Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Τεχνικές ανάλυσης συναισθήματος κειμένου στα Ελληνικά με χρήση Δικτύων Μετασχηματιστών

Διπλωματικη Εργασια

της

ΠΑΥΛΙΝΑΣ Β. ΧΑΤΖΗΑΝΤΩΝΙΟΥ

Επιβλέπων: Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής ΕΜΠ

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Ε.ΔΙ.Π. ΕΜΠ

Εργαστηριο Τεχνητης Νοημοσυνης και Συστηματων Μαθησης

Αθήνα, Μάρτιος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης

Τεχνικές ανάλυσης συναισθήματος κειμένου στα Ελληνικά με χρήση Δικτύων Μετασχηματιστών

Διπλωματικη Εργασια

της

ΠΑΥΛΙΝΑΣ Β. ΧΑΤΖΗΑΝΤΩΝΙΟΥ

Επιβλέπων:	Αθανάσιος Βουλόδημος
	Επίκουρος Καθηγητής Ε.Μ.Π.
Συνεπιβλέπων :	Γεώργιος Αλεξανδρίδης
	Εργαστηριακό Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Μαρτίου 2023.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αθανάσιος Βουλόδημος	Γεώργιος Στάμου	Στέφανος Κόλλιας
Επίκουρος Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2023

(Υπογραφή)

.....

Παυλινα Β. Χατζηαντωνιου

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω θερμά τον κ. Ανδρέα-Γεώργιο Σταφυλοπάτη, Καθηγητή Ε.Μ.Π, και τον κ. Αθανάσιο Βουλόδημο, Επίκουρο Καθηγητή Ε.Μ.Π, για την επίβλεψη αυτής της διπλωματικής εργασίας, καθώς και τους κ.κ. Γεώργιο Στάμου και Στέφανο Κόλλια, Καθηγητές Ε.Μ.Π., για την τιμή που μου έκαναν να συμμετέχουν στην εξεταστική επιτροπή της εργασίας. Επίσης, ευχαριστώ ιδιαίτερα τον κ. Γεώργιο Αλεξανδρίδη, Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π., για την καθοδήγησή του και τη συνεργασία που είχαμε καθώς και τον κ. Παναγιώτη Τσαντίλα, διευθύνοντα σύμβουλο της εταιρίας ΠΑΛΟ ΨΗΦΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Ε.Π.Ε., για την παραχώρηση των δεδομένων που χρησιμοποιήθηκαν στο πειραματικό μέρος της εργασίας. Τέλος, ένα μεγάλο ευχαριστώ στους γονείς μου, στον αδερφό μου και στους φίλους μου, για τη βοήθεια και τη στήριξη που μου προσέφεραν όλα αυτά τα χρόνια.

Περίληψη

Η ανάλυση συναισθήματος αποτελεί ένα διαδεδομένο πεδίο έρευνας της επεξεργασίας φυσικής γλώσσας, το οποίο πραγματεύεται την αναγνώριση του συναισθήματος που αποτυπώνεται σε ένα κείμενο. Οι εφαρμογές ενός αποτελεσματικού μοντέλου ανάλυσης συναισθήματος, μπορούν να αποτελέσουν ένα χρήσιμο εργαλείο για την καλύτερη κατανόηση των αναγκών, απόψεων, συμπεριφορών και προτιμήσεων του ευρύτερου κοινού. Αντικείμενο της παρούσας διπλωματικής εργασίας, αποτελεί η υλοποίηση, εκπαίδευση και αξιολόγηση μοντέλων μηχανικής μάθησης, ικανά να κατατάσσουν το περιεχόμενο ενός κειμένου ως θετικό, αρνητικό ή ουδέτερο. Για τον σκοπό αυτό, τα μοντέλα που αναπτύχθηκαν βασίστηκαν στην αρχιτεκτονική των δικτύων μετασχηματιστών, ενός υπερσύγχρονου γλωσσικού μοντέλου βαθιάς μάθησης, εξειδικευμένο στην μοντελοποίηση των σημασιολογικών εννοιών των φυσικών γλωσσών. Τόσο η προ-εκπαίδευση των μοντέλων, όσο και η εφαρμογή της τεχνικής της ανάλυσης συναισθήματος σε αυτά, αφορά αποσπάσματα κειμένου στην ελληνική γλώσσα από μέσα κοινωνικής δικτύωσης. Για τη βελτιστοποίηση της αποδοτικότητας των μοντέλων, πραγματοποιήθηκαν διάφοροι πειραματισμοί όσον αφορά την αρχιτεκτονική τους, την επιλογή των υπερ-παραμέτρων τους, καθώς και την προεπεξεργασία των δεδομένων πάνω στα οποία γίνεται η εκπαίδευσή τους. Τα τελικά μοντέλα της μελέτης είναι διαθέσιμα στην κοινότητα Τεχνητής Νοημοσύνης του HuggingFace.

Λέξεις Κλειδιά

Ανάλυση Συναισθήματος, Εξόρυξη Γνώμης, Βαθιά Μηχανική Μάθηση, Μετασχηματιστές, Γλωσσικό Μοντέλο, Επεξεργασία Φυσικής Γλώσσας, Ελληνικά Μέσα Κοινωνικής Δικτύωσης

Abstract

Sentiment analysis is a widely researched field in natural language processing, which deals with identifying the emotion expressed in a given text. The applications of an effective sentiment analysis model are considered to be a useful tool for better understanding the needs, opinions, attitudes and preferences of the general public. The subject of this thesis is to implement, train and evaluate machine learning models that can correctly classify a piece of text as positive, negative or neutral. The models developed for this objective are based on the Transformer architecture, a state-of-the-art deep learning language model, specialized in capturing accurately the semantic concepts of various natural languages. The corpus used in both the pre-training of the models and their fine-tuning for sentiment analysis, is gathered from social media in the greek language. For the optimization of the models' efficiency, a series of experiments were conducted, regarding their architecture, the selection of their hyperparameters, as well as the text pre-processing techniques on the training data. The resulting models are available to the AI community of HuggingFace.

Keywords

Sentiment Analysis, Opinion Mining, Deep Learning, Transformers, Language Model, Natural Language Processing, Greek Social Media

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος (sentiment analysis) είναι ερευνητικό πεδίο της επεξεργασίας φυσικής γλώσσας (natural language preprocessing) και αποτελεί μία υπολογιστική μελέτη των απόψεων, των συναισθημάτων, των προτιμήσεων και των συμπεριφορών των ανθρώπων όσον αφορά διάφορες οντότητες, όπως για παράδειγμα προϊόντα, υπηρεσίες, οργανώσεις, φυσικά πρόσωπα, συζητήσεις, γεγονότα κ.ο.κ. Η διεργασία αυτή, μπορεί να κατατάξει τα αποτελέσματά τις με πολλούς και διαφορετικούς τρόπους, όπως για παράδειγμα με την αντιστοιχία τους σε μια απλή δυαδική κατανομή (αρνητικό - θετικό), την επέκταση την ταξινόμησης σε 3 κλάσεις αρνητικών, θετικών και ουδέτερων συναισθημάτων, ακόμα και την πολυεπίπεδη διαβάθμισή τους σε ένα ευρύ φάσμα συναισθημάτων όπως «χαρά», «ενθουσιασμός», «θυμός», «λύπη», «απογοήτευση» κ.α. Η ταχεία ανάπτυξη του πεδίου, συμπίπτει με την ανάπτυξη των μέσων κοινωνικής δικτύωσης (online social media), μέσω του οποίου κάθε άτομο έχει άμεση πρόσβαση σε αναρίθμητες πηγές πληροφορίας, οι οποίες αφορούν κάθε πιθανό φάσμα της ζωής μας. Τα τελευταία χρόνια, η ανάλυση συναισθημάτων έχει εξελιχθεί σε ένα ευρύτατο πεδίο έρευνας και μπορεί να «βρει» εφαρμογή σε οποιοδήποτε πλευρά της καθημερινότητάς μας. Συγκεκριμένα, η ψυχαγωγία, η ενημέρωση, οι καταναλωτικές συμπεριφορές, καθώς και η διαμόρφωση και ανάπτυξη επιχειρηματικών και πολιτικών στρατηγικών και κατευθύνσεων, είναι ορισμένοι βασικοί τομείς που επηρεάζονται από τις διάφορες εφαρμογές της ανάλυσης συναισθήματος σε ατομικό, αλλά και συλλογικό επίπεδο. Αυτό το φαινόμενο παρατηρείται καθώς, οι αντιλήψεις και επιλογές μας διαμορφώνονται σε σημαντικό βαθμό και από την κοινή γνώμη. Επομένως, η λήψη μιας, ακόμα και ασήμαντης φαινομενικά, απόφασης έχει, είτε συνειδητά, είτε υποσυνείδητα, επηρεαστεί από αυτή.

Στη σημερινή εποχή, τα μέσα κοινωνικής δικτύωσης αποτελούν το κυρίαρχο μέσο επικοινωνίας, αλληλεπίδρασης και ενημέρωσης. Οι διάφορες πλατφόρμες συγκεντρώνουν καθημερινά ένα τεράστιο όγκο πληροφορίας, η οποία μπορεί να αντιπροσωπεύσει κάθε πλευρά της κοινής γνώμης πάνω σε οποιοδήποτε θέμα συζήτησης. Μέσω τεχνικών ανάλυσης συναισθήματος και με την κατάλληλη διαχείριση του μεγάλου αυτού όγκου δεδομένων, μπορεί να καταστεί εφικτή η όσο το δυνατόν βαθύτερη κατανόηση των απόψεων, συμπεριφορών και προτιμήσεων

του κοινού και δημιουργείται η ευκαιρία για μετέπειτα αξιοποίησή της σε ένα ευρύτατο φάσμα εφαρμογών.

1.2 Πρακτικές Εφαρμογές

Αρχικά, σε ατομικό επίπεδο, η πληροφορία που αντλείται από την ανάλυση συναισθήματος μπορεί να λειτουργήσει συμβουλευτικά. Μέσω της παρακολούθησης των τάσεων και των απόψεων που διαδίδονται στα μέσα κοινωνικής δικτύωσης, ο κάθε χρήστης λαμβάνει σημαντική πληροφορία που μπορεί να κατευθύνει τις επιλογές του ως προς τα αγαθά και τις υπηρεσίες που τον ενδιαφέρουν, ή ακόμα και να αποτραπεί από αυτά, βασιζόμενος στις εμπειρίες που έχουν δημοσιευτεί. Κατ΄ αυτόν τον τρόπο, δίνεται η ευκαιρία στο κάθε άτομο να επιλέξει την επιχείρηση ή τον επαγγελματία ή το προϊόν στο οποίο θα επενδύσει και θα εμπιστευτεί, πραγματοποιώντας τη δική του έρευνα, βάσει των δικών του προσωπικών κριτηρίων. Οι επιλογές αυτές βέβαια, πολλές φορές μπορεί να έχουν μικρή σημασία αν εστιάσουμε το αντίκτυπο της ενέργειας αυτής σε μια μονάδα, ωστόσο, σε συλλογικό επίπεδο, το δίκτυο που αναπτύσσεται όταν κάθε ένας από τους πολλούς χρήστες του Διαδικτύου εκφράζει με θετικό ή αρνητικό τρόπο την άποψή του για ένα θέμα, αυξάνεται εκθετικά. Η ανάλυση λοιπόν των συναισθημάτων αυτών που διατυπώνονται στα σχόλια, τελικά δημιουργούν μια ιδιαίτερα αντιπροσωπευτική εικόνα της κοινής γνώμης, η οποία σε μεγαλύτερη κλίμακα, μπορεί να χρησιμοποιηθεί ως κατευθυντήρια γραμμή για την λήψη σημαντικότατων αποφάσεων που αφορούν οργανισμούς και επιχειρήσεις.

Ιδιαίτερα διαδεδομένες εφαρμογές της ανάλυσης συναισθήματος στα μέσα κοινωνικής δικτύωσης, αφορούν την παρακολούθηση των συζητήσεων που αναπτύσσονται σε αυτά από εταιρείες και οργανισμούς, γύρω από τα προϊόντα ή τις υπηρεσίες που παρέχουν. Εκμεταλλευόμενοι τους αλγόριθμους αναγνώρισης συναισθήματος, έχουν τη δυνατότητα να διαμορφώσουν μια καθαρή και αντιπροσωπευτική εικόνα ως προς την απήχησή τους και να λαμβάνουν τις κατάλληλες αποφάσεις και μέτρα, προκειμένου να εξελίξουν τη δράση τους, να προσαρμόσουν κατάλληλα τις καμπάνιες προώθησης προϊόντων, υπηρεσιών και φυσικών προσώπων. Επιπλέον, οι μηχανισμοί της τεχνικής αυτής μπορούν να επιφέρουν την εποικοδομητική αναθεώρηση ορισμένων παραμέτρων, συντελώντας στην διόρθωση και βελτίωση των όποιων ανεπιθύμητων χαρακτηριστικών που υπογραμμίζονται στις πλατφόρμες αυτές. Παράλληλα, με την άμεση και γρήγορη επεξεργασία των συνεχώς μεταβαλλόμενων δεδομένων, δίνεται η δυνατότητα να εντοπιστούν όσο το δυνατόν συντομότερα οι αστοχίες και να εξασφαλιστεί η εποικοδομητική αλληλεπίδραση με τους πελάτες/ καταναλωτές, προστατεύοντας συγχρόνως την ακεραιότητα και τη δημοτικότητά τους. Λαμβάνοντας λοιπόν υπόψη τα παραπάνω, δημιουργείται μια σαφής και κατατοπιστική εικόνα ως προς τις ανάγκες, τις προτιμήσεις και τις παρατηρήσεις του κοινού, την οποία μπορούν να αξιοποιήσουν κατάλληλα, βασίζοντας την ανάπτυξη στρατηγικών και τη λήψη αποφάσεων σε πραγματικά, ενημερωμένα και αξιόπιστα δεδομένα.

Η εύρεση, παρακολούθηση και απομόνωση συζητήσεων και σχολίων στο Διαδίκτυο και ειδικότερα στα μέσα κοινωνικής δικτύωσης, καθώς και η μετέπειτα διαχείριση και επεξεργασία του μεγάλου όγκου πληροφοριών που περιέχονται σε αυτά, είναι πλέον αδύνατο να γίνεται από

τον ανθρώπινο παράγοντα. Επομένως, η ανάγκη για αυτοματοποιημένα συστήματα τεχνητής νοημοσύνης και πιο συγκεκριμένα ανάλυσης συναισθήματος είναι η μοναδική και προφανής λύση. Με τη βοήθεια του ερευνητικού αυτού πεδίου, επιτυγχάνεται η αναγνώριση και ταξινόμηση των συναισθημάτων που αποτυπώνονται σε ένα κείμενο, ενώ εξασφαλίζεται η συνεχής ενημέρωση των μηχανισμών αυτών προκειμένου τα αποτελέσματα να συμβαδίζουν πάντα με την επικαιρότητα και να παραμένουν αντιπροσωπευτικά και αξιόπιστα. Οι εκθετικοί ρυθμοί με τους οποίους εξαπλώνονται οι χρήστες των μέσων κοινωνικής δικτύωσης στη σύγχρονη εποχή, σε συνδυασμό με τις απαραίτητες πρακτικές εφαρμογές που περιγράφηκαν παραπάνω, δημιουργούν εξαιρετικά μεγάλη ζήτηση και ανάγκη για νεοσύστατες εταιρείες που εστιάζουν στην παροχή υπηρεσιών ανάλυσης συναισθήματος. Αυτές οι πρακτικές εφαρμογές λοιπόν, παρέχουν ισχυρά κίνητρα για έρευνα στην ανάλυση του συναισθήματος και γνώμης.

1.3 Αντικείμενο Διπλωματικής Εργασίας

Το αντικείμενο της παρούσας διπλωματικής εργασίας αφορά την εκπαίδευση και αξιολόγηση γλωσσικών μοντέλων, με σκοπό την προσαρμογή τους σε εργασίες ανάλυσης συναισθήματος σε κείμενα που αναρτώνται στην ελληνική γλώσσα στα μέσα κοινωνικής δικτύωσης. Η διαδικασία υλοποίησης, περιλαμβάνει σε πρώτο στάδιο την προ-εκπαίδευσή των μοντέλων αυτών και στη συνέχεια την κατάλληλη προσαρμογή τους έτσι ώστε να μπορούν ταξινομούν το συναίσθημα που αποτυπώνεται σε ένα κείμενο ως θετικό, αρνητικό ή ουδέτερο. Ειδικότερα, το σώμα κειμένου και το σύνολο δεδομένων προς ανάλυση συναισθήματος, συγκεντρώθηκαν από την εταιρία ΠΑΛΟ ΨΗΦΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Ε.Π.Ε. [1], αποσκοπώντας στην στοχευμένη εξυπηρέτηση των εμπορικών εταιριών με τις οποίες συνεργάζεται. Τα δείγματα λοιπόν πάνω στα οποία θα εφαρμοστούν οι αλγόριθμοι βαθιάς μηχανικής μάθησης, θέτουν την πειραματική διαδικασία σε περιβάλλον πραγματικών συνθηκών, το οποίο επιφέρει και τις ανάλογες προκλήσεις. Συγκεκριμένα, λαμβάνοντας υπόψη τις, σχετικά με την παγκόσμια κλίμακα, περιορισμένες πηγές κειμένων στα ελληνικά, τόσο σε ποσοτικό όσο και ποιοτικό επίπεδο, αλλά και τα ιδιαίτερα μορφολογικά, συντακτικά και νοηματικά χαρακτηριστικά της γλώσσας (συντακτικά, γραμματικά και ορθογραφικά λάθη, εκτεταμένη χρήση αργκό, νοηματικά περίπλοκες έννοιες, συντομεύσεις κ.α), η μελέτη μας αποκτά ιδιαίτερο ενδιαφέρον και απαιτεί εξειδικευμένες τεχνικές βελτιστοποίησης προκειμένου να αντιμετωπίσει τις δυσκολίες αυτές. Η μελέτη μας πραγματεύεται τους πιθανούς τρόπους ανάπτυξης όσο το δυνατόν αποτελεσματικότερων ταξινομητών, ικανών να διαχειριστούν την ιδιαίτερη μορφή πληροφορίας που προέρχεται από τα μέσα κοινωνικής δικτύωσης. Για τον σκοπό αυτό, εξετάστηκε ένα ευρύ φάσμα μεθόδων βελτιστοποίησης των μοντέλων μας, όσον αφορά την αρχιτεκτονική αλλά και την ειδική προ-επεξεργασία των δεδομένων εκπαίδευσης. Συμπληρωματικά, οι ταξινομητές εφαρμόστηκαν στο δημοφιλές προ-εκπαιδευμένο γλωσσικό μοντέλο GreekBERT [17], κατασκευάζοντας το GreekSocialBERT [7].

1.4 Διάρθρωση Εργασίας

Η Διπλωματική Εργασία διαρθρώνεται σε 5 κεφάλαια. Στο παρόν Κεφάλαιο 1, πραγματοποιήθηκε μια εισαγωγή στο θέμα, τις πρακτικές εφαρμογές αλλά και τον λόγο που είναι απαραίτητη η έρευνα στο αντικείμενο. Στο Κεφάλαιο 2, καλύπτεται το θεωρητικό υπόβαθρο που απαιτείται ώστε να είναι δυνατή η κατανόηση εννοιών γύρω από την επεξεργασία φυσικής γλώσσας καθώς και την μοντελοποίηση των εφαρμογών της από υπολογιστικά συστήματα. Στο Κεφάλαιο 3, επεξηγούνται βασικές έννοιες αναφορικά με την ανάλυση συναισθήματος, ενώ στο Κεφάλαιο 4, ακολουθεί η αναλυτική περιγραφή της πειραματικής διαδικασίας που πραγματοποιήθηκε, καθώς και η αξιολόγηση των αποτελεσμάτων που προέκυψαν από αυτή. Το Κεφάλαιο 5, ολοκληρώνει την παρούσα εργασία με τα τελικά συμπεράσματα που προέκυψαν και τις μελλοντικές κατευθύνσεις στις οποίες θα μπορούσε να επεκταθεί η συγκεκριμένη μελέτη.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

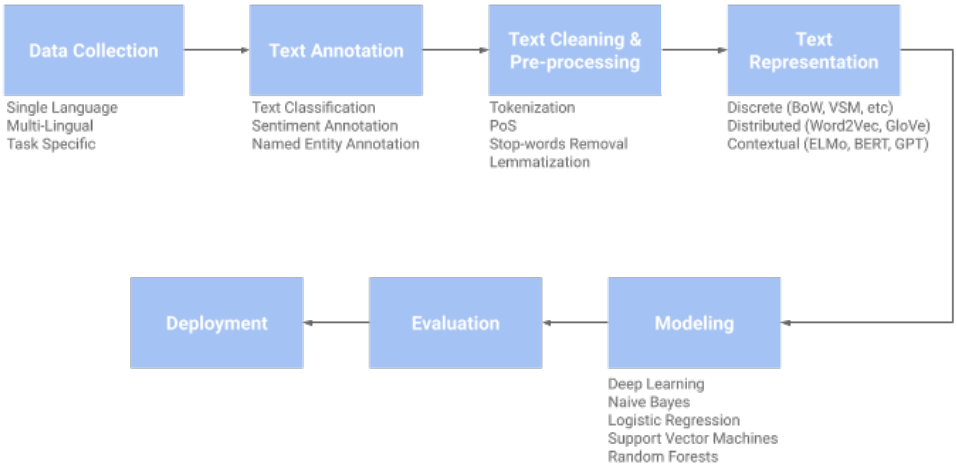
2.1 Επεξεργασία Φυσικής Γλώσσας

Η επεξεργασία φυσικής γλώσσας (ΕΦΓ) αποτελεί κλάδο της τεχνητής νοημοσύνης και αφορά την αποκωδικοποίηση, κατανόηση και επεξεργασία της ανθρώπινης γλώσσας από υπολογιστικές μηχανές. Με την εφαρμογή τεχνικών ΕΦΓ, από την ανάλυση συναισθήματος (sentiment analysis) έως και την αυτόματη αναγνώριση ομιλίας (speech recognition) , ο υπολογιστής είναι σε θέση να κατανοήσει την γραμματική, το συντακτικό και το «νόημα» (context) των δεδομένων φυσικής γλώσσας που αναλύει. Τα συχνότερα πεδία έρευνας της ΕΦΓ ακολουθούν ενδεικτικά:

- **Ανάλυση συναισθήματος - Εξόρυξη γνώμης** (Sentiment Analysis - Opinion Mining): κατανομή του συναισθήματος που αποτυπώνει ένα σώμα κειμένου ως θετικό, αρνητικό, ουδέτερο κλπ.
- **Απάντηση Ερωτήσεων** (Question Answering - Q&A): αναζήτηση της σωστής απάντησης σε μια δεδομένη ερώτηση διατυπωμένη σε ανθρώπινη γλώσσα
- **Εξαγωγή Επώνυμων Οντοτήτων** (Named Entity Recognition - NER): εντοπισμός και ταξινόμηση πληροφορίας σε διακριτές κατηγορίες οντοτήτων
- **Μηχανική Μετάφραση** (Language Translation): μετάφραση κειμένου από μία ανθρώπινη γλώσσα σε μία άλλη
- **Εξαγωγή Κειμένου** (Text Extraction): ανάκτηση πληροφοριών από κείμενα γραμμένα σε φυσική γλώσσα

Η βασική μεθοδολογία για την υλοποίηση εφαρμογών ΕΦΓ, όπως φαίνεται και στο παρακάτω σχήμα (Σχήμα 2.1), ξεκινά από την συλλογή (Data Collection) ενός όσο το δυνατότερο πιο αντιπροσωπευτικού σώματος κειμένου, το οποίο περιέχει την κατάλληλη πληροφορία για τον εκάστοτε σκοπό. Οι πηγές, το μέγεθος και τα γλωσσολογικά χαρακτηριστικά (συντακτικό, γραμματική, ορθογραφία κλπ) του σώματος κειμένου, έχουν πολύ βασικό ρόλο στην

αποτελεσματική μοντελοποίησή του, καθώς στην περίπτωση που ορισμένοι παράγοντες υστερούν, η απόδοση του μοντέλου ΕΦΓ δεν μπορεί να είναι η βέλτιστη. Σε περιπτώσεις που η ΕΦΓ πραγματοποιείται με σκοπούς ανάλυσης συναισθήματος, εξαγωγής επώνυμων οντοτήτων κ.λ.π., σημαντικό είναι γίνει η κατάλληλη επεξεργασία έτσι ώστε τα δεδομένα να κατηγοριοποιηθούν σωστά (Text Annotation). Το ατόφιο κειμενικό σώμα μπορεί φυσικά να βελτιωθεί, με την διόρθωση ορθογραφικών λαθών, την αφαίρεση διπλότυπων εισαγωγών, κ.α. Στη συνέχεια, η προεπεξεργασία (Preprocessing) επεκτείνεται πέρα από την ανθρώπινη αντίληψη ενός κειμένου, και επικεντρώνεται σε τρόπους μετατροπής του έτσι ώστε να είναι εύκολα διαχειρίσιμο από υπολογιστικές μηχανές. Αφού λοιπόν το σώμα κειμένου έχει λάβει την τελική του μορφή, γίνεται η αριθμητική του αναπαράστασή (Text Representation), έτσι ώστε να γίνει η συντακτική και σημασιολογική του ανάλυση με σκοπό να αποκωδικοποιηθεί μορφολογικά και λεξικολογικά η φυσική γλώσσα από τον υπολογιστή. Τέλος, ακολουθεί η εισαγωγή του στο κατάλληλο γλωσσικό μοντέλο όπου γίνεται η εκπαίδευση και η αξιολόγησή του.



Σχήμα 2.1: Διαδικασία υλοποίησης εφαρμογών ΕΦΓ

2.2 Τεχνικές Προεπεξεργασίας

Η προεπεξεργασία του κειμένου σχετίζεται με δύο βασικές τεχνικές· τη συντακτική ανάλυση και τη σημασιολογική ανάλυση.

2.2.1 Συντακτική Ανάλυση

Στη συντακτική ανάλυση πραγματοποιείται ανάλυση κειμένου χρησιμοποιώντας βασικούς γραμματικούς κανόνες προκειμένου να αποσαφηνιστεί η δομή της πρότασης, να γίνει κατανοητός ο τρόπος οργάνωσης των λέξεων και να εξεταστεί ο συσχετισμός μεταξύ τους όταν χρησιμοποιούνται σε μία πρόταση. Βασική μέθοδος που χρησιμοποιείται για τον παραπάνω σκοπό είναι αυτή της ανάλυσης σε σύμβολα (tokenization), δηλαδή της διάσπασης σε επι-

μέρους κομμάτια (tokens), τα οποία συνήθως είναι ολόκληρες λέξεις, κομμάτια λέξεων ή ακόμα και προτάσεις. Ακόμα, χρησιμοποιείται ευρέως και η πρακτική της επισημείωσης μερών του λόγου (Part-of-Speech (PoS) tagging, η οποία κατηγοριοποιεί τα σύμβολα ανάλογα με τον συντακτικό τους ρόλο, διακρίνοντας έτσι τα μέρη του λόγου, παραδείγματος χάρη προσδιορίζοντας την λέξη που αντιπροσωπεύουν ως ρήμα, ουσιαστικό, επίθετο κ.ο.κ. Κατ' αυτόν τον τρόπο, ο υπολογιστής μπορεί να κατανοήσει βαθύτερα την σημασία μιας λέξης ανάλογα με το πώς χρησιμοποιείται μέσα στην πρόταση. Εφαρμόζεται επίσης και η μέθοδος της λημματοποίησης (τλλεμματιζατιον / στεμμινγ), η οποία έχει ως σκοπό να επαναφέρει τη λέξη όσο το δυνατόν πιο κοντά στη ρίζα της, αφαιρώντας προθέσεις και καταλήξεις, ανάλογα με το πλαίσιο της πρότασης στην οποία βρίσκεται. Τέλος, η αφαίρεση κοινών λέξεων (stop-word removal), επιτυγχάνει την απαλοιφή των χωρίς σημασιολογική αξία λέξεων από το κείμενο, οι οποίες κατά βάση χρησιμοποιούνται έτσι ώστε η προτάσεις να έχουν συνοχή μεταξύ τους και να αποδίδεται νόημα. Οι παραπάνω μέθοδοι ωστόσο, δεν μπορούν να εφαρμοστούν πάντα με ευκολία κατά την ΕΦΓ, καθώς η εφαρμογή τους περιορίζεται σημαντικά σε γλώσσες με λίγους πόρους (low-resource languages), όπως τα Ελληνικά, δεδομένου ότι λόγω της περιορισμένης χρήσης τους στο Διαδίκτυο, δεν υπάρχει αφθονία υλικού. Σε τέτοιες περιπτώσεις, η έρευνα πολλές φορές βασίζεται σε γλώσσες με πολλούς πόρους (high-resource languages), π.χ Αγγλικά, έτσι ώστε να γίνει μετάφραση μεγάλων σωμάτων κειμένων σε άλλες φυσικές γλώσσες και να δημιουργηθούν επιπλέον δεδομένα προς ανάλυση για τα διάφορα πεδία έρευνας της ΕΦΓ.

2.2.2 Σημασιολογική Ανάλυση

Η σημασιολογική ανάλυση επικεντρώνεται στην αποκωδικοποίηση του νοήματος του κειμένου. Η αρχική προσέγγισή είναι η κατανόηση της σημασίας της κάθε επιμέρους λέξης, ενώ στη συνέχεια η ανάλυση εξετάζει το νόημα που προκύπτει από τον συνδυασμό λέξεων, και την σημασία που προσδίδουν στο κείμενο λαμβάνοντας υπόψη κάθε φορά τα συμφραζόμενά τους. Οι βασικές μέθοδοι που χρησιμοποιούνται είναι η επίλυση αμφίσημων λέξεων (Word sense disambiguation - WSD) και η εξαγωγή σχέσεων. Όσον αφορά το WSD, αποτελεί ένα από τα βασικά ζητούμενα που χρήζουν επίλυση στον τομέα της επεξεργασίας φυσικής γλώσσας και σκοπός είναι η ανάπτυξη τεχνικών έτσι ώστε να προσδιοριστεί σωστά το νόημα που προσδίδει μια λέξη στο κείμενο ανάλογα με τα συμφραζόμενά της. Η δυσκολία στην κατανόηση κειμένου συγκεκριμένα, αποδίδεται στη φύση πολλών λέξεων οι οποίες είτε έχουν διττή σημασία, είτε μπορούν να έχουν παραπάνω από ένα συντακτικό ρόλο στην πρόταση. Σε γλώσσες με πολλούς πόρους, η χρήση υψηλής ακριβείας επισημείωσης PoS μπορεί να αποτελέσει λύση στο πρόβλημα αυτό. Στο πεδίο της εξαγωγής σχέσεων, απαιτείται η ανίχνευση και κατηγοριοποίηση των σημασιολογικών σχέσεων ανάμεσα σε ένα σύνολο οντοτήτων (τοποθεσίες, πρόσωπα κλπ.).

2.3 Μοντέλα Εξαγωγής Χαρακτηριστικών

Μια σημαντική παράμετρος της επεξεργασίας φυσικής γλώσσας είναι ο τρόπος αναπαράστασης κειμένου σε ένα γλωσσικό μοντέλο. Προκειμένου να γίνει η εκπαίδευση το μοντέλου πάνω σε ένα σώμα κειμένου, απαιτείται η αριθμητική αναπαράσταση των όρων που το

αποτελούν. Σε μια ιδανική υλοποίηση, οι λέξεις που σχετίζονται μεταξύ τους αναπαρίστανται με διανύσματα κωδικοποιημένων όρων τοποθετημένα το ένα κοντά στο άλλο, ενώ οι λέξεις με μεγάλη σημασιολογική διαφορά αντιστοιχούν σε απομακρυσμένα στοιχεία του διανυσματικού χώρου. Ανάλογα με την φυσική γλώσσα του κειμένου, τα σύμβολα που προκύπτουν από την προεπεξεργασία του μπορεί να αντιστοιχούν είτε σε ολόκληρες λέξεις, είτε σε αποσπάσματα λέξεων. Ειδικά σε γλώσσες με λίγους πόρους, το ενδεχόμενο οι λέξεις να διασπώνται σε περισσότερα από ένα σύμβολα είναι αρκετά σύνηθες. Για την περιγραφή των παρακάτω μοντέλων αναπαράστασης κειμένου, θεωρούμε την αντιστοιχία λέξης-συμβόλου δεδομένη, έτσι ώστε να γίνει πιο κατανοητή η περιγραφή τους.

Η αναπαράσταση κειμένου μπορεί να είναι είτε Διακριτή (Discrete Text Representation), είτε Κατανεμημένη (Distributed Text Representation). Η βασική διαφορά ανάμεσα τους είναι ότι στην περίπτωση της Διακριτής Αναπαράστασης θεωρείται πως η κάθε λέξη του κειμένου είναι ανεξάρτητη από τις υπόλοιπες, με αποτέλεσμα το διάνυσμα που της αντιστοιχεί στην αριθμητική της αναπαράσταση να συμπληρώνεται ανάλογα με τις δικές ιδιότητες, όπως για παράδειγμα η αν συναντάται μέσα σε ένα κείμενο και με τι συχνότητα εμφανίζεται. Αντίθετα, η Κατανεμημένη Αναπαράσταση αναλύει τις σημασιακές σχέσεις ανάμεσα στις λέξεις λαμβάνοντας υπόψη τα συμφραζόμενά τους και τις αποτυπώνει αριθμητικά σε κάθε διάνυσμα. Τα μοντέλα που χρησιμοποιούνται για τη υλοποίηση των παραπάνω αναπαραστάσεων αναλύονται παρακάτω:

2.3.1 Διακριτή Αναπαράσταση Κειμένου

Κωδικοποίηση One-hot

Η κωδικοποίηση one-hot, μετατρέπει τις V λέξεις του κειμένου σε ένα V-διάστατο διανυσματικό χώρο, ο οποίος αποτελείται από την κωδικοποιημένη αναπαράσταση όλου του λεξιλογίου που περιλαμβάνει το σώμα κειμένου που εξετάζουμε. Η κάθε λέξη αναπαρίσταται από ένα τέτοιο σημείο - διάνυσμα, w_{vx1} , στον διανυσματικό χώρο, και έχει δυαδική αριθμητική μορφή. Συγκεκριμένα, οι θέσεις του κάθε διανύσματος συμπληρώνονται από μηδενικά, εκτός από μία· την θέση που αντιστοιχεί στη λέξη που κωδικοποιείται κάθε φορά και συμπληρώνεται με τον αριθμό 1. Για παράδειγμα, εφαρμόζουμε ενδεικτικά one-hot κωδικοποίηση στην πρόταση "Machine Learning is fun". Η αναπαράσταση της κάθε λέξης της παραπάνω πρότασης θα είναι:

$$\text{Machine} \rightarrow [1000], \text{Learning} \rightarrow [0100], \text{is} \rightarrow [0010], \text{fun} \rightarrow [0001]$$

Επομένως, η αριθμητική αναπαράσταση ολόκληρης της πρότασης θα είναι:

$$\text{sentence} = [[1,0,0,0],[0,1,0,0],[0,0,1,0],[0,0,0,1]]$$

Γενικεύοντας το παραπάνω παράδειγμα, καταλήγουμε στο συμπέρασμα πως με τη one-hot κωδικοποίηση σε μεγάλα σώματα κειμένου απαιτούνται πολύ μεγάλου μεγέθους διανύσματα V τάξης, καθώς κάθε στοιχείο ενός διανύσματος αντιστοιχεί σε μία μοναδική λέξη. Έτσι, για ένα λεξικό $100,000$ λέξεων, η κωδικοποίηση μίας μόνο λέξης απαιτεί διάνυσμα διάστασης $100,000$, επιβαρύνοντας κατά πολύ την υπολογιστική ισχύ και τη μνήμη ενός υπολογιστή. Παράλληλα,

πρόβλημα εντοπίζεται και στην εισαγωγή των one-hot διανυσμάτων στο γλωσσικό μοντέλο, δεδομένου ότι για κάθε κείμενο σχηματίζεται ένας 2-διάστατος πίνακας (n, m), όπου n ο αριθμός των συμβόλων του κάθε κειμένου και m το μέγεθος του λεξικού. Ανάλογα λοιπόν με την παράμετρο n, η οποία διαφέρει από κείμενο σε κείμενο, αλλάζουν και οι διαστάσεις της εισόδου, κάτι το οποίο δεν είναι διαχειρίσιμο από τα μοντέλα Μηχανικής Μάθησης. Μία λύση στο παραπάνω πρόβλημα είναι η μετατροπή των δειγμάτων σε διανύσματα με σταθερά ορισμένο μέγεθος L. Στη περίπτωση που τα σύμβολα του κειμένου ξεπερνούν την τιμή L, μπορεί να εφαρμοστεί περικοπή (truncation) στα περιττά σύμβολα, ενώ όταν υπάρχουν μικρότερα σε μέγεθος διανύσματα, να αυξηθεί το μήκος τους με την προσθήκη επιπλέον συμβόλων "γεμίσματος" (padding).

Η one-hot κωδικοποίηση ωστόσο, δεν παρουσιάζει αδυναμία μόνο στον τρόπο με τον οποίο θα αναπαρασταθούν οι λέξεις έτσι ώστε να είναι διαχειρίσιμες από τα γλωσσικά μοντέλα. Δεδομένου ότι το κάθε στοιχείο αντιμετωπίζεται ως ανεξάρτητη οντότητα, με μοναδικό κριτήριο τη θέση που του αντιστοιχεί στο σώμα κειμένου, η αναπαράσταση των λέξεων γίνεται αγνοώντας τη μεταξύ τους σημασιολογική σχέση, ενώ δεν λαμβάνεται υπόψη ούτε το εκάστοτε εννοιολογικό πλαίσιο. Συνεπώς, αυτή η πρακτική δεν εμβαθύνει στη σημασιολογία της γλώσσας την οποία εξετάζει και δεν μπορεί να αποδώσει αποτελεσματικά μια αντιπροσωπευτικη αριθμητική αναπαράσταση του κειμένου. Σε μια ιδανική υλοποίηση, θα έπρεπε οι λέξεις που σχετίζονται μεταξύ τους να αναπαρίστανται με διανύσματα τοποθετημένα το ένα κοντά στο άλλο, ενώ οι λέξεις με μεγάλη σημασιολογική διαφορά να αντιστοιχούν σε απομακρυσμένα στοιχεία του διανυσματικού χώρου, κάτι το οποίο όμως δεν μπορεί να επιτευχθεί με την one-hot κωδικοποίηση.

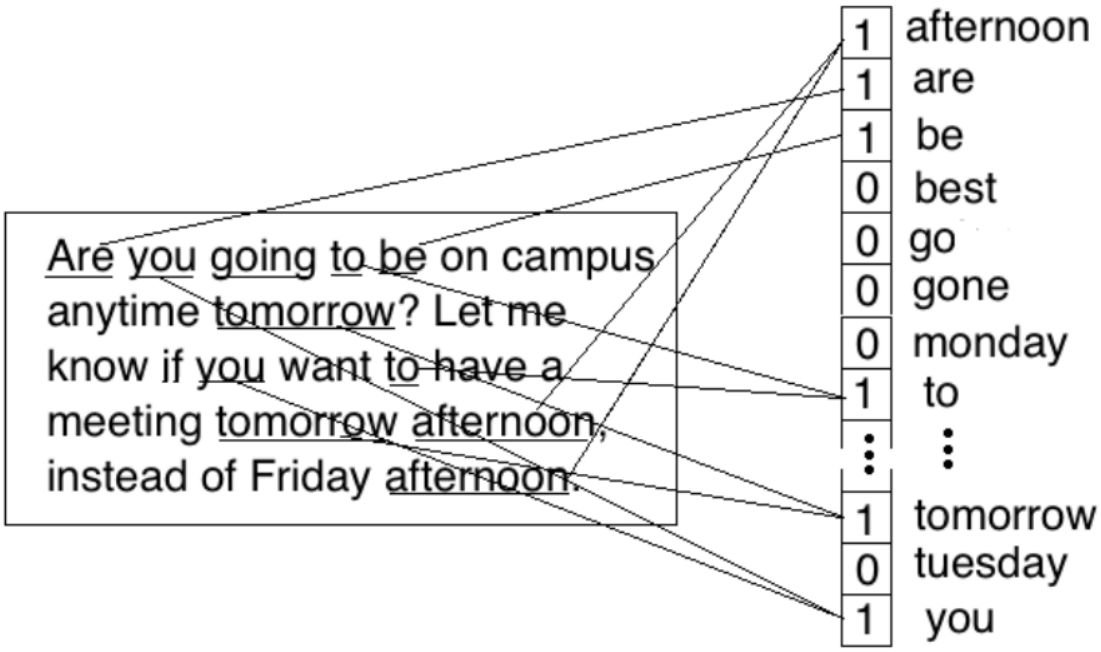
Μια από τις πιο διαδεδομένες τεχνικές αριθμητικής αναπαράστασης κειμένου είναι η μέθοδος του "σακουλιού λέξεων" (Bag-of-Words - BoW). Το μοντέλο αυτό επικεντρώνεται στον υπολογισμό της συχνότητας με την οποία εμφανίζονται οι λέξεις σε ένα κείμενο. Η ονομασία "σακκούλι λέξεων" προκύπτει από το γεγονός ότι λαμβάνονται υπόψη μόνο οι λέξεις που αποτελούν το σώμα κειμένου σαν ξεχωριστές οντότητες, ανεξάρτητα από τη δομή και την διάταξή τους στις προτάσεις που σχηματίζουν. Παρακάτω, αναλύονται οι διαφορετικές υλοποιήσεις του συγκεκριμένου μοντέλου:

Σακούλι Λέξεων

1. **Συχνότητα εμφάνισης λέξης**

   Μια από τις απλούστερες μορφές της παραπάνω τεχνικής αναπαράστασης είναι αυτή της συχνότητας εμφάνισης λέξεων. Όπως και στην one-hot κωδικοποίηση, η εισαγωγή των λέξεων στο γλωσσικό μοντέλο γίνεται με διανυσματική αναπαράσταση. Το μοντέλο αρχικά δημιουργεί ένα λεξικό μεγέθους V, το οποίο περιλαμβάνει όλες τις μοναδικές λέξεις που περιέχονται στο σώμα κειμένου. Σκοπός της υλοποίησης, είναι να αποδοθεί στην κάθε λέξη μια αριθμητική τιμή, η οποία θα αντιστοιχεί στη συχνότητα εμφάνισης αυτής της λέξης στο εκάστοτε κείμενο (Σχήμα 2.2). Έτσι, το σώμα κειμένου αντιστοιχίζεται σε ένα σύνολο από διανύσματα μεγέθους V, το καθένα από τα οποία αναπαριστά

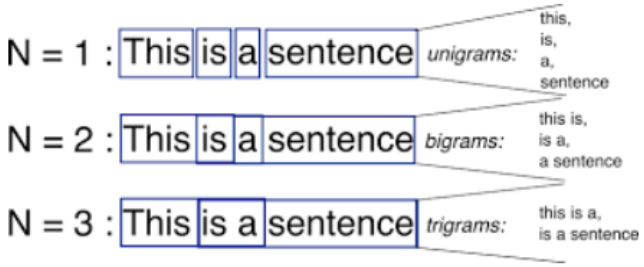
αριθμητικά τα επιμέρους κείμενα που το αποτελούν.



Σχήμα 2.2: Παράδειγμα BoW υλοποίησης με χρήση συχνότητας εμφανίσεων λέξεων [16]

2. **N-Grams**

Όπως και στην περίπτωση της συχνότητας εμφάνισης λέξης, υπολογίζεται η συχνότητα εμφάνισης των όρων ενός λεξικού. Ωστόσο, αντί το λεξικό που δημιουργείται να περιλαμβάνει όλες τις μοναδικές λέξεις του σώματος κειμένου, κάθε όρος του λεξικού αυτού αποτελείται από έναν συνδυασμό N λέξεων, όπου $N = 2, 3, 4, \ldots$. Προκειμένου λοιπόν να εφαρμοστεί η τεχνική N-Gram, πρέπει να οριστεί κατάλληλα η παράμετρος N έτσι ώστε να προκύψει, όταν αυτό είναι δυνατό, ένα δίγραμμα ($N = 2$), τρίγραμμα ($N = 3$) κ.ο.κ. (Σχήμα 2.3). Με την υλοποίηση αυτή, μπορεί να επιτευχθεί σημαντική μείωση στο μέγεθος του λεξικού σε σχέση με την παραπάνω απλή εφαρμογή BoW, καθώς και βαθύτερη κατανόηση της σημασιολογίας της γλώσσας, μέσω της συσχέτισης των λέξεων που παρουσιάζονται μαζί στο κείμενο. Έτσι, ένα, για παράδειγμα, μοντέλο "σακουλιού διγραμμάτων" πολλές φορές μπορεί να έχει αρκετά καλύτερη απόδοση από την απλή εφαρμογή της συχνότητας εμφάνισης μοναδικών λέξεων



Σχήμα 2.3: Παράδειγμα N-Gram υλοποίησης [2]

3. **TF-IDF**

Η δημοφιλέστερη προσέγγιση για την υλοποίηση του μοντέλου σακκουλιού λέξεων, είναι η κωδικοποίηση μέσω της συνδιασμένης μετρικής της συχνότητας όρου-αντίστροφης συχνότητας κειμένου (Term Frequency-Inverse Document Frequency - TF-IDF) . Σκοπός της TF-IDF είναι να αξιολογήσει κατά πόσο σημαντική είναι μία λέξη μέσα σε ένα κείμενο/πρόταση ως προς το νόημα που αποδίδει, λαμβάνοντας υπόψη τη συχνότητά με την οποία εμφανίζεται όχι μόνο στο κάθε κείμενο, αλλά σε ολόκληρο το σώμα κειμένων. Ειδικότερα:

- **Συχνότητα όρου** (Term Frequency - TF): ο όρος TF αφορά τη συχνότητα εμφάνισης μιας λέξης σε ένα απόσπασμα από το σώμα κειμένου, σε αναλογία με το μέγεθός του. Η σχέση που περιγράφει τη συχνότητα είναι:

$$TF(t) = \frac{\text{συχνότητα εμφάνισης λέξης t στο κείμενο}}{\text{συνολικός αριθμός λέξεων στο κειμένο}} \tag{2.1}$$

  Από την παραπάνω Εξίσωση 2.1 καταλήγουμε στο συμπέρασμα ότι στις περισσότερες περιπτώσεις ο TF όρος θα δώσει βαρύτητα κυρίως σε λέξεις που χρησιμοποιούνται με μεγάλη συχνότητα, όπως λ.χ. οι κοινές λέξεις.

- **Αντίστροφη συχνότητα κειμένου** (Inverse Document Frequency - IDF): ο όρος IDF αναλογεί στο πόσο σπάνια εμφανίζεται μία λέξη μέσα στο σώμα κειμένων

$$IDF(t) = \ln \frac{\text{συνολικός αριθμός κειμένων}}{\text{αριθμός κειμένων που περιέχουν την λέξη t}} \tag{2.2}$$

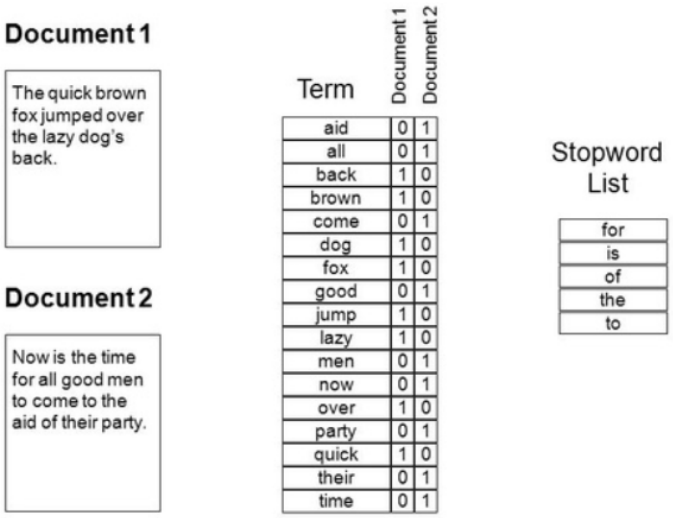
  Από την παραπάνω Εξίσωση 2.2 που περιγράφει τον IDF, φαίνεται πως όσο πιο συχνά εμφανίζεται η λέξη στο κείμενο, τόσο μικρότερη τιμή IDF έχει. Σε αντίθεση με τον όρο TF, ο όρος IDF αποφεύγει να αποδώσει μεγάλη βαρύτητα σε κοινές λέξεις, θεωρώντας πως οι σπανιότερες λέξεις θα έχουν μεγαλύτερη σημασιολογική βαρύτητα. Έτσι, μπορούν να ληφθούν υπόψη ορολογίες που χρησιμοποιούνται κατά βάση σε πιο εξειδικευμένο σώμα κειμένων, για παράδειγμα με ιατρικό περιεχόμενο.

Ο πολλαπλασιασμός των δύο προαναφερόμενων όρων μας δίνει το TF-IDF της λέξης t (Εξίσωση 2.3)

$$\text{TF-IDF}(t) = \text{TF}(t)\text{IDF}(t) \tag{2.3}$$

Οι παραπάνω υλοποιήσεις του μοντέλου σακκουλιού λέξεων, αν και δημιουργούν αριθμητικές αναπαραστάσεις που, σε αντίθεση με την one-hot κωδικοποίηση, μπορούν να εισαχθούν με ευκολία στα γλωσσικά μοντέλα, έχουν περιορισμούς ως προς την αποτελεσματικότητά τους. Ένα από τα βασικά μειονεκτήματα του BoW και των παραλλαγών του, είναι πως αγνοεί την διάταξη και τη δομή των λέξεων μέσα στο κείμενο και συνεπώς το νόημα που προκύπτει από τα συμφραζόμενα μπορεί πολύ εύκολα να αλλοιωθεί. Η συγκεκριμένη μέθοδος αναπαράστασης επίσης, εξαρτάται κατά πολύ από το σώμα κειμένου στο οποίο εφαρμόζεται και δεν μπορεί εύκολα να χρησιμοποιηθεί σε άλλης θεματολογίας δεδομένα. Για παράδειγμα, ένα σώμα κειμένου που αφορά τη Βιολογία, δεν μπορεί να διαχειριστεί κείμενα με θέμα την Επιστήμη Υπολογιστών, καθώς η υλοποίηση επεξεργάζεται μη διατεταγμένες λίστες όρων, ανεξάρτητα από τα συμφραζόμενά τους. Εκτός από την αδυναμία να κωδικοποιηθεί σωστά η σημασιολογία της γλώσσας, όσο πιο μεγάλο είναι το μέγεθος του λεξικού που δημιουργείται από το σώμα κειμένου, τόσο μεγαλύτερη είναι και η αριθμητική αναπαράσταση της κάθε λέξης, με αποτέλεσμα να προκύπτουν αραιά διανύσματα. Κατά τη μοντελοποίηση επομένως, απαιτείται η επεξεργασία ενός πολύ μεγάλου όγκου δεδομένων, του οποίου η χρήσιμη πληροφορία είναι αναλογικά ελάχιστη, που εξαντλεί τη μνήμη και την υπολογιστική ισχύ.

Για την βελτίωση της απόδοσης του BoW, θα μπορούσαν να εφαρμοστούν διάφορες τεχνικές επεξεργασίας κειμένου, με σκοπό την μείωση, όσο αυτό είναι δυνατό, του μεγέθους λεξικού που δημιουργείται. Συγκεκριμένα, η μετατροπή κεφαλαίων χαρακτήρων σε πεζούς και η αφαίρεση των σημείων στίξης, μπορεί να ελαττώσει σημαντικά την έκταση του λεξικού, ενώ σε γλώσσες με πολλούς πόρους, η εφαρμογή λημματοποίησης και η απαλοιφή των κοινών λέξεων, περιορίζει σε μεγάλο βαθμό την εισαγωγή περιττών λέξεων στο γλωσσικό μοντέλο (Σχήμα 2.4).



Σχήμα 2.4: Παράδειγμα BoW υλοποίησης με απαλοιφή κοινών λέξεων [8]

Μοντέλο Διανυσματικού Χώρου

Το μοντέλο διανυσματικού χώρου (Vector Space Model - VSM) αποτελεί μία εξέλιξη των τεχνικών κωδικοποίησης one-hot και σακκουλιού λέξεων, η οποία αναπαριστά το κάθε απόσπασμα/κείμενο ή την εκάστοτε ξεχωριστή λέξη του σώματος κειμένου, σε διανυσματική μορφή.

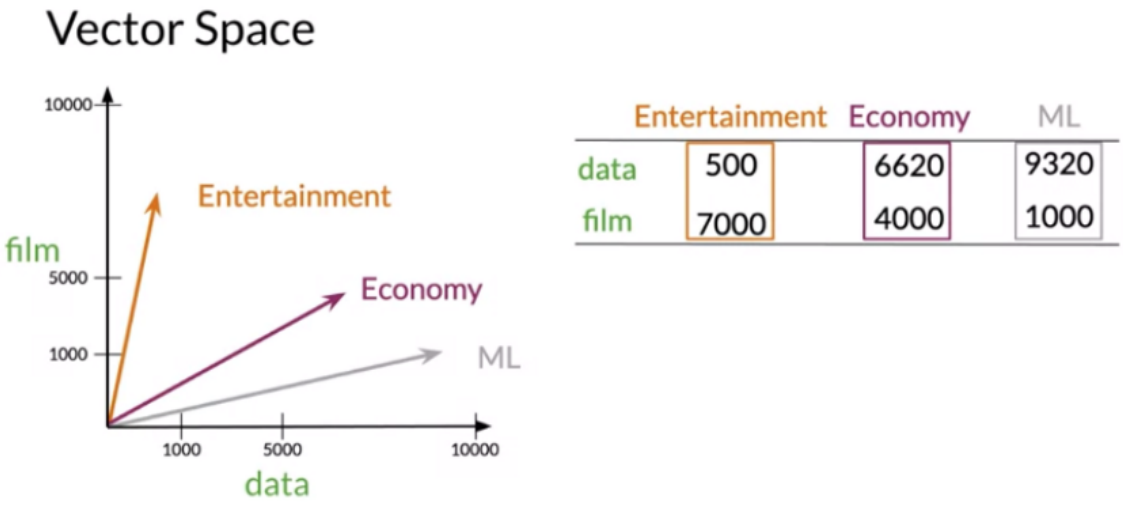
1. **Αναπαράσταση αποσπάσματος**

   Σε αυτή την αναπαράσταση κάθε λέξη αντιπροσωπεύεται από την συχνότητα εμφάνισής της σε κάθε απόσπασμα κειμένου. Έτσι, για ένα λεξικό μεγέθους V και σώμα κειμένου αποτελούμενο από D κείμενα, προκύπτουν DV-διάστατα διανύσματα, των οποίων οι όροι αντιστοιχούν στη συχνότητα εμφάνισης της κάθε λέξης $V_i, i = 0, 1, \ldots, V-1$, στο συγκεκριμένο απόσπασμα $D_j, j = 0, 1, \ldots, D-1$. Επομένως, πρακτικά προκύπτει πίνακας αριθμητικής αναπαράστασης διαστάσεων $V \times D$, ο οποίος συμπληρώνεται με την τιμή συχνότητας εμφάνισης της i-οστής λέξης, στο j-οστό απόσπασμα κειμένου, στην ij-οστή του θέση.

2. **Αναπαράσταση λέξης**

   Η συγκεκριμένη αναπαράσταση εξετάζει τις φορές που μια συγκεκριμένη λέξη εμφανίζεται σε κοντινή απόσταση από μια άλλη, μέσα σε ένα σώμα κειμένου. Η απόσταση αυτή, αντιστοιχεί στον αριθμό λέξεων που μεσολαβούν από την μία λέξη στην άλλη και ανάλογα με την υλοποίηση μπορεί να λάβει διάφορες τιμές. Προκύπτει λοιπόν ένας πίνακας αριθμητικής αναπαράστασης ο οποίος σε αυτή την περίπτωση για κάθε ζεύγος λέξεων i, j του λεξικού και απόσταση k, συμπληρώνεται στην θέση i, j με τη συχνότητα εμφάνισης της λέξης i σε εύρος απόστασης $[j-k, j+k]$ από την λέξη j. Για μέγεθος λεξικού V, ο πίνακας αναπαράστασης έχει διαστάσεις $V \times V$.

Η υλοποίηση του VSM, είτε με την πρώτη είτε με τη δεύτερη μέθοδο, αποδίδει ένα διανυσματικό χώρο που αντιπροσωπεύει τις σημασιολογικές σχέσεις μέσα σε ένα σώμα κειμένου. Στο παρακάτω Σχήμα 2.5, απεικονίζεται η κατασκευή του IDF με σχεδιασμό αναπαράστασης αποσπάσματος, ο οποίος εξετάζει τη σχέση των λέξεων με τις κατηγορίες των κειμένων που περιέχονται σε ένα σώμα κειμένου.



Σχήμα 2.5: Παράδειγμα VSM υλοποίησης [10]

Όπως ήταν αναμενόμενο, το μοντέλο καταλήγει στο συμπέρασμα ότι ο όρος "δεδομένα" (data), είναι πιο σχετικός με την "Οικονομία" (Economy) και τη "Μηχανική Μάθηση" (Machine Learning - ML), απ' ότι με την "Ψυχαγωγία" (Entertainment), η οποία είναι συναφής με τη λέξη "ταινία" (film). Στο συγκεκριμένο παράδειγμα, οι σχέσεις των όρων είναι εμφανείς. Ωστόσο, στις περισσότερες περιπτώσεις το μέγεθος του λεξικού είναι ιδιαίτερα μεγάλο και δεν είναι απαραίτητα εύκολη η διάκριση των σχέσεων μεταξύ των όρων. Για να πραγματοποιηθούν σωστά λοιπόν οι συγκρίσεις ανάμεσα στις διανυσματικές αναπαραστάσεις του μοντέλου, μπορούν να εφαρμοστούν αλγεβρικές μέθοδοι όπως ο υπολογισμός της Ευκλείδειας Απόστασης και της ομοιότητας συνημιτόνου

Η VSM αναπαράσταση, αν και μπορεί να επιφέρει αξιόλογα αποτελέσματα σε διάφορα πεδία έρευνας της ΕΦΓ, όπως και στις περιπτώσεις του σακκουλιού λέξεων και της κωδικοποίησης one-hot, αντιμετωπίζει το πρόβλημα εξάντλησης υπολογιστικών πόρων. Αυτό συμβαίνει εξαιτίας των πολυδιάστατων και αραιών πινάκων που δημιουργούνται κατά την εφαρμογή του, καθώς ένας πίνακας διαστάσεων MN έχει κόστος $\mathcal{O}(NM^2)$.

Οι αδυναμίες που παρουσιάζουν τα παραπάνω μοντέλα αναπαράστασης όσον αφορά τους υπολογιστικούς πόρους και την ελλιπή κατανόηση της σημασιολογίας της γλώσσας, μπορούν να καλυφθούν από την κατανεμημένη αναπαράσταση, η οποία αξιοποιεί μια διαφορετική προσέγγιση, αυτή των διανυσμάτων ενσωμάτωσης λέξεων (Word Embeddings).

2.3.2 Κατανεμημένη Αναπαράσταση Κειμένου

Η κατανεμημένη αναπαράσταση, χρησιμοποιεί διανύσματα ενσωμάτωσης λέξεων, τα οποία λαμβάνοντας υπόψη τα συμφραζόμενα της κάθε λέξης, μπορούν να αποδώσουν τη σημασιολογία του κάθε όρου αριθμητικά. Η τεχνική αυτή, βασίζεται στην υπόθεση ότι οι λέξεις που βρίσκονται η μία κοντά στην άλλη, έχουν υψηλότερη πιθανότητα να σχετίζονται μεταξύ τους. Με τη χρήση των διανυσμάτων ενσωμάτωσης λέξεων, κάθε λέξη αναπαρίσταται ως ένα πυκνό διάνυσμα συγκεκριμένων διαστάσεων. Κατ' αυτό τον τρόπο, αποφεύγεται η εισαγωγή πολυδιάστατων αραιών διανυσμάτων στο γλωσσικό μοντέλο και επομένως εξοικονομούνται σημαντικά οι υπολογιστικοί πόροι. Οι πιο δημοφιλής αλγόριθμοι για την παραγωγή των εν λόγω διανυσμάτων ακολουθούν παρακάτω:

Word2Vec

Ο αλγόριθμος Word2Vec [22], βασίζει την αναπαράσταση κάθε λέξης στους γειτονικούς όρους μέσα στο κείμενο, δηλαδή, στα συμφραζόμενά της. Όταν το σώμα κειμένου είναι αρκετά μεγάλο, μπορεί να αποτυπώσει διανυσματικά, όχι μόνο το συντακτικό της γλώσσας, αλλά και τις σημασιολογικές σχέσεις ανάμεσα στις λέξεις που την αποτελούν. Τα διανύσματα ενσωμάτωσης λέξεων που δημιουργούνται από αυτόν τον αλγόριθμο, τοποθετούνται με τέτοιο τρόπο μέσα στον διανυσματικό χώρο, έτσι ώστε οι σημασιολογικά σχετικές μεταξύ τους λέξεις να απεικονίζονται σε κοντινή απόσταση. Η αρχιτεκτονική του Word2Vec, αποτελείται από ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης 3 επιπέδων, που περιλαμβάνει ένα επίπεδο εισόδου, ένα κρυφό επίπεδο και τέλος ένα επίπεδο εξόδου.

Υπάρχουν δύο διαφορετικές εκδοχές του αλγόριθμου Word2Vec· το μοντέλο συνεχούς σακουλιού λέξεων ((Continuous Bag of Words - CBOW) και το μοντέλο πρόβλεψης N-άδεων (Skip-Gram) (Σχήμα 2.6). Ανάλογα με την υλοποίηση του αλγόριθμου Word2Vec, το νευρωνικό δίκτυο λαμβάνει τις παρακάτω μορφές:

- **CBOW** Στην περίπτωση του CBOW, γίνεται πρόβλεψη της αναπαράστασης της κεντρικής λέξης που εξετάζεται από το μοντέλο βάσει των γειτονικών όρων που την περιβάλλουν (συμφραζόμενα). Το μοντέλο λαμβάνει ως είσοδο ένα παράθυρο N γειτονικών λέξεων το οποίο περιέχει τα one-hot κωδικοποιημένα διανύσματα τους, και αναπαρίσταται με ένα διανυσματικό πίνακα $P_{|V|xN}$, όπου V το μέγεθος του λεξικού. Γειτονικές λέξεις θεωρούνται οι $n = \frac{N}{2}$ λέξεις που προηγούνται της κεντρικής, και n λέξεις που έπονται αυτής. Στη συνέχεια, το επίπεδο εισόδου προβάλλεται στο κρυφό επίπεδο που ακολουθεί. Στο δεύτερο επίπεδο, υπολογίζεται ο μέσος όρος της διανυσματικής αναπαράστασης P των συμφραζόμενων λέξεων του παραθύρου. Τέλος, αφού εφαρμοστεί συνάρτηση softmax στο γινόμενο του μέσου όρου με τα βάρη εξόδου του κρυφού επιπέδου, ως έξοδος, προκύπτει η διανυσματική αναπαράσταση της πιθανής λέξης που έχει ως συμφραζόμενα τις γειτονικές λέξεις εισόδου. Σε περίπτωση λανθασμένης εξόδου, η διαδικασία επαναλαμβάνεται και τα διανύσματα ενσωμάτωσης ενημερώνονται κατάλληλα μέχρι να βρεθεί η σωστή απάντηση. Στην αρχιτεκτονική του CBOW μοντέλου, η σειρά με την οποία εισάγονται οι προηγούμενες και οι επόμενες λέξεις δεν επηρεάζει

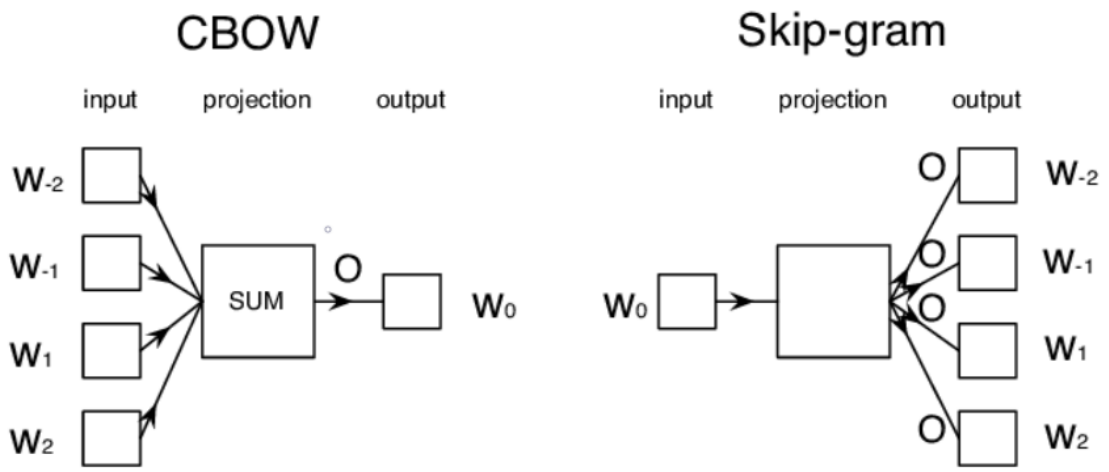
την προβολή στο κρυφό επίπεδο. Γι' αυτό το λόγο λοιπόν έλαβε την ονομασία Continuous Bag-of-Words. Ο όρος "Continuous" ("Συνεχής"), αποδίδεται στο γεγονός ότι δημιουργούνται συνεχείς κατανεμημένες αναπαραστάσεις λέξεων.

- **Skip-Gram**

Το μοντέλο Skip-Gram, σε αντίθεση με το CBOW, προβλέπει τις γειτονικές λέξεις χρησιμοποιώντας την κεντρική λέξη. Κατ' αυτόν τον τρόπο, εισάγεται στο πρώτο επίπεδο το αραιό διάνυσμα της one-hot κωδικοποίησης της λέξης, της οποίας το μέγεθος είναι ίσο με το μέγεθος του λεξικού του σώματος κειμένου. Τα διανύσματα ενσωμάτωσης αποτελούν τα βάρη του δευτέρου επιπέδου, το οποίο είναι ένα πλήρως διασυνδεδεμένο επίπεδο, όπως και στην περίπτωση του CBOW μοντέλου. Στην έξοδο, προκύπτουν, ανάλογα με το μέγεθος παραθύρου και την κατανομή της πιθανότητας που τους αποδίδεται, N one-hot αναπαραστάσεις των πιθανών γειτονικών λέξεων της εισόδου.

Το μοντέλο Word2Vec, μπορεί να αναπαραστήσει ακόμα και λέξεις που θεωρούνται σπάνιες, αποδίδοντας τόσο τον συντακτικό τους ρόλο, όσο και την σημασία τους ανάλογα με τα συμφραζόμενα ενός κειμένου. Ειδικά ο Skip-gram αλγόριθμος, έχει μεγάλη απόδοση όταν πρόκειται για την αναπαράσταση σπάνιων λέξεων, ωστόσο υστερεί αισθητά στον χρόνο εκπαίδευσής του, σε σχέση με το CBOW μοντέλο. Τα διανύσματα ενσωμάτωσης λέξεων που παράγει ο Word2Vec, σε αντίθεση με τα μοντέλα διακριτής αναπαράστασης που αναλύθηκαν στην προηγούμενη ενότητα, είναι πολύ πιο εύκολα διαχειρίσιμα από το γλωσσικό μοντέλο, καθώς το μέγεθός τους δεν είναι ανάλογο με το μέγεθος του λεξικού του σώματος κειμένου. Έτσι, αποφεύγεται η δημιουργία αραιών διανυσμάτων που επιβαρύνουν σημαντικά την υπολογιστική απόδοση και τον χρόνο επεξεργασίας.

Οι αλγόριθμοι Word2Vec διαθέτουν από ορισμένα χαρακτηριστικά τα οποία τους καθιστούν σε κάποιες περιπτώσεις μη βέλτιστους. Συγκεκριμένα, η υλοποίησή τους βασίζεται σε μεγάλο βαθμό στο παράθυρο γειτονικών λέξεων (window-based models) με αποτέλεσμα οι σημασιολογικές αναπαραστάσεις των λέξεων να περιορίζονται σε τοπικό επίπεδο. Επομένως, ενώ παρατηρείται υψηλή απόδοση του μοντέλου Word2Vec σε μικρότερα σύνολα δεδομένων, η εφαρμογή του σε μεγάλο όγκο κειμένων δεν είναι ανάλογη. Μια ακόμα βασική αδυναμία που παρουσιάζεται κατά την μοντελοποίηση, είναι η δυσκολία αναπαράστασης λέξεων που δεν ανήκουν στα δεδομένα εκπαίδευσης (training dataset), καθώς οι αλγόριθμοι όταν συναντούν μια λέξη εκτός του λεξικού (out-of-vocabulary, OOV), αγνοούν τη δομή και το εννοιολογικό της πλαίσιο και της αποδίδουν μια τυχαία διακριτή αναπαράσταση.



Σχήμα 2.6: Αρχιτεκτονική Word2Vec αλγορίθμων. Το μοντέλο CBOW προβλέπει την κεντρική λέξη wt βάσει των w(t-k) γειτονικών της και το μοντέλο Skip-Gram προβλέπει τις γειτονικές λέξεις λαμβάνοντας ως είσοδο την κεντρική. [18]

fastText

Το μοντέλο fastText [23] αναπτύχθηκε ως μια πιο εξελιγμένη εκδοχή του Word2Vec. Σκοπός του είναι να βρει λύση στο πρόβλημα των OOV λέξεων έτσι ώστε να μπορέσει να αναπαραστήσει με ευελιξία οποιαδήποτε λέξη. Η εναλλακτική στρατηγική που ακολουθεί το μοντέλο fastText, αναπαριστά την κάθε λέξη ως σακούλι χαρακτήρων n-gram, τα οποία οριοθετούνται από τα σύμβολα ‘ < ’ και ‘ > ‘ στην αρχή και το τέλος των συνόλων. Αναλυτικότερα, η τεχνική σακουλιού χαρακτήρα αποτελεί παραλλαγή της αναπαράστασης σακουλιού λέξεων και αντιπροσωπεύει αυτή τη φορά χαρακτήρες (κυρίως γράμματα), η οποία αξιοποιώντας την n-gram προσέγγιση αποδίδει την αναπαράσταση της λέξης ως συνδυασμό n-χαρακτήρων. Για παράδειγμα, η λέξη “what”, για παράμετρο $n = 3$ (τρίγραμμα), αναπαρίσταται ως εξής: <wh, wha, hat, at>. Στη συγκεκριμένη περίπτωση, κατά την αναπαράσταση της λέξης προκύπτει ο όρος “hat”, ο οποίος αν και στην παραπάνω ανάλυση αποτελεί απλά ένα στοιχείο τριγράμματος, θα μπορούσε να προκύψει ως αυτοτελής λέξη σε μια πρόταση. Η αναπαράσταση της λέξης “hat” ωστόσο δεν θα είναι η ίδια με αυτή που συναντάται στην αναπαράσταση της λέξης “what”. Αφού λοιπόν το μοντέλο fastText αναπαραστήσει την κάθε λέξη με χαρακτήρες n-gram, ένα Word2Vec μοντέλο εκπαιδεύεται και παράγει τα ανάλογα διανύσματα ενσωμάτωσης λέξεων. Σε μερικές περιπτώσεις μάλιστα, μέσω της διάσπασης των λέξεων σε επιμέρους n-grams, το μοντέλο μπορεί να εντοπίσει ευκολότερα τις σημασιολογικές σχέσεις ανάμεσα στη λέξη και στα επιμέρους συνθετικά της (π.χ “kingdom”, “king”). Ακολουθώντας αυτή τη λογική, ο αλγόριθμος μπορεί να παράγει με μεγάλη ταχύτητα οποιαδήποτε αναπαράσταση λέξης, χωρίς να είναι αναγκαίο αυτή να συμπεριλαμβάνεται στο σύνολο δεδομένων που έχει εκπαιδευτεί. Ωστόσο, η δημιουργία άπειρων συνδυασμών χαρακτήρων που απαιτεί αυτή η διαδικασία, προϋποθέτει πολύ μεγάλη υπολογιστική πολυπλοκότητα.

GloVe

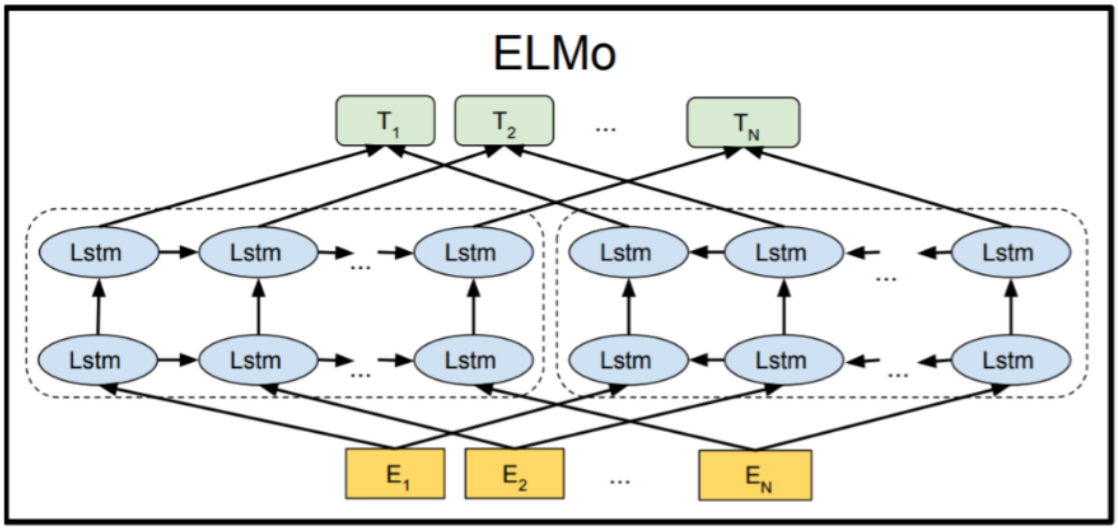
Όπως αναφέρθηκε και παραπάνω, οι μέθοδοι παραθύρου λέξεων αδυνατούν να αξιοποιήσουν σωστά μεγάλο όγκο πληροφορίας και περιορίζουν τις αναπαραστάσεις σε τοπικό επίπεδο. Αυτό συμβαίνει καθώς τα παράθυρα που επιλέγονται αφορούν τα συμφραζόμενα συγκεκριμένα γύρω από την λέξη που αναπαρίσταται και δεν μπορούν να αποδώσουν το γενικότερο νόημά της σε όλο το κείμενο. Το μοντέλο GloVe [25], σε αντίθεση με τους Word2Vec αλγορίθμους, δεν βασίζεται μόνο στην τοπική πληροφορία που δίνουν οι γειτονικές λέξεις, αλλά ενσωματώνει στατιστικά από όλη την έκταση του σώματος κειμένου. Ειδικότερα, τα διανύσματα ενσωμάτωσης δημιουργούνται βάσει της πιθανότητας συνεμφάνισης των λέξεων, η οποία καθορίζεται από το πόσο συχνά μια λέξη i εμφανίζεται στις λέξεις που περιβάλλουν μια άλλη λέξη j, στα πλαίσια ολόκληρης της συλλογής κειμένων. Τα καθολικά στατιστικά που υπολογίζονται για κάθε λέξη κατά τη μοντελοποίηση, βοηθούν τον αλγόριθμο να αναπαραστήσει σωστά ακόμα και σπανιότερες λέξεις, ενώ μπορεί να εφαρμοστεί τόσο σε μικρά σύνολα δεδομένων, όσο και σε μεγαλύτερα. Βέβαια, για την αξιοποίηση όλης αυτής της πληροφορίας, η πολυπλοκότητα είναι μεγάλη.

2.3.3 Προ-εκπαιδευμένα Μοντέλα

Τα προ-εκπαιδευμένα μοντέλα, με την παραγωγή διανυσμάτων ενσωμάτωσης συμφραζομένων (contextualized embeddings), αναπαριστούν το κείμενο λαμβάνοντας υπόψη και τα συμφραζόμενα της κάθε λέξης, έτσι ώστε να αποδοθεί κατάλληλα η εννοιολογική σημασία που της αντιστοιχεί στο συγκεκριμένο απόσπασμα. Παρακάτω, ακολουθούν μοντέλα ενσωμάτωσης (contextual models) που χρησιμοποιούνται στα πεδία έρευνας της ΕΦΓ

Μοντέλο ELMo

Το μοντέλο ELMo[31] αποτελεί είδος γλωσσικού μοντέλου βαθιάς αναπαράστασης λέξεων βάσει συμφραζομένων (deep contextualized word representation). Κατά τη μοντελοποίηση, τα συντακτικά και σημασιολογικά χαρακτηριστικά μιας φυσικής γλώσσας κωδικοποιούνται και βάσει αυτών, οι αναπαραστάσεις προσαρμόζονται ανάλογα με τη χρήση και το νόημα που αποδίδουν σε ένα συγκεκριμένο κείμενο. Το μοντέλο ELMo χρησιμοποιεί ένα αμφίδρομο νευρωνικό δίκτυο μακράς βραχυχρόνιας μνήμης (bidirectional long short term memory, biLSTM), έτσι ώστε οι αναπαραστάσεις να δημιουργούνται λαμβάνοντας υπόψη τόσο τις επόμενες λέξεις ενός όρου, όσο και την προηγούμενες στα πλαίσια μιας πρότασης (Σχήμα 2.7). Το καθένα από τα 2 biLSTM επίπεδα που αποτελούν το δίκτυο, επεξεργάζεται διαφορετικά την λέξη όσον αφορά το συντακτικό και το νόημά της και στο τέλος παράγει ένα διάνυσμα αναπαράστασης σύμφωνα με τη δική του προσέγγιση. Συγκεκριμένα, τα υψηλότερα LSTM επίπεδα επικεντρώνονται στη σημασιολογία των όρων, ενώ τα χαμηλότερα τείνουν να αναλύουν το συντακτικό ρόλο κάθε λέξης. Τελικά, το γλωσσικό μοντέλο αναπαριστά τη λέξη ως γραμμικό συνδυασμό των επιμέρους διανυσμάτων, εξασφαλίζοντας πως η λέξη μπορεί να έχει παραπάνω από μία αναπαράσταση, ανάλογα με τα συμφραζόμενα που την περιβάλλουν.



Σχήμα 2.7: Αρχιτεκτονική ELMo [31]

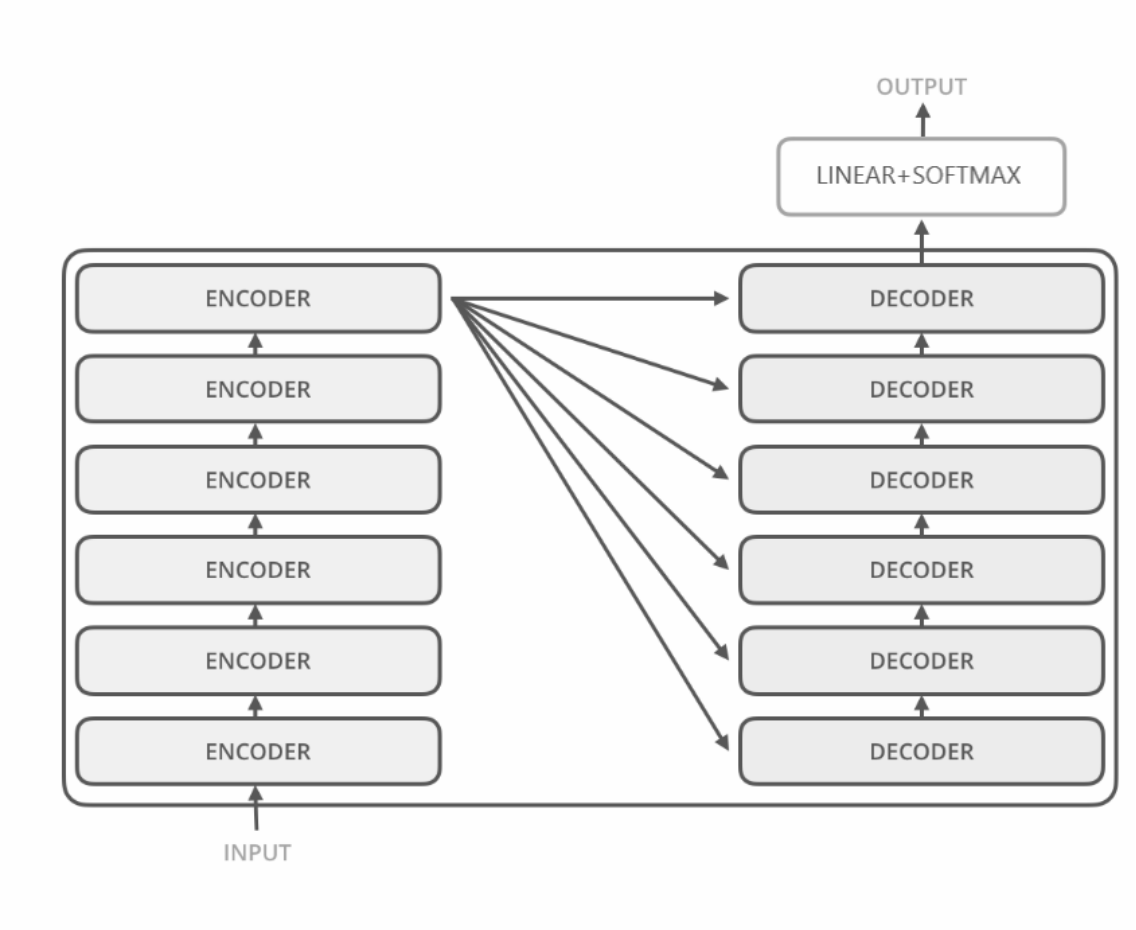
Μετασχηματιστές

Οι μετασχηματιστές (transformers) [32] αποτελούν υπερσύγχρονα γλωσσικά μοντέλα βαθιάς μηχανικής μάθησης, τα οποίο βασίζονται στην αρχιτεκτονική κωδικοποιητή - αποκωδικοποιητή (encoder - decoder) των αναδρομικών νευρικών δικτύων (Recurrent Neural Networks - RNNs) ακολουθίας-σε-ακολουθία (sequence-to-sequence - seq2seq). Βασική παράμετρος στην εκπαίδευση των μετασχηματιστών, είναι η έννοια της "προσοχής" (Attention). Ο μηχανισμός αυτός, λαμβάνοντας ως είσοδο κωδικοποιημένα διανύσματα, βοηθά το μοντέλο να ξεχωρίσει και να αξιοποιήσει τη χρήσιμη πληροφορία που περιέχουν, ανάλογα με τη σχέση των συμφραζομένων. Έτσι, οι αναπαραστάσεις δημιουργούνται με μεγάλη ακρίβεια, ενώ παράλληλα ο χρόνος επεξεργασίας μειώνεται σημαντικά.

- **Μηχανισμός Προσοχής**

  Ο μηχανισμός της προσοχής δίνει τη δυνατότητα στο μοντέλο να επικεντρωθεί στις λέξεις που είναι πιο σχετικές με τη λέξη που επεξεργάζεται. Η αρχιτεκτονική του μετασχηματιστή συγκεκριμένα, τόσο στη περίπτωση του κωδικοποιητή, όσο και στη περίπτωση το αποκωδικοποιητή, χρησιμοποιεί, όπως περιγράφεται παρακάτω και αποτυπώνεται στο Σχήμα 2.8, την τεχνική της αυτο-προσοχής (Self-Attention). Η αυτο-προσοχή, υπολογίζει την σημασιολογική σχέση της λέξης με κάθε έναν από τους όρους της ακολουθίας, εξετάζοντας όλους τους πιθανούς τρόπους που μπορεί να σχετίζεται με αυτές. Τελικά, επιλέγονται αυτές με τις υψηλότερες βαθμολογίες. Όσον αφορά τον αποκωδικοποιητή, εφαρμόζεται στην υλοποίηση του και μια επιπλέον παραλλαγή της προσοχής, η προσοχή κωδικοποιητή-αποκωδικοποιητή (Encoder-Decoder Attention). Ο μηχανισμός αυτός, λαμβάνει την έξοδο των κωδικοποιητών του μετασχηματιστή και σε συνδυασμό με τα αποτελέσματα του επιπέδου αυτο-προσοχής που προηγείται στον αποκωδικοποιητή, διαμορφώνει την δική του βαθμολογία για τη σχετικότητα των συμφραζόμενων όρων,

προσθέτοντάς τη με τη σειρά του στην αναπαράσταση της λέξης.



Σχήμα 2.8: Αρχιτεκτονική Μετασχηματιστή [5]

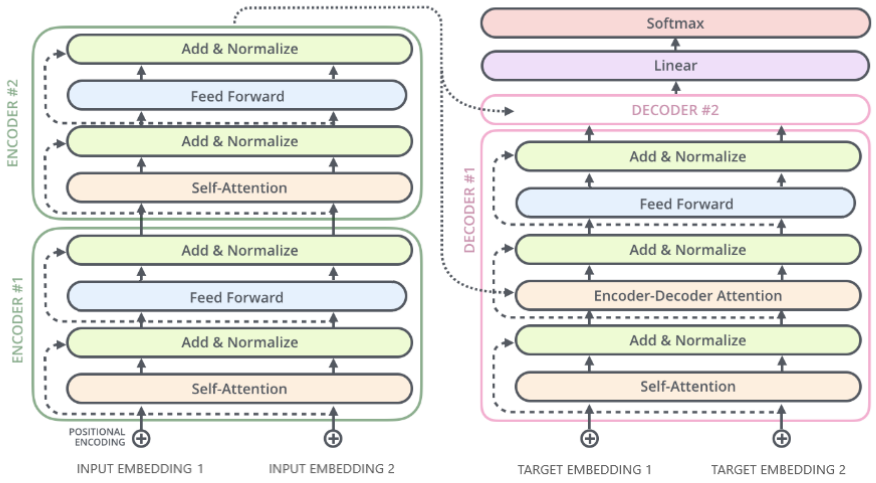
- **Πολλαπλές Κεφαλές Προσοχής**

  Κατά την υλοποίηση του Μετασχηματιστή, ο μηχανισμός της προσοχής επαναλαμβάνει τους υπολογισμούς του πολλές φορές, σε παράλληλο χρόνο. Κάθε φορά που συμβαίνει αυτό, οι διάφορες εκδοχές που παράγονται αντιστοιχούν σε μία κεφαλή προσοχής (Attention Head). Οι έξοδοι που προκύπτουν από την εκάστοτε κεφαλή, συνενώνονται και παράγουν μια τελική βαθμολογία προσοχής (Attention score). Σε αυτή τη περίπτωση, ο μετασχηματιστής χρησιμοποιεί μηχανισμό προσοχής πολλαπλών κεφαλών (Multiple Attention Heads) και κατ' αυτόν τον τρόπο ενισχύει την αποτελεσματικότητα της σημασιολογικής κωδικοποίησης της λέξης με τα συμφραζόμενά της, διακρίνοντας τις λέξεις με τις οποίες σχετίζεται.

- **Αρχιτεκτονική**

  Η βασική αρχιτεκτονική του μετασχηματιστή, αποτελείται από πολλαπλούς κωδικοποιητές και τους αντίστοιχους αποκωδικοποιητές τους. Όσον αφορά τον κωδικοποιητή, όπως φαίνεται και στο Σχήμα 2.9, ως είσοδο δέχεται τα διανύσματα ενσωμάτωσης της

ακολουθίας εισόδου (Embedding Layer), σε συνδυασμό με την κωδικοποιημένη αναπαράσταση της θέσης της λέξης μέσα στο κείμενο (Position Encoding Layer), ενώ η δομή κάθε επιπέδου κωδικοποίησης είναι πανομοιότυπη. Αναλυτικότερα, ο κάθε κωδικοποιητής περιλαμβάνει ένα επίπεδο αυτο-προσοχής που εκτιμά τις σχέσεις ανάμεσα στις διαφορετικές λέξεις της ακολουθίας εισόδου που δέχεται, τα αποτελέσματα του οποίου τροφοδοτούν το δίκτυο πρόσθιας τροφοδότησης που ακολουθεί. Η διαδικασία αυτή επαναλαμβάνεται σειριακά, με τον κάθε επόμενο κωδικοποιητή να δέχεται ως είσοδο την έξοδο του προηγούμενου. Κατά την μετάδοση πληροφορίας από το ένα επίπεδο στο άλλο, προκειμένου να σταθεροποιηθεί, η έξοδος των υπο-επιπέδων κανονικοποιείται (layer normalization). Στην περίπτωση των αποκωδικοποιητών, ως αρχική είσοδος λαμβάνεται το διάνυσμα ενσωμάτωσης που αναπαριστά την επιθυμητή ακολουθία εξόδου του μοντέλου (target sequence). Η αρχιτεκτονική, όπως και στην περίπτωση των κωδικοποιητών, είναι ίδια για κάθε αποκωδικοποιητή. Ανάμεσα στο επίπεδο αυτο-προσοχής και το νευρωνικό δίκτυο πρόσθιας τροφοδότησης που περιλαμβάνει, παρεμβάλλεται ένα επιπλέον επίπεδο κωδικοποίησης-αποκωδικοποίησης, ο ρόλος του οποίου είναι ανάλογος με αυτόν του επιπέδου αυτο-προσοχής, με τη διαφορά ότι βασίζει τη λειτουργία του όχι μόνο στην έξοδο του προηγούμενου επιπέδου, αλλά και στην έξοδο που προκύπτει από το σύνολο των κωδικοποιητών. Στην συνέχεια, τα αποτελέσματά του λειτουργούν ως είσοδος στο επίπεδο πρόσθιας τροφοδότησης που ακολουθεί. Όπως και στην αρχιτεκτονική των κωδικοποιητών, είσοδος για κάθε επόμενο αποκωδικοποιητή θεωρείται το αποτέλεσμα του προηγούμενου, ενώ σε κάθε υπό-επίπεδο προηγείται η κανονικοποίησή (layer normalization) των επιμέρους εξόδων που προκύπτουν.



Σχήμα 2.9: Αρχιτεκτονική Μετασχηματιστή με 2 κωδικοποιητές και 2 αποκωδικοποιητές [5]

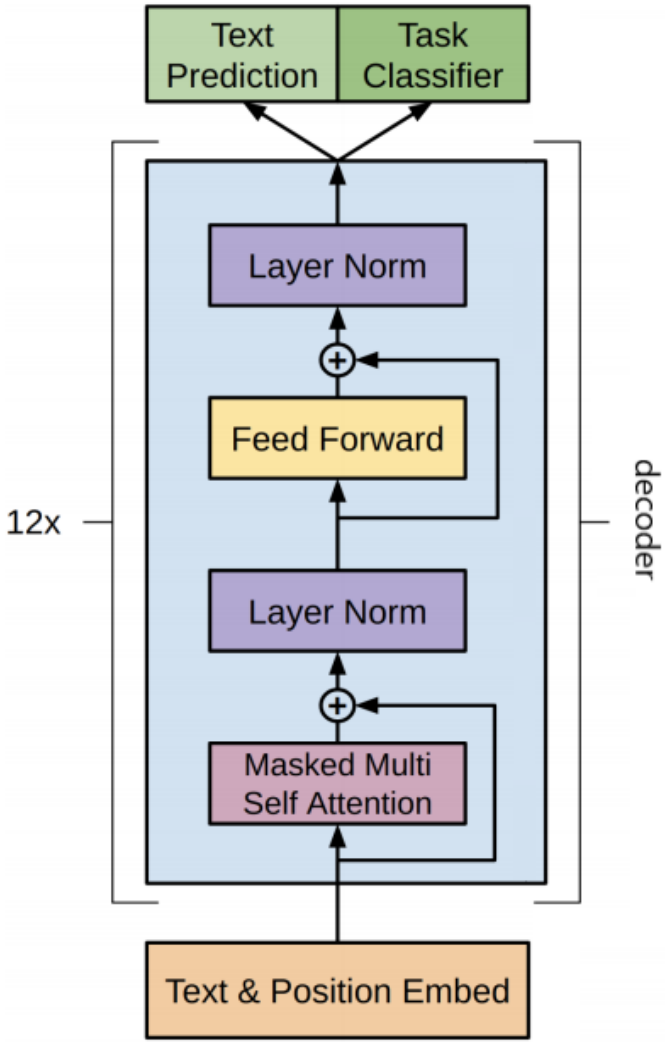
Ακολουθούν τα δημοφιλέστερα μοντέλα προ-εκπαίδευσης:

1. **GPT**

   Ο παραγωγικός προ-εκπαιδευμένος μετασχηματιστής (Generative Pre-Training Transformer - GPT) [28], αποτελεί μια ημι-εποπτευόμενη προσέγγιση στην κατανόηση και επεξεργασία φυσικής γλώσσας, η οποία συνδυάζει τη μη-εποπτευόμενη προ-εκπαίδευση (unsupervised pre-training) του μοντέλου σε μεγάλα σώματα κειμένου, με την εποπτευόμενη προσαρμογή των παραμέτρων του (supervised fine-tuning) σε συγκεκριμένες εργασίες. Βασικό πλεονέκτημα του GPT είναι ο όγκος πληροφορίας πάνω στην οποία έχει πραγματοποιήσει την προ-εκπαίδευσή του. Κάθε «νέα γενιά» GPT μοντέλων, όπως το GPT-2 και GPT-3, εκπαιδεύεται σε σώματα κειμένου τα οποία περιλαμβάνουν μέχρι και 10 φορές περισσότερες παραμέτρους. Το μεγάλο μέγεθος του μοντέλου επομένως, επιτρέπει την εύκολη εκπαίδευσή του πάνω σε εξειδικευμένα σύνολα δεδομένων (fine-tuning), χωρίς να απαιτείται μεγάλη ποσότητα νέας πληροφορίας.

   Συγκεκριμένα, δεδομένου ότι τα περισσότερα διαθέσιμα σώματα κειμένου δεν είναι προσαρμοσμένα πάνω σε κάποια ειδική εργασία, η υλοποίηση του GPT εκμεταλλεύεται την αφθονία πληροφορίας που περιέχουν, εκπαιδεύοντας σε πρώτο επίπεδο το γλωσσικό μοντέλο πάνω τους με στόχο την εκμάθηση των αρχικών παραμέτρων. Στη συνέχεια, η εποπτευόμενη εκπαίδευση που ακολουθεί, βασίζεται σε επισημειωμένα (annotated/labelled) σύνολα δεδομένων ειδικού περιεχομένου, τα οποία αν και περιορισμένου μεγέθους, βοηθούν το μοντέλο να προσαρμοστεί ανάλογα με την ζητούμενη εργασία. Η τεχνική αυτή ονομάζεται μεταφορά μάθησης (transfer learning) και μέσω αυτής επιτυγχάνεται η παραγωγή καθολικών αναπαραστάσεων κειμένου, οι οποίες με μικρές προσαρμογές μπορούν να εφαρμοστούν σε ένα ευρύ φάσμα πεδίων έρευνας της ΕΦΓ.

   Η αρχιτεκτονική του GPT, όπως φαίνεται και στο Σχήμα 2.10, είναι μια παραλλαγή της αρχιτεκτονικής των μετασχηματιστή. Συγκεκριμένα, ενώ η δομή των μετασχηματιστή αποτελείται, όπως αναφέρθηκε παραπάνω, από συνδυασμό κωδικοποιητών και αποκωδικοποιητών, το μοντέλο GPT αξιοποιεί μια στοίβα αποκωδικοποιητών 12 επιπέδων, κάθε ένα από τα οποία περιλαμβάνει μηχανισμό αυτο-προσοχής, ο οποίος ακολουθείται από ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης. Ως είσοδος στο πρώτο επίπεδο, όπως και στην περίπτωση των μετασχηματιστών, λαμβάνεται μια ακολουθία αναπαραστάσεων, η οποία είναι προσαρμοσμένη κατάλληλα για ειδική επεξεργασία. Λόγω της αρχιτεκτονικής του GPT, η επεξεργασία των συμφραζομένων για την πρόβλεψη της μελλοντικής λέξης που ακολουθεί γίνεται μόνο από τα αριστερά προς τα δεξιά μια πρότασης. Επομένως, η πληροφορία στην οποία βασίζεται το μοντέλο για την πρόβλεψή του, προκύπτει μόνο από τις προηγούμενες λέξεις και την ίδια τη λέξη που εξετάζει.



Σχήμα 2.10: Αρχιτεκτονική GPT [14]

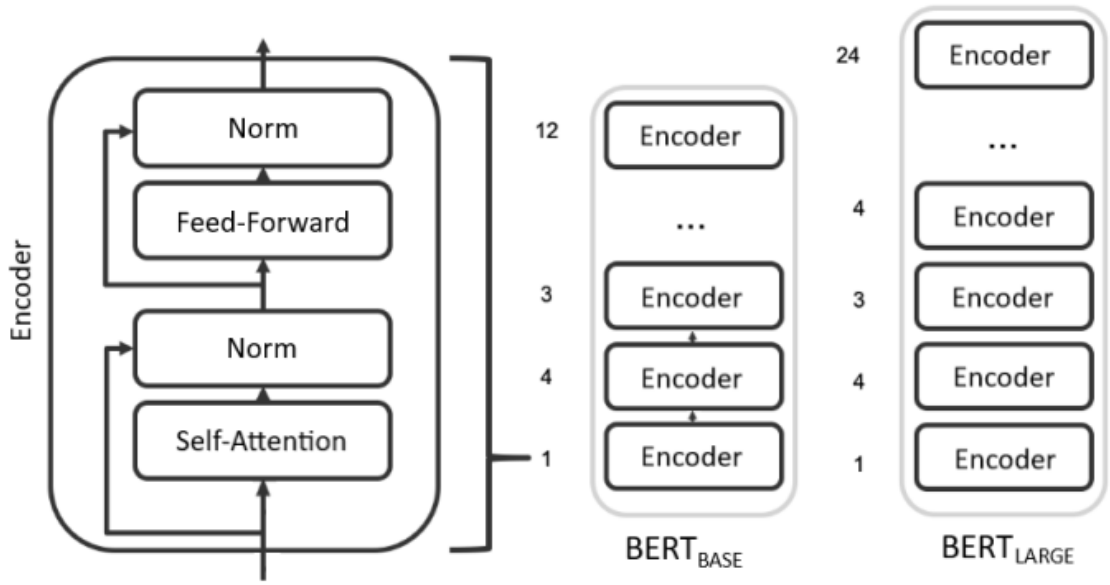
2. **BERT**

Το μοντέλο BERT (Bidirectional Encoder Representations from Transformers) [12], είναι ένα γλωσσικό μοντέλο το οποίο είναι σχεδιασμένο να εκπαιδεύει βαθιές αμφίδρομες αναπαραστάσεις. Σε αντίθεση με τον GPT μετασχηματιστή που αναφέρθηκε παραπάνω, το μοντέλο έχει τη δυνατότητα να διαβάζει την ακολουθία εισόδου στο σύνολό της, λαμβάνοντας υπόψη όχι μόνο τις λέξεις που προηγούνται ενός όρου, αλλά και αυτές που έπονται. Επομένως, η "αριστερά προς τα δεξιά" επεξεργασία του GPT μοντέλου, αντικαθίσταται από αμφίδρομη προσέγγιση των συμφραζομένων με αποτέλεσμα να αποκωδικοποιείται σωστά το νόημα που προσδίδουν τα συμφραζόμενα της πρότασης στη λέξη.

Η αρχιτεκτονική του BERT βασίζεται σε αυτή των μετασχηματιστών, ωστόσο στη συγκεκριμένη περίπτωση αξιοποιείται μόνο ο κωδικοποιητής του μοντέλου αναπαράστα-

σης, ο οποίος όπως αναλύθηκε και στην ενότητα , αποτελείται από έναν μηχανισμό αυτο-προσοχής και ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης που τον ακολουθεί. Συγκεκριμένα, πρόκειται για μια πολυεπίπεδη δομή κωδικοποιητών, η οποία, βάσει των προηγούμενων και των επόμενων λέξεων μιας λέξης που ανήκει στην ακολουθία εισαγωγής, εκπαιδεύεται να ερμηνεύει σημασιολογικά την λέξη και να αποδίδει την ακριβή αναπαράστασή της. Υπάρχουν δύο βασικές εκδοχές του BERT μοντέλου, των οποίων οι δομές, όπως αναλύεται παρακάτω, ξεπερνούν σε μέγεθος την απλή έκδοση του μετασχηματιστή (6 επίπεδα κωδικοποιητών, 512 κρυμμένες μονάδες, 8 κεφαλές), καθώς διαθέτουν περισσότερα επίπεδα κωδικοποιητών και μεγαλύτερα δίκτυα πρόσθιας τροφοδότησης (Σχήμα 2.11):

- BERT_{BASE}: Πρόκειται για το βασικό μοντέλο BERT, ανάλογο σε μέγεθος με τον GPT μετασχηματιστή. Αποτελείται από 12 κεφαλές προσοχής και 12 κωδικοποιητές, των οποίων το κάθε δίκτυο πρόσθιας τροφοδότησης περιέχει 768 κρυφές μονάδες.
- BERT_{LARGE}: Ένα μοντέλο BERT εξαιρετικά μεγάλων διαστάσεων με 16 κεφαλές προσοχής, 24 κωδικοποιητές 1024 κρυμμένων μονάδων.



Σχήμα 2.11: Αρχιτεκτονική BERT

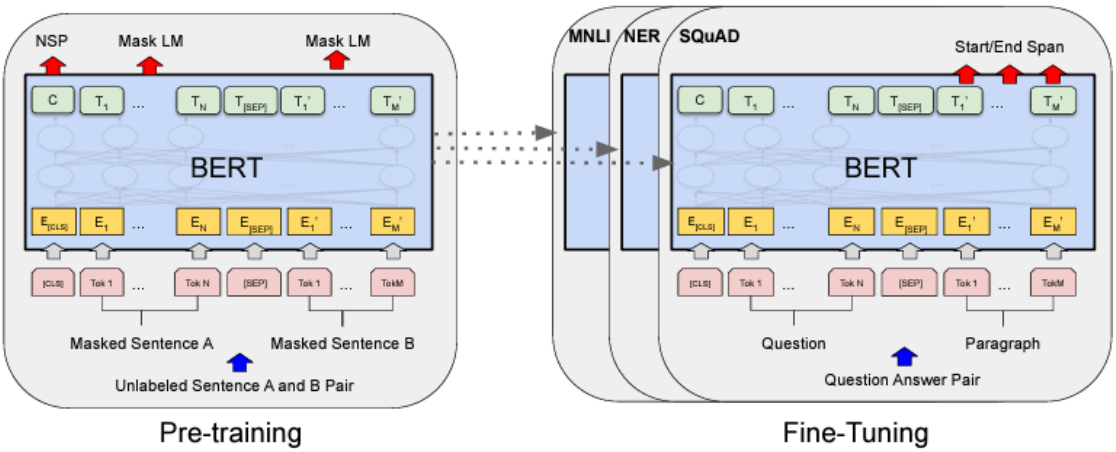
Για την προ-εκπαίδευση του BERT, χρησιμοποιούνται 2 βασικές στρατηγικές χωρίς επίβλεψη· το γλωσσικό μοντέλο απόκρυψης (Masked Language Model - MLM) και η πρόβλεψη της επόμενης πρότασης (Next Sentence Prediction - NSP).

- **Γλωσσικό μοντέλο απόκρυψης**: Τα περισσότερα γλωσσικά μοντέλα εκπαιδεύονται είτε "από αριστερά προς τα δεξιά" είτε "από δεξιά προς τα αριστερά",

καθώς η αμφίδρομη επεξεργασία θα είχε ως αποτέλεσμα η κάθε λέξη να "δει τον εαυτό της" και συνεπώς η πρόβλεψη της λέξης αυτής να μην έχει πλέον νόημα. Προκειμένου λοιπόν να γίνει η εκπαίδευση και από τις δύο κατευθύνσεις, γίνεται απόκρυψη του 15% των λέξεων της ακολουθίας εισόδου. Από το σύνολο των επιλεγμένων λέξεων, το 80% αυτών αντικαθίσταται από το ειδικό σύμβολο [MASK] του μετασχηματιστή. Ένα 10% των υπόλοιπων όρων αντικαθίσταται από μια άλλη διαφορετική λέξη και οι λέξεις που απομένουν (10%), παραμένουν ίδιες. Κατ' αυτό τον τρόπο, το μοντέλο μπορεί να εξετάσει αμφίδρομα όλα τα συμφραζόμενα της πρότασης και να επιχειρήσει να προβλέψει τη λέξη που λείπει. Η πρόβλεψη αυτή, θα αποτελέσει και την έξοδο του μοντέλου.

- **Πρόβλεψη της επόμενης πρότασης**: Με την χρήση της τεχνικής της πρόβλεψης της επόμενης πρότασης, σκοπός του μοντέλου είναι να προβλέψει την ύπαρξη σχέσης (ή μη) ανάμεσα σε 2 προτάσεις. Για την εκπαίδευση του μοντέλου χρησιμοποιείται ένα σύνολο δεδομένων από ζεύγη προτάσεων. Στο 50% των περιπτώσεων, η δεύτερη πρόταση είναι ακριβώς η ίδια με αυτή που διαδέχεται την πρώτη στο πρωτότυπο κείμενο, ενώ στις υπόλοιπες περιπτώσεις η δεύτερη πρόταση επιλέγεται τυχαία.

Το προ-εκπαιδευμένο μοντέλο BERT που προκύπτει, μπορεί να χρησιμοποιηθεί σε πολλές εφαρμογές της ΕΦΓ (Σχήμα 2.12. Η βαθιά κατανόηση της γλώσσας που προσφέρει ο μετασχηματιστής, διευκολύνει κατά πολύ την εκ νέου εκπαίδευση μοντέλων ειδικού σκοπού. Τα μοντέλα, μπορούν να αρχικοποιηθούν με τις παραμέτρους του BERT και στη συνέχεια να εκπαιδευτούν περαιτέρω πάνω σε σύνολο δεδομένων ειδικού περιεχομένου, που αφορά συγκεκριμένα τον σκοπό που ορίζει η ζητούμενη εργασία.



Σχήμα 2.12: Διαδικασία προ-εκπαίδευσης και του fine-tuning του BERT για το πρόβλημα της ερωταπόκρισης πάνω στο σύνολο δεδομένων SQuaD (Stanford Question Answering Dataset) [30]

3. **RoBERTa**

Η αποτελεσματικότητα του BERT, το καθιστά ένα από τα κυρίαρχα συστήματα επεξεργασίας φυσικής γλώσσας. Στα πλαίσια βελτίωσής του, έγινε προσαρμογή των διάφορων υπερ-παραμέτρων του γλωσσικού μοντέλου και του μεγέθους των δεδομένων εκπαίδευσής του. Τα αποτελέσματα της παραπάνω έρευνας, οδήγησαν στην ανάπτυξη του νέου μοντέλου RoBERTa (Robustly Optimized BERT) Pretraining από την Facebook AI [19], του οποίου οι επιδόσεις μπορούν να συναγωνιστούν με αυτές του GPT, ακόμα και να τις ξεπεράσουν σε ορισμένες περιπτώσεις. Η σύγκριση των μοντέλων, γίνεται εξετάζοντας ένα ευρύ πεδίο εργασιών της επεξεργασίας φυσικής γλώσσας, κάθε μια εκ των οποίων αντιστοιχίζεται σε ανάλογη συλλογή πληροφορίας, σύμφωνα με τη φύση του προβλήματός της. Συγκεκριμένα, προκειμένου να αξιολογηθεί η απόδοση του εκάστοτε μοντέλου στο κάθε πρόβλημα, γίνεται η εκπαίδευσή του πάνω σε δημοφιλείς συλλογές συνόλων δεδομένων μεγάλου μεγέθους (π.χ SQuAD, GLUE, MNLI, CoLA κ.α).

Το RoBERTa ενώ διατηρεί την βασική αρχιτεκτονική του GPT μοντέλου, διαφοροποιεί τον σχεδιασμό του με τις ακόλουθες αλλαγές: (1) περαιτέρω εκπαίδευση του μοντέλου σε επιπλέον δεδομένα και για μεγαλύτερο χρονικό διάστημα, (2) απαλοιφή της τεχνικής πρόβλεψης της επόμενης πρότασης κατά την εκπαίδευση, (3) μεγαλύτερες ακολουθίες εισόδου, (4) δυναμική απόκρυψη των λέξεων της ακολουθίας εισόδου (dynamic masking). Η διαφορετική αυτή προσέγγιση στον σχεδιασμό του GPT, απέδειξε το πόσο σημαντικό ρόλο έχουν οι υπερ-παράμετροι που καθορίζουν τον Μετασχηματιστή.

Το γλωσσικό μοντέλο RoBERTa, αποτελεί μια ιδιαίτερα αποτελεσματική, τελευταίας τεχνολογίας προσέγγιση του μετασχηματιστή GPT. Με ανάλογο τρόπο, έχουν αναπτυχθεί πολλές ακόμα εκδοχές του γλωσσικού αυτού μοντέλου (DistilBERT, XLNet κ.α), των οποίων η συμβολή στην βελτίωση και διεύρυνση των πεδίων έρευνας της ΕΦΓ είναι αξιοσημείωτη. Όσον αφορά την παρούσα διπλωματική εργασία, όπως αναλύεται παρακάτω, η πειραματική διαδικασία βασίστηκε στο γλωσσικό μοντέλο RoBERTa, ενώ έγινε σύγκριση των αποτελεσμάτων και με την προ-εκπαιδευμένη παραλλαγή του GPT στην ελληνική γλώσσα.

Κεφάλαιο 3

Ανάλυση Συναισθήματος

3.1 Ανάλυση Συναισθήματος και Αναπαραστάσεις

3.1.1 Αναπαράσταση σε κατηγορίες

Η κατανομή συναισθήματος μπορεί να γίνει σε πολλαπλά επίπεδα και με διαφορετικές προσεγγίσεις ανάλογα με το ζητούμενο που απαιτείται. Η απλούστερη μορφή αναπαράστασης έχει δυαδική μορφή και διακρίνει την συναισθηματική φόρτιση σε δύο βασικές κατηγορίες· θετική και αρνητική. Μια από τις συνηθέστερες κατηγοριοποιήσεις επίσης, είναι η κατάταξη κειμένου σε τρεις αντί για δύο συνιστώσες, δηλαδή σε θετική, αρνητική και ουδέτερη.

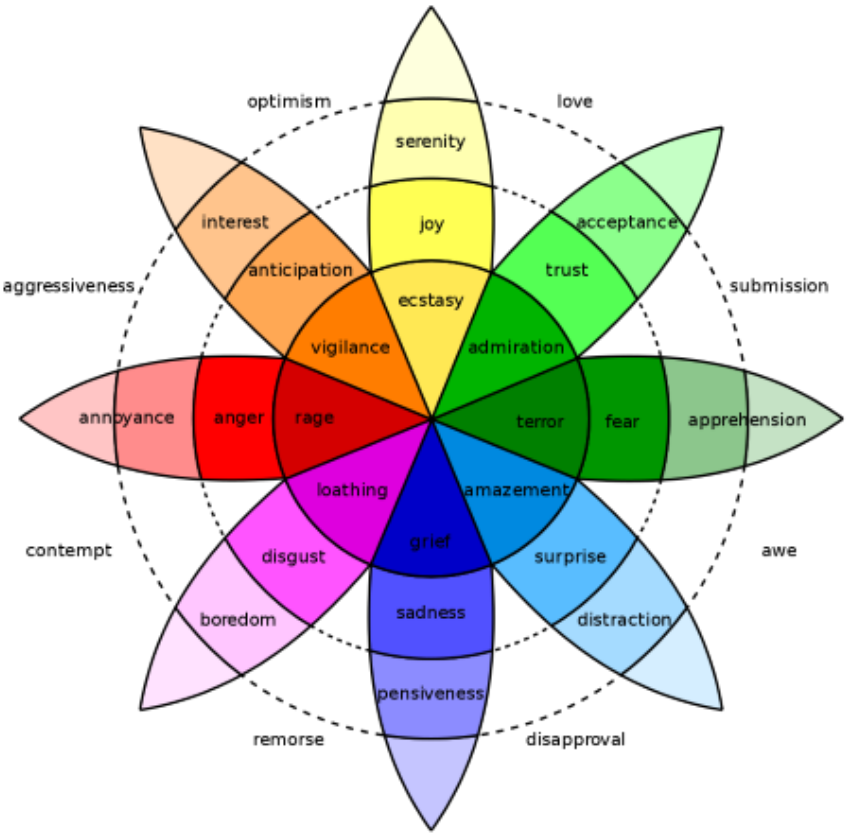
Στο πεδίο έρευνας συναντάται επίσης και η διαβάθμιση συναισθήματος σε 5 κατηγορίες (5-rating system) η οποία εστιάζει στην πολικότητα των συναισθημάτων και κατατάσσει τα συναισθήματα κλιμακωτά, χρησιμοποιώντας παραπάνω από τρεις βαθμίδες (Σχήμα 3.1).

Σχήμα 3.1: Διαβάθμιση συναισθήματος σε 5 κατηγορίες

Η παραπάνω διακύμανση στα συναισθήματα μπορεί να είναι είτε διακριτή, με απόλυτη αντιστοιχία σε μία από τις κατηγορίες, είτε να επεκτείνεται πέρα από τους κλασικούς όρους, χρησιμοποιώντας αριθμητικούς χαρακτήρες σε μορφή συνεχούς κλίμακας, όπου η ταξινόμηση μπορεί να λάβει και ενδιάμεσες τιμές, π.χ έναν βαθμό από -100 έως 100, με -100 να παριστάνει το «πολύ αρνητικό», το 0 το «ουδέτερο» και το 100 το «πολύ θετικό». Ομοίως, συχνή πηγή πληροφορίας σχετικά με την εντυπώσεις για ένα προϊόν αποτελεί και η απόδοση 1 έως 5 «αστεριών» (5-star rating) σε διάφορες online κριτικές.

Στο πεδίο έρευνας στην κοινότητα της Ψυχολογίας, η δημοφιλής μελέτη [27] προτείνει πως το εύρος των ανθρώπινων συναισθημάτων μπορεί να κατηγοριοποιηθεί με διακριτό τρόπο σε **8 βασικά συναισθήματα**: φόβο (fear), εμπιστοσύνη (trust), χαρά (joy), θλίψη (sadness), προσμονή (anticipation), αηδία (disgust) και έκπληξη (surprise), ο συνδυασμός των οποίων είναι δυνατόν να οδηγήσει στη δημιουργία περισσότερο πολύπλοκων συναισθημάτων (Σχήμα

3.2).



Σχήμα 3.2: Ο τροχός των συναισθημάτων του Plutchik [27]

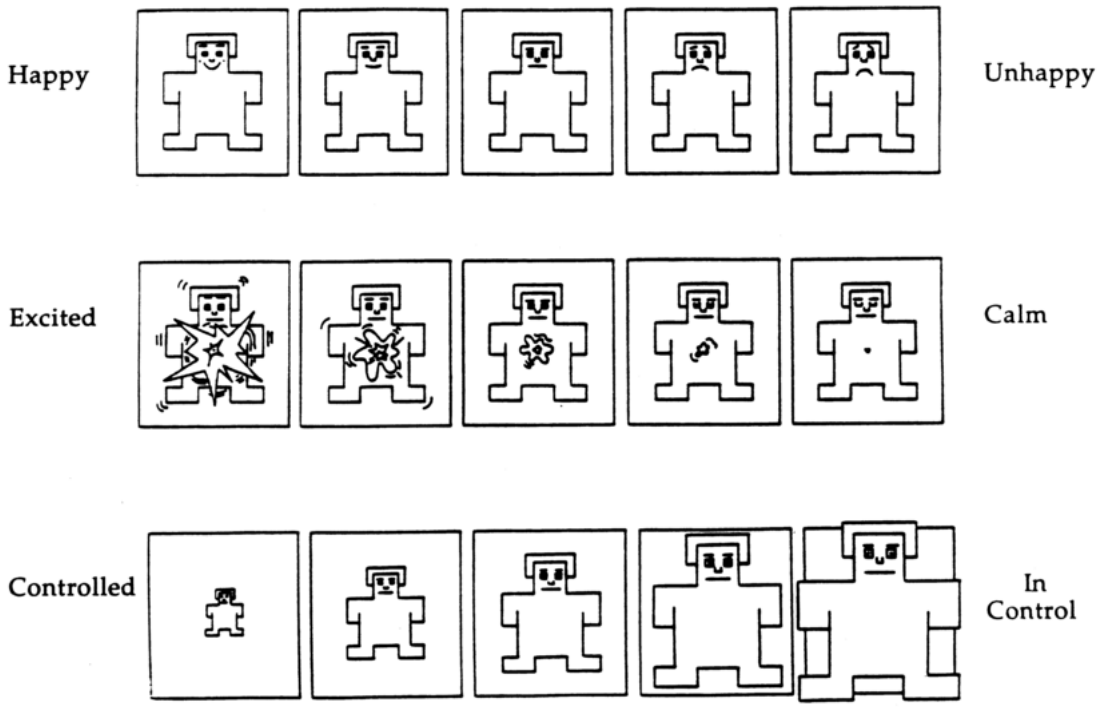
Ως επέκταση των παραπάνω, μια επίσης διαδεδομένη έρευνα [13], κατατάσσει τα συναισθήματα σε **6 βασικές κατηγορίες**: θυμό (anger), αηδία (disgust), φόβο (fear), ευτυχία (happiness), θλίψη (sadness) και έκπληξη (surprise).

Βάσει των παραπάνω, στο πεδίο έρευνας της ανάλυσης συναισθήματος συχνά υιοθετούνται οι συγκεκριμένες προσεγγίσεις και τα δεδομένα επισημειώνονται σύμφωνα με αυτές. Ωστόσο, από υπολογιστικής άποψης, τα συναισθήματα που επιλέγονται προς εξέταση κατά το σχεδιασμό ενός συστήματος αναγνώρισης συναισθήματος, εξαρτώνται συχνά από την εφαρμογή και το σκοπό της, οπότε στην περίπτωση αυτή θα πρέπει να λαμβάνεται υπόψιν και το πλαίσιο μέσα στο οποίο θα πραγματοποιείται η μελέτη.

3.1.2 Διαστατικές Αναπαραστάσεις

Μια εναλλακτική προσέγγιση για την περιγραφή των συναισθημάτων είναι η αναπαράσταση τους σύμφωνα με χαρακτηριστικά συνεχούς μορφής, τις διαστάσεις. Οι συνηθέστερες παράμετροι υπολογιστικής αξιολόγησης στην διαστατική αναπαράσταση των συναισθημάτων είναι η ενεργοποίηση (activation), το σθένος (valence) και η κυριαρχία (dominance). Συ-

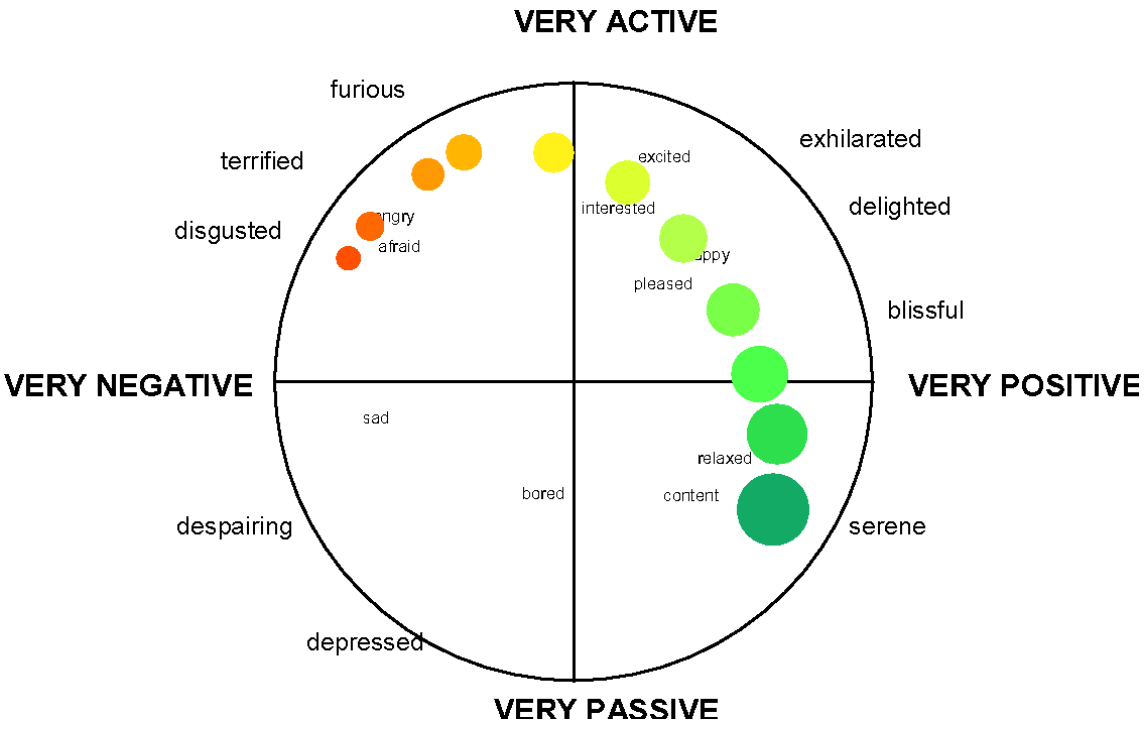
γκεκριμένα, η ενεργοποίηση περιγράφει πόσο έντονη είναι η συναισθηματική εμπειρία, ενώ το σθένος αξιολογεί το επίπεδο ευχαρίστησης που σχετίζεται με ένα συναίσθημα. Μπορεί να λάβει τόσο θετικές, όσο και αρνητικές τιμές που αντιστοιχούν σε ευχάριστα και δυσάρεστα συναισθήματα αντίστοιχα. Τέλος, η κυριαρχία περιγράφει την ικανότητα ενός ατόμου να ελέγξει τα συναισθήματά του κατά τη διάρκεια μιας συναισθηματικά φορτισμένης εμπειρίας. Τα συστήματα που χρησιμοποιούν αναπαραστάσεις διαστάσεων δεν απαιτούν τον προκαθορισμό ενός συνόλου συναισθηματικών κατηγοριών (emotional classes), όπως στην περίπτωση της διακριτής κατηγοριοποίησης (Ενότητα 3.1.1), ωστόσο απαιτούν συνήθως την προεπιλογή του αριθμού των επιπέδων της ενεργοποίησης και του σθένους που θα χρησιμοποιηθούν κατά τη διαδικασία της ταξινόμησης. Για να διευκολυνθεί η επισημείωση των συναισθηματικών εκδηλώσεων σε κλίμακες ενεργοποίησης, σθένους και κυριαρχίας αντίστοιχα, ερευνητικές μελέτες εισήγαγαν την κλίμακα αυτοαξιολόγησης (Self-Assessment Manikins - SAM)). Το μοντέλο SAM αποτελείται από διαισθητικές εικονογραφικές παραστάσεις (manikins), που περιγράφουν τα χαρακτηριστικά στις τρεις παραπάνω διαστάσεις. Ένα παράδειγμα του SAM για μία κλίμακα 5 σημείων αναφορικά με την ενεργοποίηση της αξιολόγησης, του σθένους και της κυριαρχίας απεικονίζεται Σχήμα 3.3 [9].



Σχήμα 3.3: SAM: Σθένος (σειρά 1) - Ενεργοποίση (σειρά 2) - Κυριαρχία (σειρά 3) [20]

Εναλλακτικά, οι ερευνητές προσπάθησαν να αξιοποιήσουν πλήρως τη συνεχή φύση των εν λόγω αναπαραστάσεων, συλλέγοντας συνεχείς βαθμολογήσεις (ratings) συναισθηματικών διαστάσεων και αναπτύσσοντας συστήματα που να υπολογίζουν τις συνεχείς συναισθηματικές ιδιότητες. Ένα ακόμη βήμα αποτελεί η συγκέντρωση βαθμολογιών συνεχούς μορφής των ιδιοτήτων των διαστάσεων κατά τη διάρκεια του χρόνου και η δημιουργία συστημάτων που

να αντιπροσωπεύουν συναισθηματικά χαρακτηριστικά, όχι ως σημεία, αλλά ως καμπύλες που λαμβάνουν συνεχείς τιμές και εξελίσσονται στο χρόνο. Η συγκέντρωση τέτοιων βαθμολογιών συνεχούς μορφής προτάθηκε από την μελέτη [11], όπου παρουσιάστηκε το λογισμικό επισημείωσης Feeltrace, που απεικονίζεται στο Σχήμα 3.4. Το Feeltrace δίνει τη δυνατότητα στο χρήστη να παρέχει επισημείωση συναισθηματικού περιεχομένου σε πραγματικό χρόνο μετακινώντας το δρομέα (cursor) σε μια διεπαφή που αντιπροσωπεύει το δισδιάστατο χώρο της ενεργοποίησης και του σθένους, τη στιγμή που οι συναισθηματικές εκδηλώσεις εμφανίζονται σε ξεχωριστό πρόγραμμα αναπαραγωγής βίντεο. Το αποτέλεσμα αυτής της διαδικασίας είναι μια συναισθηματική καμπύλη για κάθε συναισθηματική ιδιότητα.



Σχήμα 3.4: Feeltrace: Διάσταση σθένους (οριζόντιος άξονας) - Διάσταση ενεργοποίησης (κάθετος άξονας) [21]

Ανεξάρτητα από την εκάστοτε αναπαράσταση, η υποκειμενικότητα του έργου της ταξινόμησης αποτελεί πρόκληση ως προς την ορθότητα της συναισθηματικής επισημείωσης. Τα άτομα συχνά αντιλαμβάνονται τα συναισθήματα με βάση τη δική τους προσωπική και εσωτερική διαδικασία και κρίση, η οποία ενδέχεται να μην ταυτίζεται με την πραγματική συναισθηματική επισημείωση μιας συναισθηματικής εκδήλωσης. Φυσικά, όσο μεγαλύτερο είναι το επίπεδο λεπτομέρειας των συναισθηματικών σχολιασμών, π.χ., περισσότερες ετικέτες που αναπαριστούν το συναίσθημα σε κατηγορίες, περισσότερα επίπεδα συναισθηματικών χαρακτηριστικών, τόσο λιγότεροι θα είναι οι σχολιαστές που αναμένεται να συμφωνήσουν. Το γεγονός αυτό υπογραμμίζει μια θεμελιώδη πρόκληση του προβλήματος της αναγνώρισης των συναισθημάτων, δηλαδή το γεγονός ότι υπάρχει αβεβαιότητα όσον αφορά στην επισημείωση που σχετίζεται με ένα παράδειγμα, σε αντίθεση με τον πιο διαδεδομένο τρόπο αναγνώρισης, όπου υπάρχει σα-

φής συσχέτιση μεταξύ ενός παραδείγματος και της αντίστοιχης διακριτής και συγκεκριμένης κατηγοριοποίησής του.

3.2 Ανάλυση συναισθήματος βασισμένη στις όψεις

Μια ιδιαίτερα ενισχυτική και αποδοτική προσθήκη στις εργασίες ανάλυσης συναισθήματος, είναι ο στοχευμένος συνδυασμός της εφαρμογής της με την αναζήτηση συγκεκριμένων και επιλεγμένων χαρακτηριστικών που ανήκουν σε ένα κείμενο. Η διαδικασία εύρεσης αυτών των γνωρισμάτων ανήκει στο πεδίο της ανάλυσης συναισθήματος βασισμένη στις όψεις (Aspect-Based Sentiment Analysis - ABSA) και αφορά την ταξινόμηση σε κατηγορία συναισθημάτων εστιάζοντας σε ένα συγκεκριμένο γνώρισμα. Η τεχνική ABSA χρησιμοποιείται για κάθε ξεχωριστό αντικείμενο - χαρακτηριστικό μιας οντότητας. Για παράδειγμα, στις κριτικές ενός φορητού υπολογιστή, εάν ενδιαφερόμαστε συγκεκριμένα για την απόδοση του επεξεργαστή, μπορούμε να θεωρήσουμε μια κριτική θετική ή αρνητική ανάλογα με την γνώμη των καταναλωτών για το συγκεκριμένο χαρακτηριστικό, και να μην βασιστούμε στην γενική εικόνα, που ενδεχομένως να μην συμβαδίζει με το ειδικό κριτήριο που έχουμε θέσει. Συγκεκριμένα, το σύστημα προσπαθεί να εντοπίσει το κυριότερο χαρακτηριστικό του αντικειμένου που έχει επιλεχθεί και παράλληλα το μέσο όρο σχολιασμού (πχ. κριτική με 1 έως 5 αστέρια) που του αντιστοιχεί προκειμένου να διαμορφώσει κατάλληλα την αξιολόγησή του. Το σύστημα ABSA, λοιπόν, αποτελεί ένα εξαιρετικά χρήσιμο εργαλείο ειδικά για εταιρείες και οργανισμούς που προσφέρουν υπηρεσίες ή/και προϊόντα, καθώς επιτρέπει την άμεση ανταπόκριση και εποικοδομητική αναθεώρηση και βελτίωση των παροχών της, ανάλογα με την ανατροφοδότηση των καταναλωτών και των πελατών που εξυπηρετεί.

3.3 Ανάλυση Συναισθήματος στα Μέσα Κοινωνικής Δικτύωσης

3.3.1 Εφαρμογές

Συγκεκριμένα, ιδιαίτερο ενδιαφέρον παρουσιάζεται στα μέσα κοινωνικής δικτύωσης, όπου μέσω της τεχνικής αυτής επιτυγχάνεται η βαθύτερη κατανόηση των αναγκών, προτιμήσεων και απόψεων του ευρύτερου κοινού, προκειμένου να αξιοποιηθεί κατάλληλα από οργανισμούς, επιχειρήσεις και υπηρεσίες, σχετικά με τα προϊόντα και τις παροχές που προσφέρουν. Η παρακολούθηση της απήχησης και της δημοτικότητας ενός προϊόντος, μιας υπηρεσίας, ακόμα και ενός προσώπου, γίνεται με την ανάλυση της πληθώρας δεδομένων που προέρχονται από τις πλατφόρμες κοινωνικής δικτύωσης, σε μορφή κριτικών, σχολίων, άρθρων κ.ο.κ. Η πληροφορία που προκύπτει από την ανάλυση αυτή, αποτελεί ένα εξαιρετικά χρήσιμο εργαλείο, το οποίο μπορεί να αξιοποιηθεί για την βελτίωση και εξέλιξη προϊόντων και υπηρεσιών, ενώ παράλληλα αποτελεί βασικό παράγοντα επιρροής στη λήψη αποφάσεων και την ανάπτυξη στρατηγικών, προκειμένου να προσαρμοστούν οι καμπάνιες προώθησης προϊόντων, υπηρεσιών και φυσικών προσώπων, ανάλογα με τις νέες ανάγκες και τάσεις που προκύπτουν. Παράλληλα, η ανάλυ-

ση αυτή της κοινής γνώμης, μπορεί να λειτουργήσει και ως εποικοδομητική κριτική και να χρησιμοποιηθεί ως κίνητρο για βελτίωση και αναθεώρηση των συγκεκριμένων αστοχιών.

Με τη βοήθεια του ερευνητικού αυτού πεδίου, επιτυγχάνεται η αναγνώριση και ταξινόμηση των συναισθημάτων που αποτυπώνονται σε ένα κείμενο, ενώ εξασφαλίζεται η συνεχής ενημέρωση των μηχανισμών αυτών προκειμένου τα αποτελέσματα να συμβαδίζουν πάντα με την επικαιρότητα και να παραμένουν αντιπροσωπευτικά και αξιόπιστα. Οι εκθετικοί ρυθμοί με τους οποίους εξαπλώνονται τα μέσα κοινωνικής δικτύωσης στη σύγχρονη εποχή, σε συνδυασμό με τις απαραίτητες πρακτικές εφαρμογές που περιγράφηκαν παραπάνω, δημιουργούν εξαιρετικά μεγάλη ζήτηση και ανάγκη για νεοσύστατες εταιρείες που εστιάζουν στην παροχή υπηρεσιών ανάλυσης συναισθήματος. Αυτές οι πρακτικές εφαρμογές λοιπόν, παρέχουν ισχυρά κίνητρα για έρευνα στην ανάλυση του συναισθήματος και γνώμης.

3.3.2 Προκλήσεις

Η πληροφορία που λαμβάνουμε από τα μέσα κοινωνικής δικτύωσης αν και, σύμφωνα με την παραπάνω ανάλυση, εξαιρετικά πολύτιμη, απαιτεί ιδιαίτερη προσοχή κατά την επεξεργασία της. Συγκεκριμένα, τα κείμενα που συναντώνται σε όλες τις πλατφόρμες μέσα κοινωνικής δικτύωσης, τείνουν να διακρίνονται από μορφολογικά και ποιοτικά χαρακτηριστικά που μπορούν να δυσκολέψουν το έργο της επεξεργασίας φυσικής γλώσσας. Αυτό συμβαίνει καθώς κύριο χαρακτηριστικό των δειγμάτων αυτών είναι ο ανεπίσημος και πολλές φορές πρόχειρος τρόπος γραφής των χρηστών των μέσων κοινωνικής δικτύωσης, ο οποίος σε πολλές περιπτώσεις περιέχει σημασιολογικές αστοχίες, ορθογραφικά και συντακτικά λάθη, συντομεύσεις λέξεων, σχήματα λόγου, δυσνόητες και αυθαίρετες έννοιες, ενώ ιδιαίτερα συχνή είναι και η χρήση της αργκό. Το γεγονός αυτό, σε συνδυασμό με τα ιδιαίτερα μορφολογικά γνωρίσματα που εμφανίζουν τα κείμενα, όπως η αυθαίρετη χρήση σημείων στίξης, κεφαλαίων γραμμάτων και ειδικών χαρακτήρων (π.χ emoticons, hashtags κ.α), δημιουργεί μια μεγάλη πρόκληση κατά την μοντελοποίηση των κειμένων και συνεπώς η πιστή αναπαράσταση του νοήματός τους έχει ένα μεγαλύτερο βαθμό δυσκολίας. Επιπλέον, η περιορισμένη χρήση της ελληνικής γλώσσας στο Διαδίκτυο, περιορίζει ακόμα περισσότερο την δυνατότητα εύρεσης εύκολα διαχειρίσιμου και ποιοτικού υλικού για τις διάφορες εργασίες ΕΦΓ. Συνεπώς, προκειμένου να γίνει σωστή και όσο το δυνατόν αντιπροσωπευτικότερη εφαρμογή της ανάλυσης συναισθήματος σε δείγματα από τα μέσα κοινωνικής δικτύωσης, είναι απαραίτητη η μεθοδική επιλογή στοχευμένου, σχετικού με το αντικείμενο μελέτης υλικού, καθώς και η μετέπειτα κατάλληλη επεξεργασία των δεδομένων αυτών, ως προς την απαλοιφή των περιττών και περιοριστικών γνωρισμάτων που εμφανίζουν.

Κεφάλαιο 4

Πειραματική Διαδικασία

4.1 Περιγραφή και Εργαλεία

Στα πλαίσια της παρούσας διπλωματικής εργασίας, αξιοποιήθηκαν οι μηχανισμοί των μετασχηματισμών, προκειμένου να γίνει εκπαίδευση και αξιολόγηση γλωσσικών μοντέλων ανάλυσης συναισθήματος πάνω σε κείμενα γραμμενά στα ελληνικά, προερχόμενα από τα μέσα κοινωνικής δικτύωσης. Η αρχιτεκτονική μετασχηματιστή στην οποία βασίστηκε η μοντελοποίηση, είναι αυτή του RoBERTa [19], η δομή του οποίου αποτελείται από ένα προηγμένο νευρωνικό δίκτυο, ικανό να αποτυπώσει περίπλοκες σημασιολογικές έννοιες. Για την υλοποίηση των παραπάνω αλγορίθμων βαθιάς μηχανικής μάθησης, χρησιμοποιήθηκε η ανοιχτού κώδικα βιβλιοθήκη μετασχηματιστών HuggingFace [33]. Η συγκεκριμένη πλατφόρμα, παρέχει ένα σύνολο από χρήσιμα εργαλεία, βιβλιοθήκες και γλωσσικά μοντέλα που εξυπηρετούν τις διάφορες εφαρμογές της ΕΦΓ και διατίθενται για χρήση στις βιβλιοθήκες βαθιάς μάθησης PyTorch [24] και TensorFlow [3] της γλώσσας προγραμματισμού Python. Η πειραματική διαδικασία που ακολουθήθηκε, αφού προηγήθηκε η προεπεξεργασία των δεδομένων, αφορά τρεις βασικές πρακτικές βαθιάς μάθησης· αυτή της ανάλυσης του σώματος κειμένων σε σύμβολα (tokenization), της προ-εκπαίδευσης (pre-training) του μοντέλου καθώς και την μετέπειτα στοχευμένη προσαρμογή του (fine-tuning), έτσι ώστε να αναπτυχθεί το τελικό μοντέλο ανάλυσης συναισθήματος. Συμπληρωματικά, για να συγκρίνουμε τις επιδόσεις των ταξινομητών που υλοποιήσαμε σύμφωνα με την παραπάνω μεθοδολογία, έγινε η επιπλέον εφαρμογή τους σε ένα ιδιαίτερα δημοφιλές προ-εκπαιδευμένο μοντέλο, το GreekBERT [17]. Οι περιγραφές των παραπάνω διαδικασιών, ακολουθούν στις επόμενες υποενότητες.

4.2 Σχετικές Μελέτες

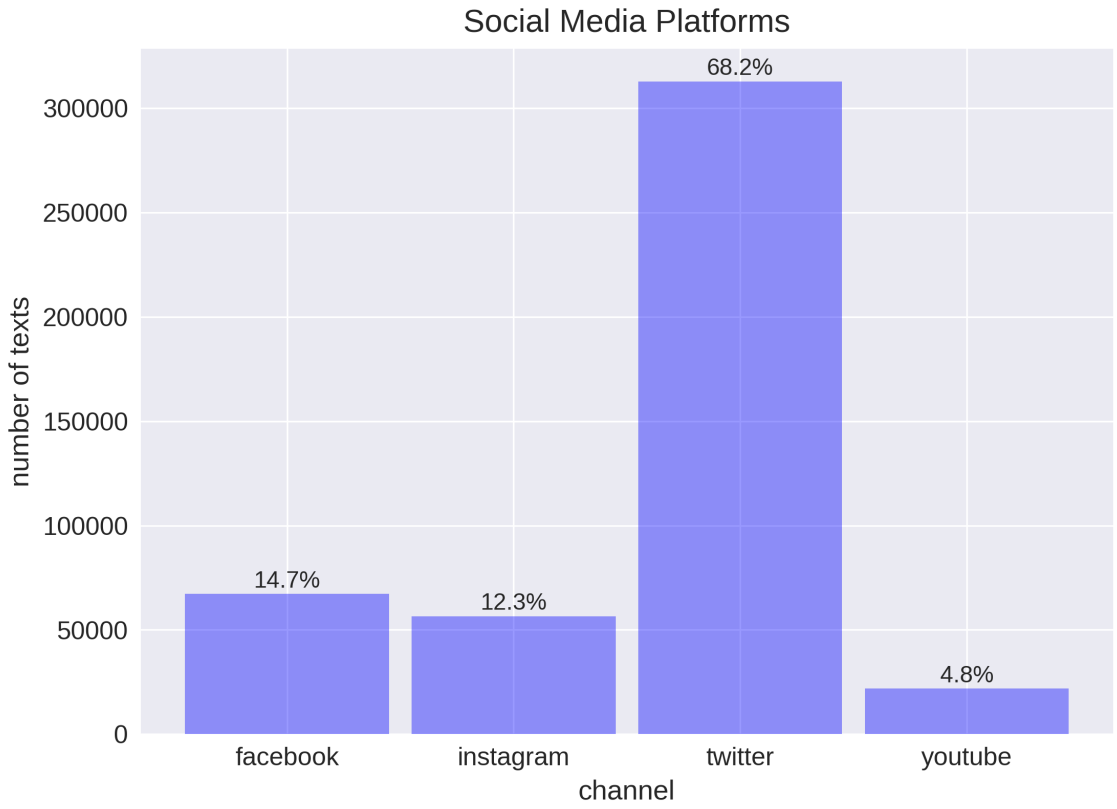
Αν και η ΕΦΓ και συγκεκριμένα η ανάλυση συναισθήματος αποτελεί ένα δημοφιλές πεδίο έρευνας τα τελευταία χρόνια, οι διατιθέμενοι πόροι για τις φυσικές γλώσσες που δεν είναι ιδιαίτερα διαδεδομένες σε παγκόσμια κλίμακα, όπως τα ελληνικά, είναι περιορισμένοι. Ωστόσο, σχετικές πρόσφατες μελέτες [7, 26, 15, 6] έχουν αρχίσει να επιφέρουν νέο ενδιαφέρον στον τομέα, καλύπτοντας ένα ευρύ φάσμα από εφαρμογές και τεχνικές ανάπτυξης γλωσσικών

μοντέλων. Το ερευνητικό τους πεδίο συγκεκριμένα εστιάζει στην υλοποίηση και εκπαίδευση μοντέλων ΕΦΓ αλλά και την προσπάθεια επέκτασης και εξέλιξης των ήδη διαδεδομένων γλωσσικών μοντέλων. Παράλληλα, γίνονται προσπάθειες ενίσχυσης του διαθέσιμου υλικού που απαιτείται για την αναβάθμιση του συγκεκριμένου τομέα, μέσω της συλλογής ικανών σωμάτων κειμένων αλλά και της κατάλληλης προσαρμογής τους προκειμένου να είναι συμβατά με εργασίες ΕΦΓ, συνήθως επιβλεπόμενης μάθησης, για το οποίο προορίζονται (π.χ επισημείωση συνόλου δεδομένων για ανάλυση δεδομένων).

4.3 Συλλογή και Προεπεξεργασία Δεδομένων

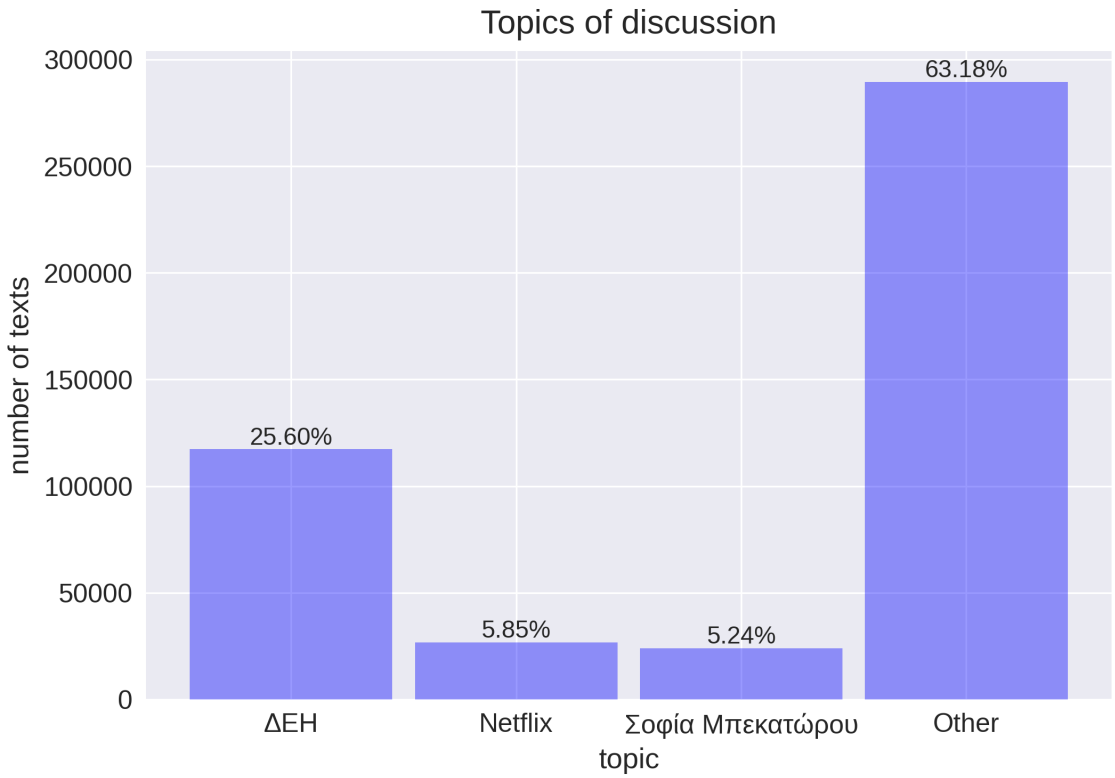
4.3.1 Σώμα κειμένου προ-εκπαίδευσης

Η επιλογή πηγών αλλά και ο όγκος του σώματος κειμένου πάνω στο οποίο θα γίνει η προεκπαίδευση του γλωσσικού μοντέλου, αποτελεί καθοριστικό παράγοντα για την ικανότητά του τόσο να αναπαραστήσει επιτυχώς μια φυσική γλώσσα, όσο και την αποδοτικότητά του στις διάφορες εργασίες ΕΦΓ που θα γίνει η προσαρμογή του. Όσον αφορά το εγχείρημα της διπλωματικής αυτής εργασίας, η ελληνική γλώσσα σαν αντικείμενο μελέτης επιφέρει και κάποιους περιορισμούς στη συγκέντρωση ενός αντιπροσωπευτικού συνόλου δεδομένων. Η δυσκολία αυτή προκύπτει καθώς οι διαθέσιμες πηγές ελληνικών κειμένων υστερούν σημαντικά, τόσο ποσοτικά, όσο και ποιοτικά σε σχέση με άλλες δημοφιλείς φυσικές γλώσσες όπως τα Αγγλικά και τα Ισπανικά. Σαφώς, η πρόκληση αυτή εντείνεται ακόμα περισσότερο στα μέσα κοινωνικής δικτύωσης, όπου η χρήση της γλώσσας είναι πολύ πιο ελεύθερη και ανεπίσημη. Για τον σκοπό αυτό, λαμβάνοντας υπόψη τις ιδιαιτερότητες που παρουσιάζονται στα συγκεκριμένα κείμενα, επιλέχθηκε η προ-εκπαίδευση του γλωσσικού μας μοντέλου να γίνει στοχευμένα πάνω σε ένα σώμα κειμένου από πηγές μέσων κοινωνικής δικτύωσης. Η κατασκευή και εκπαίδευση του γλωσσικού μοντέλου βασίστηκε στη συλλογή πληροφορίας από διάφορες πλατφόρμες κοινωνικής δικτύωσης (Twitter, Instagram, Facebook κλπ.), από την εταιρεία ΠΑΛΟ ΨΗΦΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Ε.Π.Ε. [1], η κατανομή των οποίων φαίνεται αναλυτικά στο Σχήμα 4.1



Σχήμα 4.1: Κατανομή δεδομένων ανά πλατφόρμα κοινωνικής δικτύωσης

Συγκεκριμένα, χρησιμοποιήθηκε ένα σώμα κειμένου 458,293 εγγραφών στην ελληνική γλώσσα το οποίο συλλέχθηκε από τον Απρίλιο του 2019 ως τον Απρίλιο του 2021 και από το οποίο έχουν αφαιρεθεί οι διπλότυπες εισαγωγές. Η κάθε εγγραφή που ανήκει στο σύνολο δεδομένων που μας παρέχεται, χαρακτηρίζεται από μοναδική ταυτότητα (id), την πηγή πληροφορίας (channel), την ημερομηνία δημιουργίας της (createdate), το ερώτημα (query_name) και το κείμενό της (text). Τα θέματα (topics) που πραγματεύεται το κάθε σχόλιο ποικίλλουν και μπορούν να διακριθούν σε 677 διαφορετικές κατηγορίες, οι οποίες καλύπτουν ένα ευρύ φάσμα συζητήσεων γύρω από υπηρεσίες (π.χ ΔΕΗ, Cosmote), brands (π.χ Dettol, BMW, Nivea), πρόσωπα (π.χ Σοφία Μπεκατώρου) κ.ά. Αν και υπάρχουν πολλά θέματα συζήτησης στη συλλογή κειμένων που διαθέτουμε, ένα σημαντικό ποσοστό αυτών καλύπτεται από 3 κυρίαρχες θεματικές: τη ΔΕΗ (117,318 εγγραφές), το Netflix (26,804 εγγραφές και τη Σοφία Μπεκατώρου (24,004 εγγραφές). Σημειώνεται πως η αμέσως επόμενη θεματική συγκεντρώνει μόλις 9,737 σχόλια. Αναλυτική κατανομή των θεματικών αναπαρίσταται στο Σχήμα 4.2.



Σχήμα 4.2: Θέματα συζήτησης στις εγγραφές της συλλογής

Προεπεξεργασία δεδομένων

Δεδομένου ότι οι πηγές των κειμένων που επεξεργαζόμαστε είναι τα μέσα κοινωνικής δικτύωσης, συναντάμε σε μεγάλη συχνότητα διάφορα χαρακτηριστικά τα οποία χρειάζονται τροποποιήσεις. Αναλυτικότερα, πριν εκπαιδεύσουμε το μοντέλο μας, αφού έγινε απαλοιφή των στοιχείων με κενά σχόλια, επιλέξαμε να αφαιρέσουμε από το εκάστοτε κείμενο ορισμένα περιττά χαρακτηριστικά, τα οποία πέρα από το γεγονός ότι δεν προσδίδουν κάποιο συναισθηματικό τόνο, είναι πολύ πιθανό να αποπροσανατολίσουν την επεξεργασία του. Συγκεκριμένα, η απαλοιφή αφορά τα παρακάτω γνωρίσματα

1. Ηλεκτρονικές διευθύνσεις - URLs
2. Emoticons
3. Hashtags (#)
4. Retweet χαρακτήρες (RT)
5. Αναφορές λογαριασμού - Mentions (@)

Επιπλέον αντικαταστάθηκαν τα τονισμένα φωνήεντα των λέξεων με τους αντίστοιχους μη τονισμένους χαρακτήρες και διαγράφηκαν οι περιττοί χαρακτήρες κενού που προϋπήρχαν ή

προέκυψαν μετά την επεξεργασία που πραγματοποιήσαμε. Εκτός από τη βασική προεπεξεργασία που αναφέρθηκε παραπάνω, ακολουθήσαμε τέσσερις (4) διαφορετικές προσεγγίσεις στην επεξεργασία κειμένου, οι οποίες αφορούν τους παρακάτω συνδυασμούς

p_1 Αφαίρεση διαδοχικά επαναλαμβανόμενων σημείων στίξης και μετατροπή κεφαλαίων χαρακτήρων σε πεζούς

p_2 Διαγραφή σημείων στίξης και μετατροπή κεφαλαίων χαρακτήρων σε πεζούς

p_3 Διαγραφή σημείων στίξης και διατήρηση κεφαλαίων χαρακτήρων

p_4 Αφαίρεση διαδοχικά επαναλαμβανόμενων σημείων στίξης και διατήρηση κεφαλαίων χαρακτήρων

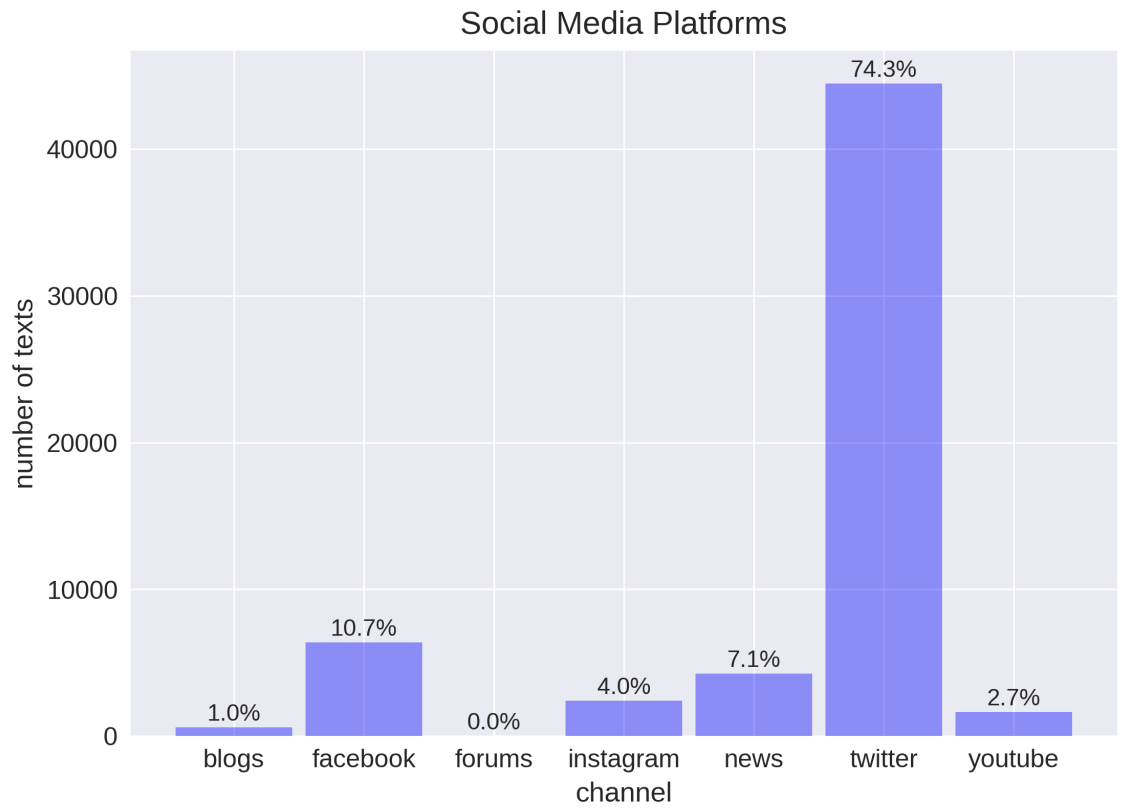
Μετά τις τροποποιήσεις αυτές, το σύνολο των εγγράφων με το οποίο τελικά θα εκπαιδεύσουμε το μοντέλο μας διαθέτει $456,980$ εγγραφές για τις περιπτώσεις p_1, p_4 και $456,926$ για τις περιπτώσεις p_2, p_3. Η μείωση αυτή στο σύνολο των δεδομένων μας, είναι αποτέλεσμα της διαγραφής των περιττών χαρακτηριστικών του κειμένου που συναντήσαμε (URLs, emoticons κλπ.) όπως αναφέρεται παραπάνω. Παρατηρήθηκε ότι υπήρχαν πάνω από $1,300$ tweets που αποτελούνταν μόνο από αυτού του είδους τα γνωρίσματα (π.χ το περιεχόμενο ενός tweet ήταν ένα URL) και το πεδίο 'text' μετά την προεπεξεργασία που πραγματοποιήθηκε έμεινε κενό με αποτέλεσμα να διαγραφεί από το συνολικό corpus.

GreekBERT

Το γλωσσικό αυτό μοντέλο [17] έχει εκπαιδευτεί σε έναν μεγάλο όγκο δεδομένων που δεν διαθέτει τις συντακτικές και σημασιολογικές ιδιομορφίες των κειμένων που εμφανίζονται στα μέσα κοινωνικής δικτύωσης. Ειδικότερα, το σώμα κειμένου της προ-εκπαίδευσής του, περιλαμβάνει πηγές της Βικιπαίδειας, το ελληνικό κομμάτι του OSCAR [4], αλλά και πηγές προερχόμενες από ελληνικές μεταφράσεις διαδικαστικών εγγράφων του Ευρωπαϊκού Κοινοβουλίου. Όσον αφορά τα μορφολογικά του χαρακτηριστικά, στο σώμα κειμένου δεν υπάρχουν τονισμένα φωνήεντα ούτε κεφαλαίοι χαρακτήρες.

4.3.2 Δεδομένα Ανάλυσης Συναισθήματος

Προκειμένου να αξιολογήσουμε την απόδοση του μοντέλου μας πάνω στο πεδίο της ανάλυσης συναισθήματος, συγκεντρώθηκαν σχόλια και τωεετς από τα μέσα κοινωνικής δικτύωσης, γραμμένα στην ελληνική γλώσσα, των οποίων το περιεχόμενο εξετάστηκε ως προς την συναισθηματική του φόρτιση και κατατάχθηκε αναλόγως σε μια από τις τρεις κατηγορίες (ουδέτερο, θετικό, αρνητικό). Η διαδικασία αυτή της συλλογής και της κατηγοριοποίησης των δεδομένων έγινε, όπως και στην περίπτωση της συλλογής δεδομένων που αναλύθηκε παραπάνω, από την ΠΑΛΟ ΨΗΦΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Ε.Π.Ε. [1]. Η συλλογή δεδομένων που διαθέτουμε, στην οποία έχει προηγηθεί απαλοιφή των διπλότυπων στοιχείων, αποτελείται από $59,815$ δείγματα, προερχόμενα από διάφορες πλατφόρμες κοινωνικής δικτύωσης, οι πηγές των οποίων φαίνονται αναλυτικά στο Σχήμα 4.3

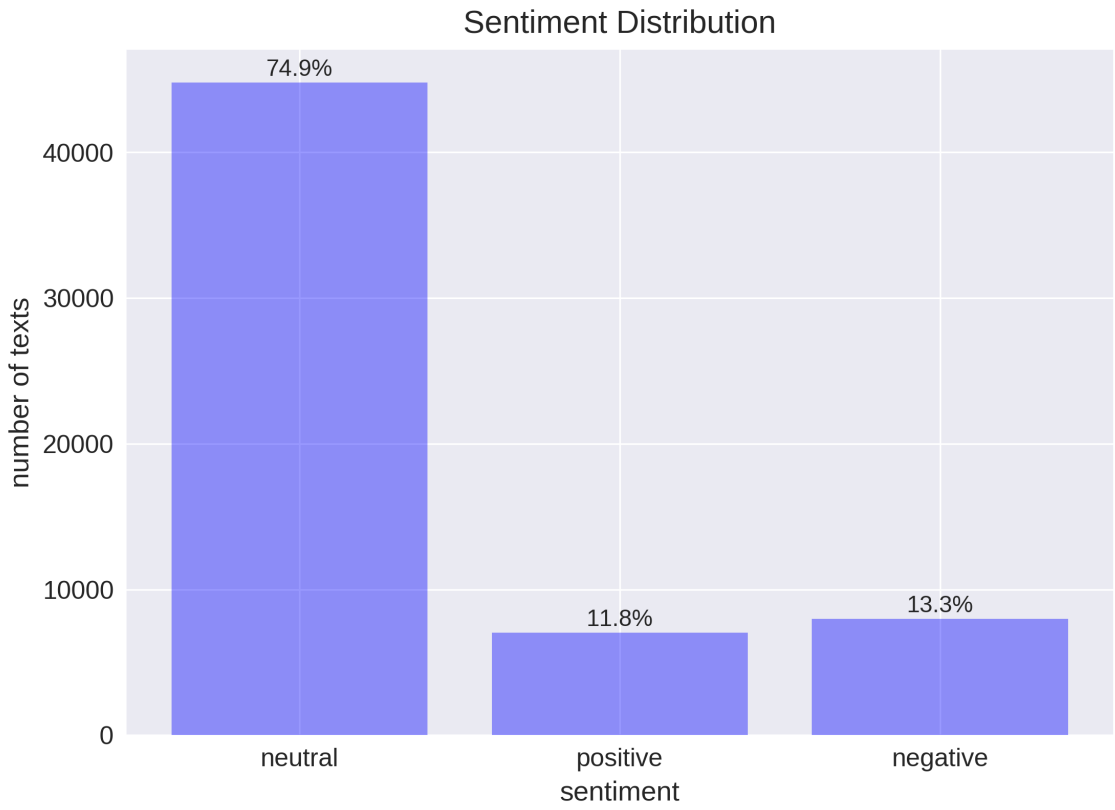


Σχήμα 4.3: Κατανομή δεδομένων ανά πλατφόρμα κοινωνικής δικτύωσης

Για το κάθε σχόλιο/tweet που συναντάμε στη συλλογή δεδομένων, μας παρέχονται οι εξής πληροφορίες: μια μοναδική ταυτότητα (id), το ερώτημα (query_name), η πηγή της πληροφορίας (channel) και το συναίσθημα που αποτυπώνεται (sentiment). Αναλυτικότερα, το σύνολο των δεδομένων μας, όπως ήταν αναμενόμενο, διαθέτει μια ισχυρή πλειοψηφία ουδέτερου περιεχομένου, ενώ οι θετικές και οι αρνητικές γνώμες ισομοιράζονται και καταλαμβάνουν αθροιστικά περίπου το 25% της συλλογής. Τα χαρακτηριστικά της κατανομής συναισθήματος φαίνονται παρακάτω, στην γενική της απεικόνισή του Σχήματος 4.4 και στον πίνακα πολικότητας ανά πλατφόρμα κοινωνικής δικτύωσης (Πίνακας 4.1). Σημειώνεται πως η συλλογή δεδομένων καλύπτει πάνω από 500 θέματα συζήτησης για brands, υπηρεσίες, πρόσωπα κ.α.

Πηγή	Ουδέτερο	Θετικό	Αρνητικό
Blogs	433	79	99
Facebook	4,575	1,271	533
Forums	14	1	3
Instagram	2,128	257	23
News	3,201	697	372
Twitter	33,297	4,295	6,869
YouTube	1,143	442	58
Σύνολο	44,791	7,042	7,977

Πίνακας 4.1: Κατανομή συναισθήματος ανά πλατφόρμα κοινωνικής δικτύωσης



Σχήμα 4.4: Κατανομή συναισθήματος

Προεπεξεργασία δεδομένων

Πρώτο βήμα στην επεξεργασία των δεδομένων μας, είναι η απαλοιφή των στοιχείων με κενά σχόλια/tweets (null text) καθώς και όσων δεν είχαν κατηγοριοποιηθεί ανάλογα με το συναίσθημα που αποτυπώνουν (null sentiment). Το μέγεθος λοιπόν του συνόλου των επισημειωμένων δεδομένων μειώθηκε από 59,815 σε 59,810. Όσον αφορά την προεπεξεργασία που πραγματοποιήθηκε στα κείμενα (σχόλια/tweets) της συλλογής, ακολουθήσαμε πανομοιότυπη

διαδικασία με αυτή της τροποποίησης των κειμένων για το γλωσσικό μας μοντέλο, η οποία αναφέρθηκε στην Ενότητα 4.3.1. Κατ΄ αυτόν τον τρόπο, αφού έγινε διαγραφή των περιττών χαρακτηριστικών (URLs, retweets (RT), mentions (@), hashtags (#), whitespaces (‘ ‘), emoticons) και αντικαταστάθηκαν με τους αντίστοιχους μη τονισμένους χαρακτήρες τους τα φωνήεντα των λέξεων, έγινε μετατροπή των συμβολοσειρών που περιγράφουν το συναίσθημα των δεδομένων (‘neutral’, ‘positive’, ‘negative’) σε αριθμητικούς χαρακτήρες. Τελικά, οι ετικέτες που αντιστοιχούν στο κάθε συναίσθημα είναι

$$\text{‘neutral’} \rightarrow 0, \text{‘positive’} \rightarrow 1, \text{‘negative’} \rightarrow 2$$

Προκειμένου να γίνει σωστή εκπαίδευση του γλωσσικού μοντέλου πάνω στην επισημειωμένη συλλογή δεδομένων που διαθέτουμε, προσαρμόσαμε τα δείγματα αναλόγως, έτσι ώστε να έχουν τα ίδια γνωρίσματα με αυτά του προ-εκπαιδευμένου μοντέλου πάνω στο οποίο θα γίνει η εκπαίδευση. Ακολουθήσαμε επομένως την ίδια διαδικασία που αναλύσαμε προηγουμένως στην Ενότητα 4.3.1 και τροποποιήσαμε τη συλλογή δεδομένων βάσει των προσεγγίσεων p_1, p_2, p_3 και p_4, εφαρμόζοντας τους αντίστοιχους συνδυασμούς διαγραφής σημείων στίξης και αντικατάστασης κεφαλαίων χαρακτήρων, όπου απαιτούνταν. Για την περίπτωση του GreekBERT, το σύνολο δεδομένων ταξινόμησης τροποποιήθηκε κατάλληλα, σύμφωνα με τις συστάσεις των δημιουργών. Συγκεκριμένα εφαρμόστηκε η επεξεργασία που αναλογεί στην p_1 προσέγγιση, αφού το μοντέλο δεν διαθέτει τονισμένα φωνήεντα και κεφαλαίους χαρακτήρες στο σώμα κειμένου του.

4.4 Υλοποίηση πειραμάτων

4.4.1 Ανάλυση σε σύμβολα

Η διαδικασία της ανάλυσης σε σύμβολα (tokenization) ενός σώματος κειμένου αποτελεί βασική προϋπόθεση για την υλοποίηση ενός αποδοτικού γλωσσικού μοντέλου, αφού η δημιουργία ενός αντιπροσωπευτικού και ολοκληρωμένου λεξικού επηρεάζει άμεσα την ικανότητα γενίκευσης του μοντέλου, δηλαδή την δυνατότητά του να προσαρμόζεται επιτυχώς σε νέα δεδομένα, τα οποία δεν έχει συναντήσει κατά την εκπαίδευσή του. Λαμβάνοντας υπόψη λοιπόν την σχετική έλλειψη υλικού στην ελληνική γλώσσα, αλλά και τα πλούσια μορφολογικά χαρακτηριστικά της, προκειμένου το τελικό λεξικό να αντιστοιχίζεται όσο το δυνατόν περισσότερο με τη φυσική γλώσσα, έγινε η επιλογή του ByteLevel BPE Tokenizer [29]. Η λειτουργία του συγκεκριμένου αναλυτή συμβόλων βασίζεται στην τεχνική του BytePair Encoding (BPE), ενός δημοφιλούς αλγόριθμου κωδικοποίησης, ο οποίος συναντάται συχνά σε διάφορα μοντέλα ΕΦΓ, όπως ο μετασχηματιστής BERT, το μοντέλο GPT-2 αλλά και συγκεκριμένα στην επιλεγμένη για την παρούσα εργασία αρχιτεκτονική, RoBERTa. Μέσω της BPE τεχνικής, οι λέξεις διασπώνται σε επιμέρους μικρότερους συνθετικούς όρους, οι οποίοι μπορεί να αντιστοιχούν σε ολόκληρες λέξεις, σε προθέματα, καταλήξεις ή ακόμα και σε ένα σύνολο διαδοχικών χαρακτήρων. Σε αντίθεση με άλλες μεθόδους, η συγκεκριμένη προσέγγιση εξετάζει τη συχνότητα με την οποία ένα ζεύγος από διαδοχικά bytes εμφανίζεται στο λεξικό και στη

συνέχεια τα συγχωνεύει αναλόγως σε ένα νέο σύμβολο. Η διαδικασία αυτή εξασφαλίζει σε μεγάλο βαθμό την αναπαράσταση συχνών λέξεων ως αυτοτελή σύμβολα, ενώ οι σπανιότερες λέξεις ενδέχεται να διασπαστούν σε επιμέρους κομμάτια. Ο ByteLevel BPE αναλυτής λοιπόν, ακολουθώντας την παραπάνω λογική, αρχικοποιεί ένα λεξικό ατομικών bytes, τα οποία ως επί το πλείστων αντιστοιχούν σε έναν χαρακτήρα της Unicode κωδικοποίησης. Στη συνέχεια, γίνεται επανειλημμένη συγχώνευση των συχνότερων ζευγών από bytes σε ένα νέο σύμβολο κ.ο.κ, έως ότου το λεξικό φτάσει ένα προκαθορισμένο μέγεθος. Η μεθοδολογία αυτή, καθιστά τον αναλυτή ικανό να διαχειριστεί λέξεις ακόμα και όταν αυτές δεν περιέχονται στο υπάρχον σώμα κειμένου (OOV), και συνεπώς αποτρέπει την ύπαρξη άγνωστων συμβόλων ([UNK]), μια ιδιαίτερα χρήσιμη ιδιότητα για την συγκεκριμένη μελέτη, δεδομένων των περιορισμένων πηγών ελληνικών κειμένων στο Διαδίκτυο. Για το διαθέσιμο σύνολο δεδομένων στην παρούσα διπλωματική εργασία, επιλέχθηκε το προτεινόμενο μέγεθος $30,522$ συμβόλων, ενώ η ελάχιστη συχνότητα ενός συμβόλου προκειμένου να συμπεριληφθεί στο τελικό λεξικό ορίστηκε στο 2. Στα πλαίσια της πειραματικής διαδικασίας, εξετάστηκε και η ελάχιστη συχνότητα να είναι ακόμα πιο αυστηρή, αυξάνοντας το κατώφλι από 2 σε 3, δεδομένου ότι τη χρήση της γλώσσας στα μέσα κοινωνικής δικτύωσης πολλές φορές περιλαμβάνει ορθογραφικά λάθη, συντομεύσεις λέξεων κ.ά. Ωστόσο δεν παρατηρήθηκε κάποια ιδιαίτερη αλλαγή στην επίδοση των μοντέλων.

4.4.2 Προ-εκπαίδευση

Μοντελοποίηση Απόκρυψης Κειμένου

Στο πρώτο στάδιο της ανάπτυξης του γλωσσικού μοντέλου, σημαντική τεχνική εκπαίδευσης μοντέλων βαθιάς μάθησης είναι αυτή της προ-εκπαίδευσης. Η διαδικασία αυτή, αφορά την μη-επιβλεπόμενη εκπαίδευση του δικτύου σε ένα μεγάλο σύνολο δεδομένων. Κατ' αυτόν τον τρόπο, το δίκτυο, έχει τη δυνατότητα να μάθει να αναπαριστά τις διάφορες σημασιολογικές και συντακτικές ιδιαιτερότητες μιας φυσικής γλώσσας, εκμεταλλευόμενο τον μεγάλο όγκο πληροφορίας που του παρέχεται. Απώτερος σκοπός της πρακτικής αυτής, είναι η ευκολότερη και αποτελεσματικότερη προσαρμογή του σε κάποια συγκεκριμένη εφαρμογή ΕΦΓ. Συγκεκριμένα, στα πλαίσια της προ-εκπαίδευσης δικτύων μετασχηματιστών, τα γλωσσικά μοντέλα εκπαιδεύονται να προβλέπουν σωστά τυχαία «κρυμμένες» λέξεις μέσα σε ένα κείμενο, οι οποίες έχουν αντικατασταθεί από ένα ειδικό σύμβολο αναφοράς, όπως το [MASK]. Η τεχνική αυτή είναι γνωστή ως μοντελοποίηση απόκρυψης κειμένου και χρησιμοποιείται ευρέως σε δημοφιλή μοντέλα μετασχηματιστών, επιφέροντας σημαντικά αποτελέσματα όσον αφορά την αποδοτική εκμάθηση αντιπροσωπευτικών γλωσσικών αναπαραστάσεων και επιτρέποντας στο μοντέλο να αποκτήσει καλύτερη κατανόηση της σημασιολογίας των λέξεων και φράσεων που αντιστοιχούν σε μία γλώσσα. Όσο μεγαλύτερο και σωστά δομημένο είναι το σώμα κειμένου, τόσο αυξάνεται και η πιθανότητα το μοντέλο να πραγματοποιήσει σωστή πρόβλεψη.

Αρχιτεκτονική

Όπως αναφέρθηκε και παραπάνω, η αρχιτεκτονική στην οποία βασίστηκε η υλοποίηση του γλωσσικού μοντέλου είναι αυτή της RoBERTa (Ενότητα 2.3.3). Πρόκειται για μια δομή 6

κρυφών επιπέδων (σε αντίθεση με τα 12 επίπεδα του αρχικού μετασχηματιστή RoBERTa), με 12 κεφαλές προσοχής το κάθε ένα. Η σχεδιαστική επιλογή για μείωση του αριθμού των επιπέδων, αποκλίνοντας από την αρχική αρχιτεκτονική, αφορά τόσο τις υψηλές υπολογιστικές απαιτήσεις, όσο και τον σχετικά περιορισμένο όγκο δεδομένων του διαθέσιμου σώματος κειμένου

Πειραματικές προσεγγίσεις

Τα ιδιόμορφα γνωρίσματα που συναντώνται στα κείμενα των μέσων κοινωνικής δικτύωσης, είναι ικανά να επηρεάσουν την ομαλή και αποτελεσματική εκπαίδευση του γλωσσικού μοντέλου. Για αυτό άλλωστε, όπως αναφέρεται και στην Ενότητα 4.3.1, η προ-εκπαίδευση περιλαμβάνει τις 4 διαφορετικές πειραματικές προσεγγίσεις, p_1, p_2, p_3, p_4 , οι οποίες τροποποιούν τα γνωρίσματα του σώματος κειμένου αναλόγως.

Παράμετροι εκπαίδευσης

Για κάθε ένα από τα παραπάνω γλωσσικά μοντέλα απόκρυψης, η διαδικασία εκπαίδευσης έγινε με ορισμένο μέγεθος δέσμης (batch size) 12, το μεγαλύτερο δυνατό μέγεθος που μπορούσε να εφαρμοστεί, λόγω του όγκου του σώματος κειμένου αλλά και των διαθέσιμων υπολογιστικών πόρων. Η επιλογή της αναλογίας συνόλου εκπαίδευσης και συνόλου επαλήθευσης είναι 90% προς 10%. Συνολικά η εκπαίδευση διαρκεί 45 εποχές, καθώς στη συνέχεια παρατηρούνται φαινόμενα υπερπροσαρμογής. Για τον λόγο αυτό, εφαρμόστηκε πρόωρος τερματισμός της εκπαίδευσης, προκειμένου να μην αλλοιωθεί η ικανότητα γενίκευσης του μοντέλου σε νέα, άγνωστα για εκείνο, δεδομένα, εξαιτίας της μεγάλης εξοικείωσης του μοντέλου με το σύνολο εκπαίδευσης.

4.4.3 Εφαρμογή Ανάλυσης Συναισθήματος

Λεπτή προσαρμογή

Η τεχνική της λεπτής προσαρμογής (fine-tuning), αφορά την προσαρμογή ενός προεκπαιδευμένου γλωσσικού μοντέλου προκειμένου να εκτελέσει μια εργασία ΕΦΓ. Η διαδικασία αυτή, περιλαμβάνει την περαιτέρω εκπαίδευση του μοντέλου αυτού πάνω σε ένα μικρότερο σύνολο δεδομένων, του οποίο όμως το περιεχόμενο είναι, ιδανικά, πιο στοχευμένο και εξειδικευμένο όσον αφορά το αντικείμενο έρευνας της συγκεκριμένης εργασίας. Το επίπεδο ομοιότητας του πεδίου έρευνας του νέου προσαρμοσμένου μοντέλου με αυτό του προεκπαιδευμένου, καθώς και τα ποιοτικά και ποσοτικά χαρακτηριστικά του συνόλου δεδομένων πάνω στο οποίο πραγματοποιείται η λεπτή προσαρμογή, μπορούν να επηρεάσουν σημαντικά τόσο την αποδοτικότητα της τεχνικής αυτής, όσο και τις απαιτήσεις του χρόνου εκπαίδευσης. Προκειμένου να μελετηθεί το αντικείμενο αυτής της διπλωματικής εργασίας, έγινε προσαρμογή των προ-εκπαιδευμένων μοντέλων της παρούσας ενότητας, έτσι ώστε να αξιολογηθεί η απόδοσή τους στην ανάλυση συναισθήματος. Πραγματοποιήθηκε επομένως περαιτέρω εκπαίδευση των μοντέλων, αυτή τη φορά χρησιμοποιώντας το επισημειωμένο σύνολο δεδομένων της

Ενότητας 4.3.2, με σκοπό τη δημιουργία ενός αποτελεσματικού ταξινομητή συναισθημάτων. Κατά την διαδικασία εκπαίδευσης των γλωσσικών μοντέλων ανάλυσης συναισθήματος, αξιολογήθηκε μια σειρά από πειραματικούς συνδυασμούς μοντελοποίησης, οι οποίοι αφορούν την αρχιτεκτονική δομή του μοντέλου, την προσαρμογή των υπερ-παραμέτρων του, αλλά και την ειδική διαχείριση του συνόλου δεδομένων πάνω στο οποίο πραγματοποιείται η εκπαίδευση.

Αρχιτεκτονική

Βασικός παράγοντας που επηρεάζει την αποτελεσματικότητα του γλωσσικού μοντέλου στην ανάλυση συναισθήματος είναι η αρχιτεκτονική δομή του. Οι σχεδιαστικές επιλογές και τροποποιήσεις που έγιναν κατά την πειραματική διαδικασία, αφορούν τις παραμέτρους που περιγράφουν τα επίπεδα, και τις συναρτήσεις ενεργοποίησης που συνιστούν τη δομή του ταξινομητή. Ειδικότερα, σε κάθε προ-εκπαιδευμένο γλωσσικό μοντέλο που αναπτύχθηκε, προκειμένου να εφαρμοστεί η τεχνική της ανάλυσης συναισθήματος, προστέθηκε μια «κεφαλή ταξινόμησης» (classification head). Η εν λόγω κεφαλή μετατρέπει τις τιμές των χαρακτηριστικών του γλωσσικού μοντέλου απόκρυψης του σε μια τελική πρόβλεψη κλάσης / κατηγορίας για τα δεδομένα εισόδου. Πιο συγκεκριμένα, αποτελείται ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα, των οποίων η είσοδος λαμβάνεται από την τελική διανυσματική αναπαράσταση του τελευταίου επιπέδου του γλωσσικού μοντέλου (last-hidden-state). Η αναπαράσταση αυτή, όσον αφορά τους μετασχηματιστές, μπορεί να έχει δύο διαφορετικές μορφές· αυτή της πλήρους ακολουθίας εξόδου (sequence output) του τελευταίου επιπέδου του γλωσσικού μοντέλου και αυτή της δειγματοληπτημένης ακολουθίας εξόδου (pooled output) του ίδιου επιπέδου. Στην πρώτη περίπτωση έχουμε την τελική αναπαράσταση της ακολουθίας εισόδου, στην οποία κάθε σύμβολο που της αντιστοιχεί, αντιπροσωπεύεται από ανάλογο ξεχωριστό διάνυσμα που αποτυπώνει τη σημασιολογική του έννοια. Στην δεύτερη περίπτωση, για κάθε ακολουθία εισόδου, η γενική της σημασιολογική έννοια περικλείεται σε μια μοναδική διανυσματική αναπαράσταση. Η συγκεκριμένη αναπαράσταση, χρησιμοποιείται ευρέως σε εργασίες ταξινόμησης αλλά και σε άλλες εφαρμογές λεπτής προσαρμογής στο πεδίο της ΕΦΓ και συγκεκριμένα στα γλωσσικά μοντέλα των μετασχηματιστών.

Η παρούσα πειραματική διαδικασία εξετάζει διάφορες σχεδιαστικές επιλογές του ταξινομητή, όσον αφορά τις συναρτήσεις ενεργοποίησης, το πλήθος των κρυφών επιπέδων αλλά και την επεξεργασία των παραμέτρων τους, προκειμένου να βελτιστοποιηθεί η απόδοσή του όσον αφορά την ορθή κατηγοριοποίηση των δεδομένων, βάσει συναισθήματος. Οι αρχιτεκτονικές της κεφαλής ταξινόμησης που αξιολογήθηκαν ακολουθούν παρακάτω:

1. **Κανένα κρυφό επίπεδο**

   Η συγκεκριμένη αρχιτεκτονική αποτελεί και την προτεινόμενη για την κεφαλή ταξινόμησης του μετασχηματιστή RoBERTa, όπως αυτός είναι υλοποιημένος στο αποθετήριο του HuggingFace. Αποτελείται από ένα πλήρως διασυνδεδεμένο επίπεδο εισόδου 768 νευρώνων, το οποίο τροφοδοτείται με τη δειγματοληπτημένη ακολουθία εξόδου. Πρόκειται για μια απλή αλλά διαδεδομένη δομή ταξινόμησης, πάνω στην οποία βασίστηκε και ο σχεδιασμός των υπόλοιπων πειραματικών αρχιτεκτονικών.

2. **Ένα κρυφό επίπεδο**

   Στη συγκεκριμένη αρχιτεκτονική, προσθέσαμε ένα κρυφό επίπεδο στην κεφαλή ταξινόμησης. Εξετάστηκε η επιπλέον εφαρμογή της ημι-γραμμικής συνάρτησης ενεργοποίησης ReLU, η οποία αποτελεί μια ευρέως διαδεδομένη μη-γραμμική συνάρτηση ενεργοποίησης στα μοντέλα βαθιάς μηχανικής μάθησης και ειδικότερα σε εφαρμογές ταξινόμησης, όπου παρατηρείται έντονα το φαινόμενο των «εξαφανιζόμενων Κλίσεων» (vanishing gradients). Σε πολλές περιπτώσεις, ενδέχεται να παρουσιαστούν περιορισμοί ως προς το βάθος της εκπαίδευσης ενός μοντέλου, καθώς η ικανότητα μετάδοσης του σφάλματος στα αρχικά επίπεδα μειώνεται σημαντικά, με αποτέλεσμα το μοντέλο να αδυνατεί να ενημερώσει επαρκώς τις παραμέτρους του και συνεπώς η εκπαίδευση να γίνεται πολύ αργά, ή ακόμα και να παραμένει στάσιμη. Μια ακόμα τεχνική που εξετάστηκε κατά την πειραματικη διαδικασία, είναι αυτή του επιπέδου κανονικοποίησης (Layer Normalization), το οποίο αφού εφαρμοστεί στην έξοδο του κρυφού επιπέδου, μπορεί να βοηθήσει και αυτό στο πρόβλημα κορεσμού των επιδόσεων του γλωσσικού μοντέλου. Οι παραπάνω δύο προσεγγίσεις δοκιμάστηκαν τόσο συνδυαστικά, όσο και ξεχωριστά.

3. **Δύο κρυφά επίπεδα**

   Σε αυτή την αρχιτεκτονική δομή, εξετάζεται η προσθήκη ενός επιπλέον κρυφού επιπέδου. Ο ταξινομητής κεφαλής, λοιπόν, συνίσταται από δύο κρυφά επίπεδα. Ομοίως με την αμέσως προηγούμενη περίπτωση, κατά την πειραματική διαδικασία δοκιμάστηκε η εφαρμογή της ReLU συνάρτησης ενεργοποίησης, καθώς και κανονικοποίηση των κρυφών επιπέδων.

Βάρη κλάσεων

Όσον αφορά το σύνολο δεδομένων προς ταξινόμηση, αν και παρατηρείται ισορροπία μεταξύ κειμένων με θετική και αρνητική επισημείωση συναισθήματος, το πλήθος τους σε σχέση με αυτό των ουδέτερων δειγμάτων είναι εξαιρετικά δυσανάλογο, όπως φαίνεται και στο γράφημα του Σχήματος 4.4. Δεδομένου ότι η πλειοψηφία των κειμένων αποτυπώνει ουδέτερο συναίσθημα, υπάρχει μεγάλη πιθανότητα η εκπαίδευση του μοντέλου να είναι "προκατειλημμένη" (biased). Ως αποτέλεσμα, ενώ το μοντέλο μαθαίνει να ταξινομεί με ακρίβεια την κυρίαρχη κλάση, η απόδοσή του στα θετικά και αρνητικά δείγματα υστερεί σημαντικά. Προκειμένου να εξασφαλιστεί η ακεραιότητα των προβλέψεων του ταξινομητή, έγινε τροποποίηση της συνάρτησης απώλειας (loss function), έτσι ώστε να λαμβάνει υπόψη την δυσαναλογία των κλάσεων. Συγκεκριμένα, κατά τον υπολογισμό της απώλειας, τα βάρη που εφαρμόζονται σε κάθε μία από τις κλάσεις, είναι αντιστρόφως ανάλογα της συχνότητας της κάθε μίας. Συνεπώς, για την ουδέτερη κλάση, το επιπρόσθετο βάρος που της αντιστοιχεί είναι σημαντικά μικρότερο από αυτό των άλλων δύο. Κατά την εκπαίδευση λοιπόν, το μοντέλο δίνει περισσότερη βαρύτητα στις κλάσεις "μειοψηφίας", με σκοπό να μειώσει την πιθανότητα λανθασμένης ταξινόμησης τους.

Προ-εκπαιδευμένα βάρη

Συχνή και σημαντική πρακτική κατά την προσαρμογή ενός προ-εκπαιδευμένου μοντέλου, είναι αυτή του παγώματος των βαρών του. Η ενέργεια αυτή μπορεί να αφορά είτε όλα τα προ-εκπαιδευμένα βάρη, είτε να περιορίζεται στα αρχικά μόνο επίπεδα της δομής, στα οποία και πραγματοποιείται το βασικό κομμάτι της μοντελοποίησης με απόκρυψη κειμένου. Ειδικότερα, οι προ-εκπαιδευμένες παράμετροι του γλωσσικού μοντέλου δεν ενημερώνονται κατά τη διαδικασία της λεπτής προσαρμογής, έτσι ώστε να πραγματοποιηθεί αποκλειστικά η εκπαίδευση του ταξινομητή πάνω στα νέα δεδομένα. Παγώνοντας τα προ-εκπαιδευμένα βάρη, το μοντέλο μπορεί να επικεντρωθεί στην εκπαίδευση του ταξινομητή, χωρίς να γίνεται εκ νέου προσπάθεια προσαρμογής των ήδη εκπαιδευμένων παραμέτρων, με αποτέλεσμα να αποφεύγεται το ενδεχόμενο της έντονης υπερπροσαρμογής, ειδικά σε περιπτώσεις όπου το σώμα κειμένου της προ-εκπαίδευσης είναι κατά πολύ μεγαλύτερο του νέου συνόλου δεδομένων. Παράλληλα, λόγω της ελαχιστοποίησης των παραμέτρων που χρήζουν εκπαίδευσης, η διάρκεια και συνεπώς το κόστος της διαδικασίας μειώνεται σημαντικά. Δεδομένου της ιδιαίτερης φύσης των δεδομένων μας αλλά του περιορισμένου όγκου δεδομένων για την εφαρμογή της ανάλυσης συναισθήματος, επιλέχθηκε να πραγματοποιηθεί πάγωμα των παραμέτρων του μοντέλου προ-εκπαίδευσης σε κάθε ένα από τα επίπεδα που συνιστούν τη δομή του.

Πειραματικές προσεγγίσεις

Η προσαρμογή ενός προ-εκπαιδευμένου γλωσσικού μοντέλου σε μια νέα εργασία, προϋποθέτει τα γνωρίσματα του νέου συνόλου δεδομένων να συμβαδίζουν με αυτά του βασικού σώματος κειμένου προ-εκπαίδευσης. Επομένως, για να πραγματοποιηθεί η σωστή επεξεργασία των κειμένων από το μοντέλο, έγινε η ανάλογη μετατροπή των μορφολογικών χαρακτηριστικών τους σύμφωνα με τις προσεγγίσεις p_1, p_2, p_3 και p_4. Κατά την πειραματική διαδικασία λοιπόν, πέρα από την γενική απόδοση των ταξινομητών σε ποιοτικό επίπεδο, θα μελετηθεί και η επίδραση των μορφολογικών χαρακτηριστικών στο πεδίο έρευνας της ΕΦΓ. Παράλληλα, για κάθε περίπτωση, θα μελετηθούν και διαφορετικές αρχιτεκτονικές δομές για τον εκάστοτε ταξινομητή συναισθήματος, προκειμένου να εξεταστούν όλα τα περιθώρια βελτίωσής.

Παράμετροι εκπαίδευσης

Η επιλογή αναλογίας συνόλου εκπαίδευσης, συνόλου επαλήθευσης και συνόλου ελέγχου είναι $70 - 15 - 15\%$. Κατά την κατανομή των δεδομένων, προκειμένου να αποτυπωθούν ορθά οι πειραματικές συνθήκες υπό τις οποίες πραγματοποιείται η ταξινόμηση, εξασφαλίστηκε η διατήρηση της αρχικής αναλογίας συναισθημάτων για κάθε σύνολο. Για κάθε ένα από τα παραπάνω γλωσσικά μοντέλα ανάλυσης συναισθήματος, η διαδικασία εκπαίδευσης έγινε με μέγεθος δέσμης 16, ενώ η διάρκεια εκπαίδευσης, όπως φαίνεται και στις απεικονίσεις της ενότητας που ακολουθεί διαφέρει για κάθε πειραματική προσέγγιση.

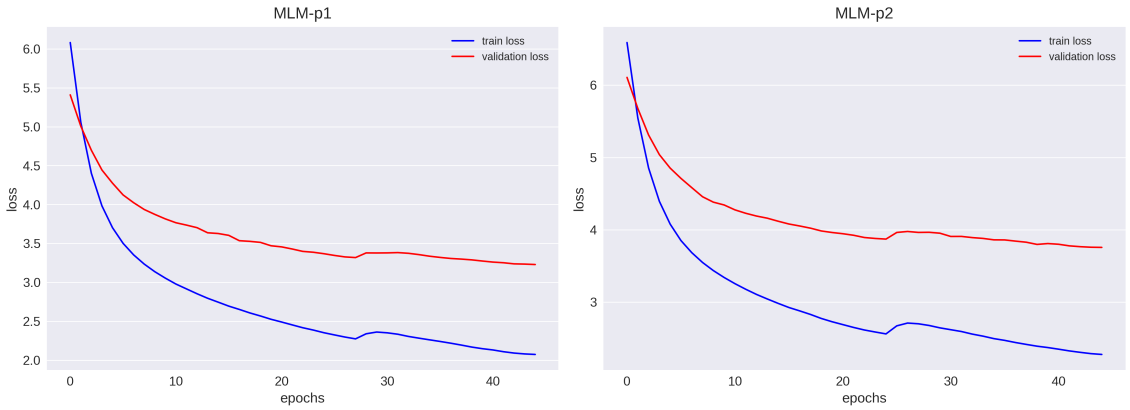
GreekSocialBERT

Το GreekBERT [17] βασίζεται στην αρχιτεκτονική του BERT, με την προ-εκπαίδευσή του να έχει γίνει πάνω σε ένα εξαιρετικά μεγάλο όγκο σώματος κειμένου, το οποίο ωστόσο, δεν περιλαμβάνει πηγές από τα ελληνικά μέσα κοινωνικής δικτύωσης. Προκειμένου να γίνει επιτυχής προσαρμογή του GreekBERT στην εργασία ανάλυσης συναισθήματος, έγινε περαιτέρω εκπαίδευσή του πάνω στο σώμα κειμένου του πειράματός μας, το οποίο τροποποιήθηκε αναλόγως, απαλείφοντας τους τόνους και μετατρέποντας τα κεφαλαία γράμματα σε πεζά (πειραματική προσέγγιση p_1. Η εκπαίδευση έγινε για 4 μόνο εποχές, προκειμένου να αποφευχθούν φαινόμενα υπερπροσαρμογής. Στη συνέχεια, κατά τη διαδικασία της λεπτής προσαρμογής, εφαρμόστηκαν οι αρχιτεκτονικές της Ενότητας 4.4.3, προκειμένου να κατασκευαστεί το τελικό γλωσσικό μοντέλο ανάλυσης συναισθήματος, το GreekSocialBERT [7].

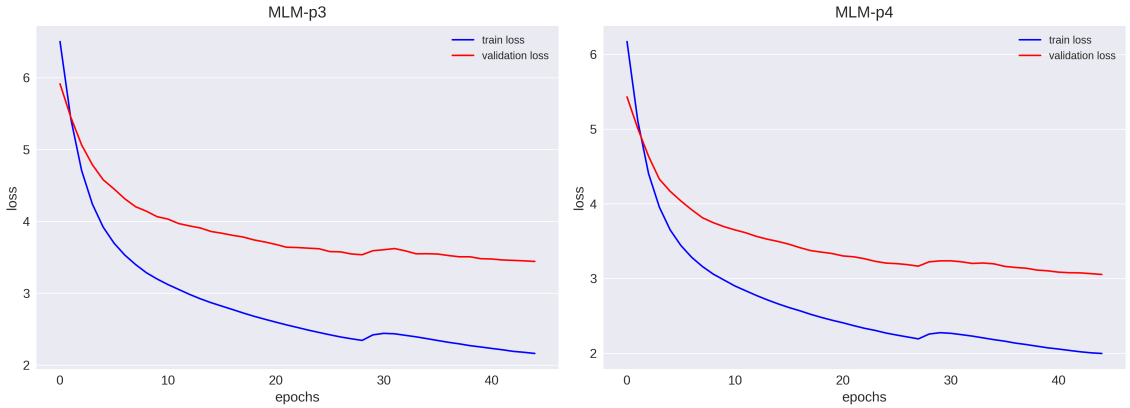
4.5 Αποτελέσματα και Αξιολόγηση

4.5.1 Προ-εκπαιδευμένα Μοντέλα

Στο Σχήμα 4.5 φαίνεται η μεταβολή της συνάρτησης απώλειας στο σύνολο δεδομένων εκπαίδευσης (μπλε γραμμή) και επαλήθευσης (κόκκινη γραμμή), κατά την εκπαίδευση των γλωσσικών μοντέλων απόκρυψης, για τις τέσσερις πειραματικές προσεγγίσεις που εξετάσαμε.



(α′) Πειραματική προσέγγιση p_1 (β′) Πειραματική προσέγγιση p_2



(γ′) Πειραματική προσέγγιση p_3 (δ′) Πειραματική προσέγγιση p_4

Σχήμα 4.5: Μεταβολή συνάρτησης απώλειας εκπαίδευσης (μπλε γραμμή) και επαλήθευσης (κόκκινη γραμμή) στα γλωσσικά μοντέλα απόκρυψης

Από τα παραπάνω αποτελέσματα διαφαίνεται ότι, παρά τις διαφορετικές τους προσεγγίσεις, τα πειράματα φαίνεται να έχουν παρόμοια απόδοση όσον αφορά τη συνάρτηση απώλειας, με τα μοντέλα που διατηρούν τα σημεία στίξης στο λεξικό τους p_1, p_4, να έχουν μια ελάχιστα καλύτερη ανταπόκριση. Βαθύτερη κατανόηση ως προς την απόδοση των γλωσσικών μοντέλων που υλοποιήθηκαν θα έχουμε στην εφαρμογή τους σε εργασίες ΕΦΓ και συγκεκριμένα στην ανάλυση συναισθήματος.

Το σώμα κειμένου αποτελεί βασική προϋπόθεση για την απόδοση του γλωσσικού μας μοντέλου. Συγκεκριμένα για την περίπτωση των δεδομένων που έχουν συλλεχθεί από τα μέσα κοινωνικής δικτύωσης, παρατηρείται πως παρουσιάζουν διάφορα γνωρίσματα που ενδέχεται να επηρεάσουν την εκπαίδευση ενός γλωσσικού μοντέλου. Ο ανεπίσημος τρόπος γραφής των χρηστών σε πολλές περιπτώσεις περιέχει σημασιολογικές αστοχίες, καθώς συχνά τα κείμενα διαθέτουν ορθογραφικά και συντακτικά λάθη, συντομεύσεις λέξεων, δυσνόητες και αυθαίρετες έννοιες, ενώ ιδιαίτερα συχνή είναι και η χρήση της αργκό. Συνεπώς, η μοντελοποίηση των κειμένων και η πιστή αναπαράσταση του νοήματός τους έχει ένα μεγαλύτερο βαθμό δυσκολίας. Τα παραπάνω χαρακτηριστικά συναντώνται σε όλες τις πλατφόρμες κοινωνικής δικτύωσης και σε κάθε φυσική γλώσσα ανεξαιρέτως. Ωστόσο, η περιορισμένη χρήση της ελληνικής γλώσσας

στα μέσα κοινωνικής δικτύωσης καθιστά το έργο της ΕΦΓ ακόμα πιο απαιτητικό για την συλλογή ποιοτικού υλικού. Σημαντικός επίσης παράγοντας για την ικανότητα γενίκευσης και προσαρμογής του μοντέλου σε νέα δεδομένα και εφαρμογές, είναι και η θεματολογία των συζητήσεων που αυτό επεξεργάζεται. Προκειμένου να είναι λοιπόν εφικτή η αποτελεσματική εφαρμογή του γλωσσικού μοντέλου σε πολλές και διαφορετικές εργασίες ΕΦΓ, οι θεματικές που καλύπτει ένα σώμα κειμένου ενδείκνυται να είναι πολλαπλές και ισορροπημένες όσον αφορά την κατανομή του περιεχομένου τους. Στην περίπτωση βέβαια που σκοπός μια εργασίας αφορά ένα πιο εξειδικευμένο αντικείμενο μελέτης, η συγκέντρωση σχετικού συνόλου δεδομένων για την εκπαίδευση του μοντέλου θα μπορέσει να ενισχύσει την απόδοσή του.

Το παρόν σύνολο δεδομένων πάνω στο οποίο έγινε η προ-εκπαίδευση, δεδομένου ότι συλλέχθηκε από τα μέσα κοινωνικής δικτύωσης με σκοπό να εξυπηρετήσει συγκεκριμένα τις εταιρείες-οργανισμούς με τις οποίες συνεργάζεται η ΠΑΛΟ ΨΗΦΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Ε.Π.Ε., αφορά ένα πιο εξειδικευμένο περιεχόμενο. Ειδικότερα, σύμφωνα με το γράφημα του Σχήματος 4.2, οι συζητήσεις που κυριαρχούν στο σώμα κειμένου αφορούν, με σημαντική διαφορά, τις υπηρεσίες της ΔΕΗ, την εταιρεία Netflix και ένα πρόσωπο της επικαιρότητας την περίοδο που συλλέχθηκαν τα δεδομένα, τη Σοφία Μπεκατώρου. Λαμβάνοντας λοιπόν υπόψη τις παραπάνω παρατηρήσεις, κατά την αξιολόγηση του προ-εκπαιδευμένου μοντέλου, αναμένουμε μεγαλύτερη ευστοχία σε παρόμοια θεματικά πλαίσια ή ακόμα και την σύνδεση των πρόβλεψεων με αυτά. Για παράδειγμα, στις παρακάτω δοκιμές αξιολόγησης του προ-εκπαιδευμένου μοντέλου (Σχήμα 4.6, φαίνεται η συνύφανση του θηλυκού γένους με τις επωνυμίες εταιρειών όπως «ΔΕΗ» κλπ. αλλά και με το όνομα «Σοφία Μπεκατώρου».

```
fill(f'η {fill.tokenizer.mask_token} ειναι καθηγητρια')
```

```
[{'score': 0.584001362323761,
  'token': 370,
  'token_str': ' δεη',
  'sequence': 'η δεη ειναι καθηγητρια'},
 {'score': 0.03319411352276802,
  'token': 633,
  'token_str': ' μπεκατωρου',
  'sequence': 'η μπεκατωρου ειναι καθηγητρια'},
 {'score': 0.02136019989848137,
  'token': 1719,
  'token_str': ' αλλη',
  'sequence': 'η αλλη ειναι καθηγητρια'},
 {'score': 0.021027807146310806,
  'token': 7411,
  'token_str': ' αστυνομια',
  'sequence': 'η αστυνομια ειναι καθηγητρια'},
 {'score': 0.019276263192296028,
  'token': 3012,
  'token_str': ' γυναικα',
  'sequence': 'η γυναικα ειναι καθηγητρια'}]
```

Σχήμα 4.6: Παραδείγματα δοκιμών αξιολόγησης του προ-εκπαιδευμένου μοντέλου

Επομένως, επιβεβαιώνεται και πειραματικά πως η απόδοση ενός γλωσσικού μοντέλου είναι άμεσα εξαρτώμενη από τα ποιοτικά χαρακτηριστικά του σώματος κειμένου. Το μοντέλο που αναπτύχθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας, αν και σε ορισμένες περιπτώσεις παρουσιάζει μια μικρή δυσκολία πρόβλεψης, φαίνεται να ανταποκρίνεται με υψηλή απόδοση, όταν χρησιμοποιείται σε περιπτώσεις που ταιριάζουν με τους σκοπούς για τους οποίους συγκεντρώθηκε το παρόν σύνολο δεδομένων.

4.5.2 Μοντέλα Ανάλυσης Συναισθήματος

Στην παρούσα μελέτη, εξετάστηκε ένα ευρύ πεδίο παραμέτρων που επηρεάζουν την απόδοση ενός μοντέλου ταξινόμησης συναισθήματος. Τα αποτελέσματα που επέφεραν οι πειραματικές προσεγγίσεις που περιγράφηκαν στην Ενότητα 4.4.3, αξιολογήθηκαν σε ποσοτικό αλλά και ποιοτικό επίπεδο, λαμβάνοντας υπόψη τους εξής παράγοντες:

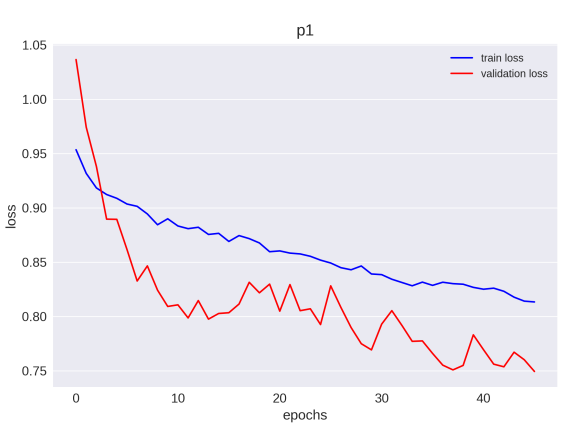
- Απώλειας

  Ο υπολογισμός της απώλειας ορίζεται στα δεδομένα εκπαίδευσης (απώλεια εκπαίδευσης) και στα δεδομένα επαλήθευσης (απώλεια επαλήθευσης). Αξιοποιώντας τις δύο αυτές συνιστώσες, μπορούμε, συνήθως, να λάβουμε μια αντιπροσωπευτική εικόνα σχετικά με την απόδοση του μοντέλου. Στόχος του κάθε πειράματος ήταν η ελαχιστοποίηση τόσο και των δύο προαναφερόμενων απωλειών, έτσι ώστε να μην εμφανιστούν φαινόμενα υπερπροσαρμογής κατά τη διάρκεια της εκπαίδευσης

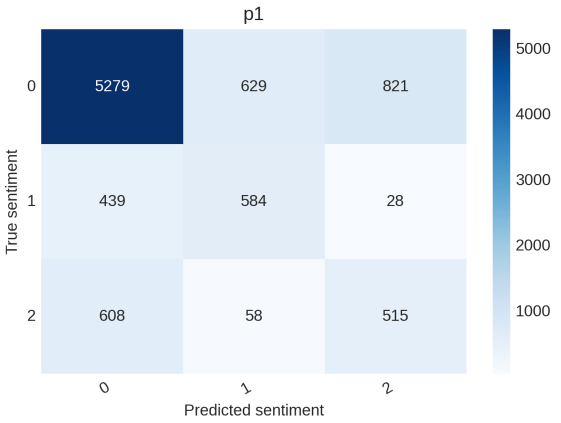
- Μετρικές Αξιολόγησης

  Οι μετρικές αξιολόγησης εξετάζουν την ικανότητα του μοντέλου να κατηγοριοποιήσει σωστά νέα δεδομένα. Συγκεκριμένα, εστιάζουμε στα αποτελέσματα που επιφέρει ο ταξινομητής όσον αφορά την ακρίβεια (Precision), την ανάκληση Recall και τον αρμονικό μέσο όρο των δύο, δηλαδή το F1-score, αλλά και τον πίνακα σύγχυσης (Confusion Matrix) που προκύπτει από την αξιολόγησή του. Σημειώνεται ότι τα δεδομένα ελέγχου, πάνω στα οποία πραγματοποιείται η διαδικασία αυτή προκειμένου να εξασφαλιστεί η ακεραιότητα του κάθε πειράματος, διατηρούν την αναλογία κατανομής συναισθημάτων της συνολικής συλλογής δεδομένων.

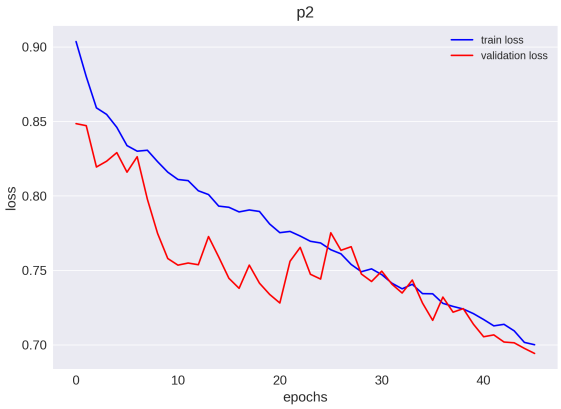
Στα αποτελέσματα που ακολουθούν, αξιολογούνται σε πρώτο επίπεδο οι διαφορετικές προσεγγίσεις που αφορούν κεφαλή ταξινόμησης ενός επιπέδου εξόδου. Στη συνέχεια, πραγματοποιώντας τις κατάλληλες τροποποιήσεις, έγινε επιλογή αρχιτεκτονικών που περιλαμβάνουν 1 και 2 κρυφά επίπεδα και εφαρμογή τους στο GreekSocialBERT. Στο Σχήμα 4.7 παρουσιάζονται τα διαγράμματα των απωλειών εκπαίδευσης (μπλε γραμμή) και επαλήθευσης (κόκκινη γραμμή) (Σχήματα 4.7α΄, 4.7γ΄, 4.7ε΄ και 4.7ζ΄) για την απλή αρχιτεκτονική ενός επιπέδου (εξόδου), καθώς και οι αντίστοιχοι πίνακες σύγχυσης (Σχήματα 4.7β΄, 4.7δ΄, 4.7ϛ΄, 4.7η΄) για κάθε μια από τις προσεγγίσεις p_1, p_2, p_3 και p_4,



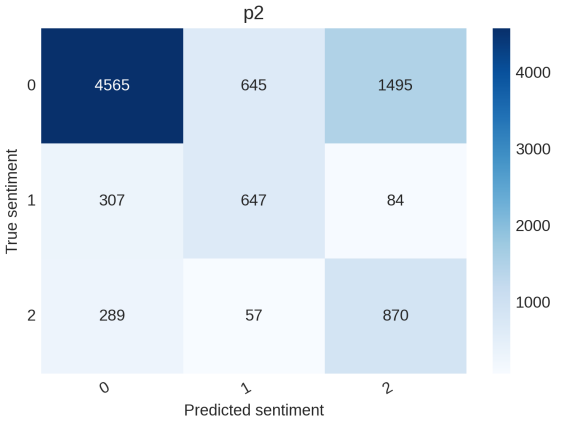
(α′)



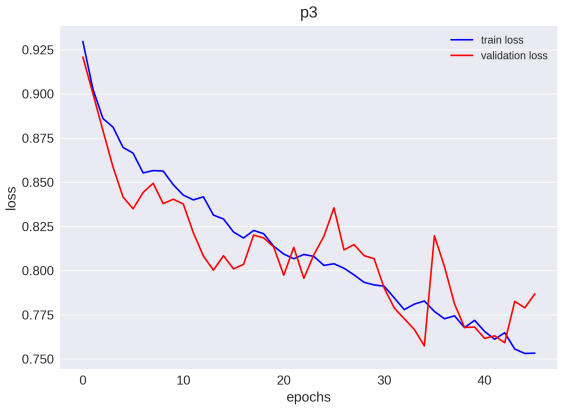
(β′)



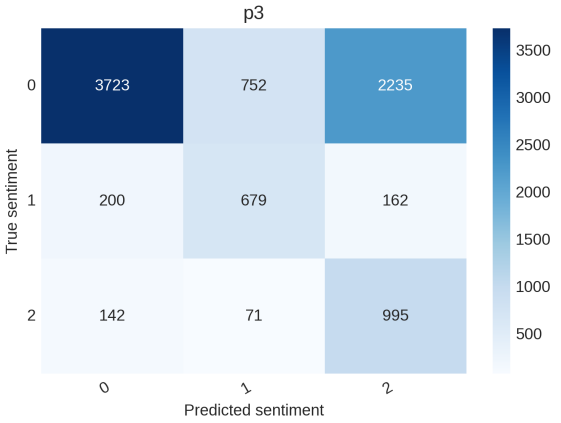
(γ′)



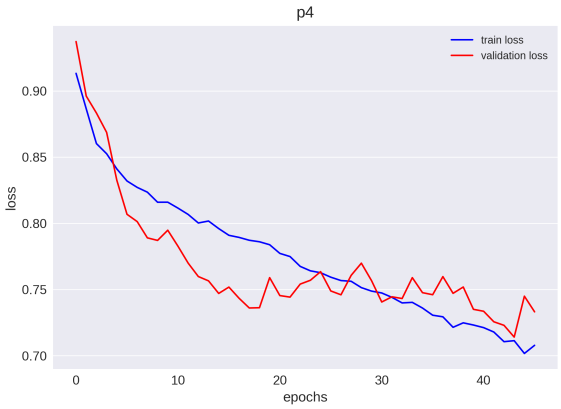
(δ′)



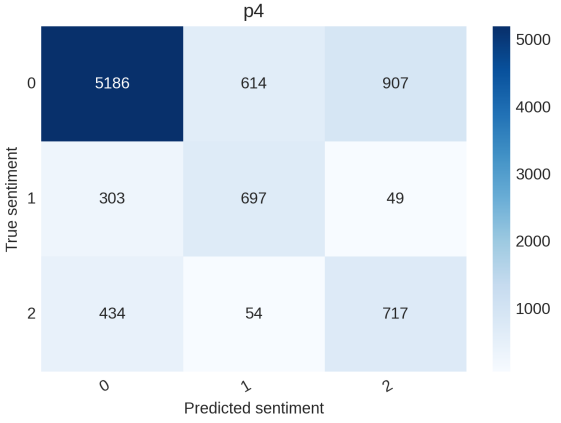
(ε′)



(ϛ′)



(ζ′)



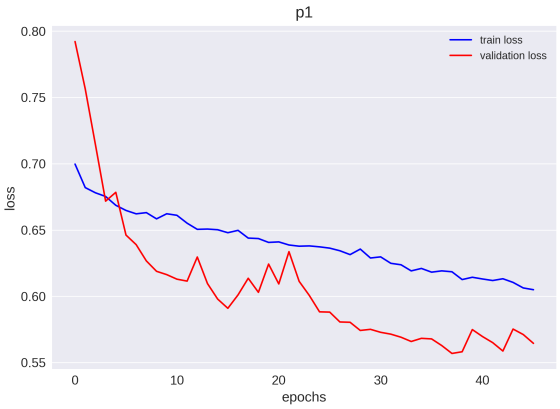
(η′)

Σχήμα 4.7: Κεφαλή ταξινόμησης ενός επιπέδου (εξόδου) και εκπαίδευση με βάρη κλάσεων

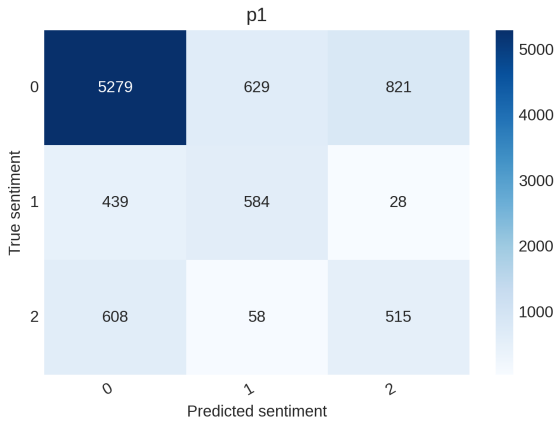
Από τις παραπάνω αξιολογήσεις απόδοσης, αν και τα αποτελέσματα δεν απέχουν κατά πολύ μεταξύ τους, η καλύτερη δυνατή προσέγγιση φαίνεται να είναι αυτή των p_2 και p_3. Στο συμπέρασμα αυτό καταλήγουμε, καθώς οι ταξινομητές φαίνεται να είναι σε θέση να διακρίνουν με μικρό περιθώριο λάθους τα θετικά από τα αρνητικά παραδείγματα. Παράλληλα, συγκριτικά με τις περιπτώσεις p_1 και p_4, ενώ κι εκείνες μπορούν να κατατάξουν αποδοτικά τα συναισθήματα, οι λανθασμένες προβλέψεις των θετικών και αρνητικών κειμένων ως ουδέτερα είναι σημαντικά περισσότερες. Στην περίπτωση των πελατών της ΠΑΛΟ ΨΗΦΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ Ε.Π.Ε., βασικό κριτήριο απόδοσης των μοντέλων ταξινόμησης συναισθήματος αποτελεί η εύρεση του αρνητικού συναισθήματος, προσέγγιση που ακολουθούμε και στο πλαίσιο της παρούσας διπλωματικής εργασίας. Συνεπώς, θεωρούμε προτιμότερη την αστοχία στην κατηγοριοποίηση ουδέτερων δειγμάτων, παρά αυτών που ανήκουν στις άλλες δύο κλάσεις.

Τα συγκεκριμένα μοντέλα χρησιμοποιούν βάρη κλάσεων κατά τον υπολογισμό της συνάρτησης απώλειας. Εξετάστηκε επίσης η συμπεριφορά των μοντέλων απαλείφοντας τα βάρη αυτά. Συγκεκριμένα, οι τιμές τις απώλειας αλλά και ο πίνακας σύγχυσης που προκύπτουν στην περίπτωση αυτή, σε αντιπαραβολή με τον συνυπολογισμό των βαρών κλάσεων, ακολουθούν ενδεικτικά για την περίπτωση p_1 στο Σχήμα 4.8

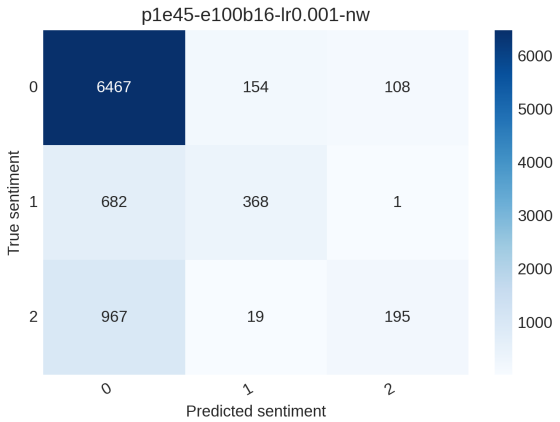
(α′) Συναρτήσεις απώλειας (με βάρη)



(β′) Συναρτήσεις απώλειας (χωρίς βάρη)



(γ′) Πίνακας σύγχυσης (με βάρη)



(δ′) Πίνακας σύγχυσης (χωρίς βάρη)

Σχήμα 4.8: Σύγκριση εκπαίδευσης με βάρη κλάσεων και χωρίς βάρη κλάσεων

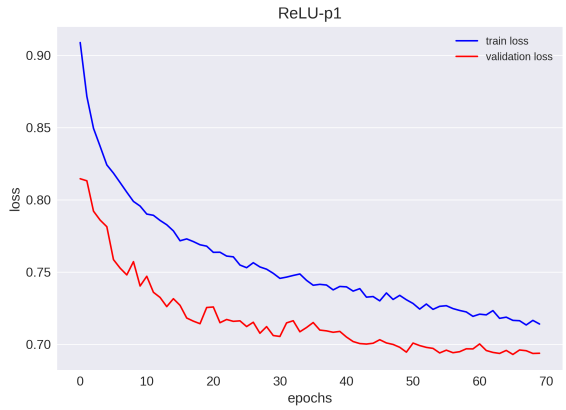
Παρατηρείται ότι, αν και η εκπαίδευση του ταξινομητή παρουσιάζει σχετικά μικρότερες τιμές απώλειας στην περίπτωση που δεν λαμβάνεται υπόψη η μη-ισορροπημένη κατανομή συναισθήματος της συλλογής δεδομένων, στην πραγματικότητα φαίνεται ότι η ικανότητά του να κατατάσσει σωστά τα δείγματα είναι περιορισμένη. Πιο συγκεκριμένα, από το Σχήμα 4.8δ΄ συμπεραίνουμε πως όταν το μοντέλο δεν λαμβάνει υπόψη ότι τα ουδέτερα (κλάση 0) κείμενα υπερτερούν σημαντικά στο σύνολο δεδομένων, τείνει να γενικεύει τις προβλέψεις του ως προς το αντίστοιχο συναίσθημα, καθώς η πιθανότητα το δείγμα να είναι πράγματι ουδέτερο είναι σαφώς μεγαλύτερη. Με αυτόν τον τρόπο οι κλάσεις μειονότητας, δηλαδή η θετική (κλάση 1) και η αρνητική (κλάση 2) παραμελούνται αισθητά, παρά το γεγονός ότι στα πλαίσια της ανάλυσης συναισθήματος οι 2 αυτές κατηγορίες αποτελούν το βασικό αντικείμενο μελέτης. Όσον αφορά ιδιαίτερα τα μέσα κοινωνικής δικτύωσης και τον ρόλο τους στην αντιπροσώπευση της κοινής γνώμης, η αποτυχία ανίχνευσης των αρνητικών σχολίων υποβαθμίζει σημαντικά την απόδοση του μοντέλου. Για τα δεδομένα του πειράματος, ο ταξινομητής κατάφερε να κατατάξει σωστά μόλις 195 αρνητικά δείγματα από τα 1.181 θεωρώντας την πλειοψηφία αυτών ως ουδέτερα. Ανάλογη συμπεριφορά παρατηρείται και στην περίπτωση των θετικών κειμένων.

Αντίθετα, όταν η συνάρτηση σφάλματος συμπεριλαμβάνει τα βάρη που προτείνει η μελέτη μας, φαίνεται ξεκάθαρα η βελτίωση της απόδοσης του ταξινομητή, καθώς η υπεροχή των ουδέτερων κειμένων δεν επηρεάζει την ικανότητά του να κατατάξει ορθά ένα μεγάλο ποσοστό αρνητικών, αλλά και θετικών δειγμάτων. Όπως ήταν βέβαια αναμενόμενο, τα βάρη κλάσεων περιορίζουν σε ένα βαθμό την ικανότητα του μοντέλου να προβλέψει σωστά την κλάση πλειοψηφίας (κλάση 0), ταξινομώντας πολύ περισσότερα δείγματα ως θετικά ή αρνητικά. Ωστόσο, όπως φαίνεται και στον αντίστοιχο πίνακα σύγχυσης (Πίνακας 4.8α΄, λόγω του πλήθους των ουδέτερων κειμένων, ο αντίκτυπος, ως προς την γενική του απόδοση είναι αμελητέος, αφού και πάλι επιτυγχάνει να κατηγοριοποιήσει σωστά την πλειοψηφία αυτών. Η ποιοτική σημασία των ουδέτερων δειγμάτων άλλωστε, σε εφαρμογές ανάλυσης συναισθήματος, δεν επηρεάζει ουσιαστικά την ικανότητα του γλωσσικού μοντέλου να ανιχνεύει συναισθηματικά φορτισμένο λόγο. Σημειώνεται επίσης, πως και στις δύο περιπτώσεις, η αναλογία θετικών - αρνητικών δειγμάτων στο σύνολο δεδομένων, επιτρέπει στον ταξινομητή να ξεχωρίσει με μεγάλο ποσοστό επιτυχίας τον θετικό από τον αρνητικό τόνο, αφού οι περιπτώσεις στις οποίες γίνεται εσφαλμένη κατηγοριοποίηση μεταξύ των δύο είναι ελάχιστες. Πράγματι, από τις μετρήσεις του Πίνακα 4.2, ο ταξινομητής φαίνεται να παρουσιάζει μεγάλο ποσοστό ακρίβειας και χαμηλή ανάκληση, αφού προτεραιότητά του είναι να προβλέψει σωστά όσο το δυνατόν περισσότερα δείγματα, ανεξάρτητα από την κλάση στην οποία ανήκουν. Αντίθετα, με την χρήση βαρών κλάσεων, η ανάκληση αυξάνεται, καθώς ο ταξινομητής προσπαθεί να εξασφαλίσει σωστές προβλέψεις ειδικά για κάθε κλάση.

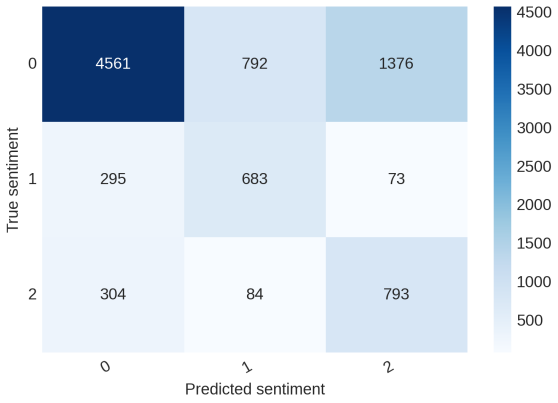
A	Ακρίβεια	Ανάκληση	F1-score
Ουδέτερο	$87,30\%$	$71,82\%$	$78,81\%$
Θετικό	$42,82\%$	$66,32\%$	$52,47\%$
Αρνητικό	$39,01\%$	$61,90\%$	$47,86\%$
B	Ακρίβεια	Ανάκληση	F1-score
Ουδέτερο	$79,68\%$	$96,11\%$	$87,13\%$
Θετικό	$68,02\%$	$35,01\%$	$46,23\%$
Αρνητικό	$64,14\%$	$16,51\%$	$26,26\%$

Πίνακας 4.2: (A). Με βάρη κλάσεων - (B). Χωρίς βάρη κλάσεων

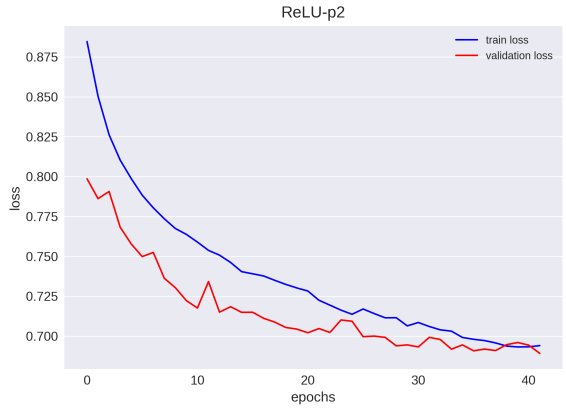
Με σκοπό την βελτίωση της αρχιτεκτονικής της κεφαλής ταξινόμησης, αναπτύχθηκαν νέες αρχιτεκτονικές δομές, οι οποίες επεκτείνουν τον αρχικό σχεδιασμό του ταξινομητή με την προσθήκη ενός κρυφού επιπέδου. Στα πλαίσια της πειραματικής λοιπόν διαδικασίας, οι αρχιτεκτονικές αυτές εφαρμόστηκαν στα γλωσσικά μοντέλα των προσεγγίσεων p_1, p_2, p_3 και p_4 αντίστοιχα. Δοκιμάστηκε επίσης και η εφαρμογή της συνάρτησης Softmax στην έξοδο του ταξινομητή. Ως είσοδος του ταξινομητή, όπως περιγράφηκε και στην Ενότητα 4.4.3, μπορεί να χρησιμοποιηθεί τόσο η τελική αναπαράσταση εξόδου του γλωσσικού μοντέλου απόκρυψης, όσο και η παραλλαγή αυτής, η δειγματοληπτημένη ακολουθία εξόδου. Στα Σχήματα 4.9 και 4.10 παρουσιάζονται τα αποτελέσματα που λάβαμε για κάθε συνδυασμό σχεδιαστικής επιλογής και εκδοχής εισόδου:



(α′)

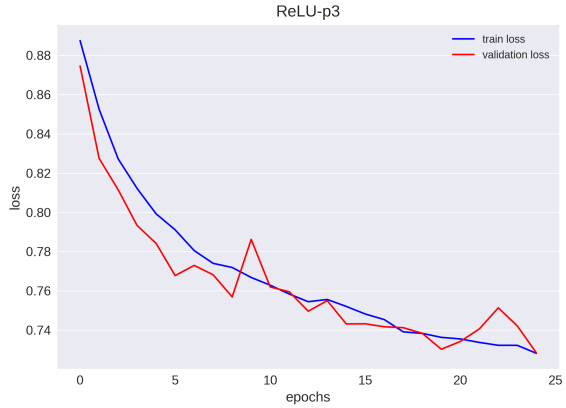


(β′)

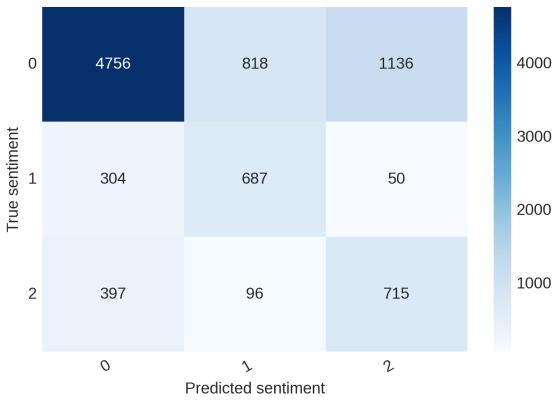


(γ′)

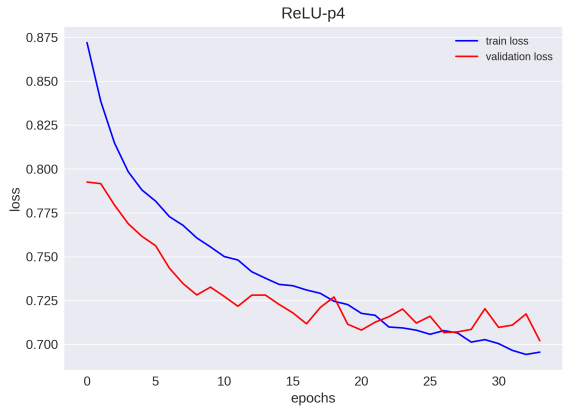
(δ′)



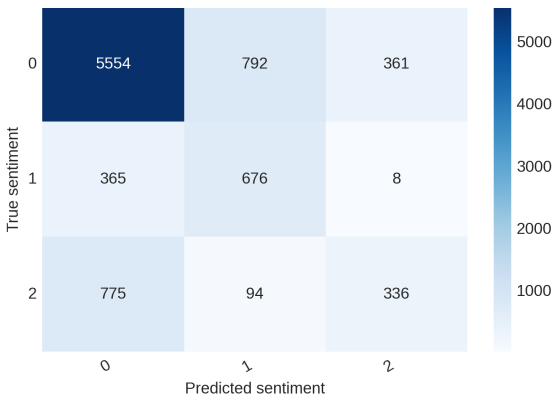
(ε′)



(ϛ′)

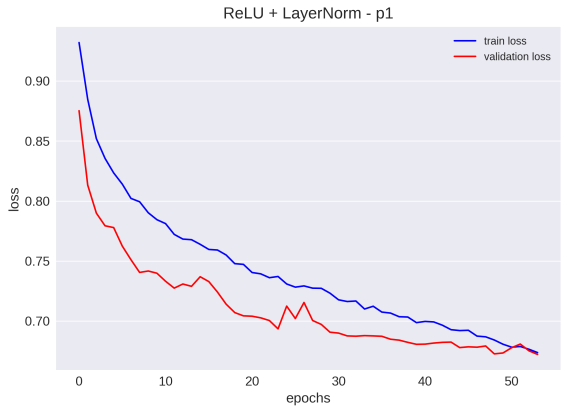


(ζ′)

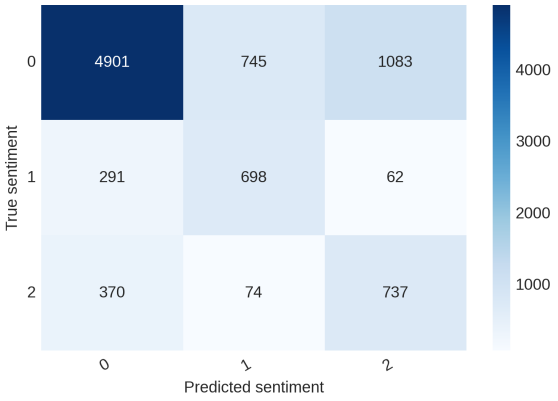


(η′)

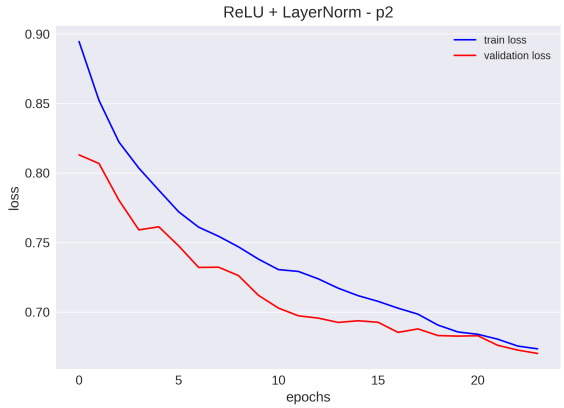
Σχήμα 4.9: Κεφαλή ταξινόμησης ενός κρυφού επιπέδου με χρήση συνάρτησης ενεργοποίησης ReLU



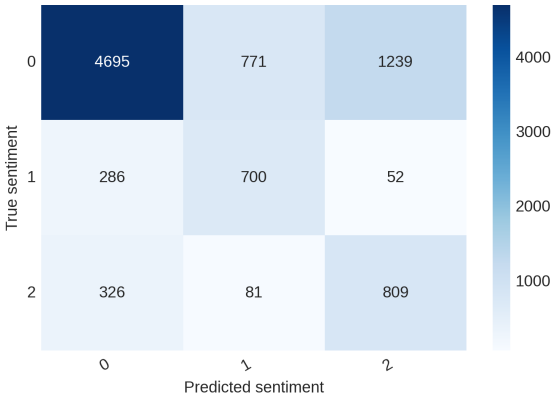
(α′)



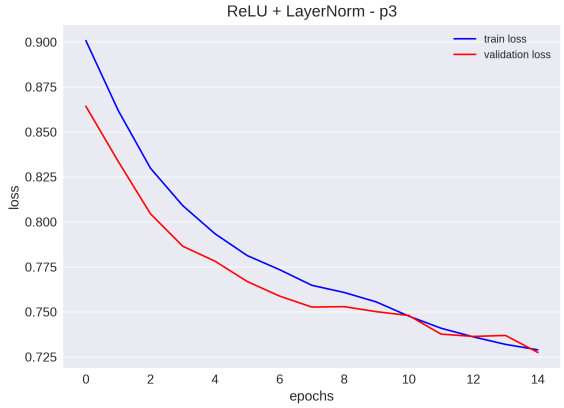
(β′)



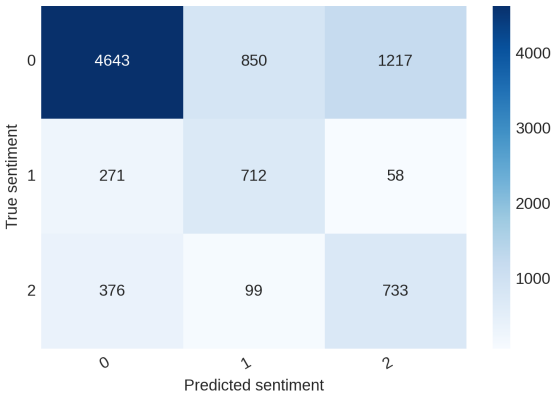
(γ′)



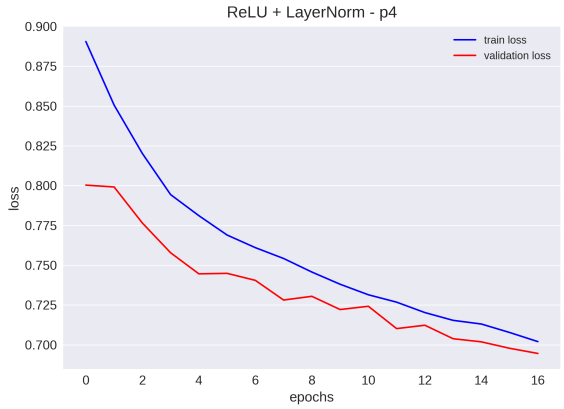
(δ′)



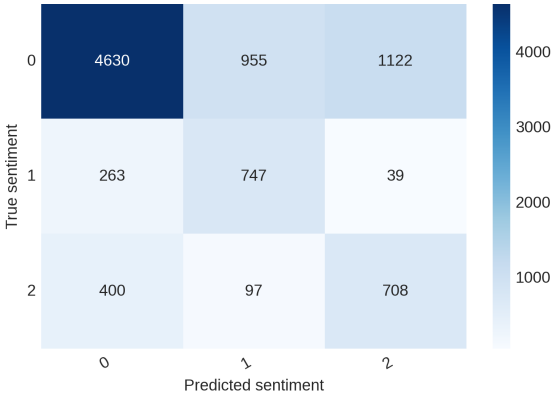
(ε′)



(ϝ′)



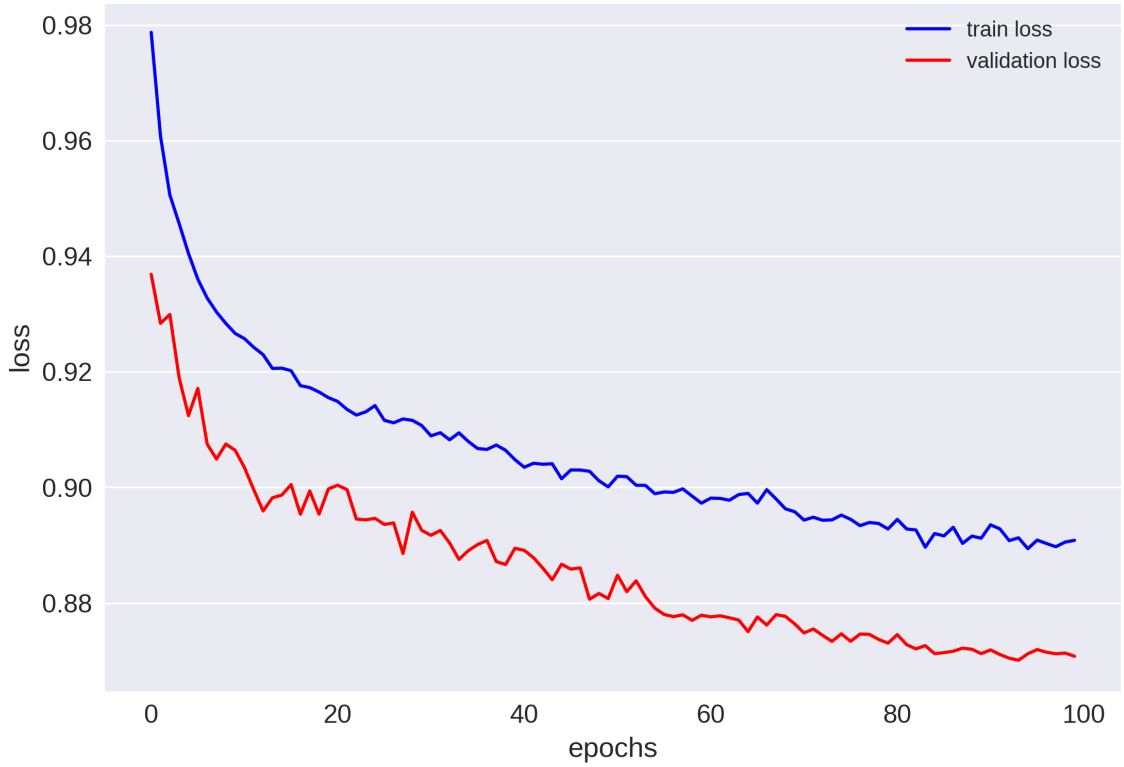
(ζ′)



(η′)

Σχήμα 4.10: Κεφαλή ταξινόμησης ενός κρυφού επιπέδου με χρήση συνάρτησης ενεργοποίησης ReLU και επιπέδου κανονικοποίησης

Όσον αφορά το σφάλμα που αντιστοιχεί στις σχεδιαστικές επιλογές, με τη χρήση της μη-γραμμικής συνάρτησης ενεργοποίησης ReLU, παρατηρείται βελτίωση σε σχέση με την απόδοση των ταξινομητών που δεν τη χρησιμοποιούν (Σχήμα 4.7), με τις τιμές απώλειας να μειώνονται πιο ομοιόμορφα. Παράλληλα σημειώνεται μείωση της απαιτούμενης διάρκειας εκπαίδευσης, μέχρι να παρουσιαστεί υπερπροσαρμογή στα δεδομένα. Συνεπώς, φαίνεται πως τα δεδομένα προσαρμόζονται καλύτερα και ταχύτερα στη συγκεκριμένη αρχιτεκτονική. Στη συνέχεια, με την προσθήκη ενός επιπέδου κανονικοποίησης ανάμεσα στο κρυφό επίπεδο και το επίπεδο εξόδου, τα μοντέλα σημειώνουν περαιτέρω βελτίωση όσον αφορά τις τιμές απώλειας αλλά και τη χρονική διάρκεια εκπαίδευσης (Σχήμα 4.10). Σε μία προσπάθεια να βελτιωθούν ακόμα περισσότερο τα αποτελέσματα, πραγματοποιήθηκε εφαρμογή της softmax συνάρτησης ενεργοποίησης στην έξοδο του ταξινομητή, όπως φαίνεται στο Σχήμα 4.11, ωστόσο από την πειραματική διαδικασία φαίνεται να επηρεάζει την εκπαίδευση του μοντέλου σημαντικά, επομένως δεν προτιμάται.

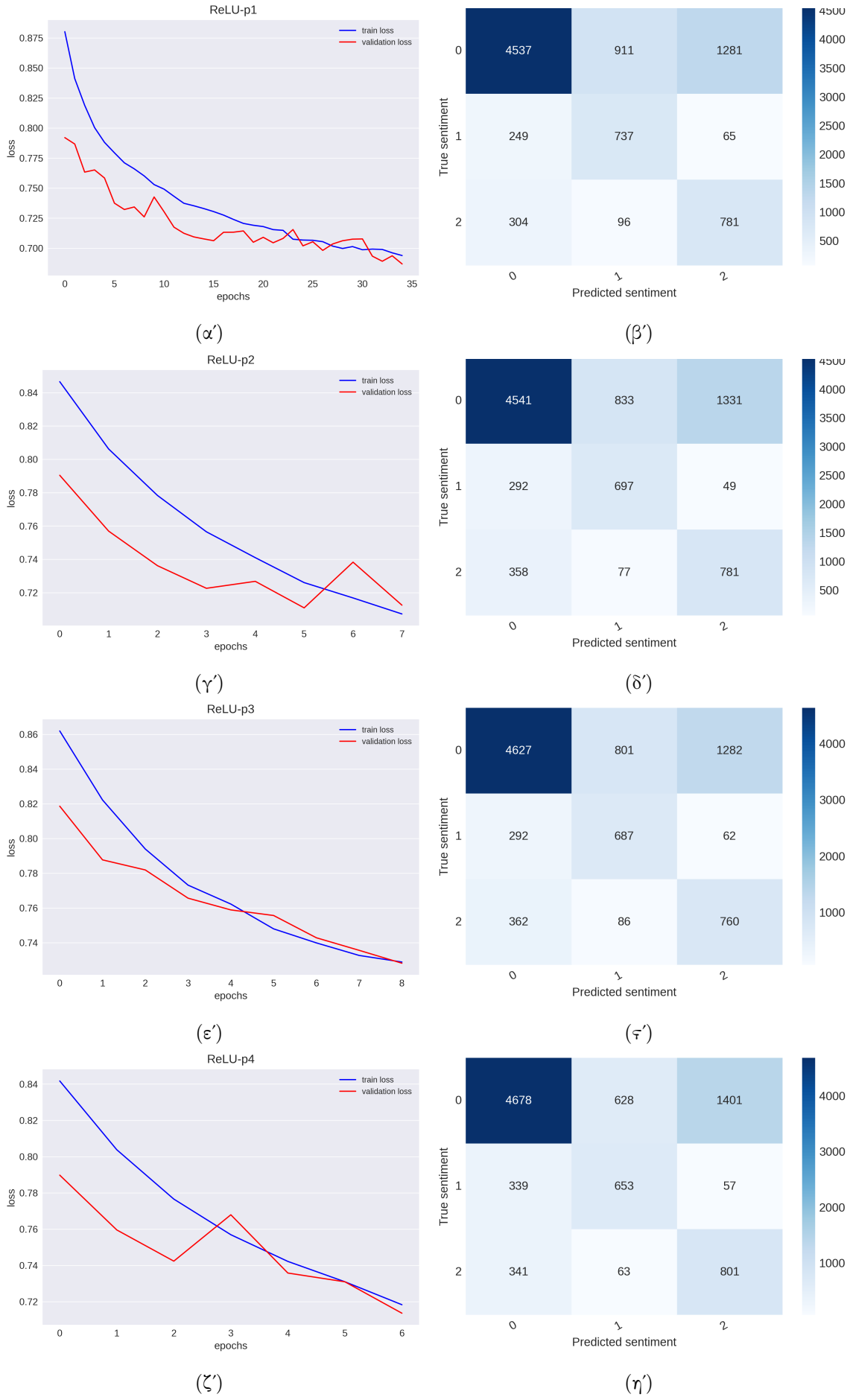


Σχήμα 4.11: Εφαρμογή συνάρτηση ενεργοποίησης softmax στο επίπεδο εξόδου του ταξινομητή

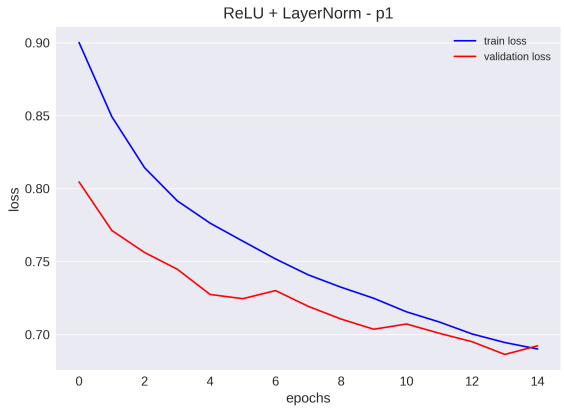
Με τη δοκιμή των ταξινομητών πάνω στα δεδομένα ελέγχου, μπορούμε να διαμορφώσουμε μια καλύτερη εικόνα για τις επιδόσεις τους. Βασική προτεραιότητα της ανάπτυξης ενός αποτελεσματικού μοντέλου ανάλυσης συναισθήματος είναι σαφώς η ιδιότητά του να διαχωρίζει με όσο το δυνατόν μεγαλύτερη βεβαιότητα τα αρνητικά από τα θετικά σχόλια. Το χαρακτηριστικό αυτό είναι ιδιαίτερα σημαντικό, καθώς κατά τις εφαρμογές ανάλυσης συναισθήματος, οι λανθασμένες ταξινομήσεις αρνητικών δειγμάτων ως θετικά, μειώνουν σε μεγάλο βαθμό την

αξιοπιστία του μοντέλου. Εστιάζοντας λοιπόν στις προβλέψεις που αφορούν τις κλάσεις 1 (θετικό συναίσθημα) και 2 (αρνητικό συναίσθημα), από τους παραπάνω πίνακες σύγχυσης βλέπουμε πως χάρη στη χρήση των βαρών κλάσεων, οι ταξινομητές, στην πλειονότητα των περιπτώσεων, επιτυγχάνουν την αποφυγή της σύγχυσης μεταξύ των δύο, ανεξάρτητα της αρχιτεκτονικής τους. Συγκεκριμένα, για την περίπτωση της αρνητικής κλάσης, από τα 1.181 παραδείγματα, στην χειρότερη περίπτωση έχουμε το πολύ 106 εσφαλμένες ταξινομήσεις ως θετικές. Ομοίως, οι προβλέψεις θετικών δειγμάτων ως αρνητικά, είναι επίσης ελάχιστες και δεν ξεπερνούν σε καμία περίπτωση τις 73, από τις συνολικά 1.051.

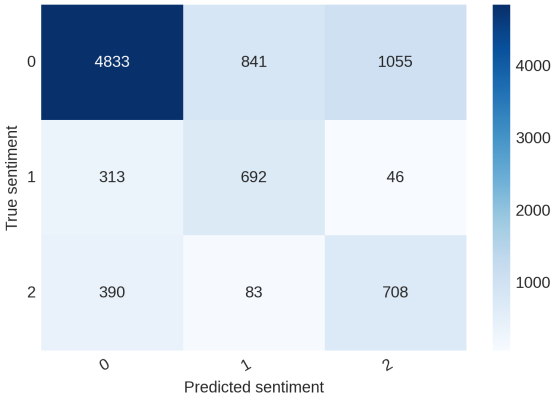
Αν και σύμφωνα με την παραπάνω ανάλυση, όλοι οι ταξινομητές φαίνεται να ανταποκρίνονται στο ζητούμενο της έρευνάς μας, βασικός επίσης παράγοντας στην επιλογή του κατάλληλου ταξινομητή, είναι και η απαιτούμενη διάρκεια εκπαίδευσης προκειμένου να λάβουμε τα επιθυμητά αποτελέσματα. Η διαδικασία εκπαίδευσης μπορεί σε πολλές περιπτώσεις να είναι εξαιρετικά χρονοβόρα και κατ' επέκταση ιδιαίτερα δαπανηρή, εξαντλώντας τους υπολογιστικούς πόρους. Συνεπώς, η επιλογή ενός αποτελεσματικού και συγχρόνως γρήγορου μοντέλου είναι ιδιαίτερα σημαντική. Όπως αναφέρθηκε και προηγουμένως, ο χρόνος εκπαίδευσης του εκάστοτε ταξινομητή, φαίνεται να μειώνεται αισθητά όταν προστίθεται στην αρχιτεκτονική δομή η συνάρτηση ενεργοποίησης ReLU, ενώ με την επιπλέον κανονικοποίηση του κρυφού επιπέδου, παρατηρείται ακόμα μεγαλύτερη βελτίωση (Σχήμα 4.10). Μέσω της επεξεργασίας αυτής, τροποποιώντας την αρχική αρχιτεκτονική του ταξινομητή κεφαλής, τελικά καταφέραμε να μειώσουμε σε μεγάλο βαθμό σχεδόν για κάθε προσέγγιση (p_1, p_2, p_3 και p_4) τον χρόνο εκπαίδευσης, βελτιώνοντας παράλληλα και την απόδοση των μοντέλων. Η αναβάθμιση αυτή ωστόσο, δεν καθιστά απαραίτητα όλες τις πειραματικές προσεγγίσεις το ίδιο αποτελεσματικές. Ειδικότερα, στην προσέγγιση p_1, παρατηρείται ότι δεν φαίνεται να μπορεί να μειώσει τον απαιτούμενο χρόνο εκπαίδευσής της, παρόλο που οι υπόλοιποι ταξινομητές δείχνουν σαφή βελτίωση όταν εφαρμόζεται τόσο η ReLU συνάρτηση, όσο και η κανονικοποίηση του κρυφού επιπέδου. Όσον αφορά τους υπόλοιπους ταξινομητές, καταλήγουμε στο συμπέρασμα ότι, εκτός από την ταχύτερη εκπαίδευσή τους, με την επιλογή αρχιτεκτονικής δομής που συνδυάζει τη συνάρτηση ReLU αλλά και κανονικοποίηση του κρυφού επιπέδου, σημειώνεται και μεγαλύτερη ακρίβεια ως προς την αποφυγή ταξινόμησης θετικών και αρνητικών δειγμάτων ως ουδέτερα. Το συγκεκριμένο σφάλμα, αν και κρίνουμε πως δεν διαστρεβλώνει ιδιαίτερα την τελική εικόνα που διαμορφώνεται από τον ταξινομητή ως προς την δυσαρέσκεια της κοινής γνώμης, μπορεί να αποτρέψει το μοντέλο από το να αξιοποιήσει επιπλέον σημαντική πληροφορία. Επομένως, παρ' ότι δεν παραποιούνται άμεσα τα τελικά αποτελέσματα, ενδέχεται η ταξινόμηση που προκύπτει να μην αντιπροσωπεύει πιστά την πραγματική πόλωση γύρω από ένα θέμα συζήτησης, αφού δεν λαμβάνεται υπόψη η ποιοτική σημασία ενός ποσοστού θετικών και αντίστοιχα αρνητικών δειγμάτων. Στα πλαίσια της πειραματικής διαδικασίας, έγινε δοκιμή εφαρμογής της δειγματοληπτημένης ακολουθίας εξόδου ως είσοδο στον κάθε ταξινομητή κεφαλής. Στα Σχήματα 4.12 και 4.13 παρουσιάζονται αποτελέσματα που προέκυψαν



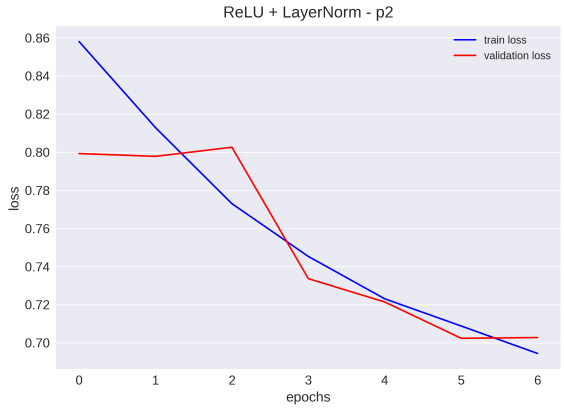
Σχήμα 4.12: Κεφαλή ταξινόμησης ενός κρυφού επιπέδου με χρήση δειγματοληπτημένης ακολουθίας και συνάρτηση ενεργοποίησης ReLU



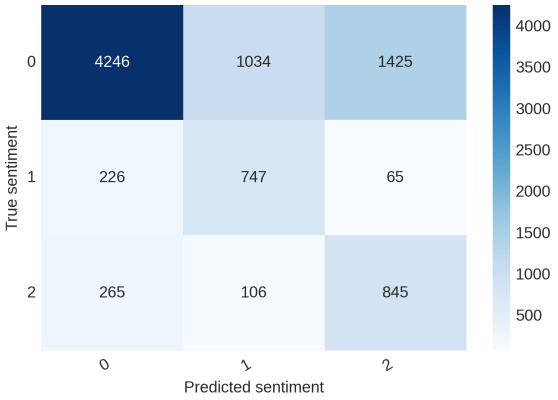
(α′)



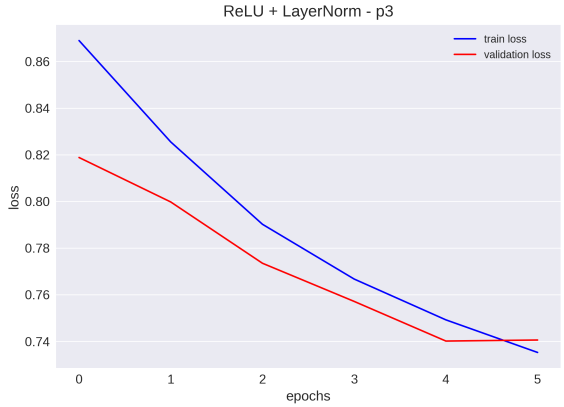
(β′)



(γ′)

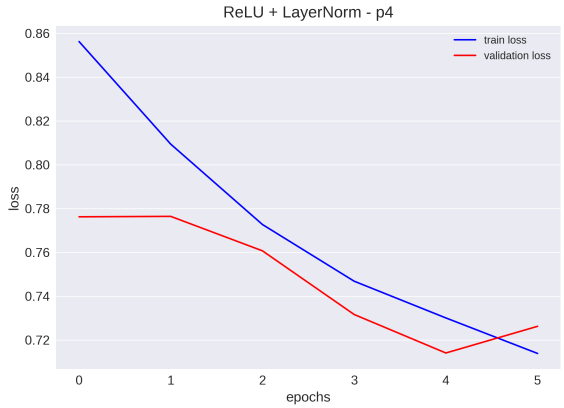


(δ′)

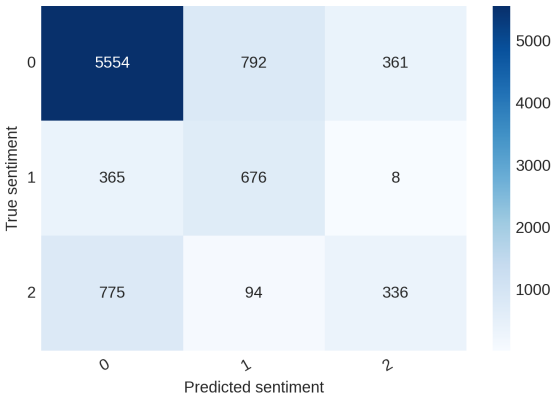


(ε′)

(ϛ′)



(ζ′)



(η′)

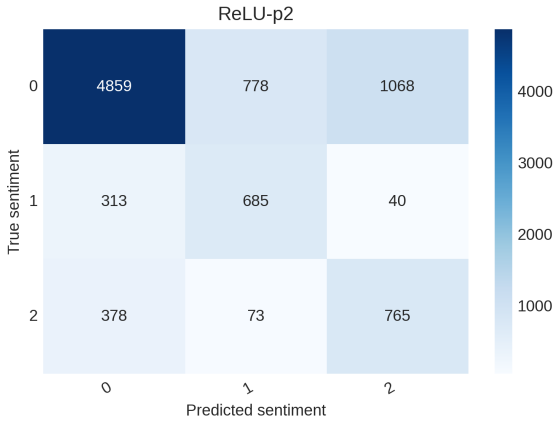
Σχήμα 4.13: Κεφαλή ταξινόμησης ενός κρυφού επιπέδου με χρήση δειγματοληπτημένης ακολουθίας, συνάρτηση ενεργοποίησης ReLU και επιπέδου κανονικοποίησης

Σχετικά με τη δειγματοληπτημένη ακολουθία εξόδου, φαίνεται πως η επιλογή της αναπαράστασης της εισόδου του ταξινομητή επηρεάζει σημαντικά την διάρκεια εκπαίδευσης του κάθε μοντέλου. Ενώ η απώλεια που παρουσιάζεται σε κάθε περίπτωση διατηρείται περίπου στο ίδια επίπεδα, οι εποχές που απαιτούνται προκειμένου να επιτευχθούν παρόμοια αποτελέσματα διαφέρουν σημαντικά. Ειδικότερα, με χρήση της δειγματοληπτημένης ακολουθίας εξόδου του γλωσσικού μοντέλου με απόκρυψη, ο χρόνος μειώνεται σχεδόν στο $\frac{1}{3}$ σε κάθε πείραμα. Ωστόσο, κατά την εφαρμογή των μοντέλων που χρησιμοποιούν τη δειγματοληπτημένη ακολουθία εξόδου ως είσοδο, παρατηρούμε μια μικρή αλλοίωση των αποτελεσμάτων όσον αφορά την ικανότητα να αναγνωρίζει και να ταξινομεί σωστά συναισθηματικά πολωμένα δείγματα. Η επιδείνωση αυτή φαίνεται τόσο στις άστοχες προβλέψεις μεταξύ των αρνητικών και θετικών κειμένων, όσο και στην τάση των ταξινομητών να τα κατηγοριοποιούν ως ουδέτερα. Εξαίρεση αποτελεί η περίπτωση p_2, στην οποία αν και σημειώνεται μικρή αύξηση στην ταξινόμηση των δειγμάτων ως ουδέτερα, η ικανότητά του να ξεχωρίζει τις θετικές από τις αρνητικές περιπτώσεις, παρουσιάζει μια μικρή βελτίωση.

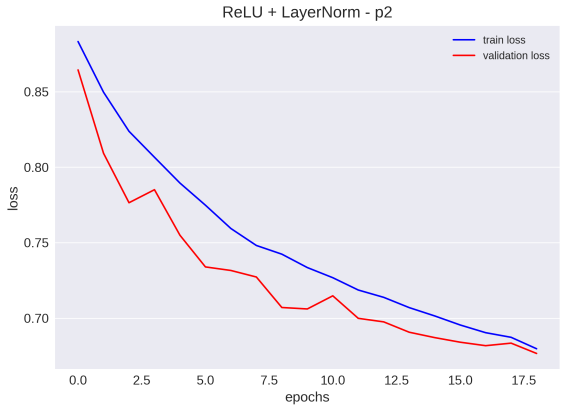
Βάσει των παραπάνω συμπερασμάτων που προκύπτουν από την πειραματική διαδικασία, φαίνεται πώς στην προσέγγιση p_2, ο ταξινομητής που χρησιμοποιεί στην αρχιτεκτονική δομή του την μη-γραμμική συνάρτηση ενεργοποίησης ReLU και παράλληλα εφαρμόζει κανονικοποίηση του κρυφού επιπέδου, λαμβάνοντας ως είσοδο τη δειγματοληπτημένη ακολουθία εξόδου, παρουσιάζει μικρό προβάδισμα. Η επιλογή του συγκεκριμένου ταξινομητή, βασίζεται στην ακρίβεια ταξινόμησης των θετικών και αρνητικών δειγμάτων, αποφεύγοντας όσο το δυνατόν περισσότερο την μεταξύ τους σύγχυση. Παράλληλα, οι περιπτώσεις στις οποίες τα δείγματα που ανήκουν στις κλάσεις 1 και 2 κατατάσσονται εσφαλμένα στην ουδέτερη κατηγορία, είναι συγκριτικά λιγότερες με αυτές που προκύπτουν με άλλους ταξινομητές. Βασικός παράγοντας στην επιλογή του συγκεκριμένου μοντέλου, είναι το γεγονός ότι επιτυγχάνει τα συγκεκριμένα αποτελέσματα με την απαιτούμενη εκπαίδευσή του να περιορίζεται στις 6 εποχές.

Σύμφωνα με τις παραπάνω παρατηρήσεις, επιλέξαμε συγκεκριμένες προσεγγίσεις ώστε να εξετάσουμε το περιθώριο βελτίωσής τους. Τα παρακάτω πειράματα αφορούν την προσθήκη δεύτερου κρυφού επιπέδου στην αρχιτεκτονική δομή του εκάστοτε ταξινομητή της περίπτωσης p_2 και την υλοποίηση του GreekSocialBERT. Συγκεκριμένα για κάθε περίπτωση λαμβάνουμε τα εξής αποτελέσματα (Σχήματα 4.14 και 4.15)

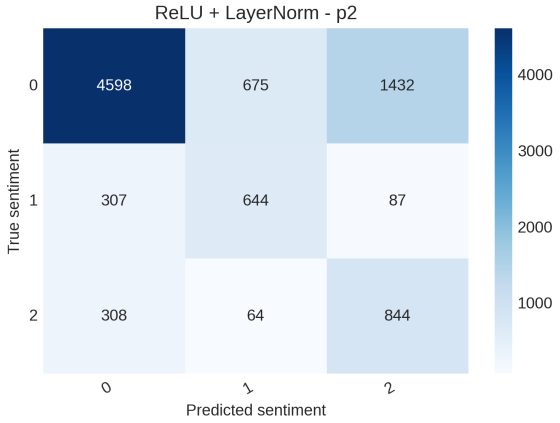
(α′) Συνάρτηση ενεργοποίησης ReLU



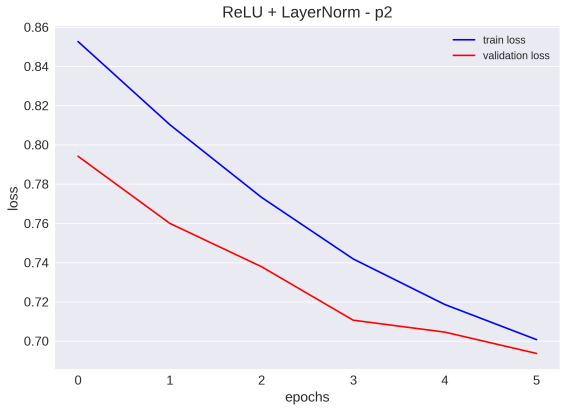
(β′) Συνάρτηση ενεργοποίησης ReLU



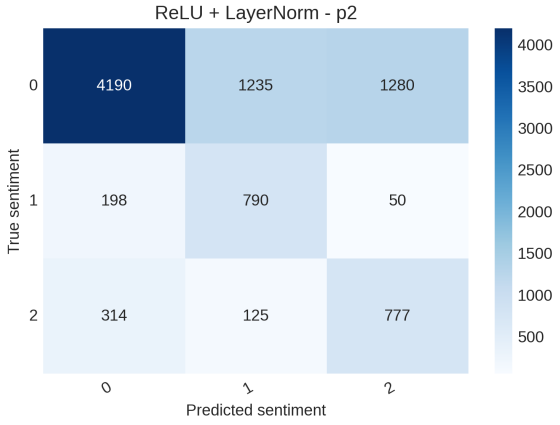
(γ′) Συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση επιπέδου



(δ′) Συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση επιπέδου



(ε′) Συνάρτηση ενεργοποίησης ReLU, κανονικοποίηση επιπέδου και δειγματοληπτιμένη ακολουθία



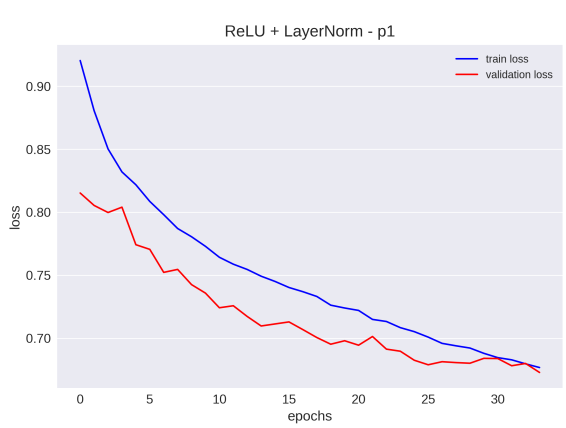
(ϛ′) Συνάρτηση ενεργοποίησης ReLU, κανονικοποίηση επιπέδου και δειγματοληπτιμένη ακολουθία

Σχήμα 4.14: Ταξινομητές με 2 κρυφά επίπεδα - περίπτωση p_2

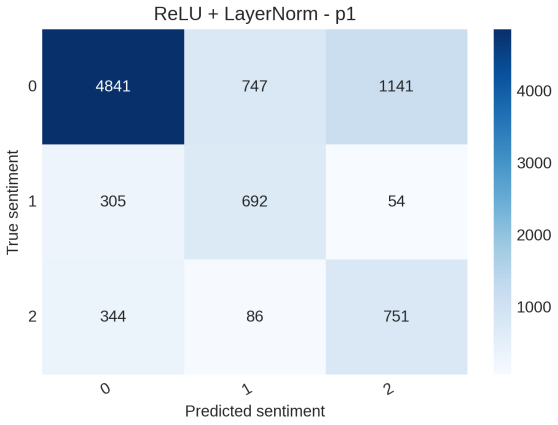
Βλέπουμε ότι η χρήση δεύτερου κρυφού επιπέδου επισπεύδει την εκπαίδευση των ταξινομητών, παρουσιάζοντας συγχρόνως και μια μικρή βελτίωση όσον αφορά το σφάλμα. Επιπλέον, κατά την διαδικασία ελέγχου, στις περιπτώσεις που χρησιμοποιείται ως είσοδος του ταξινο-

μητή η δειγματοληπτημένη ακολουθία εξόδου του προ-εκπαιδευμένου μοντέλου, παρατηρείται βελτίωση των αποτελεσμάτων, ιδιαίτερα στη βασική κλάση που απασχολεί το πείραμα, την αρνητική. Αντίθετα, όσον αφορά τη δειγματοληπτημένη ακολουθία εξόδου, το δεύτερο κρυφό επίπεδο φαίνεται να χειροτερεύει την απόδοση του ταξινομητή.

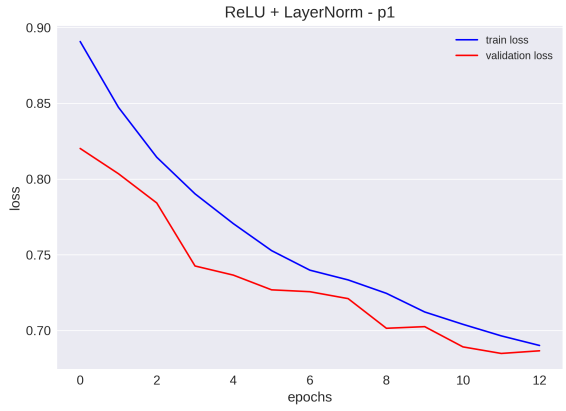
Λαμβάνοντας υπόψη τη μείωση του χρόνου εκπαίδευσης αλλά και τη βελτίωση της αποδοτικότητας των μοντέλων, επιχειρήσαμε να προσθέσουμε δεύτερο κρυφό επίπεδο και για την πειραματική προσέγγιση p_1, η οποία αν και επιφέρει αξιόλογα αποτελέσματα κατά την πειραματική διαδικασία, η εκμάθηση των ταξινομητών καθυστερεί σημαντικά σε σχέση με τις υπόλοιπες εκδοχές. Ειδικότερα, επιλέξαμε να επικεντρωθούμε στην αρχιτεκτονική που περιλαμβάνει την ReLU συνάρτηση ενεργοποίησης αλλά και την κανονικοποίηση επιπέδου, καθώς ο συνδυασμός αυτός επιφέρει καλύτερα αποτελέσματα.



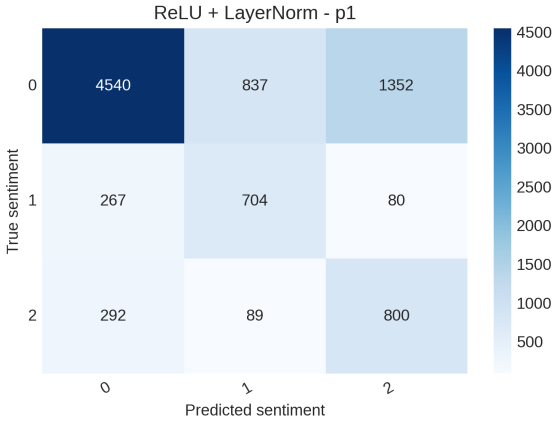
(α′) Συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση επιπέδου



(β′) Συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση επιπέδου



(γ′) Συνάρτηση ενεργοποίησης ReLU, κανονικοποίηση επιπέδου και δειγματοληπτιμένη ακολουθία



(δ′) Συνάρτηση ενεργοποίησης ReLU, κανονικοποίηση επιπέδου και δειγματοληπτιμένη ακολουθία

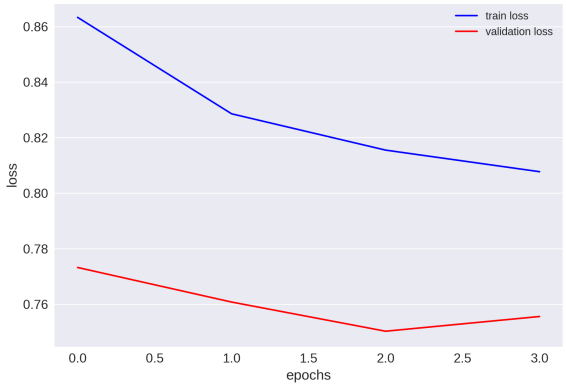
Σχήμα 4.15: Ταξινομητές με 2 κρυφά επίπεδα - περίπτωση p_1

Από το παραπάνω πείραμα, πράγματι το δεύτερο κρυφό επίπεδο αυξάνει σημαντικά την ταχύτητα εκπαίδευσης το μοντέλου, ενώ παράλληλα οι σωστές προβλέψεις αρνητικού συναι-

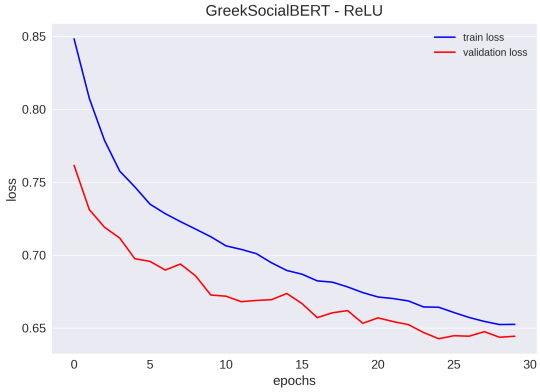
σθήματος είναι περισσότερες. Αν και παρουσιάζεται μια μικρή αύξηση στην αστοχία μεταξύ της κλάσης 1 και 2, τα δείγματα που την προκαλούν είναι ελάχιστα. Ομοίως με την περίπτωση της προσέγγισης p_2, η εφαρμογή επιπλέον επιπέδου στα μοντέλα με είσοδο τη δειγματοληπτημένη ακολουθία εξόδου δεν επιφέρει καλύτερα αποτελέσματα.

GreekSocialBERT

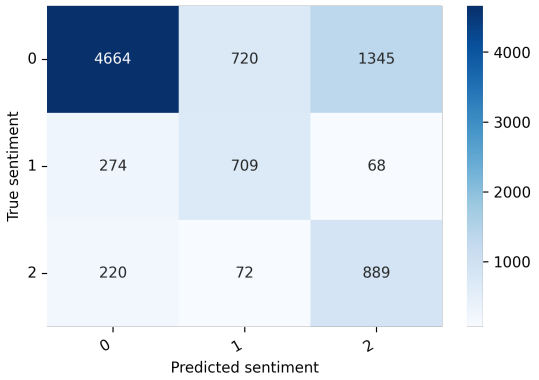
Το σώμα κειμένου για την ανάπτυξη προ-εκπαιδευμένων μοντέλων, διακρίνεται από ιδιόμορφα χαρακτηριστικά, ικανά να επηρεάσουν την ομαλή και αποτελεσματική εκπαίδευση του ταξινομητή. Αν και τα αποτελέσματα που λάβαμε από την παραπάνω πειραματική διαδικασία ανταποκρίνονται σε μεγάλο βαθμό στις προσδοκίες μας, επιχειρήσαμε να συγκρίνουμε τις επιδόσεις του ταξινομητή, όταν αυτός εφαρμοστεί σε ένα μεγάλο και ιδιαίτερα δημοφιλές προ-εκπαιδευμένο γλωσσικό μοντέλο, το GreekBERT [17]. Η προ-εκπαίδευσή του συγκεκριμένου μοντέλου, έχει γίνει πάνω σε ένα εξαιρετικά μεγάλο όγκο σώματος κειμένου, το οποίο ωστόσο, δεν περιλαμβάνει πηγές από μέσα κοινωνικής δικτύωσης στην ελληνική γλώσσα. Προκειμένου να γίνει επιτυχής προσαρμογή του GreekBERT στην εργασία ανάλυσης συναισθήματος, έγινε περαιτέρω εκπαίδευσή του συγκεκριμένα πάνω στο σώμα κειμένου της συλλογής δεδομένων μας, το οποίο τροποποιήθηκε αναλόγως, απαλείφοντας τους τόνους και μετατρέποντας τα κεφαλαία γράμματα σε πεζά (πειραματική προσέγγιση p_1). Η εκπαίδευση έγινε για 4 μόνο εποχές, προκειμένου να αποφευχθούν φαινόμενα υπερπροσαρμογής. Στη συνέχεια, κατά τη διαδικασία της λεπτής προσαρμογής, εφαρμόστηκαν οι ακόλουθες αρχιτεκτονικές, έτσι ώστε να εξεταστούν τα περιθώρια βελτίωσης του τελικού γλωσσικού μοντέλου ανάλυσης συναισθήματος, του GreekSocialBERT [7] (Σχήμα 4.16)



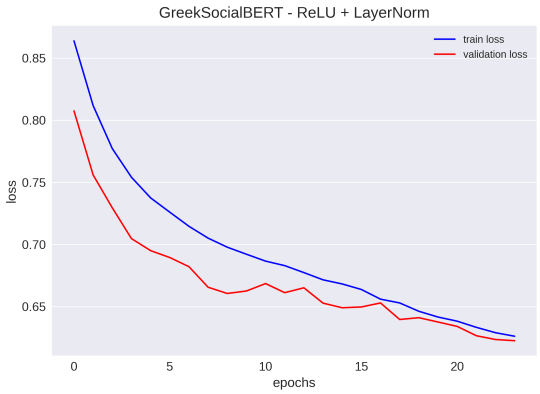
(α′) Κεφαλή ταξινόμησης επιπέδου εξόδου



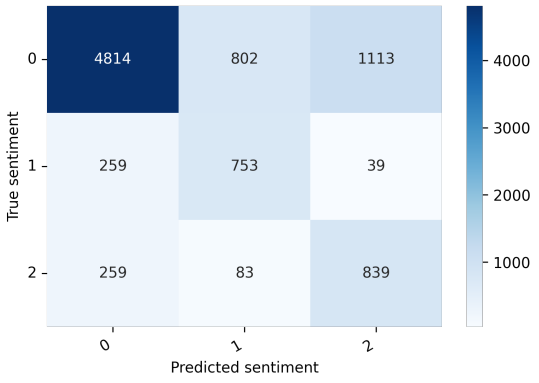
(β′) GreekSocialBERT με συνάρτηση ενεργοποίησης ReLU



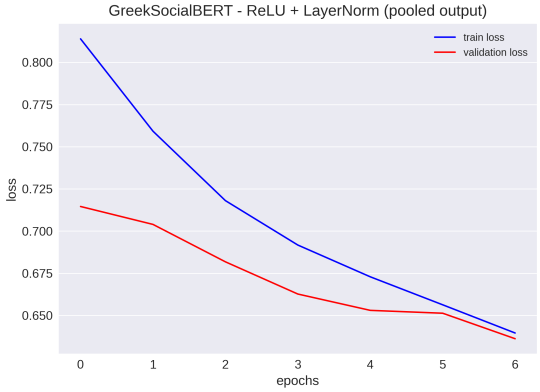
(γ′) GreekSocialBERT με συνάρτηση ενεργοποίησης ReLU



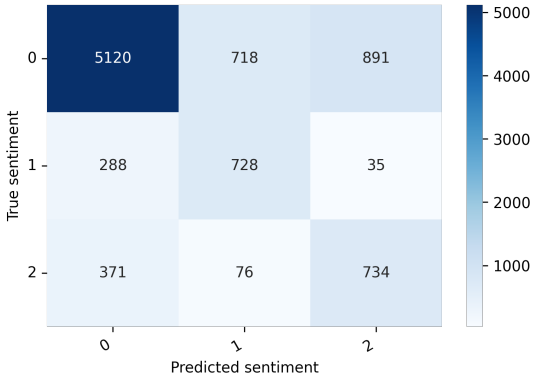
(δ′) GreekSocialBERT με συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση επιπέδου



(ε′) GreekSocialBERT με συνάρτηση ενεργοποίησης ReLU και κανονικοποίηση επιπέδου



(στ′) GreekSocialBERT με συνάρτηση ενεργοποίησης ReLU, κανονικοποίηση επιπέδου και δειγματοληπτημένη ακολουθία



(ζ′) GreekSocialBERT με συνάρτηση ενεργοποίησης ReLU, κανονικοποίηση επιπέδου και δειγματοληπτημένη ακολουθία

Σχήμα 4.16: Αρχιτεκτονικές διαμορφώσεις του μοντέλου GreekSocialBERT

Αρχικά φαίνεται πως η εφαρμογή της κεφαλής ταξινόμησης αποτυγχάνει κατά τη λεπτή προσαρμογής του γλωσσικού μας μοντέλου, καθώς παρατηρείται έντονο φαινόμενο υπό-προσαρμογής από την αρχή της εκπαίδευσης. Ωστόσο, όταν εφαρμόζονται οι επιπλέον αρχιτεκτονικές που αναπτύχθηκαν στα πλαίσια της πειραματικής διαδικασίας, το μοντέλο φαίνεται να επιτυγχάνει αξιόλογα αποτελέσματα, αναβαθμίζοντας τις επιδόσεις των υφισταμένων ταξινομητών. Παρατηρούμε ότι οι ταξινομητές εκπαιδεύονται επιτυχώς σε σύντομο χρονικό διάστημα, με τα αποτελέσματά τους να είναι αρκετά κοντά με αυτά των μοντέλων που εκπαιδεύτηκαν πάνω στο σώμα κειμένου από τα μέσα κοινωνικής δικτυώσης

Όπως ήταν αναμενόμενο, η εκτενής προ-εκπαίδευση του GreekBERT εξασφαλίζει την καλύτερη και ευκολότερη προσαρμογή του στις διάφορες εργασίες ΕΦΓ και πιο συγκεκριμένα στην ανάλυση συναισθήματος. Ωστόσο, λαμβάνοντας υπόψη τον περιορισμένο όγκο του πειραματικού σώματος κειμένου, οι επιδόσεις των αρχικών ταξινομητών είναι αξιοσημείωτες. Η σύγκριση των μοντέλων με αυτά του GreekSocialBERT, επαληθεύει ότι εκτός από τις σχεδιαστικές επιλογές του πειράματος, ιδιαίτερα σημαντικό ρόλο στην επιτυχία τους παίζει και το εξειδικευμένο περιεχόμενο του συνόλου δεδομένων προ-εκπαίδευσης. Συγκεκριμένα, λόγω των ιδιαίτερων γνωρισμάτων που παρουσιάζουν τα κείμενα στα μέσα κοινωνικής δικτύωσης, η επιλογή ενός σώματος προ-εκπαίδευσης με ανάλογο περιεχόμενο, αποτελεί βασική προϋπόθεση για την ανάπτυξη ενός αποτελεσματικού μοντέλου ταξινόμησης, καθώς σε διαφορετική περίπτωση θα παρουσιαζόταν έντονη δυσκολία ως προς την επεξεργασία και την σημασιολογική αποτύπωση των όρων.

Κεφάλαιο 5

Σύνοψη

5.1 Συμπεράσματα

Συνοψίζοντας, η έρευνα που πραγματοποιήθηκε στην παρούσα διπλωματική εργασία, πραγματεύεται την ανάπτυξη μοντέλων ανάλυσης συναισθήματος, των οποίων η δομή βασίζεται στο μοντέλο μετασχηματιστή RoBERTa. Η διαδικασία υλοποίησης περιλαμβάνει σε πρώτο στάδιο την προ-εκπαίδευσή τους πάνω σε κείμενα γραμμένα στην ελληνική γλώσσα, που προέρχονται από τα μέσα κοινωνικής δικτύωσης και στη συνέχεια την προσαρμογή τους, προκειμένου να εκτελούν εργασίες ανάλυσης συναισθήματος. Κατά την πειραματική διαδικασία εξετάστηκαν 4 διαφορετικές τροποποιήσεις των συνόλων δεδομένων ως προς τα μορφολογικά τους χαρακτηριστικά, ενώ παράλληλα για την υλοποίηση των ταξινομητών έγινε μελέτη ως προς την καταλληλότερη αρχιτεκτονική δομή που αντιστοιχεί στον καθένα.

Όσον αφορά τα προ-εκπαιδευμένα μοντέλα, κατά την πειραματική διαδικασία επιβεβαιώθηκε και πειραματικά πως το σώμα κειμένου επηρεάζει σε μεγάλο βαθμό την απόδοση ενός γλωσσικού μοντέλου. Συγκεκριμένα, η ικανότητα γενίκευσης του είναι άμεσα εξαρτώμενη από τα ποιοτικά και ποσοτικά χαρακτηριστικά των δειγμάτων, τα οποία στην προκειμένη περίπτωση, επειδή συγκεντρώθηκαν από τα μέσα κοινωνικής δικτύωσης, περιλαμβάνουν πολλές συντακτικές και νοηματικές αστοχίες, ενώ παράλληλα το εύρος των θεματικών που καλύπτουν περιορίζεται σημαντικά λόγω των εξειδικευμένων πηγών από τις οποίες προέρχονται. Λαμβάνοντας λοιπόν υπόψη τις παραπάνω παρατηρήσεις, κατά την αξιολόγηση του προ-εκπαιδευμένου μοντέλου, η απόδοση των μοντέλων ήταν σαφώς καλύτερη σε παραδείγματα που αφορούν παρόμοια θεματικά πλαίσια με αυτά του σώματος κειμένου. Παρ' όλα αυτά, όπως φάνηκε κατά την διαδικασία της λεπτής προσαρμογής, τα αποτελέσματα που προέκυψαν φαίνεται να είναι ικανά να εξυπηρετήσουν τα ζητούμενα της συγκεκριμένης μελέτης.

Ειδικότερα, σχεδόν σε κάθε πειραματική προσέγγιση τα μοντέλα φαίνεται να είναι σε θέση να διακρίνουν με μεγάλη ακρίβεια τον θετικό από τον αρνητικό τόνο που αποτυπώνει ένα κείμενο. Ωστόσο, παρατηρείται μια σχετική αστοχία όσον αφορά την εσφαλμένη ταξινόμηση ορισμένων δειγμάτων στην ουδέτερη κλάση, τα οποία όμως αποτελούν την μειοψηφία των αποτελεσμάτων. Καθοριστικός παράγοντας που επηρεάζει την εσφαλμένη ταξινόμηση των κειμένων ως ουδέτερα, είναι το εξαιρετικά μη-ισορροπημένο σύνολο δεδομένων. Συγκεκριμένα,

τα ουδέτερα δείγματα καταλαμβάνουν σχεδόν το 75% του συνόλου, με αποτέλεσμα να μην αφήνουν μεγάλα περιθώρια βελτίωσης στα γλωσσικά μοντέλα. Προκειμένου να εξισορροπηθεί έως ένα σημείο το πρόβλημα, ιδιαίτερα σημαντική ήταν η χρήση των βαρών κλάσεων κατά τον υπολογισμό του σφάλματος. Αξιοποιώντας λοιπόν την παράμετρο αυτή, η εκπαίδευση των μοντέλων σημείωσε σε ένα μεγάλο βαθμό επιτυχία, παρά τις σχετικές αστοχίες που αναφέρονται παραπάνω. Σε αντίθετη περίπτωση, όπως φάνηκε και κατά την πειραματική διαδικασία, εάν δεν ληφθεί υπόψη η ανισορροπία της κατανομής συναισθημάτων στη συλλογή δεδομένων, οι ταξινομητές, ανεξάρτητα από την αρχιτεκτονική τους δομή, αποτυγχάνουν πλήρως στην αναγνώριση και κατηγοριοποίηση των κλάσεων μειοψηφίας, δηλαδή των θετικών και αρνητικών συναισθημάτων. Όσον αφορά την αρχιτεκτονική των μοντέλων ανάλυσης συναισθήματος, οι διαφορές στις επιδόσεις τους ήταν αρκετά μικρές. Ωστόσο, πέρα από την ικανότητα ορθής ταξινόμησης, σημαντική συνιστώσα που καθιστά ένα μοντέλο αποδοτικό είναι και αυτή της ταχύτητας εκπαίδευσής του. Σε γενικές γραμμές, η χρήση της συνάρτησης ενεργοποίησης ReLU φάνηκε να βελτιώνει τόσο τον χρόνο εκμάθησης, όσο και τα αποτελέσματα του μοντέλου, ενώ με την επιπλέον εφαρμογή κανονικοποίησης επιπέδου, τα αποτελέσματα που προκύπτουν είναι ακόμα καλύτερα, με την προσέγγιση p_2 να ξεχωρίζει. Εξετάστηκε επιπλέον και η εφαρμογή δεύτερου κρυφού επιπέδου στις υφιστάμενες αρχιτεκτονικές δομές. Η προσθήκη αυτή, φάνηκε να αναβαθμίζει σε γενικές γραμμές την απόδοση του ταξινομητή. Στα πλαίσια της μελέτης μας, αντί της ακολουθίας εξόδου του γλωσσικού μοντέλου απόκρυψης, ως είσοδος στην κεφαλή ταξινόμησης χρησιμοποιήθηκε επιπλέον και η δειγματοληπτημένη έξοδος ακολουθίας των προ-εκπαιδευμένων μοντέλων. Το βασικό συμπέρασμα αυτού του πειράματος ήταν πως η ταχύτητα εκπαίδευσης των παραπάνω μοντέλων αυξήθηκε κατά πολύ, μειώνοντας το πλήθος απαιτούμενες εποχές στο $\frac{1}{3}$. Όσον αφορά την απόδοση, τα αποτελέσματα που λαμβάνουμε κυμαίνονται σε παρόμοια επίπεδα με τις προηγούμενες περιπτώσεις (δεν παρατηρείται, δηλαδή, κάποια επιπλέον βελτίωση). Τέλος, όσον αφορά την ανάπτυξη του GreekSocialBERT, αν και τα αποτελέσματα που απέδωσε ήταν σχετικά καλύτερα από αυτά του αρχικού μοντέλου, οι διαφορές στην αποδοτικότητά τους δεν ήταν ιδιαίτερα μεγάλες, παρά το γεγονός ότι η προ-εκπαίδευση του GreekBERT έχει γίνει σε ένα κατά πολύ μεγαλύτερο σώμα κειμένου. Συνεπώς, επαληθεύεται για ακόμα μία φορά η ιδιαίτερη σημασία της συμβατότητας του αντικειμένου της έρευνας που διεξάγεται με τα δείγματα εισαγωγής. Για το συγκεκριμένο πεδίο έρευνας, πιθανή αναβάθμιση θα μπορούσε να είναι και η δοκιμή διαφορετικών μοντέλων μετασχηματιστή, πέρα από τον RoBERTa.

5.2 Μελλοντικές Επεκτάσεις

Η μελέτη που πραγματοποιήθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας, εξετάζει διάφορους πιθανούς τρόπους βελτιστοποίησης ενός γλωσσικού μοντέλου ανάλυσης συναισθήματος. Η περιορισμένη χρήση της ελληνικής γλώσσας σε παγκόσμια κλίμακα, ενδέχεται πολλές φορές να αποτελεί εμπόδιο για έρευνες σχετικές με ΕΦΓ. Για τον λόγο αυτό, τα αποτελέσματα της πειραματικής διαδικασίας που πραγματοποιήθηκε μπορούν να αποτελέσουν ένα χρήσιμο σημείο αναφοράς για μελλοντικές εργασίες. Η έρευνά μας σαφώς, θα μπορούσε

να επεκταθεί και ακόμα περισσότερο. Ενδεικτικά, όπως διαπιστώθηκε και από την ανάλυσή μας, η βασικότερη παράμετρος που επηρεάζει τις επιδόσεις του γλωσσικού μοντέλου, είναι τα χαρακτηριστικά των δεδομένων πάνω στα οποία γίνεται η εκπαίδευσή του. Επομένως, ένα επόμενο βήμα στην εξέλιξη της προσέγγισής μας, είναι να εμπλουτίσουμε ακόμα παραπάνω τη συλλογή δεδομένων αλλά και το σώμα κειμένου που αφορά την προ-εκπαίδευση. Εκτός από την αύξηση του όγκου των δειγμάτων, μια ακόμα κατεύθυνση που θα μπορούσαμε να ακολουθήσουμε, είναι να προσπαθήσουμε να «εντείνουμε» την συναισθηματική φόρτιση των κειμένων, εκμεταλλευόμενοι την χρήση των emoticons που συναντάμε συχνά στα μέσα κοινωνικής δικτύωσης. Ειδικότερα, σε αντίθεση με την παρούσα διαδικασία προ-επεξεργασίας των δεδομένων που ακολουθήσαμε, κατά την οποία έγινε απαλοιφή των συγκεκριμένων γνωρισμάτων, θα μπορούσαμε να αντιστοιχίσουμε το νόημά αυτών με λεκτικούς όρους που το αποτυπώνουν. Πέρα από την επεξεργασία των δεδομένων, εναλλακτικοί τρόποι αναβάθμισης των μοντέλων είναι και ο περαιτέρω πειραματισμός όσον αφορά την δομή τους. Το συγκεκριμένο πεδίο έρευνας μπορεί να ξεκινήσει από την απλή δοκιμή μιας εναλλακτικής συνάρτησης ενεργοποίησης και να επεκταθεί μέχρι και την ολική αλλαγή αρχιτεκτονικής μετασχηματιστή.

Βιβλιογραφία

[1] Palo services ltd. [Online; accessed March 1, 2023].

[2] What exactly is an n gram?, 2014. [Online; accessed March 1, 2023].

[3] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[4] J. Abadji, P. Ortiz Suarez, L. Romary, and B. Sagot. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, page arXiv:2201.06642, Jan. 2022.

[5] J. Alammar. The illustrated transformer, 2018. [Online; accessed March 1, 2023].

[6] G. Alexandridis, K. Michalakis, J. Aliprantis, P. Polydoras, P. Tsantilas, and G. Caridakis. A deep learning approach to aspect-based sentiment prediction. In I. Maglogiannis, L. Iliadis, and E. Pimenidis, editors, *Artificial Intelligence Applications and Innovations*, pages 397–408, Cham, 2020. Springer International Publishing.

[7] G. Alexandridis, I. Varlamis, K. Korovesis, G. Caridakis, and P. Tsantilas. A survey on sentiment analysis and opinion mining in greek social media. *Information*, 12(8), 2021.

[8] A. Bhardwaj. Natural language processing: The method behind understanding humans and machines, 2020. [Online; accessed March 1, 2023].

[9] M. M. Bradley and P. J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[10] A. Chadha. Word embeddings and vector spaces. *Distilled Notes for the Natural Language Processing Specialization on Coursera (offered by deeplearning.ai)*, 2020. `https://aman.ai`.

[11] R. Cowie, E. Douglas-Cowie, S. Savvidou, and E. McMahon. Feeltrace: an instrument for recording perceived emotion in real time. 2000.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[13] P. Ekman. Facial expressions of emotion: New findings, new questions. *Psychological Science*, 3(1):34–38, 1992.

[14] R. et al. Improving language understanding by generative pre-training, 2018. [Online; accessed March 1, 2023].

[15] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224, 2017.

[16] A. Hoonlor, J. Mohammed, M. Zaki, and W. Wallace. Sequential patterns and temporal patterns for text mining. 03 2023.

[17] J. Koutsikakis, I. Chalkidis, P. Malakasiotis, and I. Androutsopoulos. Greek-bert: The greeks visiting sesame street. In *11th Hellenic Conference on Artificial Intelligence*, SETN 2020, page 110–117, New York, NY, USA, 2020. Association for Computing Machinery.

[18] W. Ling, C. Dyer, A. Black, and I. Trancoso. Two/too simple adaptations of word2vec for syntax problems. 05 2015.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[20] M. Lombard, R. Reich, M. Grabe, C. Bracken, and T. Bolmarcich. Presence and television. *Human Communication Research*, 26:75 – 98, 01 2000.

[21] D. Lottridge, M. Chignell, and A. Jovicic. *Affective Interaction Understanding, Evaluating, and Designing for Human Emotion*, volume 7, pages 197–217. 08 2011.

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

[23] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pretraining distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch:

An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019.

[25] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[26] K. Perifanos and D. Goutsos. Multimodal hate speech detection in greek social media. *Multimodal Technologies and Interaction*, 5(7):34, 2021.

[27] R. Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350, 2001.

[28] A. Radford and K. Narasimhan. Improving language understanding by generative pre-training. 2018.

[29] R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[30] S.-H. Tsang. Review — bert: Pre-training of deep bidirectional transformers for language understanding, 2022. [Online; accessed March 1, 2023].

[31] S.-H. Tsang. Review — elmo: Deep contextualized word representations, 2022. [Online; accessed March 1, 2023].

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[33] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019.

Γλωσσάριο

Ελληνικός όρος	Ξενόγλωσσος όρος
ακολουθιακή έξοδος	sequence output
ακρίβεια	precision
αμφίδρομο	bidirectional
ανάκληση	recall
ανάλυση συναισθήματος	sentiment analysis
αναπαράσταση	representation
αναφορά	mention
αντίστροφη συχνότητα κειμένου	inverse document frequency
απόκρυψη	masking
αρνητικό	negative
αρχιτεκτονική	architecture
αυτο-προσοχή	self-attention
βάρη	weights
βαθιά μάθηση	deep learning
βραχυχρόνια	short-term
γλωσσικό μοντέλο	language model
γνώρισμα	attribute
δεδομένα	data
δυναμικός	dynamic
εκπαίδευση	training
εκτός λεξικού	out of vocabulary
έλεγχος	test
ελληνικά	greek
εμπρόσθια τροφοδότηση	feed-forward
ενσωμάτωνση συμφραζομένων	contextualized embeddings
εξαγωγή οντοτήτων	named entity recognition
εξαγωγή πληροφορίας	text extraction
επαλήθευση	validation
επεξεργασία φυσικής γλώσσας	natural language processing
επιβλεπόμενη	supervised

επισημειωμένο	labeled-annotated
εποχές	epochs
ερωταπόκριση	question answering
εταιρίες-μάρκες	brands
θέματα συζήτησης	topics
θετικό	positive
κανονικοποίηση επιπέδου	layer normalization
κείμενο	text
κενοί χαρακτήρες	null
κλάση	class
κωδικοποίηση	encoding
μέσα κοινωνικής δικτύωσης	social media
μακρυχρόνια	long-term
μετασχηματιστής	transformer
μεταφορά μάθησης	transfer learning
μη-γραμμική συνάρτηση ενεργοποίησης	non-linear activation function
μη-επιβλεπόμενη	unsupervised
μηχανική μάθηση	machine learning
μηχανική μετάφραση	language translation
μνήμη	memory
νευρωνικό δίκτυο	neural network
ουδέτερο	neutral
Παραγωγικός Προ-Εκπαιδευμένος Μετασχηματιστής	Generative Pre-Training Transformer
πίνακας σύγχυσης	confusion matrix
πηγή πληροφορίας	source
πλήρως συνδεδεμένο επίπεδο	fully-connected layer
πλατφόρμα	platform
πολλαπλή κεφαλή προσοχής	multi-head attention
προεκπαίδευση	pretraining
προεπεξεργασία	preprocessing
προκατειλημμένος	biased
προσαρμογή	fine-tuning
προσοχή	attention
πρόωρος τερματισμός	early stopping
πρόβλεψη επόμενης λέξης	next sentence prediction
πρόβλημα εξαφανιζόμενων κλίσεων	vanishing gradient problem
σακούλι λέξεων	bag of words
συνεχής κλίμακα	continuous scale
συντακτική ανάλυση	parsing
συχνότητα όρων	term frequency
σφάλμα	loss

σύνολο	set
σύνολο δεδομένων	dataset
σώμα κειμένου	corpus
ταξινομητής	classifier
ταξινόμηση	classification
τεχνητή νοημοσύνη	artificial intelligence
χαμηλής πυκνότητας	low density

Συντομεύσεις - Αρκτικόλεξα

βλπ	βλέπε
Ε.Μ.Π	Εθνικό Μετσόβιο Πολυτεχνείο
ΕΦΓ	Επεξεργασία Φυσικής Γλώσσας
κ.ά	και άλλα
κ.λ.π	και λοιπά
κ.ο.κ	και ούτω καθεξής
ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
biLSTM	bidirectional long short term memory
BoW	Bag-of-Words
BPE	Byte pair encoding
CBOW	Continuous Bag-of-Words
ELMo	Embedding from Language Models
GloVe	Global Vectors of Word Representation
GLUE	General Language Understanding Evaluation
GPT	Generative Pre-Training Transformer
IDF	Inverse Document Frequency
LayerNorm	Layer Normalization
MLM	Masked Language Modeling
NER	Named-Entity Recognition
NLP	Natural Language Processing
NSP	Next Sentence Prediction
OOV	Out-of-vocabulary
PoS	Part-of-Speech
Q&A	Question & Answering
ReLU	Rectified Linear Unit
RNN	Recurring Neural Network
RoBERTa	Robustly Optimized BERT
RT	Retweet
SQuaD	Standford Question Answering Dataset

TF	Term Frequency
UNK	Unkown
URL	Uniform Resource Locator
VSM	Vector Space Model