**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ**

**ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**

## Πτυχιακή Εργασία

| | |
|---|---|
| Τίτλος Πτυχιακής Εργασίας | **Ανάλυση Συναισθήματος σε κειμενικά δεδομένα με χρήση ταξινομητών BERT**<br><br>**Sentiment analysis on textual data using BERT classification Models** |
| Ονοματεπώνυμο Φοιτητή | **Αιμιλιανός Κουρπάς-Δανάς** |
| Πατρώνυμο | **Θεόδωρος** |
| Αριθμός Μητρώου | **Π-20100** |
| Επιβλέπων | **Διονύσιος Σωτηρόπουλος, Επίκουρος Καθηγητής** |

Ημερομηνία Παράδοσης **Ιούλιος 2024**

# Copyright ©

# Ευχαριστίες

Θέλω θερμά να ευχαριστήσω τον επιβλέποντα καθηγητή μου, Διονύση Σωτηρόπουλο, για την ανεκτίμητη υποστήριξη και την επιμονή του καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας. Εκφράζω επίσης τις ευχαριστίες μου προς όλους τους καθηγητές του τμήματος για τις αξιοσημείωτες γνώσεις που μου προσέφεραν κατά τη διάρκεια της φοίτησής μου.Επιπλέον, εκφράζω τις ειλικρινείς μου ευγνωμοσύνες προς τους γονείς μου για την συνεχή ενθάρρυνση τους όλα αυτά τα χρόνια, παρά τα εμπόδια που παρουσιάστηκαν στο δρόμο.

# Περίληψη

Η έρευνα στον τομέα της ανάλυσης συναισθήματος αποτελεί καθιερωμένο πεδίο στην επεξεργασία της φυσικής γλώσσας, επιδιώκοντας την αναγνώριση των συναισθημάτων που εκφράζονται σε ένα κείμενο. Η αποτελεσματική εφαρμογή μοντέλων ανάλυσης συναισθήματος μπορεί να αποτελέσει ένα χρήσιμο εργαλείο για την καλύτερη κατανόηση των αναγκών, απόψεων, συμπεριφορών και προτιμήσεων του ευρύτερου κοινού.

Για τον σκοπό αυτό, η ανάπτυξη των μοντέλων βασίστηκε στην προηγμένη αρχιτεκτονική των δικτύων μετασχηματιστών, με έμφαση στο εξειδικευμένο γλωσσικό μοντέλο βαθιάς μάθησης Greek-BERT. Αυτή η προσέγγιση επιτρέπει την υλοποίηση, εκπαίδευση και αξιολόγηση μοντέλων μηχανικής μάθησης με στόχο την ακριβή κατάταξη του περιεχομένου ενός κειμένου ως θετικό, αρνητικό ή ουδέτερο. Κεντρικό σημείο της έρευνας αποτελεί η εφαρμογή της τεχνικής ανάλυσης συναισθήματος σε αποσπάσματα κειμένου στην ελληνική γλώσσα, καταγεγραμμένα από μέσα κοινωνικής δικτύωσης.

*Λέξεις Κλειδιά:* Ανάλυση Συναισθήματος, Εξόρυξη Γνώμης, Βαθιά Μηχανική Μάθηση, Μετασχηματιστές, Επεξεργασία Φυσικής Γλώσσας

# Abstract

Research in the field of sentiment analysis is an established field in natural language processing, seeking to identify the emotions expressed in a text. The effective application of sentiment analysis models can be a useful tool to better understand the needs, opinions, attitudes and preferences of the public.

To this end, the development of the models was based on the advanced transformer network architecture, with a focus on the specialized deep learning language model Greek-BERT. This approach enables the implementation, training and evaluation of machine learning models to accurately classify the content of a text as positive, negative or neutral. The central focus of the research is the application of the sentiment analysis technique to text excerpts in Greek, recorded from social media.

*Key Words:* Sentiment Analysis, Opinion Mining, Deep Learning, Transformers, Natural Language Processing

# Πίνακας Περιεχομένων

# Κατάλογος Εικόνων

# Κατάλογος Πινάκων

# Κατάλογος Διαγραμμάτων

# Κεφάλαιο 1: Εισαγωγή στην Ανάλυση Συναισθήματος

## 1.1 Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος αποτελεί έναν κλάδο της επεξεργασίας φυσικής γλώσσας που επικεντρώνεται στην εξέταση απόψεων, συναισθημάτων, προτιμήσεων και συμπεριφορών ανθρώπων σχετικά με διάφορα θέματα, όπως προϊόντα, υπηρεσίες, οργανώσεις, φυσικά πρόσωπα, συζητήσεις και γεγονότα. Κατά τη διαδικασία αυτή, τα αποτελέσματα μπορούν να ταξινομούνται με διάφορους τρόπους, όπως με τη χρήση απλών δυαδικών κατανομών (αρνητικό - θετικό), τη διαίρεση σε τρεις κλάσεις (αρνητικά, θετικά και ουδέτερα συναισθήματα), ή ακόμα και την κατηγοριοποίηση σε ένα ευρύ φάσμα συναισθημάτων, όπως χαρά, ενθουσιασμός, θυμός, λύπη και άλλα.

Η ανάπτυξη του πεδίου συνδέεται με την άνοδο των κοινωνικών δικτύων, μέσα από τα οποία κάθε άτομο έχει πρόσβαση σε πληθώρα πηγών πληροφορίας που καλύπτουν κάθε πτυχή της ζωής. Η ανάλυση συναισθήματος έχει εξελιχθεί σε ευρύ πεδίο έρευνας και μπορεί να εφαρμοστεί σε πολλούς τομείς της καθημερινότητάς μας, όπως ψυχαγωγία, ενημέρωση, καταναλωτικές συμπεριφορές, επιχειρηματικές και πολιτικές στρατηγικές.

Στη σύγχρονη εποχή, τα κοινωνικά δίκτυα κυριαρχούν ως μέσο επικοινωνίας και ενημέρωσης, συγκεντρώνοντας τεράστιο όγκο πληροφοριών που αντιπροσωπεύουν διάφορες απόψεις της κοινής γνώμης. Με τη χρήση τεχνικών ανάλυσης συναισθήματος και αποτελεσματικής διαχείρισης των δεδομένων, επιτυγχάνεται βαθύτερη κατανόηση των απόψεων και προτιμήσεων του κοινού, προσφέροντας ευκαιρίες για ευρεία εφαρμογή σε ποικίλους τομείς.

## 1.2 Πρακτικές Εφαρμογές

Σε προσωπικό επίπεδο, η ανάλυση συναισθημάτων παρέχει συμβουλευτική πληροφορία. Παρακολουθώντας τις τάσεις και απόψεις στα κοινωνικά μέσα, ο κάθε χρήστης αποκτά σημαντική πληροφορία που καθοδηγεί τις επιλογές του σχετικά με αγαθά και υπηρεσίες. Η ανάλυση αυτή δημιουργεί μια αντιπροσωπευτική εικόνα της κοινής γνώμης, η οποία επηρεάζει αποφάσεις σε συλλογικό επίπεδο.

Εφαρμογές της ανάλυσης συναισθήματος στα μέσα κοινωνικής δικτύωσης παρακολουθούν συζητήσεις για προϊόντα και υπηρεσίες, παρέχοντας στις επιχειρήσεις ενημερωμένες πληροφορίες για την ανταπόκριση του κοινού. Οι αλγόριθμοι αναγνώρισης συναισθήματος επιτρέπουν τη διαμόρφωση αντιπροσωπευτικής εικόνας και τη λήψη αποφάσεων πάνω σε προϊόντα, υπηρεσίες και πρόσωπα.

Η ανάλυση συναισθημάτων βελτιώνει την ανταπόκριση σε σχόλια, εντοπίζοντας αστοχίες και διασφαλίζοντας εποικοδομητική αλληλεπίδραση με πελάτες/καταναλωτές. Οι εφαρμογές αυτές παρέχουν στρατηγικά δεδομένα για επιχειρηματικές αποφάσεις, βασισμένες σε πραγματικά και αξιόπιστα στοιχεία.

Η ανάγκη για αυτοματοποιημένα συστήματα τεχνητής νοημοσύνης, όπως η ανάλυση συναισθήματος, αυξάνεται λόγω του ραγδαίου εξελισσόμενου αριθμού χρηστών στα κοινωνικά μέσα. Αυτές οι πρακτικές εφαρμογές καλούνται να καλύψουν τη ζήτηση για αξιόπιστες υπηρεσίες ανάλυσης συναισθήματος, προσφέροντας στις επιχειρήσεις ενημερωμένες και εξειδικευμένες επιλογές.

## 1.3 Αντικείμενο Διπλωματικής Εργασίας

Το αντικείμενο της παρούσας πτυχιακής εργασίας εστιάζεται στην εκπαίδευση και αξιολόγηση του ελληνικού BERT μοντέλου, με στόχο την προσαρμογή του για την ανάλυση συναισθήματος σε ελληνικό κείμενο στο Twitter. Η διαδικασία περιλαμβάνει την προ-εκπαίδευση του μοντέλου και την προσαρμογή του για την ταξινόμηση του συναισθήματος σε θετικό, αρνητικό ή ουδέτερο επίπεδο.

Κατανοούμε τις προκλήσεις που προκύπτουν από τον περιορισμένο όγκο κειμένων στα ελληνικά, την ποικιλία των χαρακτηριστικών της γλώσσας, καθώς και τις ειδικές παραμέτρους που επηρεάζουν τα κείμενα στο Twitter.

## 1.4 Διάρθρωση Εργασίας

Η Διπλωματική Εργασία διαρθρώνεται σε 5 κεφάλαια. Στο Κεφάλαιο 1 προσφέρει μια εισαγωγή στον κλάδο. Στη συνέχεια, αναλύονται με επιμέλεια τα κεφάλαια που ακολουθούν. Στο Κεφάλαιο 2 ,ως θεωρητικό υπόβαθρο, εξετάζεται η επεξεργασία φυσικής γλώσσας και οι τεχνικές προεπεξεργασίας, με επιμέρους αναφορές σε συντακτική και σημασιολογική ανάλυση, καθώς και μοντέλα εξαγωγής χαρακτηριστικών, όπως η διακριτή αναπαράσταση κειμένου και η διανυσματική αναπαράσταση

κειμένου με τη χρήση μοντέλων όπως το Word2Vec και το GloVe. Στο Κεφάλαιο 3 εξετάζονται λεπτομερώς οι μέθοδοι αναπαράστασης συναισθημάτων, με ενότητες για αναπαράσταση σε κατηγορίες και διαστατικές αναπαραστάσεις. Περιλαμβάνονται επίσης θέματα όπως η ανάλυση συναισθημάτων στα μέσα κοινωνικής δικτύωσης και η σύνδεση της ανάλυσης συναισθήματος με τη μηχανική μάθηση. Στο Κεφάλαιο 4 παρουσιάζονται και αναλύονται προ-εκπαιδευμένα μοντέλα όπως το ELMo, οι Μετασχηματιστές (Transformers), GPT, BERT, και RoBERTa. Στο Κεφάλαιο 5 περιγράφονται λεπτομερώς οι εργαλεία και η μεθοδολογία που χρησιμοποιήθηκαν στην πειραματική διαδικασία, περιλαμβάνοντας περιγραφή δεδομένων, προεπεξεργασία, λεπτή προσαρμογή, και αξιολόγηση των αποτελεσμάτων. Το Κεφάλαιο 6, 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

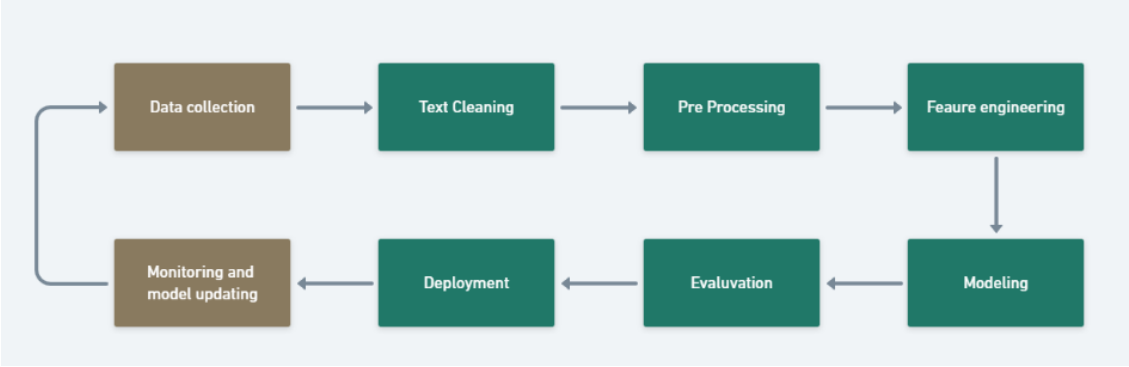
Η επεξεργασία φυσικής γλώσσας (ΕΦΓ) ανήκει στον τομέα της τεχνητής νοημοσύνης και ασχολείται με την αναγνώριση, κατανόηση και επεξεργασία της ανθρώπινης γλώσσας από υπολογιστικές μηχανές. Εφαρμόζοντας τεχνικές ΕΦΓ, από την ανάλυση συναισθήματος (sentiment analysis) έως την αυτόματη αναγνώριση ομιλίας (speech recognition), ο υπολογιστής εκμεταλλεύεται τη γραμματική, το συντακτικό, και το "νόημα" (context) των δεδομένων φυσικής γλώσσας.

Στα πλαίσια της ΕΦΓ, ενδεικτικά πεδία έρευνας περιλαμβάνουν:

- Ανάλυση συναισθήματος - Εξόρυξη γνώμης (Sentiment Analysis - Opinion Mining): κατανόηση του συναισθήματος που εκφράζεται σε ένα κείμενο ως θετικό, αρνητικό, ουδέτερο κ.λπ.
- Εξαγωγή Κειμένου (Text Extraction): ανάκτηση πληροφοριών από κείμενα γραμμένα σε φυσική γλώσσα.
- Απάντηση Ερωτήσεων (Question Answering - Q&A): εύρεση σωστής απάντησης σε ερώτηση που διατυπώνεται σε ανθρώπινη γλώσσα.
- Εξαγωγή Επώνυμων Οντοτήτων (Named Entity Recognition - NER): εντοπισμός και κατηγοριοποίηση πληροφορίας σε κατηγορίες οντοτήτων.
- Μηχανική Μετάφραση (Language Translation): μετάφραση κειμένου από μια ανθρώπινη γλώσσα σε μια άλλη.

Η βασική προσέγγιση για την ανάπτυξη εφαρμογών επεξεργασίας φυσικής γλώσσας (ΕΦΓ), όπως παρουσιάζεται στην Εικόνα 1, ξεκινά με τη συλλογή ενός όσο το δυνατόν αντιπροσωπευτικού συνόλου κειμένου. Αυτό το κείμενο πρέπει να περιέχει την κατάλληλη πληροφορία για τον εκάστοτε σκοπό. Σημαντικοί παράγοντες, όπως οι πηγές, το μέγεθος και τα γλωσσολογικά χαρακτηριστικά του κειμένου, επηρεάζουν την αποτελεσματικότητα της μοντελοποίησης.

Στην περίπτωση που η επεξεργασία φυσικής γλώσσας στοχεύει σε ανάλυση συναισθήματος, εξαγωγή επώνυμων οντοτήτων κ.λπ., η κατάλληλη επεξεργασία των δεδομένων είναι ζωτικής σημασίας για την σωστή κατηγοριοποίησή τους. Το αρχικό κείμενο μπορεί να βελτιωθεί με διορθώσεις ορθογραφικών λαθών, αφαίρεση διπλότυπων εισαγωγών, κ.α. Στη συνέχεια, η προεπεξεργασία επικεντρώνεται στη μετατροπή του κειμένου για ευκολότερη χειραγώγηση από υπολογιστικές μηχανές. Τέλος, το κείμενο εισάγεται σε κατάλληλο γλωσσικό μοντέλο για εκπαίδευση και αξιολόγηση.



**Εικόνα 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

Κατά τη συντακτική ανάλυση, γίνεται η εξέταση του κειμένου με βάση τους βασικούς γραμματικούς κανόνες. Σκοπός είναι να αποκαλυφθεί η δομή της πρότασης, να διασαφηνιστεί ο τρόπος οργάνωσης των λέξεων και να εξεταστεί ο συσχετισμός τους κατά τη χρήση σε μια πρόταση. Η ανάλυση σε σύμβολα (tokenization) αποτελεί βασική μέθοδο, διασπώντας το κείμενο σε μικρότερα κομμάτια, γνωστά ως "tokens," που μπορεί να είναι είτε ολόκληρες λέξεις, κομμάτια λέξεων ή ακόμα και προτάσεις. Επίσης, χρησιμοποιείται η επισήμανση μερών του λόγου (Part-of-Speech (PoS) tagging), κατηγοριοποιώντας τα σύμβολα ανάλογα με το συντακτικό τους ρόλο.

### 2.2.2 Σημασιολογική Ανάλυση

Στη σημασιολογική ανάλυση, εφαρμόζονται μέθοδοι που επιτρέπουν στον υπολογιστή να κατανοήσει βαθύτερα τη σημασία των λέξεων στο πλαίσιο μιας πρότασης. Αυτές περιλαμβάνουν τη λημματοποίηση, που επαναφέρει τις λέξεις στη ρίζα τους, και την αφαίρεση κοινών λέξεων (stop-word removal) που δεν προσφέρουν σημασιολογική αξία. Η σημασιολογική ανάλυση επιτρέπει στον υπολογιστή να κατανοήσει το περιεχόμενο με μεγαλύτερη ευαισθησία και ενδιαφέρον.

## 2.3 Μοντέλα Εξαγωγής Χαρακτηριστικών

Ένα κρίσιμο στάδιο στην επεξεργασία φυσικής γλώσσας είναι ο τρόπος που το κείμενο αναπαρίσταται σε ένα γλωσσικό μοντέλο. Για να εκπαιδευτεί ένα μοντέλο πάνω σε ένα σώμα κειμένου, είναι απαραίτητη η αριθμητική αναπαράσταση των όρων που το αποτελούν. Σε μια ιδανική υλοποίηση, οι λέξεις που σχετίζονται μεταξύ τους αναπαρίστανται με διανύσματα κωδικοποιημένων όρων που τοποθετούνται κοντά το ένα στο άλλο. Συγχρόνως, λέξεις με μεγάλη σημασιολογική διαφορά αντιστοιχούν σε απομακρυσμένα στοιχεία του διανυσματικού χώρου.

Η αναπαράσταση του κειμένου μπορεί να υιοθετεί δύο βασικές μορφές: τη Διακριτή (Discrete Text Representation) και την Κατανεμημένη (Distributed Text Representation). Στη Διακριτή Αναπαράσταση, κάθε λέξη θεωρείται ως ανεξάρτητη, και το σχετικό διάνυσμα συμπληρώνεται με βάση τις ιδιότητές της, όπως η συχνότητα εμφάνισης στο κείμενο. Αντίθετα, η Κατανεμημένη Αναπαράσταση λαμβάνει υπόψη τα συμφραζόμενα και αναπαριστά τις σημασιακές σχέσεις μεταξύ των λέξεων.

Η αναπαράσταση του κειμένου με αυτούς τους τρόπους εξαρτάται από τη φύση της φυσικής γλώσσας, καθώς 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. Τα μοντέλα εξαγωγής χαρακτηριστικών που χρησιμοποιούνται για αυτές τις αναπαραστάσεις αναλύονται παρακάτω.

### 2.3.1 Διακριτή Αναπαράσταση Κειμένου

#### 2.3.1.1 Κωδικοποίηση One-hot

Η τεχνική της κωδικοποίησης one-hot αναπαριστά τις λέξεις ενός κειμένου σε έναν διανυσματικό χώρο V διαστάσεων. Αυτό το διάνυσμα αποτελείται από την κωδικοποιημένη αναπαράσταση όλου του λεξιλογίου που περιλαμβάνει το κείμενο που μελετούμε. Κάθε λέξη αναπαρίσταται από ένα διανυσματικό σημείο, $w_{vx1}$ , στον χώρο, με δυαδική αριθμητική μορφή. Συγκεκριμένα, κάθε διάνυσμα συμπληρώνεται με μηδενικά σε όλες τις θέσεις εκτός από μία, η οποία αντιστοιχεί στη λέξη που κωδικοποιείται, και συμπληρώνεται με τον αριθμό 1.

Ας υποθέσουμε τη φράση "Natural Language Processing empowers machines." Η one-hot κωδικοποίηση για αυτήν τη φράση θα είναι:

Natural → [100000]
Language → [010000]
Processing → [001000]

empowers → [000100]
machines → [000010]

Κάθε λέξη αναπαρίσταται από ένα διάνυσμα, με το "1" στη θέση που αντιστοιχεί στην κωδικοποιημένη λέξη. Έτσι, το συνολικό διανυσματικό αναπαραστατικό της πρότασης θα είναι:

sentence = [[1, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0], [0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 1, 0]]

Αυτή η προσέγγιση, ωστόσο, αντιμετωπίζει προβλήματα όταν εφαρμόζεται σε μεγάλα σώματα κειμένου. Απαιτούνται πολύ μεγάλα διανύσματα V τάξης, καθώς κάθε λέξη αντιστοιχεί σε ένα μοναδικό διάνυσμα. Για παράδειγμα, ένα λεξικό με 100,000 λέξεις θα απαιτούσε ένα διάνυσμα διάστασης 100,000 για την κωδικοποίηση μιας μόνο λέξης, επιβαρύνοντας την υπολογιστική ισχύ και τη μνήμη. Παράλληλα, το πρόβλημα εντοπίζεται και στην εισαγωγή one-hot διανυσμάτων στα γλωσσικά μοντέλα, καθώς οι διαστάσεις της εισόδου διαφέρουν ανάλογα με τον αριθμό των συμβόλων του κάθε κειμένου και το μέγεθος του λεξικού.

Μια εναλλακτική λύση είναι η μετατροπή των δειγμάτων σε διανύσματα με σταθερό μέγεθος L. Σε περίπτωση υπέρβασης του μέγιστου μήκους, μπορεί να γίνει περικοπή (truncation), ενώ σε περίπτωση μικρότερου μήκους, μπορεί να γίνει επέκταση με πρόσθεση συμβόλων γεμίσματος (padding).

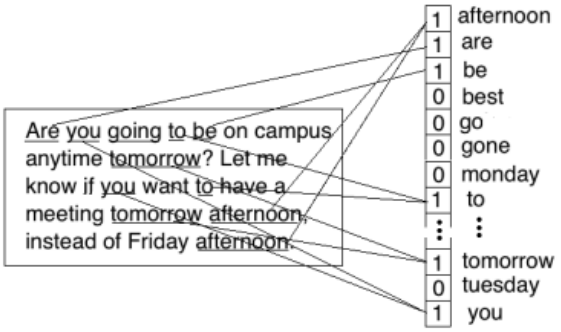
Ωστόσο, η one-hot κωδικοποίηση παρουσιάζει περιορισμούς, καθώς αγνοεί τις σημασιολογικές σχέσεις μεταξύ των λέξεων και δεν προσφέρει αντιπροσωπευτική αριθμητική αναπαράσταση του κειμένου. Για να αντιμετωπίσουμε αυτά τα προβλήματα, εξετάζουμε την τεχνική του Bag-of-Words (BoW). Αυτό το μοντέλο επικεντρώνεται στον υπολογισμό της συχνότητας εμφάνισης λέξεων σε ένα κείμενο. Ο όρος "σακούλι λέξεων" προέρχεται από το ότι λαμβάνονται υπόψη μόνο οι λέξεις ως ανεξάρτητες μονάδες, χωρίς να λαμβάνεται υπόψη η δομή ή η διάταξή τους στις προτάσεις που ανήκουν.

Παρακάτω περιγράφονται διάφορες υλοποιήσεις αυτού του μοντέλου(BoW).

### 2.3.1.2 Συχνότητα εμφάνισης λέξης

Μια από τις πιο απλές τεχνικές αναπαράστασης κειμένου είναι η χρήση της συχνότητας εμφάνισης λέξεων. Σε αντίθεση με την one-hot κωδικοποίηση, όπου κάθε λέξη αναπαρίσταται ως ένα διανυσματικό σύνολο, εδώ η εισαγωγή των λέξεων στο γλωσσικό μοντέλο γίνεται με διανυσματική αναπαράσταση, λαμβάνοντας υπόψη τη συχνότητα εμφάνισής τους.

Η υλοποίηση αρχικά δημιουργεί ένα λεξικό μεγέθους V, που περιλαμβάνει όλες τις μοναδικές λέξεις στο κείμενο. Κάθε λέξη αντιστοιχεί σε μια αριθμητική τιμή, που αντιπροσωπεύει τη συχνότητα εμφάνισής της στο κείμενο. Έτσι, το κείμενο αναπαρίσταται ως ένα σύνολο διανυσμάτων μεγέθους V, και κάθε διάνυσμα αντιπροσωπεύει αριθμητικά τα επιμέρους κείμενα που το συνθέτουν.




**Εικόνα 2: Παράδειγμα BoW υλοποίησης με χρήση συχνότητας εμφανίσεων λέξεων [1]**

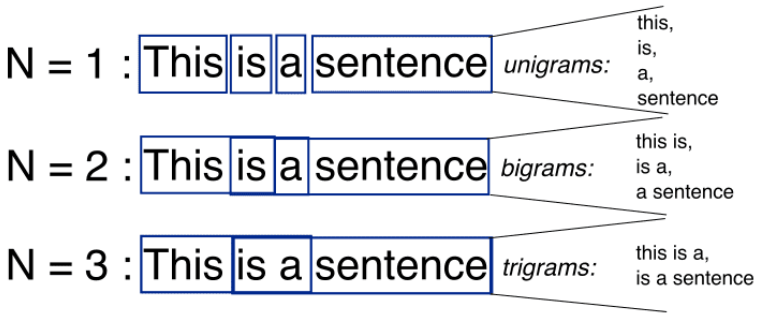
### 2.3.1.3 N-Grams

Σε αυτήν την προσέγγιση, όπως και με τη συχνότητα εμφάνισης λέξεων, υπολογίζεται η συχνότητα εμφάνισης όρων, αλλά με μια παραλλαγή. Αντί να δημιουργείται ένα λεξικό που περιλαμβάνει όλες τις μοναδικές λέξεις του κειμένου, κάθε όρος του λεξικού αποτελείται από έναν συνδυασμό N λέξεων, όπου N = 2, 3, 4, κλπ.

Για την εφαρμογή της τεχνικής N-Gram, πρέπει να οριστεί η παράμετρος N κατάλληλα, ώστε να προκύπτει ένα δίγραμμα (N = 2), τρίγραμμα (N = 3), και ούτω καθεξής.

Αυτή η υλοποίηση επιτυγχάνει σημαντική μείωση στο μέγεθος του λεξικού σε σχέση με την απλή εφαρμογή Bag-of-Words, καθώς επίσης προσφέρει βαθύτερη κατανόηση της σημασιολογίας της γλώσσας. Αυτό συμβαίνει μέσω της συσχέτισης των λέξεων που εμφανίζονται μαζί στο κείμενο.

Ένα "σακούλι διαγραμμάτων," για παράδειγμα, μπορεί να παρουσιάζει καλύτερη απόδοση από την απλή χρήση της συχνότητας εμφάνισης μοναδικών λέξεων, καθώς λαμβάνει υπόψη τη σειρά και τη σημασιολογική σχέση των λέξεων στο κείμενο.




**Εικόνα 3: Παράδειγμα N-Gram[2]**

### 2.3.1.4 Συχνότητα όρου - Αντίστροφη συχνότητα κειμένου(TF-IDF)

Στην περίπτωση του TF-IDF, χρησιμοποιείται μια προηγμένη προσέγγιση για την αναπαράσταση του κειμένου σε μορφή "σακούλας λέξεων." Αντί να περιοριστεί στη συχνότητα εμφάνισης μιας λέξης, εφαρμόζει την συνδυασμένη μετρική TF-IDF, που λαμβάνει υπόψη τη σημασία μιας λέξης σε ένα κείμενο σε σχέση με ολόκληρο το σώμα των κειμένων.

Πιο συγκεκριμένα:

- Η συχνότητα όρου (TF) αφορά το πόσο συχνά εμφανίζεται μια λέξη σε ένα κείμενο, σε σχέση με το συνολικό αριθμό λέξεων στο κείμενο.

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

- Η αντίστροφη συχνότητα κειμένου (IDF) μετράει το πόσο σπάνια εμφανίζεται μια λέξη σε όλο το σώμα των κειμένων, δίνοντας μικρότερη βαρύτητα σε συχνά εμφανιζόμενες λέξεις.

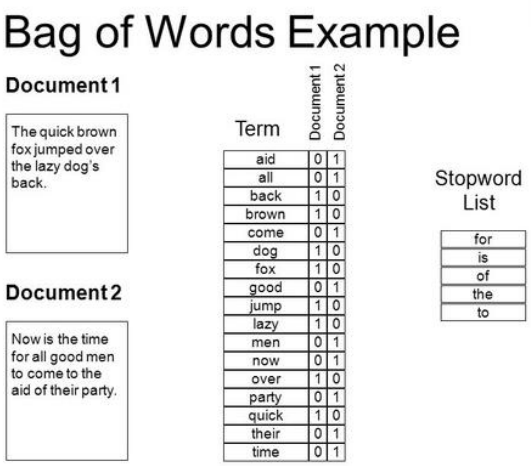
$$\mathrm{IDF}(t) = \ln\left(\frac{\text{συνολικός αριθμός κειμένων}}{\text{αριθμός κειμένων που περιέχουν τη λέξη } t}\right)$$

Η απλούστερη αναπαράσταση μιας λέξης είναι το TF-IDF. Αυτό προκύπτει από τον πολλαπλασιασμό δύο αριθμών: του TF, που μετράει πόσο συχνά εμφανίζεται η λέξη σε ένα κείμενο, και του IDF, που λαμβάνει υπόψη πόσο σπάνια είναι η λέξη σε όλο το σύνολο των κειμένων.

$$\text{TF-IDF}(t) = \mathrm{TF}(t) \times \mathrm{IDF}(t)$$

Ενώ οι υλοποιήσεις αυτές παρέχουν αριθμητικές αναπαραστάσεις που μπορούν να ενσωματωθούν εύκολα σε γλωσσικά μοντέλα, έχουν περιορισμούς. Το μοντέλο σακούλι λέξεων (BoW) αγνοεί τη δομή και τη διάταξη των λέξεων στο κείμενο, και η μεγάλη διάσταση του λεξικού μπορεί να οδηγήσει σε αραιά διανύσματα, επηρεάζοντας τη μνήμη και την υπολογιστική ισχύ.

Για τη βελτίωση της απόδοσης του BoW, μπορούν να εφαρμοστούν τεχνικές επεξεργασίας κειμένου, όπως η μετατροπή σε πεζούς χαρακτήρες και η αφαίρεση των σημείων στίξης, προκειμένου να μειωθεί το μέγεθος του λεξικού.



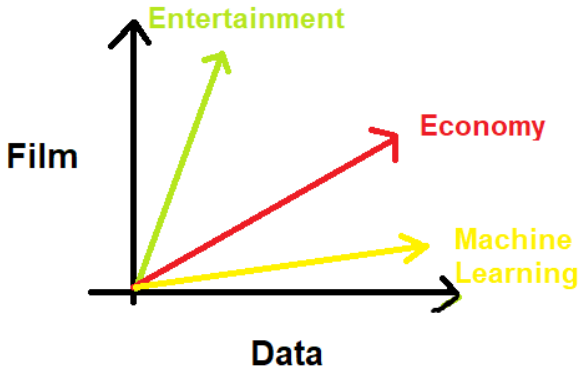

**Εικόνα 4: Υλοποίηση με απαλοιφή κοινών λέξεων [3]**

### 2.3.1.5 Μοντέλο Διανυσματικού Χώρου

Το προτεινόμενο μοντέλο διανυσματικού χώρου (Vector Space Model - VSM) αποτελεί μια προχωρημένη εκδοχή των τεχνικών κωδικοποίησης one-hot και σακκουλιού λέξεων. Σε αυτό το μοντέλο, προσφέρεται μια αναπαράσταση σε διανυσματική μορφή για κάθε απόσπασμα κειμένου ή ακόμη και για κάθε ξεχωριστή λέξη στο σώμα του κειμένου.

- Αναπαράσταση Αποσπάσματος: Σε αυτήν την αναπαράσταση, κάθε λέξη αναπαρίσταται από τη συχνότητα εμφάνισής της σε κάθε απόσπασμα κειμένου. Αποτελούνται διανύσματα με διαστάσεις DV, όπου οι όροι αντιστοιχούν στη συχνότητα εμφάνισης κάθε λέξης σε ένα συγκεκριμένο απόσπασμα.

- Αναπαράσταση Λέξης: Σε αυτήν την αναπαράσταση εξετάζονται οι φορές που μια λέξη εμφανίζεται κοντά σε μια άλλη εντός ενός σώματος κειμένου. Η απόσταση αυτή αντιστοιχεί στον αριθμό των λέξεων που μεσολαβούν, και ο πίνακας αναπαράστασης έχει διαστάσεις V×V , όπου V είναι το μέγεθος του λεξικού.




**Εικόνα 5:Παράδειγμα VSM**

Η υλοποίηση του VSM παρέχει έναν διανυσματικό χώρο που αντιπροσωπεύει τις σημασιολογικές σχέσεις εντός ενός σώματος κειμένου. Παραδείγματος χάριν, οι σχέσεις μεταξύ λέξεων, όπως "δεδομένα," "οικονομία," και "μηχανική μάθηση," είναι εμφανείς.

Ωστόσο, η VSM αντιμετωπίζει προκλήσεις όσον αφορά τους υπολογιστικούς πόρους λόγω των πολυδιάστατων και αραιών πινάκων που προκύπτουν, έχοντας κόστος $O(NM^2)$, όπου N είναι το μέγεθος του σώματος κειμένου και M το μέγεθος του λεξικού.

Για να αντιμετωπιστούν αυτά τα προβλήματα, προτείνεται η χρήση κατανεμημένης αναπαράστασης, που βασίζεται στην τεχνική των διανυσμάτων ενσωμάτωσης λέξεων (Word Embeddings).

### 2.3.2 Διανυσματική Αναπαράσταση Κειμένου

Η προσέγγιση της διανυσματικής αναπαράστασης κειμένου βασίζεται στη χρήση διανυσμάτων ενσωμάτωσης λέξεων, τα οποία λαμβάνουν υπόψη τα συμφραζόμενα κάθε λέξης, προσφέροντας αριθμητική αναπαράσταση της σημασιολογίας κάθε όρου. Αυτή η τεχνική βασίζεται στην υπόθεση ότι λέξεις που βρίσκονται κοντά η μία στην άλλη έχουν υψηλότερη πιθανότητα να σχετίζονται μεταξύ τους. Με τη χρήση διανυσμάτων ενσωμάτωσης λέξεων, κάθε λέξη αναπαρίσταται ως ένα πυκνό διάνυσμα συγκεκριμένων διαστάσεων. Αυτή η προσέγγιση αποφεύγει τη χρήση πολυδιάστατων και αραιών διανυσμάτων στο γλωσσικό μοντέλο, επιτρέποντας την εξοικονόμηση σημαντικών υπολογιστικών πόρων.

#### 2.3.2.1 Word2Vec

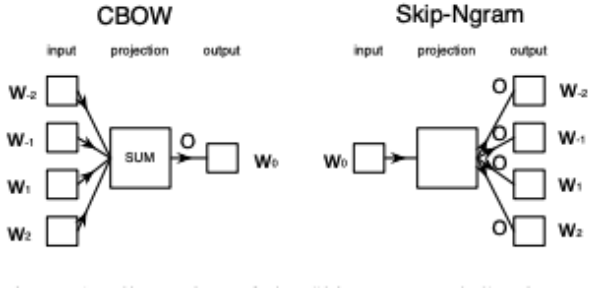
Ο αλγόριθμος Word2Vec [4]είναι μια μέθοδος για την αναπαράσταση λέξεων σε έναν διανυσματικό χώρο, βασιζόμενος στις συναφείς λέξεις που εμφανίζονται στο ίδιο κείμενο, δηλαδή στα συμφραζόμενα τους. Σε μεγάλα κείμενα, μπορεί να αναπαριστά όχι μόνο τη συντακτική δομή, αλλά και τις σημασιολογικές σχέσεις μεταξύ των λέξεων.

Ο αλγόριθμος δημιουργεί διανύσματα ενσωμάτωσης λέξεων, τα οποία τοποθετούνται σε έναν διανυσματικό χώρο έτσι ώστε σημασιολογικά συναφείς λέξεις να απεικονίζονται κοντά μεταξύ τους. Η αρχιτεκτονική του Word2Vec περιλαμβάνει ένα νευρωνικό δίκτυο τροφοδότησης με τρία επίπεδα: επίπεδο εισόδου, κρυφό επίπεδο και επίπεδο εξόδου.

Υπάρχουν δύο εκδοχές του αλγορίθμου Word2Vec: το μοντέλο συνεχούς σακούλας λέξεων (CBOW) και το μοντέλο πρόβλεψης νάνων (Skip-Gram). Στην περίπτωση του CBOW, προβλέπεται η αναπαράσταση της κεντρικής λέξης με βάση τα γειτονικά συμφραζόμενα. Στο μοντέλο Skip-Gram, προβλέπονται οι γειτονικές λέξεις με βάση την κεντρική λέξη.

Το Word2Vec μπορεί να αντιμετωπίσει ακόμα και σπάνιες λέξεις, αναπαριστώντας τόσο το συντακτικό τους ρόλο όσο και τη σημασία τους στο πλαίσιο ενός κειμένου. Ιδιαίτερα το μοντέλο Skip-Gram είναι αποτελεσματικό για την αναπαράσταση σπάνιων λέξεων, αν και χρειάζεται περισσότερο χρόνο για εκπαίδευση σε σύγκριση με το CBOW μοντέλο.

Το Word2Vec, σε αντίθεση με άλλες μεθόδους αναπαράστασης, διαχειρίζεται πιο αποδοτικά τα διανύσματα ενσωμάτωσης, αποφεύγοντας τη δημιουργία αραιών διανυσμάτων που μπορούν να επηρεάσουν την υπολογιστική απόδοση και τον χρόνο επεξεργασίας. Παρόλα αυτά, οι αλγόριθμοι Word2Vec έχουν ορισμένα μειονεκτήματα, όπως η περιορισμένη αναπαράσταση σημασιολογικών σχέσεων σε τοπικό επίπεδο και η δυσκολία αναπαράστασης λέξεων εκτός του συνόλου εκπαίδευσης.




**Εικόνα 6:Αρχιτεκτονική Word2Vec[5]**

#### 2.3.2.2 GloVe

Όπως αναφέρθηκε προηγουμένως, οι μέθοδοι παραθύρου λέξεων αντιμετωπίζουν προκλήσεις στο να αξιοποιήσουν αποτελεσματικά μεγάλου όγκου πληροφορίας και περιορίζουν τις αναπαραστάσεις σε τοπικό επίπεδο. Αυτό συμβαίνει επειδή τα παράθυρα επιλέγονται γύρω από μια λέξη και δεν μπορούν να αντικατοπτρίσουν πλήρως το γενικό νόημα της σε ολόκληρο το κείμενο. Σε αντίθεση, το μοντέλο GloVe [6]δεν επικεντρώνεται μόνο στην τοπική πληροφορία που προκύπτει από τις γειτονικές λέξεις, αλλά ενσωματώνει στατιστικά από ολόκληρο το σώμα του κειμένου.

Συγκεκριμένα, τα διανύσματα ενσωμάτωσης δημιουργούνται βάσει της πιθανότητας συνεμφάνισης των λέξεων. Αυτή η πιθανότητα καθορίζεται από το πόσο συχνά μια λέξη εμφανίζεται στις λέξεις που περιβάλλουν μια άλλη λέξη, σε όλο το κείμενο. Τα καθολικά στατιστικά που υπολογίζονται για κάθε λέξη κατά τη μοντελοποίηση βοηθούν τον αλγόριθμο να αναπαραστήσει ακόμα και σπάνιες λέξεις, επιτρέποντας την εφαρμογή του σε μικρά και μεγάλα σύνολα δεδομένων.

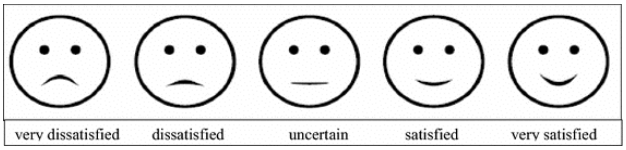
Παρόλα αυτά, η αξιοποίηση όλης αυτής της πληροφορίας απαιτεί υψηλή πολυπλοκότητα. Το μοντέλο GloVe αντιμετωπίζει αυτήν την πρόκληση, παρέχοντας ένα πλαίσιο που επιτρέπει την πιο πλήρη αναπαράσταση των λέξεων, αξιοποιώντας πληροφορίες από όλο το κείμενο."

# Κεφάλαιο 3: Ανάλυση Συναισθήματος

## 3.1 Ανάλυση Συναισθήματος και Αναπαραστάσεις

### 3.1.1 Αναπαράσταση σε κατηγορίες

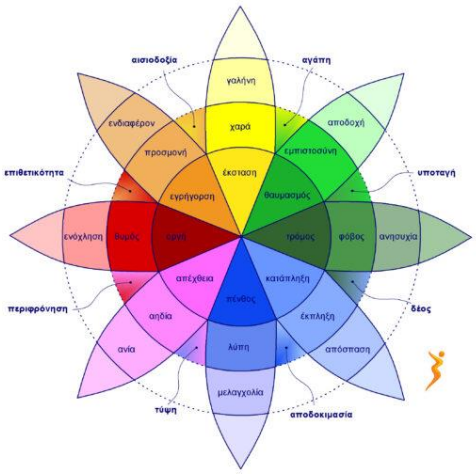
Η κατανομή των συναισθημάτων μπορεί να πραγματοποιηθεί σε διάφορα επίπεδα με διάφορες προσεγγίσεις, ανάλογα με το συγκεκριμένο πλαίσιο που απαιτείται. Μια απλή μορφή αναπαράστασης είναι η δυαδική, που χωρίζει τα συναισθήματα σε θετικά και αρνητικά. Επίσης, η κατηγοριοποίηση του κειμένου σε τρεις συνιστώσες, θετική, αρνητική και ουδέτερη, αποτελεί συχνή πρακτική.




**Εικόνα 7: Διαβάθμιση συναισθήματος σε 5 κατηγορίες**

Στον τομέα της έρευνας, χρησιμοποιείται επίσης η διαβάθμιση συναισθήματος σε 5 κατηγορίες, επικεντρωμένη στην πολικότητα των συναισθημάτων και την κατάταξη τους σε κλίμακα. Η διακύμανση αυτή μπορεί να είναι είτε διακριτική, με απόλυτη αντιστοιχία σε μία από τις κατηγορίες, είτε να επεκτείνεται σε αριθμητικούς χαρακτήρες με συνεχή κλίμακα, προσφέροντας ενδιάμεσες τιμές όπως ένας βαθμός από -100 έως 100, όπου το -100 αντιπροσωπεύει το "πολύ αρνητικό", το 0 το "ουδέτερο" και το 100 το "πολύ θετικό". Αντίστοιχα, η αξιολόγηση με βάση τα "αστέρια" (star rating) σε online κριτικές παρέχει πληροφορίες σχετικά με την απόδοση ενός προϊόντος.

Στον χώρο της ψυχολογίας, μια μελέτη[7] προτείνει την κατηγοριοποίηση των συναισθημάτων σε 8 βασικά είδη, όπως ο φόβος, η εμπιστοσύνη, η χαρά, η θλίψη, η προσμονή, η αηδία και η έκπληξη. Επιπλέον, υπάρχει μια έρευνα [8]που τα κατατάσσει σε 6 βασικές κατηγορίες: θυμό, αηδία, φόβο, ευτυχία, θλίψη και έκπληξη.




**Εικόνα 8: Ο τροχός των συναισθημάτων του Plutchik[7]**

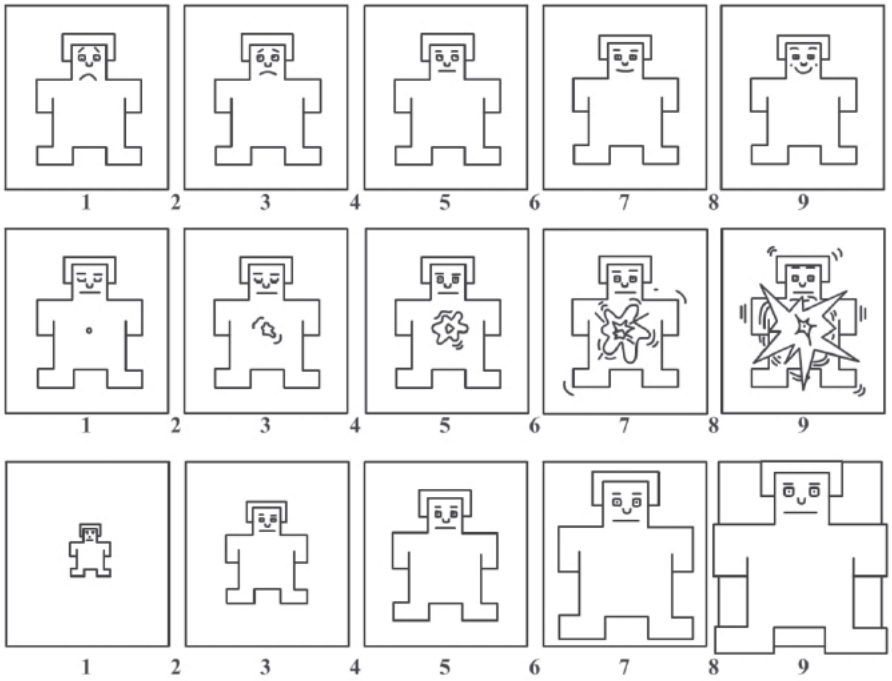
Βάση τα προηγούμενα, στον χώρο της έρευνας σχετικά με την ανάλυση συναισθήματος, επικρατούν συχνά οι παραπάνω προσεγγίσεις, με τα δεδομένα να καταγράφονται ανάλογα. Ωστόσο, από άποψη υπολογιστικής προσέγγισης, η επιλογή των συναισθημάτων που εξετάζονται κατά τον σχεδιασμό ενός συστήματος αναγνώρισης συναισθήματος συχνά εξαρτάται από τη συγκεκριμένη εφαρμογή και τον σκοπό της. Σε αυτήν την περίπτωση, πρέπει επίσης να λαμβάνεται υπόψιν το πλαίσιο εντός του οποίου πραγματοποιείται η μελέτη.

### 3.1.2 Διαστατικές Αναπαραστάσεις

Μια εναλλακτική προσέγγιση για την έκφραση των συναισθημάτων είναι η αναπαράστασή τους με βάση χαρακτηριστικά συνεχούς μορφής, γνωστά και ως διαστατικές αναπαραστάσεις. Στο πλαίσιο αυτό, οι βασικοί υπολογιστικοί παράμετροι είναι η ενεργοποίηση, το σθένος και η κυριαρχία. Πιο συγκεκριμένα, η ενεργοποίηση αντιπροσωπεύει την ένταση της συναισθηματικής εμπειρίας, το σθένος αξιολογεί το επίπεδο ευχαρίστησης που σχετίζεται με το συναίσθημα, και η κυριαρχία περιγράφει την ικανότητα ελέγχου των συναισθημάτων κατά τη διάρκεια μιας συναισθηματικά φορτισμένης εμπειρίας.

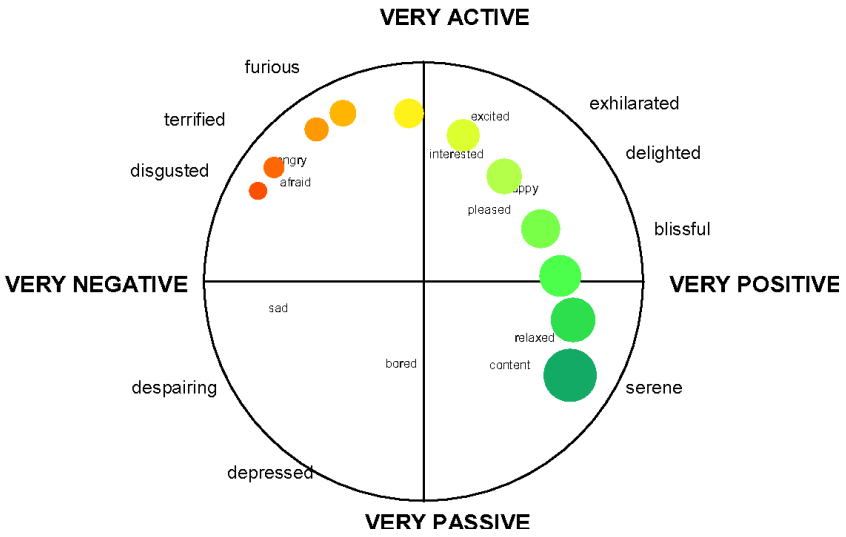
Τα συστήματα που χρησιμοποιούν αναπαραστάσεις διαστάσεων δεν απαιτούν τον προκαθορισμό ενός συγκεκριμένου συναισθηματικού κατηγοριών συνόλου, αλλά συνήθως απαιτούν την προεπιλογή του αριθμού των επιπέδων ενεργοποίησης και σθένους που θα χρησιμοποιηθούν κατά τη διαδικασία ταξινόμησης.

Για να διευκολυνθεί η επισημείωση των συναισθηματικών εκδηλώσεων σε κλίμακες ενεργοποίησης, σθένους και κυριαρχίας, ερευνητικές μελέτες εισήγαγαν την κλίμακα αυτοαξιολόγησης (Self-Assessment Manikins - SAM). Το μοντέλο SAM αποτελείται από διαισθητικές εικονογραφικές παραστάσεις που αντιπροσωπεύουν τις διαστάσεις ενεργοποίησης, σθένους και κυριαρχίας. Ένα παράδειγμα του SAM για μια κλίμακα 5 σημείων αναφορικά με την ενεργοποίηση της αξιολόγησης, του σθένους και της κυριαρχίας παρουσιάζεται στην Εικόνα 3 [9].




**Εικόνα 9: SAM: Σθένος, Διέγερση και Κυριαρχία [10]**

Εναλλακτικά, έχει γίνει προσπάθεια εκμετάλλευσης πλήρως της συνεχούς φύσης αυτών των αναπαραστάσεων, συλλέγοντας συνεχείς βαθμολογίες σε συναισθηματικές διαστάσεις και αναπτύσσοντας συστήματα που υπολογίζουν τις συνεχείς συναισθηματικές ιδιότητες. Αυτό συμπεριλαμβάνει τη συγκέντρωση βαθμολογιών συνεχούς μορφής για τα χαρακτηριστικά των διαστάσεων κατά τη διάρκεια του χρόνου, δημιουργώντας καμπύλες που αντιπροσωπεύουν τις συναισθηματικές ιδιότητες[11]. Ένα παράδειγμα αυτού του τύπου αναπαράστασης είναι το λογισμικό επισημείωσης Feeltrace, το οποίο επιτρέπει στους χρήστες να παρέχουν επισημειώσεις σε πραγματικό χρόνο μετακινώντας ένα δρομέα σε μια διεπαφή που αντιπροσωπεύει τον δισδιάστατο χώρο της ενεργοποίησης και του σθένους. Το αποτέλεσμα είναι μια συναισθηματική καμπύλη για κάθε συναισθηματική ιδιότητα, όπως παρουσιάζεται στην Εικόνα 4 [11].




**Εικόνα 10: Feeltrace [12]**

Ανεξαρτήτως της συγκεκριμένης αναπαράστασης, το πρόβλημα της υποκειμενικότητας κατά την ταξινόμηση αποτελεί πρόκληση για τη σωστή ανάθεση συναισθημάτων. Οι άνθρωποι συχνά αντιλαμβάνονται συναισθήματα μέσα από τη δική τους προσωπική διαδικασία και κρίση, που μπορεί να μην ταυτίζεται απόλυτα με την πραγματική συναισθηματική αντίληψη μιας συναισθηματικής εκδήλωσης. Καθώς η λεπτομέρεια αυξάνεται, δηλαδή ο αριθμός των ετικετών που αντιστοιχούν σε συναισθηματικές κατηγορίες και τα επίπεδα συναισθηματικών χαρακτηριστικών, μειώνεται η πιθανότητα συμφωνίας μεταξύ των αναλυτών. Αυτό αποτελεί βασική πρόκληση στην αναγνώριση συναισθημάτων, διότι δημιουργεί αβεβαιότητα όσον αφορά την αναγνώριση που σχετίζεται με ένα παράδειγμα, σε αντίθεση με την πιο διαδεδομένη πρακτική, όπου υπάρχει αυστηρή σύνδεση μεταξύ ενός παραδείγματος και της αντίστοιχης διακριτικής και καθορισμένης κατηγοριοποίησής του.

## 3.2 Ανάλυση Συναισθήματος στα Μέσα Κοινωνικής Δικτύωσης

### 3.2.1 Εφαρμογές

Επικεντρωνόμενοι στις διάφορες περιπτώσεις εφαρμογής, έρχεται στο προσκήνιο μια σειρά από ενδιαφέροντα σενάρια, με ιδιαίτερη έμφαση στα κοινωνικά μέσα δικτύωσης. Σε αυτό το πλαίσιο, η τεχνική ανάλυσης ανοίγει πόρτες για βαθύτερη κατανόηση των αναγκών, προτιμήσεων και απόψεων του ευρύτερου κοινού. Αυτό επιτρέπει την αποδοτική χρήση των δεδομένων από οργανισμούς, επιχειρήσεις και υπηρεσίες, όσον αφορά τα προϊόντα και τις παροχές τους.

Στον κόσμο των μέσων κοινωνικής δικτύωσης, η παρακολούθηση της απήχησης και της δημοτικότητας ενός προϊόντος, μιας υπηρεσίας, ακόμα και ενός προσώπου, γίνεται μέσω της ανάλυσης ποικίλων δεδομένων. Αυτά μπορεί να είναι κριτικές, σχόλια, άρθρα, κ.ά. Τα αποτελέσματα αυτής της ανάλυσης παρέχουν χρήσιμες πληροφορίες για τη βελτίωση προϊόντων και υπηρεσιών, ενώ ταυτόχρονα επηρεάζουν αποφάσεις και στρατηγικές, προσαρμόζοντας τις καμπάνιες προώθησης ανάλογα με τις νέες ανάγκες και τάσεις.

Η ανάλυση της κοινής γνώμης λειτουργεί επίσης ως εποικοδομητική κριτική, αναδεικνύοντας περιθώρια βελτίωσης. Τα αποτελέσματα αυτής της διαδικασίας αναγνώρισης και ταξινόμησης συναισθημάτων σε κείμενα παρέχουν σταθερή ενημέρωση, εξασφαλίζοντας ότι παραμένουν συναρπαστικά και αντιπροσωπευτικά. Με την ταχεία εξάπλωση των μέσων κοινωνικής δικτύωσης, αυτές οι πρακτικές εφαρμογές προκαλούν αυξημένη ζήτηση για εταιρείες που εξειδικεύονται στην ανάλυση συναισθημάτων, προσφέροντας έναν δυναμικό τομέα έρευνας.

### 3.2.2 Προκλήσεις

Τα δεδομένα που αντλούμε από το Twitter, παρότι αξιόλογα σύμφωνα με την προηγούμενη ανάλυση, απαιτούν ειδική προσοχή κατά την επεξεργασία τους. Συγκεκριμένα, τα κείμενα που εμφανίζονται σε όλες τις πλατφόρμες κοινωνικής δικτύωσης ξεχωρίζουν λόγω του ανεπίσημου και πολλές φορές αόριστου τρόπου γραφής των χρηστών. Αυτό δημιουργεί προκλήσεις στην επεξεργασία της φυσικής γλώσσας, καθώς τα δείγματα εμπεριέχουν σημασιολογικές αστοχίες, ορθογραφικά και συντακτικά

λάθη, συντομεύσεις λέξεων, σχήματα λόγου, δυσνόητες και αυθαίρετες έννοιες. Η συχνή χρήση αργκό επίσης προσθέτει στην πολυπλοκότητα.

Η περιορισμένη χρήση της ελληνικής γλώσσας στο Διαδίκτυο περιορίζει τη δυνατότητα εύρεσης ποιοτικού υλικού για τις εργασίες ΕΦΓ. Επιπλέον, η μορφολογική ιδιαιτερότητα των κειμένων, με την αυθαίρετη χρήση σημείων στίξης, κεφαλαίων γραμμάτων, hashtags, κ.ά., δημιουργεί πρόκληση στην μοντελοποίηση των κειμένων. Η επιλεκτική, στοχευμένη συλλογή υλικού, και η κατάλληλη επεξεργασία δεδομένων είναι κρίσιμες για τη σωστή αναπαράσταση του νοήματος και την αποφυγή περιττών ή περιοριστικών γνωρισμάτων.

## 3.3 Μηχανική Μάθηση

Στον τομέα της ανάλυσης συναισθήματος στο κείμενο, υπάρχουν τρεις βασικές προσεγγίσεις που εξετάζονται:

- Προσέγγιση βασισμένη σε μηχανική μάθηση: Σε αυτήν την προσέγγιση, χρησιμοποιούνται τεχνικές μηχανικής μάθησης για την εκπαίδευση μοντέλων που θα είναι σε θέση να αναγνωρίζουν και να αναλύουν συναισθήματα σε κείμενα. Αυτές οι τεχνικές συχνά χρησιμοποιούν αλγορίθμους μηχανικής μάθησης όπως τα νευρωνικά δίκτυα για την επίλυση του προβλήματος. Επιπλέον, επικεντρώνονται στην ανάπτυξη μοντέλων που είναι ικανά να προβλέπουν την τάση των συναισθημάτων σε κείμενα.

- Προσέγγιση βασισμένη σε λεξικό: Σε αυτήν την προσέγγιση, χρησιμοποιούνται λεξικά που περιλαμβάνουν λέξεις και τα συναφή τους συναισθηματικά βάρη. Κάθε λέξη αξιολογείται με βάση το συναισθηματικό της χαρακτήρα, και η συνολική αξιολόγηση του κειμένου προκύπτει από το άθροισμα των συναισθηματικών βαρών των λέξεων που περιέχει.

- Προσέγγιση βασισμένη σε υβριδικές μεθόδους: Σε αυτήν την προσέγγιση, συνδυάζονται στοιχεία και από τις δύο προηγούμενες προσεγγίσεις. Χρησιμοποιούνται είτε συγχρόνως είτε διαδοχικά, με σκοπό να επιτευχθεί καλύτερη απόδοση στην ανάλυση συναισθημάτων κειμένου.

Στο πλαίσιο αυτό, εστιάζουμε στην πρώτη προσέγγιση, δηλαδή τη βασισμένη σε μηχανική μάθηση. Περιγράφουμε τον όρο "μηχανική μάθηση", αναλύουμε τα προβλήματα που αντιμετωπίζει, και εξηγούμε τους διάφορους αλγορίθμους που ευρέως χρησιμοποιούνται σε αυτήν την προσέγγιση, εστιάζοντας ιδιαίτερα στο πλαίσιο της ανάλυσης συναισθημάτων.

### 3.3.1 Ερμηνεία της Μηχανικής Μάθησης και Βασικές Έννοιες

Κατά την αναφορά στη Μηχανική Μάθηση, αναφερόμαστε στην χρήση ενός αλγορίθμου που εκτελείται σε μια υπολογιστική μηχανή, επιδιώκοντας να βελτιώσει την επίδοσή του στη διαδικασία εκτέλεσης μιας λειτουργίας. Οι λειτουργίες αυτές αντιστοιχούν σε αυτές της ανθρώπινης νοημοσύνης και ενδέχεται να περιλαμβάνουν:

- Μηχανική κατανόηση της γλώσσας και παραγωγή ομιλίας (NLP/NLU): Επιτρέπει στις μηχανές να κατανοούν και να παράγουν φυσική γλώσσα.

- Μηχανική αναγνώριση προτύπων: Επικεντρώνεται στην αναγνώριση προτύπων και των σχέσεων τους σε δεδομένα.

- Ανάπτυξη στρατηγικής σε διάφορες καταστάσεις (π.χ., παιχνίδια): Επιτρέπει τη δημιουργία στρατηγικής για την αντιμετώπιση διαφόρων καταστάσεων, όπως συμβαίνει σε παιχνίδια.

Η Μηχανική Μάθηση αποτελεί βασικό στοιχείο της Τεχνητής Νοημοσύνης και αξιοποιεί έννοιες από τη στατιστική, τη θεωρία πληροφορίας και τη γνωσιακή επιστήμη.

Σε γενικές γραμμές, ένα πρόβλημα μάθησης επεξεργάζεται ένα σύνολο δειγμάτων από δεδομένα, συμβολίζουμε ως $D = \{x_1, x_2, \dots, x_n\}$, προσπαθώντας να ανακαλύψει άγνωστες ιδιότητες των δεδομένων αυτών. Το ανωτέρω σύνολο ονομάζεται σύνολο εκπαίδευσης και χρησιμοποιείται για την εκπαίδευση του συστήματος μηχανικής μάθησης. Κάθε δείγμα $x_i$ ονομάζεται χαρακτηριστικό και μπορεί να είναι είτε βαθμωτό (single feature) είτε διάνυσμα (feature vector).

Οι τρεις βασικοί τύποι μηχανικής μάθησης είναι:

- Επίβλεψη (Supervised Learning): Στην επίβλεψη, κάθε δείγμα-χαρακτηριστικό $x_i$ συνοδεύεται από μια ετικέτα-στόχο, που αποτελεί τη μεταβλητή που το σύστημα μηχανικής μάθησης πρέπει να προβλέψει όπως ,για παράδειγμα, προβλήματα ταξινόμησης (classification) και παλινδρόμησης (regression) ανήκουν σε αυτήν την κατηγορία. Η ταξινόμηση, για παράδειγμα, μπορεί να εφαρμοστεί στην αναγνώριση προσώπων, στην αναγνώριση ομιλίας, στην ιατρική/βιολογία, και στην επεξεργασία φυσικής γλώσσας (NLP).

- Χωρίς Επίβλεψη (Unsupervised Learning): Στην αντίθετη πλευρά, στην μάθηση χωρίς επίβλεψη, τα δεδομένα εκπαίδευσης δεν έχουν συνοδευτικές ετικέτες. Οι εφαρμογές περιλαμβάνουν συσταδοποίηση (clustering), εκτίμηση πυκνότητας, και συμπίεση δεδομένων. Αυτές οι τεχνικές βρίσκουν χρήση, για παράδειγμα, στον τομέα της αναγνώρισης ομοιοτήτων μεταξύ δεδομένων εισόδου.

- Ενισχυτική Μάθηση (Reinforcement Learning): Στην ενισχυτική μάθηση, το σύστημα μαθαίνει μέσω αλληλεπίδρασης με το περιβάλλον και κρίσης που τιμωρεί ή επιβραβεύει. Εφαρμογές περιλαμβάνουν την ανάπτυξη στρατηγικής σε παιχνίδια, τον ρομποτικό έλεγχο, και την αλληλεπίδραση με ανθρώπους.

Κάθε κατηγορία παρέχει μοναδικές λύσεις για ποικίλα προβλήματα, ενισχύοντας την εφαρμογή της μηχανικής μάθησης σε διάφορους τομείς.

Στον χώρο της μηχανικής μάθησης, εστιάζουμε σε δύο βασικά στοιχεία: την εκπαίδευση και την ανάκληση. Κατά την εκπαίδευση, παρουσιάζουμε πολλά παραδείγματα στο σύστημα, είτε με επιδιωκόμενους στόχους είτε χωρίς, με σκοπό τη ρύθμιση των παραμέτρων του. Αυτό συμβάλλει στη βελτίωση της απόδοσής του σε συγκεκριμένες λειτουργίες, όπως η αναγνώριση.

Από την άλλη πλευρά, η ανάκληση αφορά την εισαγωγή ενός ή περισσοτέρων προτύπων χωρίς προηγούμενη εκπαίδευση, με σκοπό την ανάκτηση της απόκρισης του συστήματος. Μετά την εκπαίδευση, αξιολογούμε την απόδοσή του χρησιμοποιώντας τη συνολική ακρίβεια. Αυτό μετρά το ποσοστό των σωστών προβλέψεων από το σύνολο ελέγχου, δηλαδή από παραδείγματα που δεν είχαν χρησιμοποιηθεί κατά τη διάρκεια της εκπαίδευσης.

$$\text{Accuracy} = \frac{\text{σωστά ταξινομημένα πρότυπα που ανηκουν στο T}}{\text{πληθικός αριθμός συνόλου T}}$$

Εάν επιδιώκουμε να αξιολογήσουμε την απόδοση του συστήματός μας σε κάθε κατηγορία, θα πρέπει να χρησιμοποιήσουμε διαφορετικές μετρικές απόδοσης, που καθορίζονται για κάθε κατηγορία και ονομάζονται ακρίβεια (precision) και ανάκληση (recall). Για τον ορισμό αυτών των μετρικών, ας εξετάσουμε την απλή περίπτωση της δυαδικής ταξινόμησης σε μία από δύο κατηγορίες, ας πούμε positive και negative.

Οι όροι που χρησιμοποιούνται είναι οι εξής:

- True Positives (TP): Ο αριθμός των δειγμάτων που ταξινομήθηκαν σωστά στην κατηγορία positive.
- False Positives (FP): Ο αριθμός των δειγμάτων που ταξινομήθηκαν λανθασμένα στην κατηγορία positive.
- True Negatives (TN): Ο αριθμός των δειγμάτων που ταξινομήθηκαν σωστά στην κατηγορία negative.
- False Negatives (FN): Ο αριθμός των δειγμάτων που ταξινομήθηκαν λανθασμένα στην κατηγορία negative.

Οι μετρικές ακρίβεια και ανάκληση ορίζονται ως εξής:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

Η ακρίβεια μετρά το πόσα από τα θετικά παραδείγματα που ταξινομήθηκαν ως θετικά ήταν πράγματι θετικά, ενώ η ανάκληση μετρά το πόσα από τα πραγματικά θετικά παραδείγματα αναγνωρίστηκαν σωστά. Οι δύο αυτές μετρικές, ακρίβεια και ανάκληση, βρίσκονται σε αντίφαση μεταξύ τους: όταν μια αυξάνεται, η άλλη μειώνεται και αντίστροφα. Όταν τις συνδυάζουμε σε μία μόνο τιμή, αυτή ονομάζεται F-measure (ή F1-measure):

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Τελικά, ποια μετρική θα χρησιμοποιηθεί για τον έλεγχο της απόδοσης του συστήματος αναγνώρισης εξαρτάται από την εφαρμογή. Σε γενικές γραμμές, όταν το σύνολο ελέγχου είναι ισορροπημένο, δηλαδή περιλαμβάνει ισορροπημένο αριθμό δειγμάτων για κάθε κατηγορία, ώστε τα αποτελέσματα της συνολικής ακρίβειας να μην παραπλανούν και τα λάθη ταξινόμησης να έχουν ίση σημασία ανεξάρτητα από την κατηγορία, τότε χρησιμοποιείται η συνολική ακρίβεια (accuracy) ως μέτρο απόδοσης. Σε επόμενα στάδια, όταν αναφερόμαστε στην "ακρίβεια", εννοούμε το μέγεθος της συνολικής ακρίβειας.

## 3.4 Ανάλυση Συναισθήματος – Σύνδεση με Μηχανική Μάθηση

Όπως αναφέραμε προηγουμένως, η πρόκληση της ανάλυσης συναισθήματος, δηλαδή τον προσδιορισμό της θετικής ή αρνητικής κατάταξης της άποψης ενός συγγραφέα για ένα θέμα, μπορεί να αντιμετωπιστεί με τεχνικές μηχανικής μάθησης, και έχουν υπάρξει πολλές προσπάθειες σε αυτό τον τομέα. Ένα παράδειγμα αποτελεί η μέθοδος που προτείνει ο Turney[13] το 2002, ο οποίος χρησιμοποίησε ένα σύστημα αυτόματης ταξινόμησης κριτικών ταινιών βασιζόμενο σε αλγόριθμο unsupervised learning που χρησιμοποιεί τον δείκτη Pointwise Mutual Information (PMI). Επίσης, την ίδια περίοδο, οι Pang, Lee και Vaithyanathan[14] εξέτασαν τη χρήση αλγορίθμων επιβλεπόμενης μάθησης, όπως Naive Bayes, Maximum Entropy Classification και SVMs, για την ταξινόμηση στο πρόβλημα του sentiment analysis για κριτικές ταινιών. Και άλλες παρόμοιες μελέτες στον τομέα της ανάλυσης συναισθημάτων ακολούθησαν μια παρόμοια πορεία, χρησιμοποιώντας όλο και περισσότερα πολύπλοκα μοντέλα μηχανικής μάθησης. Αυτά περιλάμβαναν μοντέλα όπως τα Hidden Markov Models (HMMs)[15] ή τα Conditional Random Fields (CRFs)[16], τα οποία λάμβαναν υπόψη τη σειρά των λέξεων στο κείμενο, χρησιμοποιώντας διάφορα σύνολα χαρακτηριστικών.

Ωστόσο, τα πιο πρόσφατα έργα επικεντρώνονται στην εφαρμογή μοντέλων βαθιάς μάθησης, όπως τα συνελικτικά νευρωνικά δίκτυα και τα επαναλαμβανόμενα νευρωνικά δίκτυα. Αυτά τα μοντέλα έχουν επιδείξει εντυπωσιακά αποτελέσματα και αποτελούν το κύριο αντικείμενο έρευνας τα τελευταία χρόνια.

Ένα ζήτημα που προκύπτει κατά την υλοποίηση ενός συστήματος αναγνώρισης συναισθήματος από κείμενο είναι η αντιμετώπιση της ουδέτερης κλάσης. Παρόλο που πολλοί ερευνητές τείνουν να αγνοούν την ουδέτερη κλάση και επικεντρώνονται μόνο στις θετικές και αρνητικές κλάσεις, είναι σημαντικό να σημειωθεί ότι όχι όλες οι προτάσεις περιέχουν συναίσθημα. Η εκπαίδευση του ταξινομητή στην ανίχνευση μόνο αυτών των δύο κλάσεων μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting) του μοντέλου στο σύνολο εκπαίδευσης.

Για την αναπαράσταση του κειμένου απαιτείται ένα διάνυσμα χαρακτηριστικών $\boldsymbol{x}$. Μία από τις πιο απλές αναπαραστάσεις είναι η αναπαράσταση Bag of Words (BoW), η οποία αντιμετωπίζει το κείμενο ως ένα σύνολο ανεξάρτητων λέξεων, αγνοώντας τη σειρά τους. Παρότι αυτή η αναπαράσταση είναι απλή, χρησιμοποιείται ευρέως σε προβλήματα text classification, όπως το spam filtering, και επιτυγχάνει ικανοποιητικά αποτελέσματα.

Η BoW αναπαράσταση μπορεί να επεκταθεί σε bigrams ή n-grams για να ενισχύσει την απόδοση του ταξινομητή, λαμβάνοντας υπόψη τον συνδυασμό διαδοχικών λέξεων. Ωστόσο, η επιλογή του n εξαρτάται από την εφαρμογή, με τα bigrams ή trigrams να αποτελούν συνήθως καλή επιλογή. Είναι σημαντικό να σημειωθεί ότι η χρήση n-grams μπορεί να αυξήσει τον χώρο των χαρακτηριστικών, αλλά μπορεί να βελτιώσει τα αποτελέσματα.

Σε προεπεξεργασία του κειμένου, συνήθως εφαρμόζονται τεχνικές όπως η αφαίρεση stopwords (λέξεις που δεν συνεισφέρουν στο νόημα), η αφαίρεση σημείων στίξης και η λημματοποίηση ή αποκοπή λέξεων για να μειωθεί ο χώρος των χαρακτηριστικών. Τέλος, κατά την κατασκευή των χαρακτηριστικών, μπορεί να δηλωθεί ο αριθμός εμφανίσεων ενός όρου ή ενός n-gram στο κείμενο αντί να δηλωθεί απλά η ύπαρξή του. Στο πρόβλημα του sentiment analysis, όμως, ο αριθμός των εμφανίσεων μιας λέξης δεν καθορίζει συνήθως το συναίσθημα του κειμένου, και επομένως, συνήθως χρησιμοποιούνται δυαδικοποιημένες εκδοχές των αλγορίθμων που εστιάζουν στην ύπαρξη ή μη του όρου και όχι στον αριθμό εμφανίσεών του.

### 3.4.1 Ο αλγόριθμος Naive Bayes

Ο αλγόριθμος Naive Bayes ανήκει στην κατηγορία των αλγορίθμων επιβλεπόμενης μηχανικής μάθησης. Πρόκειται για έναν απλό πιθανοτικό ταξινομητή που βασίζεται στο θεώρημα του Bayes και στην αφελή υπόθεση της υπό συνθήκη ανεξαρτησίας μεταξύ των χαρακτηριστικών $x_i, x_j$, όπου $i \neq j$, δεδομένης της κλάσης $c_k$

Ο Naive Bayes αποτελεί μια από τις βασικές τεχνικές ταξινόμησης κειμένου. Παρά την απλότητά του και τις υποθέσεις που κάνει, επιδίδεται αποτελεσματικά σε πολλά προβλήματα. Η καλή απόδοσή του συνδυάζεται με χαλαρές απαιτήσεις σε CPU και μνήμη, ενώ ο χρόνος εκπαίδευσης είναι σημαντικά μικρότερος σε σχέση με άλλες μεθόδους. Ωστόσο, ο Naive Bayes είναι μερικές φορές κακός εκτιμητής, καθώς συχνά υπερεκτιμά τις πιθανότητες εξόδου.

Στο πρόβλημα της αναγνώρισης προτύπων, επιλέγουμε μια κλάση $c_j$ για ένα πρότυπο $x$ με βάση το διάνυσμα χαρακτηριστικών του. Η επιλογή γίνεται μέσω Ν πιθανών κλάσεων $c_1, c_2, \ldots, c_N$.

Ορίζουμε τη δεσμευμένη πιθανότητα $P(c_j|x)$ ως την πιθανότητα το πρότυπο $x$ να ανήκει στην κλάση $c_j$, γνωστή και ως εκ των υστέρων πιθανότητα (a posteriori probability). Έτσι, επιλέγουμε για το $x$ την κλάση που μεγιστοποιεί την παραπάνω a posteriori πιθανότητα, έστω την κλάση $k$. Δηλαδή, θεωρούμε τον ακόλουθο κανόνα απόφασης:

$$k = \arg\max P\left(c_j|x\right), \quad j = 1, \ldots, N$$

Αυτός ακριβώς είναι ο κανόνας απόφασης στον ταξινομητή Naive Bayes, γνωστός και ως Maximum A Posteriori (MAP) ταξινομητής.

Η πιθανότητα $P(c_j|x)$ υπολογίζεται εφαρμόζοντας το θεώρημα του Bayes, ως εξής:

$$P(c_j|x) = \frac{P(c_j, x)}{P(x)} = \frac{P(x|c_j)P(c_j)}{P(x)}$$

Οι όροι $P(c_j)$ και $P(x|c_j)$ υπολογίζονται μέσω της εκτίμησης μέγιστης πιθανοφάνειας (MLE -Maximum Likelihood Estimation ) πάνω στο training set. Στον υπολογισμό συμμετέχει και η υπόθεση ανεξαρτησίας μεταξύ των χαρακτηριστικών δεδομένης της κλάσης $c_j$, οπότε η πιθανότητα $P(x|c_j)$υπολογίζεται ως το γινόμενο των επιμέρους πιθανοτήτων $P(x_i|c_j)$. Συνεπώς:

$$P(x|c_j) = \prod_{i=1}^{n} P(x_i|c_j)$$

Για την απόφαση του ταξινομητή, αρκεί ο υπολογισμός των πιθανοτήτων $P(c_j)$ και $P(x_i|c_j)$. Οι πιθανότητες αυτές εκτιμώνται κατά την MLE, με το $P(c_j)$ να είναι το ποσοστό των προτύπων στο training set που ανήκουν στην κλάση $c_j$, και το $P(x_i|c_j)$ να είναι οι συχνότητες των χαρακτηριστικών στο ίδιο training set.

Υπάρχουν διάφορες παραλλαγές του αλγορίθμου Naive Bayes, με τις διαφορές τους να εξαρτώνται από την υπόθεση που κάνουν σχετικά με την κατανομή $P(x_i|c_j)$. Αναφέρονται ορισμένες από αυτές:

- Gaussian Naive Bayes: Σε αυτήν την έκδοση, υποθέτει ότι η κατανομή $P(x_i|c_j)$ είναι συνεχής και ακολουθεί την κανονική κατανομή. Ουσιαστικά, χρησιμοποιεί το γνωστό μοντέλο Gaussian για τον υπολογισμό των πιθανοτήτων.

- Multinomial Naive Bayes: Αυτή η έκδοση υλοποιεί τον αλγόριθμο για πολυωνυμικά κατανεμημένα δεδομένα και χρησιμοποιείται ευρέως στην ταξινόμηση κειμένου. Τα δεδομένα αναπαριστώνται ως μετρήσεις λέξεων ή n-grams, και η κατανομή παραμετροποιείται από διανύσματα. Οι παράμετροι υπολογίζονται μέσω εξομαλυμένης εκδοχής MLE. Επιπλέον, λειτουργεί καλά με διανύσματα Term Frequency-Inverse Document Frequency (TF-IDF), εκτός από την αναπαράσταση Bag of Words (BoW).

- Complement Naive Bayes: Ο CNB είναι τροποποίηση του Multinomial Naive Bayes, σχεδιασμένη για ανισορροπημένα δεδομένα. Χρησιμοποιεί στατιστικά στοιχεία από το σύνολο της αντίστοιχης κατηγορίας για τον υπολογισμό των βαρών του μοντέλου.

- Bernoulli Naive Bayes: Αυτή η έκδοση εφαρμόζει το μοντέλο Bernoulli στον Naive Bayes. Κάθε όρος του λεξιλογίου ισούται με 1 αν εμφανίζεται στο κείμενο, αλλιώς με 0. Η πιθανότητα υπολογίζεται λαμβάνοντας υπόψη τους όρους που δεν εμφανίζονται στο κείμενο.

Ο αλγόριθμος Multinomial Naive Bayes συνήθως επιλέγεται όταν η συχνή εμφάνιση των λέξεων έχει σημαντικό ρόλο στην ταξινόμηση, όπως σε προβλήματα κατηγοριοποίησης θεμάτων. Για παράδειγμα, όταν προσπαθούμε να κατηγοριοποιήσουμε κείμενα βασιζόμενοι στο θέμα τους, η συχνότητα εμφάνισης των λέξεων είναι σημαντική πληροφορία. Αντίθετα, ο δυαδικοποιημένος Multinomial Naive Bayes χρησιμοποιείται όταν η συχνότητα εμφάνισης των λέξεων δεν παίζει καθοριστικό ρόλο στην ταξινόμηση. Σε περιπτώσεις όπως η αναγνώριση παραπλανητικής διαφήμισης στον κυβερνοχώρο, ενδιαφερόμαστε περισσότερο για το εάν μια λέξη όπως "ανεπιθύμητη" εμφανίζεται παρά για το πόσες φορές αναφέρεται. Τέλος, ο αλγόριθμος Bernoulli Naive Bayes είναι χρήσιμος όταν η απουσία μιας συγκεκριμένης λέξης είναι σημαντική. Για παράδειγμα, χρησιμοποιείται συνήθως στην ανίχνευση spam ή στην ανίχνευση περιεχομένου για ανηλίκους, όπου η απουσία συγκεκριμένων λέξεων μπορεί να είναι ενδεικτική.

### 3.4.2 Μέγιστη Εντροπία

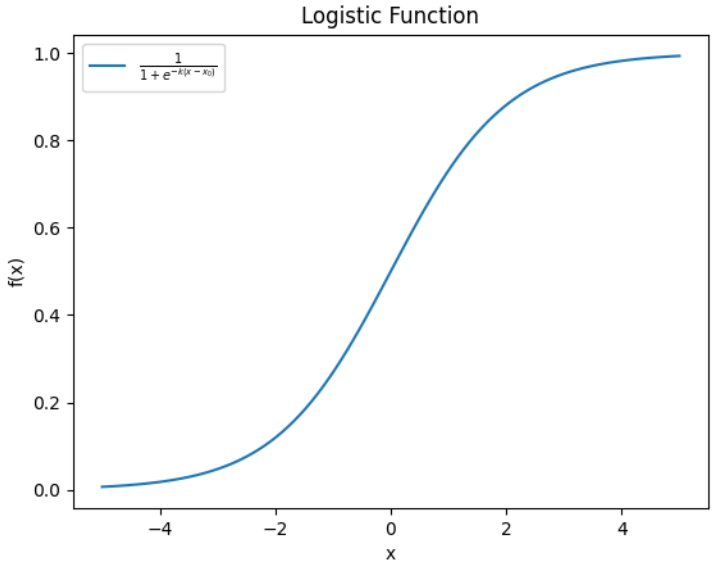
Ο αλγόριθμος Maximum Entropy, γνωστός επίσης ως αλγόριθμος Λογιστικής Παλινδρόμησης, υλοποιεί ένα γραμμικό μοντέλο με σκοπό την ταξινόμηση, αντί για την παλινδρόμηση, παρόλο που ονομάζεται αλλιώς. Είναι ένας πιθανοτικός ταξινομητής, όπου οι πιθανότητες εξόδου μοντελοποιούνται με τη χρήση μιας λογιστικής συνάρτησης. Στην περίπτωση πολλών εισόδων, η γενίκευση της λογιστικής συνάρτησης ονομάζεται softmax function. Η softmax συνάρτηση για μια σειρά αποτελεσμάτων$z_1, z_2, \ldots, z_k$ ορίζεται ως εξής:

Ο τύπος της λογιστικής συνάρτησης (logistic function) που χρησιμοποιεί ο αλγόριθμος Maximum Entropy είναι ο εξής:

$$f(x) = \frac{1}{1 + e^{-k(x-x_0)}}$$

Όπου:

- $f(x)$ είναι η εξόδος της λογιστικής συνάρτησης για την είσοδο $x$.
- $k$ είναι ένας παράγοντας κλίσης (συνήθως θετικός).
- $x_0$ είναι η οριζόντια θέση της καμπύλης.




**Διάγραμμα 1: Συνάρτηση SoftMax.**

Αυτή η συνάρτηση χρησιμοποιείται για να μοντελοποιήσει τις πιθανότητες εξόδου στο πλαίσιο της ταξινόμησης που επιδιώκει ο αλγόριθμος Maximum Entropy.

Ο Maximum Entropy βασίζεται στην αρχή της μέγιστης εντροπίας, επιλέγοντας το μοντέλο που κάνει τη λιγότερη υπόθεση πέρα από τα περιοριστικά δεδομένα του σετ εκπαίδευσης. Έτσι, η κατανομή είναι όσο το δυνατόν ομοιόμορφη. Αυτός ο ταξινομητής χρησιμοποιείται σε πολλά προβλήματα ταξινόμησης κειμένου, όπως ανίχνευση γλώσσας, ταξινόμηση κειμένου με βάση το θέμα, sentiment analysis, κ.λπ.

Σε αντίθεση με τον πιθανοτικό ταξινομητή Naive Bayes, ο Maximum Entropy δεν υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών. Αυτό το χαρακτηριστικό τον καθιστά αποδοτικό σε προβλήματα ταξινόμησης κειμένου, όπου τα χαρακτηριστικά-λέξεις δεν είναι προφανώς ανεξάρτητα. Ωστόσο, το μειονέκτημά του είναι ο μεγαλύτερος χρόνος εκπαίδευσης λόγω του προβλήματος βελτιστοποίησης που πρέπει να επιλυθεί για την προσδιορισμό των παραμέτρων του.

Ας χρησιμοποιήσουμε την παραδοσιακή αναπαράσταση BoW, με τις λέξεις $w1, w2, \ldots, wn$ να αποτελούν το λεξιλόγιο. Ο στόχος είναι να κατασκευάσουμε ένα πιθανοτικό μοντέλο που, δεδομένου ένα κείμενο $\boldsymbol{x}$, θα το αντιστοιχίζει σε μία κατηγορία $cj$ (θετική ή αρνητική για το πρόβλημά μας).

Αρχικά, από το training set υπολογίζουμε με τη μέθοδο Μέγιστης Πιθανοφάνειας (Maximum Likelihood Estimation) την εμπειρική πιθανότητα το τυχαίο κείμενο $\boldsymbol{x}$ να ανήκει στην κατηγορία $c$:

$$\mathrm{P}(x, c) = \frac{\text{αριθμός φορών που το δείγμα}(x, c)\text{εμφανίζεται στο training set}}{\text{μέγεθος training set}}$$

Ορίζουμε στη συνέχεια την παρακάτω Boolean συνάρτηση:

$$f_i(x, c) = \begin{cases} 1, & \text{if } c = c_j \text{ and x include the word wi} \\ 0, & \text{otherwise} \end{cases}$$

η οποία στην βιβλιογραφία του Max Entropy ονομάζεται χαρακτηριστικό.

Συγκεκριμένα, ορίζουμε δύο προσδοκίες χαρακτηριστικών (feature expectations):

- Αναμενόμενη τιμή χαρακτηριστικού ως προς την εμπειρική κατανομή $P(x, c)$:

$$E(f_i) = \sum_{\mathrm{x},c} P(\mathrm{x}, c) \cdot f_i(\mathrm{x}, c)$$

- Αναμενόμενη τιμή χαρακτηριστικού ως προς το μοντέλο $P(c|\mathrm{x})$:

$$E(f_i) = \sum_{\mathrm{x},c} P(\mathrm{x}) \cdot P(c|x) \cdot f_i(\mathrm{x}, c)$$

όπου $P(\mathrm{x})$ αντιπροσωπεύει την εμπειρική κατανομή του \(\tilde{x}\) στο σύνολο εκπαίδευσης, και συνήθως ορίζεται ως:

$$P(\tilde{x}) = \frac{1}{\text{σύνολο εκπαίδευσης}}$$

Επιβάλλοντας τον περιορισμό ότι η αναμενόμενη τιμή χαρακτηριστικού είναι ίση με την εμπειρική τιμή, προκύπτει η παρακάτω εξίσωση:

$$\sum_{\tilde{x},c} P(x) \cdot P(c|\mathrm{x}) \cdot f_i(\mathrm{x}, c) = \sum_{\tilde{x},c} P(\mathrm{x}, c) \cdot f_i(\mathrm{x}, c)$$

Η εξίσωση (3) χαρακτηρίζεται ως περιορισμός, με κάθε χαρακτηριστικό $f_i$ να έχει έναν αντίστοιχο περιορισμό. Αυτοί οι περιορισμοί μπορούν να ικανοποιηθούν από άπειρα στοχαστικά μοντέλα.

Χρησιμοποιώντας την αρχή της μέγιστης εντροπίας, ο αλγόριθμος επιλέγει το μοντέλο που είναι όσο το δυνατόν ομοιόμορφο. Συγκεκριμένα, επιλέγει το μοντέλο $P^*$:

$$P^* = \arg\max\left(-\sum P(x)P(c|\mathrm{x})\log P\,(c|\mathrm{x})\right)$$

υπό τους περιορισμούς:

- $P(c|\mathrm{x}) \geq 0, \forall \mathrm{x}, c)$
- $\sum_c P(c|\mathrm{x}) = 1, \forall \mathrm{x}$
- $\sum P(\mathrm{x})P(c|\mathrm{x})f_i(x,c) = \sum P(\mathrm{x},c)f_i(\mathrm{x},c)\,, \forall \tilde{x}, c, i = 1, \ldots, \eta$

Το παραπάνω πρόβλημα αναδιατυπώνεται στο δυικό πρόβλημα χωρίς περιορισμούς, χρησιμοποιώντας πολλαπλασιαστές Lagrange $\lambda_1, \ldots, \lambda_n$. Η εκτίμηση των παραμέτρων $\lambda_i$ απαιτεί τη χρήση ενός επαναληπτικού αλγορίθμου κλιμάκωσης, όπως ο GIS (Generalized Iterative Scaling) ή ο IIS (Improved Iterative Scaling).

Παρατηρείται ότι μόλις εντοπιστούν οι πολλαπλασιαστές Lagrange, η εκ των υστέρων πιθανότητα το κείμενο $x$ να ανήκει στην κατηγορία $c_j$ δίνεται από τη συνάρτηση softmax:

$$P(c_j|\mathrm{x}) = \frac{\exp\left(\sum_i \lambda_i f_i(\mathrm{x}, c_j)\right)}{\sum_c \exp(\sum_i \lambda_i f_i(x,c))}$$

Εδώ, ο πολλαπλασιαστής Lagrange $\lambda_i$ αναπαριστά το βάρος του χαρακτηριστικού $i$ στην επιλογή της κλάσης $c_i$. Ένα υψηλό θετικό βάρος υποδηλώνει ότι η λέξη $i$ πιθανόν σχετίζεται με την κλάση $c_j$, ενώ ένα υψηλό αρνητικό βάρος υποδηλώνει ότι η λέξη $i$ πιθανόν δεν σχετίζεται με την κλάση $c_j$.

### 3.4.3 Μηχανές Διανυσμάτων Υποστήριξης

Οι μηχανές διανυσμάτων υποστήριξης (SVMs) αποτελούν ένα δημοφιλές σύνολο μεθόδων για την επιβλεπόμενη μάθηση. Κάποια από τα βασικά τους πλεονεκτήματα περιλαμβάνουν την αποτελεσματικότητά τους σε χώρους πολλών διαστάσεων, ακόμη και όταν τα χαρακτηριστικά είναι περισσότερα από τα δείγματα, αποφεύγοντας έτσι το overfitting. Επιπλέον, χρησιμοποιούν ένα υποσύνολο των παραδειγμάτων εκπαίδευσης για την κατασκευή του συνόρου απόφασης, διαφοροποιούντας τους από άλλους αλγορίθμους που απαιτούν μεγαλύτερη μνήμη. Τέλος, μπορούν να αντιμετωπίσουν γραμμικά και μη γραμμικά προβλήματα διαχωρισμού, επιτρέποντας στον χρήστη να χρησιμοποιήσει προκαθορισμένες ή δικές του συναρτήσεις πυρήνα. Ωστόσο, ένα από τα βασικά μειονεκτήματα των SVMs είναι ότι δεν υπολογίζουν άμεσα πιθανότητες, όπως κάνουν άλλοι αλγόριθμοι.

Για να το καταλάβουμε καλύτερα, ας υποθέσουμε ότι αντιμετωπίζουμε ένα πρόβλημα διαχωρισμού δύο κλάσεων, όπως στο sentiment analysis. Οι δύο κλάσεις συμβολίζονται ως $C_0$ (η αρνητική) και $C_1$ (η θετική), ενώ οι ετικέτες των σημείων είναι -1 ή 1. Το σύνολο εκπαίδευσης αποτελείται από πολλά συμβολίζουμε το πλήθος τους ως N επισημειωμένα παραδείγματα $(x_i, d_i)$
που ανήκουν σε μία από τις δύο κλάσεις.

#### 3.4.3.1 Γραμμικά Διαχωρίσιμα Προβλήματα

Σε περίπτωση που το πρόβλημά μας είναι γραμμικά διαχωρίσιμο, δηλαδή κάθε παράδειγμα που ανήκει στην κλάση $C_0$ μπορεί να διαχωριστεί από ένα υπερεπίπεδο από τα παραδείγματα που ανήκουν στην κλάση $C_1$, η εξίσωση αυτού του υπερεπιπέδου μπορεί να εκφραστεί ως:

$$w^T x + b = 0$$

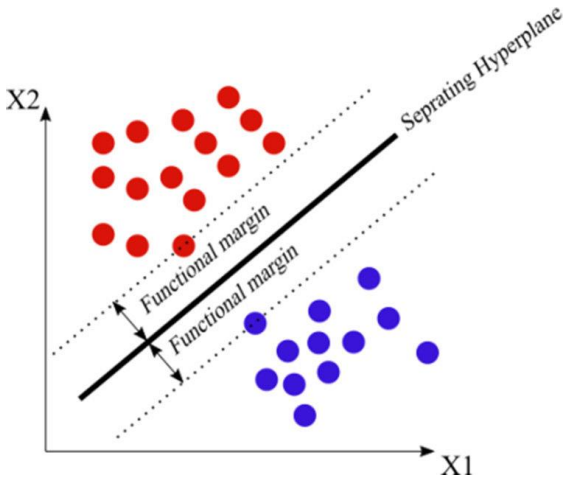
όπου $x$ αναπαριστά ένα διάνυσμα εισόδου, $w$ είναι ένα προσαρμόσιμο διάνυσμα βαρών, και $b$ είναι η πόλωση.

Ο στόχος της μηχανής διανυσμάτων υποστήριξης είναι να βρει το συγκεκριμένο υπερεπίπεδο που μεγιστοποιεί το περιθώριο διαχωρισμού, το οποίο είναι η απόσταση μεταξύ του υπερεπιπέδου και του πλησιέστερου σημείου δεδομένων. Αυτό το περιθώριο διαχωρισμού, συμβολίζεται ως ρ. Επομένως,

η επιφάνεια απόφασης αναφέρεται ως βέλτιστο υπερεπίπεδο, καθώς η μεγιστοποίηση του περιθωρίου συνάδει με την ιδέα ότι μεγαλύτερο περιθώριο συνεπάγεται μικρότερο σφάλμα γενίκευσης.

Μπορείτε να δείτε μια γεωμετρική αναπαράσταση αυτού του βέλτιστου υπερεπιπέδου στην εικόνα 11 για ένα δισδιάστατο χώρο εισόδου.

Συνοπτικά, στο πλαίσιο αυτού του γραμμικά διαχωρίσιμου προβλήματος, η μηχανή διανυσμάτων υποστήριξης προσδίδει μεγάλη σημασία στην εύρεση του βέλτιστου υπερεπιπέδου για την επίτευξη μέγιστου περιθωρίου διαχωρισμού.



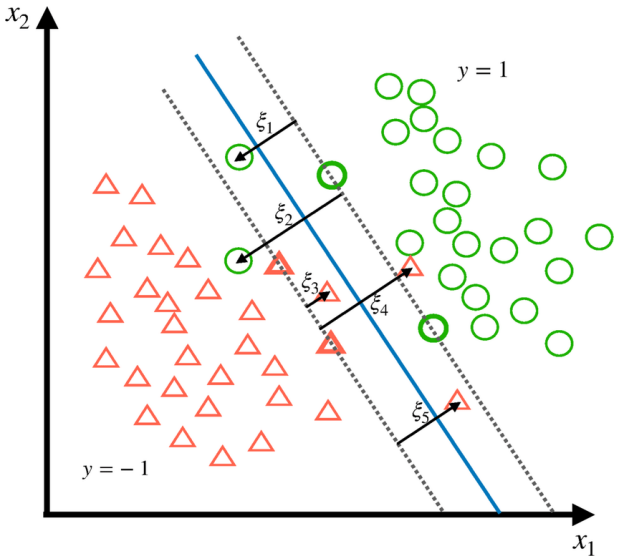

**Εικόνα 11: Βέλτιστο υπερεπίπεδο για γραμμικά προβλήματα**

### 3.4.3.2 Μη Γραμμικά Διαχωρίσιμα Προβλήματα – Μεταβλητές Χαλαρότητας

Όταν το πρόβλημα δεν είναι γραμμικά διαχωρίσιμο, εισάγουμε για κάθε πρότυπο μια μεταβλητή χαλαρότητας $\xi_i \geq 0$ έτσι ώστε:

$$d_i(w^T x_i + b) \geq 1 - \xi_i$$

Η μεταβλητή χαλαρότητας $x_i$ επιτρέπει λάθους ταξινόμησης, με $\xi_i > 1$ σημαίνοντας ότι το πρότυπο $x_i$ έχει ταξινομηθεί λανθασμένα, ενώ αν $0 < \xi_i \leq 1$, το πρότυπο ταξινομείται σωστά αλλά παραμένει εντός της περιοχής διαχωρισμού. Ουσιαστικά, η μεταβλητή χαλάρωσης $\xi_i$ επιτρέπει μικρές παραβάσεις από τον αυστηρό κανόνα του γραμμικού διαχωρισμού, καθιστώντας τον αλγόριθμο πιο ευέλικτο σε πραγματικά δεδομένα.




**Εικόνα 12: Βέλτιστο υπερεπίπεδο για μη γραμμικά προβλήματα**

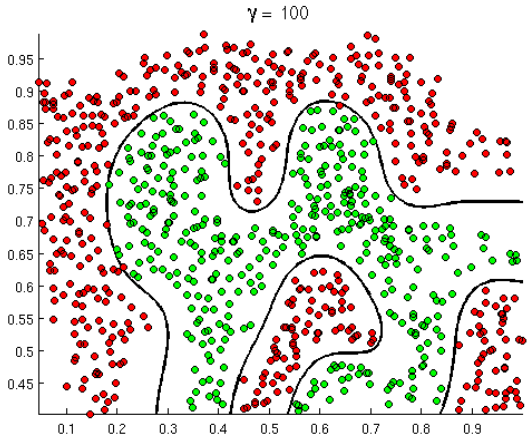
### 3.4.3.3 Μη Γραμμικά Διαχωρίσιμα Προβλήματα –Συναρτήσεις Πυρήνα

Όταν το πρόβλημα δεν είναι γραμμικά διαχωρίσιμο, μια προσέγγιση είναι η αναγωγή του σε έναν χώρο μεγαλύτερης διάστασης, ενδεχομένως άπειρης, μέσω του μετασχηματισμού $x \to \Phi(x)$. Σε αυτόν το νέο χώρο, τα παραδείγματα εκπαίδευσης γίνονται "αραιά", και το πρόβλημα μετατρέπεται σε ένα γραμμικά διαχωρίσιμο. Ο υπολογισμός του $\Phi(x)$ μπορεί να είναι εξαιρετικά πολύπλοκος όσο αυξάνεται ο αριθμός των διαστάσεων. Ευτυχώς, όμως, δεν χρειάζεται να υπολογίσουμε το $\Phi(x)$, αλλά το εσωτερικό γινόμενο $K(x, y) = \Phi(x)^T \Phi(y)$, το οποίο απλά είναι ένας αριθμός.

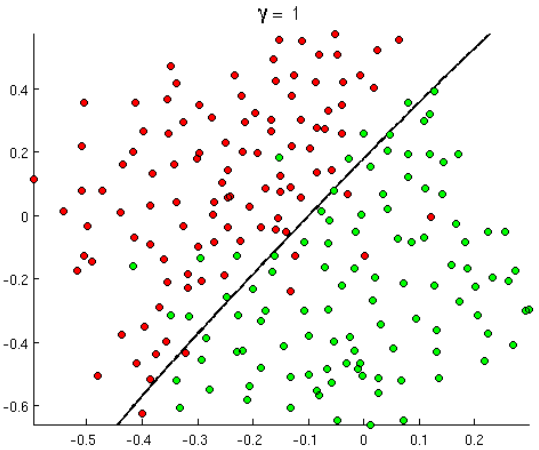
Αυτή η συνάρτηση $K(x, y)$ καλείται πυρήνας (kernel) και μπορεί να είναι μία από τις εξής:

- Γκαουσιανή Rbf: $(K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma||\boldsymbol{x} - \boldsymbol{y}||^2))$
- Πολυωνυμική: $(K(\boldsymbol{x}, \boldsymbol{y}) = \left(\boldsymbol{x^T y} + r\right)^d)$
- Σιγμοειδής: $(K(\boldsymbol{x}, \boldsymbol{y}) = \tanh\left(\gamma \boldsymbol{x^T y} + r\right))$

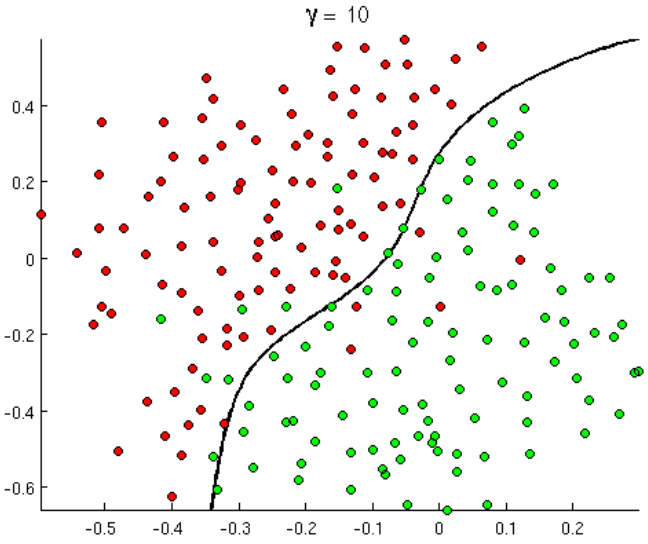
Για παράδειγμα, στην συνάρτηση Rbf, η παράμετρος $\gamma$ καθορίζει πόσο μακριά φτάνει η επιρροή ενός παραδείγματος εκπαίδευσης, με χαμηλές τιμές να σημαίνουν "μακριά" και υψηλές τιμές "κοντά". Συνεπώς, η παράμετρος $\gamma$ μπορεί να θεωρηθεί ως το αντίστροφο της ακτίνας επιρροής των δειγμάτων που επιλέγονται από το μοντέλο ως διανύσματα υποστήριξης. Στην εικόνα 13 φαίνεται η κατασκευή του decision boundary κάνοντας χρήση του rbf πυρήνα, ενώ στις εικόνες 14,15,16,17 φαίνεται η επιρροή της παραμέτρου γ του rbf πυρήνα στην κατασκευή του decision boundary οδηγώντας τελικά σε overfitting .



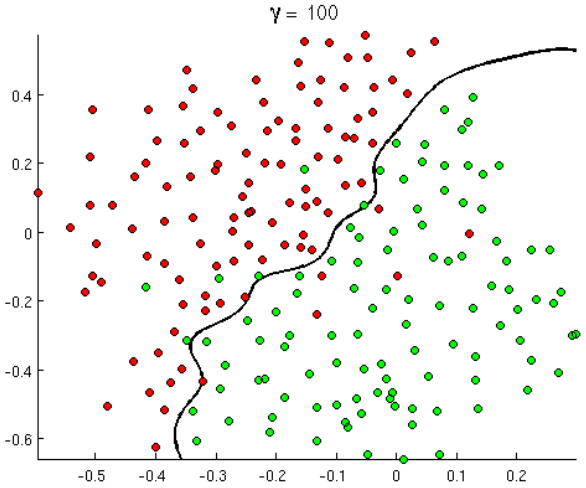

**Εικόνα 13: RBF kernel**



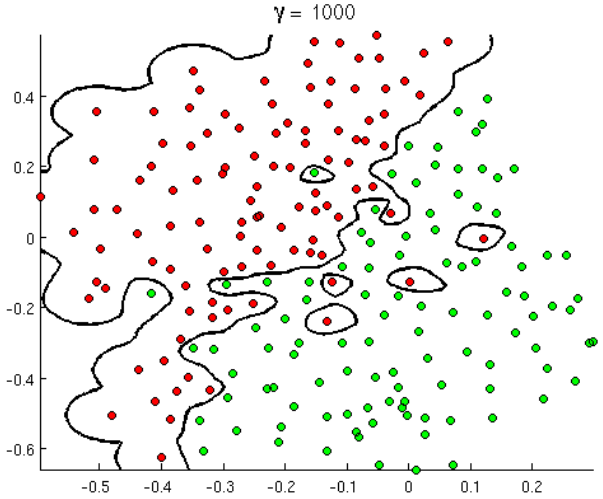

**Εικόνα 14: RBF kernel με γ=1**




**Εικόνα 15: RBF kernel με επιρροή γ=10**




**Εικόνα 16: RBF kernel με επιρροή γ=100**




**Εικόνα 17: RBF kernel με επιρροή γ=1000**

## 3.5 Νευρωνικά Δίκτυα

### 3.5.1 Τεχνητά Νευρωνικά Δίκτυα

Τα τεχνητά νευρωνικά δίκτυα [17], γνωστά και ως νευρωνικά δίκτυα, αντλούν έμπνευση από τη λειτουργία των νευρώνων και του εγκεφάλου. Εξαπλώνονται εκτενώς σε διάφορα προβλήματα επιβλεπόμενης μηχανικής μάθησης, επιτυγχάνοντας σημαντικά αποτελέσματα. Αν και αρχικά αναπτύχθηκαν για να μοντελοποιήσουν τη συμπεριφορά του ανθρώπινου εγκεφάλου, τα νευρωνικά δίκτυα έχουν ακολουθήσει μια εξέλιξη διαφορετική από τη νευροβιολογία, διατηρώντας παράλληλα κάποιες αναλογίες.

Αυτό το μοντέλο μηχανικής μάθησης εκπαιδεύεται με στόχο τη ρύθμιση των εσωτερικών του παραμέτρων, ελαχιστοποιώντας μια συνάρτηση κόστους (όπως θα εξηγηθεί). Η βασική δομική μονάδα του είναι ο τεχνητός νευρώνας, με τον πιο απλό τύπο του να είναι το perceptron του Rosenblatt.

### 3.5.2 Perceptron

Το perceptron του Rosenblatt αποτελεί τον πιο απλό νευρώνα σε ένα νευρωνικό δίκτυο, αφού αποτελείται από έναν μόνο νευρώνα. Σύμφωνα με το συγκεκριμένο μοντέλο:

- Δέχεται $m$ σήματα εισόδου, $x1$, $x2$, … , $xm$.
- Λαμβάνει μια σταθερή είσοδο $x0 = 1$, η οποία αντιστοιχεί στην πόλωση $b$.
- Υπολογίζει τον γραμμικό συνδυασμό $v = \boldsymbol{w}T\boldsymbol{x}$, όπου $\boldsymbol{w} = (b, w1, w2, \dots, wm)T$ είναι το επαυξημένο διάνυσμα βαρών, $b$ η εξωτερικά εφαρμοζόμενη πόλωση και $w1$, $w2$, … , $wm$ τα συναπτικά βάρη του perceptron που αντιστοιχούν στις εισόδους $x1$, $x2$,…, $xm$.
- Περνά τον γραμμικό συνδυασμό $v$ μέσα από μια συνάρτηση ενεργοποίησης, η οποία μπορεί να είναι είτε η μοναδιαία βηματική συνάρτηση (unit step function με τιμές 0/1), είτε η συνάρτηση προσήμου (sign function με τιμές -1/1), και τελικά παράγει το σήμα εξόδου $y$.

Οι δυνατότητες του perceptron, ωστόσο, είναι περιορισμένες. Μπορεί να αντιμετωπίσει μόνο γραμμικά διαχωρίσιμα προβλήματα, κατασκευάζοντας ένα υπερεπίπεδο $\boldsymbol{w}T\boldsymbol{x} + b = 0$, του οποίου οι παράμετροι $\boldsymbol{w}$ και $b$ καθορίζονται από τον κανόνα εκπαίδευσης του perceptron που περιγράφεται παρακάτω. Επίσης μπορεί να επιλύσει ορισμένες λογικές συναρτήσεις, όπως AND, OR και NOT, δεν μπορεί να αντιμετωπίσει τη συνάρτηση XOR.

#### 3.5.2.1 Οδηγίες Εκπαίδευσης του Perceptron

Δεδομένα: $N$ πρότυπα εισόδου $x1$, … , $xN$ μαζί με τα αντίστοιχα διανύσματα επιθυμητών αποκρίσεων $d1$, … , $dN$.

1) Ξεκινούμε αρχικοποιώντας το επαυξημένο διάνυσμα βαρών $\boldsymbol{w}(0) = 0$.
2) Εισάγουμε τα πρότυπα με τη σειρά (η κυκλική παρουσίαση όλων των προτύπων συνιστά μία εποχή). Για κάθε πρότυπο:
   - Υπολογίζουμε την απόκριση του perceptron ως $y(n) = sgn(\boldsymbol{w}\boldsymbol{T}(n)\boldsymbol{x}(n))$.
   - Ενημερώνουμε το διάνυσμα βαρών του perceptron σύμφωνα με τον κανόνα $\boldsymbol{w}(n + 1) = \boldsymbol{w}(n) + n(d(n) - y(n))\boldsymbol{x}(n)$.

Ο αλγόριθμος ολοκληρώνεται όταν όλα τα πρότυπα ταξινομούνται σωστά. Σε προβλήματα που δεν είναι γραμμικά διαχωρίσιμα, ο αλγόριθμος δεν φτάνει ποτέ στο τέλος.

Η παράμετρος $n$, γνωστή ως ρυθμός μάθησης, είναι κρίσιμη. Ένα μεγάλο $n$ μπορεί να επιταχύνει τη σύγκλιση, αλλά και να οδηγήσει σε ταλάντωση γύρω από τις βέλτιστες τιμές βαρών. Αντίθετα, ένα μικρό $n$ οδηγεί σε πιο αργή σύγκλιση.

### 3.5.3 Νευρωνικά Δίκτυα Πολλών Επιπέδων

Στην αντιμετώπιση προβλημάτων που δεν είναι γραμμικά διαχωρίσιμα, τα τεχνητά νευρωνικά δίκτυα πολλών επιπέδων, γνωστά και ως πολυστρωματικά perceptrons (MLP), αποτελούν μια αποτελεσματική λύση. Ένα MLP αποτελείται από:

- Ένα επίπεδο εισόδου, που απλώς μεταβιβάζει τα σήματα εισόδου σε όλους τους νευρώνες του κρυφού επιπέδου.
- Ένα ή περισσότερα κρυφά επίπεδα (hidden layers) που αποτελούνται από μη γραμμικούς νευρώνες.

- Ένα επίπεδο εξόδου, που αποτελείται είτε από γραμμικούς είτε από μη γραμμικούς νευρώνες. Η επιλογή αυτή εξαρτάται από την εφαρμογή, χρησιμοποιώντας γραμμικούς για προβλήματα προσέγγισης συναρτήσεων και μη γραμμικούς για προβλήματα ταξινόμησης.

Σύμφωνα με το θεώρημα του καθολικού προσεγγιστή, ένα κρυφό επίπεδο μη γραμμικών νευρώνων είναι αρκετό για την προσέγγιση οποιασδήποτε συνεχούς συνάρτησης ή αντιστοιχίας εισόδου-εξόδου. Γι' αυτό, συνήθως εξετάζονται MLP με τρία επίπεδα, ένα κρυφό επίπεδο, χωρίς περαιτέρω μελέτη πιο βαθιών αρχιτεκτονικών. Όσο πιο πολύπλοκη είναι η συνάρτηση που θέλουμε να προσεγγίσουμε, τόσο περισσότερους κρυφούς νευρώνες χρειαζόμαστε.

Η αρχιτεκτονική ενός τριών επιπέδων δικτύου με τρεις εισόδους και δύο εξόδους φαίνεται στην εικόνα 18. Η συνάρτηση ενεργοποίησης ενός τυχαίου νευρώνα του δικτύου πρέπει να είναι διαφορίσιμη για να επιτρέπεται η εκπαίδευση με τον κανόνα του gradient descent. Συνήθως χρησιμοποιούνται οι ακόλουθες συναρτήσεις:

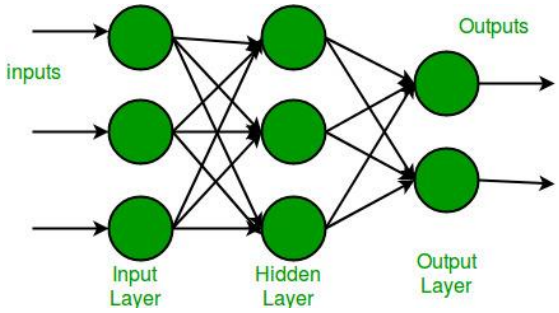
- Λογιστική Συνάρτηση: Σιγμοειδής μη γραμμικότητα.

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

- Συνάρτηση Υπερβολικής Εφαπτομένης: Άλλη σιγμοειδής μορφή μη γραμμικότητας.
-

$$tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

- Γραμμική Συνάρτηση: Για νευρώνες εξόδου για παλινδρόμηση.
  $\varphi(v) = av$
- Softmax Συνάρτηση: Για νευρώνες εξόδου για ταξινόμηση, παρέχοντας πιθανότητες εξόδου.




**Εικόνα 18: MLP 3 επιπέδων**

### 3.5.4 Εκπαίδευση MLP – Αλγόριθμος Backpropagation

Αφού καθορίσουμε την αρχιτεκτονική του δικτύου, επιλέγοντας τον αριθμό των επιπέδων, τον αριθμό των νευρώνων ανά επίπεδο και τις συναρτήσεις ενεργοποίησης, προχωρούμε στην εκπαίδευσή του, χρησιμοποιώντας ένα επισημειωμένο σύνολο παραδειγμάτων εκπαίδευσης στη μορφή:

$$\{(x_1, d_1), (x_2, d_2), \dots, (x_N, d_N)\}$$

Για ένα πρόβλημα ταξινόμησης με $k$ κατηγορίες, οι επιθυμητές έξοδοι $d_i$ κωδικοποιούνται ως one-hot vectors διάστασης $k$ με τη μονάδα να υποδηλώνει την αντίστοιχη κατηγορία. Έτσι, το νευρωνικό θα έχει $k$ νευρώνες εξόδου, ενώ οι νευρώνες εισόδου είναι ίσοι με τη διάσταση του feature vector $x$. Για το πλήθος των κρυφών νευρώνων, δεν υπάρχει συγκεκριμένος κανόνας, και συνήθως γίνεται κάποιος πειραματισμός.

Η εκπαίδευση του νευρωνικού δικτύου στοχεύει στην κατάλληλη επιλογή των συναπτικών βαρών και πολώσεων για όλους τους νευρώνες, ώστε να παράγονται οι σωστές έξοδοι για κάθε είσοδο. Αυτή η επιλογή βασίζεται στην ελαχιστοποίηση μιας κατάλληλης συνάρτησης κόστους $J$, και η διόρθωση των βαρών κατευθύνεται από το επίπεδο εξόδου προς το επίπεδο εισόδου, χρησιμοποιώντας έναν κανόνα που ονομάζεται backpropagation (οπισθοδιάδοση σφάλματος). Ο αλγόριθμος διεξάγεται σε 2 φάσεις:

- Φάση εμπρόσθιας διάδοσης (forward propagation)
- Φάση ανάστροφης διάδοσης (backward propagation)

### 3.5.5 Κύκλος και Τερματισμός στη Μάθηση

Η επαναλαμβανόμενη παρουσίαση των διαφόρων παραδειγμάτων εκπαίδευσης, δηλαδή των διανυσμάτων εισόδου $x_1, \dots, x_N$, στο νευρωνικό δίκτυο κατά τη διάρκεια της εκπαίδευσης, ονομάζεται "εποχή". Η διαδικασία εκπαίδευσης μπορεί να τερματιστεί με διάφορα κριτήρια, όπως η επιλογή ενός κατωφλίου για την ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος ή του μέσου κόστους διεντροπίας. Μπορεί επίσης να επιλεγεί ο μέγιστος αριθμός επαναλήψεων. Για την αντιμετώπιση του φαινομένου υπερεκπαίδευσης, που σχετίζεται με την αδυναμία γενίκευσης του μοντέλου σε νέα δεδομένα, μπορούμε να χρησιμοποιήσουμε ένα υποσύνολο των δεδομένων εκπαίδευσης, γνωστό ως "σύνολο ελέγχου" (validation set), για την παρακολούθηση της απόδοσης του μοντέλου κατά τη διάρκεια κάθε εποχής. Η διαδικασία εκπαίδευσης μπορεί να τερματίζεται όταν η απόδοση στο σύνολο επικύρωσης δεν βελτιώνεται για κάποιο συνεχόμενο αριθμό επαναλήψεων, γνωστό ως "Early Stopping".

Υπάρχουν δύο κύρια είδη μάθησης, ανάλογα με τη συνάρτηση που ελαχιστοποιείται και το πότε γίνονται οι ανανεώσεις των βαρών. Αναφέραμε προηγουμένως την "on-line" μάθηση, όπου οι διορθώσεις βαρών πραγματοποιούνται για κάθε παράδειγμα ξεχωριστά, με συνάρτηση κόστους προς ελαχιστοποίηση ως το στιγμιαίο τετραγωνικό σφάλμα ή το στιγμιαίο κόστος διεντροπίας. Η δεύτερη μέθοδος είναι η "μαζική" μάθηση, όπου οι προσαρμογές στα συναπτικά βάρη εκτελούνται μετά την παρουσίαση όλων των παραδειγμάτων εκπαίδευσης. Εδώ, η συνάρτηση κόστους προς ελαχιστοποίηση είναι το μέσο τετραγωνικό σφάλμα ή το μέσο κόστος διεντροπίας. Ενώ η "on-line" μάθηση είναι απλή στην υλοποίηση και συγκλίνει πιο γρήγορα, η "μαζική" μάθηση παρουσιάζει πλεονεκτήματα στην παραλληλοποίηση, αν και απαιτεί περισσότερο χώρο αποθήκευσης και περισσότερες εποχές εκπαίδευσης.

### 3.5.6 Αρχιτεκτονικές Νευρικών Δικτύων

Τα νευρικά δίκτυα διακρίνονται σε συγκεκριμένους τύπους που εξειδικεύονται σε διάφορες εργασίες. Για παράδειγμα, μια εργασία που περιλαμβάνει την αναγνώριση προσώπων εφαρμόζεται με συνελικτικά νευρικά δίκτυα (CNN), ενώ μια εργασία μετάφρασης μηχανής χρησιμοποιεί αναδρομικά νευρωνικά δίκτυα (RNN). Στις επόμενες ενότητες θα παρουσιάσουμε μια σύντομη επισκόπηση των πιο κοινών αρχιτεκτονικών νευρικών δικτύων και των χρήσεων τους.

Ωστόσο, αξίζει να σημειώσουμε ότι οι αρχιτεκτονικές αυτές δεν είναι στατικές και εξελίσσονται συνεχώς με την έρευνα και την τεχνολογική πρόοδο. Συνεπώς, υπάρχει πάντα χώρος για καινοτόμες προσεγγίσεις και τη δημιουργία νέων τύπων νευρικών δικτύων που να προσαρμόζονται σε συγκεκριμένα προβλήματα και εφαρμογές.

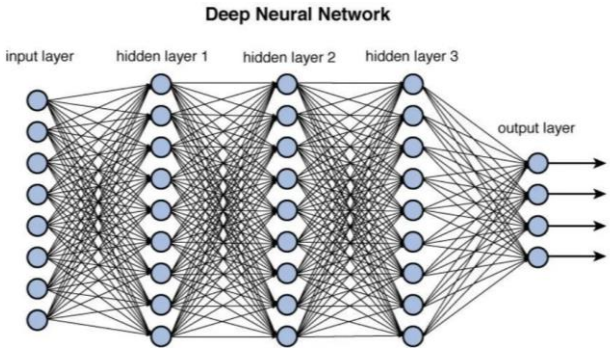
#### 3.5.6.1 Επιφανειακά Νευρικά Δίκτυα

Γενικά, ένα επιφανειακό νευρικό δίκτυο (SNN) αποτελείται από ένα κρυφό επίπεδο. Ορισμένες από τις κύριες χρήσεις του περιλαμβάνουν μοντέλα μηχανικής μάθησης που είδαμε στις προηγούμενες ενότητες, όπως η λογιστική παλινδρόμηση και οι μηχανές υποστήριξης διανυσμάτων.

Παρόλο που τα ΕΝΔ είναι απλοί σε δομή, μπορούν να χρησιμοποιηθούν για πολλές βασικές εργασίες ταξινόμησης και πρόβλεψης. Ενδεικτικά παραδείγματα περιλαμβάνουν μοντέλα όπως η λογιστική παλινδρόμηση και οι μηχανές υποστήριξης διανυσμάτων.

#### 3.5.6.2 Βαθιά Νευρικά Δίκτυα

Ένα βαθύ νευρικό δίκτυο (DNN) αποτελεί ένα δίκτυο με περισσότερα από ένα κρυφό επίπεδο. Χρησιμοποιούνται για πιο πολύπλοκες εργασίες, καθώς τα πολλά επίπεδα προσθέτουν περισσότερη πολυπλοκότητα στο δίκτυο. Τα ΒΝΔ διακρίνονται σε δύο κατηγορίες, προώθησης (feedforward) και αντίστροφης (backward), ανάλογα με τη διαδικασία ροής τους [18]. Στην εικόνα 19, μπορούμε να δούμε μια τυπική δομή ΒΝΔ. Εκτός από το επίπεδο εισόδου, όπου τα χαρακτηριστικά τροφοδοτούνται στο δίκτυο, αποτελείται από τρία κρυφά επίπεδα πλήρως συνδεδεμένα με τα κόμβους του επόμενου και του προηγούμενου επιπέδου. Ως επίπεδο εξόδου του δικτύου χρησιμοποιεί ένα επίπεδο με τέσσερις κόμβους, δηλαδή τέσσερες εξόδους.



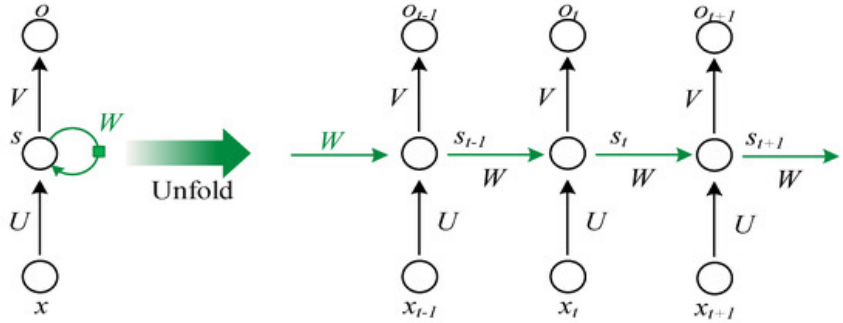

**Εικόνα 19: Δομή βαθιάς νευρωνικού δικτύου**

### 3.5.6.3 Συνελικτικά Νευρωνικά Δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα (CNN) αρχικά αναπτύχθηκαν για εφαρμογές σχετικές με την όραση των υπολογιστών [19]και είναι το κυρίαρχο είδος νευρωνικού δικτύου για αυτόν τον τομέα. Σήμερα, σήμερα, η χρήση των CNN έχει επεκταθεί επίσης στην επεξεργασία φυσικής γλώσσας, με πολλές νέες μελέτες να επιτυγχάνουν υποσχόμενα αποτελέσματα. Για παράδειγμα, στη μελέτη [20]εξηγείται πώς χρησιμοποιούνται προ-εκπαιδευμένα διανύσματα λέξεων για την εκπαίδευση ενός CNN στην κατηγοριοποίηση στο επίπεδο προτάσεων. Ουσιαστικά, ένα CNN αποτελείται από πολλά συνελικτικά επίπεδα στα κρυφά του επίπεδα, τα οποία αναλαμβάνουν διάφορες εργασίες, όπως η εξαγωγή χαρακτηριστικών, η μείωση της ανάλυσης για την ανεξαρτησία των χαρακτηριστικών από θόρυβο και ασήμαντες αλλαγές [18].

### 3.5.6.4 Επαναληπτικά Νευρωνικά Δίκτυα

Τα επαναληπτικά νευρωνικά δίκτυα (RNN) αποτελούν δημοφιλή επιλογή στην ανάλυση συναισθημάτων και στη φυσική γλώσσα γενικότερα. Το κύριο πλεονέκτημα αυτού του είδους νευρωνικών δικτύων είναι η ικανότητά τους να χρησιμοποιούν πληροφορίες από το προηγούμενο χρονικό βήμα για να εκτιμήσουν το τρέχον βήμα[21]. Αυτό ονομάζεται "μνήμη" και επιτρέπει στα RNN να διαβάζουν τις εισόδους με σειρά και να θυμούνται πληροφορίες από τα προηγούμενα επίπεδα. Ένα παράδειγμα RNN απεικονίζεται στην παρακάτω εικόνα. Εδώ παρατηρούμε τρία επίπεδα ενός τυπικού RNN παραδείγματος. Καθώς τα RNN λειτουργούν σε ακολουθίες, το πλήθος των επιπέδων (χρονικά βήματα) συσχετίζεται με τον αριθμό των εισόδων (λέξεις σε μια πρόταση). Τα δεδομένα στοιχεία είναι το διάνυσμα $Xt$ που αντιπροσωπεύει την είσοδο στο χρονικό βήμα $t$ και το επίπεδο $ht$ που αντιπροσωπεύει το κρυφό επίπεδο του ίδιου χρονικού βήματος. Οι εισαγωγές του $ht$ βασίζονται στους υπολογισμούς του προηγούμενου επιπέδου και την είσοδο του τρέχοντος $Xt$. Οι προβλέψεις $y$ λαμβάνουν χώρα σε κάθε χρονικό βήμα.




**Εικόνα 20: Τυπικό παράδειγμα RNN**

Καθώς τα επαναληπτικά νευρωνικά δίκτυα είναι πολύ βαθιά, το δίκτυο ενδέχεται να αντιμετωπίσει προβλήματα σχετικά με την εξαφάνιση ή την έκρηξη των κλίσεων. Αυτό σημαίνει ότι οι παράγωγοι μπορεί να μειώνονται ή να αυξάνονται εκθετικά. Μια λύση για αυτό το πρόβλημα είναι τα νευρωνικά δίκτυα μακράς και σύντομης μνήμης (LSTM). Αυτά τα δίκτυα χρησιμοποιούν πύλες σε κάθε κόμβο που καθορίζουν ποια τιμή θα εξαχθεί και ποια όχι: • Πύλη λύσης, η οποία υπογράφει μια τιμή βάσει της κατανομής Bernoulli και καθορίζει εάν οι πληροφορίες θα διατηρηθούν ή θα ξεχαστούν. • Πύλη εισαγωγής που δείχνει ποιες από τις τιμές θα ενημερωθούν. • Πύλη εξόδου, που καθορίζει την πρόβλεψη του τρέχοντος χρονικού βήματος.

# Κεφάλαιο 4: Προ-εκπαιδευμένα Μοντέλα

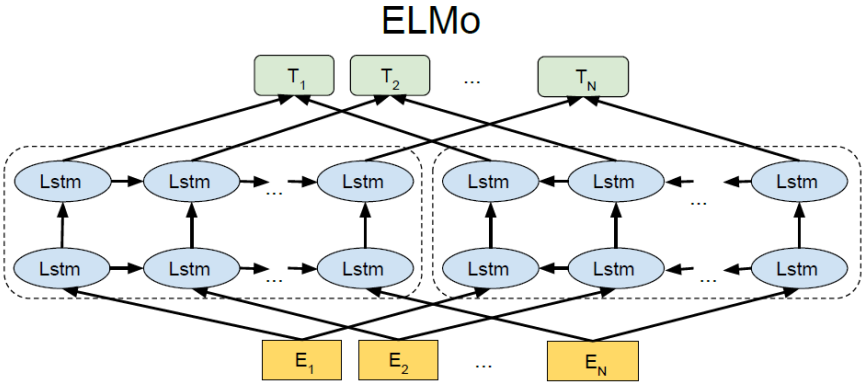
Τα προ-εκπαιδευμένα μοντέλα λειτουργούν με τον τρόπο της παραγωγής διανυσμάτων ενσωμάτωσης συμφραζομένων, δηλαδή παρουσιάζουν το κείμενο λαμβάνοντας υπόψη τα περιβάλλοντά τους και τα συγκεκριμένα πλαίσια στα οποία χρησιμοποιούνται. Αυτό επιτρέπει την κατανόηση της εννοιολογικής σημασίας κάθε λέξης στο συγκεκριμένο πλαίσιο. Στη συνέχεια, παρατίθενται μοντέλα ενσωμάτωσης (contextual models) που χρησιμοποιούνται στον τομέα έρευνας της ΕΦΓ.

## 4.1 Μοντέλο ELMo

Το μοντέλο ELMo[22] αναπαριστά λέξεις βασισμένο σε συμφραζόμενα, παρέχοντας βαθιές και πλούσιες αναπαραστάσεις λέξεων. Κατά την επεξεργασία, συντακτικά και σημασιολογικά χαρακτηριστικά της φυσικής γλώσσας κωδικοποιούνται, και οι αναπαραστάσεις προσαρμόζονται ανάλογα με τη χρήση και το νόημα σε ένα συγκεκριμένο κείμενο.

Το μοντέλο χρησιμοποιεί ένα αμφίδρομο νευρωνικό δίκτυο μακράς βραχυχρόνιας μνήμης (biLSTM). Οι αναπαραστάσεις δημιουργούνται λαμβάνοντας υπόψη τόσο τις επόμενες όσο και τις προηγούμενες λέξεις ενός όρου μέσα στα πλαίσια μιας πρότασης. Κάθε biLSTM επίπεδο επεξεργάζεται διαφορετικά τη λέξη σχετικά με το συντακτικό και το νόημα της, παράγοντας ένα διάνυσμα αναπαράστασης.

Τα υψηλότερα επίπεδα του biLSTM επικεντρώνονται στη σημασιολογία των όρων, ενώ τα χαμηλότερα αναλύουν το συντακτικό ρόλο κάθε λέξης. Τελικά, η λέξη αναπαρίσταται ως γραμμικός συνδυασμός των διανυσμάτων, εξασφαλίζοντας πολλαπλές αναπαραστάσεις ανάλογα με τα συμφραζόμενα που την περιβάλλουν.




**Εικόνα 21: Αρχιτεκτονική ELMo[31]**

## 4.2 Μετασχηματιστές

Οι μετασχηματιστές (transformers) αποτελούν προηγμένα γλωσσικά μοντέλα βαθιάς μηχανικής μάθησης, εγκαταλείποντας την αρχιτεκτονική κωδικοποιητή-αποκωδικοποιητή (encoder-decoder) που χρησιμοποιούν τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks - RNNs) σε ακολουθίες-σε-ακολουθίες (sequence-to-sequence - seq2seq)[23]. Σημαντική παράμετρος στην εκπαίδευση των μετασχηματιστών είναι η έννοια της "προσοχής". Αυτός ο μηχανισμός λαμβάνει κωδικοποιημένα διανύσματα ως είσοδο και βοηθά το μοντέλο να αναγνωρίσει και να εκμεταλλευτεί τη χρήσιμη πληροφορία που περιέχουν, λαμβάνοντας υπόψη τη σχέση των συμφραζομένων. Έτσι, οι αναπαραστάσεις δημιουργούνται με ακρίβεια, ενώ παράλληλα μειώνεται σημαντικά ο χρόνος επεξεργασίας.

### 4.2.1 Ανάλυση συναισθημάτων με τη χρήση Μετασχηματιστών

Μια εφαρμογή των μοντέλων Transformers είναι η αναγνώριση των συναισθημάτων μέσα από το κείμενο [24]. Τα μοντέλα αυτά διαθέτουν τη δυνατότητα να αναγνωρίζουν λέξεις που εκφράζουν συναίσθημα και, ουσιαστικά, να καταλήγουν σε ένα συμπέρασμα για το περιεχόμενο του κειμένου. Σκοπός είναι να ανιχνεύουν αυτόματα και να εξάγουν πληροφορίες από το κείμενο, όπως το συναίσθημα που εκφράζει ο συγγραφέας ή η άποψή του για το συγκεκριμένο κείμενο.

Η ανάλυση συναισθημάτων με τη χρήση των Transformers προσφέρει μια πολύτιμη δυνατότητα αναγνώρισης συναισθηματικών στοιχείων σε κείμενα. Αυτή η τεχνική μπορεί να εφαρμοστεί σε διάφορους τομείς:

- Επιχειρήσεις: Η αναγνώριση των συναισθημάτων των πελατών μέσω σχολίων και κριτικών σε προϊόντα διευκολύνει τις επιχειρήσεις να κατανοήσουν τις ανάγκες και τις προτιμήσεις τους.

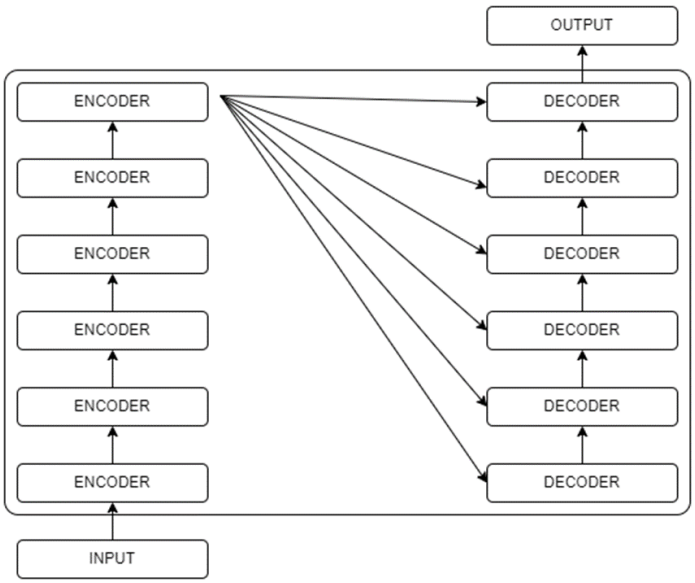
- Πολιτική: Η ανάλυση της κοινής γνώμης και των συναισθημάτων συμβάλλει στην καλύτερη κατανόηση των πολιτικών απαιτήσεων και επιδιώξεων του κοινού.

- Υγειονομική περίθαλψη: Η χρήση αυτοματοποιημένων συστημάτων μπορεί να συμβάλει στη λήψη αποφάσεων για τη βελτίωση του συστήματος υγειονομικής περίθαλψης με βάση τις ανάγκες των ασθενών.

### 4.2.2 Μηχανισμός Προσοχής

Ο μηχανισμός προσοχής παρέχει στο μοντέλο τη δυνατότητα να επικεντρωθεί σε λέξεις που έχουν στενή σχέση με την επεξεργαζόμενη λέξη. Στην αρχιτεκτονική του μετασχηματιστή, ειδικά στον κωδικοποιητή και τον αποκωδικοποιητή, χρησιμοποιείται η τεχνική της αυτό-προσοχής (Self-Attention), που περιγράφεται και απεικονίζεται στην εικόνα 22. Στο πλαίσιο αυτού του μηχανισμού, υπολογίζεται η σημασιολογική σχέση της λέξης με κάθε άλλη λέξη στην ακολουθία, εξετάζοντας όλους τους πιθανούς τρόπους σύνδεσής της. Στη συνέχεια, επιλέγονται αυτές με τις υψηλότερες βαθμολογίες.

Όσον αφορά τον αποκωδικοποιητή, εφαρμόζεται επίσης μια παραλλαγή της προσοχής, γνωστή ως προσοχή κωδικοποιητή-αποκωδικοποιητή (Encoder-Decoder Attention). Σε αυτόν τον μηχανισμό, λαμβάνεται η έξοδος των κωδικοποιητών του μετασχηματιστή και, σε συνδυασμό με τα αποτελέσματα της αυτό-προσοχής που προηγείται στον αποκωδικοποιητή, διαμορφώνεται μια βαθμολογία για τη σχετικότητα των συμφραζομένων όρων. Αυτή η βαθμολογία ενσωματώνεται στην αναπαράσταση της λέξης, επιτρέποντας στο μοντέλο να αντιληφθεί τις σημαντικές πληροφορίες.




**Εικόνα 22: Αρχιτεκτονική Μετασχηματιστή[25]**

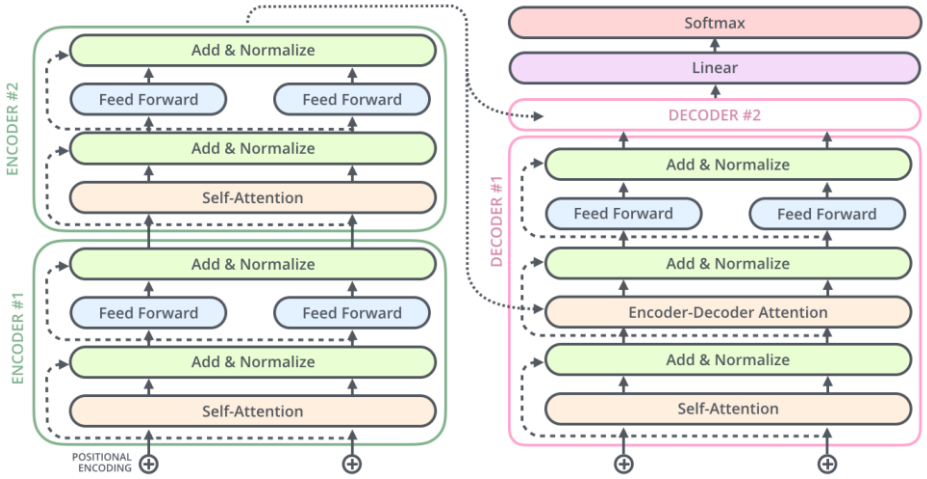
### 4.2.3 Πολλαπλες Κεφαλές Προσοχής

Κατά την υλοποίηση του Μετασχηματιστή, ο μηχανισμός προσοχής εφαρμόζει τους υπολογισμούς του πολλές φορές παράλληλα. Κάθε εφαρμογή αυτής της διαδικασίας αντιστοιχεί σε μία κεφαλή προσοχής (Attention Head). Οι εξόδοι από κάθε κεφαλή συνδυάζονται, δημιουργώντας έτσι μια τελική βαθμολογία προσοχής (Attention Score). Ο Μετασχηματιστής χρησιμοποιεί μηχανισμό προσοχής με πολλαπλές κεφαλές (Multiple Attention Heads), ενισχύοντας έτσι την αποτελεσματικότητα της σημασιολογικής κωδικοποίησης της λέξης. Αυτό επιτρέπει στο μοντέλο να διακρίνει με ακρίβεια τις λέξεις που σχετίζονται με το επεξεργαζόμενο συμφραζόμενο, ενισχύοντας την κατανόηση της συνολικής σημασιολογικής δομής

### 4.2.4 Αρχιτεκτονική Μετασχηματιστή

Η βασική δομή του μοντέλου Μετασχηματιστή αποτελείται από πολλαπλούς κωδικοποιητές και αποκωδικοποιητές. Όσον αφορά τον κωδικοποιητή, όπως παρουσιάζεται στην εικόνα 23, δέχεται τα διανύσματα ενσωμάτωσης της ακολουθίας εισόδου, σε συνδυασμό με την κωδικοποιημένη αναπαράσταση της θέσης της κάθε λέξης στο κείμενο. Κάθε επίπεδο κωδικοποίησης περιλαμβάνει ένα επίπεδο αυτό-προσοχής για την εκτίμηση των σχέσεων μεταξύ των λέξεων της εισόδου. Τα αποτελέσματα του αυτό-προσοχής τροφοδοτούν το δίκτυο πρόσθιας τροφοδότησης.

Αυτή η διαδικασία επαναλαμβάνεται σειριακά, με κάθε κωδικοποιητή να λαμβάνει ως είσοδο την έξοδο του προηγούμενου. Κατά τη μεταφορά πληροφορίας από ένα επίπεδο στο άλλο, η έξοδος κανονικοποιείται για σταθεροποίηση. Οι αποκωδικοποιητές, από την άλλη πλευρά, δέχονται το διάνυσμα ενσωμάτωσης της επιθυμητής ακολουθίας εξόδου ως αρχική είσοδο. Η αρχιτεκτονική είναι πανομοιότυπη με τους κωδικοποιητές και περιλαμβάνει επίσης επίπεδα αυτο-προσοχής και νευρωνικό δίκτυο πρόσθιας τροφοδότησης. Ένα επιπλέον επίπεδο κωδικοποίησης-αποκωδικοποίησης παρεμβάλλεται μεταξύ του αυτο-προσοχής και του δικτύου πρόσθιας τροφοδότησης, λαμβάνοντας είσοδο την έξοδο των κωδικοποιητών.




**Εικόνα 23: Αρχιτεκτονική Μετασχηματιστή με 2 κωδικοποιητές και 2 αποκωδικοποιητές [5]**

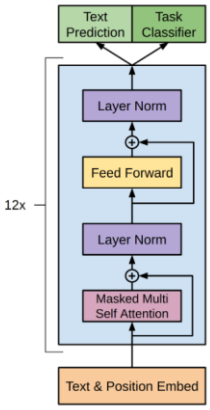
## 4.3 GPT

Το Generative Pre-Training Transformer (GPT) [26]είναι μια προσέγγιση στην κατανόηση και επεξεργασία φυσικής γλώσσας που συνδυάζει την μη-εποπτευόμενη προ-εκπαίδευση σε μεγάλα σώματα κειμένου με την εποπτευόμενη προσαρμογή σε συγκεκριμένες εργασίες. Το κύριο πλεονέκτημα του GPT είναι ο όγκος της πληροφορίας που έχει απορροφήσει κατά την προ-εκπαίδευσή του σε μεγάλα σώματα κειμένου.

Κάθε νέα γενιά του GPT, όπως το GPT-2 και το GPT-3, εκπαιδεύεται σε σώματα κειμένου με πολλαπλάσιες παραμέτρους. Η μεγάλη κλίμακα του μοντέλου επιτρέπει εύκολη εκπαίδευση σε εξειδικευμένα σύνολα δεδομένων με μικρές προσαρμογές, χωρίς την ανάγκη για μεγάλη ποσότητα νέας πληροφορίας.

Το GPT χρησιμοποιεί μια μορφή μεταφοράς μάθησης, εκπαιδεύοντας αρχικά το γλωσσικό μοντέλο σε μεγάλα σώματα κειμένου και στη συνέχεια προσαρμόζοντάς το σε συγκεκριμένες εργασίες. Η αρχιτεκτονική του GPT περιλαμβάνει μια στοίβα αποκωδικοποιητών με 12 επίπεδα, καθένα από τα οποία περιλαμβάνει μηχανισμό αυτο-προσοχής και ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης.

Κατά την επεξεργασία, το GPT βασίζεται στις προηγούμενες λέξεις και την ίδια τη λέξη που εξετάζει, χρησιμοποιώντας πληροφορίες από το παρελθόν προς το παρόν μιας πρότασης.




**Εικόνα 24: Αρχιτεκτονική GPT[27]**

## 4.4 BERT

Το BERT (Bidirectional Encoder Representations from Transformers) [28]είναι ένα γλωσσικό μοντέλο που σχεδιάστηκε για να εκπαιδεύει προφορικές αναπαραστάσεις εξετάζοντας τόσο το περιεχόμενο που προηγείται όσο και αυτό που ακολουθεί έναν όρο στην ακολουθία εισόδου. Σε αντίθεση με τον μετασχηματιστή GPT που περιγράφηκε προηγουμένως, το BERT υιοθετεί μια διπλή κατεύθυνση προσέγγιση, επιτρέποντας την ανάγνωση της ακολουθίας εισόδου στο σύνολό της. Αυτό σημαίνει ότι λαμβάνει υπόψη τόσο τις λέξεις πριν από έναν όρο όσο και αυτές που ακολουθούν.

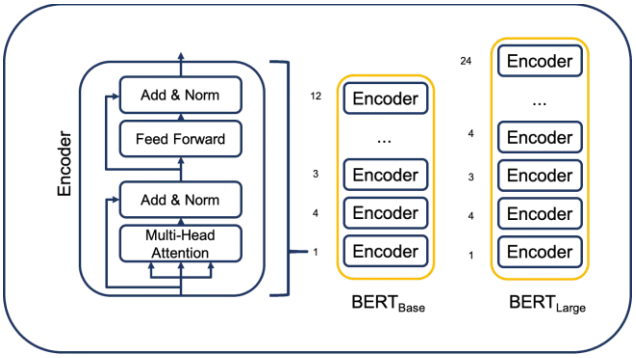
Συνεπώς, η προσέγγιση του BERT διαφέρει από την "αριστερά προς τα δεξιά" επεξεργασία του μοντέλου GPT. Αντί για την μονόπλευρη προσέγγιση προς τη μία κατεύθυνση, το BERT αναγνωρίζει και αξιοποιεί τα συμφραζόμενα αμφίδρομα, βοηθώντας έτσι στην ακριβή αποκωδικοποίηση της σημασίας που παρέχουν τα συμφραζόμενα στην κάθε λέξη της πρότασης.

### 4.4.1 Αρχιτεκτονική BERT

Η δομή του BERT (Bidirectional Encoder Representations from Transformers) [12]βασίζεται στην αρχιτεκτονική των μετασχηματιστών. Σε αντίθεση με τον GPT, που χρησιμοποιεί και τους δύο κωδικοποιητές και αποκωδικοποιητές, το BERT εκμεταλλεύεται μόνο τον κωδικοποιητή του μοντέλου, όπως αναλύθηκε στην προηγούμενη ενότητα. Ο κωδικοποιητής αποτελείται από έναν μηχανισμό αυτο-προσοχής και ένα νευρωνικό δίκτυο πρόσθιας τροφοδότησης.

Πιο συγκεκριμένα, το BERT χρησιμοποιεί μια πολυεπίπεδη δομή κωδικοποιητών. Κάθε επίπεδο κωδικοποιητή εκπαιδεύεται να ερμηνεύει σημασιολογικά τις λέξεις μιας λέξης στην ακολουθία εισόδου, λαμβάνοντας υπόψη τις προηγούμενες και τις επόμενες λέξεις. Υπάρχουν δύο εκδοχές του BERT μοντέλου:

- BERT BASE: Αντιστοιχεί σε μέγεθος με τον GPT μετασχηματιστή και αποτελείται από 12 κεφαλές προσοχής και 12 κωδικοποιητές, με κάθε δίκτυο πρόσθιας τροφοδότησης περιλαμβάνοντας 768 κρυφές μονάδες.
- BERT LARGE: Είναι ένα μοντέλο BERT μεγάλων διαστάσεων, με 16 κεφαλές προσοχής και 24 κωδικοποιητές με 1024 κρυφές μονάδες.




**Εικόνα 25: Αρχιτεκτονική BERT**

Για την προ-εκπαίδευση του BERT, χρησιμοποιούνται δύο βασικές στρατηγικές χωρίς επίβλεψη: το γλωσσικό μοντέλο απόκρυψης (Masked Language Model - MLM) και η πρόβλεψη της επόμενης πρότασης (Next Sentence Prediction - NSP). Αυτές οι στρατηγικές συνεισφέρουν στην αποτελεσματική εκπαίδευση του μοντέλου για να κατανοήσει τις σημασιολογικές σχέσεις μεταξύ των λέξεων στο κείμενο.

#### 4.4.1.1 Γλωσσικό μοντέλο απόκρυψης

Η εκπαίδευση περισσότερων γλωσσικών μοντέλων γίνεται συνήθως με δύο βασικές κατευθύνσεις: είτε "από αριστερά προς τα δεξιά" είτε "από δεξιά προς τα αριστερά". Η αμφίδρομη επεξεργασία, δηλαδή η δυνατότητα της κάθε λέξης να "δει τον εαυτό της," θα μπορούσε να οδηγήσει σε ένα αναποτελεσματικό μοντέλο, όπου η πρόβλεψη της κάθε λέξης χάνει το νόημα.

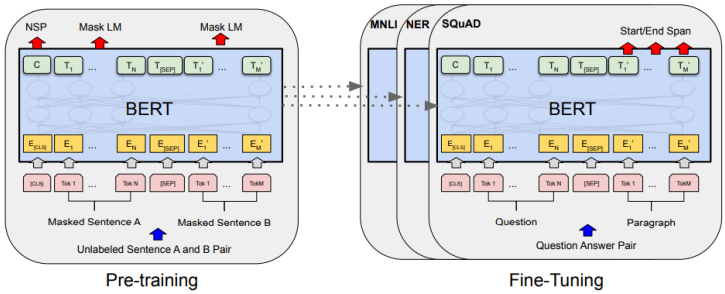
Για να αντιμετωπιστεί αυτό το ζήτημα και να επιτραπεί η εκπαίδευση και από τις δύο κατευθύνσεις, πραγματοποιείται απόκρυψη του 15% των λέξεων της ακολουθίας εισόδου. Από το σύνολο των επιλεγμένων λέξεων, το 80% αντικαθίσταται από το ειδικό σύμβολο [MASK] του μετασχηματιστή. Το 10% των υπόλοιπων όρων αντικαθίσταται από μια διαφορετική λέξη, ενώ οι υπόλοιπες 10% παραμένουν αναλλοίωτες. Με αυτόν τον τρόπο, το μοντέλο μπορεί να εξετάσει αμφίδρομα όλα τα συμφραζόμενα της πρότασης, προσπαθώντας να προβλέψει τη λέξη που λείπει. Αυτή η πρόβλεψη αποτελεί την έξοδο του μοντέλου.

#### 4.4.1.2 Πρόβλεψη επόμενης πρότασης

Χρησιμοποιώντας την τεχνική πρόβλεψης της επόμενης πρότασης, το μοντέλο στοχεύει να προβλέψει εάν υπάρχει σχέση μεταξύ δύο προτάσεων ή όχι. Κατά την εκπαίδευση του μοντέλου, χρησιμοποιείται ένα σύνολο δεδομένων που περιλαμβάνει ζεύγη προτάσεων. Σε περίπτωση που αντιστοιχεί στο 50% των περιπτώσεων, η δεύτερη πρόταση είναι ακριβώς ίδια με αυτή που ακολουθεί την πρώτη στο αρχικό κείμενο. Στα υπόλοιπα 50%, η δεύτερη πρόταση επιλέγεται τυχαία.

### 4.4.2 Εκτεταμένη Εφαρμογή BERT στην ΕΦΓ

Το προ-εκπαιδευμένο μοντέλο BERT αποτελεί ένα ισχυρό εργαλείο που μπορεί να εφαρμοστεί σε διάφορες περιπτώσεις στον τομέα της ΕΦΓ (εικόνα 26). Η βαθιά κατανόηση της γλώσσας που παρέχει ο μετασχηματιστής απλοποιεί σημαντικά την επανακατάρτιση μοντέλων ειδικού σκοπού. Τα μοντέλα μπορούν να αρχικοποιηθούν με τις παραμέτρους του BERT και, στη συνέχεια, να υποστούν επιπλέον εκπαίδευση σε ένα ειδικό σύνολο δεδομένων που αφορά αποκλειστικά τον τομέα εργασίας που καθορίζεται από τη συγκεκριμένη εφαρμογή.




**Εικόνα 26: Διαδικασία προ-εκπαίδευσης και του fine-tuning του BERT[29]**

## 4.4 RoBERTa

Η αποτελεσματικότητα του BERT το έχει καθιερώσει ως ένα από τα κυρίαρχα συστήματα επεξεργασίας φυσικής γλώσσας. Προκειμένου να βελτιωθεί ακόμα περισσότερο, αναπτύχθηκε το μοντέλο RoBERTa (Robustly Optimized BERT) Pre training από την Facebook AI [30]. Οι επιδόσεις του RoBERTa έχουν τη δυνατότητα να ανταγωνιστούν και να ξεπεράσουν το GPT σε ορισμένες περιπτώσεις.

Η σύγκριση των δύο μοντέλων γίνεται σε ένα ευρύ φάσμα εργασιών της επεξεργασίας φυσικής γλώσσας, με κάθε εργασία να αντιστοιχεί σε αντίστοιχη συλλογή πληροφορίας. Η προσέγγιση του RoBERTa διατηρεί τη βασική αρχιτεκτονική του GPT μοντέλου, αλλά επιφέρει αλλαγές στον σχεδιασμό του, όπως η εκπαίδευση σε επιπλέον δεδομένα και για μεγαλύτερο χρονικό διάστημα, η απαλοιφή της πρόβλεψης της επόμενης πρότασης, η χρήση μεγαλύτερων ακολουθιών εισόδου, και η δυναμική απόκρυψη των λέξεων της ακολουθίας εισόδου (dynamic masking).

Η επιτυχία του RoBERTa ανοίγει νέους ορίζοντες στην έρευνα της επεξεργασίας φυσικής γλώσσας, ενώ πολλές παραλλαγές του (όπως DistilBERT, XLNet κ.ά.) συνεχίζουν να επεκτείνουν τα πεδία της ΕΦΓ. Στο πλαίσιο αυτής της διπλωματικής εργασίας, το RoBERTa χρησιμοποιήθηκε ως γλωσσικό μοντέλο, συγκρινόμενο με την προ-εκπαιδευμένη έκδοση του GPT στην ελληνική γλώσσα.

# Κεφάλαιο 5: Πειραματική Διαδικασία

## 5.1 Περιγραφή-Εργαλεία

Στα πλαίσια της παρούσας διπλωματικής εργασίας, χρησιμοποιήθηκε 10-fold cross-validation για την εκπαίδευση και αξιολόγηση γλωσσικών μοντέλων ανάλυσης συναισθήματος σε κείμενα που είναι γραμμένα στα ελληνικά και προέρχονται από το twitter.

Για την υλοποίηση των παραπάνω αλγορίθμων βαθιάς μηχανικής μάθησης, χρησιμοποιήθηκε η ανοιχτού κώδικα βιβλιοθήκη μετασχηματιστών HuggingFace [31]. Η πλατφόρμα αυτή παρέχει εργαλεία και βιβλιοθήκες για εφαρμογές στην επεξεργασία φυσικής γλώσσας και διατίθεται για χρήση στις βιβλιοθήκες βαθιάς μάθησης PyTorch [32] και TensorFlow [33] της γλώσσας προγραμματισμού Python.

Η πειραματική διαδικασία που ακολουθήθηκε περιλάμβανε την προεπεξεργασία των δεδομένων και την υλοποίηση του 10-πλου σταυρωτού ελέγχου(cross-validation).

Για να συγκρίνουμε τις επιδόσεις των ταξινομητών που αναπτύξαμε, εφαρμόσαμε επιπλέον τη μεθοδολογία σε ένα ήδη προ-εκπαιδευμένο μοντέλο, το GreekBERT [34]. Οι λεπτομερείς περιγραφές αυτών των διαδικασιών παρουσιάζονται στις επόμενες υποενότητες.

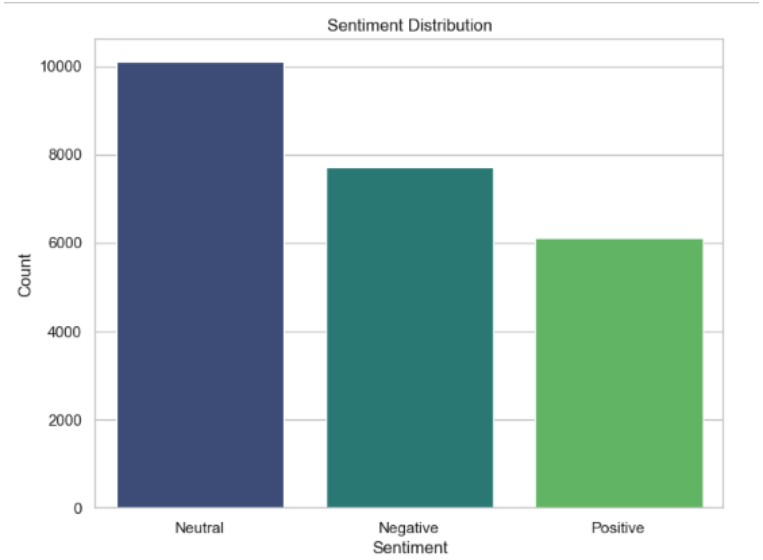
## 5.2 Περιγραφή Δεδομένων

Το dataset που χρησιμοποιήθηκε για την εκπόνηση της πτυχιακής εργασίας αναφέρεται σε δεδομένα από τον τομέα της τραπεζικής βιομηχανίας, συλλεγμένα από το Twitter στα ελληνικά. Το dataset αποτελείται από συνολικά 23,927 εγγραφές, ενώ κάθε εγγραφή περιλαμβάνει δύο βασικές στήλες: το "sentiment value" και το "text".

Η στήλη "sentiment value" αντιπροσωπεύει το συναίσθημα κάθε εγγραφής, παρέχοντας έναν δείκτη του συναισθηματικού χαρακτήρα του περιεχομένου. Αυτό το χαρακτηριστικό είναι ιδιαίτερα χρήσιμο για την ανάλυση της στάσης των χρηστών του Twitter έναντι θεμάτων που αφορούν τον τραπεζικό τομέα στο ελληνικό περιβάλλον.

Η στήλη "text" περιέχει το πραγματικό κείμενο που καταγράφηκε στο Twitter. Το κείμενο αυτό είναι στα ελληνικά και αντιπροσωπεύει τις δημοσιεύσεις των χρηστών σχετικά με θέματα που σχετίζονται με τον τραπεζικό τομέα. Η χρήση της ελληνικής γλώσσας προσφέρει ένα επιπρόσθετο στοιχείο πολυπολιτισμικής κατανόησης, ενισχύοντας την ερευνητική αξία του dataset.

Το παρόν dataset παρέχει μια ευκαιρία για την εξερεύνηση της σχέσης μεταξύ των συναισθηματικών αντιδράσεων των χρηστών στο Twitter και των θεμάτων που επηρεάζουν τον τραπεζικό τομέα στην ελληνική κοινωνία. Μέσω αναλυτικών μεθόδων επεξεργασίας φυσικής γλώσσας και εξόρυξης δεδομένων, το dataset αυτό μπορεί να αποτελέσει βασικό εργαλείο για την αναγνώριση των προτεραιοτήτων, των αντιδράσεων και των συναισθημάτων του κοινού έναντι του τραπεζικού τομέα.




**Διάγραμμα 2: Κατανομή Συναισθήματος**

## 5.3 Greek-BERT

Το γλωσσικό μοντέλο [34] έχει εκπαιδευτεί σε έναν μεγάλο όγκο δεδομένων που δεν περιλαμβάνει τις συντακτικές και σημασιολογικές ιδιομορφίες των κειμένων που εμφανίζονται στα μέσα κοινωνικής δικτύωσης. Συγκεκριμένα, το σώμα κειμένου της προ-εκπαίδευσής του περιλαμβάνει πηγές από τη Βικιπαίδεια καθώς και πηγές από ελληνικές μεταφράσεις διαδικαστικών εγγράφων του Ευρωπαϊκού Κοινοβουλίου.

Όσον αφορά τα μορφολογικά χαρακτηριστικά, στο σώμα κειμένου δεν εμφανίζονται τόνοι στα φωνήεντα, ούτε χρησιμοποιούνται κεφαλαίοι χαρακτήρες.

## 5.4 Προεπεξεργασία Δεδομένων

Στην αρχή της επεξεργασίας των δεδομένων μας, πραγματοποιήσαμε τη διαγραφή των στοιχείων με κενά σχόλια/tweets (null text), καθώς και των δεδομένων που δεν είχαν κατηγοριοποιηθεί ανάλογα με το συναίσθημα που εκφράζουν (null sentiment). Αυτή η ενέργεια οδήγησε στη διατήρηση του συνολικού αριθμού των επισημειώσεων δεδομένων από 23,927 σε 23,927. Επίσης έγινε διαγραφή των περιττών χαρακτηριστικών:

- Αφαίρεση ηλεκτρονικών διευθύνσεων (URLs)
- Αφαίρεση emoticons
- Αφαίρεση hashtags (#)
- Αφαίρεση χαρακτήρων Retweet (RT)
- Αφαίρεση αναφορών λογαριασμού (Mentions @)

Επιπλέον, μετατρέψαμε τις συμβολοσειρές που υποδηλώνουν το συναίσθημα των δεδομένων ('neutral', 'positive', 'negative') σε αριθμητικούς χαρακτήρες. Τελικά, οι ετικέτες που αντιστοιχούν σε κάθε συναίσθημα ορίστηκαν ως εξής:

'neutral' → 0, 'positive' → 1, 'negative' → 2.

Επιπλέον, αντικαταστήσαμε τα τονισμένα φωνήεντα με τους αντίστοιχους μη τονισμένους χαρακτήρες και διαγράψαμε περιττούς χαρακτήρες κενού που προϋπήρχαν. Ο στόχος αυτών των ενεργειών ήταν να βελτιστοποιήσουν την ποιότητα των δεδομένων προτού προχωρήσουμε στην εκπαίδευση του μοντέλου.

Εκτός από τη βασική προεπεξεργασία που περιγράφηκε παραπάνω, ακολουθήσαμε μια συνολική προσέγγιση στην επεξεργασία του κειμένου με σκοπό τη βελτίωση της συνολικής απόδοσης του μοντέλου όπου αυτή είναι διαγραφή σημείων στίξης και μετατροπή κεφαλαίων χαρακτήρων σε πεζούς.

## 5.5 Ανάλυση σε Σύμβολα

Η διαδικασία της ανάλυσης σε σύμβολα, γνωστή και ως tokenization, αποτελεί ένα βασικό στάδιο για την ανάπτυξη αποδοτικών γλωσσικών μοντέλων. Καθώς δημιουργείται ένα εκφραστικό και πλήρες λεξικό, επηρεάζει αμέσως την ικανότητα γενίκευσης του μοντέλου. Η γενίκευση αφορά τη δυνατότητα του μοντέλου να προσαρμόζεται επιτυχώς σε νέα δεδομένα, τα οποία δεν είχε συναντήσει κατά την εκπαίδευσή του.

Εν ολίγοις, η διαδικασία αυτή αποτελεί βασικό προαπαιτούμενο για την απόδοση του μοντέλου, καθώς επηρεάζει τον τρόπο που αντιλαμβάνεται το κείμενο και τις λέξεις. Η δημιουργία ενός καλού λεξικού εξασφαλίζει την αντιπροσώπευση των λέξεων με τρόπο που επιτρέπει στο μοντέλο να αντιμετωπίζει προκλήσεις, όπως την προσαρμογή σε νέα, μη οικεία δεδομένα.

## 5.6 Λεπτή Προσαρμογή

Η τεχνική της λεπτής προσαρμογής (fine-tuning) αναφέρεται στην προσαρμογή ενός προ εκπαιδευμένου γλωσσικού μοντέλου με σκοπό την εκτέλεση μιας συγκεκριμένης εργασίας, όπως η ανάλυση συναισθημάτων (ΕΦΓ). Αυτή η διαδικασία περιλαμβάνει την επιπλέον εκπαίδευση του μοντέλου σε ένα μικρότερο σύνολο δεδομένων, το οποίο, ειδικά, είναι 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 προ εκπαιδευμένου, καθώς και τα ποιοτικά και ποσοτικά χαρακτηριστικά του συνόλου δεδομένων πάνω στο οποίο πραγματοποιείται η λεπτή προσαρμογή, μπορούν να επηρεάσουν σημαντικά την αποδοτικότητα της τεχνικής αυτής, καθώς και τις απαιτήσεις του χρόνου εκπαίδευσης.

Για τη μελέτη του αντικειμένου αυτής της διπλωματικής εργασίας, πραγματοποιήθηκε προσαρμογή των προπαιδευμένων μοντέλων της ενότητας, προκειμένου να αξιολογηθεί η απόδοσή τους στην ανάλυση συναισθήματος. Έγινε, συνεπώς, περαιτέρω εκπαίδευση των μοντέλων, χρησιμοποιώντας το επισημειωμένο σύνολο δεδομένων, με στόχο τη δημιουργία ενός αποτελεσματικού ταξινομητή συναισθημάτων.

Κατά τη διαδικασία εκπαίδευσης των γλωσσικών μοντέλων ανάλυσης συναισθήματος, αξιολογήθηκε μια σειρά από πειραματικούς συνδυασμούς μοντελοποίησης. Αυτοί αφορούσαν την αρχιτεκτονική δομή του μοντέλου, την προσαρμογή των υπερ-παραμέτρων του, αλλά και την ειδική διαχείριση του συνόλου δεδομένων.

### 5.6.1 Αρχιτεκτονική

Ένας κρίσιμος παράγοντας που επηρεάζει την αποτελεσματικότητα του γλωσσικού μοντέλου στην ανάλυση συναισθήματος είναι η αρχιτεκτονική δομή του. Οι επιλογές που γίνονται κατά την πειραματική διαδικασία αφορούν τις παραμέτρους που περιγράφουν τα επίπεδα και τις συναρτήσεις ενεργοποίησης που διαμορφώνουν τη δομή του ταξινομητή.

Συγκεκριμένα, κάθε προ εκπαιδευμένο γλωσσικό μοντέλο προσαρμόζεται για την ανάλυση συναισθήματος με την προσθήκη μιας "κεφαλής ταξινόμησης" (classification head). Αυτή η κεφαλή μετατρέπει τις τιμές των χαρακτηριστικών του γλωσσικού μοντέλου σε μια τελική πρόβλεψη κλάσης/κατηγορίας για τα δεδομένα εισόδου.

Πιο συγκεκριμένα, η "κεφαλή ταξινόμησης" αποτελείται από ένα ή περισσότερα πλήρως συνδεδεμένα επίπεδα. Η είσοδος αυτών των επιπέδων προέρχεται από την τελική διανυσματική αναπαράσταση του τελευταίου επιπέδου του γλωσσικού μοντέλου (last-hidden-state). Η αναπαράσταση αυτή μπορεί να πάρει δύο μορφές: την πλήρη ακολουθία εξόδου (sequence output) του τελευταίου επιπέδου του γλωσσικού μοντέλου και τη δειγματοληπτημένη ακολουθία εξόδου (pooled output) του ίδιου επιπέδου.

Στην πρώτη περίπτωση, έχουμε την τελική αναπαράσταση της ακολουθίας εισόδου, ενώ στη δεύτερη περίπτωση, για κάθε ακολουθία εισόδου, η γενική σημασιολογική έννοια περιλαμβάνεται σε μια μοναδική διανυσματική αναπαράσταση. Αυτή η αναπαράσταση χρησιμοποιείται ευρέως σε εργασίες ταξινόμησης και άλλες εφαρμογές λεπτής προσαρμογής στο πεδίο της ανάλυσης συναισθήματος και συγκεκριμένα στα γλωσσικά μοντέλα των μετασχηματιστών.

## 5.7 Προ-εκπαιδευμένα βάρη

Ένα συχνό και σημαντικό πρόγραμμα κατά την προσαρμογή ενός προ-εκπαιδευμένου μοντέλου είναι να παγώνονται τα βάρη του. Αυτή η πρακτική μπορεί να εφαρμοστεί είτε σε όλα τα προ-εκπαιδευμένα βάρη είτε να περιορίζεται στα αρχικά επίπεδα της δομής, όπου πραγματοποιείται το βασικό κομμάτι της μοντελοποίησης με απόκρυψη κειμένου.

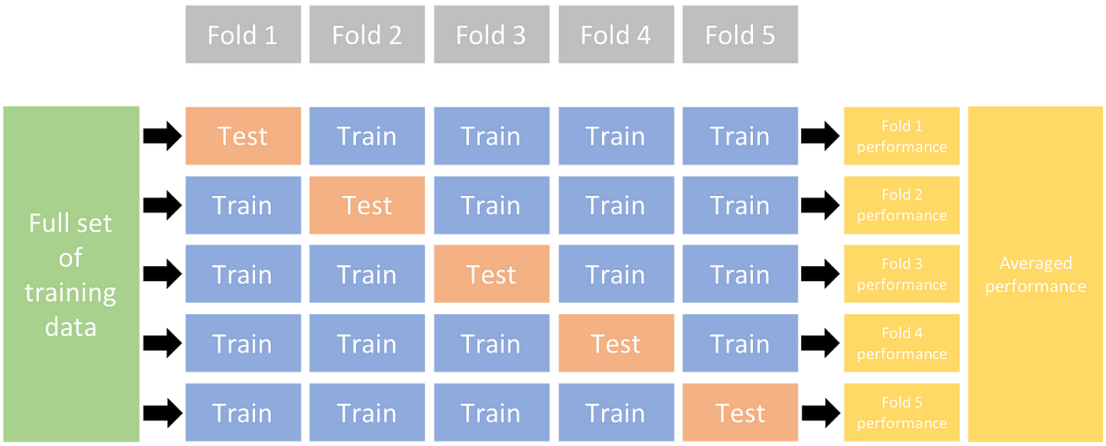
Συγκεκριμένα, οι προ-εκπαιδευμένες παράμετροι του γλωσσικού μοντέλου δεν ενημερώνονται κατά τη διαδικασία της λεπτής προσαρμογής. Κατά τον παγωμένο χρόνο, μόνο ο ταξινομητής εκπαιδεύεται στα νέα δεδομένα. Αυτό επιτυγχάνεται παγώνοντας τα προ-εκπαιδευμένα βάρη, επιτρέποντας στο μοντέλο να επικεντρωθεί αποκλειστικά στην εκπαίδευση του ταξινομητή χωρίς να προσαρμόζει ξανά τις ήδη εκπαιδευμένες παραμέτρους.

Η παγωμένη προσέγγιση είναι χρήσιμη για την αποφυγή υπερπροσαρμογής, ιδίως όταν το σώμα κειμένου της προ-εκπαίδευσης είναι σημαντικά μεγαλύτερο από το νέο σύνολο δεδομένων. Επιπλέον, λόγω της ελαχιστοποίησης των παραμέτρων που απαιτούν εκπαίδευση, μειώνεται σημαντικά η διάρκεια και, κατά συνέπεια, το κόστος της διαδικασίας.

Δεδομένου του περιορισμένου όγκου δεδομένων μας για την ανάλυση συναισθήματος, επιλέξαμε το πάγωμα των παραμέτρων του μοντέλου προ-εκπαίδευσης σε κάθε ένα από τα επίπεδα που αποτελούν τη δομή του.

## 5.8 K – Fold Cross Validation

Στην προσέγγιση της K-Fold Cross Validation, το αρχικό σύνολο δεδομένων χωρίζεται σε k υποσύνολα, γνωστά ως "folds". Αυτά τα folds είναι αμοιβαία αποκλειόμενα και έχουν περίπου ίδιο μέγεθος. Σε κάθε επανάληψη, ο κατηγοριοποιητής εκπαιδεύεται επαναληπτικά, εξετάζοντας ένα fold ως σύνολο δοκιμής και τα υπόλοιπα ως σύνολα εκπαίδευσης. Ο αριθμός των folds (k) καθορίζεται από τον χρήστη και συνήθως επιλέγεται ως 5 ή 10.




**Εικόνα 27: Παράδειγμα μεθόδου K - Fold Cross Validation με αριθμό folds = 5**

Η Εικόνα 7 επεξηγεί τη διαδικασία K-Fold Cross Validation με ένα παράδειγμα όπου υπάρχουν 5 folds. Κατά κάθε επανάληψη, ένα από τα υποσύνολα χρησιμοποιείται ως σύνολο δοκιμής, ενώ τα υπόλοιπα αποτελούν το σύνολο εκπαίδευσης. Αυτή η διαδικασία επαναλαμβάνεται k φορές, εξασφαλίζοντας ότι κάθε fold χρησιμοποιείται τουλάχιστον μία φορά ως σύνολο δοκιμής.

## 5.9 Αποτελέσματα και Αξιολόγηση

Με προσεκτική εφαρμογή της τεχνικής StratifiedKFold για τον διαχωρισμό των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, προχωρήσαμε σε ένα πειραματικό πλαίσιο, εκτελώντας την τεχνική BERT tokenization και εκπαίδευσης στο BERT μοντέλο για κάθε fold. Καταγράψαμε προσεκτικά τα αποτελέσματα, συμπεριλαμβανομένων του loss και της ακρίβειας, προσφέροντας ολοκληρωμένη αξιολόγηση της απόδοσης του μοντέλου.

Τα στατιστικά αποτελέσματα αναδεικνύουν ξεκάθαρα την ικανότητα του BERT μοντέλου στην αντιμετώπιση του προβλήματος ανάλυσης συναισθήματος στην ελληνική γλώσσα. Επιπλέον, προσθέσαμε τον πίνακα σύγχυσης, παρέχοντας μια οπτική αναπαράσταση της απόδοσης του μοντέλου σε κάθε κατηγορία. Αυτή η προσθήκη επιτρέπει μια πιο λεπτομερή κατανόηση της απόδοσης και της συμπεριφοράς του μοντέλου σε διαφορετικές κλάσεις. Οι συνολικοί υπολογισμοί των στατιστικών, όπως ο μέσος όρος της ακρίβειας και του λάθους, μαζί με τις μετρικές precision, recall, και F1-score, προσφέρουν μια πλήρη και εμπεριστατωμένη αξιολόγηση της απόδοσης του BERT μοντέλου στον τομέα της ανάλυσης συναισθήματος για την ελληνική γλώσσα.

### 5.9.1 Precision ανά fold

| FOLD | Precision |
|---|---|
| 1 | 65.70% |
| 2 | 65.37% |
| 3 | 65.69% |
| 4 | 65.24% |
| 5 | 64.73% |
| 6 | 66.01% |
| 7 | 64.90% |
| 8 | 65.63% |
| 9 | 65.21% |
| 10 | 65.27% |
| Avg | 65.06% |

**Πίνακας 1: Precision ανά Fold**

Οι τιμές προσέγγισης, που κυμαίνονται από 64,73% έως 66,01%, αποτυπώνουν τη συνοχή του μοντέλου στο να αποφεύγει τα ψευδή θετικά. Ο μέσος όρος προσέγγισης του 65,06% παρέχει μια συνολική αξιολόγηση, υποδεικνύοντας την αποτελεσματικότητα του μοντέλου στην κατάλληλη πρόβλεψη θετικών καταστάσεων σε διάφορα υποσύνολα του συνόλου δεδομένων.

### 5.9.2 Recall ανά fold

| FOLD | Recall |
|---|---|
| 1 | 65.55% |
| 2 | 65.26% |
| 3 | 65.59% |
| 4 | 65.13% |
| 5 | 64.63% |
| 6 | 65.93% |
| 7 | 64.72% |
| 8 | 65.55% |
| 9 | 65.05% |
| 10 | 65.09% |
| Avg | 65.27% |

**Πίνακας 2: Recall ανά Fold**

Η μέση τιμή του Recall ανά υποσύνολο (Fold) εκφράζει την ικανότητα του μοντέλου να ανιχνεύει σωστά τα θετικά περιγραφόμενα παραδείγματα. Με τιμές που κυμαίνονται από 64,63% έως 65,93%, το μοντέλο διατηρεί σταθερή απόδοση.

### 5.9.3 F1-Score ανά fold

| FOLD | F1-score |
|---|---|
| 1 | 65.49% |
| 2 | 65.19% |
| 3 | 65.51% |
| 4 | 65.04% |
| 5 | 64.53% |
| 6 | 65.83% |
| 7 | 64.61% |
| 8 | 65.45% |
| 9 | 65.95% |
| 10 | 64.95% |
| Avg | 65.13% |

**Πίνακας 3: F1-Score ανά Fold**

Ο πίνακας αυτός παρουσιάζει τα αποτελέσματα του F1-Score για κάθε υποσύνολο (Fold) καθώς και τον μέσο όρο αυτών των μετρικών. Το F1-Score είναι μια συνολική μετρική που λαμβάνει υπόψη τόσο την ακρίβεια όσο και την ανάκληση του μοντέλου. Με μια μέση τιμή F1-Score περίπου 65,13%, παρατηρούμε μια σταθερή απόδοση στα διάφορα υποσύνολα, υποδεικνύοντας τη συνολική ισορροπία του μοντέλου στην επίτευξη ακριβών και πλήρων προβλέψεων.

### 5.9.4 Accuracy ανά fold

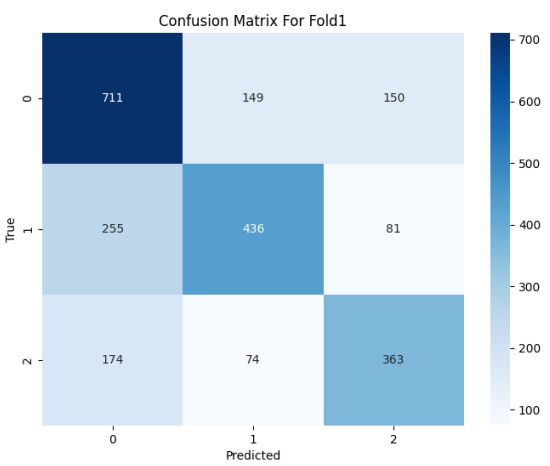
| FOLD | ACCURACY |
|---|---|
| 1 | 63.06% |
| 2 | 63.39% |
| 3 | 63.89% |
| 4 | 64.94% |
| 5 | 64.90% |
| 6 | 63.44% |
| 7 | 64.61% |
| 8 | 63.42% |
| 9 | 62.83% |
| 10 | 65.30% |
| Avg | 64.00% |

**Πίνακας 4: Ακρίβεια ανά Fold**

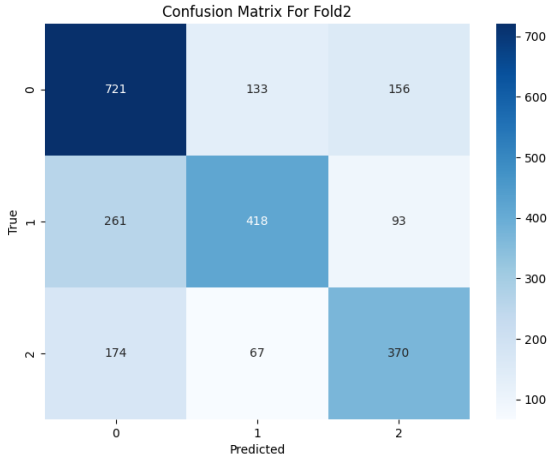
Κατά τη διαδικασία 10-fold Cross Validation, μετρήσαμε τα ποσοστά ακρίβειας σε κάθε επανάληψη για να αξιολογήσουμε τη συνολική επίδοση του μοντέλου σε διαφορετικά υποσύνολα δεδομένων. Τα αποτελέσματα καταδεικνύουν ένα σταθερό επίπεδο ακρίβειας, κυμαινόμενο από 62.83% έως 65.30%. Αυτή η σταθερότητα υποδεικνύει την αξιοπιστία και τη γενικευτικότητα του μοντέλου σε διάφορες συνθήκες εκπαίδευσης και δοκιμής

### 5.9.5 Πίνακες Σύγχυσης

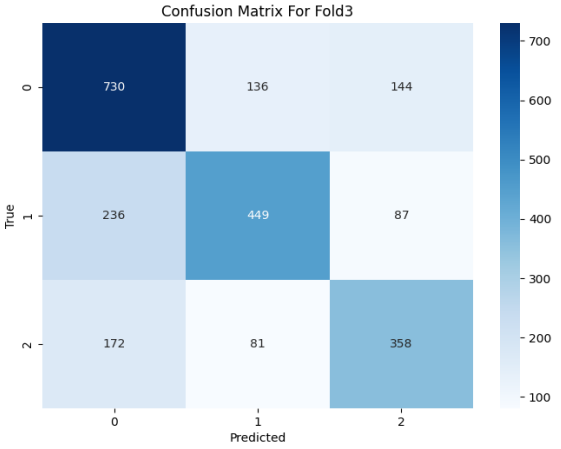
Ο πίνακας σύγχυσης (Confusion Matrix) παρέχει πληροφορίες σχετικά με τις πραγματικές και προβλεπόμενες κατηγοριοποιήσεις που πραγματοποιούνται από ένα σύστημα κατηγοριοποίησης. Με βάση τον πίνακα σύγχυσης μπορεί να υπολογιστούν η ορθότητα, η ακρίβεια, η ανάκληση, ο αρμονικός μέσος και η ειδικότητα.



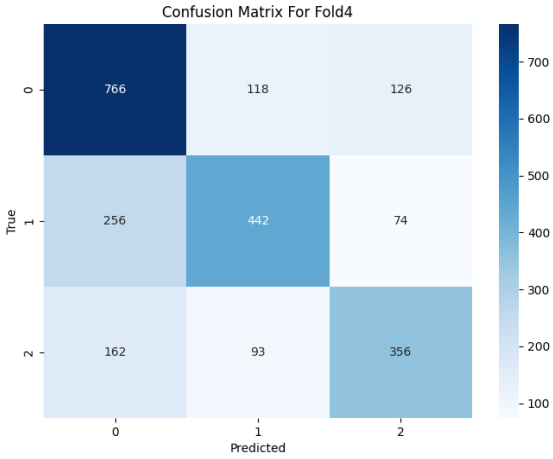

**Διάγραμμα 3: Πίνακας Σύγχυσης για Fold 1**



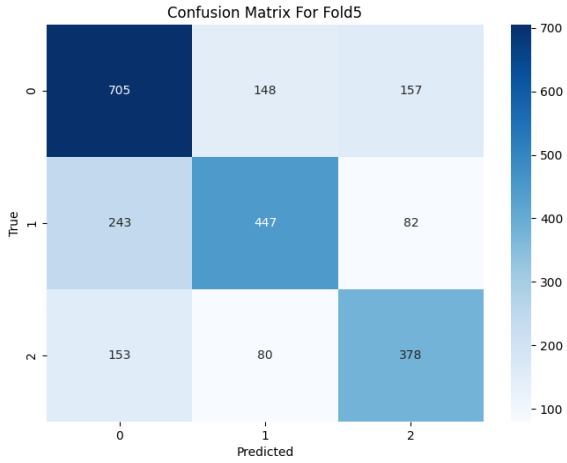

**Διάγραμμα 4: Πίνακας Σύγχυσης για Fold 2**



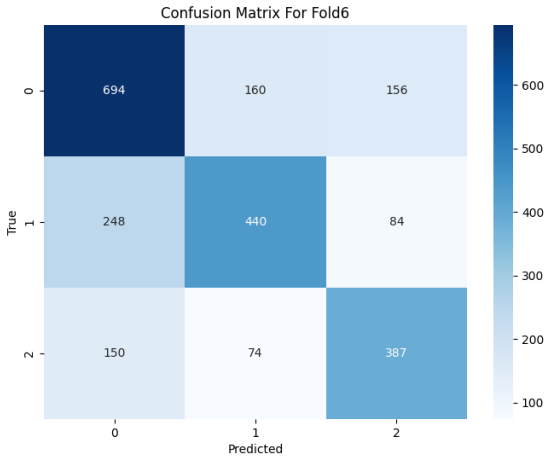

**Διάγραμμα 5: Πίνακας Σύγχυσης για Fold 3**



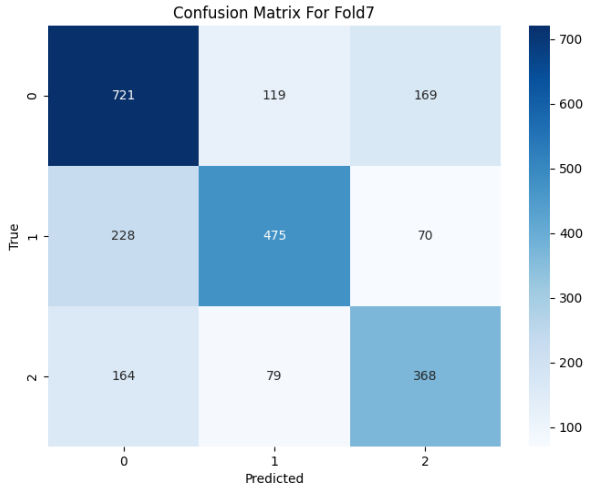

**Διάγραμμα 6: Πίνακας Σύγχυσης για Fold 4**



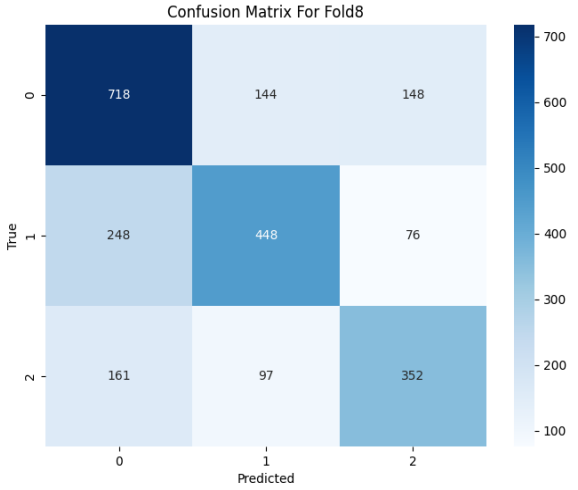

**Διάγραμμα 7: Πίνακας Σύγχυσης για Fold 5**



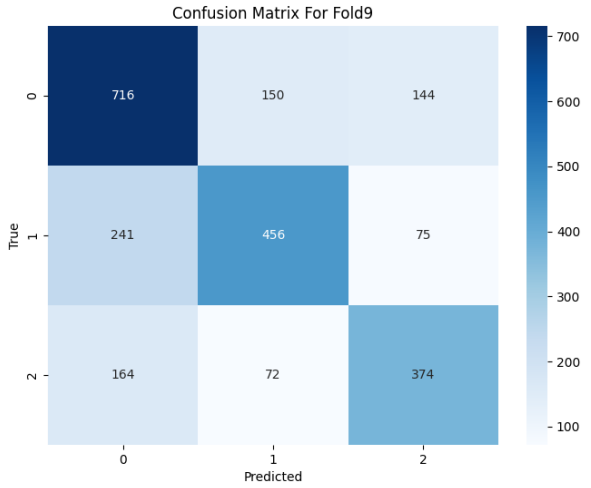

**Διάγραμμα 8: Πίνακας Σύγχυσης για Fold 6**



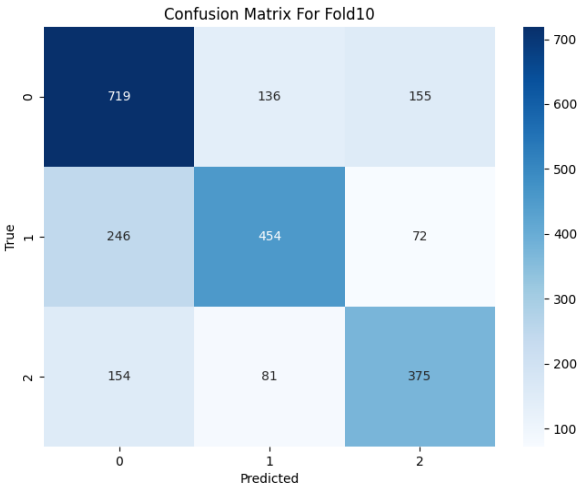

**Διάγραμμα 9: Πίνακας Σύγχυσης για Fold 7**




**Διάγραμμα 10: Πίνακας Σύγχυσης για Fold 8**

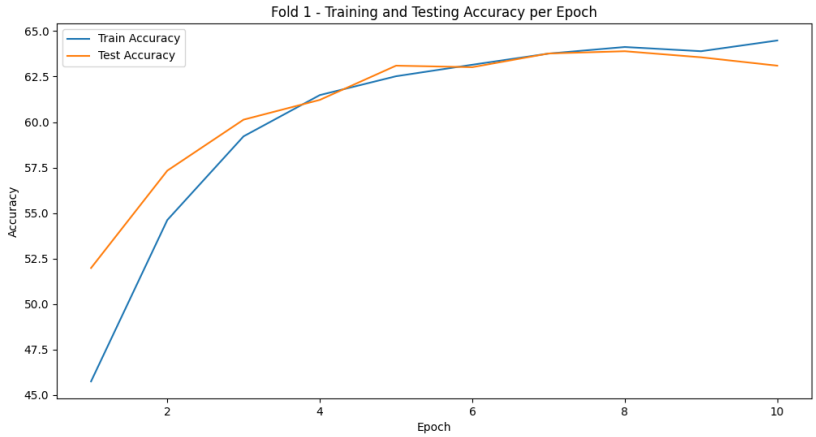



**Διάγραμμα 11: Πίνακας Σύγχυσης για Fold 9**

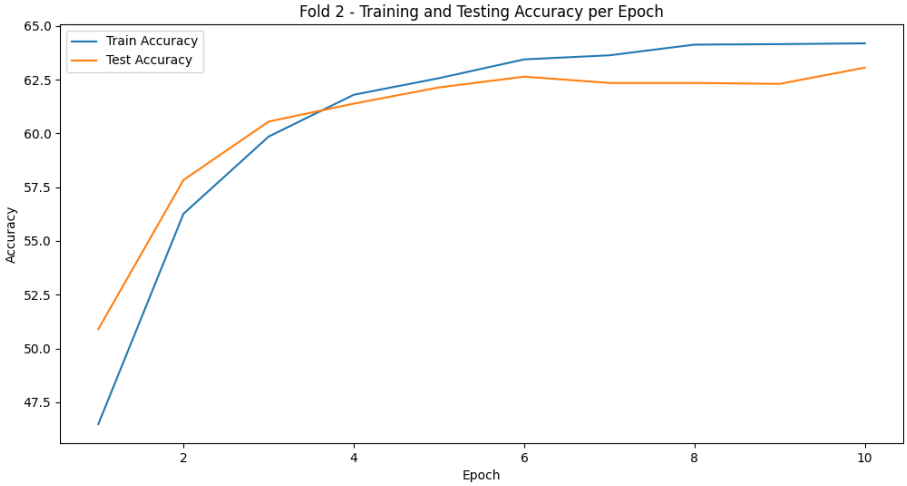



**Διάγραμμα 12: Πίνακας Σύγχυσης για Fold 10**

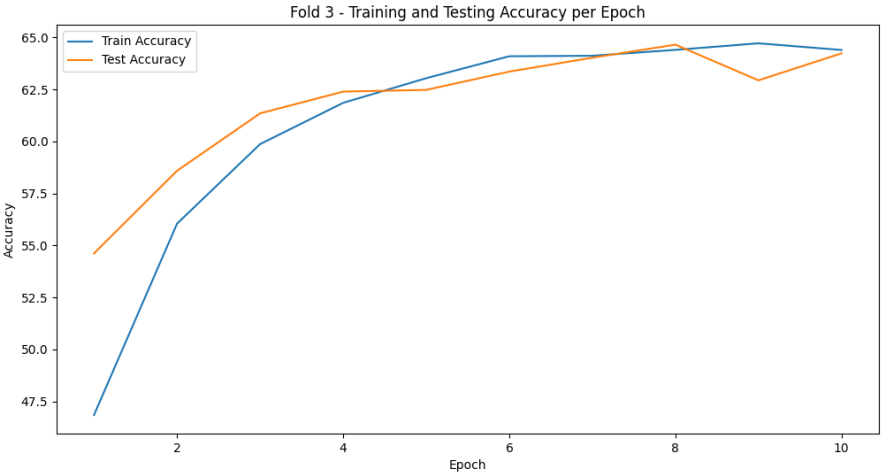
## 5.9.6 Διαγράμματα Ακρίβειας και Εκπαίδευσης



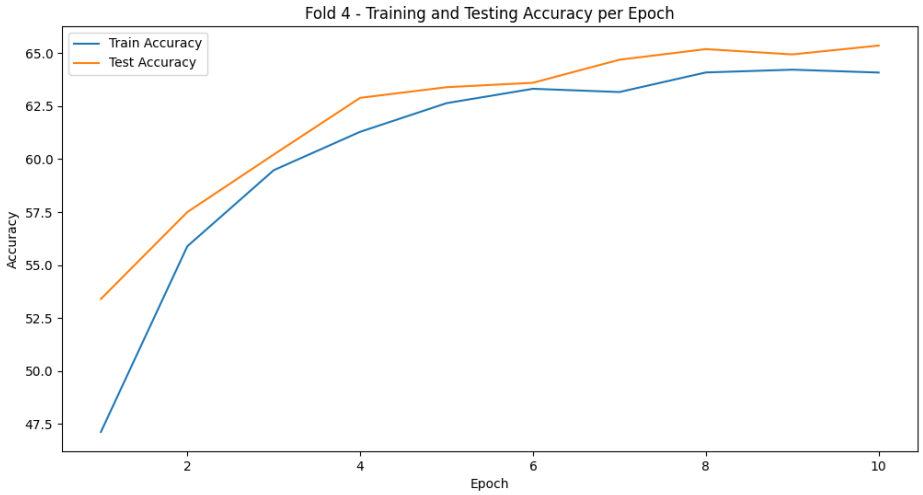

**Διάγραμμα 13: Γράφημα ακρίβειας Fold 1**



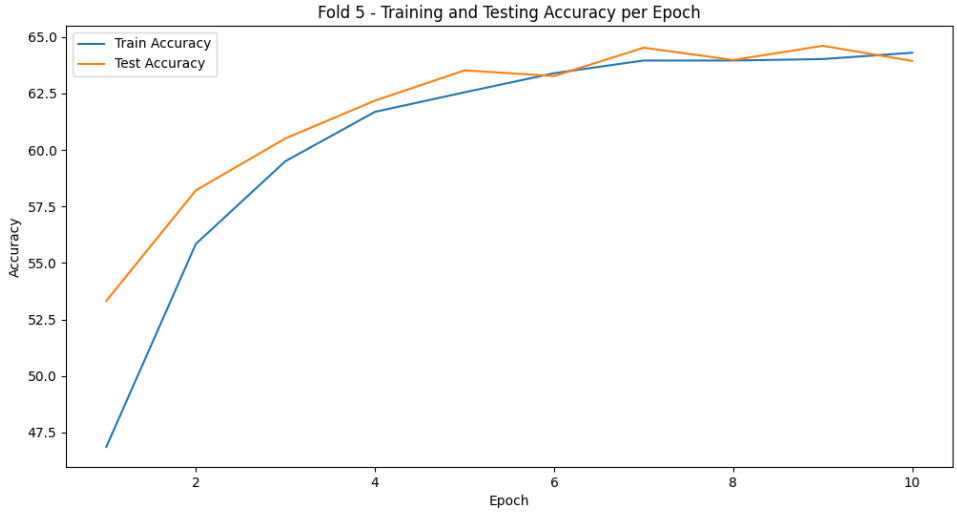

**Διάγραμμα 14: Γράφημα ακρίβειας Fold 2**



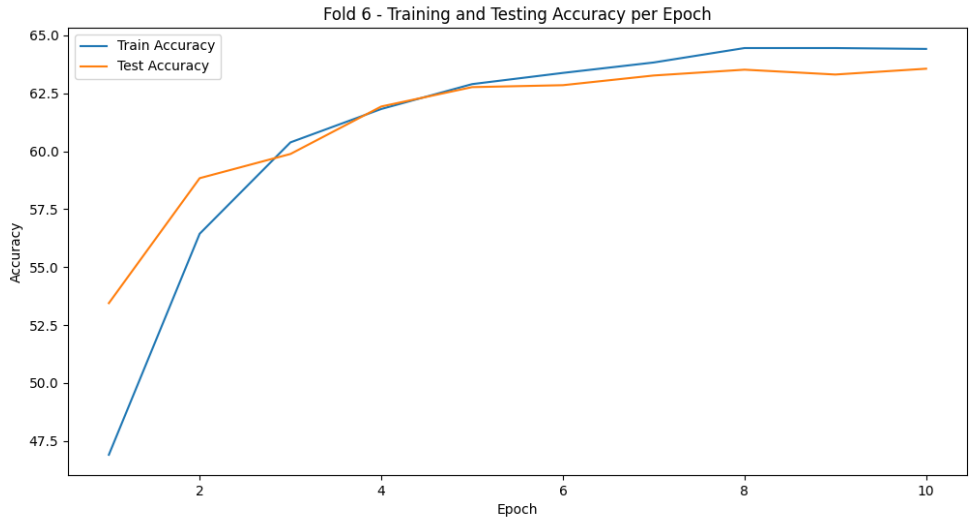

**Διάγραμμα 15: Γράφημα ακρίβειας Fold 3**



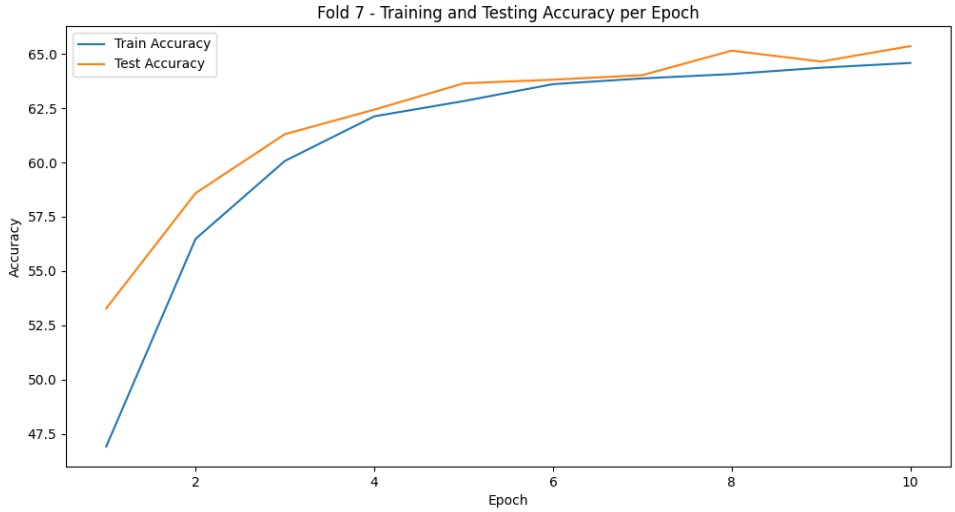
**Διάγραμμα 16: Γράφημα ακρίβειας Fold 4**



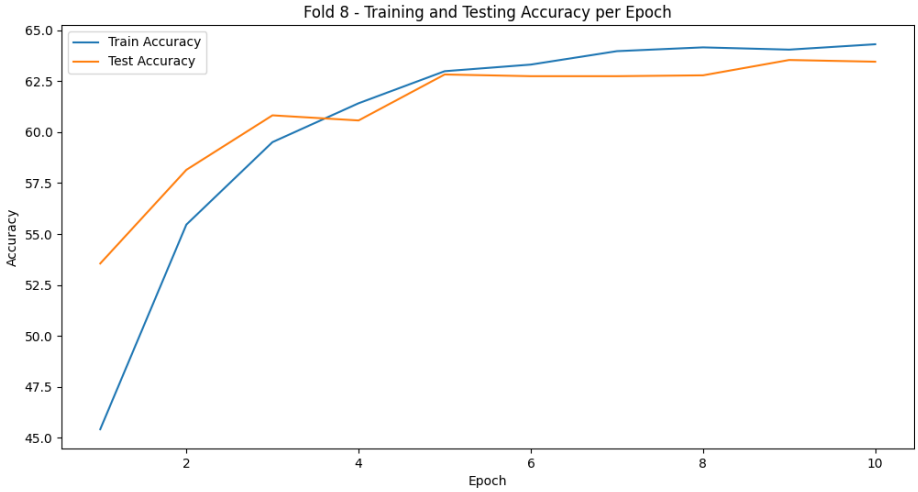

**Διάγραμμα 17: Γράφημα ακρίβειας Fold 5**



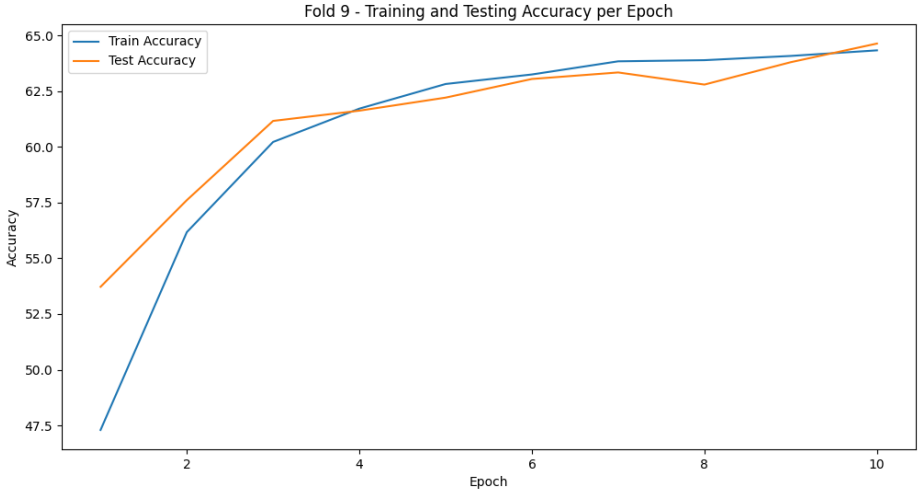

**Διάγραμμα 18: Γράφημα ακρίβειας Fold 6**



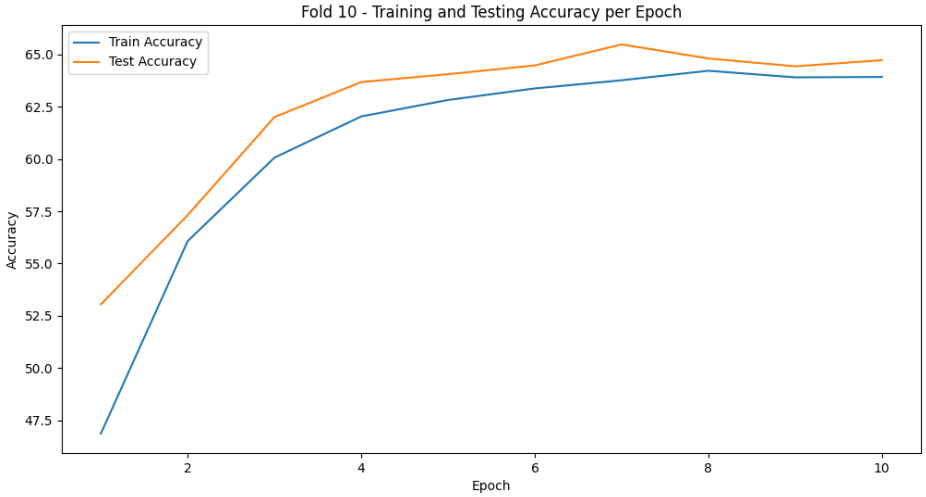

**Διάγραμμα 19: Γράφημα ακρίβειας Fold 7**




**Διάγραμμα 20: Γράφημα ακρίβειας Fold 8**

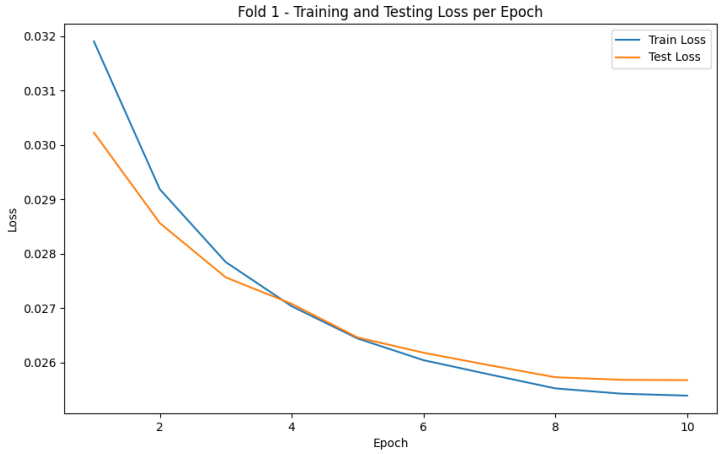



**Διάγραμμα 21: Γράφημα ακρίβειας Fold 9**

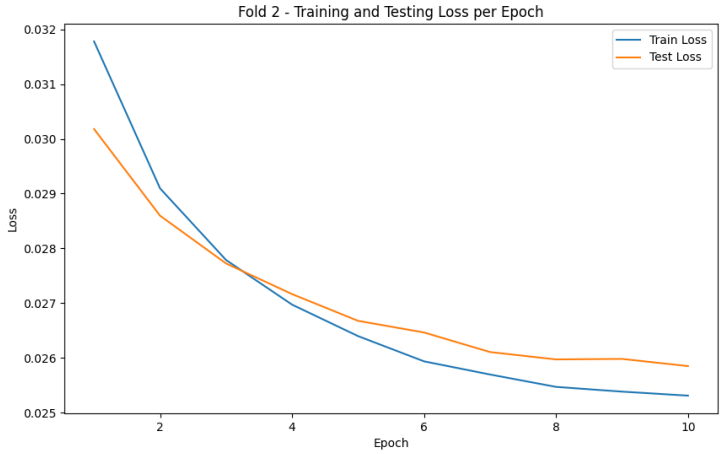



**Διάγραμμα 22: Γράφημα ακρίβειας Fold 10**

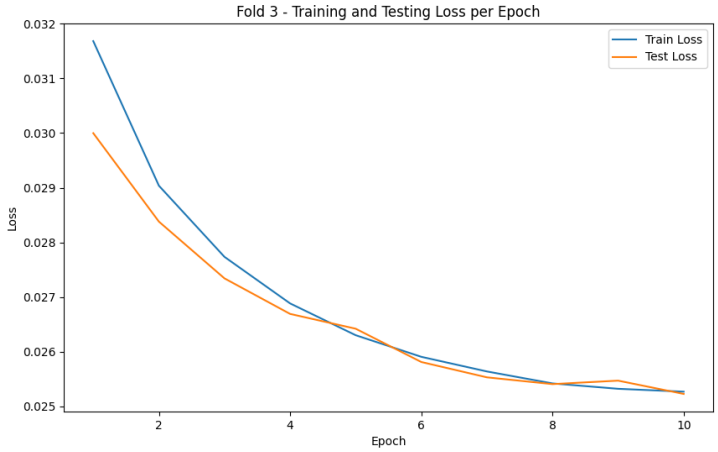
### 5.9.6 Διαγράμματα Απώλειας



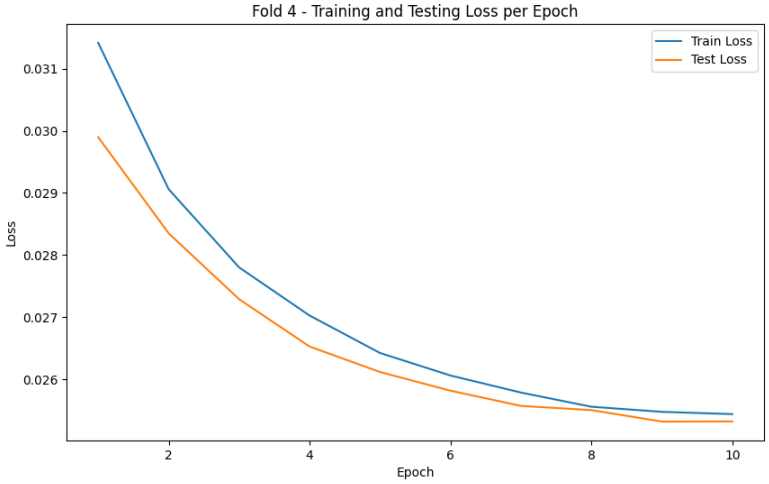

**Διάγραμμα 23: Γράφημα απώλειας Fold 1**



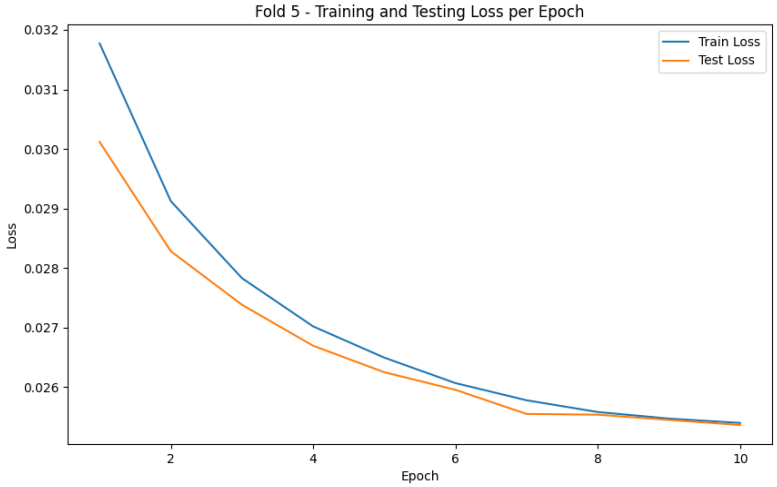

**Διάγραμμα 24: Γράφημα απώλειας Fold 2**



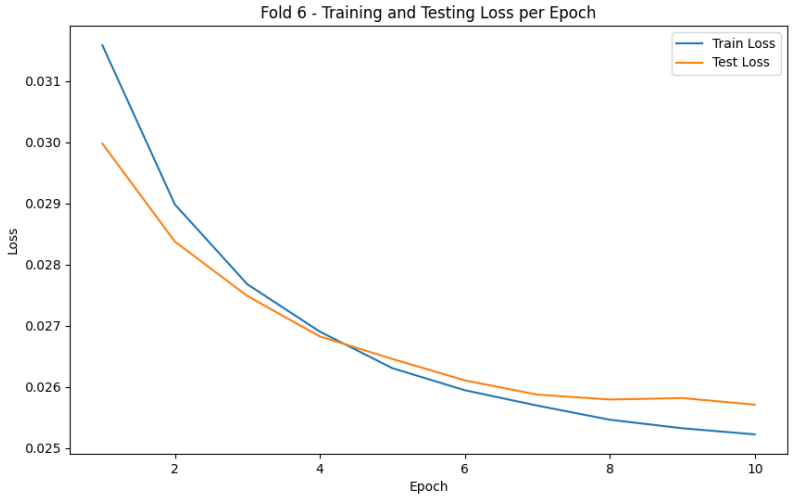

**Διάγραμμα 25: Γράφημα απώλειας Fold 3**



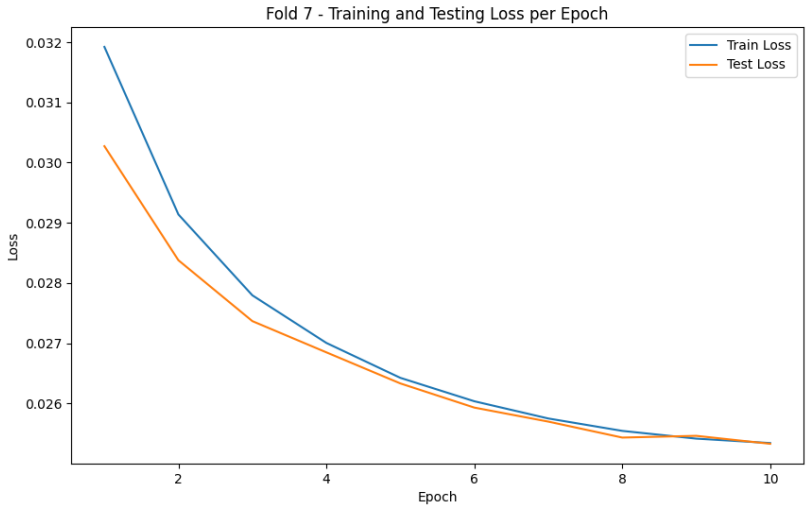

**Διάγραμμα 26: Γράφημα απώλειας Fold 4**



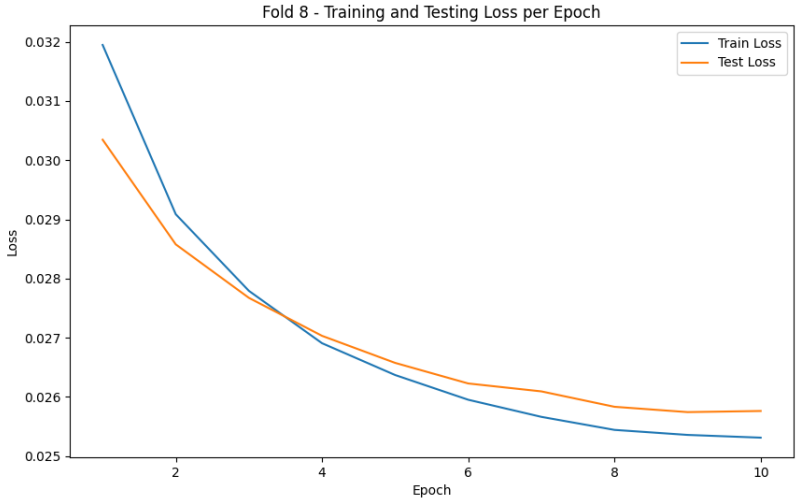

**Διάγραμμα 27: Γράφημα απώλειας Fold 5**



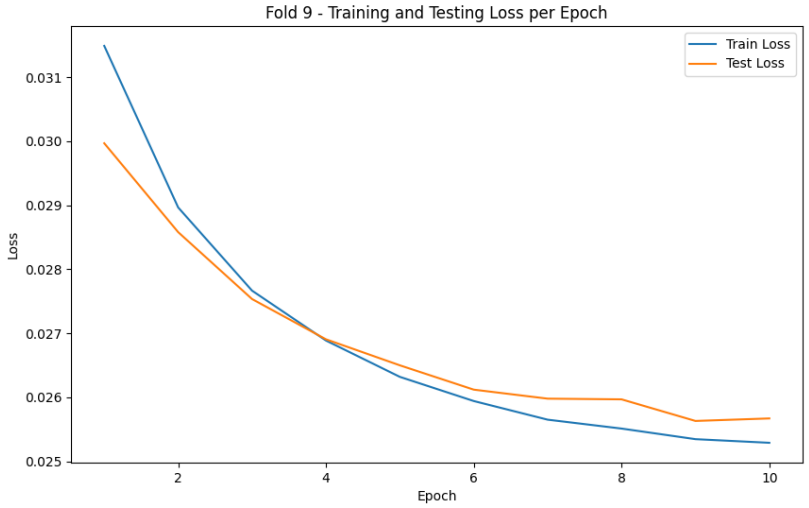

**Διάγραμμα 28: Γράφημα απώλειας Fold 6**



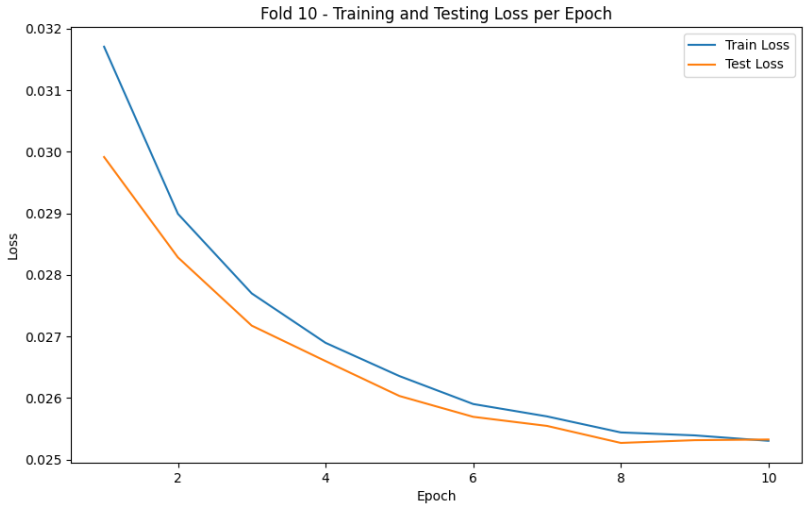

**Διάγραμμα 29: Γράφημα απώλειας Fold 7**




**Διάγραμμα 30: Γράφημα απώλειας Fold 8**




**Διάγραμμα 31: Γράφημα απώλειας Fold 9**




**Διάγραμμα 32: Γράφημα απώλειας Fold 10**

# Κεφάλαιο 6: Συμπεράσματα και Μελλοντικές Κατευθύνσεις

## 6.1 Συμπεράσματα

Συνολικά, η εκπαίδευση και αξιολόγηση του μοντέλου BERT για την ανάλυση συναισθημάτων στην ελληνική γλώσσα, χρησιμοποιώντας μια προσέγγιση 10-fold Cross Validation, παρέχει ενθαρρυντικά αποτελέσματα. Οι μετρικές Precision, Recall, F1-Score, και Ακρίβεια καταδεικνύουν σταθερή απόδοση του μοντέλου σε διάφορα υποσύνολα δεδομένων, ενώ οι μέσες τιμές τους (περίπου 65,06%, 65,27%, 65,13%, και 64,00%, αντίστοιχα) αναδεικνύουν μια γενικευμένη ικανότητα του μοντέλου να διατηρεί ισορροπία μεταξύ ακρίβειας και ανάκλησης.

Η διαδικασία K-Fold Cross Validation επιτρέπει την εκτίμηση της απόδοσης του μοντέλου σε διάφορες συνθήκες, βοηθώντας στη διασφάλιση της γενικευτικότητάς του. Ενώ το μοντέλο εκπαιδεύεται στα δεδομένα Twitter σε περιβάλλοντα 10-fold, οι παραμετρικές μετρικές παραμένουν σταθερές, αποδεικνύοντας την αξιοπιστία και την αποτελεσματικότητα του BERT μοντέλου στην ανάλυση συναισθημάτων στην ελληνική γλώσσα. Η ικανότητα του μοντέλου να διατηρεί σταθερή απόδοση διασφαλίζει την αξιοπιστία των προβλέψεών του, ενισχύοντας τη χρησιμότητα του στον τομέα της ανάλυσης συναισθημάτων.

## 6.2 Μελλοντικές Κατευθύνσεις

Σχετικά με τις μελλοντικές κατευθύνσεις της έρευνάς μου, εξετάζονται πολλαπλές προοπτικές για τη βελτίωση και την εξέλιξη του μοντέλου. Μια προσέγγιση θα μπορούσε να είναι η επέκταση του δείγματος δεδομένων, περιλαμβάνοντας περισσότερα και πιο ποικίλα δείγματα για να βελτιωθεί η γενικοποίηση του μοντέλου.

Επιπλέον, εξετάζεται η βελτίωση της επεξεργασίας του κειμένου, ερευνώντας βελτιωμένες μεθόδους επεξεργασίας ή προηγμένες τεχνικές προεπεξεργασίας για την αύξηση της απόδοσης του μοντέλου.

Ενδιαφέρον παρουσιάζει και η ενσωμάτωση νέων χαρακτηριστικών, σκεπτόμενοι τον τρόπο ενσωμάτωσης νέων πληροφοριών ή χαρακτηριστικών στο μοντέλο, όπως η χρήση εξωτερικών πηγών δεδομένων.

Παράλληλα, προτείνεται η προσαρμογή των υπερπαραμέτρων του μοντέλου, όπως το ρυθμός μάθησης, για την εξαγωγή βέλτιστων αποτελεσμάτων.

Σε μια διαφορετική προοπτική, εξετάζεται η αναζήτηση άλλων αρχιτεκτονικών μοντέλων πέρα από το BERT, με στόχο τον εντοπισμό πιο κατάλληλου μοντέλου για την ανάλυση συναισθημάτων στην ελληνική γλώσσα.

Τέλος, υποστηρίζεται η εξέταση νέων τομέων στην έρευνα, όπως η επέκταση σε άλλα συναισθηματικά κείμενα πέραν του Twitter, προκειμένου να ενισχυθεί η εφαρμοστική αξία του μοντέλου.

Συνολικά, αυτές οι προτάσεις θα μπορούσαν σημαντικά να συνεισφέρουν στη βελτίωση και την εξέλιξη της έρευνας στον τομέα της ανάλυσης συναισθημάτων για την ελληνική γλώσσα.

# Πίνακας Ορολογίας

| 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** |
| **ενσωμάτωση 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** | pre-processing |
| **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

| | |
|---|---|
| ΕΦΓ | Επεξεργασία Φυσικής Γλώσσας |
| κ.ά | και άλλα |
| κ.λ.π | και λοιπά |
| BERT | Bidirectional Encoder Representations from Transformers |
| biLSTM | bidirectional long short-term memory |
| BoW | Bag-of-Words |
| CBOW | Continuous Bag-of-Words |
| CNN | Convolutional Neural Networks |
| CRF | Conditional Random Fields |
| DNN | Deep Neural Networks |
| ELMo | Embedding from Language Models |
| GloVe | Global Vectors of Word Representation |
| GPT | Generative Pre-Training Transformer |
| HMM | Hidden Markov Models |
| IDF | Inverse Document Frequency |
| MLE | Maximum Likelihood Estimation |
| MLP | Multilayer Perceptron |
| MLM | Masked Language Modeling |
| NER | Named-Entity Recognition |
| NLP | Natural Language Processing |
| NLU | Natural Language Understanding |
| NSP | Next Sentence Prediction |
| PMI | Pointwise Mutual Information |
| PoS | Part-of-Speech |
| Q&A | Question & Answering |
| RBF | Radial Basis Function |
| RNN | Recurring Neural Network |
| RoBERTa | Robustly Optimized BERT |
| SNN | Spiking Neural Network |
| SVM | Support Vector Machine |
| TF | Term Frequency |
| URL | Uniform Resource Locator |
| VSM | Vector Space Model |

# Βιβλιογραφία

[1] Hoonlor, A., Mohammed, J., Zaki, M., & Wallace, W. Sequential patterns and temporal patterns for text mining.

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

[3] Bhardwaj, A. (2020). Natural language processing: The method behind understanding humans and machines. [Online; accessed October 1, 2023].

[4] 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.

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

[6] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pp. 1530–1545.

[7] Plutchik, R. (2001). 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), pp. 342–350.

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

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

[10] Lombard, M., Reich, R., Grabe, M., Bracken, C., & Bolmarcich, T. (2000). Presence and television. Human Communication Research, 26, pp.75–98.

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

[12] Lottridge, D., Chignell, M., & Jovicic, A. (2011). Affective Interaction Understanding, Evaluating, and Designing for Human Emotion, 7, pp. 197–217.

[13] Turney, P. D. (2002). “Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews”. In Proceedings of the 40th annual meeting on the association for computational linguistics, pp. 417–424.

[14] Pang, B., Lee, L., & Vaithyanathan, S. (2002). “Thumbs up? Sentiment classification using Machine learning Techniques”. In Proceedings of the conference on empirical methods in natural language processing, pp. 79–86.

[15] Rustamov, S., Mustafayev, E., Clements, M. A. (2013). “Sentiment analysis using Neuro-Fuzzy and Hidden Markov models of text”. In Proceedings of IEEE 2013, pp. 1-6.

[16] Nakagawa, T., Inui, K., & Kurohashi, S. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In Proceedings of the 2010 annual conference of the North American chapter of the association for computational linguistics, pp. 786–794.

[17] Simon Haykin (2009). “Neural Networks and Learning Machines”. 3rd edition.

[18] Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges, and trends. Knowledge-Based Systems, Volume 226, pp. 107-134.

[19] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE, 12, Volume 86, pp. 2278 - 2324.

[20] Kim, S.-M., & Hovy, E. (2004). Determining the Sentiment of Opinions. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 1367–es.

[21] Wankhade, M., Rao, A., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. Artificial Intelligence Review, 55(7), pp. 1-50.

[22] Tsang, S.-H. (2022). Review — elmo: Deep contextualized word representations. [Online; accessed October 1, 2023].

[23] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. CoRR, abs/1706.03762.

[24] (Unknown author). (2022). [Online; accessed October 1, 2023]. Available at https://aclanthology.org/2022.wassa-1.25.pdf.

[25] Alammar, J. (2018). The illustrated transformer. [Online; accessed October 1, 2023].

[26] Radford, A., & Narasimhan, K. (2018). Improving language understanding by generative pre-training.

[27] (Unknown author). (2018). Improving language understanding by generative pre-training. [Online; accessed October 1, 2023].

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

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

[30] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

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

[32] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. K¨opf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: Βιβλιογραφία 85 An imperative style, high-performance deep learning library. CoRR, abs/1912.01703, 2019.

[33] 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. Man´e, 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. Vi´egas, 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.

[34] Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). Greek-bert: The Greeks visiting Sesame Street. In 11th Hellenic Conference on Artificial Intelligence, SETN 2020, page 110–117.