ΠΟΛΥΤΕΧΝΕΙΟ
ΚΡΗΤΗΣ

Σχολή Μηχανικών
Παραγωγής και Διοίκησης

Διπλωματική Εργασία

ΣΥΓΚΡΙΣΗ ΑΛΓΟΡΙΘΜΩΝ ΕΠΕΞΕΡΓΑΣΙΑΣ ΦΥΣΙΚΗΣ ΓΛΩΣΣΑΣ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΙΚΑΝΟΠΟΙΗΣΗΣ ΠΕΛΑΤΩΝ

ΧΑΡΑΛΑΜΠΟΣ ΤΣΑΠΑΚΟΣ

Χανιά, June 2025

Η παρούσα διπλωματική εργασία
αφιερώνεται στους παππούδες και
τις γιαγιάδες μου.

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον κ. Στέλιο Τσαφαράκη, Αναπληρωτή Καθηγητή της Σχολής Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης, για την ανάθεση του θέματος και την εποπτεία του καθ' όλη τη διάρκεια της εργασίας.

Ιδιαίτερες ευχαριστίες οφείλω στον κ. Αναστάσιο Κυριακίδη, Υποψήφιο Διδάκτορα της Σχολής, για την ουσιαστική καθοδήγηση, την τεχνική υποστήριξη και τη διαθεσιμότητά του σε κάθε στάδιο της διαδικασίας. Η συμβολή του υπήρξε καθοριστική για την ολοκλήρωση της διπλωματικής μου.

Τέλος, ευχαριστώ θερμά την οικογένειά μου, και τους φίλους μου για τη συνεχή στήριξη.

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται την ανάλυση της ικανοποίησης πελατών μέσα από δεδομένα σχολίων/κριτικών που προέρχονται από διαδικτυακές πηγές. Βασίζεται σε τεχνικές επεξεργασίας φυσικής γλώσσας (Natural Language Processing – NLP), με στόχο την εξαγωγή χρήσιμων συμπερασμάτων για το πώς αξιολογούν οι χρήστες ένα προϊόν ή μία εμπειρία. Τα εμπειρικά δεδομένα περιλαμβάνουν σχόλια χρηστών για ένα συγκεκριμένο προϊόν ή σύνολο προϊόντων.

Ο σκοπός της εργασίας είναι διπλός: αφενός να αξιολογηθεί η απόδοση διαφορετικών αλγορίθμων στην εξαγωγή θεμάτων, συναισθημάτων και ταξινόμησης σχολίων, και αφετέρου να αποδειχθεί η χρησιμότητα των μεθόδων NLP για επιχειρηματική αξιολόγηση της ικανοποίησης πελατών.

Η μέθοδος που θα εφαρμοστεί περιλαμβάνει τη χρήση του MATLAB και συγκεκριμένα του Text Analytics Toolbox για εφαρμογή βασικών τεχνικών NLP: sentiment analysis, topic modeling και text classification. Επιπλέον, για λόγους σύγκρισης και αξιολόγησης, θα χρησιμοποιηθούν σύγχρονα προχωρημένα μοντέλα όπως ο BERT και το GPT-4.

Η επιλογή αυτών των μεθόδων δικαιολογείται από τη σημασία που αποκτούν οι τεχνικές NLP στην εποχή της πληθώρας δεδομένων κειμένου, αλλά και από την ανάγκη για επιστημονική αξιολόγηση των εργαλείων που παρέχει το MATLAB έναντι πιο πρόσφατων λύσεων, όπως τα μεγάλα γλωσσικά μοντέλα (LLMs). Η συγκριτική ανάλυση ενισχύει τη δυνατότητα επιλογής κατάλληλων μεθόδων σε μελλοντικές εφαρμογές.

Λέξεις κλειδιά: επεξεργασία φυσικής γλώσσας, ανάλυση συναισθήματος, θεματική μοντελοποίηση, ταξινόμηση κειμένου

Abstract

This thesis explores the application of Natural Language Processing (NLP) techniques for analyzing customer satisfaction through online review data. The study begins by presenting the theoretical background and capabilities of the MATLAB Text Analytics Toolbox, followed by the implementation of core NLP tasks such as sentiment analysis, topic modeling, and text classification. To assess the effectiveness of different approaches, a comparative analysis is conducted between traditional methods available in the MATLAB environment and state-of-the-art language models, including GPT-4 and BERT.

The research aims to evaluate the accuracy, interpretability, and practical applicability of each technique by extracting insights from customer feedback. By leveraging real-world data, this work highlights how advanced NLP can be integrated into consumer analytics, contributing to more informed business decisions and enhancing user experience evaluation. The findings are expected to inform future implementations of text analysis tools in both academic research and industry practices.

Key words: natural language processing, sentiment analysis, topic modeling, text classification

Περιεχόμενα

1. Εισαγωγή

1.1. Παρουσίαση Προβλήματος

Η **ανάλυση συναισθήματος** αποτελεί έναν από τους σημαντικότερους κλάδους της επεξεργασίας φυσικής γλώσσας (**Natural Language Processing - NLP**), καθώς επιτρέπει την εξαγωγή συναισθηματικών πληροφοριών από δεδομένα κειμένου. Η συνεχής ανάπτυξη των ψηφιακών πλατφορμών και των μέσων κοινωνικής δικτύωσης έχει οδηγήσει στη συλλογή τεράστιου όγκου δεδομένων, δημιουργώντας την ανάγκη για αυτοματοποιημένες μεθόδους ανάλυσης, οι οποίες θα μπορούν να προσδιορίζουν αν το περιεχόμενο ενός κειμένου εκφράζει θετικό, αρνητικό ή ουδέτερο συναίσθημα. Οι σύγχρονες τεχνικές ανάλυσης συναισθήματος βασίζονται είτε σε λεξικογραφικές μεθόδους είτε σε μοντέλα μηχανικής μάθησης, επιτρέποντας την εξαγωγή ποιοτικών πληροφοριών από αδόμητα δεδομένα.

Η παρούσα εργασία εστιάζει στην **ανάλυση συναισθήματος σε κριτικές χρηστών του επιτραπέζιου παιχνιδιού «7 Wonders»**. Οι κριτικές που αφήνουν οι χρήστες στις διάφορες διαδικτυακές πλατφόρμες περιέχουν σημαντικές πληροφορίες για τη συνολική εμπειρία τους με το παιχνίδι, ωστόσο η επεξεργασία τους με μη αυτόματο τρόπο είναι ανέφικτη λόγω του μεγάλου όγκου δεδομένων. Μέσω αυτοματοποιημένων τεχνικών επεξεργασίας φυσικής γλώσσας, η εργασία αυτή επιδιώκει να καθαρίσει, να αναλύσει και να κατηγοριοποιήσει τα σχόλια των χρηστών, ώστε να εξαχθούν πολύτιμες πληροφορίες σχετικά με τα χαρακτηριστικά του παιχνιδιού που απολαμβάνουν ή προβληματίζουν τους παίκτες.

1.2. Ιστορική Αναδρομή στην Ανάλυση Συναισθήματος και τις Τεχνικές NLP

Η ανάλυση συναισθήματος (Sentiment Analysis), ως πεδίο μελέτης της επεξεργασίας φυσικής γλώσσας (Natural Language Processing - NLP), έχει εξελιχθεί ραγδαία τις τελευταίες δύο δεκαετίες. Η ανάγκη κατανόησης των συναισθημάτων που εκφράζονται σε κείμενο αποτέλεσε βασικό κίνητρο για την ανάπτυξη τεχνικών που να μπορούν να επεξεργάζονται τη γλώσσα υπολογιστικά. Η σημερινή πρόοδος σε αυτόν τον τομέα βασίζεται σε ένα ισχυρό υπόβαθρο από έρευνες, εργαλεία και τεχνολογικές καινοτομίες, που αξίζει να αναλυθούν.

Πρώιμα στάδια και λεξικογραφικές προσεγγίσεις

Οι πρώτες συστηματικές απόπειρες για την ταξινόμηση συναισθήματος σε κείμενο εμφανίζονται στις αρχές της δεκαετίας του 2000. Μια από τις πρώτες σημαντικές δημοσιεύσεις ήταν των **Bo Pang, Lillian Lee και Shivakumar Vaithyanathan (2002)**, οι οποίοι πρότειναν τη χρήση μηχανικής μάθησης για την ταξινόμηση κινηματογραφικών κριτικών ως θετικές ή αρνητικές [1]. Παράλληλα, αναπτύσσονται

λεξικογραφικές μέθοδοι όπως το **SentiWordNet**, οι οποίες αντιστοιχούν λέξεις σε πολικότητα (θετική, αρνητική, ουδέτερη), προσφέροντας ένα πρώτο εργαλείο ανάλυσης χωρίς ανάγκη εκπαίδευσης.

Ταξινομητές και εξαγωγή χαρακτηριστικών

Κατά τη δεκαετία του 2000, η εφαρμογή κλασικών αλγορίθμων ταξινόμησης, όπως **Naive Bayes**, **Support Vector Machines (SVM)** και **k-Nearest Neighbors (KNN)**, κυριαρχεί στην έρευνα. Αυτοί οι αλγόριθμοι χρησιμοποιούν **σαφώς ορισμένα χαρακτηριστικά** (features), όπως **Bag-of-Words (BoW)** και **TF-IDF**, για την αναπαράσταση κειμένου σε αριθμητική μορφή [2]. Το 2001, ο **Leo Breiman** εισήγαγε τα **Random Forests**, τα οποία βελτίωσαν την αξιοπιστία σε προβλήματα ταξινόμησης [3].

Μείωση διαστατικότητας και οπτικοποίηση δεδομένων

Παράλληλα με την ανάπτυξη μοντέλων, αναπτύσσονται τεχνικές για τη μείωση της διαστατικότητας των δεδομένων, όπως η **Ανάλυση Κύριων Συνιστωσών (PCA)** [4]. Η ανάγκη οπτικοποίησης των embeddings υψηλής διάστασης οδήγησε στην ανάπτυξη της **t-SNE** από τους **van der Maaten και Hinton (2008)**, μια τεχνική που επιτρέπει την οπτικοποίηση της δομής των δεδομένων σε δύο ή τρεις διαστάσεις [5].

Αναπαραστάσεις λέξεων και η εποχή του deep learning

Η έλευση του **Word2Vec** από τον **Mikolov et al. (2013)** αποτέλεσε τομή, καθώς μετέφερε την αναπαράσταση των λέξεων από σπάνιους διανύσματα (sparse vectors) σε συνεχείς, σημασιολογικά πλούσιες αναπαραστάσεις (dense embeddings) [6]. Ακολούθησαν τα **GloVe**, **FastText** και άλλες μορφές embeddings, οι οποίες ενίσχυσαν τη δυνατότητα κατανόησης των συμφραζόμενων.

Το 2017, οι **Vaswani et al.** παρουσίασαν τον **Transformer**, έναν αρχιτεκτονικό σχεδιασμό βασισμένο σε μηχανισμό **Self-Attention**, ο οποίος επέτρεψε τη μαζική παραλληλοποίηση και την αποτελεσματική επεξεργασία μεγάλων ακολουθιών [7]. Αυτό οδήγησε στη δημιουργία του **BERT (Bidirectional Encoder Representations from Transformers)** το 2018 από την Google [8], που αποτέλεσε ορόσημο στην επεξεργασία φυσικής γλώσσας, καθώς αξιοποιεί προεκπαίδευση και μετεκπαίδευση (fine-tuning) σε downstream[1] εργασίες, μεταξύ των οποίων και η ανάλυση συναισθήματος.

[1] Η μεταγενέστερη εργασία (downstream task) εξαρτάται από την έξοδο μιας προηγούμενης εργασίας ή διαδικασίας. Περιλαμβάνει την εφαρμογή της γνώσης του προεκπαιδευμένου μοντέλου σε ένα νέο πρόβλημα. Η έξοδος της προηγούμενης εργασίας χρησιμεύει ως είσοδος στην μεταγενέστερη εργασία και το μοντέλο μπορεί να εκτελέσει την μεταγενέστερη εργασία μόνο αφού ολοκληρώσει την προηγούμενη.

Η ανάγκη για γρήγορα και «ελαφριά» μοντέλα οδήγησε στην ανάπτυξη παραλλαγών όπως τα **Tiny, Mini και Small BERT**, με λιγότερα στρώματα και κεφαλές προσοχής (attention headers), κατάλληλα για συστήματα χαμηλών πόρων. Αυτά τα μοντέλα εκπαιδεύονται συχνά με **Knowledge Distillation** [9], όπου ένα μεγαλύτερο μοντέλο (teacher) καθοδηγεί την εκπαίδευση ενός μικρότερου (student).

VADER και λεξικά συναισθήματος

Ταυτόχρονα, αναπτύσσονται εργαλεία όπως το **VADER (Hutto & Gilbert, 2014)**, τα οποία χρησιμοποιούν προκαθορισμένα λεξικά και απλούς κανόνες για να αναλύσουν συναισθηματικά την πολικότητα ενός σύντομου κειμένου [10]. Λόγω της ταχύτητάς του και της ακρίβειάς του σε μικρά κείμενα, χρησιμοποιείται ευρέως σε πλατφόρμες κοινωνικής δικτύωσης.

Εποχή των Μεγάλων Γλωσσικών Μοντέλων (LLMs)

Η είσοδος των **GPT-2, GPT-3** και μετέπειτα του **GPT-4 και GPT-4 Turbo** από την **OpenAI**, σηματοδότησε μια νέα εποχή όπου **γενικά προεκπαιδευμένα γλωσσικά μοντέλα** μπορούν να επιτελέσουν πολλαπλές εργασίες χωρίς ρητό fine-tuning, μεταξύ των οποίων και η ανάλυση συναισθήματος [11].

1.3. Σημασία της ανάλυσης συναισθήματος σε κριτικές χρηστών

Η ανάγκη για ανάλυση συναισθήματος σε κριτικές χρηστών προκύπτει από το γεγονός ότι οι αριθμητικές βαθμολογίες που συνήθως συνοδεύουν τις κριτικές παρέχουν μόνο μια γενική εκτίμηση της εμπειρίας των παικτών, χωρίς να αναδεικνύουν τις συγκεκριμένες πτυχές που εκτιμήθηκαν θετικά ή αρνητικά. Οι παραδοσιακές μέθοδοι ανάλυσης δεδομένων, που βασίζονται στην απλή ποσοτική αξιολόγηση, δεν είναι επαρκείς για την εξαγωγή ουσιαστικών συμπερασμάτων, καθώς αγνοούν το περιεχόμενο των κριτικών και την υποκειμενικότητα της γλώσσας των χρηστών. Αντιθέτως, η εφαρμογή τεχνικών ανάλυσης συναισθήματος επιτρέπει την κατανόηση της γνώμης των χρηστών με μεγαλύτερη ακρίβεια, επιτρέποντας τον εντοπισμό συγκεκριμένων θεμάτων και συναισθηματικών τάσεων.

1.4. Εφαρμογές και παραδείγματα χρήσης

Η ανάλυση συναισθήματος έχει αναδειχθεί ως ένα από τα πλέον χρήσιμα εργαλεία στην επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP), καθώς βρίσκει εφαρμογή σε ένα ευρύ φάσμα πεδίων, από το **ηλεκτρονικό εμπόριο** και τα **μέσα κοινωνικής δικτύωσης** έως τη **βιομηχανία των βιντεοπαιχνιδιών και των επιτραπέζιων παιχνιδιών**. Οι σύγχρονες τεχνικές ανάλυσης συναισθήματος επιτρέπουν στις επιχειρήσεις να αξιολογούν την ικανοποίηση των καταναλωτών, στις

διαδικτυακές πλατφόρμες να προσφέρουν πιο στοχευμένες προτάσεις στους χρήστες τους και στους ερευνητές να κατανοούν τη δυναμική της κοινής γνώμης σε πραγματικό χρόνο.

Στον τομέα του **ηλεκτρονικού εμπορίου**, πλατφόρμες όπως το **Amazon και το eBay** χρησιμοποιούν τεχνικές NLP για την **αυτόματη κατηγοριοποίηση των κριτικών προϊόντων**, επιτρέποντας στους καταναλωτές να λαμβάνουν καλύτερα ενημερωμένες αγοραστικές αποφάσεις. Παράλληλα, οι επιχειρήσεις αξιοποιούν την ανάλυση σχολίων για να κατανοήσουν ποια χαρακτηριστικά των προϊόντων τους είναι πιο δημοφιλή ή ποια προκαλούν δυσαρέσκεια στους χρήστες. Στον χώρο της **ψυχαγωγίας**, η ανάλυση σχολίων επιτραπέζιων και ψηφιακών παιχνιδιών επιτρέπει στους δημιουργούς να αναγνωρίζουν **πλεονεκτήματα και αδυναμίες** ενός παιχνιδιού, οδηγώντας σε βελτιώσεις που ενισχύουν τη συνολική εμπειρία των χρηστών.

Οι πιο σύγχρονες μέθοδοι ανάλυσης συναισθήματος αξιοποιούν **προηγμένες τεχνικές NLP**, όπως το **Latent Dirichlet Allocation (LDA), τα word embeddings, το BERT και το VADER**, προσφέροντας μεγαλύτερη ακρίβεια στην ανάλυση συναισθήματος σε μεγάλες κλίμακες δεδομένων. Για παράδειγμα, σε μια **σύγκριση του VADER και του BERT** σε ανάλυση συναισθήματος σε tweets για την πανδημία COVID-19, τα αποτελέσματα έδειξαν ότι το BERT βελτίωσε την απόδοση της ταξινόμησης, με τον αλγόριθμο **Support Vector Machine (SVM)** να επιτυγχάνει **ακρίβεια 92%** στο σύνολο δεδομένων Omicron (ResearchGate).

Μια άλλη μελέτη παρουσίασε το **Opinion-BERT**, ένα υβριδικό μοντέλο BERT ενισχυμένο με γνώμες, το οποίο έχει σχεδιαστεί για πολυ-εργασιακή μάθηση, επιτυγχάνοντας **ταυτόχρονη κατηγοριοποίηση συναισθήματος και κατάστασης** (Nature). Ο συνδυασμός **LDA και ανάλυσης συναισθήματος** έχει επίσης μελετηθεί σε πλαίσια όπως η ανάλυση σχολίων για το **ChatGPT**, όπου χρησιμοποιήθηκε το LDA για την εξαγωγή θεμάτων, ενώ το BERT εφαρμόστηκε για την κατηγοριοποίηση συναισθημάτων με υψηλή ακρίβεια (MDPI).

Στο πεδίο της οικονομίας, η χρήση του **BERTopic** έχει διερευνηθεί για την **πρόβλεψη χρηματιστηριακών τιμών**, ενσωματώνοντας ανάλυση συναισθήματος σε δεδομένα χρηματιστηριακής αγοράς, συνδυάζοντας διαφορετικά μοντέλα βαθιάς μάθησης και παρέχοντας πιο ακριβείς προβλέψεις (arXiv).

Αυτές οι μελέτες αναδεικνύουν τη σημασία της ενσωμάτωσης προηγμένων τεχνικών NLP, όπως το **LDA, τα embeddings, το BERT και το VADER**, στην ανάλυση συναισθήματος. Η ικανότητα αυτών των τεχνικών να εντοπίζουν τάσεις και να παρέχουν ακριβείς συναισθηματικές κατηγοριοποιήσεις έχει αποδειχθεί πολύτιμη σε τομείς όπως η **επιχειρηματική ευφυΐα, η κοινωνική ανάλυση και η στρατηγική βελτίωση προϊόντων**.

1.5. Στόχοι της διπλωματικής

Η παρούσα εργασία στοχεύει στην ανάλυση κριτικών του επιτραπέζιου παιχνιδιού **"7 Wonders"** με τη χρήση **τεχνικών NLP και μηχανικής μάθησης**. Οι βασικοί στόχοι περιλαμβάνουν:

- **Την επεξεργασία και τον καθαρισμό των δεδομένων**, εφαρμόζοντας τεχνικές προεπεξεργασίας όπως η **αφαίρεση stopwords, το lemmatization, η αναγνώριση οντοτήτων και η εξαγωγή βασικών θεμάτων μέσω topic modeling (LDA, LSA)**. Στο πλαίσιο της εργασίας, διερευνήθηκε επίσης η **βελτιστοποίηση του LDA**, εξετάζοντας διαφορετικές παραμέτρους του μοντέλου, όπως ο αριθμός των θεμάτων, με στόχο τη βελτίωση της συνοχής των θεμάτων που προέκυψαν από τις κριτικές των χρηστών.

- **Την ανάλυση του συναισθήματος στις κριτικές**, με στόχο την κατηγοριοποίησή τους ως **θετικές, αρνητικές ή ουδέτερες**, χρησιμοποιώντας τόσο **λεξικογραφικές μεθόδους (VADER)** όσο και **μοντέλα μηχανικής μάθησης (BERT, GPT-based models)**.

- **Την εξαγωγή πληροφοριών σχετικά με τα κύρια χαρακτηριστικά που επηρεάζουν την αντίληψη των παικτών για το παιχνίδι**, επιτρέποντας την κατανόηση των παραγόντων που συμβάλλουν στη συνολική εμπειρία των χρηστών.

- **Τη σύγκριση διαφορετικών προσεγγίσεων ανάλυσης συναισθήματος**, αξιολογώντας την ακρίβεια και την απόδοση των μοντέλων.

- **Την οπτικοποίηση των αποτελεσμάτων**, ώστε να αποτυπωθούν οι συχνότερες θεματικές ενότητες που αναφέρονται στις κριτικές μέσω **word clouds και γραφημάτων κατανομής συναισθήματος**.

1.6. Δομή της εργασίας

Η εργασία αποτελείται από **πέντε κεφάλαια**. Στην **Εισαγωγή**, περιγράφεται το πρόβλημα της ανάλυσης συναισθήματος σε κριτικές χρηστών και παρουσιάζεται η σημασία της στη βιομηχανία των επιτραπέζιων παιχνιδιών. Στο **Κεφάλαιο 2 - Θεωρητικό Υπόβαθρο**, αναλύονται οι βασικές τεχνικές NLP που χρησιμοποιούνται για την προεπεξεργασία κειμένου και την ανάλυση συναισθήματος. Το **Κεφάλαιο 3 - Μεθοδολογία** επικεντρώνεται στη **μεθοδολογία**, περιγράφοντας τη διαδικασία συλλογής δεδομένων, τις τεχνικές καθαρισμού των κριτικών και τις μεθόδους ανάλυσης συναισθήματος. Επίσης, παρουσιάζονται τα **αποτελέσματα της έρευνας**, με συγκρίσεις μεταξύ των διαφορετικών μεθόδων, οπτικοποιήσεις των δεδομένων και στατιστικές αναλύσεις. Τέλος, στο **Κεφάλαιο 4 - Συμπεράσματα** συνοψίζονται τα βασικά ευρήματα της μελέτης και προτείνονται

κατευθύνσεις για μελλοντική έρευνα, όπως η ενσωμάτωση νεότερων τεχνολογιών deep learning και η εφαρμογή της μεθοδολογίας σε μεγαλύτερα σύνολα δεδομένων.

Η μελέτη αυτή προσφέρει μια **δομημένη και αναλυτική προσέγγιση στην ανάλυση συναισθήματος κριτικών χρηστών**, συνδυάζοντας διαφορετικές μεθόδους NLP και μηχανικής μάθησης. Η αξιολόγηση των αποτελεσμάτων επιτρέπει την **κατανόηση των αντιλήψεων των καταναλωτών**, συμβάλλοντας στη βελτίωση των προϊόντων στον χώρο των επιτραπέζιων παιχνιδιών και προσφέροντας ένα γενικότερο μοντέλο που μπορεί να εφαρμοστεί και σε άλλες μορφές κριτικών χρηστών.

2. Θεωρητικό Υπόβαθρο

2.1. Συναφή Έργα / Ανασκόπηση Βιβλιογραφίας

Η ανάλυση συναισθήματος έχει εξελιχθεί σημαντικά τα τελευταία χρόνια, με την εισαγωγή τεχνικών που συνδυάζουν λεξικογραφικές μεθόδους, μοντέλα μηχανικής μάθησης και βαθιάς μάθησης. Στην παρούσα ενότητα, εξετάζονται προηγούμενες μελέτες και έρευνες που σχετίζονται με το αντικείμενο της διπλωματικής, με έμφαση στις τεχνικές ανάλυσης συναισθήματος, στις μεθόδους προεπεξεργασίας κειμένου, στη θεματολογική μοντελοποίηση (Topic Modeling), στη χρήση embeddings και στην εφαρμογή των νεότερων μοντέλων βαθιάς μάθησης όπως το BERT.

2.1.1. Ανάλυση Συναισθήματος (Sentiment Analysis)

Η ανάλυση συναισθήματος είναι η διαδικασία κατηγοριοποίησης του συναισθήματος που εκφράζεται σε ένα κείμενο σε θετικό, αρνητικό ή ουδέτερο. Οι πρώτες προσεγγίσεις βασίζονταν σε **λεξικογραφικές μεθόδους**, όπου προκαθορισμένες λίστες λέξεων (π.χ. **VADER, AFINN, SentiWordNet**) χρησιμοποιούνται για την ανάλυση της συναισθηματικής διάθεσης ενός κειμένου. Παρότι αυτές οι μέθοδοι είναι γρήγορες και ελαφριές, παρουσιάζουν περιορισμούς καθώς δεν λαμβάνουν υπόψη το συμφραζόμενο (context) (Hutto & Gilbert, 2014) [10].

Με την πρόοδο της **μηχανικής μάθησης**, η ανάλυση συναισθήματος βελτιώθηκε με **μοντέλα επιβλεπόμενης μάθησης** όπως οι **Naive Bayes, Support Vector Machines (SVMs) και Random Forests**, τα οποία εκπαιδεύονται σε ετικετοποιημένα δεδομένα για να ταξινομήσουν συναισθήματα με μεγαλύτερη ακρίβεια (Medhat et al., 2014) [21]. Ωστόσο, η ανάγκη για μοντέλα που μπορούν να κατανοούν τη σημασιολογία του κειμένου οδήγησε στη χρήση **βαθιάς μάθησης**, με την ανάπτυξη μοντέλων **Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) και Transformers**, όπως το **BERT και το GPT** (Devlin et al., 2018) [8].

2.1.2. Τεχνικές Προεπεξεργασίας Κειμένου (Text Preprocessing Techniques)

Η προεπεξεργασία κειμένου αποτελεί κρίσιμο στάδιο στην ανάλυση συναισθήματος, καθώς επηρεάζει άμεσα την ποιότητα των χαρακτηριστικών που εξάγονται για τα μοντέλα μηχανικής μάθησης. Οι πιο κοινές τεχνικές περιλαμβάνουν τον **διαχωρισμό λέξεων (Tokenization)**, την **αφαίρεση stopwords**, την **κανονικοποίηση λέξεων (Lemmatization και Stemming)** και την **αναγνώριση οντοτήτων (Named Entity Recognition - NER)**.

Σύμφωνα με τους Webster & Kit (1992) [36], το tokenization διαφέρει ανάλογα με τη γλώσσα του κειμένου, με τα αγγλικά να έχουν σαφέστερους κανόνες διαχωρισμού σε σύγκριση με άλλες γλώσσες. Ο Kumhar et al. (2023) [37] επισημαίνουν ότι το stemming είναι λιγότερο ακριβές από τη

lemmatization, καθώς συχνά αφαιρεί καταλήξεις χωρίς να λαμβάνει υπόψη τη σημασιολογική πληροφορία.

2.1.3. Θεματολογική Μοντελοποίηση (Topic Modeling) με LDA και BERTopic

Η θεματολογική μοντελοποίηση χρησιμοποιείται για την **εξαγωγή θεμάτων από μη δομημένα δεδομένα κειμένου**. Το **Latent Dirichlet Allocation (LDA)** είναι μια από τις πιο δημοφιλείς τεχνικές, η οποία βασίζεται σε **κατανομή πιθανοτήτων** για την ανάθεση λέξεων σε διαφορετικά θέματα (Blei et al., 2003) [22].

Σύμφωνα με μελέτες, το **BERTopic**, μια νεότερη προσέγγιση που χρησιμοποιεί embeddings και clustering για την αναγνώριση θεμάτων, έχει αποδειχθεί πιο αποτελεσματικό από το LDA σε ορισμένες περιπτώσεις, καθώς αξιοποιεί τη σημασιολογική πληροφορία που περιέχουν τα word embeddings (Grootendorst, 2022) [38].

2.1.4. Word Embeddings και Χαρακτηριστικά Κειμένου

Τα Word Embeddings έχουν αντικαταστήσει τις παραδοσιακές τεχνικές αναπαράστασης κειμένου, όπως το **Bag-of-Words (BoW) και το TF-IDF**, καθώς διατηρούν **σημασιολογικές σχέσεις** μεταξύ των λέξεων. Τεχνικές όπως το **Word2Vec** (Mikolov et al., 2013) [6], **FastText** (Bojanowski et al., 2017) [39] **και BERT embeddings** επιτρέπουν την καλύτερη αναπαράσταση του κειμένου για χρήση σε μοντέλα ταξινόμησης συναισθήματος.

Σύμφωνα με έρευνα που πραγματοποιήθηκε για την ανάλυση σχολίων στο Twitter, η χρήση BERT embeddings σε συνδυασμό με CNNs και LSTMs αύξησε την ακρίβεια της ανάλυσης συναισθήματος κατά **15% σε σχέση με τα παραδοσιακά μοντέλα** (Sun et al., 2019) [40].

2.1.5. Σύγκριση Παραδοσιακών και Σύγχρονων Μοντέλων (VADER, BERT, GPT-based models)

Η χρήση λεξικογραφικών μεθόδων, όπως το **VADER (Valence Aware Dictionary and sEntiment Reasoner)**, είναι ιδανική για σύντομα κείμενα και έχει εφαρμοστεί ευρέως σε αναλύσεις μέσων κοινωνικής δικτύωσης (Hutto & Gilbert, 2014). Ωστόσο, τα τελευταία χρόνια, η **υπολογιστική ισχύς και η πρόσβαση σε μεγάλα σύνολα δεδομένων** έχουν επιτρέψει την ανάπτυξη πιο προηγμένων μοντέλων όπως το **BERT και τα GPT-based μοντέλα**, τα οποία κατανοούν το συναισθηματικό περιεχόμενο του κειμένου με μεγαλύτερη ακρίβεια (Devlin et al., 2018) [8].

Έρευνες έχουν δείξει ότι η χρήση του BERT για ανάλυση συναισθήματος σε συνδυασμό με Transfer Learning αυξάνει την ακρίβεια σε μεγάλα σύνολα δεδομένων έως **92%** (Liu et al., 2019) [41].

2.1.6. Εφαρμογές Ανάλυσης Συναισθήματος σε Επιτραπέζια και Ψηφιακά Παιχνίδια

Η ανάλυση συναισθήματος έχει χρησιμοποιηθεί στη βιομηχανία των **ψηφιακών και επιτραπέζιων παιχνιδιών** για να κατανοηθεί η εμπειρία των παικτών και να βελτιωθούν τα προϊόντα. Για παράδειγμα, έρευνες σε **κριτικές Steam και BoardGameGeek** έχουν δείξει ότι η ανάλυση συναισθήματος μπορεί να αποκαλύψει κρίσιμες πληροφορίες σχετικά με το **gameplay, τη στρατηγική και την ποιότητα των στοιχείων του παιχνιδιού** (Loria et al., 2021) [42].

Συγκεκριμένα, η χρήση του LDA για εξαγωγή θεμάτων από κριτικές χρηστών επιτρέπει την **απομόνωση των κύριων θεμάτων** που επηρεάζουν τη συνολική εμπειρία των παικτών. Παράλληλα, η ενσωμάτωση deep learning μοντέλων, όπως το BERT και τα Transformer-based architectures, βελτιώνει την κατανόηση της σημασιολογίας των σχολίων και ενισχύει την ακρίβεια των αποτελεσμάτων.

2.2. Προεπεξεργασία Κειμένου

Η προεπεξεργασία κειμένου αποτελεί ένα από τα σημαντικότερα στάδια στην Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing - NLP), καθώς συμβάλλει στη μετατροπή του ακατέργαστου κειμένου σε μια δομημένη και αναλύσιμη μορφή. Πρόκειται για ένα απαραίτητο βήμα προτού τα δεδομένα κειμένου χρησιμοποιηθούν σε αλγορίθμους μηχανικής μάθησης ή βαθιάς μάθησης, καθώς το αρχικό κείμενο περιέχει θόρυβο, περιττές πληροφορίες και μη τυποποιημένες εκφράσεις που μπορούν να επηρεάσουν την ακρίβεια της ανάλυσης (Webster & Kit, 1992) [36].

Στην παρούσα εργασία, η προεπεξεργασία πραγματοποιήθηκε μέσω του Text Analytics Toolbox του MATLAB, το οποίο παρέχει ένα σύνολο εργαλείων για την κανονικοποίηση και τον καθαρισμό των δεδομένων. Ο καθαρισμός έγινε στην στήλη comment του πίνακα T που φιλοξενεί τα δεδομένα. Οι τεχνικές που εφαρμόστηκαν, περιλαμβάνουν τον διαχωρισμό των κριτικών σε λέξεις (Tokenization), την κανονικοποίηση των λέξεων (Lemmatization), την αναγνώριση γραμματικών χαρακτηριστικών (POS Tagging), την αφαίρεση ακατάλληλων όρων (Stopword Removal), καθώς και την απομάκρυνση σημείων στίξης, αριθμών και διευθύνσεων URL.

2.2.1. Αφαίρεση URLs

Ένα από τα προκαταρκτικά βήματα που ενσωματώθηκε στον καθαρισμό του κειμένου είναι η **αφαίρεση διευθύνσεων URL** από τα σχόλια των χρηστών.

Οι χρήστες που αφήνουν κριτικές συχνά περιλαμβάνουν συνδέσμους είτε προς εξωτερικές ιστοσελίδες είτε προς άλλα προϊόντα. Αυτές οι πληροφορίες δεν είναι χρήσιμες για την ανάλυση

συναισθήματος ή την εξαγωγή θεμάτων, καθώς οι διευθύνσεις URL δεν περιέχουν συναισθηματική πληροφορία και μπορεί να προσθέσουν θόρυβο στο dataset.

Η αφαίρεση των URL πραγματοποιήθηκε με τη χρήση της εντολής **eraseURLs()** του MATLAB. Η συνάρτηση αυτή σαρώνει το κείμενο και αφαιρεί οποιαδήποτε στοιχεία που έχουν τη μορφή διαδικτυακών διευθύνσεων (π.χ. "https://www.example.com").

Στο MATLAB, η εντολή που εφαρμόστηκε είναι η εξής:

```
text = eraseURLs(text);
```

2.2.2. Tokenization (Διαχωρισμός Κειμένου σε Λέξεις)

Το πρώτο στάδιο της προεπεξεργασίας είναι το tokenization, δηλαδή ο διαχωρισμός του κειμένου σε μικρότερα τμήματα που ονομάζονται tokens. Ο **διαχωρισμός αυτός** αποτελεί μια θεμελιώδη διαδικασία στην επεξεργασία φυσικής γλώσσας και επηρεάζεται από τη γλώσσα στην οποία είναι γραμμένο το κείμενο (Webster & Kit, 1992) [36]. Το **Tokenization** επιτρέπει τη διάσπαση του κειμένου σε **λέξεις, προτάσεις ή χαρακτήρες**, διευκολύνοντας έτσι την κατανόηση της δομής και της πληροφορίας του.

Η πιο συνηθισμένη προσέγγιση στην ανάλυση κειμένου είναι η **κατακερμάτιση του κειμένου σε λέξεις**, καθώς οι περισσότερες τεχνικές NLP βασίζονται σε μεμονωμένες λέξεις ως **βασικές μονάδες ανάλυσης** (Kumhar et al., 2023).

Μαθηματικά, το tokenization μπορεί να περιγραφεί ως μια συνάρτηση f που λαμβάνει ένα κείμενο T και το μετατρέπει σε μια ακολουθία tokens $\{t_1, t_2, \ldots, t_n\}$:

$$f(T)=\{t_1, t_2, \ldots, t_n\}$$

όπου κάθε t_i αντιπροσωπεύει ένα token.

Στο MATLAB, το tokenization πραγματοποιείται με την ακόλουθη εντολή:

```
T = tokenizedDocument(T, Language="en");
```

2.2.3. Προσθήκη Σημασιολογικών Χαρακτηριστικών (POS Tagging & Named Entity Recognition - NER)

Ένα από τα σημαντικά βήματα στη γλωσσική ανάλυση είναι η κατανόηση του ρόλου που παίζει κάθε λέξη μέσα στο κείμενο. Αυτό επιτυγχάνεται με την εφαρμογή του **Part of Speech (POS) Tagging**, που προσδιορίζει τη γραμματική κατηγορία των λέξεων (ουσιαστικό, ρήμα, επίθετο κ.λπ.), καθώς και της

Named Entity Recognition (NER), που εντοπίζει και κατηγοριοποιεί ειδικές οντότητες, όπως ονόματα, τοποθεσίες και ημερομηνίες.

Η ενσωμάτωση αυτών των χαρακτηριστικών επιτρέπει την ακριβέστερη κατανόηση του περιεχομένου του κειμένου και μπορεί να βοηθήσει στη βελτίωση των μοντέλων ταξινόμησης.

Στο MATLAB, η διαδικασία υλοποιείται ως εξής:

```
T = addPartOfSpeechDetails(T);

T = addEntityDetails(T);
```

2.2.4. Κανονικοποίηση Λέξεων: Lemmatization vs Stemming

Η **κανονικοποίηση των λέξεων** βοηθά στην αναγνώριση διαφορετικών μορφών της ίδιας λέξης, μειώνοντας το μέγεθος του λεξιλογίου και βελτιώνοντας τη συνέπεια του κειμένου. Οι δύο βασικές μέθοδοι είναι το **stemming** και η **lemmatization**.

Το **stemming** αποκόπτει τα επιθήματα των λέξεων ώστε να επιστρέψει μια απλοποιημένη μορφή τους. Για παράδειγμα, η λέξη "running" μπορεί να μετατραπεί σε "run". Ωστόσο, η μέθοδος αυτή μπορεί να οδηγήσει σε μη φυσιολογικές μορφές λέξεων, όπως "comput" αντί "computer".

Η **lemmatization**, αντίθετα, βασίζεται στη λεξικογραφική ανάλυση της γλώσσας και μετατρέπει τις λέξεις στη βασική λεξική τους μορφή (λήμμα). Για παράδειγμα, η λέξη "better" μετατρέπεται σε "good".

Στην παρούσα εργασία, επιλέχθηκε η lemmatization αντί του stemming, καθώς θεωρείται προτιμότερη για ανάλυση συναισθήματος και κατηγοριοποίηση κειμένου, διότι διατηρεί τη νοηματική σχέση των λέξεων και αποφεύγει τα σφάλματα που μπορεί να προκαλέσει το stemming (Uysal & Gunal, 2014).

Στο MATLAB, η κανονικοποίηση των λέξεων πραγματοποιείται με την εντολή:

```
T = normalizeWords(T, Style="lemma");
```

2.2.5. Μετατροπή Κεφαλαίων σε Πεζά (Lowercasing)

Η τροποποίηση των κεφαλαίων γραμμάτων σε πεζά αποτελεί ένα βασικό βήμα προεπεξεργασίας κειμένου, το οποίο εφαρμόζεται ώστε να μην αποδίδεται διαφορετική σημασία σε λέξεις που διαφέρουν μόνο ως προς τη χρήση κεφαλαίων γραμμάτων (π.χ. "Game" και "game"). Οι κύριοι λόγοι

για τη μετατροπή σε πεζά είναι η εξομάλυνση της ανάλυσης και η αποφυγή περιττού πολυπλοκότητας στο λεξιλόγιο του μοντέλου.

Η συγκεκριμένη τεχνική έχει αποδειχθεί ότι βελτιώνει την ακρίβεια της ταξινόμησης κειμένου, καθώς μειώνει την πιθανότητα δημιουργίας περιττών διακριτών όρων που ουσιαστικά αναφέρονται στην ίδια έννοια (Uysal & Gunal, 2014), (HaCohen-Kerner, Miller, & Yigal, 2020).

Στο MATLAB, η μετατροπή σε πεζά γράμματα πραγματοποιείται με την εντολή:

```
T = lower(T);
```

2.2.6. Αφαίρεση Λέξεων χωρίς Πληροφοριακή Αξία (Stopword Removal)

Ορισμένες λέξεις εμφανίζονται πολύ συχνά σε ένα κείμενο, αλλά δεν προσφέρουν χρήσιμη πληροφορία στην ανάλυση. Αυτές ονομάζονται **stopwords**.

Εκτός από τη στάνταρ λίστα του MATLAB, αφαιρέθηκαν επιπλέον λέξεις που επαναλαμβάνονταν συχνά στις κριτικές, όπως "game", "play", "board", "player", καθώς και αριθμητικές αναφορές.

Η αφαίρεση έγινε με τη χρήση των εντολών:

```
T = removeWords(T, ["game", "play", "board",…,  "player"], IgnoreCase=true);
```

```
T = removeStopWords(T, IgnoreCase=true);
```

2.2.7. Αφαίρεση Αριθμών και Συμβόλων

Οι αριθμοί και τα σημεία στίξης συνήθως δεν προσφέρουν ουσιαστική πληροφορία στην ανάλυση συναισθήματος και επομένως αφαιρέθηκαν από το dataset, χρησιμοποιώντας τις εντολές:

```
T = erasePunctuation(T);
```

```
T = regexprep(T, '\d+', '');
```

2.2.8. Αφαίρεση Πολύ Μικρών και Πολύ Μεγάλων Λέξεων

Για την αποφυγή περιττών λέξεων, αφαιρέθηκαν όσες είχαν λιγότερους από **δύο χαρακτήρες** ή περισσότερους από **15**.

Αυτό έγινε στο MATLAB χρησιμοποιώντας τις εντολες:

```
T = removeShortWords(T, 2);
```

T = removeLongWords(T, 15);

2.3. Μέτρα Αξιολόγησης Μοντέλων

Η αξιολόγηση της απόδοσης των μοντέλων μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας, αποτελεί ένα από τα πιο κρίσιμα στάδια σε κάθε ερευνητική μελέτη, καθώς καθορίζει την αξιοπιστία και τη γενίκευση των αποτελεσμάτων. Στην ανάλυση συναισθήματος, τα μοντέλα ταξινόμησης χρησιμοποιούνται για να κατηγοριοποιήσουν κείμενα σε θετικά, αρνητικά ή ουδέτερα συναισθήματα. Ωστόσο, για να διαπιστωθεί η αποτελεσματικότητα ενός τέτοιου μοντέλου, είναι απαραίτητο να εφαρμοστούν **κατάλληλα μέτρα αξιολόγησης – ταξινόμησης**, οι οποίες θα αναδείξουν την ικανότητά του να εντοπίζει τις σωστές κατηγορίες.

Η επιλογή των μέτρων αξιολόγησης δεν είναι τυχαία. Κάθε μέτρο παρέχει διαφορετική πληροφορία σχετικά με την απόδοση του μοντέλου και επιλέγεται με βάση τη φύση του προβλήματος. Ένα βασικό στοιχείο που επηρεάζει τα μέτρα αξιολόγησης είναι η ποιότητα και η ισορροπία των δεδομένων. Αν ένα dataset περιλαμβάνει περισσότερα παραδείγματα μιας κλάσης από κάποιας άλλης (ανισορροπία κλάσεων), η ακρίβεια μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα. Συνεπώς, αντιλαμβανόμαστε ότι η κατανομή των δεδομένων, η καθαρότητα τους και η αυθεντικότητα τους είναι πολύ σημαντικά χαρακτηριστικά για την καλή απόδοση των μοντέλων.

Επίσης, είναι ουσιαστικό να χρησιμοποιούνται **πολλαπλά μέτρα αξιολόγησης** για να αποκτηθεί μια πληρέστερη εικόνα της απόδοσης του μοντέλου. Τα μέτρα απόδοσης που χρησιμοποιούνται συνήθως είναι η **ακρίβεια (accuracy), η ευστοχία (precision), η ανάκληση (recall) και το F1-score**. Επιπλέον, όταν το πρόβλημα αφορά ταξινόμηση σε περισσότερες από δύο κλάσεις, είναι σημαντικό να χρησιμοποιούνται προσαρμογές αυτών των μέτρων, όπως η **macro και weighted εκδοχή τους**, καθώς και ο **πίνακας σύγχυσης (confusion matrix)**, που αποτελεί βασικό εργαλείο κατανόησης της απόδοσης του μοντέλου.

2.3.1. Βασικά Μέτρα Αξιολόγησης

Ακρίβεια (Accuracy)

Η **ακρίβεια** είναι το πιο διαδεδομένο μέτρο αξιολόγησης και υπολογίζει το ποσοστό των σωστά προβλεφθέντων περιπτώσεων σε σχέση με το συνολικό αριθμό των προβλέψεων. Ορίζεται ως εξής:

$$Accuracy = \frac{1}{N}\sum_{k=1}^{N} 1_{\left[d^{(k)}=\hat{d}^{(k)}\right]},$$

όπου:

- N το συνολικό πλήθος των δειγμάτων που ανήκουν στο προς μελέτη σύνολο
- $d^{(k)}$ η ετικέτα (label) – στόχος που αντιστοιχεί στην είσοδο του δείγματος k με $d^{(k)} \epsilon \{1, 2, \ldots, Cl\}$
- $\hat{d}^{(k)}$ η ετικέτα που προβλέπει ο εκάστοτε ταξινομητής για την είσοδο του δείγματος k
- Cl το συνολικό πλήθος των δυνατών κλάσεων ταξινόμησης των εγγράφων

Η ακρίβεια είναι χρήσιμη όταν οι κλάσεις είναι ισορροπημένες, αλλά μπορεί να αποδειχθεί **παραπλανητική σε περιπτώσεις ανισόρροπων δεδομένων**. Για παράδειγμα, αν ένα dataset περιλαμβάνει 90% θετικές κριτικές και 10% αρνητικές, ένα μοντέλο που ταξινομεί όλα τα δείγματα ως θετικά θα έχει ακρίβεια 90%, παρά το γεγονός ότι **δεν έχει προβλέψει σωστά καμία από τις αρνητικές κριτικές**. Αυτό καθιστά απαραίτητη τη χρήση επιπλέον μέτρων απόδοσης.

Πίνακας Σύγχυσης (Confusion Matrix)

Ο **πίνακας σύγχυσης** είναι ένας τετραγωνικός πίνακας διάστασης ***Cl×Cl***, όπου ***Cl*** είναι το πλήθος των κλάσεων του ταξινομητή. Ο πίνακας (Εικόνα 1.1) αυτός χρησιμοποιείται για την αποτίμηση της απόδοσης ενός μοντέλου, καθώς απεικονίζει τη συσχέτιση μεταξύ των πραγματικών και των προβλεπόμενων ετικετών των δεδομένων. Το στοιχείο $C_{i,j}$ αντιπροσωπεύει το πλήθος των δειγμάτων που ανήκουν στην κλάση ***i*** αλλά ταξινομήθηκαν λανθασμένα ως ***j*** από το μοντέλο. Έτσι η κύρια διαγώνιος του πίνακα περιλαμβάνει το συνολικό αριθμό δειγμάτων που ταξινομούνται σωστά στην κάθε κλάση, και κατά συνέπεια όσα περισσότερα στοιχεία είναι μηδενικά, εκτός της κύριας διαγωνίου, τόσο σωστότερη είναι η ταξινόμηση.

Σε ένα πρόβλημα δυαδικής ταξινόμησης, με 2 δυνατές κλάσεις (συνήθως 0/1), ο πίνακας σύγχυσης έχει την ακόλουθη μορφή:

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Εικόνα 1.1. Πίνακας Σύγχυσης Δυαδικής Ταξινόμησης (Binary Classification Confusion Matrix)

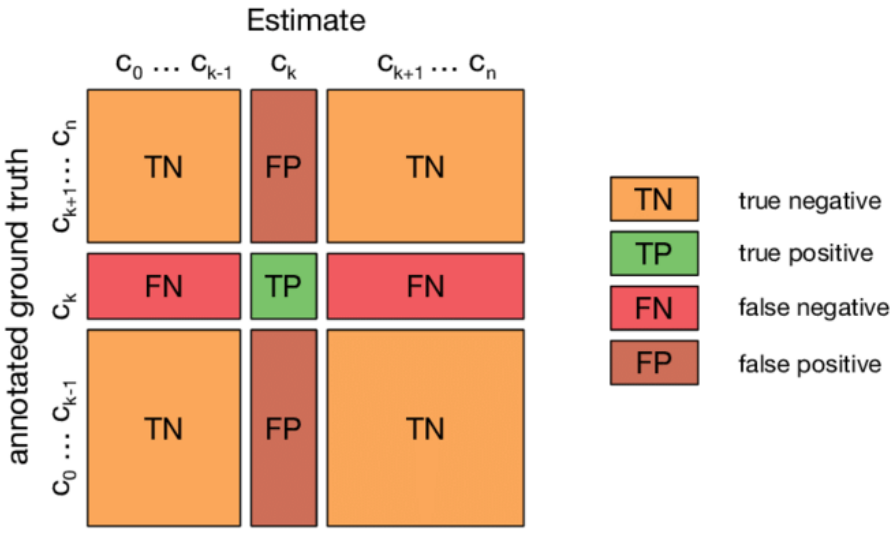
Τα στοιχεία του πίνακα ορίζονται ως:

- TP (True Positives): Τα δείγματα που ανήκουν στη θετική κλάση και προβλέφθηκαν σωστά.
- TN (True Negatives): Τα δείγματα που ανήκουν στην αρνητική κλάση και προβλέφθηκαν σωστά.
- FP (False Positives): Τα δείγματα που ανήκουν στην αρνητική κλάση αλλά ταξινομήθηκαν λανθασμένα ως θετικά.
- FN (False Negatives): Τα δείγματα που ανήκουν στη θετική κλάση αλλά ταξινομήθηκαν λανθασμένα ως αρνητικά.

Με τη βοήθεια του πίνακα σύγχυσης για δυαδική ταξινόμηση μπορούμε να υπολογίσουμε την **ακρίβεια** ως:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Στην περίπτωση της πολυκλασικής ταξινόμησης (multi-class classification), όταν κάθε διάνυσμα εισόδου έχει περισσότερες από δύο υποψήφιες κατηγορίες ταξινόμησης, έστω *Cl* σε πλήθος, στις οποίες μπορεί να ανήκει η αντίστοιχη ετικέτα του, τότε ο Πίνακας Σύγχυσης έχει την μορφή που φαίνεται στην (Εικόνα 1.2), για τυχαία κλάση *k*.



Εικόνα 1.2. Πίνακας Σύγχυσης Πολυκλασικής Ταξινόμησης (Multi-class Classification Confusion Matrix)

Η διαφορά με τη δυαδική ταξινόμηση είναι ότι ο χαρακτηρισμός ενός δείγματος με ένα από τα 4 δυνατά ονόματα (TP, TN, FP, FN) εξαρτάται από την υπό μελέτη κλάση. Έτσι, αν ορίσουμε F={1,2,…Cl} το σύνολο των δυνατών κλάσεων, για την κλάση k ($k \in F$), προκύπτουν:

- πλήθος των ψευδώς αρνητικών δειγμάτων

$$FN_k = \sum_{\substack{j \epsilon F \\ j \neq k}} C_{k,j}$$

- πλήθος των ψευδώς θετικών δειγμάτων

$$FP_k = \sum_{\substack{i \epsilon F \\ i \neq k}} C_{i,k}$$

- πλήθος των αληθώς αρνητικών δειγμάτων

$$TN_k = \sum_{i,j \epsilon F/_{\{k\}}} C_{i,j}$$

Και έτσι προκύπτει η συνολική ακρίβεια ταξινόμησης για όλες τις δυνατές κλάσεις:

$$Accuracy = \frac{\sum_{i \epsilon F} C_{i,i}}{\sum_{i \epsilon F} \sum_{j \epsilon F} C_{i,j}}$$

Ευστοχία (Precision)

Η **ευστοχία (precision)** μετρά το ποσοστό των προβλέψεων που ταξινομήθηκαν ως θετικές και ήταν όντως θετικές:

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

Στην περίπτωση της πολυκλασικής ταξινόμησης για μια κλάση *k*ϵ*F* η σχέση γράφεται:

$$Precision_k = \frac{C_{k,k}}{C_{k,k} + \sum_{\substack{i \epsilon F \\ i \neq k}} C_{i,k}} = \frac{C_{k,k}}{\sum_{i \epsilon F} C_{i,k}}$$

με τη συνολική τιμή της ευστοχίας (precision) της ταξινόμησης να προκύπτει αθροίζοντας τις επιμέρους τιμές precision για κάθε κλάση *k*. Επομένως, ισχύει:

$$Precision = \sum_{k \epsilon F} Precision_k = \sum_{k \epsilon F} \frac{C_{k,k}}{\sum_{i \epsilon F} C_{i,k}}$$

Η ευστοχία είναι ιδιαίτερα σημαντική σε περιπτώσεις όπου το κόστος των False Positives (FP) είναι υψηλό. Για παράδειγμα, σε ένα μοντέλο που ανιχνεύει ανεπιθύμητα email (spam detection), η χαμηλή ευστοχία σημαίνει ότι πολλά μη ανεπιθύμητα μηνύματα (non-spam) ταξινομούνται λανθασμένα ως spam, γεγονός που μπορεί να έχει αρνητικές επιπτώσεις στη χρηστικότητα του μοντέλου.

Ανάκληση (Recall)

Η **ανάκληση (recall)**, γνωστή και ως **ευαισθησία (sensitivity)** ή αλλιώς **producer's accuracy**, μετρά την ικανότητα του μοντέλου να εντοπίζει σωστά όλες τις θετικές περιπτώσεις και ορίζεται ως ο λόγος του πλήθους των σωστά ταξινομημένων δειγμάτων σε δεδομένη κλάση, προς το πλήθος των δειγμάτων που ήταν γνωστό εξαρχής ότι ανήκουν στη συγκεκριμένη κλάση.

Για δυαδική ταξινόμηση:

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

Στην περίπτωση της πολυκλασικής ταξινόμησης για μια κλάση *kϵF* η σχέση γίνεται:

$$Recall_k = \frac{C_{k,k}}{C_{k,k} + \sum_{\substack{j \epsilon F \\ j \neq k}} C_{k,j}} = \frac{C_{k,k}}{\sum_{j \epsilon F} C_{k,j}}$$

Αντίστοιχα με τη συνολική τιμή της ευστοχίας (precision), η συνολική τιμή της ανάκλησης (recall) της ταξινόμησης προκύπτει αθροίζοντας τις επιμέρους τιμές recall για κάθε κλάση *k*. Επομένως, ισχύει:

$$Recall = \sum_{k \epsilon F} Recall_k = \sum_{k \epsilon F} \frac{C_{k,k}}{\sum_{j \epsilon F} C_{k,j}}$$

Η ανάκληση είναι σημαντική όταν η παράβλεψη των θετικών περιπτώσεων έχει σοβαρές επιπτώσεις. Για παράδειγμα, σε ένα σύστημα ανίχνευσης απάτης σε τραπεζικές συναλλαγές, η χαμηλή ανάκληση σημαίνει ότι το μοντέλο αδυνατεί να εντοπίσει πολλές περιπτώσεις απάτης, κάτι που είναι ιδιαίτερα προβληματικό. Εύκολα μπορεί να συναχθεί το συμπέρασμα, ιδιαίτερα από τους τύπους για τις $Precision_k$ και $Recall_k$ αντίστοιχα ότι όσο πιο κοντά στο 1 είναι οι τιμές precision και recall για κάθε κλάση, τόσο πιο ακριβής είναι η συνολική ταξινόμηση.

F1-Score

Το **F1-score** είναι ο αρμονικός μέσος της ευστοχίας και της ανάκλησης και χρησιμοποιείται όταν χρειάζεται να υπάρχει μια **ισορροπημένη αξιολόγηση μεταξύ των δύο αυτών μέτρων**:

$$F1 - score = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \Leftrightarrow$$

$$F1 - score \ = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Όταν οι διαθέσιμες κατηγορίες ταξινόμησης είναι περισσότερες από δύο, το συνολικό F1-score δίνεται από το άθροισμα των F1-score που πετυχαίνει κάθε κατηγορία ξεχωριστά. Δηλαδή, ισχύει ότι

$$F1 - score = \sum_{k \epsilon F} (F1 - score)_k = \sum_{k \epsilon F} \frac{2 \cdot C_{k,k}}{\sum_{i \epsilon F} C_{i,k} + \sum_{j \epsilon F} C_{k,j}}$$

Το F1-score είναι ιδιαίτερα χρήσιμο όταν υπάρχει **ανισορροπία στις κλάσεις**, καθώς επιτυγχάνει μια καλύτερη εξισορρόπηση μεταξύ των False Positives και False Negatives. Όταν αυτό υπολογίζεται για κάθε κλάση, είναι το διάστημα [0,1], με τη μέγιστη τιμή, όταν αυτή επιτυγχάνεται, να δηλώνει τέλεια ταξινόμηση των δειγμάτων στη εκάστοτε κλάση.

2.3.2. Στατιστικά Τεστ Σύγκρισης Μοντέλων

Η χρήση των κλασικών μέτρων αξιολόγησης, όπως το **accuracy, precision, recall και F1-score**, παρέχει μια γενική εικόνα της απόδοσης ενός ταξινομητή. Ωστόσο, όταν επιθυμούμε να συγκρίνουμε δύο διαφορετικά μοντέλα μεταξύ τους, οι απλές αριθμητικές συγκρίσεις μπορεί να είναι ανεπαρκείς. Για την καλύτερη ποσοτική εκτίμηση των διαφορών μεταξύ δύο μοντέλων, εφαρμόζονται **στατιστικά τεστ σύγκρισης**, τα οποία επιτρέπουν την εξαγωγή ασφαλέστερων συμπερασμάτων ως προς το αν οι διαφορές στις επιδόσεις είναι **στατιστικά σημαντικές** ή απλώς τυχαίες.

Στην παρούσα εργασία, για τη **σύγκριση των αποτελεσμάτων του VADER και του GPT-4**, εφαρμόστηκαν δύο από τις πιο διαδεδομένες στατιστικές δοκιμές:

Paired t-test

Το **paired t-test** (εξαρτημένο t-test) είναι μια **παραμετρική στατιστική δοκιμή** που χρησιμοποιείται για να συγκρίνει τις μέσες τιμές δύο συσχετισμένων συνόλων δεδομένων. Στην περίπτωση της ανάλυσης συναισθήματος, το τεστ αυτό εφαρμόστηκε για να διαπιστωθεί αν η **μέση διαφορά στις επιδόσεις του VADER και του GPT-4 είναι στατιστικά σημαντική**.

Ο μαθηματικός τύπος του paired t-test είναι:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

όπου:

- $\bar{d}$ είναι η μέση διαφορά μεταξύ των τιμών των δύο μοντέλων,
- s_d είναι η τυπική απόκλιση των διαφορών,
- n είναι ο αριθμός των παρατηρήσεων.

Το αποτέλεσμα της δοκιμής συγκρίνεται με μια κρίσιμη τιμή από την κατανομή t-Student για να προσδιοριστεί αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των μοντέλων.

Wilcoxon Signed-Rank Test

Το **Wilcoxon signed-rank test** είναι μια **μη παραμετρική στατιστική δοκιμή** που χρησιμοποιείται όταν **δεν μπορεί να θεωρηθεί ότι οι διαφορές μεταξύ των δύο συνόλων δεδομένων ακολουθούν κανονική κατανομή**. Σε αντίθεση με το t-test, το οποίο προϋποθέτει ότι τα δεδομένα προέρχονται από κατανομή Gaussian, το Wilcoxon signed-rank test εξετάζει τη **διάμεσο των διαφορών** και είναι πιο ανθεκτικό στην παρουσία ακραίων τιμών (outliers).

Ο τύπος του Wilcoxon signed-rank test υπολογίζεται ως εξής:

$$W = \sum R^+ - \frac{n(n+1)}{4}$$

όπου $\boldsymbol{R^+}$ είναι το άθροισμα των θετικών βαθμολογιών των διαφορών μεταξύ των δύο μεθόδων και $\boldsymbol{n}$ το πλήθος των ζευγών δεδομένων.

Το Wilcoxon test χρησιμοποιήθηκε στη σύγκριση του VADER και του GPT-4 για να διαπιστωθεί αν οι διαφορές μεταξύ των δύο τεχνικών είναι **στατιστικά σημαντικές, ακόμα και όταν η κατανομή των διαφορών δεν είναι κανονική**.

Ερμηνεία των Στατιστικών Τεστ

Η **έξοδος** των στατιστικών τεστ επιστρέφει μια **p-value**, η οποία υποδεικνύει την πιθανότητα να έχουν προκύψει οι παρατηρούμενες διαφορές τυχαία, δηλαδή χρησιμοποιείται για να προσδιοριστεί εάν η διαφορά απόδοσης μεταξύ των δύο μοντέλων είναι στατιστικά σημαντική. Η ερμηνεία της **p-value**

εξαρτάται από το **κατώφλι στατιστικής σημαντικότητας** (α), το οποίο ορίζεται εκ των προτέρων και συνήθως λαμβάνει τιμές **0.05 ή 0.1**.

- Αν **p-value < 0.05**, τότε απορρίπτουμε τη μηδενική υπόθεση (H_0), πράγμα που σημαίνει ότι υπάρχει **στατιστικά σημαντική διαφορά** μεταξύ των δύο μεθόδων. Αυτό σημαίνει ότι η διαφορά δεν είναι τυχαία και πιθανώς οφείλεται στην πραγματική ανωτερότητα του ενός μοντέλου έναντι του άλλου.

- Αν **p-value < 0.1**, τότε η διαφορά θεωρείται **οριακά στατιστικά σημαντική**. Αυτό υποδεικνύει ότι υπάρχει κάποια ένδειξη διαφοροποίησης μεταξύ των δύο μοντέλων, αλλά η ισχυρότητα της απόδειξης είναι μικρότερη σε σχέση με το όριο του 0.05.

- Αν **p-value ≥ 0.1**, τότε δεν υπάρχουν επαρκή στατιστικά στοιχεία για να συμπεράνουμε ότι η διαφορά είναι σημαντική. Σε αυτή την περίπτωση, η διαφορά που παρατηρήθηκε μπορεί να οφείλεται σε **τυχαιότητα** και δεν μπορούμε να υποστηρίξουμε με σιγουριά ότι ένα μοντέλο υπερτερεί του άλλου.

Η εφαρμογή των **paired t-test και Wilcoxon signed-rank test** επέτρεψε την αντικειμενική αξιολόγηση της **στατιστικής σημαντικότητας** της διαφοράς μεταξύ των αποτελεσμάτων του **VADER και του GPT-4** στην ανάλυση συναισθήματος. Η ενσωμάτωση τέτοιων μεθόδων είναι ζωτικής σημασίας όταν συγκρίνονται διαφορετικές τεχνικές και μοντέλα, καθώς αποτρέπει τη λανθασμένη εξαγωγή συμπερασμάτων που βασίζονται αποκλειστικά στις αριθμητικές διαφορές των μέτρων αξιολόγησης.

2.4. Μείωση Διαστατικότητας (Dimensionality Reduction)

Η μείωση διαστατικότητας είναι μια σημαντική διαδικασία στην επεξεργασία δεδομένων, καθώς επιτρέπει τη μείωση του αριθμού των χαρακτηριστικών (features) διατηρώντας τη μέγιστη δυνατή πληροφορία. Αυτό διευκολύνει την εκπαίδευση των μοντέλων, μειώνει την πολυπλοκότητα και αντιμετωπίζει προβλήματα όπως η υπερπροσαρμογή (overfitting). Οι δύο πιο διαδεδομένες τεχνικές μείωσης διαστατικότητας είναι η **Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis - PCA)** και η **t-Distributed Stochastic Neighbor Embedding (t-SNE)**.

2.4.1. Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis-PCA)

Η **Ανάλυση Κύριων Συνιστωσών (PCA)** είναι μια γραμμική τεχνική μείωσης διαστατικότητας που χρησιμοποιείται για να βρει νέες διαστάσεις (κύριες συνιστώσες) που εξηγούν τη μέγιστη διασπορά στα δεδομένα. Οι κύριες συνιστώσες είναι ορθογώνιες μεταξύ τους και προκύπτουν από τον διαχωρισμό των διαστάσεων των δεδομένων με βάση την κατανομή της διασποράς τους.

Ο υπολογισμός των κύριων συνιστωσών γίνεται μέσω της **ιδιοτιμής** και των **ιδιοδιανυσμάτων** του πίνακα συνδιακύμανσης:

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$

όπου C είναι ο πίνακας συνδιακύμανσης, είναι τα δεδομένα εισόδου και x_i είναι ο μέσος όρος των δεδομένων.

Οι **ιδιοτιμές** καθορίζουν τη σημασία κάθε κύριας συνιστώσας, ενώ οι **ιδιοδιανύσματα** καθορίζουν τους νέους άξονες στο μειωμένο χώρο.

2.4.2. t-Distributed Stochastic Neighbor Embedding (t-SNE)

Η **t-SNE** είναι μια μη γραμμική τεχνική μείωσης διαστατικότητας, η οποία χρησιμοποιείται κυρίως για την οπτικοποίηση δεδομένων υψηλών διαστάσεων. Η βασική αρχή του αλγορίθμου είναι ότι διατηρεί τις τοπικές σχέσεις μεταξύ των σημείων, μετατρέποντας αποστάσεις σε πιθανότητες ομοιότητας και προβάλλοντας τα δεδομένα σε δύο ή τρεις διαστάσεις.

Η πιθανότητα ενός σημείου x_j να είναι γείτονας του x_i στο χώρο υψηλών διαστάσεων υπολογίζεται ως εξής:

$$p_{j|i} = \frac{exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq i} exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right)}$$

Στον «χαμηλοδιάστατο» χώρο, χρησιμοποιείται η κατανομή **t-Student** για την αναπαράσταση των δεδομένων:

$$q_{j|i} = \frac{(1 + \left\|y_i - y_j\right\|^2)^{-1}}{\sum_{k \neq i}(1 + \left\|y_i - y_j\right\|^2)^{-1}}$$

Η βελτιστοποίηση πραγματοποιείται με ελαχιστοποίηση της **KL απόκλισης (Kullback-Leibler divergence)** μεταξύ των κατανομών $p_{j|i}$ και $q_{j|i}$.

2.5. Διαδικασίες Διαχωρισμού Δεδομένων

Σε κάθε εφαρμογή μηχανικής μάθησης, είναι απαραίτητο να γίνεται διαχωρισμός των διαθέσιμων δεδομένων σε **εκπαιδευτικά (training)** και **δοκιμαστικά (testing)** σύνολα. Ο κύριος λόγος για αυτό είναι ότι τα μοντέλα θα πρέπει να μπορούν **να γενικεύουν σωστά σε νέα, άγνωστα δεδομένα** και όχι απλώς να απομνημονεύουν το σύνολο εκπαίδευσης. Αν δεν γίνει σωστά ο διαχωρισμός, είναι πιθανό το μοντέλο να παρουσιάσει υψηλή απόδοση στην εκπαίδευση, αλλά να αποτύχει πλήρως σε νέα δεδομένα — ένα φαινόμενο γνωστό ως **υπερπροσαρμογή (overfitting)**.

Ο διαχωρισμός επιτρέπει επίσης την **αντικειμενική αξιολόγηση της απόδοσης** του μοντέλου. Χωρίς ένα ανεξάρτητο test set ή κάποιο είδος επικύρωσης, τα αποτελέσματα του μοντέλου μπορεί να είναι υπερεκτιμημένα. Επομένως, οι τεχνικές διαχωρισμού αποτελούν ουσιαστικό βήμα για την ανάπτυξη έγκυρων και αξιόπιστων μοντέλων.

2.5.1. Hold-Out Split

Η **μέθοδος Hold-Out** είναι η πιο απλή τεχνική διαχωρισμού δεδομένων για εκπαίδευση και αξιολόγηση μοντέλων. Το dataset χωρίζεται σε:

- **Training set**: Το υποσύνολο αυτό χρησιμοποιείται για την εκπαίδευση του μοντέλου, δηλαδή για την προσαρμογή των παραμέτρων του.
- **Test set**: Αυτό το σύνολο παραμένει αχρησιμοποίητο κατά την εκπαίδευση και χρησιμοποιείται μόνο για την τελική αξιολόγηση της απόδοσης.

Η πιο συνηθισμένη κατανομή είναι **80%-20%**, όπου το 80% των δεδομένων χρησιμοποιείται για εκπαίδευση και το 20% για τεστ. Άλλες συχνές αναλογίες περιλαμβάνουν 70%-30% ή 90%-10%, ανάλογα με το μέγεθος του dataset και τις απαιτήσεις της εφαρμογής.

Η μέθοδος Hold-Out είναι γρήγορη και εύκολη στην εφαρμογή, ωστόσο μπορεί να οδηγήσει σε **μεροληπτικά ή ασταθή αποτελέσματα**, ιδιαίτερα σε περιπτώσεις μικρών συνόλων δεδομένων, καθώς η αξιολόγηση βασίζεται μόνο σε ένα υποσύνολο των παραδειγμάτων.

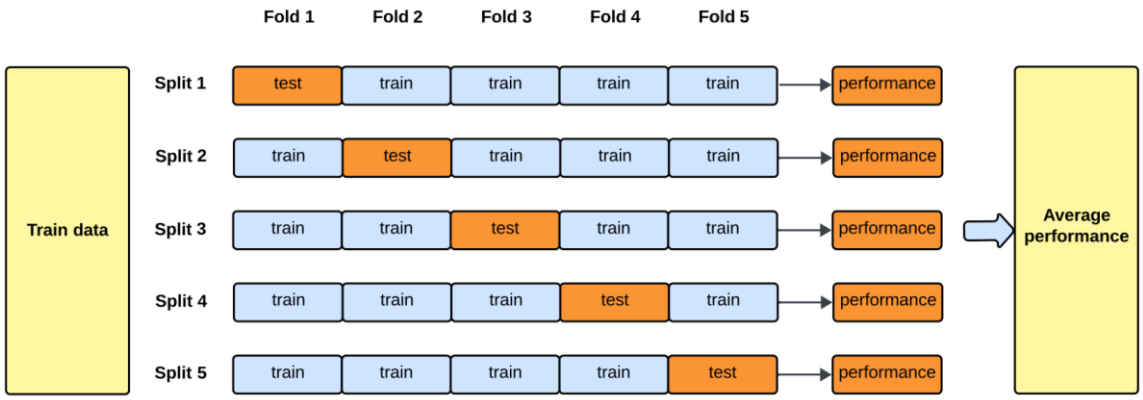
2.5.2. Cross-Validation

Η **διασταυρούμενη επικύρωση (cross-validation)** είναι μια πιο αξιόπιστη τεχνική διαχωρισμού δεδομένων, καθώς εξασφαλίζει ότι όλα τα δεδομένα χρησιμοποιούνται τόσο για εκπαίδευση όσο και για αξιολόγηση. Η πιο διαδεδομένη τεχνική είναι το **k-fold cross-validation**, όπου τα δεδομένα χωρίζονται σε ***k*** υποσύνολα:

1. Κάθε φορά, ένα υποσύνολο *k* χρησιμοποιείται για δοκιμή, ενώ τα υπόλοιπα *k-1* υποσύνολα χρησιμοποιούνται για εκπαίδευση.

2. Η διαδικασία επαναλαμβάνεται *k* φορές, και το τελικό σκορ προκύπτει ως ο μέσος όρος των αποτελεσμάτων.

Η cross-validation βοηθά στην **ελαχιστοποίηση της μεταβλητότητας** που προκαλείται από την τυχαία κατανομή των δεδομένων και προσφέρει μια πιο **γενικευμένη εκτίμηση** της απόδοσης του μοντέλου. Στην (Εικόνα 1.3) φαίνεται ένα παράδειγμα για 5 υποσύνολα.



Εικόνα 1.3. Γραφική Αναπαράσταση εκπαίδευσης με 5-fold Cross-Validation

2.6. Αλγόριθμοι Εκμάθησης ως Ταξινομητές (Classifiers)

Στην επιστήμη δεδομένων, ένας ταξινομητής (classifier) είναι ένας τύπος αλγόριθμου μηχανικής μάθησης που χρησιμοποιείται για την ανάθεση μιας ετικέτας κλάσης (class label) σε μια είσοδο δεδομένων. Ένα παράδειγμα είναι ένας ταξινομητής αναγνώρισης εικόνας για την επισήμανση μιας εικόνας (π.χ. "αυτοκίνητο", "φορτηγό" ή "πρόσωπο"). Οι αλγόριθμοι ταξινομητών εκπαιδεύονται χρησιμοποιώντας δεδομένα με ετικέτες (labeled data). στο παράδειγμα αναγνώρισης εικόνας, ο ταξινομητής λαμβάνει δεδομένα εκπαίδευσης που επισημαίνουν εικόνες. Μετά από επαρκή εκπαίδευση, ο ταξινομητής μπορεί στη συνέχεια να λάβει εικόνες χωρίς ετικέτα ως εισόδους και θα εξάγει ετικέτες ταξινόμησης για κάθε εικόνα.

Οι αλγόριθμοι εκμάθησης χρησιμοποιούν εξελιγμένες μαθηματικές και στατιστικές μεθόδους για να δημιουργήσουν προβλέψεις σχετικά με την πιθανότητα ταξινόμησης μιας εισαγωγής δεδομένων με δεδομένο τρόπο. Στο παράδειγμα αναγνώρισης εικόνας, ο ταξινομητής προβλέπει στατιστικά εάν μια εικόνα είναι πιθανό να είναι αυτοκίνητο, φορτηγό ή άνθρωπος, ή και κάποια άλλη ταξινόμηση την οποία ο ταξινομητής έχει εκπαιδευτεί να αναγνωρίζει.

Η παρούσα εργασία χρησιμοποιεί διάφορους αλγορίθμους εκμάθησης, δημιουργώντας ταξινομητές, για την ανάλυση συναισθήματος και την ταξινόμηση των σχολίων των χρηστών, που επιτρέπουν τη διαχωριστική μοντελοποίηση δεδομένων και την κατηγοριοποίησή τους σε διαφορετικές κλάσεις που αντιστοιχούν στην βαθμολογία (rating). Παρακάτω αναλύονται οι σημαντικότεροι αλγόριθμοι εκμάθησης που ορίζουν ταξινομητές.

2.6.1. Γραμμικός Ταξινομητής (Linear Classifier)

Ο γραμμικός ταξινομητής διαχωρίζει τα δεδομένα χρησιμοποιώντας μια ευθεία γραμμή (σε δύο διαστάσεις) ή ένα υπερεπίπεδο (hyperplane) σε υψηλότερες διαστάσεις. Η βασική μορφή ενός γραμμικού ταξινομητή είναι:

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

Όπου:

- x είναι το διάνυσμα χαρακτηριστικών,
- w είναι το διάνυσμα βαρών (*weights*) του μοντέλου,
- b είναι η προκατάληψη (*bias*), και
- y είναι η προβλεπόμενη ετικέτα (*label*).

Ένας γραμμικός ταξινομητής προσπαθεί να βρει τα **βέλτιστα βάρη** που διαχωρίζουν τις κλάσεις, συνήθως μέσω ενός αλγορίθμου βελτιστοποίησης όπως η **Καθοδική Κλίση (Gradient Descent)**, που είναι ένας επαναληπτικός αλγόριθμος βελτιστοποίησης πρώτης τάξης, για την εύρεση ενός τοπικού ελάχιστου/μέγιστου μιας δεδομένης συνάρτησης.

2.6.2. Υποστηρικτικές Διανυσματικές Μηχανές (SVM - Support Vector Machines)

Οι υποστηρικτικές διανυσματικές μηχανές (SVMs) είναι ένας ισχυρός αλγόριθμος ταξινόμησης που χρησιμοποιεί ένα υπερεπίπεδο (hyperplane) για να διαχωρίσει τις κλάσεις με το **μέγιστο περιθώριο (margin)**. Το βασικό κριτήριο εκπαίδευσης του SVM είναι η επίλυση της ακόλουθης συνθήκης βελτιστοποίησης:

$$min_{w,b} \frac{1}{2} \|w\|^2$$

υπό τους περιορισμούς: $y_i(w^T x_i + b) \geq 1, \ \forall \ i$

Ο **γραμμικός SVM (Linear SVM)** χρησιμοποιεί ένα **γραμμικό υπερεπίπεδο** για τη διάκριση των κλάσεων, ενώ ο **μη γραμμικός SVM (RBF SVM)** χρησιμοποιεί **πυρηνικές συναρτήσεις (kernel functions)**, όπως η **Radial Basis Function (RBF)**:

$$K(x_i, x_j) = exp\left(-\gamma \left\| x_i - x_j \right\|^2\right)$$

όπου είναι $\gamma = \frac{1}{2\sigma^2}$ μια υπερ-παράμετρος που καθορίζει την επιρροή κάθε δείγματος.

2.6.3. κ-Κοντινότεροι Γείτονες (k-Nearest Neighbors - KNN)

Ο αλγόριθμος **k-Nearest Neighbors (KNN)** βασίζεται στην απόσταση μεταξύ των δεδομένων. Για κάθε νέο δείγμα, το μοντέλο αναζητά τους **k πλησιέστερους γείτονες** και εκχωρεί την κλάση που έχει την πλειοψηφία μεταξύ αυτών.

Η απόσταση μεταξύ των σημείων συχνά υπολογίζεται μέσω της **Ευκλείδειας απόστασης**:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M} (x_{i,m} - x_{j,m})^2}$$

όπου M είναι ο αριθμός των χαρακτηριστικών.

2.6.4. Δέντρο Απόφασης (Decision Tree)

Τα δέντρα απόφασης δημιουργούν ένα **ιεραρχικό μοντέλο κανόνων** για την ταξινόμηση των δεδομένων. Κάθε κόμβος αντιπροσωπεύει μια **απόφαση βασισμένη σε κάποιο χαρακτηριστικό** και τα «φύλλα» του «δέντρου» περιέχουν την **τελική ταξινόμηση**. Σε αυτές τις δομές δέντρων (Εικόνα 1.4), τα «φύλλα» αντιπροσωπεύουν ετικέτες κλάσεων (class labels) και τα «κλαδιά» αντιπροσωπεύουν συνδέσμους χαρακτηριστικών που οδηγούν σε αυτές τις ετικέτες κλάσεων. Έτσι, προκύπτει ένα σύστημα γονέα (parent) και παιδιού (children). Η βέλτιστη διαίρεση των δεδομένων επιλέγεται με βάση κριτήρια όπως η **εντροπία**:

$$H(T) = I_E(p_1, p_2, \dots, p_J) = -\sum_{i=1}^{J} p_i log_2(pi)$$

όπου $(p_1, p_2, \dots, p_J)$ είναι κλάσματα που αθροίζουν στο 1 και παριστάνουν την πιθανότητα εμφάνισης κάθε κλάσης στα «φύλλα» που προκύπτει από μια διακλάδωση του δέντρου.

$$\overbrace{IG(T,a)}^{information\ gain} = \overbrace{H(T)}^{entropy\ (parent)} - \overbrace{H(T|a)}^{sum\ of\ entropies\ (children)}$$

Λαμβάνοντας τον μέσο όρο επί των δυνατών τιμών του *A*

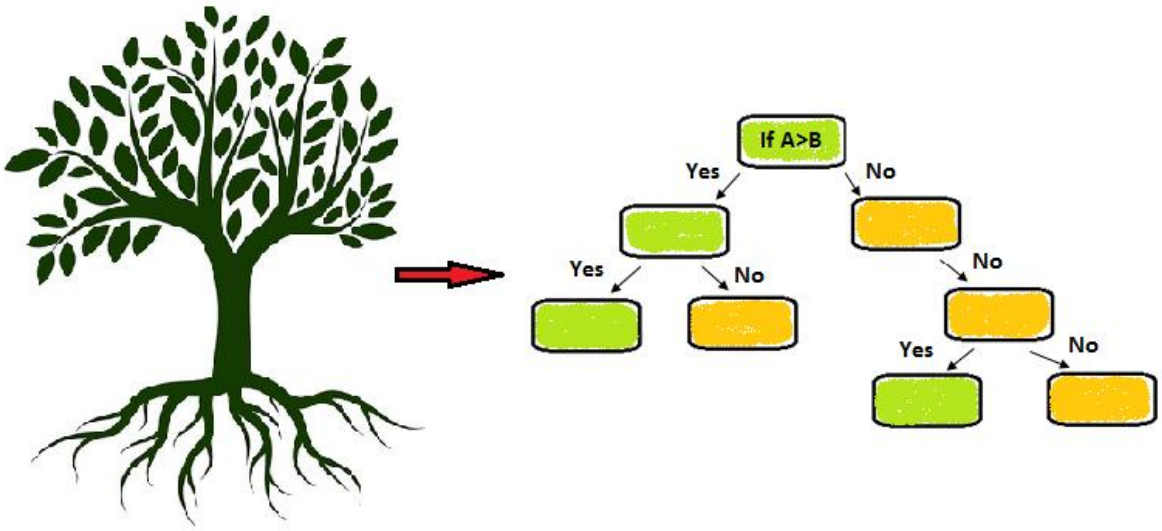
$$\overbrace{E_A\big(\overbrace{IG(T,a)}^{expected\ information\ gain}\big)}^{} = \overbrace{I(T;A)}^{mutual\ information\ between\ T\ and\ A}$$

$$= \overbrace{H(T)}^{entropy\ (parent)} - \overbrace{H(T|A)}^{weighted\ sum\ of\ entropies\ (children)}$$

$$= -\sum_{i=1}^{J} p_i log_2(pi) - \sum_{a} p(a) \sum_{i=1}^{J} \Pr(i|a) log_2 \Pr(i|a)$$

Δηλαδή, το αναμενόμενο κέρδος πληροφοριών E_A είναι η αμοιβαία πληροφορία (μέτρο της αμοιβαίας εξάρτησης μεταξύ των δύο μεταβλητών), που σημαίνει ότι κατά μέσο όρο, η μείωση της εντροπίας του *T* είναι η αμοιβαία πληροφορία.

Το κέρδος πληροφοριών IG ($information\ gain$) χρησιμοποιείται για να αποφασίσει ποιο χαρακτηριστικό θα διαχωριστεί σε κάθε βήμα στην κατασκευή του δέντρου. Στη θεωρία πληροφοριών και τη μηχανική μάθηση, το κέρδος πληροφοριών είναι συνώνυμο της απόκλισης Kullback–Leibler: ο όγκος των πληροφοριών που αποκτήθηκαν για μια τυχαία μεταβλητή ή ένα σήμα από την παρατήρηση μιας άλλης τυχαίας μεταβλητής.



Εικόνα 1.4. Απλή Μορφή Δέντρου Αποφάσεων

Το απλούστερο είναι το καλύτερο, επομένως είναι επιθυμητό να διατηρείται μικρό το δέντρο. Για να γίνει αυτό, σε κάθε βήμα θα πρέπει να επιλέγεται η διαίρεση που έχει ως αποτέλεσμα τους πιο συνεπείς θυγατρικούς κόμβους. Ένα ευρέως χρησιμοποιούμενο μέτρο συνέπειας ονομάζεται *πληροφορία* που μετράται σε bit. Για κάθε κόμβο του δέντρου, η τιμή πληροφοριών "αντιπροσωπεύει τον αναμενόμενο όγκο πληροφοριών που θα χρειαζόταν για να καθοριστεί εάν μια νέα παρουσία θα

πρέπει να ταξινομηθεί ναι ή όχι, δεδομένου ότι το παράδειγμα έφτασε σε αυτόν τον κόμβο" (Witten et al. 2011) [29].

2.6.5. Naive Bayes

Ο αλγόριθμος **Naive Bayes** βασίζεται στο **Θεώρημα του Bayes**, το οποίο περιγράφει την πιθανότητα ενός γεγονότος δεδομένης κάποιας πληροφορίας:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Η Naive Bayes είναι μια απλή τεχνική για την κατασκευή ταξινομητών, μοντέλων που εκχωρούν ετικέτες κλάσεων (class labels) σε στιγμιότυπα προβλημάτων (problem instances), που αναπαρίστανται ως διανύσματα τιμών χαρακτηριστικών (vectors of feature values), όπου οι ετικέτες κλάσεων προέρχονται από κάποιο πεπερασμένο σύνολο. Δεν υπάρχει ένας μόνο αλγόριθμος για την εκπαίδευση τέτοιων ταξινομητών, αλλά μια οικογένεια αλγορίθμων που βασίζεται σε μια κοινή αρχή: όλοι οι ταξινομητές Naive Bayes υποθέτουν ότι η τιμή ενός συγκεκριμένου χαρακτηριστικού είναι ανεξάρτητη από την τιμή οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της μεταβλητής κλάσης. Για παράδειγμα, ένα φρούτο μπορεί να θεωρηθεί ότι είναι μήλο εάν είναι κόκκινο, στρογγυλό και έχει διάμετρο περίπου 10 cm. Ένας ταξινομητής Naive Bayes θεωρεί ότι καθένα από αυτά τα χαρακτηριστικά συμβάλλει ανεξάρτητα στην πιθανότητα ότι αυτό το φρούτο είναι μήλο, ανεξάρτητα από τυχόν πιθανούς συσχετισμούς μεταξύ των χαρακτηριστικών χρώματος, στρογγυλότητας και διαμέτρου.

2.6.6. Τυχαίο Δάσος (Random Forest)

Το **Random Forest** είναι ένας αλγόριθμος συνόλου (ensemble) ταξινόμησης και παλινδρόμησης που αποτελείται από **πολλαπλά δέντρα απόφασης** (Εικόνα 1.5). Ο βασικός στόχος του είναι να μειώσει την υπερεκπαίδευση (overfitting) που μπορεί να προκύψει από τη χρήση ενός μεμονωμένου δέντρου απόφασης, συνδυάζοντας πολλαπλά δέντρα για τη βελτίωση της ακρίβειας του μοντέλου. Γενικά οι μέθοδοι συνόλου χρησιμοποιούν πολλαπλούς αλγόριθμους μάθησης για να επιτύχουν καλύτερη προγνωστική απόδοση από ό,τι θα μπορούσε να επιτευχθεί από οποιονδήποτε από τους αλγόριθμους μάθησης μεμονωμένα (Opitz, Maclin, et al. 1999).

Η πρόβλεψη ενός δείγματος γίνεται με **πλειοψηφική ψήφο (majority voting) στις ταξινομήσεις** ή με τον μέσο όρο στην περίπτωση της παλινδρόμησης.

Η βασική διαδικασία κατασκευής ενός **Random Forest** περιλαμβάνει:

1. **Bootstrap Sampling**: Επιλογή τυχαίων υποσυνόλων δεδομένων για την εκπαίδευση κάθε δέντρου.

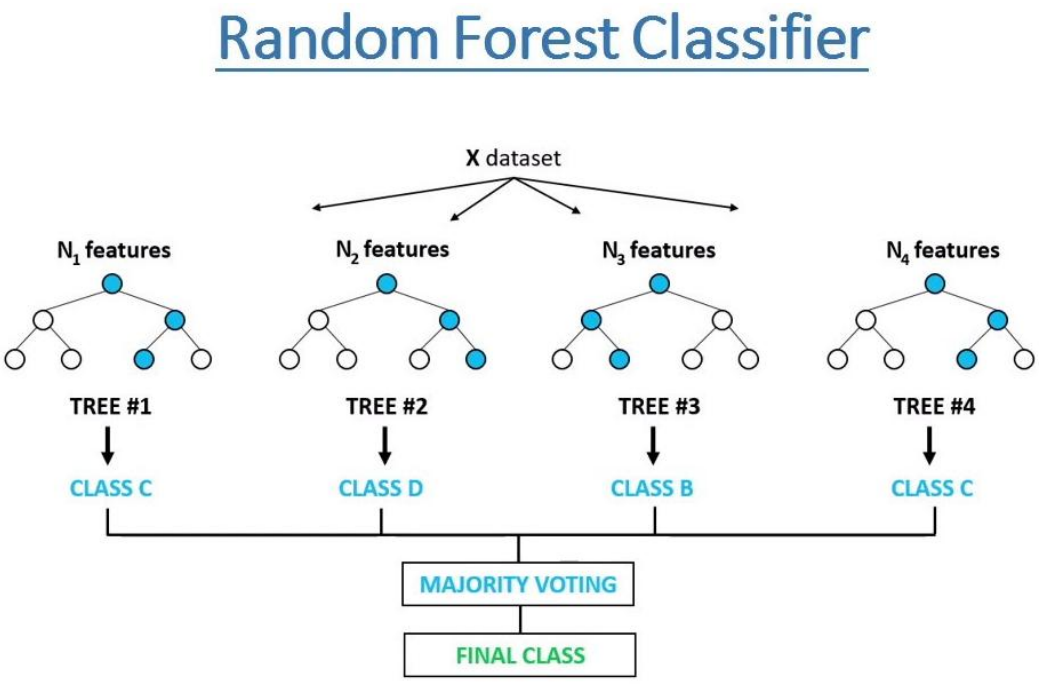
2. **Τυχαία επιλογή χαρακτηριστικών** σε κάθε διαχωρισμό (split) του δέντρου.

3. **Συνδυασμός των αποτελεσμάτων** όλων των δέντρων μέσω πλειοψηφικής ψήφου για ταξινόμηση ή μέσου όρου για παλινδρόμηση.

Η πιθανότητα ενός δείγματος *x* να ανήκει σε μια κλάση *c* υπολογίζεται ως:

$$P(c|x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Όπου, T είναι ο αριθμός των δέντρων στο δάσος και $h_t(x)$ είναι η πρόβλεψη του t-οστού δέντρου απόφασης, με την πολυπλοκότητα του μοντέλου καθορίζεται από το βάθος των δέντρων και τον αριθμό των χαρακτηριστικών που επιλέγονται τυχαία σε κάθε διαχωρισμό.



Εικόνα 1.5. Δομή Τυχαίου Δάσους αποτελούμενο από τέσσερα Δέντρα Αποφάσεων

2.7. Προηγμένα Γλωσσικά Μοντέλα

2.7.1. BERT

Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα γλωσσικό μοντέλο που παρουσιάστηκε τον Οκτώβριο του 2018 από ερευνητές της Google (Devlin, Jacob, Chang et al. 2018) [8]. Μαθαίνει να αναπαριστά το κείμενο ως ακολουθία διανυσμάτων μέσω αυτοεπιβλεπόμενης μάθησης (self-supervised learning) και βασίζεται αποκλειστικά στην αρχιτεκτονική του encoder των transformers. Η εμφάνισή του σηματοδότησε μια σημαντική πρόοδο στην απόδοση των μεγάλων γλωσσικών μοντέλων. Από το 2020 και έπειτα, το BERT αποτελεί ευρέως αποδεκτή βασική γραμμή (baseline) για πειράματα επεξεργασίας φυσικής γλώσσας (NLP).

Η εκπαίδευσή του γίνεται μέσω **πρόβλεψης καλυμμένων λέξεων** (masked token prediction) και **πρόβλεψης επόμενης πρότασης** (next sentence prediction). Μέσω αυτής της διαδικασίας, το BERT μαθαίνει λανθάνουσες και συμφραζόμενες αναπαραστάσεις των λέξεων στο περιβάλλον τους, παρόμοια με τα μοντέλα ELMo και GPT-2. Έχει εφαρμοστεί με επιτυχία σε πλήθος εργασιών NLP, όπως η επίλυση αντωνυμιών και η αποσαφήνιση λέξεων με πολλαπλές σημασίες. Αποτελεί εξέλιξη του ELMo και η αποτελεσματικότητά του οδήγησε στην ανάπτυξη ενός ολόκληρου ερευνητικού κλάδου, γνωστού ως "BERTology", που εστιάζει στην ερμηνεία των εσωτερικών λειτουργιών του μοντέλου.

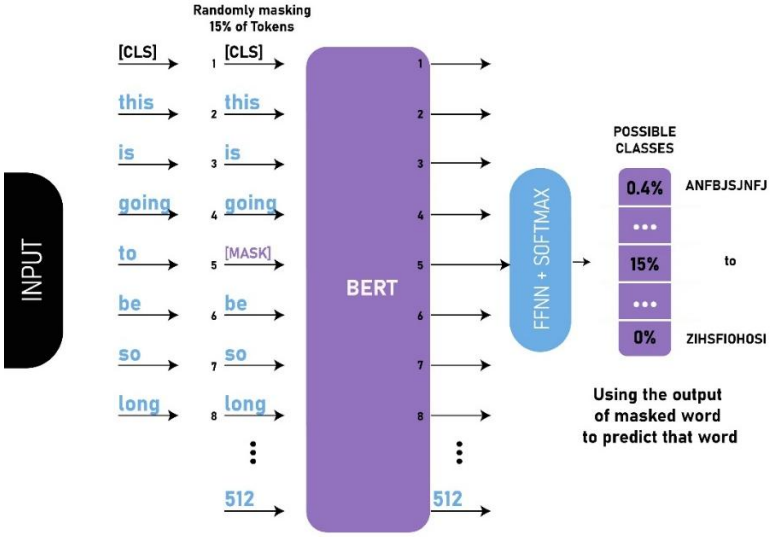
Τα προεκπαιδευμένα μοντέλα **Tiny BERT, Mini BERT και Small BERT** είναι συμπυκνωμένες εκδόσεις του **BERT (Bidirectional Encoder Representations from Transformers)**, που στοχεύουν στη **μείωση του υπολογιστικού κόστους** διατηρώντας παράλληλα ικανοποιητική απόδοση.

Η εκπαίδευση αυτών των μοντέλων πραγματοποιείται συχνά μέσω της διαδικασίας **Knowledge Distillation**, κατά την οποία ένα πλήρες μοντέλο BERT (Teacher Model) «μεταδίδει» τη γνώση του σε ένα μικρότερο μοντέλο (Student Model), μέσω μιας συνάρτησης απώλειας που ελαχιστοποιεί τη διαφορά μεταξύ των προβλέψεών τους.

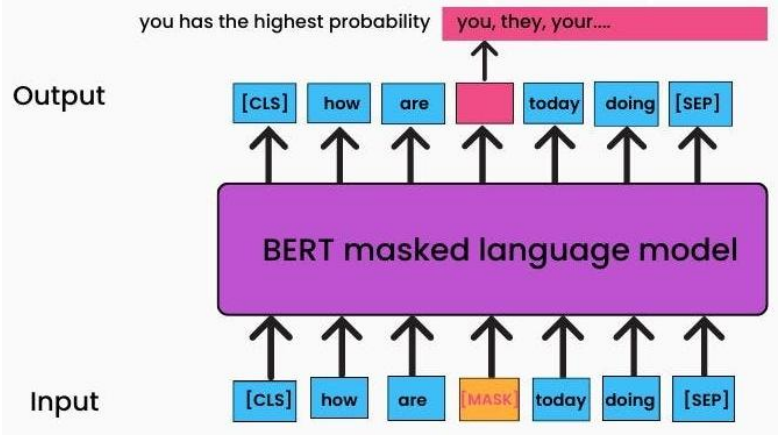
Στην παρούσα εργασία χρησιμοποιούνται οι **προεκπαιδευμένες συμπυκνωμένες εκδόσεις του BERT**, που παρέχονται μέσω του περιβάλλοντος MATLAB, αποκλειστικά ως **εξαγωγείς χαρακτηριστικών** (feature extractors), **χωρίς να εφαρμόζεται περαιτέρω εκπαίδευση ή βελτιστοποίηση (fine-tuning)**. Δεν έγινε χρήση του βασικού μοντέλου (base model) ούτε των εκτενέστερων εκδόσεών του, λόγω περιορισμένων υπολογιστικών πόρων. Τα εν λόγω μοντέλα απαιτούν σημαντική υπολογιστική ισχύ και, συνεπώς, παρατεταμένο χρόνο εκτέλεσης, κάτι που υπερέβαινε τις διαθέσιμες δυνατότητες υπό ρεαλιστικές συνθήκες.

Η βασική αρχιτεκτονική τους, (Εικόνα 1.6) & (Εικόνα 1.7), ακολουθεί τον ίδιο σχεδιασμό με το BERT base model, το οποίο έχει 108.8 εκατομμύρια παραμέτρους εκμάθησης, αλλά με **λιγότερες «κεφαλές προσοχής» (attention heads) και στρώματα (transformer layers)**.

- **Tiny BERT**: Περιλαμβάνει **2 ή 4 στρώματα** transformer και είναι ιδιαίτερα ελαφρύ, με 4.3 εκατομμύρια παραμέτρους εκμάθησης.
- **Mini BERT**: Περιλαμβάνει **4 ή 6 στρώματα** και διατηρεί ισορροπία μεταξύ απόδοσης και ταχύτητας, με 11.1 εκατομμύρια παραμέτρους εκμάθησης.
- **Small BERT**: Περιλαμβάνει **8 ή περισσότερα στρώματα** και προσφέρει βελτιωμένη απόδοση έναντι των μικρότερων εκδόσεων, με 28.5 εκατομμύρια παραμέτρους εκμάθησης.



Εικόνα 1.6. Αρχιτεκτονική Προεκπαίδευσης BERT με χρήση Masked Language Modeling (MLM) | Παράδειγμα 1



Εικόνα 1.7. Αρχιτεκτονική Προεκπαίδευσης BERT με χρήση Masked Language Modeling (MLM) | Παράδειγμα 2

2.7.2. VADER (Valence Aware Dictionary and sEntiment Reasoner)

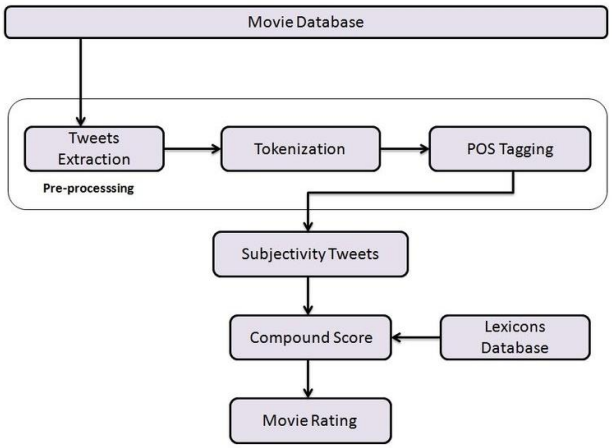
Το **VADER** είναι ένα **λεξικογραφικό και βασισμένο σε κανόνες** εργαλείο ανάλυσης συναισθήματος, ειδικά σχεδιασμένο για σύντομα κείμενα όπως tweets και σχόλια σε κοινωνικά δίκτυα (Εικόνα 1.8).

Η συνολική συναισθηματική βαθμολογία ενός κειμένου υπολογίζεται ως εξής:

$$S = \sum_{i=1}^{N} w_i v_i$$

Όπου, w_i είναι το βάρος της λέξης *i* και v_i είναι η συναισθηματική βαθμολογία της λέξης *i*.

Το VADER χρησιμοποιεί μια λίστα λέξεων με προεπιλεγμένες συναισθηματικές βαθμολογίες, ονόματι compound scores, τις οποίες αναπτύσσει μέσω ανθρώπινης επισήμανσης (human-annotated lexicon). Παράλληλα, διαθέτει ενσωματωμένους κανόνες για την ανίχνευση ενισχυτικών ή αποδυναμωτικών στοιχείων (όπως κεφαλαία γράμματα, σημεία στίξης, επιρρήματα και εικονίδια έκφρασης – emojis).



Εικόνα 1.8. Διάγραμμα Ροής Ανάλυσης Συναισθήματος με VADER

2.7.3. GPT-4 Turbo

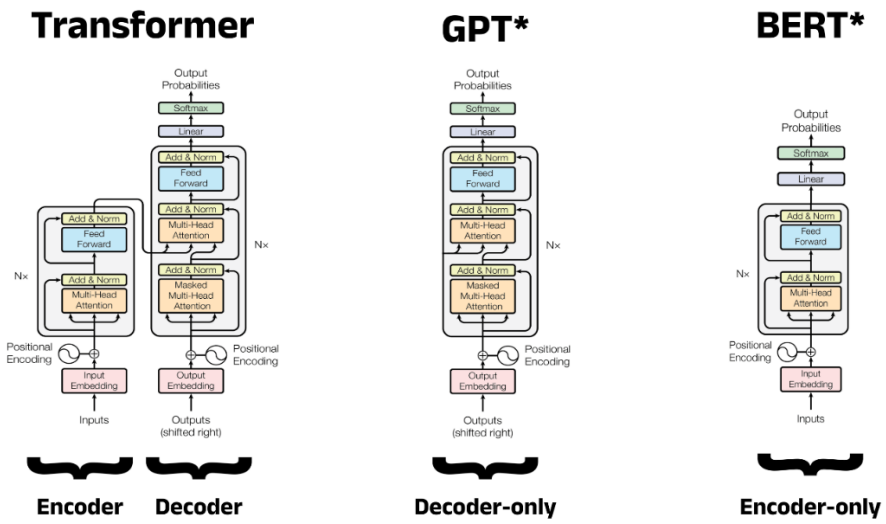
Το GPT-4 Turbo αποτελεί βελτιστοποιημένη εκδοχή του μοντέλου GPT-4, σχεδιασμένη για **υψηλότερη υπολογιστική αποδοτικότητα**, **μειωμένο κόστος ανά αίτημα** και **ταχύτερη απόκριση**, διατηρώντας παράλληλα το ίδιο επίπεδο ακρίβειας και απόδοσης σε εργασίες επεξεργασίας φυσικής γλώσσας (NLP).

Παραμένει πιστό στη **Transformer-based αρχιτεκτονική** (Εικόνα 1.9), με τεχνικές βελτιώσεις που ενισχύουν την **παράλληλη επεξεργασία δεδομένων**, την **απόδοση σε περιβάλλοντα υψηλού φόρτου** και την **αποτελεσματική διαχείριση μνήμης**. Η προεκπαίδευσή του βασίζεται στην

αυτοεπιβλεπόμενη μάθηση (self-supervised learning), μέσω πρόβλεψης του επόμενου token σε μια ακολουθία, αξιοποιώντας μεγάλες ποσότητες μη επισημασμένων δεδομένων.

Το GPT-4 Turbo υποστηρίζει ευρύ φάσμα εφαρμογών NLP, όπως ανάλυση συναισθήματος, κατανόηση και παραγωγή κειμένου, συνοπτική απόδοση κειμένου, μετάφραση και απάντηση σε ερωτήματα ελεύθερου τύπου.

Το μοντέλο διατίθεται προς χρήση μέσω του **OpenAI API**, επιτρέποντας την απομακρυσμένη πρόσβαση σε προεκπαιδευμένα LLMs υψηλής απόδοσης. Μέσω κατάλληλων HTTP requests στη MATLAB ή client βιβλιοθηκών (π.χ. Python SDK), παρέχεται η δυνατότητα ενσωμάτωσης του μοντέλου σε pipelines ανάλυσης φυσικής γλώσσας, χωρίς την ανάγκη τοπικής εγκατάστασης και εκπαίδευση.



Εικόνα 1.9. Παράδειγμα Αρχιτεκτονικής Transformers, μοντέλων GPT και BERT

3. Μεθοδολογία

Η ανάλυση δεδομένων που πραγματοποιείται στην παρούσα εργασία ακολουθεί μια δομημένη προσέγγιση, όπου κάθε στάδιο επεξεργασίας και εφαρμογής μοντέλων έχει υλοποιηθεί ως ξεχωριστή συνάρτηση στο MATLAB. Για την αποτελεσματική οργάνωση και εκτέλεση όλων των διαδικασιών, δημιουργήθηκε το κύριο αρχείο εκτέλεσης (main_code.mlx), το οποίο περιλαμβάνει όλα τα βήματα της ανάλυσης σε διακριτά sections.

Το *main_code.mlx*, λειτουργεί ως κεντρικό σημείο εκκίνησης, λαμβάνοντας τα δεδομένα από το dataset, και καλώντας τις επιμέρους συναρτήσεις που έχουν αναπτυχθεί για την επεξεργασία δεδομένων, την εφαρμογή μοντέλων ανάλυσης συναισθήματος, την ταξινόμηση των κριτικών και τη θεματική μοντελοποίηση. Με αυτόν τον τρόπο, επιτυγχάνεται μια οργανωμένη και αυτοματοποιημένη ροή εργασιών, επιτρέποντας στον χρήστη να εκτελέσει ολόκληρη την ανάλυση με μία μόνο εκτέλεση του αρχείου.

3.1. Περιγραφή Βάσης Δεδομένων (Dataset)

Προέλευση των Δεδομένων

Τα δεδομένα που χρησιμοποιούνται στην παρούσα εργασία προέρχονται από το Kaggle, συγκεκριμένα από τη βάση δεδομένων (dataset) “BoardGameGeek Reviews”, το οποίο είναι διαθέσιμο στη διεύθυνση: bgg-15m-reviews.csv. Το αρχείο αυτό περιέχει περισσότερες από 15 εκατομμύρια κριτικές χρηστών για επιτραπέζια παιχνίδια από την πλατφόρμα BoardGameGeek (BGG), έναν από τους μεγαλύτερους διαδικτυακούς κόμβους για τους λάτρεις των επιτραπέζιων παιχνιδιών.

Η συγκεκριμένη βάση δεδομένων αποτελείται από κριτικές που έχουν αφήσει οι χρήστες του BGG, βαθμολογώντας διάφορα επιτραπέζια παιχνίδια σε μία κλίμακα από το 1 έως το 10. Εκτός από τη βαθμολογία, καταγράφονται και τα σχόλια των χρηστών, τα οποία παρέχουν ποιοτική πληροφόρηση σχετικά με την εμπειρία τους από το παιχνίδι. Αυτά τα δεδομένα κρίνονται ιδιαίτερα χρήσιμα για την ανάλυση συναισθήματος, καθώς επιτρέπουν τη συσχέτιση των γραπτών κριτικών με τις αριθμητικές βαθμολογίες.

Επιλογή του Παιχνιδιού «7 Wonders»

Αν και το αρχικό dataset περιλαμβάνει δεδομένα για έναν μεγάλο αριθμό επιτραπέζιων παιχνιδιών, επιλέχθηκε να περιοριστεί η ανάλυση αποκλειστικά στο παιχνίδι «7 Wonders». Το «7 Wonders» είναι ένα πολυβραβευμένο στρατηγικό παιχνίδι καρτών, το οποίο θεωρείται ένα από τα πιο δημοφιλή

επιτραπέζια των τελευταίων ετών. Ο βασικός λόγος για αυτήν την επιλογή είναι η μεγάλη ποσότητα διαθέσιμων κριτικών που υπάρχουν για το συγκεκριμένο παιχνίδι, επιτρέποντας την εκτέλεση μιας αξιόπιστης στατιστικής και γλωσσικής ανάλυσης.

Για την εξαγωγή των σχετικών δεδομένων, εφαρμόστηκε ένα φιλτράρισμα στο αρχικό dataset, με σκοπό να διατηρηθούν μόνο οι εγγραφές που αναφέρονται στο «7 Wonders». Μετά τη διαδικασία φιλτραρίσματος, παρέμειναν 10.369 κριτικές που περιλαμβάνουν σχόλια και αντίστοιχες βαθμολογίες χρηστών.

Αρχικός Καθαρισμός των Δεδομένων

Η διαδικασία καθαρισμού των δεδομένων πραγματοποιήθηκε αρχικά εντός του Microsoft Excel, χρησιμοποιώντας τα κατάλληλα φίλτρα και τεχνικές φιλτραρίσματος δεδομένων. Δεδομένου ότι το αρχικό dataset περιείχε πάνω από 15 εκατομμύρια κριτικές, ήταν απαραίτητο να επιλεχθούν μόνο οι σχετικές εγγραφές και να απομακρυνθούν οι περιττές ή μη χρήσιμες καταχωρήσεις. Προκειμένου να διασφαλιστεί η ποιότητα των δεδομένων που θα χρησιμοποιηθούν για την ανάλυση, εφαρμόστηκαν τα ακόλουθα κριτήρια καθαρισμού:

1. **Διατήρηση μόνο αγγλόφωνων κριτικών**

    - Το αρχικό dataset περιείχε κριτικές σε πολλές διαφορετικές γλώσσες. Για τη διασφάλιση της ομοιογένειας των δεδομένων, επιλέχθηκε να διατηρηθούν μόνο οι κριτικές που είναι γραμμένες στα αγγλικά. Η γλώσσα των κριτικών ανιχνεύθηκε με χρήση κατάλληλων γλωσσικών εργαλείων, και όλες οι μη αγγλικές κριτικές αφαιρέθηκαν από το dataset.

2. **Αφαίρεση ανούσιων ή ακατανόητων σχολίων**

    - Ορισμένες κριτικές περιείχαν άχρηστο περιεχόμενο, όπως τυχαίους χαρακτήρες (π.χ., "φξδησξφασ"), άσχετα σύμβολα ή ακολουθίες αριθμών που δεν προσέφεραν ουσιαστική πληροφορία. Αυτές οι κριτικές απομακρύνθηκαν, καθώς δεν είχαν γλωσσική σημασία και δεν θα συνέβαλαν ουσιαστικά στην ανάλυση του συναισθήματος.

3. **Απομάκρυνση σχολίων που περιείχαν μόνο αριθμούς**

    - Σε ορισμένες περιπτώσεις, οι χρήστες είχαν υποβάλει σχόλια που περιείχαν μόνο αριθμούς, χωρίς συνοδευτικό κείμενο. Αυτές οι εγγραφές διαγράφηκαν, καθώς δεν

θα μπορούσαν να χρησιμοποιηθούν αποτελεσματικά στην ανάλυση φυσικής γλώσσας.

Δομή του Τελικού Dataset

Μετά τη διαδικασία φιλτραρίσματος και καθαρισμού, το dataset διαμορφώθηκε ως εξής:

- **Συνολικός αριθμός κριτικών**: **10.369**
- **Στήλες δεδομένων**:
    1. **Χρήστης (User)** – Το όνομα του χρήστη που άφησε την κριτική.
    2. **Βαθμολογία (Rating)** – Η αριθμητική βαθμολογία που έδωσε ο χρήστης, σε κλίμακα από το 1 έως το 10.
    3. **Σχόλιο (Comment)** – Το κείμενο της κριτικής που περιέχει τη γνώμη του χρήστη για το παιχνίδι.

Κανονικοποίηση και Κατηγοριοποίηση των Βαθμολογιών

Στο αρχικό dataset, οι βαθμολογίες των χρηστών καταγράφονται σε μια διακριτή κλίμακα με δεκαδικές τιμές, δηλαδή περιλαμβάνουν τιμές όπως 3.6, 4.9, 9.5 κ.λπ. Ωστόσο, προκειμένου να καταστούν πιο διαχειρίσιμες και να διευκολυνθεί η ανάλυση, εφαρμόστηκε στρογγυλοποίηση. Αυτή η στρογγυλοποίηση διασφαλίζει ότι οι αριθμητικές τιμές των αξιολογήσεων είναι συνεπείς και ευκολότερες στην επεξεργασία κατά την περαιτέρω ανάλυση.

Μετά την επιλογή και τον καθαρισμό των δεδομένων και την στρογγυλοποίηση, προέκυψε η ανάγκη για μια ομαδοποίηση των βαθμολογιών σε τρεις διακριτές κατηγορίες, ώστε να διευκολυνθεί η ανάλυση και η σύγκριση των συναισθημάτων των χρηστών. Η αρχική βαθμολογία των χρηστών, η οποία κυμαινόταν από 1 έως 10, κατηγοριοποιήθηκε ως εξής:

- **Χαμηλές Βαθμολογίες (1-4)**: Περιλαμβάνουν τις αξιολογήσεις των χρηστών που αποτυπώνουν μια σαφώς αρνητική εμπειρία από το παιχνίδι. Σε αυτήν την κατηγορία ανήκουν οι χρήστες που έδωσαν βαθμολογία από 1 έως 4, γεγονός που υποδηλώνει ότι το παιχνίδι δεν ανταποκρίθηκε στις προσδοκίες τους ή δεν τους άρεσε καθόλου.
- **Μέτριες Βαθμολογίες (5-7)**: Αυτή η κατηγορία αντιπροσωπεύει χρήστες που είχαν ουδέτερη ή ανάμεικτη εμπειρία με το παιχνίδι. Οι βαθμολογίες 5, 6 και 7 δείχνουν ότι οι χρήστες δεν

ήταν ούτε ενθουσιασμένοι ούτε εντελώς απογοητευμένοι, αλλά πιθανώς θεώρησαν ότι το παιχνίδι έχει κάποια θετικά και κάποια αρνητικά στοιχεία.

- **Υψηλές Βαθμολογίες (8-10)**: Η τελευταία κατηγορία περιλαμβάνει τις βαθμολογίες 8, 9 και 10, οι οποίες υποδηλώνουν ιδιαίτερα θετική εμπειρία από το παιχνίδι. Οι χρήστες αυτοί θεώρησαν ότι το παιχνίδι είναι ποιοτικό, διασκεδαστικό και άξιζε μια υψηλή βαθμολογία.

Η κατηγοριοποίηση αυτή επιτρέπει τη σαφέστερη ανάλυση των συναισθημάτων των παικτών, καθώς καθιστά πιο εύκολη τη σύγκριση των λεκτικών κριτικών με τις αντίστοιχες βαθμολογίες. Επιπλέον, αυτή η ομαδοποίηση διευκολύνει την εκπαίδευση των μοντέλων ταξινόμησης, καθώς μετατρέπει μια κλίμακα 10 τιμών σε μια πιο διαχειρίσιμη κλίμακα τριών κατηγορικών κατηγοριών.

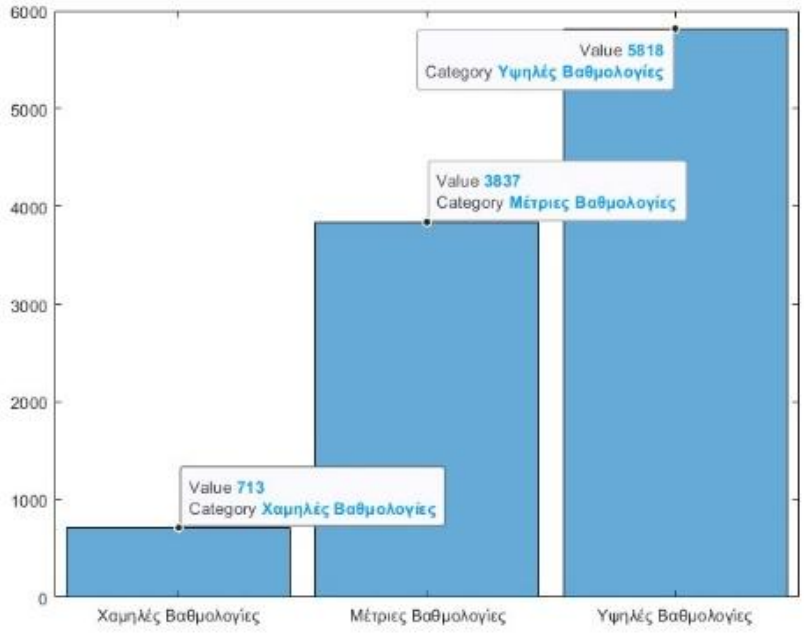
Κατανομή των Βαθμολογιών

Για την καλύτερη κατανόηση του τρόπου με τον οποίο οι χρήστες αξιολογούν το παιχνίδι, πραγματοποιήθηκε μια στατιστική ανάλυση της κατανομής των βαθμολογιών.

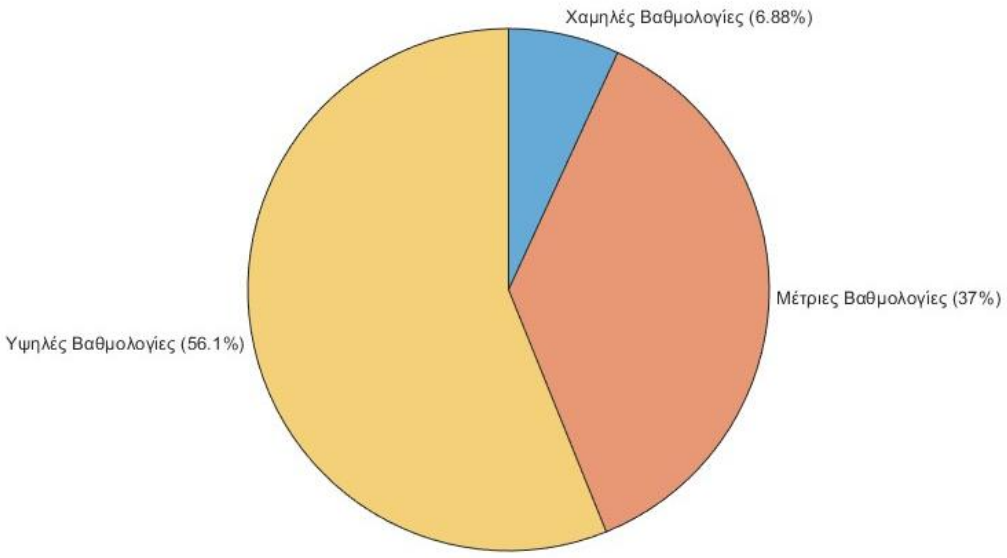
Από τα δεδομένα προέκυψε ότι:

- Οι **υψηλές βαθμολογίες (8-10)** αποτελούν το **56.1%** του συνόλου των κριτικών.
- Οι **μέτριες βαθμολογίες (5-7)** καταλαμβάνουν το **37%** των εγγραφών.
- Οι **χαμηλές βαθμολογίες (1-4)** είναι περιορισμένες, μόλις **6.88%**.

Αυτά τα στοιχεία απεικονίζονται στις παρακάτω γραφικές παραστάσεις, οι οποίες παρέχουν μια οπτική αναπαράσταση της κατανομής των αξιολογήσεων:



Εικόνα 1.10. Ραβδόγραμμα (Bar Chart):Παρουσιάζει τη συχνότητα εμφάνισης των διαφορετικών κατηγοριών βαθμολογιών.



Εικόνα 1.11. Διάγραμμα Πίτας (Pie Chart): Εμφανίζει το ποσοστό κάθε κατηγορίας, δίνοντας μια πιο συνοπτική εικόνα της αναλογίας των θετικών, ουδέτερων και αρνητικών κριτικών.

Οι αναλύσεις αυτές είναι σημαντικές, καθώς μας δίνουν μια πρώτη εικόνα της γενικής τάσης των χρηστών απέναντι στο παιχνίδι και μας επιτρέπουν να κατανοήσουμε αν υπάρχει ισορροπία ή αν η πλειοψηφία των κριτικών κλίνει προς μια συγκεκριμένη κατεύθυνση.

3.2. Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων αποτελεί ένα από τα πιο κρίσιμα στάδια στη διαδικασία ανάλυσης φυσικής γλώσσας (NLP), καθώς επιτρέπει την εξαγωγή καθαρών και ομοιογενών κειμένων, απαλλαγμένων από θόρυβο, τα οποία είναι κατάλληλα για περαιτέρω ανάλυση. Στην παρούσα εργασία, η διαδικασία αυτή πραγματοποιήθηκε χρησιμοποιώντας το **Text Analytics Toolbox του MATLAB**, το οποίο παρέχει μια σειρά από εργαλεία για την κανονικοποίηση και τον καθαρισμό των δεδομένων.

Η προεπεξεργασία εφαρμόστηκε στα σχόλια των χρηστών που είχαν συλλεχθεί από το dataset, με στόχο τη μείωση των περιττών πληροφοριών και τη διατήρηση μόνο των στοιχείων που είναι χρήσιμα για την ανάλυση συναισθήματος και την ταξινόμηση των κριτικών. Για την υλοποίηση της διαδικασίας δημιουργήθηκε η συνάρτηση ***text_preprocessing.mlx***, η οποία δέχεται ως **είσοδο** το σύνολο των σχολίων που βρίσκονται στη στήλη **comment** και επιστρέφει ως **έξοδο** το προεπεξεργασμένο κείμενο σε μορφή tokenized document. Η συνάρτηση καλείται στο κύριο αρχείο εκτέλεσης **main_code.mlx**, επιτρέποντας την αυτοματοποίηση της διαδικασίας με την εντολή:

```
[preprocessedText] = text_preprocessing(T.comment);
```

Η προεπεξεργασία περιλαμβάνει μια σειρά από στάδια, καθένα από τα οποία συμβάλλει στον καθαρισμό και την ετοιμότητα των δεδομένων για περαιτέρω επεξεργασία.

Το πρώτο στάδιο αφορά την **αφαίρεση συνδέσμων** (links, URLs) που μπορεί να έχουν απομείνει στα δεδομένα, και έπειτα τον **διαχωρισμό του κειμένου σε tokens** (**tokenization**), δηλαδή τη μετατροπή των κριτικών σε μεμονωμένες λέξεις , επιτρέποντας την ανάλυση της γλώσσας σε μικρότερες μονάδες. Με τον τρόπο αυτό, το κείμενο μετατρέπεται σε μορφή που μπορεί να χρησιμοποιηθεί από τα μοντέλα επεξεργασίας κειμένου. Στη συνέχεια, προκειμένου να εμπλουτιστεί η γλωσσική πληροφορία του dataset, εφαρμόστηκε **Part of Speech (POS) tagging**, το οποίο προσδιορίζει τη γραμματική κατηγορία κάθε λέξης (ουσιαστικό, ρήμα, επίθετο κ.λπ.), καθώς και **Named Entity Recognition (NER)**, το οποίο ανιχνεύει και ταξινομεί συγκεκριμένες οντότητες, όπως ονόματα και τοποθεσίες, χαρακτηριστικά που μπορούν να αξιοποιηθούν σε πιο προχωρημένες αναλύσεις, όπως **topic modeling.**

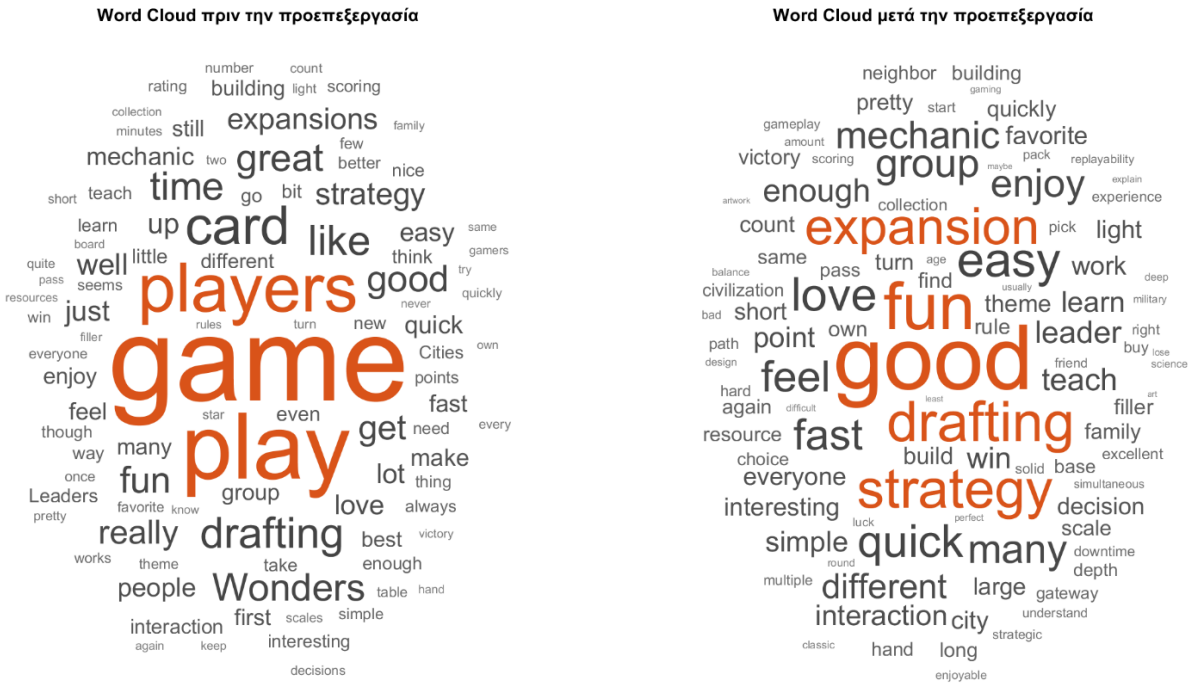
Ακολούθως, εφαρμόστηκε **lemmatization**, μια διαδικασία κανονικοποίησης των λέξεων, κατά την οποία διαφορετικές μορφές της ίδιας λέξης ενώνονται σε μία βασική, λεξικογραφική μορφή. Για παράδειγμα, οι λέξεις «*playing*, *played* και *plays»* μετατρέπονται όλες στη ρίζα *play*, επιτρέποντας τη μείωση του αριθμού των μοναδικών λέξεων και τη βελτίωση της ανάλυσης. Η επιλογή της **lemmatization** αντί του **stemming** έγινε προκειμένου να διατηρηθεί το νοηματικό περιεχόμενο των λέξεων χωρίς παραμορφώσεις. Επίσης, πραγματοποιήθηκε **μετατροπή όλων των χαρακτήρων σε πεζά** (lowercasing), προκειμένου να ενοποιηθούν οι περιπτώσεις λέξεων που διαφέρουν μόνο ως προς τη χρήση κεφαλαίων γραμμάτων. Η διαδικασία αυτή εξασφαλίζει ότι όροι όπως «Game», «game» και «GAME» αντιμετωπίζονται ως ένα ενιαίο λεκτικό στοιχείο, μειώνοντας τον αριθμό των μοναδικών tokens.

Ένα σημαντικό βήμα στην προεπεξεργασία ήταν η **αφαίρεση λέξεων που δεν προσφέρουν πληροφοριακή αξία (stopwords)**. Εκτός από τη λίστα των τυπικών stopwords που παρέχεται από το MATLAB, δημιουργήθηκε μια **προσαρμοσμένη λίστα λέξεων** που εμφανίζονταν συχνά στα σχόλια χωρίς να συνεισφέρουν ουσιαστικά στην ανάλυση. Αυτές οι λέξεις περιλάμβαναν γενικούς όρους όπως *«game*, *play*, *board»*, το όνομα του επιτραπέζιου καθώς και αριθμητικές αναφορές. Ο στόχος αυτής της διαδικασίας ήταν να αφαιρεθούν επαναλαμβανόμενες λέξεις που δεν σχετίζονται άμεσα με το συναίσθημα ή την άποψη του χρήστη για το παιχνίδι.

Επιπλέον, για τη μείωση του θορύβου, αφαιρέθηκαν **όλοι οι αριθμοί και τα σημεία στίξης**. Οι αριθμοί αντικαταστάθηκαν με κενό, καθώς συχνά εμφανίζονταν σε μορφές όπως *10/10* ή *5 stars*, οι οποίες δεν προσέφεραν χρήσιμη πληροφορία για την ανάλυση συναισθήματος. Αντίστοιχα, τα σημεία στίξης, όπως κόμματα και τελείες, αφαιρέθηκαν, καθώς δεν είχαν γλωσσική σημασία στη συγκεκριμένη περίπτωση.

Τέλος, εφαρμόστηκαν **περιορισμοί μήκους λέξεων**, προκειμένου να εξαλειφθούν πιθανές λανθασμένες εγγραφές ή τυπογραφικά σφάλματα. Συγκεκριμένα, αφαιρέθηκαν λέξεις που είχαν λιγότερους από 2 χαρακτήρες ή περισσότερους από 15, καθώς κρίθηκαν είτε άχρηστες είτε αποτέλεσμα σφαλμάτων στην πληκτρολόγηση των κριτικών.

Για την αξιολόγηση της αποτελεσματικότητας της προεπεξεργασίας, δημιουργήθηκε ένα **Word Cloud**, (Εικόνα 1.12) το οποίο απεικονίζει τις πιο συχνά εμφανιζόμενες λέξεις στις κριτικές, πριν και μετά τον καθαρισμό. Το Word Cloud επιτρέπει την οπτική επιβεβαίωση ότι έχουν αφαιρεθεί οι περιττές λέξεις και ότι οι σημαντικοί όροι που σχετίζονται με το περιεχόμενο των κριτικών εξακολουθούν να είναι παρόντες.



Εικόνα 1.12. Word Cloud, το οποίο απεικονίζει τις πιο συχνά εμφανιζόμενες λέξεις στις κριτικές, πριν και μετά τον καθαρισμό

3.3. Ανάλυση Συχνών Φράσεων ανά Κατηγορία Βαθμολογίας

Στο πλαίσιο της ανάλυσης των σχολίων των χρηστών, κρίθηκε σκόπιμο να διερευνηθεί ποιες φράσεις επαναλαμβάνονται συχνότερα στις κριτικές, ανάλογα με τη βαθμολογία που έχουν αποδώσει οι χρήστες. Η προσέγγιση αυτή βασίστηκε στην υπόθεση ότι οι φράσεις που χρησιμοποιούνται σε αρνητικές κριτικές πιθανόν να διαφοροποιούνται σημαντικά από αυτές που απαντώνται σε θετικές αξιολογήσεις, τόσο ως προς το περιεχόμενο όσο και ως προς τη συναισθηματική φόρτιση. Κάτι, που όπως θα γίνει εμφανές παρακάτω, δεν ισχύει πάντα.

Για την υλοποίηση της ανάλυσης δημιουργήθηκε η συνάρτηση **multiword_phrases**, η οποία κατηγοριοποιεί τις κριτικές σε τρεις ομάδες βάσει της στρογγυλοποιημένης βαθμολογίας:

- **Χαμηλές Βαθμολογίες**: από 1 έως και 5,
- **Μέτριες Βαθμολογίες**: από 5.1 έως 7.9,
- **Υψηλές Βαθμολογίες**: από 8 έως 10.

Αρχικά, οι βαθμολογίες μετατράπηκαν σε κατηγορικές ετικέτες, με στόχο την ανεξάρτητη επεξεργασία των σχολίων ανά ομάδα. Στη συνέχεια, για κάθε ομάδα εφαρμόστηκε η συνάρτηση

text_preprocessing που είχε υλοποιηθεί προηγουμένως, ώστε οι κριτικές να καθαριστούν από θόρυβο και να μετατραπούν σε μορφή tokenizedDocument. Αφού προεπεξεργάστηκαν τα δεδομένα, δημιουργήθηκαν **bag-of-ngrams** για κάθε ομάδα, περιορίζοντας την ανάλυση σε φράσεις που αποτελούνται από **τρεις διαδοχικές λέξεις (trigrams)**.

Για την οπτική αναπαράσταση των συχνότερων φράσεων, δημιουργήθηκαν **γραφικές παραστάσεις τύπου Word Cloud**, όπου οι πιο συχνές φράσεις εμφανίζονται με μεγαλύτερο μέγεθος, επιτρέποντας την άμεση οπτική σύγκριση μεταξύ των τριών κατηγοριών. Οι παραστάσεις δημιουργούνται με τη χρήση της εντολής wordcloud() σε συνδυασμό με bagOfNgrams, για την διατήρηση τριών λέξεων (N=3) και έτσι την δημιουργία και οπτικοποίηση φράσεων σε τρία υποδιαγράμματα (subplots), (Εικόνα 1.13).

Εικόνα 1.13. Word Clouds των πιο συχνών φράσεων τριών λέξεων (trigrams) ανά κατηγορία βαθμολογίας

Τέλος, η συνάρτηση επιστρέφει τις **δέκα πιο συχνές φράσεις** για κάθε κατηγορία (μέσω της εντολής topkngrams), παρέχοντας τη δυνατότητα αποθήκευσης ή περαιτέρω ανάλυσής τους. Αυτό γίνεται εκτυπώνοντας έναν πίνακα με στήλες, για την φράση (ngram) που δημιουργείται από τις 3 λέξεις (ngram length), βάσει της συχνότητας εμφάνισης της στα δεδομένα (count).

Πίνακας 1.1. Top 10 φράσεις στις χαμηλές βαθμολογίες.

Ngram	Count	NgramLength
multiple path victory	6	3
race galaxy glory	2	3
galaxy glory rome	2	3
long term strategy	2	3
pick pass rest	2	3
fun social experience	2	3
path victory pretty	2	3
pick hope pass	2	3
gratification victory point	2	3
victory point others	2	3

Πίνακας 1.2. Top 10 φράσεις στις μέτριες βαθμολογίες.

Ngram	Count	NgramLength
multiple path victory	18	3
good large group	13	3
many path victory	12	3
different path victory	10	3
simultaneous action selection	8	3
leader city expansion	7	3
easy learn fast	7	3
easy teach learn	7	3
expansion city leader	6	3
fairly easy learn	6	3

Πίνακας 1.3. Top 10 φράσεις στις υψηλές βαθμολογίες.

Ngram	Count	NgramLength
multiple path victory	42	3
leader city expansion	27	3
different path victory	21	3
many path victory	20	3
expansion leader city	19	3
love drafting mechanic	18	3
many different strategy	16	3
leader city pack	15	3
quick easy learn	14	3
microbadge microbadge microbadge	14	3

Η συνάρτηση καλείται στο κύριο αρχείο main_code.mlx ως:

```
[toplow, topmed, tophigh] = multiword_phrases(T);
```

και τα αποτελέσματα εκτυπώνονται στην κονσόλα για κάθε κατηγορία ξεχωριστά. Με τον τρόπο αυτό, επιτυγχάνεται η σύνδεση της **ποιοτικής φύσης των φράσεων με το ποσοτικό πλαίσιο της βαθμολογίας**, ενισχύοντας την ερμηνευτική ισχύ της ανάλυσης συναισθήματος.

Αυτή η τεχνική επιτρέπει την ανάδειξη θεματικών προτύπων που σχετίζονται με θετικές ή αρνητικές εμπειρίες των χρηστών, συμβάλλοντας έτσι στη βαθύτερη κατανόηση της ψυχολογίας και των προτιμήσεών των χρηστών. Σε πρακτικό επίπεδο, τα αποτελέσματα αυτής της ανάλυσης μπορούν να χρησιμοποιηθούν για την εξαγωγή πολύτιμων συμπερασμάτων ως προς τα στοιχεία που επηρεάζουν θετικά ή αρνητικά την εμπειρία του παίκτη.

Παρατηρώντας τα αποτελέσματα των word clouds και των πινάκων με τις δέκα συχνότερες φράσεις, εντοπίζεται αξιοσημείωτη ομοιότητα μεταξύ των κατηγοριών. Φράσεις όπως «multiple path victory», «different path victory» και «leader city expansion» εμφανίζονται με συνέπεια σε όλες τις βαθμολογικές ομάδες, γεγονός που υποδηλώνει ότι τα κοινά θεματικά μοτίβα – όπως η στρατηγική, οι διαφορετικοί τρόποι νίκης και οι επεκτάσεις – είναι κομβικά σημεία στην εμπειρία του παιχνιδιού, ανεξαρτήτως της αξιολόγησης. Αυτή η ομοιογένεια ενδέχεται να προκύπτει είτε επειδή οι παίκτες αναγνωρίζουν αυτά τα στοιχεία ως ουδέτερες βασικές μηχανικές του παιχνιδιού, είτε επειδή

πρόκειται για χαρακτηριστικά που λειτουργούν θετικά για κάποιους και αρνητικά για άλλους, ανάλογα με τις προτιμήσεις τους.

3.4. Απλός Εποπτευόμενος Ταξινομητής με χρήση Bag-of-Words

Σε αυτό το βήμα, υλοποιήθηκε ένας **απλός εποπτευόμενος πολυκλασικός ταξινομητής**, με σκοπό την κατηγοριοποίηση των σχολίων σε τρεις ομάδες βάσει της βαθμολογίας. Ο στόχος ήταν η εκπαίδευση ενός γραμμικού μοντέλου που θα μπορεί να προβλέψει σωστά την κατηγορία βαθμολογίας για νέα, αδημοσίευτα σχόλια, τα οποία είναι αρχικά αποθηκευμένα σε ένα διάνυσμα new_reviews = […].

Συγκεκριμένα, τα νέα σχόλια που χρησιμοποιήθηκαν είναι τα παρακάτω:

1. «7 Wonders is an exceptional game that brilliantly captures the thrill of civilization building. The drafting mechanic keeps all players engaged, and the diverse strategies make each game unique. With beautiful artwork and well-designed components, it’s a visual treat as well. It’s a perfect choice for both casual and serious gamers looking to have a great time together.»

2. «While 7 Wonders has its fans, I found it somewhat underwhelming. The game can feel a bit repetitive after a few plays, as the strategies can become predictable. Additionally, the scoring system can be confusing for new players, leading to a frustrating experience. It’s not the engaging civilization-building game I hoped it would be.»

3. «7 Wonders is not a great game. I felt that the learning curve can be steep for newcomers, which might hinder the enjoyment for some. Overall, in my opinion, it is not a game for those who don't want to think very much.»

Η διαδικασία ξεκινά με τη μετατροπή των αριθμητικών βαθμολογιών σε κατηγορικές, ακολουθώντας την ίδια λογική με προηγούμενες ενότητες: Χαμηλές (1–5), Μέτριες (5.1–7.9) και Υψηλές (8–10). Στη συνέχεια, εφαρμόζεται διαχωρισμός δεδομένων μέσω της τεχνικής Hold-Out Split, για τρία διαφορετικά ποσοστά εκπαίδευσης-δοκιμής (90-10, 80-20 και 70-30), ώστε να αξιολογηθεί η σταθερότητα και η γενικευσιμότητα του μοντέλου σε διαφορετικά σενάρια.

Αφού προεπεξεργαστούν τα δεδομένα κειμένου μέσω της συνάρτησης text_preprocessing, ακολουθεί η δημιουργία ενός Bag-of-Words (BoW) μοντέλου, το οποίο μετατρέπει τα σχόλια σε αριθμητική αναπαράσταση. Η εκπροσώπηση αυτή επιτρέπει τη χρήση κλασικών αλγορίθμων μηχανικής μάθησης. Για το σκοπό αυτό, χρησιμοποιήθηκε η συνάρτηση **fitcecoc** με γραμμικό αλγόριθμο εκμάθησης-learner (Linear SVM), το οποίο ενδείκνυται για προβλήματα πολυκλασικής ταξινόμησης.

Αφού εκπαιδευτεί το μοντέλο με τα δεδομένα εκπαίδευσης, πραγματοποιείται πρόβλεψη για το σύνολο δοκιμής και **υπολογίζεται η ακρίβεια** (accuracy) για κάθε περίπτωση. Παράλληλα, το μοντέλο εφαρμόζεται και σε τρία νέα σχόλια, τα οποία αξιολογούνται με βάση την εκπαιδευμένη γνώση του ταξινομητή.

Αξίζει να σημειωθεί πως η δομή του κώδικα περιλαμβάνει πρόβλεψη τόσο σε δεδομένα δοκιμής όσο και σε νέα ανεξάρτητα δεδομένα, καθιστώντας το εργαλείο άμεσα επαναχρησιμοποιήσιμο και επεκτάσιμο. Τα αποτελέσματα παρουσιάζονται σε μορφή ακρίβειας για κάθε Holdout Split και συνοδεύονται από τις προβλέψεις κατηγοριοποίησης των νέων σχολίων.

Η ακρίβεια που επιτυγχάνεται σε κάθε Holdout Ratio λειτουργεί ως βασικό σημείο αναφοράς για τη σύγκριση με πιο εξελιγμένες μεθόδους ταξινόμησης, όπως τα μετασχηματιστικά μοντέλα τύπου BERT ή τα embeddings. Η απλότητα της μεθόδου αυτής και η αποδοτικότητά της σε μικρές και μεσαίες κλίμακες δεδομένων την καθιστούν κατάλληλη ως αφετηρία αξιολόγησης μοντέλων κατηγοριοποίησης σχολίων.

Η συνάρτηση simple_SC καλείται από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

```
[acc_SC] = simple_SC(T, new_reviews);
```

Και δίνει την παρακάτω έξοδο:

```
Ακρίβεια για Holdout Ratio 0.1: 0.62452

Προβλεπόμενες κατηγορίες για τα νέα σχόλια (Holdout 0.1):

1. Υψηλές Βαθμολογίες

2. Μέτριες Βαθμολογίες

3. Μέτριες Βαθμολογίες

Ακρίβεια για Holdout Ratio 0.2: 0.58562

Προβλεπόμενες κατηγορίες για τα νέα σχόλια (Holdout 0.2):

1. Μέτριες Βαθμολογίες

2. Μέτριες Βαθμολογίες

3. Μέτριες Βαθμολογίες
```

Ακρίβεια για Holdout Ratio 0.3: 0.58296

Προβλεπόμενες κατηγορίες για τα νέα σχόλια (Holdout 0.3):

1. Υψηλές Βαθμολογίες
2. Μέτριες Βαθμολογίες
3. Μέτριες Βαθμολογίες

Κατά την εκτέλεση της συνάρτησης για τα ποσοστά Holdout 10%, 20% και 30%, η ακρίβεια κυμάνθηκε αντίστοιχα στο 62.45%, 58.56% και 58.29%. Παρατηρείται ότι η ακρίβεια μειώνεται ελαφρώς καθώς αυξάνεται το ποσοστό του test set, κάτι που είναι αναμενόμενο καθώς μειώνονται τα δεδομένα εκπαίδευσης. Αξίζει να σημειωθεί ότι τα αποτελέσματα ενδέχεται να διαφέρουν ελαφρώς κάθε φορά που εκτελείται ο κώδικας, λόγω του τυχαίου διαχωρισμού των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, γεγονός που επηρεάζει τόσο τις προβλέψεις όσο και την ακρίβεια.

Ωστόσο, οι προβλέψεις για τα τρία νέα σχόλια που δοκιμάστηκαν δείχνουν περιορισμένη διαφοροποίηση, καθώς κυριαρχεί η κατηγορία «Μέτριες Βαθμολογίες», ακόμα και για σχόλια που εκ πρώτης όψεως φαίνονται θετικά ή αρνητικά. Αυτό πιθανόν οφείλεται στο ότι το BoW μοντέλο δεν ενσωματώνει σημασιολογική πληροφορία ή συντακτικό πλαίσιο, αλλά στηρίζεται σε μετρήσεις συχνότητας λέξεων.

Η κυριαρχία της μέσης κατηγορίας στις προβλέψεις ίσως αντανακλά τη σχετική ισορροπία του συνόλου δεδομένων, αλλά ταυτόχρονα υποδεικνύει και περιορισμό στην ικανότητα του ταξινομητή να διακρίνει ξεκάθαρα τις διαφορετικές συναισθηματικές αποχρώσεις των σχολίων. Το φαινόμενο αυτό ενισχύει την ανάγκη για πιο πλούσια μοντέλα ανάλυσης κειμένου, όπως τα word embeddings ή τα deep learning μοντέλα που ακολουθούν στα επόμενα στάδια της εργασίας.

3.5. Σύγκριση Ταξινομητών για χρήση σε Document Embeddings

Στην ενότητα αυτή παρουσιάζεται μια πιο ολοκληρωμένη προσέγγιση στην ταξινόμηση κριτικών με βάση την ανάλυση συναισθήματος, αξιοποιώντας τεχνικές ενσωμάτωσης κειμένου (document embeddings) και πληθώρα αλγορίθμων εκμάθησης για την δημιουργία ταξινομητών. Για την υλοποίηση της διαδικασίας δημιουργήθηκε η συνάρτηση **classification_with_embeddings**, η οποία περιλαμβάνει όλα τα απαραίτητα βήματα για την προετοιμασία των δεδομένων, την εκπαίδευση και αξιολόγηση των μοντέλων, καθώς και τη σύγκριση των αποτελεσμάτων. Η συγκεκριμένη υλοποίηση στοχεύει στην αξιολόγηση της αποτελεσματικότητας διαφορετικών μοντέλων μηχανικής μάθησης,

τόσο μέσω απλής διάσπασης των δεδομένων (Hold-Out), όσο και μέσω διασταυρούμενης επικύρωσης (K-Fold Cross-Validation).

Η συνάρτηση classification_with_embeddings καλείται από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

[holdoutAccuracies, crossvalAccuracies] = classification_with_embeddings(T);

Ως είσοδο δέχεται έναν πίνακα δεδομένων τύπου table (T) που περιέχει τις στήλες comment (σχόλια σε μορφή κειμένου) και rating (βαθμολογίες). Ως έξοδο επιστρέφει δύο διανύσματα: το **holdoutAccuracies**, το οποίο περιλαμβάνει την ακρίβεια κάθε ταξινομητή με τη μέθοδο Hold-Out, και το **crossvalAccuracies**, το οποίο περιλαμβάνει την ακρίβεια από την 5-Fold διασταυρούμενη επικύρωση. Οι τιμές αυτές χρησιμεύουν για τη σύγκριση της αποδοτικότητας των διαφορετικών ταξινομητών και την ανάδειξη του πιο αποτελεσματικού μοντέλου. Επίσης, η διαδικασία κατηγοριοποίησης των βαθμολογιών πραγματοποιείται όπως και στις προηγούμενες ενότητες.

Για τη μετατροπή του κειμένου σε μορφή κατάλληλη για εκπαίδευση μοντέλων, χρησιμοποιήθηκε το προεκπαιδευμένο μοντέλο "all-MiniLM-L6-v2" της πλατφόρμας Hugging Face, το οποίο αποτελεί μία ελαφριά και αποδοτική παραλλαγή του BERT. Το μοντέλο αυτό βασίζεται στη χρήση attention μηχανισμών και έχει εκπαιδευτεί σε μεγάλους όγκους δεδομένων ώστε να εξάγει συμπυκνωμένες αλλά πλούσιες σημασιολογικά αναπαραστάσεις των κειμένων. Η ενσωμάτωση κάθε κριτικής ως διανύσματος επιτρέπει στα μοντέλα ταξινόμησης να κατανοούν πιο βαθιές σχέσεις στο κείμενο και όχι απλώς λέξεις-κλειδιά.

Η αξιολόγηση των ταξινομητών πραγματοποιήθηκε με δύο τρόπους διαχωρισμού των δεδομένων:

1. Hold-Out Split: Τα δεδομένα χωρίστηκαν σε σύνολο εκπαίδευσης (80%) και δοκιμής (20%). Εκπαιδεύτηκαν επτά διαφορετικοί ταξινομητές:

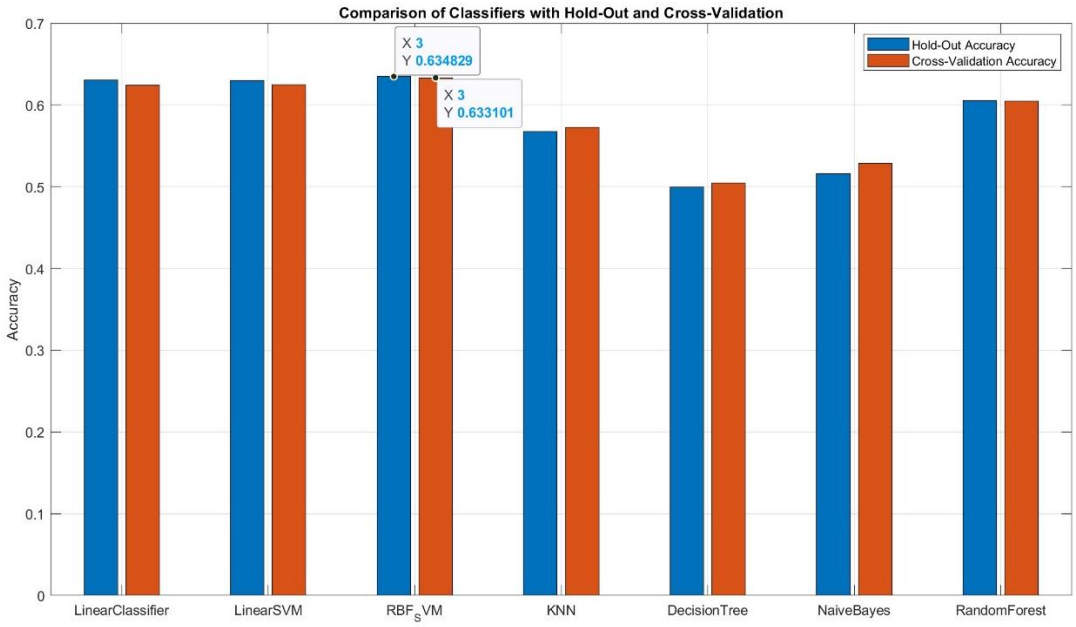
- Linear Classifier
- Linear SVM
- RBF SVM
- K-Nearest Neighbors (KNN)
- Decision Tree

- Naive Bayes
- Random Forest

Για κάθε ταξινομητή υπολογίστηκε η ακρίβεια στο test set, παρέχοντας μια πρώτη εκτίμηση της γενίκευσης του κάθε μοντέλου.

2. Cross-Validation (5-Fold): Η ίδια ομάδα ταξινομητών αξιολογήθηκε μέσω 5-πλής διασταυρούμενης επικύρωσης. Τα δεδομένα χωρίστηκαν σε 5 υποσύνολα, και σε κάθε επανάληψη το μοντέλο εκπαιδευόταν στα 4 και δοκιμαζόταν στο 1. Ο μέσος όρος της ακρίβειας από τα 5 folds δίνει μια πιο στατιστικά έγκυρη εκτίμηση της απόδοσης.

Τα αποτελέσματα παρουσιάστηκαν με τη μορφή γραφήματος (bar chart) – (Εικόνα 1.14), συγκρίνοντας την ακρίβεια του κάθε μοντέλου και για τις δύο μεθόδους αξιολόγησης.



Εικόνα 1.14. Ακρίβεια Ταξινομητών με Document Embeddings

Έξοδος:

Καλύτερη ακρίβεια στο Hold-Out: 0.63483 από τον ταξινομητή RBF_SVM

Καλύτερη ακρίβεια στο Cross-Validation: 0.6331 από τον ταξινομητή RBF_SVM

Όπως φαίνεται στο παραπάνω γράφημα, οι διαφορές μεταξύ των ταξινομητών είναι σχετικά μικρές, ωστόσο διακρίνεται μια σταθερά καλύτερη απόδοση του μοντέλου **RBF SVM** σε σχέση με τους

υπόλοιπους, και στις δύο μεθόδους αξιολόγησης. Συγκεκριμένα, σημειώνει **ακρίβεια 63,48% με Hold-Out split** και **63,31% με Cross-Validation**, τιμές που το καθιστούν ως την πιο αξιόπιστη επιλογή ανάμεσα στους επτά ταξινομητές που εξετάστηκαν. Η σχεδόν ταυτόσημη απόδοση του RBF SVM στις δύο μεθόδους αξιολόγησης (Hod-Out και Cross-Validation) υποδηλώνει καλή ικανότητα γενίκευσης και σταθερότητα του μοντέλου απέναντι σε διαφορετικές κατανομές των δεδομένων. Αντίθετα, πιο απλές μέθοδοι όπως ο Naive Bayes, και τα Δέντρα Απόφασης εμφανίζουν αισθητά χαμηλότερες επιδόσεις, κάτι που αποκαλύπτει την αδυναμία τους να αξιοποιήσουν αποτελεσματικά τις πλούσιες σημασιολογικές αναπαραστάσεις του κειμένου που προκύπτουν από τα document embeddings.

Η μέθοδος Random Forest διακρίνεται για τη σχετική σταθερότητα που παρουσιάζει, χωρίς ωστόσο να καταφέρνει να ξεπεράσει την απόδοση των μοντέλων SVM. Αυτή η σύγκριση υπογραμμίζει ότι, παρόλο που τα embeddings παρέχουν ένα ισχυρό υπόβαθρο αναπαράστασης, η επιλογή του κατάλληλου αλγορίθμου εκμάθησης παραμένει κρίσιμη για την επίτευξη υψηλής ακρίβειας. Επίσης, αξίζει να σημειωθεί και εδώ, ότι τα αποτελέσματα ενδέχεται να διαφέρουν ελαφρώς κάθε φορά που εκτελείται ο κώδικας, λόγω του τυχαίου διαχωρισμού των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, γεγονός που επηρεάζει τόσο τις προβλέψεις όσο και την ακρίβεια.

Συνολικά, τα αποτελέσματα αναδεικνύουν τη σημασία όχι μόνο της επιλογής μοντέλου, αλλά και της μεθόδου αξιολόγησης που χρησιμοποιείται. Παράλληλα, επιβεβαιώνουν τη χρησιμότητα των document embeddings ως ενδιάμεση αναπαράσταση, καθώς επιτρέπουν ακόμα και σε παραδοσιακά μοντέλα να επιτυγχάνουν ικανοποιητικές επιδόσεις. Η παρούσα υλοποίηση προσφέρει μια σφαιρική αποτίμηση των επιδόσεων διαφορετικών αλγορίθμων εκμάθησης σε πρόβλημα ταξινόμησης συναισθήματος, καταδεικνύοντας τη σημασία μιας πολυεπίπεδης προσέγγισης στην επιλογή της τελικής μεθοδολογίας.

3.6. Ταξινόμηση με Embeddings Εγγράφων και Οπτικοποίηση με PCA/t-SNE

Στην παρούσα ενότητα περιγράφεται η διαδικασία ταξινόμησης σχολίων μέσω document embeddings και της χρήσης γραμμικών και μη γραμμικών μοντέλων μηχανικής μάθησης. Η υλοποίηση πραγματοποιήθηκε με τη δημιουργία της συνάρτησης **doc_embeddings**, η οποία αναλαμβάνει να μετατρέψει τα κείμενα σε διανύσματα, να εκπαιδεύσει ταξινομητές, να αξιολογήσει την απόδοσή τους και να απεικονίσει τη δομή των δεδομένων μέσω τεχνικών μείωσης διαστάσεων.

Η συνάρτηση doc_embeddings καλείται από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

```
[acc_DE_linear, acc_DE_rbf, categoryCounts] = doc_embeddings(T);
```

Δέχεται ως είσοδο έναν πίνακα δεδομένων T, που περιέχει τις στήλες comment (σχόλια σε μορφή κειμένου) και rating (αριθμητικές βαθμολογίες). Η έξοδος της συνάρτησης περιλαμβάνει:

- acc_DE_linear: η ακρίβεια του γραμμικού μοντέλου (Linear SVM),
- acc_DE_rbf: η ακρίβεια του μη γραμμικού μοντέλου (RBF SVM),
- categoryCounts: η συχνότητα κάθε κατηγορίας σχολίων.

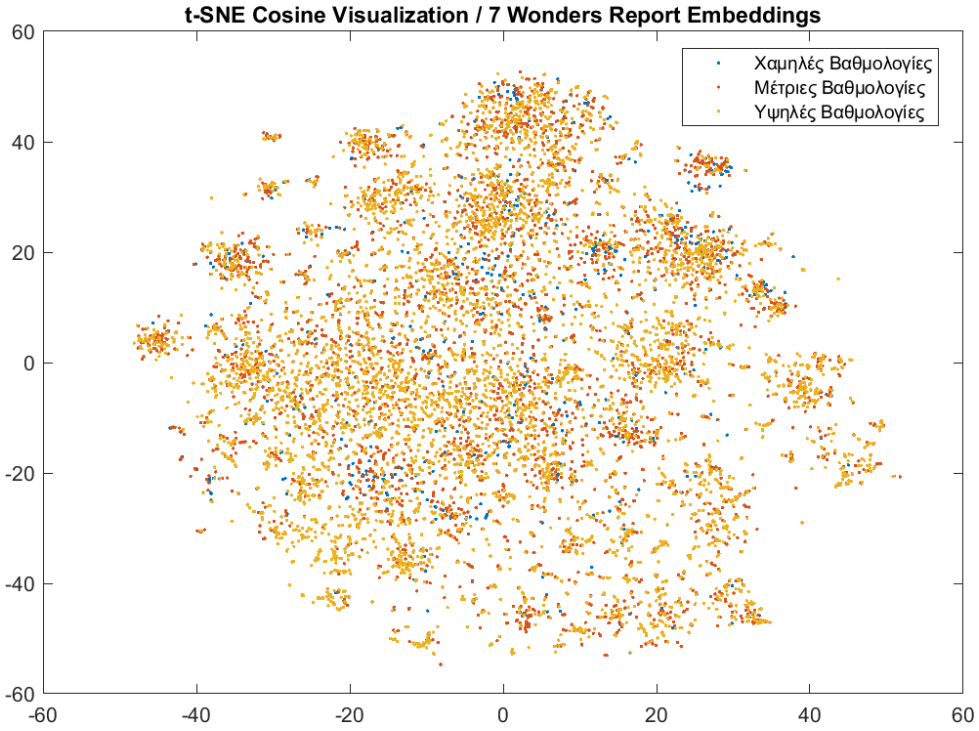
Η κατηγοριοποίηση των σχολίων πραγματοποιείται όπως στις προηγούμενες ενότητες, χωρίζοντας τα σχόλια σε χαμηλές, μέτριες και υψηλές βαθμολογίες.

Για τη μετατροπή των σχολίων σε ενσωματωμένα διανύσματα, χρησιμοποιείται το προεκπαιδευμένο μοντέλο all-MiniLM-L6-v2, που χρησιμοποιήθηκε και στην προηγούμενη ενότητα. Είναι ένα αποδοτικό και ελαφρύ μοντέλο βασισμένο στη δομή transformer, εκπαιδευμένο σε ευρύ φάσμα δεδομένων και βελτιστοποιημένο για ταχύτητα και ποιότητα σημασιολογικής αναπαράστασης. Τα ενσωματωμένα αυτά διανύσματα αποτελούν τις εισόδους για τους ταξινομητές.

Ακολουθεί ο διαχωρισμός των δεδομένων σε σύνολο εκπαίδευσης (80%) και δοκιμής (20%) με χρήση Hold-Out Split. Για την καλύτερη κατανόηση της κατανομής των ενσωματωμένων δεδομένων, εφαρμόζονται δύο μέθοδοι μείωσης διαστάσεων:

- **Principal Component Analysis (PCA):** Μειώνει τη διάσταση των ενσωματώσεων κρατώντας τις 50 σημαντικότερες συνιστώσες.
- **t-distributed Stochastic Neighbor Embedding (t-SNE):** Εφαρμόζεται στα αποτελέσματα του PCA για απεικόνιση σε δύο διαστάσεις, με σκοπό την οπτικοποίηση της δομής των σχολίων ανά κατηγορία.

Για τον υπολογισμό των αποστάσεων στο t-SNE χρησιμοποιήθηκε η συνάρτηση cosine, καθώς είναι πιο κατάλληλη για δεδομένα υψηλής, εστιάζοντας στη γωνιακή σχέση μεταξύ των διανυσμάτων παρά στο ευκλείδειο μήκος τους.



Εικόνα 1.15. Scatter Plot των Κατηγοριών Βαθμολογιών μέσω t-SNE

Το παραπάνω διάγραμμα (Εικόνα 1.15) απεικονίζει τις ενσωματώσεις των κριτικών (document embeddings) σε δύο διαστάσεις, χρησιμοποιώντας t-SNE μετά από PCA μείωση διαστάσεων. Παρατηρείται ότι, παρόλο που υπάρχει γενική διασπορά και κάποια αλληλοεπικάλυψη μεταξύ των κατηγοριών, οι υψηλές βαθμολογίες (κίτρινο) τείνουν να συγκεντρώνονται σε διακριτές περιοχές, ενώ οι μέτριες και χαμηλές βαθμολογίες (κόκκινο και μπλε αντίστοιχα) εμφανίζουν μεγαλύτερη διασπορά. Αυτό δείχνει ότι οι σημασιολογικές διαφορές μεταξύ των κριτικών υψηλής βαθμολογίας είναι πιο έντονες, γεγονός που εξηγεί και τη βελτιωμένη απόδοση του ταξινομητή στην κατηγορία αυτή.

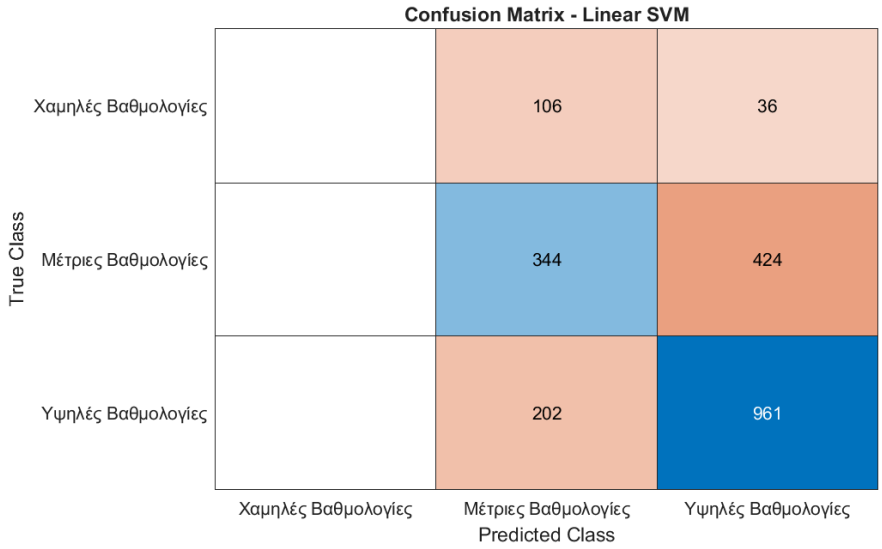
Έπειτα, εκπαιδεύεται ένας **γραμμικός ταξινομητής (Linear SVM)**, κατάλληλος για γραμμικά διαχωρίσιμα δεδομένα, και ένας **SVM με πυρήνα RBF**. Αυτός είναι ο πιο αποδοτικός που επιλέχθηκε από την εφαρμογή του κώδικα της προηγούμενης ενότητας, ο οποίος μπορεί να συλλάβει πιο πολύπλοκα μοτίβα και μη γραμμικές σχέσεις.

Η απόδοση κάθε μοντέλου αξιολογείται με χρήση των μέτρων **ακρίβεια** (accuracy), **ευστοχία** (precision), **ανάκληση** (recall), **F1 Score**, και **Πίνακας Σύγχυσης** (Confusion Matrix) για τον κάθε ταξινομητή.

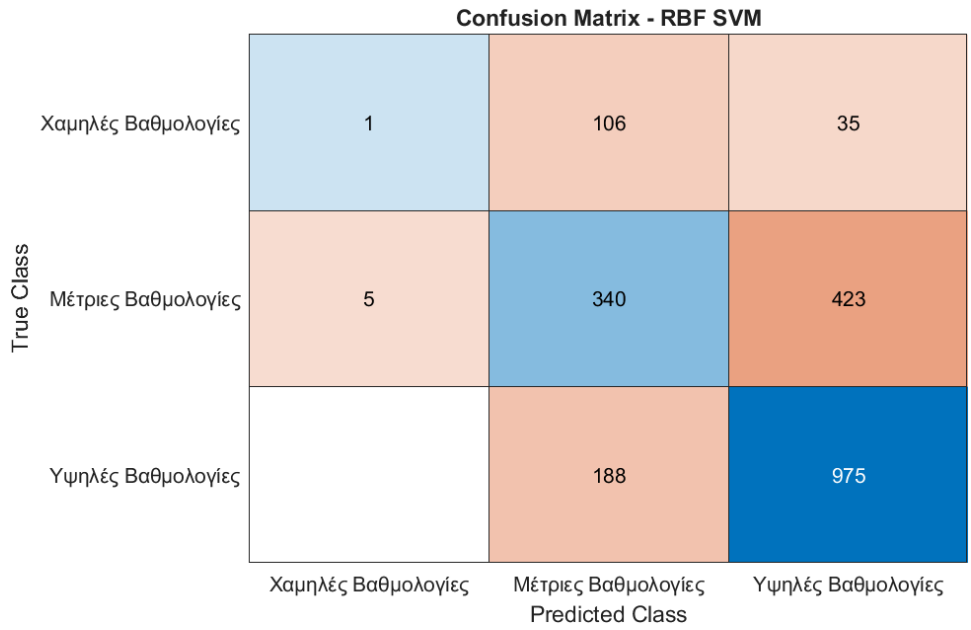
Συγκεκριμένα, η **ακρίβεια του Linear SVM υπολογίστηκε 0.62952**, ενώ **του RBF SVM είναι ελαφρώς υψηλότερη, στο 0.63483**. Όλα τα μέτρα απόδοσης φαίνονται στον παρακάτω πίνακα:

Πίνακας 1.4. Μέτρα Απόδοσης Ταξινομητών Linear και RBF

	Accuracy	***Precision***	***Recall***	***F1 Score***
Linear SVM	0.63	0.40	0.42	0.41
RBF SVM	0.64	0.46	0.43	0.44



Εικόνα 1.16. Πίνακας Σύγχυσης του γραμμικού ταξινομητή



Εικόνα 1.17. Πίνακας Σύγχυσης του ταξινομητή RBF

Εύκολα παρατηρείται ο RBF ταξινομητής έχει καλύτερη απόδοση, καθώς όσο πιο κοντά στο 1 βρίσκονται τα μέτρα απόδοσης, τόσο καλύτερη θεωρείται η απόδοση του ταξινομητή ή μοντέλου. Οι αντίστοιχες confusion matrices, (Εικόνα 1.16) & (Εικόνα 1.17), αποκαλύπτουν ότι ο RBF ταξινομητής τείνει να είναι πιο προσεκτικός στις προβλέψεις για τη χαμηλή κατηγορία, με λιγότερες εσφαλμένες ταξινομήσεις σε σχέση με το γραμμικό μοντέλο. Παρά το γεγονός ότι και τα δύο μοντέλα δυσκολεύονται στη διάκριση των μέτριων βαθμολογιών, η συνολική συμπεριφορά του RBF SVM δείχνει ότι μπορεί να αποδώσει καλύτερα όταν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα. Για ακόμη μία φορά, είναι σημαντικό να σημειωθεί ότι τα αποτελέσματα ενδέχεται να διαφέρουν ελαφρώς κάθε φορά που εκτελείται ο κώδικας, λόγω του τυχαίου διαχωρισμού των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, γεγονός που επηρεάζει τόσο τις προβλέψεις όσο και τα μέτρα απόδοσης.

Η συγκεκριμένη υλοποίηση αποδεικνύει πως η χρήση embeddings σε συνδυασμό με τεχνικές οπτικοποίησης και διαφορετικούς ταξινομητές προσφέρει μια πλήρη εικόνα της συμπεριφοράς των μοντέλων. Οι t-SNE απεικονίσεις βοηθούν στην κατανόηση της πυκνότητας και αλληλοεπικάλυψης των κατηγοριών, ενώ τα μέτρα απόδοσης επιτρέπουν την αντικειμενική σύγκριση των μεθόδων. Έτσι, η ενότητα αυτή επιβεβαιώνει τη δύναμη των ενσωματώσεων κειμένου (document embeddings) και την ανάγκη κατάλληλης επιλογής μοντέλου ανάλογα με τη φύση του προβλήματος.

3.7. Επιλογή Βέλτιστου Solver για LDA

Για την εξαγωγή θεμάτων από τις κριτικές των χρηστών, είναι ιδιαίτερα σημαντική η επιλογή του κατάλληλου αλγορίθμου επίλυσης (solver) στο πλαίσιο του μοντέλου Latent Dirichlet Allocation (LDA). Η παρούσα ενότητα έχει ως στόχο τη διερεύνηση της απόδοσης διαφορετικών μεθόδων εκπαίδευσης του LDA, ώστε να προσδιοριστεί ποια αποδίδει καλύτερα στο συγκεκριμένο σύνολο δεδομένων. Για την υλοποίηση της διαδικασίας δημιουργήθηκε και χρησιμοποιήθηκε η συνάρτηση lda_solvers_comp, η οποία λαμβάνει ως είσοδο το σύνολο σχολίων των χρηστών (T.comment) και εκτελεί μια σειρά από στάδια αξιολόγησης.

Αρχικά, τα δεδομένα χωρίζονται με τη μέθοδο Hold-Out σε δύο υποσύνολα: ένα σύνολο εκπαίδευσης που χρησιμοποιείται για την εκμάθηση των θεμάτων και ένα σύνολο επικύρωσης για την εκτίμηση της γενικευσιμότητας του μοντέλου. Πριν την εκπαίδευση, εφαρμόζεται προεπεξεργασία κειμένου όπως και στους προηγούμενους κώδικες.

Στη συνέχεια, εκπαιδεύονται τέσσερα διαφορετικά μοντέλα LDA, χρησιμοποιώντας τους εξής solvers:

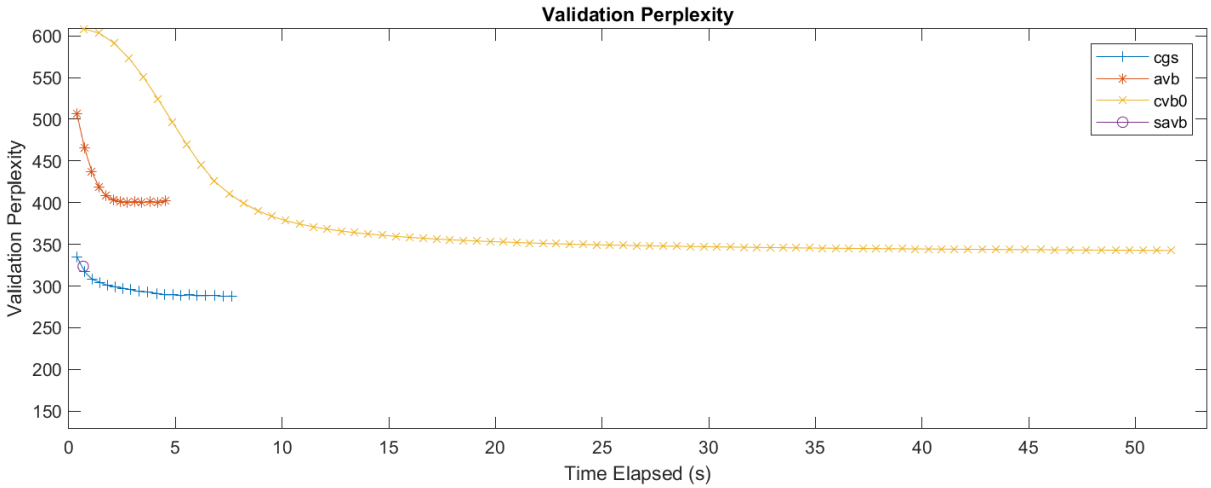
Ο **Collapsed Gibbs Sampling (cgs)** βασίζεται σε τεχνικές δειγματοληψίας, είναι υπολογιστικά πιο απαιτητικός, αλλά παράγει υψηλής ποιότητας αποτελέσματα. Ο **Approximate Variational Bayes (avb)** εφαρμόζει μια παραμετρική προσέγγιση, προσφέροντας ταχύτερη εκπαίδευση με ενδεχόμενη απώλεια ακρίβειας. Ο **Collapsed Variational Bayes 0 (cvb0)** είναι μια πιο γρήγορη και σταθερή παραλλαγή της μεθόδου VB, συχνά κατάλληλη για μεγάλα datasets. Τέλος, ο **Stochastic Approximate Variational Bayes (savb)** χρησιμοποιεί στοχαστική προσέγγιση και είναι προσαρμοσμένος σε σενάρια online μάθησης και επεξεργασίας πολύ μεγάλων συνόλων δεδομένων.

Για την αξιολόγηση των παραπάνω μοντέλων, αξιοποιούνται δύο μέτρα απόδοσης. Το πρώτο είναι η **Περιπλοκή** (**Perplexity)**, η οποία μετρά την ικανότητα του μοντέλου να προβλέπει νέα δεδομένα, με τις χαμηλότερες τιμές υποδηλώνουν καλύτερη γενίκευση. Το δεύτερο μέτρο είναι η **Λογαριθμική Πιθανότητα (Log-Probability)**, η οποία εκφράζει το πόσο πιθανό είναι να έχουν παραχθεί τα δεδομένα επικύρωσης από το μοντέλο, με τις υψηλότερες τιμές να υποδηλώνουν καλύτερη προσαρμογή του μοντέλου στα δεδομένα.

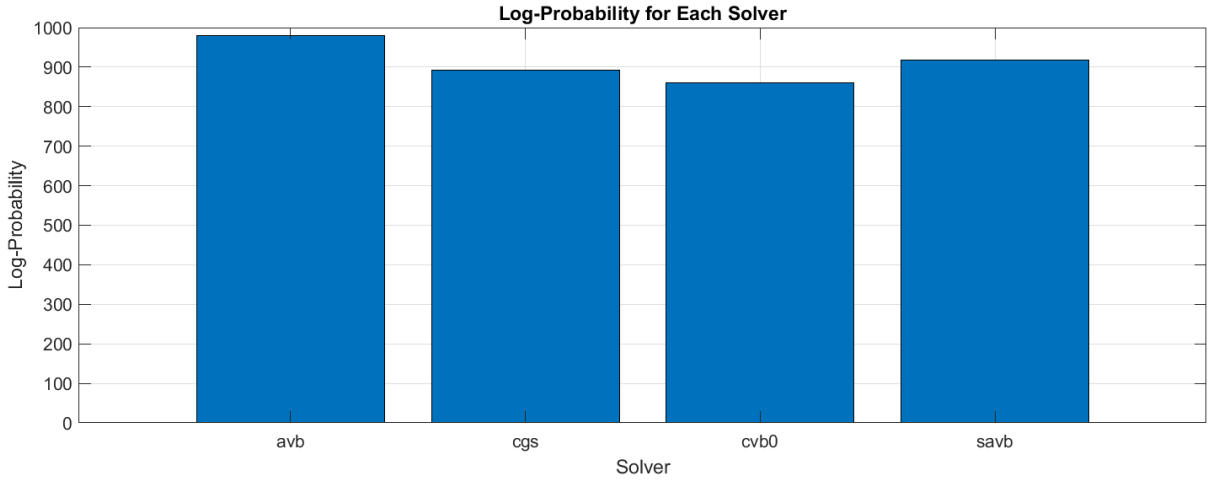
Η συνάρτηση καλείται από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

lda_solvers_comp(T.comment);

Η συνάρτηση lda_solvers_comp παράγει δύο γραφήματα. Στο πρώτο παρουσιάζεται η εξέλιξη της perplexity σε συνάρτηση με τον χρόνο εκπαίδευσης για κάθε έναν από τους solvers, προσφέροντας οπτική πληροφόρηση για την ταχύτητα σύγκλισης και την απόδοση κάθε μεθόδου. Στο δεύτερο γράφημα απεικονίζονται οι τιμές της log-probability ανά solver σε μορφή ραβδογράμματος – (Εικόνα 1.19), διευκολύνοντας την άμεση σύγκριση μεταξύ τους.



Εικόνα 1.18. Εξέλιξη της Validation-Perplexity κατά τη διάρκεια της εκπαίδευσης για κάθε Solver του LDA



Εικόνα 1.19. Μέση Log-Probability ανά Solver του LDA - Μέτρο απόδοσης στο validation set

Παρατηρώντας τα αποτελέσματα των δύο διαγραμμάτων, διαπιστώνεται ότι δεν υπάρχει ένας απόλυτα κυρίαρχος solver. Ο **AVB** εμφανίζει τη μεγαλύτερη τιμή στη λογαριθμική πιθανότητα, γεγονός που υποδηλώνει πολύ καλή προσαρμογή του μοντέλου στα δεδομένα επικύρωσης. Ωστόσο, η τιμή της perplexity για τον AVB παραμένει σχετικά υψηλή. Αντίθετα, ο **CGS** παρουσιάζει τη χαμηλότερη perplexity, γεγονός που υποδηλώνει καλύτερη γενίκευση, και παράλληλα διατηρεί υψηλή log-probability, χωρίς όμως να φτάνει το επίπεδο του AVB.

Ο **SAVB**, αν και δεν κυριαρχεί απόλυτα σε καμία από τα δύο μέτρα απόδοσης, διατηρεί μια **ισορροπία** με **ικανοποιητικά υψηλή log-probability και συγκριτικά χαμηλότερη perplexity** από τον AVB. Επιπλέον, η σταθερότητα και η ταχύτητα εκπαίδευσής του τον καθιστούν ιδανική επιλογή για μεγάλα σύνολα δεδομένων και σενάρια online μάθησης.

Καθώς λοιπόν η υπεροχή κάποιου solver δεν είναι απολύτως ξεκάθαρη, η τελική επιλογή βασίστηκε στη συνολική ποιότητα και συμπεριφορά του μοντέλου. Ως βέλτιστος επιλέχθηκε ο **SAVB**, διότι προσφέρει έναν καλό συμβιβασμό μεταξύ των δύο βασικών μέτρων απόδοσης, ενώ παρουσιάζει και σταθερή και ταχύτατη εκπαίδευση στο διάγραμμα perplexity. Ωστόσο, αξίζει να σημειωθεί ότι κάποιος άλλος χρήστης θα μπορούσε επίσης να δικαιολογήσει την επιλογή του CGS ως βέλτιστου solver, δεδομένων των επιδόσεών του.

Η συγκριτική αξιολόγηση των solvers είναι κρίσιμη, καθώς επηρεάζει τόσο την ποιότητα όσο και την αξιοπιστία των εξαγόμενων θεμάτων. Η επιλογή του κατάλληλου solver αποτελεί ένα ουσιώδες βήμα για την επιτυχή εφαρμογή του LDA, ειδικά σε περιβάλλοντα πραγματικού κόσμου όπου απαιτείται ισορροπία μεταξύ ακρίβειας και υπολογιστικού κόστους.

3.8. Επιλογή Βέλτιστου Αριθμού Θεμάτων για LDA

Η εύρεση του κατάλληλου αριθμού θεμάτων (topics) αποτελεί θεμελιώδες βήμα στην εφαρμογή του μοντέλου Latent Dirichlet Allocation (LDA), καθώς επηρεάζει άμεσα τόσο τη νοηματική συνοχή των θεμάτων όσο και το πόσο εύκολα μπορούν να αναλυθούν και να ερμηνευθούν τα αποτελέσματα. Σε συνέχεια της προηγούμενης ενότητας, το μοντέλο LDA εκπαιδεύεται με τη χρήση της συνάρτησης fitlda και του solver SAVB, ο οποίος επιλέχθηκε ως ο πλέον κατάλληλος ύστερα από συγκριτική αξιολόγηση.

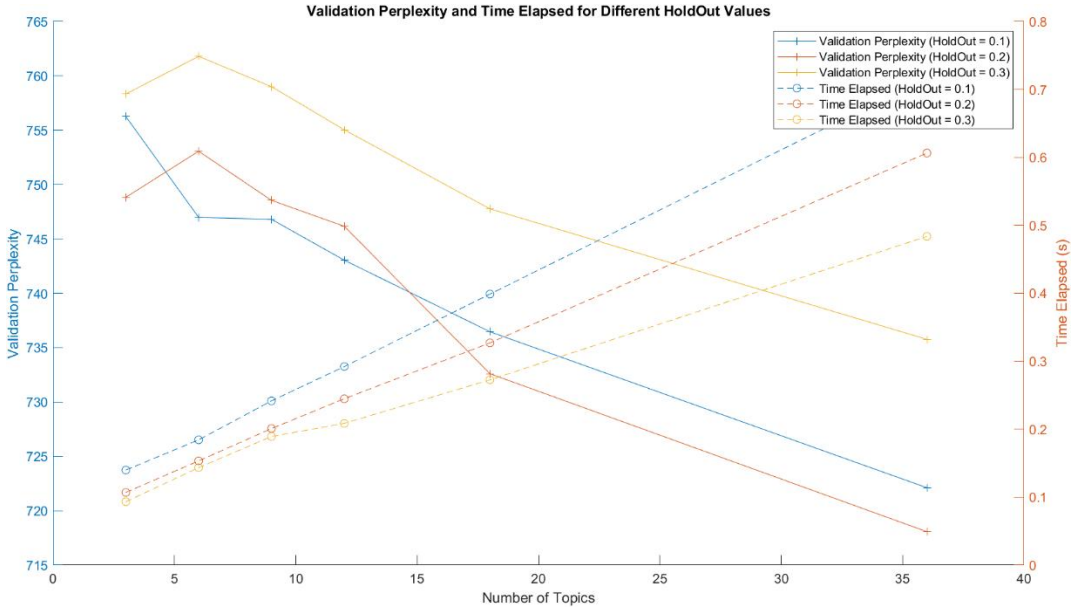
Στην παρούσα ενότητα αξιοποιούνται δύο ανεξάρτητες μεθοδολογικές προσεγγίσεις για την επιλογή του αριθμού θεμάτων: η μέθοδος Hold-Out και η μέθοδος k-fold Cross-Validation. Για την υλοποίηση των δύο προσεγγίσεων αναπτύχθηκαν οι συναρτήσεις nof_topics_lda.mlx και nof_topics_lda_cv.mlx, αντίστοιχα. Κάθε μία από αυτές αξιολογεί την απόδοση του μοντέλου LDA για διαφορετικούς αριθμούς θεμάτων με βάση δύο βασικές μεταβλητές: την perplexity, που εκφράζει την ικανότητα του μοντέλου να γενικεύει σε άγνωστα δεδομένα, και τον χρόνο εκπαίδευσης, που αντικατοπτρίζει το υπολογιστικό κόστος.

Οι συναρτήσεις αυτές καλούνται από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

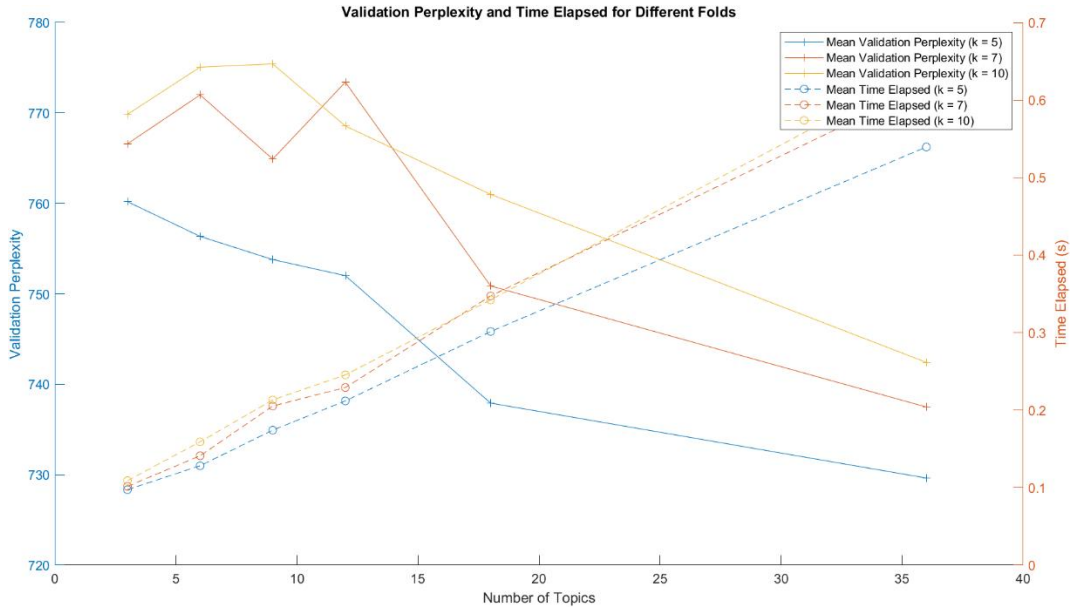
```
[results] = nof_topics_lda(T.comment);

[results_cv] = nof_topics_lda_cv(T.comment);
```

Η πρώτη προσέγγιση βασίζεται στην τεχνική Hold-Out, όπου το σύνολο των κειμένων διαχωρίζεται σε σύνολα εκπαίδευσης και επικύρωσης με επιλεγμένα ποσοστά 10%, 20% και 30%. Για κάθε μια από αυτές τις περιπτώσεις, το μοντέλο εκπαιδεύεται με διαφορετικούς αριθμούς θεμάτων (3, 6, 9, 12, 18 και 36). Για κάθε εκπαιδευόμενο μοντέλο, υπολογίζεται η perplexity, ένα μέτρο που αντανακλά την ικανότητα του μοντέλου να προβλέπει νέα, «αόρατα» δεδομένα. Χαμηλότερες τιμές perplexity σημαίνουν καλύτερη γενίκευση. Ταυτόχρονα, υπολογίζεται και ο χρόνος εκπαίδευσης, καθώς σε εφαρμογές με μεγάλο όγκο δεδομένων είναι σημαντική η υπολογιστική αποδοτικότητα.

Η δεύτερη προσέγγιση βασίζεται στο k-fold Cross-Validation, εφαρμόζοντας την ίδια λογική αξιολόγησης σε σύνολα δεδομένων που έχουν κατανεμηθεί σε 5, 7 και 10 folds. Η τεχνική αυτή είναι πιο ανθεκτική σε παραμορφώσεις λόγω τυχαίας κατανομής των δεδομένων και προσφέρει πιο σταθερές εκτιμήσεις, καθώς κάθε σημείο του συνόλου χρησιμοποιείται και για εκπαίδευση και για επικύρωση. Για κάθε fold και αριθμό θεμάτων, καταγράφονται οι αντίστοιχες τιμές perplexity και χρόνος εκπαίδευσης, και στο τέλος υπολογίζονται οι μέσες τιμές ώστε να καταστούν συγκρίσιμες μεταξύ τους.



Εικόνα 1.20. Απόδοση μοντέλου LDA σε όρους perplexity και χρόνου εκτέλεσης για διαφορετικούς αριθμούς θεμάτων, με την χρήση Hold-Out



Εικόνα 1.21. Απόδοση μοντέλου LDA σε όρους perplexity και χρόνου εκτέλεσης για διαφορετικούς αριθμούς θεμάτων, με την χρήση Cross-Validation

Οι δύο αυτές προσεγγίσεις παράγουν διαγράμματα που απεικονίζουν την εξέλιξη της perplexity (στον αριστερό άξονα) και του χρόνου εκτέλεσης (στον δεξιό άξονα) σε συνάρτηση με τον αριθμό των θεμάτων. Από την ανάλυση των γραφημάτων, (Εικόνα 1.20) & (Εικόνα 1.21), παρατηρείται ότι με την αύξηση του αριθμού των θεμάτων η perplexity αρχικά μειώνεται, γεγονός που δείχνει ότι το μοντέλο

επιτυγχάνει καλύτερη θεματική διαφοροποίηση. Ωστόσο, μετά από ένα σημείο η βελτίωση σταματά ή αντιστρέφεται, ενώ ο χρόνος εκπαίδευσης αυξάνεται απότομα.

Η ανάλυση υποδεικνύει ότι **δεν είναι αποδοτική η χρήση υπερβολικά μεγάλου αριθμού θεμάτων**, καθώς αυξάνει σημαντικά το υπολογιστικό κόστος χωρίς να οδηγεί απαραίτητα σε βελτίωση της ακρίβειας ή της νοηματικής συνοχής. Αυτή η παρατήρηση είναι σημαντική, ειδικότερα σε προβλήματα όπου ο όγκος δεδομένων είναι τεράστιος. Αντιθέτως, η επιλογή ενός αριθμού θεμάτων που ισορροπεί μεταξύ της απόδοσης και της πολυπλοκότητας του μοντέλου είναι η βέλτιστη στρατηγική.

Με βάση τα παραπάνω ευρήματα, **η επιλογή των έξι ή εννέα θεμάτων κρίνεται ως η πλέον κατάλληλη για την ανάλυση**, καθώς προσφέρει έναν ικανοποιητικό συμβιβασμό μεταξύ ακρίβειας και πολυπλοκότητας. Συγκεκριμένα, για αυτούς τους αριθμούς θεμάτων παρατηρούνται σχετικά χαμηλές τιμές perplexity, γεγονός που υποδηλώνει καλή ικανότητα του μοντέλου να γενικεύει σε νέα δεδομένα. Παράλληλα, ο χρόνος εκπαίδευσης παραμένει σε αποδεκτά επίπεδα, χωρίς να επιβαρύνει σημαντικά την υπολογιστική διαδικασία.

Επιπλέον, τα παραγόμενα θέματα εμφανίζουν θεματική συνοχή και διαφοροποίηση, καθιστώντας ευκολότερη την κατανόηση, την ερμηνεία και τη χρηστική αξιοποίησή τους σε επόμενο στάδιο της ανάλυσης. Η τελική **απόφαση για τον αριθμό θεμάτων** βασίστηκε **όχι μόνο στις ποσοτικές μετρήσεις**, **αλλά και σε ποιοτικά χαρακτηριστικά** όπως η σαφήνεια των θεματικών ομάδων και η συνάφειά τους με τις έννοιες που αποτυπώνονται στα κείμενα. Έτσι, επιλέγεται ένας αριθμός θεμάτων που δεν είναι ούτε υπερβολικά περιορισμένος — ώστε να αποφεύγεται η απώλεια πληροφορίας — ούτε υπερβολικά μεγάλος, γεγονός που θα οδηγούσε σε περιττό κατακερματισμό της πληροφορίας και μείωση της συνοχής.

Συνεκτιμώντας όλα τα παραπάνω, αποφασίστηκε η χρήση **έξι θεμάτων** στο τελικό μοντέλο LDA. Η επιλογή αυτή συνδυάζει σαφήνεια, θεματική διαφοροποίηση και ικανοποιητική απόδοση τόσο σε ποσοτικό όσο και σε ποιοτικό επίπεδο, ενώ παράλληλα εξυπηρετεί την ανάγκη για ευανάγνωστη και πρακτικά αξιοποιήσιμη ερμηνεία των αποτελεσμάτων στην επόμενη φάση της ανάλυσης.

3.9. Ανάλυση και Οπτικοποίηση των Θεμάτων με LDA

Σε αυτή την ενότητα παρουσιάζεται η πλήρης διαδικασία ανάλυσης θεμάτων μέσω της τεχνικής Latent Dirichlet Allocation (LDA), με βάση τον αριθμό θεμάτων που επιλέχθηκε στην προηγούμενη ενότητα. Όπως τεκμηριώθηκε, επιλέχθηκαν έξι θέματα, καθώς προσφέρουν ικανοποιητική θεματική διαφοροποίηση, μειωμένες τιμές perplexity και χαμηλό υπολογιστικό κόστος. Η ανάλυση

πραγματοποιήθηκε με τη χρήση της συνάρτησης fitlda σε ένα σύνολο προεπεξεργασμένων σχολίων χρηστών, με στόχο την αναγνώριση θεματικών μοτίβων και την κατανόηση των κυρίαρχων εννοιών που αποτυπώνονται στο κείμενο. Για την υλοποίηση της διαδικασίας αυτής, δημιουργήθηκε ο κώδικας my_lda ο οποίος παίρνει ως εισόδους, από τον πίνακα δεδομένων (T) την στήλη με τα σχόλια, και ένα διάνυσμα (new_reviews) με τρία νέα σχόλια όπως έγινε και σε προηγούμενους κώδικες. Τα αποτελέσματα-έξοδος του κώδικα αναλύονται στην συνέχεια.

Ο κώδικας αυτός καλείται ως συνάρτηση από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

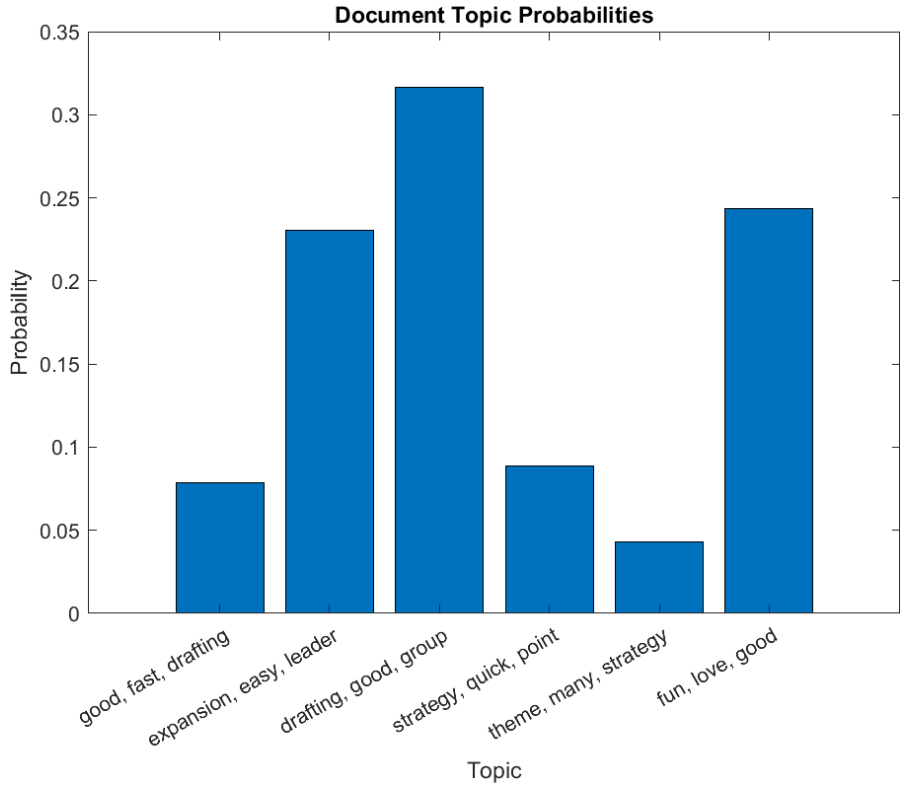
```
[tbl, top] = my_lda(T.comment, new_reviews);
```

Η διαδικασία ξεκινά με την προεπεξεργασία των δεδομένων κειμένου, μέσω της συνάρτησης text_preprocessing, η οποία αφαιρεί τα stop words, εφαρμόζει stemming και μετατρέπει το κείμενο σε tokens. Στη συνέχεια, δημιουργείται ένα μοντέλο bag-of-words από τα καθαρισμένα έγγραφα. Λέξεις που εμφανίζονται σπάνια αφαιρούνται ώστε να μειωθεί η διάσταση του προβλήματος και να αυξηθεί η σταθερότητα του μοντέλου. Έγγραφα χωρίς περιεχόμενο απορρίπτονται.

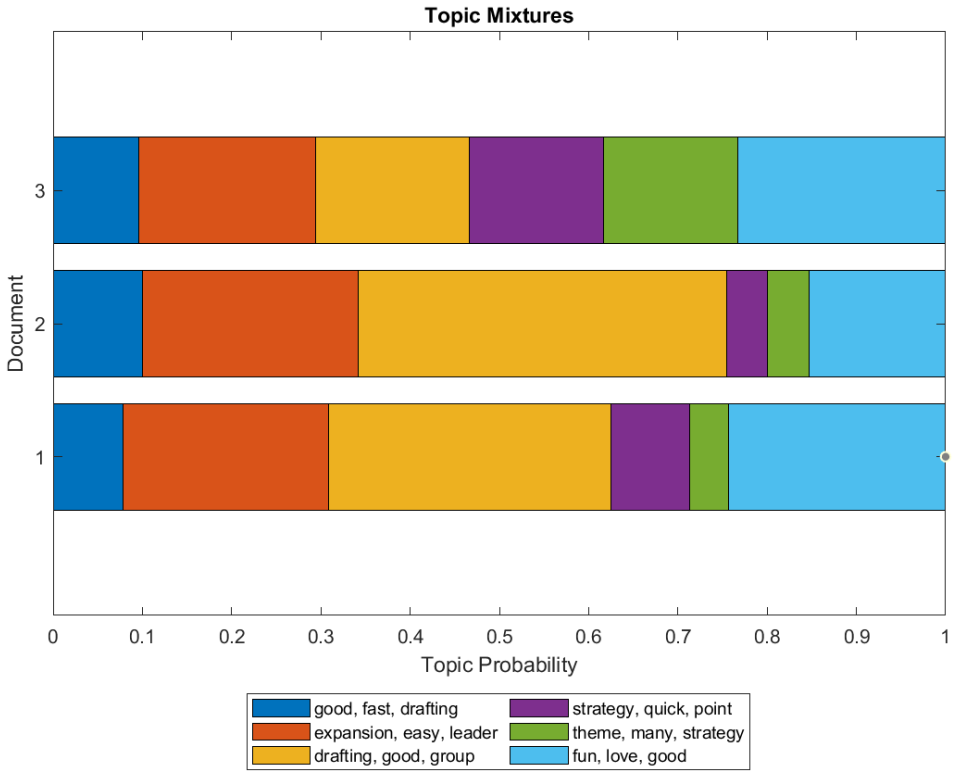
Έπειτα, εφαρμόζεται η συνάρτηση fitlda για την εκπαίδευση του μοντέλου LDA με έξι θέματα, χρησιμοποιώντας ως solver τον SAVB. Το μοντέλο που προκύπτει επιτρέπει την εξαγωγή θεματικών ομάδων, οι οποίες αναπαρίστανται με τις πιο χαρακτηριστικές λέξεις ανά θέμα. Η οπτικοποίηση των λέξεων πραγματοποιείται με τη χρήση διαγραμμάτων τύπου wordcloud (Εικόνα 1.22), όπου η σχετική συχνότητα εμφάνισης των λέξεων αποδίδεται γραφικά μέσω του μεγέθους τους. Η τεχνική αυτή προσφέρει άμεση κατανόηση της λέξη-κεντρικής πληροφορίας για κάθε θέμα. Η παρατήρηση των wordclouds δείχνει θεματική συνοχή και διαφοροποίηση, καθώς κάθε θέμα χαρακτηρίζεται από διαφορετικό λεξιλόγιο, γεγονός που ενισχύει την αξιοπιστία του μοντέλου.

Εικόνα 1.22. Word Clouds για κάθε θεματική ομάδα

Για να αξιολογηθεί η ικανότητα του μοντέλου να κατανέμει θεματικά νέες κριτικές, πραγματοποιείται η εφαρμογή του εκπαιδευμένου μοντέλου σε νέα εισερχόμενα κείμενα. Τα κείμενα αυτά μετατρέπονται σε tokenizedDocument και αντιστοιχίζονται στο λεξιλόγιο του αρχικού μοντέλου. Με τη συνάρτηση transform, υπολογίζονται οι πιθανότητες κάθε θέματος για κάθε νέο έγγραφο, δηλαδή το topic mixture. Οι πιθανότητες αυτές απεικονίζονται με ραβδόγραμμα (Εικόνα 1.23) και διαγράμματα τύπου stacked bar (Εικόνα 1.24), τα οποία αναδεικνύουν τη θεματική σύνθεση κάθε εγγράφου. Παρατηρείται ότι σε όλα τα παραδείγματα υπάρχει σαφής κατανομή των πιθανοτήτων με έντονη επικράτηση συγκεκριμένων θεμάτων, κάτι που επιβεβαιώνει την καλή διαχωρισιμότητα του μοντέλου.



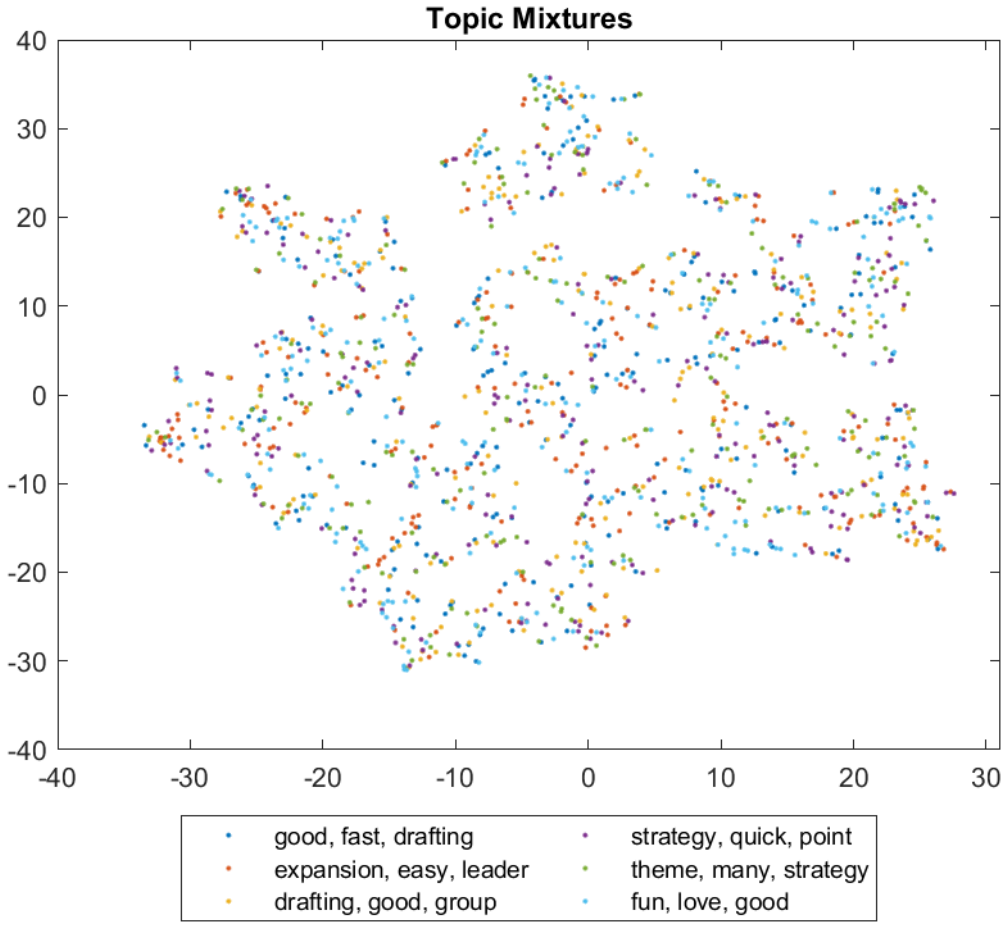
Εικόνα 1.23. Ραβδόγραμμα πιθανοτήτων θεματικών ομάδων



Εικόνα 1.24. Stacked Bar Διάγραμμα Θεματική Σύνθεσης Εγγράφων

Η κατανομή των εγγράφων ως προς τις θεματικές ομάδες συνοψίζεται με ένα γράφημα των μέσων πιθανοτήτων ανά θέμα (Εικόνα 1.23), προσφέροντας μια γενική εικόνα της επικράτησης συγκεκριμένων θεμάτων στο σύνολο του dataset. Παρατηρείται ότι τα θέματα με λέξεις όπως "drafting, good, group" και "fun, love, good" έχουν υψηλή πιθανότητα, κάτι που επιβεβαιώνει τη συνοχή του μοντέλου και την παρουσία συστηματικών εννοιών στις κριτικές.

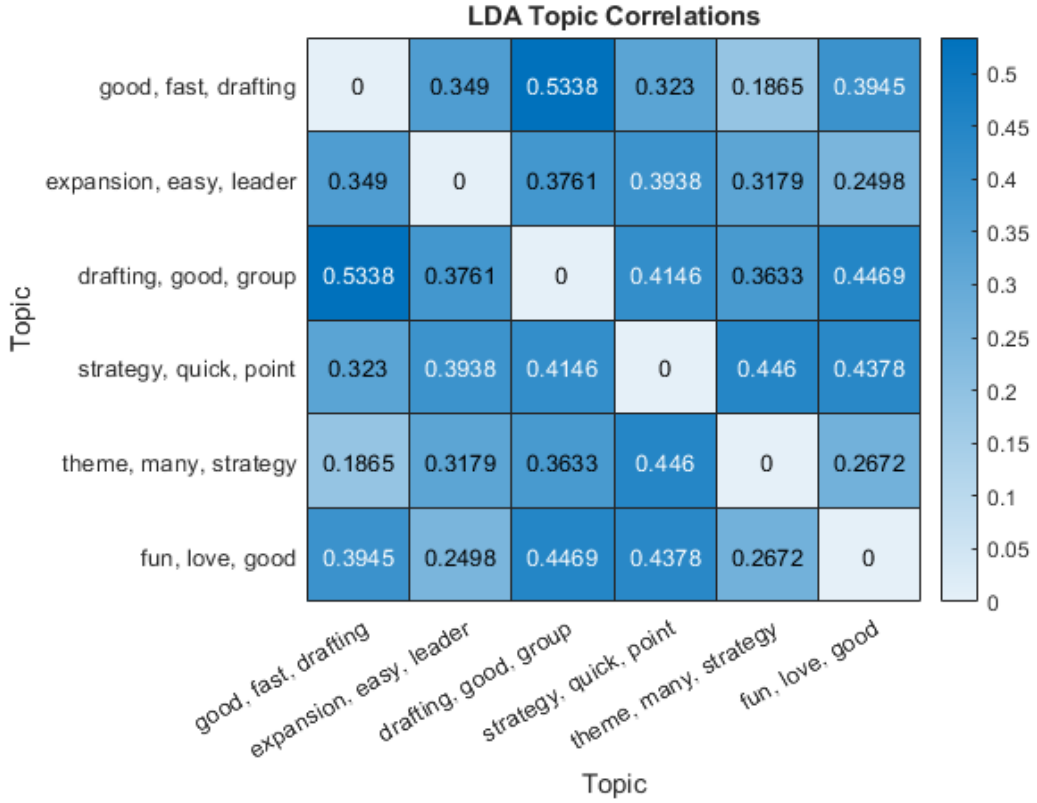
Στη συνέχεια, εφαρμόζεται ανάλυση κύριων συνιστωσών (PCA) για τη μείωση της διάστασης των πιθανοτήτων θεμάτων ανά έγγραφο. Η μείωση επιτρέπει την προβολή των δεδομένων σε δύο διαστάσεις, η οποία πραγματοποιείται μέσω του αλγορίθμου t-SNE. Το αποτέλεσμα είναι μία γραφική απεικόνιση του συνόλου των εγγράφων (Εικόνα 1.25), ομαδοποιημένων με βάση την κυρίαρχη θεματική τους. Οι διαφορετικές ομάδες αποδίδονται με χρώματα, αντιστοιχώντας στις κορυφαίες λέξεις κάθε θέματος. Η απεικόνιση αυτή ενισχύει την οπτική επιβεβαίωση της θεματικής διαφοροποίησης που επιτυγχάνει το μοντέλο.



Εικόνα 1.25. Scatter Plot των Θεματικών Μειγμάτων Ομάδων μέσω t-SNE

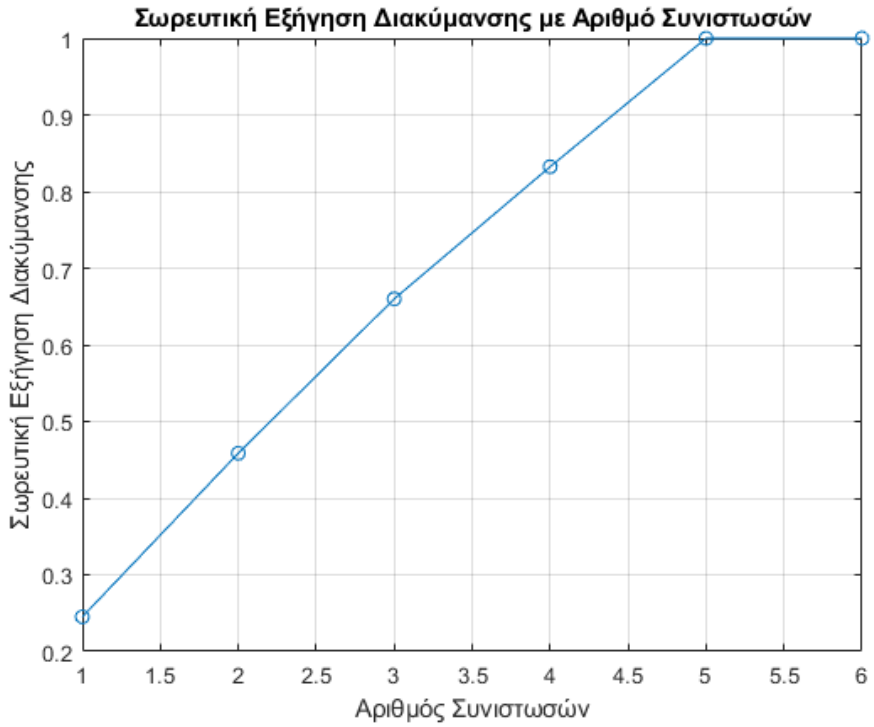
Είναι σημαντικό να διευκρινιστεί ότι παρόλο που η κατανομή των σημείων στην Εικόνα 1.25 εμφανίζει οπτικό διαχωρισμό σε ορισμένες περιοχές, παρατηρείται ταυτόχρονα και επικάλυψη μεταξύ διαφορετικών ομάδων. Το φαινόμενο αυτό είναι απόλυτα αναμενόμενο σε εφαρμογές topic modeling, καθώς πολλές θεματικές ενότητες μοιράζονται κοινό λεξιλόγιο ή σχετικές έννοιες. Επομένως, έγγραφα που σχετίζονται με περισσότερα από ένα θέματα μπορούν να εντοπίζονται σε ενδιάμεσες περιοχές στον τρισδιάστατο χώρο που προβάλλεται σε δύο διαστάσεις με t-SNE. Η επικάλυψη αυτή αντικατοπτρίζει τη ρευστότητα της γλώσσας και την πολυθεματικότητα πολλών σχολίων, χωρίς να μειώνει την εγκυρότητα της ανάλυσης.

Η αναλυτική μελέτη των θεμάτων συμπληρώνεται με την απεικόνιση του πίνακα συσχετίσεων μεταξύ των θεμάτων (Εικόνα 1.26), ο οποίος προκύπτει από τις συντεταγμένες λέξεων στο μοντέλο. Η απεικόνιση του πίνακα με τη χρήση heatmap επιτρέπει τον εντοπισμό πιθανών επικαλύψεων ή αλληλεπιδράσεων μεταξύ των θεμάτων. Παρατηρείται ότι ορισμένα ζεύγη θεμάτων, όπως «drafting, good, group» και «fun, love, good», εμφανίζουν σημαντικά θετική συσχέτιση, κάτι που είναι αναμενόμενο δεδομένης της θεματικής τους εγγύτητας.



Εικόνα 1.26. Heatmap Πίνακας Συσχέτισης Θεματικών Ομάδων

Τέλος, εφαρμόστηκε PCA όχι μόνο για απεικόνιση, αλλά και για τη διερεύνηση της πληροφορίας που διατηρείται με διαφορετικό αριθμό κύριων συνιστωσών. Η σχετική γραφική παράσταση (Εικόνα 1.27) παρουσιάζει τη σωρευτική εξήγηση διακύμανσης, η οποία φτάνει σχεδόν το 100% με έξι συνιστώσες, δηλαδή όσες και τα θέματα του μοντέλου. Αυτό δείχνει ότι οι έξι θεματικές ομάδες καλύπτουν πλήρως το φάσμα της πληροφορίας, επιβεβαιώνοντας την καταλληλόλητα του αριθμού θεμάτων και από πλευράς διακριτικής ικανότητας του μοντέλου.



Εικόνα 1.27. Σωρευτική Εξήγηση Διακύμανσης

Επιπλέον, ο πίνακας συσχέτισης θεμάτων (Πίνακας 1.5) παρέχει ποσοτική πληροφόρηση σχετικά με το πόσο ισχυρά σχετίζονται μεταξύ τους τα επιμέρους θέματα, βάσει της ομοιότητας των λέξεων που τα συγκροτούν. Για παράδειγμα, το θέμα με λέξεις "drafting, good, group" παρουσιάζει ισχυρή θετική συσχέτιση (0.5338) με το "good, fast, drafting", υποδηλώνοντας επικαλυπτόμενα συμφραζόμενα στις κριτικές. Τέτοιες αλληλεπιδράσεις ενισχύουν την εικόνα μιας συνεκτικής θεματικής χαρτογράφησης.

Πίνακας 1.5. Πίνακας Συσχέτισης Θεμάτων

Topic Index	Topic	Top Correlated Topic Index	Top Correlated Topic	Correlation Coefficient
1	"good, fast, drafting"	3	"drafting, good, group"	0.5338
2	"expansion, easy, leader"	4	"strategy, quick, point"	0.3938
3	"drafting, good, group"	1	"good, fast, drafting"	0.5338
4	"strategy, quick, point"	5	"theme, many, strategy"	0.4460
5	"theme, many, strategy"	4	"strategy, quick, point"	0.4460
6	"fun, love, good"	3	"drafting, good, group"	0.4469

Τέλος, η απεικόνιση των κορυφαίων λέξεων που σχετίζονται με τα θέματα των νέων σχόλιων (Πίνακας 1.6) αποκαλύπτει λέξεις με υψηλή βαρύτητα, όπως "fun", "love" και "good", οι οποίες παρουσιάζουν τις μεγαλύτερες τιμές score. Το γεγονός αυτό καταδεικνύει ότι οι θεματικές του μοντέλου είναι εστιασμένες σε θετικές εμπειρίες και συναισθήματα, κάτι που συμφωνεί με τη φύση των περισσότερων σχολίων που εξετάζονται.

Πίνακας 1.6. Λέξεις υψηλής βαρύτητας από τα νέα σχόλια

Word	Score
fun	0.071748
love	0.032613
good	0.028562

Συνοψίζοντας, η ολοκληρωμένη διαδικασία εφαρμογής της τεχνικής Latent Dirichlet Allocation (LDA), όπως παρουσιάστηκε στις προηγούμενες υποενότητες, αποτέλεσε βασικό στάδιο στην προσπάθεια εντοπισμού και ανάλυσης θεματικών μοτίβων μέσα σε ένα μεγάλο σύνολο κειμένων. Από την επιλογή του κατάλληλου αριθμού θεμάτων, την αξιολόγηση των επιδόσεων του μοντέλου με βάση ποσοτικά και ποιοτικά κριτήρια, έως την ερμηνεία και οπτικοποίηση των αποτελεσμάτων, κάθε στάδιο συνέβαλε καθοριστικά στην εξαγωγή έγκυρων και χρήσιμων συμπερασμάτων.

Η τελική επιλογή έξι θεμάτων αποδείχθηκε ιδιαίτερα αποτελεσματική. Το μοντέλο κατάφερε να αναδείξει με σαφήνεια διαφορετικές θεματικές περιοχές που εμφανίζονται στο σύνολο των κριτικών, διατηρώντας παράλληλα θεματική συνοχή και καλή διαφοροποίηση μεταξύ των ομάδων. Τα wordclouds και οι πίνακες κορυφαίων λέξεων για κάθε θέμα επιβεβαίωσαν τη σαφήνεια και χρηστικότητα της θεματικής εκπροσώπησης. Παράλληλα, τα διαγράμματα πιθανοτήτων και θεματικών μειγμάτων απέδειξαν την ικανότητα του μοντέλου να αναθέτει με ακρίβεια θέματα σε νέα κείμενα, ενισχύοντας την εμπιστοσύνη στα αποτελέσματα.

Ο πίνακας συσχέτισης θεμάτων και η οπτικοποίηση των εγγράφων με t-SNE ανέδειξαν τη θεματική γειτνίαση ορισμένων ομάδων, όπως αναμενόταν λόγω αλληλεπικαλυπτόμενου λεξιλογίου, χωρίς ωστόσο να παραβλέπεται η γενικότερη διακριτότητα των θεμάτων. Η PCA-ανάλυση και η απεικόνιση της σωρευτικής εξήγησης διακύμανσης παρείχαν επιπλέον τεκμηρίωση για την πληρότητα του μοντέλου, καθώς οι έξι συνιστώσες κατάφεραν να καλύψουν το σύνολο της σημαντικής πληροφορίας.

Συνοψίζοντας, η εφαρμογή του LDA αποτέλεσε ένα πολύτιμο εργαλείο για την κατανόηση των βασικών νοηματικών αξόνων που κυριαρχούν στο εξεταζόμενο σώμα κειμένων. Με συνδυασμό τεχνικών επιλογών, ορθής παραμετροποίησης και στοχευμένων οπτικοποιήσεων, το μοντέλο ανέδειξε κρυφές θεματικές τάσεις με τρόπο συνεκτικό και κατανοητό. Η εμπειρική αξιολόγηση του αριθμού θεμάτων, η επεξήγηση των παραγόμενων αποτελεσμάτων και η διασταύρωσή τους με

ποσοτικούς δείκτες διασφάλισαν την αξιοπιστία της μεθοδολογίας. Έτσι, η θεματική ανάλυση μέσω LDA ολοκληρώθηκε με επιτυχία, προσφέροντας ουσιαστική γνώση για τη δομή και το περιεχόμενο των σχολίων.

3.10. Ταξινόμηση Σχολίων με Χρήση BERT (Tiny, Mini, Small)

Σε αυτή την ενότητα παρουσιάζεται η εφαρμογή και αξιολόγηση τριών παραλλαγών του μοντέλου BERT (Tiny, Mini και Small) για την πολυκλασική ταξινόμηση σχολίων σε κατηγορίες συναισθήματος ("Χαμηλές", "Μέτριες", "Υψηλές" Βαθμολογίες). Τα μοντέλα και οι αντίστοιχοι ταξινομητές σχεδιάστηκαν και υλοποιήθηκαν αποκλειστικά για τις ανάγκες της παρούσας εργασίας, ενώ λήφθηκαν υπόψη τόσο οι υπολογιστικοί περιορισμοί όσο και η ανάγκη για ισορροπία μεταξύ ταχύτητας και απόδοσης. Η εκπαίδευση πραγματοποιήθηκε σε υπολογιστικό σύστημα με κάρτα γραφικών NVIDIA RTX 4060 των 8GB VRAM, γεγονός που καθόρισε σημαντικά τις παραμέτρους του training process, όπως το batch size και την υπομονή (patience) για early stopping.

Και στα τρία μοντέλα ακολουθήθηκε κοινή προεπεξεργασία: οι βαθμολογίες στρογγυλοποιήθηκαν και κατηγοριοποιήθηκαν σε τρεις τάξεις, δημιουργήθηκε οπτική αναπαράσταση της κατανομής τους, και τα δεδομένα διαχωρίστηκαν σε σύνολο εκπαίδευσης (90%) και δοκιμής (10%) μέσω της μεθόδου Holdout. Δεν εφαρμόστηκε ο κώδικας προεπεξεργασίας κειμένου όπως έγινε στους προηγούμενους κώδικες, καθώς αυτά τα μοντέλα χρειάζονται το πλήρες περιεχόμενο της πρότασης. Οι τελικοί ταξινομητές αξιολογήθηκαν βάσει των μέτρων απόδοσης accuracy, precision, recall και F1-score.

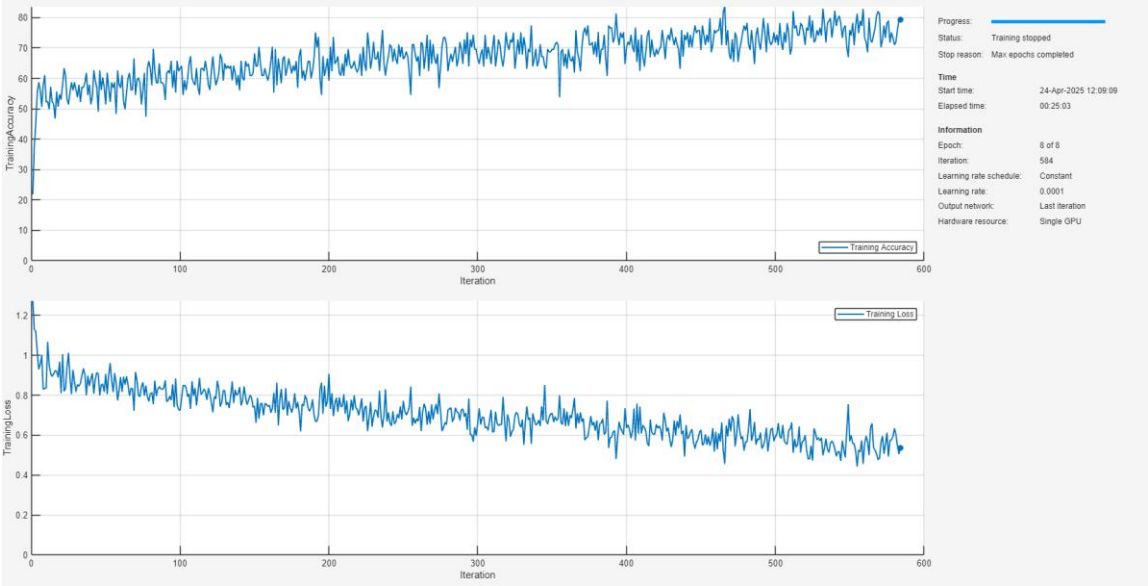
Tiny BERT

Το Tiny BERT επιλέχθηκε ως η ταχύτερη και ελαφρύτερη έκδοση, με μόλις 4.3 εκατομμύρια παραμέτρους εκπαίδευσης, για αρχική αξιολόγηση και σύγκριση. Έτσι, δημιουργήθηκε ο κώδικας tiny_bert.mlx ο οποίος δέχεται ως είσοδο τον πίνακα των δεδομένων T. Εκπαιδεύτηκε με optimizer Adam, ρυθμό εκμάθησης 0.0001 (1e-4) και 8 εποχές εκπαίδευσης, χωρίς early stopping ή mini-batching, ώστε να αξιοποιηθεί πλήρως η υπολογιστική μνήμη VRAM της NVIDIA RTX 4060. Η απουσία mini-batch επιτρέπει την ταχύτερη επεξεργασία μικρών datasets, ενώ η σταθερή διάρκεια εκπαίδευσης διευκολύνει τη σύγκριση απόδοσης. Εκτελώντας τον κώδικα, η ταξινόμηση σχολίων βασίζεται στο εκπαιδευμένο μοντέλο και γίνεται σε πραγματικό χρόνο δίνοντας ως έξοδο τα μέτρα απόδοσης accuracy, precision, recall, F1-score και ένα πίνακα σύγχυσης.

Ο κώδικας αυτός καλείται ως συνάρτηση από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

[tiny_bert_accuracy, tiny_bert_precision, tiny_bert_recall, tiny_bert_f1_score] = tiny_bert(T);

Μετά από περίπου **25 λεπτά εκπαίδευσης** το μοντέλο δίνει τις παρακάτω εξόδους:



Εικόνα 1.28. Εκπαίδευση και αποτύπωση ζωντανής Accuracy και Training Loss της Tiny BERT

Εικόνα 1.29. Πίνακας Σύγχυσης για Tiny BERT

Πίνακας 1.7. Ακρίβεια και Μέτρα Απόδοσης Tiny BERT

Accuracy 0.61 ή 61%			
	Precision	Recall	F1-score
Χαμηλές	0.3898	0.3239	0.3538
Μέτριες	0.5083	0.4792	0.4933
Υψηλές	0.6943	0.7349	0.7140

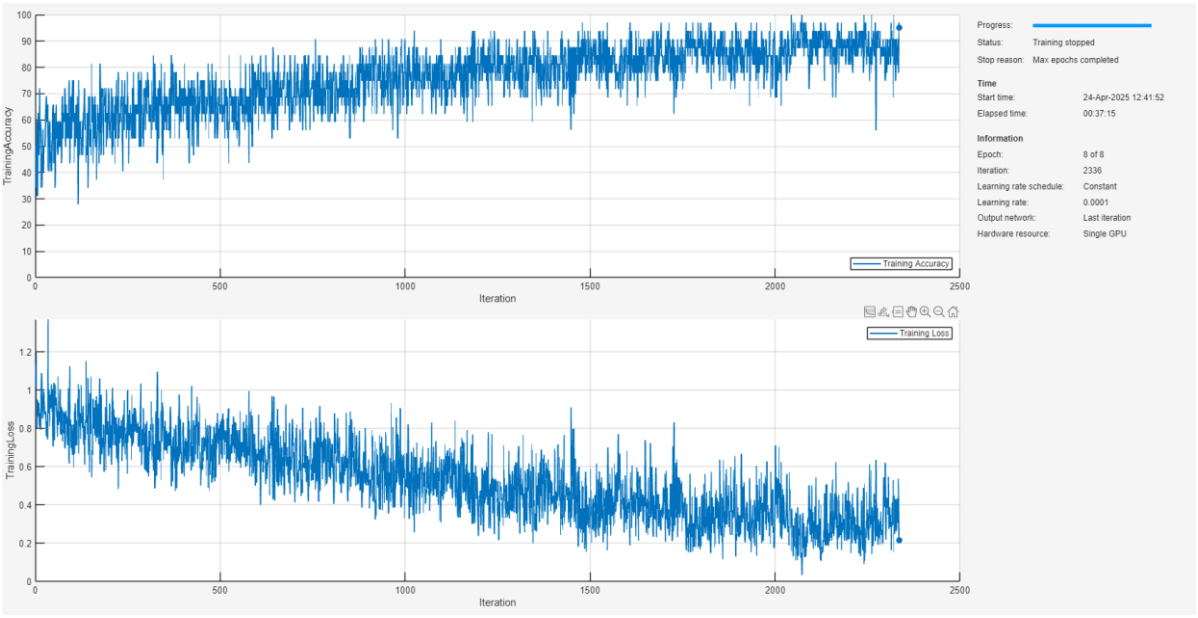
Mini BERT

Το Mini BERT παρέχει καλύτερη ισορροπία μεταξύ πολυπλοκότητας και ταχύτητας, με μόλις 11.1 εκατομμύρια παραμέτρους εκπαίδευσης. Σε αυτό το μοντέλο χρησιμοποιήθηκε ο Adam optimizer με learning rate 1e-4, 8 εποχές εκπαίδευσης, mini-batch size 32 και early stopping με υπομονή 3 εποχών, δηλαδή διακοπή εκπαίδευσης αν δεν βελτιωθεί η απόδοση για 3 συνεχόμενες εποχές. Το mini-batch επιτρέπει τη σταδιακή ενημέρωση του μοντέλου και την αξιοποίηση του GPU memory bandwidth της RTX 4060, ενώ το early stopping προστατεύει από overfitting σε περίπτωση κορεσμού απόδοσης. Παρόμοια με την εκδοχή tiny δημιουργήθηκε ο κώδικας mini_bert.mlx με ίδιες εισόδους και εξόδους.

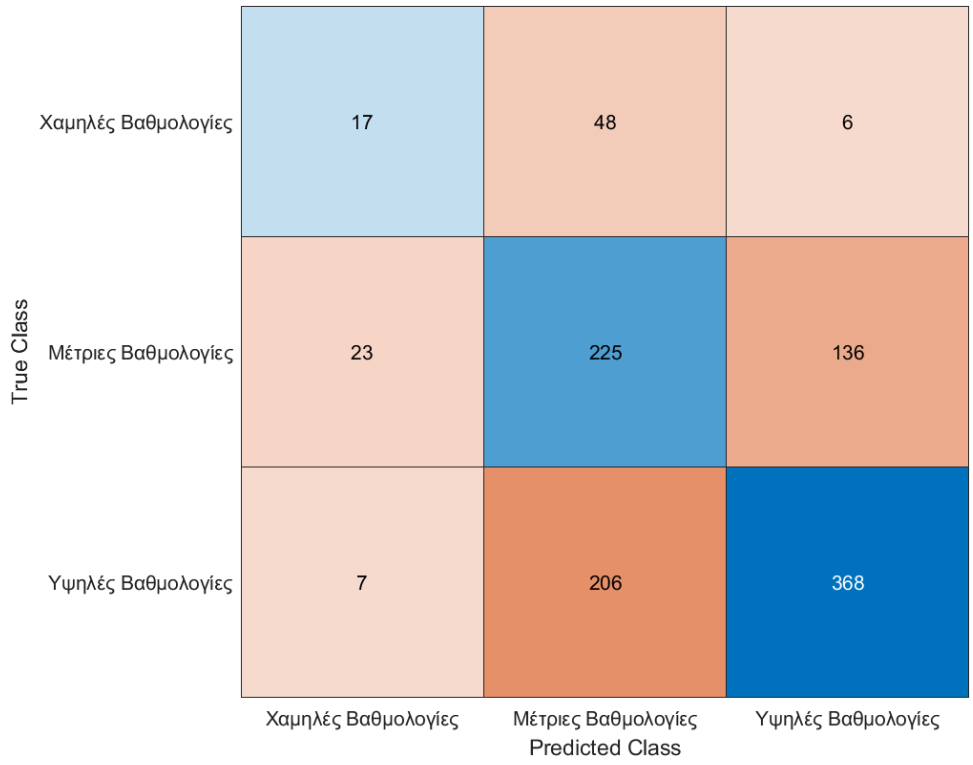
Ο κώδικας αυτός καλείται ως συνάρτηση από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

[mini_bert_accuracy, mini_bert_precision, mini_bert_recall, mini_bert_f1_score] = mini_bert(T);

Μετά από περίπου **35 λεπτά εκπαίδευσης** το μοντέλο δίνει τις παρακάτω εξόδους:



Εικόνα 1.30. Εκπαίδευση και αποτύπωση ζωντανής Accuracy και Training Loss της Mini BERT



Εικόνα 1.31. Πίνακας Σύγχυσης για Mini BERT

Πίνακας 1.8. Ακρίβεια και Μέτρα Απόδοσης Mini BERT

	Accuracy	**0.59 ή 59%**	
	Precision	Recall	F1-score
Χαμηλές	0.3617	0.2394	0.2881
Μέτριες	0.4697	0.5859	0.5214
Υψηλές	0.7216	0.6334	0.6746

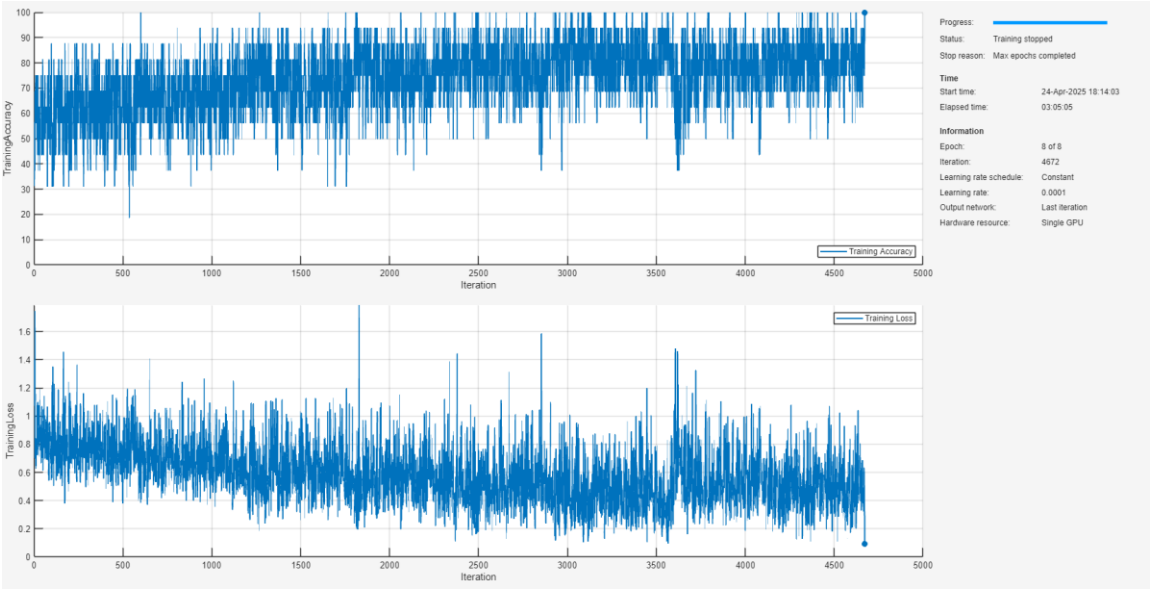
Small BERT

Η έκδοση Small BERT είναι η πιο απαιτητική υπολογιστικά και συνεπώς σχεδιάστηκε με πιο προσεκτικές - 28.5 εκατομμύρια παραμέτρους εκπαίδευσης. Όπως και στο Mini BERT, χρησιμοποιήθηκε Adam optimizer με learning rate 1e-4, 8 εποχές και early stopping. Ωστόσο, λόγω του αυξημένου αριθμού παραμέτρων του μοντέλου, το mini-batch size περιορίστηκε σε 16, ώστε να αποφευχθεί υπέρβαση της διαθέσιμης GPU μνήμης. Αυτή η επιλογή είναι κρίσιμη σε συστήματα με περιορισμένους πόρους όπως η RTX 4060, και εξασφαλίζει σταθερότητα κατά την εκπαίδευση. Το μοντέλο είναι ικανό να καταγράφει πιο σύνθετα μοτίβα από τα κείμενα, με αντίτιμο αυξημένο χρόνο εκπαίδευσης. Όπως και στις προηγούμενες εκδόσεις, δημιουργήθηκε ο κώδικας small_bert.mlx με ίδιες εισόδους και εξόδους.

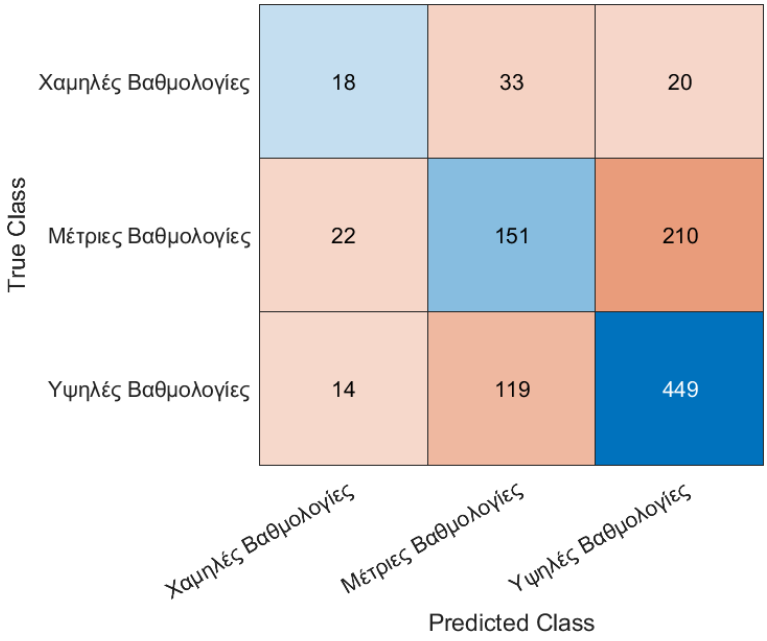
Ο κώδικας αυτός καλείται ως συνάρτηση από τον βασικό εκτελέσιμο κώδικα main_code.mlx ως εξής:

[small_bert_accuracy, small_bert_precision, small_bert_recall, small_bert_f1_score] = small_bert(T);

Μετά από περίπου **3 ώρες εκπαίδευσης** το μοντέλο δίνει τις παρακάτω εξόδους:



Εικόνα 1.32. Εκπαίδευση και αποτύπωση ζωντανής Accuracy και Training Loss της Small BERT



Εικόνα 1.33. Πίνακας Σύγχυσης για Small BERT

Πίνακας 1.9. Ακρίβεια και Μέτρα Απόδοσης Small BERT

	Accuracy	**0.6 ή 60%**	
	Precision	Recall	F1-score
Χαμηλές	0.3333	0.2535	0.2880
Μέτριες	0.4983	0.3943	0.4402
Υψηλές	0.6613	0.7715	0.7121

Αναλύοντας τα αποτελέσματα που προέκυψαν από την εφαρμογή των μοντέλων Tiny, Mini και Small BERT, (Εικόνα 1.28)-(Πίνακας 1.7) & (Εικόνα 1.30)-(Πίνακας 1.8) & (Εικόνα 1.32)-(Πίνακας 1.9) αντίστοιχα, προκύπτουν ορισμένα κρίσιμα συμπεράσματα τόσο ως προς την απόδοση όσο και ως προς την πρακτική χρησιμότητα κάθε έκδοσης.

Η έκδοση **Tiny BERT** παρουσίασε τη συνολικά καλύτερη συμπεριφορά, με **accuracy 61%** και υψηλότερη επίδοση σε F1-score για τις «Υψηλές Βαθμολογίες» (**0.7140**), που αποτελούν την πολυπληθέστερη τάξη. Η σχετική σταθερότητα στο training accuracy και η σταδιακή μείωση του training loss υποδεικνύουν ότι το μοντέλο εκπαιδεύτηκε αποτελεσματικά, χωρίς σημαντικά φαινόμενα overfitting. Η confusion matrix δείχνει ότι οι «Μέτριες» και «Υψηλές» τάξεις διαχωρίζονται ικανοποιητικά, αν και παραμένει αλληλοεπικάλυψη.

Το **Mini BERT**, παρά την πιο προσεγμένη διαδικασία εκπαίδευσης με mini-batch και early stopping, κατέγραψε **ελαφρώς χαμηλότερη ακρίβεια (59%)**. Αν και η precision στην τάξη «Υψηλές» ήταν η υψηλότερη (**0.7216**), η recall αυτής της τάξης (**0.6334**) υπολείπεται της αντίστοιχης του Tiny BERT. Ενδιαφέρον έχει η καλή επίδοση στις «Μέτριες» (recall 0.5859), που συνοδεύεται όμως από πιο αδύναμη F1-score (**0.5214**). Το confusion matrix δείχνει σημαντική σύγχυση μεταξύ «Μέτριων» και «Υψηλών», κάτι που θα μπορούσε να αποδοθεί σε **μικρή διαφοροποίηση γλωσσικού ύφους** ανάμεσα στις δύο κατηγορίες.

Τέλος, το **Small BERT**, αν και πιο «βαρύ» μοντέλο, δεν κατάφερε να υπερέχει στα μέτρα απόδοσης. Κατέγραψε **accuracy 60%**, με υψηλή recall στις «Υψηλές» (**0.7715**), αλλά και σημαντικά μειωμένες επιδόσεις στην τάξη «Μέτριες» (F1-score **0.4402**) και «Χαμηλές» (F1-score **0.2880**). Η confusion matrix επιβεβαιώνει τη μεγάλη δυσκολία του μοντέλου να διακρίνει τις «Μέτριες» από τις «Υψηλές» κριτικές, καθώς υπάρχει εκτεταμένη σύγχυση μεταξύ τους. Αν και η απόδοση είναι σχετικά

ισορροπημένη, η επιπλέον πολυπλοκότητα του μοντέλου δεν αποτυπώθηκε σε αξιοσημείωτη βελτίωση.

Είναι σημαντικό να σημειωθεί ότι δεν έγινε προσπάθεια βελτιστοποίησης των μοντέλων ως προς τις υπερπαραμέτρους ή το validation strategy. Ο **στόχος της παρούσας ενότητας** ήταν η δίκαιη **σύγκριση διαφορετικών εκδόσεων του BERT** κάτω από κοινές ρυθμίσεις εκπαίδευσης (ίδιο learning rate, epochs και split). Μια πιθανή επέκταση θα μπορούσε να εξετάσει την επίδραση διαφορετικών μεθόδων διαχωρισμού δεδομένων (π.χ. 80/20 ή cross-validation) ή τη χρήση τεχνικών όπως weighted loss, ώστε να αντιμετωπιστεί η ανισορροπία μεταξύ τάξεων.

Συνολικά, τα αποτελέσματα (Πίνακας 1.10) υποδεικνύουν ότι το **Tiny BERT είναι η πιο αποτελεσματική επιλογή** για το συγκεκριμένο πρόβλημα ταξινόμησης, προσφέροντας ικανοποιητικά αποτελέσματα με σημαντικά μικρότερο υπολογιστικό κόστος. Οι πιο «βαριές» εκδοχές, αν και εν δυνάμει ισχυρότερες, δεν κατάφεραν να υπερισχύσουν της Tiny, πιθανώς λόγω της περιορισμένης ποικιλομορφίας του dataset και της εγγενούς αμφισημίας της ενδιάμεσης τάξης-κλάσης (Μέτριες Βαθμολογίες).

Οι παρατηρήσεις αυτές ενισχύουν τη **σημασία της επιλογής κατάλληλου μοντέλου** με βάση όχι μόνο την αρχιτεκτονική του, αλλά και τα χαρακτηριστικά των δεδομένων εισόδου.

Πίνακας 1.10. Μέτρα Απόδοσης όλων των Εκδόσεων BERT που υπολογίστηκαν

Tiny Bert			
Accuracy		**0.61 ή 61%**	
	Precision	Recall	F1-score
Χαμηλές	**0.3898**	**0.3239**	**0.3538**
Μέτριες	**0.5083**	0.4792	0.4933
Υψηλές	0.6943	0.7349	**0.7140**
Mini Bert			
Accuracy		0.59 ή 59%	
	Precision	Recall	F1-score

Χαμηλές	0.3617	0.2394	0.2881
Μέτριες	0.4697	**0.5859**	**0.5214**
Υψηλές	**0.7216**	0.6334	0.6746
Small Bert			
Accuracy		0.6 ή 60%	
	Precision	Recall	F1-score
Χαμηλές	0.3333	0.2535	0.2880
Μέτριες	0.4983	0.3943	0.4402
Υψηλές	0.6613	**0.7715**	0.7121

3.11. Συγκριτική Αξιολόγηση Ταξινομητών

Σε αυτή την ενότητα επιχειρείται μια συγκριτική ανάλυση των τριών κύριων εποπτευόμενων ταξινομητών που υλοποιήθηκαν στο πλαίσιο της εργασίας: του απλού ταξινομητή με χρήση Bag-of-Words, του Document Embeddings με SVM και του Tiny BERT. Κάθε μία από αυτές τις προσεγγίσεις αξιολογήθηκε με βάση την ακρίβεια (accuracy) και τα τυπικά μέτρα απόδοσης precision, recall και F1-score. Η σύγκριση αποσκοπεί στο να αναδείξει τις διαφορές τόσο στην υπολογιστική πολυπλοκότητα όσο και στην απόδοση κάθε μεθόδου, προκειμένου να εξαχθούν χρήσιμα συμπεράσματα για τη μελλοντική βελτιστοποίηση και αξιοποίησή τους.

Bag-of-Words Εποπτευόμενος Ταξινομητής

Η πιο βασική προσέγγιση ήταν αυτή που βασίστηκε στη μετατροπή των κειμένων σε διάνυσμα χαρακτηριστικών μέσω του Bag-of-Words, και την εφαρμογή ενός απλού ταξινομητή. Τα αποτελέσματα για τις τιμές Holdout 10%, 20% και 30% έδειξαν ακρίβειες 62.45%, 58.56% και 58.29% αντίστοιχα. Παρατηρείται ότι όσο μειώνεται το σύνολο εκπαίδευσης, η απόδοση του μοντέλου υποχωρεί, γεγονός που αναδεικνύει τη σημασία του όγκου των δεδομένων σε τέτοιες μεθόδους. Παρά την απλότητά της, η μέθοδος αυτή προσφέρει μια καλή γραμμή βάσης (baseline) για τη σύγκριση με πιο σύνθετες τεχνικές.

Document Embeddings με RBF SVM

Η δεύτερη προσέγγιση αξιοποίησε αναπαραστάσεις εγγράφων (embeddings), οι οποίες αποτυπώνουν πιο σύνθετες σχέσεις μεταξύ λέξεων και εννοιών. Οι αναπαραστάσεις αυτές ταξινομήθηκαν με χρήση SVM με μη γραμμικό πυρήνα RBF. Τα αποτελέσματα έδειξαν ακρίβεια 64% και μέτριες τιμές για τα υπόλοιπα μέτρα απόδοσης: precision 0.46, recall 0.43 και F1-score 0.44. Η μικρή βελτίωση στην ακρίβεια σε σχέση με τον απλό ταξινομητή, χωρίς όμως σαφή βελτίωση στα υπόλοιπα μέτρα απόδοσης, υποδηλώνει ότι η ποιότητα των embeddings παίζει κρίσιμο ρόλο και ίσως απαιτούνται πιο εξειδικευμένες τεχνικές ή επιπλέον βελτιστοποίηση.

Tiny BERT Ταξινομητής

Η τρίτη και πιο εξελιγμένη προσέγγιση βασίστηκε στο προεκπαιδευμένο μοντέλο Tiny BERT, το οποίο ενσωματώνει σημασιολογική κατανόηση του κειμένου μέσω μετασχηματιστών. Παρά την υπολογιστική του «ελαφρότητα» σε σχέση με μεγαλύτερες εκδοχές του BERT, το Tiny BERT επέδειξε εξαιρετική ισορροπία στην ταξινόμηση των σχολίων. Συγκεκριμένα, παρουσίασε ακρίβεια 61%, με τιμές F1-score 0.3538 για τις «Χαμηλές», 0.4933 για τις «Μέτριες» και 0.7140 για τις «Υψηλές» Βαθμολογίες. Το μοντέλο υπερέχει σαφώς στην αναγνώριση των θετικών κριτικών, ενώ η απόδοσή του στις υπόλοιπες τάξεις παραμένει ικανοποιητική δεδομένου του μεγέθους του και της πολυπλοκότητας του προβλήματος.

Συμπεράσματα

Συγκρίνοντας συνολικά τους ταξινομητές, παρατηρείται ότι:

- Η μέθοδος με Bag-of-Words λειτουργεί ικανοποιητικά για μικρά ή λιγότερο σύνθετα datasets.
- Η προσέγγιση με embeddings προσφέρει βελτίωση στην ακρίβεια, αλλά περιορίζεται στην ερμηνεία της σημασιολογίας.
- Το Tiny BERT, παρόλο που δεν υπερέχει καθολικά στα μέτρα απόδοσης, επιτυγχάνει την καλύτερη απόδοση του στην πιο σημαντική κατηγορία, αφού διατηρεί υψηλό F1-score για τα θετικά σχόλια, που αποτελούν και το 56.1% των δεδομένων.

Συνεπώς, η επιλογή μεθόδου εξαρτάται άμεσα από το διαθέσιμο υπολογιστικό δυναμικό, τον επιθυμητό χρόνο εκπαίδευσης και τον βαθμό κατανόησης των αποτελεσμάτων που απαιτείται. Για εφαρμογές όπου είναι σημαντική η ακρίβεια στην αναγνώριση θετικών σχολίων ή η κατανόηση νοήματος από τον κώδικα, το Tiny BERT αναδεικνύεται ως η καταλληλότερη επιλογή. Όμως, βάσει της

υπολογισμένης απόδοσης, δεδομένου του dataset που χρησιμοποιήθηκε, και του hardware στο οποίο εκτελέστηκαν οι κώδικες, **ο καλύτερος ταξινομητής είναι ο Document Embeddings με RBF SVM** αλγόριθμο εκμάθησης.

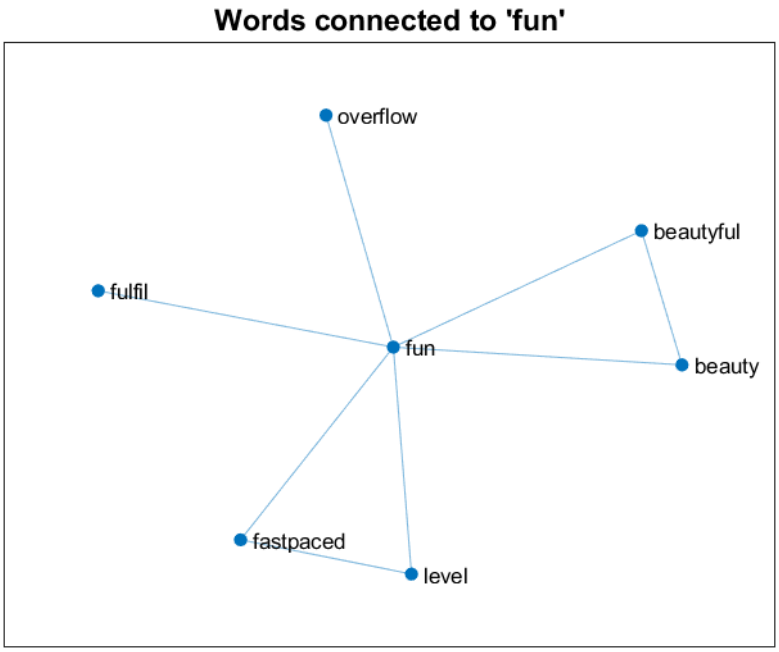
3.12. Ανάλυση Συναισθήματος με Χρήση Λεξιλογικού Μοντέλου VADER

Η ενότητα αυτή παρουσιάζει μια εποπτευόμενη προσέγγιση ανάλυσης συναισθήματος που αξιοποιεί τη μεθοδολογία VADER (Valence Aware Dictionary and sEntiment Reasoner) σε συνδυασμό με μοντέλα λέξεων και γραφήματα λέξεων. Η συνάρτηση darth_vader_train_test που υλοποιήθηκε αποκλειστικά στο πλαίσιο της παρούσας εργασίας αναλαμβάνει τόσο την εκπαίδευση όσο και τη δοκιμή ενός προσαρμοσμένου λεξικού συναισθημάτων, εφαρμόζοντάς το σε πραγματικά δεδομένα κριτικών.

Η διαδικασία ξεκινά με την επεξεργασία του πίνακα εισόδου T, ο οποίος περιέχει δύο βασικές στήλες: τα κείμενα των σχολίων (comment) και τις αντίστοιχες βαθμολογίες (rating). Οι βαθμολογίες στρογγυλοποιούνται και τα δεδομένα διαχωρίζονται σε σύνολο εκπαίδευσης (80%) και δοκιμής (20%) με στρωματοποιημένο HoldOut ώστε να διατηρηθεί η κατανομή των κατηγοριών.

Κατά την εκπαίδευση, εφαρμόζεται προεπεξεργασία στα σχόλια του training set και δημιουργείται ένα word embedding μοντέλο με ρυθμίσεις: παράθυρο λέξεων (Window) ίσο με 15, ελάχιστη εμφάνιση λέξεων (MinCount) στο 1, και αριθμό εποχών (NumEpochs) ίσο με 10. Το παραθύρου συμφραζόμενων (context window size), δηλαδή είναι το πλήθος λέξεων πριν και μετά από κάθε κεντρική λέξη που θα εξετάζει το μοντέλο για να μάθει τις σχέσεις μεταξύ λέξεων. Έτσι, το word embedding μετατρέπει τις λέξεις σε διανύσματα ώστε να επιτρέπει τη σύγκρισή τους σε έναν πολυδιάστατο σημασιολογικό χώρο.

Στη συνέχεια, κατασκευάζεται γράφος λέξεων (Εικόνα 1.34), όπου κάθε λέξη συνδέεται με τους πλησιέστερους σημασιολογικά γείτονές της, με βάση την ευκλείδεια απόσταση. Οι κόμβοι του γράφου είναι οι λέξεις και οι ακμές φέρουν βάρη που αντικατοπτρίζουν τις μεταξύ τους αποστάσεις.



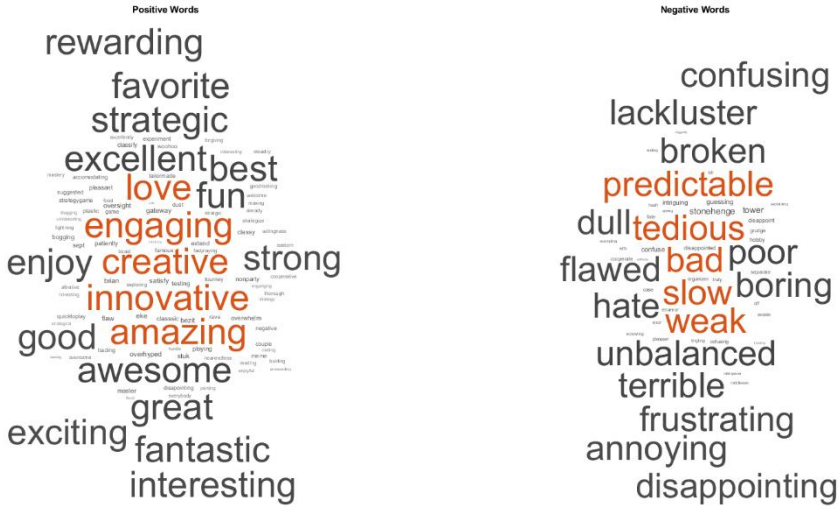
Εικόνα 1.34. Λέξεις με άμεση συνάφεια με την λέξη “fun”

Ο υπολογισμός των συναισθηματικών βαθμολογιών γίνεται με τη χρήση seed λέξεων: 19 θετικές (όπως "fun", "excellent", "awesome") και 19 αρνητικές (όπως "boring", "disappointing", "worst"). Μέσω της βοηθητικής συνάρτησης polarityScores, υπολογίζεται πόσο «έντονα» σχετίζεται κάθε λέξη με τις θετικές και αρνητικές λέξεις-σπόρους, λαμβάνοντας υπόψη τη δομή του γράφου και το βάθος διάχυσης. Η διαδικασία αυτή επαναλαμβάνεται για βάθη απόστασης (depth) από 1 έως 5 (τιμή του maxPathLength), και σε κάθε βήμα κανονικοποιούνται οι τιμές και υπολογίζεται ο συνδυασμένος συναισθηματικός δείκτης για κάθε λέξη. Ο τελικός sentiment score για κάθε λέξη προκύπτει από τον μέσο όρο των επιμέρους βημάτων.

Το αποτέλεσμα αυτής της διαδικασίας είναι ένα λεξικό (lexiconTable) που περιέχει τις λέξεις του embedding και τη συναισθηματική τους βαθμολογία, όπως φαίνεται ενδεικτικά στον (Πίνακας 1.11). Οι λέξεις με πολύ χαμηλή ένταση συναισθήματος (κάτω από το 0.01 σε απόλυτη τιμή) απορρίπτονται. Επίσης, δημιουργούνται σύννεφα λέξεων (Εικόνα 1.35) για την αναπαράσταση των θετικά και αρνητικά συναισθηματικά φορτισμένων λέξεων.

Πίνακας 1.11. Λεξικό με Βαθμολογίες Συναισθήματος

Token	Sentiment Score
innovative	4
repetitive	0.2275
balanced	0.0883
complex	-0.0731
unbalanced	-4

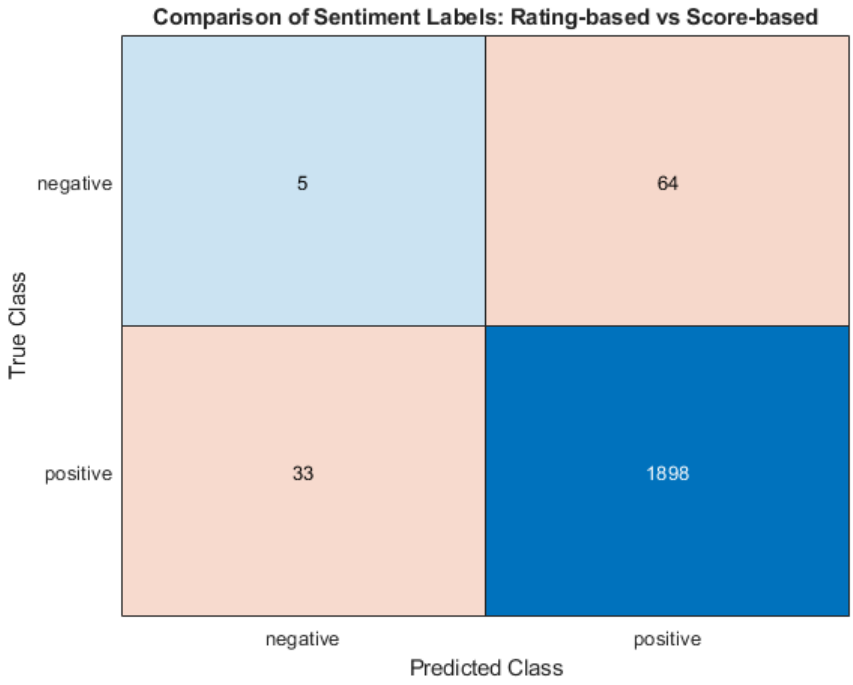


Εικόνα 1.35. Σύννεφα Θετικών και Αρνητικών Λέξεων

Στο στάδιο της δοκιμής, εφαρμόζεται η ίδια διαδικασία προεπεξεργασίας στα σχόλια του test set. Στη συνέχεια, το λεξικό εφαρμόζεται μέσω της συνάρτησης vaderSentimentScores και υπολογίζονται οι compound scores για κάθε σχόλιο. Ως compound score ορίζεται η συνολική συναισθηματική ένταση που αντιστοιχίζεται σε ένα σχόλιο από το μοντέλο. Δηλαδή, βάσει των sentiment scores που φιλοξενεί το lexicon Table και την συχνότητα εμφάνισης των λέξεων αυτού σε ένα σχόλιο, υπολογίζεται ένας αριθμός ορισμένος στο [-1, 1], όπως φαίνεται στον (Πίνακας 1.12).

Τα αποτελέσματα συγκρίνονται με τις πραγματικές βαθμολογίες των σχολίων, με στόχο την εκτίμηση της απόδοσης του λεξικού. Οι κριτικές χαρακτηρίζονται θετικές ή αρνητικές τόσο με βάση τη

βαθμολογία που έβαλε ο χρήστης όσο και με βάση το sentiment score που τους αποδίδεται από το μοντέλο, και εμφανίζεται ένα confusion matrix (Εικόνα 1.36) που απεικονίζει την αντιστοιχία αυτών των δύο χαρακτηρισμών.

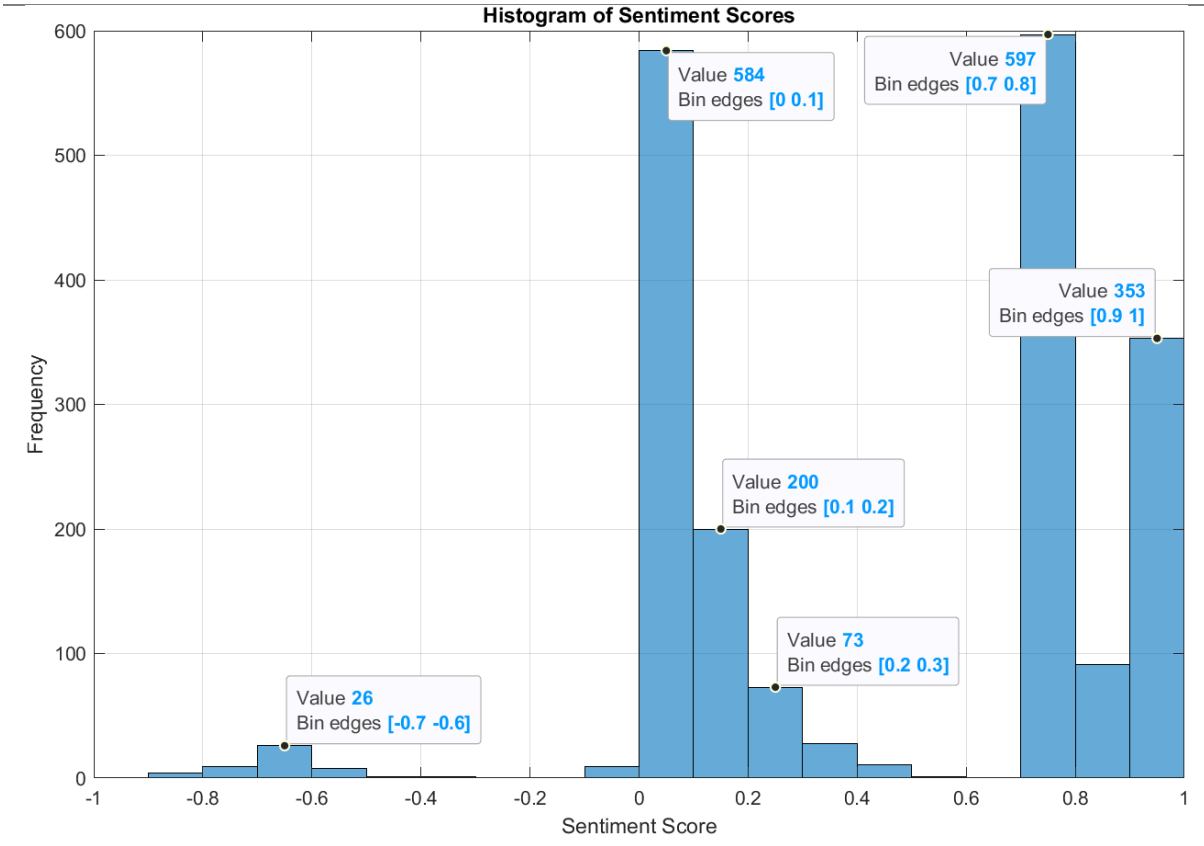


Εικόνα 1.36. Πίνακας Σύγχυσης Βαθμολογίας Χρηστών και Sentiment Score

Τέλος, η συνάρτηση επιστρέφει το compoundScores, ένα διάνυσμα με τις συναισθηματικές βαθμολογίες των σχολίων του test set, το lexiconTable, το τελικό λεξικό συναισθημάτων με λέξεις και sentiment scores, και το testDataResults, τον πίνακα με τις δοκιμαστικές κριτικές, τα sentiment scores και τις ταξινομήσεις τους. Επίσης, εκτυπώνει και ένα ακόμα διάγραμμα, το ιστόγραμμα των compoundScores που αποδίδονται τελικά στα σχόλια των χρηστών.

Πίνακας 1.12. Ενδεικτικά Αποτελέσματα VADER στον πίνακα testDataResults

comment	rating round	compound score	rating label	score label
love artwork amazing fast mechanism	10	0.9012	"positive"	"positive"
randomness complex version monopoly	5	0.0614	"negative"	"positive"
tedious drafting process start	5	-0.7036	"negative"	"negative"
broken token organizer	9	-0.7226	"positive"	"negative"
frustrating due lack control approach lighthearted involved filler	7	-0.6789	"positive"	"negative"



Εικόνα 1.37. Ιστόγραμμα των Compound Scores από το Sentiment Analysis των σχολίων των χρηστών

Το ιστόγραμμα των compound scores (Εικόνα 1.37) δείχνει ξεκάθαρη **μετατόπιση προς θετικές τιμές**, με το μεγαλύτερο πλήθος κριτικών να συγκεντρώνεται στις περιοχές 0.7–0.8 και 0.9–1.0. Αυτό φανερώνει ότι **το σετ είχε θετικό φορτίο**, στοιχείο που συνάδει με την ποσοστιαία υπερίσχυση θετικών σχολίων στο σύνολο των δεδομένων.

Όπως παρατηρείται στον Πίνακας 8, αν και ορισμένες λέξεις λαμβάνουν εύλογες βαθμολογίες (π.χ. innovative = 4, unbalanced = -4), υπάρχουν και περιπτώσεις όπου οι τιμές είναι αμφιλεγόμενες. Ενδεικτικά, η λέξη balanced έχει σκορ μόλις **0.0883**, και η complex εμφανίζει ελαφρώς αρνητικό πρόσημο (**-0.0731**), γεγονός που δεν ευθυγραμμίζεται απαραίτητα με την αντικειμενική ή γενικώς αποδεκτή συναισθηματική φόρτιση των λέξεων αυτών.

Το παραπάνω φανερώνει μια **σημαντική πρόκληση στη διαδικασία κατασκευής λεξικών συναισθήματος**: η απόδοση βαθμολογιών σε tokens βάσει μεθόδων αυτόματης εξαγωγής (π.χ. word embeddings, γραφηματικοί αλγόριθμοι) ενδέχεται να παράγει **σκορ που δεν είναι απόλυτα ευθυγραμμισμένα με την ανθρώπινη αντίληψη**. Αυτή η ασυμφωνία, αν και αναμενόμενη, επηρεάζει την ποιότητα της τελικής ανάλυσης συναισθήματος.

Είναι σημαντικό να σημειωθεί ότι η δημιουργία ενός πλήρους και αξιόπιστου λεξικού απαιτεί **εκτεταμένη προσπάθεια, πολλαπλούς κύκλους επανελέγχου** και συνήθως **συνδυασμό αυτόματων μεθόδων και ανθρώπινης επιμέλειας**, διαδικασία η οποία ξεφεύγει από τα χρονικά και ερευνητικά όρια της παρούσας εργασίας.

Η αξιολόγηση της ταξινόμησης φαίνεται στον συγκριτικό πίνακα (Πίνακας 1.12), στον οποίο παρατηρείται ότι παρά την υψηλή βαθμολογία ενός σχολίου, το σύστημα μπορεί να κατατάξει το συναίσθημα ως αρνητικό όταν περιέχονται έντονα αρνητικές λέξεις (π.χ. "broken", "frustrating"). Αυτό **αναδεικνύει τη σημασία του γενικού πλαισίου** (context) της κριτικής, αλλά και τον περιορισμό των lexicon-based μεθόδων. Να σημειωθεί ότι στον (Πίνακας 1.12) φαίνονται ενδεικτικά μερικές γραμμές από τον πλήρη πίνακα που δίνει ο κώδικας. Η τελική απόδοση του συστήματος φαίνεται στο confusion matrix (Εικόνα 35), όπου η ακρίβεια της θετικής κλάσης είναι υψηλή, αλλά υπάρχει χαμηλή ανάκληση (recall) στην αρνητική, καθώς η συντριπτική πλειοψηφία των σχολίων κατηγοριοποιήθηκαν ως θετικά, ακόμα και όταν είχαν οριακές ή ελαφρώς αρνητικές βαθμολογίες.

Συνολικά, το μοντέλο VADER με προσαρμοσμένο λεξικό **προσφέρει γρήγορη και διαφανή συναισθηματική ανάλυση-ταξινόμηση, με δυνατότητα ερμηνείας των αποτελεσμάτων**. Παρότι υστερεί στην αναγνώριση πιο «ουδέτερων» ή «λεπτών» αποχρώσεων συναισθήματος, αποτελεί μια σταθερή βάση για συγκρίσεις με πιο σύνθετα μοντέλα, όπως το GPT-4 Turbo που θα εξεταστεί στην επόμενη ενότητα.

3.13. Συγκριτική Ανάλυση Συναισθήματος μεταξύ VADER και GPT-4 Turbo

Η παρούσα ενότητα επικεντρώνεται στη σύγκριση δύο διαφορετικών προσεγγίσεων ανάλυσης συναισθήματος: του custom μοντέλου βασισμένου στη μεθοδολογία VADER (Valence Aware Dictionary and sEntiment Reasoner) που παρουσιάστηκε στην προηγούμενη ενότητα, και του ισχυρού γλωσσικού μοντέλου GPT-4 Turbo της OpenAI. Ο στόχος είναι η αξιολόγηση της ακρίβειας πρόβλεψης σε σχόλια προϊόντων, ταξινομημένα σε τρεις κατηγορίες: «Χαμηλές», «Μέτριες» και «Υψηλές» βαθμολογίες.

Έτσι, δημιουργήθηκε ο κώδικας compare_sentiment_models.mlx με τρόπο τέτοιο ώστε να μπορεί να εκτελεστεί ανεξάρτητα από την προηγούμενη ενότητα, επαναλαμβάνοντας όλα τα απαραίτητα βήματα για την εκπαίδευση του custom VADER-based μοντέλου και την δημιουργία του απαραίτητου λεξικού.

Η συνάρτηση καλείται από τον κύριο κώδικα ως:

[compoundScores, lexiconTable, testDataResults] = compare_sentiment_models(T);

Ο κώδικας δέχεται ως είσοδο τον πλήρη πίνακα δεδομένων T, και η διαδικασία ξεκινά με την προεπεξεργασία των δεδομένων, όπου οι αριθμητικές βαθμολογίες στρογγυλοποιούνται και μετατρέπονται σε κατηγορικές όπως και σε όλες τις προηγούμενες ενότητες. Μετατρέποντας τις βαθμολογίες σε κατηγορικές, δημιουργήθηκε και το διάνυσμα ground truth το οποίο αποτελεί την «επίγεια αλήθεια» κατηγοριοποίησης της κριτικής, και θα χρησιμοποιηθεί στην συνέχεια για τον υπολογισμό της ακρίβειας των προβλέψεων του κάθε μοντέλου.

Ακολουθεί διαχωρισμός των δεδομένων σε σύνολο εκπαίδευσης (80%) και δοκιμής (20%), ώστε να διατηρηθεί η αναλογία των τάξεων. Στη συνέχεια, υλοποιείται το custom μοντέλο sentiment analysis που παρουσιάστηκε στην προηγούμενη ενότητα του οποίου τα sentiment scores αποθηκεύονται σε ένα λεξικό, το οποίο φιλτράρεται με πιο αυστηρό κατώφλι 0.05, και χρησιμοποιείται για την ανάλυση των σχολίων του test set. Για την κάθε κριτική του test set, υπολογίζεται compound score και πραγματοποιείται κατηγοριοποίηση σε τρεις κλάσεις, βάσει ορισμένου κατωφλίου 0.1 ως εξής (threshold): σκορ > 0.1 → «Υψηλές», < -0.1 → «Χαμηλές», αλλιώς «Μέτριες».

Τα ίδια σχόλια δίνονται και στο μοντέλο GPT-4 Turbo μέσω API της OpenAI. Για κάθε σχόλιο, αποστέλλεται μήνυμα με οδηγία ταξινόμησης και καταγράφεται η απάντηση. Αξίζει να σημειωθεί ότι η αξιοποίηση του μοντέλου μέσω API απαιτεί την ύπαρξη προσωπικού API key από την πλατφόρμα της OpenAI, το οποίο είναι μοναδικό για κάθε χρήστη και δεν μπορεί να κοινοποιηθεί στη διπλωματική εργασία για λόγους ασφαλείας. Αυτό διαβάζεται από τον κώδικα μέσω ενός apikey.txt

αρχείου που πρέπει να τοποθετήσει ο χρήστης εντός του φακέλου εργασίας του MATLAB. Η χρέωση για την χρήση της υπηρεσίας και του μοντέλου GPT-4 Turbo είναι της τάξεως των 4.5€ για την ανάλυση 2000 σχολίων, ανά εκτέλεση κώδικα.

Οι προβλέψεις του κάθε μοντέλου τοποθετούνται στον πίνακα testDataResults όπως φαίνεται ενδεικτικά στον (Πίνακας 1.13), χωρίς αυτά ωστόσο να αποτελούν παράδειγμα για την απόδοση του κάθε μοντέλου, αφού επιλέχθηκαν τυχαία. Αυτές, συγκρίνονται με το ground truth των αξιολογήσεων, και υπολογίζεται η ακρίβεια πρόβλεψης. Για τη στατιστική αξιολόγηση της διαφοράς στην απόδοση μεταξύ των μοντέλων χρησιμοποιούνται δύο τεστ: το paired t-test (ελέγχει αν υπάρχει σημαντική διαφορά στους μέσους όρους ακρίβειας) και το Wilcoxon signed-rank test (μη παραμετρικό τεστ για τις ίδιες παρατηρήσεις), τα οποία είναι επιθυμητό να δώσουν τιμές μικρότερες από 0.05 ή 0.10 για να θεωρηθεί στατιστικά σημαντική η διαφορά της ανάλυσης των μοντέλων. Τα αποτελέσματα της σύγκρισης φαίνονται στον (Πίνακας 1.14).

Επίσης, ο κώδικας επιστρέφει τις προβλέψεις, αποθηκεύει τα αποτελέσματα σε αρχείο Excel, και το εκπαιδευμένο μοντέλο (.mat).

Πίνακας 1.13. Πίνακας Προβλέψεων των Μοντέλων σε σχέση με τις Πραγματικές Βαθμολογίες Χρηστών

Comment	Rating Round	Ground Truth	Custom VADER Prediction	GPT 4 Turbo Prediction
just excellent	10	Υψηλές Βαθμολογίες	Υψηλές Βαθμολογίες	Υψηλές Βαθμολογίες
Very good game with cities and leaders	10	Υψηλές Βαθμολογίες	Μέτριες Βαθμολογίες	Υψηλές Βαθμολογίες
fun filler game	9	Υψηλές Βαθμολογίες	Υψηλές Βαθμολογίες	Μέτριες Βαθμολογίες
good one, because it is quick to teach and new folks seem to be in the swing of things after a game or two	7	Μέτριες Βαθμολογίες	Μέτριες Βαθμολογίες	Υψηλές Βαθμολογίες
Ok so this is DIFFICULT. Like very difficult. Complex and badly written rules	5	Χαμηλές Βαθμολογίες	Υψηλές Βαθμολογίες	Χαμηλές Βαθμολογίες
7 Wonder isn't a terrible game, but it's not a great one. I applaud the game for the fact that it's relatively strategic and can be completed with a high number of players in a short time, but I have a hard time getting excited about this one.	5	Χαμηλές Βαθμολογίες	Χαμηλές Βαθμολογίες	Μέτριες Βαθμολογίες

Πίνακας 1.14. Αποτελέσματα Σύγκρισης VADER και GPT-4 Turbo

Ακρίβεια Custom Μοντέλου	54.95%
Ακρίβεια GPT-4 Turbo	60.55%
Paired t-test	Στατιστικά σημαντική διαφορά (p = 0.0002)
Wilcoxon signed-rank test	Στατιστικά σημαντική διαφορά (p = 0.0002)

Τα αποτελέσματα της σύγκρισης ανέδειξαν την υπεροχή του GPT-4 Turbo, το οποίο πέτυχε ακρίβεια 60.55% έναντι 54.95% του custom VADER-based μοντέλου. Η διαφορά αυτή κρίθηκε στατιστικά σημαντική και από τα δύο τεστ (p = 0.0002), γεγονός που υποδηλώνει την ανώτερη απόδοση του GPT 4 στην ταξινόμηση συναισθήματος με βάση τα συμφραζόμενα. Η υπεροχή αυτή αποδίδεται στην εγγενή δυνατότητα των LLMs (Large Language Models) να αντιλαμβάνονται τα συμφραζόμενα και «λεπτές αποχρώσεις» συναισθήματος, σε αντίθεση με τα παραδοσιακά μοντέλα που βασίζονται σε προκαθορισμένα λεξικά και γραφικές σχέσεις λέξεων.

Ωστόσο, είναι σημαντικό να σημειωθεί ότι σε κάποια σχόλια το GPT μοντέλο δεν έκανε κάποια πρόβλεψη, τοποθετώντας στο κελί της πρόβλεψης «Δεν περιέχεται κείμενο για ανάλυση» ή «Δεν παρέχει αρκετές πληροφορίες για να κατατάξω το κείμενο». Αυτά τα φαινόμενα συσχετίζονται πιθανώς με σχόλια εξαιρετικά σύντομα, ουδέτερα ή εκτός θέματος, τα οποία μπορούν να αντιμετωπιστούν μέσω βελτιωμένων τεχνικών προεπεξεργασίας του κειμένου των σχολίων.

Παρόλα αυτά, και το custom μοντέλο αποτελεί μια αποδοτική, γρήγορη και πλήρως επεκτάσιμη λύση, η οποία μπορεί να βελτιωθεί περαιτέρω μέσα από βελτιστοποίηση παραμέτρων όπως με tuning των seeds, δημιουργία καταλληλότερου λεξικού συναισθημάτων, αλλαγή embedding παραμέτρων ή threshold τιμών. Επίσης, έχει σημαντικό πλεονέκτημα κόστους, καθώς δεν απαιτεί σύνδεση με εξωτερικό API. Σε κάθε περίπτωση, η παρούσα εργασία, όπως έχει προαναφερθεί, δεν εστιάζει στη βελτιστοποίηση των μοντέλων, αλλά στην υλοποίηση και συγκριτική αξιολόγηση διαφορετικών μοντέλων και μεθοδολογιών.

4. Συμπεράσματα και Προοπτικές

Η παρούσα διπλωματική εργασία είχε ως στόχο τη μελέτη και εφαρμογή τεχνικών Ανάλυσης Συναισθήματος (Sentiment Analysis) σε δεδομένα σχολίων χρηστών, αξιοποιώντας τόσο παραδοσιακές όσο και σύγχρονες μεθόδους Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP). Μέσω αυτής της ερευνητικής διαδρομής, εξετάστηκε η απόδοση και η ερμηνευσιμότητα διαφορετικών μοντέλων ταξινόμησης συναισθήματος, με σκοπό την αξιολόγηση της αξιοπιστίας τους και τη διερεύνηση των δυνατοτήτων αξιοποίησής τους σε πραγματικά δεδομένα.

Αρχικά, υλοποιήθηκαν και συγκρίθηκαν διαφορετικοί εποπτευόμενοι ταξινομητές, ξεκινώντας από απλά bag-of-words μοντέλα και φτάνοντας σε προσεγγίσεις με document embeddings και πολυπλοκότερους αλγορίθμους όπως SVM. Η απόδοσή τους κρίθηκε ικανοποιητική για βασικά επίπεδα ταξινόμησης, ωστόσο κρίθηκε αναγκαία η ενσωμάτωση μοντέλων βασισμένων σε μετασχηματιστές (transformers), όπως τα Tiny, Mini και Small BERT. Μέσω της σύγκρισης των τριών εκδοχών του BERT, διαπιστώθηκε ότι ακόμη και οι μικρότερες εκδόσεις του μοντέλου μπορούν να αποδώσουν ανταγωνιστικά, εφόσον η κατάλληλη παραμετροποίηση επιλεγεί με βάση το διαθέσιμο υπολογιστικό σύστημα και τον όγκο των δεδομένων.

Στο πλαίσιο της λεξικοκεντρικής προσέγγισης, δόθηκε ιδιαίτερη έμφαση στην υλοποίηση μιας παραμετροποιημένης εκδοχής του VADER, μέσω word graphs και custom λεξικών. Η χρήση propagation-based αλγορίθμων στον γράφο των λέξεων προσέφερε μια ερμηνεύσιμη, διαφανή μέθοδο υπολογισμού του συναισθηματικού φορτίου λέξεων, χωρίς την ανάγκη μεγάλων συνόλων εκπαίδευσης ή black-box μοντέλων. Αν και η ακρίβεια του μοντέλου αυτού δεν ξεπέρασε εκείνη του GPT 4 Turbo, τα αποτελέσματα ήταν αξιοπρεπή, ιδίως αν αναλογιστεί κανείς την ελαφρότητα και επεκτασιμότητα της υλοποίησης.

Η τελική συγκριτική αξιολόγηση μεταξύ του custom μοντέλου και του GPT 4 Turbo, μέσω στατιστικών τεστ (paired t-test και Wilcoxon signed-rank), ανέδειξε τη σαφή υπεροχή του δεύτερου ως προς την ακρίβεια πρόβλεψης. Το GPT 4 πέτυχε ποσοστό 60.55% έναντι 54.95% του λεξικού μοντέλου, επιβεβαιώνοντας τη δυνατότητα των LLMs να χειρίζονται πολυδιάστατα νοήματα και συμφραζόμενα με υψηλή απόδοση.

Ωστόσο, θα πρέπει να σημειωθεί ότι στόχος της εργασίας δεν ήταν η βελτιστοποίηση κάθε μοντέλου ξεχωριστά αλλά η αξιολόγηση και σύγκριση διαφορετικών προσεγγίσεων υπό ενιαίο και ελεγχόμενο πειραματικό πλαίσιο. Η περαιτέρω βελτίωση της ακρίβειας (ιδίως στα παραδοσιακά μοντέλα) είναι

εφικτή, μέσα από καλύτερο fine-tuning, επιλογή χαρακτηριστικών και ενίσχυση του training pipeline, αλλά ξεπερνά το εύρος και τη στόχευση της παρούσας μελέτης.

Η συνολική προσέγγιση που ακολουθήθηκε αναδεικνύει τη σημασία της αξιολόγησης πολλαπλών τεχνικών στην Επεξεργασία Φυσικής Γλώσσας, αλλά και την ανάγκη για ισορροπία μεταξύ ακρίβειας, ερμηνευσιμότητας και υπολογιστικής αποδοτικότητας. Η εμπειρία που αποκτήθηκε από την ανάλυση σχολίων προϊόντων μεθοδολογικά διαφορετικών εργαλείων, ενίσχυσε σημαντικά την κατανόηση της φύσης του προβλήματος της Ανάλυσης Συναισθήματος και δημιούργησε γόνιμο έδαφος για μελλοντική έρευνα σε ακόμη πιο ρεαλιστικά ή στοχευμένα περιβάλλοντα εφαρμογής.

Μελλοντικές επεκτάσεις της παρούσας εργασίας θα μπορούσαν να περιλαμβάνουν τη δοκιμή μεγαλύτερων και πιο εξειδικευμένων μετασχηματιστικών μοντέλων, όπως RoBERTa ή DistilBERT, καθώς και τη χρήση pre-trained embeddings ή multi-lingual παραλλαγών. Επιπλέον, θα είχε ενδιαφέρον η εφαρμογή της ανάλυσης συναισθήματος σε δεδομένα άλλου τύπου (π.χ. μέσα κοινωνικής δικτύωσης, ιατρικά reviews, πολιτικές ομιλίες), προκειμένου να εξεταστεί η γενικευσιμότητα των μοντέλων. Τέλος, η ενσωμάτωση εξωτερικών γνώσεων (π.χ. ontology-based sentiment reasoning) και η ανάπτυξη υβριδικών συστημάτων θα μπορούσαν να βελτιώσουν σημαντικά την ακρίβεια και προσαρμοστικότητα τέτοιων συστημάτων σε συνθήκες πραγματικού κόσμου.

Σε γενικότερο επίπεδο, η εργασία αυτή ανέδειξε τη χρησιμότητα της Ανάλυσης Συναισθήματος στη σύγχρονη ανάλυση δεδομένων, τόσο στον χώρο του μάρκετινγκ όσο και σε τομείς όπως η εξυπηρέτηση πελατών, η ψυχολογία και η πολιτική επιστήμη. Η δυνατότητα αυτόματης κατανόησης του συναισθηματικού τόνου χιλιάδων σχολίων ή κειμένων, η ομαδοποίηση και κατηγοριοποίηση αυτών, με τρόπο επεκτάσιμο και ερμηνεύσιμο, δημιουργεί προοπτικές για ταχύτερη λήψη αποφάσεων και πληρέστερη κατανόηση της συμπεριφοράς και των επιθυμιών των χρηστών. Το αποθετήριο κώδικα MATLAB που αναπτύχθηκε στο πλαίσιο αυτής της διπλωματικής μπορεί να αποτελέσει τη βάση για μελλοντικές μελέτες, που θα εφαρμόζουν πιο εξελιγμένες και προσαρμοσμένες τεχνικές βελτιστοποίησης, θα αξιοποιούν μεγαλύτερα μοντέλα και εκτενέστερα datasets. Με αυτόν τον τρόπο, η παρούσα εργασία δεν σηματοδοτεί τη λήξη, αλλά συνεισφέρει με τρόπο ουσιαστικό στη συνέχιση της έρευνας στον τομέα της Επεξεργασίας Φυσικής Γλώσσας και της Ανάλυσης Δεδομένων.

Βιβλιογραφία

1. Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of EMNLP. 2002; pp.79–86.

2. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv. 2002; 34(1): 1–47.

3. Breiman L. Random forests. Mach Learn. 2001; 45(1): 5–32.

4. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci. 2016; 374(2065): 20150202.

5. van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008; 9(Nov): 2579–2605.

6. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.

7. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS). 2017.

8. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.

9. Turc I, Chang MW, Lee K, Toutanova K. Well-read students learn better: On the importance of pre-training compact models. arXiv preprint arXiv:1908.08962. 2019.

10. Hutto CJ, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the International AAAI Conference on Web and Social Media. 2014; 8(1): 216–225.

11. OpenAI. GPT-4 Technical Report. OpenAI; 2023.

12. Bartzoka D. Παραγωγή Κειμένου με τη Χρήση Νευρωνικών Δικτύων [master's thesis]. Πανεπιστήμιο Μακεδονίας; 2022. https://dspace.lib.uom.gr/bitstream/2159/28514/4/BartzokaDimitraMsc2022.pdf

13. MathWorks. Text Data Preparation. [Internet]. https://www.mathworks.com/help/textanalytics/text-data-preparation.html

14. Glass Box Medicine. Measuring performance: The confusion matrix. [Internet]. 2019. https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

15. ResearchGate. Confusion matrix for multi-class classification. [Internet]. https://www.researchgate.net/

16. Kaggle. BoardGameGeek Reviews. [Internet]. https://www.kaggle.com/datasets/jvanelteren/boardgamegeek-reviews

17. MathWorks. Preprocess Text Data in Live Editor. [Internet]. https://www.mathworks.com/help/textanalytics/ug/preprocess-text-data-in-live-editor.html

18. Microsoft. Filter data in a range or table in Excel. [Internet]. https://support.microsoft.com/...

19. Microsoft. FILTER function in Excel. [Internet]. https://support.microsoft.com/...

20. c3.ai. What is a Classifier? [Internet]. https://c3.ai/glossary/data-science/classifier/

21. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng J. 2014; 5(4): 1093–1113.

22. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. J Mach Learn Res. 2003; 3: 993–1022.

23. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009; 45(4): 427–437.

24. Demšar J. Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res. 2006

25. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405: 442–451.

26. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960; 20(1): 37–46.

27. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bull. 1945; 1(6): 80–83.

28. Student WS. The probable error of a mean. Biometrika. 1908; 6(1): 1–25.

29. Witten I, Frank E, Hall M. Data Mining. Burlington, MA: Morgan Kaufmann; 2011. pp.101–103.

30. Opitz D, Maclin R. Popular ensemble methods: An empirical study. J Artif Intell Res. 1999; 11: 169–198.

31. Jurafsky D, Martin JH. Speech and Language Processing. Draft 3rd ed. Stanford University. https://web.stanford.edu/~jurafsky/slp3/

32. Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press; 1999.

33. Bird S, Klein E, Loper E. Natural Language Processing with Python. O'Reilly Media; 2009.

34. Wolf T, Debut L, Sanh V, et al. Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020). 2020; pp.38–45.

35. Powers DMW. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. J Mach Learn Technol. 2011.

36. Webster J, Kit C. Tokenization as the initial phase in NLP. In: Proceedings of the 14th conference on Computational linguistics; 1992; Nantes, France. Association for Computational Linguistics; p. 1106–1110.

37. Kumhar N, Pradhan R, Goyal S, Verma OP. A comprehensive survey of text preprocessing techniques for data mining and NLP. Arch Comput Methods Eng. 2023.

38. Grootendorst M. BERTopic: Neural topic modeling with class-based TF-IDF and BERT embeddings [Internet]. 2022. https://github.com/MaartenGr/BERTopic

39. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist. 2017;5:135–146.

40. Sun C, Qiu X, Xu Y, Huang X. How to fine-tune BERT for text classification? In: Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer; 2019. p. 194–206.

41. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692. 2019.39

42. Loria S, Keen P, Honnibal M, Yankovsky R, Karesh D, Dempsey E. TextBlob: Simplified Text Processing [Internet]. 2021. https://textblob.readthedocs.io/

Παράρτημα Ι – Κώδικες στο GitHub

cTsapakos/NLP-Thesis: This repository contains MATLAB code live scripts for sentiment analysis and topic modeling using customer review data. It includes a main script for full execution (main_code.mlx) and modular functions for each NLP model, designed for flexibility and reuse.