ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Εξατομικευμένο Σύστημα Συγκέντρωσης Ειδήσεων και Προτάσεων με χρήση Επεξεργασίας Φυσικής Γλώσσας: Συγκριτική Μελέτη Μοντέλων Βασισμένων σε κανόνες και Μοντέλων Μηχανικής Μάθησης

Ηλίας Κωνσταντινίδης
ΑΜ. 711171214

Επιβλέπων:
Χρήστος Τρούσσας
Επίκουρος Καθηγητής

Μέλη εξεταστικής Επιτροπής

Η Διπλωματική Εργασία έγινε αποδεκτή και βαθμολογήθηκε από την εξής τριμελή επιτροπή :

Ονοματεπώνυμο	Βαθμίδα	Υπογραφή
Χρήστος Τρούσσας	**Επ.Καθηγητής**	Christos Troussas Digitally signed by Christos Troussas Date: 2024.10.09 12:42:51 +03'00'
Ακριβή Κρούσκα	**Μέλος ΕΔΙΠ**	Akrivi Krouska Digitally signed by Akrivi Krouska Date: 2024.10.09 19:21:38 +03'00'
Παναγιώτα Τσελέντη	**Μέλος ΕΔΙΠ**	Panagiota Tselenti Digitally signed by Panagiota Tselenti Date: 2024.10.09 18:20:53 +03'00'

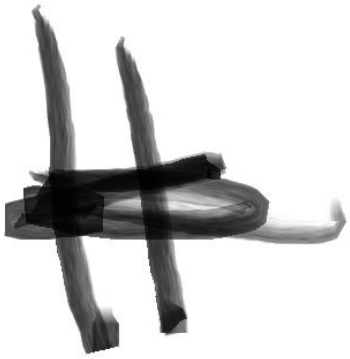
ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Ηλίας Κωνσταντινίδης του Αντώνη, με αριθμό μητρώου 711171214 φοιτητής του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών Πληροφορικής του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Βεβαιώνω ότι είμαι συγγραφέας αυτής της Διπλωματικής εργασίας και κάθε βοήθεια την οποία είχα για την προετοιμασία της, είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο/Η Δηλών/ούσα



(Υπογραφή)

Περιεχόμενο

ΠΕΡΙΛΗΨΗ

Η παρούσα διατριβή αποσκοπεί στη σύγκριση δύο μοντέλων για την ανάπτυξη ενός συστήματος συστάσεων ειδήσεων. Το πρώτο μοντέλο βασίζεται σε κανόνες και χρησιμοποιεί ανάλυση συναισθήματος μέσω του αλγορίθμου Vader, ενώ το δεύτερο μοντέλο χρησιμοποιεί μηχανική μάθηση και λογιστική παλινδρόμηση για την ανάλυση συναισθήματος. Στόχος είναι να διερευνηθεί η απόδοση των δύο προσεγγίσεων και να αξιολογηθεί η ακρίβειά τους στην κατηγοριοποίηση και σύσταση ειδήσεων με βάση το συναίσθημα και τον τύπο του άρθρου. Η επεξεργασία φυσικής γλώσσας μπορεί να βελτιώσει την ακρίβεια των εξατομικευμένων προτάσεων, επιτρέποντας μια πιο προηγμένη κατανόηση του περιεχομένου του κειμένου, εντοπίζοντας με ακρίβεια τις προθέσεις και τα συναισθήματα πίσω από τις λέξεις και παρέχοντας προτάσεις που ανταποκρίνονται περισσότερο στα ενδιαφέροντα των χρηστών.

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

- **ML = Machine Learning**
- **NLP = Natural Language Processing**
- **TF-IDF = Term Frequency – Inverse Document Frequency**
- **NLTK = Natural Language Toolkit**

ΕΙΣΑΓΩΓΗ

Στην ψηφιακή εποχή, ο μεγάλος όγκος των πληροφοριών που είναι διαθέσιμες στο διαδίκτυο έχει καταστήσει όλο και πιο δύσκολο για τους χρήστες να βρουν περιεχόμενο που ανταποκρίνεται στα ενδιαφέροντά τους. Για την αντιμετώπιση αυτής της πρόκλησης, τα συστήματα συστάσεων έχουν γίνει αναπόσπαστο μέρος των πλατφορμών που παρέχουν άρθρα, ειδήσεις και άλλες υπηρεσίες πλούσιες σε περιεχόμενο. Τα συστήματα αυτά καθοδηγούν τους χρήστες προς το περιεχόμενο που ταιριάζει στις προτιμήσεις τους, βελτιώνοντας τη δέσμευση και την ικανοποίηση των χρηστών.

Παραδοσιακά, τα συστήματα συστάσεων βασίζονται σε μοντέλα βασισμένα σε κανόνες, τα οποία λειτουργούν με βάση ένα σύνολο προκαθορισμένων κανόνων και λογικής. Αυτά τα μοντέλα είναι σχετικά απλά στη σχεδίαση και την υλοποίηση, γεγονός που τα καθιστά δημοφιλή επιλογή στα πρώτα στάδια της ανάπτυξης συστημάτων συστάσεων. Ωστόσο, με την έλευση της μηχανικής μάθησης, εμφανίστηκε μια νέα κατηγορία μοντέλων, ικανών να μαθαίνουν από τη συμπεριφορά των χρηστών και να προσαρμόζονται με την πάροδο του χρόνου. Αυτά τα μοντέλα μηχανικής μάθησης έχουν τη δυνατότητα να παρέχουν πιο εξατομικευμένες και ακριβείς συστάσεις, αποκαλύπτοντας πρότυπα και σχέσεις στα δεδομένα που τα μοντέλα που βασίζονται σε κανόνες μπορεί να παραβλέπουν.

Στόχος αυτής της διπλωματικής άσκησης είναι να διερευνήσει και να συγκρίνει την αποτελεσματικότητα των μοντέλων που βασίζονται σε κανόνες και των μοντέλων μηχανικής μάθησης στο πλαίσιο συστημάτων συστάσεων άρθρων. Η σύγκριση αυτή θα στηριχθεί τόσο σε θεωρητικές προοπτικές όσο και σε πρακτικές εφαρμογές, προσφέροντας πληροφορίες σχετικά με τα πλεονεκτήματα και τους περιορισμούς κάθε προσέγγισης. Με την εξέταση βασικών πτυχών όπως η ακρίβεια, η προσαρμοστικότητα και η πολυπλοκότητα της υλοποίησης, η μελέτη αυτή αποσκοπεί στην παροχή μιας ολοκληρωμένης κατανόησης του τρόπου με τον οποίο αυτοί οι δύο τύποι μοντέλων αποδίδουν σε πραγματικές συνθήκες.

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ ΚΑΙ ΑΝΑΣΚΟΠΗΣΗ ΤΗΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

1.1 Επισκόπηση των συστημάτων συστάσεων (Recommendation Systems)

Τα συστήματα συστάσεων είναι μηχανές λογισμικού που προτείνουν αντικείμενα στους χρήστες με βάση τις προηγούμενες προτιμήσεις τους, την εμπλοκή τους με το προϊόν και την αλληλεπίδρασή τους, μεταξύ άλλων παραγόντων. Τα συστήματα συστάσεων διατηρούν το ενδιαφέρον των χρηστών για ό,τι συνιστά ο ιστότοπος. Οι μηχανές συστάσεων δημιουργούν μια εξατομικευμένη εμπειρία χρήστη βοηθώντας κάθε καταναλωτή ξεχωριστά να εντοπίσει και να ανακαλύψει τις αγαπημένες του ταινίες, τηλεοπτικές εκπομπές, ψηφιακά προϊόντα, βιβλία, άρθρα, υπηρεσίες και άλλα. Τα συστήματα αυτά βοηθούν τις επιχειρήσεις να αυξήσουν τις πωλήσεις τους, ενώ παράλληλα ωφελούν τους καταναλωτές. Τα συστήματα συστάσεων επιτρέπουν στους καταναλωτές να βρίσκουν εύκολα προϊόντα, προωθούν την ευκολία χρήσης και τους αναγκάζουν να παραμείνουν στον ιστότοπο αντί να τον εγκαταλείψουν. [1]

1.2 Η λειτουργία ενός συστήματος σύστασης

1.2.1 Συλλογή δεδομένων χρήστη

Τα συστήματα συστάσεων συλλέγουν δεδομένα με την παρακολούθηση ενεργειών του χρήστη, όπως κλικ, προβολές και αγορές. Λαμβάνουν επίσης υπόψη τα σχόλια των χρηστών, όπως αξιολογήσεις και κριτικές, μαζί με δημογραφικές πληροφορίες και συνήθειες περιήγησης. Τα δεδομένα αυτά βοηθούν στην κατανόηση των προτιμήσεων και της συμπεριφοράς των χρηστών. [2]

1.2.2. Ανάλυση δεδομένων

Αναλύοντας τα δεδομένα που συλλέγονται, τα συστήματα συστάσεων προβλέπουν τι μπορεί να αρέσει στους χρήστες. Εξετάζουν διάφορους παράγοντες, όπως το δημοφιλές περιεχόμενο, τα σχόλια των χρηστών και τα μοτίβα στη συμπεριφορά των χρηστών, για να κάνουν ακριβείς προτάσεις. Αυτή η ανάλυση διασφαλίζει ότι οι συστάσεις είναι σχετικές και εξατομικευμένες. [2]

1.2.3 Φιλτράρισμα

Τα συστήματα συστάσεων χρησιμοποιούν πολύπλοκους αλγορίθμους για την επεξεργασία των δεδομένων και τη δημιουργία συστάσεων. Εφαρμόζονται διαφορετικές μαθηματικές τεχνικές για τη βελτίωση των προτάσεων ανάλογα με τον τύπο του μοντέλου σύστασης που χρησιμοποιείται. Ο στόχος είναι να φιλτράρονται οι άσχετες επιλογές και να παρουσιάζονται στους χρήστες οι πιο κατάλληλες συστάσεις. [2]

1.2.4 Παραγωγή συστάσεων

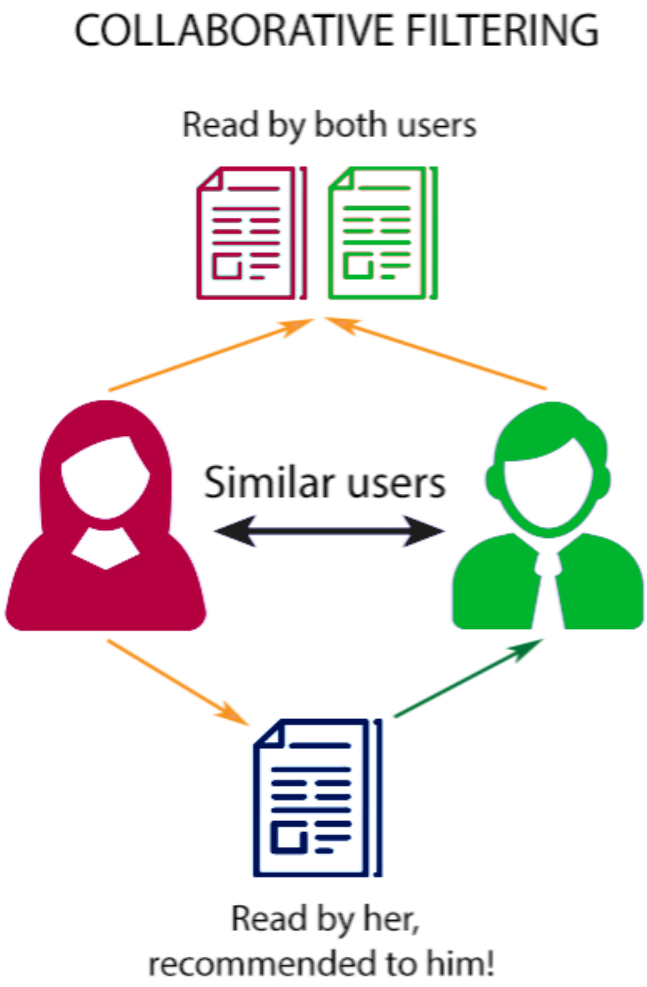
Τέλος, το σύστημα συστάσεων παράγει έναν κατάλογο πιθανών επιλογών με βάση την εισαγωγή του χρήστη. Στη συνέχεια, οι επιλογές αυτές κατατάσσονται με βάση τη συνάφεια τους με τις προτιμήσεις του χρήστη. Η τεχνητή νοημοσύνη παίζει καθοριστικό ρόλο σε αυτή τη διαδικασία, καθώς μαθαίνει και προσαρμόζεται συνεχώς για να παρέχει καλύτερες συστάσεις με την πάροδο του χρόνου. [2]

1.3 Τύποι συστημάτων συστάσεων

Υπάρχουν κυρίως τρεις μεθοδολογίες για τα συστήματα συστάσεων: συνεργατικό φιλτράρισμα (**collaborative filtering)**, φιλτράρισμα βάσει περιεχομένου (**content-based filtering**) και υβριδικά συστήματα (**hybrid Systems**).

1.3.1 Συνεργατικό φιλτράρισμα (collaborative filtering)

Η προσέγγιση του συνεργατικού φιλτραρίσματος συλλέγει δεδομένα σχετικά με τη συμπεριφορά των χρηστών και τα αναλύει για να προβλέψει τι θα αρέσει στους μεμονωμένους χρήστες. Αυτό γίνεται με την εύρεση ομοιοτήτων μεταξύ των χρηστών και τη χρήση αυτών για την πραγματοποίηση προβλέψεων. [3]



Εικόνα 1. Επισκόπηση του συνεργατικού φιλτραρίσματος [3]

Για παράδειγμα, στον χρήστη Α αρέσουν οι ταινίες του είδους δράση, περιπέτεια και επιστημονική φαντασία. Στον χρήστη Β αρέσουν οι ταινίες του είδους δράση, περιπέτεια και φαντασία. Έχουν παρόμοια ενδιαφέροντα. Έτσι, είναι πολύ πιθανό ο Α να του αρέσουν ταινίες του είδους φαντασία και ο Β να του αρέσουν ταινίες του είδους επιστημονική φαντασία. Με αυτόν τον τρόπο πραγματοποιείται το συνεργατικό φιλτράρισμα.

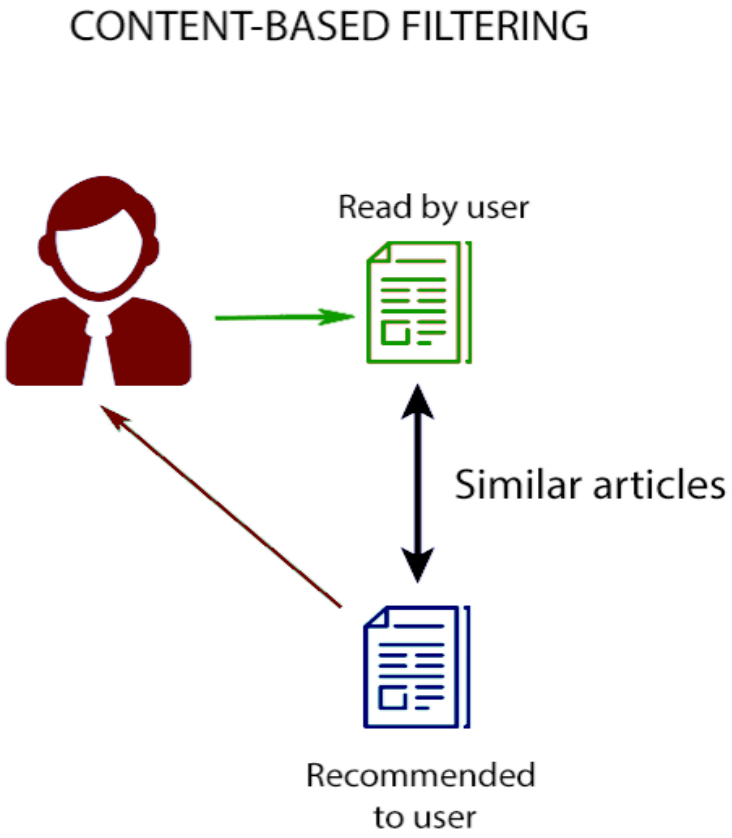
Τα δύο είδη τεχνικών συνεργατικού φιλτραρίσματος που χρησιμοποιούνται είναι:

- Συνεργατικό φιλτράρισμα χρήστη-χρήστη (**user-user**)
- Συνεργατικό φιλτράρισμα στοιχείο-στοιχείο (**item-item**)

Η ικανότητα αυτού του συστήματος συστάσεων να κάνει ακριβείς συστάσεις χωρίς να έχει καμία προηγούμενη γνώση του συνιστώμενου αντικειμένου είναι ένα από τα κύρια πλεονεκτήματά του. Δεν υπάρχει εξάρτηση από περιεχόμενο που μπορεί να αναλυθεί από υπολογιστές. [3]

1.3.2 Φιλτράρισμα βάσει περιεχομένου (Content-based filtering)

Οι μέθοδοι φιλτραρίσματος με βάση το περιεχόμενο βασίζονται στην περιγραφή του προϊόντος και στο προφίλ του χρήστη για τον προσδιορισμό των προτιμώμενων επιλογών του χρήστη. Τα προϊόντα περιγράφονται με λέξεις-κλειδιά σε αυτό το σύστημα συστάσεων και δημιουργείται ένα προφίλ χρήστη που εκφράζει τον τύπο του αντικειμένου που προτιμά ο συγκεκριμένος χρήστης. [3]

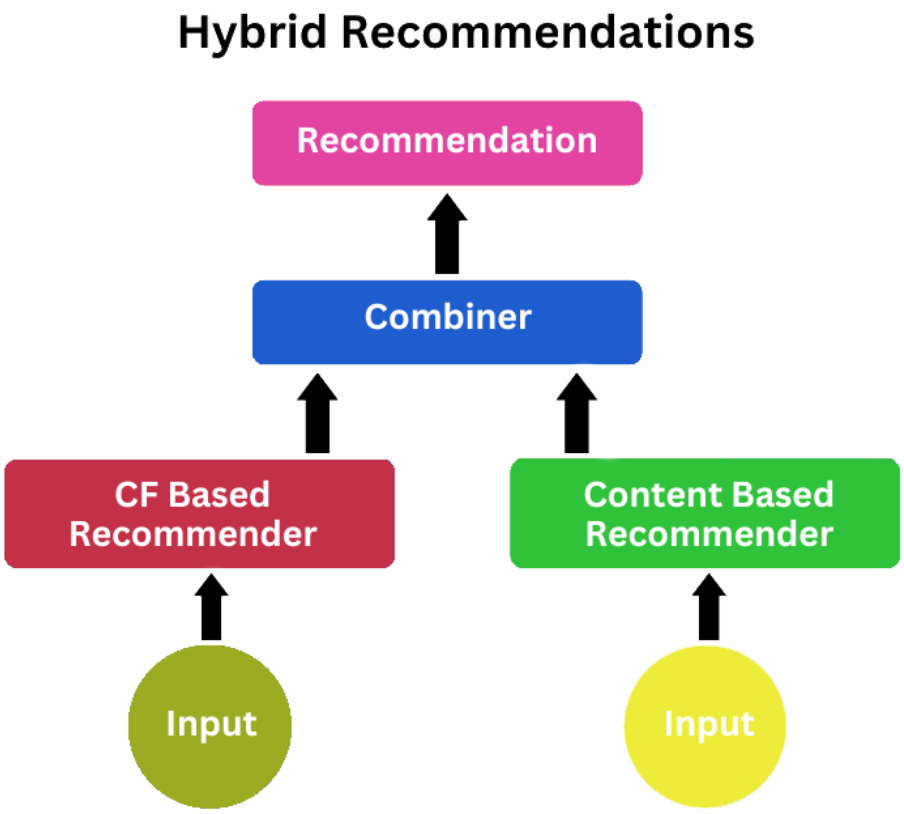


Εικόνα 2. Επισκόπηση του φιλτραρίσματος βάσει περιεχομένου [3]

Για παράδειγμα, εάν ένας χρήστης επιθυμεί να παρακολουθεί ταινίες όπως η ταινία "*The Shawshank Redemption*", το σύστημα συστάσεων προτείνει ταινίες του είδους δράμα/έγκλημα ή ταινίες που βασίζονται σε μυθιστορήματα του Stephen King. Η κεντρική υπόθεση του φιλτραρίσματος βάσει περιεχομένου είναι ότι θα σας αρέσει ένα στοιχείο αν σας αρέσουν και παρόμοια στοιχεία.

1.3.3 Υβριδικά συστήματα συστάσεων (Hybrid systems)

Για να προτείνουν ένα ευρύτερο φάσμα προϊόντων στους πελάτες, τα υβριδικά συστήματα συστάσεων χρησιμοποιούν ταυτόχρονα τόσο το φιλτράρισμα βάσει περιεχομένου όσο και το συνεργατικό φιλτράρισμα. Πρόκειται για ένα νέο σύστημα συστάσεων που λέγεται ότι παρέχει πιο ακριβείς συστάσεις από άλλα συστήματα συστάσεων. [3]
Το **Amazon** είναι ένα εξαιρετικό παράδειγμα υβριδικού συστήματος συστάσεων. Προβαίνει σε συστάσεις αντιπαραβάλλοντας τις αγοραστικές και ερευνητικές συνήθειες των χρηστών και βρίσκοντας παρόμοιους χρήστες στην εν λόγω πλατφόρμα. Με αυτόν τον τρόπο το **Amazon** χρησιμοποιεί το συνεργατικό φιλτράρισμα. [3]



Εικόνα 3. Επισκόπηση του υβριδικού συστήματος συστάσεων [3]

Συνιστώντας τέτοια προϊόντα που μοιράζονται παρόμοια χαρακτηριστικά με εκείνα που έχουν αξιολογηθεί υψηλά από τον χρήστη, το amazon.com χρησιμοποιεί φιλτράρισμα βάσει περιεχομένου. Μπορούν επίσης να ασκήσουν βέτο στα κοινά προβλήματα των συστημάτων συστάσεων, όπως τα προβλήματα ανεπάρκειας δεδομένων. [3]

1.4 Συναισθηματική ανάλυση (Sentiment Analysis)

Ένα σύστημα συστάσεων με βάση το συναίσθημα λαμβάνει υπόψη τις προτιμήσεις των χρηστών μαζί με το συναίσθημα ή τη γνώμη που εκφράζουν οι χρήστες για τα προϊόντα ή τις υπηρεσίες. Χρησιμοποιεί τεχνικές επεξεργασίας φυσικής γλώσσας **(NLP)** για την εξαγωγή χαρακτηριστικών που σχετίζονται με το συναίσθημα από κριτικές, αξιολογήσεις ή σχόλια χρηστών. Αυτή η προσέγγιση βοηθά στην εξατομίκευση των συστάσεων με βάση τις συναισθηματικές αντιδράσεις των χρηστών, οδηγώντας σε μια πιο προσαρμοσμένη και ικανοποιητική εμπειρία χρήστη. Οι σημερινοί δικτυακοί τόποι ηλεκτρονικού εμπορίου μπορούν να βελτιώσουν τις στρατηγικές τους για την ενίσχυση της ικανοποίησης των πελατών τους χρησιμοποιώντας συστάσεις που βασίζονται στο συναίσθημα. Αυτές οι καινοτομίες βοηθούν στη δημιουργία μιας πιο εξατομικευμένης και ικανοποιητικής εμπειρίας χρήστη, προσαρμόζοντας συστάσεις με βάση τις συναισθηματικές αντιδράσεις των χρηστών. Η μέθοδος αυτή λαμβάνει υπόψη τις ψυχολογικές και συναισθηματικές πτυχές της συμπεριφοράς και των επιλογών των χρηστών, υπερβαίνοντας τις απλές αλγοριθμικές συστάσεις. Στο παρακάτω κείμενο, θα αναλύσουμε 2 τρόπους που μπορούμε να επιτύχουμε την συναισθηματική ανάλυση οι οποίοι είναι, τα μοντέλα κανόνων και τα μοντέλα μηχανικής μάθησης. [5]

1.4.1 Μοντέλα κανόνων (Rule-based models)

Η ανάλυση συναισθήματος βάσει κανόνων είναι μια μέθοδος προσδιορισμού του συναισθήματος ή του συναισθηματικού τόνου ενός κειμένου, όπως μια πρόταση, μια παράγραφος ή ένα έγγραφο, με τον καθορισμό ενός συνόλου κανόνων ή προτύπων. Αυτοί οι κανόνες ή τα μοτίβα βασίζονται συνήθως σε συγκεκριμένες λέξεις-κλειδιά, φράσεις ή γλωσσικά χαρακτηριστικά που σχετίζονται με το συναίσθημα και χρησιμοποιούνται για τον προσδιορισμό της πολικότητας του συναισθήματος, όπως θετική, αρνητική ή ουδέτερη, του κειμένου. Η ανάλυση συναισθήματος βάσει κανόνων βασίζεται σε προκαθορισμένους κανόνες και όχι στη χρήση τεχνικών στατιστικής ή μηχανικής μάθησης για την ανάλυση δεδομένων κειμένου. [6]

Οι προσεγγίσεις ανάλυσης συναισθήματος βάσει κανόνων είναι συχνά απλούστερες και περισσότερο ερμηνεύσιμες σε σύγκριση με άλλες μεθόδους που περιλαμβάνουν την εκπαίδευση μοντέλων μηχανικής μάθησης σε μεγάλα σύνολα δεδομένων με ετικέτες. Μπορούν να είναι ιδιαίτερα χρήσιμες όταν υπάρχει περιορισμένος αριθμός δεδομένων με ετικέτες για εκπαίδευση ή όταν η εστίαση είναι σε συγκεκριμένους τομείς ή γλώσσες όπου τα υπάρχοντα προ-εκπαιδευμένα μοντέλα μπορεί να μην είναι τόσο αποτελεσματικά. Οι προσεγγίσεις που βασίζονται σε κανόνες παρέχουν επίσης περισσότερο έλεγχο και διαφάνεια στη διαδικασία ανάλυσης συναισθήματος, καθώς οι κανόνες μπορούν να οριστούν ρητά και να τροποποιηθούν από ειδικούς ή αναλυτές του τομέα με βάση τις γνώσεις ή τις απαιτήσεις του τομέα τους. [6]

Η ανάλυση συναισθήματος βάσει κανόνων μπορεί να υλοποιηθεί με τη χρήση διαφορετικών τεχνικών, όπως κανονικές εκφράσεις, αντιστοίχιση λέξεων-κλειδιών ή αντιστοίχιση προτύπων. Αυτές οι τεχνικές περιλαμβάνουν τον ορισμό ενός συνόλου κανόνων ή μοτίβων με βάση λέξεις-κλειδιά ή άλλα γλωσσικά χαρακτηριστικά που είναι ενδεικτικά του συναισθήματος. Για παράδειγμα, μια απλή προσέγγιση βασισμένη σε κανόνες μπορεί να

περιλαμβάνει την καταμέτρηση των εμφανίσεων θετικών και αρνητικών λέξεων-κλειδιών σε ένα κομμάτι κειμένου και τον προσδιορισμό του συναισθήματος με βάση την καταμέτρηση των λέξεων-κλειδιών. Πιο σύνθετες προσεγγίσεις βασισμένες σε κανόνες μπορούν να περιλαμβάνουν τη χρήση κανονικών εκφράσεων ή άλλων γλωσσικών προτύπων για την καταγραφή του πλαισίου, της άρνησης ή άλλων γλωσσικών αποχρώσεων που επηρεάζουν το συναίσθημα. Στο αναγνωστικό κοινό και τη δέσμευση, κατηγοριοποιώντας άρθρα με βάση τον συναισθηματικό τόνο. Οι συντάκτες μπορούν να χρησιμοποιήσουν αυτές τις γνώσεις για να βελτιώσουν τη στρατηγική περιεχομένου τους, διασφαλίζοντας ότι οι ιστορίες τους έχουν τον επιθυμητό συναισθηματικό αντίκτυπο στους αναγνώστες. [6]

1.4.2 Μοντέλα Μηχανικής Μάθησης (Machine Learning Models)

Η μηχανική μάθηση έχει μεταμορφώσει ριζικά την κατανόηση των ανθρώπινων συναισθημάτων που εκφράζονται σε κείμενο. Σε αντίθεση με τα προηγούμενα συστήματα που βασίζονταν σε προσεγγίσεις με βάση λέξεις-κλειδιά για την ταξινόμηση των συναισθημάτων ως θετικά, αρνητικά ή ουδέτερα, τα μοντέλα μηχανικής μάθησης μαθαίνουν απευθείας από τα δεδομένα. Είναι ικανά να εντοπίζουν αποχρώσεις της γλώσσας, όπως ο σαρκασμός ή η ειρωνεία, οι οποίες είναι δύσκολο να εντοπιστούν με τη χρήση απλών κανόνων. Με την εκπαίδευση αυτών των μοντέλων σε ένα τεράστιο φάσμα παραδειγμάτων, η μηχανική μάθηση υπερέχει στην πρόβλεψη συναισθημάτων σε διάφορα πλαίσια, συμπεριλαμβανομένων των αναρτήσεων στα μέσα κοινωνικής δικτύωσης, των κριτικών και των αναφορών πελατών. Αυτή η ικανότητα την καθιστά εξαιρετικά πολύτιμη τόσο για τις επιχειρήσεις όσο και για τους ερευνητές.

Ένα σημαντικό πλεονέκτημα της χρήσης της μηχανικής μάθησης για την ανάλυση συναισθήματος έγκειται στην ικανότητά της να διαχειρίζεται αποτελεσματικά τεράστιες ποσότητες δεδομένων. Καθώς εμφανίζονται συνεχώς νέα κείμενα στο διαδίκτυο, τα μοντέλα αυτά βελτιώνονται σταδιακά όσον αφορά την κατανόηση των γλωσσικών προτύπων και των συναισθημάτων. Αυτή η ικανότητα εξέλιξης τους επιτρέπει να αναλύουν δεδομένα σε πραγματικό χρόνο, όπως ροές **Twitter** ή κριτικές προϊόντων. Κατά συνέπεια, παρέχουν γρήγορες γνώσεις σχετικά με τις απόψεις του κοινού και τα επίπεδα ικανοποίησης των πελατών. Επιπλέον, δεδομένου ότι αυτά τα μοντέλα μπορούν να μαθαίνουν από νέα δεδομένα με την πάροδο του χρόνου, παραμένουν επίκαιρα με τις εξελισσόμενες τάσεις και τους νέους τρόπους με τους οποίους εκφράζονται τα άτομα.

Επιπλέον, η ανάλυση συναισθήματος με χρήση μηχανικής μάθησης διαδραματίζει κρίσιμο ρόλο στην ενίσχυση των συστημάτων συστάσεων. Εξετάζοντας τις αξιολογήσεις και τα σχόλια των χρηστών, τα μοντέλα αυτά μπορούν να διακρίνουν τα συναισθήματα των ατόμων απέναντι σε διάφορα προϊόντα ή περιεχόμενο. Για παράδειγμα, εάν ένας χρήστης εκφράζει συχνά θετικά συναισθήματα για συγκεκριμένα είδη ταινιών ή τύπους προϊόντων, ένα σύστημα συστάσεων μπορεί να προτείνει παρόμοια στοιχεία. Αυτή η προσέγγιση παρέχει στους χρήστες μια πιο εξατομικευμένη εμπειρία. Αυτό όχι μόνο κρατά τους χρήστες δεσμευμένους με σχετικές συστάσεις, αλλά δίνει επίσης τη δυνατότητα στις επιχειρήσεις να βελτιώσουν τις στρατηγικές μάρκετινγκ τους. Τελικά, αυτό έχει ως αποτέλεσμα πιο ικανοποιημένους και πιστούς πελάτες.

1.5 Σχετικά Έργα

Με βάση τη συγκριτική μελέτη του **A.Y.Zhubatkhan** (2021) σχετικά με διάφορους αλγορίθμους μηχανικής μάθησης για συστήματα σύστασης ταινιών, η έρευνα αξιολόγησε την απόδοση αυτών των αλγορίθμων χρησιμοποιώντας το σύνολο δεδομένων **MovieLens 100k**, εστιάζοντας σε δύο βασικές μετρήσεις απόδοσης: το μέσο απόλυτο σφάλμα **(MAE)** και το μέσο τετραγωνικό σφάλμα **(RMSE)**. Η μελέτη του **Zhubatkhan** εξέτασε εννέα διαφορετικούς αλγορίθμους για να καθορίσει την αποτελεσματικότητά τους στην ακριβή πρόβλεψη των προτιμήσεων των χρηστών. Τα αποτελέσματα έδειξαν ότι η αποσύνθεση μοναδικών τιμών **(SVD)** και οι K-κοντινότεροι γείτονες με μέσα **(KNNwM)** υπερείχαν όσον αφορά το **MAE**, αποδεικνύοντας την ικανότητά τους να μειώνουν τα απόλυτα σφάλματα πρόβλεψης. Αντίθετα, όταν αξιολογήθηκαν με τη χρήση RMSE, οι αλγόριθμοι **BaselineOnly** και **SlopeOne** υπερείχαν έναντι των άλλων, υποδεικνύοντας την ευρωστία τους στην αντιμετώπιση τετραγωνικών σφαλμάτων πρόβλεψης. Αυτή η συγκριτική ανάλυση υπογραμμίζει τη σημασία της επιλογής των κατάλληλων αλγορίθμων με βάση τις συγκεκριμένες επιδόσεις. [7]

Μία άλλη μελέτη που διεξήχθη από τον **Raja Marappan** το 2022, αναπτύχθηκε ένα υβριδικό σύστημα σύστασης ταινιών που χρησιμοποιεί τόσο την ομοιότητα συνημίτονου **(cosine similarity)** όσο και την ανάλυση συναισθήματος για τη βελτίωση της ακρίβειας των συστάσεων. Ο **Marappan** δημιούργησε το σύστημα για να ανταποκρίνεται στα ενδιαφέροντα των χρηστών αναλύοντας κριτικές ταινιών, προσδιορίζοντας αυτόματα το συναίσθημα που εκφράζεται σε αυτές και ενσωματώνοντας το συναίσθημα αυτό στη συνολική βαθμολογία της ταινίας. Συνδυάζοντας την ομοιότητα συνημίτονου -η οποία μετρά την ομοιότητα μεταξύ των προτιμήσεων των χρηστών και των χαρακτηριστικών των ταινιών- με την ανάλυση συναισθήματος, το σύστημα παρέχει εξατομικευμένες συστάσεις ταινιών που λαμβάνουν υπόψη όχι μόνο τις προτιμήσεις των χρηστών αλλά και τις ποιοτικές απόψεις άλλων θεατών. Αυτή η προσέγγιση επιτρέπει στη μηχανή συστάσεων να προσαρμόζει τις προτάσεις της με βάση τον συναισθηματικό τόνο των κριτικών, με αποτέλεσμα μια πιο διαφοροποιημένη και επικεντρωμένη στον χρήστη εμπειρία σύστασης ταινιών. [8]

Το 2017, οι **C. Golian** και **Jaroslav Kuchař** διεξήγαγαν μια μελέτη σχετικά με τη σύσταση ειδησεογραφικών άρθρων με τη χρήση ενός ταξινομητή βασισμένου σε κανόνες στο πλαίσιο της πρόκλησης **CLEF NewsREEL 2017**. Η έρευνά τους επικεντρώθηκε στη βελτιστοποίηση του συστήματος βασισμένου σε κανόνες για τη μείωση του αριθμού των απαιτούμενων κανόνων με παράλληλη βελτίωση της συνολικής απόδοσης, η οποία είναι κρίσιμη για συστάσεις σε πραγματικό χρόνο στο σημερινό ταχέως εξελισσόμενο ειδησεογραφικό τοπίο. Η μελέτη εστιάζει στις μοναδικές προκλήσεις που αντιμετωπίζουν τα συστήματα σύστασης ειδησεογραφικών άρθρων, όπως η συχνή δημιουργία νέου περιεχομένου, γεγονός που τα διαφοροποιεί από τα συστήματα που έχουν σχεδιαστεί για άλλους τύπους περιεχομένου. Οι **Golian** και **Kuchař** βελτίωσαν τον μοντέλο τους που βασίζεται σε κανόνες για να βελτιώσουν την αποδοτικότητα και την αποτελεσματικότητα, παρέχοντας μια λύση στην υπερφόρτωση πληροφοριών και αποδεικνύοντας την πρακτική εφαρμογή του για την παροχή σχετικών συστάσεων για ειδήσεις. [9]

Επιπλέον, η εργασία του **Dhruv Khandelwal (2018)** παρουσιάζει το **«LeMeNo»**, ένα εξατομικευμένο σύστημα σύστασης ειδήσεων που χρησιμοποιεί τεχνικές μηχανικής μάθησης για να προσαρμόζει τις προτάσεις ειδησεογραφικών άρθρων στους χρήστες με βάση τα

ενδιαφέροντά τους και το περιεχόμενο των άρθρων. Το **LeMeNo** κάνει εξατομικευμένες συστάσεις με βάση τη συμπεριφορά και τις προτιμήσεις των χρηστών, μαθαίνοντας από τις αλληλεπιδράσεις τους για να βελτιώσει τη συνάφεια των ειδησεογραφικών άρθρων που συνιστώνται. Με τη συνεχή παρακολούθηση και προσαρμογή στα ενδιαφέροντα των χρηστών, το σύστημα διασφαλίζει ότι οι χρήστες λαμβάνουν ειδησεογραφικό περιεχόμενο προσαρμοσμένο στις συγκεκριμένες προτιμήσεις τους, αντιμετωπίζοντας αποτελεσματικά την πρόκληση της παροχής σχετικών και ελκυστικών ειδήσεων σε ένα δυναμικό τοπίο πληροφοριών. [10]

ΜΕΘΟΔΟΛΟΓΙΑ

2.1 Ορισμός του Προβλήματος

Το πρόβλημα είναι η βελτίωση της αποτελεσματικότητας των συστημάτων σύστασης ειδήσεων με το συνδυασμό μοντέλων μηχανικής μάθησης (ML), συστημάτων βασισμένων σε κανόνες και ανάλυσης συναισθήματος. Οι παραδοσιακές προσεγγίσεις σύστασης ειδήσεων, όπως το φιλτράρισμα βάσει περιεχομένου και το συνεργατικό φιλτράρισμα, ενώ είναι ικανές να αξιοποιούν τη συμπεριφορά και τις προτιμήσεις των χρηστών, συχνά αποτυγχάνουν να συλλάβουν το πλήρες φάσμα των ενδιαφερόντων των χρηστών, ιδίως σε ταχέως μεταβαλλόμενα ειδησεογραφικά περιβάλλοντα. Αυτός ο περιορισμός μπορεί να οδηγήσει σε λιγότερο σχετικές ή υπερβολικά γενικές συστάσεις ειδήσεων, γεγονός που μειώνει τη δέσμευση και την ικανοποίηση των χρηστών. Με την ενσωμάτωση της ανάλυσης συναισθήματος, τα συστήματα αυτά μπορούν να κατανοήσουν καλύτερα τον συναισθηματικό τόνο τόσο του ειδησεογραφικού περιεχομένου όσο και των αντιδράσεων των χρηστών, επιτρέποντάς τους να βελτιώσουν τις συστάσεις ώστε να ανταποκρίνονται καλύτερα στα συναισθήματα των χρηστών.

Δίνοντας στους χρήστες τη δυνατότητα να φιλτράρουν και να επιλέγουν ειδησεογραφικά άρθρα σύμφωνα με τις κατηγορίες και τα συναισθήματα που προτιμούν, η προτεινόμενη λύση επιδιώκει να καλύψει αυτό το κενό. Οι χρήστες μπορούν να επιλέγουν ενεργά αν θα βλέπουν άρθρα με θετικά ή αρνητικά συναισθήματα εντός συγκεκριμένων κατηγοριών, ενσωματώνοντας άμεσα την ανάλυση συναισθημάτων στη διαδικασία συστάσεων. Ένας χρήστης που ενδιαφέρεται για την τεχνολογία, για παράδειγμα, μπορεί να θέλει να διαβάζει άρθρα σχετικά με τις νέες εξελίξεις στον τομέα αυτό, ενώ ένας άλλος χρήστης μπορεί να αναζητά αντίθετες απόψεις για πολιτικά θέματα. Το σύστημα συστάσεων γίνεται πιο φιλικό προς τον χρήστη, επιτρέποντας τους να επιλέξουν το συναίσθημα και την κατηγορία που προτιμούν, παρέχοντας εξατομικευμένο ειδησεογραφικό περιεχόμενο που ανταποκρίνεται στις τρέχουσες πληροφοριακές και συναισθηματικές ανάγκες τους.

2.2 Ανασκόπηση

Όπως είδαμε και στο προηγούμενο κεφάλαιο, όλες οι προαναφερθείσες μελέτες χρησιμοποιούν τεχνικές μηχανικής μάθησης καθώς και μοντέλα βασισμένα σε κανόνες για τη διατύπωση προτάσεων. Ωστόσο, καμία από αυτές δεν επικεντρώνεται στη συναισθηματική

ανάλυση για τη δημιουργία προτάσεων σε ειδησεογραφικά θέματα. Αυτός είναι και ο κύριος στόχος της παρούσας εργασίας, να συμβάλει και να βελτιώσει τον τρόπο με τον οποίο γίνονται οι συστάσεις στις ειδήσεις.

2.3 Ανάλυση

Σκοπός της διπλωματικής αυτής, είναι η διεξοδική αξιολόγηση ενός μοντέλου μηχανικής μάθησης και ενός μοντέλου βασισμένο σε κανόνες με την χρήση ανάλυσης συναισθήματος , εστιάζοντας σε τίτλους ειδησεογραφικών άρθρων. Ο πρωταρχικός στόχος ήταν να δω πόσο καλά κάθε μοντέλο ταξινόμησε το συναίσθημα ενός τίτλου ως θετικό, αρνητικό ή ουδέτερο. Μετά τον προσδιορισμό του συναισθήματος, το σύστημα έκανε συστάσεις για ειδησεογραφικά άρθρα που ταίριαζαν τόσο με το συναίσθημα όσο και με την κατηγορία του τίτλου εισόδου. Οι επιδόσεις του μοντέλου μηχανικής μάθησης και της προσέγγισης που βασίζεται σε κανόνες συγκρίθηκαν με τη χρήση βασικών μετρικών όπως η ακρίβεια, η ανάκληση και το F1-score για να προσδιοριστεί ποιο από τα δύο παρείχε πιο ακριβείς και αξιόπιστες ταξινομήσεις συναισθήματος. Η ανάλυση αυτή όχι μόνο ανέδειξε τα συγκριτικά πλεονεκτήματα και τις αδυναμίες των δύο μεθόδων, αλλά προσέφερε επίσης χρήσιμες πληροφορίες για τη βελτίωση των συστημάτων σύστασης ειδήσεων.

2.4 Σχεδίαση

2.4.1 Ανάγκες χρήσης

Οι ανάγκες χρήσης αυτού του συστήματος είναι να παρέχει στους χρήστες σχετικές συστάσεις ειδησεογραφικών άρθρων με βάση το συναίσθημα του τίτλου που εισάγουν. Αντί για χειροκίνητη επιλογή προτιμήσεων συναισθήματος, το σύστημα αναλύει το συναίσθημα (θετικό, αρνητικό ή ουδέτερο) και το συγκρίνει με άλλα άρθρα της ίδιας κατηγορίας. Αυτό ανταποκρίνεται στην ανάγκη για εξατομικευμένη κατανάλωση περιεχομένου, επιτρέποντας στους χρήστες να βρίσκουν ειδήσεις που ταιριάζουν με το ύφος και το θέμα του αρχικού τους ενδιαφέροντος. Το σύστημα απλοποιεί την εμπειρία του χρήστη παρέχοντας απρόσκοπτες συστάσεις με βάση το συναίσθημα, χωρίς να απαιτείται πρόσθετη προσπάθεια από τον χρήστη, επιτρέποντάς του να έχει γρήγορη πρόσβαση σε ειδησεογραφικά άρθρα που σχετίζονται με το τρέχον συναισθηματικό πλαίσιο ή τα ενδιαφέροντά του.

2.4.2 Αρχιτεκτονική

Το σύστημα έχει μια απλή αρχιτεκτονική ενός επιπέδου, με τα πάντα να εκτελούνται στο ίδιο περιβάλλον, συγκεκριμένα στο **Google Colab**. Αυτό περιλαμβάνει την προεπεξεργασία δεδομένων, την ανάλυση συναισθήματος (είτε με μοντέλο μηχανικής μάθησης είτε με μοντέλο βασισμένο σε κανόνες) και τη μηχανή συστάσεων. Όλα τα στοιχεία λειτουργούν σε μία τοποθεσία, εξαλείφοντας την ανάγκη για εξωτερικούς διακομιστές ή βάσεις δεδομένων, καθιστώντας το σύστημα απλό και αυτάρκες.

2.4.3 Βασικές μονάδες συστήματος

Οι βασικές μονάδες του συστήματος, τόσο για τη μηχανική μάθηση όσο και για τα μοντέλα που βασίζονται σε κανόνες, είναι η προεπεξεργασία δεδομένων, η ταξινόμηση συναισθημάτων και η μηχανή συστάσεων.

Η μονάδα προεπεξεργασίας δεδομένων στο σύστημα που βασίζεται στην Μηχανική Μάθηση χειρίζεται τη τμηματοποίηση **(tokenization)**, την αφαίρεση των σταθερών λέξεων και την εξαγωγή χαρακτηριστικών με την μεθοδο **TF-IDF** για να προετοιμάσει τα δεδομένα εισόδου για την ταξινόμηση συναισθήματος. Το μοντέλο ανάλυσης συναισθήματος είναι ένας αλγόριθμος μηχανικής μάθησης που κατηγοριοποιεί τους τίτλους των ειδησεογραφικών άρθρων ως θετικούς, αρνητικούς ή ουδέτερους. Τέλος, η μηχανή συστάσεων επιλέγει και παραδίδει άρθρα που αντιστοιχούν στο ταξινομημένο συναίσθημα και στην κατηγορία του τίτλου εισόδου.

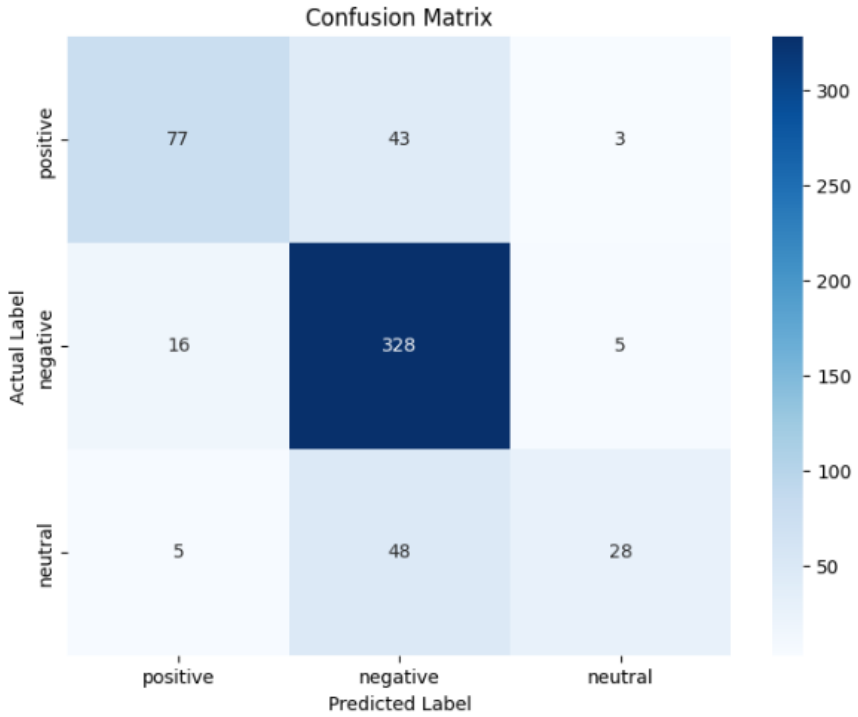
Παρομοίως, στο σύστημα που βασίζεται σε κανόνες, η ενότητα προεπεξεργασίας δεδομένων είναι υπεύθυνη για τον καθαρισμό και την προετοιμασία του τίτλου εισόδου, η οποία περιλαμβάνει την κανονικοποίηση των συμβόλων και την κανονικοποίηση του κειμένου. Ο ταξινομητής συναισθήματος βάσει κανόνων χρησιμοποιεί προκαθορισμένους γλωσσικούς κανόνες και λεξικά συναισθήματος για τον υπολογισμό του συναισθήματος. Ο μηχανισμός συστάσεων λειτουργεί με παρόμοιο τρόπο, συνιστώντας άρθρα με βάση το αναγνωρισμένο συναίσθημα και την κατηγορία, διασφαλίζοντας όσο το περισσότερο δυνατόν ότι οι συστάσεις παράγονται με συνέπεια και στα δύο συστήματα.

2.5 Υλοποίηση

Το έργο αυτό θα υλοποιηθεί σε δύο ξεχωριστά συστήματα: το ένα με ένα μοντέλο μηχανικής μάθησης την λογιστική παλινδρόμηση και το άλλο με ένα μοντέλο βασισμένο σε κανόνες τον αλγόριθμο **Vader** για ανάλυση συναισθήματος. Η προεπεξεργασία των δεδομένων και στα δύο συστήματα αποτελείται από **tokenization**, αφαίρεση **stopwords** και εξαγωγή χαρακτηριστικών με χρήση **TF-IDF**. Για την ταξινόμηση των τίτλων των ειδησεογραφικών άρθρων ως θετικών, αρνητικών ή ουδέτερων, το σύστημα μηχανικής μάθησης χρησιμοποιεί λογιστική παλινδρόμηση που εκπαιδεύεται σε δεδομένα με ετικέτες. Στο σύστημα που βασίζεται σε κανόνες, ο αλγόριθμος Vader χρησιμοποιείται για τον προσδιορισμό του συναισθήματος χρησιμοποιώντας προκαθορισμένους γλωσσικούς κανόνες και λεξικά συναισθήματος. Μετά την ταξινόμηση του συναισθήματος, μια μηχανή συστάσεων προτείνει άρθρα που αντιστοιχούν στο συναίσθημα και την κατηγορία του τίτλου εισόδου. Και τα δύο συστήματα υλοποιούνται ανεξάρτητα, επιτρέποντας τη σύγκριση των επιδόσεων.

2.6 Testing

2.6.1 Απόδοση μοντέλων



Εικόνα 4.πίνακας σύγχυσης μοντέλου μηχανικής μαθήσεως

	precision	recall	f1-score	support
positive	0.79	0.63	0.70	123
negative	0.78	0.94	0.85	349
neutral	0.78	0.35	0.48	81
accuracy			0.78	553
macro avg	0.78	0.64	0.68	553
weighted avg	0.78	0.78	0.76	553

Εικόνα 5. Classification report

	precision	recall	f1-score	support
Negative	0.33	0.19	0.24	407
Neutral	0.31	0.66	0.42	516
Positive	0.79	0.57	0.66	1528
accuracy			0.53	2451
macro avg	0.47	0.47	0.44	2451
weighted avg	0.61	0.53	0.54	2451

Εικόνα 6. Classification report

2.6.2 Αποτελέσματα

Τα αποτελέσματα της ανάλυσης συναισθήματος δείχνουν ότι τα δύο μοντέλα έχουν σημαντικά διαφορετικές επιδόσεις. Το μοντέλο μηχανικής μάθησης, το οποίο χρησιμοποίησε λογιστική παλινδρόμηση **(logistic regression)**, πέτυχε ακρίβεια 78% κατά την ταξινόμηση των τίτλων των ειδησεογραφικών άρθρων ως θετικών, αρνητικών ή ουδέτερων. Αντίθετα, το μοντέλο που βασίζεται σε κανόνες και χρησιμοποίησε τον αλγόριθμο **Vader** πέτυχε ακρίβεια μόλις 53%. Το κύριο πρόβλημα με το μοντέλο που βασίστηκε στο **Vader** ήταν η αδυναμία του να διακρίνει μεταξύ θετικών και ουδέτερων συναισθημάτων, γεγονός που οδήγησε σε χαμηλότερη συνολική ακρίβεια. Αυτή η σύγκριση καταδεικνύει την ικανότητα του μοντέλου μηχανικής μάθησης να ανιχνεύει αποχρώσεις συναισθήματος στους τίτλους ειδησεογραφικών άρθρων.

2.6.3 Πεδία χρησιμότητας του συστήματος

Αυτό το σύστημα σύστασης ειδήσεων έχει ευρύ φάσμα εφαρμογών στον κλάδο των ειδήσεων και των μέσων ενημέρωσης. Η ικανότητά του να ταξινομεί και να συστήνει άρθρα με βάση το θετικό, αρνητικό ή ουδέτερο συναίσθημα, σε συνδυασμό με συγκεκριμένες κατηγορίες, το καθιστά ιδιαίτερα χρήσιμο σε ποικίλα πλαίσια.

Πλατφόρμες όπως το **Google News,** το **Flipboard** και το **Apple News** μπορούν να χρησιμοποιήσουν αυτό το σύστημα για να παρέχουν μια πιο εξατομικευμένη εμπειρία ειδήσεων. Οι χρήστες μπορούν να λαμβάνουν συστάσεις με βάση τον συναισθηματικό τόνο του περιεχομένου με το οποίο ασχολούνται συχνά. Για παράδειγμα, εάν ένας χρήστης διαβάζει άρθρα με θετικό συναίσθημα στην πολιτική, μπορεί να του συνιστώνται παρόμοια θετικά περιεχόμενα στην ίδια κατηγορία, επιτρέποντάς του να εξερευνήσει ειδησεογραφικά άρθρα που είναι συναισθηματικά κατάλληλα.

Τα μέσα ενημέρωσης μπορούν να χρησιμοποιήσουν αυτό το σύστημα για να παρέχουν προσαρμοσμένες συστάσεις περιεχομένου στους αναγνώστες τους. Η πλατφόρμα θα προσφέρει μια πιο διαδραστική και εξατομικευμένη εμπειρία ανάγνωσης, επιτρέποντας στους χρήστες να αναζητούν ειδησεογραφικά άρθρα με βάση τον συναισθηματικό τους τόνο. Αυτό θα μπορούσε να ενισχύσει τη διατήρηση των χρηστών με την προσαρμογή στις συναισθηματικές τους προτιμήσεις, όπως ουδέτερο, επικριτικό (αρνητικό) ή θετικό περιεχόμενο για συγκεκριμένα θέματα όπως η υγεία, οι επιχειρήσεις ή ο αθλητισμός.

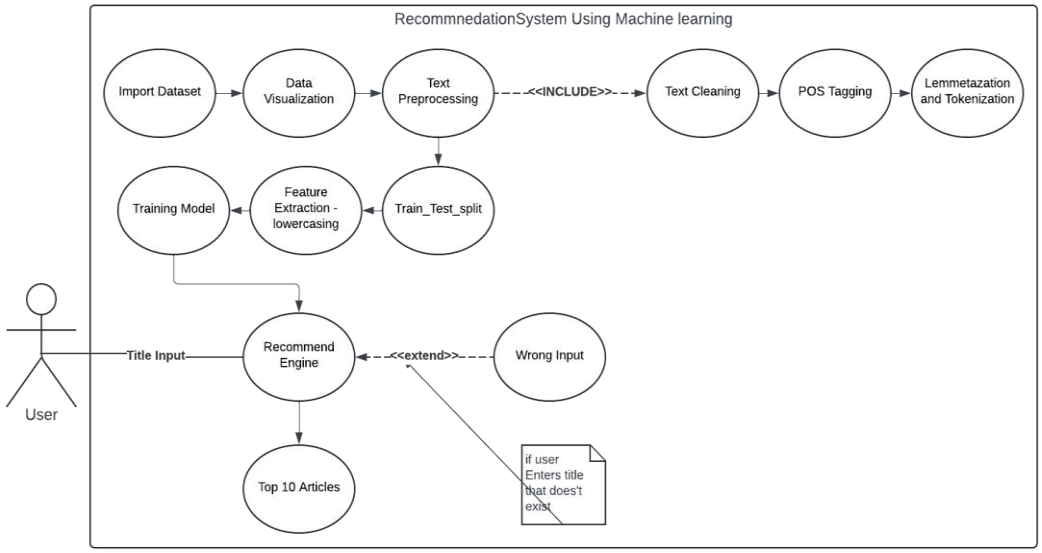
Το σύστημα αυτό μπορεί να χρησιμοποιηθεί από ειδησεογραφικούς οργανισμούς που επιμελούνται περιεχόμενο για ενημερωτικά δελτία ή ιστορίες για την αρχική σελίδα για να προτείνουν άρθρα με βάση τις τάσεις των συναισθημάτων των αναγνωστών τους. Εάν ένας χρήστης ασχολείται κυρίως με ουδέτερες ή θετικές ιστορίες, οι αλγόριθμοι επιμέλειας μπορούν να δώσουν προτεραιότητα σε παρόμοια άρθρα, με αποτέλεσμα μια πιο εξατομικευμένη επιλογή περιεχομένου που ανταποκρίνεται στις συναισθηματικές προτιμήσεις κάθε αναγνώστη.

Οι δημοσιογράφοι και οι συντάκτες μπορούν να χρησιμοποιήσουν αυτό το σύστημα για να εντοπίσουν τάσεις στην ειδησεογραφία. Μπορούν να αξιολογήσουν τον αντίκτυπο του συναισθήματος στο αναγνωστικό κοινό και τη δέσμευση, κατηγοριοποιώντας τα άρθρα με

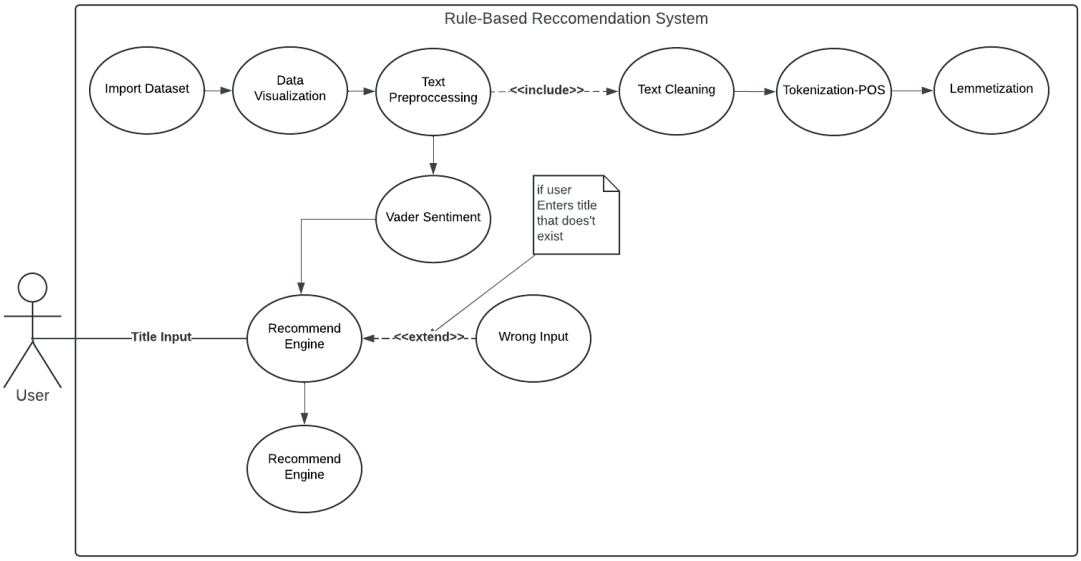
βάση τον συναισθηματικό τόνο. Οι συντάκτες μπορούν να χρησιμοποιήσουν αυτές τις γνώσεις για να βελτιώσουν τη στρατηγική περιεχομένου τους, διασφαλίζοντας ότι οι ιστορίες τους έχουν τον επιθυμητό συναισθηματικό αντίκτυπο στους αναγνώστες.

Συνοψίζοντας, το σύστημα συστάσεων με βάση το συναίσθημα είναι ιδανικό για την ενίσχυση της εξατομίκευσης του περιεχομένου, της δέσμευσης και της λήψης στρατηγικών αποφάσεων σε διάφορα πλαίσια ειδήσεων και μέσων ενημέρωσης. Δίνοντας έμφαση στον τρόπο με τον οποίο οι ειδήσεις παρουσιάζονται συναισθηματικά και όχι στο θέμα τους, το σύστημα παρέχει πολύτιμες πληροφορίες και βελτιώνει τις εμπειρίες των χρηστών σε όλες τις πλατφόρμες.

2.6.4 UML διάγραμμα (use case)



Εικόνα 7.Uml Διάγραμμα του συστήματος συστάσεως με την Χρήση Μηχανικής Μάθησης



Εικόνα 8. Uml Διάγραμμα του συστήματος συστάσεως με την Μοντέλου Κανόνων

ΣΥΣΤΗΜΑ

3.1 Γλώσσα Προγραμματισμού

Για την υλοποίηση των συστημάτων χρησιμοποιούμε τη γλώσσα προγραμματισμού **Python**. Η **Python** είναι μια γλώσσα προγραμματισμού υπολογιστών που χρησιμοποιείται συνήθως για τη δημιουργία δικτυακών τόπων και λογισμικού, την αυτοματοποίηση εργασιών και την ανάλυση δεδομένων. Η **Python** είναι μια γλώσσα προγραμματισμού γενικού σκοπού, που σημαίνει ότι δεν περιορίζεται στην επίλυση συγκεκριμένων προβλημάτων αλλά μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα εργασιών. [11]

Θεωρείται ευρέως ως μία από τις ευκολότερες γλώσσες προγραμματισμού για να γράψει, να διαβάσει και να μάθει κανείς, επειδή το συντακτικό της είναι παρόμοιο με αυτό της αγγλικής γλώσσας. Πολλοί επαγγελματίες με ελάχιστη έως καθόλου εμπειρία στον προγραμματισμό υπολογιστών μπορούν να μάθουν βασικές δεξιότητες **Python** και να προχωρήσουν σε πιο προηγμένες τεχνικές κωδικοποίησης. [11]

Ένας άλλος λόγος που η **Python** είναι τόσο δημοφιλής είναι ότι είναι ανοικτού κώδικα, πράγμα που σημαίνει ότι είναι δωρεάν, εύκολα προσβάσιμη και τροποποιήσιμη από οποιονδήποτε. Υπάρχει μια μεγάλη βιβλιοθήκη δημόσια διαθέσιμου κώδικα **Python**, γεγονός που διευκολύνει πολύ τους προγραμματιστές να δημιουργήσουν και να κλιμακώσουν τα δικά τους έργα. Αυτό σημαίνει ότι αντί να ξεκινάτε από το μηδέν με κάθε νέο έργο, μπορείτε να ανατρέξετε και να τροποποιήσετε υπάρχοντα κώδικα από προηγούμενα δημόσια έργα **Python**. [11]

Η **Python** έχει εξελιχθεί σε βασικό πυλώνα της επιστήμης των δεδομένων, επιτρέποντας σε επιστήμονες, αναλυτές δεδομένων και άλλους ειδικούς να εκτελούν χωρίς κόπο ένα ευρύ φάσμα εργασιών που σχετίζονται με δεδομένα, όπως περίπλοκους στατιστικούς υπολογισμούς, σχολαστικές απεικονίσεις δεδομένων, προηγμένα μοντέλα μηχανικής μάθησης, χειρισμό μεγάλων συνόλων δεδομένων και πολλά άλλα. Τόσο οι αρχάριοι όσο και οι ειδικοί το προτιμούν λόγω της αναγνωσιμότητας και της απλότητάς του. Επιπλέον, η ευελιξία της **Python** την καθιστά χρήσιμη για ένα ευρύ φάσμα εργασιών, συμπεριλαμβανομένης της ανάπτυξης μοντέλων μηχανικής μάθησης σε περιβάλλοντα παραγωγής και της διερευνητικής ανάλυσης δεδομένων. [11]

Πολλοί διαφορετικοί τύποι οπτικοποιήσεων δεδομένων, όπως διαγράμματα διασποράς **(scatter plots)**, διαγράμματα πίτας **(pie charts)**, γραμμικά και ραβδογράμματα **(line and bar graphs)**, ιστογράμματα **(histograms)** ακόμη και περίπλοκες τρισδιάστατες οπτικοποιήσεις, μπορούν να κατασκευαστούν με την **Python.** Συνεπώς, αποτελεί ένα αποτελεσματικό μέσο για τη μεταφορά γνώσεων. Τα μοτίβα και οι τάσεις των δεδομένων μπορούν να βρεθούν στα δεδομένα με τη χρήση βιβλιοθηκών όπως οι **Matplotlib**, **Seaborn** και **Plotly**, οι οποίες επιτρέπουν την κατασκευή τόσο βασικών όσο και διαδραστικών αναπαραστάσεων. [11]

Στον τομέα της επιστήμης των δεδομένων, το τεράστιο οικοσύστημα βιβλιοθηκών και πλαισίων της **Python** είναι ένα από τα μεγαλύτερα πλεονεκτήματά της. Ενώ ορισμένες βιβλιοθήκες, όπως οι **Pandas** και **NumPy,** προσφέρουν αποτελεσματικά εργαλεία για τον χειρισμό και την ανάλυση δεδομένων, άλλες, όπως οι **TensorFlow, Keras** και **PyTorch,** διευκολύνουν την ανάπτυξη μοντέλων μηχανικής μάθησης και βαθιάς μάθησης. Αυτές οι ενότητες διασφαλίζουν ότι η Python μπορεί να διαχειριστεί αποτελεσματικά και με λίγη

προσπάθεια έργα δεδομένων μεγάλης κλίμακας, ενώ παράλληλα επιταχύνουν τη διαδικασία ανάπτυξης.

3.2 Περιβάλλον Ανάπτυξης

Το **Google Colaboratory** (επίσης γνωστό ως **Colab**) στοχεύει στην προώθηση της εκπαίδευσης και της έρευνας στον τομέα της μηχανικής μάθησης. Τα σημειωματάρια του **Colaboratory** χρησιμοποιούν το **Jupyter** και λειτουργούν ως αντικείμενο του **Google Docs**, επιτρέποντας στους χρήστες να μοιράζονται και να συνεργάζονται στο ίδιο σημειωματάριο. Το **Colaboratory** προσφέρει χρόνους εκτέλεσης **Python 2** και **Python 3** προ-ρυθμισμένους με βασικές βιβλιοθήκες μηχανικής μάθησης και τεχνητής νοημοσύνης, όπως οι **TensorFlow**, **Matplotlib** και **Keras**. Η εικονική μηχανή **(VM)** απενεργοποιείται μετά από ένα ορισμένο χρονικό διάστημα και όλα τα δεδομένα και οι διαμορφώσεις του χρήστη χάνονται. Ωστόσο, το σημειωματάριο διατηρείται και τα αρχεία μπορούν να μεταφερθούν από το σκληρό δίσκο της VM στο λογαριασμό **Google Drive** του χρήστη. Τέλος, αυτή η υπηρεσία της **Google** προσφέρει ένα περιβάλλον εκτέλεσης με επιτάχυνση κάρτας γραφικών **(GPU-Accelerated)**, το οποίο είναι πλήρως διαμορφωμένο με το λογισμικό που αναφέρθηκε προηγουμένως. Η υποδομή του **Google Colaboratory** φιλοξενείται στην πλατφόρμα **Google Cloud.** [24]

3.2.1 Διαφορές Jupiter Notebook και Google Collab

Το Google Colab και το Jupyter Notebook είναι και τα δύο δημοφιλή εργαλεία εκτέλεσης κώδικα Python, ιδίως για έργα επιστήμης δεδομένων και μηχανικής μάθησης, αλλά διαφέρουν σημαντικά. Το Google Colab είναι μια πλατφόρμα βασισμένη στο cloud που επιτρέπει στους χρήστες να γράφουν και να εκτελούν κώδικα Python απευθείας στο πρόγραμμα περιήγησης ιστού, εξαλείφοντας την ανάγκη τοπικής εγκατάστασης οποιουδήποτε λογισμικού. Προσφέρει δωρεάν πρόσβαση σε **GPUs** και **TPUs,** καθιστώντας την ιδανική για την εκτέλεση εργασιών έντασης πόρων, όπως μοντέλα βαθιάς μάθησης. Επειδή βασίζεται στο cloud, η κάθε εργασία αποθηκεύεται στο διαδίκτυο και μπορεί εύκολα να μοιραστεί και να συνεργαστεί με άλλους στέλνοντας έναν σύνδεσμο. [25]

Από την άλλη πλευρά, το **Jupyter Notebook** είναι ένα τοπικό εργαλείο που πρέπει να εγκατασταθεί σε υπολογιστή. Αποτελεί μέρος του οικοσυστήματος Jupyter, το οποίο περιλαμβάνει άλλα εργαλεία εκτέλεσης κώδικα, αλλά δεν προσφέρει δωρεάν πρόσβαση σε GPU/TPU από το κουτί. Το **Jupyter** είναι ιδιαίτερα παραμετροποιήσιμο και προσαρμόσιμο, επιτρέποντάς σας να εκτελείτε κώδικα, να προβάλλετε τα αποτελέσματα και να γράφετε κείμενο, όλα σε ένα μέρος. Ωστόσο, επειδή είναι τοπικό, η υπολογιστική ισχύς περιορίζεται στο μηχάνημά σας. Ενώ το **Colab** είναι ιδανικό για χρήστες που θέλουν ευκολία, δωρεάν πόρους cloud και εύκολη συνεργασία, το Jupyter Notebook παρέχει περισσότερο έλεγχο του περιβάλλοντός σας και είναι καταλληλότερο για έργα που απαιτούν προσαρμοσμένες ρυθμίσεις ή εργασία εκτός σύνδεσης. [25]

3.3 Βιβλιοθήκες

3.3.1 Pandas

Το **Pandas** είναι ένα ισχυρό και προσαρμόσιμο πακέτο **Python** που διευκολύνει τις εργασίες χειρισμού δεδομένων. Η εργασία με δεδομένα σε μορφή πίνακα σε λογιστικά φύλλα ή πίνακες SQL ταιριάζει απόλυτα στο **Pandas**. Όταν εργάζονται με δομημένα δεδομένα στην **Python**, οι αναλυτές δεδομένων, οι επιστήμονες και οι μηχανικοί χρειάζονται το πακέτο **Pandas.** [12]

3.3.2 Numpy

Η λέξη **«Numerical Python»** συντομεύεται σε **«NumPy»**. Πρόκειται για μια βιβλιοθήκη **Python** που διατίθεται δωρεάν. Χρησιμοποιείται για τον επιστημονικό προγραμματισμό **Python**, δηλαδή για την Επιστήμη Δεδομένων, τη Μηχανική, τα Μαθηματικά και τον προγραμματισμό θετικών επιστημών. [13]
Το θεμέλιο της επιστήμης δεδομένων είναι οι εξαιρετικά περίπλοκοι επιστημονικοί υπολογισμοί. Οι επιστήμονες δεδομένων θέλουν ισχυρά εργαλεία προκειμένου να πραγματοποιήσουν αυτούς τους υπολογισμούς. Η χρήση αυτής της ενότητας για την εκτέλεση στατιστικών και μαθηματικών διαδικασιών στην **Python** είναι αρκετά χρήσιμη. Όταν πολλαπλασιάζει πίνακες ή πολυδιάστατους πίνακες, αποδίδει θαυμάσια. [13]

3.3.3 Natural Language Toolkit (NLTK)

Όταν εργάζεστε με δεδομένα ανθρώπινης γλώσσας, ένα από τα πιο συχνά χρησιμοποιούμενα πακέτα **Python** είναι το **Natural Language Toolkit (NLTK)**. Εκτός από λεξιλογικούς πόρους όπως το **WordNet** και περισσότερα από 50 σώματα δεδομένων (σύνολα δεδομένων κειμένου), προσφέρει φιλικές προς το χρήστη διεπαφές σε μια σειρά εργαλείων επεξεργασίας κειμένου για εργασίες όπως τμηματοποίηση **(tokenization),** ανάλυση **(parsing),** επισήμανση **(tagging)**, ταξινόμηση και σημασιολογική συλλογιστική. Λόγω των πλούσιων χαρακτηριστικών του και της αφθονίας των πόρων του, το NLTK χρησιμοποιείται συχνά σε ακαδημαϊκή έρευνα, εκπαιδευτικές πρωτοβουλίες, ακόμη και σε εμπορικά εγχειρήματα. Διευκολύνει εξελιγμένες δραστηριότητες Επεξεργασίας Φυσικής Γλώσσας **(NLP)**, όπως η ανάλυση συναισθήματος, η αναγνώριση ονομαστικών οντοτήτων και η επισήμανση μέρους του λόγου. Αν και πιο ισχυρές βιβλιοθήκες όπως η **SpaCy** ή η **Hugging Face Transformers** είναι καταλληλότερες για συστήματα μεγαλύτερης κλίμακας ή επιπέδου παραγωγής, ο φιλικός προς το χρήστη σχεδιασμός και η εκτενής τεκμηρίωση του **NLTK** το καθιστούν ένα εξαιρετικό μέρος για αρχάριους να αρχίσουν να εξερευνούν ιδέες **NLP** [14]. Παρακάτω θα μιλήσουμε για τις ενότητες που χρησημοποιήσαμε.

- **Word_tokenize:** Η μέθοδος **Word_tokenize** της **NLTK** χρησιμοποιείται για την εξαγωγή τμημάτων **(tokens)** από μια συμβολοσειρά χαρακτήρων. Στην πραγματικότητα επιστρέφονται οι συλλαβές μιας μεμονωμένης λέξης. Μια μεμονωμένη λέξη μπορεί να περιέχει μία ή περισσότερες συλλαβές. Παρέχεται μια **tokenized** έκδοση του κειμένου με τη συνιστώμενη φρασεολογία από το **NLTK**. Είναι η διαδικασία διαίρεσης ενός μεγάλου γραπτού εγγράφου σε διαχειρίσιμα κομμάτια γνωστά ως **tokens**. Στο **NLTK**, η λειτουργία **tokenize** της λέξης είναι ζωτικής σημασίας. [15]

- **Stopwords:** Οι λέξεις που εμφανίζονται συχνά σε οποιαδήποτε γλώσσα ή σώμα κειμένων είναι γνωστές ως **stopwords**. Δεν παρέχουν καμία νέα ή χρήσιμη πληροφορία στο κείμενο στο οποίο χρησιμοποιούνται για ορισμένες εργασίες **NLP**. Σε γενικές γραμμές, όροι όπως a, they, the, is, an, κ.λπ. είναι **stopwords**. Η βιβλιοθήκη αυτή παρέχει σε μια λίστα αυτές τις λέξεις με σκοπό να τις αφαιρέσουμε κατά την διάρκεια της προ επεξεργασίας. [16]
- **WordNetLemmatizer - Wordnet:** Ένα εργαλείο του πακέτου **Natural Language Toolkit (NLTK)** που ονομάζεται **WordNetLemmatizer** χρησιμοποιεί το **WordNet**, μια σημαντική λεξιλογική βάση δεδομένων της αγγλικής γλώσσας, για να πραγματοποιήσει τη λεμματοποίηση **(lemmetization)**. Η λεμματοποίηση αναφέρεται στη διαδικασία διάσπασης μιας λέξης στην πιο βασική της μορφή, ή «λήμμα». Η λεμματοποίηση λαμβάνει υπόψη τη σημασία της λέξης και τη μετατρέπει σε νόμιμη λέξη-ρίζα με βάση το μέρος του λόγου της, σε αντίθεση με το **stemming**, το οποίο συχνά απλώς αποκόπτει τα άκρα των λέξεων.
- **Sentiment :** Είναι ενότητα της βιβλιοθήκης **(NLTK)** της **Python** προσφέρει εργαλεία για την ανάλυση του συναισθήματος του κειμένου. Περιλαμβάνει μοντέλα βασισμένα σε λεξικό, όπως το μοντέλο **VADER (Valence Aware Dictionary and sEntiment Reasoner)**, το οποίο μπορεί να ταξινομήσει κείμενο ως θετικό, αρνητικό ή ουδέτερο. Αυτά τα μοντέλα λειτουργούν συγκρίνοντας τις λέξεις ενός κειμένου με προκαθορισμένα λεξικά λέξεων που σχετίζονται με βαθμολογίες συναισθήματος. Αυτό τα καθιστά χρήσιμα για εργασίες όπως η ανάλυση κριτικών πελατών, αναρτήσεων στα μέσα κοινωνικής δικτύωσης ή οποιουδήποτε άλλου περιεχομένου κειμένου για τον προσδιορισμό του συναισθηματικού του τόνου.

3.3.4 Scikit-learn

Το **Scikit-learn** είναι μια βιβλιοθήκη **Python** ανοικτού κώδικα που υλοποιεί μια σειρά αλγορίθμων μηχανικής μάθησης, προεπεξεργασίας, διασταυρούμενης επικύρωσης και οπτικοποίησης χρησιμοποιώντας μια ενοποιημένη διεπαφή. Πρόκειται για μια βιβλιοθήκη μηχανικής μάθησης ανοικτού κώδικα που παρέχει πληθώρα εργαλείων για διάφορες εργασίες μηχανικής μάθησης, όπως ταξινόμηση, παλινδρόμηση, ομαδοποίηση και πολλά άλλα. Για να μπορέσει κάποιος να δουλέψει με αυτήν την βιβλιοθήκη, είναι απαραίτητη προϋπόθεση να χρησιμοποιούμε την **Numpy** που αναφέραμε πιο πάνω.

- **TfidfVectorizer:** Στο **Scikit-learn,** ο **TF-IDF vectorizer** μετατρέπει μια συλλογή εγγράφων κειμένου σε έναν πίνακα χαρακτηριστικών TF-IDF, όπου κάθε γραμμή αντιπροσωπεύει ένα έγγραφο και κάθε στήλη αντιστοιχεί στη βαθμολογία TF-IDF ενός όρου. Σε αντίθεση με τον διανυσματοποιητή **Count Vectorizer**, ο οποίος μετρά μόνο τις εμφανίσεις λέξεων, ο διανυσματοποιητής **TF-IDF Vectorizer** λαμβάνει επίσης υπόψη τη σημασία των όρων σε όλα τα έγγραφα.
- **Logistic Regression:** Η λογιστική παλινδρόμηση στο **Scikit-learn** είναι ένα γραμμικό μοντέλο που χρησιμοποιείται για δυαδικές εργασίες ταξινόμησης, το οποίο προβλέπει την πιθανότητα μιας κλάσης με βάση τα χαρακτηριστικά εισόδου εφαρμόζοντας μια λογιστική συνάρτηση. Περισσότερες λεπτομέρειες θα αναφερθούν παρακάτω.

- **Metrics:** Στην ενότητα **sklearn.metrics** του **Scikit-learn** διατίθεται μια σειρά από συναρτήσεις για την αξιολόγηση της αποτελεσματικότητας των μοντέλων μηχανικής μάθησης. Είναι δυνατή η αξιολόγηση μοντέλων παλινδρόμησης και ταξινόμησης με τη χρήση αυτών των μετρικών.
- **Train_test_split:** Είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για την αξιολόγηση της απόδοσης ενός μοντέλου. Το σύνολο δεδομένων χωρίζεται σε δύο μέρη: το σύνολο εκπαίδευσης, το οποίο το μοντέλο χρησιμοποιεί για να μάθει πρότυπα, και το σύνολο δοκιμής, το οποίο διατηρείται χωριστά και χρησιμοποιείται για να αξιολογηθεί πόσο καλά το μοντέλο γενικεύει σε αόρατα δεδομένα. Με τον διαχωρισμό των δεδομένων, το μοντέλο μπορεί να εκπαιδευτεί στο ένα τμήμα, ενώ η ακρίβεια και η απόδοσή του αξιολογούνται στο άλλο. Αυτό βοηθά στην αποφυγή της υπερπροσαρμογής, διασφαλίζοντας ότι το μοντέλο δεν απομνημονεύει απλώς τα δεδομένα εκπαίδευσης, αλλά μπορεί επίσης να αποδώσει καλά σε νέα, αθέατα παραδείγματα.

3.3.5 Seaborn

Το **Seaborn** είναι μια γνωστή βιβλιοθήκη οπτικοποίησης δεδομένων **Python** που βασίζεται στην **Matplotlib** και παρέχει μια εύχρηστη διεπαφή για τη δημιουργία οπτικά ελκυστικών και εκπαιδευτικών στατιστικών γραφικών. Λόγω της συμβατότητάς της με τα πλαίσια δεδομένων **Pandas**, διευκολύνει την ταχεία και αποτελεσματική οπτικοποίηση και εξερεύνηση δεδομένων. [20]

Ένα ευρύ φάσμα αποτελεσματικών εργαλείων οπτικοποίησης δεδομένων, όπως χάρτες θερμότητας, διαγράμματα διασποράς, γραμμικά διαγράμματα και ραβδογράμματα, είναι διαθέσιμα από το **Seaborn**. Επιπλέον, προσφέρει βοήθεια σε πιο σύνθετες στατιστικές αναλύσεις, συμπεριλαμβανομένων των διαγραμμάτων κατανομής, των κατηγορικών διαγραμμάτων και της ανάλυσης παλινδρόμησης. [20]

Το κύριο πλεονέκτημα του **Seaborn** είναι η ικανότητά του να παράγει εντυπωσιακά διαγράμματα με λίγη εργασία κωδικοποίησης. Μπορείτε να αλλάξετε γρήγορα την ποικιλία των προεγκατεστημένων θεμάτων και χρωματικών σχημάτων για να ταιριάζει στα γούστα σας. Μια ποικιλία ενσωματωμένων στατιστικών λειτουργιών παρέχεται επίσης από το **Seaborn**, επιτρέποντας στους χρήστες να πραγματοποιούν γρήγορα και απλά περίπλοκες στατιστικές αναλύσεις χρησιμοποιώντας τις απεικονίσεις τους. [20]

Η ικανότητα του **Seaborn** να παράγει περίπλοκες απεικονίσεις πολλαπλών πλάνων είναι ένα άλλο αξιοσημείωτο χαρακτηριστικό. Οι χρήστες μπορούν εύκολα να συγκρίνουν διαφορετικές μεταβλητές ή υποσύνολα δεδομένων δημιουργώντας πλέγματα γραφικών παραστάσεων με το **Seaborn**. Εξαιτίας αυτού, είναι το τέλειο εργαλείο για διερευνητική ανάλυση και οπτικοποίηση δεδομένων. [20]

Με τη φιλική προς το χρήστη διεπαφή του, το **Seaborn** είναι μια ισχυρή και ευέλικτη βιβλιοθήκη οπτικοποίησης δεδομένων **Python** που καθιστά απλή τη δημιουργία οπτικά ελκυστικών και εκπαιδευτικών στατιστικών γραφικών. Διευκολύνει τη δημιουργία περίπλοκων οπτικοποιήσεων πολλαπλών πλάνων και προσφέρει μια ποικιλία εργαλείων για την οπτικοποίηση δεδομένων, συμπεριλαμβανομένης της εξελιγμένης στατιστικής ανάλυσης.[20]

3.3.6 Matplotlib

Η **Matplotlib** είναι μια δημοφιλής βιβλιοθήκη **Python** για τη δημιουργία στατικών, διαδραστικών και κινούμενων γραφικών. Σας επιτρέπει να δημιουργήσετε μια ποικιλία γραφικών παραστάσεων, όπως γραμμικά διαγράμματα, ραβδογράμματα, διαγράμματα διασποράς, ιστογράμματα και πολλά άλλα. Η Matplotlib είναι ιδιαίτερα παραμετροποιήσιμη, καθιστώντας την εξαιρετική επιλογή για χρήστες που απαιτούν ακριβή έλεγχο των οπτικοποιήσεών τους. Συνεργάζεται επίσης καλά με άλλες βιβλιοθήκες όπως η **NumPy** και η **Pandas**, καθιστώντας την ιδανική για την ανάλυση δεδομένων και την επιστημονική έρευνα. Συνολικά, είναι ένα απαραίτητο εργαλείο για κάθε εργασία οπτικοποίησης δεδομένων.

3.3.7 imblearn

Είναι μια βιβλιοθήκη της **Python** που βοηθάει στο χειρισμό συνόλων δεδομένων όπου ορισμένες κλάσεις έχουν πολύ περισσότερα δεδομένα από άλλες, κάτι που είναι ένα συνηθισμένο πρόβλημα στη μηχανική μάθηση. Χωρίς την αντιμετώπιση αυτής της ανισορροπίας, τα μοντέλα μπορεί να μεροληπτούν προς την πλειοψηφούσα κλάση, οδηγώντας σε κακή απόδοση για τη μειοψηφούσα κλάση. Το **Imbalanced-Learn** προσφέρει μεθόδους όπως η υπερδειγματοληψία (αύξηση των παραδειγμάτων από την κλάση της μειονότητας) ή η υποδειγματοληψία (μείωση των παραδειγμάτων από την κλάση της πλειοψηφίας) για την εξισορρόπηση του συνόλου δεδομένων. Αυτό βελτιώνει την ικανότητα του μοντέλου να αναγνωρίζει και να προβλέπει πιο δίκαια όλες τις κλάσεις, ακόμη και σε μη ισορροπημένα δεδομένα.

3.3.8 Regex

Στην **Python**, η βιβλιοθήκη **re** υποστηρίζει την εργασία με κανονικές εκφράσεις **(regex)** [21], οι οποίες είναι ένα εργαλείο για την αναζήτηση, την αντιστοίχιση και τον χειρισμό συμβολοσειρών με βάση συγκεκριμένα μοτίβα. Οι **Regex** μπορούν να χρησιμοποιηθούν για την εύρεση μοτίβων κειμένου, όπως αριθμοί τηλεφώνου, email ή URL, καθώς και για τον καθαρισμό και μετασχηματισμό δεδομένων με αντικατάσταση ή διαχωρισμό συμβολοσειρών σύμφωνα με συγκεκριμένα κριτήρια.

3.3.9 IPython.display

Είναι μια ενότητα της βιβλιοθήκης **IPython** που παρέχει εργαλεία για την προβολή περιεχομένου πλούσιων μέσων (όπως εικόνες, ήχο ή βίντεο) απευθείας σε ένα σημειωματάριο **Jupyter** ή σε ένα κέλυφος **IPython**. Σας επιτρέπει να εμφανίζετε αντικείμενα όπως **HTML**, **Markdown** ή **LaTeX**, διευκολύνοντας την παρουσίαση διαδραστικών και οπτικά ελκυστικών αποτελεσμάτων. Αυτό είναι ιδιαίτερα χρήσιμο για τη δημιουργία αναφορών, οπτικοποιήσεων δεδομένων ή εκπαιδευτικού περιεχομένου μέσα σε ένα διαδραστικό περιβάλλον.

3.4 Σύνολο Δεδομένων

Αυτό το σύνολο δεδομένων, αποτελείται από μια μεγάλη ποικιλία ειδησεογραφικών άρθρων, το καθένα με μια σειρά χαρακτηριστικών που αποτυπώνουν τόσο το περιεχόμενο όσο και το πλαίσιο της είδησης. Το σύνολο δεδομένων αυτό είναι σχεδιασμένο για ανάλυση συναισθήματος και ταξινόμηση θεμάτων, γεγονός που το καθιστά εξαιρετικά σημαντικό για τη μελέτη του τρόπου με τον οποίο διαμορφώνονται τα ειδησεογραφικά άρθρα και του τρόπου

με τον οποίο τα αντιλαμβάνονται οι αναγνώστες. Το σύνολο αυτό αποτελείται από της εξής κολώνες :

1. **Source :** Η πλατφόρμα ή το μέσο ενημέρωσης που δημοσίευσε το άρθρο
2. **Author :** Ο συγγραφέας του άρθρου. Ορισμένες φορές, το πεδίο αυτό αφήνεται κενό **(NaN)** εάν ο συντάκτης δεν προσδιορίζεται.
3. **Title :** Ο Τίτλος του Άρθρου.
4. **Description:** Παρέχει μια σύντομη εισαγωγή στο περιεχόμενο του άρθρου, επεκτείνοντας ελαφρώς τον τίτλο για να προσφέρει ένα στιγμιότυπο του τι θα καλύψει το άρθρο με περισσότερες λεπτομέρειες.
5. **URL :** Εδώ παρέχεται ο διαδικτυακός σύνδεσμος για το πλήρες άρθρο, επιτρέποντας στους χρήστες να έχουν πρόσβαση στην αρχική πηγή και να διαβάσουν ολόκληρο το άρθρο, αν θέλουν να εξερευνήσουν πέρα από την περίληψη και το συναίσθημα.
6. **Published at :** Αυτή η στήλη καταγράφει την ακριβή ημερομηνία και ώρα που δημοσιεύτηκε το άρθρο, συμπεριλαμβανομένων των πληροφοριών για τη ζώνη ώρας. Αυτό βοηθά στην παρακολούθηση της επικαιρότητας της είδησης, η οποία μπορεί να είναι ιδιαίτερα σημαντική για ιστορίες που κινούνται γρήγορα.
7. **Sentiment :** Ένα από τα πιο σημαντικά χαρακτηριστικά αυτού του συνόλου δεδομένων είναι η ανάλυση συναισθήματος, όπου κάθε άρθρο ταξινομείται ως θετικό, ουδέτερο ή αρνητικό με βάση το περιεχόμενο του.

Το σύνολο δεδομένων προήλθε από το **Kaggle** [23], μια ευρέως χρησιμοποιούμενη πλατφόρμα για έργα επιστήμης δεδομένων και μηχανικής μάθησης. Το **Kaggle** προσφέρει μια ποικιλία συνόλων δεδομένων που χρησιμοποιούνται τόσο για ακαδημαϊκή έρευνα όσο και για πρακτικές εφαρμογές στην επιστήμη των δεδομένων.
Το συγκεκριμένο σύνολο δεδομένων θα χρησιμοποιηθεί και στα δύο συστήματα, παρέχοντας την απαραίτητη πληροφορία για την κατηγοριοποίηση και την ανάλυση συναισθημάτων των ειδησεογραφικών άρθρων. Σκοπός είναι η δημιουργία ενός συστήματος συστάσεων που θα προσφέρει στον χρήστη σχετικές προτάσεις άρθρων, συνδυάζοντας το συναίσθημα και την κατηγορία κάθε άρθρου.

3.5 Οπτικοποίηση Δεδομένων

3.5.1 Ορισμός της οπτικοποίησης Δεδομένων

Η οπτικοποίηση δεδομένων είναι ο μετασχηματισμός των ακατέργαστων δεδομένων σε οπτικές αναπαραστάσεις. Συνήθως, οι απεικονίσεις αυτές έχουν τη μορφή διαγραμμάτων και γραφικών παραστάσεων. Η οπτικοποίηση δεδομένων αποσκοπεί στο να καταστήσει τα δεδομένα ευκολότερα και ταχύτερα κατανοητά, ακόμη και για άτομα που δεν έχουν εκπαιδευτεί στην ανάλυση ή δεν έχουν φυσική κλίση στους αριθμούς**.** [52]

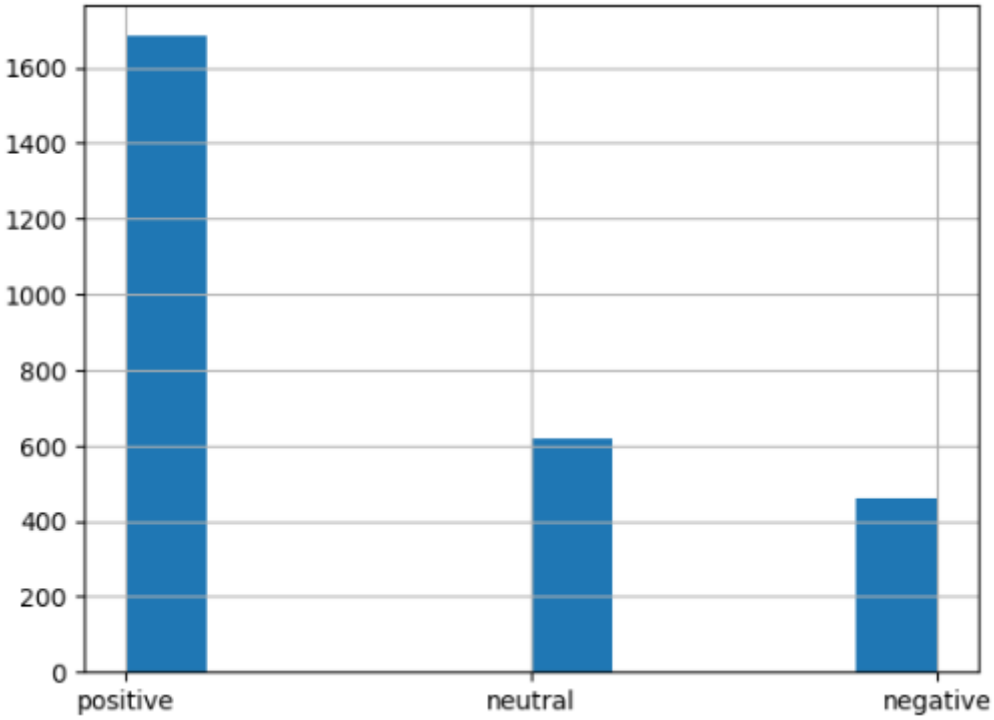
3.5.2 Η σημασία της οπτικοποίησης Δεδομένων

Η οπτικοποίηση δεδομένων λειτουργεί ως εργαλείο αφήγησης, μετατρέποντας τα δεδομένα σε κατανοητή μορφή που αναδεικνύει τάσεις, μοτίβα και ακραίες τιμές. Μια συναρπαστική οπτικοποίηση αφηγείται μια ιστορία αφαιρώντας την ακαταστασία από τα δεδομένα και αναδεικνύοντας πολύτιμες πληροφορίες. Η οπτικοποίηση δεδομένων είναι ζωτικής σημασίας για τους αναλυτές μηχανικής μάθησης επειδή τους επιτρέπει να κατανοούν και να αναλύουν

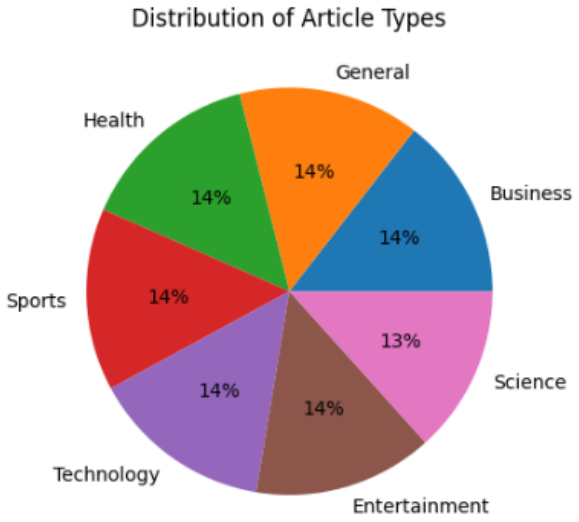
πολύπλοκα σύνολα δεδομένων και να τα παρουσιάζουν με σαφή τρόπο. Αποτελεί βασικό συστατικό της προετοιμασίας και της ανάλυσης δεδομένων, βοηθώντας στην ανίχνευση ακραίων τιμών, τάσεων και μοτίβων που άλλες αναλυτικές μέθοδοι μπορεί να χάσουν. [53]

Με την αυξανόμενη διαθεσιμότητα μεγάλων δεδομένων, είναι πιο σημαντικό από ποτέ να χρησιμοποιούνται τεχνικές οπτικοποίησης δεδομένων για την εξερεύνηση και την κατανόησή τους. Οι αλγόριθμοι μηχανικής μάθησης αποδίδουν καλύτερα με υψηλής ποιότητας, καθαρά δεδομένα και η οπτικοποίηση δεδομένων μπορεί να βοηθήσει στον εντοπισμό και την εξάλειψη ασυνεπειών ή ανωμαλιών.

3.5.3 Χρήση εργαλείων οπτικοποίησης στην εργασία



Εικόνα 9. Ιστόγραμμα για την στήλη Sentiment



Εικόνα 10. Κυκλικό διάγραμμα για τις κατηγορίες των άρθρων

3.6 Προεπεξεργασία κείμενου

Τόσο τα μοντέλα που βασίζονται σε κανόνες όσο και τα μοντέλα μηχανικής μάθησης αποδίδουν σημαντικά καλύτερα όταν τα δεδομένα τους είναι υψηλής ποιότητας. Ανεπαρκή ή χαμηλής ποιότητας δεδομένα μπορούν να οδηγήσουν σε κακή λήψη αποφάσεων, μειωμένη ακρίβεια και χαμηλότερη αποτελεσματικότητα για κάθε προσέγγιση. Όταν πρόκειται για δεδομένα κειμένου, τα οποία είναι συχνά αδόμητα και θορυβώδη, η πρόκληση αυξάνεται. Η προεπεξεργασία κειμένου είναι ένα σημαντικό βήμα για τη μετατροπή ακατάστατων, μη δομημένων δεδομένων κειμένου σε δομημένη μορφή που μπορεί να χρησιμοποιηθεί αποτελεσματικά από μοντέλα που βασίζονται σε κανόνες και αλγορίθμους μηχανικής μάθησης. Η προεπεξεργασία βελτιώνει την ποιότητα και τη δομή των δεδομένων κειμένου, με αποτέλεσμα καλύτερη απόδοση των μοντέλων, ακριβέστερες προβλέψεις και αξιοποιήσιμες γνώσεις σε διάφορες εφαρμογές, όπως η ανάλυση συναισθήματος, η ταξινόμηση κειμένου και τα συστήματα συστάσεων. Τα σωστά προετοιμασμένα δεδομένα επιτρέπουν και στα δύο μοντέλα να αποδίδουν βέλτιστα, επιτρέποντάς τους να εξάγουν πολύτιμα μοτίβα από το κείμενο. Παρακάτω θα αναφέρουμε τα βήματα τα οποία είναι απαραίτητα για την σωστή επεξεργασία κειμένου. [26]

3.6.1 Βήμα 1: Καθαρισμός Κειμένου (Text Cleaning)

Σε αυτό το βήμα, θα κάνουμε τα βασικά βήματα για να καθαρίσουμε το κείμενο. Οι ενέργειες αυτές περιλαμβάνουν τη μετατροπή όλου του κειμένου σε πεζά, την αφαίρεση μη λεκτικών χαρακτήρων και κενών σημείων και την αφαίρεση τυχόν αριθμητικών ψηφίων.

```python
# Define a function to clean the text
def clean(text):
# Removes all special characters and numericals leaving the alphabets
    text = re.sub('[^A-Za-z]+', ' ', text).lower()
    return text

# Cleaning the text in the review column
df['Cleaned Reviews'] = df['Description'].apply(clean)
df.head()
```

Εικόνα 11. Συνάρτηση Καθαρισμού κειμένου

```python
# Define a function to remove URLs from text
def remove_urls(text):
    # Define a regex pattern to match URLs
    url_pattern = re.compile(r'\bmarketscreener\scom\s\b')
    return url_pattern.sub('', text)

# Apply the function to the 'text' column and create a new column 'clean Reviews'
df['Cleaned Reviews'] = df['Cleaned Reviews'].apply(remove_urls)
```

Εικόνα 12. Συνάρτηση Καθαρισμού URL

Στη στήλη που αναφέρεται στην Περιγραφή του άρθρου, οι λέξεις *"marketscreener"* και *"com"* είναι άσχετες για την ανάκτηση συναισθήματος, καθώς δεν προσφέρουν ουσιαστική πληροφορία για την ανάλυση του περιεχομένου του κειμένου. Αυτές οι λέξεις σχετίζονται με

τεχνικούς όρους ή ονόματα ιστοσελίδων και δεν έχουν συναισθηματικό βάρος. Επομένως, κατά την προεπεξεργασία του κειμένου, μπορούν να αφαιρεθούν, καθώς αποτελούν "θόρυβο" και δεν συμβάλλουν στην ανάλυση του συναισθήματος του κειμένου.

3.6.2 Βήμα 2: Τμηματοποίηση (Tokenization)

Η τμηματοποίηση είναι η διαδικασία διαίρεσης ενός κειμένου σε μικρότερες μονάδες που ονομάζονται **tokens**. Στην επεξεργασία φυσικής γλώσσας, οι μάρκες είναι συνήθως λέξεις ή υπολέξεις. Η τμηματοποίηση είναι ένα σημαντικό βήμα σε πολλές εργασίες NLP, όπως η επεξεργασία κειμένου, η μοντελοποίηση γλώσσας και η μηχανική μετάφραση. Η διαδικασία περιλαμβάνει τη διαίρεση μιας συμβολοσειράς ή ενός κειμένου σε μια λίστα από tokens. Τα σημεία μπορούν να θεωρηθούν ως μέρη, όπως μια λέξη σε μια πρόταση ή μια πρόταση σε μια παράγραφο. [28]

Η τμηματοποίηση είναι η διαδικασία τμηματοποίησης μη δομημένων δεδομένων και κειμένου φυσικής γλώσσας σε διακριτά κομμάτια πληροφοριών και η αντιμετώπισή τους ως ξεχωριστά στοιχεία. Οι μάρκες μέσα σε ένα έγγραφο μπορούν να χρησιμοποιηθούν ως διανύσματα, μετατρέποντας ένα μη δομημένο έγγραφο κειμένου σε μια αριθμητική δομή δεδομένων κατάλληλη για ανάλυση μηχανικής μάθησης. Αυτή η ταχεία μετατροπή επιτρέπει σε έναν υπολογιστή να χρησιμοποιήσει αυτά τα στοιχεία που έχουν μετατραπεί σε **token** αμέσως για να δρομολογήσει πρακτικές ενέργειες και απαντήσεις. Εναλλακτικά, θα μπορούσαν να χρησιμοποιηθούν ως χαρακτηριστικά σε έναν αγωγό μηχανικής μάθησης, προτρέποντας σε πιο προηγμένες διαδικασίες λήψης αποφάσεων ή συμπεριφορές. [28]

Η κωδικοποίηση μπορεί να ταξινομηθεί σε διάφορους τύπους με βάση τον τρόπο κατάτμησης του κειμένου. οι τύποι αυτοί θα αναφερθούν στις παρακάτω παραγράφους.

Η τμηματοποίηση των λέξεων **(Word tokenization)** σε λέξεις διαχωρίζει το κείμενο σε μεμονωμένες λέξεις. Πολλές εργασίες **NLP** ακολουθούν αυτή την προσέγγιση, αντιμετωπίζοντας τις λέξεις ως τις θεμελιώδεις μονάδες νοήματος. [28]

```
Input: "Tokenization is an important NLP task."
Output: ["Tokenization", "is", "an", "important", "NLP", "task", "."]
```

*Εικόνα 13. Παράδειγμα **Word Tokenization [28]***

Η τμηματοποίηση προτάσεων **(Sentence tokenization)**, περιλαμβάνει τη διαίρεση μιας παραγράφου ή ενός μεγαλύτερου μπλοκ κειμένου σε μεμονωμένες προτάσεις. Αυτό είναι ένα σημαντικό βήμα σε εργασίες όπως η περίληψη κειμένου ή η μετάφραση, όπου η κατανόηση της δομής των προτάσεων είναι ζωτικής σημασίας.

```
Input: "Tokenization is an important NLP task. It helps break down text into smaller units."
Output: ["Tokenization is an important NLP task.", "It helps break down text into smaller units."]
```

*Εικόνα 14. Παράδειγμα **Sentence Tokenization [28]***

Η τμηματοποίηση υπολέξεων **(Subword tokenization)** είναι η διαδικασία διάσπασης των λέξεων σε μικρότερες μονάδες, η οποία είναι ιδιαίτερα χρήσιμη όταν εργάζεστε με μορφολογικά πλούσιες γλώσσες ή σπάνιες λέξεις.[28]

```
Input: "tokenization"
Output: ["token", "ization"]
```

*Εικόνα 15. Παράδειγμα **Subword Tokenization [28]***

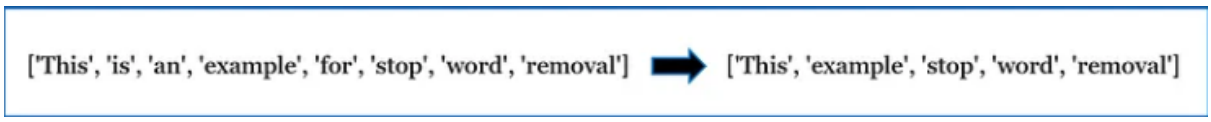
Η τμηματοποίηση χαρακτήρων **(Character tokenization)** διαχωρίζει το κείμενο σε μεμονωμένους χαρακτήρες. Αυτό μπορεί να είναι χρήσιμο κατά τη μοντελοποίηση γλώσσας σε επίπεδο χαρακτήρων. [28]

```
Input: "Tokenization"
Output: ["T", "o", "k", "e", "n", "i", "z", "a", "t", "i", "o", "n"]
```

*Εικόνα 16. Παράδειγμα **Character Tokenization [28]***

3.6.3 Βήμα 3: Αφαίρεση σταθερών λέξεων (Stopwords Removal)

Οι σταθερές λέξεις είναι οι λέξεις που συναντάμε συχνότερα σε κάθε φυσική γλώσσα. Αυτές οι λέξεις μπορεί να προσθέσουν μικρή αξία στο νόημα του εγγράφου κατά την ανάλυση δεδομένων κειμένου και τη δημιουργία μοντέλων **NLP**. Ως αποτέλεσμα, η αφαίρεση των **stopwords** μας επιτρέπει να επικεντρωθούμε στις πιο σημαντικές πληροφορίες του κειμένου και βελτιώνει την ακρίβεια της ανάλυσής μας. Ένα πλεονέκτημα της αφαίρεσης των stopwords είναι ότι μειώνεται το μέγεθος του συνόλου δεδομένων, γεγονός που μειώνει το χρόνο που απαιτείται για την εκπαίδευση των μοντέλων επεξεργασίας φυσικής γλώσσας.[27]



*Εικόνα 17. Παράδειγμα **Stopwords Removal [27]***

3.6.4 Βήμα 4: Επισήμανση Μέρος λόγου (POS tagging)

Η επισήμανση μέρους του λόγου **(POS)** είναι μια θεμελιώδης εργασία στην Επεξεργασία Φυσικής Γλώσσας **(NLP)** που αποδίδει μια γραμματική κατηγορία (π.χ. ουσιαστικό, ρήμα, επίθετο) σε κάθε λέξη μιας πρότασης. Στόχος είναι η κατανόηση της συντακτικής δομής μιας πρότασης και ο προσδιορισμός των γραμματικών ρόλων των μεμονωμένων λέξεων. Η επισήμανση **POS** είναι απαραίτητη για μια ποικιλία εφαρμογών **NLP**, όπως η ανάλυση κειμένου, η μηχανική μετάφραση και η ανάκτηση πληροφοριών. [29]

```
pos = nltk.pos_tag(tokens)
pos
```

```
[('This', 'DT'),
 ('is', 'VBZ'),
 ('an', 'DT'),
 ('article', 'NN'),
 ('on', 'IN'),
 ('Sentiment', 'NN'),
 ('Analysis', 'NN')]
```

*Εικόνα 18. Παράδειγμα **POS tagging** με χρήση της βιβλιοθήκης **NLTK** [29]*

3.6.5 Βήμα 5: Περιστολή/Λημματοποίηση (Stemming/Lemmatization)

Η Περιστολή και η Λημματοποίηση είναι τεχνικές προεπεξεργασίας κειμένου που χρησιμοποιούνται στην επεξεργασία φυσικής γλώσσας **(NLP)**. Συγκεκριμένα, μειώνουν τις κλιτές μορφές των λέξεων σε ένα σύνολο δεδομένων κειμένου σε μία μόνο κοινή ρίζα λέξης ή μορφή λεξικού, γνωστή επίσης ως «λήμμα» στην υπολογιστική γλωσσολογία. [32]

Η Περιστολή και η Λημματοποίηση είναι ιδιαίτερα χρήσιμα σε συστήματα ανάλυσης συναισθήματος, όπου ο στόχος είναι να εντοπιστεί το συναίσθημα που κρύβεται πίσω από ένα κείμενο, ανεξάρτητα από τις συγκεκριμένες κλίσεις των λέξεων. Για παράδειγμα, ένα σύστημα που αναλύει το συναίσθημα μιας αναθεώρησης που περιέχει τη λέξη *«happy»* θα πρέπει να αναγνωρίσει ότι οι λέξεις *«happiness»*, *«happiest»* ή *«happily»* φέρουν παρόμοιο θετικό συναίσθημα. Η Περιστολή και η Λημματοποίηση επιτρέπουν στο σύστημα να μειώσει αυτές τις παραλλαγές σε μια κοινή ρίζα ή βασική μορφή, βελτιώνοντας την ακρίβεια της ανίχνευσης συναισθήματος. Βοηθούν επίσης στη βελτίωση της απόδοσης των αλγορίθμων μηχανικής μάθησης και κανόνων απλοποιώντας το κείμενο, μειώνοντας το μέγεθος του λεξιλογίου και διασφαλίζοντας ότι οι διαφορετικές μορφές της ίδιας λέξης αντιμετωπίζονται με συνέπεια. [32]

Οι αλγόριθμοι Περιστολής χρησιμοποιούν έναν κατάλογο κοινών προθημάτων και επιθημάτων για την αφαίρεση της αρχής ή του τέλους μιας λέξης που μπορεί να έχει κλίση. Αυτή η διαδικασία είναι γενικά αδιάκριτη και μπορεί να παράγει βασικές μορφές μιας λέξης με λανθασμένη ορθογραφία ή σημασία. Η Περιστολή λειτουργεί χωρίς γνώση των συμφραζομένων, οπότε δεν μπορεί να διακρίνει μεταξύ παρόμοιων λέξεων με διαφορετικές σημασίες. [31]

Από την άλλη πλευρά, η Λημματοποίηση είναι πιο πολύπλοκη από τη Περιστολή, επειδή απαιτεί την ταξινόμηση των λέξεων τόσο με βάση το μέρος του λόγου όσο και με βάση την κλίση. Αυτό μπορεί να γίνει αρκετά περίπλοκο σε άλλες γλώσσες εκτός από την αγγλική, όπου οι μόνες κλιτές μορφές είναι ο ενικός ή ο πληθυντικός αριθμός, ο χρόνος του ρήματος και οι συγκριτικές ή υπερθετικές μορφές των επιρρημάτων και των επιθέτων. [31]

```
Original: There is nothing either good or bad but thinking makes it so.
Tokenized: ['There', 'is', 'nothing', 'either', 'good', 'or', 'bad', 'but', 'thinkin
g', 'makes', 'it', 'so', '.']
Stemmed: ['there', 'is', 'noth', 'either', 'good', 'or', 'bad', 'but', 'think', 'mak
e', 'it', 'so', '.']
```

*Εικόνα 19. Παράδειγμα χρήσης Αλγόριθμου **Stemming** [31]*

```
Original: There is nothing either good or bad but thinking makes it so.
Tokenized: ['There', 'is', 'nothing', 'either', 'good', 'or', 'bad', 'but', 'thinkin
g', 'makes', 'it', 'so', '.']
Lemmatized: There be nothing either good or bad but think make it so .
```

*Εικόνα 20. Παράδειγμα χρήσης Αλγορίθμου **Lemmatization** [31]*

Για τους σκοπούς της εργασίας, θα εφαρμόσουμε έναν αλγόριθμο Λημματοποίησης για να βελτιώσουμε την απόδοση των μοντέλων ανάλυσης συναισθήματος, είτε αυτά βασίζονται σε κανόνες είτε σε μηχανική μάθηση. Με τη λημματοποίηση θα μπορούμε να επεξεργαζόμαστε λέξεις σε όλες τις μορφές τους με πιο αποτελεσματικό τρόπο, βοηθώντας έτσι το σύστημα να αναγνωρίζει τα συναισθήματα με μεγαλύτερη ακρίβεια, είτε το κείμενο χρησιμοποιεί διαφορετικές παραλλαγές της ίδιας λέξης είτε όχι.

3.6.6 Υλοποίηση βημάτων 2-4 Για το Μοντέλο Κανόνων

```
pos_dict = {'J':wordnet.ADJ, 'V':wordnet.VERB, 'N':wordnet.NOUN, 'R':wordnet.ADV}


def token_stop_pos(text):
    tags = nltk.pos_tag(word_tokenize(text))
    newlist = []
    for word, tag in tags:
        if word.lower() not in set(stopwords.words('english')):
          newlist.append(tuple([word, pos_dict.get(tag[0])]))
    return newlist

df['POS tagged'] = df['Cleaned Reviews'].apply(token_stop_pos)
df.head()
```

Εικόνα 21. Υλοποίηση κώδικα Για τα βήματα 2-4

Η συνάρτηση **token_stop_pos** λαμβάνει το κείμενο και εκτελεί τμηματοποίηση, αφαιρεί τις σταθερές λέξεις και κάνει επισήμανση του μέρους λόγου. Την εφαρμόσαμε στη στήλη «**Cleaned Reviews**» και δημιουργήσαμε μια νέα στήλη για τα δεδομένα «**POS tagged**». Όπως αναφέρθηκε προηγουμένως, για να ληφθεί το ακριβές Λήμμα ο αλγόριθμος Λημματοποίησης **(WordNetLemmatizer)** απαιτεί ετικέτες **POS** με τη μορφή 'n', 'a', κ.λπ. Όμως οι ετικέτες POS που λαμβάνονται από το **pos_tag** έχουν τη μορφή 'NN', 'ADJ', κ.λπ. Για να αντιστοιχίσουμε το **pos_tag** σε ετικέτες του wordnet, δημιουργήσαμε μια δομή δεδομένων λεξικού **pos_dict**. Κάθε **pos_tag** που αρχίζει με J αντιστοιχίζεται στο **wordnet.ADJ**, κάθε **pos_tag** που αρχίζει με R αντιστοιχίζεται στο wordnet.**ADV** και ούτω κάθε εξής. Οι ετικέτες που μας ενδιαφέρουν είναι: ουσιαστικό, επίθετο, επίρρημα, ρήμα. Οτιδήποτε εκτός αυτών των τεσσάρων αντιστοιχίζεται στο **None**.

3.6.7 Υλοποίηση Βήματος 5 για Μοντέλο Κανόνων

```
wordnet_lemmatizer = WordNetLemmatizer()
def lemmatize(pos_data):
    lemma_rew = " "
    for word, pos in pos_data:
      if not pos:
        lemma = word
        lemma_rew = lemma_rew + " " + lemma
      else:
        lemma = wordnet_lemmatizer.lemmatize(word, pos=pos)
        lemma_rew = lemma_rew + " " + lemma
    return lemma_rew

df['Lemma'] = df['POS tagged'].apply(lemmatize)
df.head()
```

Εικόνα 22. Συνάρτηση Λημματοποίησης με την Χρήση του αλγορίθμου ***WordNetLemmatizer***

Η συνάρτηση **lemmatize** που δέχεται πλειάδες pos_tag και δίνει το Λήμμα για κάθε λέξη στο pos_tag με βάση το μέρος του λόγου αυτής της λέξης. Την εφαρμόσαμε στη στήλη **«POS tagged»** και δημιουργήσαμε μια στήλη **«Lemma»** για να αποθηκεύσουμε το αποτέλεσμα.

3.6.8 Υλοποίηση βήματος 4 για το Μοντέλο Μηχανικής Μάθησης

```
def get_wordnet_pos(treebank_tag):
  if treebank_tag.startswith('J'):
    return wordnet.ADJ
  elif treebank_tag.startswith('V'):
    return wordnet.VERB
  elif treebank_tag.startswith('N'):
    return wordnet.NOUN
  elif treebank_tag.startswith('R'):
    return wordnet.ADV
  else:
    return wordnet.NOUN
```

Εικόνα 23. Συνάρτηση Επισήμανσης μέρος λόγου

Η συνάρτηση **get_wordnet_pos** παίρνει ένα συντακτικό μέρος του λόγου **(POS tag)**, που συνήθως προέρχεται από έναν αναλυτή γλώσσας, και το μετατρέπει σε μορφή που μπορεί να καταλάβει το WordNet, μια λεξιλογική βάση δεδομένων. Τα **POS tags** ξεκινούν με συγκεκριμένα γράμματα για να αντιπροσωπεύσουν τον τύπο της λέξης: για παράδειγμα, το 'J' σημαίνει επίθετο, το 'V' ρήμα, το 'N' ουσιαστικό, και το 'R' επίρρημα. Η συνάρτηση ελέγχει το πρώτο γράμμα του **tag** και επιστρέφει τον σωστό τύπο για το **WordNet** (π.χ., **wordnet.ADJ** για επίθετα ή **wordnet.VERB** για ρήματα). Αν δεν ταιριάζει με αυτά, επιστρέφει **wordnet.NOUN**, δηλαδή το θεωρεί ως ουσιαστικό. Αυτό είναι χρήσιμο σε διαδικασίες όπως

η λημματοποίηση, όπου οι λέξεις πρέπει να επεξεργαστούν ανάλογα με το γραμματικό τους ρόλο.

3.6.9 Υλοποίηση βήματος 2 και 5 για το Μοντέλο Μηχανικής Μάθησης

```
class LemmaTokenizer:
  def __init__(self):
    self.wnl = WordNetLemmatizer()
  def __call__(self, doc):
    tokens = word_tokenize(doc)
    words_and_tags = nltk.pos_tag(tokens)
    return [self.wnl.lemmatize(word, pos=get_wordnet_pos(tag)) \
            for word, tag in words_and_tags]
```

*Εικόνα 24. Κλάση **Lemmatokenizer** για την χρήση Λημματοποίησης*

Η κλάση **LemmaTokenizer** είναι υπεύθυνη για τη λημματοποίηση κειμένου, δηλαδή τη μετατροπή των λέξεων στη βασική τους μορφή. Όταν δημιουργείται ένα αντικείμενο από αυτή την κλάση, αρχικοποιείται ένας λημματοποιητής από το WordNet. Όταν η κλάση καλείται με ένα κείμενο, πρώτα το χωρίζει σε λέξεις (tokens) και στη συνέχεια, με τη βοήθεια του nltk.pos_tag, αναγνωρίζει το συντακτικό ρόλο κάθε λέξης (όπως ρήμα, ουσιαστικό, επίθετο). Τέλος, για κάθε λέξη, χρησιμοποιεί τη μέθοδο lemmatize για να την επαναφέρει στη βασική της μορφή, λαμβάνοντας υπόψη το γραμματικό της ρόλο.

3.7 Εξαγωγή Χαρακτηριστικών (Feature Extraction)

Η εξαγωγή χαρακτηριστικών είναι μια τεχνική μηχανικής μάθησης και ανάλυσης δεδομένων που εντοπίζει και εξάγει σχετικά χαρακτηριστικά από ακατέργαστα δεδομένα. Αυτά τα χαρακτηριστικά χρησιμοποιούνται αργότερα για τη δημιουργία ενός πιο κατατοπιστικού συνόλου δεδομένων, το οποίο μπορεί στη συνέχεια να χρησιμοποιηθεί για διάφορες εργασίες όπως Ταξινόμηση **(Classification)**, Πρόβλεψη **(Prediction)**, Ομαδοποίηση **(Clustering).** [33]

Η εξαγωγή χαρακτηριστικών επιδιώκει να μειώσει την πολυπλοκότητα των δεδομένων (γνωστή και ως «διαστατικότητα των δεδομένων»), διατηρώντας ταυτόχρονα όσο το δυνατόν περισσότερες σχετικές πληροφορίες. Αυτό βελτιώνει την απόδοση και την αποδοτικότητα των αλγορίθμων μηχανικής μάθησης και απλοποιεί τη διαδικασία ανάλυσης. Η εξαγωγή χαρακτηριστικών μπορεί να περιλαμβάνει την ανάπτυξη νέων χαρακτηριστικών το λεγόμενο "**feature engineering**", καθώς και την επεξεργασία δεδομένων για τη διάκριση και την απλούστευση της χρήσης των σημαντικών χαρακτηριστικών από τα λιγότερο σημαντικά. [33]

3.7.1 Η Σημασία της Εξαγωγής Χαρακτηριστικών

Πολλές πρακτικές εφαρμογές βασίζονται σε μεγάλο βαθμό στην εξαγωγή χαρακτηριστικών. Η εξαγωγή χαρακτηριστικών είναι απαραίτητη για την αναγνώριση εικόνας και ομιλίας, την προγνωστική μοντελοποίηση και την επεξεργασία φυσικής γλώσσας **(NLP)**. Σε αυτές τις περιπτώσεις, τα ακατέργαστα δεδομένα μπορεί να περιλαμβάνουν πολυάριθμα άσχετα ή περιττά χαρακτηριστικά. Ως αποτέλεσμα, οι αλγόριθμοι δυσκολεύονται να επεξεργαστούν τα δεδομένα με ακρίβεια. [33]

Κατά την εξαγωγή χαρακτηριστικών, τα σχετικά χαρακτηριστικά διαχωρίζονται από τα άσχετα. Με λιγότερα χαρακτηριστικά προς επεξεργασία, το σύνολο δεδομένων γίνεται απλούστερο και η ακρίβεια και η αποτελεσματικότητα της ανάλυσης βελτιώνονται. [33]

3.7.2 Μέθοδοι εξαγωγής χαρακτηριστικών για δεδομένα κειμένου

Οι μέθοδοι εξαγωγής χαρακτηριστικών σε κείμενο αποτελούν βασικό βήμα στην αναπαράσταση των δεδομένων για εφαρμογές μηχανικής μάθησης. Ένα από τα πιο διαδεδομένα μοντέλα είναι το **Bag of Words (BoW)**, το οποίο μετατρέπει το κείμενο σε διανύσματα βασισμένα στη συχνότητα των λέξεων, χωρίς να λαμβάνεται υπόψη η σειρά των λέξεων. Ένα άλλο σημαντικό μοντέλο είναι ο **TF-IDF Vectorizer**, ο οποίος θα αναλυθεί με περισσότερες λεπτομέρειες στο παρακάτω κείμενο. Αυτές οι μέθοδοι βοηθούν στην καλύτερη αναπαράσταση των κειμένων με τρόπο που μπορεί να αξιοποιηθεί από αλγορίθμους μηχανικής μάθησης. [34]

3.7.3 TF-IDF (Term Frequency Inverse Document Frequency)

Το TF-IDF είναι μια αριθμητική στατιστική που μετρά τη σημασία μιας λέξης σε ένα έγγραφο σε σχέση με μια συλλογή εγγράφων γνωστή ως σώμα κειμένων **(corpus)**. Ο στόχος του TF-IDF είναι να ποσοτικοποιήσει τη σημασία ενός όρου σε ένα έγγραφο με βάση τη συχνότητά του στο έγγραφο και τη σπανιότητά του σε πολλαπλά έγγραφα. Η τεχνική αυτή αποτελείται απο 2 σημεία, την συχνότητα του όρου **(term frequency)** και την αντιστροφή συχνότητα εγγράφου **(Inverse Document Frequency).** [35]

Το **term frequency** καθορίζει τη συχνότητα ενός όρου σε ένα έγγραφο. Υπολογίζεται ως ο αριθμός των φορών που εμφανίζεται ένας όρος σε ένα έγγραφο διαιρούμενος με τον συνολικό αριθμό των όρων στο έγγραφο. Ο στόχος είναι να δοθεί έμφαση στις λέξεις που εμφανίζονται συχνά σε ένα έγγραφο. [35]

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

*Εικόνα 25. Μαθηματική εξίσωση του **term frequency [35]***

Το **IDF** υπολογίζει τη σπανιότητα ενός όρου σε ένα σύνολο εγγράφων. Υπολογίζεται λαμβάνοντας τον λογάριθμο του συνολικού αριθμού των εγγράφων διαιρούμενο με τον αριθμό των εγγράφων που περιέχουν τον όρο. Ο στόχος είναι να τιμωρούνται οι κοινές λέξεις που βρίσκονται σε όλα τα έγγραφα. [35]

$$IDF(t,D) = \log\left(\frac{\text{Total number of documents in the corpus } N}{\text{Number of documents containing term } t}\right)$$

*Εικόνα 26. Μαθηματική εξίσωση του **IDF [35]***

3.7.4 Υλοποίηση TF-IDF σε κώδικα

```
vectorizer = TfidfVectorizer(max_features=4000, tokenizer=LemmaTokenizer(), stop_words='english', lowercase = True)
```

*Εικόνα 27. Δημιουργία Αντικειμένου της κλάσης **TF-IDF vectorizer***

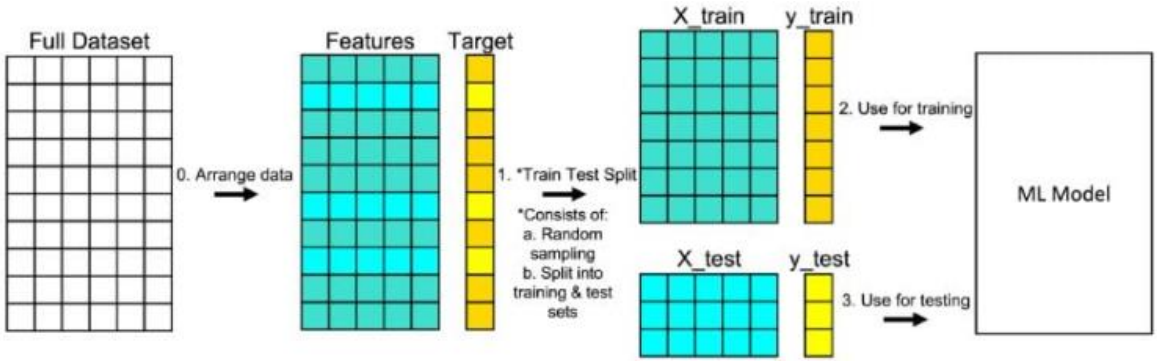
Αυτή η γραμμή κώδικα δημιουργεί έναν TfidfVectorizer που μετατρέπει δεδομένα κειμένου σε αριθμητικά χαρακτηριστικά με βάση τη σημασία τους στο έγγραφο. Περιορίζει τον αριθμό των χαρακτηριστικών (λέξεις ή tokens) στα 4000 πιο σημαντικά, διασφαλίζοντας ότι μόνο οι πιο σχετικές λέξεις χρησιμοποιούνται για ανάλυση. Αφαιρεί επίσης τις κοινές αγγλικές λέξεις στάσης, όπως το «the» ή το «and», οι οποίες συνήθως δεν προσθέτουν μεγάλη αξία στην κατανόηση του νοήματος του κειμένου. Επιπλέον, μετατρέπει όλες τις λέξεις σε πεζά γράμματα, έτσι ώστε παρόμοιες λέξεις με διαφορετική κεφαλαιοποίηση να αντιμετωπίζονται το ίδιο. Αυτό βοηθά στην προετοιμασία του κειμένου για τα μοντέλα μηχανικής μάθησης. Η παράμετρος tokenizer, για το οποίο μιλήσαμε προηγουμένως, χειρίζεται τη διαδικασία τμηματοποίησης και της λημματοποίησης των λέξεων.

3.8 Διαχωρισμός σε εκπαίδευση και δοκιμή (Train Test split)

Ο διαχωρισμός εκπαίδευσης-δοκιμής είναι μια μέθοδος αξιολόγησης της απόδοσης ενός αλγορίθμου μηχανικής μάθησης. Μπορεί να εφαρμοστεί σε προβλήματα ταξινόμησης και παλινδρόμησης, καθώς και σε οποιονδήποτε αλγόριθμο μάθησης με επίβλεψη.

Η διαδικασία περιλαμβάνει τη διαίρεση ενός συνόλου δεδομένων σε δύο υποσύνολα. Το πρώτο υποσύνολο χρησιμοποιείται για την προσαρμογή του μοντέλου και είναι γνωστό ως σύνολο δεδομένων εκπαίδευσης. Το δεύτερο υποσύνολο δεν χρησιμοποιείται για την εκπαίδευση του μοντέλου- αντίθετα, το στοιχείο εισόδου του συνόλου δεδομένων τροφοδοτείται στο μοντέλο και γίνονται προβλέψεις και συγκρίνονται με τα αναμενόμενα αποτελέσματα. Αυτό το δεύτερο σύνολο δεδομένων είναι γνωστό ως σύνολο δεδομένων δοκιμής.

Ο στόχος είναι να εκτιμηθεί η απόδοση του μοντέλου μηχανικής μάθησης σε νέα δεδομένα που δεν έχουν χρησιμοποιηθεί προηγουμένως για την εκπαίδευσή του. Με αυτόν τον τρόπο σκοπεύουμε να εφαρμόσουμε το μοντέλο στην πράξη. Συγκεκριμένα, να το προσαρμόσουμε σε υπάρχοντα δεδομένα με γνωστές εισόδους και εξόδους και στη συνέχεια να προβλέψουμε νέα παραδείγματα στο μέλλον, όπου δεν έχουμε τις αναμενόμενες τιμές εξόδου ή στόχου. Όταν είναι διαθέσιμο ένα αρκετά μεγάλο σύνολο δεδομένων, μπορεί να χρησιμοποιηθεί η διαδικασία εκπαίδευσης δοκιμής.



*Εικόνα 28. Διαδικασία **Train test Split***

3.8.1 Υλοποίηση Διαχωρισμού Εκπαίδευσης και Δοκιμής

```
X_train, X_test, Y_train, Y_test = train_test_split(x, y, test_size=0.2, random_state = 50)
```

Εικόνα 29. Υλοποίηση συνάρτησης ***train_test_split***

Οι μεταβλητές **x** και **y** αντιπροσωπεύουν τα χαρακτηριστικά εισόδου και τις αντίστοιχες τιμές-στόχους. Η παράμετρος **test_size** καθορίζει το ποσοστό των δεδομένων που πρέπει να διατεθεί για δοκιμές. Σε αυτή την περίπτωση, **test_size=0,20** σημαίνει ότι το 20% των δεδομένων θα χρησιμοποιηθεί για δοκιμή, ενώ το υπόλοιπο 80% θα χρησιμοποιηθεί για εκπαίδευση.

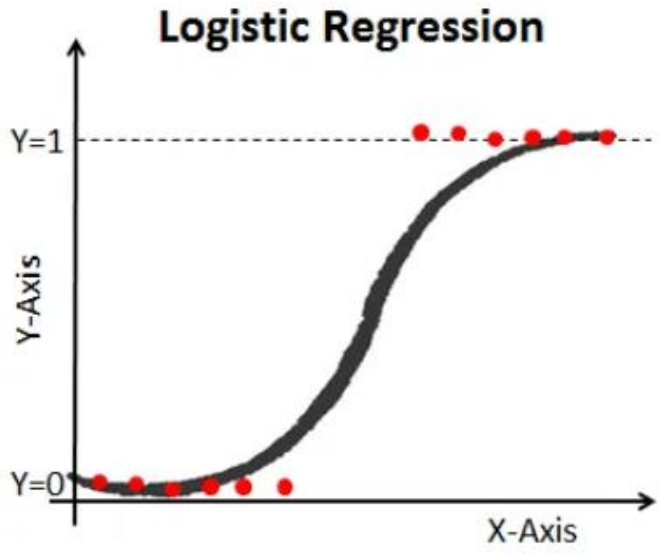
Η παράμετρος **random_state** είναι ένα προαιρετικό όρισμα που σας επιτρέπει να ορίσετε μια τιμή σπόρου για τη γεννήτρια τυχαίων αριθμών. Παρέχοντας μια συγκεκριμένη τιμή **random_state** (π.χ., **random_state = 50**), εξασφαλίζετε ότι τα δεδομένα χωρίζονται με αναπαραγωγικό τρόπο.

Η συνάρτηση train_test_split επιστρέφει τέσσερις ξεχωριστούς πίνακες: **X_train**, **X_test, Y_train** και **Y_test**. Τα **X_train** και **Y_train** αντιπροσωπεύουν τα δεδομένα εκπαίδευσης, ενώ τα **X_test** και **Y_test** αντιπροσωπεύουν τα δεδομένα δοκιμής.

3.9 Μοντέλο Μηχανικής μάθησης: Λογιστική Παλινδρόμηση (Logistic Regression)

3.9.1 Ορισμός της Λογιστικής Παλινδρόμησης

Η λογιστική παλινδρόμηση είναι μια στατιστική τεχνική που χρησιμοποιείται ευρέως στη μηχανική μάθηση και τη στατιστική, ιδίως για προβλήματα δυαδικής ταξινόμησης. Σε αντίθεση με τη γραμμική παλινδρόμηση **(linear regression)**, η οποία χρησιμοποιείται για την πρόβλεψη συνεχών αριθμητικών τιμών, η λογιστική παλινδρόμηση έχει σχεδιαστεί ειδικά για την πρόβλεψη δυαδικών αποτελεσμάτων, όπου η μεταβλητή-στόχος λαμβάνει μία από δύο πιθανές τιμές, που συχνά αναπαρίστανται ως 0 και 1.



Εικόνα 30. Γραφική αναπαράσταση της Λογιστικής Παλινδρόμησης

3.9.2 Λογιστική συνάρτηση (Sigmoid function)

Η σιγμοειδής συνάρτηση, που αναφέρεται επίσης ως λογιστική συνάρτηση, είναι μια μαθηματική καμπύλη που χαρακτηρίζεται από το σχήμα **S** ή τη σιγμοειδή μορφή της. Η συνάρτηση αυτή δέχεται οποιονδήποτε πραγματικό αριθμό ως είσοδο και επιστρέφει μια τιμή μεταξύ 0 και 1. Πλησιάζει το μηδέν καθώς η είσοδος γίνεται αρνητική και το ένα καθώς η είσοδος γίνεται θετική. Όταν η είσοδος είναι μηδέν, η σιγμοειδής συνάρτηση επιστρέφει 0,5. [40]

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Εικόνα 31. Μαθηματική Εξίσωση της Λογιστικής Συνάρτησης

Η λογιστική συνάρτηση εκφράζεται μαθηματικά με τον παραπάνω τύπο της εικόνας 26, όπου **e** είναι η βάση του φυσικού λογαρίθμου (περίπου 2.71828), **x** είναι η τιμή εισόδου, και το αποτέλεσμα, **σ(x)**, είναι μια τιμή μεταξύ 0 και 1. Αυτό σημαίνει ότι οποιαδήποτε τιμή εισόδου **x** υπολογίζεται έτσι ώστε το αποτέλεσμα να είναι πάντα μέσα σε αυτό το εύρος, το οποίο μπορούμε να ερμηνεύσουμε ως πιθανότητα. Όταν **x** είναι μεγάλο, το αποτέλεσμα πλησιάζει το 1, ενώ όταν **x** είναι μικρό ή αρνητικό, πλησιάζει το 0. [40]

Η συνάρτηση αυτή μας βοηθάει να καταλάβουμε πόσο πιθανό είναι ένα αποτέλεσμα, κάνοντάς το πιο εύκολο να ερμηνευτεί. Αν η τιμή που δίνει η συνάρτηση είναι κοντά στο 1, τότε το μοντέλο είναι αρκετά σίγουρο για τη θετική κατηγορία, ενώ αν είναι κοντά στο 0, είναι σίγουρο για την αρνητική. [40]

3.9.3 Εξίσωση της Λογιστικής Παλινδρόμησης

Η εξίσωση της λογιστικής παλινδρόμησης σχηματίζεται συνδυάζοντας τη λογιστική συνάρτηση με μια γραμμική εξίσωση. Η γραμμική εξίσωση αποτελείται από τα χαρακτηριστικά της εισόδου (τα δεδομένα) και τα βάρη (συντελεστές) του μοντέλου. Ο σκοπός είναι να μοντελοποιήσει την πιθανότητα ότι μια περίπτωση ανήκει σε μια συγκεκριμένη κατηγορία, συνήθως τη θετική κλάση. Με άλλα λόγια, η λογιστική συνάρτηση μετατρέπει την έξοδο της γραμμικής εξίσωσης, η οποία μπορεί να παίρνει οποιαδήποτε τιμή, σε μια πιθανότητα που κυμαίνεται από **0 έως 1**, η οποία αντιπροσωπεύει την πιθανότητα να ανήκει η περίπτωση στη θετική κλάση. [40]

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x$$

$$=> P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Εικόνα 32. Μαθηματική εξίσωση της λογιστικής Παλινδρόμησης [40]

Στην εικόνα που δείχνει τη λογιστική παλινδρόμηση, η εξίσωση εξηγεί πώς υπολογίζουμε την πιθανότητα μιας τιμής-στόχου (Y) να είναι 1, δεδομένων των εισόδων (X). Συγκεκριμένα, το P(Y=1|X) αναπαριστά την πιθανότητα το αποτέλεσμα να ανήκει στην θετική κλάση (Y=1), λαμβάνοντας υπόψη τα χαρακτηριστικά εισόδου, όπως X_1, X_2 κ.λπ. Τα β_0 και β_1 είναι οι συντελεστές που μαθαίνονται από τα δεδομένα κατά τη διαδικασία εκπαίδευσης και αντιστοιχούν στη γραμμική σχέση μεταξύ των χαρακτηριστικών εισόδου και του αποτελέσματος. [40]

3.9.4 Κατηγορίες Λογιστικής Παλινδρόμησης

Η λογιστική παλινδρόμηση ταξινομείται σε τρεις τύπους: δυαδική, τακτική και πολυωνυμική λογιστική παλινδρόμηση. Κάθε τύπος είναι προσαρμοσμένος σε συγκεκριμένα δεδομένα και ερευνητικά ερωτήματα, παρέχοντας στους ερευνητές ισχυρά εργαλεία για την προγνωστική μοντελοποίηση. Η δυαδική λογιστική παλινδρόμηση χρησιμοποιείται για δυαδικές μεταβλητές αποτελεσμάτων, η διατακτική λογιστική παλινδρόμηση για διατεταγμένα κατηγορικά αποτελέσματα και η πολυωνυμική λογιστική παλινδρόμηση για ονομαστικά αποτελέσματα με πολλές κατηγορίες. Η κατανόηση αυτών των κατηγοριών και των εφαρμογών τους είναι κρίσιμη στην ανάλυση δεδομένων. [40]

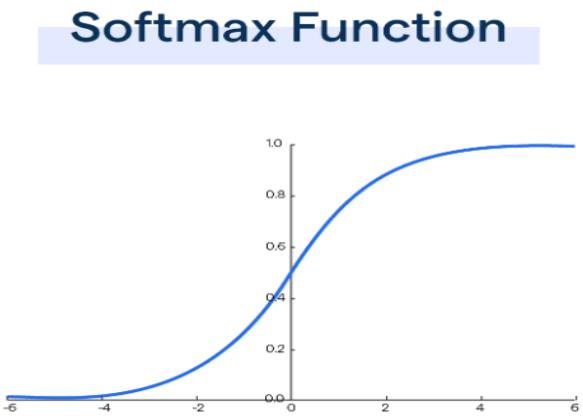
a) Η δυαδική λογιστική παλινδρόμηση **(Binary Logistic Regression)**, ο πιο συνηθισμένος από τους τρεις τύπους λογιστικής παλινδρόμησης, χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι δυαδική. Μπορεί να εξετάσει μόνο δύο αποτελέσματα. Για παράδειγμα, αυτή η μέθοδος μπορεί να καθορίσει αν ένα μήνυμα ηλεκτρονικού ταχυδρομείου είναι spam ή όχι, ή αν ένας όγκος είναι κακοήθης ή καλοήθης. Αυτός ο τύπος λογιστικής παλινδρόμησης είναι ένα αποτελεσματικό εργαλείο σε διάφορους τομείς, συμπεριλαμβανομένης της ιατρικής έρευνας, του μάρκετινγκ και των κοινωνικών επιστημών. [40]
b) Ο δεύτερος τύπος λογιστικής παλινδρόμησης, Τακτική λογιστική παλινδρόμηση **(Ordinal Logistic Regression)**, χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι ταξινομημένη. Μια τακτική μεταβλητή μπορεί να είναι λογικά ταξινομημένη, αλλά τα διαστήματα μεταξύ των τιμών δεν είναι πάντα ομοιόμορφα. Παραδείγματα περιλαμβάνουν την πρόβλεψη των επιπέδων ικανοποίησης των πελατών (πολύ δυσαρεστημένοι, δυσαρεστημένοι, ουδέτεροι, ικανοποιημένοι, ευχαριστημένοι). Αυτός ο τύπος παλινδρόμησης αποδίδει πιο διαφοροποιημένα αποτελέσματα και είναι χρήσιμος σε τομείς όπως η έρευνα αγοράς και ο έλεγχος ποιότητας. [40]
c) Η πολυωνυμική λογιστική παλινδρόμηση **(Multinomial Logistic Regression)** είναι ο τρίτος τύπος λογιστικής παλινδρόμησης. Χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι ονομαστική και έχει περισσότερα από δύο επίπεδα χωρίς σειρά ή προτεραιότητα. Για παράδειγμα, η πολυωνυμική λογιστική παλινδρόμηση θα μπορούσε να χρησιμοποιηθεί για να προβλεφθεί αν κάποιος θα αγοράσει ένα SUV, ένα sedan ή ένα hatchback. Αυτή η τεχνική παλινδρόμησης είναι χρήσιμη σε μια ποικιλία εφαρμογών, όπως η ανάλυση μάρκετινγκ και οι κοινωνικές επιστήμες. [40]

Για τον σκοπό της εργασίας μας, θα εφαρμόσουμε την τρίτη κατηγορία της λογιστικής παλινδρόμησης την Πολυωνυμική, διότι στην προκειμένη περίπτωση έχουμε πάνω απο 2 κλάσεις συναισθήματος. Οι τρεις κλάσεις που θα εξετάσουμε είναι θετικό, αρνητικό, και

ουδέτερο συναίσθημα. Κατά τη διαδικασία εύρεσης ενός αποδοτικού αλγορίθμου μηχανικής μάθησης, δοκιμάσαμε και την επέκταση της δυαδικής κατηγοριοποίησης, γνωστή ως **One-vs-Rest (OvR),** η οποία είναι και αυτή κατάλληλη όταν έχουμε περισσότερες από δύο κλάσεις. Η προσέγγιση αυτή δημιουργεί ένα ξεχωριστό μοντέλο για κάθε μία από τις τρεις κλάσεις. Για παράδειγμα, για την κλάση του θετικού συναισθήματος, το μοντέλο θα εκπαιδευτεί ώστε να διακρίνει το θετικό από το αρνητικό και το ουδέτερο, και το ίδιο θα γίνει και για τις άλλες δύο κλάσεις. Έτσι, το σύστημα μπορεί να ταξινομεί σωστά τα δεδομένα σε μία από τις τρεις αυτές κατηγορίες, χρησιμοποιώντας πολλαπλά μοντέλα δυαδικής ταξινόμησης. Βάσει των δύο προσεγγίσεων, η πολυωνυμική λογιστική παλινδρόμηση υπερτερεί της One-vs-Rest (OvR), όπως προκύπτει από τη σύγκριση των μετρήσεων ακρίβειας, καθώς και από δημοσιευμένη μελέτη του Rukshan Pramoditha, αναφέροντας την πολυωνυμική παλινδρόμηση, ως την πιο αξιόπιστη. [41]

3.9.5 Συνάρτηση Softmax

Η συνάρτηση Softmax χρησιμοποιείται για τη γενίκευση της λογιστικής παλινδρόμησης ώστε να υποστηρίζει πολλαπλές κλάσεις. Η συνάρτηση αυτή δέχεται ένα διάνυσμα εισόδου (z = [z1,z2,..., zk]) και παράγει ένα άλλο διάνυσμα εξόδου απο κατανομές πιθανοτήτων. [43]



Εικόνα 33. Γραφική παράσταση της softmax συνάρτησης [45]

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Εικόνα 34. Μαθηματική εξίσωση της συνάρτησης softmax [43]

Στην Εικόνα 31 βλέπουμε την μαθηματική εξίσωση της συνάρτησης Softmax όπου η μεταβλητή **z** αναφέρεται στο διάνυσμα εισόδου συγκεκριμένων δεδομένων, η μεταβλητή **e^(zi)** αναφέρεται στην εκθετική συνάρτηση του διανύσματος εισόδου, η μεταβλητή **e^(zj)**

αναφέρεται στην εκθετική συνάρτηση του διανύσματος εξόδου και η μεταβλητή **k** στο πλήθος των κλάσεων.

3.9.6 Εξίσωση της Πολυωνυμικής Παλινδρόμησης

Συνδιάζοντας την λογιστική παλινδρόμηση και την συνάρτηση softmax, προκυπτει η μαθηματική εξίσωση της πολυωνυμικής παλινδρόμησης. Όταν εφαρμόζουμε το softmax στη λογιστική παλινδρόμηση, οι είσοδοι θα είναι το τετραγωνικό γινόμενο του διανύσματος βάρους **(w)** και του διανύσματος εισόδου **(x)** συν έναν όρο μεροληψίας **(b)**, οπότε ο συνολικός όρος είναι **w.x + b**. [43]

$$p(y_k = 1|\mathbf{x}) = \frac{e^{(\mathbf{w}_k\mathbf{x}+\mathbf{b}_k)}}{\sum_{j=1}^{K} e^{(\mathbf{w}_j\mathbf{x}+\mathbf{b}_j)}}$$

Εικόνα 35. *Μαθηματική εξίσωσης της πολυωνυμικής λογιστικής συνάρτησης [43]*

3.9.7 Συνάρτηση απώλειας (cost function)

Στόχος της εκπαίδευσης είναι η δημιουργία ενός μοντέλου που προβλέπει υψηλή πιθανότητα για την κλάση-στόχο αλλά χαμηλή πιθανότητα για τις άλλες κλάσεις. Έχοντας αυτόν τον στόχο κατά νου, θα ελαχιστοποιήσουμε τη συνάρτηση κόστους, γνωστή και ως διασταυρούμενη εντροπία **(cross entropy)**. Η συνάρτηση κόστους cross τιμωρεί το μοντέλο όταν εκτιμά χαμηλή πιθανότητα για μια κλάση-στόχο. Η διασταυρούμενη εντροπία είναι ένα μέτρο του πόσο καλά ένα σύνολο εκτιμώμενων πιθανοτήτων κλάσεων ταιριάζει με τις κλάσεις-στόχους. [44]

$$J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)} \log\left(\hat{p}_k^{(i)}\right)$$

Εικόνα 36. Συνάρτηση Απώλειας [44]

Στην παραπάνω εικόνα η πιθανότητα-στόχος **yk(i)**, καθορίζει αν η περίπτωση i ανήκει στην κατηγορία k ή όχι. Η πιθανότητα θα είναι είτε 1 είτε 0, ανάλογα με το σε ποια κλάση ανήκει η περίπτωση. [44]

3.9.8 Υλοποίηση Μοντέλου Μηχανικής Μάθησης σε κώδικα

```
lg = LogisticRegression(multi_class='multinomial', solver = 'saga')
lg.fit(X_train, Y_train)
```

Εικόνα 37. Προπόνηση και τεστάρισμα μοντέλου Μηχανικής Μάθησης

Στον παραπάνω κώδικα δημιουργούμε ένα αντικείμενο της κλάσης logistic regression και επιλέγουμε την επιλογή **multinomial** για το λόγο που αναφέραμε παραπάνω. Ύστερα εκπαιδεύουμε το μοντέλο με τα δεδομένα εκπαίδευσης **X_train** και **Y_train.**

3.10 Μοντέλο βασισμένο σε κανόνες

3.10.1 Αλγόριθμος VADER (Valence Aware Dictionary and sEntiment Reasoner)

Οι Hutto και Gilbert [46] ανέπτυξαν το VADER, έναν αλγόριθμο βασισμένο σε λεξικό, το 2014 για να λύσουν το πρόβλημα της ανάλυσης της γλώσσας, των συμβόλων και του ύφους του κειμένου στην ανάλυση συναισθήματος. Χρησιμοποιείται ευρέως για μια ποικιλία εφαρμογών, όπως η παρακολούθηση των μέσων κοινωνικής δικτύωσης, η ανάλυση των ανατροφοδοτήσεων των πελατών και η διαχείριση της φήμης των εμπορικών σημάτων. Κάνει χρήση ενός προκατασκευασμένου λεξικού που περιλαμβάνει λέξεις που έχουν βαθμολογηθεί για το συναίσθημα, καθώς και γραμματικούς και συντακτικούς κανόνες για τις αρνήσεις, τους εντατικούς και τους τροποποιητές. [47]

Το VADER χρησιμοποιεί ένα προκατασκευασμένο λεξικό λέξεων ή φράσεων με βαθμολογίες συναισθήματος που κυμαίνονται από -1 (εξαιρετικά αρνητικό) έως +1 (εξαιρετικά θετικό). Το λεξικό περιέχει επίσης λέξεις με ουδέτερη βαθμολογία συναισθήματος. Οι βαθμολογίες συναισθήματος του VADER βασίζονται σε ανθρώπινες σημειώσεις και λαμβάνουν υπόψη την ένταση του συναισθήματος που σχετίζεται με κάθε λέξη. Η κύρια έξοδος του VADER είναι μια βαθμολογία πολικότητας συναισθήματος, η οποία αντιπροσωπεύει το συνολικό συναίσθημα που εκφράζεται στο κείμενο. Η βαθμολογία είναι μια συνεχής τιμή που κυμαίνεται από -1 έως +1, με αρνητικές τιμές που υποδηλώνουν αρνητικό συναίσθημα, θετικές τιμές που υποδηλώνουν θετικό συναίσθημα και τιμές κοντά στο μηδέν που υποδηλώνουν ουδέτερο συναίσθημα. [46]

3.10.2 Υλοποίηση Μοντέλου βασισμένο σε Κανόνες

```
analyzer = SentimentIntensityAnalyzer()
# function to calculate vader sentiment
def vadersentimentanalysis(review):
    vs = analyzer.polarity_scores(review)
    return vs['compound']
df['Vader Sentiment'] = df['Lemma'].apply(vadersentimentanalysis)
```

*Εικόνα 38. Συνάρτηση υπολογισμού συναισθήματος με την χρήση του αλγορίθμου **VADER***

Η συνάρτηση **vadersentimentanalysis** υπολογίζει το συνολικό συναίσθημα ενός δεδομένου κειμένου χρησιμοποιώντας το εργαλείο ανάλυσης συναισθήματος VADER. Λαμβάνει ένα κομμάτι κειμένου (π.χ. μια κριτική ή ένα άρθρο ειδήσεων) ως είσοδο, εφαρμόζει τη μέθοδο **polarity_scores()** του Vader για τη δημιουργία βαθμολογίας συναισθήματος και επιστρέφει τη σύνθετη βαθμολογία, η οποία είναι μια κανονικοποιημένη τιμή μεταξύ -1 (πιο αρνητική) και +1 (πιο θετική). Στη συνέχεια, η συνάρτηση εφαρμόζεται σε κάθε στοιχείο στήλης **Description** και οι βαθμολογίες συναισθήματος που προκύπτουν αποθηκεύονται σε μια νέα στήλη **'Vader Sentiment'**, συνοψίζοντας το συνολικό συναίσθημα του κειμένου στο σύνολο δεδομένων.

```
# function to analyse
def vader_analysis(compound):
    if compound >= 0.5:
        return 'Positive'
    elif compound <= -0.5 :
        return 'Negative'
    else:
        return 'Neutral'
df['Vader Analysis'] = df['Vader Sentiment'].apply(vader_analysis)
```

Εικόνα 39. Συνάρτηση προσδιορισμού συναισθήματος

Η συνάρτηση **vader_analysis** ταξινομεί το συναίσθημα ενός κειμένου με βάση τη σύνθετη βαθμολογία του. Εάν η σύνθετη βαθμολογία είναι μεγαλύτερη ή ίση με 0,5, ταξινομεί το συναίσθημα ως «θετικό»- εάν η βαθμολογία είναι μικρότερη ή ίση με -0,5, το ταξινομεί ως «αρνητικό»- και για σύνθετες βαθμολογίες μεταξύ -0,5 και 0,5, το συναίσθημα ταξινομείται ως «ουδέτερο». Η συνάρτηση αυτή εφαρμόζεται στη συνέχεια στη στήλη **Vader Sentiment** και το αποτέλεσμα της ταξινόμησης του συναισθήματος αποθηκεύεται σε μια νέα στήλη, **Vader Analysis**, που συνοψίζει το συναίσθημα ως θετικό, αρνητικό ή ουδέτερο.

3.11 Αξιολόγηση Μοντέλων

3.11.1 Τι είναι η αξιολόγηση Μοντέλου

Η αξιολόγηση μοντέλου ειναι η διαδικασία εκτίμησης αποτελεσματικότητας και τις ακρίβειας ενός συστήματος στην εκτέλεση μια συγκεκριμένης διαδικασίας, όπως ειναι η κατηγοριοποίηση ή η πρόβλεψη. Για την ανάπτυξη των μοντέλων, οι αξιολογήσεις βασίστηκαν σε διάφορες μετρικές, προκειμένου να εκτιμηθεί η απόδοση του μοντέλου μηχανικής μάθησης σε δεδομένα που δεν είχε χρησιμοποιήσει κατά την διαδικασία της εκπαίδευσής του και το μοντέλο κανόνων στην αποτελεσματικότητα και την ευστοχία των προκαθορισμένων κανόνων του. Στόχος της αξιολόγησης είναι να προσδιοριστεί ποσό καλά μπορούν τα μοντέλα να αντιμετωπίσουν τα προβλήματα για τα οποία σχεδιάστηκαν, καθώς να εντοπιστούν αδυναμίες και περιθώρια βελτίωσης.

3.11.2 Μετρικές αξιολόγησης

Οι μετρικές Αξιολόγησης ειναι ποσοτικά μέτρα που χρησιμοποιούνται για την αξιολόγηση της απόδοσης μοντέλων μηχανικής μάθησης αλλά και κανόνων. Αυτές οι μετρικές παρέχουν πληροφορίες σχετικά με το πόσο καλά αποδίδει ένα μοντέλο και επίσης βοηθούν και στην σύγκριση διαφορετικών μοντέλων ή αλγορίθμων. Η επιλογή της μετρικής αξιολόγησης εξαρτάται πλήρως από τον τύπο του μοντέλου και το σχέδιο εφαρμογής του μοντέλου. [48]

3.11.3 Τύποι μετρικών αξιολόγησης

Οι μετρικές αξιολόγησης διαφέρουν ανάλογα με το τύπο του προβλήματος που έχουμε να αντιμετωπίσουμε. Υπάρχουν 2 κατηγορίες τύπων μετρικών αξιολόγησης, οπού η πρώτη έχει να κάνει με την ταξινόμηση και η δεύτερη με την παλινδρόμηση. Παρακάτω θα αναφερθούμε συνοπτικά για τις 2 αυτές κατηγορίες.

a) Οι μετρικές ταξινόμησης είναι μια συλλογή μετρικών που χρησιμοποιούνται για την αξιολόγηση της απόδοσης των μοντέλων ταξινόμησης. Αυτές οι μετρικές αξιολογούν την ακρίβεια του μοντέλου, την ακρίβεια, την ανάκληση και άλλους παράγοντες. Χρησιμοποιείται συχνά για τη σύγκριση διαφορετικών μοντέλων ή για τη βελτιστοποίηση ενός μεμονωμένου μοντέλου για μέγιστη απόδοση. Οι μετρικές ταξινόμησης μπορούν να χωριστούν σε τρεις κατηγορίες: ακρίβεια, ευαισθησία και ειδικότητα. Η ακρίβεια, η οποία μετρά τη συνολική απόδοση του μοντέλου, είναι συνήθως η πιο σημαντική μετρική. Η ευαισθησία και η εξειδίκευση δείχνουν πόσο καλά ένα μοντέλο διακρίνει μεταξύ των κλάσεων. Τέλος, άλλες μετρικές όπως η **AUC**, η **F1** αξιολογούν την ακρίβεια και την αναγνώριση του μοντέλου. [49]

b) Οι μετρικές παλινδρόμησης είναι απαραίτητες για την αξιολόγηση της απόδοσης των μοντέλων παλινδρόμησης ειδικότερα. Οι μετρικές αυτές βοηθούν στην αξιολόγηση της ικανότητας ενός μοντέλου παλινδρόμησης να προβλέπει συνεχή αποτελέσματα. Οι συνήθεις μετρικές αξιολόγησης της παλινδρόμησης περιλαμβάνουν το μέσο απόλυτο σφάλμα **(MAE)**, το μέσο τετραγωνικό σφάλμα **(MSE)**, τη ρίζα του μέσου τετραγωνικού σφάλματος **(RMSE)**, ο συντελεστής προσδιορισμού **(R-squared)** και το μέσο απόλυτο ποσοστιαίο σφάλμα **(MAPE**). Οι επιστήμονες δεδομένων και οι μηχανικοί μηχανικής μάθησης μπορούν να χρησιμοποιήσουν αυτές τις ειδικές για την παλινδρόμηση μετρικές για να αξιολογήσουν την ακρίβεια και την αποτελεσματικότητα των προβλέψεων των μοντέλων παλινδρόμησης. [50]

Το πρόβλημα που καλούμαστε να λύσουμε είναι ένα πρόβλημα ταξινόμησης ειδήσεων, οπότε για την αξιολόγηση των μοντέλων μας θα χρησιμοποιήσουμε μετρικές της ταξινόμησης.

3.11.4 Μετρικές Ταξινόμησης

Ο πίνακας σύγχυσης **(confusion matrix)**, επίσης γνωστός ως πίνακας σφάλματος ή πίνακας ενδεχομένων, είναι ένα βασικό εργαλείο για την αξιολόγηση της απόδοσης των αλγορίθμων ταξινόμησης. Ο πίνακας σύγχυσης είναι ένας πίνακας που αθροίζει τα αποτελέσματα ταξινόμησης ενός δυαδικού ταξινομητή. Παρέχει λεπτομερή ανάλυση της απόδοσης ταξινόμησης ενός μοντέλου, συμπεριλαμβανομένου του αριθμού των αληθώς θετικών **(TP)**, των αληθώς αρνητικών **(TN)**, των ψευδώς θετικών **(FP)** και των ψευδώς αρνητικών **(FN)**. [51]

Τις περισσότερες φορές ο πίνακας αυτός ειναι 2χ2 αλλά στην δικιά μας περίπτωση έχουμε πάνω απο 2 κλάσεις συναισθημάτων οπότε ο πίνακας μας θα ειναι 3χ3. Παρακάτω θα αναλύσουμε τα στοιχεία απο τα οποία αποτελείται ο πίνακας σύγχυσης.

1. True Positive **(TP)**: Ο αριθμός των θετικών περιπτώσεων που ταξινομούνται σωστά ως θετικές από το μοντέλο.
2. True Negative **(TN)**: Ο αριθμός των αρνητικών περιπτώσεων που ταξινομούνται σωστά ως αρνητικές από το μοντέλο.
3. True Neutral **(TNt):** Ο αριθμός των ουδέτερων περιπτώσεων που ταξινομούνται σωστά ως ουδέτερες από το μοντέλο.
4. False Positive **(FP)**: Ο αριθμός των αρνητικών περιπτώσεων που ταξινομούνται εσφαλμένα ως θετικές από το μοντέλο.

5. False Negative **(FN)**: Ο αριθμός των θετικών περιπτώσεων που ταξινομούνται εσφαλμένα ως αρνητικές από το μοντέλο.
6. False Neutral **(FNt):** ο αριθμός των ουδέτερων περιπτώσεων που ταξινομούνται εσφαλμένα ως αρνητικές ή θετικές απο το μοντέλο.

Η ακρίβεια είναι η πιο συχνά χρησιμοποιούμενη μετρική απόδοσης για την αξιολόγηση μοντέλων δυαδικής και πολλαπλής ταξινόμησης. Υπολογίζει το ποσοστό των σωστών προβλέψεων που πραγματοποιεί το μοντέλο επί του συνόλου των προβλέψεων. Μια υψηλή βαθμολογία ακρίβειας υποδηλώνει ότι το μοντέλο πραγματοποιεί μεγάλο ποσοστό σωστών προβλέψεων, ενώ μια χαμηλή βαθμολογία ακρίβειας υποδηλώνει ότι το μοντέλο πραγματοποιεί υπερβολικά μεγάλο αριθμό εσφαλμένων προβλέψεων. Η ακρίβεια υπολογίζεται με τον ακόλουθο τύπο. [51]

$$\boldsymbol{Accuracy = \frac{TP+TN+TNt}{TP + TN+TNt+ FP + FN+FNt}} \quad \textbf{(1)}$$

Ο όρος Precision είναι μια μετρική που καθορίζει το ποσοστό των αληθώς θετικών περιπτώσεων μεταξύ εκείνων που προβλέπονται ως θετικές από το μοντέλο. Αξιολογεί τις θετικές προβλέψεις του μοντέλου ως προς την ακρίβεια. Αντίστοιχα, η ακρίβεια στις αρνητικές και ουδέτερες προβλέψεις μετράει με πόση ακρίβεια το μοντέλο εντοπίζει αληθινά αρνητικές και αληθινά ουδέτερες περιπτώσεις, αντίστοιχα. Μια υψηλή βαθμολογία ακρίβειας για οποιαδήποτε κλάση (θετική, αρνητική ή ουδέτερη) υποδηλώνει ότι το μοντέλο μπορεί να αναγνωρίσει σωστά τις περιπτώσεις αυτής της κλάσης, ενώ παράγει ελάχιστα ψευδώς θετικά αποτελέσματα (FP) από άλλες κλάσεις. Αντίθετα, ένα χαμηλό σκορ ακρίβειας για οποιαδήποτε κλάση υποδηλώνει ότι το μοντέλο κάνει πάρα πολλές λανθασμένες προβλέψεις από τις άλλες κλάσεις, με αποτέλεσμα μεγάλο αριθμό λανθασμένων θετικών αποτελεσμάτων για τη συγκεκριμένη.[51]

$$\boldsymbol{Precision_{Positive} = \frac{TP}{TP+FP}} \quad \textbf{(2)}$$

$$\boldsymbol{Precision_{Negative} = \frac{TN}{TN+FN}} \quad \textbf{(3)}$$

$$\boldsymbol{Precision_{Neutral} = \frac{TNt}{TNt+FNt}} \quad \textbf{(4)}$$

Η ανάκληση **(recall)** είναι μια μετρική απόδοσης που μετρά το ποσοστό των περιπτώσεων μιας δεδομένης κλάσης (θετικής, αρνητικής ή ουδέτερης) που αναγνωρίζονται σωστά από ένα μοντέλο ταξινόμησης σε σχέση με όλες τις πραγματικές περιπτώσεις της κλάσης αυτής. Αποτελεί σημαντική μετρική για την αξιολόγηση της ικανότητας του μοντέλου να συλλαμβάνει περιπτώσεις σε διάφορες κλάσεις και χρησιμοποιείται συχνά σε συνδυασμό με άλλες μετρικές όπως η ακρίβεια, το F1-score και η ακρίβεια. [51]

Στην περίπτωση των θετικών προβλέψεων, η ανάκληση - επίσης γνωστή ως ευαισθησία ή ποσοστό αληθώς θετικών περιπτώσεων **(TPR)** - μετρά το ποσοστό των αληθώς θετικών περιπτώσεων **(TP)** που αναγνωρίζονται σωστά από το μοντέλο επί όλων των πραγματικών θετικών περιπτώσεων. Ομοίως, η ανάκληση για τις αρνητικές προβλέψεις μετρά το ποσοστό

των αληθώς αρνητικών **(TN)** περιπτώσεων που αναγνωρίζονται σωστά από το μοντέλο σε όλες τις πραγματικές αρνητικές περιπτώσεις. Για τις ουδέτερες προβλέψεις, η ανάκληση αξιολογεί το ποσοστό των αληθώς ουδέτερων **(TNt)** περιπτώσεων που αναγνωρίστηκαν σωστά από όλες τις πραγματικές ουδέτερες περιπτώσεις.

Ένα υψηλό σκορ ανάκλησης **(recall)** για οποιαδήποτε κλάση, είτε θετική, είτε αρνητική, είτε ουδέτερη, υποδηλώνει ότι το μοντέλο μπορεί να αναγνωρίσει ένα μεγάλο ποσοστό των πραγματικών περιπτώσεων της συγκεκριμένης κλάσης. Από την άλλη πλευρά, ένα χαμηλό σκορ ανάκλησης υποδηλώνει ότι το μοντέλο χάνει πολλές περιπτώσεις, με αποτέλεσμα υψηλό αριθμό ψευδώς αρνητικών **(FN)**. Η υψηλή ανάκληση είναι ιδιαίτερα σημαντική σε καταστάσεις όπου είναι ζωτικής σημασίας να μην χάνονται σημαντικές περιπτώσεις, όπως ο εντοπισμός θετικών ή αρνητικών συναισθημάτων σε μια εργασία ανάλυσης συναισθήματος. [51]

$$\boldsymbol{Recall_{Positive} = \frac{TP}{TP+FN}} \quad \textbf{(5)}$$

$$\boldsymbol{Recall_{Negative} = \frac{TN}{TN+FN}} \quad \textbf{(6)}$$

$$\boldsymbol{Recall_{Neutral} = \frac{TNt}{TNt+FN}} \quad \textbf{(7)}$$

Το **F1-score** είναι μια μετρική απόδοσης που συνδυάζει την ακρίβεια και την ανάκληση για να παρέχει μια ολοκληρωμένη αξιολόγηση της απόδοσης ενός δυαδικού μοντέλου ταξινόμησης. Υπολογίζει τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης, αποδίδοντας ίση βαρύτητα και στις δύο μετρικές. [51]

Το F1-score είναι μια μετρική απόδοσης που συνδυάζει την ακρίβεια και την ανάκληση για να παρέχει μια πλήρη αξιολόγηση ενός μοντέλου δυαδικής ταξινόμησης. Υπολογίζει τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης, αποδίδοντας ίση βαρύτητα και στις δύο μετρικές. Ένα υψηλό F1-score υποδεικνύει ότι το μοντέλο αποδίδει καλά τόσο στην ακρίβεια όσο και στην ανάκληση, ενώ ένα χαμηλό F1-score υποδεικνύει ότι το μοντέλο αποδίδει άσχημα σε ένα από τα δύο ή και στα δύο. [51]

$$\boldsymbol{F1-score = \frac{2*(precision*recall)}{precision*recall}} \quad \textbf{(8)}$$

3.12 Μηχανές Συστάσεων (Recommend Engines)

3.12.1 Μηχανή σύστασης Μοντέλου Μηχανικής Μάθησης

```
def recommend_articles(title):
    article2idx = pd.Series(df.index,index=df['Title'])
    # Get the index of the article with the given title
    article_index = article2idx[title]
    if type(article_index) == pd.Series:
        article_index = article_index.iloc[0]

    # Get the category of the query article
    query_category = df['Category'].iloc[article_index]

    # Predict the sentiment of the query article
    query_sentiment = lg.predict(vectorizer.transform([df['Description'].iloc[article_index]]))#[0]

    # Calculate sentiment predictions for all articles
    sentiment_predictions = lg.predict(vectorizer.transform(df['Description']))

    # Filter articles by sentiment and category
    recommended_articles = []
    for idx in range(len(df)):
        if (sentiment_predictions[idx] == query_sentiment) and (df['Category'].iloc[idx] == query_category) and (idx != article_index):
            recommended_articles.append(df['Title'].iloc[idx])

    # Limit to top 10 recommended articles
    top_recommended_articles = recommended_articles[:10]

    # Display recommended articles
    display(HTML(f"<h3>Recommended Articles for {query_sentiment} sentiment:</h3>"))
    if top_recommended_articles:
        for article in top_recommended_articles:
            display(HTML(f"<p>{article}</p>"))
    else:
        display(HTML("<p>No recommended articles found.</p>"))
```

Εικόνα 40. Αλγόριθμος Σύστασης Ειδησεων με την χρηση Μηχανικής Μάθησης

Η συνάρτηση **recommend_articles** είναι υπεύθυνη για την πρόταση άρθρων με βάση έναν συγκεκριμένο τίτλο που παρέχεται ως είσοδος. Ο βασικός στόχος της συνάρτησης είναι να αναλύσει το συναίσθημα και την κατηγορία του αρχικού άρθρου και στη συνέχεια να βρει και να προτείνει άλλα άρθρα που ταιριάζουν σε αυτά τα χαρακτηριστικά.

Αρχικά, η συνάρτηση δημιουργεί έναν πίνακα που αντιστοιχεί τους τίτλους των άρθρων με τους δείκτες τους στον πίνακα δεδομένων **df**. Αυτό επιτρέπει τη γρήγορη αναζήτηση του άρθρου με βάση τον τίτλο που παρέχεται. Αφού βρεθεί ο δείκτης του άρθρου, η συνάρτηση εξάγει την κατηγορία του άρθρου από τη στήλη Category και τη χρησιμοποιεί αργότερα για να φιλτράρει τα άρθρα που θα προτείνει.

Στη συνέχεια, χρησιμοποιώντας ένα προεκπαιδευμένο μοντέλο λογιστικής παλινδρόμησης **(lg)**, η συνάρτηση προβλέπει το συναίσθημα του αρχικού άρθρου, βασισμένη στην περιγραφή του από τη στήλη **Description**. Το μοντέλο χρησιμοποιείται επίσης για να προβλέψει το συναίσθημα όλων των υπόλοιπων άρθρων στη βάση δεδομένων.

Μετά από αυτήν την ανάλυση, η συνάρτηση φιλτράρει όλα τα άρθρα για να βρει εκείνα που έχουν το ίδιο συναίσθημα και ανήκουν στην ίδια κατηγορία με το αρχικό άρθρο. Επιπλέον, εξασφαλίζεται ότι το αρχικό άρθρο δεν περιλαμβάνεται στα προτεινόμενα. Από τα άρθρα που πληρούν αυτά τα κριτήρια, επιλέγονται τα 10 πρώτα ως προτάσεις.

Στο τέλος, η συνάρτηση εμφανίζει τα προτεινόμενα άρθρα σε μορφή **HTML**. Εάν δεν βρεθούν κατάλληλες προτάσεις, η συνάρτηση εμφανίζει ένα μήνυμα που ενημερώνει ότι δεν υπάρχουν διαθέσιμα άρθρα για πρόταση. Η συνολική διαδικασία επικεντρώνεται στην παροχή

προτάσεων που είναι σχετικές τόσο από άποψη συναισθήματος όσο και κατηγορίας, διασφαλίζοντας έτσι μια πιο στοχευμένη εμπειρία για τον χρήστη.

3.12.2 **Μηχανη Σύστασης μοντέλου βασισμένο σε Κανόνες**

```
def recommend(title):
    # Get the index of the input news article based on its title
    movie2idx = pd.Series(df.index, index=df['Title'])
    if title not in movie2idx:
        print(f"Article titled '{title}' not found in the dataset.")
        return

    idx = movie2idx[title]

    if type(idx) == pd.Series:
        idx = idx.iloc[0]

    # Get the sentiment and category of the input article
    input_sentiment = df.loc[idx, 'Vader Analysis']
    input_category = df.loc[idx, 'Type']

    # Filter articles that match the sentiment and category
    recommended_articles = df[(df['Vader Analysis'] == input_sentiment) & (df['Type'] == input_category)]

    # If there are no matching articles, return a message
    if recommended_articles.empty:
        print(f"No articles found with the sentiment '{input_sentiment}' and category '{input_category}'.")
        return

    # Remove the input article from the recommendations
    recommended_articles = recommended_articles[recommended_articles.index != idx]

    # Limit to top 10 recommendations
    recommended_titles = recommended_articles['Title'].head(10)

    # Create HTML output for display
    html_output = "<h3>Recommended Articles:</h3>"
    html_output += "<ul>"
    for i, title in enumerate(recommended_titles):
        html_output += f"<li>{i+1}. {title}</li>"
    html_output += "</ul>"

    # Display the recommendations
    display(HTML(html_output))
```

Εικόνα 41. Αλγόριθμος Σύστασης με χρήση μοντέλου βασισμένο σε Κανόνες

Η συνάρτηση **recommend** προτείνει άρθρα με βάση τον τίτλο που παρέχεται ως είσοδος και συγκρίνει το συναίσθημα και την κατηγορία του άρθρου με αυτά των άλλων άρθρων στη βάση δεδομένων. Ακολουθεί μια περιγραφή βήμα-βήμα με παραγράφους:

Αρχικά, η συνάρτηση βρίσκει τον δείκτη του άρθρου με βάση τον τίτλο που παρέχεται από τον χρήστη. Αυτό επιτυγχάνεται δημιουργώντας έναν πίνακα που αντιστοιχεί τους τίτλους των άρθρων με τους δείκτες τους στο σύνολο δεδομένων **df**. Αν ο τίτλος δεν υπάρχει στο dataset, η συνάρτηση εμφανίζει ένα μήνυμα λάθους που ενημερώνει τον χρήστη ότι το άρθρο δεν βρέθηκε, και διακόπτει την εκτέλεση.

Μόλις βρεθεί ο δείκτης του άρθρου, η συνάρτηση ελέγχει αν υπάρχει κάποια διπλή καταχώρηση για τον συγκεκριμένο τίτλο. Αν υπάρχουν πολλαπλές εγγραφές με τον ίδιο τίτλο, χρησιμοποιείται η πρώτη από αυτές. Στη συνέχεια, η συνάρτηση εξάγει το συναίσθημα του άρθρου από την στήλη **Vader Analysis** και την κατηγορία του από την στήλη **Type**.

Με βάση το συναίσθημα και την κατηγορία του αρχικού άρθρου, η συνάρτηση φιλτράρει τα υπόλοιπα άρθρα για να βρει εκείνα που έχουν το ίδιο συναίσθημα και ανήκουν στην ίδια κατηγορία. Αν δεν υπάρχουν άρθρα που να ταιριάζουν με αυτά τα κριτήρια, η συνάρτηση ενημερώνει τον χρήστη ότι δεν βρέθηκαν σχετικά άρθρα και διακόπτει την εκτέλεση.

Αν υπάρχουν άρθρα που ταιριάζουν, η συνάρτηση αφαιρεί το αρχικό άρθρο από τις προτεινόμενες επιλογές για να μην εμφανίζεται στην ίδια λίστα. Στη συνέχεια, επιλέγονται μέχρι 10 άρθρα από τις σχετικές προτάσεις.

Τέλος, η συνάρτηση δημιουργεί μια **HTML** αναπαράσταση των προτεινόμενων άρθρων. Η λίστα με τους τίτλους των προτεινόμενων άρθρων εμφανίζεται με τη σειρά τους, και κάθε πρόταση αριθμείται. Αυτή η λίστα εμφανίζεται στον χρήστη με τη χρήση της βιβλιοθήκης **HTML**, προσφέροντας έτσι μια φιλική και καλαίσθητη εμφάνιση των προτεινόμενων άρθρων.

ΣΥΜΠΕΡΑΣΜΑ

Συμπερασματικά, η συγκριτική μελέτη των δύο μοντέλων για την ανάπτυξη ενός συστήματος σύστασης ειδήσεων έδειξε ότι η χρήση της μηχανικής μάθησης προσφέρει μεγαλύτερη ακρίβεια στην ανάλυση συναισθήματος και στις προτάσεις, σε σύγκριση με το μοντέλο που βασίζεται σε κανόνες. Ωστόσο, και τα δύο μοντέλα έχουν πλεονεκτήματα, καθώς το μοντέλο με κανόνες είναι πιο απλό στην εφαρμογή και εξηγησιμό, ενώ το μοντέλο μηχανικής μάθησης μπορεί να ενσωματώσει πιο σύνθετα μοτίβα και να προσαρμόζεται καλύτερα σε διαφορετικά δεδομένα. Τέλος, η χρήση φυσικής επεξεργασίας γλώσσας (NLP) αποδεικνύεται κρίσιμη για τη βελτίωση της ακρίβειας και της ποιότητας των εξατομικευμένων προτάσεων, προσφέροντας πιο στοχευμένες και σχετικές ειδήσεις με βάση το συναίσθημα και την κατηγορία κάθε άρθρου.

ΑΝΑΦΟΡΕΣ

[1] Pham, K. (2022, July 13). What are Recommendation Systems? - Khang Pham Medium. *Medium*. https://medium.com/@khang.pham.exxact/what-are-recommendation-systems-6bb5036042db

[2] Jonna, V. (2024, April 22). Types of Recommendation Systems - ellow.io. Ellow Talent. https://ellow.io/types-of-recommendation-systems/

[3] Types of Recommendation Systems & Their Use Cases - Maruti Techlabs - Medium. (2022, January 5). *Medium*. https://marutitech.medium.com/what-are-the-types-of-recommendation-systems-3487cbafa7c9

[4] Vaishnavi, N., & Kalpana, B. (2024). Sentiment Based Product Recommendation System Using Machine Learning Techniques. *Journal of Engineering Science and Technology Review*, *17*(1), 16–23. https://doi.org/10.25103/jestr.171.03

[5] Vaishnavi, N., & Kalpana, B. (2024b). Sentiment Based Product Recommendation System Using Machine Learning Techniques. *Journal of Engineering Science and Technology Review*, *17*(1), 16–23. https://doi.org/10.25103/jestr.171.03

[6] Vaj, T. (2023, April 14). Rule-base sentiment analysis - Tiya Vaj - Medium. *Medium*. https://vtiya.medium.com/rule-base-sentiment-analysis-adfad898470b

[7] Zhubatkhan, A., Buribayev, Z., Aubakirov, S., Dilmagambetova, M., & Ryskulbek, S. (2021). COMPARISON MODELS OF MACHINE LEARNING FOR MOVIE RECOMMENDATION SYSTEMS. *NEWS OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN*, *335*(1), 26–31. https://doi.org/10.32014/2021.2224-5294.4

[8] Marappan, R., & Bhaskaran, S. (2022). Movie Recommendation System Modeling Using Machine Learning. Trends Journal of Sciences Research, 1(1), 12–16. https://doi.org/10.31586/ijmebac.2022.291

[9] Golian, C., & Kuchař, J. (2017). Recommending news articles using rule-based classifier. https://dspace5.zcu.cz/bitstream/11025/26335/1/Golian.pdf

[10] Khandelwal, D., Shanbhag, D., Shriyan, A., Thorve, R., & Borse, Y. (2018). LeMeNo: Personalised News Using Machine Learning. https://doi.org/10.1109/iccubea.2018.8697560

[11] Staff, C. (2024, April 3). *What Is Python Used For? A Beginner's Guide*. Coursera. https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python

[12] GeeksforGeeks. (2024, July 30). *Pandas Introduction*. GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-pandas-in-python/

[13] Daniel. (2023, October 30). *NumPy : the most used Python library in Data Science*. Data Science Courses | DataScientest. https://datascientest.com/en/numpy-the-python-library-in-data-science

[14] *NLTK :: Natural Language Toolkit*. (n.d.). https://www.nltk.org/

[15] Salunke, R. (2023, March 29). *NLTK word_tokenize*. EDUCBA. https://www.educba.com/nltk-word_tokenize/

[16] Judah, B. (2024, May 2). Removing stop words with NLTK library in Python - Analytics Vidhya - Medium. *Medium*. https://medium.com/analytics-vidhya/removing-stop-words-with-nltk-library-in-python-f33f53556cc1

[17] Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media. https://www.nltk.org/book/

[18] Omar. (2023, August 15). The sklearn.metrics module. - Omar - Medium. *Medium*. https://medium.com/@bouimouass.o/the-sklearn-metrics-c568e0abcf03

[19] *Sklearn Train Test Split | Leadpages*. (n.d.). Leadpages | Landing Page & Website Software for Businesses. https://www.leadpages.com/blog/sklearn-train-test-split

[20] DataCamp. (n.d.). *Python Seaborn tutorial for beginners: Start visualizing data*. DataCamp. https://www.datacamp.com/tutorial/seaborn-python-tutorial

[21] *Regex*. (2024, August 16). https://www.computerhope.com/jargon/r/regex.htm

[22] GeeksforGeeks. (2020, December 11). *ImbalancedLearn module in Python*. GeeksforGeeks. https://www.geeksforgeeks.org/imbalanced-learn-module-in-python/

[23] *Kaggle: Your Machine Learning and Data Science Community*. (n.d.). https://www.kaggle.com/

[24] Carneiro, T., Da Nobrega, R. V. M., Nepomuceno, T., Bian, G., De Albuquerque, V. H. C., & Filho, P. P. R. (2018). Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications. *IEEE Access*, *6*, 61677–61685. https://doi.org/10.1109/access.2018.2874767

[25] Saeed, H. (2024, August 23). *Google Colab vs Jupyter Notebooks: A Guide to Data Enthusiasts*. CodeInterview Blog. https://codeinterview.io/blog/google-colab-vs-jupyter-notebooks-a-guide-to-data-enthusiasts/

[26] Aydin, A. (2023, October 5). 1 — Text Preprocessing Techniques for NLP - Aysel Aydin - Medium. *Medium*. https://ayselaydin.medium.com/1-text-preprocessing-techniques-for-nlp-37544483c007

[27] De Silva, M. (2023, August 29). Preprocessing Steps for Natural Language Processing (NLP): A Beginner's Guide. *Medium*. https://medium.com/@maleeshadesilva21/preprocessing-steps-for-natural-language-processing-nlp-a-beginners-guide-d6d9bf7689c9

[28] GeeksforGeeks. (2024a, January 31). *NLP | How tokenizing text, sentence, words works*. GeeksforGeeks. https://www.geeksforgeeks.org/nlp-how-tokenizing-text-sentence-words-works/

[29] Mudadla, S. (2023, November 10). What is Parts of Speech (POS) Tagging Natural Language Processing?In what kind of applications we can use Parts of Speech (POS) Tagging in Natural Language Processing. *Medium*. https://medium.com/@sujathamudadla1213/what-is-parts-of-speech-pos-tagging-natural-language-processing-in-2b8f4b07b186

[30] Harika. (2024, February 20). *Rule-Based Sentiment Analysis in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/rule-based-sentiment-analysis-in-python/

[31] Murel, J., PhD, & Kavlakoglu, E. (2024, August 13). Stemming and Lemmatization. *What are stemming and lemmatization?* https://www.ibm.com/topics/stemming-lemmatization

[32] Gillis, A. S. (2023, March 13). lemmatization. Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/lemmatization

[33] What is Feature Extraction? | Domino Data Lab. (n.d.). Domino Data Lab. https://domino.ai/data-science-dictionary/feature-extraction

[34] GeeksforGeeks. (2024c, May 23). *What is Feature Extraction?* GeeksforGeeks. https://www.geeksforgeeks.org/what-is-feature-extraction/

[35] Jain, A. (2024, February 4). TF-IDF in NLP (Term Frequency Inverse Document Frequency). *Medium*. https://medium.com/@abhishekjainindore24/tf-idf-in-nlp-term-frequency-inverse-document-frequency-e05b65932f1d

[36] Brownlee, J. (2019, August 30). *Train-test split for evaluating machine learning algorithms*. Machine Learning Mastery. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms

[37] Galarnyk, M. (2022, July 28). *Understanding Train Test Split*. Built In. https://builtin.com/data-science/train-test-split

[38] Viswa. (2024, March 24). Logistic Regression - Viswa - Medium. *Medium*. https://medium.com/@vk.viswa/logistic-regression-d001d0bce6c7

[39] Piduguralla, S. (2023, September 8). *Understanding the Sigmoid Function in Logistic Regression: Mapping Inputs to Probabilities*. https://www.linkedin.com/pulse/understanding-sigmoid-function-logistic-regression-piduguralla/

[40] Viswa. (2024b, March 24). Logistic Regression - Viswa - Medium. *Medium*. https://medium.com/@vk.viswa/logistic-regression-d001d0bce6c7

[41] Easily, L. S. (2024, April 14). *What Are The 3 Types of Logistic Regression?* LEARN STATISTICS EASILY. https://statisticseasily.com/types-of-logistic-regression/

[42] Pramoditha, R. (2023, December 1). Logistic Regression for Multiclass Classification — 3 Strategies You Need to Know. *Medium*. https://rukshanpramoditha.medium.com/logistic-regression-for-multiclass-classification-3-strategies-you-need-to-know-0a3e74574b96

[43] Sidharth. (2023, February 3). Multinomial Logistic Regression: Defintion, Math, and Implementation. *PyCodeMates*. https://www.pycodemates.com/2022/03/multinomial-logistic-regression-definition-math-and-implementation.html

[44] Thakur, P. (2022, October 23). What is Softmax Regression? - Preethi Thakur - Medium. *Medium*. https://medium.com/@tpreethi/softmax-regression-93808c02e6ac

[45] BotPenguin. (2024, September 3). *Softmax Function: Advantages and Applications | BotPenguin*. https://botpenguin.com/glossary/softmax-function

[46] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Proceedings-International AAAI Conference on Weblogs and Social Media, Association for the Advancement of Artificial Intelligence, 2014, pp. 216–225.

[47] S. Panchal, (2020, March 7), "Sentiment Analysis with VADER- Label the Unlabelled Data," Medium Website, Analytics Vidhya. https://medium.com/analytics-vidhya/sentiment-analysis-with-vader-label-the-unlabeled-data-8dd785225166

[48] Tavish. (2024, September 20). *12 Important Model Evaluation Metrics for Machine Learning Everyone Should Know (Updated 2024)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

[49] Metrics for Classification Model. (2023, June 22). AlmaBetter. https://www.almabetter.com/bytes/tutorials/data-science/classification-metrics

[50] Agrawal, R. (2024, September 19). *Know The Best Evaluation Metrics for Your Regression Model !* Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

[51] Programmer, P. (2023, May 17). Evaluation Metrics for Classification - Python Programmer - Medium. *Medium*. https://medium.com/@mlmind/evaluation-metrics-for-classification-fc770511052d

[52] Infragistics. (n.d.). *What Is Data Visualization? | Reveal Business Intelligence Glossary*. Reveal Embedded Analytics. https://www.revealbi.io/glossary/data-visualization

[53] Ajala, E. (2024, May 9). Data Visualization in Machine Learning - Eniola Ajala - Medium. *Medium*. https://medium.com/@ajalaeniola454/data-visualization-in-machine-learning-84641c95a759