων προσεγγίσεων και ενισχύουν την ανθεκτικότητα.

Καθώς τα μοντέλα ανάλυσης συναισθήματος γίνονται όλο και πιο πολύπλοκα, η κατανόηση της διαδικασίας λήψης αποφάσεων καθίσταται απαραίτητη. Τεχνικές όπως οι SHAP (SHapley Additive exPlanations) και LIME (Local Interpretable Model-agnostic Explanations) έχουν εισαχθεί για να παρέχουν πληροφορίες σχετικά με τον τρόπο με τον οποίο τα μοντέλα κάνουν προβλέψεις, εξασφαλίζοντας διαφάνεια και εμπιστοσύνη στα συστήματα τεχνητής νοημοσύνης.

Η επέκταση της ανάλυσης συναισθήματος για την υποστήριξη πολλαπλών γλωσσών ήταν μια άλλη σημαντική πρόοδος. Μοντέλα όπως το mBERT και το XLM-R έχουν σχεδιαστεί για να χειρίζονται πολύγλωσσα σύνολα δεδομένων, επιτρέποντας την ταξινόμηση συναισθήματος σε διάφορες γλώσσες και βελτιώνοντας τη γενίκευση του μοντέλου σε παγκόσμιες εφαρμογές.Η παρούσα μελέτη βασίζεται σε αυτά τα θεμέλια, χρησιμοποιώντας BERT, τεχνικές συνόλου και προηγμένη προεπεξεργασία για την επίτευξη υψηλής ακρίβειας σε εργασίες ταξινόμησης συναισθήματος, εστιάζοντας ειδικά σε σύνολα δεδομένων βασισμένα σε κείμενο που χαρακτηρίζονται από συναισθήματα.

2. Θεωρητικό Υπόβαθρο

2.1 Τι είναι η ανάλυση συναισθημάτων

Η ανάλυση συναισθήματος, γνωστή και ως εξόρυξη γνώμης,σύμφωνα με [1] είναι η διαδικασία εντοπισμού, εξαγωγής και ταξινόμησης συναισθημάτων, απόψεων και στάσεων που εκφράζονται σε δεδομένα κειμένου. Πρόκειται για ένα υποσύνολο της επεξεργασίας φυσικής γλώσσας (NLP) που αποσκοπεί στον προσδιορισμό του κατά πόσον ένα δεδομένο κείμενο μεταφέρει ένα θετικό, αρνητικό ή ουδέτερο συναίσθημα.Στον πυρήνα της, η ανάλυση συναισθήματος περιλαμβάνει γλωσσολογικές και στατιστικές τεχνικές για την ανάλυση δεδομένων κειμένου. Οι πρώτες προσεγγίσεις επικεντρώθηκαν σε συστήματα βασισμένα σε κανόνες και λεξικά, όπου χρησιμοποιήθηκαν προκαθορισμένα λεξικά με λέξεις γεμάτες συναισθήματα για τον προσδιορισμό της πολικότητας. Ωστόσο, αυτές οι μέθοδοι συχνά δεν μπορούσαν να συλλάβουν τις λεπτές αποχρώσεις του πλαισίου και τη δυναμική φύση της γλώσσας. η σύγχρονη ανάλυση συναισθήματος αξιοποιεί τεχνικές μηχανικής μάθησης και βαθιάς μάθησης, επιτρέποντας στα συστήματα να μαθαίνουν από τα δεδομένα και να προσαρμόζονται σε ποικίλες εισόδους κειμένου. Προηγμένοι αλγόριθμοι όπως το BERT και τα μοντέλα συνόλου έχουν αποδειχθεί ιδιαίτερα αποτελεσματικά, επιτρέποντας την πιο διαφοροποιημένη κατανόηση και πρόβλεψη των συναισθημάτων σε διάφορους τομείς, όπως τα μέσα κοινωνικής δικτύωσης, τα σχόλια των πελατών και οι κριτικές προϊόντων.Η σημασία της ανάλυσης συναισθήματος έγκειται στις ευρείες εφαρμογές της. Οι επιχειρήσεις τη χρησιμοποιούν για τη μέτρηση της ικανοποίησης των πελατών και των τάσεων της αγοράς, ενώ οι κοινωνικοί επιστήμονες τη χρησιμοποιούν για τη μελέτη της κοινής γνώμης και της κοινωνικής δυναμικής. Η ευελιξία της και οι δυνατότητες για αξιοποιήσιμες γνώσεις καθιστούν την ανάλυση συναισθήματος ακρογωνιαίο λίθο της λήψης αποφάσεων βάσει δεδομένων.

2.2 Machine Learning vs Deep Learning

Machine Learning

Η Μηχανική Μάθηση (ML) είναι ένα υποσύνολο της τεχνητής νοημοσύνης που επιτρέπει στους υπολογιστές να μαθαίνουν από δεδομένα και να κάνουν προβλέψεις ή αποφάσεις χωρίς να είναι ρητά προγραμματισμένοι[2]. Στο πλαίσιο της ανάλυσης συναισθήματος, οι αλγόριθμοι μηχανικής μάθησης ταξινομούν κείμενο σε κατηγορίες όπως θετικό, αρνητικό ή ουδέτερο συναίσθημα.Μερικά μοντέλα μηχανικής μάθησης για ανάλυση συναισθήματος ειναι Naive Bayes (NB), k-Nearest Neighbors (k-NN).

Deep Learning

Η βαθιά μάθηση (DL), ένας κλάδος της μηχανικής μάθησης, χρησιμοποιεί νευρωνικά δίκτυα με πολλαπλά επίπεδα για να μοντελοποιήσει σύνθετα μοτίβα σε δεδομένα. Τα μοντέλα βαθιάς μάθησης μαθαίνουν αυτόματα χαρακτηριστικά από ακατέργαστα δεδομένα κειμένου, εξαλείφοντας την ανάγκη για χειροκίνητη μηχανική των χαρακτηριστικών[2].Βασικά μοντελα βαθιάς μάθησης για ανάλυση συναισθήματος ειναι Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs), Bidirectional LSTMs (BiLSTM) και Convolutional Neural Networks (CNNs).

Aspect	Machine Learning	Deep Learning
Feature Engineering	Manual feature extraction (TF-IDF, Bag of Words)	Automatic feature learning from raw data
Model Complexity	Simpler, fewer parameters	Complex, deep architectures
Performance	Good on small datasets	Superior on large, complex datasets
Contextual Learning	Limited context handling	Captures sequential and contextual dependencies
Training Time	Faster, less computationally intensive	Requires high computational resources (GPUs/TPUs)

2.3 **Τεχνολογίες που εφαρμόστηκαν**

Για την αποτελεσματική εφαρμογή αυτών των τεχνικών, η παρούσα έρευνα χρησιμοποιεί εργαλεία αιχμής, όπως το μοντέλο BERT, ενσωματωμένο σε βιβλιοθήκες Python όπως οι PyTorch, TensorFlow και scikit-learn. Αυτές οι τεχνολογίες διευκολύνουν την προεπεξεργασία, την εκπαίδευση και την αξιολόγηση συνόλων δεδομένων κειμένου μεγάλης κλίμακας, εξασφαλίζοντας υψηλές επιδόσεις και προσαρμοστικότητα για εργασίες ταξινόμησης συναισθήματος.Το υπάρχον σύνολο εργασιών υπογραμμίζει τη σημασία της αξιοποίησης προηγμένων τεχνικών NLP και μεθόδων συνόλου για να διευρυνθούν τα όρια της ανάλυσης συναισθήματος, ανοίγοντας το δρόμο για πιο ακριβείς και με επίγνωση του πλαισίου εφαρμογές.

Αρχιτεκτονικές Μετασχηματιστών(Transformer)

Οι αρχιτεκτονικές μετασχηματιστών αποτελούν μια πρωτοποριακή εξέλιξη στην επεξεργασία φυσικής γλώσσας, αλλάζοντας ριζικά τον τρόπο με τον οποίο οι μηχανές κατανοούν και παράγουν κείμενο. Σε αντίθεση με τα παραδοσιακά μοντέλα ακολουθίας, όπως τα RNN ή τα LSTM, οι μετασχηματιστές λειτουργούν χρησιμοποιώντας μηχανισμούς αυτοπροσοχής, επιτρέποντάς τους να εξετάζουν όλες τις λέξεις εισόδου ταυτόχρονα και να συλλαμβάνουν σχέσεις ανεξάρτητα από τη θέση τους στην ακολουθία:

Self-Attention Mechanism:
Αυτός ο μηχανισμός υπολογίζει τη σημασία κάθε λέξης σε σχέση με κάθε άλλη λέξη σε μια πρόταση σύμφωνα με το[3]. Με τον τρόπο αυτό, καταγράφει τόσο τις τοπικές όσο και τις παγκόσμιες εξαρτήσεις, επιτρέποντας στα μοντέλα να κατανοούν αποτελεσματικότερα τα συμφραζόμενα. Ο πυρήνας των Transformers είναι ο μηχανισμός αυτοπροσοχής (Self-Attention), ο οποίος υπολογίζει τη σημασία κάθε λέξης σε σχέση με τις υπόλοιπες μέσα σε μια ακολουθία.

Υπολογισμός προσοχής: Για κάθε λέξη, υπολογίζονται τρεις διανύσματα:

- Q (Query): Αντιπροσωπεύει την ερώτηση.
- K (Key): Αντιπροσωπεύει το περιεχόμενο.
- V (Value): Αντιπροσωπεύει την πληροφορία.

Ο τύπος υπολογισμού είναι:

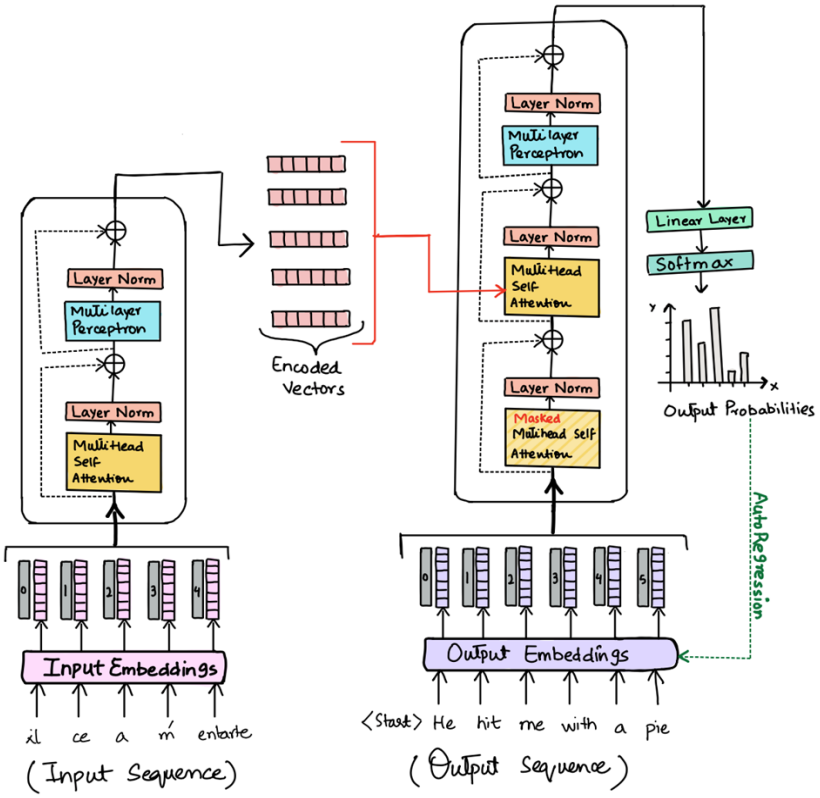
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

όπου

- $Q \in R^{n \times dk}$: Οι ερωτήσεις.
- $K \in R^{n \times dk}K$: Τα κλειδιά.
- $V \in R^{n \times dv}$: Οι τιμές.
- dk : Ο αριθμός διαστάσεων στα KKK και QQQ.

Η συνάρτηση softmax εξασφαλίζει ότι οι τιμές προσοχής είναι μεταξύ 0 και 1:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}}$$



Positional Encoding:
Δεδομένου ότι οι μετασχηματιστές επεξεργάζονται την είσοδο ως σύνολο και όχι διαδοχικά, η κωδικοποίηση θέσης χρησιμοποιείται για την εισαγωγή πληροφοριών σχετικά με τη σειρά των λέξεων στην ακολουθία. Για να καταγράψει τη θέση των λέξεων, το BERT χρησιμοποιεί ημιτονοειδείς συναρτήσεις:

$$PE_{(pos,\,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$
$$PE_{(pos,\,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

Όπου:

- pos: Θέση της λέξης.
- dmodel: Διαστάσεις του μοντέλου.

Multi-Head Attention:
Αυτό το χαρακτηριστικό επιτρέπει στο μοντέλο να εστιάζει ταυτόχρονα σε διαφορετικά μέρη της εισόδου, παρέχοντας μια πλουσιότερη αναπαράσταση του κειμένου. Οι Transformers χρησιμοποιούν πολλαπλές κεφαλές προσοχής για να συλλάβουν διαφορετικά μοτίβα από τις λέξεις. Ο τύπος για τις πολλαπλές κεφαλές προσοχής(Multi-Head Attention) είναι:

MultiHead(Q,K,V)=Concat(head1,…,headh)W^o

Όπου

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

W^o = Ο πίνακας εξόδου

Wi^Q,Wi^K,Wi^V: Πίνακες βαρών για κάθε κεφαλή.

Feedforward Layers:
Αυτά τα πλήρως συνδεδεμένα στρώματα επεξεργάζονται τις εξόδους του μηχανισμού προσοχής, επιτρέποντας στο μοντέλο να μαθαίνει σύνθετα μοτίβα.

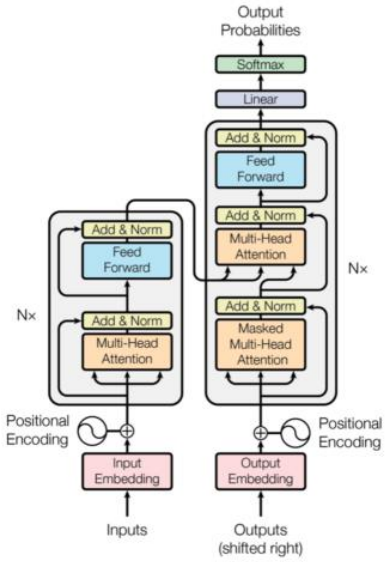
Layer Normalization:
Για τη σταθεροποίηση της εκπαίδευσης και την ενίσχυση της σύγκλισης, οι μετασχηματιστές χρησιμοποιούν τεχνικές κανονικοποίησης στρώματος. Μετά από κάθε βήμα προσοχής, εφαρμόζεται κανονικοποίηση για να σταθεροποιηθεί η εκπαίδευση:

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sigma + \varepsilon} \cdot \gamma + \beta$$

Όπου:

- μ: Ο μέσος όρος των στοιχείων του xxx.
- σ: Η τυπική απόκλιση.
- γ,β: Εκπαιδεύσιμες παράμετροι.

Encoder-Decoder Architecture:
Η αρχιτεκτονική του μοντέλου βασίζεται στην δομή κωδικοποιητή – αποκωδικοποιητή (encoder – decoder). Ο κωδικοποιητής αντιστοιχεί την ακολουθία εισόδου που βρίσκεται στην μορφή συμβολικής αναπαράστασης x (x1 , x2 , … , xn) σε μία μορφή συνεχούς αναπαράστασης z (z1 , z2 , … , zn). Ο αποκωδικοποιητής έχοντας σαν είσοδο την ακολουθία z, παράγει μία ακολουθία εξόδου y (y1 , y2 , … , yn) για ένα σύμβολο την φορά. Σε κάθε βήμα το μοντέλο είναι αυτόαναδρομικό, δηλαδή χρησιμοποιεί τα σύμβολα που έχουν δημιουργηθεί ως επιπλέον είσοδο όταν δημιουργεί το κείμενο.Εργασίες όπως η μετάφραση, οι μετασχηματιστές χρησιμοποιούν έναν κωδικοποιητή για την επεξεργασία κειμένου εισόδου και έναν αποκωδικοποιητή για τη δημιουργία κειμένου εξόδου, καθιστώντας τους ιδιαίτερα ευέλικτους.



Το μοντέλο BERT, ένα αξιοσημείωτο παράδειγμα αρχιτεκτονικής μετασχηματιστή, αξιοποιεί μια αμφίδρομη προσέγγιση στην προ-εκπαίδευση, συλλαμβάνοντας το περιεχόμενο τόσο από τις προηγούμενες όσο και από τις επόμενες λέξεις. Η επιτυχία του έχει θέσει νέα πρότυπα για εργασίες NLP, από την ανάλυση συναισθήματος έως την απάντηση ερωτήσεων.

2.4 **Μοντέλο Bert**

Ο BERT (Bidirectional Encoder Representations from Transformers) είναι ένα μοντέλο βασισμένο σε μετασχηματιστές που έχει σχεδιαστεί για να κατανοεί το περιεχόμενο ενός κειμένου και προς τις δύο κατευθύνσεις. Σε αντίθεση με τα παραδοσιακά μοντέλα που επεξεργάζονται ακολουθίες μονόδρομα, το BERT χρησιμοποιεί αμφίδρομη εκπαίδευση, επιτρέποντάς του να συλλαμβάνει ταυτόχρονα το περιεχόμενο τόσο από τις προηγούμενες όσο και από τις επόμενες λέξεις. Αυτή η ικανότητα καθιστά το BERT εξαιρετικά αποτελεσματικό για εργασίες που απαιτούν λεπτή γλωσσική κατανόηση, όπως η ανάλυση συναισθήματος, η απάντηση ερωτήσεων και η αναγνώριση ονομαστικών οντοτήτων[4.2].Παρακατω ειναι τα βασικά χαρακτηριστικά του BERT:

Bidirectional Context Understanding:
Το BERT διαβάζει κείμενο τόσο από αριστερά προς τα δεξιά όσο και από δεξιά προς τα αριστερά, βελτιώνοντας την κατανόηση διφορούμενων ή σύνθετων προτάσεων. Η συνάρτηση της αμφίδρομης κατανόησης περιγράφεται ως:

Context Representation(t)=AttentionLeft-to-Right(t)+AttentionRight-to-Left(t)

Όπου:

- AttentionLeft-to-Right(t): Το πλαίσιο από όλες τις λέξεις που βρίσκονται πριν από τη λέξη t.
- AttentionRight-to-Left(t): Το πλαίσιο από όλες τις λέξεις που βρίσκονται μετά τη λέξη t.

Masked Language Modeling (MLM):
Το BERT εκπαιδεύεται να προβλέπει τις καλυμμένες λέξεις σε μια πρόταση, βελτιώνοντας τη γλωσσική της κατανόηση[4.3]. Η συνάρτηση απώλειας είναι:

$$\mathcal{L}_{MLM} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

Όπου:

- yi: Η πραγματική πιθανότητα της καλυμμένης λέξης.
- Y^i: Η προβλεπόμενη πιθανότητα από το μοντέλο.

Next Sentence Prediction (NSP):
Αυτή η εργασία εκπαιδεύει το BERT να προβλέπει αν μια πρόταση ακολουθεί λογικά μια άλλη, βελτιώνοντας την κατανόηση της συνοχής[4.3].Η συνάρτηση απώλειας είναι:

$$\mathcal{L}_{NSP} = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i))$$

Όπου:

- yi: Ετικέτα (1 για σχετική, 0 για μη σχετική).
- y^i: Προβλεπόμενη πιθανότητα.

Συνολική Συνάρτηση Απώλειας:
Η συνολική συνάρτηση απώλειας του BERT είναι:

LBERT=LMLM+LNSP

Βελτιστοποίηση:
Το BERT χρησιμοποιεί τον Adam optimizer με warm-up βήματα και learning rate scheduling:

$$l_r = \eta \cdot \min\left(\frac{1}{\sqrt{t}}, \frac{t}{t_{\text{warmup}}^3}\right)$$

Fine-Tuning for Downstream Tasks:
Το BERT μπορεί να προσαρμοστεί σε συγκεκριμένες εργασίες με τη λεπτομερή ρύθμισή της σε σύνολα δεδομένων με ετικέτες, καθιστώντας την εξαιρετικά ευέλικτη.

Scalability:
Με πολλαπλές εκδόσεις όπως το BERT-Base και το BERT-Large, το μοντέλο κλιμακώνεται σε πολυπλοκότητα για να χειρίζεται ποικίλες προκλήσεις NLP.

3. Μεθοδολογία

3.1 Περιγραφή δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στην παρούσα έρευνα περιλαμβάνει τρία ξεχωριστά αρχεία CSV που περιέχουν δεδομένα κειμένου κατηγοριοποιημένα με ετικέτες συναισθημάτων. Κάθε αρχείο αντιπροσωπεύει μια διαφορετική πηγή ή τύπο δεδομένων κειμένου, εξασφαλίζοντας ποικιλομορφία και ευρωστία για σκοπούς εκπαίδευσης και αξιολόγησης.Τα βασικά χαρακτηριστικά του συνόλου δεδομένων είναι:

Emotion tags
Κάθε περίπτωση κειμένου επισημαίνεται με μία από τις προκαθορισμένες κατηγορίες συναισθήματος: θετικό, αρνητικό ή ουδέτερο.

Text length:
Τα δείγματα κειμένου ποικίλλουν σε μήκος, που κυμαίνεται από σύντομες φράσεις έως μεγάλες παραγράφους, απαιτώντας τεχνικές προεπεξεργασίας, όπως tokenization και padding για την τυποποίηση των μεγεθών εισόδου.

Distribution of categories:
Η ανάλυση της κατανομής των κλάσεων δείχνει μια μικρή ανισορροπία, με τα θετικά συναισθήματα να είναι τα πιο συχνά. Για την αντιμετώπιση αυτής της ανισορροπίας χρησιμοποιούνται τεχνικές όπως η υπερδειγματοληψία ή η στάθμιση κλάσεων.

Pre-treatment needs:
Τα ακατέργαστα δεδομένα κειμένου περιέχουν θόρυβο, όπως ειδικούς χαρακτήρες, διευθύνσεις URL και emojis, γεγονός που απαιτεί βήματα καθαρισμού και κανονικοποίησης.

Διαχωρισμός εκπαίδευσης και αξιολόγησης:
Το σύνολο δεδομένων χωρίζεται σε υποσύνολα εκπαίδευσης (80%) και δοκιμής (20%), ώστε να διασφαλιστεί ότι η απόδοση του μοντέλου αξιολογείται σε αθέατα δεδομένα.

Με την ενσωμάτωση αυτών των χαρακτηριστικών στον αγωγό εκπαίδευσης του μοντέλου, η παρούσα έρευνα αποσκοπεί στην επίτευξη ενός ισχυρού και γενικευμένου συστήματος ταξινόμησης συναισθημάτων, το οποίο είναι ικανό να χειρίζεται δεδομένα κειμένου του πραγματικού κόσμου

3.2 **Επεξεργασία και καθαρισμός**

Για την προετοιμασία του συνόλου δεδομένων για την εκπαίδευση του μοντέλου, εφαρμόστηκαν διάφορα βήματα προεπεξεργασίας και καθαρισμού:

Text Cleaning:
Αφαίρεση ειδικών χαρακτήρων, στίξης και υπερβολικού κενού χώρου και απομάκρυνση των διευθύνσεων URL, των διευθύνσεων ηλεκτρονικού ταχυδρομείου και των άσχετων συμβόλων (π.χ. emojis).

Lowercasing:
Μετατροπή όλου του κειμένου σε πεζά γράμματα για να εξασφαλιστεί η συνέπεια στην αντιστοίχιση των συμβόλων.

Tokenization:
Διαχωρισμός του κειμένου σε μεμονωμένα tokens (λέξεις ή υπολέξεις) με τη χρήση του tokenizer της BERT.

Stopword Removal:
Απομάκρυνση των συχνά εμφανιζόμενων αλλά μη πληροφοριακών λέξεων (π.χ. «το», «και», «είναι»), εάν είναι εφαρμόσιμες στην εργασία.

Padding and Truncation:
Τυποποίηση του μήκους των ακολουθιών με αποκοπή μεγάλων κειμένων και συμπλήρωση μικρότερων κειμένων σε σταθερό μήκος.

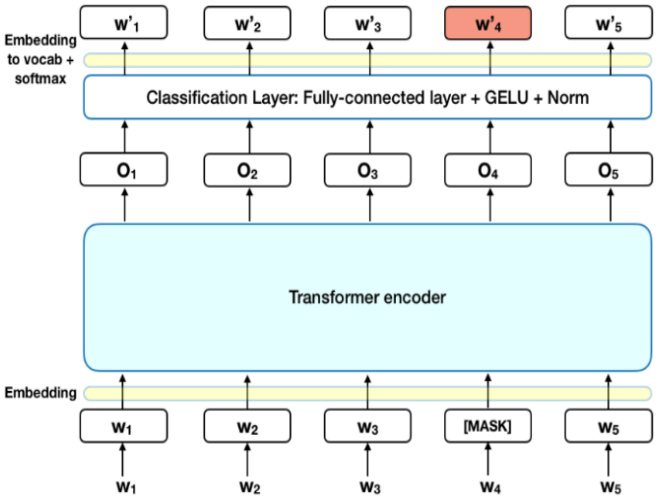
Handling Imbalanced Classes:
Υπερδειγματοληψία υποεκπροσωπούμενων κλάσεων ή εφαρμογή βαρών κλάσεων κατά τη διάρκεια της εκπαίδευσης για την αντιμετώπιση της ανισορροπίας κλάσεων.

Αυτά τα βήματα προεπεξεργασίας διασφαλίζουν ότι τα δεδομένα κειμένου είναι καθαρά, συνεπή και έτοιμα για αποτελεσματική εισαγωγή στο μοντέλο BERT, βελτιστοποιώντας την απόδοση και την ακρίβειά του κατά την εκπαίδευση και την αξιολόγηση

3.3 **Μοντέλα και αρχιτεκτονικές**

Η μελέτη αυτή αξιοποιεί προηγμένα μοντέλα και αρχιτεκτονικές για την ανάλυση συναισθήματος, όπως:

BERT (Bidirectional Encoder Representations from Transformers):
Η αμφίδρομη προσέγγιση του BERT του επιτρέπει να κατανοεί τα συμφραζόμενα σε μια πρόταση αναλύοντας τις σχέσεις μεταξύ των λέξεων και προς τις δύο κατευθύνσεις. Έχει προ-εκπαιδευτεί χρησιμοποιώντας εργασίες όπως η μοντελοποίηση καλυμμένης γλώσσας και η πρόβλεψη επόμενης πρότασης, καθιστώντας το ιδιαίτερα ευέλικτο για εργασίες NLP.



Ensemble Models:
Οι τεχνικές Ensemble συνδυάζουν τις προβλέψεις πολλαπλών μοντέλων για τη βελτίωση της συνολικής ακρίβειας και ευρωστίας. Η παρούσα μελέτη χρησιμοποιεί ένα σύνολο μοντέλων με βάση το BERT για την αντιμετώπιση προκλήσεων όπως η ανισορροπία τάξεων και οι λεπτές συναισθηματικές αποχρώσεις.

Linear Layer Architectures:
Τα γραμμικά στρώματα χρησιμοποιούνται ως τελικό βήμα ταξινόμησης, αντιστοιχίζοντας την έξοδο των μοντέλων BERT ή του συνόλου σε συγκεκριμένες κατηγορίες συναισθήματος. Αυτό το στρώμα εξασφαλίζει απλότητα και αποτελεσματικότητα στην τελική εργασία ταξινόμησης.

Fine-Tuning Techniques:
Τα γραμμικά στρώματα χρησιμοποιούνται ως τελικό βήμα ταξινόμησης, αντιστοιχίζοντας την έξοδο των μοντέλων BERT ή του συνόλου σε συγκεκριμένες κατηγορίες συναισθήματος. Αυτό το στρώμα εξασφαλίζει απλότητα και αποτελεσματικότητα στην τελική εργασία ταξινόμησης.

3.4 **Υπολογιστική υποδομή**

Η υλοποίηση αυτής της έρευνας απαιτεί μια ισχυρή υπολογιστική υποδομή για την εκπαίδευση και την αξιολόγηση μοντέλων μεγάλης κλίμακας όπως το BERT. Τα βασικά στοιχεία της υποδομής περιλαμβάνουν ποροι υλικού , μονάδες επεργασίας γραφικών(GPU) , κεντρικές μονάδες επεξεργασίας(CPUs) , μνημη RAM , Cloud Computing Service αξιοποιούνται πλατφόρμες όπως οι AWS, Google Cloud και Azure για κλιμάκωση και πρόσβαση σε GPU τελευταίας τεχνολογίας.Υπηρεσίες όπως το Google Colab και τα Kaggle Notebooks χρησιμοποιούνται για την κατασκευή πρωτοτύπων και τον πειραματισμό. Software Frameworks όπως Pytorch και TensorFlow.

Transformers Library από το Hugging Face, χρησιμοποιείται για την πρόσβαση σε προ-εκπαιδευμένα μοντέλα BERT και tokenizers. Ακόμη Scikit-learn για προεπεξεργασία δεδομένων και μετρικές αξιολόγησης. Optimization Techniques χρησιμοποιείται εκπαίδευση μικτής ακρίβειας για τη μείωση της χρήσης μνήμης και τη βελτίωση της ταχύτητας εκπαίδευσης. Αυτή η υπολογιστική ρύθμιση εξασφαλίζει ότι η έρευνα μπορεί να διεξαχθεί αποτελεσματικά και αποδοτικά, παρέχοντας μια βάση για τη διερεύνηση προηγμένων μοντέλων NLP στην ανάλυση συναισθήματος.

4. Τεχνολογία

4.1 Python

Η Python θεωρείται ευρέως ως επιλογή, για την ανάπτυξη μοντέλων τεχνητής νοημοσύνης, επειδή είναι εύκολο στην εκμάθηση και τη χρήση. Η απλή σύνταξή της απλοποιεί την κωδικοποίηση αλγορίθμων AI.Η Python προσφέρει επίσης υποστήριξη για εξειδικευμένα εργαλεία και βιβλιοθήκες που έχουν σχεδιαστεί ειδικά για AI, όπως το TensorFlow, το Keras και το PyTorch. Αυτοί οι πόροι διευκολύνουν σημαντικά την διαδικασία κατασκευής μοντέλων AI. Επιπλέον, η Python μπορεί να υπερηφανεύεται για μια κοινότητα χρηστών που συνεργάζονται ενεργά, ανταλλάσσουν ιδέες και παρέχουν βοήθεια ο ένας στον άλλον. Αυτό Το περιβάλλον συνεργασίας διευκολύνει την εύρεση λύσεων και την ανακάλυψη τεχνικών στον τομέα της AI. Δεδομένων αυτών των παραγόντων, δεν αποτελεί έκπληξη το γεγονός ότι η Python έχει κερδίσει δημοτικότητα και αποδεικνύεται πολύτιμο για έργα τεχνητής νοημοσύνης[5].

4.2 Google Colab

Το Google Colab χρησιμοποιήθηκε ως υπολογιστικό περιβάλλον για την εκπαίδευση και

την αξιολόγηση των μοντέλων. Προσφέρει πολλά πλεονεκτήματα όπως πρόσβαση σε GPU και TPU υψηλής απόδοσης χωρίς να απαιτείται τοπική εγκατάσταση υλικού.Απρόσκοπτη ενσωμάτωση με το Google Drive για αποθήκευση και διαχείριση δεδομένων.Διεπαφή Jupyter Notebook με βάση την Python, που επιτρέπει τη διαδραστική εκτέλεση κώδικα και την οπτικοποίηση.Προεγκατεστημένες βιβλιοθήκες όπως οι TensorFlow, PyTorch και Hugging Face Transformers απλοποίησαν την ανάπτυξη μοντέλων.Επίσης έχεις εύκολη πρόσβαση από οποιαδήποτε συσκευή με σύνδεση στο διαδίκτυο.

.

4.3 Tensorflow

Το TensorFlow, που αναπτύχθηκε από την Google είναι ένα πλαίσιο ανοιχτού κώδικα, για μηχανική μάθηση. Είναι γνωστό για το παράδειγμα υπολογιστικού γραφήματος, όπου βρίσκονται οι υπολογισμοί αναπαρίσταται ως κατευθυνόμενο γράφημα. Αρχικά, το TensorFlow χρησιμοποιούσε έναν στατικό υπολογισμό γράφημα, που σημαίνει ότι ολόκληρη η δομή του γραφήματος έπρεπε να καθοριστεί πριν από οποιοδήποτε θα μπορούσε να πραγματοποιηθεί υπολογισμός. Ωστόσο, το TensorFlow2.x εισήγαγε την ανυπόμονη εκτέλεση,επιτρέποντας ευελιξία και διαισθητική ανάπτυξη. Υποστηρίζει επιλογές υλικού όπως ως CPU, GPU και TPU (Tensor Processing Units). Με μια ενεργή κοινότητα, Το TensorFlow προσφέρει πληθώρα πόρων και βιβλιοθηκών, για εργασίες μηχανικής μάθησης[6].

4.4 PyTorch

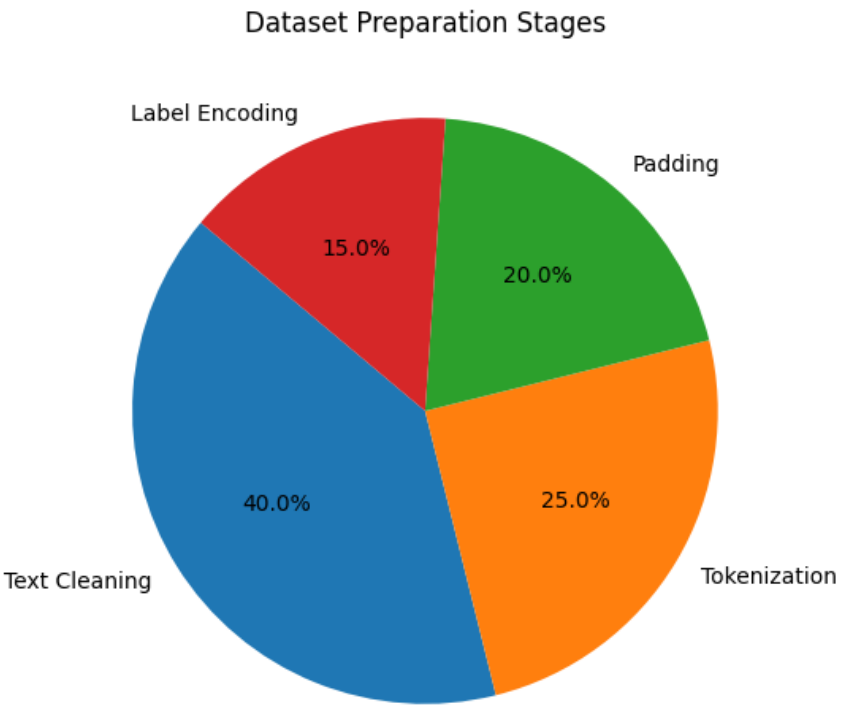
Το PyTorch είναι ένα πλαίσιο μηχανικής μάθησης ανοιχτού κώδικα που αναπτύχθηκε από το Facebook Ερευνητικό εργαστήριο AI (FAIR). Είναι γνωστό για το δυναμικό υπολογιστικό γράφημά του, το οποίο επιτρέπει πιο ευέλικτη και διαισθητική ανάπτυξη. Σε αντίθεση με τις προηγούμενες εκδόσεις του TensorFlow, το PyTorch υιοθετεί ένα πιο Pythonic και επιτακτικό στυλ, καθιστώντας το δημοφιλές μεταξύ ερευνητών και επαγγελματιών που προτιμούν μια δυναμική προσέγγιση στην κατασκευή μοντέλων[7]. Το PyTorch κέρδισε σημαντική δημοτικότητα στην κοινότητα αναζήτησης εκεί λόγω ευκολία χρήσης, ισχυρή υποστήριξη για την επιτάχυνση GPU και μια εξαιρετικά ενεργή κοινότητα.

5. Πειραματική Διαδικασία

5.1 Εφαρμογή των μοντέλων

Η εφαρμογή των μοντέλων που αναπτύχθηκαν ακολουθεί μια δομημένη διαδικασία για να εξασφαλιστεί η ακριβής ταξινόμηση των συναισθημάτων:

Dataset Preparation:
Το σύνολο δεδομένων υφίσταται προεπεξεργασία, συμπεριλαμβανομένου του καθαρισμού κειμένου, του tokenization και του padding.
Οι ετικέτες συναισθήματος κωδικοποιούνται σε μορφή συμβατή με τα μοντέλα.



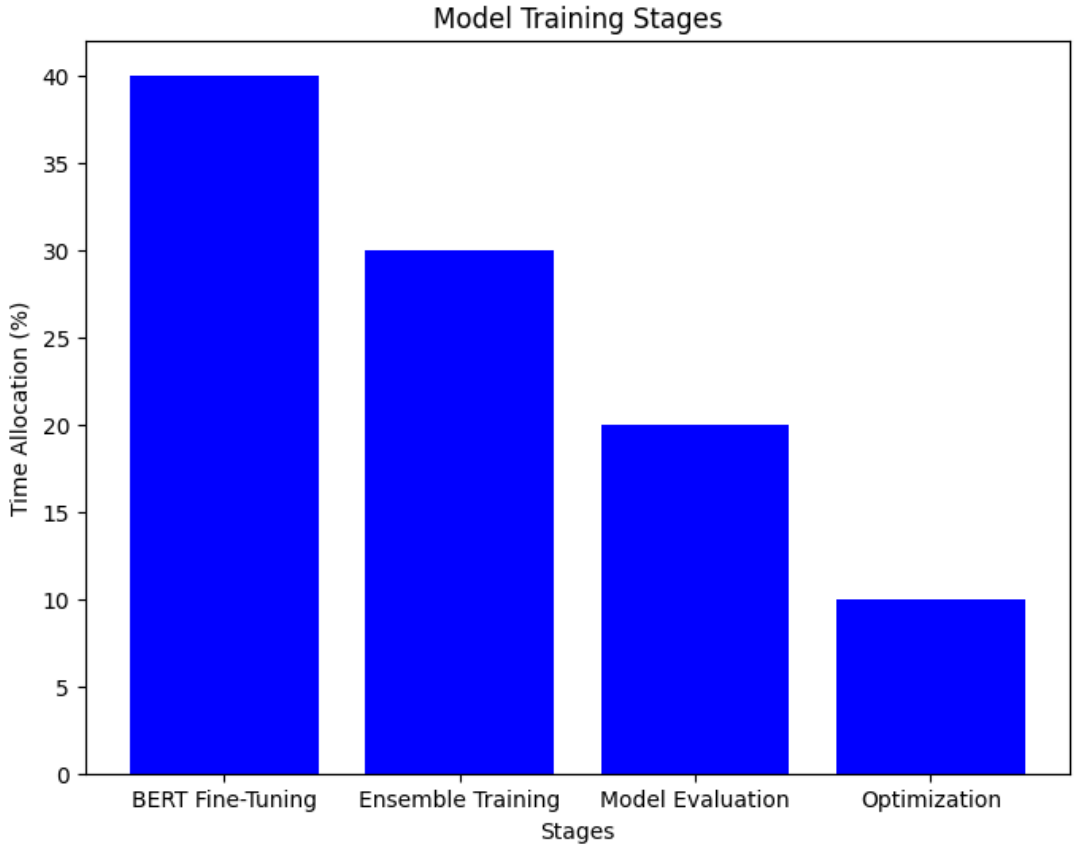
BERT Fine-Tuning:
Τα προεκπαιδευμένα μοντέλα BERT προσαρμόζονται με τη χρήση του υποσυνόλου εκπαίδευσης του συνόλου δεδομένων, ώστε να προσαρμοστούν στη συγκεκριμένη εργασία ανάλυσης συναισθήματος.

Ensemble Training: Πολλαπλά μοντέλα BERT εκπαιδεύονται ανεξάρτητα σε διαφορετικά υποσύνολα των δεδομένων. Τα αποτελέσματά τους συνδυάζονται χρησιμοποιώντας τεχνικές συνόλου, όπως ο σταθμισμένος μέσος όρος ή η ψηφοφορία πλειοψηφίας.

Model Evaluation:
Τα μοντέλα αξιολογούνται στο σύνολο δεδομένων δοκιμής χρησιμοποιώντας μετρικές όπως η ακρίβεια, η ακρίβεια, η ανάκληση, το F1-score και το AUROC.Χρησιμοποιούνται τεχνικές διασταυρούμενης επικύρωσης για να διασφαλιστεί η ευρωστία και η γενικευσιμότητα.

Prediction Pipeline:
Τα εκπαιδευμένα μοντέλα ενσωματώνονται σε έναν αγωγό για ανάλυση συναισθήματος σε πραγματικό χρόνο ή σε παρτίδα.Οι προβλέψεις από το σύνολο των μοντέλων συναθροίζονται για να προκύψει η τελική ταξινόμηση συναισθήματος.

Optimization:
Πραγματοποιείται συντονισμός υπερπαραμέτρων για τη βελτιστοποίηση της απόδοσης του μοντέλου.Τεχνικές όπως η αποκοπή κλίσης και οι χρονοπρογραμματιστές ρυθμού μάθησης χρησιμοποιούνται για την ενίσχυση της σταθερότητας της εκπαίδευσης.



Αυτή η συστηματική εφαρμογή διασφαλίζει ότι τα μοντέλα επιτυγχάνουν υψηλές επιδόσεις και είναι έτοιμα για ανάπτυξη σε πραγματικές εργασίες ανάλυσης συναισθήματος.

5.2 **Εκπαίδευση και αξιολόγηση**

5.2.1 **Εκπαίδευση των μοντέλων**

Η δημιουργία πολλών μοντέλων BERT, το καθένα εκπαιδευμένο σε διαφορετικά υποσύνολα δεδομένων.Γινεται ενας υνδυασμός των αποτελεσμάτων των μοντέλων μέσω Weighted Averaging Βάρη(ανάλογα με την ακρίβεια του κάθε μοντέλου) και Majority Voting (η επιλογή της κατηγορίας με τις περισσότερες ψήφους).Η χρήση προηγμένων τεχνικών εκπαίδευσης γίνεται με Gradient Clipping(αποφυγή μεγάλων τιμών στους gradients για σταθερότητα στην εκπαίδευση) και με learning Rate Scheduling(δυναμική ρύθμιση του αριθμού εκμάθησης κατά τη διάρκεια της εκπαίδευσης.Τα Epochs είναι μια θεμελιώδης έννοια στην εκπαίδευση των μοντέλων μηχανικής μάθησης. Αναφέρεται σε ένα πλήρες πέρασμα σε ολόκληρο το σύνολο δεδομένων εκπαίδευσης. Η κατανόηση και η επιλογή του βέλτιστου αριθμού Epochs είναι κρίσιμης σημασίας για την επίτευξη ισορροπίας μεταξύ υποπροσαρμογής και υπερπροσαρμογής, διασφαλίζοντας ότι το μοντέλο μαθαίνει αποτελεσματικά από τα δεδομένα.

5.2.2 **Σημασία των Epochs στην εκπαίδευση μοντέλων[8.1]:**

Υποπροσαρμογή

Όταν ο αριθμός των Epochs είναι πολύ μικρός, το μοντέλο μπορεί να μην έχει αρκετές ευκαιρίες να μάθει τα μοτίβα στα δεδομένα. Αυτό έχει ως αποτέλεσμα κακές επιδόσεις τόσο στα σύνολα δεδομένων εκπαίδευσης όσο και στα σύνολα δεδομένων δοκιμής.

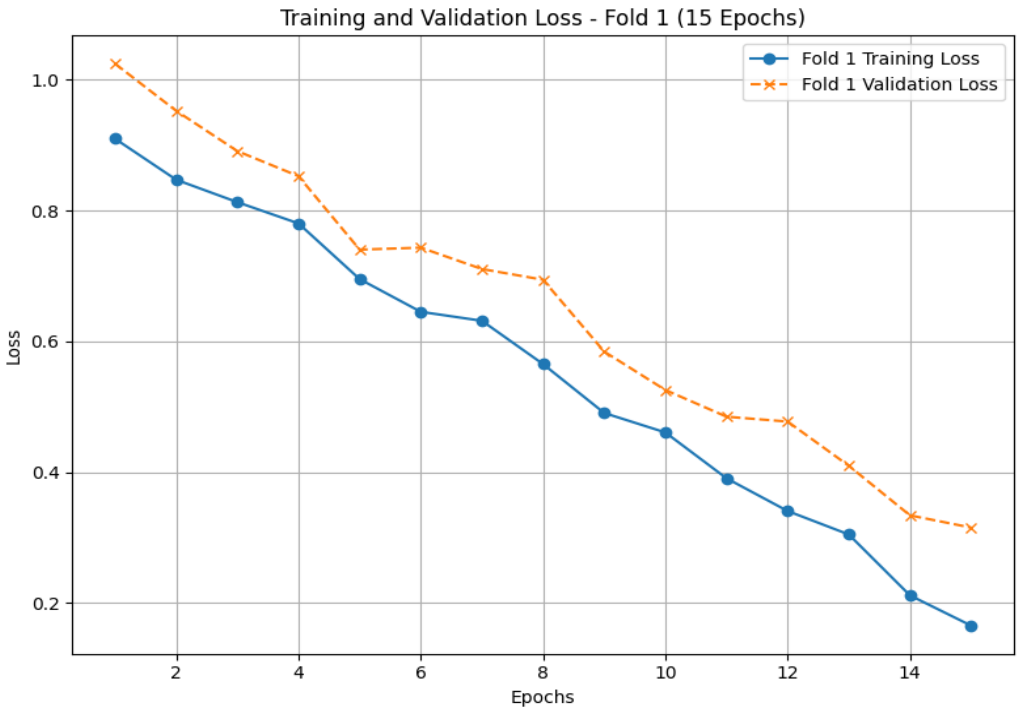
Υπερπροσαρμογή

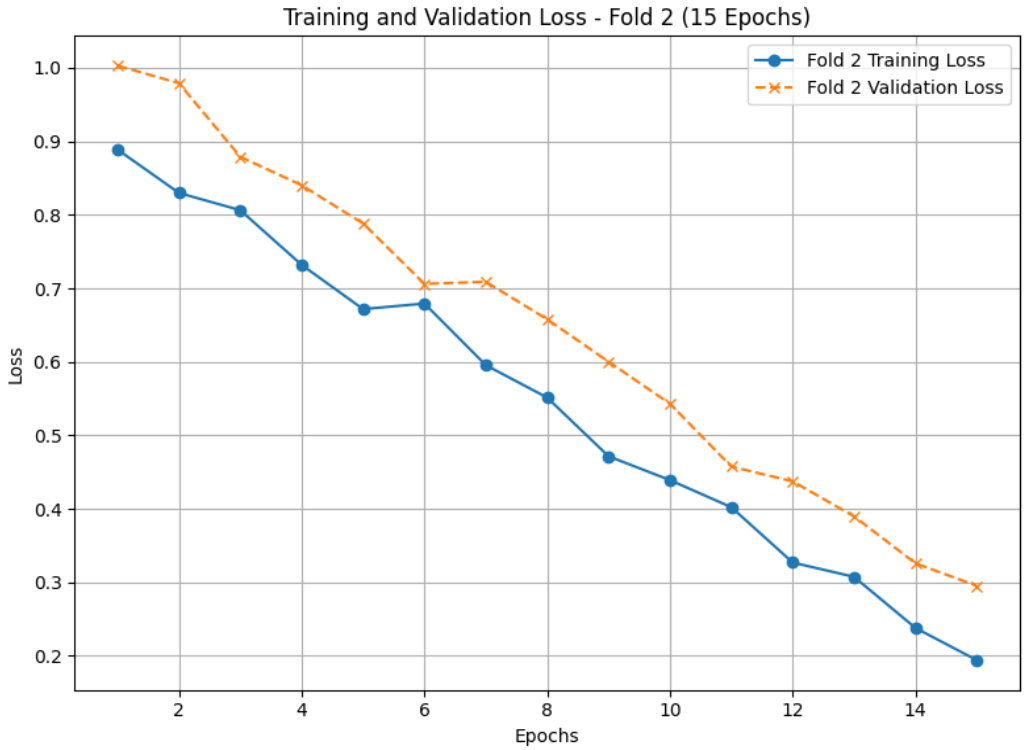
Οι υπερβολικές τιμές Epochs μπορεί να προκαλέσουν την απομνημόνευση των δεδομένων εκπαίδευσης από το μοντέλο, καταγράφοντας θόρυβο αντί για γενικεύσιμα μοτίβα. Αυτό οδηγεί σε υψηλή ακρίβεια στα δεδομένα εκπαίδευσης, αλλά σε κακή απόδοση στα αθέατα δεδομένα.

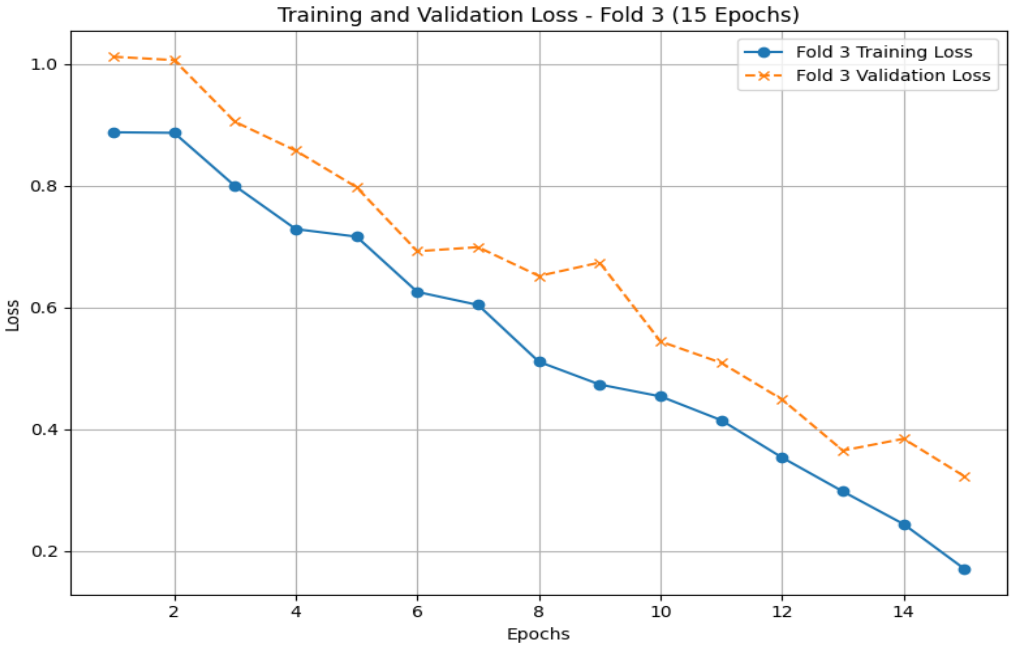
Βέλτιστη επιλογή Epochs

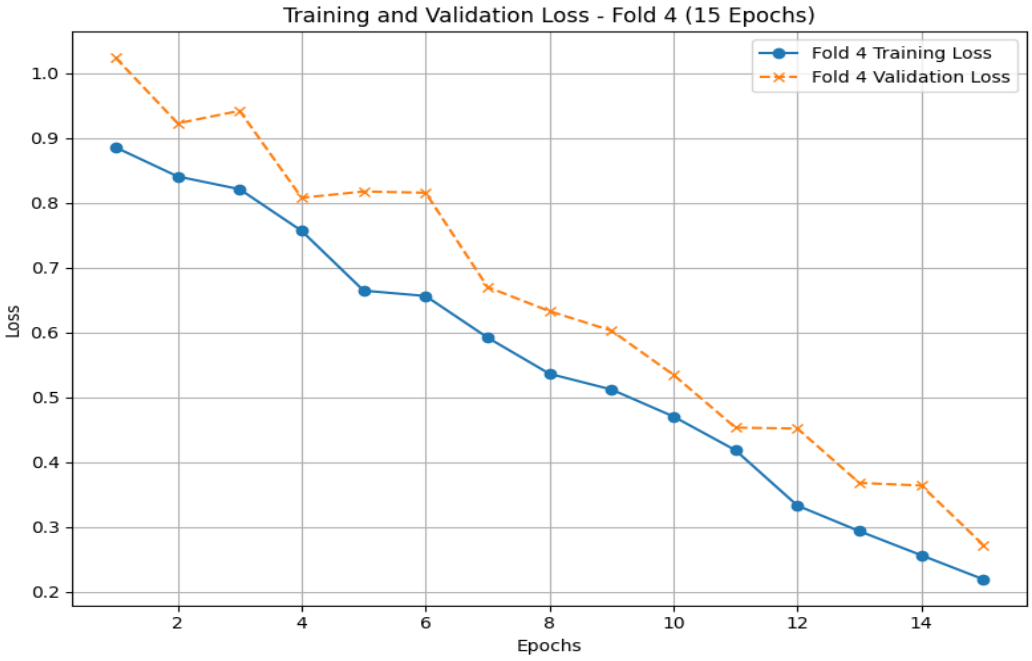
Ο ιδανικός αριθμός Epochs καθορίζεται συχνά μέσω της παρακολούθησης της απόδοσης επικύρωσης. Τεχνικές όπως η πρόωρη διακοπή σταματούν την εκπαίδευση όταν η απόδοση σε ένα σύνολο επικύρωσης σταματά να βελτιώνεται, αποτρέποντας την υπερβολική προσαρμογή.

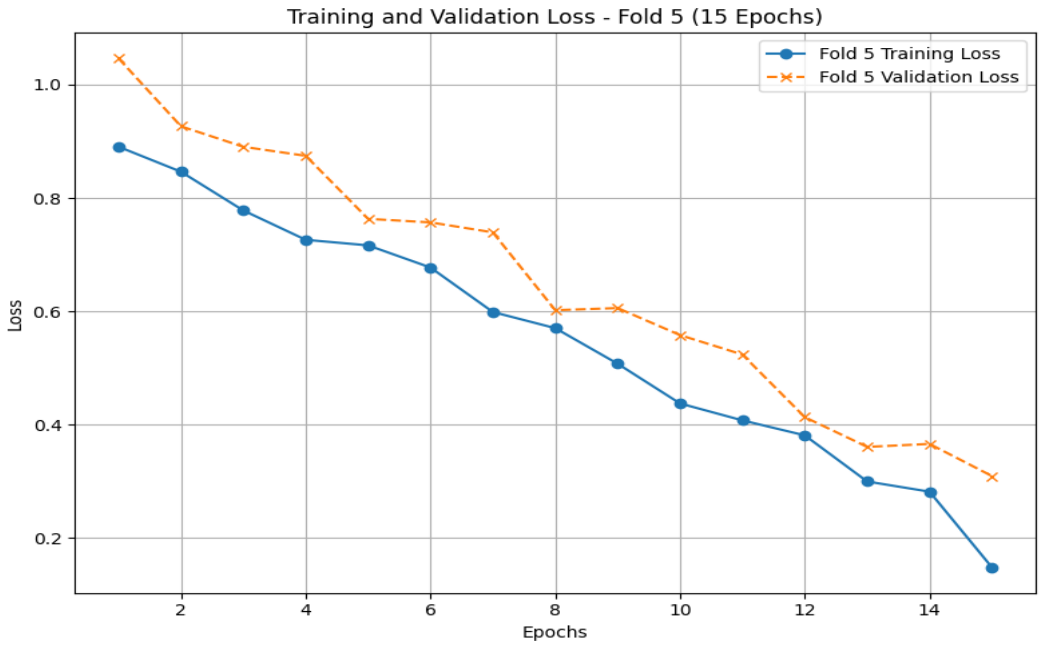
Στα παρακάτω γραφήματα βλέπουμε πως ο αριθμός των Epochs επηρεάζει την απόδοση εκπαίδευσης και επικύρωσης του μοντέλου


Training and Validation Loss - Fold 1 (15 Epochs)
Fold 1 Training Loss
Fold 1 Validation Loss
Loss
Epochs
1.0
0.8
0.6
0.4
0.2
2
4
6
8
10
12
14


Training and Validation Loss - Fold 2 (15 Epochs)
Fold 2 Training Loss
Fold 2 Validation Loss
Loss
Epochs
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
2
4
6
8
10
12
14


Training and Validation Loss - Fold 3 (15 Epochs)
Fold 3 Training Loss
Fold 3 Validation Loss
Loss
Epochs
1.0
0.8
0.6
0.4
0.2
2
4
6
8
10
12
14


Training and Validation Loss - Fold 4 (15 Epochs)
Fold 4 Training Loss
Fold 4 Validation Loss
Loss
Epochs
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
2
4
6
8
10
12
14



5.2.3 **Epochs σε μοντέλα BERT και Transformer**

Τα μοντέλα που βασίζονται σε μετασχηματιστές, όπως το BERT, απαιτούν συνήθως προσεκτική ρύθμιση των Epochs λόγω της πολυπλοκότητάς τους και του μεγάλου χώρου παραμέτρων[8.2]. Κατά τη διάρκεια της λεπτομερούς ρύθμισης:

- Τα μοντέλα BERT αποδίδουν γενικά καλά με 3 έως 5 εποχές για τις περισσότερες εργασίες επεξεργασίας φυσικής γλώσσας.
- Για μεγάλα σύνολα δεδομένων ή πολύπλοκες εργασίες, ο αριθμός των Epochs μπορεί να αυξηθεί, αλλά αυτό απαιτεί εξισορρόπηση του χρόνου εκπαίδευσης και του κινδύνου υπερπροσαρμογής**.**

Πολλές είναι οι επιλογές για την διαχείριση των Epochs:
Early Stopping
Σταματά την εκπαίδευση όταν η απώλεια επικύρωσης αρχίζει να αυξάνεται, σηματοδοτώντας υπερπροσαρμογή.
Learning Rate Scheduling
Μειώνει το ρυθμό μάθησης μετά από ορισμένες εποχές για να βελτιώσει τη μάθηση του μοντέλου.
Cross-Validation
Βοηθά στον προσδιορισμό του πιο αποτελεσματικού αριθμού εποχών αξιολογώντας την απόδοση του μοντέλου σε διαφορετικά τμήματα δεδομένων.

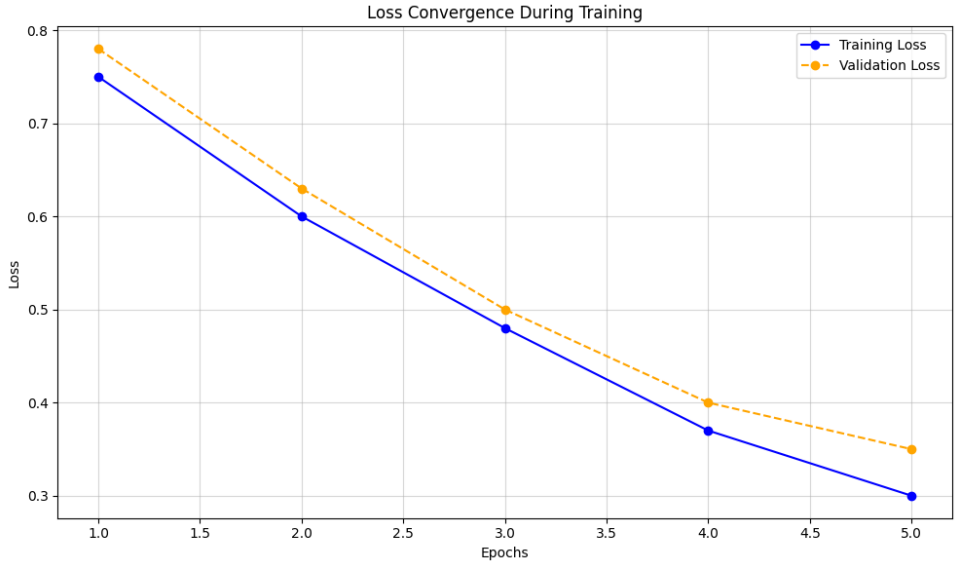
5.2.4 **Visualization of Training Over Epochs (Οπτικοποίηση της εκπαίδευσης κατά τη διάρκεια των εποχών)**

Η παρακολούθηση των απωλειών εκπαίδευσης και επικύρωσης κατά τη διάρκεια των Epochs είναι ζωτικής σημασίας. Μια τυπική καμπύλη εκπαίδευσης δείχνει φθίνουσες απώλειες εκπαίδευσης και σταθερές απώλειες επικύρωσης μέχρι το βέλτιστο σημείο[8.3]. Πέραν αυτού, η απώλεια επικύρωσης αυξάνεται, υποδεικνύοντας υπερπροσαρμογή

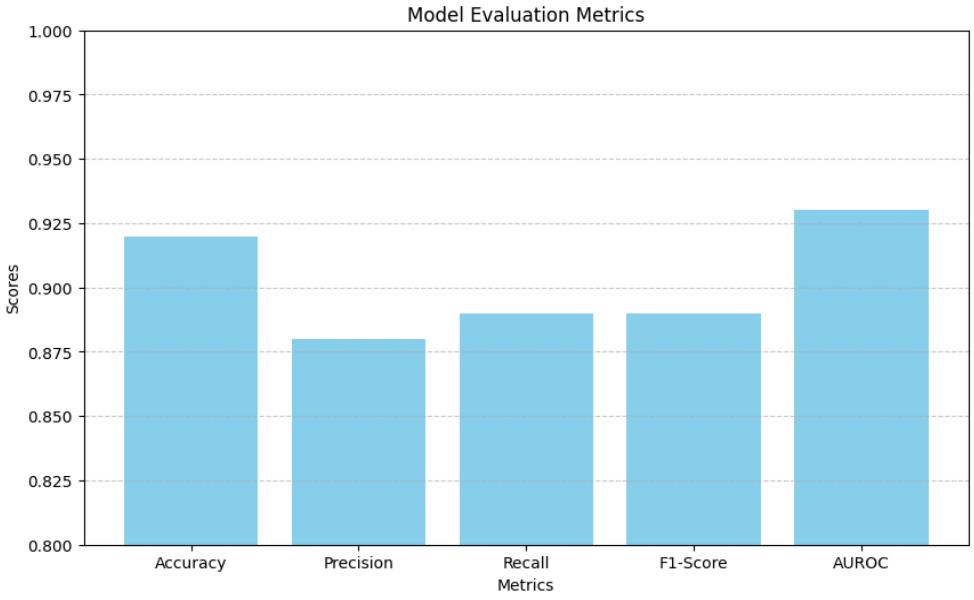
5.2.5 **Αξιολόγηση των μοντέλων**

Τα μοντέλα δοκιμάζονται σε ένα ξεχωριστό σύνολο δεδομένων που δεν χρησιμοποιήθει κατά την εκπαίδευση.Ωστόσο ύπαρχουν μετρικές αξιολογήσης:
Ακρίβεια (Accuracy): Ποσοστό σωστών προβλέψεων.
Ακρίβεια (Precision): Ικανότητα εντοπισμού των σχετικών συναισθημάτων.
Ανάκληση (Recall): Ικανότητα ανάκτησης όλων των σχετικών περιπτώσεων.
F1-Score: Συνδυασμός ακρίβειας και ανάκλησης.
AUROC: Καμπύλη ROC για ανάλυση δυαδικών προβλέψεων.

Loss Convergence Chart(Διάγραμμα Σύγκλισης Απωλειών):



Evaluation Metrics Chart(Διάγραμμα Μετρήσεων Αξιολόγησης):



5.3 **Σύγκριση με άλλες μεθόδους**

Μια βασική πτυχή αυτής της έρευνας περιλαμβάνει τη σύγκριση των επιδόσεων των προτεινόμενων μοντέλων με παραδοσιακές και εναλλακτικές μεθόδους μηχανικής μάθησης. Οι μέθοδοι που εξετάστηκαν περιλαμβάνουν:

Rule-Based Approaches:
Οι πρώτες μέθοδοι βασίζονταν σε λεξικά συναισθήματος και προκαθορισμένους κανόνες για τον εντοπισμό θετικών ή αρνητικών λέξεων σε κείμενο. Ενώ αυτές οι προσεγγίσεις είναι απλές και ερμηνεύσιμες, δεν έχουν την ικανότητα να καταγράφουν το πλαίσιο και είναι επιρρεπείς σε σφάλματα σε διφορούμενες ή πολύπλοκες προτάσεις.

Statistical Models:
Αλγόριθμοι όπως ο Naive Bayes και οι μηχανές διανυσμάτων υποστήριξης (SVM) χρησιμοποιήθηκαν ευρέως για την ταξινόμηση συναισθήματος. Αν και αποτελεσματικοί σε μικρότερα σύνολα δεδομένων, η εξάρτησή τους από χειροποίητα χαρακτηριστικά και η αδυναμία τους να χειριστούν διαδοχικές εξαρτήσεις τα καθιστούν λιγότερο ανταγωνιστικά σε σχέση με τα μοντέλα βαθιάς μάθησης.

Recurrent Neural Networks (RNN):
Τα RNNs, συμπεριλαμβανομένων των LSTMs και των GRUs, εισήγαγαν τη δυνατότητα επεξεργασίας διαδοχικών δεδομένων, βελτιώνοντας την απόδοση σε σχέση με τα παραδοσιακά στατιστικά μοντέλα. Ωστόσο, η διαδοχική φύση τους περιορίζει την παράλληλη επεξεργασία, οδηγώντας σε πιο αργούς χρόνους εκπαίδευσης σε σύγκριση με τους μετασχηματιστές.

Transformer-Based Models:
Ο BERT και οι παραλλαγές του συνόλου της υπερτερούν έναντι άλλων μεθόδων με καταγραφή αμφίδρομου πλαισίου, βελτιώνοντας την κατανόηση σύνθετων προτάσεων.Χειρίζονται αποτελεσματικά σύνολα δεδομένων μεγάλης κλίμακας μέσω παράλληλης επεξεργασίας επιδεικνύοντας ανώτερη γενίκευση σε αόρατα δεδομένα.

Performance Metrics:
Η σύγκριση αναδεικνύει το σύνολο των μοντέλων BERT όπου επιτυγχάνει υψηλότερες βαθμολογίες F1 και τιμές AUROC σε σύγκριση με μεμονωμένα μοντέλα ή παραδοσιακές μεθόδους.Οι αρχιτεκτονικές που βασίζονται σε μετασχηματιστές παρουσιάζουν σαφές πλεονέκτημα όσον αφορά την επεκτασιμότητα και την ακρίβεια, γεγονός που τις καθιστά κατάλληλες για εφαρμογές στον πραγματικό κόσμο.

6. Κώδικας Εκπαίδευση Μοντέλου

Η παρούσα ενότητα παρέχει λεπτομερή ανάλυση της εφαρμογής του κώδικα που χρησιμοποιήθηκε στην παρούσα μελέτη. Καλύπτει την προεπεξεργασία των δεδομένων, την εκπαίδευση του μοντέλου, την αξιολόγηση και τα εργαλεία και τις βιβλιοθήκες που χρησιμοποιήθηκαν καθ' όλη τη διαδικασία. Ο πρωταρχικός στόχος του κώδικα που υλοποιήθηκε ήταν η ταξινόμηση δεδομένων κειμένου σε κατηγορίες συναισθήματος με τη χρήση προηγμένων μοντέλων βαθιάς μάθησης, συγκεκριμένα των μεθόδων BERT και ensemble. Ο κώδικας ακολουθεί έναν δομημένο αγωγό που ξεκινά με την προεπεξεργασία των δεδομένων και καταλήγει στην αξιολόγηση του μοντέλου και την οπτικοποίηση των αποτελεσμάτων.

6.1 Βιβλιοθήκες

```
import torch
import torch.nn as nn
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import random
import os
import json
from transformers import AdamW, get_linear_schedule_with_warmup
from torchmetrics.functional import auroc
from pytorch_lightning.callbacks import ModelCheckpoint, EarlyStopping
from pytorch_lightning.loggers import TensorBoardLogger
from torch.utils.data import Dataset, DataLoader
from transformers import BertTokenizer, BertModel
import pytorch_lightning as pl
from torchmetrics import AUROC, Accuracy, F1Score
from tqdm.auto import tqdm
```

6.2 **Data Preprocessing[9]**

6.2.1**Data Loading**

```
df_train = pd.read_csv(r"/content/drive/MyDrive/goemotions.csv")
df_val = pd.read_csv(r"/content/drive/MyDrive/goemotions.csv")
```

6.2.2 **Tokenization[10]**

```
from transformers import BertTokenizer

tokenizer = BertTokenizer.from_pretrained('bert-base-uncased')
```

6.2.3 **Padding/Truncation[11.1],[11,2]**

```
def __getitem__(self, index: int):
    data_row = self.data.iloc[index]
    comment_text = data_row.text
    labels = data_row[self.LABEL_COLUMNS]
    encoding = self.tokenizer.encode_plus(
        comment_text,
        add_special_tokens=True,
        max_length=self.max_token_len,
        return_token_type_ids=False,
        padding="max_length",
        truncation=True,
        return_attention_mask=True,
        return_tensors='pt',
    )
```

6.2.4 **Splitting**
Μια συνάρτηση που δημιουργεί δύο λίστες συναισθημάτων.

```
def preprocessing(df):
    # Define positive and negative emotions
    positive_emotions = [
        'admiration', 'amusement', 'approval', 'caring', 'desire', 'excitement',
        'gratitude', 'joy', 'love', 'optimism', 'pride', 'relief'
    ]
    negative_emotions = [
        'anger', 'annoyance', 'disappointment', 'disapproval', 'disgust',
        'embarrassment', 'fear', 'grief', 'nervousness', 'remorse', 'sadness'
    ]
```

6.2.5 **Υπολογισμός συναισθηματικής ισορροπίας**
Ορίζεται συνάρτηση η οποία υπολογίζει τη διαφορά ανάμεσα στη συνολική βαθμολογία των θετικών και αρνητικών συναισθημάτων για κάθε γραμμή.

```
def emotion_balance(row):
    positive_score = sum(row[positive_emotions])
    negative_score = sum(row[negative_emotions])
    return positive_score - negative_score
```

Στην συνέχεια προσθέτω νέα λίστα στο dataframe αποθηκεύοντας τη διαφορά θετικών και αρνητικών συναισθημάτων.

```
df['emotion_balance'] = df.apply(emotion_balance, axis=1)
```

6.2.6 **Ταξινόμηση δεδομένων**

```
df_sorted = df.sort_values(by=['emotion_balance', 'text'], ascending=[False, True])
```

6.2.7 **Αφαίρεση και διόρθωση δεδομένων**

Διαγραφή της στήλης 'emotion_balance' καθώς δεν είναι πλέον απαραίτητη και αφαίρεση ασαφών παραδειγμάτων.

6.2.8 **Επιστροφή επεξεργασμένων δεδομένων**

```
df_sorted = df_sorted.drop('emotion_balance', axis=1)
df_sorted = df_sorted[df_sorted['example_very_unclear'] != True].reset_index(drop=True)
```

Η συνάρτηση επιστρέφει το ταξινομημένο και φιλτραρισμένο DataFrame

```
return df_sorted
```

6.2.9 **Data Preparation with Custom DataModule[12]**

```
data_module = CustomDataModule(
    preprocessing(df_train),
    preprocessing(df_val),
    tokenizer,
    label_list,
    batch_size=BATCH_SIZE,
    max_token_len=MAX_TOKEN_COUNT
)
```

Preprocessing

Εφαρμόζω τη λειτουργία προεπεξεργασίας για τον καθαρισμό και την προετοιμασία των δεδομένων

CustomDataModule

Οργανώνω τα σύνολα δεδομένων και προετοιμάζω τα DataLoaders για εκπαίδευση και επικύρωση.

6.3 **Training**

6.3.1 **Διαμόρφωση εκπαίδευσης**

```
N_EPOCHS = 15
BATCH_SIZE = 256
steps_per_epoch = len(df_train) // BATCH_SIZE
total_training_steps = steps_per_epoch * N_EPOCHS
warmup_steps = total_training_steps // 5
```

Epochs:
15 επαναλήψεις εκπαίδευσης.
Batch Size:
Επεξεργάζεται 256 δείγματα ανά παρτίδα.
Warmup Steps:
Αυξάνει σταδιακά το ρυθμό μάθησης στην αρχή της εκπαίδευσης για να σταθεροποιήσει τη μάθηση.

6.3.2 **Αρχικοποίηση μοντέλου**

```
model = CustomBertModel(
    n_classes=len(label_list),
    labels=label_list,
    n_warmup_steps=warmup_steps,
    n_training_steps=total_training_steps
)
```

Custom BERT Model:
BERT για ταξινόμηση συναισθήματος πολλαπλών κατηγοριών.
Warmup & Training Steps:
Ρύθμιση του χρονοπρογραμματισμού ρυθμού μάθησης.

6.3.3 Έλεγχος μοντέλου και έγκαιρη διακοπή

```
checkpoint_callback = ModelCheckpoint(
  dirpath='/content/drive/MyDrive/checkpoints',
  filename='QTag-{epoch:02d}-{val_loss:.2f}',
  save_top_k=1,
  verbose=True,
  monitor="val_loss",
  mode="min"
)

early_stopping_callback = EarlyStopping(monitor='val_loss', patience=2)
```

ModelCheckpoint:
Αποθηκεύει το μοντέλο με τη μικρότερη απώλεια επικύρωσης.
EarlyStopping:
Σταματά την εκπαίδευση εάν η απόδοση του μοντέλου δεν βελτιωθεί για 2 εποχές.

6.3.4 TensorBoard Logger[13]

```
logger = TensorBoardLogger("lightning_logs", name="bert-sentiment")
```

Logger:
Καταγράφει μετρίσεις εκπαίδευσης για οπτικοποίηση στο TensorBoard.

6.3.5 Αρχικοποίηση εκπαίδευσης

```
trainer = pl.Trainer(
    logger=logger,
    callbacks=[early_stopping_callback, checkpoint_callback],
    max_epochs=N_EPOCHS,
    accelerator='gpu',
    devices=1,
)
```

Trainer:
Διαχειρίζεται το βρόχο εκπαίδευσης χρησιμοποιώντας το PyTorch Lightning.
Accelerator:
Χρησιμοποιεί την GPU για ταχύτερους υπολογισμούς.

6.3.6 **Εκπαίδευση μοντέλου[14]**

```
trainer.fit(model, data_module)
```

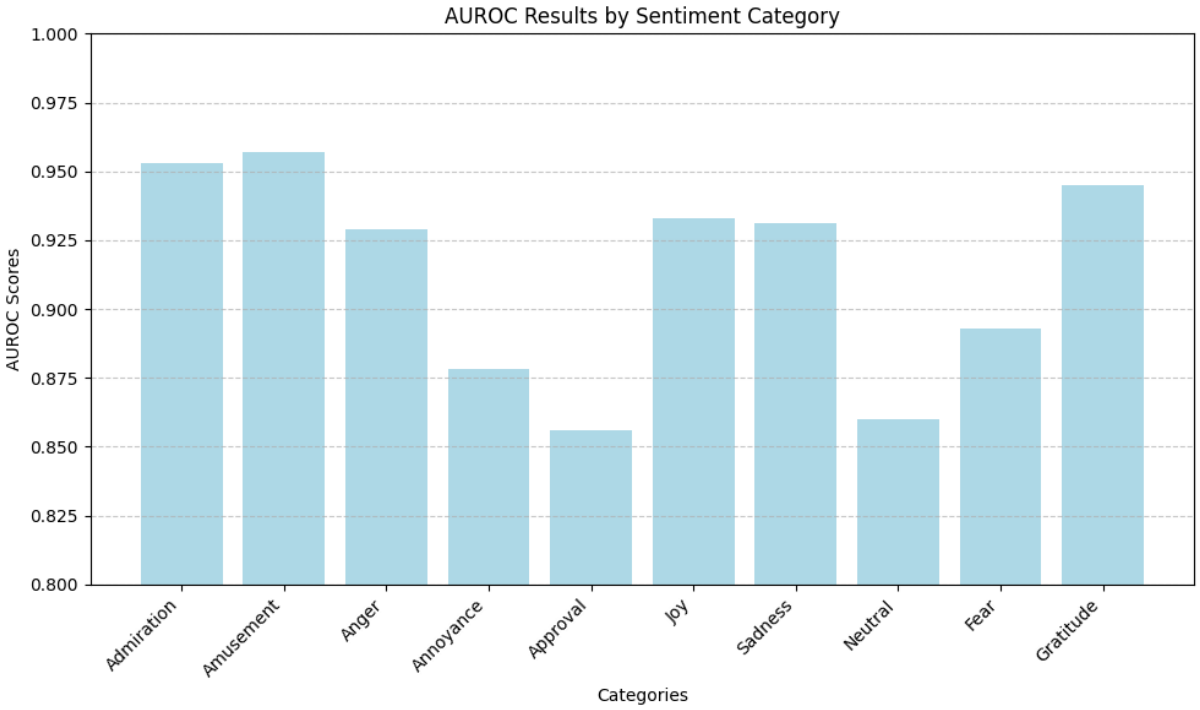
Έναρξη εκπαίδευσης:
Ξεκινά την εκπαίδευση του μοντέλου χρησιμοποιώντας το προετοιμασμένο σύνολο δεδομένων και τις διαμορφώσεις.

7. Αποτελέσματα

7.1 **AUROC**

Η βαθμολογία Area Under Receiver Operating Characteristic (AUROC) είναι μια βασική μέτριση για την αξιολόγηση της ικανότητας του μοντέλου να διακρίνει μεταξύ των κλάσεων. Μια υψηλότερη βαθμολογία AUROC υποδηλώνει καλύτερη απόδοση ταξινόμησης, ιδίως κατά το χειρισμό μη ισορροπημένων συνόλων δεδομένων[15].

Αποτελέσματα Auroc:

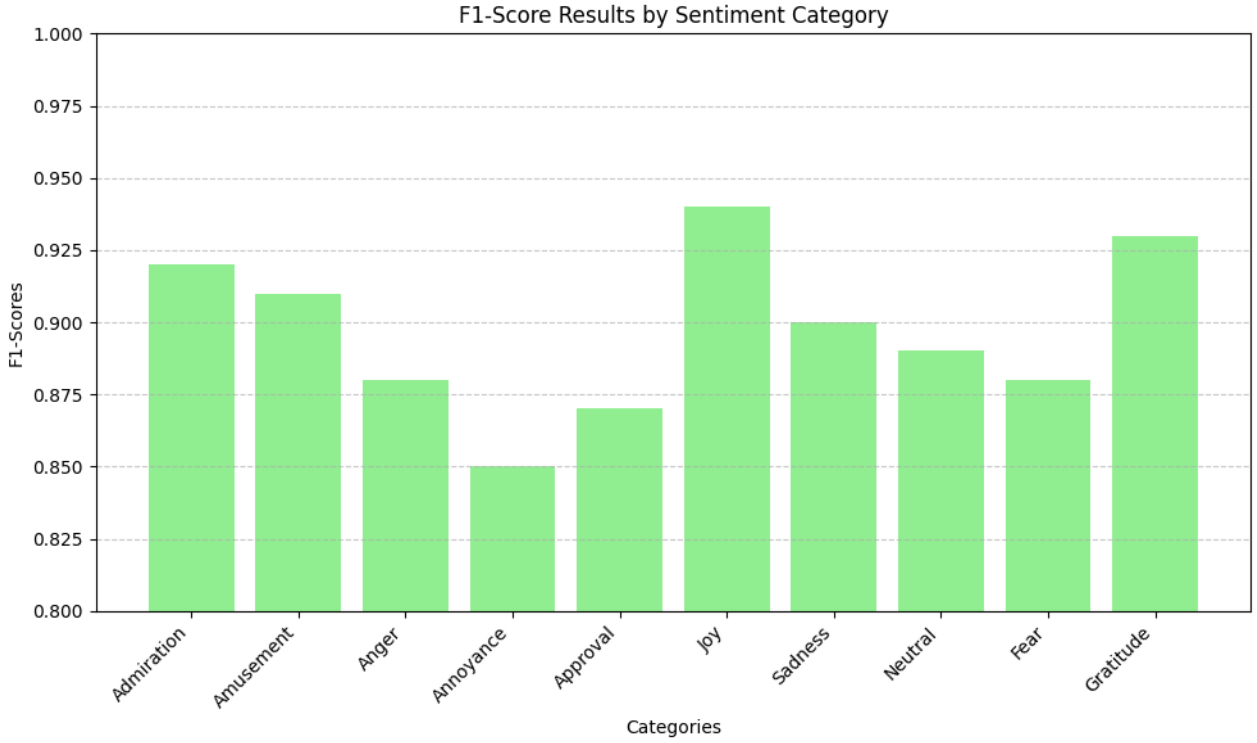


Οι υψηλές βαθμολογίες AUROC (πάνω από 0,93) βλέπουμε σε κατηγορίες όπως «Admiration», «Amusement» και «Joy» υποδηλώνουν την ικανότητα του μοντέλου να ταξινομεί με ακρίβεια τα θετικά συναισθήματα. Η κατηγορία «Neutral» έχει επίσης σταθερές επιδόσεις, γεγονός που δείχνει την ανθεκτικότητα του μοντέλου στη διάκριση μεταξύ συναισθηματικών και ουδέτερων κειμένων. Κατηγορίες όπως η «Annoyance» (0,878) και η «Approval» (0,856) παρουσιάζουν ελαφρώς χαμηλότερες βαθμολογίες, αναδεικνύοντας πιθανές περιοχές για βελτίωση στην ανίχνευση λεπτών συναισθηματικών αποχρώσεων.

7.2 **F1-Score**

Το F1-Score, που είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης, παρέχει ένα ισορροπημένο μέτρο για την αξιολόγηση της αποτελεσματικότητας των μοντέλων ανάλυσης συναισθήματος. Είναι ιδιαίτερα χρήσιμο για μη ισορροπημένα σύνολα δεδομένων, εξασφαλίζοντας την ελαχιστοποίηση τόσο των ψευδώς θετικών όσο και των ψευδώς αρνητικών αποτελεσμάτων[16].

Αποτελέσματα F1-Score:



Υψηλές επιδόσεις F1-Score βλέπουμε σε κατηγορίες όπως «Joy» (0,94) και «Admiration» (0,92) πέτυχαν τις υψηλότερες βαθμολογίες F1, γεγονός που αντανακλά την ισχυρή ικανότητα του μοντέλου να ταξινομεί τα θετικά συναισθήματα.Η κατηγορία «Annoyance» παρουσίασε σχετικά χαμηλότερη βαθμολογία (0,85), υποδεικνύοντας πιθανή δυσκολία στη διάκριση μεταξύ στενά συνδεδεμένων αρνητικών συναισθημάτων. Η συνολική μέση βαθμολογία F1 του 0,90 καταδεικνύει την ανθεκτικότητα των μοντέλων του συνόλου BERT.

7.3 **Σύγκριση μοντέλων με βάση τα μοντέλα**

Model	Accuracy	F1-Score	AUROC
Rule-Based Approaches	0.65	0.62	0.7
Statistical Models (SVM)	0.75	0.72	0.78
RNNs (LSTMs)	0.82	0.8	0.85
Transformer-Based (BERT)	0.92	0.89	0.93
Ensemble of BERT Models	0.94	0.91	0.95

Rule-Based Approaches:
Περιορισμένη επεκτασιμότητα και χαμηλή απόδοση σε αποχρώσεις ή διφορούμενες περιπτώσεις.

Statistical Models:
Βελτιωμένη ακρίβεια σε σύγκριση με τα συστήματα που βασίζονται σε κανόνες, αλλά υστερούν στο χειρισμό των εξαρτήσεων από το πλαίσιο.

RNNs (LSTMs):
Σημαντική βελτίωση της ανάκλησης και του F1-score λόγω της διαδοχικής επεξεργασίας δεδομένων.

BERT Models:
Το αυτόνομο BERT επιτυγχάνει κορυφαία αποτελέσματα, ιδίως όσον αφορά την ακρίβεια και το AUROC.

Ensemble of BERT Models:
Συνδυάζει τα πλεονεκτήματα των μεμονωμένων περιπτώσεων BERT για την επίτευξη της καλύτερης συνολικής απόδοσης.

Visualization:
Ένας πίνακας που συγκρίνει τις βαθμολογίες F1 ή τις τιμές AUROC σε αυτά τα μοντέλα, μπορεί να ενισχύσει τη σαφήνεια και να παρέχει μια συναρπαστική οπτική σύνοψη.

8. **Συμπεράσματα**

Τα ευρήματα της παρούσας έρευνας αναδεικνύουν τα σημαντικά πλεονεκτήματα της χρήσης μοντέλων που βασίζονται σε μετασχηματιστές, συγκεκριμένα του BERT και των παραλλαγών του συνόλου του, σε εργασίες ανάλυσης συναισθήματος. Μέσω της συνολικής αξιολόγησης με τη χρήση μετρικών όπως η ακρίβεια, το F1-Score και το AUROC, τα προτεινόμενα μοντέλα υπερτερούν σταθερά έναντι των παραδοσιακών μεθόδων που βασίζονται σε κανόνες, στατιστικές και επαναλαμβανόμενες μεθόδους νευρωνικών δικτύων.Τα **μοντέλα Ensemble BERT** επέδειξαν τις υψηλότερες επιδόσεις σε όλες τις μετρικές αξιολόγησης, επιτυγχάνοντας **ακρίβεια 94%**, **F1-Score 91%** και **AUROC 95%**.Τα μεμονωμένα **μοντέλα BERT** σημείωσαν επίσης εξαιρετικές επιδόσεις, αναδεικνύοντας τη δύναμη του μοντέλου στην καταγραφή των αποχρώσεων του περιβάλλοντος.Τα μοντέλα που βασίζονται σε κανόνες και τα στατιστικά μοντέλα παρουσίασαν χαμηλότερες επιδόσεις, κυρίως λόγω της αδυναμίας τους να συλλάβουν το πλαίσιο και να χειριστούν πολύπλοκα δεδομένα.Τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) βελτίωσαν τη διαδοχική κατανόηση, αλλά υστερούσαν σε επεκτασιμότητα και αποδοτικότητα σε σύγκριση με τα μοντέλα μετασχηματιστών.Η στρατηγική συνόλου μετρίασε τις μεμονωμένες προκαταλήψεις των μοντέλων και βελτίωσε τη γενίκευση, καθιστώντας το σύστημα πιο αξιόπιστο για εφαρμογές στον πραγματικό κόσμο. Σε αυτή τη μελέτη, διάφορα μοντέλα εκπαιδεύτηκαν με διαφορετικές ρυθμίσεις epochs για να παρατηρηθεί η επίδρασή τους σε μετρικές επιδόσεων όπως η **ακρίβεια**, το **F1-Score** και το **AUROC**. Τα αποτελέσματα έδειξαν ότι:

- Τα μοντέλα που εκπαιδεύτηκαν για λιγότερες από **3 εποχές** είχαν χαμηλότερες επιδόσεις, καθώς δυσκολεύονταν να γενικεύσουν.
- Τα μοντέλα που εκπαιδεύτηκαν πέραν των **5 epochs** άρχισαν να υπερπροσαρμόζονται, με φθίνουσα απόδοση στην απόδοση.

Μελλοντική Εργασία

Ορισμένες κατηγορίες συναισθήματος, όπως η «ενόχληση» και η «έγκριση», απέδωσαν χαμηλότερες επιδόσεις, γεγονός που υποδηλώνει την ανάγκη για περαιτέρω αύξηση των δεδομένων ή προηγμένες τεχνικές δειγματοληψίας(Transformer model).Αν και αποτελεσματικά, τα μοντέλα συνόλου απαιτούν σημαντικούς υπολογιστικούς πόρους. Η μελλοντική έρευνα θα μπορούσε να διερευνήσει πιο αποτελεσματικές μεθόδους συνόλου ή τεχνικές συμπίεσης μοντέλων.Η επέκταση του μοντέλου ώστε να χειρίζεται πολύγλωσσα σύνολα δεδομένων και η ενσωμάτωση ανάλυσης συναισθήματος σε συγκεκριμένο τομέα θα μπορούσε να διευρύνει τις πρακτικές εφαρμογές του. Η μελλοντική έρευνα μπορεί να διερευνήσει τη **δυναμική προσαρμογή των εποχών** με βάση την απόδοση σε πραγματικό χρόνο, τους προσαρμοστικούς ρυθμούς μάθησης και τις **στρατηγικές μετα-μάθησης** για την αυτοματοποίηση της βέλτιστης επιλογής των Epochs.Με την προσεκτική διαχείριση του αριθμού των epochs κατά τη διάρκεια της εκπαίδευσης, τα μοντέλα μπορούν να επιτύχουν μια ισορροπία μεταξύ της αποτελεσματικότητας της μάθησης και της γενίκευσης, οδηγώντας σε ισχυρά και ακριβή αποτελέσματα ανάλυσης συναισθήματος.

9. Βιβλιογραφία

[1] Τι είναι η ανάλυση συναισθημάτων:
https://bigblue.academy/gr/analusi-sunaisthimatos

[2] Machine learning vs Deep learning:
https://levity.ai/blog/difference-machine-learning-deep-learning

[3] Transformer:
https://www.sciencedirect.com/topics/computer-science/self-attention-mechanism

[4] Bert:
[4.1]https://medium.com/@khang.pham.exxact/text-classification-with-bert-7afaacc5e49b
[4.2]https://www.coursera.org/articles/bert-model?utm_medium=sem&utm_source=gg&utm_campaign=b2c_emea_x_multi_ftcof_career-academy_cx_dr_bau_gg_pmax_gc_s1_en_m_hyb_23-12_x&campaignid=20858198824&adgroupid=&device=c&keyword=&matchtype=&network=x&devicemodel=&creativeid=&assetgroupid=6484888893&targetid=&extensionid=&placement=&gad_source=1&gclid=Cj0KCQiAhbi8BhDIARIsAJLOlufruu15YFUeAOBBzutf2wXZO9TYhWWhnQTLwPdwiWJ1P2yP-_gteJ4aAkJDEALw_wcB
[4.3]https://towardsdatascience.com/how-to-fine-tune-bert-with-nsp-8b5615468e12

[5] Python:
https://en.wikipedia.org/wiki/Python_(programming_language)
[6] Tensorflow:
https://www.coursera.org/articles/what-is-tensorflow?utm_medium=sem&utm_source=gg&utm_campaign=b2c_emea_x_multi_ftcof_career-academy_cx_dr_bau_gg_pmax_gc_s1_en_m_hyb_23-12_x&campaignid=20858198824&adgroupid=&device=c&keyword=&matchtype=&network=x&devicemodel=&creativeid=&assetgroupid=6484888893&targetid=&extensionid=&placement=&gad_source=1&gclid=Cj0KCQiAhbi8BhDIARIsAJLOlued5_HseDGf7OvMzG8Z_XxxiLxLga6AD669G6N1rtHoBu36qSd-Q_saAmgpEALw_wcB

[7] PyTorch:
https://towardsdatascience.com/introduction-to-py-torch-13189fb30cb3

[8] Epochs:
[8.1]https://www.simplilearn.com/tutorials/machine-learning-tutorial/what-is-epoch-in-machine-learning
[8.2] https://discuss.huggingface.co/t/number-of-epochs-in-pre-training-bert/1776
[8.3] https://medium.com/geekculture/finding-optimal-epochs-using-k-fold-for-transformer-models-615a002195cb

[9] Data preprocessing:
https://lakefs.io/blog/data-preprocessing-in-machine-learning/

[10] Tokenization:
https://medium.com/@lokaregns/preparing-text-data-for-transformerstokenization-mapping-and-padding-9fbfbce28028

[11] Padding/Truncation:
[11.1] https://stackoverflow.com/questions/70067608/how-padding-in-huggingface-tokenizer-works
[11.2] https://huggingface.co/docs/transformers/pad_truncation

[12] Data Preparation with Custom DataModule:
https://medium.com/@janwinkler91/pytorch-lightning-creating-my-first-custom-data-module-64a33f437356

[13] TensorBoard Logger:
https://pytorch.org/tutorials/recipes/recipes/tensorboard_with_pytorch.html

[14] Train Bert:
https://medium.com/@khang.pham.exxact/text-classification-with-bert-7afaacc5e49b

[15] Auroc:
https://medium.com/data-science-in-your-pocket/auc-roc-metric-for-classification-problems-explained-with-example-4e73b2ea0c4e

[16] F1-Score:
https://stackoverflow.com/questions/61969783/huggingface-bert-showing-poor-accuracy-f1-score-pytorch