**Εθνικο Μετσοβιο Πολυτεχνειο**
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
ΔΠΜΣ Επιστημη Δεδομενων και Μηχανικη Μαθηση

# ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΒΑΘΙΑΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΑΝΑΓΝΩΡΙΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΚΕΙΜΕΝΟ

## Μεταπτυχιακη Διπλωματικη Εργασια

του

**Γεώργιου Παναγιώτη Σ. Κόλλια**

Διπλωματούχου Ηλεκτρολόγου Μηχανικού &
Μηχανικού Υπολογιστών Ε.Μ.Π.

**Επιβλέπων:** Γεώργιος Στάμου, Αναπληρωτής Καθηγητής Ε.Μ.Π.

**Συνεπίβλεψη :** Παρασκευή Τζούβελη, Προσωπικό Ε.Ε.Δ.Ι.Π.

Αθήνα, Οκτώβριος 2021

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

# ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΒΑΘΙΑΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΑΝΑΓΝΩΡΙΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΕ ΚΕΙΜΕΝΟ

## Μεταπτυχιακη Διπλωματικη Εργασια

του

**Γεώργιου Παναγιώτη Σ. Κόλλια**
Διπλωματούχου Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών Ε.Μ.Π.

**Συμβουλευτική Επιτροπή:** Γεώργιος Στάμου, Αναπληρωτής Καθηγητής
Σταφυλοπάτης Ανδρέας-Γεώργιος, Καθηγητής
Κόλλιας Στέφανος, Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την Οκτωβρίου 2021.

| *(Υπογραφή)* | *(Υπογραφή)* | *(Υπογραφή)* |
|---|---|---|
| ..... | ..... | ..... |
| Γεώργιος Στάμου | Σταφυλοπάτης Ανδρέας-Γεώργιος | Κόλλιας Στέφανος |
| Αναπληρωτής Καθηγητής | Καθηγητής | Καθηγητής |
| Ε.Μ.Π. | Ε.Μ.Π. | Ε.Μ.Π. |

Αθήνα, Οκτώβριος 2021

*(Υπογραφή)*

.....
**Γεώργιος Παναγιώτης Σ. Κόλλιας**
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright ©–All rights reserved Γεώργιος Παναγιώτης Σ. Κόλλιας, 2021.
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ΄ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν το συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

# Περίληψη

Η Επεξεργασία Φυσικής Γλώσσας αποτελεί ένα υποπεδίο της Τεχνητής Νοημοσύνης που έχει γνωρίσει αλματώδη πρόοδο τα τελευταία χρόνια. Ειδικότερα, μετά την εισαγωγή των Transformers και κυρίως του μοντέλου BERT την τελευταία πενταετία, οι επιδόσεις των μηχανών σε διάφορες εργασίες έχουν βελτιωθεί και σε πολλές περιπτώσεις προσεγγίζουν ή και ξεπερνούν το ανθρώπινο επίπεδο. Τα μοντέλα αυτά είναι γενικού σκοπού και έχουν αυξήσει σημαντικά τις επιδόσεις σε πολλές διαφορετικές εργασίες που εντάσσονται στο υποπεδίο αυτό. Μια από αυτές τις εργασίες είναι και η Ανάλυση Συναισθήματος στην οποία επικεντρώνεται και η παρούσα πτυχιακή εργασία.

Η επίτευξη ωστόσο των προαναφερθέντων επιδόσεων συνακολουθείται από αύξηση στον αριθμό των παραμέτρων των εκάστοτε μοντέλων και τελικά σε αύξηση του μεγέθους, του βάθους και της εν γένει πολυπλοκότητάς τους. Είναι χαρακτηριστικό για τα πρόσφατα SOTA μοντέλα των τελευταίων χρόνων αριθμούν εκατομμύρια παραμέτρους. Αυτό στην πράξη οδηγεί σε αυξημένους χρόνους εκπαίδευσης και ανάγκες για αυξημένη υπολογιστική ισχύ. Σε περιβάλλοντα περιορισμένων υπολογιστικών πόρων ή σε περιβάλλοντα όπου απαιτείται μικρή καθυστέρηση απόκρισης, αυτές οι αυξημένες απαιτήσεις μπορούν δυνητικά να καταστήσουν τη χρήση αυτών των μοντέλων προβληματική. Για την αντιμετώπιση αυτού του προβλήματος έχουν προταθεί διάφορες τεχνικές, που επικεντρώνονται ως επί το πλείστον στη συμπίεση των πολύπλοκων μοντέλων. Μια από τις πλέον υποσχόμενες είναι και η Απόσταξη Γνώσης, όπου ένα μεγάλο μοντέλο εκπαιδεύεται εκ των προτέρων και εν συνεχεία μεταφέρεται η γνώση του σε μικρότερα συμπιεσμένα μοντέλα, τα οποία παρουσιάζουν με τον τρόπο αυτό σημαντικά αυξημένες επιδόσεις.

Στόχος της παρούσας πτυχιακής εργασίας είναι η σύγκριση συμπιεσμένων μοντέλων BERT με το αντίστοιχο μοντέλο βάσης και η συνεπακόλουθη εξαγωγή συμπερασμάτων αναφορικά με τις επιδόσεις τους σε σχέση με την υποκείμενη πολυπλοκότητά τους. Αφού πραγματοποιηθεί μια γενική βιβλιογραφική ανασκόπηση του πεδίου της Επεξεργασίας Φυσικής Γλώσσας και της Ανάλυσης Συναισθήματος ειδικότερα, εκτελούνται πειράματα με τέσσερα συμπιεσμένα μοντέλα BERT στο σετ δεδομένων του GoEmotions, για αναγνώριση συναισθημάτων από γραπτό λόγο. Παρουσιάζονται και αντιπαρατίθενται οι διάφορες μετρικές που χρησιμοποιήθηκαν για την αξιολόγησή τους και εκτιμάται κατά πόσον η χρήση τους σε περιβάλλοντα με περιορισμούς ειναι μια πραγματιστική επιλογή.

**Λέξεις Κλειδιά:** Επεξεργασία Φυσικής Γλώσσας, Ανάλυση Συναισθήματος, Συμπιεσμένα Μοντέλα, Απόσταξη Γνώσης, Μοντέλα Γλώσσας, BERT, Transformers

# Abstract

Natural Language Processing constitutes a subfield of Artificial Intelligence where immense progress has been achieved over the course of the last few years. More specifically, after the introduction of Transformers, and mainly of the BERT language model four years ago, machine performance on various tasks has been greatly improved, with the emerging results, being on many occasions, on par or even superior to the corresponding human level. Those models are general-purpose, thus capable of achieving considerable performance on various tasks of this specific field. The current dissertation is focused on one of those tasks, known as Sentiment Analysis.

However, achievement of the aforementioned performance is followed by an increase on the number of model parameters and eventually on model depth, size and overall complexity. It is common for recent SOTA models to contain millions of parameters. This can lead to increasing training times and the need of more computational resources. On limited-resources or low-latency environments, those increased requirements can potentially render the use of such models problematic. In order to mitigate this problem, various techniques have been proposed, with most of them focusing on compression of the complex models. One of the most promising ones is Knowledge Distillation, where a large model is trained in advance, and its obtained knowledge is being transferred afterwards on smaller compressed models, achieving that way significantly improved performance.

The main purpose of the current dissertation is the comparison between compact BERT Models and the BERT base Model and the subsequent reaching on meaningful insights, regarding their overall performance in relation to their underlying complexity. After performing a systematic literature review of the field of Natural Language Processing and of the more specific subfield of Sentiment Analysis, we conduct experiments with four compact BERT Models and the base BERT one, on the GoEmotions dataset, used for Emotion Recognition based on text. Consequently, various resulting metrics used for model evaluation are presented and compared and we conclude whether the use of the aforementioned models on limited-resources computational environments constitutes a pragmatic choice.

**Keywords:** Natural Language Processing, Sentiment Analysis, Compact Models, Knowledge Distillation, Language Models, BERT, Transformers

# Ευχαριστίες

Θα ήθελα κατάρχάς να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή Γεώργιο Στάμου για την ευκαιρία που μου έδωσε να ασχοληθώ με το συγκεκριμένο θέμα και για τη γενικότερη καθοδήγησή του καθ΄όλη τη διάρκεια των μεταπτυχιακών μου σπουδών.

Ιδιαίτερες ευχαριστίες θα ήθελα να αποδώσω και στην Ερευνήτρια κα. Τζούβελη Παρασκευή για τη διαρκή καθοδήγηση της, για τις υποδείξεις και τις συμβουλές της, χωρίς τις οποίες η ολοκλήρωση της παρούσας εργασίας θα ήταν αδύνατη. Το ενδιαφέρον της για το θέμα της διπλωματικής εργασίας, μου έδωσε σημαντικό κίνητρο για την επιτυχή ολοκλήρωσή της.

Τέλος ευχαριστώ τα κοντινά μου πρόσωπα για την αμέριστη στήριξη τους και για την ανάταση που προσφέρει η διαρκής παρουσία τους στη ζωή μου.

# Contents

**Contents** **12**

**List of Figures** **13**

**1 Εισαγωγή** **15**

**2 Επεξεργασία Φυσικής Γλώσσας** **17**
2.1 Εισαγωγή . . . . . 17
2.2 Σύντομη ιστορική αναδρομή . . . . . 17
2.3 Οι διάφορες προσεγγίσεις στην Επεξεργασία Φυσικής Γλώσσας . . . . . 18
2.4 Οι επιμέρους εφαρμογές στην Επεξεργασία Φυσικής Γλώσσας . . . . . 18

**3 Οι σημαντικότερες εξελίξεις στην Επεξεργασία Φυσικής Γλώσσας: από τα one-hot encodings στο BERT** **21**
3.1 One-hot encodings . . . . . 21
3.2 Word2vec . . . . . 21
3.2.1 Continuous bag-of-words(CBOW) . . . . . 22
3.2.2 Continuous skip-gram . . . . . 22
3.3 GloVe . . . . . 22
3.4 Seq2seq . . . . . 23
3.5 FastText . . . . . 24
3.6 RNN και LSTM . . . . . 24
3.7 Μηχανισμός Προσοχής . . . . . 25
3.8 Transformer . . . . . 27
3.9 Transformer-XL . . . . . 28
3.10 Η Μεταφορά Μάθησης . . . . . 29
3.11 Το μοντέλο ULMFiT (Universal Language Model Fine-tuning for Text Classification) . . . . . 31
3.12 Το μοντέλο ELMo (Embeddings from Language Models) . . . . . 31
3.13 Τα μοντέλα GPT (Generative Pre-Training) . . . . . 32
3.14 Το μοντέλο BERT (Bidirectional Encoder Representations from Transformers) . . . . . 34
3.14.1 Το μοντέλο και η γενική αρχιτεκτονική του . . . . . 34
3.14.2 Η είσοδος του μοντέλου . . . . . 34
3.14.3 Προ-εκπαίδευση του BERT: Μοντέλα Γλώσσας με Μάσκα(Masked Language Models) . . . . . 34
3.14.4 Προ-εκπαίδευση του BERT: Πρόβλεψη επόμενης πρότασης(Next Sentence Prediction) . . . . . 35

**4 Οι σημαντικότερες εξελίξεις στην Επεξεργασία Φυσικής Γλώσσας από το BERT και έπειτα 37**
4.1 Το μοντέλο XLNet . . . 37
4.2 Το μοντέλο RoBERTa . . . 38
4.3 Το μοντέλο ALBERT . . . 39
4.4 Το μοντέλο ELECTRA . . . 40
4.5 Το μοντέλο DeBERTa . . . 41
4.6 Το μοντέλο StructBERT . . . 41
4.7 Συμπίεση μοντέλων με Απόσταξη Γνώσης . . . 43

**5 Η Ανάλυση Συναισθήματος 45**

**6 Σετ Δεδομένων και Benchmarks 51**
6.1 Το σετ δεδομένων SNLI(Stanford Natural Language Inference) . . . 51
6.2 Το σετ δεδομένων MultiNLI(Multi-Genre Natural Language Inference) . . . 51
6.3 GLUE Benchmark . . . 52
6.4 SuperGLUE Benchmark . . . 54
6.5 Το σετ δεδομένων SQuAD . . . 56
6.6 Το σετ δεδομένων SQuAD 2.0 . . . 56
6.7 Το σετ δεδομένων GoEmotions . . . 57

**7 Πειράματα 63**
7.1 Εισαγωγή . . . 63
7.2 Σχεδιασμός και Εκτέλεση του Πειράματος . . . 63
7.3 Τα μοντέλα που θα συγκριθούν . . . 65
7.4 Αποτελέσματα . . . 66
7.5 Συμπεράσματα . . . 70

# List of Figures


3.1 Η one-hot αναπαράσταση της φράσης "i ate an apple and played the piano" . 22
3.2 Οι δύο αρχιτεκτονικές . . . . . 23
3.3 Η αρχιτεκτονική του μοντέλου Seq2seq. Στο παράδειγμα δίνεται ως είσοδος η ακολουθία "ABC" και παράγεται στην έξοδο η ακολουθία "WXYZ". Το μοντέλο παύει να κάνει προβλέψεις όταν δώσει ως έξοδο το token ¡EOS¿ . . . 23
3.4 Η παραγωγή των κρυφών καταστάσεων . . . . . 24
3.5 Το κελί μνήμης του LSTM . . . . . 25
3.6 Παράδειγμα υπολογισμού attention για την πρόταση "Thinking Machines" . . 27
3.7 Η αρχιτεκτονική του μοντέλου Transformer . . . . . 28
3.8 Το μοντέλο Transformer-XL με μήκος τμήματος 4 . . . . . 29
3.9 Η βασική ιδεά της Μεταφοράς Μάθησης . . . . . 30
3.10 Επισκόπηση του μοντέλου . . . . . 31
3.11 Τα τρία στάδια της εκπαίδευσης του μοντέλου . . . . . 32
3.12 Το μοντέλο ELMo χρησιμοποιεί δύο χωριστά δίκτυα για να επεξεργαστεί το κείμενο αμφίδρομα . . . . . 33
3.13 Τροποποιήσεις στην είσοδο για να ακολουθήσει μετέπειτα το fine-tuning του μοντέλου . . . . . 33
3.14 Αναπαράσταση μια ακολουθίας εισόδου του BERT . . . . . 35

4.1 Η μετάθεση στο XLNet . . . . . 38
4.2 Παράδειγμα της μεταθετικής γλωσσικής μοντελοποίησης . . . . . 39
4.3 Η διαδικασία εκπαίδευσης του ELECTRA . . . . . 41
4.4 Οι δύο διαφορετικές προσεγγίσεις σε αντιπαραβολή . . . . . 42
4.5 Οι δύο νέες εργασίες για την προ-εκπαίδευση . . . . . 43
4.6 Η διαδικασία εκπαίδευσης με την τεχνική της Απόσταξης Γνώσης . . . . . 44

5.1 Το πεδίο της Ανάλυσης Συναισθήματος . . . . . 46

6.1 Οι διαφορετικές συχνότητες με τις οποίες εμφανίζονται τα διάφορα ιδιαίτερα φαινόμενα ανάμεσα στα SNLI και το MNLI . . . . . 52
6.2 Οι τύποι των γλωσσικών φαινομένων που επισημαίνονται στο διαγνωστικό σετ δεδομένων, οργανωμένοι σε τέσσερις βασικές κατηγορίες . . . . . 54
6.3 Οι κατηγορίες της συλλογιστικής . . . . . 57
6.4 Οι κατηγορίες των αναπάντητων ερωτήσεων . . . . . 58
6.5 Η συχνότητα των εμφανισθέντων labels, μαζί με τη συσχέτιση ανά των αξιολογητών . . . . . 60
6.6 Οι συσχετίσεις ανάμεσα στις αξιολογήσεις των συναισθημάτων . . . . . 61

7.1 Η νέα τροποποιημένη διαδικασία εκπαίδευσης . . . . . 64

7.2 Οι υπερπαράμετροι που χρησιμοποιήθηκαν για το αρχικό πείραμα . . . . . 65
7.3 Οι χρόνοι για την πραγματοποίηση των προβλέψεων. . . . . 68
7.4 Το Macro-F1 score των μοντέλων. . . . . 68
7.5 Το Micro-F1 score των μοντέλων. . . . . 68
7.6 Οι χρόνοι για την πραγματοποίηση των προβλέψεων. . . . . 69
7.7 Το Macro-F1 score των μοντέλων. . . . . 69
7.8 Το Micro-F1 score των μοντέλων. . . . . 69

# 1

# Εισαγωγή

Τις τελευταίες δεκαετίες, όλο και αυξανόμενα γίνεται χρήση του όρου Τεχνητή Νοημοσύνη από ειδικούς και μη. Έχουν δοθεί διάφοροι ορισμοί για το τι συντελεί "Τεχνητή Νοημοσύνη" ωστόσο θα μπορούσε αδρομερώς να ειπωθεί ότι είναι ένας κλάδος της Πληροφορικής, στόχος του οποίου είναι η εμφύσηση ανθρώπινης ευφυΐας σε υπολογιστικά συστήματα, με χρήση ποικίλων μεθόδων και τεχνικών. Το σκεπτικό πίσω από αυτό είναι ότι η υπολογιστική ισχύς και ταχύτητα των υπολογιστικών συστημάτων σε συνδυασμό με τα ποιοτικά χαρακτηριστικά της ανθρώπινης ή και υπερανθρώπινης ευφυΐας μπορεί να οδηγήσει στην αντιμετώπιση δυσεπίλυτων για την ανθρωπότητα προβλημάτων και να την οδηγήσει έτσι σε περαιτέρω πρόοδο και ευμάρεια. Φυσικά, εγείρονται παράλληλα και πολλές ανησυχίες φιλοσοφικού τύπου ως προς το τι καθίσταται συνείδηση αλλά και ως προς τον κίνδυνο αφανισμού των ανθρώπων από μια μηχανή που θα έχει αναπτύξει υπερανθρώπινη ευφυΐα.

Ένα πεδίο της ευρύτερης ΤΝ είναι και η Μηχανική Μάθηση, που ορίζεται ως η μελέτη αλγορίθμων των οποίων οι επιδόσεις βελτιώνονται αυτόματα μέσω της εμπειρίας και της χρήσης μεγάλου συνήθως όγκου δεδομένων. Δομείται συνήθως ένα μοντέλο, βασισμένο σε ένα γνωστό σετ δεδομένων, στο οποίο εκπαιδεύεται για την μετέπειτα πραγματοποίηση προβλέψεων και συναγωγών σε νέα άγνωστα δεδομένα. Αξίζει να σημειωθεί ότι Μηχανική Μάθηση είναι γενικά ένα αρκετά ευρύ πεδίο, στο οποίο γίνεται χρήση ποικίλων και ετερόκλητων τεχνικών και μεθόδων, μεταξύ των οποίων είναι η υπολογιστική στατιστική και τα νευρωνικά δίκτυα. Στην κατεύθυνση των τελευταίων έχει αναπτυχθεί ευρέως τα τελευταία χρόνια και το υποπεδίο της Βαθιάς Μηχανικής Μάθησης, στο οποίο γίνεται χρήση νευρωνικών δικτύων πολλαπλών επιπέδων και παραμέτρων, τα οποία έχει παρατηρηθεί ότι με τη χρήση πολύ μεγάλου όγκου δεδομένων αποδίδουν πολύ καλά σε προβλήματα συγκεκριμένου τύπου, όπως η αναγνώρισή εικόνας, με αποτέλεσμα να είναι το SOTA.

Η θεωρία και οι τεχνικές που αναπτύσσονται στα παραπάνω πεδία χρησιμοποιούνται για την επίλυση και αντιμετώπιση διαφορετικού τύπου προκλήσεων και προβλημάτων. Η συγκεκριμένη εργασία καταπιάνεται με την Επεξεργασία Φυσικής Γλώσσας, που αναφέρεται στη διάδραση μεταξύ των υπολογιστικών συστημάτων και της ανθρώπινης φυσικής γλώσσας. Πιο συγκεκριμένα, θα πραγματοποιηθεί αρχικά μια εκτενής βιβλιογραφική ανασκόπηση του πεδίου εν γένει, με την παρουσίαση της εξέλιξης του πεδίου ανά τα χρόνια, την περιγραφή των επικρατουσών τεχνικών και μεθόδων που χρησιμοποιούνται, τη συνοπτική εξέταση των SOTA μοντέλων, καθώς και των σετ δεδομένων που χρησιμοποιούνται στην αξιολόγηση τους. Μετά από αυτό, θα εστιάσουμε σε ένα πιο συγκεκριμένο πεδίο της Επεξεργασίας Φυσικής Γλώσσας, την Ανάλυση Συναισθήματος που αναφέρεται στην εξαγωγή συμπερασμάτων για τη συναισθηματική κατάσταση ενός υποκειμένου, με βάση το γραπτό ή προφορικό λόγο. Στο πλαίσιο αυτού του πεδίου θα καταπιαστούμε με το πρόβλημα της χρήσης SOTA μοντέλων-που συνήθως αποτελούνται από εκατομμύρια παραμέτρους και απαιτούν μεγάλους χρόνους εκπαίδευσης- σε

περιβάλλοντα περιορισμένων υπολογιστικών πόρων. Για την αντιμετώπιση αυτού του προβλήματος γίνεται χρήση συμπιεσμένων μοντέλων χαμηλότερης πολυπλοκότητας. Στην παρούσα πτυχιακή εκτελούμε πειράματα πάνω σε τέσσερα τέτοια μοντέλα κλιμακούμενης πολυπλοκότητας και συγκρίνουμε την απόδοση τους για να αποφανθούμε στο κατά πόσον η χρήση τους αποτελεί μια πραγματιστική επιλογή.

# 2

# Επεξεργασία Φυσικής Γλώσσας

## 2.1 Εισαγωγή

Ως Επεξεργασία Φυσικής Γλώσσας αναφέρεται στη βιβλιογραφία το διεπιστημονικό υποπεδίο της Τεχνητής Νοημοσύνης, της Γλωσσολογίας και της Επιστήμης των Υπολογιστών που ασχολείται με την αλληλεπίδραση του Υπολογιστή με την ανθρώπινη φυσική γλώσσα. Απώτερος στόχος είναι ο υπολογιστής να "κατανοεί" την ανθρώπινη γλώσσα και να διαχειρίζεται και επεξεργάζεται αυτόματα γραπτά κείμενα και προφορικό λόγο. Απότοκος αυτού είναι η αυτοματοποίηση και διευκόλυνση διαφόρων διεργασιών στην υπηρεσία του ανθρώπου [1].

## 2.2 Σύντομη ιστορική αναδρομή

Οι απαρχές του πεδίου εντοπίζονται ήδη στον 17ο αιώνα, οπότε και οι φιλόσοφοι και μαθηματικοί Καρτέσιος και Λάιμπνιτζ πρότειναν κάποιους κώδικες για το συσχετισμό λέξεων ανάμεσα σε διαφορετικές γλώσσες. Οι προτάσεις αυτές παρέμειναν σε θεωρητικό επίπεδο και μόνο αρκετά χρόνια αργότερα ,στη δεκαετία του 1930 έγιναν προσπάθειες για τη δημιουργία μεταφραστικών μηχανών με περιορισμένη ωστόσο επιτυχία. Ειδική μνεία πρέπει να γίνει στο σημείο αυτό και στον γλωσσολόγο, σημειολόγο και φιλόσοφο Ferdinard de Saussure[2], του οποίου οι ιδέες για το νόημα της γλώσσας και το πώς αυτό ανακύπτει από σχέσεις και αντιθέσεις στο πλαίσιο αμοιβαία διαμοιρασθέντων ενυπαρχόντων δομών, οδήγησαν στην εξέλιξη της γλωσσολογίας και της σημειολογίας στις αρχές του 20ου αιώνα και αποτέλεσαν προπομπό του στρουκτουραλισμού. Σημαντικός σταθμός για το πεδίο αποτελεί η δημοσίευση του Άγγλου μαθηματικού και θεωρητικού της επιστήμης των υπολογιστών Alan Turing τη δεκαετία του 1950 με τίτλο "Computing Machinery and Intelligence"[3]. Ο Turing θα εισαγάγει σε αυτή τη δημοσίευση το περίφημο Turing test ως κριτήριο ευφυίας ενός υπολογιστικού συστήματος. Για την επιτυχία στο τεστ το υπολογιστικό σύστημα θα πρέπει να συνδιαλαγεί γραπτώς με έναν άνθρωπο κατά τέτοιο τρόπο ώστε να νομίζει ο άνθρωπος ότι πρόκειται για συνάνθρωπό του. Εν συνεχεία, το 1957 ο Αμερικάνος γλωσσολόγος και φιλόσοφος Noam Chomsky με το βιβλίο του "Συντακτικές Δομές" [4] θα εισαγάγει ένα σύστημα κανόνων βασιζόμενο σε συντακτικές δομές, που θα βρεθεί στο επίκεντρο για τις επόμενες δεκαετίες και που θα επηρεάσει την πρακτική προσέγγιση στο πεδίο.

## 2.3 Οι διάφορες προσεγγίσεις στην Επεξεργασία Φυσικής Γλώσσας

Τον 20ο και 21ο αιώνα διακρίνονται τρεις επιμέρους προσεγγίσεις στον τομέα. Συγκεκριμένα [5]:

- **Symbolic NLP**, όπου χρησιμοποιούνται αφενός σύμβολα για την αναπαράσταση φυσικών οντοτήτων και αφετέρου Λογική για τη δημιουργία κανόνων, που θα ακολουθηθούν από τη μηχανή για την περάτωση της εκάστοτε εργασίας. Η παραπάνω προσέγγιση ήταν η κυρίαρχη για σχεδόν τέσσερις δεκαετίες(1950 - 1990).
- **Statistical NLP**, όπου γίνεται χρήση Στατιστικής και αλγορίθμων Μηχανικής Μάθησης σε πραγματικά σώματα κειμένων που υπάρχουν στη βιβλιογραφία, με σκοπό την εξαγωγή στατιστικών συμπερασμάτων.Η προσέγγιση αυτή εφορμάται εν μέρει και από τη "στατιστική επανάσταση" που έλαβε χώρα τις δεκαετίες 1980 και 1990, αλλά και από την αύξηση της υπολογιστικής ισχύος. Ένας ακόμα σημαντικός παράγοντας είναι και η σταδιακή ελάττωση της επιρροής της Γλωσσολογίας του Chomsky, που μέχρι εκείνο το σημείο υπήρξε κραταιά και περιόριζε τη χρήση πραγματικών σωμάτων κειμένου.
- **Neural NLP**, όπου γίνεται χρήση Αναπαραστάσεων και Βαθιάς Μηχανικής Μάθησης, στο πλαίσιο της γενικότερης εξέλιξης του πεδίου της Μηχανικής Μάθησης, εν μέρει λόγω της αυξημένης υπολογιστικής ισχύος και εν μέρει λόγω την ανάπτυξης νέων αποδοτικότερων αλγορίθμων. Μια τέτοια προσέγγιση αποδείχθηκε ότι μπορεί να οδηγήσει σε πολύ καλά αποτελέσματα και στον τομέα της Επεξεργασίας Φυσικής Γλώσσας. Ένα σημαντικό πλεονέκτημα είναι ότι πλέον δεν απαιτείται η εξαντλητική εύρεση ή δημιουργία χαρακτηριστικών, κάτι που αφήνεται πλέον στο δίκτυο.

## 2.4 Οι επιμέρους εφαρμογές στην Επεξεργασία Φυσικής Γλώσσας

Κατά την αρχική ενασχόληση των ερευνητών με το πεδίο το ενδιαφέρον επικεντρώθηκε στην Μηχανική Μετάφραση μεταξύ δύο γλωσσών. Ωστόσο στην πορεία και καθώς η θεωρία εξελισσόταν έγινε αντιληπτό ότι το παρόν πεδίο έχει δυνητικά πάρα πολλές εφαρμογές. Ενδεικτικά αναφέρονται οι ακόλουθες[5]:

- **Machine Translation**, η μετάφραση δηλαδή λέξεων, προτάσεων και ολόκληρων κειμένων ανάμεσα σε δύο ή περισσότερες γλώσσες
- **Sentiment Analysis**, όπου μια πρόταση επιχειρείται να ταξινομηθεί ανάλογα με το εάν έχει θετική, αρνητική ή ουδέτερη χροιά
- **Automatic Speech Recognition(ASR)**, η αυτόματη αναγνώριση του προφορικού λόγου από τη μηχανή
- **Named Entity Recognition**, η αυτόματη δηλαδή αναγνώριση οντοτήτων σε ένα κείμενο και η κατηγοριοποίηση τους(όνομα, πόλη , χώρα μεταξύ άλλων)
- **Speech-to-text and text-to-speech**, μεταγραφή προφορικού λόγου σε γραπτό και τούμπαλιν
- **Optical Character Recognition**, όπου στόχος είναι η αυτόματη αναγνώριση του κειμένου από μια φωτογραφία

- **Document Summarization**, όπου δοθέντος ενός μεγάλου κειμένου δημιουργείται μια αυτόματη σύνοψη

# 3

# Οι σημαντικότερες εξελίξεις στην Επεξεργασία Φυσικής Γλώσσας: από τα one-hot encodings στο BERT

Οι περισσότεροι αλγόριθμοι Μηχανικής Μάθησης και όλες σχεδόν οι αρχιτεκτονικές Βαθιάς Μηχανικής Μάθησης δεν δύνανται να επεξεργαστούν απλό κείμενο στην αυθεντική μορφή του. Συνήθως για τη διενέργεια οποιασδήποτε εργασίας, απαιτούνται ως είσοδοι αριθμοί. Ως εκ τούτου καθίσταται αναγκαία η αναπαράσταση του κειμένου υπό μορφή αριθμών. Αυτό επιτυγχάνεται γενικά μέσω των word embeddings. Ουσιαστικά πρόκειται για την αναπαράσταση μιας λέξης -συνήθως- με ένα αριθμητικό διάνυσμα. Αυτό μπορεί να επιτευχθεί με διαφορετικούς τρόπους, οι κυριότεροι εκ των οποίων θα παρουσιαστούν στη συνέχεια. Απώτερος στόχος είναι να αποτυπωθεί το εννοιολογικό και συντακτικό περιεχόμενο της εκάστοτε λέξης, με τις λέξεις που έχουν παρόμοια έννοια και χρήση να βρίσκονται όσο κατά το δυνατόν εγγύτερα στο διανυσματικό χώρο.

Η εξέλιξη στις τεχνικές κατασκευής των embeddings οδήγησε εν συνεχεία στην ανάπτυξη νέων τεχνικών και μοντέλων, με σημαντικά καλύτερες επιδόσεις. Τα σημαντικότερα από αυτά παρουσιάζονται με τη σειρά στο κεφάλαιο αυτό.

## 3.1 One-hot encodings

Μια από τις αρχικές προσεγγίσεις αποτέλεσε η χρήση one-hot διανυσμάτων-που είναι εύκολο να κατασκευαστούν- για την αραιή αναπαράσταση των λέξεων. Η προσέγγιση αυτή είχε σημαντικά μειονεκτήματα που περιόρισαν τη χρήση τους σε πολύ ορισμένες μόνο εργασίες, όπως η ομοιότητα μεταξύ αρχείων. Κυριότερο μειονέκτημα αποτελεί το γεγονός ότι δεν προκύπτει ομοιότητα μεταξύ των λέξεων καθότι το εσωτερικό γινόμενο είναι πάντοτε μηδενικό. Στο Σχήμα φαίνεται ο πίνακας της one-hot αναπαράστασης της φράσης "i ate an apple and played the piano"[6].

## 3.2 Word2vec

Η τεχνική αυτή εισήχθη το 2013 με τη δημοσίευση "Efficient Estimation of Word Representations in Vector Space" [7]. Χρησιμοποιούνται συγκεκριμένα νευρωνικά δίκτυα που εκπαιδεύονται σε ένα μεγάλο σώμα κειμένων προκειμένου να παράγουν ένα διανυσματικό χώρο

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| I | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ate | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| an | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| apple | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| and | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| played | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| the | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| piano | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Σχήμα 3.1: Η one-hot αναπαράσταση της φράσης "i ate an apple and played the piano"

και να αντιστοιχίσουν κάθε λέξη σε ένα αριθμητικό διάνυσμα. Αποδεικνύεται ότι οι διανυσματικές αυτές αναπαραστάσεις είναι ικανές να αποτυπώσουν τόσο το σημασιολογικό όσο και το συντακτικό περιεχόμενο των λέξεων. Έτσι παρεμφερείς εκφράσεις βρίσκονται κοντά στο διανυσματικό χώρο με βάση τη συνημιτονική ομοιότητα και με τη χρήση απλών αλγεβρικών εκφράσεων μπορούν να καταδείξουν ομοιότητες ή αναλογίες.

Τα νευρωνικά δίκτυα που χρησιμοποιούνται είναι ρηχά και συγκεκριμένα δύο επιπέδων. Προτείνονται γενικά δύο διαφορετικές αρχιτεκτονικές οι οποίες και θα παρουσιαστούν ακολούθως.

### 3.2.1 Continuous bag-of-words(CBOW)

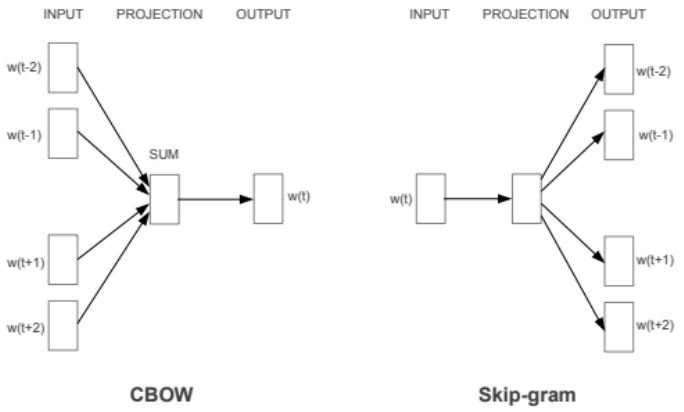
Στο μοντέλο αυτό αντικείμενο της εκπαίδευσης αποτελεί η πρόβλεψη μιας κεντρικής λέξης από τις γειτονικές της εκατέρωθεν. Χρησιμοποιείται ένα παράθυρο για να προσδιοριστεί ο αριθμός των λέξεων -δεξιά και αριστερά- που θα συμπεριληφθούν στην πρόβλεψη. Η σειρά των λέξεων δεν επηρεάζει την πρόβλεψη. Έτσι λέξεις που είναι εξίσου πιθανό να προβλεφθούν, καταλήγουν να έχουν και περεμφερή αναπαράσταση, άρα να βρίσκονται "κοντά" διανυσματικά. Βασικό πλεονέκτημα αυτού του μοντέλου είναι ότι είναι αρκετά γρήγορο στην εκπαίδευση.

### 3.2.2 Continuous skip-gram

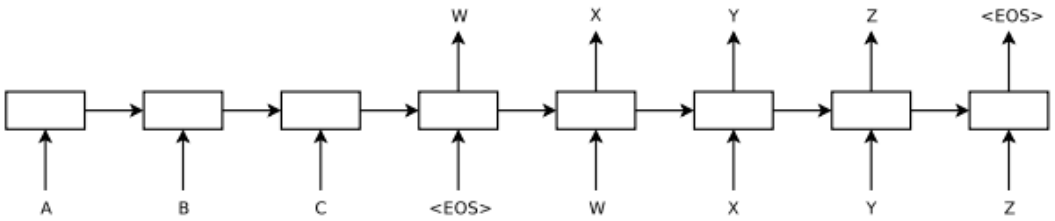
Αντίθετα με το παραπάνω μοντέλο, εδώ αντικείμενο της εκπαίδευσης αποτελεί η πρόβλεψη των γειτονικών λέξεων από μια κεντρική λέξη. Το μοντέλο αυτό είναι πιο αργό σε σχέση με το παραπάνω, ωστόσο δίνει καλύτερα αποτελέσματα σε λέξεις που δε χρησιμοποιούνται συχνά, κάτι που αποτελεί και το δυνατό του σημείο.

## 3.3 GloVe

Ένα χρόνο αργότερα ακολούθησε η δημοσίευση "GloVe: Global Vectors for Word Representation"[8] από το πανεπιστήμιο του Stanford, που εφάρμοσε μια λίγο διαφορετική προσέγγιση.




Σχήμα 3.2: Οι δύο αρχιτεκτονικές

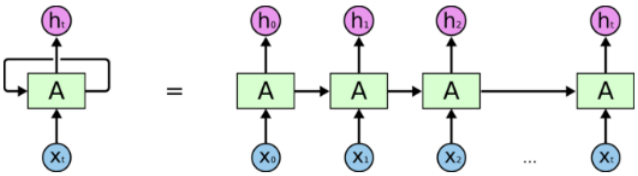



Σχήμα 3.3: Η αρχιτεκτονική του μοντέλου Seq2seq. Στο παράδειγμα δίνεται ως είσοδος η ακολουθία "ABC" και παράγεται στην έξοδο η ακολουθία "WXYZ". Το μοντέλο παύει να κάνει προβλέψεις όταν δώσει ως έξοδο το token ¡EOS¿

Συγκεκριμένα, ενώ το Word2vec κατασκευάζει τα embeddings συσχετίζοντας λέξεις στόχους με τις εκατέρωθεν λέξεις περιεχομένου, δεν δίνει σημασία στο πόσο συχνά εμφανίζονται οι λέξεις περιεχομένου. Έτσι λέξεις που εμφανίζονται συχνά από κοινού παράγουν περισσότερα δείγματα εκπαίδευσης, αλλά δεν παρέχουν καμία επιπλέον πληροφορία. Οι ερευνητές του Stadfort κατέληξαν στο ότι αυτή η συχνή από κοινού εμφάνιση λέξεων θα πρέπει να αξιοποιηθεί, με επακόλουθο τα διανύσματα που παράγονται με αυτή την τεχνική να αντιστοιχίζονται κατευθείαν στην πιθανότητα συνύπαρξης των αντίστοιχων λέξεων στο κείμενο.

## 3.4 Seq2seq

Εισήχθη το 2014 από την Google και βασικό αντικείμενο, όπως μπορεί να συναχθεί και από το όνομα του μοντέλου, είναι η μετατροπή μιας ακολουθίας λέξεων σε μια νέα, σε αντιδιαστολή με τις προηγούμενες μεθόδους όπου η επεξεργασία γινόταν ανά λέξη. Πιο συγκεκριμένα, γίνεται η χρήση ενός LSTM δικτύου για την κωδικοποίηση της ακολουθίας εισόδου σε ένα ενδιάμεσο διάνυσμα σταθερού μεγέθους. Εν συνεχεία, χρησιμοποιείται ένα ακόμα βαθύ LSTM δίκτυο για την αποκωδικοποίηση του διανύσματος εκ νέου σε μια νέα ακολουθία. Η παραπάνω προσέγγιση οδήγησε σε θεαματικά αποτελέσματα, ειδικά στο πεδίο της αυτόματης μετάφρασης ανάμεσα σε δύο γλώσσες. Χρησιμοποιήθηκε και σε άλλες εφαρμογές όπως σε chatbots και σε σύνοψη κειμένου με παρόμοια επιτυχία. Είναι αξιοσημείωτο ότι δεν χρειάζονται επιμέρους συμπεράσματα για τη φύση των προς επεξεργασία ακολουθιών, συνεπώς η ικανότητα γενίκευσης σε πολλές διαφορετικές εργασίες αυξάνεται.




Σχήμα 3.4: Η παραγωγή των κρυφών καταστάσεων

## 3.5 FastText

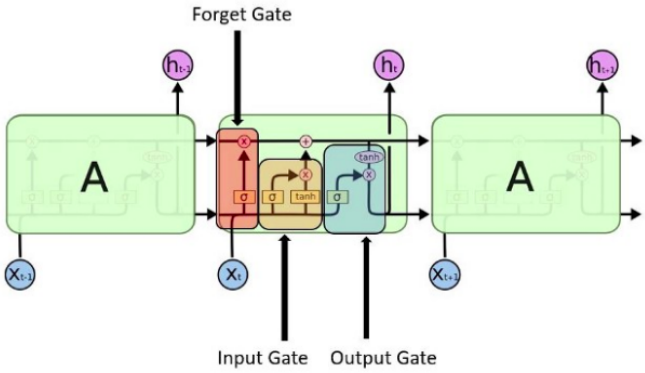
Το 2016 το Facebook βελτίωσε τις ήδη υπάρχουσες τεχνικές του Word2vec. Συγκεκριμένα με την νέα προσέγγιση, τα embeddings κατασκευάζονται με βάση έναν συνδυασμό από μικρότερες δομικές μονάδες, δηλαδή τμήματα λέξεων ή και σκέτους χαρακτήρες. Με αυτόν τον τρόπο καθίσταται δυνατή η γενίκευση σε καινούριες λέξεις, αρκεί αυτές να έχουν ίδιους χαρακτήρες με ήδη γνωστές. Επιπρόσθετα, για την παραγωγή των embeddings απαιτούνται πλέον αρκετά λιγότερα δεδομένα εκπαίδευσης καθώς αντλείται σημαντικά περισσότερη πληροφορία από την ίδια ποσότητα κειμένου. Αυτός είναι και ο λόγος που υπάρχουν περισσότερα προ-εκπαιδευμένα δίκτυα για διάφορες γλώσσες σε σχέση με άλλους παρόμοιους αλγορίθμους.

## 3.6 RNN και LSTM

Τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (Recurrent Neural Networks) είναι σχεδιασμένα για το χειρισμό και επεξεργασία ακολουθιακής πληροφορίας. Οι παρελθοντικές πληροφορίες αποθηκεύονται σε μεταβλητές και δίνονται ως είσοδοι σε επόμενα στάδια του δικτύου, ώστε μαζί με την τρέχουσα είσοδο να καθορίζουν και την επόμενη έξοδο. Σε αντίθεση με τα Δίκτυα Πρόσθιας Τροφοδότησης, τα RNN περιλαμβάνουν λοιπόν εσωτερικές, "κρυφές" καταστάσεις εν είδει μνήμης, για την επεξεργασία ακολουθιακών εισόδων. Σε κάθε χρονικό βήμα επαναλαμβάνεται ο υπολογισμός της κρυφής κατάστασης και από εκεί προέρχεται και το όνομα του Δικτύου. Συγκεκριμένα συνενώνεται η είσοδος του βήματος με την προηγούμενη κρυφή κατάσταση, δηλαδή με την κρυφή κατάσταση του προηγούμενου χρονικού βήματος, και το αποτέλεσμα τροφοδοτείται σε ένα πλήρως συνδεδεμένο δίκτυο με μια συνάρτηση ενεργοποίησης. Παράγεται έτσι η κρυφή κατάσταση του τρέχοντος χρονικού βήματος και εντελώς ανάλογα ακολουθείται η ίδια διαδικασία στα επόμενα βήματα. Τα RNNs χρησιμεύουν πολύ όταν δουλεύουμε με ακολουθιακά δεδομένα όπως για παράδειγμα είναι και ένα γραπτό κείμενο. Ωστόσο έχουν κάποια σημαντικά μειονεκτήματα. Κυριότερο εξ αυτών είναι ότι δεν μπορούν να διατηρήσουν πληροφορίες πολύ μακροπρόθεσμα. Όταν μια ακολουθία είναι πολύ μεγάλη το δίκτυο τείνει να "ξεχνάει" την πληροφορία των πρώτων χρονικών βημάτων και άρα να επικεντρώνεται στα τελευταία, άρα να εστιάζει σε συγκεκριμένο τμήμα της ακολουθίας.

Μια μικρή τροποποίηση πάνω στη λογική των RNNs, για την αντιμετώπιση και του παραπάνω προβλήματος είναι τα δίκτυα LSTM (Long Short Term Memory). Τα δίκτυα αυτά διαθέτουν κελί μνήμης παρόμοιας λογικής με την κρυφή κατάσταση των RNN δικτύων. Παρόλα αυτά, για τον έλεγχο του κελιού εισάγονται οι ακόλουθες τρεις πύλες:

- **Πύλη εισόδου**. Καθορίζει ποια τιμή από την είσοδο θα πρέπει να χρησιμοποιηθεί για να τροποποιήσει τη μνήμη. Η σιγμοειδής συνάρτηση αποφαίνεται για το ποιες τιμές θα περάσουν και εν συνεχεία μέσω της συνάρτησης υπερβολικής εφαπτομένης tanh δίνεται βάρος στις τιμές αυτές από -1 μέχρι 1, ανάλογα με τη σημαντικότητά τους. Η εξίσωσή της έχει τη μορφή:




Σχήμα 3.5: Το κελί μνήμης του LSTM

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

- **Πύλη λήθης**. Καθορίζει ποιες πληροφορίες του κελιού θα αποπεμφθούν. Η απόφαση λαμβάνεται με βάση τη σιγμοειδή συνάρτηση, στην οποία δίνεται ως είσοδος η προηγούμενη κρυφή κατάσταση και η τρέχουσα κατάσταση εισόδου και λαμβάνεται ως έξοδος μια τιμή μεταξύ του 0(διαγραφή) και του 1(αποθήκευση) για κάθε αριθμό της κατάστασης του κελιού. Η εξίσωσή της έχει τη μορφή:

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

- **Πύλη εξόδου**. Η είσοδος και η μνήμη του κελιού χρησιμοποιούνται για τη δημιουργία της εξόδου. Η σιγμοειδής συνάρτηση αποφαίνεται για το ποιες τιμές θα περάσουν και εν συνεχεία μια tanh συνάρτηση ζυγίζει τις τιμές αυτές, δίνοντάς τους τιμές από -1 μέχρι 1, καθορίζοντας έτσι και το επίπεδο της σημαντικότητάς τους. Οι τιμές πολλαπλασιάζονται εν συνεχεία με την έξοδο της σιγμοειδούς. Η εξίσωσή της έχει τη μορφή:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

## 3.7 Μηχανισμός Προσοχής

Ο Μηχανισμός Προσοχής(Attention Mechanism) εμπνεύστηκε εν μέρει από τον τρόπο που εφιστάται η προσοχή του ανθρώπου σε ένα συγκεκριμένο τμήμα μιας εικόνας ή από το πώς συσχετίζουμε λέξεις σε μια πρόταση. Ιδιαίτερα η οπτική προσοχή εστιάζει σε συγκεκριμένη περιοχή της εικόνας την οποία και βλέπουμε σε "υψηλή ανάλυση", με την υπόλοιπη εικόνα να γίνεται αντιληπτή σε "χαμηλή ανάλυση". Το σημείο εστίασης δύναται να αλλάζει κατά το δοκούν, με το υπόλοιπο της εικόνας να συνάγεται. Αντίστοιχα, σε μια πρόταση μπορούμε να εξηγήσουμε τη σχέση ανάμεσα σε δύο λέξεις. Βλέποντας για παράδειγμα το ρήμα "τρώω" είμαστε έτοιμοι να εστιάσουμε σε μια λέξη-οντότητα της κατηγορίας φαγητό παρά σε κάποια άλλη λέξη της πρότασης. Σε αδρές γραμμές, η Προσοχή στο πλαίσιο της Βαθιάς Μηχανικής Μάθησης μεταφράζεται ως ένα αριθμητικό διάνυσμα με βάρη σημαντικότητας. Για να προβλεφθεί λόγου χάρη μια λέξη σε μια πρόταση, εκτιμάται με τη χρήση του διανύσματος Προσοχής

πόσο ισχυρά συσχετίζεται με τις υπόλοιπες λέξεις και λαμβάνεται το άθροισμα των τιμών τους πολλαπλασιασμένο με το διάνυσμα Προσοχής ως η προσέγγιση της λέξης.

Το πρώτο μοντέλο με χρήση Μηχανισμού Προσοχής προτάθηκε το 2015 στη δημοσίευση "Neural Machine Translation by Jointly Learning to Align and Translate" [9]. Αποτελούσε μια βελτίωση στην αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή με χρήση RNNs και LSTMs που χρησιμοποιείτο στη Νευρωνική Μηχανική Μετάφραση έως τότε. Συγκεκριμένα, ο κωδικοποιητής και ο αποκωδικοποιητής αποτελούνταν από στρώσεις LSTMs. Ο κωδικοποιητής ήταν υπεύθυνος για τη μετατροπή της ακολουθίας εισόδου σε ένα διάνυσμα περιεχομένου, που αποτελεί και την ενδιάμεση αναπαράσταση. Αυτό το διάνυσμα είναι η έξοδος του τελευταίου LSTM και αποτελεί μια επαρκή σύνοψη του διανύσματος εισόδου. Ο αποκωδικοποιητής αντίστοιχα παράγει τις λέξεις της εξόδου τη μια μετά την άλλη. Το μεγάλο μειονέκτημα αυτής της προσέγγισης είναι ότι αντιμετωπίζει προβλήματα με μακρύτερες ακολουθίες, δυσκολεύεται δηλαδή να κατασκευάσει μια επαρκή αναπαράσταση τους καθώς δε θυμάται όλες τις λέξεις τους, ειδικά αυτές που ήταν πολλά time-steps νωρίτερα. Έτσι εφιστάται μεγαλύτερη προσοχή στα τελευταία τμήματα μιας ακολουθίας, σε αντίθεση με το πως επεξεργάζεται την πληροφορία ο ανθρώπινος εγκέφαλος. Αυτό οφείλεται στο πολύ γνωστό πρόβλημα του vanishing/exploding gradient. Με το μηχανισμό της Προσοχής αυτοί οι σκόπελοι ξεπερνώνται, καθώς το διάνυσμα περιεχομένου έχει πλέον πρόσβαση σε όλη την ακολουθία εισόδου, αντί μονάχα στο τελευταίο μέρος της κάθε φορά. Ακολουθεί μια σύντομη περιγραφή του Μηχανισμού της Προσοχής σε πιο χαμηλό επίπεδο, βασισμένο κυρίως στη δημοσίευση [10].

- Στο πρώτο βήμα, κατασκευάζονται τρία διανύσματα για κάθε embedding της εισόδου, συνήθως δηλαδή για κάθε λέξη εισόδου. Τα διανύσματα αυτά ονομάζονται διάνυσμα Ερώτημα(Query Vector), διάνυσμα Κλειδί(Key Vector) και διάνυσμα Τιμή(Value Vector) και εμπνέονται με τη σειρά τους από το πώς αντιλαμβάνεται ο άνθρωπος τα σήματα στα οποία δίνει προσοχή(θελληματικά και μη θελληματικά). Τα διανύσματα αυτά κατασκευάζονται από τον πολλαπλασιασμό του εκάστοτε embedding εισόδου με τρεις αντίστοιχους πίνακες των οποίων τα βάρη αποκτώνται διαμέσου της διαδικασίας εκπαίδευσης.

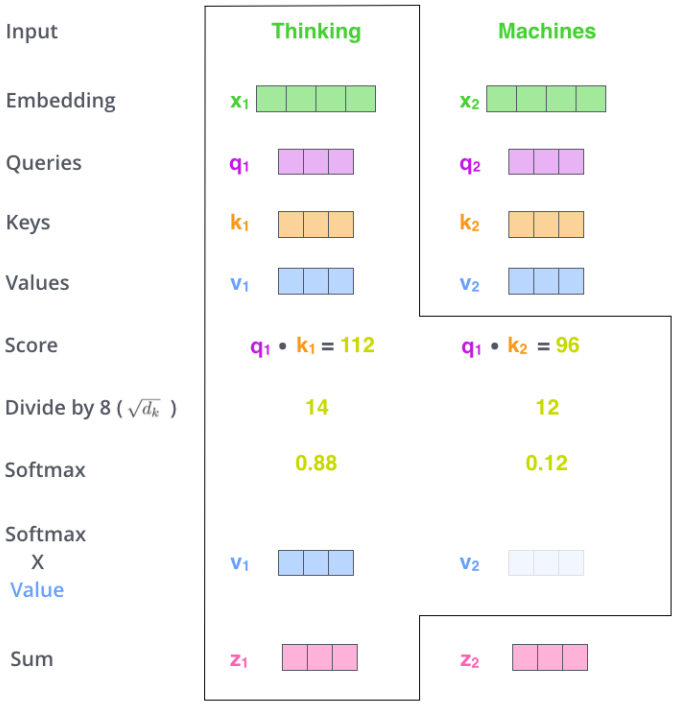
- Στο δεύτερο βήμα, για κάθε λέξη εισόδου υπολογίζεται ένα σκορ για όλες τις υπόλοιπες λέξεις της ακολουθίας. Το σκορ αυτό υποδηλώνει το πόσο και σε ποιες λέξεις της ακολουθίας θα πρέπει να επικεντρωθούμε για τη δεδομένη λέξη, τι συσχέτιση εν ολίγοις υπάρχει. Υπολογίζεται λαμβάνοντας το εσωτερικό γινόμενο μεταξύ του διανύσματος Ερώτημα της λέξης στόχου και του διανύσματος Κλειδί της εκάστοτε λέξης της ακολουθίας.

- Στο τρίτο βήμα, διαιρούνται τα σκορ με έναν αριθμό, για την απόκτηση πιο σταθερών παραγώγων.

- Στο τέταρτο βήμα, δίνονται τα αποτελέσματα σε μια softmax για να καταστούν όλα θετικά και να αθροιστούν στη μονάδα. Στις λέξεις με τα μεγαλύτερα σκορ θα πρέπει να εφισταθεί περισσότερο η προσοχή για τη λέξη στόχο.

- Στο πέμπτο βήμα, πολλαπλασιάζεται κάθε διάνυσμα Τιμής με το σκορ του παραπάνω βήματος, ώστε να δοθεί σημασία στις τιμές που έχουν υψηλότερα σκορ και να υποβαθμιστεί η σημασία των υπολοίπων.

- Στο έκτο βήμα, αθροίζονται τα ζυγισμένα διανύσματα Τιμής και παράγεται το αποτέλεσμα της στρώσης για μια δεδομένη λέξη-θέση.




Σχήμα 3.6: Παράδειγμα υπολογισμού attention για την πρόταση "Thinking Machines"

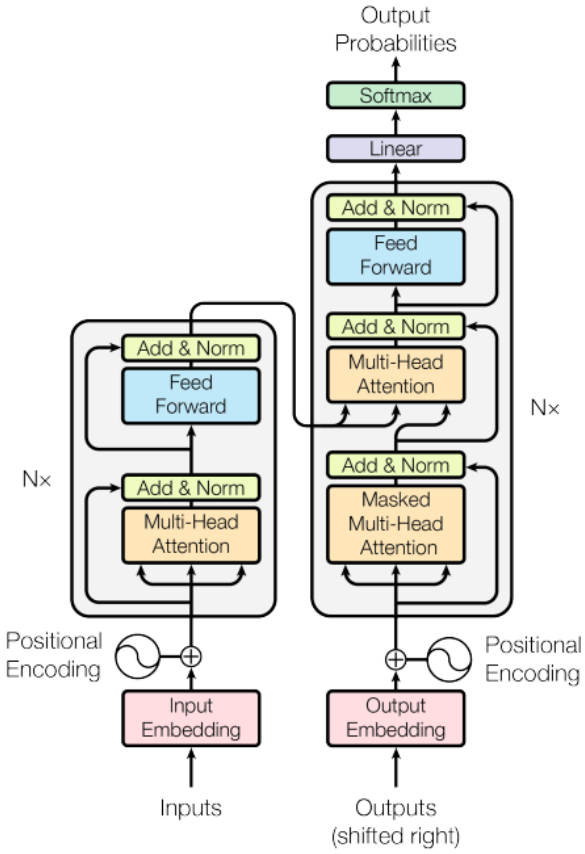
Αξίζει να σημειωθεί ότι μπορούν να χρησιμοποιηθούν και πολλές τριάδες τέτοιων διανυσμάτων, παρέχοντας έτσι περισσότερους χώρους Αναπαράστασης και δίνοντας ακόμα ακριβέστερα αποτελέσματα. Το σύστημα της Προσοχής καθίσταται γενικά πολύ αποτελεσματικό και άνοιξε το δρόμο για νέες επαναστατικές αρχιτεκτονικές και βελτίωση του SOTA.

## 3.8 Transformer

Το 2017 με τη δημοσίευση "Attention is all you need" [10] της εταιρίας Google εισάγεται το μοντέλο Transformer. Το μοντέλο αυτό, ενώ διατηρεί την κωδικοποίηση και αποκωδικοποίηση της εισόδου που έχει παρατηρηθεί και σε παρελθοντικά μοντέλα, ακολουθεί μια διαφορετική προσέγγιση. Συγκεκριμένα, χρησιμοποιεί για πρώτη φορά αποκλειστικά μηχανισμούς attention χωρίς την από κοινού χρήση RNN δικτύων. Αυτό είναι σημαντικό, καθώς παρά την ευρεία χρήση των RNN δικτύων και τις θεαματικές τους επιδόσεις σε σχέση με προηγούμενες προσεγγίσεις, έγινε γρήγορα αντιληπτό ότι υπόκεινται και αυτά σε περιορισμούς. Συγκεκριμένα παρατηρείται ότι για μεγάλες ακολουθίες συνεχούς κειμένου τα RNN δίκτυα δεν μπορούν να συσχετίσουν εύκολα λέξεις με μεγάλη απόσταση, καθώς επεξεργάζονται το αποτέλεσμα ακολουθιακά. Στην πράξη, μια βαριάντα τους, τα LSTMs περιόρισε αυτό το πρόβλημα σε σημαντικό βαθμό αλλά και πάλι δεν το εξάλειψε.

Στον αντίποδα με τη χρήση στρωμάτων attention η είσοδος επεξεργάζεται γραμμικά εξ ολοκλήρου. Περισσότερες πληροφορίες μαθαίνονται απλά με χρήση πολλαπλών στρώσεων attention ή περισσότερων κεφαλών ανά στρώση. Ουσιαστικά σε κάθε στρώση μπορεί να δίνεται έμφαση σε άλλο τμήμα της ακολουθίας εισόδου και να εμπλουτίζονται με αυτόν τον τρόπο οι σημασιολογικές και συντακτικές λεπτομέρειες.

Η αρχιτεκτονική του μοντέλου για μια στρώση του φαίνεται στο Σχήμα 3.7. Παρατηρούμε ότι αφενός η είσοδος εμπλουτίζεται με πληροφορίες θέσεως υπό μορφή διανύσματος καθώς

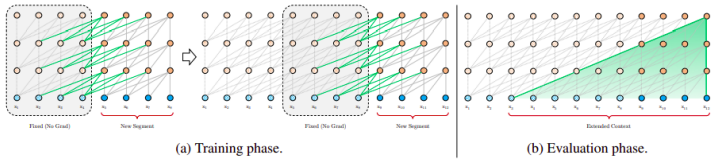



Σχήμα 3.7: Η αρχιτεκτονική του μοντέλου Transformer

αυτή δίνεται όλη μαζί, άρα δεν μπορούν να αντληθούν πληροφορίες θέσης όπως στα ακολουθιακά μοντέλα. Εν συνεχεία, βλέπουμε ότι η κάθε στρώση του κωδικοποιητή αποτελείται από δύο μέρη, αρχικά από μια υποστρώση attention πολλαπλών κεφαλών και ακολούθως από ένα απλό νευρωνικό δίκτυο πρόσθιας τροφοδότησης που παράγει τις εξόδους και τις προωθεί στο επόμενο επίπεδο ή στον αποκωδικοποιητή. Επιπρόσθετα, χρησιμοποιείται και μια υπολειμματική διασύνδεση μαζί με ομαλοποίηση στρώσης. Αυτό εφαρμόζεται για τεχνικούς λόγους καθώς στην πράξη το μοντέλο εκπαιδεύεται πιο γρήγορα και εύκολα. Στην πλευρά του αποκωδικοποιητή κάθε στρώση αποτελείται από μια επιπλέον υποστρώση attention πολλαπλών κεφαλών με χρήση μάσκας. Αυτό συμβαίνει γιατί ο αποκωδικοποιητής σχεδιάζεται ώστε να παράγει στην έξοδο ένα token τη φορά σε βήματα, οπότε η έξοδος κάθε σταδίου δεν θα πρέπει να σχετίζεται με τα επόμενα tokens. Έτσι η μάσκα αποκρύπτει ουσιαστικά τα tokens που ακολουθούν μετατρέποντας τον αποκωδικοποιητή σε αυτοπαλινδρομικό, όπως οι αποκωδικοποιητές που βασίζονταν σε RNN δίκτυα. Έτσι υποστηρίζονται εργασίες όπως αυτόματη μετάφραση που ήταν και η κυριότερη εργασία στην οποία χρησιμοποιήθηκε το μοντέλο αυτό.

## 3.9 Transformer-XL

Το μοντέλο Transformer-XL εισάγεται το 2019 με τη δημοσίευση "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context" [11] από ερευνητές του Πανεπιστημίου του Carnegie Mellon και της Google Brain. Βασικό κίνητρο της ανάπτυξής του αποτέλε-




Σχήμα 3.8: Το μοντέλο Transformer-XL με μήκος τμήματος 4

σαν τα προβλήματα που αντιμετωπίζει ο απλός Transformer στη μοντελοποίηση αλληλοεξαρτήσεων ανάμεσα σε ακολουθιακά δεδομένα πολύ μακριών ακολουθιών, στο πλαίσιο της ανάπτυξης Γλωσσικών Μοντέλων. Συγκεκριμένα το πρόβλημα εντοπίζεται στο ότι ο Transformer εκπαιδεύει το Γλωσσικό Μοντέλο σε χωριστά τμήματα σταθερού μεγέθους μερικών εκατοντάδων χαρακτήρων, χωρίς να υπάρχει ροή πληροφορίας ανάμεσα τους. Αυτό έχει ως αποτέλεσμα:

- Να μην μπορεί το μοντέλο να συλλάβει εξαρτήσεις με μέγεθος μεγαλύτερο από το σταθερό που έχει οριστεί.
- Τα τμήματα που προκύπτουν ως απλές ακολουθίες συμβόλων να μη λαμβάνουν υπόψη τα όρια των προτάσεων ή γενικά οποιαδήποτε σημασιολογικά όρια. Με αυτόν τον τρόπο χάνονται σημαντικές πληροφορίες περιεχομένου, απαραίτητες για την πρόβλεψη των αρχικών συμβόλων. Το πρόβλημα αυτό αναφέρεται ως κατακερματισμός περιεχομένου.

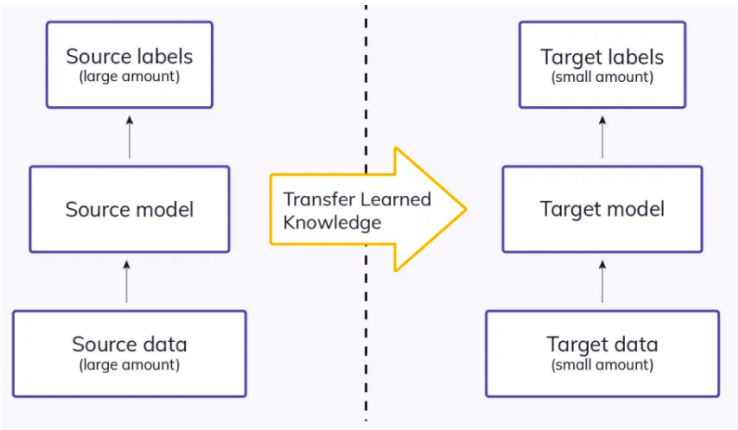
Για την αντιμετώπιση των παραπάνω οι ερευνητές ακολούθησαν με το νέο μοντέλο δύο καινοτόμες προσεγγίσεις. Συγκεκριμένα:

- Εισάγεται η επαναληψιμότητα, υπό την έννοια ότι πλέον, αναπαραστάσεις από προηγούμενα τμήματα καθίστανται προσβάσιμες και από τα αμέσως επόμενα. Αυτή η επιπλέον διασύνδεση αυξάνει το μήκος της μεγαλύτερης ακολουθίας κατά N φορές, όπου N είναι το βάθος του δικτύου. Επιπρόσθετα επιλύεται το πρόβλημα του Κατακερματισμού Περιεχομένου αφού παρέχεται περιεχόμενο στα tokens του εκάστοτε νέου τμήματος.
- Εισάγονται σχετικά encodings θέσεως, σε αντιδιαστολή με τα απόλυτα που χρησιμοποιούνται στον αρχικό Transformer. Έτσι επιτρέπεται η επαναχρησιμοποίηση κατάστασης χωρίς την πρόκληση χρονικής σύγχυσης.

Τα παραπάνω φαίνονται και στο Σχήμα 3.8. Το νέο μοντέλο πέτυχε βελτίωση του SOTA σε διάφορα benchmarks. Επιπρόσθετα ήταν πιο γρήγορο από τον απλό Transformer, μάθαινε εξαρτήσεις από μεγαλύτερες ακολουθίες και ως αποτέλεσμα σημείωνε καλύτερες επιδόσεις σε αυτές.

## 3.10 Η Μεταφορά Μάθησης

Η γενική ιδέα της Μεταφοράς Μάθησης είναι να χρησιμοποιηθεί γνώση που αποκτάται από εφαρμογές για τις οποίες υπάρχει μια αφθονία δεδομένων εκπαίδευσης, σε μια άλλη περιοχή για την οποία αντίστοιχα δεν υπάρχουν πολλά δεδομένα με labels. Έτσι η πιο γενική αυτή γνώση συμβάλλει στην αύξηση της ικανότητας γενίκευσης και στη μείωση του χρόνου εκπαίδευσης ενός συγκεκριμένου μοντέλου στην περιοχή αυτή. Αυτή η μέθοδος είναι πολλές φορές κεφαλαιώδους σημασίας στην εκπαίδευση ενός μοντέλου καθώς για πολλές εφαρμογές η εύρεση των -μεγάλου όγκου- αξιόπιστων δεδομένων που απαιτούνται για την εκπαίδευση




Σχήμα 3.9: Η βασική ιδεά της Μεταφοράς Μάθησης

του, καθίσταται μια αρκετά δύσκολη διαδικασία χρονικά και οικονομικά. Επιπρόσθετα μπορεί πολλές φορές να μην υπάρχουν οι υπολογιστικοί πόροι για την εκπαίδευση ενός τόσο μεγάλου σετ δεδομένων. Η διαδικασία αυτή φαίνεται και στο Σχήμα 3.9

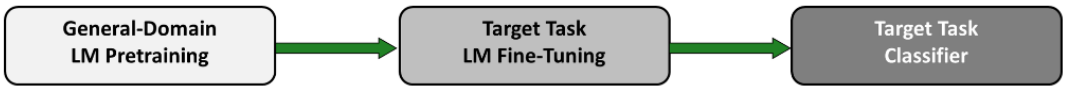
Το μοντέλο λοιπόν μαθαίνει μετά από εντατική εκπαίδευση ενδιάμεσες αναπαραστάσεις των δεδομένων, έχει δηλαδή ήδη αποκτήσει κάποια βάρη. Εν συνεχεία μπορεί να χρησιμοποιηθεί ως έχει σε μια υποπεριοχή, με την προσθήκη κάποιων επιπλέον στρώσεων εάν αυτό καθίσταται απαραίτητο από την επιμέρους εφαρμογή. Συνήθως οι επιπλέον στρώσεις μπορεί να είναι για παράδειγμα ένα απλό δίκτυο πρόσθιας τροφοδότησης με μια συνάρτηση softmax ώστε να λειτουργήσει ως ένας τελικός ταξινομητής. Για την τελική εκπαίδευση δύναται να παγώσουν τα βάρη του αρχικού δικτύου και να πραγματοποιηθεί εκπαίδευση μόνο του τελικού ταξινομητή σε έναν μικρό όγκο εξειδικευμένων δεδομένων. Επιπλέον προαιρετικά μπορεί να ακολουθηθεί και το λεγόμενο fine-tuning. Στην περίπτωση αυτή, κατά την τελική εκπαίδευση δεν παγώνουν τα βάρη του αρχικού προεκπαιδευμένου μοντέλου, αλλά εκπαιδεύονται περαιτέρω και αυτά. Καθώς αυτό μπορεί να οδηγήσει σε overfitting, τα αρχικά βάρη εκπαιδεύονται με μικρότερο ρυθμό εκμάθησης, με τις παραγώγους να ανανεώνονται έτσι σημαντικά πιο αργά. Εν συνόψει, μια αρκετά γενική και ενδεικτική διαδικασία μεταφορά μάθησης απαρτίζεται από τα ακόλουθα βήματα:

- Εκπαίδευση ενός αρχικού μοντέλου σε ένα μεγάλο σετ δεδομένων.

- Δημιουργία ενός δεύτερου μοντέλου που αποτελεί αντιγραφή του πρώτου, μαζί με τα βάρη στα οποία αυτό κατέληξε κατά την προεκπαίδευση, μόνο που στο τελευταίο στάδιο προστίθεται συνήθως μια επιπλέον στρώση, που είναι στις περισσότερες περιπτώσεις ένας ταξινομητής.

- Εκπαίδευση του νέου μοντέλου σε ένα μικρότερο σετ δεδομένων, πιο εξειδικευμένων, είτε με πάγωμα των αρχικών βαρών είτε με εκπαίδευση όλων, με χαμηλότερο όμως ρυθμό εκμάθησης για λόγους αποφυγής overfitting.

Η μέθοδος της Μεταφοράς Μάθησης χρησιμοποιήθηκε αρχικά στο πεδίο της Όρασης Υπολογιστών και τα τελευταία χρόνια άρχισε να χρησιμοποιείται ευρέως και στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας, με μοντέλα όπως το ULMFit.




Σχήμα 3.10: Επισκόπηση του μοντέλου

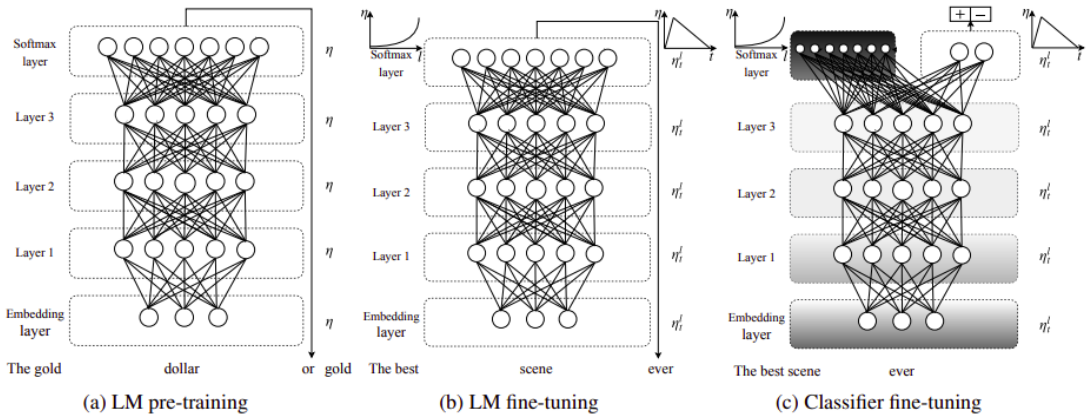
## 3.11 Το μοντέλο ULMFiT (Universal Language Model Fine-tuning for Text Classification)

Το μοντέλο παρουσιάστηκε στη δημοσίευση "Universal Language Model Fine-tuning for Text Classification" του 2018 [12]. Θεωρείται ίσως το πρώτο μοντέλο που εφαρμόζει αποδοτικά τη μεταφορά μάθησης στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας. Βασιζόμενοι στο παράδειγμα άλλων πεδίων της Μηχανικής Μάθησης- και ιδιαίτερα στην Όραση Υπολογιστών- οι ερευνητές γρήγορα συνειδητοποίησαν ότι μια προσέγγιση χρήσης προ-εκπαιδευμένων δικτύων θα μπορούσε να ωφελήσει σημαντικά και στην Επεξεργασία Φυσικής Γλώσσας. Συγκεκριμένα, η εύρεση μεγάλων σετ δεδομένων με labels καθώς και η εκπαίδευση μοντέλων σε αυτά καθίσταται από ένα σημείο και μετά δύσκολη, λόγω χρόνου και περιορισμένων οικονομικών και υπολογιστικών πόρων. Έτσι μόνο πολύ συγκεκριμένοι οργανισμοί και εταιρίες δύνανται να εκπαιδεύσουν τέτοια μοντέλα, με αποτέλεσμα την ανάσχεση της εξέλιξης στο πεδίο. Αντίθετα, με τα προεκπαιδευμένα δίκτυα μπορεί ένας μικρότερος οργανισμός να εκμεταλλευτεί ένα πολύ ισχυρό δίκτυο, απλά εκπαιδεύοντας το σε ένα μικρό -εξειδικευμένο στις ανάγκες του- σετ δεδομένων, κάνοντας απλά fine-tuning στα αρχικά εκπαιδευμένα βάρη. Με το ULMFit καταδεικνύεται για πρώτη φορά ότι κάτι τέτοιο μπορεί να γίνει αποδοτικά. Όπως φαίνεται και στο Σχήμα 3.10, στο πρώτο στάδιο εκπαιδεύεται ένα μοντέλο γλώσσας σε ένα μεγάλο σώμα εγγράφων γενικού πεδίου.

Το μοντέλο μπορεί να προβλέψει πλέον την επόμενη λέξη σε μια ακολουθία με ένα συγκεκριμένο βαθμό βεβαιότητας. Διαισθητικά, είναι σαν να μαθαίνει τα γενικά χαρακτηριστικά της γλώσσας. Στο δεύτερο στάδιο, η γνώση που αποκτήθηκε προηγουμένως χρησιμοποιείται για την εκπαίδευση σε ένα πιο εξειδικευμένο σετ δεδομένων που ανταποκρίνεται στην επιμέρους εφαρμογή. Επειδή ωστόσο τα νέα δεδομένα μπορεί να προέρχονται από διαφορετική κατανομή σε σχέση με τα αρχικά, στο δεύτερο στάδιο πραγματοποιείται fine-tuning δηλαδή εκπαιδεύεται και το αρχικό δίκτυο αλλά με μικρότερο βήμα. Αντικείμενο της εκπαίδευσης και σε αυτό το στάδιο είναι η πρόβλεψη της επόμενης λέξης σε μια ακολουθία. Έτσι το τελικό μοντέλο αποκτά πλέον γνώση και για τα δεδομένα της επιμέρους εφαρμογής. Στο τρίτο στάδιο προστίθενται στρώσεις ανάλογα με το τι πρέπει να επιτευχθεί με την εκάστοτε εφαρμογή. Για παράδειγμα, μπορεί να χρησιμοποιηθεί το δίκτυο για την αναγνώριση συναισθήματος, οπότε στο τελευταίο στάδιο θα προστεθεί σε αυτή την περίπτωση ένα δίκτυο πρόσθιας τροφοδότησης συνοδευόμενο από μια συνάρτηση softmax, προκειμένου η έξοδος να είναι μια κατανομή πιθανότητας ανά τα διαφορετικά αποτελέσματα. Τα τρία στάδια φαίνονται στο Σχήμα 3.11. Με την παραπάνω προσέγγιση, οι ερευνητές βελτίωσαν το SOTA για διάφορες εφαρμογές και άνοιξαν το δρόμο για περαιτέρω αξιοποίηση της μεταφοράς μάθησης.

## 3.12 Το μοντέλο ELMo (Embeddings from Language Models)

Εισήχθη το 2018 με τη δημοσίευση "Deep contextualized word representations" [13]. Το παρόν μοντέλο εξέλιξε σε σημαντικό βαθμό την κατασκευή embeddings κάτι που ήταν κεφαλαιώδες για την περαιτέρω ανάπτυξη του πεδίου. Συγκεκριμένα, τα embeddings των GloVe

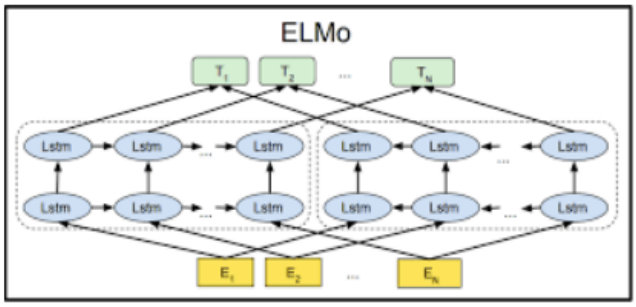



Σχήμα 3.11: Τα τρία στάδια της εκπαίδευσης του μοντέλου

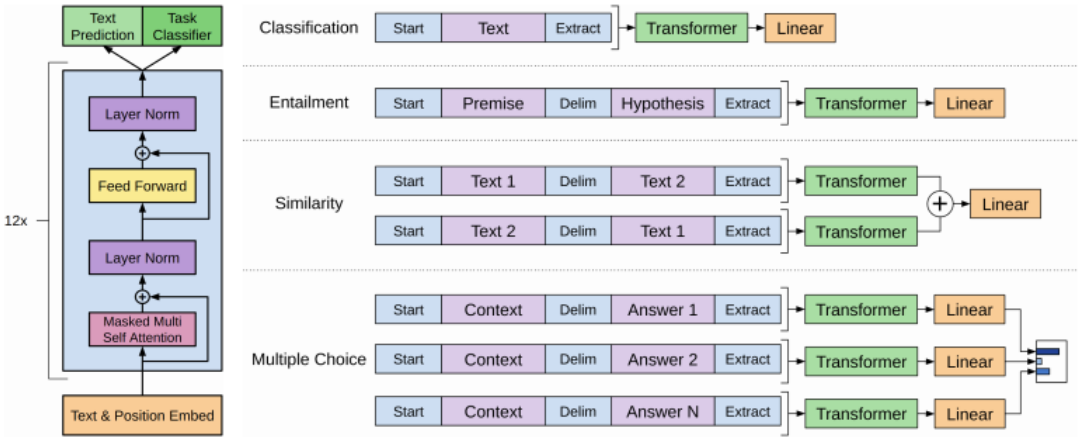
και Word2vec ήταν στατικά υπό την έννοια ότι σε κάθε λέξη μπορούσε να αντιστοιχιστεί μονάχα το ίδιο embedding. Στην περίπτωση του ELMo περνάμε σε πιο δυναμικά embeddings. Έτσι μια λέξη ανάλογα με το περιεχόμενο στο οποίο χρησιμοποιείται μπορεί να έχει διαφορετική έννοια και άρα να αναπαρίσταται με διαφορετικό embedding. Έτσι λαμβάνεται υπόψιν η πολυσημία η οποία από την πρακτική μας εμπειρία ξέρουμε ότι έχει ουσιαστική σημασία στην κατανόηση του λόγου. Ένα επιπλέον χαρακτηριστικό που έκανε το ELMo δημοφιλές είναι και η αμφιδρομικότητά του, όπως φαίνεται και στο Σχήμα 3.12. Μια ακολουθία δηλαδή επεξεργάζεται και από την αρχή και από το τέλος με χρήση δύο παράλληλων LSTMs και τα αποτελέσματα συγχωνεύονται. Αυτό ονομάζεται "ρηχή" αμφιδρομικότητα καθώς στην πραγματικότητα γίνονται δύο παράλληλες μονόδρομες επεξεργασίες. Η αμφιδρομικότητα καθίσταται σημαντικό χαρακτηριστικό καθώς προσομοιάζει πώς διαβάζει στην πραγματικότητα και ο ίδιος ο άνθρωπος, αντλώντας έννοια και από λέξεις που βρίσκονται πιο μπροστά στο κείμενο. Έτσι ένα μονόδρομο μοντέλο χάνει σημαντική ποσότητα πληροφορίας με συνέπεια να έχει χειρότερες επιδόσεις. Το μοντέλο αυτό είναι και από τα πρώτα που δίνει εμπειρικά αποτελέσματα προς την κατεύθυνση των προ-εκπαιδευμένων μοντέλων, σε αντιδιαστολή με την κατασκευή κάθε φορά των embeddings από την αρχή. Αφού προ-εκπαιδευτεί και αποκτήσει μια βασική αντίληψη της γλώσσας, εκπαιδεύεται εν συνεχεία επιβλεπόμενα για επιμέρους εργασίες, με την προσθήκη επιπλέον στρώσεων. Αξίζει να σημειωθεί ότι κατά την εκπαίδευση τα βάρη του προ-εκπαιδευμένου δικτύου παγώνουν. Εν κατακλείδι, το ELMo βελτίωσε σημαντικά το SOTA σε διάφορες εφαρμογές. Ωστόσο η "ρηχή" αμφιδρομικότητα του καθώς και το γεγονός ότι η αρχιτεκτονική του έπρεπε να προσαρμόζεται σε σημαντικό βαθμό στην εκάστοτε εφαρμογή-κάτι που σε πολλές περιπτώσεις δεν είναι εύκολο- άφηνε περιθώρια για περαιτέρω βελτιώσεις στο SOTA.

## 3.13 Τα μοντέλα GPT (Generative Pre-Training)

Πρόκειται για μια οικογένεια μοντέλων που πρωτοεισήχθησαν με τη δημοσίευση "Improving Language Understanding by Generative Pre-Training" [14] το 2018 από το ερευνητικό εργαστήριο της OpenAI. Βασίστηκαν στο μοντέλο του Transformer με χρήση μόνο των αποκωδικοποιητών που αυτό εισήγαγε. Βασική επιδίωξη των ερευνητών κατά τη σχεδίαση του μοντέλου ήταν η προσαρμογή του σε διαφορετικές εφαρμογές με αμελητέες μετατροπές στη γενική αρχιτεκτονική του, προσφέροντας έτσι ένα γενικό πλαίσιο για την επίτευξη ισχυρών επιδόσεων στην κατανόηση της Φυσικής Γλώσσας. Παρατηρείται έτσι ότι για πολλές εφαρμογές απαιτεί-




Σχήμα 3.12: Το μοντέλο ELMo χρησιμοποιεί δύο χωριστά δίκτυα για να επεξεργαστεί το κείμενο αμφίδρομα




Σχήμα 3.13: Τροποποιήσεις στην είσοδο για να ακολουθήσει μετέπειτα το fine-tuning του μοντέλου

ται ένα απλό fine-tuning, ενώ για κάποιες άλλες, για τις οποίες χρειάζεται η προσαρμογή της εισόδου, γίνονται μικρές και πρακτικές τροποποιήσεις, όπως αυτές που παρουσιάζονται στο Σχήμα 3.13. Το μοντέλο εν γένει προ-εκπαιδεύται σε ένα μεγάλο και ετερόκλητο σώμα κειμένων για την κατασκευή embeddings βασισμένων και αυτά στο περιεχόμενο. Αυτή η συνεπακόλουθη αντίληψη της γλώσσας που απορρέει από την παραπάνω εργασία, αξιοποιείται στη συνέχεια σε επιμέρους εφαρμογές όπως η απάντηση ερωτημάτων και η ταξινόμηση προτάσεων μεταξύ άλλων. Αυτό επιτυγχάνεται με την επιβλεπόμενη εκπαίδευση για την επιμέρους εφαρμογή, κατά την διάρκεια της οποίας βελτιώνονται με μικρό βήμα και τα βάρη του προεκπαιδευμένου δικτύου(fine-tuning). Ωστόσο, το μοντέλο αυτό βάσει σχεδίασης είναι αυτο-παλινδρομικό, το οποίο σημαίνει ότι επεξεργάζεται τις γλωσσικές ακολουθίες από τα αριστερά προς τα δεξιά χάνοντας έτσι τα πλεονεκτήματα της αμφιδρομικότητας. Αυτό συμβαίνει γιατί γίνεται χρήση μόνο των αποκωδικοποιητών του Transformer καθώς το μοντέλο σχεδιάστηκε αρχικά με γνώμονα την παραγωγή κειμένου για διάφορες εφαρμογές και αποτελεί ένα σχεδιαστικό trade-off.

## 3.14 Το μοντέλο BERT (Bidirectional Encoder Representations from Transformers)

### 3.14.1 Το μοντέλο και η γενική αρχιτεκτονική του

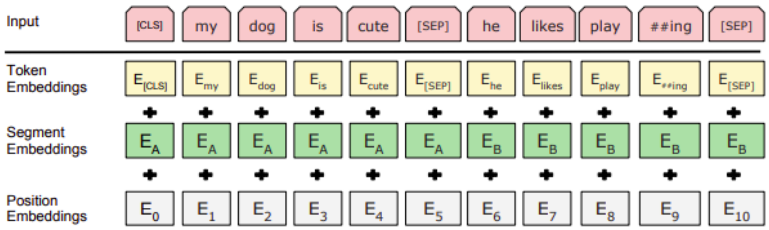
Το μοντέλο BERT παρουσιάστηκε το 2019 από την επιστημονική δημοσίευση "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [15] του εργαστηρίου της Google. Βασίστηκε και αυτό στο μοντέλο Transformer, με τους ερευνητές ωστόσο να εστιάζουν στη δημιουργία ενός μοντέλου "βαθιάς" αμφιδρομικότητας, το οποίο να προ-εκπαιδεύεται σε ένα μεγάλο σώμα κειμένων, να κατασκευάζει embeddings και με αμελητέες αλλαγές στην αρχιτεκτονική του να μπορεί να εκπαιδεύεται επιβλεπόμενα για συγκεκριμένες εφαρμογές σε μικρότερα σετ δεδομένων, εξειδικευμένα για την εκάστοτε εφαρμογή. Κατάφεραν έτσι να συγκεράσουν τα πλεονεκτήματα του ELMo(αμφιδρομικότητα) και του GPT(γενική αρχιτεκτονική που εφαρμόζεται σε κάθε εφαρμογή χωρίς μεγάλες τροποποιήσεις). Συγκεκριμένα το BERT χρησιμοποιεί στρώσεις κωδικοποιητών όπως αυτοί ορίζονται στο πλαίσιο του Transformer. Παράγονται έτσι στο τέλος της προ-εκπαίδευσης οι κατάλληλες ενδιάμεσες αναπαραστάσεις που θα χρησιμοποιηθούν στη συνέχεια σε επιπλέον στρώσεις ανάλογα με την εκάστοτε εφαρμογή. Συνήθως προστίθεται μόνο μια επιπλέον στρώση που εκπαιδεύεται επιβλεπόμενα, με τα βάρη του προ-εκπαιδευμένου δικτύου να αλλάζουν και αυτά, με πολύ μικρό όμως βήμα(fine-tuning). Υπό αυτή τη σκοπιά προσομοιάζει στο GPT. Στην αρχική δημοσίευση παρουσιάστηκαν δύο βαριάντες του μοντέλου. Ένα μοντέλο βάσης με 12 στρώσεις, 12 κεφαλές attention, 768 μέγεθος του κρυφού διανύσματος και 110 εκατομμύρια παραμέτρους συνολικά και ένα μεγαλύτερο μοντέλο με 24 στρώσεις, 16 κεφαλές attention, 1024 μέγεθος του κρυφού διανύσματος και 340 εκατομμύρια παραμέτρους. Το πρώτο μοντέλο σχεδιάστηκε για τη σύγκριση των επιδόσεων σε σχέση με το GPT, ενώ το δεύτερο για να επιφέρει τα SOTA αποτελέσματα για τα οποία έγινε γνωστό το BERT. Μια βασική καινοτομία του BERT είναι και ο τρόπος με τον οποίο γίνεται η προεκπαίδευση εξ αιτίας και του οποίου επιτυγχάνεται και η "βαθιά" αμφιδρομικότητά του. Σε αντίθεση με προηγούμενες προσεγγίσεις όπου τα μοντέλα εκπαιδεύονταν με μοντέλα γλώσσας από αριστερά προς τα δεξιά ή το ανάποδο, ή ακόμα και προς τις δύο διευθύνσεις, παράλληλα και ανεξάρτητα, με τα αποτελέσματα να συγχωνεύονται στο τέλος, το μοντέλο BERT ακολουθεί διαφορετική στρατηγική. Η προεκπαίδευση του πραγματοποιείται με δύο επιμέρους δοκιμασίες.

### 3.14.2 Η είσοδος του μοντέλου

Για να μπορεί το BERT να χρησιμοποιείται σε ένα εύρος εφαρμογών, η αναπαράσταση της εισόδου του έχει σχεδιαστεί ώστε να υποστηρίζει την είσοδο τόσο μιας πρότασης όσο και ενός ζεύγους προτάσεων με ενιαίο τρόπο ως μια ακολουθία εισόδου. Η έναρξη κάθε ακολουθίας γίνεται με το token [CLS], ενώ ο διαχωρισμός των προτάσεων γίνεται αφενός με ένα ενδιάμεσο token [SEQ] και αφετέρου με την εισαγωγή ενός embedding τμήματος που διαφέρει ανάλογα στο ποια πρόταση ανήκει το εκάστοτε token. Για την τελική είσοδο αθροίζονται τρία embeddings: το embedding θέσης, το embedding τμήματος και το embedding του token. Τα παραπάνω φαίνονται και στο Σχήμα 3.14.

### 3.14.3 Προ-εκπαίδευση του BERT: Μοντέλα Γλώσσας με Μάσκα(Masked Language Models)

Για την επίτευξη μιας "βαθιάς" αμφιδρομικότητας το BERT χρησιμοποιεί ένα μοντέλο γλώσσας με μάσκα. Συγκεκριμένα, ένα τυχαίο 15 % δείγμα των tokens αποκρύπτεται, αντικαθίσταται δηλαδή με ένα token-μάσκα. Σκοπός της εκπαίδευσης γίνεται η πρόβλεψη αυτών




Σχήμα 3.14: Αναπαράσταση μια ακολουθίας εισόδου του BERT

των tokens. Σε περίπτωση που γινόταν απλή πρόβλεψη υπό συνθήκη λαμβάνοντας υπόψιν το περιεχόμενο και προς τις δύο κατευθύνσεις, το μοντέλο θα μπορούσε να δει κάθε λέξη έμμεσα και επομένως δε θα μάθαινε κάτι καινούριο. Αυτό γίνεται αντιληπτό και διαισθητικά αφού η ακολουθία εισόδου δίνεται όλη μαζί σε αντίθεση με παλαιότερες προσεγγίσεις όπου εισαγόταν ακολουθιακά ένα token σε κάθε βήμα. Η παραπάνω λύση αποτελεί μια καινοτομία και οδηγεί σε παραγωγή αμφίδρομων αναπαραστάσεων. Ωστόσο, δημιουργεί μια αναντιστοιχία ανάμεσα στο στάδιο της προ-εκπαίδευσης και του fine-tuning καθώς οι μάσκες δεν εμφανίζονται στο δεύτερο. Για να αντιμετωπιστεί αυτό το πρόβλημα, από το 15 % των επιλεχθέντων tokens αποκρύπτονται με μάσκα το 80 % , ενώ ένα 10 % αντικαθίσταται με μια τυχαία λέξη και το τελευταίο 10 % αντικαθίσταται με την ίδια τη λέξη. Είναι σημαντικό στο σημείο αυτό να τονιστεί ότι η εκπαίδευση σε αυτό το στάδιο γίνεται μη επιβλεπόμενα, κάτι που την καθιστά αποτελεσματική στην πράξη για πολύ μεγάλα σετ δεδομένων.

### 3.14.4 Προ-εκπαίδευση του BERT: Πρόβλεψη επόμενης πρότασης(Next Sentence Prediction)

Η δοκιμασία της πρόβλεψης επόμενης πρότασης κατά την εκπαίδευση γίνεται ώστε το BERT να μπορεί να υποστηρίξει αποτελεσματικά μεγαλύτερο εύρος εφαρμογών. Συγκεκριμένα κάποιες εφαρμογές όπως η Απάντηση Ερωτημάτων(Question Answering) απαιτούν γνώση ανάμεσα στη λογική σχέση μεταξύ δύο προτάσεων. Για την εξαγωγή αυτής της πληροφορίας δίνονται κάθε φορά δύο προτάσεις ως είσοδος. Τις μισές φορές η δεύτερη πρόταση ακολουθεί λογικά την πρώτη, ενώ τις υπόλοιπες οι δύο προτάσεις επιλέγονται τυχαία και άρα δεν σχετίζονται μεταξύ τους. Σκοπός της εκπαίδευσης είναι να προβλέψει το μοντέλο κάθε φορά εάν η δεύτερη πρόταση είναι συνέχεια της πρώτης, ανάγεται δηλαδή έτσι η διαδικασία σε ένα πρόβλημα δυαδικής ταξινόμησης. Παρά την απλότητα της η παραπάνω δοκιμασία οδηγεί σε σημαντική βελτίωση των επιδόσεων του μοντέλου σε ορισμένες εργασίες, όπως προκύπτει πειραματικά.

# 4

# Οι σημαντικότερες εξελίξεις στην Επεξεργασία Φυσικής Γλώσσας από το BERT και έπειτα

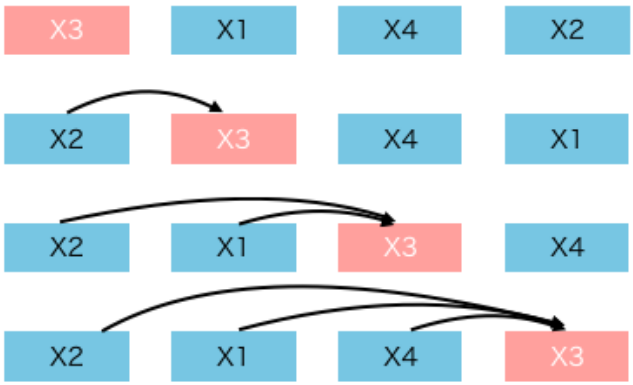
Το μοντέλο BERT με τις επιδόσεις του και με τις καινοτομίες που εισήγαγε θεωρήθηκε από πολλούς ότι οδήγησε το πεδίο της Επεξεργασίας Φυσικής Γλώσσας σε μια νέα εποχή, πυροδοτώντας παράλληλα περαιτέρω έρευνα προς αυτήν την κατεύθυνση την τελευταία διετία. Πολλές νέες δημοσιεύσεις επικεντρώθηκαν στη βελτιστοποίηση του ίδιου του μοντέλου, ενώ άλλες εφάρμοσαν κάποιες εναλλακτικές προσεγγίσεις, πολλές φορές για να αντιμετωπίσουν προβλήματα που ανέκυπταν στο BERT μετά από όψιμη και ενδελεχή μελέτη και χρήση του. Στη συνέχεια αυτού του κεφαλαίου, παρουσιάζονται λοιπόν οι πιο εμβληματικές εξελίξεις στο χώρο, από το BERT και μέχρι τη συγγραφή αυτής της πτυχιακής εργασίας.

## 4.1 Το μοντέλο XLNet

Εισήχθη στη βιβλιογραφία με τη δημοσίευση "XLNet: Generalized Autoregressive Pre-training for Language Understanding" το 2019 [16], σε σύμπραξη του εργαστηρίου Google AI Brain και του Αμερικανικού Πανεπιστημίου Carnegie Mellon. Ακολουθείται γενικά η στρατηγική του BERT, υπό την έννοια ότι αποτελεί μια μέθοδο κατασκευής αμφίδρομων Embeddings βασιζόμενα σε ένα στάδιο προεκπαίδευσης. Παρόλα αυτά οι ερευνητές προσπάθησαν να αντιμετωπίσουν τα εξής δύο ζητήματα που είχαν ανακύψει με το BERT:

- Το BERT κατατάσσεται εν γένει στα μοντέλα αυτοκωδικοποιητή. Στα μοντέλο αυτά ανακατασκευάζεται η είσοδος τους μετά από τυχαία παραλλαγή της, ως τμήμα της διαδικασίας εκπαίδευσης. Συγκεκριμένα στο BERT, όπως παρουσιάστηκε και παραπάνω, ένα τυχαίο ποσοστό της εισόδου καλύπτεται από [MASK] tokens. Αυτό οδηγεί στην κατασκευή embeddings λαμβάνοντας υπόψιν το αμφίδρομο περιεχόμενο της εκάστοτε ακολουθίας. Ωστόσο, αυτές οι μάσκες εμφανίζονται μονάχα κατά το στάδιο της προεκπαίδευσης δημιουργώντας μια αναντιστοιχία ανάμεσα στην προ-εκπαίδευση και στο μετέπειτα fine-tuning.

- Επιπρόσθετα, μια παραδοχή του BERT είναι ότι οι μάσκες, οι κρυφές δηλαδή λέξεις, είναι ανεξάρτητες η μια από την άλλη και εξαρτώνται αποκλειστικά από τις υπόλοιπες φανερές λέξεις. Κάτι τέτοιο, όπως γίνεται εύκολα αντιληπτό διαισθητικά, τις περισσότερες φορές δεν ανταποκρίνεται στην πραγματικότητα.



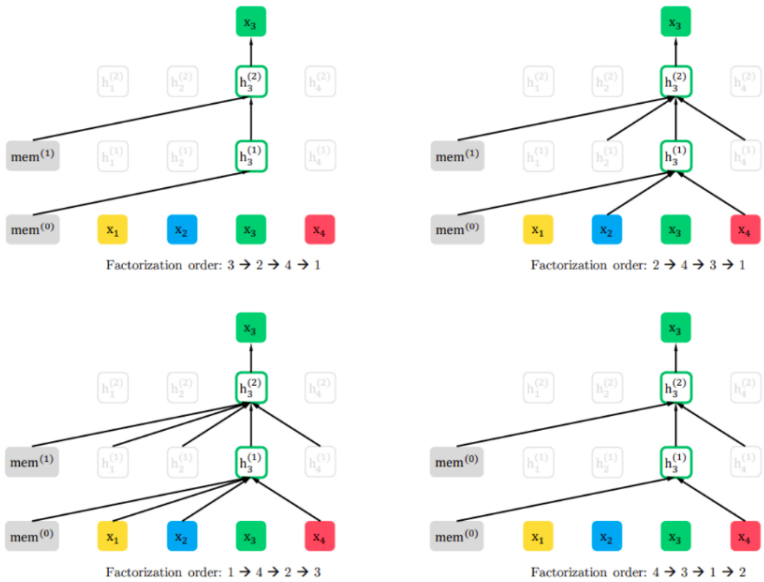

Σχήμα 4.1: Η μετάθεση στο XLNet

Στη βάση αυτή οι ερευνητές ακολούθησαν μια διαφορετική προσέγγιση. Χρησιμοποίησαν συγκεκριμένα ένα αυτο-παλινδρομικό μοντέλο. Είναι βέβαια γνωστό ότι αυτά τα μοντέλα προβλέπουν λέξεις μόνο προς τη μια κατεύθυνση και άρα δεν δύνανται να παρέχουν πληροφορίες αμφιδρομικότητας. Για να συνδυάσουν επομένως τα καλά και των δύο κόσμων, οι ερευνητές χρησιμοποίησαν ένα γλωσσικό μοντέλο μετάθεσης(Permutation Language Model). Η ουσία αυτής της προσέγγισης εντοπίζεται στη λέξη μετάθεση. Έστω για παράδειγμα ότι έχουμε μια ακολουθία τεσσάρων tokens και θέλουμε να προβλέψουμε το τρίτο. Με μετάθεση του token αυτού σε κάθε επιμέρους θέση(Σχήμα 4.1), μπορούν να προκύψουν τέσσερα δυνατά σχήματα όπως φαίνεται και στο Σχήμα 4.2. Έτσι με τη χρήση της αυτο-παλινδρόμισης προβλέπεται σε κάθε θέση το token και συνδυάζονται τα αποτελέσματα. Με τον τρόπο αυτό λαμβάνονται πληροφορίες για όλες τις περιβάλλουσες λέξεις εκατέρωθεν της λέξης στόχου και άρα η κατασκευή του αντίστοιχου embedding γίνεται κατ'ουσίαν με "βαθιά" αμφιδρομική επεξεργασία της ακολουθίας. Τα αποτελέσματα αυτής της προσέγγισης υπήρξαν θεαματικά. Το μοντέλο κατάφερε να ξεπεράσει το BERT σε 20 διαφορετικές εφαρμογές, συχνά μάλιστα κατά πολύ. Ξεπέρασε γενικά και τα υπόλοιπα SOTA μοντέλα σε 18 εφαρμογές, ενώ έκτοτε έχουν γίνει απόπειρες να εφαρμοστεί η προσέγγιση του και σε άλλα πεδία της Μηχανικής Μάθησης.

## 4.2 Το μοντέλο RoBERTa

Δημοσιεύθηκε το 2019 με τίτλο "RoBERTa: A Robustly Optimized BERT Pretraining Approach" [17] από το ερευνητικό εργαστήριο Facebook AI σε σύμπραξη με ερευνητές από το πανεπιστήμιο της Washington. Κατόπιν ενδελεχούς ανάλυσης της διαδικασίας εκπαίδευσης του BERT οι ερευνητές κατέληξαν ότι είναι υποεκπαιδευμένο. Έτσι, μετά από δοκιμές πρότειναν μια σειρά εναλλακτικών τροποποιήσεων που θα μπορούσαν να ενισχύσουν την απόδοση του BERT χωρίς να αλλάξουν τα βασικά δομικά στοιχεία του μοντέλου. Έτσι προέκυψε το μοντέλο RoBERTa. Συγκεκριμένα:

- Αυξήθηκε το σετ δεδομένων εκπαίδευσης σε 160 GB από τα 16 GB που ήταν αρχικά στο BERT
- Αυξήθηκε το μέγεθος των batches σε 8K από 256 που ήταν στην αρχική δημοσίευση
- Αυξήθηκε ο αριθμός των iterations σε 300K από 100K




Σχήμα 4.2: Παράδειγμα της μεταθετικής γλωσσικής μοντελοποίησης

- Αφαιρέθηκε εξ ολοκλήρου η δοκιμασία προ-εκπαίδευσης της πρόβλεψης επόμενης πρότασης
- Τα [MASK] tokens ανατίθενται πλέον με δυναμικό τρόπο για κάθε ακολουθία εισόδου, σε αντίθεση με την αρχική δημοσίευση που η ανάθεση γινόταν στατικά
- Κωδικοποίηση μεγάλου λεξιλογίου με 50K μονάδες υπολέξεων σε αντίθεση με την αρχική δημοσίευση όπου η κωδικοποίηση γινόταν σε επίπεδο χαρακτήρων, με το μέγεθος του λεξιλογίου να είναι 30K

Με αυτές τις τροποποιήσεις το μοντέλο κατάφερε να ξεπεράσει το BERT σε όλες τις δοκιμασίες του benchmark GLUE, ενώ παράλληλα κατάφερε να επιτύχει αντίστοιχα υψηλές επιδόσεις σε σχέση με την πρόσφατη βιβλιογραφία, καταδεικνύοντας ότι το μοντέλο BERT χωρίς σημαντικές αλλαγές μπορεί να συνεχίσει να είναι ανταγωνιστικό σε σχέση με τα νεότερα SOTA.

## 4.3 Το μοντέλο ALBERT

Δημοσιεύθηκε και αυτό το 2019 με τίτλο "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations" [18] από το ερευνητικό εργαστήριο της Google σε συνεργασία με το Τεχνολογικό Ινστιτούτο της Toyota στο Chicago. Οι ερευνητές διαπίστωσαν ότι η αύξηση του μεγέθους των προ-εκπαιδευμένων μοντέλων πέρα από ένα σημείο οδηγεί σε προβλήματα όπως υπερβολικά αυξημένο χρόνο εκπαίδευσης λόγω και του επιπλέον χρόνου που απαιτείται για την επικοινωνία με αυξημένο αριθμό παραμέτρων, περιορισμούς λόγω πεπερασμένης υπολογιστικής μνήμης και γενικότερα μια τάση για μείωση των επιδόσεων παρότι θεωρητικά θα έπρεπε να επιτυγχάνει το ανάποδο. Εν συνεχεία πρότειναν μια "ελαφρύτερη" εκδοχή του BERT χρησιμοποιώντας τις ακόλουθες δύο τεχνικές μείωσης παραμέτρων:

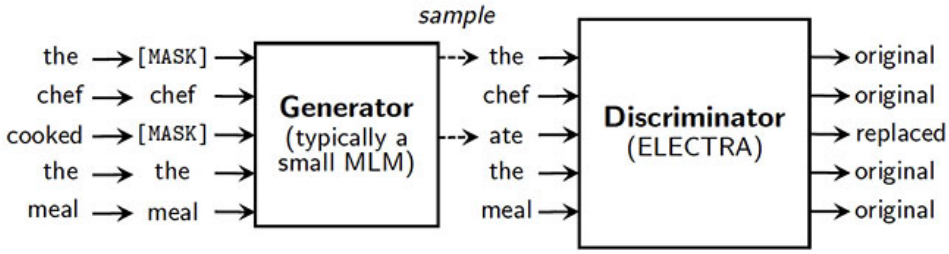
- Παραμετροποίηση παραγοντοποιημένων embeddings (Factorized embedding parameterization), όπου το μέγεθος των κρυφών στρώσεων αποσυνδέεται από το μέγεθος των embeddings λεξιλογίου με την αποσύνθεση του πίνακα embeddings-λεξιλογίου σε δύο μικρότερους.

- Διαμοιρασμός όλων των παραμέτρων-attention και πρόσθιας τροφοδότησης- ανάμεσα στις στρώσεις, για την αποτροπή της υπερβολικής αύξησης του αριθμού των παραμέτρων με την αύξηση του βάθους του δικτύου

Επιπρόσθετα μια ακόμη τροποποίηση εφαρμόσθηκε στο κομμάτι της πρόβλεψης επόμενης πρότασης του BERT. Συγκεκριμένα- και σε συνέχεια κατοπινών ερευνών που αφαίρεσαν αυτό το τμήμα της εκπαίδευσης γιατί θεώρησαν ότι ίσως οδηγεί σε πτώση των συνολικών επιδόσεων του μοντέλου- οι ερευνητές της παρούσας δημοσίευσης απέδωσαν αυτή τη χαμηλή απόδοση στη γενικότερη ευκολία περαίωσης αυτής της εργασίας σε σχέση με τα μοντέλα γλώσσας με μάσκα. Πιο ειδικά, η πρόβλεψη επόμενης πρότασης συγχωνεύει την πρόβλεψη θέματος με την πρόβλεψη συνοχής. Όμως η εκπαίδευση στην πρόβλεψη θέματος είναι αφενός ευκολότερη και αφετέρου επικαλύπτεται με την εκπαίδευση της εργασίας του μοντέλου γλώσσας με μάσκα. Προτείνεται συνεπώς η επικέντρωση στην πρόβλεψη συνοχής ανάμεσα στις προτάσεις με χρήση μιας μετρικής σφάλματος που βασίζεται στην ακριβή σειρά των προτάσεων κάτι που οδηγεί στη λεπτομερέστερη εκμάθηση της. Συγκεκριμένα το μοντέλο καλείται να προβλέψει εάν οι προτάσεις είναι δοσμένες με τη σωστή ή την αντίθετη σειρά. Με τα παραπάνω το μοντέλο ALBERT με σημαντικά λιγότερες παραμέτρους και ταχύτερη διαδικασία εκπαίδευσης επιτυγχάνει συγκρίσιμα αποτελέσματα με το αρχικό BERT, ενώ το μοντέλο με παρεμφερή αριθμό παραμέτρων βελτιώνει το SOTA σε όλες τις βασικές εφαρμογές.

## 4.4 Το μοντέλο ELECTRA

Δημοσιεύθηκε το 2020 με τίτλο "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators" [19] με τη σύμπραξη του πανεπιστημίου του Stanford με το ερευνητικό εργαστήριο Google Brain. Αφορμή αποτέλεσε η διαπίστωση ότι το BERT και όσα άλλα μοντέλα χρησιμοποιούν το γλωσσικό μοντέλο με μάσκα για την προ-εκπαίδευση τους, παρότι πετυχαίνουν αξιοσημείωτες επιδόσεις, απαιτούν πολύ μεγάλο αριθμό υπολογισμών. Συγκεκριμένα, το BERT αποκρύπτει με μάσκα ένα μόνο 15% των δεδομένων και εκπαιδεύεται στην πρόβλεψή τους. Κάτι τέτοιο δεν είναι ιδιαίτερα αποδοτικό από άποψη χρήσης των συνολικών δεδομένων όπως γίνεται φανερό, αφού το μοντέλο μαθαίνει από ένα πολύ μικρό μόνο τμήμα των δεδομένων. Σε αυτό το πλαίσιο οι ερευνητές προτείνουν έναν εναλλακτικό τρόπο εκπαίδευσης που ονόμασαν "replaced token detection"(αναγνώρισης αντικατασταθέντος token). Συγκεκριμένα, αντί για χρήση μάσκας, κάποια tokens αντικαθίστανται με εναλλακτικά tokens που παράγονται από ένα μικρό γλωσσικό μοντέλο-γεννήτρια. Εν συνεχεία το ELECTRA προσπαθεί για κάθε token να ανιχνεύσει εάν είναι το αυθεντικό ή εάν έχει παραχθεί από τη γεννήτρια. Έτσι μαθαίνει από όλη την είσοδο αντί μόνο από ένα μικρό τμήμα της και συνεπώς καθίσταται υπολογιστικά πιο αποδοτικό. Τα παραπάνω επιβεβαιώνει και ο ταχύτερος χρόνος εκπαίδευσής του. Η διαδικασία εκπαίδευσης φαίνεται και στο Σχήμα 4.3. Εν γένει, το μοντέλο επιτυγχάνει SOTA αποτελέσματα με χρήση μέχρι και του 25% των δεδομένων που απαιτούντο με τις προηγούμενες προσεγγίσεις και με μικρότερο χρόνο εκπαίδευσης. Μελλοντικά αυτό το μοντέλο λόγω της αποδοτικότητας του θα μπορούσε να κάνει τα προεκπαιδευμένα μοντέλα κωδικοποιητών πιο προσιτά σε εταιρίες και επιχειρήσεις.




Σχήμα 4.3: Η διαδικασία εκπαίδευσης του ELECTRA

## 4.5 Το μοντέλο DeBERTa

Το μοντέλο DeBERTa παρουσιάζεται το 2021 στο πλαίσιο της ερευνητικής δημοσίευσης "DeBERTa: Decoding-enhanced BERT with Disentangled Attention" [20] από το ερευνητικό εργαστήριο της Microsoft. Οι ερευνητές εισάγουν δύο νέες καινοτομίες-τροποποιήσεις σε σχέση με την αρχική δημοσίευση του BERT με απώτερο στόχο την περαιτέρω βελτίωση της απόδοσής του. Συγκεκριμένα προτείνεται:

- Ένας μηχανισμός disentangled attention. Στην αρχική δημοσίευση του BERT κάθε λέξη τη εισόδου αναπαρίσταται με τη χρήση ενός διανύσματος που είναι το άθροισμα του embedding θέσης και του embedding περιεχομένου. Αντιθέτως, στην παρούσα δημοσίευση κάθε λέξη αναπαρίσταται με χρήση δύο διανυσμάτων που κωδικοποιούν τη θέση και το περιεχόμενο της αντίστοιχα. Έτσι τα αντίστοιχα βάρη υπολογίζονται από τους αντίστοιχα απεμπλεγμένους πίνακες. Η παραπάνω προσέγγιση ακολουθείται κατόπιν της παρατήρησης ότι το βάρος του attention ανάμεσα σε ένα ζευγάρι λέξεων εξαρτάται όχι μόνο από το περιεχόμενό τους αλλά και από τη σχετική τους θέση.

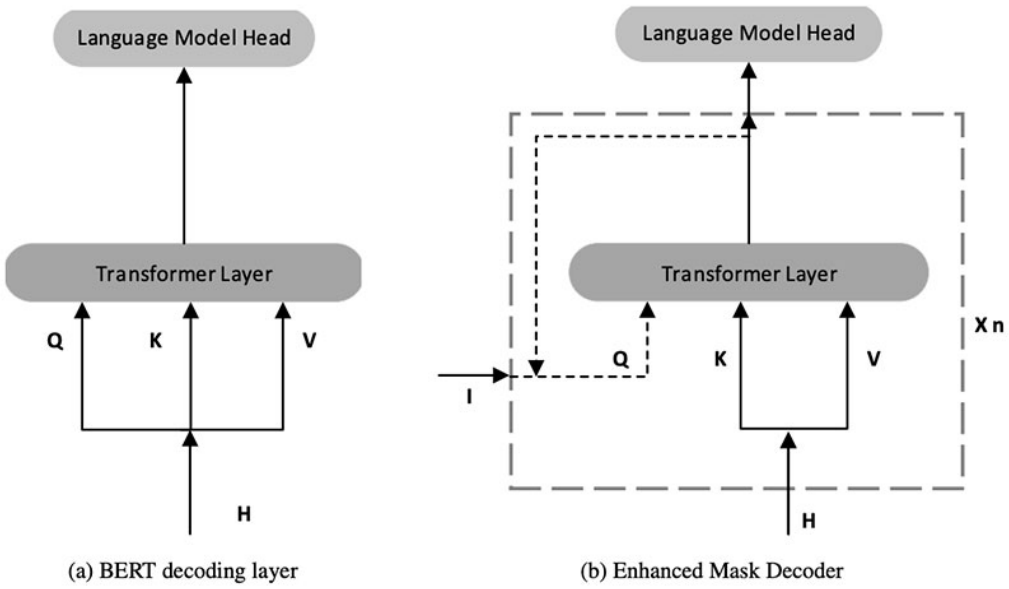
- Χρήση ενισχυμένου αποκωδικοποιητή μάσκας(Enhanced Mask Decoder). Το μοντέλο DeBERTa χρησιμοποιεί και αυτό ένα γλωσσικό μοντέλο με μάσκα για την εκπαίδευση του μοντέλου. Ενισχύει περαιτέρω το μοντέλο του BERT λαμβάνοντας υπόψιν και την απόλυτη θέση της κάθε λέξης στην εκάστοτε πρόταση. Για να επιτευχθεί αυτό χρησιμοποιείται ένα embedding με πληροφορίες για την απόλυτη θέση της λέξης ακριβώς πριν τη στρώση του softmax. Η διαφορά των δύο προσεγγίσεων παρουσιάζεται στο Σχήμα 4.4.

Τέλος το μοντέλο χρησιμοποιεί μια νέα τεχνική Regularization για την αύξηση της ικανότητας γενίκευσης του μοντέλου. Ονομάζεται Scale-invariant fine-tuning και κάνει χρήση ενός αντιπαραθετικού αλγορίθμου εκμάθησης.

Με την εφαρμογή των παραπάνω καινοτομιών το δίκτυο καταφέρνει να επιτύχει καλύτερες επιδόσεις από το SOTA με χρήση των μισών δεδομένων. Επιπρόσθετα ένα ελαφρώς τροποποιημένο δίκτυο πέτυχε καλύτερες επιδόσεις από τις αναμενόμενες ανθρώπινες στο SuperGLUE benchmark.

## 4.6 Το μοντέλο StructBERT

Η δημοσίευση έγινε το 2019 με τίτλο "StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding" [21] από την ερευνητική ομάδα της εταιρίας Alibaba. Το μοντέλο δομείται επάνω στο BERT, αποτελεί δηλαδή ένα προ-εκπαιδευμένο



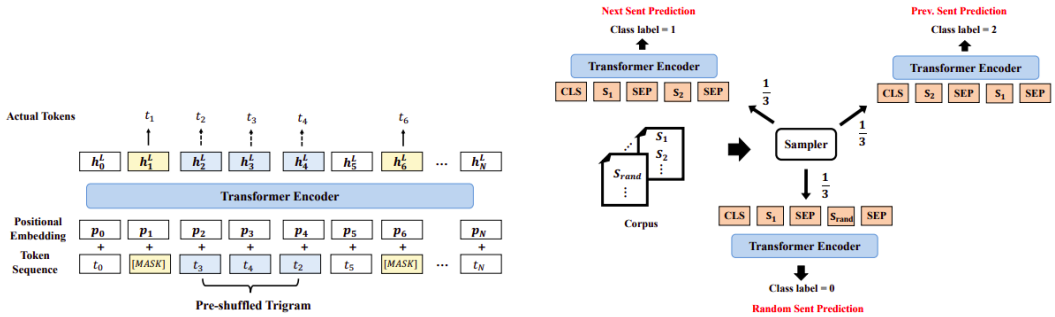

Σχήμα 4.4: Οι δύο διαφορετικές προσεγγίσεις σε αντιπαραβολή

γλωσσικό μοντέλο που παράγει διανυσματικές αναπαραστάσεις λέξεων. Οι ερευνητές σε αυτή την περίπτωση προσπάθησαν να ενσωματώσουν και τις ενυπάρχουσες γλωσσικές δομές κατά τη διαδικασία κατασκευής των διανυσμάτων, υπό την εικασία ότι κάτι τέτοιο θα οδηγήσει σε μεγαλύτερη ικανότητα γενίκευσης και προσαρμοστικότητα του μοντέλου. Αυτό επιτυγχάνεται με την εισαγωγή δύο νέων εργασιών προεκπαίδευσης. Συγκεκριμένα εισάγονται:

- Η εύρεση της σωστής ακολουθίας των λέξεων σε μια πρόταση. Οι λέξεις της εισόδου αναμειγνύονται με ένα τυχαίο τρόπο και το γλωσσικό μοντέλο καλείται να τις επαναδιατάξει στην αρχική τους σωστή σειρά. Αυτή η εργασία πραγματοποιείται από κοινού με την αρχική εργασία της εύρεσης των κρυμμένων με μάσκα λέξεων που χρησιμοποιείται και στο BERT.

- Η εύρεση της σωστής ακολουθίας προτάσεων. Σε αντιδιαστολή με το BERT όπου χρησιμοποιείται για την προεκπαίδευση η Πρόβλεψη Επόμενης Πρότασης, το StructBERT επεκτείνει την εργασία αυτή σε πρόβλεψη τόσο της επόμενης όσο και της προηγούμενης πρότασης. Έτσι το γλωσσικό μοντέλο μαθαίνει την ακολουθιακή σειρά των προτάσεων με έναν αμφίδρομο τρόπο και άρα αποκτά καλύτερη εποπτεία της δομής της γλώσσας. Ειδικότερα δίνονται ως είσοδος δύο προτάσεις. Για την πρόταση στόχο δίνεται το 1/3 των φορών η επόμενη πρόταση που την ακολουθεί, το άλλο 1/3 η πρόταση που προηγείται και το άλλο 1/3 μια τυχαία πρόταση από ένα άλλο έγγραφο. Το μοντέλο καλείται να αποφανθεί ποια από τις τρεις περιπτώσεις ισχύει σε κάθε περίπτωση και έτσι μαθαίνει.

Οι δύο νέες εργασίες προεκπαίδευσης φαίνονται και στο Σχήμα 4.5 και μαζί με την εργασία του γλωσσικού μοντέλου με μάσκα, εκτελούνται από κοινού, με ενοποιημένο τρόπο. Το μοντέλο επιτυγχάνει SOTA αποτελέσματα σε πολλές εργασίες των σημαντικότερων benchmarks και ανοίγει το δρόμο για περαιτέρω αξιοποίηση της γλωσσικής δομής στην κατασκευή διανυσματικών αναπαραστάσεων.



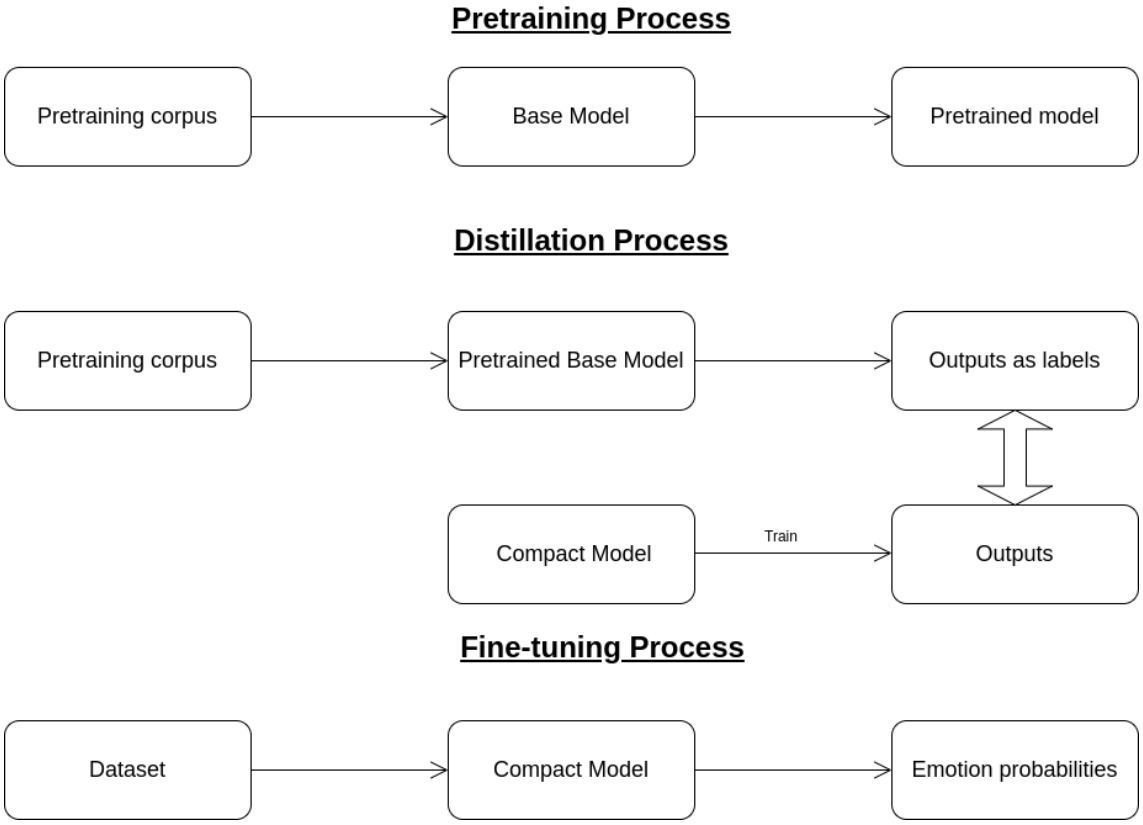

(a) Word Structural Objective (b) Sentence Structural Objective

Σχήμα 4.5: Οι δύο νέες εργασίες για την προ-εκπαίδευση

## 4.7 Συμπίεση μοντέλων με Απόσταξη Γνώσης

Από την εισαγωγή των Transformers και έπειτα κατέστη σαφές στην επιστημονική κοινότητα, ότι η εκπαίδευση δύο σταδίων, η προ-εκπαίδευση δηλαδή των μοντέλων σε ένα μεγάλο σώμα δεδομένων για την κατασκευή embeddings και η εφαρμογή εν συνεχεία επιβλεπόμενης μάθησης σε μικρά σύνολα δεδομένων για εξειδικευμένες εργασίες, επιφέρει θεαματικά αποτελέσματα στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας. Οι ερευνητές προσανατολίστηκαν λοιπόν προς αυτή την κατεύθυνση τα επόμενα χρόνια, βελτιώνοντας κάποια στάδια, πάνω όμως σε αυτό το βασικό μοτίβο. Ωστόσο, τα SOTA μοντέλα, χρόνο με το χρόνο, μαζί με τις βελτιωμένες επιδόσεις τους γίνονται όλο και πιο πολύπλοκα και με μεγαλύτερο βάθος, με αποτέλεσμα να έχουν συνήθως εκατομμύρια παραμέτρους και η προ-εκπαίδευση τους να απαιτεί πολύ μεγάλο υπολογιστικό κόστος. Η χρήση αυτών των μοντέλων μπορεί να μην αποτελεί πρόβλημα σε ερευνητικά περιβάλλοντα με την συνεπακόλουθη αύξηση και της υπολογιστής ισχύος, ωστόσο μπορεί να καθίσταται απαγορευτική σε περιβάλλοντα με περιορισμένη μνήμη ή με απαιτήσεις περιορισμένης καθυστέρησης εξυπηρέτησης, παραδείγματος χάριν σε φορητές, κινητές συσκευές. Για να αντιμετωπιστεί αυτό το πρόβλημα έχουν ακολουθηθεί διάφορες προσεγγίσεις κυρίως προς την κατεύθυνση της συμπίεσης των μεγάλων μοντέλων. Μεταξύ άλλων έχει προταθεί μια τοπική μέθοδος για κλάδεμα βαρών βασισμένη στο σφάλμα του μοντέλου[22], κλάδεμα ολόκληρων φίλτρων[23][24], κατασκευή pipeline συμπίεσης[25] και επικέντρωση σε μετρικές που έχουν να κάνουν με το υλικό[26] όπως οι πράξεις κινητής υποδιαστολής. Μια από τις κεφαλαιώδεις προσεγγίσεις για την συμπίεση των μοντέλων είναι η Απόσταξη Γνώσης(Knowledge Distillation) που πρωτοεισήχθη ως ιδέα από τον Geoffrey Hinton το 2015 με τη δημοσίευση "Distilling the Knowledge in a Neural Network" [27] για Συνελικτικά κυρίως δίκτυα. Η προσέγγιση αυτή προσαρμόστηκε στο μοντέλο BERT και στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας εν γένει, με τη δημοσίευση του 2019 "Distilling Task-Specific Knowledge from BERT into Simple Neural Networks" [28]. Ουσιαστικά γίνεται χρήση ενός μεγάλου μοντέλου-δασκάλου, που αποτελεί συνήθως το SOTA μοντέλο του οποίου τις επιδόσεις θέλουμε να διατηρήσουμε, και ενός μικρότερου συμπιεσμένου μοντέλου-μαθητή, στο οποίο προσδοκάται η μεταφορά γνώσης από το μεγάλο μοντέλο.

Το μεγάλο μοντέλο προεκπαιδεύεται κανονικά σε περιβάλλον που δεν υπόκειται σε περιορισμούς υπολογιστικών πόρων και χρόνου, δηλαδή σε ισχυρές κάρτες γραφικών και offline. Εν συνεχεία τα αποτελέσματα, δηλαδή οι συναγωγές που προβλέπει, χρησιμοποιούνται ως labels για την εκπαίδευση του μοντέλου-μαθητή. Με τον τρόπο αυτό το μοντέλο-μαθητής επωφελείται από τη γνώση που έχει αποκομιστεί από την εκπαίδευση του μεγάλου δικτύου και στη θεωρία παρουσιάζει βελτιωμένες επιδόσεις, ανακόλουθες με τη χωρητικότητα που




Σχήμα 4.6: Η διαδικασία εκπαίδευσης με την τεχνική της Απόσταξης Γνώσης

υποδηλωνόταν με το μέγεθος και την πολυπλοκότητά του. Στη συνέχεια χρησιμοποιείται ένα συγκεκριμένο σετ δεδομένων για το fine-tuning του μοντέλου μαθητή, σχετιζόμενο με την επιμέρους εργασία στην οποία θα γίνει η χρήση του μοντέλου. Τέλος γίνεται η πραγματοποίηση συναγωγών και αξιολογούνται οι επιδόσεις του. Τα παραπάνω παρουσιάζονται στο Σχήμα 4.6.

# 5

# Η Ανάλυση Συναισθήματος

Ως Ανάλυση Συναισθήματος ορίζεται το πεδίο εκείνο του οποίου σκοπός είναι η εξαγωγή συμπερασμάτων, η αναγνώριση, η συστηματική μελέτη και η ποσοτικοποίηση της συναισθηματικής κατάστασης και της υποκειμενικής αντιμετώπισης που απορρέει από ένα κείμενο ή από τον προφορικό λόγο. Αποτελεί ένα διεπιστημονικό πεδίο που αντλεί πόρους από την επιστήμη υπολογιστών, την επεξεργασία φυσικής γλώσσας, την ψυχολογία και την υπολογιστική γλωσσολογία μεταξύ άλλων. Σε αδρές γραμμές η Ανάλυση Συναισθήματος βοηθάει στο να αποφανθούμε για το κατά πόσο ένα κείμενο εκφράζει συναισθήματα που είναι θετικά, αρνητικά ή ουδέτερα. Βρίσκει εφαρμογές σε σημαντικό βαθμό στον εμπορικό κόσμο, όπου και χρησιμοποιείται για να διαπιστωθεί πώς αισθάνονται οι πελάτες για συγκεκριμένες υπηρεσίες ή προϊόντα. Οι απαρχές του πεδίου εντοπίζονται ήδη στη δεκαετία του 1950 όπου επιχειρήθηκε η άντληση συμπερασμάτων από γραπτά έγγραφα σε χαρτί. Στη σύγχρονη εποχή, με την έλευση του Διαδικτύου, η Ανάλυση Συναισθήματος εφαρμόζεται σε σημαντικό βαθμό σε περιεχόμενό του, όπως κοινωνικά δίκτυα, tweets, blogs και σχόλια μεταξύ άλλων. Οι διάφοροι Οργανισμοί χρησιμοποιούν τα συμπεράσματα που εξάγονται για να εντοπίσουν νέες ευκαιρίες και να επικεντρωθούν αποδοτικότερα προς τις δημογραφικές ομάδες που στοχεύουν.

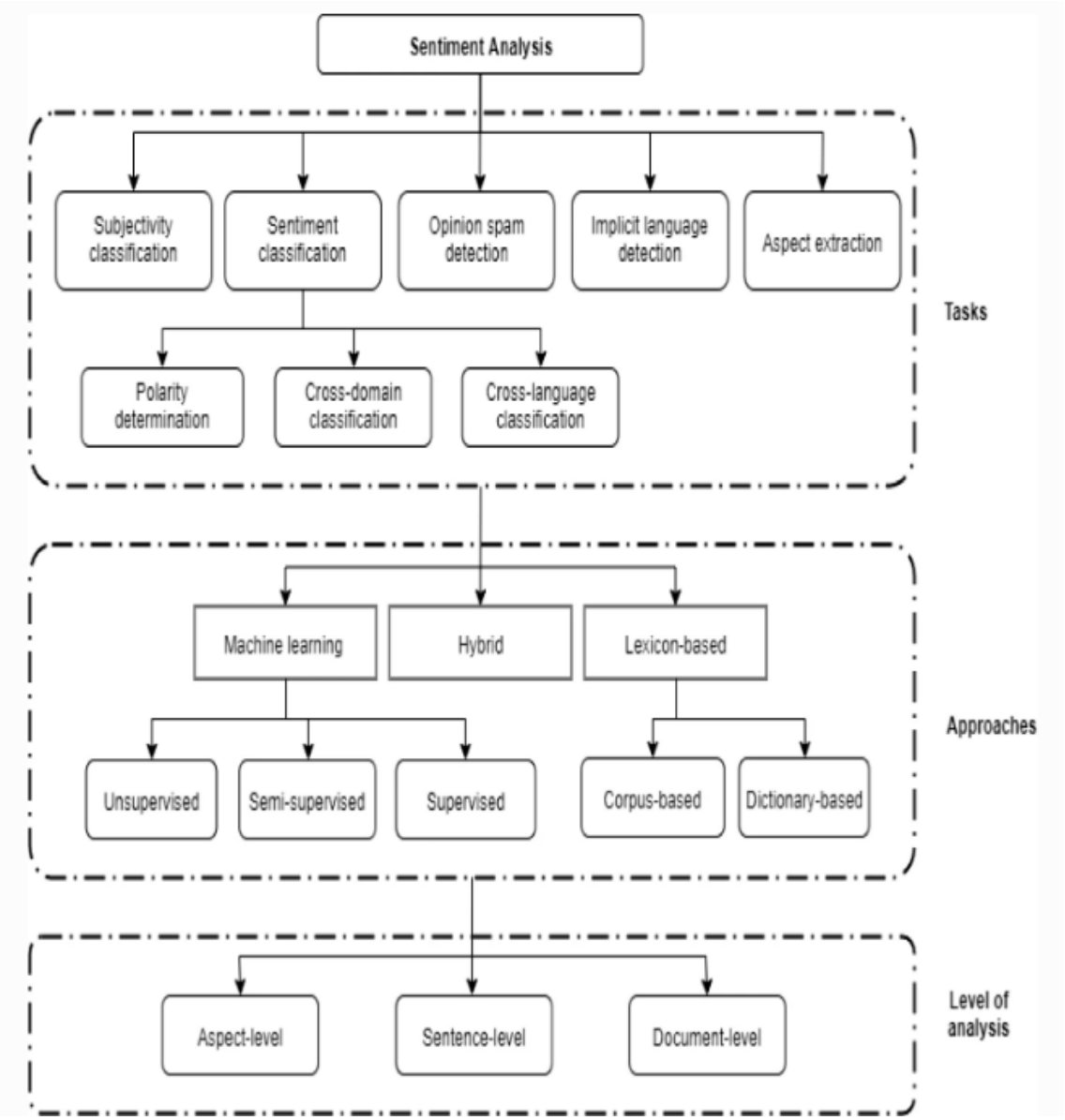
Οι δύο κύριοι τύποι Ανάλυσης Συναισθήματος είναι:

- **Αναγνώριση Υποκειμενικότητας/Αντικειμενικότητας**. Επιχειρείται η ταξινόμηση μιας πρότασης ή ενός τμήματος ενός κειμένου σε δύο κατηγορίες: υποκειμενικό και αντικειμενικό. Αυτού του είδους η ανάλυση παρουσιάζει σημαντικές προκλήσεις καθώς πολλές φορές η σημασία μιας λέξης ή και μιας ολόκληρης φράσης εξαρτάται από το περιεχόμενό της.

- **Αναγνώριση Χαρακτηριστικών/Πτυχών**. Αναφέρεται στην απόφανση για τις γνώμες και τα συναισθήματα που εκφράζονται αναφορικά με διαφορετικά χαρακτηριστικά ή πτυχές μιας οντότητας, για παράδειγμα μιας ηλεκτρονικής συσκευής ή μιας τράπεζας. Αυτού του τύπου η ανάλυση καθιστά δυνατή μια αρκετά πιο λεπτομερή και αναλυτική επισκόπηση των γνωμών και των υποκείμενων συναισθημάτων.

Οι υπάρχουσες προσεγγίσεις στην Ανάλυση Συναισθήματος μπορούν να ομαδοποιηθούν στις ακόλουθες επιμέρους κατηγορίες, σύμφωνα με το [29]:

- **Τεχνικές βασισμένες σε Συστήματα Γνώσης**. Με βάση τις τεχνικές αυτές το κείμενο κατηγοριοποιείται σε κατηγορίες Συναισθημάτων ανάλογα με την ύπαρξη λέξεων ξεκάθαρου συναισθηματικού προσήμου, όπως χαρούμενος, λυπημένος, θυμωμένος, φοβισμένος κ.α. Σε κάποια μάλιστα πιο προηγμένα Συστήματα Γνώσης δίνονται και πιθανότητες για κάθε συναίσθημα για τις εκάστοτε λέξεις.




Σχήμα 5.1: Το πεδίο της Ανάλυσης Συναισθήματος

- **Προσεγγίσεις που βασίζονται στη Μηχανική Μάθηση**. Αυτές χωρίζονται σε τρεις επιμέρους υποκατηγορίες: τη μη επιβλεπόμενη μάθηση, την επιβλεπόμενη μάθηση και την ημι-επιβλεπόμενη μάθηση. Στη μη επιβλεπόμενη μάθηση ακολουθούνται μέθοδοι που ομαδοποιούν τα δεδομένα-τα οποία δεν έχουν labels- σε συστάδες ανάλογα με την ομοιότητα που παρουσιάζουν. Στην ημι-επιβλεπόμενη μάθηση χρησιμοποιούνται από κοινού τόσο δεδομένα με labels όσο και χωρίς. Σε αυτή την περίπτωση υπάρχει συνήθως ένα σύνολο από δεδομένα χωρίς labels, το οποίο συμπληρώνεται με ένα μικρότερο-συνήθως περιορισμένο-σύνολο με δεδομένα που διαθέτουν labels, για την ανάπτυξη ενός ταξινομητή. Έχει παρατηρηθεί ότι αυτού του τύπου η προσέγγιση συνδυάζει καλές επιδόσεις σε ακρίβεια και απαιτεί σημαντικότερα λιγότερη προσπάθεια σε σχέση με αυτή της επιβλεπόμενης μάθησης. Τέλος στην επιβλεπόμενη μάθηση το εκάστοτε μοντέλο εκπαιδεύεται σε ένα σετ δεδομένων με labels και χρησιμοποιείται ακολούθως για τη διενέργεια προβλέψεων σε άγνωστα δεδομένα. Γενικά αυτή η τελευταία προσέγγιση επιφέρει τις περισσότερες φορές τα καλύτερα αποτελέσματα, αλλά εξαρτάται σε μεγάλο βαθμό από τα δεδομένα εκπαίδευσης, τα οποία σε πολλές περιπτώσεις είναι δύσκολο να συγκεντρωθούν και απαιτούν σημαντικό κόπο.

- **Προσεγγίσεις βασισμένες στη Βαθιά Μηχανική Μάθηση**. Η Βαθιά Μηχανική Μάθηση είναι ένα υπο-πεδίο της Μηχανικής Μάθησης που χρησιμοποιεί βαθιά νευρωνικά δίκτυα. Κατά καιρούς έχει γίνει χρήση διαφόρων τέτοιου τύπου δικτύων και συνδυασμοί αυτών με πολύ υποσχόμενα αποτελέσματα. Ενδεικτικά, έχει γίνει χρήση CNN, LSTM, GRU, ANN, UPN, Bi-LSTM και συνδυασμοί αυτών. Το πεδίο αυτό τα τελευταία χρόνια παρουσιάζει έντονη ερευνητική δραστηριότητα με πληθώρα δημοσιεύσεων σε ετήσια βάση.

- **Προσεγγίσεις βασισμένες σε αρχεία λέξεων**. Αυτού του τύπου οι προσεγγίσεις σκανάρουν το κείμενο για λέξεις που εκφράζουν θετικό ή αρνητικό περιεχόμενο. Παράδειγμα αρνητικών λέξεων είναι οι "κακός", "άσχημος", "τρομακτικός". Παράδειγμα θετικών λέξεων είναι οι "καλός" και "όμορφος". Οι τιμές αυτών των λέξεων καταγράφονται σε ένα αρχείο λέξεων. Παρατηρείται ότι οι λέξεις με μεγάλες θετικές ή αρνητικές τιμές είναι συνήθως επίθετα ή επιρρήματα. Το μειονέκτημα της προσέγγισης αυτής είναι ότι για κάθε πεδίο ενδιαφέροντος, το αρχείο των λέξεων θα πρέπει να προσαρμόζεται, καθώς η ανάλυση συναισθήματος εξαρτάται γενικά σε μεγάλο βαθμό από αυτό. Αυτή η διαδικασία δυστυχώς μπορεί να αποβεί εξαιρετικά χρονοβόρα. Ωστόσο το βασικό πλεονέκτημα αυτών των προσεγγίσεων είναι ότι δεν απαιτούν δεδομένα εκπαίδευσης, η συλλογή των οποίων, όπως έχει ήδη αναφερθεί είναι συνήθως μια εξαιρετικά χρονοβόρα και δαπανηρή διαδικασία. Εν γένει, οι προσεγγίσεις αυτές διαχωρίζονται σε δύο κατηγορίες, τις βασισμένες σε λεξικό και τις βασισμένες σε σωρό κειμένων (corpus). Για την πρώτη κατηγορία υπάρχει στην αρχή ένα μικρό σύνολο λέξεων, το οποίο επεκτείνεται επαναληπτικά με συνώνυμα και αντώνυμα αυτών από υπάρχοντα λεξικά. Για εφαρμογές γενικού σκοπού αυτή η προσέγγιση αποδίδει καλύτερα αποτελέσματα. Η δεύτερη προσέγγιση προσαρμόζεται σε συγκεκριμένες εφαρμογές. Στην αρχή υπάρχει ένα σύνολο λέξεων γενικού σκοπού το οποίο επεκτείνεται μέσω λέξεων που ανακαλύπτονται σε ένα σωρό κειμένων, στη βάση μοτίβων συνύπαρξης.

- **Υβριδικές προσεγγίσεις**. Στις προσεγγίσεις αυτές γίνεται χρήση αφενός τεχνικών Μηχανικής Μάθησης (και Βαθιάς Μηχανικής Μάθησης) και αφετέρου τεχνικών βασισμένων σε αρχεία λέξεων, όπως η χρήση Οντολογιών και Σημασιολογικών Δικτύων, για την ανίχνευση σημασιών που εκφράζονται πλαγίως στο κείμενο. Απώτερος σκοπός είναι η επίτευξη σημαντικών επιδόσεων, μέσω της αποδοτικής χρήσης χαρακτηριστικών

και των δύο προσεγγίσεων, ξεπερνώντας έτσι τους περιορισμούς και τις ελλείψεις που παρουσιάζουν μεμονωμένα.

Κάποιες από τις βασικές επιμέρους εφαρμογές της Ανάλυσης Συναισθήματος είναι:

- **Ταξινόμηση Συναισθήματος(Sentiment**. Αποτελεί μια από τις πιο δημοφιλείς και ευρέως ερευνηθείσες εφαρμογές της Ανάλυσης Συναισθήματος. Πιο γνωστό είναι το υποπεδίο της απόφανσης περί της συναισθηματικής πόλωσης στο εκάστοτε κείμενο. Παραδοσιακά, η πόλωση ταξινομείται ως θετική ή αρνητική, με κάποιες έρευνες να εισάγουν και μια τρίτη ουδέτερη κατηγορία.

- **Ταξινόμηση Υποκειμενικότητας(Subjectivity Classification)**. Είναι μια εφαρμογή για την απόφανση περί της ύπαρξης ή μη υποκειμενικότητας στο κείμενο. Κύριος στόχος είναι ο αποκλεισμός ανεπιθύμητων αντικειμενικών δεδομένων από περαιτέρω επεξεργασία, για αυτό και συνήθως εφαρμόζεται ως πρώτο βήμα στην Ανάλυση Συναισθήματος. Επιτυγχάνεται με τον εντοπισμό υποκειμενικών στοιχείων, λέξεων δηλαδή με συναισθηματικό φορτίο ή με υποκειμενικές έννοιες όπως "ακριβός", "εύκολος", "καλός" μεταξύ άλλων. Αυτά τα στοιχεία χρησιμοποιούνται για να ταξινομήσουν τα τμήματα του κειμένου ως αντικειμενικά ή υποκειμενικά.

- **Εντοπισμός παραπλανητικής γνώμης(Opinion Spam Detection)**. Η έρευνα πάνω σε αυτή την εφαρμογή άνθισε με την εξέλιξη του λεγόμενου ηλεκτρονικού εμπορίου. Οι παραπλανητικές γνώμες εντοπίζονται συνήθως υπό μορφή έξυπνα γραμμένων σχολίων τα οποία είτε προωθούν είτε υποτιμούν ένα προϊόν, χωρίς η γνώμη αυτή να ανταποκρίνεται στην αλήθεια. Η έρευνα επικεντρώνεται στην αναγνώριση τριών τύπων χαρακτηριστικών που σχετίζονται με μια ψεύτικη κριτική: περιεχόμενο της κριτικής, μεταδεδομένα της κριτικής και γνώση για το προϊόν στο φυσικό κόσμο. Συγκεκριμένα, το περιεχόμενο της κριτικής αναλύεται συνήθως με τεχνικές Μηχανικής Μάθησης για τον εντοπισμό εξαπάτησης. Αντίστοιχα, τα μεταδεδομένα που περιλαμβάνουν τη διεύθυνση IP, την αξιολόγηση αστεριών, την τοποθεσία και το id του χρήστη μεταξύ άλλων, δεν είναι πολλές φορές άμεσα διαθέσιμα για περαιτέρω επεξεργασία και εξαγωγή συμπερασμάτων. Το τρίτο χαρακτηριστικό είναι η γνώση του φυσικού κόσμου. Για παράδειγμα εάν ένα προϊόν έχει καλή φήμη και ξαφνικά σε μια χρονική περίοδο ένα υποδεέστερο προϊόν αξιολογείται ως ανώτερο, τότε οι κριτικές αυτής της περιόδου καθίστανται αυτόματα ύποπτες.

- **Εντοπισμός έμμεσου λόγου(Implicit Language Detection)**. Ο έμμεσος λόγος αναφέρεται σε χιούμορ, ειρωνεία ή σαρκασμό. Γενικά υπάρχει σημαντική ασάφεια ως προς τη χρησιμοποίησή τους, κάτι που καθιστά την αναγνώριση τους δύσκολη κάποιες φορές ακόμα και από ανθρώπους. Ωστόσο η αναγνώριση αυτή είναι κομβικής σημασίας, καθώς μια τέτοια χρήση του λόγου μπορεί να ανατρέψει την πόλωση ενός κειμένου. Πολλές μέθοδοι αξιοποιούν τη χρήση emoticons, τις εκφράσεις που δηλώνουν έντονο γέλιο και την έντονη χρήση σημείων στίξης για την αναγνώριση τέτοιων χρήσεων της γλώσσας.

- **Εξαγωγή Πτυχών(Aspect Extraction)**. Αναφέρεται στην ανάκτηση οντοτήτων-στόχων και πτυχών αυτών στο κείμενο. Η οντότητα-στόχος μπορεί να είναι ένα προϊόν, ένα πρόσωπο, ένας οργανισμός κ.α. Οι γνώμες των ατόμων για τις διάφορες πτυχές της οντότητας-στόχου πρέπει να αναγνωρισθούν για τη διενέργεια Ανάλυσης Συναισθήματος. Υπάρχουν διάφορες μέθοδοι για την εξαγωγή πτυχών. Μια από τις πρώτες που χρησιμοποιήθηκαν βασίζεται στην ανάλυση συχνότητας. Συγκεκριμένα, εντοπίζονται

συχνά χρησιμοποιημένα ουσιαστικά που είναι πιθανόν να αποτελούν πτυχές. Βάσει ενός γνωστού εμπειρικού κανόνα, εάν ένα ρήμα χρησιμοποιείται σε τουλάχιστον 1% των προτάσεων τότε θεωρείται ως πτυχή. Επιπρόσθετα, υπάρχουν μέθοδοι που βασίζονται στη σύνταξη. Οι μέθοδοι αυτές επικεντρώνονται στις συντακτικές σχέσεις που βρίσκονται οι εκάστοτε πτυχές, για τον εντοπισμό τους.

Η Ανάλυση Συναισθήματος γίνεται τέλος σε διαφορετικά επίπεδα. Πιο συγκεκριμένα:

- **Ανάλυση σε επίπεδο εγγράφου**. Σε αυτού του τύπου την ανάλυση το έγγραφο θεωρείται μονάδα. Είναι μια απλουστευμένη εφαρμογή στην οποία γίνεται η παραδοχή ότι όλο το έγγραφο περιέχει την αμιγή γνώμη ενός ατόμου. Στην πράξη αυτό βέβαια δε συμβαίνει πάντα.

- **Ανάλυση σε επίπεδο πρότασης**. Σε αυτού του τύπου την ανάλυση μονάδα ανάλυσης θεωρείται η πρόταση. Χρησιμοποιείται ιδιαίτερα στην ταξινόμηση της υποκειμενικότητας, όπου αναλύεται η εκάστοτε πρόταση για να διαπιστωθεί εάν περιέχει γεγονότα ή συναισθήματα/γνώμες.

- **Ανάλυση σε επίπεδο πτυχής**. Αφορά την ανάλυση συναισθημάτων για συγκεκριμένες οντότητες και των πτυχών αυτών, σε ένα έγγραφο κειμένου. Είναι επίσης γνωστή ως ανάλυση σε επίπεδο χαρακτηριστικών ή ανάλυση σε επίπεδο οντοτήτων. Η πρόκληση σε αυτού του τύπου την ανάλυση έγκειται κυρίως στο ότι ενώ το γενικότερο συναίσθημα που αποπνέει ένα έγγραφο μπορεί να είναι θετικό ή αρνητικό, ο συγγραφέας του δύναται να έχει διαφορετικές απόψεις για διαφορετικές πτυχές μιας επιμέρους οντότητας.

Όλα τα παραπάνω παρουσιάζονται συνοπτικά και στο Σχήμα 5.1.

# 6

# Σετ Δεδομένων και Benchmarks

Για την αξιολόγηση των επιδόσεων των μοντέλων σε διάφορες εφαρμογές χρησιμοποιούνται benchmarks και σετ δεδομένων, τα σημαντικότερα εκ των οποίων παρουσιάζονται στις επόμενες ενότητες.

## 6.1 Το σετ δεδομένων SNLI(Stanford Natural Language Inference)

Παρουσιάζεται το 2015 στο πλαίσιο της δημοσίευσης "A large annotated corpus for learning natural language inference" [30] από την ερευνητική ομάδα του Πανεπιστημίου του Stanford. αποτελείται από 570000 ζεύγη προτάσεων για καθένα από τα οποία παρέχονται labels για το κατά πόσο η μια πρόταση συνάγεται από την προηγούμενη, εάν δημιουργεί αντίφαση ή εάν είναι ουδέτερη. Τα ζεύγη αποτελούνται από μια εισαγωγική πρόταση-προκείμενο και ένας εργαζόμενος που έχει χαρακτηριστεί υψηλών επιδόσεων, κλήθηκε να συμπληρώσει το ζεύγος με υποθέσεις που είτε συνάγονται από την προκείμενη πρόταση, είτε έρχονται σε αντίφαση με αυτήν είτε είναι ουδέτερες ως προς αυτή.

## 6.2 Το σετ δεδομένων MultiNLI(Multi-Genre Natural Language Inference)

Παρουσιάζεται το 2018 με τη δημοσίευση "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference" [31] από ερευνητές του New York University. Περιλαμβάνει 433000 ζεύγη προτάσεων και η διαδικασία συλλογής των δεδομένων μοντελοποιήθηκε με βάση το σετ δεδομένων SNLI. Δημιουργήθηκε για να αντιμετωπιστούν προβλήματα που είχαν ανακύψει με το SNLI, τα οποία το καθιστούσαν τελικά όχι αρκετά απαιτητικό για την αξιολόγηση των επιδόσεων των μοντέλων. Ιδιαίτερα το SNLI χρησιμοποιούσε πηγές από ένα μόνο είδος-λεζάντες εικόνων- και λόγω αυτού ήταν αρκετά περιορισμένες σε περιγραφές συγκεκριμένων οπτικών σκηνών. Συνεπακόλουθα και οι διατυπωθείσες υποθέσεις ήταν σύντομες και απλές, καθιστώντας την εμφάνιση ορισμένων φαινομένων αρκετά σπάνια για να χρησιμοποιηθούν αποτελεσματικά στην αξιολόγηση των επιδόσεων των μοντέλων. Στον αντίποδα, το νέο σετ δεδομένων αντλεί προτάσεις από δέκα διαφορετικές πηγές, αρκετά ετερόκλητες μεταξύ τους, προκειμένου να αναπαριστούν όλο το εύρος της Αγγλικής Γλώσσας. Ενδεικτικά συμπεριλήφθηκαν κυβερνητικές πηγές, καθομιλουμένη σε πρόσωπο με πρόσωπο, γράμματα, τηλεφωνικές συνομιλίες και ταξιδιωτικοί οδηγοί μεταξύ άλλων. Επιπρόσθετα έγινε προσπάθεια να συμπεριληφθούν μεγαλύτερες σε μήκος προτάσεις σε σχέση με το αρχικό SNLI. Η

| Tag | Dev. Freq. SNLI | Dev. Freq. MultiNLI | Diff. | Most Frequent Label Label | Most Frequent Label % | Model Acc. CBOW | Model Acc. BiLSTM | Model Acc. ESIM |
|---|---|---|---|---|---|---|---|---|
| Entire Corpus | 100 | 100 | 0 | entailment | ∼35 | ∼65 | ∼67 | ∼72 |
| Pronouns (PTB) | 34 | 68 | 34 | entailment | 34 | 66 | 68 | 73 |
| Quantifiers | 33 | 63 | 30 | contradiction | 36 | 66 | 68 | 73 |
| Modals (PTB) | <1 | 28 | 28 | entailment | 35 | 65 | 67 | 72 |
| Negation (PTB) | 5 | 31 | 26 | contradiction | **48** | 67 | 70 | 75 |
| WH terms (PTB) | 5 | 30 | 25 | entailment | 35 | 64 | 65 | 72 |
| Belief Verbs | <1 | 19 | 18 | entailment | 34 | 64 | 67 | 71 |
| Time Terms | 19 | 36 | 17 | neutral | 35 | 64 | 66 | 71 |
| Discourse Mark. | <1 | 14 | 14 | neutral | 34 | 62 | 64 | 70 |
| Presup. Triggers | 8 | 22 | 14 | neutral | 34 | 65 | 67 | 73 |
| Compr./Supr.(PTB) | 3 | 17 | 14 | neutral | 39 | 61 | 63 | 69 |
| Conditionals | 4 | 15 | 11 | neutral | 35 | 65 | 68 | 73 |
| Tense Match (PTB) | 62 | 69 | 7 | entailment | 37 | 67 | 68 | 73 |
| Interjections (PTB) | <1 | 5 | 5 | entailment | 36 | 67 | 70 | 75 |
| >20 words | <1 | 5 | 5 | entailment | **42** | 65 | 67 | 76 |

Σχήμα 6.1: Οι διαφορετικές συχνότητες με τις οποίες εμφανίζονται τα διάφορα ιδιαίτερα φαινόμενα ανάμεσα στα SNLI και το MNLI

διαδικασία διαμόρφωσης των υποθέσεων είναι η ίδια που ακολουθήθηκε για το SNLI, καθότι αυτός ο τρόπος ήταν αρκετά επιτυχημένος και δεν δημιούργησε στην πορεία κάποιο πρόβλημα. Εν κατακλείδι, το MultiNLI είναι ένα πιο σύνθετο και απαιτητικό σετ δεδομένων, με μεγάλος εύρος θεμάτων και με πολλά ιδιαίτερα φαινόμενα, κρίσιμα για την αξιολόγηση της επίδοσης του εκάστοτε μοντέλου να συνάγει πάνω σε Φυσική Γλώσσα, να παρουσιάζονται σε μεγάλο ποσοστό των προτάσεων όπως φαίνεται και στο Σχήμα 6.1.

## 6.3 GLUE Benchmark

Το GLUE(General Language Understanding Evaluation) benchmark δημοσιεύθηκε το 2019 από κοινού από το ερευνητικό εργαστήριο Courant της Νέας Υόρκης, από το πανεπιστήμιο της Washington και από το εργαστήριο της DeepMind [32]. Στόχος των ερευνητών ήταν να ωθήσουν τη μελλοντική έρευνα στο πεδίο προς την κατεύθυνση της ανάπτυξης ενοποιημένων μοντέλων που θα δύνανται να χρησιμοποιούνται σε ένα εύρος γλωσσικών εφαρμογών σε διαφορετικά πεδία. Σε αντιδιαστολή, τα έως τότε μοντέλα Κατανόησης Φυσικής Γλώσσας πάνω από το επίπεδο λέξης, μπορούσαν να φέρουν εις πέρας μια συγκεκριμένη εφαρμογή αλλά αποτύγχαναν με δεδομένα εκτός του πεδίου. Για το σκοπό αυτό δημιουργείται το GLUE που περιλαμβάνει αφενός μια συλλογή εφαρμογών Κατανόησης Φυσικής Γλώσσας -μεταξύ των οποίων είναι η Απάντηση Ερωτημάτων και η Ανάλυση Συναισθήματος- και αφετέρου μια διαδικτυακή πλατφόρμα για την αξιολόγηση των μοντέλων, την ανά μεταξύ τους σύγκριση και την ανάλυσή τους. Το GLUE δεν επιβάλλει περιορισμούς στην αρχιτεκτονική των επιμέρους μοντέλων, πέρα από την απαίτηση να μπορούν να επεξεργάζονται προτάσεις μεμονωμένα αλλά και κατά ζεύγη. Επιπλέον τα σετ δεδομένων που χρησιμοποιεί δεν δημιουργήθηκαν εξ αρχής, άλλα επιλέχθηκαν από ήδη ενυπάρχοντα στην κοινότητα, τα οποία είναι ευρέως αποδεκτά για το ενδιαφέρον που παρουσιάζουν καθώς και για το επίπεδο δυσκολίας τους. Τέσσερα από αυτά διαχειρίζονται από ιδιώτες, προκειμένου να αποφευχθούν φαινόμενα αθέμιτης χρήσης του benchmark.

Επιπρόσθετα παρέχεται και ένα μικρό σετ δεδομένων κατασκευασμένο από ειδικούς για διαγνωστικούς σκοπούς, για να διαπιστώνεται δηλαδή εάν το μοντέλο δύναται να αντιμετωπίσει

κάποια πολύ συχνά εμφανιζόμενα γλωσσικά φαινόμενα, που φαίνονται στο σχήμα 6.2. Συνολικά παρέχονται εννιά διαφορετικές εφαρμογές και τα ανάλογα σετ δεδομένων. Αυτές ανήκουν σε διαφορετικά επιμέρους πεδία με την αξιολόγηση του μοντέλου να σχεδιάζεται έτσι ώστε οι επιδόσεις του να είναι υψηλές όταν εκείνο τα πηγαίνει αρκούντως καλά σε όλα. Χρησιμοποιείται ένας macro-μέσος όρος για τη συνολική επίδοση, ενώ παρέχονται και τα αποτελέσματα ανά επιμέρους εφαρμογή. Οι εφαρμογές παρουσιάζονται συνοπτικά παρακάτω[32].

- **CoLA(The Corpus of Linguistic Acceptability)**. Αποτελείται από προτάσεις λέξεων ειλημμένες από βιβλία και επιστημονικά άρθρα. Κάθε πρόταση περιέχει και ένα label για το εάν είναι μια γραμματικά σωστή πρόταση στην αγγλική γλώσσα.

- **SST-2(The Stanford Sentiment Treebank)**. Αποτελείται από προτάσεις ειλημμένες από κριτικές ταινιών με τα labels να περιγράφουν το συναίσθημα που αποπνέουν. Το μοντέλο καλείται να προβλέψει αυτό το συναίσθημα.

- **MRPC(The Microsoft Research Paraphrase Corpus)**. Αποτελείται από ζεύγη προτάσεων που εξάγονται αυτόματα από διαδικτυακές ειδησεογραφικές ιστοσελίδες. Περιλαμβάνουν labels με την κρίση του κατά πόσο οι προτάσεις του ζεύγους είναι σημασιολογικά ισοδύναμες, κάτι που αποτελεί και το αντικείμενο μάθησης του μοντέλου.

- **QQP(The Quora Question Pairs)**. Περιλαμβάνει ζεύγη προτάσεων από την ιστοσελίδα απάντησης ερωτημάτων της κοινότητας Quora. Περιλαμβάνουν, ομοίως με το παραπάνω, labels με την κρίση του κατά πόσο οι προτάσεις του ζεύγους είναι σημασιολογικά ισοδύναμες.

- **STS-B(The Semantic Textual Similarity Benchmark)**. Περιλαμβάνει ζεύγη προτάσεων, αντλούμενα από τίτλους ειδήσεων, από λεζάντες από εικόνες και βίντεο και από δεδομένα που προέκυψαν από συναγωγή φυσικής γλώσσας. Κάθε ζεύγος έχει ως label έναν αριθμό από το 1 έως το 5 ανάλογα με την ομοιότητα των επιμέρους προτάσεων. Στόχος της εκπαίδευσης είναι η πρόβλεψη αυτού του αριθμού.

- **MNLI(The Multi-Genre Natural Language Inference Corpus)**. Αποτελεί μια συλλογή από ζεύγη προτάσεων που έχουν συλλεχθεί από δέκα διαφορετικές πηγές, όπως κυβερνητικές αναφορές, έργα φαντασίας και μεταγεγραμμένο λόγο μεταξύ άλλων. Δεδομένης μιας εισαγωγικής πρότασης και μιας πρότασης υπόθεσης, το μοντέλο καλείται να αποφανθεί για το εάν η υπόθεση συνεπάγεται της εισαγωγής, εάν την αντικρούει ή εάν είναι άσχετη προς αυτήν.

- **QNLI(The Stanford Question Answering Dataset)**. Είναι ένα σετ δεδομένων για την εκπαίδευση μοντέλων στην απάντηση ερωτημάτων. Αποτελείται από ζεύγη ερωτήσεων-παραγράφων, όπου η μια από τις προτάσεις της παραγράφου περιέχει την απάντηση στην ερώτηση όπου έχει διατυπωθεί από κάποιον άνθρωπο. Οι παράγραφοι έχουν αντληθεί από άρθρα της Wikipedia. Ο σκοπός της εκπαίδευσης είναι να αποφανθεί το μοντέλο εάν η πρόταση περιέχει όντως την απάντηση στη διατυπωθείσα ερώτηση.

- **RTE(The Recognizing Textual Entailment)**. Προέρχεται από δεδομένα που συλλέγονται από διάφορες ετήσιες προκλήσεις πάνω σε συνεπαγωγές ανάμεσα στις προτάσεις. Τα δείγματα κατασκευάζονται από ειδήσεις ή άρθρα της Wikipedia και ο στόχος της εκπαίδευσης είναι ο ίδιος με το παραπάνω σετ δεδομένων.

- **WNLI(The Winograd Schema Challenge)**. Είναι μια εφαρμογή κατανόησης αναγνωσθέντος κειμένου. Το μοντέλο διαβάζει μια πρόταση με μια αντωνυμία και

| Coarse-Grained Categories | Fine-Grained Categories |
|---|---|
| Lexical Semantics | Lexical Entailment, Morphological Negation, Factivity, Symmetry/Collectivity, Redundancy, Named Entities, Quantifiers |
| Predicate-Argument Structure | Core Arguments, Prepositional Phrases, Ellipsis/Implicits, Anaphora/Coreference Active/Passive, Nominalization, Genitives/Partitives, Datives, Relative Clauses, Coordination Scope, Intersectivity, Restrictivity |
| Logic | Negation, Double Negation, Intervals/Numbers, Conjunction, Disjunction, Conditionals, Universal, Existential, Temporal, Upward Monotone, Downward Monotone, Non-Monotone |
| Knowledge | Common Sense, World Knowledge |

Σχήμα 6.2: Οι τύποι των γλωσσικών φαινομένων που επισημαίνονται στο διαγνωστικό σετ δεδομένων, οργανωμένοι σε τέσσερις βασικές κατηγορίες

καλείται να επιλέξει από μια λίστα επιλογών που αναφέρεται αυτή. Για τη μετατροπή της εφαρμογής αυτής σε εφαρμογή ταξινόμησης ζεύγους προτάσεων, κατασκευάζονται ζεύγη προτάσεων με την αντικατάσταση κάθε φορά της αντωνυμίας με τις υποψήφιες λέξεις που αυτή αναφέρεται. Έτσι στην πράξη το μοντέλο αποφαίνεται για το κατά πόσο η πρόταση με την αντικατεστημένη αντωνυμία συνάγεται από την αρχική πρόταση.

## 6.4 SuperGLUE Benchmark

Δημοσιεύθηκε το 2019 στο πλαίσιο του "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems" [33] από τους ίδιους ερευνητές που δημιούργησαν το πρώτο GLUE benchmark. Η ανάγκη για τη δημιουργία αυτού του νέου benchmark προέκυψε καθώς έπειτα από την κυκλοφορία του πρώτου, δημιουργήθηκαν πολλά εξελιγμένα μοντέλα τα οποία σιγά σιγά όλο και βελτίωναν τις επιδόσεις τους σε αυτό, κάτι που είχε ως συνέπεια, αν και μόλις ένα χρόνο μετά την κυκλοφορία του, να μην αποτελεί ισχυρή πρόκληση για την αποτελεσματική αξιολόγηση των επιδόσεων των νέων αυτών μοντέλων. Το νέο benchmark σχεδιάστηκε με γνώμονα:

- **Την ουσία των εφαρμογών**. Οι εφαρμογές θα πρέπει να αξιολογούν την ικανότητα ενός μοντέλου να κατανοεί και να εξάγει συμπεράσματα.
- **Τη δυσκολία των εφαρμογών**. Οι νέες εφαρμογές επιλέχθηκαν ώστε να είναι αρκούντως δύσκολες για τα νέα SOTA μοντέλα, αλλά επιλύσιμες από τους περισσότερους Αγγλόφωνους με σπουδές τριτοβάθμιας εκπαίδευσης ή ανώτερες.
- **Την παροχή μιας αυτόματης μετρικής απόδοσης** που να αντιστοιχίζεται επαρκώς στην ανθρώπινη κρίση για ένα ικανοποιητικό αποτέλεσμα.
- **Την αξιολόγηση σε ήδη υπάρχοντα σετ δεδομένων**, τα οποία να έχουν δοκιμαστεί από την κοινότητα. Επιπρόσθετα, προτιμήθηκαν τεστ σετ με μη ιδιωτικά προσβάσιμα labels.
- **Το φορμά των εφαρμογών**. Προτιμήθηκαν εφαρμογές με σχετικά απλό φορμά εισόδου και εξόδου, για να αποφευχθεί η ανάπτυξη αρχιτεκτονικών εξειδικευμένων στην εκάστοτε εφαρμογή. Επιπλέον, το νέο benchmark σχεδιάστηκε για να υποστηρίζει

μεγαλύτερες εισόδους σε σχέση με τη μονή ακολουθία ή τα ζεύγη ακολουθιών του GLUE.

- **Την άδεια χρήσης**, συγκεκριμένα προτιμήθηκαν δεδομένα διαθέσιμα υπό άδειες που επιτρέπουν την αναδιανομή τους για ερευνητικούς σκοπούς.

Στο πνεύμα του πρώτου GLUE οι ερευνητές εισάγουν οχτώ νέες προκλήσεις για την αξιολόγηση των μελλοντικών μοντέλων. Συγκεκριμένα:

- **BoolQ (Boolean Questions)**. Κάθε δείγμα αποτελείται από ένα μικρό απόσπασμα και μια ερώτηση πάνω σε αυτό που απαντάται με ναι ή όχι. Οι ερωτήσεις παρέχονται από χρήστες της μηχανής αναζήτησης της Google και συνταιριάζονται εν συνεχεία με μια παράγραφο που περιέχει την απάντηση, ειλημμένη από τη Wikipedia.

- **CB (CommitmentBank)**. Κάθε δείγμα εκπαίδευσης περιλαμβάνει ένα εισαγωγικό απόσπασμα και μια υπόθεση πάνω σε αυτό. Το μοντέλο καλείται να απαντήσει στο κατά πόσο ο συγγραφέας του εισαγωγικού αποσπάσματος δεσμεύεται για την αλήθεια της υπόθεσης. Κάθε δείγμα κατηγοριοποιείται σε τρεις επιμέρους κατηγορίες: Συμπέρασμα, Αντιπαράθεση και Ουδέτερο.

- **COPA (Choice of Plausible Alternatives)**. Είναι μια εφαρμογή για την συλλογιστική πάνω στην αιτιότητα στο κείμενο. Συγκεκριμένα, στο μοντέλο δίνεται μια εισαγωγική πρόταση και καλείται να αποφανθεί την αιτία ή το αποτέλεσμα από δύο πιθανές επιλογές.

- **MultiRC (Multi-Sentence Reading Comprehension)**. Αποτελεί μια εφαρμογή Απάντησης Ερωτήσεων όπου κάθε δείγμα εκπαίδευσης αποτελείται από μια παράγραφο, μια ερώτηση για την παράγραφο και μια λίστα πιθανών απαντήσεων. Το μοντέλο καλείται να προβλέψει ποιες απαντήσεις είναι αληθείς και ποιες ψευδείς.

- **ReCoRD (Reading Comprehension with Commonsense Reasoning Dataset**). αποτελεί μια εφαρμογή Απάντησης Ερωτήσεων πολλαπλής επιλογής. Κάθε δείγμα αποτελείται από ένα ειδησεογραφικό άρθρο και μια ερώτηση επί αυτού, με μια οντότητά της να διατηρείται κρυφή με χρήση μάσκας. Το μοντέλο θα πρέπει να προβλέψει την κρυφή οντότητα από μια λίστα πιθανών οντοτήτων από το αρχικό άρθρο.

- **RTE (Recognizing Textual Entailment)**. Αυτή η εφαρμογή είναι η ίδια με αυτήν του αρχικού GLUE.

- **WiC (Word-in-Context)**. Σε αυτή την εφαρμογή παρέχονται δύο αποσπάσματα κειμένου και μια πολύσημη λέξη που εμφανίζεται και στα δύο. Το μοντέλο καλείται να αποφανθεί για το κατά πόσον η λέξη χρησιμοποιείται με την ίδια σημασία και στα δύο αποσπάσματα.

- **WSC (Winograd Schema Challenge)**. Περιλαμβάνει δείγματα εκπαίδευσης που αποτελούνται από μια πρόταση με αντωνυμία και μια λίστα με φράσεις που περιλαμβάνουν ουσιαστικό από την πρόταση. Το μοντέλο θα πρέπει να επιλέξει τη φράση με το σωστό ρήμα στο οποίο αναφέρεται η πρόταση.

Εν κατακλείδι, το νέο benchmark παρέχει ένα πλούσιο και απαιτητικό πεδίο εργασίας για την ανάπτυξη νέων αλγορίθμων μηχανικής μάθησης γενικού σκοπού, για την κατανόηση της Φυσικής Γλώσσας.

## 6.5 Το σετ δεδομένων SQuAD

Το σετ δεδομένων SQuAD(Stanford Question Answering Dataset) παρουσιάστηκε το 2016 στο πλαίσιο της δημοσίευσης "SQuAD: 100,000+ Questions for Machine Comprehension of Text" [34] από τους ερευνητές του τμήματος της Επιστήμης Υπολογιστών του Πανεπιστημίου του Stanford. Περιέχει εν γένει ένα σύνολο από ζεύγη ερωτήσεων και απαντήσεων και αποτελεί μια ισχυρή πρόκληση για τα μοντέλα Επεξεργασίας Φυσικής Γλώσσας, στην εφαρμογή της Απάντησης Ερωτημάτων. Στην εφαρμογή αυτή εξετάζεται η ικανότητα ενός μοντέλου, αφού "διαβάσει" ένα τμήμα κειμένου, να μπορεί να απαντήσει ερωτήσεις πάνω σε αυτό. Οι δημιουργοί του σετ δεδομένων άντλησαν πληροφορίες -με τυχαίο τρόπο- από 536 από τα 10000 κορυφαία άρθρα της Wikipedia. Από αυτά τα άρθρα-και αφού εξαιρέθηκαν σχήματα, πίνακες και πολύ μικρές παράγραφοι- επιλέχθηκαν 23215 παράγραφοι που καλύπτουν ένα μεγάλο εύρος διαφορετικών θεμάτων. Τα δείγματα αυτά χωρίστηκαν τυχαία εν συνεχεία σε σετ εκπαίδευσης(80%), σετ ανάπτυξης(10%) και σετ αξιολόγησης(10%). Εν συνεχεία, επιστρατεύθηκαν εργαζόμενοι υψηλών απόδοσης για τη δημιουργία ερωτήσεων. Συγκεκριμένα, για κάθε επιμέρους παράγραφο κλήθηκαν να διατυπώσουν πέντε ερωτήσεις με δικά τους λόγια και να επισημάνουν που αυτές απαντώνται μέσα στην παράγραφο. Για την κατανόηση των ιδιοτήτων του σετ δεδομένων που προέκυψε και για την αξιολόγηση της ποιότητάς του, οι ερευνητές διερεύνησαν τις ακόλουθες πτυχές:

- **Ποικιλία στις κατηγορίες των απαντήσεων**. Κάθε απάντηση κατηγοριοποιήθηκε αναλόγως σε κάποια από τις κατηγορίες "ημερομηνία", "άλλο αριθμητικό", "άτομο", "τοποθεσία", "άλλη οντότητα", "φράση με ρήμα", "φράση με επίθετο", "υποπρόταση" και "άλλο".

- **Συλλογιστική που απαιτείται για την απάντηση στις ερωτήσεις**. Για το σκοπό αυτό, αφού λήφθηκε ένα μικρό δείγμα από το σετ ανάπτυξης, κατατάχθηκαν χειροκίνητα οι ερωτήσεις σε κατηγορίες ανάλογα με το είδος του συλλογισμού που απαιτείται για την απάντησή τους. Αυτές φαίνονται στο Σχήμα 6.3Για παράδειγμα, στην κατηγορία "συντακτική τροποποίηση" ανήκουν ερωτήσεις που είναι παραφρασμένες και άρα που η εύρεση της απάντησης απαιτεί την αναδιάταξη λέξεων στην παράγραφο. Οι επιμέρους κατηγορίες και τα ποσοστά αυτών φαίνονται στον πίνακα.

- **Συντακτική απόκλιση**. Μετράται ανάμεσα σε μια ερώτηση και στην πρόταση που περιέχει την απάντηση για να αξιολογηθεί η δυσκολία της ερώτησης. Πρόκειται στην ουσία για μια μετρική που αποτιμά των αριθμό των τροποποιήσεων που πρέπει να γίνουν ώστε να μετατραπεί μια ερώτηση στην πρόταση που δίνει την απάντησή της.

Εν γένει, το σετ δεδομένων SQuAD είναι από τα πλέον δημοφιλή και ευρέως χρησιμοποιούμενα στο χώρο, κάτι που οφείλεται μεταξύ άλλων στο μέγεθός του, στο επίπεδο δυσκολίας του-καθώς η απάντηση στην εκάστοτε ερώτηση βρίσκεται μόνο σε ένα σημείο- και στην απαίτηση ανάπτυξης συλλογιστικής για την απάντηση των ερωτήσεων.

## 6.6 Το σετ δεδομένων SQuAD 2.0

Εισάγεται το 2018 με τη δημοσίευση "Know What You Don't Know: Unanswerable Questions for SQuAD" [35] από την ερευνητική ομάδα που εισήγαγε και το πρώτο SQuAD και αποτελεί στην ουσία επέκτασή του. Οι ερευνητές διαπίστωσαν ότι τα μοντέλα που εμφανίζουν υψηλές επιδόσεις στο πρώτο SQuAD απέχουν αρκετά από μια πραγματική κατανόηση της Φυσικής Γλώσσας. Ένας από τους βασικούς λόγους είναι ότι το μοντέλο επικεντρώνεται

| Reasoning | Description | Example | Percentage |
|---|---|---|---|
| Lexical variation (synonymy) | Major correspondences between the question and the answer sentence are synonyms. | Q: What is the Rankine cycle sometimes **called**? Sentence: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle. | 33.3% |
| Lexical variation (world knowledge) | Major correspondences between the question and the answer sentence require world knowledge to resolve. | Q: Which **governing bodies** have veto power? Sen.: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process. | 9.1% |
| Syntactic variation | After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence even after local modifications. | Q: What Shakespeare scholar **is currently on the faculty**? Sen.: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington. | 64.1% |
| Multiple sentence reasoning | There is anaphora, or higher-level fusion of multiple sentences is required. | Q: What collection does **the V&A Theatre & Performance galleries** hold? Sen.: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance. | 13.6% |
| Ambiguous | We don't agree with the crowdworkers' answer, or the question does not have a unique answer. | Q: What is the main goal of criminal punishment? Sen.: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment. | 6.1% |

Σχήμα 6.3: Οι κατηγορίες της συλλογιστικής

σε ερωτήσεις για τις οποίες η ύπαρξη μιας απάντησης είναι εγγυημένη. Έτσι ψάχνει στην ουσία για την πρόταση που φαίνεται πιο σχετική προς την ερώτηση αντί να ελέγχει εάν η απάντηση προκύπτει στο κείμενο. Για το σκοπό αυτό το αρχικό σετ δεδομένων εμπλουτίζεται με 53775 επιπλέον ερωτήσεις για τις οποίες δεν προκύπτει απάντηση στο αντίστοιχο κείμενο. Οι νέες αυτές ερωτήσεις υποβλήθηκαν από εργαζόμενους όπως και στο πρώτο σετ δεδομένων και βασικός γνώμονας ήταν να είναι αφενός σχετικές με το θέμα της παραγράφου, ώστε να μην είναι πολύ εύκολο να αποφανθεί το μοντέλο εάν είναι αναπάντητες, και αφετέρου να υπάρχουν πιθανές απαντήσεις στην εκάστοτε παράγραφο, υπό την έννοια ότι θα πρέπει να υπάρχει κάποιο σημείο του οποίου ο τύπος να είναι ίδιος με τον τύπο της προσδοκώμενης απάντησης όπως αυτή αναμένεται από την ερώτηση. Για την ποιοτική ανάλυση του νέου σετ δεδομένων μελετήθηκαν κάποιες από αυτές τις νέες ερωτήσεις και κατηγοριοποιήθηκαν όπως φαίνεται στον Σχήμα 6.4 ανάλογα με τα φαινόμενα που παρουσιάζουν. Εν γένει το νέο σετ δεδομένων επεκτείνει το παλαιότερο και οδηγεί δυνητικά τα καινούρια μοντέλα σε μια κατανόηση της Φυσικής Γλώσσας σε βαθύτερο επίπεδο.

## 6.7 Το σετ δεδομένων GoEmotions

Παρουσιάστηκε το 2020 στο πλαίσιο της δημοσίευσης "GoEmotions: A Dataset of Fine-Grained Emotions" [36], υπό τη σύμπραξη του Πανεπιστημίου του Stanford, της Google και της Amazon. Όπως μπορεί να συναχθεί και από το όνομά του, το σετ δεδομένων αυτό δημιουργήθηκε για την εκπαίδευση μοντέλων πάνω στην Ανάλυση Συναισθήματος. Ενώ στο παρελθόν είχαν δημιουργηθεί διάφορα σετ δεδομένων για αυτό το υποπεδίο της Επεξεργασίας Φυσικής Γλώσσας, για διάφορες εφαρμογές όπως tweets και επικεφαλίδες ειδήσεων, αυτά αντιμετώπιζαν τα ακόλουθα δύο προβλήματα:

- **Μικρός αριθμός δειγμάτων εκπαίδευσης**, συνήθως μέχρι μερικές χιλιάδες. Αυτό δυσχεραίνει την εκπαίδευση των μοντέλων.

| Reasoning | Description | Example | Percentage |
|---|---|---|---|
| Negation | Negation word inserted or removed. | Sentence: "*Several hospital pharmacies have decided to outsource high risk preparations . . .*" Question: "*What types of pharmacy functions have **never** been outsourced?*" | 9% |
| Antonym | Antonym used. | S: "*the extinction of the dinosaurs. . . allowed the tropical rainforest to spread out across the continent.*" Q: "*The extinction of what led to the **decline** of rainforests?*" | 20% |
| Entity Swap | Entity, number, or date replaced with other entity, number, or date. | S: "*These values are much greater than the 9–88 cm as projected . . . in its Third Assessment Report.*" Q: "*What was the projection of sea level increases in the **fourth assessment report**?*" | 21% |
| Mutual Exclusion | Word or phrase is mutually exclusive with something for which an answer is present. | S: "*BSkyB. . . waiv[ed] the charge for subscribers whose package included two or more premium channels.*" Q: "*What service did BSkyB **give away for free unconditionally**?*" | 15% |
| Impossible Condition | Asks for condition that is not satisfied by anything in the paragraph. | S: "*Union forces left Jacksonville and confronted a Confederate Army at the Battle of Olustee. . . Union forces then retreated to Jacksonville and held the city for the remainder of the war.*" Q: "*After what battle did Union forces leave Jacksonville **for good**?*" | 4% |
| Other Neutral | Other cases where the paragraph does not imply any answer. | S: "*Schuenemann et al. concluded in 2011 that the Black Death. . . was caused by a variant of Y. pestis. . .*" Q: "*Who **discovered** Y. pestis?*" | 24% |
| Answerable | Question is answerable (i.e. dataset noise). | | 7% |

Σχήμα 6.4: Οι κατηγορίες των αναπάντητων ερωτήσεων

- **Μικρός αριθμός διαφορετικών συναισθημάτων στα οποία ταξινομούνται τα δείγματα εκπαίδευσης**. Έτσι η εκπαίδευση δεν παράγει αρκούντως λεπτομερή συμπεράσματα για το υποκείμενο συναίσθημα του κάθε δείγματος. Συγκεκριμένα, το σετ δεδομένων με τα περισσότερα έως τότε διαφορετικά συναισθήματα αριθμούσε 14.

Προκειμένου να ξεπεραστούν αυτοί οι σκόπελοι, οι ερευνητές συγκέντρωσαν 58000 προσεκτικά επιλεγμένα σχόλια από την ιστοσελίδα δημόσιας συζήτησης Reddit. Εν συνεχεία τα κατηγοριοποίησαν χειροκίνητα σε 27 διαφορετικές κατηγορίες συναισθημάτων και σε μια ουδέτερη. Η συλλογή των σχολίων έγινε από δημοφιλή subreddits με τουλάχιστον 10000 σχόλια το καθένα. Επιπρόσθετα, ακολουθήθηκαν οι παρακάτω ενέργειες:

- **Μείωση χυδαιοτήτων**. Για το σκοπό αυτό συγκεκριμένα subreddits αποκλείστηκαν ολοκληρωτικά. Διατηρήθηκαν όμως σχόλια με αγενές περιεχόμενο, καθώς αυτό θεωρήθηκε κομβικό για την απόφανση υπέρ των αρνητικών σχολίων

- **Χειροκίνητη επισκόπηση**, για την αφαίρεση προσβλητικών σχολίων ενάντια σε συγκεκριμένες εθνότητες, φύλο, σεξουαλικό προσανατολισμό και άτομα με αναπηρίες

- **Φιλτράρισμα μήκους σχολίων**. Επιλέγονται σχόλια μήκους 3-30 tokens

- **Εξισορρόπηση σε επίπεδο subreddits**. Συγκεκριμένα, αποφεύγεται η υπερεκπροσώπηση των πιο δημοφιλών subreddits

- **Χρήση μάσκας σε ονόματα συγκεκριμένων ανθρώπων και σε θρησκευτικούς όρους**

- **Εξισορρόπηση συναισθήματος**. Μειώθηκε η μεροληψία στο συναίσθημα με την αφαίρεση subreddits με χαμηλή εκπροσώπηση σε θετικό, αρνητικό, ουδέτερο ή διφορούμενο συναίσθημα

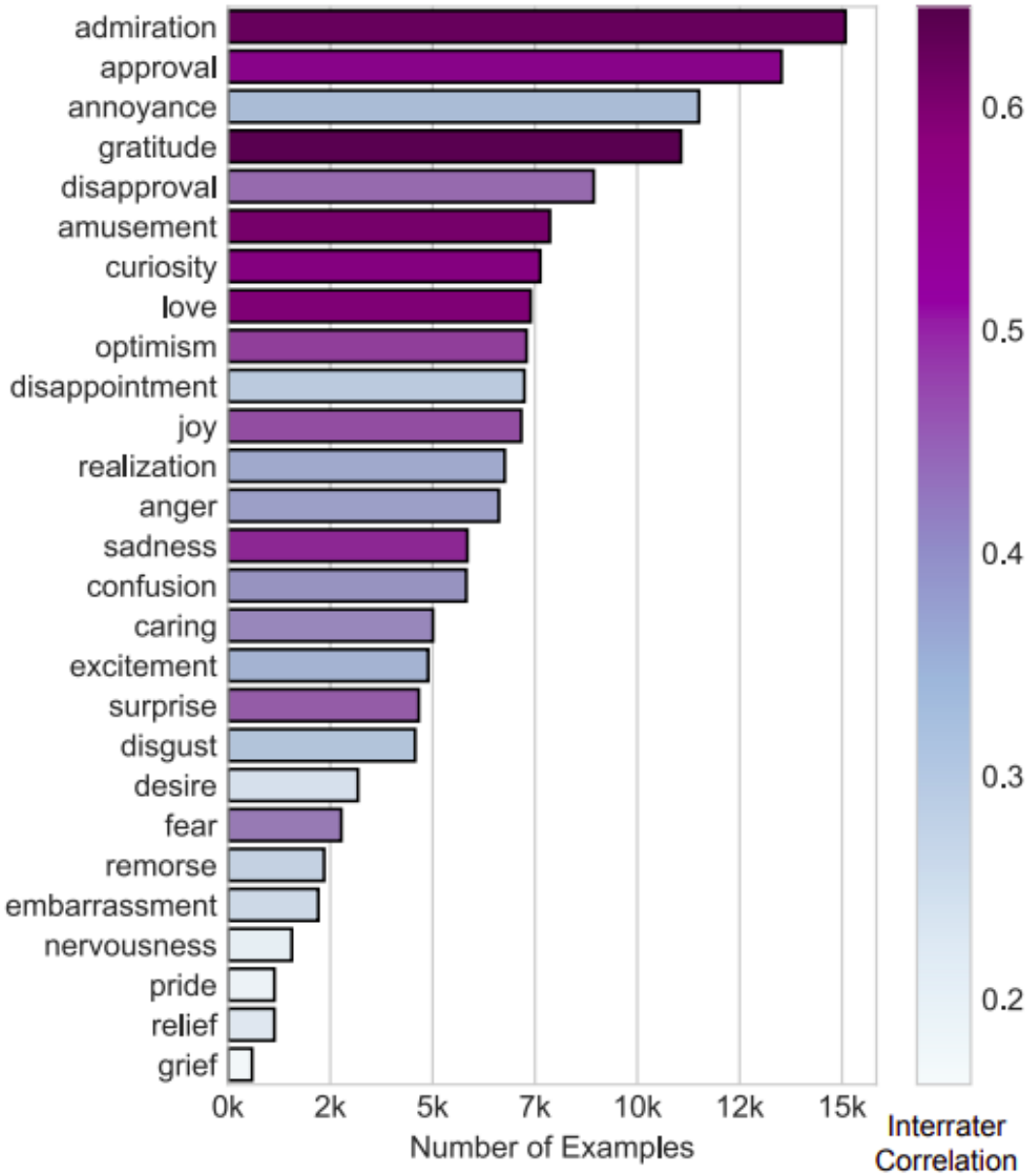
Αναφορικά με την ταξινομία, επιχειρήθηκε να ακολουθηθούν οι παρακάτω στόχοι:

- **Η παροχή της μεγαλύτερης δυνατής κάλυψης αναφορικά με τα συναισθήματα που εκφράζονται στα δεδομένα**
- **Η παροχή της μεγαλύτερης δυνατής κάλυψης αναφορικά με τα είδη της συναισθηματικής έκφρασης**. Για το σκοπό αυτό οι ερευνητές συμβουλεύτηκαν τη βιβλιογραφία της ψυχολογίας στα πεδία της έκφρασης συναισθήματος και της αναγνώρισης συναισθήματος
- **Ο περιορισμός της επικάλυψης ανάμεσα στα συναισθήματα καθώς και ο περιορισμός στον αριθμό των δυνατών συναισθημάτων**. Αποφεύγεται το να περιλαμβάνονται πολύ παρεμφερή συναισθήματα, αφού αυτό κάνει την επισήμανση πιο δύσκολη

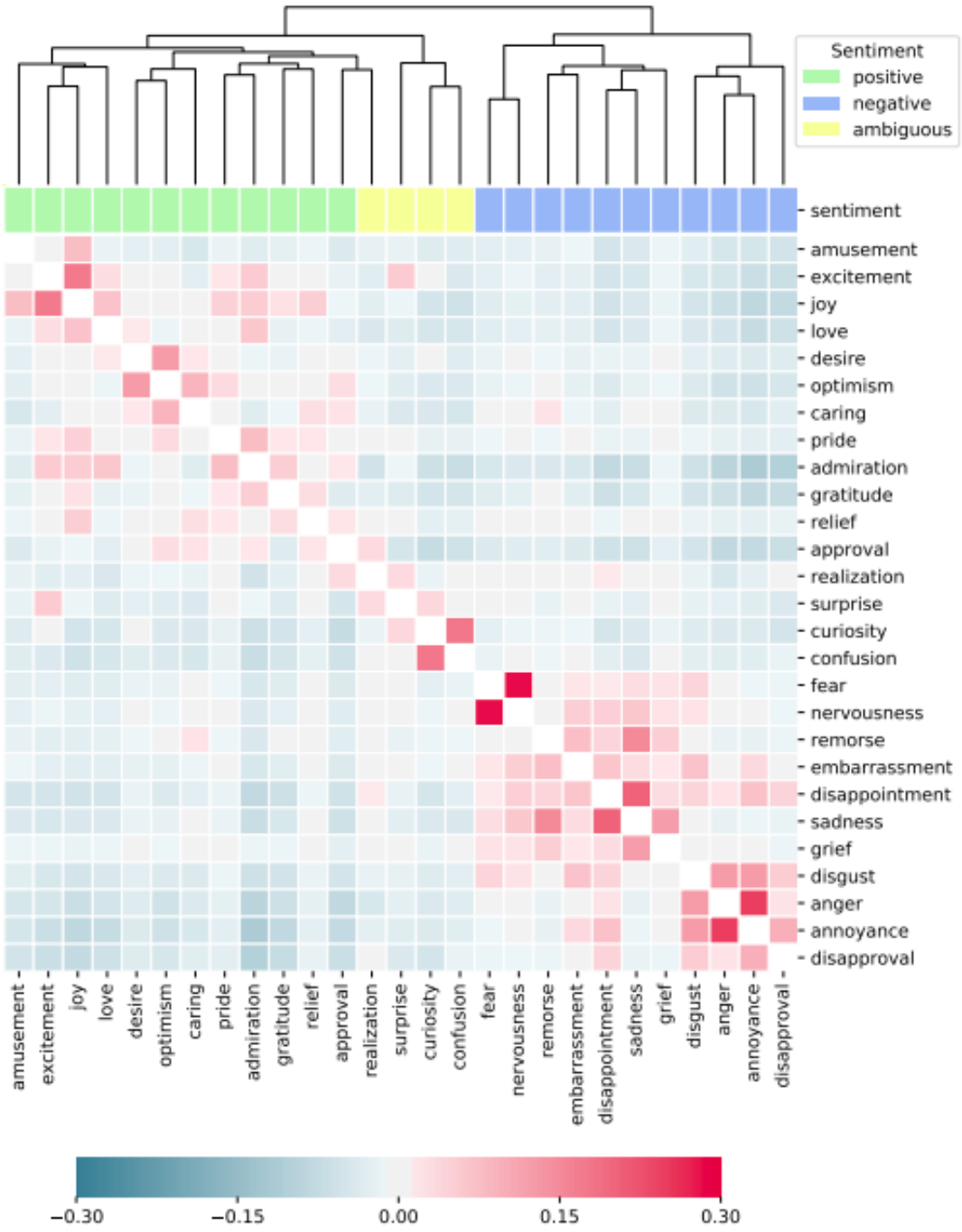
Όσο αφορά την επισήμανση των δεδομένων, σε κάθε παράδειγμα ανατίθενται τρεις αξιολογητές. Για εκείνα τα παραδείγματα στα οποία δεν συμφωνούν σε τουλάχιστον ένα label, ανατίθενται δύο επιπλέον αξιολογητές. Ζητήθηκε από αυτούς να αναγνωρίσουν το συναίσθημα που εκφράζεται από το συγγραφέα του κειμένου, αφού τους είχε δοθεί ένας προκαθορισμένος ορισμός για κάθε συναίσθημα μαζί με αντίστοιχα κείμενα-παραδείγματα. Οι αξιολογητές είναι ελεύθεροι να επιλέξουν περισσότερα του ενός συναισθήματα, αρκεί να είναι σίγουροι σε σημαντικό βαθμό. Σε περίπτωση που δεν μπορούν να καταλήξουν σε κάποιο συγκεκριμένο συναίσθημα θα επιλέγεται η κατηγορία "Ουδέτερο".

Σχετικά με την ανάλυση των δεδομένων, παρατηρείται ότι το 83 % των παραδειγμάτων έχει ένα μοναδικό label, με τουλάχιστον δύο αξιολογητές να συμφωνούν επί αυτού στο 94 % των περιπτώσεων. Το 26 % χαρακτηρίζεται ως "Ουδέτερο", με τα παραδείγματα αυτά να αφαιρούνται καθώς κρίνεται ότι δεν παρουσιάζουν σημασιολογικό ενδιαφέρον. Στο σχήμα φαίνεται η κατανομή των labels. Υπάρχει σημαντική διαφορά στις συχνότητες εμφάνισης των συναισθημάτων, κάτι που κρίνεται ως λογικό καθώς κάτι τέτοιο αντικατοπτρίζει τις διαφορετικές συχνότητες των συναισθημάτων στην ανθρώπινη έκφραση. Επιπρόσθετα, με τη συσχέτιση ανάμεσα στους αξιολογητές φαίνεται το κατά πόσο ήταν εύκολο να συμφωνήσουν οι ερευνητές στο εκάστοτε συναίσθημα, άρα στην ουσία το πόσο "ξεκάθαρο" είναι αυτό(Σχήμα 6.5). Τα συναισθήματα για τα οποία οι αξιολογητές αποφάνθηκαν με μεγαλύτερη ευκολία είναι η ευγνωμοσύνη, η εκτίμηση και η διασκέδαση, ενώ αυτά που παρουσίασαν τη μεγαλύτερη δυσκολία είναι το πένθος και η νευρικότητα.

Ακόμα, για να διερευνηθεί η σχέση ανάμεσα στα συναισθήματα δημιουργείται ένας heatmap με τις συσχετίσεις τους όπως φαίνεται στο Σχήμα 6.6. Φαίνεται ότι τα συναισθήματα παρεμφερούς έντασης, όπως η ενόχληση και ο θυμός παρουσιάζουν ισχυρά θετική συσχέτιση. Αντιθέτως, συναισθήματα με το ανάποδο αίσθημα έχουν αρνητική συσχέτιση.




Σχήμα 6.5: Η συχνότητα των εμφανισθέντων labels, μαζί με τη συσχέτιση ανά των αξιολογητών




Σχήμα 6.6: Οι συσχετίσεις ανάμεσα στις αξιολογήσεις των συναισθημάτων

# 7

# Πειράματα

## 7.1 Εισαγωγή

Στόχος της συγκεκριμένης εργασίας είναι η σύγκριση της απόδοσης διαφορετικών, "ελαφρύτερων" παραλλαγών του BERT. Πιο συγκεκριμένα, όπως έχει ήδη αναλυθεί και στις προηγούμενες ενότητες, η εκπαίδευση δύο σταδίων, η προ-εκπαίδευση δηλαδή των μοντέλων σε ένα μεγάλο σώμα δεδομένων, η κατασκευή embeddings και η εφαρμογή εν συνεχεία επιβλεπόμενης μάθησης σε μικρά σύνολα δεδομένων για εξειδικευμένες εργασίες, επέφερε θεαματικά αποτελέσματα στο πεδίο της Επεξεργασίας Φυσικής Γλώσσας. Ωστόσο, τα SOTA μοντέλα έχουν συνήθως εκατομμύρια παραμέτρους και η προ-εκπαίδευση τους απαιτεί πολύ μεγάλο υπολογιστικό κόστος. Αυτό μπορεί να καθιστά απαγορευτική τη χρήση τους σε περιβάλλοντα με περιορισμένη μνήμη ή με απαιτήσεις περιορισμένης καθυστέρησης εξυπηρέτησης(latency). Για να αντιμετωπιστεί αυτό το πρόβλημα έχουν ακολουθηθεί διάφορες προσεγγίσεις κυρίως προς την κατεύθυνση της συμπίεσης των μεγάλων μοντέλων, με χρήση μεταξύ άλλων και Απόσταξης Γνώσης όπως ήδη έχει περιγραφεί.

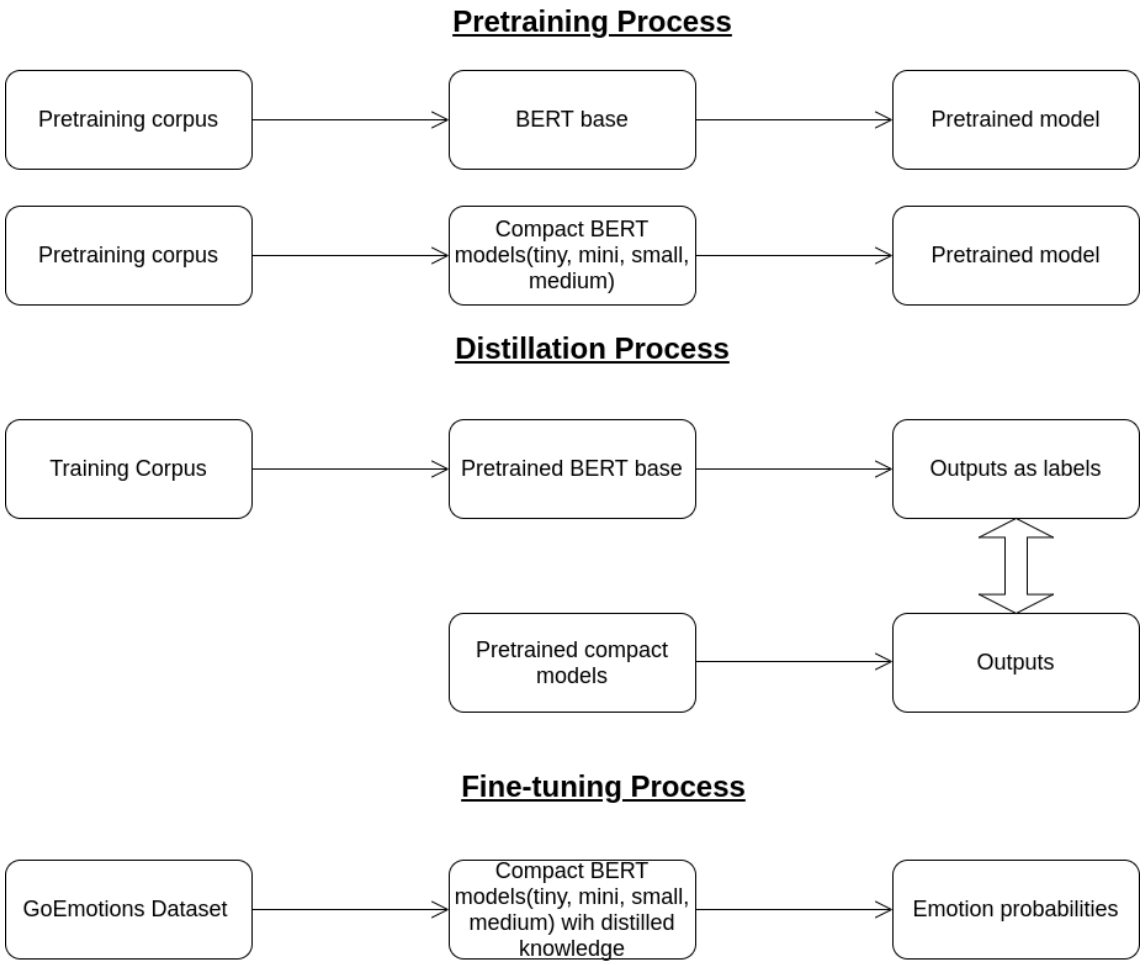
Με τη δημοσίευση [37] αντλούνται πέντε BERT μοντέλα, τέσσερα συμπιεσμένα και το μοντέλο βάσης του BERT. Τα μοντέλα αυτά έχουν ήδη προεκπαιδευτεί και εν συνεχεία πραγματοποιείται Απόσταξη Γνώσης από το μοντέλο βάσης στα συμπιεσμένα μοντέλα. Στο τέλος γίνεται fine-tuning και αξιολογούνται οι επιδόσεις του μοντέλου στο σετ δεδομένων GoEmotions, το οποίο έχει περιγραφεί αναλυτικά σε προηγούμενη ενότητα. Μετά τη σύγκριση των μοντέλων καταλήγουμε σε συμπεράσματα και προτάσεις για μελλοντική έρευνα.

## 7.2 Σχεδιασμός και Εκτέλεση του Πειράματος

Η διαδικασία που ακολουθήθηκε στο [37] είναι ίδια με αυτή που περιγράφηκε στην ενότητα 4.6. Η μόνη διαφορά είναι ότι τα μοντέλα αυτά έχουν ήδη προεκπαιδευτεί και εν συνεχεία πραγματοποιείται η Απόσταξη Γνώσης από το μοντέλο βάσης στα συμπιεσμένα μοντέλα. Το σχήμα της ενότητας 4.6 τροποποιείται λοιπόν για να συμπεριλάβει αυτό το επιπλέον στάδιο της προεκπαίδευσης των συμπιεσμένων μοντέλων και παρουσιάζεται στο σχήμα 7.1. Εν συνεχεία, γίνεται η εκπαίδευση με βάση τα labels που παρήγαγε το μοντέλο βάσης και τέλος γίνεται fine-tuning και αξιολογούνται οι επιδόσεις του μοντέλου στο σετ δεδομένων GoEmotions, το οποίο έχει περιγραφεί αναλυτικά σε προηγούμενη ενότητα.

Το πείραμα στη GPU GeForce RTX 2080 της NVIDIA και στη CPU Intel® Xeon® Silver 4215R Processor. Η υλοποίηση έγινε με χρήση της βιβλιοθήκης Pytorch και βασίστηκε στο repository [38].

Για την εκτέλεση των πειραμάτων έγινε χρήση των βασικών υπερπαραμέτρων που φαίνον-




Σχήμα 7.1: Η νέα τροποποιημένη διαδικασία εκπαίδευσης

| Parameter | |
|---|---|
| Learning rate | 5e-5 |
| Warmup proportion | 0.1 |
| Epochs | 10 |

Σχήμα 7.2: Οι υπερπαράμετροι που χρησιμοποιήθηκαν για το αρχικό πείραμα

ται στο Σχήμα 7.2, που βασίστηκαν στο αρχίκό cross-validation του [38].

Εν συνεχεία εκτελέστηκαν διαφορετικά πειράματα με τα μοντέλα να τροποποιούνται ως προς το batch size και το maximum sequence length κάθε φορά με σκοπό την επίτευξη όσο το δυνατών καλύτερων αποτελεσμάτων.

## 7.3 Τα μοντέλα που θα συγκριθούν

Τα μοντέλα-παραλλαγές του BERT που θα χρησιμοποιηθούν για τα πειράματα της συγκεκριμένης εργασίας είναι 5 και περιγράφονται παρακάτω, σε αύξουσα σειρά πολυπλοκότητας:

- **BERT-tiny**: Είναι το μικρότερο μοντέλο που θα χρησιμοποιηθεί στα πειράματα. Αποτελείται από περίπου 4.4 εκατομμύρια παραμέτρους. Επιπρόσθετα έχει 2 transformer block επίπεδα και κρυφό μέγεθος αναπαράστασης 128.
- **BERT-mini**: Το αμέσως μεγαλύτερο μοντέλο αποτελείται από 11.3 εκατομμύρια παραμέτρους και έχει 4 transformer block επίπεδα. Το κρυφό μέγεθος αναπαράστασης της εκάστοτε θέσης είναι σε αυτή την περίπτωση 256
- **BERT-small**: Το μοντέλο αυτό αποτελείται από 29.1 εκατομμύρια παραμέτρους με επίσης 4 transformer block επίπεδα. Διαφοροποιείται από το BERT-mini στο κρυφό μέγεθος αναπαράστασης που στην περίπτωση αυτή είναι 512
- **BERT-medium**: Το μοντέλο αποτελείται από 41.7 εκατομμύρια παραμέτρους. Έχει και αυτό κρυφό μέγεθος αναπαράστασης 512, ενώ έχει 8 transformer block επίπεδα
- **BERT-base**: Είναι το μοντέλο βάσης του BERT που περιγράφεται και στην αρχική δημοσίευσή του [15]. Αποτελείται από 110.1 εκατομμύρια παραμέτρους και έχει 12 transformer block επίπεδα και κρυφό μέγεθος αναπαράστασης 768

Στον ακόλουθο πίνακα παρουσιάζονται συγκεντρωτικά τα 24 μοντέλα της αρχικής δημοσίευσης. Τα 5 μοντέλα που χρησιμοποιήθηκαν στα πειράματα είναι τα υπογραμμισμένα.

| | **H=128** | **H=256** | **H=512** | **H=768** |
|---|---|---|---|---|
| **L=2** | <u>4.4</u> | 9.7 | 22.8 | 39.2 |
| **L=4** | 4.8 | <u>11.3</u> | <u>29.1</u> | 53.4 |
| **L=6** | 5.2 | 12.8 | 35.4 | 67.5 |
| **L=8** | 5.6 | 14.4 | <u>41.7</u> | 81.7 |
| **L=10** | 6.0 | 16.0 | 48.0 | 95.9 |
| **L=12** | 6.4 | 17.6 | 54.3 | <u>110.1</u> |

Επιπρόσθετα, όπως αναφέρθηκε και παραπάνω έτρεξαν διαφορετικές εκτελέσεις με τροποποιημένες παραμέτρους. Ουσιαστικά δημιουργούνται 2 οικογένειες πειραμάτων, μια με μεταβλητό batch size και μια με μεταβλητό maximum sequence length . Περιγράφονται παρακάτω:

- Χρήση των αρχικών υπερπαραμέτρων του σχήματος με batch size 16 και maximum sequence length 7.2.
- Τροποποίηση του αρχικού batch size σε 32
- Τροποποίηση του αρχικού batch size σε 64
- Τροποποίηση του maximum sequence length σε 20
- Τροποποίηση του αρχικού maximum sequence length σε 100

## 7.4 Αποτελέσματα

Μετά την εκτέλεση των πειραμάτων προκύπτει οι ακόλουθοι συγκεντρωτικοί πίνακες των αποτελεσμάτων. Παρατηρείται, όπως αναμενόταν ότι σε όλα τα πειράματα, καθώς αυξάνει η πολυπλοκότητα και η χωρητικότητα των μοντέλων, βελτιώνονται και οι αντίστοιχες επιδόσεις τους. Επειδή το σετ δεδομένων GoEmotions στο οποίο πραγματοποιήθηκε το fine-tuning είναι μη ισορροπημένο, δηλαδή κάποια συναισθήματα απαντώνται συχνότερα, δίνεται μεγαλύτερη έμφαση στις macro μετρικές σε σχέση με τις micro. Αυτό συμβαίνει καθώς οι micro μετρικές(precision, recall, F1) επηρεάζονται σε μεγάλο βαθμό από τα συναισθήματα που κυριαρχούν στο σετ δεδομένων. Μαζί αυξάνεται και το Accuracy, ενώ παράλληλα το loss μειώνεται σταδιακά έστω και ανεπαίσθητα.

| BS=16,Seq=50 | BERT-tiny | BERT-mini | BERT-small | BERT-medium | BERT-base |
|---|---|---|---|---|---|
| accuracy | 0.365 | 0.400 | 0.398 | 0.397 | 0.406 |
| loss | 0.114 | 0.105 | 0.116 | 0.115 | 0.145 |
| macro_f1 | 0.195 | 0.343 | 0.426 | 0.421 | 0.510 |
| macro_precision | 0.210 | 0.383 | 0.445 | 0.438 | 0.494 |
| macro_recall | 0.195 | 0.340 | 0.428 | 0.427 | 0.544 |
| micro_f1 | 0.490 | 0.534 | 0.532 | 0.525 | 0.571 |
| micro_precision | 0.589 | 0.564 | 0.515 | 0.510 | 0.529 |
| micro_recall | 0.419 | 0.508 | 0.551 | 0.542 | 0.620 |
| weighted_f1 | 0.383 | 0.491 | 0.526 | 0.521 | 0.573 |
| weighted_precision | 0.368 | 0.499 | 0.508 | 0.507 | 0.534 |
| weighted_recall | 0.419 | 0.508 | 0.551 | 0.542 | 0.620 |

| BS=32,Seq=50 | BERT-tiny | BERT-mini | BERT-small | BERT-medium | BERT-base |
|---|---|---|---|---|---|
| accuracy | 0.279 | 0.394 | 0.406 | 0.399 | 0.423 |
| loss | 0.131 | 0.106 | 0.108 | 0.109 | 0.134 |
| macro_f1 | 0.054 | 0.267 | 0.416 | 0.420 | 0.507 |
| macro_precision | 0.052 | 0.365 | 0.445 | 0.447 | 0.515 |
| macro_recall | 0.059 | 0.271 | 0.418 | 0.423 | 0.530 |
| micro_f1 | 0.362 | 0.518 | 0.537 | 0.533 | 0.577 |
| micro_precision | 0.547 | 0.593 | 0.530 | 0.525 | 0.542 |
| micro_recall | 0.270 | 0.459 | 0.545 | 0.541 | 0.617 |
| weighted_f1 | 0.223 | 0.437 | 0.529 | 0.526 | 0.579 |
| weighted_precision | 0.195 | 0.516 | 0.523 | 0.519 | 0.547 |
| weighted_recall | 0.270 | 0.459 | 0.545 | 0.541 | 0.617 |

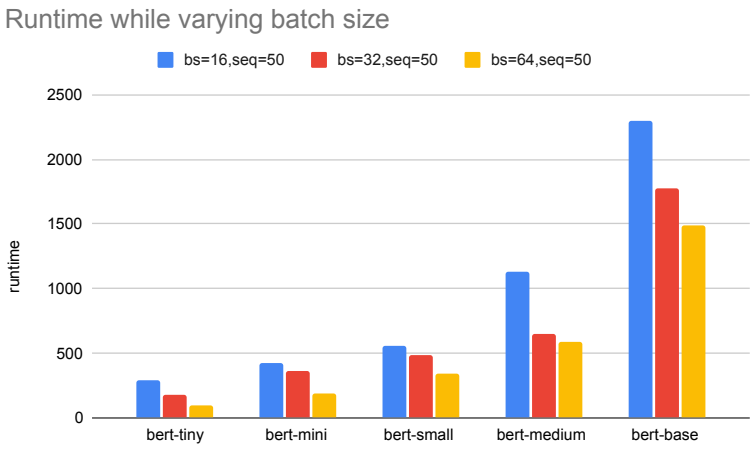
| BS=64,Seq=50 | BERT-tiny | BERT-mini | BERT-small | BERT-medium | BERT-base |
|---|---|---|---|---|---|
| accuracy | 0.296 | 0.365 | 0.397 | 0.403 | 0.422 |
| loss | 0.148 | 0.115 | 0.105 | 0.105 | 0.118 |
| macro_f1 | 0.018 | 0.201 | 0.388 | 0.390 | 0.491 |
| macro_precision | 0.012 | 0.209 | 0.422 | 0.422 | 0.495 |
| macro_recall | 0.036 | 0.200 | 0.385 | 0.393 | 0.509 |
| micro_f1 | 0.304 | 0.490 | 0.532 | 0.531 | 0.569 |
| micro_precision | 0.329 | 0.618 | 0.542 | 0.537 | 0.541 |
| micro_recall | 0.282 | 0.406 | 0.521 | 0.525 | 0.599 |
| weighted_f1 | 0.140 | 0.388 | 0.516 | 0.518 | 0.570 |
| weighted_precision | 0.093 | 0.379 | 0.531 | 0.528 | 0.547 |
| weighted_recall | 0.282 | 0.406 | 0.521 | 0.525 | 0.599 |

| BS=32,Seq=100 | BERT-tiny | BERT-mini | BERT-small | BERT-medium | BERT-base |
|---|---|---|---|---|---|
| accuracy | 0.278 | 0.388 | 0.398 | 0.403 | 0.414 |
| loss | 0.133 | 0.107 | 0.109 | 0.111 | 0.135 |
| macro_f1 | 0.054 | 0.263 | 0.393 | 0.407 | 0.498 |
| macro_precision | 0.052 | 0.306 | 0.406 | 0.440 | 0.469 |
| macro_recall | 0.058 | 0.266 | 0.400 | 0.413 | 0.534 |
| micro_f1 | 0.361 | 0.516 | 0.530 | 0.525 | 0.570 |
| micro_precision | 0.546 | 0.601 | 0.523 | 0.519 | 0.534 |
| micro_recall | 0.269 | 0.452 | 0.537 | 0.531 | 0.612 |
| weighted_f1 | 0.223 | 0.430 | 0.520 | 0.518 | 0.571 |
| weighted_precision | 0.195 | 0.446 | 0.510 | 0.513 | 0.537 |
| weighted_recall | 0.269 | 0.452 | 0.537 | 0.531 | 0.612 |

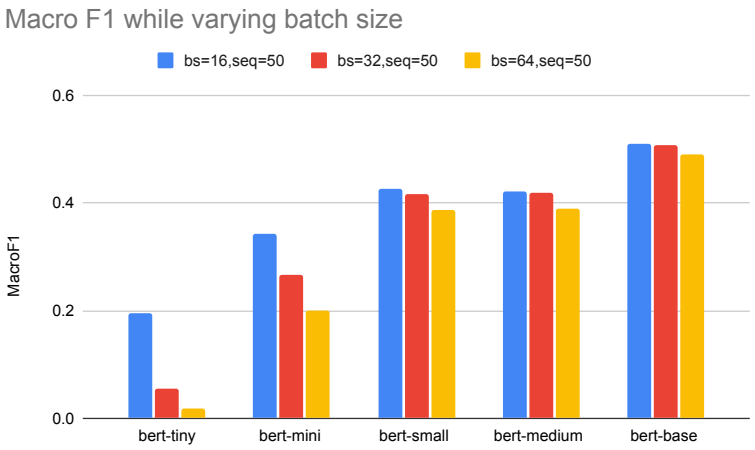
| BS=32,Seq=20 | BERT-tiny | BERT-mini | BERT-small | BERT-medium | BERT-base |
|---|---|---|---|---|---|
| accuracy | 0.267 | 0.371 | 0.378 | 0.373 | 0.392 |
| loss | 0.135 | 0.113 | 0.117 | 0.119 | 0.148 |
| macro_f1 | 0.052 | 0.249 | 0.366 | 0.381 | 0.448 |
| macro_precision | 0.052 | 0.288 | 0.382 | 0.402 | 0.444 |
| macro_recall | 0.055 | 0.243 | 0.362 | 0.375 | 0.461 |
| micro_f1 | 0.346 | 0.489 | 0.500 | 0.497 | 0.531 |
| micro_precision | 0.528 | 0.582 | 0.505 | 0.498 | 0.504 |
| micro_recall | 0.258 | 0.421 | 0.496 | 0.496 | 0.562 |
| weighted_f1 | 0.215 | 0.409 | 0.490 | 0.487 | 0.532 |
| weighted_precision | 0.190 | 0.427 | 0.491 | 0.484 | 0.509 |
| weighted_recall | 0.258 | 0.421 | 0.496 | 0.496 | 0.562 |

Στο Σχήμα 7.6 παρατηρούμε ότι ο χρόνος εκτέλεσης αυξάνεται σχεδόν εκθετικά, με την αύξηση της πολυπλοκότητας των μοντέλων. Αυτό στην πράξη σημαίνει ότι με κάθε περαιτέρω αύξηση της πολυπλοκότητας των μοντέλων, αυξάνεται όλο και περισσότερο ο χρόνος εκτέλεσης, συνεπώς συμπεραίνεται ότι και οι ανάγκες συμπίεσης καθίστανται όλο και πιο επιτακτικές. Επιπρόσθετα, παρατηρείται ότι οι Micro-F1 επιδόσεις των μοντέλων(Σχήμα 7.8) είναι παρεμφερείς παρά την αύξηση της πολυπλοκότητας. Οι Macro-F1 επιδόσεις(Σχήμα 7.7) ακολουθούν μια εμφανέστερη αυξητική τάση, με τις επιδόσεις και στις δύο μετρικές να μην παρουσιάζουν ωστόσο μεγάλη διασπορά.

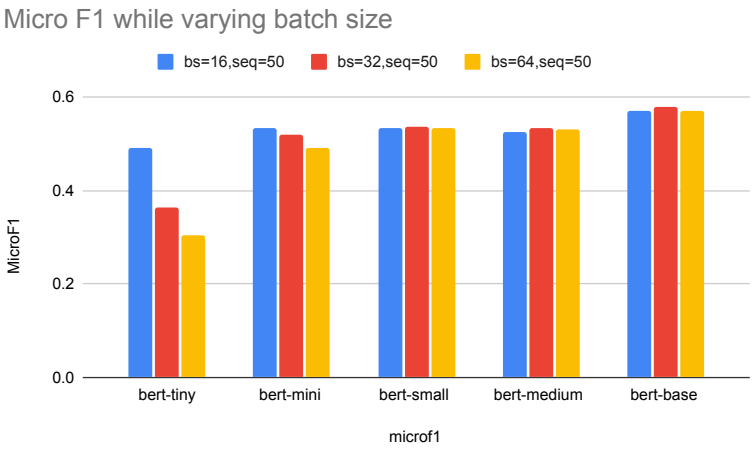
Αναφορικά με τις τροποποιημένες παραμέτρους του μοντέλου παρατηρείται ότι οι καλύτερες επιδόσεις επιτυγχάνονται με batch size 16 και με maximum sequence length 50.



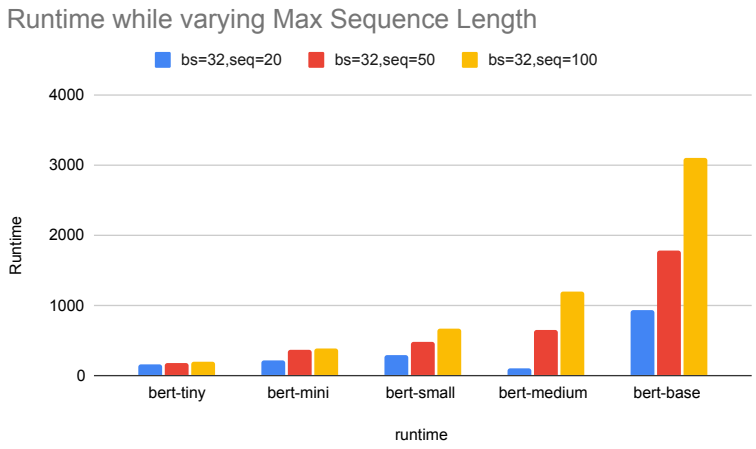

Σχήμα 7.3: Οι χρόνοι για την πραγματοποίηση των προβλέψεων.



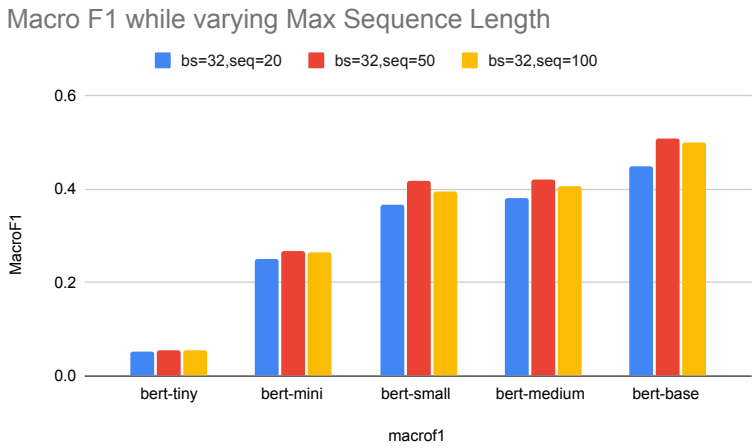

Σχήμα 7.4: Το Macro-F1 score των μοντέλων.



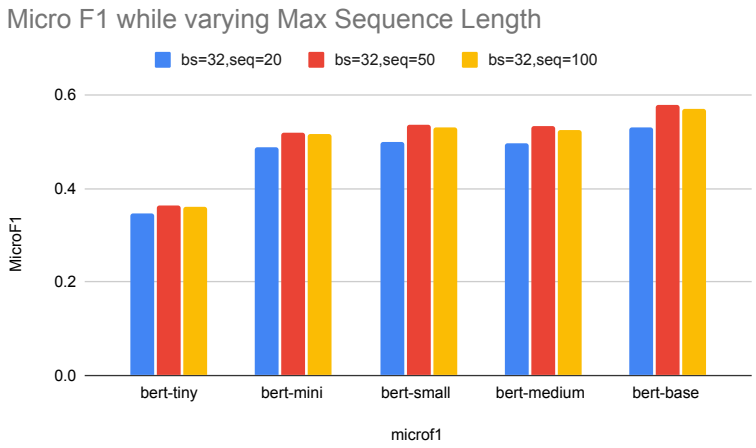

Σχήμα 7.5: Το Micro-F1 score των μοντέλων.




Σχήμα 7.6: Οι χρόνοι για την πραγματοποίηση των προβλέψεων.




Σχήμα 7.7: Το Macro-F1 score των μοντέλων.




Σχήμα 7.8: Το Micro-F1 score των μοντέλων.

## 7.5 Συμπεράσματα

Μετά από ανάλυση των αποτελεσμάτων παρατηρείται, όπως ήταν και διαισθητικά αναμενόμενο, ότι τα μεγαλύτερα, πιο σύνθετα μοντέλα παρουσιάζουν και τις καλύτερες επιδόσεις. Ωστόσο, οι διαφορές στις επιδόσεις αν και αξιοσημείωτες δεν καθιστούν απαγορευτική τη χρήση των μικρότερων μοντέλων. Αυτό στην πράξη σημαίνει ότι σε περιβάλλοντα με περιορισμούς είτε υπολογιστικών πόρων είτε χρονικούς, η χρήση ενός μικρότερου συνεπτυγμένου μοντέλου αποτελεί μια αρκετά υποσχόμενη εναλλακτική . Ουσιαστικά υπάρχει ένα trade-off ανάμεσα στις επιδόσεις του μοντέλου και στους περιορισμούς που υπόκειται λόγω υπολογιστικού περιβάλλοντος. Σε μια φορητή συσκευή για παράδειγμα, θα είχε νόημα να θυσιαστεί λίγη από την αξιοπιστία των αποτελεσμάτων για να καταστεί εφικτή η χρήση του BERT μοντέλου. Σε κάθε περίπτωση, καταδεικνύεται ότι η χρήση μικρότερων-συμπιεσμένων μοντέλων με προεκπαίδευση αυτών, αποτελεί μια στρατηγική ικανή να μετριάσει ή να αντιμετωπίσει το πρόβλημα της εκπαίδευσης και χρήσης του απαιτητικού BERT και παρεμφερών SOTA μοντέλων σε περιβάλλοντα με περιορισμούς υπολογιστικών πόρων. Μελλοντικά θα μπορούσε να διερευνηθεί η ανάπτυξη νέων τεχνικών για την αποτελεσματικότερη συμπίεση των μοντέλων ως προς το μέγεθος και ως προς τις επιδόσεις τους. Στόχος είναι η ανάπτυξη ακόμα απλούστερων μοντέλων και η βελτίωση των επιδόσεων τους ώστε να προσεγγίζουν, στο μέτρο του δυνατού τις επιδόσεις των αντίστοιχων SOTA μοντέλων βάσης.

# Bibliography


[1] Albert Einstein. Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]. *Annalen der Physik*, 322(10):891–921, 1905.

[2] Ferdinand de saussure. *Britannica, https://www.britannica.com/biography/Ferdinand-de-Saussure.*

[3] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(October):433–60, 1950.

[4] Noam Chomsky. *Syntactic Structures.* Mouton and Co., The Hague, 1957.

[5] Natural language processing. *Wikipedia, https://en.wikipedia.org/wiki/Natural_language_processing.*

[6] Word embedding in nlp: One-hot encoding and skip-gram neural network. *towardsdatascience, https://towardsdatascience.com/word-embedding-in-nlp-one-hot-encoding-and-skip-gram-neural-network-81b424da58f2.*

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[11] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context, 2019.

[12] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification, 2018.

[13] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[14] OpenAI. Improving language understanding by generative pre-training. 2018.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[16] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

[17] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[18] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

[19] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

[20] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[21] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. Structbert: Incorporating language structures into pre-training for deep language understanding, 2019.

[22] Yann Lecun, John Denker, and Sara Solla. Optimal brain damage. volume 2, pages 598–605, 01 1989.

[23] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets, 2017.

[24] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming, 2017.

[25] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2016.

[26] Raphael Tang, Ashutosh Adhikari, and Jimmy Lin. Flops as a direct optimization objective for learning sparse neural networks, 2018.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[28] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks, 2019.

[29] Bedir Tekinerdogan Alexander Ligthart, Cagatay Catal. Systematic reviews in sentiment analysis: a tertiary study. 2021.

[30] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference, 2015.

[31] Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference, 2018.

[32] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[33] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems, 2020.

[34] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[35] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad, 2018.

[36] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions, 2020.

[37] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019.

[38] Goemotions-pytorch. *Github, https://github.com/monologg/GoEmotions-pytorch.*