ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

Σχεδιασμός και Ανάπτυξη Εφαρμογής Ηλεκτρονικής Διακυβέρνησης για τη βελτίωση της ποιότητας υπηρεσιών

Αναστάσιος Σχίζας

Υποβάλλεται
για την εκπλήρωση των προϋποθέσεων λήψης
Μεταπτυχιακού Διπλώματος
στην ειδίκευση «Προηγμένα Πληροφοριακά Συστήματα»
του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”
στο
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Φεβρουάριος 2024

Επιβλέπουσα: Ανδριάνα Πρέντζα
Ακαδημαϊκή Θέση: Καθηγήτρια

Συγγραφέας Αναστάσιος Σχίζας

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή: Αναστάσιος Σχίζας

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας: Σχεδιασμός και Ανάπτυξη Εφαρμογής Ηλεκτρονικής Διακυβέρνησης για τη βελτίωση της ποιότητας υπηρεσιών

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών "Πληροφοριακά Συστήματα & Υπηρεσίες" του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις 29/2/2024 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπουσα (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς): Καθ. Ανδριάνα Πρέντζα

Μέλος Εξεταστικής Επιτροπής: Καθ. Δημοσθένης Κυριαζής

Μέλος Εξεταστικής Επιτροπής: Καθ. Μιχαήλ Φιλιππάκης

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Ο Αναστάσιος Σχίζας γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Σχεδιασμός και Ανάπτυξη Εφαρμογής Ηλεκτρονικής Διακυβέρνησης για τη βελτίωση της ποιότητας υπηρεσιών», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.
Επιπλέον *δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.*
Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο ΔΗΛΩΝ

Ονοματεπώνυμο: Αναστάσιος Σχίζας
Αριθμός Μητρώου: ΜΕ2149

Υπογραφή:

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σχεδιασμός και Ανάπτυξη Εφαρμογής Ηλεκτρονικής Διακυβέρνησης για τη βελτίωση της ποιότητας υπηρεσιών

ΑΝΑΣΤΑΣΙΟΣ ΣΧΙΖΑΣ

Α.Μ.: ΜΕ2149

ΠΕΡΙΛΗΨΗ

Η εργασία αυτή έχει σαν θέμα τη χρήση τεχνικών Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP) και Μηχανικής Μάθησης (Machine Learning - ML) για την κατηγοριοποίηση κειμένου (text classification) και την ανάλυση συναισθήματος (sentiment analysis), με στόχο τη δημιουργία μιας εφαρμογής ηλεκτρονικής διακυβέρνησης για δημόσιες υπηρεσίες. Θα εξεταστούν διαφορετικά μοντέλα ML για την κατηγοριοποίηση του κειμένου και την ανάλυση του συναισθήματος των χρηστών, σε ένα σύνολο δεδομένων που αποτελείται από κείμενα που σχετίζονται με αξιολογήσεις χρηστών και αφορούν διάφορες περιπτώσεις. Με αυτά τα δεδομένα θα εκπαιδευτούν τα μοντέλα και θα γίνει σύγκρισή τους για να επιλεχθεί ποιο ή ποια από αυτά έχουν καλά αποτελέσματα ώστε να χρησιμοποιηθούν στην εφαρμογή. Εν κατακλείδι η εφαρμογή αυτή θα μπορεί να παρέχει μέσω διαγραμμάτων αποτελέσματα για να παρακολουθείται η απόκριση του κοινού και να ανιχνεύεται η διάθεση και οι ανάγκες των πολιτών, έτσι ώστε αξιοποιηθούν από δημόσιες υπηρεσίες.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Συναισθηματική ανάλυση και κατηγοριοποίηση κειμένου με χρήση επεξεργασίας φυσικής γλώσσας.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: NLP, SENTIMENT ANALYSIS, TEXT CLASSIFICATION, MACHINE LEARNING, PYTHON

ABSTRACT

This work focuses on the use of Natural Language Processing (NLP) and Machine Learning (ML) techniques for text classification and sentiment analysis. The goal is to create an e-governance application for public services. Various machine learning models will be examined for text classification and user sentiment analysis using a dataset consisting of texts related to user reviews across different cases. The models will be trained on this dataset, compared, and the one(s) with the best results will be selected for integration into the application. In conclusion, the application will provide results through charts to monitor public response and detect the mood and needs of citizens, enabling utilization by public services.

SUBJECT AREA: Sentiment analysis and text categorization using natural language processing.

KEYWORDS: NLP, SENTIMENT ANALYSIS, TEXT CLASSIFICATION, MACHINE LEARNING, PYTHON

ΕΥΧΑΡΙΣΤΙΕΣ

Για τη διεκπεραίωση της παρούσας Πτυχιακής Εργασίας, θα ήθελα να ευχαριστήσω τη κ. Ανδριάνα Πρέντζα για τη συνεργασία και την καθοδήγηση της.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1. ΕΙΣΑΓΩΓΗ

1.1 Ηλεκτρονική διακυβέρνηση

Ηλεκτρονική διακυβέρνηση (eGovernment) ορίζεται η αξιοποίηση των Τεχνολογιών Πληροφορικής και Επικοινωνιών (ΤΠΕ) στις δημόσιες υπηρεσίες (κεντρικές και περιφερειακές, κεντρικής διοίκησης ή αυτοδιοίκησης), σε συνδυασμό με τις οργανωτικές αλλαγές και τις νέες δεξιότητες του προσωπικού [1]. Σκοπός του e-Government είναι η μεγαλύτερη συμμετοχή των πολιτών στη βελτίωση της ποιότητας των υπηρεσιών με την χρήση του διαδικτύου. Η Ελλάδα έχει διαμορφώσει μία στρατηγική e-Government μέσω του Υπουργείου Ψηφιακής Διακυβέρνησης, αναλύοντας τις κατευθύνσεις και τους στόχους για το μέλλον. Η γραφειοκρατία, η δυσπιστία των πολιτών προς τη Δημόσια Διοίκηση λόγω ελλιπούς αποτελεσματικότητας, οι καθυστερήσεις επικοινωνίας μεταξύ των διαφόρων υπηρεσιών, η χρήση χαρτιού, η υποχρεωτική σε πολλές περιπτώσεις ξεχωριστή επικοινωνία με κάθε φορέα για να συλλεχθούν τα απαραίτητα έγγραφα είναι μερικά μόνο από τα προβλήματα που μπορούν να λυθούν με τη χρήση ΤΠΕ [2]. Παραδείγματα που μπορούν να αναφερθούν είναι το gov.gr, η νέα ενιαία ψηφιακή πύλη της δημόσιας διοίκησης όπου πολίτες και επιχειρήσεις μπορούν να βρουν τις ψηφιακές υπηρεσίες που θέλουν εύκολα και γρήγορα [3]. Και η πιο πρόσφατη προσθήκη είναι το mAIgov, ένας ψηφιακός βοηθός που χρησιμοποιεί τεχνολογίες Τεχνητής Νοημοσύνης (Artificial Intelligence - AI) και Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing – NLP). Σκοπό έχει να διευκολύνει την επικοινωνία με τους πολίτες, κατανοώντας τα ερωτήματα που του υποβάλλουν, αναζητώντας τις σχετικές πληροφορίες και παρέχοντας απαντήσεις με απλό και κατανοητό τρόπο. Οι χρήστες έχουν τη δυνατότητα να υποβάλλουν ερωτήματα τόσο γραπτά όσο και προφορικά [4].

1.2 Στόχοι διπλωματικής εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη ενός NLP συστήματος το οποίο θα κάνει κατηγοριοποίηση και ανάλυση συναισθημάτων σε κείμενο, για τις ανάγκες μίας οντότητας δημόσιας διοίκησης, σε συνδυασμό με τη χρήση μοντέλων Μηχανικής Μάθησης (Machine Learning – ML). Με αυτή την ηλεκτρονική εφαρμογή θα μπορεί να βελτιωθεί η διαχείριση των σχέσεων με τους πολίτες, η κατανόηση των αναγκών των πολιτών καθώς και η λήψη αποφάσεων. Με την κατηγοριοποίηση γίνεται πιο εύκολη η γενικότερη διαχείριση και η οργάνωση της πληροφορίας επιτρέποντας πιο αποτελεσματική πρόσβαση στα δεδομένα. Με την αναγνώριση συναισθημάτων γίνεται πιο εύκολη η επίλυση προβλημάτων των πολιτών γρηγορότερα και αποτελεσματικότερα. Γενικότερα μπορεί να βελτιώσει την επικοινωνία μεταξύ της δημόσιας διοίκησης και των πολιτών, παρέχοντας πιο αποτελεσματικές υπηρεσίες. Επίσης μπορεί να οδηγήσει σε πιο αποτελεσματική διοίκηση και καλύτερη εξυπηρέτηση των πολιτών. Η χρήση της ηλεκτρονικής εφαρμογής θα μπορεί να είναι 24 ώρες το 24ωρο και 7 ημέρες την εβδομάδα, αποτελώντας ένα πρακτικό εργαλείο που μπορεί να χρησιμοποιηθεί ευρέως. Ελέγχοντας διαφορετικά μοντέλα ML και καταλήγοντας στο καταλληλότερο βάσει αποτελεσμάτων στόχος είναι η καλύτερη πρόβλεψη σε ένα πρόβλημα ταξινόμησης. Πιο συγκεκριμένα θα ελεγχθούν οι προεπιλεγμένες τιμές των αλγορίθμων και θα γίνει προσαρμογή παραμέτρων τους για να αυξηθεί το επίπεδο ακρίβειας (accuracy). Η ταξινόμηση θα γίνει στις ακόλουθες κατηγορίες:

- Οδικό Δίκτυο
- Αθλητισμός/Ψυχαγωγία
- Περιβάλλον
- Τουρισμός/Φιλοξενία

Για την συναισθηματική ανάλυση το πρόβλημα ταξινόμησης θα αφορά δύο περιπτώσεις:

- Θετική
- Αρνητική

Τα δεδομένα που χρησιμοποιήθηκαν συγκεντρώθηκαν μέσω της αρχικής εφαρμογής καταχώρησης, όπου δόθηκε πρόσβαση σε τυχαίους χρήστες καταγράφοντας τις απόψεις τους. Στη συνέχεια έγινε ταξινόμηση στις παραπάνω κατηγορίες χειροκίνητα. Τα κείμενα επεξεργάστηκαν και μετασχηματίστηκαν στη ρίζα της κάθε λέξης για να εισαχθούν μετά στα μοντέλα ML προς εκπαίδευσή τους.

Η εφαρμογή δίνει τη δυνατότητα στον διαχειριστή να δει τη βάση δεδομένων και να κάνει εξαγωγή σε αρχείο χρησιμοποιώντας φίλτρα ημερομηνίας, κατηγορίας και συναισθήματος για περεταίρω ανάλυση. Επιπλέον μπορεί να δει σε έξι διαφορετικά διαγράμματα τα συγκεντρωτικά αποτελέσματα και να φιλτράρει ανά ημερομηνία.

1.3 Δομή διπλωματικής εργασίας

Η διπλωματική εργασία αποτελείται από εφτά κεφάλαια. Το πρώτο κεφάλαιο περιέχει τη γενική περιγραφή του θέματος της εργασίας και περιλαμβάνει γενική αναφορά στο e-Government, τους στόχους και τη δομή της εργασίας.

Το δεύτερο κεφάλαιο περιλαμβάνει βιβλιογραφική ανασκόπηση, περιγράφοντας την NLP μαζί με τις κατηγορίες προσέγγισής της καθώς και την ML με τις κατηγορίες που υπάρχουν με βάση την μέθοδο εκπαίδευσης. Αναλύεται η ταξινόμηση κειμένου (Text Classification) με χρήση NLP και Ανάλυση Συναισθήματος (sentiment analysis). Στη συνέχεια παρουσιάζονται προοπτικές και παραδείγματα που αφορούν την Ευρωπαϊκή Ένωση (ΕΕ) σχετικά με AI, NLP και δημόσια διοίκηση που έχουν ερευνηθεί.

Στο τρίτο κεφάλαιο παρουσιάζεται η μεθοδολογία ανάλυσης, σχεδίασης, και ανάπτυξης του συστήματος. Περιγράφεται ο τρόπος συλλογής δεδομένων και αναλυτικά η δομή τους στη βάση δεδομένων. Αναλύεται ο τρόπος καθαρισμού και προεπεξεργασίας των δεδομένων καθώς και ο τρόπος εκπαίδευσης με χρήση έξι μοντέλων ML:

- ΝΑΪVE BAYES
- SVM
- DECISION TREES
- RANDOM FOREST
- K-NEAREST NEIGHBORS
- LOGISTIC REGRESSION

Στο τέλος του τρίτου κεφαλαίου αναλύονται οι μετρικές αξιολόγησης των μοντέλων.

Στο τέταρτο κεφάλαιο παρουσιάζονται τα εργαλεία και οι βιβλιοθήκες της python που χρησιμοποιήθηκαν για την υλοποίηση της εφαρμογής.

Στο πέμπτο κεφάλαιο παρουσιάζεται η τελική εφαρμογή και η εμφάνισή της στο χρήστη.

Στο έκτο κεφάλαιο αναλύονται τα αποτελέσματα της εκπαίδευσης των μοντέλων και τι συμπεράσματα βγήκαν από αυτά.

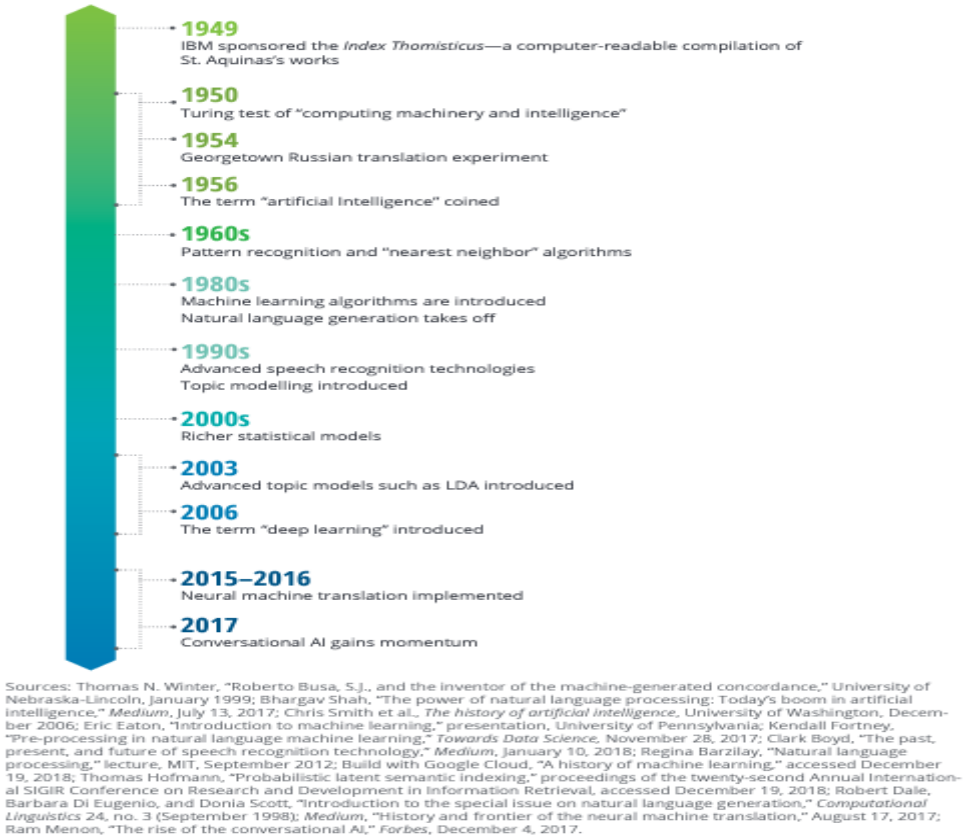
Στο έβδομο και τελευταίο κεφάλαιο παρουσιάζονται τα γενικά συμπεράσματα της εργασίας καθώς και μελλοντικές επεκτάσεις.

2. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

2.1 Επεξεργασία Φυσικής Γλώσσας

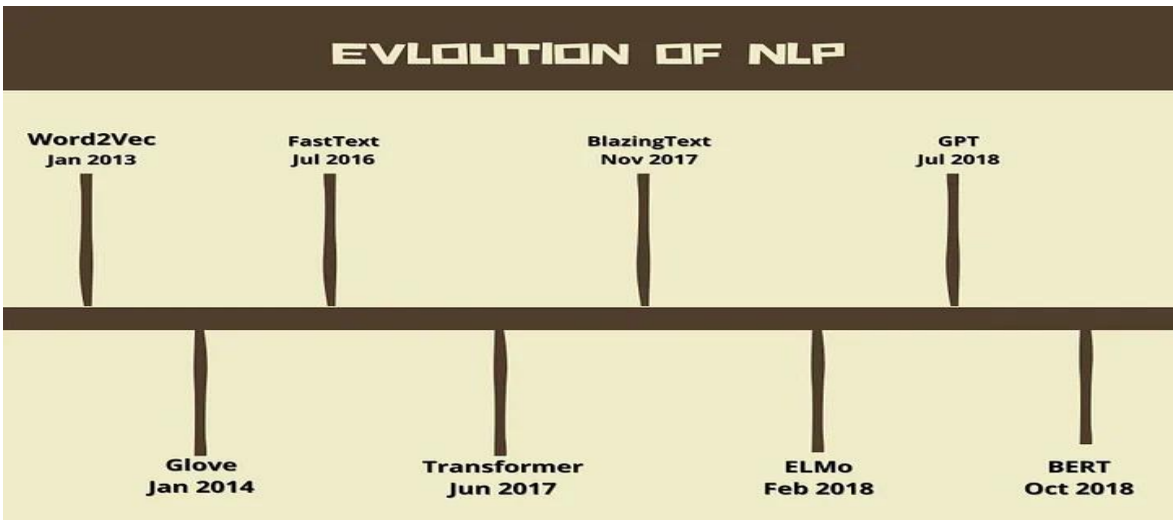
Η NLP είναι ένας τομέας έρευνας στην επιστήμη των υπολογιστών και στην AI που ασχολείται με την επεξεργασία φυσικών γλωσσών όπως αγγλικά ή μανδαρινικά. Αυτή η επεξεργασία γενικά περιλαμβάνει τη μετάφραση της φυσικής γλώσσας σε δεδομένα (αριθμούς) που μπορεί να χρησιμοποιήσει ένας υπολογιστής για να μάθει για τον κόσμο. Και αυτή η κατανόηση του κόσμου χρησιμοποιείται μερικές φορές για τη δημιουργία κειμένου φυσικής γλώσσας που αντικατοπτρίζει αυτή την κατανόηση [5]. Ένας πιο σύντομος ορισμός είναι ότι η NLP είναι η χρήση ανθρώπινων γλωσσών, όπως τα αγγλικά ή γαλλικά, μέσω υπολογιστή [6].

Η NLP είναι ένα κομμάτι της AI που ασχολείται με το πώς ένας ηλεκτρονικός υπολογιστής μπορεί να κατανοήσει μία γλώσσα όπως θα μπορούσε να το κάνει και ένας άνθρωπος, είτε μέσω του γραπτού είτε μέσω του προφορικού λόγου. Ιστορικά οι πρώτη προσπάθεια ξεκίνησε το 1949. Σήμερα έχει γίνει μία ‘επανάσταση’ με την τεχνολογία νευρωνικών δικτύων GPT (Generative Pre-trained Transformer) της OpenAI. Το μοντέλο GPT-3, για παράδειγμα, αποτελείται από 175 δισεκατομμύρια παραμέτρους, και αποτελεί ένα από τα μεγαλύτερα γνωστά νευρωνικά δίκτυα.



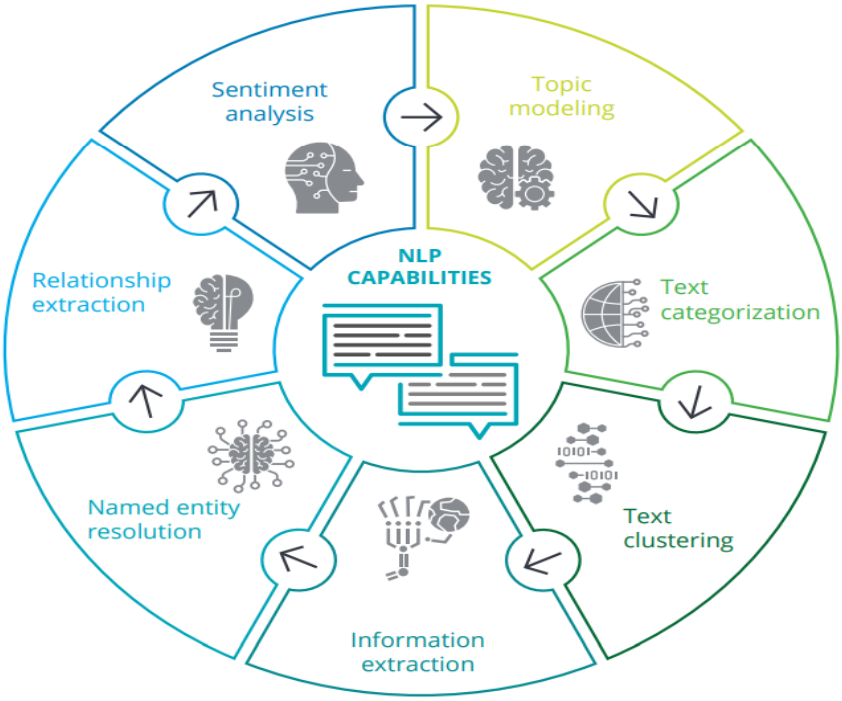
Εικόνα 1: Ιστορική αναδρομή και εξέλιξη της Επεξεργασίας Φυσικής Γλώσσας [7]

Στην εικόνα 1 υπάρχει μία ιστορική αναδρομή από το 1949 και την εξέλιξη της NLP με συγκεκριμένα γεγονότα και τη χρονιά που έγιναν.



Εικόνα 2: Μοντέλα Επεξεργασίας Φυσικής Γλώσσας με χρονολογική σειρά έκδοσης [8]

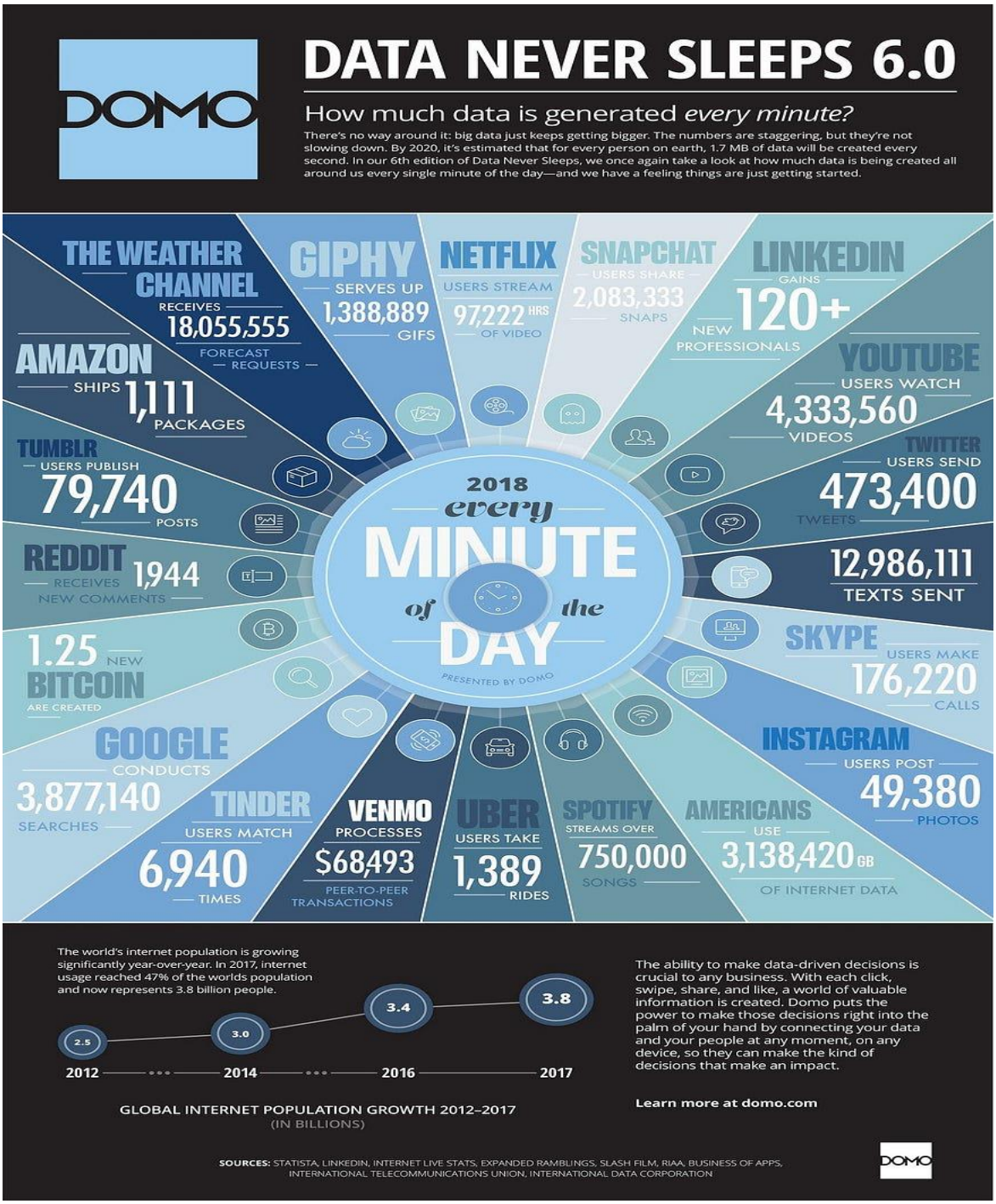
Στην εικόνα 2 φαίνονται διάφορα μοντέλα που χρησιμοποιούνται στην NLP και τη χρονολογική σειρά έκδοσής τους. Μερικές από τις δυνατότητες της NLP που βρίσκονται στην εικόνα 3 είναι η αναγνώριση ομιλίας, αποσαφήνιση της έννοιας μίας λέξης, ανάλυση συναισθήματος, περίληψη κειμένου, μετάφραση γλώσσας, ρομπότ συνομιλίας, εικονικοί πράκτορες, πρόβλεψη κειμένου, αναλύσεις κειμένου, ταξινόμηση κειμένου, μοντελοποίηση θεμάτων, εξαγωγή σχέσεων, ονομαστική ανάλυση οντότητας (named entity resolution), εξαγωγή πληροφορίας, ομαδοποίηση κειμένου.



Εικόνα 3: Τεχνικές και διαδικασίες επεξεργασίας κειμένου [9]

Η NLP συνδυάζει την υπολογιστική γλωσσολογία που βασίζεται σε κανόνες μοντελοποίησης της ανθρώπινης γλώσσας, με στατιστικά μοντέλα και μοντέλα ML και βαθιάς μάθησης (Deep Learning - DL). Μαζί, αυτές οι τεχνολογίες επιτρέπουν στους

υπολογιστές να επεξεργάζονται την ανθρώπινη γλώσσα σε μορφή κειμένου ή φωνητικών δεδομένων και να "κατανοούν" το πλήρες νόημά της, συμπεριλαμβανομένης της πρόθεσης και των συναισθημάτων του ομιλητή ή του συγγραφέα [10].



Εικόνα 4: Αποτελέσματα έρευνας παραγωγής δεδομένων ανά λεπτό [11]

Στις μέρες μας “παράγονται” τεράστιες ποσότητες δεδομένων κάθε λεπτό, εκατομμύρια κάθε μέρα. Εξαιτίας αυτής της τεράστιας ποσότητας δεδομένων και πολλών από αυτά σε μορφές μη δομημένες η NLP μπορεί να βοηθήσει σημαντικά στην κατανόηση αυτών των δεδομένων και στην αποκρυπτογράφησή τους. Στην εικόνα 4 υπάρχουν διαφορετικά παραδείγματα με το πόσα δεδομένα παράγονται σε κάθε λεπτό της ημέρας σύμφωνα με έρευνα της εταιρίας domo.com.

2.1.1. Προσεγγίσεις στην NLP

Υπάρχουν γενικά τρεις τύποι προσεγγίσεων στην NLP.

- Συστήματα βασισμένα σε κανόνες (Rule-based Systems)
- Συστήματα βασισμένα σε ML.
- Συστήματα βασισμένα σε DL με χρήση Νευρωνικών Δικτύων (Neural Networks - NN)

Τα συστήματα βασισμένα σε κανόνες αποτελούν μία από τις παλαιότερες μεθόδους στον τομέα της NLP. Χρησιμοποιούνται προκαθορισμένοι γλωσσικοί κανόνες για την ανάλυση και την επεξεργασία των δεδομένων, με βάση αυτούς τους κανόνες, για παράδειγμα για την αναγνώριση σημασιολογικών σχέσεων σε προτάσεις το σύστημα θα αναλύει τα δεδομένα, εξάγοντας πληροφορίες για τις σχέσεις μεταξύ λέξεων και φράσεων [12] [13].

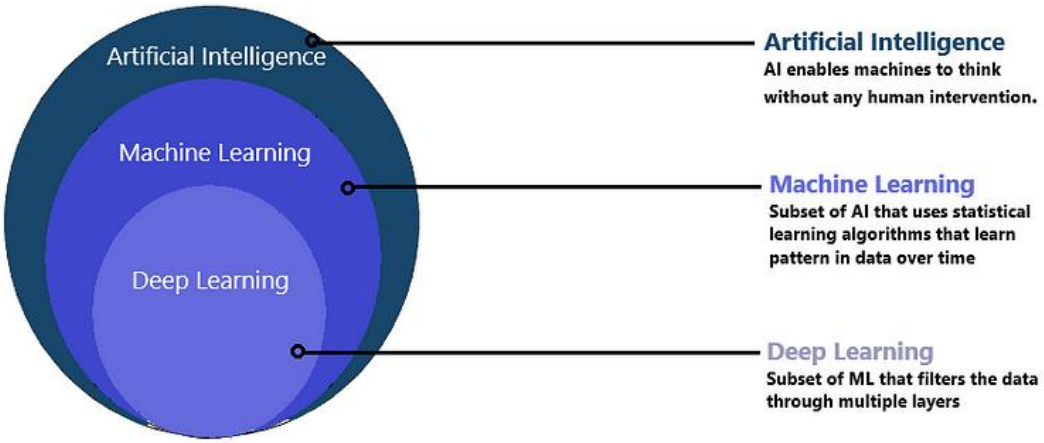
Τα συστήματα με χρήση ML περιλαμβάνουν την εκπαίδευση μοντέλων ML χρησιμοποιώντας αλγόριθμους. Παράδειγμα είναι η κατηγοριοποίηση email ως ανεπιθύμητο ή μη ανεπιθύμητο χρησιμοποιώντας αλγόριθμους ML [14].

Τα συστήματα βασισμένα σε DL, χρησιμοποιώντας NN όπου εκπαιδεύονται να αντιλαμβάνονται και να επεξεργάζονται φυσική γλώσσα με σύνθετες δομές. Κατά την εκπαίδευση, αυτά τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks - DNN), όπως τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks - RNN), τα Συνελικτικά Νευρωνικά Δίκτυα (Convolutional Neural Network - CNN) και τα Αναδρομικά Συνελικτικά Νευρωνικά Δίκτυα (Region Based Convolutional Neural Networks - RCNN), εξελίσσονται για να κατανοήσουν και να επεξεργαστούν τη φυσική γλώσσα. Παραδείγματα με χρήση NN είναι η αυτόματη μετάφραση, η δημιουργία φυσικού κειμένου, φιλτράρισμα ανεπιθύμητων μηνυμάτων, εξαγωγή πληροφοριών, σύνοψη κειμένου, σύστημα διαλόγου [15] [16].

Στα θετικά της NLP είναι ότι μπορεί να απαντάει σε ένα ερώτημα γρηγορότερα σε έναν χρήστη. Ο χρήστης λαμβάνει άμεσα απαντήσεις στις ερωτήσεις τους, ανεξαρτήτως ώρας ή ημέρας. Βοηθά τους υπολογιστές να επικοινωνούν με τους ανθρώπους στη γλώσσα τους. Μπορεί να χρησιμοποιηθεί για τον εντοπισμό πληροφοριών από μεγάλες βάσεις δεδομένων με μεγάλη ακρίβεια μειώνοντας τα κόστη [17].

Μειονεκτήματα τις NLP είναι ότι ενδέχεται να μην προσαρμόζεται εύκολα σε νέους τομείς και έχει περιορισμένη λειτουργία, γι' αυτό και κατασκευάζεται για έναν μόνο και συγκεκριμένο σκοπό. Η διαδικασία εκπαίδευσης ενδέχεται να απαιτήσει αρκετό χρονικό διάστημα. Σε περίπτωση που απαιτείται η ανάπτυξη ενός νέου μοντέλου χωρίς τη χρήση προεκπαιδευμένου μοντέλου, είναι πιθανό να απαιτηθούν εβδομάδες προτού επιτευχθεί υψηλό επίπεδο απόδοσης. Ένα άλλο μειονέκτημα είναι ότι η ML δεν είναι 100% αξιόπιστη και υπάρχει πάντα η πιθανότητα σφαλμάτων [18] [19].

2.2. Μηχανική Μάθηση



Εικόνα 5: Επίπεδα τεχνολογιών που περιλαμβάνει η τεχνητή νοημοσύνη [20]

Στην εικόνα 5 φαίνεται πώς δομείται από το γενικότερο πλαίσιο της AI, στα υποσύνολα της ML και της DL. Η ML ασχολείται με την κατασκευή αλγορίθμων για την κατανόηση δεδομένων ώστε ένας ηλεκτρονικός υπολογιστής να μπορεί με βάση αυτά τα δεδομένα να παίρνει αποφάσεις. Με τον όρο Μηχανική Μάθηση ορίζεται ένα σύστημα ικανό να αποκτά και να ενσωματώνει γνώση αυτόματα [21]. Είναι ο μηχανισμός μέσω του οποίου προσπαθούμε να φτιάξουμε μηχανές να μαθαίνουν χωρίς να τις προγραμματίζουμε ρητά να το κάνουν. Χρησιμοποιεί στατιστικές τεχνικές και μερικές φορές προηγμένους αλγορίθμους για να κάνει είτε προβλέψεις είτε να μάθει κρυφά μοτίβα μέσα από τα δεδομένα και ουσιαστικά αντικαθιστά συστήματα που βασίζονται σε κανόνες για να καταστήσει τα συστήματα που βασίζονται σε δεδομένα πιο ισχυρά [22].

Ένας άνθρωπος βλέποντας μερικές εικόνες ενός αντικειμένου μπορεί εύκολα μάθει να το ξεχωρίζει. Για έναν υπολογιστή όμως αυτό είναι ένα πολύπλοκο θέμα. Θα χρειαστεί να καταχωρηθούν στον υπολογιστή πάρα πολλές φωτογραφίες για να μπορέσει να ξεχωρίσει π.χ. ένα άλογο από μία αγελάδα. Η δυσκολία έγκειται στο ότι ο υπολογιστής χρειάζεται πάρα πολλά παραδείγματα για να μάθει τα χαρακτηριστικά εκείνα που του χρειάζονται για να πάρει την σωστή απόφαση. Βασικό χαρακτηριστικό για να δουλέψει σωστά η μηχανική μάθηση είναι τα δεδομένα. Τα δεδομένα όσο πιο σωστά παρουσιάζουν το θέμα που χρειάζεται να μάθει ο υπολογιστής τόσο πιο σωστή πρόβλεψη θα κάνει.

Εικόνα 6: Κατηγορίες μηχανικής μάθησης [23]

Η μηχανική μάθηση χωρίζεται σε τέσσερις κύριες κατηγορίες όπως φαίνεται στην εικόνα 6.

- Εποπτευόμενη Μηχανική Μάθηση (Supervised Machine Learning)
- Μη εποπτευόμενη Μηχανική Μάθηση (Unsupervised Machine Learning)
- Ημι-εποπτευόμενη Μηχανική Μάθηση (Semi-supervised Machine Learning)
- Ενισχυτική Μάθηση (Reinforcement Learning)

2.2.1 Επιβλεπόμενη Μηχανική Μάθηση

Η εποπτευόμενη μηχανική μάθηση είναι όταν υπάρχει ένα σύνολο δεδομένων και είναι γνωστό από την αρχή το αποτέλεσμα, το οποίο είναι μέσα στα δεδομένα μας. Όταν υπάρχει ένα σύνολο ανεξάρτητων δεδομένων και αναζητείται η εξαρτημένη μεταβλητή. Εκπαιδεύεται δηλαδή το μοντέλο με τα δεδομένα που υπάρχουν ώστε να μάθει να βρίσκει το επιθυμητό αποτέλεσμα, το οποίο είναι ήδη γνωστό από τα δεδομένα. Με αυτό τον τρόπο όταν εισάγονται στο μοντέλο καινούργια δεδομένα το ιδανικό σενάριο θα ήταν να μπορεί να κάνει προβλέψεις. Οι μεταβλητές που χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου έχουν ετικέτα και αυτή είναι η βασική διαφορά με την μη εποπτευόμενη μηχανική μάθηση. Για παράδειγμα στον πίνακα 1 γνωρίζοντας το ύψος και το βάρος, σαν ανεξάρτητες μεταβλητές, μπορεί να προβλεφθεί η ηλικία ενός ανθρώπου, σαν εξαρτημένη μεταβλητή.

Πίνακας 1 : Παράδειγμα συνόλου δεδομένων (ύψος-βάρος-ηλικία)

HEIGHT(cm)	WEIGHT(kg)	AGE
155	61,6	18
167	72,5	19
180	87,9	35

2.2.2 Μη Επιβλεπόμενη Μηχανική Μάθηση

Η μη εποπτευόμενη μηχανική μάθηση είναι όταν υπάρχουν δεδομένα χωρίς να υπάρχει κάποια εξαρτημένη μεταβλητή. Εκπαιδεύεται το μοντέλο χωρίς να είναι γνωστή από πριν η σωστή απάντηση. Γίνεται προσπάθεια να ανακαλυφθούν κρυφά μοτίβα ή ομοιότητες μεταξύ των δεδομένων για να τα κατηγοριοποιηθούν. Για παράδειγμα υποθέτοντας ότι, για δεδομένα υπάρχουν εικόνες που αφορούν ζώα, όπως γάτα, σκύλο, άλογο, λιοντάρι, ελέφαντας. Το μοντέλο θα προσπαθήσει να κατηγοριοποιήσει τις εικόνες χωρίς να γνωρίζει για ποιο ζώο είναι η κάθε εικόνα αναγνωρίζοντας τις ομοιότητες και θα προσπαθήσει να φτιάξει ομάδες ζώων.

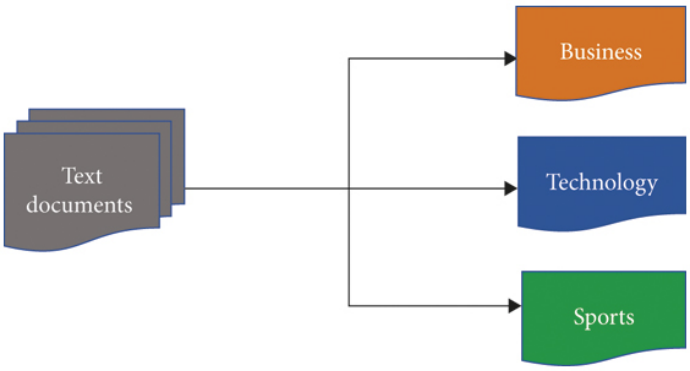
2.2.3 Ημι-Επιβλεπόμενη Μηχανική Μάθηση

Η ημι-εποπτευόμενη μηχανική μάθηση είναι κάτι ενδιάμεσα σε εποπτευόμενη και μη εποπτευόμενη μηχανική μάθηση. Υπάρχουν δηλαδή αρκετά δεδομένα χωρίς ετικέτες και λιγότερα δεδομένα με ετικέτες. Για παράδειγμα σε δεδομένα που αφορούν εικόνες ανθρώπων. Ένα μικρό υποσύνολο αυτών έχουν ετικέτες με την ηλικία τους. Το μοντέλο από τις λίγες εικόνες που έχουν ετικέτα με τις ηλικίες θα βρει τα χαρακτηριστικά που αφορούν την ηλικία και θα τα χρησιμοποιήσει στις εικόνες χωρίς ετικέτα για να προβλέψει την ηλικία στο υπόλοιπο σύνολο δεδομένων.

2.2.4 Ενισχυτική Μάθηση

Η ενισχυτική μάθηση ασχολείται με τον τρόπο μάθησης ενός αλγορίθμου να προσαρμόζεται σε ένα περιβάλλον μέσω δοκιμών και σφαλμάτων λειτουργώντας με ένα σύστημα ανταμοιβής. Ο αλγόριθμος λαμβάνει ενέργειες από το περιβάλλον, παίρνει ανταμοιβές ή ποινές και προσπαθεί να εντοπίσει τη στρατηγική που θα του επιτρέπει να μεγιστοποιήσει τη συνολική ανταμοιβή που θα λάβει. Για παράδειγμα θα ήταν εφικτό να αναλυθεί πώς ο αλγόριθμος θα μάθαινε να παίζει σκάκι. Ο αλγόριθμος μπορεί να λάβει μία τρέχουσα κατάσταση του παιχνιδιού (θέση των πιονιών, τον αριθμό των κινήσεων που έχουν γίνει κ.λπ.) και να εξετάσει τις πιθανές κινήσεις που μπορεί να πραγματοποιήσει. Μέσω προσομοιώσεων, ο αλγόριθμος μπορεί να αξιολογήσει την ποιότητα κάθε κίνησης και να επιλέξει αυτή που θεωρεί τη βέλτιστη. Καθώς παίζει περισσότερα παιχνίδια, ο αλγόριθμος μπορεί να μάθει τις βέλτιστες στρατηγικές για διάφορες καταστάσεις και να βελτιώσει την απόδοσή του στο παιχνίδι. Με τον χρόνο, μπορεί ακόμα να αντιληφθεί πιο σύνθετα μοτίβα και στρατηγικές που μπορούν να οδηγήσουν σε επιτυχημένα αποτελέσματα στο σκάκι.

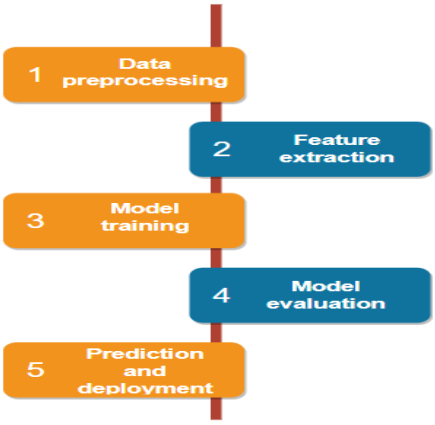
2.3 Κατηγοριοποίηση Κειμένου με χρήση NLP και Ανάλυση Συναισθήματος



Εικόνα 7: Παράδειγμα κατηγοριοποίησης σε τρεις κατηγορίες [24]

Η κατηγοριοποίηση κειμένου στη μηχανική μάθηση αφορά την αυτόματη κατηγοριοποίηση ενός κειμένου σε μία προκαθορισμένη κατηγορία, όπως το παράδειγμα της εικόνας 7. Μπορεί να εκπαιδευτεί ένα μοντέλο στην εκμάθηση μοτίβων στα δεδομένα ώστε να μπορεί να προβλέπει τη σωστή κατηγορία σε ένα καινούργιο κείμενο. Μερικές από τις περιπτώσεις που μπορούν να χρησιμοποιηθούν για την κατηγοριοποίηση κειμένου είναι η ανάλυση συναισθημάτων, το φιλτράρισμα ανεπιθύμητων μηνυμάτων, η κατηγοριοποίηση θεμάτων, η ταξινόμηση εγγράφων.

Για να κάνουμε κατηγοριοποίηση κειμένου γενικά έχουμε πέντε βήματα που φαίνονται στην εικόνα 8:



Εικόνα 8 : Βήματα για κατηγοριοποίηση κειμένου

Στο πρώτο βήμα περιλαμβάνεται η προεπεξεργασία δεδομένων (data preprocessing), όπου γίνεται η επεξεργασία των δεδομένων κειμένου. Περιλαμβάνει τον καθαρισμό, μετασχηματισμό και την προετοιμασία των δεδομένων, ώστε αμέσως μετά να μπορούν να χρησιμοποιηθούν στην εκπαίδευση του μοντέλου. Γίνεται καθαρισμός από τα κείμενα περιττών χαρακτήρων, όπως σημεία στίξης ή ειδικά σύμβολα ή html tags και μετατροπή όλων των κειμένων σε πεζά γράμματα. Μετά γίνεται το tokenization που περιλαμβάνει το διαχωρισμό του κειμένου σε μεμονωμένες λέξεις ή μικρότερες μονάδες που ονομάζονται tokens. Αφαιρώντας ενδιάμεσες λέξεις που δεν έχουν νόημα (π.χ. "και", "το", "είναι"), έτσι ελαχιστοποιείται ο λεγόμενος “θόρυβος”. Στη συνέχεια κάνουμε lemmatization όπου μετατρέπουμε τις λέξεις των κειμένων στη αρχική τους ρίζα.

Στο δεύτερο βήμα είναι η διαδικασία μετατροπής των δεδομένων σε αριθμητική αναπαράσταση (vectorization). Μερικές από τις τεχνικές vectorization είναι One-Hot Encoding, Count Vectorization, TF-IDF (Term Frequency-Inverse Document Frequency), Word Embeddings, Sentence or Document Embeddings.

Τρίτο βήμα είναι η εκπαίδευση του μοντέλου με τα δεδομένα κειμένου τα οποία έχουν συγκεκριμένες κατηγορίες που έχουν επιλεχθεί. Μερικοί αλγόριθμοι που χρησιμοποιούνται είναι Naive Bayes, Support Vector Machines (SVM), decision trees, random forests, καθώς και DL models όπως CNN ή RNN.

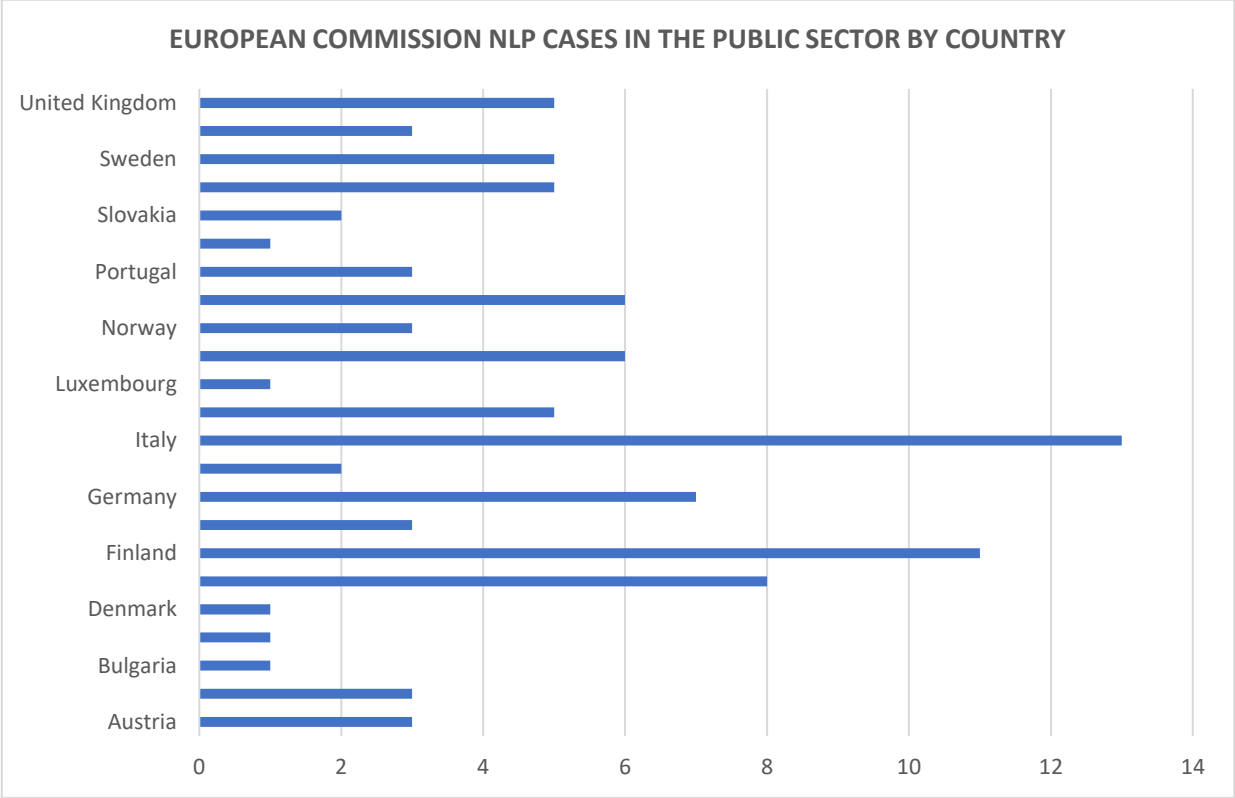
Το τέταρτο βήμα αφορά την αξιολόγηση του μοντέλου που μπορεί να περιλαμβάνει μετρήσεις όπως accuracy, ακρίβεια (precision), ανάκληση (recall), F1-score. Ανάλογα με τα δεδομένα που υπάρχουν, η αξιολόγηση μπορεί να περιλαμβάνει και ανθρώπινη αξιολόγηση. Η ανθρώπινη αξιολόγηση μπορεί να προσφέρει πολύτιμες γνώσεις, ειδικά για εργασίες που απαιτούν υποκειμενική κρίση.

Το πέμπτο βήμα αφορά την τελική πρόβλεψη του μοντέλου, μετά την εκπαίδευση και την αξιολόγησή του, όπου μπορεί πλέον να χρησιμοποιηθεί για να παράγει αποτελέσματα προβλέψεων σε κατηγοριοποίηση κειμένου.

2.3.1 Ανάλυση Συναισθήματος

Η συναισθηματική ανάλυση αφορά στον εντοπισμό αρνητικών ή θετικών συναισθημάτων σε ένα κείμενο. Μπορεί επίσης να ανιχνεύσει συναισθήματα όπως η χαρά, ο θυμός, η λύπη κ.λπ.. ή πια είναι η πρόθεση π.χ. (ενδιαφέρομαι ή δεν ενδιαφέρομαι). Ουσιαστικά είναι ένα εργαλείο ταξινόμησης κειμένου. Η σημαντικότητα της ανάλυσης συναισθήματος έγκειται στο ότι μπορεί να οδηγήσει στην βαθύτερη κατανόηση για το πως αισθάνεται ένας χρήστης, είτε είναι πελάτης ή πολίτης ή απλά κάποιος που χρησιμοποιεί μια υπηρεσία. Αυτό είναι πολύ σημαντικό για τη λήψη αποφάσεων και τη βελτίωση της ικανοποίησης των χρηστών.

2.4 Τεχνητή νοημοσύνη & επεξεργασία φυσικής γλώσσας στις δημόσιες υπηρεσίες στην ΕΕ



Εικόνα 9: Περιπτώσεις NLP στο δημόσιο τομέα ανά χώρα από την Ευρωπαϊκή Επιτροπή [25]

Η ΕΕ αναγνωρίζοντας ότι η τεχνητή νοημοσύνη θα έχει κεντρικό ρόλο στην καθημερινότητα των πολιτών και βλέποντας ήδη αυτές τις προοπτικές αλλά και για το μέλλον, έχει στοχεύσει να γίνει μία παγκόσμια ηγέτιδα δύναμη. Στον δημόσιο τομέα η τεχνητή νοημοσύνη έχει ήδη κεντρικό ρόλο και αυτό σίγουρα θα αυξηθεί στο άμεσο μέλλον αλλά και στο πιο βραχυπρόθεσμο. Επιπλέον, τα κράτη μέλη συχνά επισημαίνουν πρωτοβουλίες που σχετίζονται με τον δημόσιο τομέα στις εθνικές τους στρατηγικές [26].

Στην εικόνα 9 φαίνεται η κατανομή περιπτώσεων που αφορούν ηλεκτρονική διακυβέρνηση και συγκεκριμένα σε επεξεργασία φυσικής γλώσσας ανά χώρα.

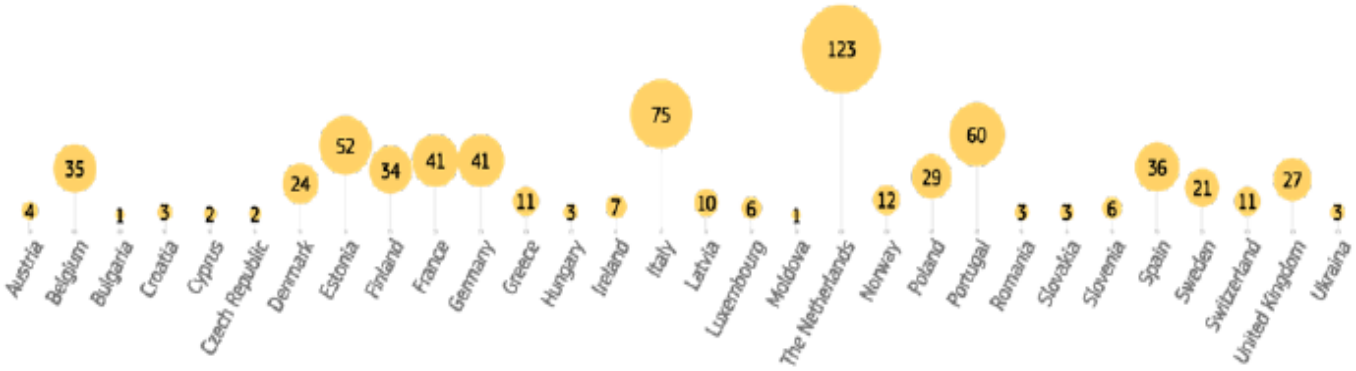
Η ΕΕ ειδικότερα, στοχεύει να αναπτύξει «trusted AI» βασισμένη σε πραγματικά ευρωπαϊκές ηθικές και κοινωνικές αξίες που έχουν δανειστεί από τον Ευρωπαϊκό Χάρτη Θεμελιωδών Δικαιωμάτων. Για το σκοπό αυτό, βασιζόμενη στη δήλωση συνεργασίας για την τεχνητή νοημοσύνη που εγκρίθηκε από όλα τα κράτη μέλη της ΕΕ, στη Νορβηγία και την Ελβετία στις 10 Απριλίου 2018, η ανακοίνωση «Artificial Intelligence for Europe» της 25ης Απριλίου 2018 πρότεινε μια στρατηγική για την τεχνητή νοημοσύνη για την Ευρώπη, η οποία εγκρίθηκε από το Ευρωπαϊκό Συμβούλιο τον Ιούνιο του 2018 [27].

Στην εικόνα 10 [28] φαίνονται τρέχουσες περιπτώσεις τεχνητής νοημοσύνης που αφορούν ηλεκτρονική διακυβέρνηση, αλλά και περιπτώσεις με τεχνολογικές προοπτικές.

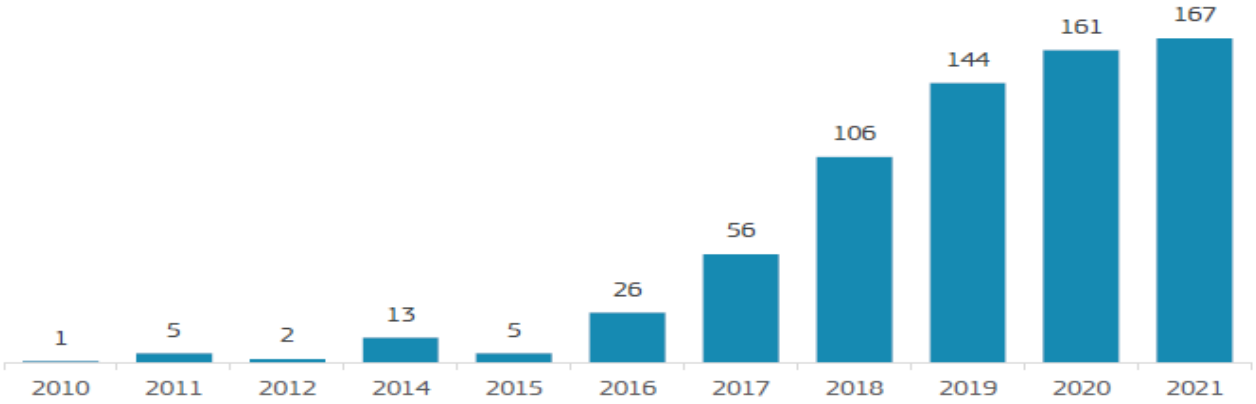
AI typology	Description	Example	No. of cases
Audio Processing	These AI applications are capable of detecting and recognizing sound, music and other audio inputs, including speech, thus enabling the recognition of voices and transcription of spoken words.	Corti in Denmark is used to process the audio of emergency calls in order to detect whether the caller could have a cardiac arrest	8
Chatbots, Intelligent Digital Assistants, Virtual Agents and Recommendation Systems	This AI typology includes virtualised assistants or online 'bots' currently used in not only to provide generic advice but also behaviour related recommendations to users.	In Latvia, the Chatbot UNA is used to help answer frequently asked questions regarding the process of registering a company	52
Cognitive Robotics, Process Automation and Connected and Automated Vehicles	The common trait of these AI technologies is process automation, which can be achieved through robotized hardware or software	The use of self-driving snowploughs in an airport in Norway in order to improve the clearing of snow on runways.	16
Computer Vision and Identity Recognition	AI applications from this list category use some form of image, video or facial recognition to gain information on the external environment and/or the identity of specific persons or objects.	In Estonia, the SATIKAS system is in used which is capable of detecting mowed (or the lack of mowed) grasslands on satellite imagery	29
Expert and Rule-based Systems, Algorithmic Decision Making	The reason why these apparently distant AI developments are joined into a single application is their prevalent orientation to facilitate or fully automate decision making processes of potential relevance not only to the private but also to the public sector.	Nursery child recruitment system used in Warsaw. The algorithm considers data provided by parents during the registration, calculates the score and automatically assigns children into individual nurseries.	29
AI-empowered Knowledge Management	The common element here is the underlying capacity of embedded AI to create a searchable collection of case descriptions, texts and other insights to be shared with experts for further analysis.	In Slovakia, an AI system is used in the government to assist in the browsing and finding of relevant semantic data	12
Machine Learning, Deep Learning	While almost all the other categories of AI use some form of Machine Learning, this residual category refers to AI solutions which are not suitable for the other classifications.	In Czechia, AI is used in social services to facilitate citizens to stay in their natural environment for as long as possible	17
Natural Language Processing, Text Mining and Speech Analytics	These AI applications are capable of recognising and analysing speech, written text and communicate back.	In Dublin, an AI system analyses citizen opinions in the Dublin Region for an overview of their most pressing concerns by analysing local twitter tweets with various algorithms.	19
Predictive Analytics, Simulation and Data Visualisation	These AI solutions learn from large datasets to identify patterns in the data that are consequently used to visualise, simulate or predict new configurations.	Since 2012, the Zurich City Police have been using software that predicts burglaries. Based on these predictions, police could be forwarded to check these areas and limit burglaries from happening.	37
Security Analytics and Threat Intelligence	These refer to AI systems which are tasked with analysing and monitoring security information and to prevent or detect malicious activities.	In the Norwegian National Security Authority a new system is used based on machine learning is enabling the automatic analysis of any malware detected to improve cybersecurity	11

Εικόνα 10 : Τρέχουσες περιπτώσεις τεχνητής νοημοσύνης μέσω της Ε.Ε.

Σε έρευνα που δημοσιεύτηκε στην ΕΕ όπου φαίνεται στην εικόνα 11, πόσες περιπτώσεις τεχνητής νοημοσύνης εξετάστηκαν και στην εικόνα 12 την ιστορική εξέλιξη των περιπτώσεων [29], όπου φαίνεται ότι η ανάπτυξη περιπτώσεων τεχνίτης νοημοσύνης που αφορούν ηλεκτρονική διακυβέρνηση έχει εκθετική αύξηση ειδικά τα τελευταία έτη.

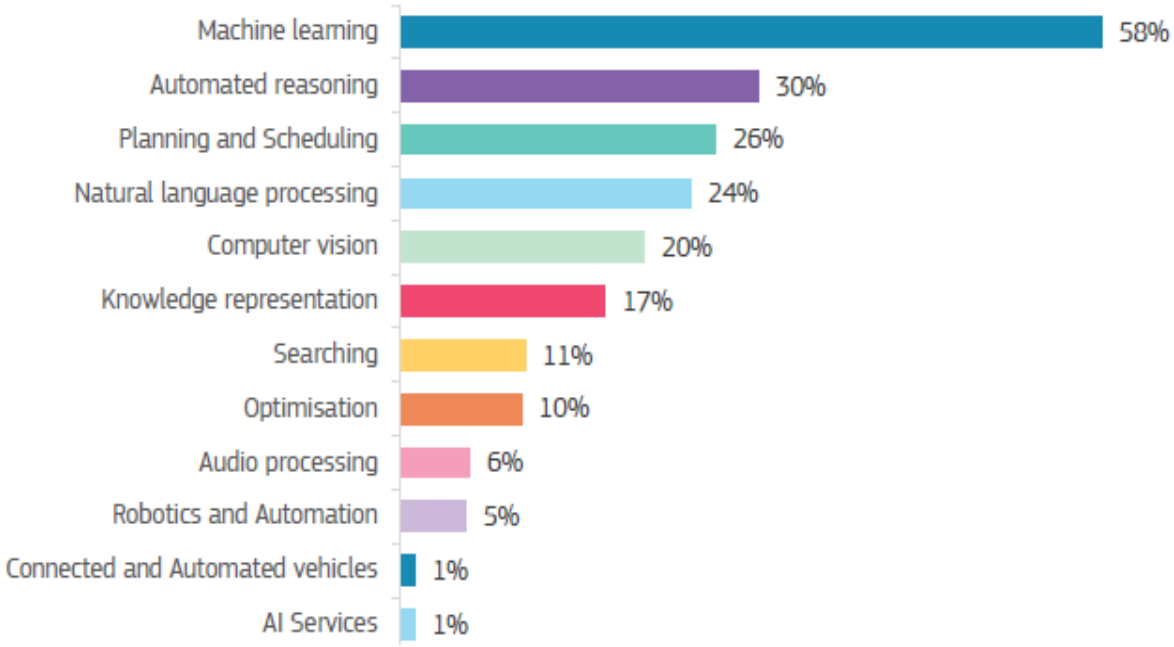


Εικόνα 11: Αριθμός ερευνών ανά χώρα Ε.Ε.



Εικόνα 12: Συνολικός αριθμός ερευνών Ε.Ε. τεχνητής νοημοσύνης ανά έτος

Στην εικόνα 13 [30] φαίνεται η κατανομή σε διαφορετικά πεδία της τεχνητής νοημοσύνης. Το μεγαλύτερο ποσοστό αφορά μηχανική μάθηση και φαίνεται ότι η επεξεργασία φυσικής γλώσσας είναι σε ποσοστό 24%.



Εικόνα 13: Ποσοστό ερευνών σε διαφορετικά επίπεδα τεχνητής νοημοσύνης

2.4.1 Παραδείγματα επεξεργασίας φυσικής γλώσσας σε δημόσια διοίκηση

Το Υπουργείο Ψηφιακής Μετάβασης και Τοποθεσίας Επιχειρήσεων (BMDW) της Αυστρίας δημιούργησε το chatbot "Mona" το οποίο μπορεί αυτόματα να απαντάει σε ερωτήσεις εταιριών σχετικές με την κρίση του κορονοϊού και την οικονομία (π.χ. θέματα που αφορούν επιδοτήσεις, θέματα εργατικού δικαίου όπως η μερική απασχόληση και η τηλεργασία).

Στο Βέλγιο αναπτύχθηκε σύστημα NLP το οποίο χρησιμοποιείται για την αυτόματη ταξινόμηση εισερχόμενων τηλεφωνικών ερωτήσεων μέσω ενός τηλεφωνικού κέντρου. Σκοπός του συστήματος είναι να ταξινομεί πολύ γρήγορα τις ερωτήσεις με σκοπό οι απαντήσεις να προωθούνται πολύ πιο γρήγορα και αν είναι εφικτό να δώσει και κάποια τυποποιημένη απάντηση για ακόμα ταχύτερη εξυπηρέτηση.

Στη Ρουμανία στην πόλη Cluz αναπτύχθηκε σύστημα με το όνομα Antonia, το οποίο αναλαμβάνει την αυτόματη επεξεργασία 64 φορμών για αιτήσεις δημόσιων υπηρεσιών και απαντά σε ερωτήσεις σχετικά με τον τρόπο συμπλήρωσης των φορμών.

Στη Φιλανδία η κεντρική βιβλιοθήκη Oodi, έφτιαξε το Obotti, ένα chatbot που προτείνει βιβλία σύμφωνα με τα ενδιαφέροντα και την ανατροφοδότηση των πελατών. Η υπηρεσία είναι χωρισμένη θεματικά σε έξι chatbots, το καθένα προτείνει περιεχόμενο ανάλογα με το θέμα του [25].

Στην Ισπανία αναπτύχθηκε από την εθνική αστυνομία το VeriPol. Η χρησιμότητα του συστήματος είναι ότι βρίσκει γρήγορα τις ψευδείς αναφορές και επιτρέπει τους πόρους τις αστυνομίας να είναι διαθέσιμοι σε άλλες εργασίες και αναφορές, ενώ ταυτόχρονα αποτρέπει τους ανθρώπους να συντάσσουν ψευδείς αναφορές εξ' αρχής. Επιπλέον μπορεί να παρέχει περισσότερη γνώση για το πως οι άνθρωποι λένε ψέματα στους αστυνομικούς και να αποκτήσουν γνώση για τον εντοπισμό αληθινών και ψευδών αναφορών [31].

Η εργασία αυτή με τη χρήση της ανάλυσης συναισθημάτων που απορρέει από τον γραπτό λόγο, προσπαθεί με τη δημιουργία ενός μοντέλου να αξιολογήσει τις υπηρεσίες ηλεκτρονικής διακυβέρνησης που βασίζονται στο Web, αλλά και να βοηθήσει στην κατανόηση γιατί οι ιστοσελίδες των κυβερνητικών υπηρεσιών που πετυχαίνουν ή αποτυγχάνουν να βοηθήσουν τους πολίτες να βρουν τις απαραίτητες πληροφορίες [32].

Στη εργασία αυτή εξετάστηκε η ανάγκη ενός Συστήματος Υποστήριξης Αποφάσεων (Decision Support System - DSS) βασισμένου σε συγκέντρωση δεδομένων (text mining) για κυβερνητικούς φορείς. Παρουσιάζονται διάφορες εφαρμογές text mining, προτείνοντας μια αρχιτεκτονική για την ανάπτυξη του συστήματος, και παρουσιάζεται ένα ολοκληρωμένο πλαίσιο που μπορεί να χρησιμοποιηθεί από κυβερνητικούς οργανισμούς για τη δημιουργία DSS βασισμένου σε text mining [33].

Στην εργασία αυτή παρουσιάζεται ένα 'VIRTUAL E-GOV WEB SITE' όπου θα εισάγονται οι απόψεις των πολιτών. Στη συνέχεια γίνεται επεξεργασία και κατηγοριοποίηση και δημιουργεί σχετικές και συνοπτικές πληροφορίες για τους αρμόδιους ώστε τελικά να γνωρίζουν πού χρειάζεται βελτίωση και πού διατηρείται η υψηλή ποιότητα υπηρεσιών. Το σύστημα κατηγοριοποιεί, από ένα σύνολο 120 εγγραφών, τις απόψεις των πολιτών σε κάθε μία από τις έξι κατηγορίες υπηρεσιών με accuracy 87,3% και recall 85,8%. Επιπλέον, το σύστημα καθορίζει τις απόψεις ως εκτίμηση με accuracy 100% για και τις δύο περιπτώσεις θετική ή αρνητική και μέση ανάκληση 90,85% και για τις δύο περιπτώσεις. [34]

Βασικός στόχος της παρούσας διπλωματικής εργασίας είναι να επιτευχθεί μεγαλύτερο accuracy στην κατηγοριοποίηση κειμένου με τη χρήση και σύγκριση

διαφορετικών αλγορίθμων ML έτσι ώστε να μπορεί να εφαρμοστεί σε πρακτικές εφαρμογές στον πραγματικό κόσμο στην ηλεκτρονική διακυβέρνηση.

3. ΜΕΘΟΔΟΛΟΓΙΑ ΑΝΑΛΥΣΗΣ, ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΑΝΑΠΤΥΞΗΣ ΣΥΣΤΗΜΑΤΟΣ

Στα πλαίσια της διπλωματικής εργασίας αναπτύχθηκε διαδικτυακή εφαρμογή ηλεκτρονικής διακυβέρνησης. Η εφαρμογή χρησιμοποιεί NLP καθώς και ML. Μετά την εισαγωγή κειμένου από κάποιον χρήστη η εφαρμογή κατηγοριοποιεί αυτόματα το κείμενο σε μία κατηγορία που έχει οριστεί και επίσης πραγματοποιεί ανάλυση συναισθήματος στο ίδιο κείμενο. Επιπλέον παρέχει στον διαχειριστή τη δυνατότητα να δει κάποιες πληροφορίες σχετικά με τις καταχωρήσεις για να λαμβάνονται υπόψιν από τους αρμόδιους ώστε να βελτιώνεται το επίπεδο των υπηρεσιών.

3.1 Μεθοδολογία

Σκοπός της εργασίας είναι να αναπτυχθεί μία εφαρμογή δημόσιας διοίκησης, που να προσφέρει μία υπηρεσία, για να διευκολύνει την επίλυση προβλημάτων που αφορούν πολίτες μίας περιοχής. Να παρέχει την ενημέρωση από τους πολίτες προς την αρμόδια αρχή τα θέματα που τους απασχολούν, για να βελτιώσει το επίπεδο παροχής πληροφοριών και των υπηρεσιών. Ενθαρρύνοντας τη συμμετοχή των πολιτών στο να έχουν άποψη για θέματα που τους απασχολούν είναι ένα θετικό βήμα για την βελτίωση της καθημερινότητας. Χρησιμοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας σε συνδυασμό με μηχανική μάθηση αναμένεται να αναπτυχθεί μία χρήσιμη εφαρμογή ώστε να προσφέρει λύσεις στο δημόσιο τομέα.

Η εφαρμογή θα δέχεται σαν είσοδο ένα κείμενο το οποίο θα κατηγοριοποιείται σε τέσσερις κατηγορίες σε πρώτη φάση και σε δεύτερη φάση θα γίνεται ανάλυση συναισθήματος. Οι κατηγορίες που επιλέχθηκαν είναι οι ακόλουθες Οδικό Δίκτυο, Αθλητισμός/Ψυχαγωγία, Περιβάλλον, Τουρισμός/Φιλοξενία. Επιλέχθηκαν κατηγορίες που σχετίζονται με περιπτώσεις που συναντώνται στην καθημερινότητα και καλύπτουν ένα ευρύ φάσμα θεμάτων, ώστε να διασφαλιστεί η συλλογή δεδομένων με ποικιλία. Επιπλέον, η επιλογή των κατηγοριών λειτούργησε ως οδηγός για την εύρεση διαφορετικών παραδειγμάτων και περιπτώσεων. Σκοπός ήταν η συλλογή επαρκούς όγκου δεδομένων για την κάλυψη των αναγκών της διπλωματικής εργασίας. Η ανάλυση συναισθήματος επιλέχθηκε να είναι σε Θετική ή Αρνητική. Επιπλέον ο αρμόδιος φορέας θα έχει τη δυνατότητα να βλέπει αναλυτικά τις καταχωρήσεις που γίνονται καθώς και την ανάλυση των δεδομένων σε έξι διαφορετικά διαγράμματα με σκοπό την καλύτερη «ανάγνωση» των δεδομένων.

3.1.1 Σύνολο Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν αποτελούνται από κείμενα που αφορούν τις τέσσερις κατηγορίες που προαναφέρθηκαν (Οδικό Δίκτυο, Αθλητισμός/Ψυχαγωγία, Περιβάλλον, Τουρισμός/Φιλοξενία) και χωρισμένα επίσης σε θετικές ή αρνητικές εγγραφές, περισσότερες λεπτομέρειες στον πίνακα 2.

Πίνακας 2: Κατηγορίες και ποσότητα εγγραφών αναλυτικά σε θετικές & αρνητικές

ΚΑΤΗΓΟΡΊΑ	ΘΕΤΙΚΕΣ	ΑΡΝΗΤΙΚΕΣ	ΣΥΝΟΛΙΚΕΣ ΕΓΓΡΑΦΕΣ
Οδικό Δίκτυο	20	23	43
Αθλητισμός/Ψυχαγωγία	12	15	27
Περιβάλλον	10	15	25
Τουρισμός/Φιλοξενία	12	16	28
ΣΥΝΟΛΟ	**54**	**69**	**123**

Τα δεδομένα κειμένου που χρησιμοποιούνται στην παρούσα διπλωματική εργασία συλλέχθηκαν μέσω της αρχικής εφαρμογής, δίνοντας πρόσβαση σε τυχαίους χρήστες οι οποίοι κατέγραψαν τα κείμενα. Επιπλέον αποτελούν και αντικείμενο προσωπικής καταγραφής και επιπλέον έρευνας που βασίζεται σε εκτεταμένη αναζήτηση στο διαδίκτυο. Στα κείμενα έγινε προσπάθεια για την όσο το δυνατόν καλύτερη προσέγγιση και αντίληψη που αφορούν τις κατηγορίες που μελετώνται. Οι κατηγορίες επιλέχθηκαν με κριτήριο, την δημοφιλία που αφορά τις εκάστοτε περιπτώσεις και την εύρεση διαφορετικών παραδειγμάτων - περιπτώσεων, για να ολοκληρωθούν τα δεδομένα που χρειάστηκαν. Η απόδοση των κατηγοριών στα δεδομένα κειμένου έγινε με προσωπική κρίση καθώς και στις κατηγορίες που αφορούν συναισθηματική ανάλυση (θετική ή αρνητική). Θα ήταν σωστό να αναφερθεί ότι δεν βρέθηκαν μαζικά επαρκή δεδομένα στο διαδίκτυο σε ποσότητα που θα μπορούσε να χρησιμοποιηθεί για τη διεκπεραίωση της εργασίας και ως εκ τούτου τα δεδομένα γράφτηκαν εξ αρχής.

Δημιουργήθηκε πίνακας στη βάση δεδομένων με όνομα comments όπου περιλαμβάνει πέντε στήλες (id, text, categories, sentiment, date). Η στήλη id είναι τύπου int και ορίστηκε ως πρωτεύον κλειδί του πίνακα. Η στήλη text είναι τύπου text και περιλαμβάνει τα κείμενα καταγραφής. Η στήλη categories είναι τύπου varchar και έχει την κατηγορία του κειμένου. Η στήλη sentiment είναι τύπου int και λαμβάνει τιμές 1 ή -1. Και τέλος η στήλη date είναι τύπου date και αποθηκεύει την ημερομηνία καταχώρησης της εγγραφής.

Πίνακας 3: Δείγμα εγγραφών από τον πίνακα comments

id	text	categories	sentiment	date
23	Είναι μεγάλο πρόβλημα τα διπλοπαρκαρίσματα στο κέντρο της πόλης. Το ένα από τα δύο μεγάλα κομμάτια της κεντρικής οδού θα έπρεπε να είναι ανοιχτό. Είναι πολύ δυσάρεστη κατάσταση. Η κυκλοφορία γίνεται όλο και πιο δύσκολη με αποτέλεσμα οι οδηγοί να έχουν χάσει την ψυχραιμία τους. Η κίνηση στους δρόμους δεν είναι εύκολη για τα οχήματα.	Οδικό Δίκτυο	-1	03/16/23

65	Θα ήθελα να σας ενημερώσω για ένα σοβαρό πρόβλημα που αντιμετωπίζουμε στην πόλη μας. Τα πάρκα και οι πράσινοι χώροι μας έχουν καταστραφεί λόγω της αδιαφορίας των πολιτικών αρχών και της κακής συντήρησης. Η έλλειψη καθαριότητας και φροντίδας επηρεάζει αρνητικά την υγεία και την ποιότητα ζωής των κατοίκων μας, ενώ ταυτόχρονα προκαλεί σοβαρές βλάβες στο περιβάλλον και την οικολογία της περιοχής. Σας ζητώ να λάβετε άμεσα μέτρα για την αντιμετώπιση αυτού του προβλήματος και τη βελτίωση της κατάστασης των πράσινων χώρων μας.	Περιβάλλον	-1	03/22/23
88	Η πόλη μας είναι ιδιαίτερα θετικό ότι έχει θέατρο και κινηματογράφο, που προσφέρουν μια μεγάλη ποικιλία ταινιών και θεατρικών παραστάσεων. Τα μουσεία της πόλης, είναι επίσης πολύ δημοφιλή για την ψυχαγωγία και την εκπαίδευση των επισκεπτών. Υπάρχει επίσης μια μεγάλη ποικιλία εστιατορίων και μπαρ, όπου μπορείτε να δοκιμάσετε την καταπληκτική τοπική κουζίνα.	Αθλητισμός/ Ψυχαγωγία	1	03/27/23

3.1.2 Καθαρισμός και προεπεξεργασία δεδομένων

Ξεκινώντας τη διαδικασία επεξεργασίας δεδομένων χρησιμοποιήθηκε η βιβλιοθήκη SpaCy [35] και συγκεκριμένα το μοντέλο “el_core_news_lg” [36], το οποίο περιλαμβάνει διαδικασίες επεξεργασίας της ελληνικής γλώσσας (tok2vec, morphologizer, parser, lemmatizer, senter, attribute_ruler, ner). Η προεπεξεργασία κειμένου διαιρεί τις λέξεις μίας πρότασης σε ξεχωριστές λέξεις, αφαιρεί λέξεις χωρίς αξία (stopwords removal), σημεία στίξης, αριθμούς και πρόσθετους κενούς χαρακτήρες, μετατρέπει τις λέξεις σε λήμματα (lemmatization) και τέλος το κείμενο σε πεζά γράμματα. Αυτή η επεξεργασία βοηθάει στην καλύτερη ανάλυση και επεξεργασία των δεδομένων σε μελλοντικά στάδια. Έγιναν δοκιμές και με διαφορετικές βιβλιοθήκες πριν επιλεγεί η SpaCy. Οι δοκιμές όμως στην Ελληνική γλώσσα δεν είχαν τα επιθυμητά αποτελέσματα καθώς δεν είναι ιδιαίτερα δημοφιλής γλώσσα και τα εργαλεία που υπάρχουν στις άλλες βιβλιοθήκες δεν λειτουργούσαν όπως αναμενόταν.

3.1.3 Εκπαίδευση μοντέλων

Για τη διαδικασία της κατηγοριοποίησης κειμένου και για την ανάλυση συναισθήματος επιλέχθηκαν έξι διαφορετικά μοντέλα εκπαίδευσης:

1. ΝΑΪVE BAYES
2. SVM
3. DECISION TREES
4. RANDOM FOREST
5. K-NEAREST NEIGHBORS
6. LOGISTIC REGRESSION

Ο Naive Bayes έχει αποδειχθεί αποτελεσματικός αλγόριθμος σε ποικίλες πρακτικές εφαρμογές, συμπεριλαμβανομένης της ταξινόμησης κειμένου, της ιατρικής διάγνωσης και της διαχείρισης της απόδοσης συστημάτων [37] [38]. Ο SVM είναι ένας από τους πιο ισχυρούς και αξιόπιστους αλγορίθμους ταξινόμησης και παλινδρόμησης σε πολλούς τομείς εφαρμογής, έχει εφαρμοστεί σε πολλά πραγματικά προβλήματα ταξινόμησης λόγω των ισχυρών θεωρητικών τους βάσεων και της καλής τους επίδοσης σε άγνωστα δεδομένα [39]. Ο Decision Tree αλγόριθμος έχει καταταχθεί εδώ και καιρό ως μία από τις πιο πρακτικές και απλές προσεγγίσεις στην ταξινόμηση. Είναι μια μέθοδος που έχει αποδειχθεί πειραματικά ότι παρέχει ανθεκτική απόδοση στην παρουσία θορύβου [40]. Ο Random Forest είναι μία κορυφαία τεχνική για τη διαχείριση μη ισορροπημένων δεδομένων και επιδεικνύει σημαντικά καλύτερη απόδοση από άλλα μοντέλα μηχανικής μάθησης λόγω της παράλληλης αρχιτεκτονικής του [41]. Παρόμοια με τον Random Forest, ο K-Nearest Neighbors είναι γνωστός κυρίως για την επίλυση προβλημάτων ταξινόμησης [42]. Τέλος ο αλγόριθμος Logistic Regression είναι χρησιμοποιείται συχνά για την ταξινόμηση κειμένου. Είναι απλός και εύκολος στην υλοποίηση, δίνοντας καλά αποτελέσματα και επίσης μπορεί να διαχειριστεί καλά τα αραιά δεδομένα [43].

Για κάθε ένα από αυτά τα μοντέλα έγιναν πειράματα εκπαίδευσης με τα δεδομένα χρησιμοποιώντας πρώτα τις προκαθορισμένες τιμές παραμέτρων που έχουν τα μοντέλα. Στη συνέχεια ακολουθεί η διαδικασία της εκπαίδευσης με τη μέθοδο βελτιστοποίησης υπερπαραμέτρων "Grid Search" (Αναζήτηση Πλέγματος), όπου πραγματοποιεί μια πλήρη έρευνα σε ένα υποσύνολο του χώρου υπερπαραμέτρων του αλγορίθμου εκπαίδευσης [44]. Και συγκεκριμένα με το εργαλείο GridSearchCV[45] της python. Είναι ένα εργαλείο που χρησιμοποιείται στη ML που βρίσκει ποιες είναι οι βέλτιστες τιμές των παραμέτρων ενός μοντέλου. Κατά την εκπαίδευση του μοντέλου το εργαλείο αυτό έχοντας ένα εύρος τιμών σε διαφορετικές παραμέτρους του, ψάχνει να βρει ποιες είναι οι βέλτιστες τιμές που καταλήγουν στην υψηλότερη απόδοση. Φροντίζει για τον διαχωρισμό των δεδομένων, την προσαρμογή μοντέλων και την αξιολόγηση της απόδοσης χρησιμοποιώντας διασταυρωμένη επικύρωση (cross-validation), που θα αναλυθεί σε επόμενο κεφάλαιο. Ο λόγος που επιλέχθηκε αυτό το εργαλείο είναι ότι έχουμε σχετικά μικρό σύνολο δεδομένων και με τη διαδικασία που μας προσφέρει αναμένεται να εξαχθούν καλύτερα αποτελέσματα.

Τα δεδομένα χωρίστηκαν σε 71% του συνόλου για να γίνει η εκπαίδευση και σε 29% για να γίνει το τεστ μετά την εκπαίδευση.

Σκοπός είναι να γίνει σύγκριση των αποτελεσμάτων που θα πάρουμε, μετά την εκπαίδευση των μοντέλων, ώστε να επιλέξουμε τον αλγόριθμο με την καλύτερη επίδοση για

να χρησιμοποιήσουμε στην εφαρμογή. Τα μοντέλα (MultinomialNB [46], SVC [47], DecisionTreeClassifier [48], RandomForestClassifier [49], KNeighborsClassifier [50], LogisticRegression [51] αλλά και το GridSearchCV που θα χρησιμοποιήσουμε είναι από την βιβλιοθήκη scikit-learn της Python.

Για να προχωρήσει η διαδικασία της εκπαίδευση των μοντέλων δημιουργήθηκαν έξι αγωγοί (pipelines), δηλαδή μία ακολουθία εργασιών μετασχηματισμού και μοντελοποίησης ML με προκαθορισμένη σειρά εκτέλεσης. Για το σκοπό αυτό χρησιμοποιήθηκε η υπό βιβλιοθήκη της scikit-learn, που ονομάζεται pipeline [52]. Μέσα σε κάθε pipeline ορίστηκε με σειρά προτεραιότητας να εκτελείται, για το κείμενο που υπάρχει για επεξεργασία, πρώτα ο μετασχηματιστής CountVectorizer [53], μετά ο TfidfTransformer [54] και τέλος το μοντέλο εκπαίδευσης που έχει οριστεί. Ο υπολογιστής δεν μπορεί να καταλάβει λέξεις όπως ένας άνθρωπος. Όταν έχουμε δεδομένα κειμένου, άρα λέξεις, πρέπει με κάποιον τρόπο να μετατραπεί – μετασχηματίσει το κείμενο σε νούμερα. Αυτό γίνεται με τον μετασχηματιστή CountVectorizer, ο οποίος μετατρέπει ένα σύνολο κειμένων σε έναν πίνακα, όπου κάθε γραμμή αντιστοιχεί σε ένα κείμενο και κάθε στήλη σε μία μοναδική λέξη που υπάρχει στο σύνολο των κειμένων και αναπαρίσταται με έναν αριθμό που δηλώνει πόσες φορές εμφανίζεται αυτή η λέξη μέσα στο κείμενο. Παράδειγμα εφαρμογής του CountVectorizer φαίνεται στην εικόνα 14. Στη μεταβλητή text υπάρχουν τέσσερις προτάσεις και εφαρμόζοντας CountVectorizer εμφανίζεται σαν αποτέλεσμα ένας πίνακας τεσσάρων γραμμών και 18 στηλών. Στις στήλες είναι οι μοναδικές λέξεις και στις γραμμές εμφανίζεται με νούμερο πόσες φορές υπάρχει η λέξη στο κείμενο της κάθε γραμμής.

```
text = [
    "Γειά, το όνομά μου είναι Νίκος",
    "Δημήτρη, αυτό είναι το τετράδιό μου",
    "Ο Θανάσης προσπαθέι να φτιάξει ένα ποδήλατο",
    "Ο καιρός σήμερα είναι βροχερός"
]
```

	ένα	αυτό	βροχερός	γειά	δημήτρη	...	σήμερα	τετράδιό	το	φτιάξει	όνομά
0	0	0	0	1	0	...	0	0	1	0	1
1	0	1	0	0	1	...	0	1	1	0	0
2	1	0	0	0	0	...	0	0	0	1	0
3	0	0	1	0	0	...	1	0	0	0	0

[4 rows x 18 columns]

Εικόνα 14: Παράδειγμα ανάλυσης με CountVectorizer

Η επόμενη εργασία κατά σειρά είναι με τον μετασχηματιστή TfidfTransformer. Ο μετασχηματιστής μετράει πόσο σημαντική είναι μία λέξη σε ένα κείμενο ελέγχοντας τη συχνότητα εμφάνισής της και πόσο σπάνια εμφανίζεται στο σύνολο των κειμένων. Στην εικόνα 15 φαίνεται στο παράδειγμα πώς θα είναι τα αποτελέσματα μετά την εφαρμογή του TfidfTransformer. Αν το αποτέλεσμα είναι προς το μηδέν ή μηδέν είναι λιγότερο σημαντική λέξη, όσο πιο μεγάλο είναι το αποτέλεσμα, τόσο πιο σημαντική είναι η λέξη. Π.χ. στην πρώτη πρόταση η λέξη “όνομα” είναι πιο σημαντική από τη λέξη “το”.

```
        ένα      αυτό  βροχερός  ...        το   φτιάξει     όνομά
0  0.000000  0.000000  0.000000  ...  0.365594  0.000000  0.463709
1  0.000000  0.463709  0.000000  ...  0.365594  0.000000  0.000000
2  0.408248  0.000000  0.000000  ...  0.000000  0.408248  0.000000
3  0.000000  0.000000  0.541736  ...  0.000000  0.000000  0.000000

[4 rows x 18 columns]
```

Εικόνα 15: Παράδειγμα ανάλυσης με TfidfTransformer

Το τελευταίο βήμα είναι η χρησιμοποίηση του μοντέλου εκπαίδευσης στα δεδομένα κειμένου που έχουμε, ώστε να προχωρήσει στην εκπαίδευσή του. Στην εικόνα 16 φαίνεται παράδειγμα pipelines σε Python για τρεις classifiers.

```python
nb_pipeline = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf', MultinomialNB())
])

svm_pipeline = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf', SVC())
])

dt_pipeline = Pipeline([
    ('vect', CountVectorizer()),
    ('tfidf', TfidfTransformer()),
    ('clf', DecisionTreeClassifier())
])
```

Εικόνα 16: Παράδειγμα pipelines σε Python

Μετά το τέλος όλων των εργασιών το μοντέλο έχει εκπαιδευτεί και μπορεί να κάνει προβλέψεις σε κείμενο που δεν έχει ξαναδεί. Στο καινούργιο κείμενο ακολουθείται η διαδικασία με τα pipelines, η οποία είναι η ίδια όπως περιεγράφηκε πριν και το μοντέλο χρησιμοποιώντας αυτά που έμαθε κατά την εκπαίδευσή του προβλέπει την κλάση με βάση τα αποτελέσματα που θα βρει με την υψηλότερη πιθανότητα.

Στη συνέχεια ορίζονται οι παράμετροι ελέγχου των μοντέλων για να χρησιμοποιηθούν στο εργαλείο GridSearchCV όπως προαναφέρθηκε. Στην εικόνα 17 φαίνεται ένα παράδειγμα διαφορετικών παραμέτρων για τρία μοντέλα.

```
svm_param_grid = {
    'vect__ngram_range': [(1, 1), (1, 2)],
    'tfidf__use_idf': (True, False),
    'clf__C': [0.01, 0.1, 1, 10, 100],
    'clf__kernel': ['linear', 'rbf', 'poly', 'sigmoid'],
}

nb_param_grid = {
    'vect__ngram_range': [(1, 1), (1, 2)],
    'tfidf__use_idf': (True, False),
    'clf__alpha': [0.001, 0.01, 0.1, 1, 10],
}

dt_param_grid = {
    'vect__ngram_range': [(1, 1), (1, 2)],
    'vect__min_df': [1, 2],
    'clf__criterion': ['gini', 'entropy'],
    'clf__max_depth': [5, 10, 15, 20, None],
    'clf__min_samples_split': [2, 5, 10, 20],
    'clf__min_samples_leaf': [1, 2, 5, 10],
}
```

Εικόνα 17: Παράδειγμα διαφορετικών παραμέτρων με χρήση GridSearchCV

Κατά τη διαδικασία εκπαίδευσης θα γίνουν δοκιμές με όλους τους πιθανούς συνδυασμούς τιμών, των παραμέτρων που ορίσαμε για το κάθε μοντέλο, ώστε να προκύψουν οι τιμές με την καλύτερη απόδοση.

Η εκπαίδευση όλων αυτών των μοντέλων και μάλιστα σε τόσες διαφορετικές παραμέτρους χρειάζεται και πολλούς υπολογιστικούς πόρους. Για να μπορέσει η διαδικασία να γίνει όσο το δυνατόν πιο γρήγορα και να αξιολογούνται τα αποτελέσματα για να προχωράει η διαδικασία εφαρμόσθηκε η παράμετρος του cross-validation της Python “n_jobs=-1”. Με αυτή την παράμετρο ενεργή, αυτόματα βρίσκονται οι πόροι που έχει το σύστημα και κατανέμει τις διεργασίες που γίνονται παράλληλα σε όλους τους πυρήνες του επεξεργαστή. Η παράλληλη αυτή επεξεργασία έχει σαν αποτέλεσμα να μειώνεται σημαντικά ο χρόνος των εργασιών άρα και της εκπαίδευσης των μοντέλων.

Μετά την ολοκλήρωση της εκπαίδευσης τα μοντέλα αποθηκεύονται ξεχωριστά σε αρχεία .pkl, μέσω της βιβλιοθήκης joblib [55], ώστε όταν τελειώσει η αξιολόγησή τους να επιλεχθούν τα καταλληλότερα, για να εισαχθούν στην εφαρμογή και να κάνει προβλέψεις αυτόματα όταν γίνεται μία νέα καταχώρηση.

3.1.4 Αξιολόγηση αποτελεσμάτων

Πίνακας 4: Μετρικές Αξιολόγησης

Μετρικές Αξιολόγησης
Classification Report
Cross-Validation
Αξιολόγηση σε άγνωστα δεδομένα

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν οι μετρήσεις με Αναφορά ταξινόμησης (classification report), με Cross-Validation και τέλος έγινε αξιολόγηση σε ένα νέο σύνολο δεδομένων που δεν ήταν στα αρχικά δεδομένα.

Πίνακας 5 : Παράδειγμα classification report

	precision	recall	f1-core
Κλάση 1	0.85	0.92	0.88
Κλάση 2	0.78	0.65	0.71
Accuracy			0.82

Το classification report παράγει έναν πίνακα όπου δίνει αποτελέσματα (accuracy, precision, recall, f1-score) σε κάθε γραμμή για την κάθε κλάση που υπάρχει.

Accuracy είναι το στοιχείο που μετράει πόσο τις εκατό σωστά ταξινομημένες περιπτώσεις υπήρχαν, δηλαδή ο αριθμός των σωστών προβλέψεων διαιρεμένος με το σύνολο όλων των δειγμάτων.

$$[\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)]$$

Είναι ένας δείκτης που δείχνει πόσο καλά έχει εκπαιδευτεί το μοντέλο. Από μόνος του δεν είναι αρκετός για να αξιολογηθεί ότι το μοντέλο μας είναι «καλό», ακόμα και αν έχει accuracy 99%. Αυτό συμβαίνει αν π.χ. σε ένα σύνολο δεδομένων που έχει πολύ μεγάλη απόκλιση ποσότητας δεδομένων σε μία από τις 2 κατηγορίες του, μπορεί να πετύχει πολύ υψηλή accuracy σε αυτή την κατηγορία αλλά στην άλλη πολύ κακή accuracy, που εν τέλει δεν θα φανεί μόνο από ένα αποτέλεσμα που θα μας δώσει το accuracy συνολικά.

Το precision μετράει την αναλογία των σωστά προβλεπόμενων θετικών περιπτώσεων από όλες τις περιπτώσεις που προβλέπονται ως θετικές.

$$[\text{Precision} = TP/(TP + FP)$$

Το recall υπολογίζει την αναλογία των σωστά προβλεπόμενων θετικών περιπτώσεων από όλες τις πραγματικές θετικές περιπτώσεις.

$$[\text{Recall} = TP/(TP + FN)]$$

Τέλος το f1-score παρέχει μία ισορροπία μεταξύ precision και recall. Συνδυάζει την accuracy και την ανάκληση σε μια ενιαία μέτρηση.

$$[\text{f1-score}=2*(\text{precision}*\text{recall}/\text{precision}+\text{recall})]$$

Όσο πιο κοντά στη μονάδα είναι το αποτέλεσμα του f1-score τόσο πιο καλή αναμένεται να είναι η απόδοση του μοντέλου [56].

Το Cross-Validation είναι μία διαδικασία όπου γίνεται διαίρεση του συνόλου των δεδομένων σε δύο τμήματα, ένα για εκπαίδευση (training set) και ένα για έλεγχο (test set). Η βασική μορφή είναι η k-fold cross-validation, σε αυτή τα δεδομένα χωρίζονται αρχικά σε k ίσα (ή σχεδόν ίσα) υποσύνολα. Στη συνέχεια, εκτελούνται k επαναλήψεις εκπαίδευσης και επικύρωσης, όπου σε κάθε επανάληψη ένα διαφορετικό υποσύνολο δεδομένων κρατείται έξω για τον έλεγχο, ενώ τα υπόλοιπα k - 1 υποσύνολα χρησιμοποιούνται για την εκπαίδευση [57]. Στη παρούσα εργασία χρησιμοποιήθηκε η υπό-βιβλιοθήκη της skicit-learn η StratifiedKFold [58]. Η StratifiedKFold χρησιμοποιεί την τεχνική Stratified K-Fold Cross-Validation (SKCV) διαχωρίζοντας τα δεδομένα σε ίσα μέρη και με όσο το δυνατόν καλύτερη κατανομή ανά κατηγορία για να ληφθούν πιο αξιόπιστες εκτιμήσεις απόδοσης [59]. Στη συγκεκριμένη περίπτωση τα δεδομένα χωρίστηκαν σε εφτά μέρη. Με τον τρόπο αυτό μπορούν να αντιμετωπιστούν τα μη ισορροπημένα σύνολα δεδομένων ή τα προβλήματα πολλαπλών κατηγοριών. Σε ορισμένες κατηγορίες, όπου τα δείγματα είναι περιορισμένα, επιτυγχάνεται μια πιο αξιόπιστη και ακριβή αξιολόγηση. Για να γίνει πιο κατανοητός ο τρόπος λειτουργίας του συγκεκριμένου εργαλείου αν για παράδειγμα οριστεί ότι σε ένα σύνολο δεδομένων το 20% ανήκει στην κατηγορία Α, το 20% στο Β και το υπόλοιπο 60% στο Γ, τότε με τη χρήση του SKCV κάθε κομμάτι που θα χωριστεί θα έχει 20% από το Α, 20% από το Β και 60% από το Γ. Σε κάθε ένα κομμάτι που χωρίζεται το σύνολο δεδομένων υποθέτοντας ότι υπάρχουν 3 κομμάτια, το πρώτο κομμάτι χρησιμοποιείται για να γίνει το τεστ και τα επόμενα δύο για να γίνει η εκπαίδευση του μοντέλου. Έτσι συνεχίζει μία επαναληπτική διαδικασία όπου στο επόμενο βήμα θα χρησιμοποιηθεί το δεύτερο κομμάτι για τεστ και στη συνέχεια το τρίτο.

Επιπλέον όταν εκπαιδευτεί το μοντέλο γίνεται και έλεγχος accuracy σε ένα νέο σύνολο δεδομένων το οποίο δεν χρησιμοποιήθηκε κατά την εκπαίδευση του μοντέλου. Αυτό θα μας δώσει μία πολύ καλή εικόνα στο πόσο καλά δουλεύει το μοντέλο σε άγνωστα, για αυτό, δεδομένα. Είναι ένα μέτρο σύγκρισης που στη περίπτωση αυτή έχει υψηλή αξία λόγω της φύσης του προβλήματος που τα δεδομένα αποτελούνται από κείμενα και μπορεί να έχουν τεράστιες αποκλίσεις μεταξύ τους. Άρα ένα καλό αποτέλεσμα σε αυτά τα δεδομένα είναι ένας καλός δείκτης ότι το μοντέλο προβλέπει σωστά.

Εικόνα 18: Ποσοστό εγγραφών ανά κατηγορία

Εικόνα 19: Αναλυτική κατανομή εγγραφών με βάση το συναίσθημα ανά κατηγορία

Στην εικόνα 18 φαίνεται η κατανομή των 31 εγγραφών στο νέο σύνολο δεδομένων που δεν γνωρίζει το μοντέλο και στην εικόνα 19 την κατανομή των εγγραφών σε αρνητικό ή θετικό συναίσθημα.

4. ΥΛΟΠΟΙΗΣΗ ΣΥΣΤΗΜΑΤΟΣ

Σε αυτό το κεφάλαιο παρουσιάζονται οι τεχνολογίες και τα εργαλεία που χρησιμοποιήθηκαν για την εκπόνηση της διπλωματικής εργασίας.

4.1 Εργαλεία που χρησιμοποιήθηκαν

Για την ανάπτυξη της εφαρμογής χρησιμοποιήθηκαν τα εξής εργαλεία:

- PYTHON
- FLASK
- XAMPP (APACHE-MySQL)

Η Python είναι μία γλώσσα προγραμματισμού υψηλού επιπέδου και μία από τις πιο δημοφιλείς πλέον γλώσσες στο κόσμο. Δημιουργήθηκε το 1989 στην Ολλανδία από τον Guido van Rossum. Η πρώτη έκδοση δημοσιεύτηκε το Φεβρουάριο του 1991. Η τελευταία έκδοση που υπάρχει σήμερα είναι η 3.11. Είναι αρκετά εύκολη στην εκμάθηση ακόμα και από έναν αρχάριο, είναι επίσης μία γρήγορη γλώσσα ειδικά στη ταχύτητα που ένας προγραμματιστής μπορεί να γράφει κώδικα με την απλή δομή της. Η Python έχει τη δυνατότητα να υποστηρίξει από web development, data analytics, ML, data science, data engineering μέχρι και AI. [60]

Το Flask είναι ένα web application framework, είναι γραμμένο σε python και ανήκει στην εργαλειοθήκη Web Server Gateway Interface (WSGI). Δημιουργήθηκε από τον Armin Ronacher και δημοσιεύτηκε τον Απρίλιο του 2010. Είναι απλό και γρήγορο στο να κατασκευάσεις ιστοσελίδες και εφαρμογές. [61] [62]

Το XAMPP είναι ένα ανοιχτού κώδικα πακέτο που περιλαμβάνει κυρίως Apache HTTP Server και MySQL(από το 2015 άλλαξε σε MariaDB) και δημιουργήθηκε από τους Apache Friends. Είναι ένας web server που επιτρέπει στους χρήστες να δοκιμάζουν τοπικά σε ένα ηλεκτρονικό υπολογιστή τον κώδικά τους. [63]

4.2 Βιβλιοθήκες python

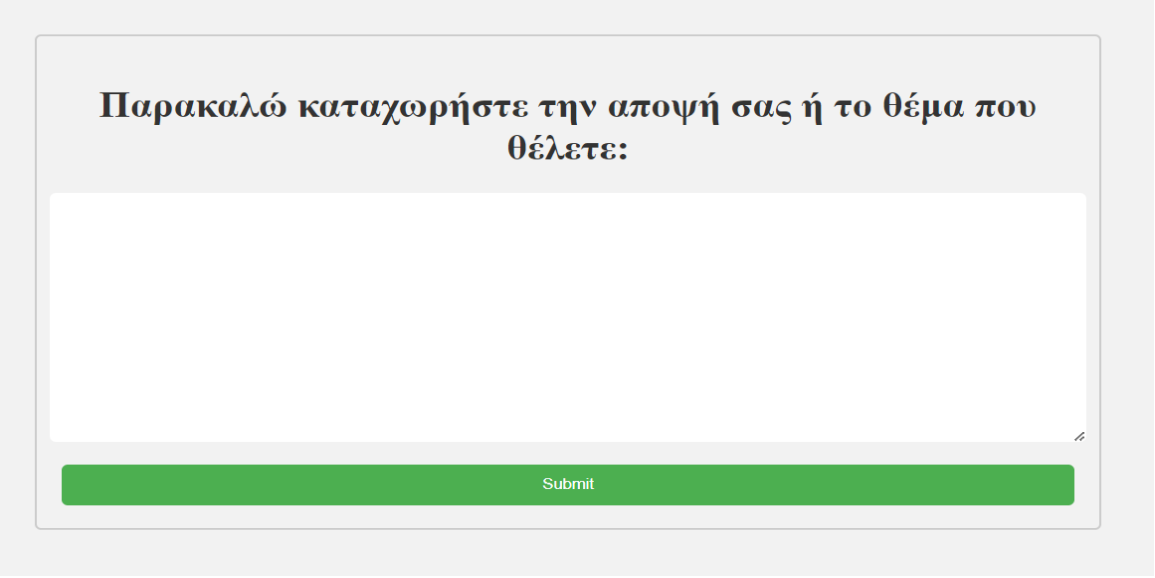
Από την Python χρησιμοποιήθηκαν οι βιβλιοθήκες pandas [64], για το διάβασμα, την επεξεργασία και την ανάλυση των δεδομένων. Για την ανάπτυξη, την εκπαίδευση και την αξιολόγηση των μοντέλων ML χρησιμοποιήθηκε η βιβλιοθήκη scikit-learn [65]. Για την απεικόνιση των αποτελεσμάτων χρησιμοποιήθηκαν οι βιβλιοθήκες matplotlib [66] και seaborn [67]. Επίσης χρησιμοποιήθηκε η βιβλιοθήκη Joblib για την αποθήκευση μοντέλων ML python σε αρχεία, ώστε να μπορούν να ανακτηθούν αργότερα χωρίς την ανάγκη να εκπαιδευτούν ξανά ή να δημιουργηθούν.

5. ΠΑΡΟΥΣΙΑΣΗ ΔΙΕΠΑΦΗΣ ΙΣΤΟΣΕΛΙΔΑΣ (WEB INTERFACE)

Παρακάτω θα παρουσιαστεί το web interface της εφαρμογής καθώς και αναλυτικά τα αποτελέσματα ανά μοντέλο. Θα γίνει σύγκριση με χρήση ή όχι διαφορετικών παραμέτρων (GridSearchCV) για κάθε μοντέλο. Θα παρουσιαστούν δώδεκα εικόνες με έξι διαγράμματα η κάθε εικόνα, συνολικά θα παρουσιαστούν εβδομήντα δύο διαγράμματα. Οι πρώτες έξι εικόνες αφορούν την κατηγοριοποίηση όπου «1»=ΘΕΤΙΚΌ & «-1»=ΑΡΝΗΤΙΚΟ και αφορά ανάλυση συναισθήματος. Σε κάθε εικόνα στην αριστερή πλευρά από πάνω προς τα κάτω υπάρχουν πρώτα τα αποτελέσματα με τις προκαθορισμένες τιμές των μοντέλων δηλαδή το classification report, cross-validation χωρισμένο σε εφτά μέρη (folds) και accuracy σε ένα καινούργιο σύνολο δεδομένων. Στην δεξιά πλευρά έχουμε τον ίδιο τύπο διαγραμμάτων μετά τη χρήση δοκιμών διαφορετικών παραμέτρων (GridSearchCV) βλέποντας τα αποτελέσματα με τις καλύτερες παραμέτρους για κάθε μοντέλο. Οι επόμενες έξι έχουν ακριβώς τους ίδιους τύπους διαγραμμάτων αλλά αφορούν την κατηγοριοποίηση κειμένου στις τέσσερις κατηγορίες (Οδικό Δίκτυο, Αθλητισμός/Ψυχαγωγία, Περιβάλλον, Τουρισμός/Φιλοξενία).

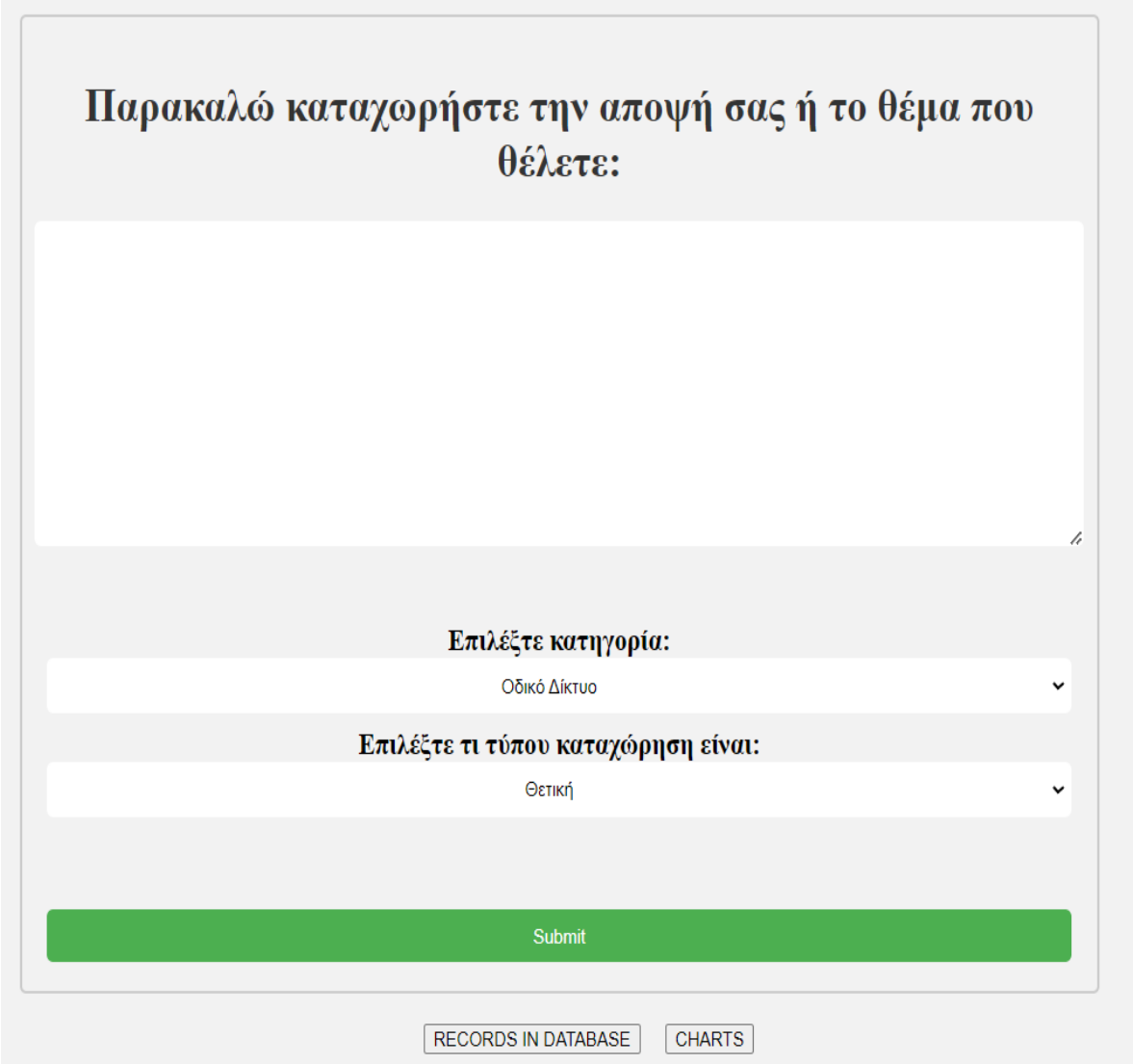
5.1 Διεπαφή ιστοσελίδας

Το Web Interface της εφαρμογής αποτελείται από 4 σελίδες. Δόθηκε έμφαση να είναι όσο το δυνατόν πιο απλό και φιλικό προς τον χρήστη για να είναι φιλικό στη χρήση του, χωρίς να χάνει σε απόδοση αποτελεσμάτων. Η πρώτη σελίδα είναι του χρήστη όπου απλά μπαίνει για να καταχωρήσει το κείμενο που επιθυμεί. Η δεύτερη σελίδα είναι για τον διαχειριστή της δημοσίας υπηρεσίας που έχει τη δυνατότητα να καταχωρήσει και αυτός κάποιο κείμενο και επιπλέον μπορεί να δει τα διαγράμματα ή τις καταχωρήσεις που υπάρχουν στη βάση.



Εικόνα 20: Αρχική οθόνη απλού χρήστη

Στην εικόνα 20 φαίνεται πώς εμφανίζεται η σελίδα στον χρήστη. Γράφοντας το κείμενο πατάει ‘Submit’ και καταχωρείται το κείμενο στη βάση δεδομένων αφού πρώτα γίνει πρόβλεψη κατηγορίας και ανάλυση συναισθήματος μεταξύ θετικού ή αρνητικού κειμένου.



Εικόνα 21: Αρχική οθόνη διαχειριστή

Στην εικόνα 21 φαίνεται η σελίδα του διαχειριστή όπου έχει τη δυνατότητα να καταχωρήσει ένα κείμενο και να επιλέξει κατηγορία και αν είναι θετικό ή αρνητικό να το καταχωρήσει στη βάση δεδομένων.

Πήγαινε Στην Αρχική Σελίδα

123 Εγγραφές !

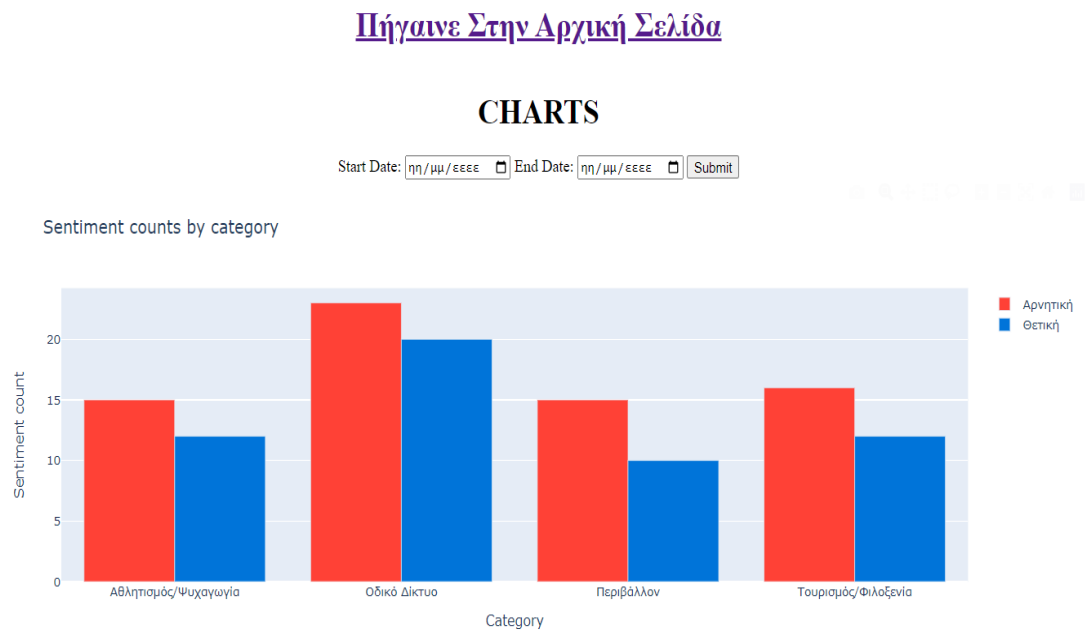
Date: ηη/μμ/εεεε Category: All Sentiment: All Filter Clear filters

Download Results in Excel File

ID	Date (Y-M-D)	Text	Categories	Sentiment
145	2023-06-09	Η υποστήριξη και προώθηση αθλητικών εκδηλώσεων και εγκαταστάσεων στον δήμο μπορεί να ενθαρρύνει τους κατοίκους να είναι ενεργοί, να επιτυγχάνουν τους στόχους τους και να ζουν έναν υγιή τρόπο ζωής.	Αθλητισμός/ Ψυχαγωγία	1
146	2023-06-09	Ένα από τα θετικά πλεονεκτήματα του αθλητισμού είναι η προώθηση της φυσικής υγείας και της ευεξίας. Οι αθλητικές δραστηριότητες προάγουν την καρδιοαναπνευστική αντοχή, την ενδυνάμωση των μυών και την ευεξία του μυαλού. Για αυτούς τους λόγους είναι πολύ θετικό ότι οι χώροι άθλησης στο δήμο μας είναι σε εξαιρετική κατάσταση.	Αθλητισμός/ Ψυχαγωγία	1
147	2023-06-09	Το κλειστό γήπεδο μπάσκετ της περιοχής μας είναι σε άριστη κατάσταση και τα παιδιά μας αλλά και οι μεγαλύτεροι το χαίρονται καθημερινά για την άθλησή τους.	Αθλητισμός/ Ψυχαγωγία	1
		Η επισκεψιμότητα από τουρίστες στο δήμο μας		

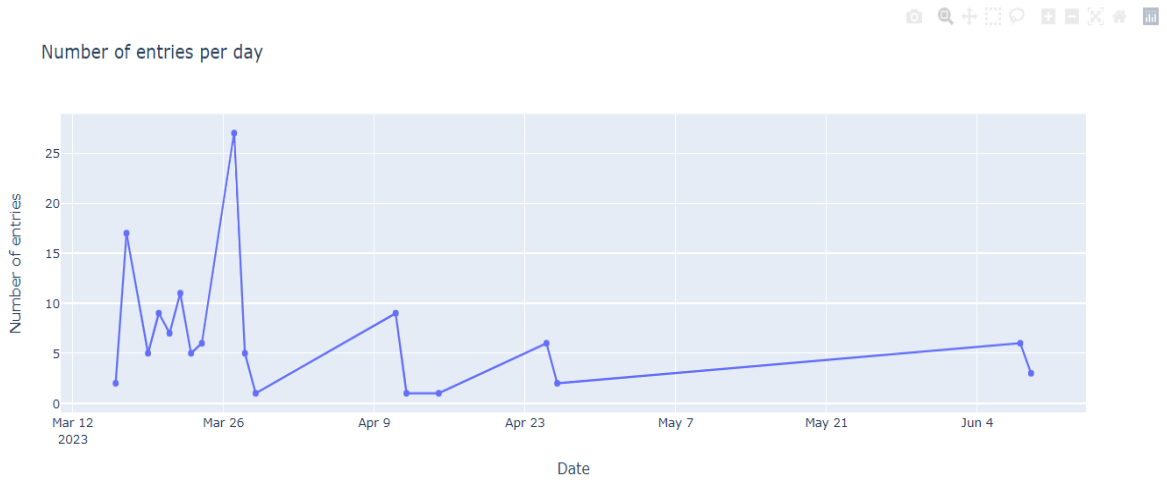
Εικόνα 22: Αρχική οθόνη εγγραφών

Στην εικόνα 22 φαίνεται η σελίδα των καταχωρήσεων από τη βάση δεδομένων όταν πατήσει το κουμπί ‘RECORDS IN DATABASE’. Υπάρχει φίλτρο για διάστημα ημερομηνιών, κατηγορία ή συναίσθημα. Ανάλογα τις επιλογές που θα γίνουν εμφανίζονται και τα αντίστοιχα αποτελέσματα. Υπάρχει επίσης η δυνατότητα να γίνει αποθήκευση των επιλεγμένων δεδομένων σε excel πατώντας το κουμπί ‘Download Results in Excel File ’, σε περίπτωση που επιθυμεί να κάνει κάποια επιπλέον επεξεργασία ή να ενημερώσει για τις καταχωρήσεις κάποιο αρμόδιο τμήμα ώστε αυτό να αναλάβει τις διεκπεραιώσεις που ίσως χρειάζονται. Για λόγους διευκόλυνσης εμφανίζονται κάθε φορά πάνω από τα φίλτρα το σύνολο των εγγραφών που επιλέχθηκαν.



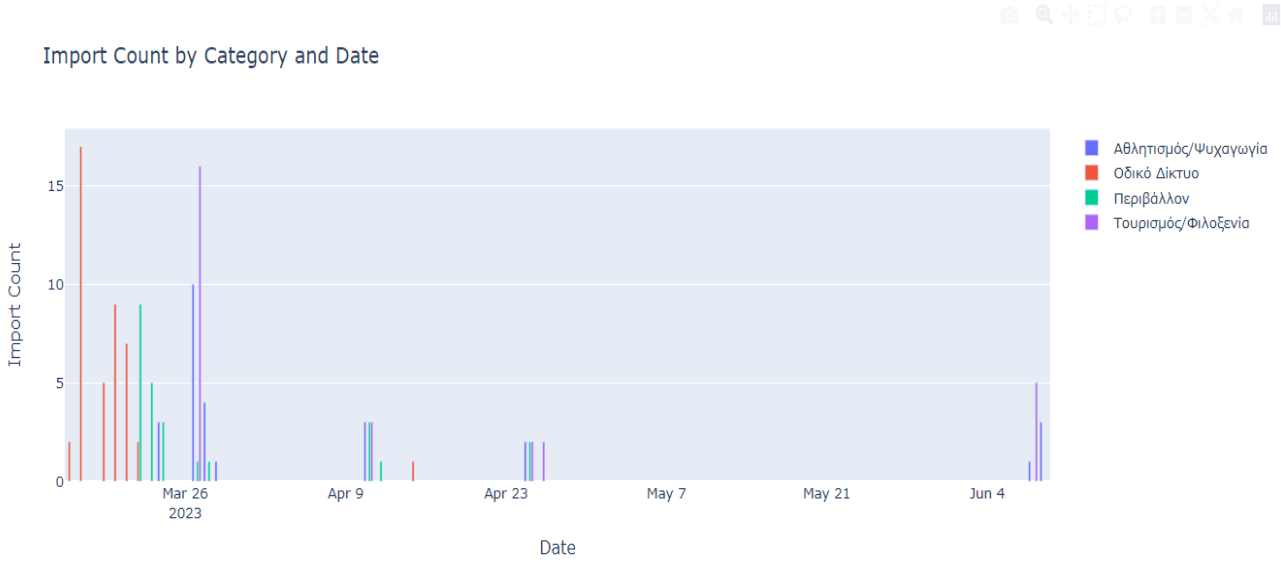
Εικόνα 23: Ραβδόγραμμα εγγραφών ανά κατηγορία και συναισθηματική ανάλυση

Στην εικόνα 23 φαίνεται το πρώτο από τα έξι διάγραμμα. Όλα τα διαγράμματα είναι δυναμικά και περνώντας το ποντίκι πάνω από κάποιο στοιχείο έχει κάθε φορά επιπλέον επιλογές σε κάθε διάγραμμα. Επίσης υπάρχει φίλτρο ημερομηνιών για πιο λεπτομερή έλεγχο. Το πρώτο διάγραμμα εμφανίζει ανά κατηγορία (Αρνητική ή Θετική) πόσες καταχωρήσεις υπάρχουν.



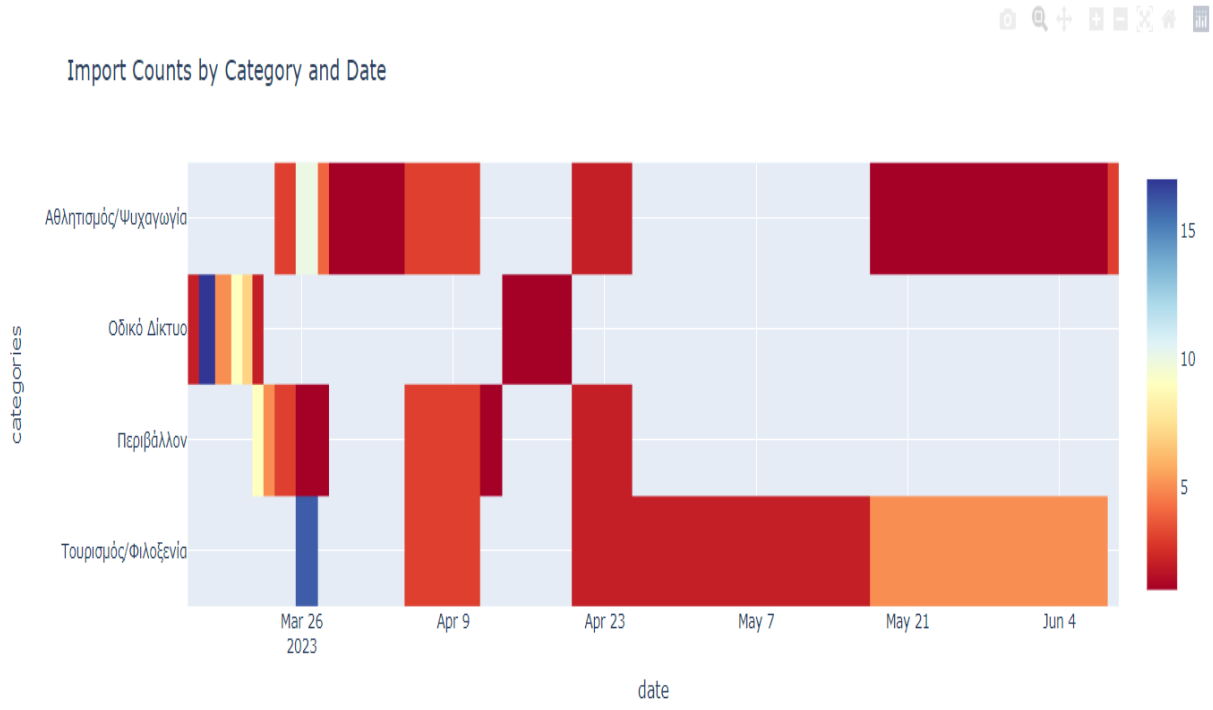
Εικόνα 24: Διάγραμμα διασποράς εγγραφών ανά ημερομηνία καταχώρησης

Στην εικόνα 24 το διάγραμμα παρουσιάζει πόσες καταχωρήσεις υπάρχουν ανά ημερομηνία.



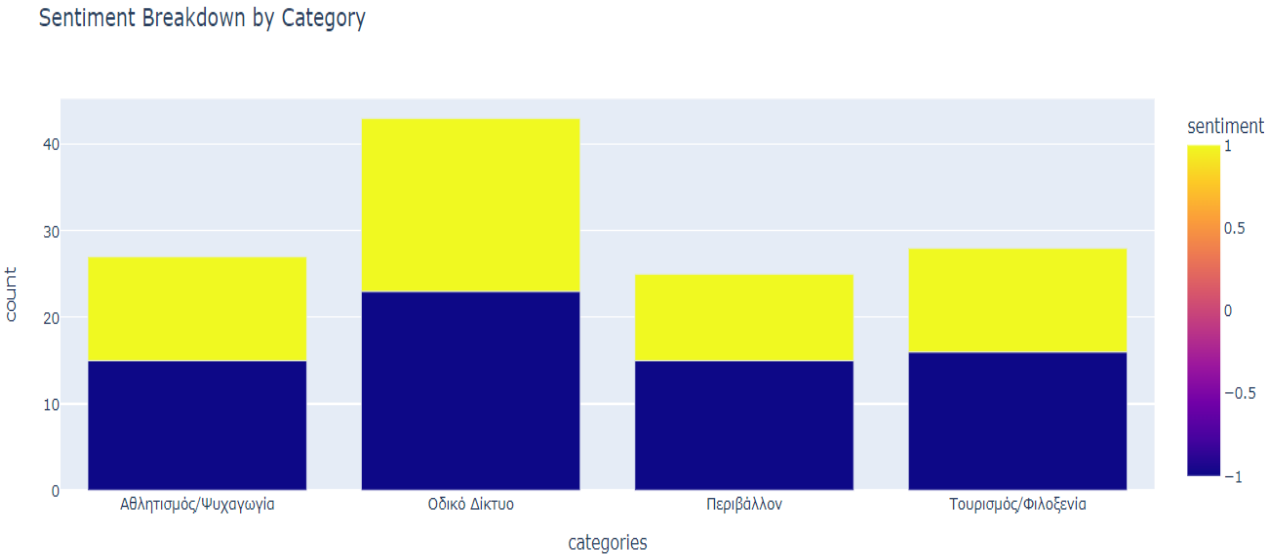
Εικόνα 25: Ραβδόγραμμα ποσότητας εγγραφώ ανά κατηγορία ανά ημερομηνία καταχώρησης

Στην εικόνα 25 το διάγραμμα παρουσιάζει πόσες καταχωρήσεις υπάρχουν ανά ημερομηνία και ανά κατηγορία.



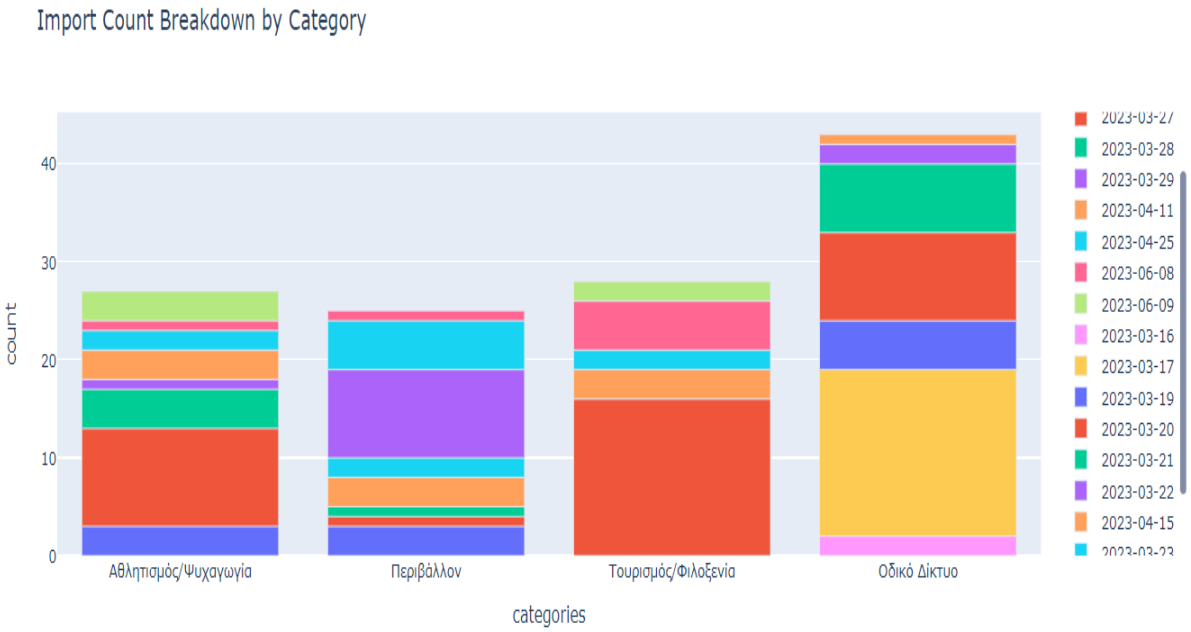
Εικόνα 26: Χάρτης θερμότητας ανά κατηγορία ανά ημερομηνία καταχώρησης

Στην εικόνα 26 παρουσιάζεται ένα διάγραμμα μορφής ‘heatmap’ το οποίο παρουσιάζει καταχωρήσεις ανά κατηγορία σε κάθε ημερομηνία.



Εικόνα 27: Ραβδόγραμμα εγγραφών ανά κατηγορία και συναισθηματική ανάλυση

Στην εικόνα 27 παρουσιάζεται διάγραμμα που δείχνει καταχωρήσεις ανά κατηγορία χωρισμένο σε Θετικές και Αρνητικές.

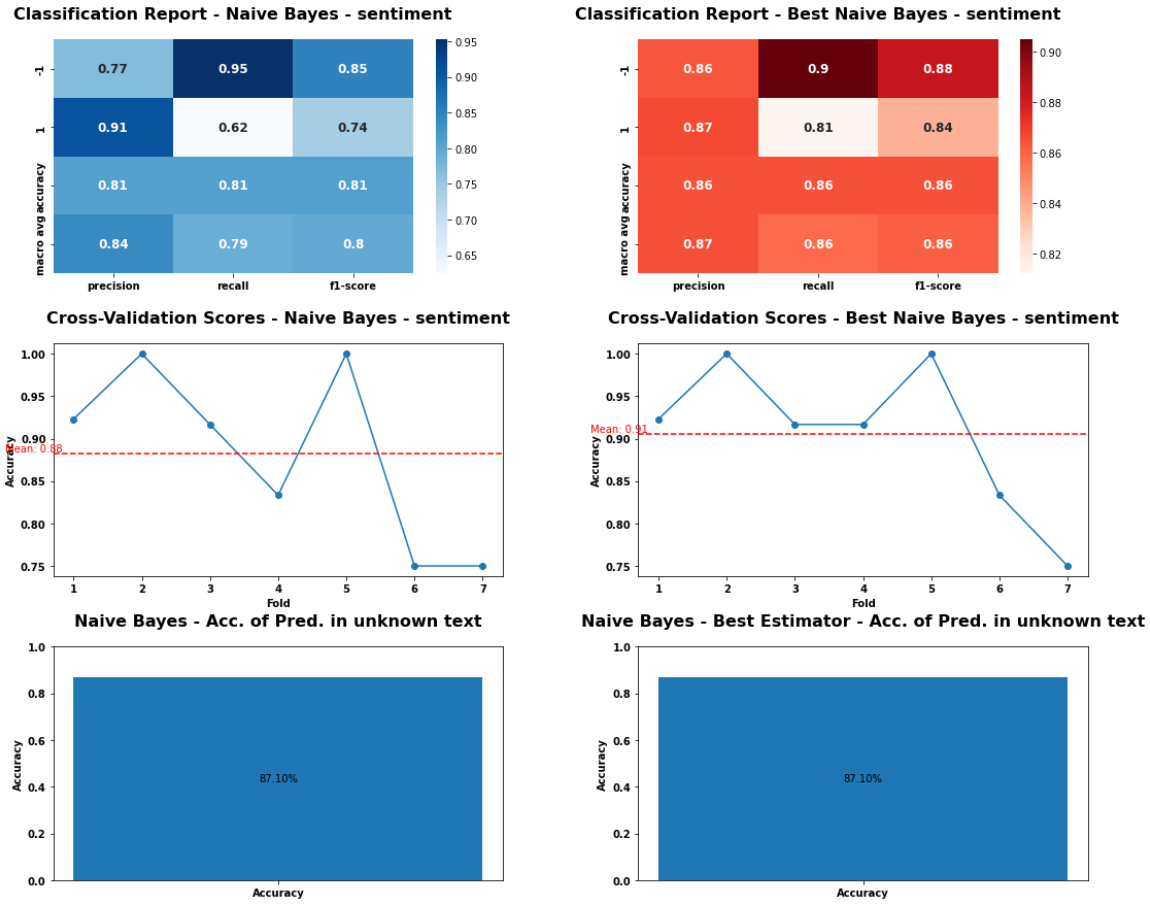


Εικόνα 28: Ραβδόγραμμα ανά κατηγορία και ανά ημερομηνία καταχώρησης

Στην εικόνα 28 (έκτο και τελευταίο διάγραμμα) παρουσιάζεται ανά κατηγορία σύνολο καταχωρήσεων σε σχέση με τις ημερομηνίες καταχώρησης.

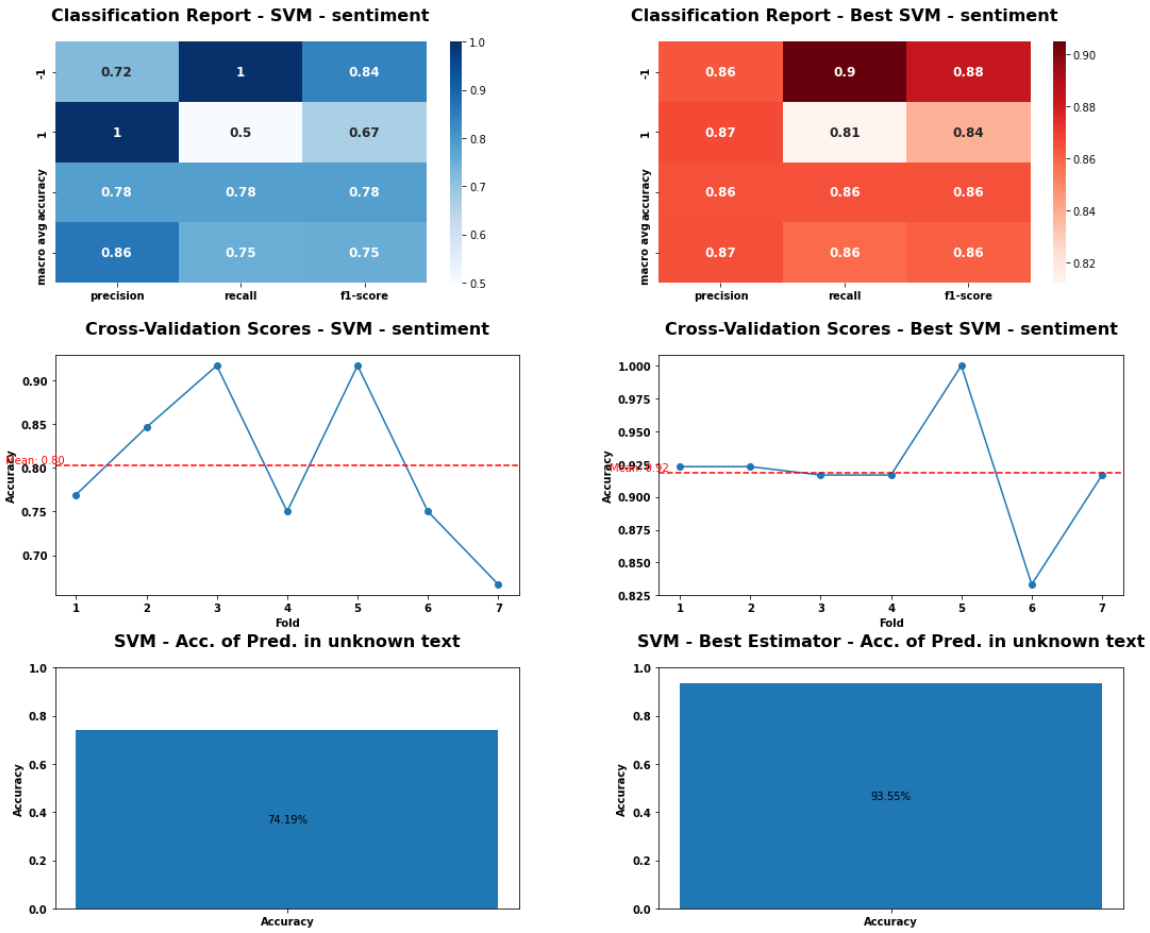
6. ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΚΠΑΙΔΕΥΣΗΣ ΜΟΝΤΕΛΩΝ

6.1 Αποτελέσματα ανάλυσης συναισθήματος



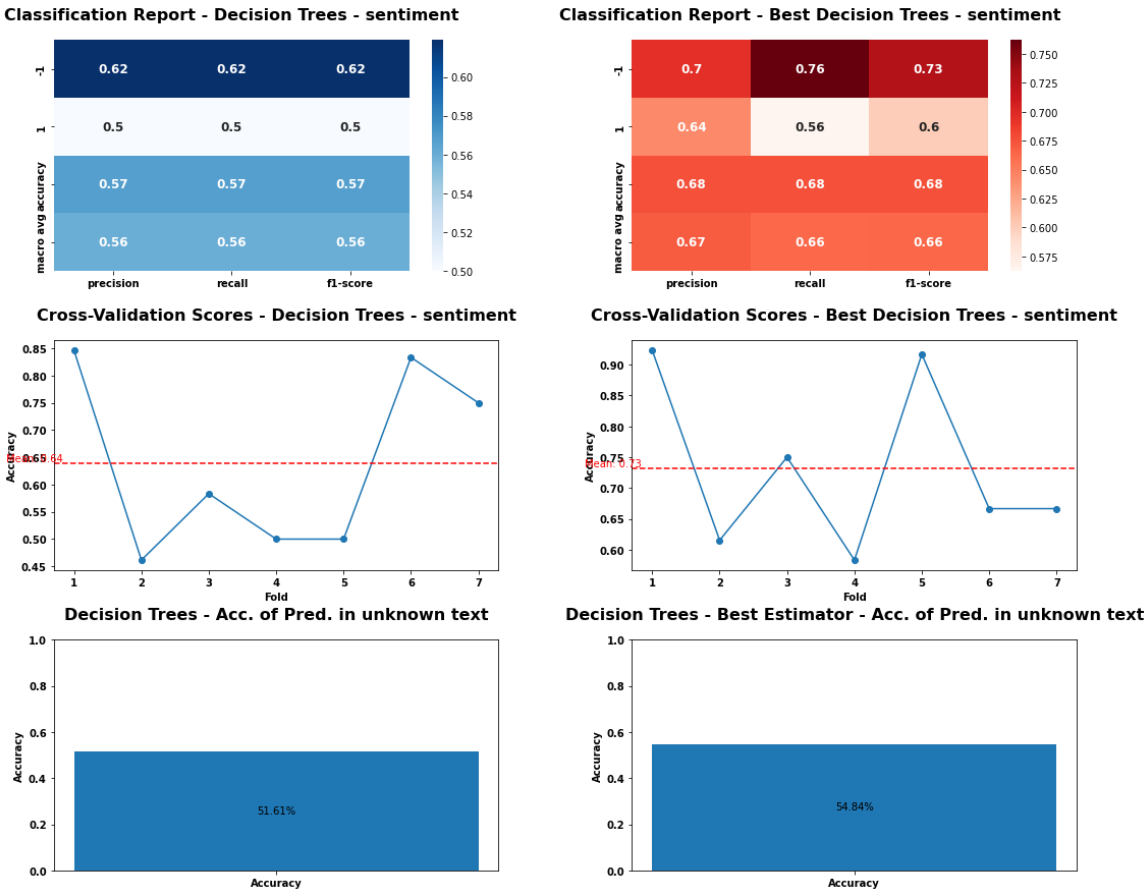
Εικόνα 29: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν ανάλυση συναισθήματος με χρήση αλγορίθμου Naive Bayes

Στην εικόνα 29 φαίνονται τα αποτελέσματα για το μοντέλο Naive Bayes. Στο classification report, το μοντέλο με GridSearchCV έχει καλύτερη απόδοση σε όλους τους δείκτες. Στο Cross-Validation διάγραμμα παρουσιάζει λίγο καλύτερο μέσο όρο στα μέρη (folds) και είναι πιο ισοκατανεμημένα τα αποτελέσματα, δείχνοντας καλύτερα αποτελέσματα στη συνολική απόδοση του μοντέλου. Στο τρίτο διάγραμμα και τα δύο μοντέλα αποδίδουν ακριβώς το ίδιο σε νέα δεδομένα και μάλιστα έχουν πολύ καλύτερο ποσοστό accuracy από τα δεδομένα εκπαίδευσης. Η χρήση GridSearchCV έδωσε ελαφρώς καλύτερα αποτελέσματα σε όλα τα επίπεδα, σε καινούργια δεδομένα όμως επιτυγχάνεται ακριβώς ίδιο αποτέλεσμα ποσοστού πρόβλεψης.



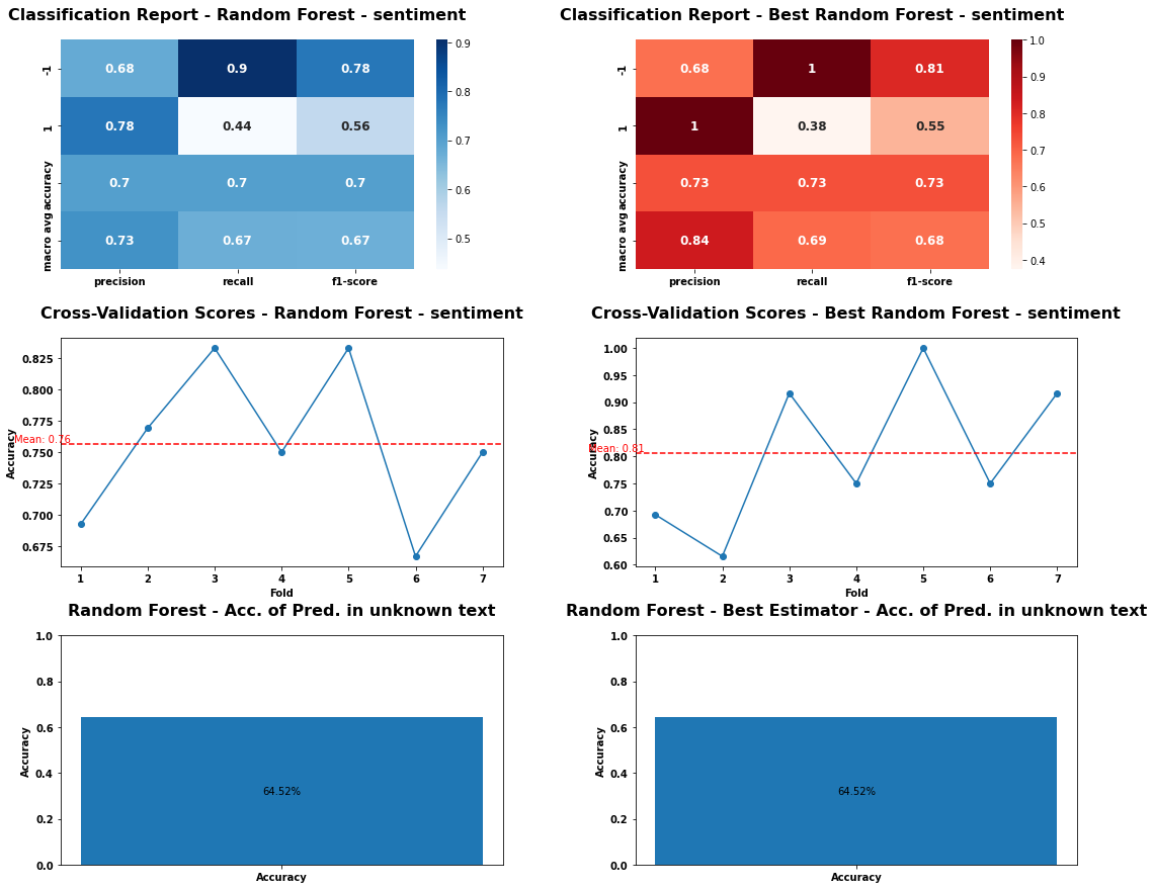
Εικόνα 30: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν ανάλυση συναισθήματος με χρήση αλγορίθμου SVM

Στην εικόνα 30 φαίνονται τα αποτελέσματα για το μοντέλο SVM. Στο classification report το μοντέλο με GridSearchCV έχει καλύτερη απόδοση σε όλους τους δείκτες. Στο Cross-Validation διάγραμμα επίσης έχει μεγαλύτερο μέσο όρο αλλά επίσης έχει και σταθερή τιμή σε πέντε από τα εφτά μέρη (folds) που αυτό δείχνει σταθερότητα πρόβλεψης. Στα αποτελέσματα πρόβλεψης σε νέα δεδομένα επιτυγχάνεται επίσης πολύ καλύτερο ποσοστό accuracy. Συνολικά φαίνεται ότι με τη χρήση GridSearchCV προκύπτουν καλύτερα αποτελέσματα σε όλα τα επίπεδα.



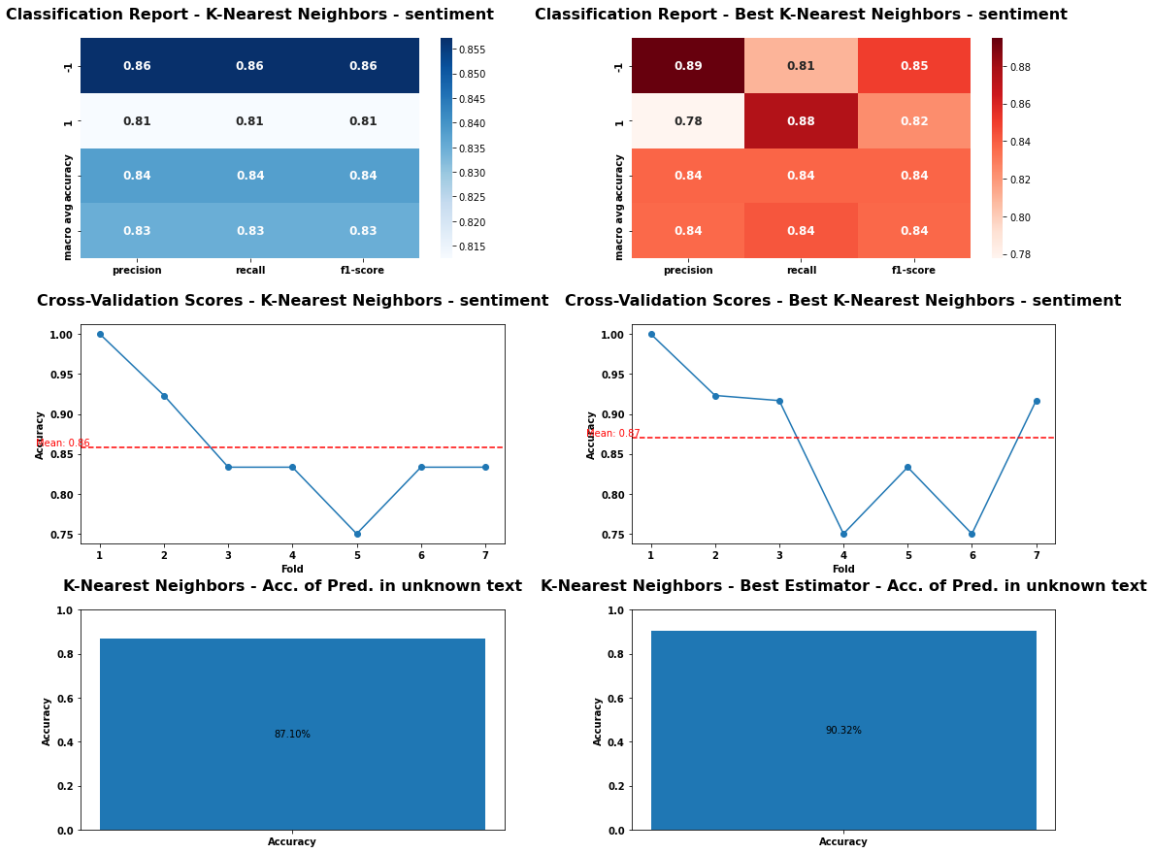
Εικόνα 31: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν ανάλυση συναισθήματος με χρήση αλγορίθμου Decision Trees

Στην εικόνα 31 φαίνονται τα αποτελέσματα για το μοντέλο Decision Trees. Γενικά το μοντέλο με GridSearchCV έχει καλύτερη απόδοση σε όλους τους δείκτες, όμως δεν είναι πολύ καλά τα αποτελέσματα. Το μοντέλο έχει χαμηλή απόδοση σε όλα τα αποτελέσματα. Αυτό μπορεί να συμβαίνει λόγω της απλότητας του μοντέλου και δεν μπορεί να δώσει τα αναμενόμενα αποτελέσματα στα συγκεκριμένα δεδομένα.



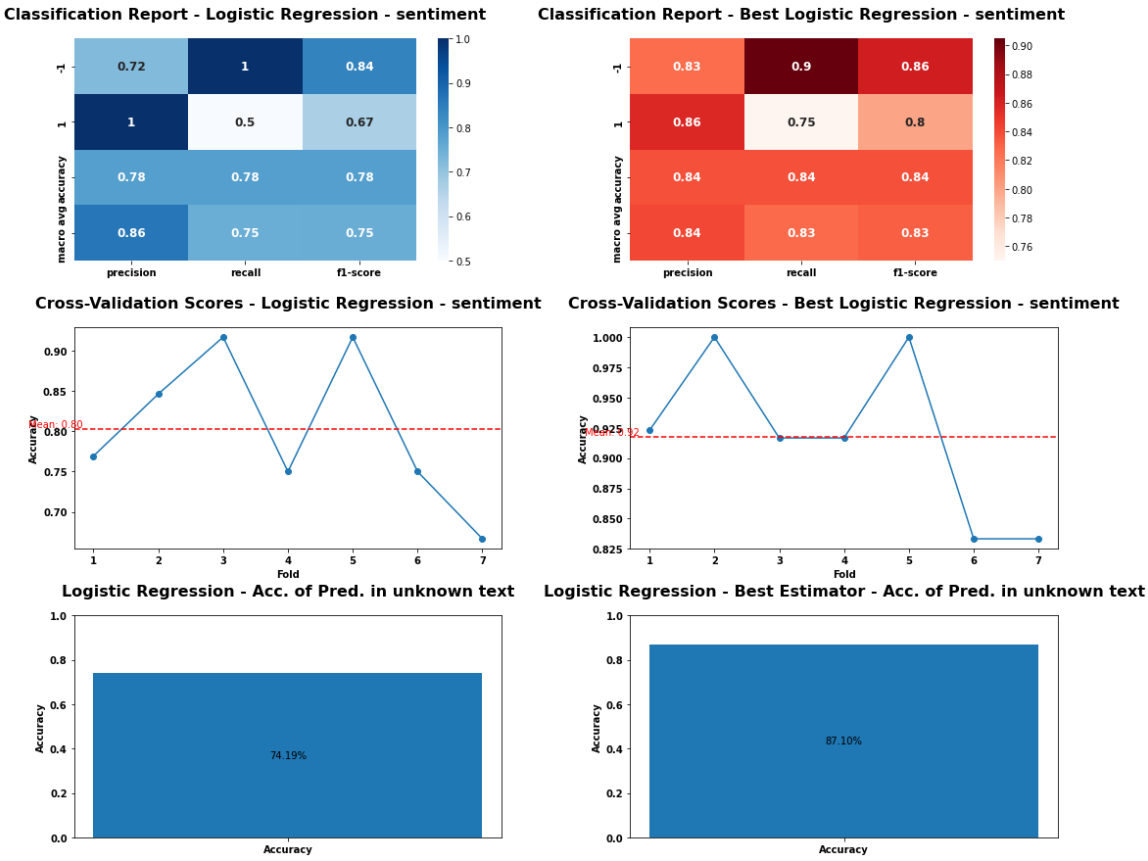
Εικόνα 32: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν ανάλυση συναισθήματος με χρήση αλγορίθμου Random Forest

Στην εικόνα 32 φαίνονται τα αποτελέσματα για το μοντέλο Random Forest. Το μοντέλο με GridSearchCV έχει ελαφρώς καλύτερη απόδοση σε όλους τους δείκτες, όμως δεν είναι πολύ καλά τα αποτελέσματα που εμφανίζονται. Το μοντέλο έχει μέτρια απόδοση σε όλα τα αποτελέσματα και στις δύο περιπτώσεις.



Εικόνα 33: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν ανάλυση συναισθήματος με χρήση αλγορίθμου KNN

Στην εικόνα 33 φαίνονται τα αποτελέσματα για το μοντέλο K-Nearest Neighbors. Το μοντέλο και στις δύο περιπτώσεις έχει παρόμοια αποτελέσματα με ελαφρώς καλύτερα αποτελέσματα στη δεύτερη περίπτωση. Ακόμα και με τη χρήση GridSearchCV δεν βελτιώθηκε σημαντικά η απόδοση του μοντέλου. Αυτό που μπορεί να παρατηρηθεί είναι ότι στη δεύτερη περίπτωση στο Cross-Validation διάγραμμα η accuracy του μοντέλου για τέσσερα από τα εφτά folds είναι πάνω από το μέσο όρο. Αυτό δείχνει ότι σε διαφορετικά μέρη των δεδομένων όταν εκπαιδεύεται το μοντέλο έχει καλύτερα αποτελέσματα και αυτό φαίνεται και στο τελευταίο διάγραμμα όπου υπάρχει μεγαλύτερο accuracy πρόβλεψης στη δεύτερη περίπτωση.



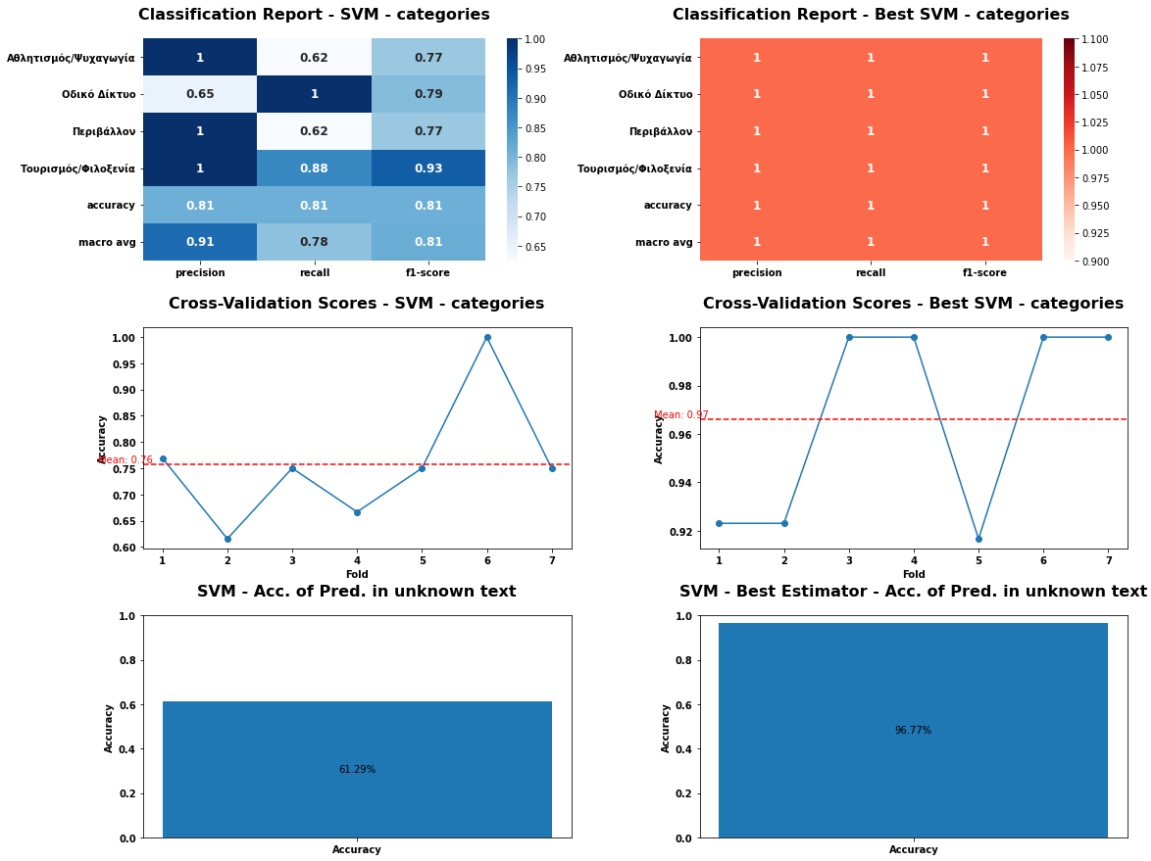
Εικόνα 34: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν ανάλυση συναισθήματος με χρήση αλγορίθμου Logistic Regression

Στην εικόνα 34 φαίνονται τα αποτελέσματα για το μοντέλο Logistic Regression. Στο classification report το μοντέλο με GridSearchCV έχει καλύτερη απόδοση σε όλους τους δείκτες. Στο Cross-Validation ο μέσος όρος βελτιώθηκε αρκετά και επίσης πέντε από τα εφτά folds έχουν ίσο ή μεγαλύτερο accuracy πρόβλεψης. Βελτίωση επίσης παρατηρείται και στο accuracy πρόβλεψης σε νέα δεδομένα.

6.2 Αποτελέσματα κατηγοριοποίησης κειμένου

Εικόνα 35: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν κατηγοριοποίηση κειμένου με χρήση αλγορίθμου Naive Bayes

Στην εικόνα 35 φαίνονται τα αποτελέσματα για το μοντέλο Naive Bayes. Στο classification report το μοντέλο με GridSearchCV έχει σημαντικά καλύτερη απόδοση σε όλους τους δείκτες. Αυτό αποτυπώνεται και στο Cross-Validation διάγραμμα με την αύξηση του μέσου όρου accuracy πρόβλεψης καθώς και στο accuracy πρόβλεψης σε νέα δεδομένα στο τελευταίο διάγραμμα.

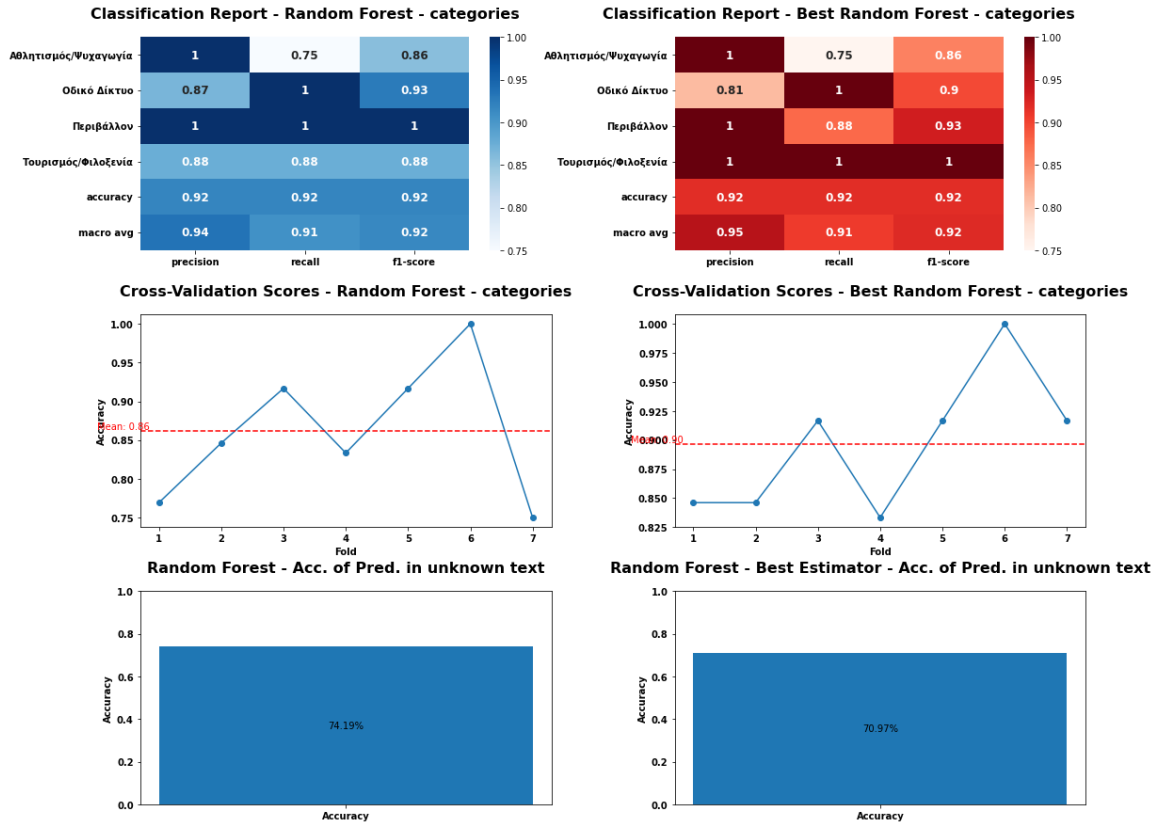


Εικόνα 36: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν κατηγοριοποίηση κειμένου με χρήση αλγορίθμου SVM

Στην εικόνα 36 φαίνονται τα αποτελέσματα για το μοντέλο SVM. Στο classification report το μοντέλο με GridSearchCV έχει σαν αποτέλεσμα την άριστη απόδοση σε όλους τους δείκτες πετυχαίνοντας σημαντική βελτίωση. Στο Cross-Vallidation διάγραμμα ο μέσος όρος επίσης αυξήθηκε σημαντικά με όλα τα folds να έχουν πάνω από 90% accuracy και τέλος σε νέα δεδομένα το μοντέλο κατάφερε να προβλέψει σωστά σε ποσοστό 96.77%.

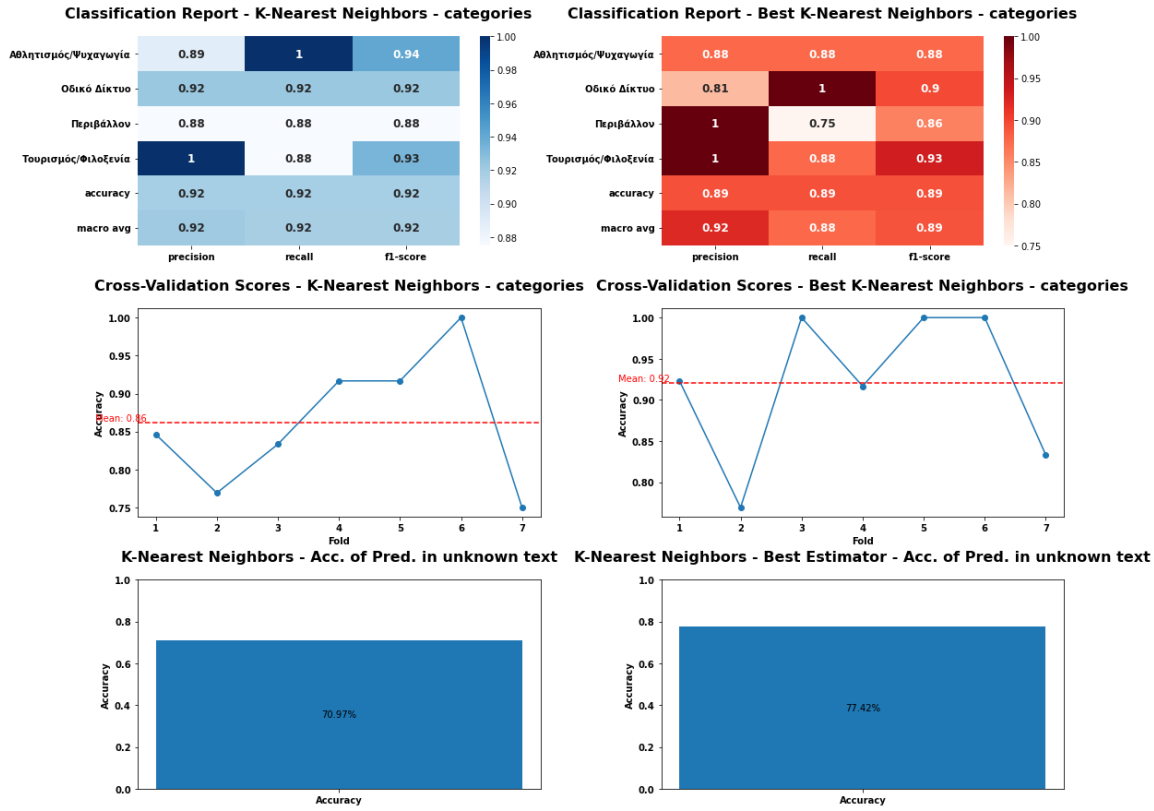
Εικόνα 37: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν κατηγοριοποίηση κειμένου με χρήση αλγορίθμου Decision Trees

Στην εικόνα 37 φαίνονται τα αποτελέσματα για το μοντέλο Decision Trees. Στο classification report το μοντέλο με GridSearchCV έχει κάποια βελτίωση όχι όμως κάτι αξιοσημείωτο. Στο Cross-Validation υπάρχει κάποια βελτίωση αλλά και πάλι δεν είναι σημαντική. Το ίδιο και στα αποτελέσματα accuracy σε νέα δεδομένα.



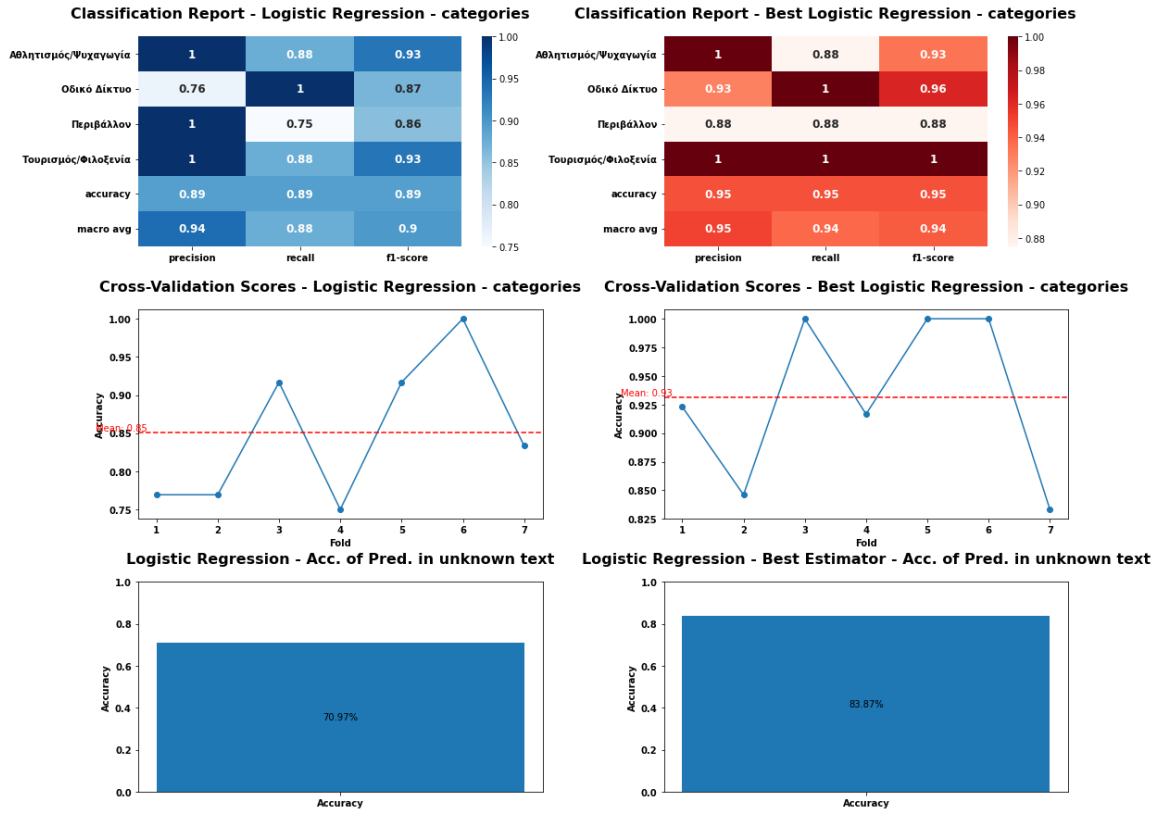
Εικόνα 38: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν κατηγοριοποίηση κειμένου με χρήση αλγορίθμου Random Forest

Στην εικόνα 38 φαίνονται τα αποτελέσματα για το μοντέλο Random Forest. Στο classification report και στις δύο περιπτώσεις παρατηρούνται σχεδόν ίδια αποτελέσματα. Δεν προέκυψε κάτι αξιόλογα καλύτερο ακόμα και μετά από τη χρήση και δοκιμή διαφορετικών παραμέτρων. Παρόλο που παρατηρείται μικρή αύξηση μέσου όρου στο Cross-vallidation διάγραμμα το αποτέλεσμα accuracy σε νέα δεδομένα είναι μικρότερα.



Εικόνα 39: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν κατηγοριοποίηση κειμένου με χρήση αλγορίθμου KNN

Στην εικόνα 39 φαίνονται τα αποτελέσματα για το μοντέλο K-Nearest Neighbors. Στο classification report στη πρώτη περίπτωση φαίνονται καλύτερα αποτελέσματα. Στο διάγραμμα Cross-Vallidation παρατηρείται καλύτερος μέσος όρος μετά τη χρήση GridSearchCV, καθώς επίσης καλύτερο accuracy και σε νέα δεδομένα.



Εικόνα 40: Classification report, cross-validation scores, accuracy σε άγνωστα δεδομένα που αφορούν κατηγοριοποίηση κειμένου με χρήση αλγορίθμου Logistic Regression

Στην εικόνα 40 φαίνονται τα αποτελέσματα για το μοντέλο Logistic Regression. Στο classification report το μοντέλο με GridSearchCV έχει καλύτερη απόδοση σε όλους τους δείκτες. Σημαντική αύξηση υπάρχει και στο μέσο όρο του διαγράμματος Cross-Validation καθώς σημαντική αύξηση υπάρχει και στο ποσοστό accuracy σε νέα δεδομένα.

6.3 Συμπεράσματα εκπαίδευσης μοντέλων

6.3.1 Συμπεράσματα συναισθηματικής ανάλυση

Πίνακας 6: Αποτελέσματα εκπαίδευσης αλγορίθμων για την περίπτωση της συναισθηματικής ανάλυσης

Grid Or Not	Target	Model	F1-score	Accuracy	Accuracy in new data
grid search	sentiment	SVM	0,86	0,86	93,55%
grid search	sentiment	Naive Bayes	0,86	0,86	87,10%
grid search	sentiment	K-Nearest Neighbors	0,84	0,84	90,32%
no grid search	sentiment	K-Nearest Neighbors	0,83	0,84	87,10%
grid search	sentiment	Logistic Regression	0,83	0,84	87,10%
no grid search	sentiment	Naive Bayes	0,80	0,81	87,10%
no grid search	sentiment	SVM	0,75	0,78	74,19%
no grid search	sentiment	Logistic Regression	0,75	0,78	74,19%
grid search	sentiment	Random Forest	0,68	0,73	64,52%
no grid search	sentiment	Random Forest	0,67	0,70	64,52%
grid search	sentiment	Decision Trees	0,66	0,68	54,84%
no grid search	sentiment	Decision Trees	0,56	0,57	51,61%

Στον πίνακα 6 φαίνονται συγκεντρωτικά στοιχεία των μοντέλων από τη μεγαλύτερη προς τη μικρότερη απόδοση για την περίπτωση του sentiment analysis.

Αναλυτικότερα τα μοντέλα SVM και Naive Bayes έχουν καλύτερη επίδοση με GridSearchCV εμφανίζοντας υψηλή απόδοση με F1-score 0,86 και accuracy 0,86. Και το accuracy στα νέα δεδομένα είναι επίσης υψηλό 93,55% και 87,10% αντίστοιχα. Το μοντέλο SVM εμφάνισε μεγαλύτερο accuracy στα νέα δεδομένα με την εφαρμογή GridSearchCV, ενώ στο Naive Bayes δεν υπήρξε βελτίωση. Το μοντέλο K-Nearest Neighbours με και χωρίς GridSearchCV επίσης, αποδίδει καλά, αλλά με ελαφρώς χαμηλότερο accuracy σε σχέση με τα προηγούμενα. Και παρουσίασε καλυτέρευση στο accuracy με 90.32% σε νέα δεδομένα στη με χρήση GridSearchCV. Το μοντέλο Logistic Regression έχει καλή γενικά απόδοση F1-score 0,83 και accuracy 0,84 και παρουσίασε βελτίωση με τη χρήση GridSearchCV σε όλες τις μετρήσεις. Τα μοντέλα Random Forest και Decision Trees έχουν τις χαμηλότερες αποδόσεις και παρουσιάζουν πάρα πολύ μικρή βελτίωση ακόμα και με τη χρήση GridSearchCV. Τα μοντέλα SVM, Naïve Bayes, K-Nearest Neighbors και Logistic Regression αναμενόταν να πάνε καλύτερα από τα υπόλοιπα μοντέλα, λόγω του ότι έχουν γενικά αποδείξει, ότι έχουν καλά αποτελέσματα σε παρόμοια θέματα κατηγοριοποίησης. Τα μοντέλα Decision Trees και Random Forest αναμενόταν να πάνε λιγότερο καλά, όχι όμως τόσο χαμηλά, λόγω του ότι μπορούν να χειριστούν πολύπλοκους κανόνες ταξινόμησης,

όπως αναμενόταν να παρουσιάσουν και μεγαλύτερη βελτίωση με τη χρήση GridSearchCV που δεν έγινε. [68] [69]

6.3.2 Συμπεράσματα κατηγοριοποίησης κειμένου

Πίνακας 7: Αποτελέσματα εκπαίδευσης αλγορίθμων για την περίπτωση της κατηγοριοποίησης κειμένου

Grid Or Not	Target	Model	F1-score	Accuracy	Accuracy in new data
grid search	categories	SVM	1,00	1,00	96,77%
grid search	categories	Naive Bayes	0,97	0,97	83,87%
grid search	categories	Logistic Regression	0,94	0,95	83,87%
grid search	categories	Random Forest	0,92	0,92	70,97%
no grid search	categories	K-Nearest Neighbors	0,92	0,92	70,97%
no grid search	categories	Random Forest	0,92	0,92	74,19%
no grid search	categories	Logistic Regression	0,90	0,89	70,97%
grid search	categories	K-Nearest Neighbors	0,89	0,89	77,42%
no grid search	categories	Naive Bayes	0,84	0,84	61,29%
no grid search	categories	SVM	0,81	0,81	61,29%
grid search	categories	Decision Trees	0,81	0,81	67,74%
no grid search	categories	Decision Trees	0,75	0,76	64,52%

Στον πίνακα 7 φαίνονται συγκεντρωτικά στοιχεία των μοντέλων από τη μεγαλύτερη προς τη μικρότερη απόδοση για την περίπτωση του text classification.

Αναλυτικότερα το μοντέλο SVM έχει την καλύτερη απόδοση με GridSearchCV εμφανίζοντας F1-score 1,00, accuracy 1,00 και σε άγνωστα δεδομένα πετυχαίνει 96,77% accuracy. Επίσης είχε αρκετά μεγάλη βελτίωση μετά τη χρήση GridSearchCV. Τα μοντέλα Naive Bayes και Logistic Regression είχαν επίσης πολύ καλή απόδοση με F1-score 0,97 accuracy 0.97 και F1-score 0,94 accuracy 0,95 αντίστοιχα και 83,87% accuracy σε άγνωστα δεδομένα. Και στα δύο μοντέλα υπήρξε βελτίωση απόδοσης μετά τη χρήση GridSearchCV. Το μοντέλο Random Forest έχει καλή απόδοση και δεν εμφανίζει βελτίωση μετά τη χρήση GridSearchCV. Επίσης το accuracy σε άγνωστα δεδομένα είναι μεγαλύτερο χωρίς τη χρήση GridSearchCV. Το μοντέλο K-Nearest Neighbors έχει καλή απόδοση και εμφανίζει μικρή βελτίωση μετά τη χρήση GridSearchCV. Επίσης το accuracy σε άγνωστα δεδομένα είναι μεγαλύτερο χωρίς τη χρήση GridSearchCV. Τέλος το μοντέλο Decision Trees εμφανίζει μικρή βελτίωση μετά τη χρήση GridSearchCV.

Συμπερασματικά θα μπορούσε να αναφερθεί ότι χρησιμοποιώντας τη μέθοδο με το εργαλείο GridSearchCV με διαφορετικές παραμέτρους στα μοντέλα, παρατηρούνται γενικά καλύτερα αποτελέσματα. Τα αποτελέσματα καθιστούν εμφανή τη σημασία της χρήσης μεθόδων βελτιστοποίησης υπερπαραμέτρων. Όσον αφορά την ανάλυση συναισθήματος, καλύτερα αποτελέσματα παρατηρήθηκαν με τη χρήση SVM & Naive Bayes. Και στην κατηγοριοποίηση κειμένου η καλύτερη επίδοση παρατηρήθηκε με SVM, αλλά και οι αλγόριθμοι Naive Bayes και Logistic Regression παρέχουν επίσης αξιόλογα αποτελέσματα. Σημαντικό ρόλο επίσης, έχει και το ποσοστό accuracy σε νέα δεδομένα. Καλύτερες αποδόσεις για την συναισθηματική ανάλυση αλλά και στην κατηγοριοποίηση κειμένου παρατηρήθηκαν με το μοντέλο SVM. Η καλή πρόβλεψη του μοντέλου σε δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευσή του, μας δείχνει ότι λειτουργεί σωστά και μπορεί να κάνει προβλέψεις σε ικανοποιητικό βαθμό.

7. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΜΕΛΛΟΝΤΙΚΕΣ ΕΠΕΚΤΑΣΕΙΣ

Στη συγκεκριμένη εργασία τα αποτελέσματα είναι γενικά θετικά. Οι προβλέψεις είναι σε ικανοποιητικό επίπεδο και σίγουρα υπάρχει χώρος για περαιτέρω βελτιώσεις. Φαίνεται με χαρακτηριστικό τρόπο η αποτελεσματικότητα χρήσης της ML, πως μπορεί να βοηθήσει και μάλιστα σε περιβάλλον που έχει κείμενο. Η αυτοματοποίηση διαδικασιών όπως της κατηγοριοποίησης κειμένου και ανάλυσης συναισθήματος προσφέρει λύσεις σε προβλήματα που παλαιότερα θα χρειαζόντουσαν πολύ χρόνο για να δείξουν αποτελέσματα. Είδαμε πως η χρήση εργαλείων που προσφέρουν δοκιμές σε διαφορετικές παραμέτρους είναι ένας καλός τρόπος να αυξηθεί η αποδοτικότητα ενός μοντέλου. Μία από τις παραμέτρους που θα βοηθούσε αρκετά θα ήταν η χρήση μίας πολύ μεγαλύτερης βάσης δεδομένων. Οπότε μία μελλοντική αύξηση καταχωρήσεων θα μπορούσε να επιφέρει βελτιώσεις στην απόδοση των μοντέλων. Αυτό θα μπορούσε να αυξήσει σημαντικά και τις κεντρικές κατηγορίες πάνω από τις τέσσερις που έχουμε τώρα, βελτιώνοντας ταυτόχρονα και τις γνώσεις της δημόσιας υπηρεσίας σχετικά με τα θέματα που έχει ένας πολίτης με αποτέλεσμα καλύτερες δράσεις και λύσεις σε προβλήματα πιο στοχευμένα από την πλευρά της δημόσιας υπηρεσίας. Με μία πολύ μεγαλύτερη βάση δεδομένων θα μπορούσε μελλοντικά να εξεταστεί η χρήση νευρωνικών δικτύων και η αποτελεσματικότερη πρόβλεψη καταχωρήσεων.

Ακόμα θα μπορούσε να εξεταστεί η αναβάθμιση της εφαρμογής για να υποστηρίζει διαφορετικές γλώσσες. Ή ακόμα και η ενσωμάτωση δια-δραστικής χρήσης με τους χρήστες δίνοντάς τους τη δυνατότητα να καταχωρούν σχόλια σχετικά με τα αποτελέσματα της καταχώρησης. Αυτή η διαδικασία ανάδρασης μπορεί να χρησιμοποιηθεί για τη συνεχή βελτίωση της απόδοσης της εφαρμογής με την επανεκπαίδευση των μοντέλων με δεδομένα που δημιουργούνται από το χρήστη και τη λεπτομερή ρύθμιση των αλγορίθμων ανάλογα.

8. ΠΙΝΑΚΑΣ ΟΡΟΛΟΓΙΑΣ

Ξενόγλωσσος όρος	**Ελληνικός Όρος**
Accuracy	Ακρίβεια
Artificial Intelligence	Τεχνητή Νοημοσύνη
Classification Report	Αναφορά Ταξινόμησης
Convolutional Neural Network	Συνελικτικά Νευρωνικά Δίκτυα
Cross-validation	Διασταυρωμένη Επικύρωση
Data Analytics	Ανάλυση Δεδομένων
Data Cleaning	Καθαρισμός Δεδομένων
Data Engineering	Μηχανική Δεδομένων
Data Preprocessing	Προεπεξεργασία Δεδομένων
Data Science	Επιστήμη Δεδομένων
Decision Support System	Συστήματος Υποστήριξης Αποφάσεων
Deep Learning	Βαθιά Μάθηση
Deep Neural Networks	Βαθιά Νευρωνικά Δίκτυα
Grid Search	Αναζήτηση Πλέγματος
Machine Learning	Μηχανική Μάθηση
Named Entity Resolution	Ονομαστική Ανάλυση Οντότητας
Natural Language Processing	Επεξεργασία Φυσικής Γλώσσας
Neural Networks	Νευρωνικά Δίκτυα
Pipeline	Αγωγός
Precision	Ακρίβεια
Recall	Ανάκληση
Recurrent Neural Networks	Αναδρομικά Νευρωνικά Δίκτυα
Region Based Convolutional Neural Networks	Αναδρομικά Συνελικτικά Νευρωνικά Δίκτυα
Reinforcement Learning	Ενισχυτική Μάθηση
Semi-supervised Machine Learning	Ημι-εποπτευόμενη Μηχανική Μάθηση
Supervised Machine Learning	Εποπτευόμενη Μηχανική Μάθηση
Text Classification	Ταξινόμηση Κειμένου
Text Mining	Συγκέντρωση Δεδομένων
Unsupervised Machine Learning	Μη εποπτευόμενη Μηχανική Μάθηση
Web Development	Ανάπτυξη Ιστού
Web Interface	Διεπαφή Ιστού
Web Server	Διακομιστή Ιστού

9. ΣΥΝΤΜΗΣΕΙΣ – ΑΡΚΤΙΚΟΛΕΞΑ – ΑΚΡΩΝΥΜΙΑ

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Networks
DSS	Συστήματος Υποστήριξης Αποφάσεων
GPT	Generative Pre-trained Transformer
ML	Machine Learning
NLP	Natural Language Processing
NN	Neural Networks
RCNN	Region Based Convolutional Neural Networks
RNN	Recurrent Neural Networks
SKCV	Stratified K-Fold Cross-Validation
SVM	Support Vector Machines
WSG	Web Server Gateway Interface
ΕΕ	Ευρωπαϊκή Ένωση
ΤΠΕ	Τεχνολογίες Πληροφορικής και Επικοινωνιών

10. ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] “Ηλεκτρονική Διακυβέρνηση”, opengov.minedu.gov.gr, Jan. 31, 2024. https://opengov.minedu.gov.gr/ηλεκτρονική-διακυβέρνηση/ (Προσπελάστηκε Jan. 31, 2024)

[2] “ΥΠΟΥΡΓΕΙΟ ΔΙΟΙΚΗΤΙΚΗΣ ΜΕΤΑΡΡΥΘΜΙΣΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΗΣ ΔΙΑΚΥΒΕΡΝΗΣΗΣ, “ΣΤΡΑΤΗΓΙΚΗ ΓΙΑ ΤΗΝ ΗΛΕΚΤΡΟΝΙΚΗ ΔΙΑΚΥΒΕΡΝΗΣΗ 2014-2020”, opengov.gr, Jan. 31, 2024. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://www.opengov.gr/minreform/wp-content/uploads/downloads/2014/02/stratigiki-ilektron.-diakyv.-teliko-pdf1.pdf, p. 9, (Προσπελάστηκε Jan. 31, 2024)

[3] “Σχετικά με το gov.gr”, www.gov.gr, Jan. 31, 2024. https://www.gov.gr/info/about-us (Προσπελάστηκε Jan. 31, 2024)

[4] “Πολιτική χρήσης mAIgov”, www.gov.gr, Jan. 31, 2024. https://www.gov.gr/info/politiki-xrisis-maigov (Προσπελάστηκε Jan. 31, 2024)

[5] Hobson Lane, Cole Howard and Hannes Max Hapke, “Natural Language Processing in Action”, 2019, Manning Publications Co., p. 4

[6] Ian Goodfellow, Yoshua Bengio and Aaron Courville, “Deep Learning”, p. 463

[7] WILLIAM D. EGGERS, NEHA MALIK and MATT GRACIE, “Using AI to unleash the power of unstructured government data”, p. 4.

[8] “Evolution Of Natural Language Processing (NLP)”, medium.com, May. 28, 2023. medium.com/analytics-vidhya/evolution-of-natural-language-processing-nlp (Προσπελάστηκε May. 28, 2023)

[9] WILLIAM D. EGGERS, NEHA MALIK and MATT GRACIE, “Using AI to unleash the power of unstructured government data”, deloitte.com, May 28, 2023, https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/public-sector/lu-ai-unstructured-government-data.pdf,p. 4, (Προσπελάστηκε May. 28, 2023)

[10] “What is natural language processing (NLP)?”, “ibm.com”, May. 28, 2023, https://www.ibm.com/topics/natural-language-processing, (Προσπελάστηκε May. 28, 2023)

[11] “Why NLP is important and it’ll be the future — our future”, towardsdatascience.com, May. 28, 2023, https://towardsdatascience.com/why-nlp-is-important-and-itll-be-the-future-our-future-59d7b1600dda (Προσπελάστηκε May. 28, 2023)

[12] “Natural language processing”, en.wikipedia.org, Feb. 1, 2024. https://en.wikipedia.org/wiki/Natural_language_processing#History (Προσπελάστηκε Feb. 1, 2024)

[13] “Rule Based Approach in NLP”, www.geeksforgeeks.org, Feb. 1, 2024. https://www.geeksforgeeks.org/rule-based-approach-in-nlp/ (Προσπελάστηκε Feb. 1, 2024)

[14] “Natural Language Processing”, www.deeplearning.ai, Feb. 1, 2024. https://www.deeplearning.ai/resources/natural-language-processing/ (Προσπελάστηκε Feb. 1, 2024)

[15] Wenpeng Yin, Katharina Kann, Mo Yu and Hinrich Schutze, “ComparativeStudyofCNNandRNNforNaturalLanguageProcessing”, arxiv.org, Feb. 2017, p.1, doi: 10.48550/arXiv.1702.01923

[16] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh, “Natural language processing: state of the art, current trends and challenges”, Springer, Jul. 2022, p.13, doi: 10.1007/s11042-022-13428-4

[17] “Processing (NLP)”, monkeylearn.com, Feb. 2, 2024. https://monkeylearn.com/blog/nlp-benefits/ (Προσπελάστηκε Feb. 2, 2024)

[18] “Natural Language Processing and Machine Learning”, www.encora.com, Feb. 2, 2024. https://www.encora.com/insights/natural-language-processing-and-machine-learning (Προσπελάστηκε Feb. 2, 2024)

[19] “NLP Tutorial”, www.javatpoint.com, Feb. 2, 2024. https://www.javatpoint.com/nlp (Προσπελάστηκε Feb. 2, 2024)

[20] “Everything you need to know about Machine Learning”, analyticsvidhya.com, May. 31, 2023, https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/ (Προσπελάστηκε May. 31, 2023)

[21] Nikolaos Spatiotis, Iosif Mporas, Michael Paraskevas, Isidoros Perikos, “Sentiment Analysis for the Greek Language”, researchgate.net, Nov. 2016, p.2, doi: 10.1145/3003733.3003769

[22] Pramod Singh, “Machine Learning with PySpark”, Apress Berkeley CA, 2019

[23] “Machine Learning: What is ML and how does it work?”, algotive.ai, May. 31, 2023, https://www.algotive.ai/blog/machine-learning-what-is-ml-and-how-does-it-work (Προσπελάστηκε May. 31, 2023)

[24] Varun Dogra, Sahil Verma, Kavita, Pushpita Chatterjee, Jana Shafi, Jaeyoung Choi and Muhammad Fazal Ijaz, “A Complete Process of Text Classification System Using State-of-the-Art NLP Models”, www.hindawi.com, June. 2022, p.2, doi: 10.1155/2022/1883698

[25] “Selected AI cases in the public sector”, ec.europa.eu, June 4, 2023, https://data.jrc.ec.europa.eu/dataset/7342ea15-fd4f-4184-9603-98bd87d8239a#publications (Προσπελάστηκε June 6, 2023)

[26] Francesco Pignatelli, “AI Watch European Landscape on the Use of Artificial Intelligence by the Public Sector”, https://ec.europa.eu/jrc, 2022, p.7, doi: 10.2760/39336

[27] Gianluca Misuraca, “AI Watch Artificial Intelligence in public services”, https://ec.europa.eu/jrc, 2020, p.7, doi: 10.2760/039619

[28] Gianluca Misuraca, “AI Watch Artificial Intelligence in public services”, https://ec.europa.eu/jrc, 2020, p.16, doi: 10.2760/039619

[29] Francesco Pignatelli, “AI Watch European Landscape on the Use of Artificial Intelligence by the Public Sector”, https://ec.europa.eu/jrc, 2022, p.36, doi: 10.2760/39336

[30] Francesco Pignatelli, “AI Watch European Landscape on the Use of Artificial Intelligence by the Public Sector”, https://ec.europa.eu/jrc, 2022, p.38, doi: 10.2760/39336

[31] Lara Quijano-Sánchez, Federico Liberatore, José Camacho-Collados, Miguel Camacho-Collados, “Applying automatic text-based detection of deceptive language to police reports: Extracting behavioral patterns from a multi-step classification model to understand how we lie to the police”, sciencedirect.com, March 2018, https://www.sciencedirect.com/science/article/abs/pii/S095070511830128X?via%3Dihub#preview-section-abstract (Προσπελάστηκε June 6, 2023)

[32] Diksha Joshi, Mayuresh Khalegaonkar, Mohini Lohikpure, Preeti Maan, Prof. R. A. Deshmukh, “SENTIMENTAL ANALYSIS ON E-GOVERNANCE”, ijirse.com, May 2017,

http://ijirse.com/wp-content/upload/2017/03/PY2087ijirse.pdf (Προσπελάστηκε June 6, 2023)

[33] G. Koteswara Rao, Shubhamoy Dey, “DECISION SUPPORT FOR E-GOVERNANCE: A TEXT MINING APPROACH”, IJMIT, Aug. 2011, p.14, doi: 10.48550/arXiv.1108.6198

[34] Tulu Tilahun, Durga Prasad Sharma, “Design and Development of E-Governance Model for Service Quality Enhancement”, www.scirp.org, 2015, doi: 10.4236/jdaip.2015.33007

[35] “spaCy”, spacy.io, May. 28, 2023, https://spacy.io/ (Προσπελάστηκε May. 28, 2023)

[36] “el_core_news_lg”, spacy.io, May. 28, 2023, https://spacy.io/models/el#el_core_news_lg (Προσπελάστηκε May. 28, 2023)

[37] I.Rish, “An empirical study of the naïve Bayes classifier”, T.J.WatsonResearchCenter, 2001, p.1

[38] Shuo Xu, Yan Li, and Zheng Wang, “Bayesian Multinomial Naïve Bayes Classifier

to Text Classification”, Institute of Scientific and Technical Information of China, 2017, p.1, doi: 10.1007/978-981-10-5041-1_57

[39] Jair Cervantes, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, Asdrubal Lopez, “A comprehensive survey on support vector machine classification: Applications, challenges and trends”, www.sciencedirect.com, 2020, p.1,21, doi: 10.1007/978-981-10-5041-1_57

[40] Orphanos Giorgos, Kalles Dimitris, Papagelis Thanasis and Christodoulakis Dimitris, “Decision Trees and NLP: A Case Study in POS Tagging”, Proceedings of Annual Conference on Artificial Intelligence (ACAI), 1999, p.4

[41] Nasir Jalal, Arif Mehmood, Gyu Sang Choi, Imran Ashraf, “A novel improved random forest for text classification using feature ranking and optimal number of trees”, www.sciencedirect.com, 2022, p.1, doι: 10.1016/j.jksuci.2022.03.012

[42] Kanish Shah, Henil Patel, Devanshi Sanghvi, Manan Shah, “A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification”, www.springer.com, 2019, p.12, doι: 10.1007/s41133-020-00032-0

[43] “How To Implement Logistic Regression Text Classification In Python With Scikit-learn and PyTorch”, spotintelligence.com, Feb. 19, 2023, https://spotintelligence.com/2023/02/22/logistic-regression-text-classification-python/#Why_use_logistic_regression (Προσπελάστηκε Feb. 19, 2023)

[44] Petro Liashchynskyi, Pavlo Liashchynskyi, “Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS”, arxiv.org, Dec. 2019, p.3, doi: 10.48550/arXiv.1912.06059

[45] “GridSearchCV”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (Προσπελάστηκε May. 28, 2023)

[46] “MultinomialNB”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html (Προσπελάστηκε May. 28, 2023)

[47] “SVC”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html (Προσπελάστηκε May. 28, 2023)

[48] “DecisionTreeClassifier”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (Προσπελάστηκε May. 28, 2023)

[49] “RandomForestClassifier”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (Προσπελάστηκε May. 28, 2023)

[50] “KNeighborsClassifier”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (Προσπελάστηκε May. 28, 2023)

[51] “LogisticRegression”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (Προσπελάστηκε May. 28, 2023)

[52] “Pipeline”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html (Προσπελάστηκε May. 28, 2023)

[53] “CountVectorizer”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html (Προσπελάστηκε May. 28, 2023)

[54] “TfidfTransformer”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html (Προσπελάστηκε May. 28, 2023)

[55] “Joblib: running Python functions as pipeline jobs”, joblib.readthedocs.io/, May. 28, 2023, https://joblib.readthedocs.io/en/stable/ (Προσπελάστηκε May. 28, 2023)

[56] Hossin M., Sulaiman M.N., “A REVIEW ON EVALUATION METRICS FOR DATA CLASSIFICATION EVALUATIONS”, IJDKP, Mar. 2015, p.4, doi: 10.5121/ijdkp.2015.520

[57] Payam Refaeilzadeh, Lei Tang, Huan Liu, “Cross-Validation”, link.springer.com, 2009, p.532, doi: 10.1007/978-0-387-39940-9_565

[58] “StratifiedKFold”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html (Προσπελάστηκε May. 28, 2023)

[59] Sashikanta Prusty, Srikanta Patnaik, Sujit Kumar Dash, “SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer”, frontiersin.org, Aug. 2022, p.2, doi: 10.3389/fnano.2022.97242

[60] “A Brief History of Python”, learnpython.com, May. 28, 2023, https://learnpython.com/blog/history-of-python/ (Προσπελάστηκε May. 28, 2023)

[61] “What is Flask Python”, pythonbasics.org, May. 28, 2023, https://pythonbasics.org/what-is-flask-python/ (Προσπελάστηκε May. 28, 2023)

[62] “Flask (web framework)”, en.wikipedia.org, May. 28, 2023, https://en.wikipedia.org/wiki/Flask_(web_framework) (Προσπελάστηκε May. 28, 2023)

[63] "XAMPP”, en.wikipedia.org, May. 28, 2023, https://en.wikipedia.org/wiki/XAMPP (Προσπελάστηκε May. 28, 2023)

[64] “pandas”, pypi.org, May. 28, 2023, https://pypi.org/project/pandas/ (Προσπελάστηκε May. 28, 2023)

[65] “scikit-learn”, scikit-learn.org, May. 28, 2023, https://scikit-learn.org/stable/ (Προσπελάστηκε May. 28, 2023)

[66] “matplotlib”, pypi.org, May. 28, 2023, https://pypi.org/project/matplotlib/ (Προσπελάστηκε May. 28, 2023)

[67] “seaborn: statistical data visualization”, seaborn.pydata.org, May. 28, 2023, https://seaborn.pydata.org/ (Προσπελάστηκε May. 28, 2023)

[68] Ahmed H. Aliwy and Esraa H. Abdul Ameer, “Comparative Study of Five Text Classification Algorithms with their Improvements”, www.researchegate.net, Jan. 2017, p.4309-4313,

[69] Ruiguang Li, Ming Liu, Dawei Xu, Jiaqi Gao, Fudong Wu & Liehuang Zhu, “A Review of Machine Learning Algorithms for Text Classification”, link.springer.com, 2022, p.228-233, doi: 10.1007/978-981-16-9229-1_14