ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ ΚΑΙ ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

ΤΜΗΜΑ ΟΡΓΑΝΩΣΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

Διπλωματική Εργασία

«ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΚΑΙ ΘΕΜΑΤΙΚΗ ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΕΙΜΕΝΩΝ ΑΞΙΟΛΟΓΗΣΕΩΝ ΗΛΕΚΤΡΟΝΙΚΟΥ ΛΙΑΝΙΚΟΥ ΕΜΠΟΡΙΟΥ»

της

ΑΝΝΑΣ ΡΙΖΟΥ ΤΟΥ ΑΝΔΡΕΑ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΛΕΩΝΙΔΑΣ ΧΑΤΖΙΘΩΜΑΣ

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

ΘΕΣΣΑΛΟΝΙΚΗ

ΟΚΤΩΒΡΙΟΣ 2022

ΑΦΙΕΡΩΣΕΙΣ

Στους γονείς μου

ΠΕΡΙΛΗΨΗ

Ο πρότερος στόχος των εταιριών είναι η προτίμηση της από το καταναλωτικό κοινό και συνεπώς η απόδοση κέρδους. Η επίτευξη αυτού του στόχου απαιτεί την αφοσίωση της στην ικανοποίηση των αναγκών των πελατών και την κάλυψη των αναγκών τους. Οι κριτικές του καταναλωτικού κοινού στο διαδίκτυο αντικατοπτρίζουν τα συναισθήματα τους για τα προϊόντα ή της υπηρεσίες της εταιρίας και συνεπώς αποτελούν μια σημαντική πηγή πληροφοριών. Η ανάλυση συναισθήματος αποτελεί ένα ισχυρό εργαλείο για την εξόρυξη και την μελέτη αυτών των απόψεων και συναισθημάτων. Στην παρούσα εργασία θα γίνει χρήση αλγορίθμων ταξινόμησης καθώς και θεματικής κατηγοριοποίησης για την κατάταξη των αξιολογήσεων και έπειτα, την εύρεση των χαρακτηριστικών του προϊόντος που δεν ικανοποίησαν τους καταναλωτές. Επιπλέον, θα γίνει χρήση τεχνικών αξιολόγησης των μοντέλων, ώστε να γίνει γνωστή η ακρίβεια των αποτελεσμάτων. Πιο συγκεκριμένα, προκειμένου να εξαχθούν τα χαρακτηριστικά ή οι λειτουργίες της συσκευής που είναι προβληματικές σύμφωνα με το σύνολο των αξιολογήσεων των καταναλωτών, σε πρώτη φάση, εφαρμόζονται οι αλγόριθμοι μηχανικής μάθησης οι οποίοι διαχωρίζουν τις αξιολογήσεις με βάση το συναίσθημα του πελάτη σε αρνητικές και θετικές. Στη συνέχεια, εφαρμόζεται η μέθοδος της θεματικής κατηγοριοποίησης στις αρνητικές εξ' αυτών. Τα πειραματικά ευρήματα αποδεικνύουν ότι οι στρατηγικές εξαγωγής χαρακτηριστικών και ρύθμισης παραμέτρων επιτρέπουν στα μοντέλα ταξινόμησης που αναπτύχθηκαν στη μελέτη να επιτύχουν ακρίβεια έως και 88%. Τα ευρήματα αυτά θα βοηθήσουν την εταιρία να προβούν σε ενέργειες βελτίωσης του προϊόντος και συνεπώς, την διατήρηση της ανταγωνιστικότητας της.

Λέξεις Κλειδιά: Ανάλυση συναισθήματος, Ταξινόμηση, Μηχανική Μάθηση, Θεματική κατηγοριοποίηση

ABSTRACT

Consumer preference and hence profitability are the companies' primary goals. It must be committed to serving and gratifying customers' requirements to achieve this goal. Online customer reviews are a valuable source of information since they capture how consumers feel about a company's goods or services. Sentiment analysis is an effective method for analysing and mining these beliefs and emotions. This study will classify customer reviews and then identify the product qualities that dissatisfied customers using classification algorithms and topic modeling. To determine the accuracy of the outcomes, model evaluation approaches will also be applied. More specifically, in a first step, machine learning algorithms are used to split the ratings based on customer sentiment into negative and positive ones in order to extract the features or functions of the device that are troublesome according to the set of consumer evaluations. Then, the negative ones are categorised using the topic modelling method. The experimental results show that the study's classification models can achieve up to 88% accuracy thanks to feature extraction and parameter tuning techniques. These insights will assist the business in taking steps to enhance the product and maintain its competitiveness.

Keywords: Sentiment analysis, Classification, Machine learning, Topic Modeling

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

1 ΕΙΣΑΓΩΓΗ

1.1 Σημαντικότητα του προβλήματος

Σήμερα, το ίντερνετ αποτελεί πηγή άπλετης πληροφορίας σε πραγματικό χρόνο όσο αναφορά την φωνή του καταναλωτή. Το 2021 εκτιμήθηκε ότι 2.14 τρισεκατομμύρια άνθρωποι παγκοσμίως αγόρασαν κάποιο προϊόν ή υπηρεσία με την βοήθεια του διαδικτύου. Είναι αξιοσημείωτο το γεγονός ότι την περασμένη χρονιά πραγματοποιήθηκαν ηλεκτρονικές αγορές αξίας 3.56 τρισεκατομμύριών δολαρίων μόνο μέσω κινητών τηλεφώνων (Woodward, 2022). Όπως είναι φυσικό, η χρήση του ίντερνετ εξαπλώνει την επιρροή του και στην αγορά, εφόσον παρέχει τεράστιο όγκο κριτικών που γράφονται καθημερινά από τους ανθρώπους για προϊόντα, υπηρεσίες αλλά και άλλα θέματα, όπως η θρησκεία και η πολιτική. Οι κριτικές αυτές βοηθούν τους καταναλωτές να επιλέγουν τα προϊόντα που θέλουν να αγοράσουν αλλά και τις εταιρίες να αποκτούν άμεσα ιδέες και κατευθύνσεις από το αγοραστικό κοινό.

Η άμεση αξιολόγηση των προϊόντων από το καταναλωτικό κοινό βοηθάει τις εταιρίες να προχωρούν σε αναπροσαρμογές που θα βελτιώσουν τα προϊόντα τους. Για τον λόγο αυτό, οι ανταγωνιστικές εταιρίες οι οποίες αφουγκράζονται την γνώμη και τις ανάγκες του καταναλωτή καταφέρνουν να επιβιώνουν και να αναπτύσσονται. Η ανάπτυξη αυτής της σχέσης μεταξύ ατόμου και εταιρίας, δημιουργεί το αίσθημα της εμπιστοσύνης και του σεβασμού και είναι δύσκολο να αναπτυχθεί εκ νέου από κάποια αναγνωστική εταιρία (Jeonga et al., 2019).

Όμως, για μια εταιρία είναι αδύνατον να διαβάσει όλες τις κριτικές των καταναλωτών για να διαμορφώσει μια γνώμη πάνω στην αποτελεσματικότητα των υπηρεσιών της ή την ποιότητα των προϊόντων της. Συνεπώς, αποτελεί μια μεγάλη πρόκληση η οποία μπορεί να αντιμετωπιστεί με την χρήση των μεθόδων της ανάλυσης συναισθήματος και της θεματικής κατηγοριοποίησης κειμένου. Η ανάλυση συναισθήματος αναλαμβάνει την εξαγωγή των απόψεων και των συναισθημάτων από τους χρήστες και είναι ευρέως διαδεδομένη μέθοδος στις μελέτες ανάλυσης κοινής γνώμης στις αρχές του 20ού αιώνα (Mantyla et al., 2018). Όμως, η εξαγωγή και η κατηγοριοποίηση των συναισθημάτων του καταναλωτικού κοινού δεν παρέχει πληροφορίες για τα «θέματα συζήτησης» αυτών για ένα προϊόν ή μια υπηρεσία. Για την εργασία αυτή γίνεται χρήση της θεματικής κατηγοριοποίησης (topic modeling), η οποία είναι ένας τύπος στατιστικής

μοντελοποίησης που αποσκοπεί στην ανακάλυψη των θεμάτων σε μια συλλογή εγγράφων. Οι μέθοδοι αυτοί αναλύονται αργότερα στην παρούσα εργασία.

1.2 Στόχοι της διπλωματικής εργασίας

Ο σκοπός της εργασίας είναι να μελετηθεί και να αναλυθεί η γνώμη και το συναίσθημα του καταναλωτή μέσω των αξιολογήσεων για ένα συγκεκριμένο προϊόν της Amazon, το Alexa Echo Auto. Για να επιτευχθεί αυτό, σε πρώτη φάση γίνεται η συλλογή των δεδομένων με την μέθοδο του web scraping και η προετοιμασία τους, Έπειτα, κατασκευάζονται μοντέλα επιβλεπόμενης μάθησης για την δυαδική ταξινόμηση των κριτικών που αντλήθηκαν από την ιστοσελίδα της Amazon, όσο αναφορά το συναίσθημα/γνώμη που εκφράζει ο αγοραστής για το εν λόγω προϊόν (sentiment analysis). Στην συνέχεια, επιλέγεται το βέλτιστο μοντέλο με χρήση τεχνικών αξιολόγησης με το οποίο θα είναι δυνατή η ταξινόμηση οποιονδήποτε νέων αξιολογήσεων που χρειάζεται να αναλυθούν για τις ανάγκες της εταιρίας. Τέλος, δημιουργούνται και συγκρίνονται δύο μοντέλα θεματικής κατηγοριοποίησης (topic modeling) των αρνητικών αξιολογήσεων, έτσι ώστε να επιλεχθεί ο κατάλληλος αλγόριθμος για την εύρεση των θεμάτων «συζήτησης» των καταναλωτών. Συνεπώς, από αυτήν την ανάλυση θα προκύψουν οι αδυναμίες του εν λόγω προϊόντος που απασχολούν τους αγοραστές, δηλαδή τα χαρακτηριστικά ή τις λειτουργίες που σύμφωνα με αυτούς επιφέρουν βελτίωση. Η ανάλυση επικεντρώθηκε στα αρνητικά σχόλια του καταναλωτικού κοινού, αφού σε αυτά μπορούν να ανακαλυφθούν ευκαιρίες βελτίωσης και ανάπτυξης των προϊόντων ή υπηρεσιών, με αποτέλεσμα την μεγιστοποίηση των πωλήσεων και την ικανοποίηση των αγοραστών.

1.3 Δομή της διπλωματικής εργασίας

Η δομή της παρούσας διπλωματικής εργασίας είναι η εξής: στο κεφάλαιο 1 περιλαμβάνεται η σημαντικότητα του προβλήματος που ασχολείται η παρούσα εργασία και ο στόχος της. Στο κεφάλαιο 2 γίνεται η ανασκόπηση της βιβλιογραφίας, η οποία επικεντρώνεται στην επεξήγηση και περιγραφή των βασικών εννοιών που εμπεριέχονται στο αντικείμενο της εργασίας. Δηλαδή, όλες τις απαραίτητες γνώσεις

που χρειάζεται ο αναγνώστης για να κατανοήσει τα εργαλεία, τις μεθόδους και τις έννοιες που χρησιμοποιούνται για την επίλυση του προβλήματος. Στο τρίτο κεφάλαιο παρουσιάζονται τα βήματα που ακολουθούνται στην επίλυση ενός προβλήματος στην Επιστήμη των Δεδομένων. Πιο συγκεκριμένα γίνεται ο προσδιορισμός του προβλήματος, η συλλογή δεδομένων, η προετοιμασία και η εξερεύνηση τους, η δημιουργία των μοντέλων καθώς και η αξιολόγηση τους. Στο κεφάλαιο 4 αφορά το Case Study όπου παρουσιάζονται τα μοντέλα που δημιουργήθηκαν για την ταξινόμηση και την θεματική κατηγοριοποίηση, καθώς και ο τρόπος κατασκευής τους. Σε αυτό το κεφάλαιο εφαρμόζονται όσα περιεγράφηκαν στο τρίτο κεφάλαιο, με χρήση τεχνικών και μοντέλων που έγιναν γνωστά από το δεύτερο κεφάλαιο. Τέλος, στο πέμπτο κεφάλαιο παρουσιάζονται τα αποτελέσματα των συγκρίσεων μεταξύ των μοντέλων που δημιουργήθηκαν για την ταξινόμηση και την θεματική κατηγοριοποίηση αντίστοιχα και γίνεται ανάλυση των αποτελεσμάτων του βέλτιστου μοντέλου θεματικής μοντελοποίησης.

2 ΗΛΕΚΤΡΟΝΙΚΟ ΕΜΠΟΡΙΟ

2.1 Τι είναι το ηλεκτρονικό εμπόριο

Το ηλεκτρονικό εμπόριο (**ecommerce**) είναι η πώληση και αγορά προϊόντων ή υπηρεσιών μέσω διαδικτύου. Αυτές οι επιχειρηματικές συναλλαγές πραγματοποιούνται μεταξύ μίας επιχείρησης προς μια άλλη επιχείρηση (B2B) ή μίας επιχείρησης προς τους καταναλωτές (B2C) ή ενός καταναλωτή προς έναν άλλον (C2C) ή ενός καταναλωτή προς μια επιχείρηση (C2B).

Οι επιχειρηματικές συναλλαγές μέσω διαδικτύου απαιτούν τη δημιουργία ενός ηλεκτρονικού καταστήματος, που θα επιτρέπει σε μια επιχείρηση να πωλεί τα προϊόντα ή τις υπηρεσίες της παγκοσμίως. Η δημιουργία του ηλεκτρονικού καταστήματος μιας επιχείρησης μπορεί να πραγματοποιηθεί με τη βοήθεια μιας πλατφόρμας ηλεκτρονικού εμπορίου, η οποία στην ουσία αποτελεί ένα λογισμικό. Κάποια παραδείγματα πλατφορμών ηλεκτρονικού εμπορίου είναι το Shopify, το BigCommerce και το Magento (Think-plus, 2020).

Κάποια από τα πλεονεκτήματα διάθεσης μιας ιστοσελίδας ηλεκτρονικού εμπορίου από μια επιχείρηση είναι τα παρακάτω (Lutkevich et al., 2022):

- **Διαθεσιμότητα**: Οι ιστότοποι ηλεκτρονικού εμπορίου είναι διαθέσιμοι κάθε ώρα της μέρας, επιτρέποντας στους καταναλωτές να κάνουν αγορές ανά πάσα στιγμή.
- **Ταχύτητα πρόσβασης**: Ο χρόνος μετάβασης και αναμονής λόγω συνωστισμού σε ένα φυσικό κατάστημα είναι πολύ μεγαλύτερος από την περιήγηση σε ένα ηλεκτρονικό κατάστημα. Ο χρόνος που ξοδεύει ο πελάτης σε ένα ηλεκτρονικό κατάστημα καθορίζεται από τη συσκευή από την οποία περιηγείται αλλά και από την ποιότητα της ιστοσελίδας.
- **Προσβασιμότητα**: Η εύρεση ενός συγκεκριμένου προϊόντος σε ένα φυσικό κατάστημα πολλές φορές δυσκολεύει τους καταναλωτές. Σε αντίθεση με το ηλεκτρονικό κατάστημα, στο οποίο μπορεί κάποιος να περιηγηθεί μέσω κατηγοριών προϊόντων ή και μέσω αναζήτησης.
- **Διεθνής εμβέλεια**: Το ηλεκτρονικό εμπόριο έχει τη δυνατότητα να επεκτείνει την πελατεία της επιχείρησης πέρα από την τοπική εμβέλεια, αφού το

ηλεκτρονικό κατάστημα μπορεί να είναι προσβάσιμο σε οποιονδήποτε χρησιμοποιεί το διαδίκτυο, σε όποια χώρα και αν κατοικεί.

- **Εξατομίκευση και συστάσεις προϊόντων**: Οι ιστότοποι ηλεκτρονικού εμπορίου έχουν στη διάθεση τους δεδομένα ιστορικών περιήγησης, αναζητήσεων και αγορών των πελατών στο κατάστημα. Τα δεδομένα αυτά είναι χρήσιμα για το πλασάρισμα προϊόντων τα οποία είναι πιο πιθανό να αρέσουν στον πελάτη.

Παρόλα αυτά οι ηλεκτρονικές αγορές έχουν και κάποια μειονεκτήματα, τα οποία περιγράφονται παρακάτω (Lutkevich et al., 2022):

- **Περιορισμένη εξυπηρέτηση πελατών**: Σε ένα φυσικό κατάστημα είναι πολύ πιο εύκολο για έναν πελάτη να κάνει μια ερώτηση ή να θέσει ένα ζήτημα σε αντίθεση με ένα ηλεκτρονικό κατάστημα. Η ηλεκτρονική εξυπηρέτηση λειτουργεί συγκεκριμένες ώρες και μπορεί να μην απαντά σε κάποια συγκεκριμένη ερώτηση και συνεπώς, είναι μη πρακτική για τους καταναλωτές.
- **Εμπειρία προϊόντος**: Όσο καλής ποιότητας και αν είναι οι εικόνες στην ιστοσελίδα, δεν μπορούν να δώσουν στον πελάτη την ίδια αίσθηση του προϊόντος, όσο η άμεση επαφή μαζί του. Έτσι, οι καταναλωτές μπορεί να αγοράσουν προϊόντα που διαφέρουν από τις προσδοκίες τους και να χρειαστεί να τα επιστρέψουν.
- **Χρόνος αναμονής**: Με το ηλεκτρονικό εμπόριο οι πελάτες περιμένουν να αποσταλεί το προϊόν που έχουν αγοράσει και συνεπώς, δεν μπορούν να το έχουν άμεσα στα χέρια τους.
- **Ασφάλεια**: Στην περίπτωση που ο πελάτης αποθηκεύσει τη πιστωτική του κάρτα στην ιστοσελίδα της επιχείρησης για την διευκόλυνση του σε μελλοντικές αγορές και η ιστοσελίδα παραβιαστεί, ενδέχεται να κλαπούν οι πληροφορίες της. Μια τέτοια παράβαση, εκτός από την ασφάλεια των χρημάτων του καταναλωτή, μπορεί επίσης να οδηγήσει σε βλάβη της φήμης της επιχείρησης.

Κάποιες από τις μεγαλύτερες πλατφόρμες ηλεκτρονικού εμπορίου στον κόσμο είναι η Amazon, το eBay, το Alibaba, το AliExpress, η Walmart κ.α. Η Amazon, η οποία ιδρύθηκε από τον Τζεφ Μπέζος, είναι η 9η πιο δημοφιλής ιστοσελίδα στον κόσμο (Think-plus, 2020). Περισσότερα από 150.6 εκατομμύρια άτομα επισκέπτονται την ιστοσελίδα της Amazon μηνιαίως και πάνω από 600 δισεκατομμύρια δολάρια

πωλήθηκαν μέσω αυτής το 2021. Η πιο δημοφιλής κατηγορία προϊόντων της Amazon είναι τα ηλεκτρονικά είδη, όμως παρέχει μια μεγάλη σειρά προϊόντων όπως εξοπλισμός σπιτιού, ένδυση, gadgets, εξοπλισμός εσωτερικού χώρου κ.α. (Woodward, 2022)

2.2 Ηλεκτρονικό εμπόριο και COVID-19

Το ηλεκτρονικό λιανικό εμπόριο αναπτύχθηκε ραγδαία τις τελευταίες δύο δεκαετίες λόγω της ευρείας χρήσης ιστοτόπων ηλεκτρονικού εμπορίου, όπως η Amazon και το eBay. Πιο συγκεκριμένα, τη χρονιά του 2011, οι πωλήσεις μέσω ηλεκτρονικού εμπορίου αποτελούσαν μονάχα το 5% των συνολικών λιανικών πωλήσεων, σε αντίθεση με το 2020, όπου με την έναρξη της πανδημίας του COVID-19, το ποσοστό αυτό ανέβηκε στο 16% (Lutkevich et al., 2022).

Η αύξηση του ποσοστού των ηλεκτρονικών πωλήσεων εν μέσω πανδημίας είναι πολύ λογική αφού οι καταναλωτές δεν επιθυμούσαν να επισκεφθούν κάποιο φυσικό κατάστημα και να προβούν σε άσκοπες μετακινήσεις. Επιπλέον, σε περιόδους απαγόρευσης μετακινήσεων και διακοπής λειτουργίας των επιχειρήσεων, οι άνθρωποι αφιέρωναν πολύ περισσότερο χρόνο στο κινητό τους τηλέφωνο, την τηλεόραση ή τον υπολογιστή, γεγονός που αύξησε την ανάγκη για ηλεκτρονικές αγορές.

Σύμφωνα με το άρθρο της Berthene (2022), εκτιμάται ότι τα δύο τελευταία χρόνια, η πανδημία του COVID-19 συνέβαλε με επιπλέον 218.53 δισεκατομμύρια δολάρια στα κέρδη του ηλεκτρονικού εμπορίου των ΗΠΑ. Όσο αναφορά την Ελλάδα, σύμφωνα με το Εργαστήριο Ηλεκτρονικού Εμπορίου και Ηλεκτρονικού Επιχειρείν του Οικονομικού Πανεπιστημίου Αθηνών, το 2021 ο τζίρος του ηλεκτρονικού εμπορίου έφτασε στα 14 δισεκατομμύρια ευρώ, ξεπερνώντας όλες τις προηγούμενες χρονιές. Επιπλέον, σύμφωνα με ερεύνα του Ελληνικού Συνδέσμου Ηλεκτρονικού εμπορίου (GRECA), το 94% των ανθρώπων θα συνεχίζει να αγοράζει μέσω ηλεκτρονικών καταστημάτων και ένα 14% θα αυξήσει τη συχνότητα των αγορών (FortuneGreece, 2022).

Μια ακόμη μεταβολή που προκλήθηκε από την πανδημία και ενίσχυσε τα κέρδη των πωλήσεων του ηλεκτρονικού εμπορίου, είναι ο πληθωρισμός. Σύμφωνα με στοιχεία της Adobe, οι καταναλωτές πλήρωσαν περισσότερα από 30 δισεκατομμύρια δολάρια περισσότερα στο διαδίκτυο για την ίδια ποσότητα αγαθών. Αυτό το γεγονός βέβαια,

φαίνεται να μην απέτρεψε το καταναλωτικό κοινό να δαπανήσει τα χρήματα του στο διαδίκτυο (Berthene, 2022).

2.3 Ηλεκτρονικό εμπόριο και Data Science

Στις μέρες μας, η εξέλιξη της επιστήμης των δεδομένων και οι εφαρμογές της βοηθούν σε πολύ μεγάλο βαθμό μικρές ή μεγάλες εταιρίες να λαμβάνουν αποφάσεις με σκοπό την κατανόηση των επιδόσεων τους και την αύξηση των κερδών τους. Πιο συγκεκριμένα, με τη χρήση του Data Science, οι εταιρίες μπορούν να προβλέψουν τα κέρδη ή τις απώλειες, να προωθήσουν στους καταναλωτές τα κατάλληλα προϊόντα με βάση τις αγορές τους αλλά και με βάση το προφίλ που δημιουργείται για αυτούς μέσω των ιστορικών τους δεδομένων κ.α. (Pant, 2019)

Κάποιες από τις συχνότερες εφαρμογές της επιστήμης των δεδομένων στο ηλεκτρονικό εμπόριο είναι οι εξής (Great Learning, 2020):

- **Ανάλυση εγγυήσεων**: Με την ανάλυση εγγυήσεων οι πωλητές ελέγχουν τη διάρκεια ζωής των προϊόντων τους, τα τυχών προβλήματα, τις επιστροφές κ.α. Η ανάλυση αυτή βασίζεται στην εκτίμηση της κατανομής των βλαβών με βάση δεδομένα τα οποία περιλαμβάνουν την ηλικία, τον αριθμό επιστροφών και την ποσότητα των προϊόντων που έχουν επιβιώσει. Έτσι, οι πωλητές ελέγχουν πόσα προϊόντα έχουν πωληθεί και πόσα έχουν επιστραφεί λόγω βλάβης. Συνεπώς, βρίσκονται σε θέση να τιμολογήσουν τις εγγυήσεις τους και να τις προσφέρουν ως πακέτο στους πελάτες με την αγορά τους.
- **Ανάλυση καλαθιού αγοράς**: Πρόκειται για μια αρκετά διαδεδομένη ανάλυση που χρησιμοποιούν οι λιανοπωλητές εδώ και χρόνια με σκοπό την εύρεση της πιθανότητας αγοράς ενός προϊόντος από έναν πελάτη με βάση τα προϊόντα που έχει ήδη στο καλάθι αγορών. Για παράδειγμα, αν ένας πελάτης αγοράσει μπύρες από ένα κατάστημα ψιλικών ειδών τότε είναι πιο πιθανό να αγοράσει και ξηρούς καρπούς. Όσο αναφορά το ηλεκτρονικό εμπόριο, τα δεδομένα των πελατών που αντλούνται είναι πολύ σημαντικά για την εύρεση των αγοραστικών τους προτιμήσεων. Η ανάλυση αυτή πραγματοποιείται με την κατασκευή αλγορίθμων Μηχανικής Μάθησης ή Βαθιάς Μάθησης.

- **Μηχανές συστάσεων**: Αποτελούν ένα από τα ισχυρότερα εργαλεία για έναν πωλητή και χρησιμοποιούνται για να οδηγήσουν έναν πελάτη στην αγορά ενός προϊόντος. Έτσι, οι πωλήσεις αυξάνονται και δημιουργούνται τάσεις. Οι μηχανές συστάσεων, οι οποίες αποτελούνται από περίπλοκους αλγορίθμους μηχανικής μάθησης και βαριάς μάθησης, έχουν τη δυνατότητα να παρακολουθούν την αγοραστική συμπεριφορά των πελατών, δηλαδή, τις καταναλωτικές τους συνήθειες και να τους προτείνουν προϊόντα ανάλογα με τα γούστα τους.
- **Ανάλυση συναισθήματος πελατών**: Αξιοποιεί την επεξεργασία φυσικής γλώσσας για να εντοπίσει λέξεις που φέρουν αρνητικό ή θετικό συναίσθημα απέναντι σε ένα προϊόν ή μια μάρκα. Έτσι, βοηθά τις επιχειρήσεις να βελτιώσουν τα προϊόντα τους. Η ανάλυση αυτή υπάρχει εδώ και χρόνια, αλλά πλέον, με την υλοποίηση αλγορίθμων μηχανικής μάθησης και άλλων τεχνικών, έχει καταστεί πιο απλή, αυτόματη και γρήγορη στην απόδοση έγκυρων αποτελεσμάτων.
- **Διαχείριση αποθεμάτων**: Πρόκειται για διαχείριση των αποθεμάτων, δηλαδή των αγαθών που αποθηκεύονται για μεταγενέστερη χρήση σε περιόδους ανάγκης, με σκοπό την αύξηση των πωλήσεων και τη βέλτιστη χρήση των πόρων. Με την βοήθεια αυτής της ανάλυσης, αν υπάρξει αύξηση των πωλήσεων, η προσφορά δε θα επηρεαστεί. Για τη διαχείριση αποθεμάτων κατασκευάζονται αλγόριθμοι μηχανικής μάθησης που αναλύουν τα δεδομένα και βρίσκουν μοτίβα και συσχετίσεις μεταξύ των αγορών.
- **Πρόβλεψη αξίας διάρκειας ζωής**: Αφορά τη συνολική αξία του κέρδους που αποκτά η επιχείρηση από έναν πελάτη, καθ' όλη τη διάρκεια της σχέσης πελάτη-επιχείρηση, η οποία ονομάζεται αξία διάρκειας ζωής πελάτη. Σε αυτήν την περίπτωση, οι προβλέψεις πραγματοποιούνται με βάση ιστορικά δεδομένα, τα οποία οδηγούν στις πρόσφατες αγορές. Ο αλγόριθμος που κατασκευάζεται αναλύει τα δεδομένα των αγορών, τα έξοδα και την αγοραστική συμπεριφορά του πελάτη και εξάγει την πιθανή αξία του για την επιχείρηση.

3 ΑΝΑΣΚΟΠΗΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

3.1 Ανάλυση κειμένου

Η ανάλυση κειμένου πρόκειται για την διαδικασία χρήσης υπολογιστών για την ανάγνωση και την κατανόηση μεγάλων ποσοτήτων μη δομημένων δεδομένων με την μορφή κειμένου (όπως για παράδειγμα αξιολογήσεις προϊόντων, μηνύματα ηλεκτρονικού ταχυδρομείου κ.α.), με σκοπό την εξαγωγή επιχειρηματικών συμπερασμάτων. Περιλαμβάνει εργασίες ταξινόμησης, εύρεσης μοτίβων, συναισθημάτων και άλλων χρήσιμων γνώσεων.

3.1.1 Επεξεργασία φυσικής γλώσσας (NLP)

Η επεξεργασία φυσικής γλώσσας (Natural language processing ή αλλιώς NLP) αποτελεί κομμάτι της τεχνητής νοημοσύνης και πρόκειται για την ικανότητα του υπολογιστή να κατανοεί την ανθρώπινη γλώσσα, σε προφορικό και γραπτό λόγο. Η NLP υπάρχει εδώ και περισσότερα από 50 χρόνια και έχει ποικίλες εφαρμογές στον πραγματικό κόσμο σε διάφορους τομείς, όπως η ιατρική, οι μηχανές αναζήτησης και η επιχειρηματική ευφυΐα (Lutkevich, 2021).

Η NLP μέθοδος έχει δύο κύριες φάσεις: την προετοιμασία των δεδομένων και την κατασκευή των μοντέλων. Η προετοιμασία των δεδομένων περιλαμβάνει την επεξεργασία και τον καθαρισμό ώστε να είναι έτοιμα να αναλυθούν. Σκοπός της είναι να δίνει λειτουργική μορφή στα δεδομένα και να επισημαίνει κάποια χαρακτηριστικά στο κείμενο με τα οποία θα μπορεί να λειτουργήσει ένας αλγόριθμος.

3.1.2 Ταξινόμηση κειμένου

Υπολογίζεται ότι περίπου το 80% όλων των πληροφοριών είναι αδόμητες, με τα δεδομένα κειμένου να θεωρούνται από τα πιο συνηθισμένα δεδομένα τέτοιου τύπου (Monkeylearn, 2022). Λόγω της ακατάστατης φύσης του κειμένου, η ανάλυση, η κατανόηση και η ταξινόμηση του είναι δύσκολη και απαιτεί χρόνο. Πράγμα που οδηγεί τις εταιρίες να αποτυγχάνουν να τα αξιοποιήσουν στο έπακρο.

Η ταξινόμηση κειμένου βοηθά στην αξιοποίηση αυτών των πληροφοριών. Χρησιμοποιείται συχνά στην ανάλυση συναισθήματος, η οποία καθορίζει αν το κείμενο εμπεριέχει θετικό ή αρνητικό συναίσθημα. Μπορεί επίσης, να είναι χρήσιμο για την ανίχνευση πρόθεσης, με την οποία προβλέπεται το επόμενο βήμα του συγγραφέα ή του ομιλητή ανάλογα με το κείμενο που παράγει [15].

Η πιο συνήθης μορφή ταξινόμησης κειμένου είναι η δυαδική ταξινόμηση δηλαδή, ο διαχωρισμός όλων των εγγράφων του σώματος κειμένου σε δύο κατηγορίες. Είναι, πολύ συχνά, το πρώτο βήμα για την επιλογή του συνόλου εγγράφων που θα υποβληθεί σε περαιτέρω επεξεργασία ή μπορεί να είναι και το μοναδικό βήμα επεξεργασίας, όπως για παράδειγμα συμβαίνει στο φιλτράρισμα ανεπιθύμητης αλληλογραφίας (Nareddy & Chakraborty, 2012). Γενικά, η βασική προσέγγιση στην ταξινόμηση κειμένου είναι η εξαγωγή χαρακτηριστικών για την περιγραφή ενός εγγράφου και στη συνέχεια η εφαρμογή ενός αλγορίθμου για την κατάταξη του συγκεκριμένου εγγράφου στην κατάλληλη κατηγορία.

3.2 Μηχανική μάθηση

3.2.1 Ορισμός

Αποτελεί ένα πεδίο της τεχνητής νοημοσύνης που αποσκοπεί στην κατασκευή αλγορίθμων πρόβλεψης που βασίζονται σε σύνολα δεδομένων. Πιο συγκεκριμένα, έχει να κάνει με την εκπαίδευση των αλγορίθμων ώστε να είναι ικανοί να ανακαλύπτουν μοτίβα χρησιμοποιώντας ιστορικά δεδομένα. Η ακρίβεια της εξόδου του αλγορίθμου εξαρτάται σε μεγάλο βαθμό από τον όγκο των δεδομένων, καθώς ο μεγάλος όγκος δεδομένων βοηθά στην κατασκευή ενός μοντέλου υψηλότερης επίδοσης που προβλέπει τα μοτίβα με μεγαλύτερη ακρίβεια. Οι εφαρμογές της Μηχανικής Μάθησης περιλαμβάνουν την ανίχνευση ανεπιθύμητης αλληλογραφίας, την ηλεκτρονική υποστήριξη πελατών, τις προτάσεις προϊόντων κ.α. (JavaPoint, 2022)

3.2.2 Κατηγορίες αλγορίθμων

Οι τύποι αλγορίθμων που εφαρμόζονται για την εκμάθηση των μοντέλων μηχανικής μάθησης χωρίζονται σε τέσσερεις κατηγορίες (SAS, 2022):

1. **Αλγόριθμοι επιβλεπόμενης μάθησης:** Εκπαιδεύονται με δεδομένα που αντιστοιχίζονται με τα επιθυμητά αποτελέσματα, οπότε τα αποτελέσματα θα έχουν τιμές που είναι ήδη γνωστές. Κάποιοι από τους πιο δημοφιλείς αλγόριθμους επιβλεπόμενης μάθησης είναι οι Naive Bayes, Random Forest και Logistic Regression.

2. **Αλγόριθμοι μη επιβλεπόμενης μάθησης:** Τα δεδομένα που δέχονται σαν είσοδο δεν αντιστοιχίζονται με δεδομένα εξόδου και για το σύστημα δεν υπάρχει ‘σωστή’ απάντηση. Σε αυτή την περίπτωση, ο αλγόριθμος ψάχνει μόνος το τρόπο να χωρίσει τα δεδομένα, χωρίς κάποιο αρχικό ερέθισμα. Οι πιο γνωστές τεχνικές μη επιβλεπόμενης μάθησης είναι οι K-Means, Nearest-Neighbor mapping, τεχνικές διάτμησης θεμάτων κειμένου, προτεινόμενων προϊόντων και εύρεσης ακραίων τιμών.

3. **Αλγόριθμοι ημι-επιβλεπόμενης μάθησης:** Δέχονται ως είσοδο μια μίξη από τον τύπο δεδομένων των δύο πρώτων αλγορίθμων. Δηλαδή, εκπαιδεύονται με δεδομένα, αντιστοιχισμένα ή μη, με τα επιθυμητά αποτελέσματα. Εδώ ο αλγόριθμος έχει την ελευθερία να εξερευνήσει τα patterns και να αναπτύξει τη δική του αντίληψη για τα δεδομένα.

4. **Αλγόριθμοι ενισχυτικής μάθησης:** Οι αλγόριθμοι αυτοί ανακαλύπτουν μέσω των δοκιμών και των σφαλμάτων. Αυτός ο τρόπος εκμάθησης έχει τρία χαρακτηριστικά: τον εκπαιδευόμενο, το περιβάλλον και τις ενέργειες. Το ζητούμενο είναι η επιλογή των ενεργειών που αποφέρουν την μεγαλύτερη ανταμοιβή μέσα σε ένα δεδομένο χρονικό διάστημα. Οι αλγόριθμοι αυτοί χρησιμοποιούνται κυρίως στην ρομποτική και τα ηλεκτρονικά παιχνίδια.

3.2.3 Χαρακτηριστικά (features)

Το κύριο πρόβλημα των προβλημάτων επεξεργασίας φυσικής γλώσσας είναι ότι οι υπολογιστές δεν μπορούν να «δουλέψουν» με την αρχική μορφή των δεδομένων,

δηλαδή το κείμενο. Οπότε χρειάζεται κάποια τεχνική μετατροπής των σωμάτων κειμένου σε πίνακες (ή διανύσματα) χαρακτηριστικών. Μια λύση αυτού του προβλήματος αποτελεί η εξαγωγή χαρακτηριστικών (feature extraction), η οποία αποτελεί μια πολύ σημαντική διαδικασία για την προετοιμασία των κειμένων πριν την κατηγοριοποίηση τους από έναν αλγόριθμο ταξινόμησης. Πιο συγκεκριμένα, αυτή η μέθοδος αποδίδει βάρη σε συγκεκριμένες λέξεις και τις αναπαριστά αριθμητικά. Αυτό έχει ως αποτέλεσμα τη βελτίωση της απόδοσης του αλγορίθμου και δίνει μια καλύτερη εικόνα των δεδομένων. Οι πιο γνωστές μέθοδοι εξαγωγής χαρακτηριστικών είναι οι Bag-of-Words και TF-IDF.

3.2.3.1 Bag-of-Words

Η μέθοδος Bag-of-Words είναι από τις πιο βασικές μεθόδους εξαγωγής και αναπαράστασης των κειμένων, η οποία αντιμετωπίζει το σύνολο αυτών σαν ένα «σάκο από λέξεις».

Αρχικά, όλο το σώμα κειμένου μετατρέπεται σε πεζά γράμματα και αφαιρούνται όλα τα σημεία στίξης. Έπειτα, το κείμενο κατακερματίζεται και εξάγεται ένα σύνολο από μοναδικές λέξεις, όπως επίσης και ο αριθμός φορών εμφάνισης τους στο σώμα κειμένου. Τέλος, γίνεται η κατασκευή του αλγορίθμου. Στην ουσία, κατασκευάζεται ένας πίνακας, στον οποίο αναθέτεται μια στήλη για κάθε λέξη, ενώ κάθε γραμμή αντιστοιχείται σε μια κριτική. Αυτή η διαδικασία είναι γνωστή ως διανυσματοποίηση κειμένου. Κάθε εγγραφή στον πίνακα σηματοδοτεί την παρουσία (ή την απουσία) της λέξης στην κριτική. Αν η λέξη υπάρχει, τότε παίρνει την τιμή 1 και αν όχι, παίρνει την τιμή 0.

Η μέθοδος Bag-of-Words είναι εύκολη στην κατανόηση και στην εφαρμογή, όμως έχει κάποια μειονεκτήματα και περιορισμούς. Το μέγεθος του λεξιλογίου έχει αντίκτυπο στην αραιότητα των αναπαραστάσεων των εγγράφων οπότε απαιτείται καλή διαχείριση. Το μοντέλο αυτό, εστιάζει στη συχνότητα εμφάνισης των λέξεων και όχι στη σημασία τους, πράγμα που αποτελεί ένα κύριο μειονέκτημα του, αφού η διάταξη των λέξεων μπορεί να αλλάξει εντελώς το νόημα μιας πρότασης. Επίσης, είναι δύσκολο να μοντελοποιήσει αραιές αναπαραστάσεις, οπότε δυσκολεύεται να αξιοποιήσει μια μικρή ποσότητα πληροφοριών σε έναν τεράστιο χώρο αναπαραστάσεων (Engati, 2021). Η διαχείριση των αραιών διανυσμάτων μπορεί να αντιμετωπιστεί με τον περιορισμό των

μεγεθών των διανυσμάτων που προκύπτουν. Μερικοί από τους τρόπους εφαρμογής αυτού του περιορισμού είναι:

- Αγνόηση του είδους των γραμμάτων (πεζών ή κεφαλαίων).
- Παράλειψη των σημείων στίξης όταν κρίνεται ότι δεν παρέχουν πληροφορίες σχετικά με το νόημα του κειμένου.
- Παράλειψη σημείων του λόγου που χρησιμοποιούνται για τη σύνδεση σημασιών και δεν περιέχουν χρήσιμες πληροφορίες.
- Ο εντοπισμός ανορθόγραφων λέξεων και η αντιστοίχιση τους με λέξεις που έχουν ήδη καταχωρηθεί στο λεξικό.
- Προσδιορισμός λέξεων που είναι γραμμένες με διαφορετικές μορφές και ομαδοποίηση συνωνύμων. (Πιζάνιας, 2018)

3.2.3.2 Term Frequency-Inverse Document Frequency (TF-IDF)

Όπως αναφέρθηκε και παραπάνω, η εξαγωγή χαρακτηριστικών είναι μια διαδικασία με την οποία εξάγουμε σημαντικά χαρακτηριστικά από δεδομένα κειμένου με σκοπό την τροφοδότηση ενός στατιστικού αλγορίθμου ή αλγορίθμου Μηχανικής Μάθησης. Η μέθοδος TF-IDF αποτελεί μια τέτοια μέθοδο και μάλιστα αναγνωρίζεται ως η καλύτερη μέθοδος εξαγωγής χαρακτηριστικών για την ανάλυση κειμένου (Nguyen et al., 2018).

Συχνότητα όρων (TF). Ο όρος TF είναι η συχνότητα εμφάνισης μιας φράσης ή μια ομάδας λέξεων σε ένα κείμενο. Ονομάζεται επίσης, μοντέλο Bag-of-Words. Σε αυτό το μοντέλο, κάθε έγγραφο αναπαρίσταται ως ένα διάνυσμα από 0 και 1. Εάν μια φράση υπάρχει σε ένα έγγραφο, η αντίστοιχη θέση της στο διάνυσμα κωδικοποιείται ως «1» και αν δεν υπάρχει, κωδικοποιείται με «0». Ο όρος TF υπολογίζεται ως εξής:

$$TF(\text{όρος}) = \frac{\textit{Συχνότητα εμφάνισης όρου στο κείμενο}}{\textit{Αριθμός φορών εμφάνισης του όρου στο κείμενο}}$$

Αντίστροφη συχνότητα εγγράφων (IDF). Το IDF μιας λέξης είναι το μέτρο της σχετικής σημασίας αυτής της λέξης σε ολόκληρο το σώμα κειμένων. Το IDF υπολογίζεται ως εξής:

$$IDF(\text{όρος}) = log\left(\frac{Συνολικός\ αριθμός\ εγγράφων}{Συνολικός\ αριθμός\ εγγράφων\ που\ περιέχουν\ τον\ όρο}\right)$$

Αντίστροφη συχνότητα όρων και εγγράφων (TF-IDF). Το TF-IDF είναι το γινόμενο της βαθμολογίας TF και IDF για συγκεκριμένες λέξεις. Κάθε εγγραφή αντιπροσωπεύεται ως διάνυσμα που αποτελείται από βαθμολογίες TF-IDF για κάθε λέξη του εγγράφου. Το TF-IDF μειώνει την επιρροή των τακτικά εμφανιζόμενων αλλά λιγότερο κατατοπιστικών χαρακτηριστικών.

3.2.3.3 Μέθοδος N-grams

Το N-grams μπορεί να οριστεί ως η ακολουθία N στοιχείων από ένα σώμα κειμένου ή ομιλίας. Τα στοιχεία μπορεί να είναι γράμματα, συλλαβές ή λέξεις, ανάλογα με την εφαρμογή. Στην επεξεργασία φυσικής γλώσσας (NLP), η μέθοδος αναφέρεται συνήθως στα N-grams ως σειρές λέξεων, όπου το n συμβολίζει τον αριθμό των λέξεων που επιλέγεται. Συνήθως διακρίνονται οι ακόλουθοι τύποι N-grams:

- **Unigram:** Ένα N-gram με απλά μια συμβολοσειρά στο εσωτερικό του κειμένου. Για παράδειγμα μπορεί να είναι η λέξη «οθόνη» στην πρόταση «Η οθόνη δεν έχει καλή ποιότητα».
- **Bigram:** Ο συνδυασμός 2 συμβολοσειρών ή λέξεων που εμφανίζονται σε ένα έγγραφο, όπως για παράδειγμα «ανάλυση οθόνης» ή «επιτραπέζιος υπολογιστής». Εδώ αξίζει να σημειωθεί ότι η σειρά τον λέξεων μένει ίδια.
- **3-gram ή Trigram:** Ένα N-gram που περιέχει έως και 3 στοιχεία που επεξεργάζονται μαζί (π.χ. «συσκευή εγγραφής βίντεο»)

Ένα από τα πλεονεκτήματα της μεθόδου N-grams είναι ότι βοηθά στην ανάλυση του κειμένου σε διαφορετικά επίπεδα (bigram, trigram, n-gram) και συνεπώς βρίσκονται βαθύτερα νοήματα και πληροφορίες για την γνώμη του συγγραφέα. Επίσης, είναι εύκολη στην εκτέλεση και στην κατανόηση.

Το κύριο πρόβλημα της μεθόδου είναι ότι αν η επιλογή του n δεν είναι σωστή τότε το αποτέλεσμα δε θα είναι ικανοποιητικό (Eisenstein, 2019). Για παράδειγμα, στην πρόταση «Το **κινητό** μετά την πτώση του από το τραπέζι, **έσπασε**», οι λέξεις που είναι γραμμένες με έντονα γράμματα εξαρτώνται μεταξύ τους. Πιο συγκεκριμένα, οι πιθανότητα της λέξης «έσπασε» εξαρτάται από τη γνώση ότι το υποκείμενο είναι ένα κινητό. Εάν το **n δεν είναι αρκετά μεγάλο** για να αποσπάσει αυτήν την πληροφορία, τότε το μοντέλο που θα κατασκευαστεί θα προσφέρει χαμηλές πιθανότητες για αυτήν την πρόταση.

Αντίστοιχα, όταν το n είναι πολύ μεγάλο, είναι δύσκολο να κάνουμε καλές εκτιμήσεις των παραμέτρων n-gram από το σύνολο δεδομένων μας, λόγω της διασποράς δεδομένων. Για να αντιμετωπιστεί το παραπάνω παράδειγμα του κινητού, πρέπει να προσομοιωθούν 9-grams, πράγμα που σημαίνει ότι πρέπει να ληφθούν υπόψη V^9 συμβάντα. Έστω ένα πολύ μικρό λεξιλόγιο $V = 10^3$, τότε θα πρέπει να εκτιμηθεί η πιθανότητα 10^{27} διαφορετικών γεγονότων.

Άρα, τα μικρά μεγέθη n-gram έχουν υψηλή προκατάληψη, ενώ τα μεγάλα μεγέθη n-gram έχουν υψηλή διακύμανση. Επίσης, είναι δυνατόν να υπάρχουν ταυτόχρονα και τα δύο αυτά προβλήματα. Η γλώσσα είναι γεμάτη από εξαρτήσεις μεγάλης εμβέλειας που δεν μπορούμε να συλλάβουμε επειδή το n είναι πολύ μικρό, ενώ τα σύνολα δεδομένων είναι γεμάτα από σπάνια φαινόμενα των οποίων οι πιθανότητες δεν μπορούμε να υπολογίσουμε με ακρίβεια επειδή το n είναι πολύ μεγάλο.

3.2.4 Μοντέλα ταξινόμησης

Τα μοντέλα ταξινόμησης είναι μια τεχνική Μηχανικής Μάθησης που χρησιμοποιείται για την κατηγοριοποίηση νέων παρατηρήσεων με βάση δεδομένα εκπαίδευσης. Συνεπώς, ένας αλγόριθμος μαθαίνει μέσω ενός συνόλου δεδομένων και στη συνέχεια ταξινομεί πρόσθετες παρατηρήσεις στον επιλεγόμενο αριθμό κλάσεων ή κατηγοριών.

Επίσης, επισημαίνεται ότι τα δεδομένα που δέχεται ο αλγόριθμος ως είσοδο είναι κατηγοριοποιημένα σε κλάσεις.

Στην πραγματική ζωή, οι αλγόριθμοι ταξινόμησης μπορούν να αξιοποιηθούν σε πολλούς τομείς και λύνουν προβλήματα αναγνώρισης προσώπου και φωνής, αναγνώριση ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου (spam emails), ιατρικών διαγνώσεων κ.α.

3.2.4.1 Λογιστική παλινδρόμηση (Logistic regression)

Η λογιστική παλινδρόμηση είναι μια στατιστική μέθοδος που χρησιμοποιείται για την κατασκευή αλγορίθμων μηχανικής μάθησης όπου η εξαρτημένη μεταβλητή είναι διχοτομική, δηλαδή δυαδική. Είναι χρήσιμη για την περιγραφή της σχέσης μιας εξαρτημένης μεταβλητής και μιας ή παραπάνω ανεξάρτητων μεταβλητών. Δεδομένου ότι οι κλάσεις είναι διακριτές στα προβλήματα ταξινόμησης, ο στόχος είναι να βρεθούν τα όρια απόφασης μεταξύ των κλάσεων, τα οποία μπορεί να είναι αρκετά περίπλοκα και μη γραμμικά, ανάλογα με την περίπτωση. Στην περίπτωση της λογιστικής παλινδρόμησης τα όρια είναι γραμμικά, όμως γενικά υπάρχουν και άλλες παραδοχές. Άρα θα μπορούσαν να περιγραφούν ως υπερεπίπεδα στον πολυδιάστατο χώρο χαρακτηριστικών, όπου η διάσταση του προσδιορίζεται από τον αριθμό των στοιχείων του διανύσματος χαρακτηριστικών ενός διανύσματος εκπαίδευσης (Rao et al., 2021).

Η έξοδος της λογιστικής παλινδρόμησης βρίσκεται στο διάστημα [0,1] αφού ουσιαστικά πρόκειται για την πιθανότητα των κλάσεων με βάση κάποιες εξαρτώμενες μεταβλητές. Στην παρακάτω εξίσωση το θ είναι η εκπαιδευόμενη παράμετρος και το X είναι η τιμή εισόδου (Jessica, 2022):

$$h_\theta(X) = \frac{1}{1 + e^{-\theta X}}$$

Η έξοδος της εξίσωσης, δίνει την τιμή της πρόβλεψης. Όσο η τιμή αυτή είναι κοντά στο 1, η παρατήρηση είναι πιθανώς θετική (ανήκει στην κλάση 1) και όσο η τιμή είναι κοντά στο 0, η παρατήρηση ανήκει στην αρνητική κλάση (κλάση 0).

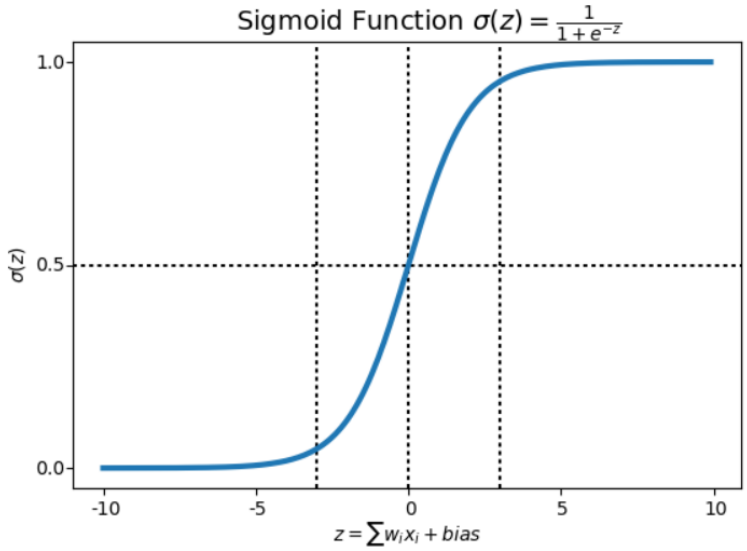
Στη λογιστική παλινδρόμηση χρησιμοποιείται η συνάρτηση απώλειας λογαριθμικής πιθανότητας για να βελτιστοποιηθεί η απόδοση της. Ο μαθηματικός τύπος της συνάρτησης αυτής δίνεται παρακάτω (Jessica, 2022):

$$J(\theta) = \left(-\frac{1}{m}\right)\sum_{m}^{i=1}(y^i \log p^i + (1-y^i)\log(1-p^i))$$

,όπου m ο αριθμός των δειγμάτων, y^i ο συμβολισμός του i δείγματος, p^i η τιμή πρόβλεψης του i δείγματος.

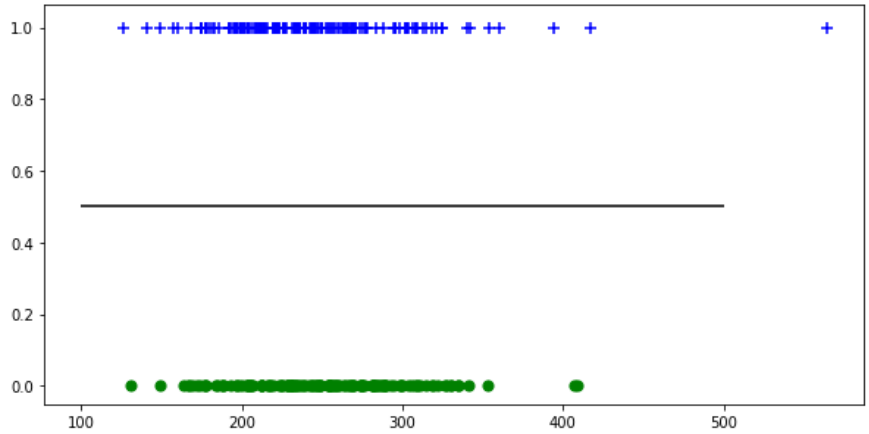
Στη συνέχεια γίνεται πρόσθεση της απώλειας όλων των θεμάτων, υπολογίζεται ο μέσος όρος και προστίθεται το αρνητικό πρόσημο για να συμβολίζει το κόστος. Ο στόχος είναι η ελαχιστοποίηση της τετραγωνικής συνάρτησης κόστους $J(\theta)$, αφού όσο η εξίσωση είναι μικρότερη, ταιριάζει καλύτερα στο σύνολο δεδομένων.

Στην πράξη, ο αλγόριθμος λογιστικής παλινδρόμησης αναλύει τις σχέσεις μεταξύ των μεταβλητών, χρησιμοποιώντας τη σιγμοειδή συνάρτηση (βλ. Εικόνα 3.1) για να εκφράσει τα αποτελέσματα με πιθανότητες. Έτσι, σε μια δυαδική κατηγοριοποίηση, ο πληθυσμός χωρίζεται σε δύο ομάδες, όπου οι παρατηρήσεις με πιθανότητα μεγαλύτερη του 0.5 ανήκουν στην κλάση 1 και οι παρατηρήσεις με πιθανότητα μικρότερη του 0.5 ανήκουν στην κλάση 0 (Jessica, 2022).



Εικόνα 3.1: Διάγραμμα σιγμοειδούς συνάρτησης (Pant, 2019)

Το γραμμικό όριο απόφασης, το οποίο χωρίζει τις δύο κλάσεις, βρίσκεται έπειτα από την κατάταξη των δεδομένων από τη σιγμοειδή συνάρτηση και είναι το ζητούμενο για την κατηγοριοποίηση των δεδομένων από τον αλγόριθμο λογιστικής παλινδρόμησης, αφού αξιοποιείται για την πρόβλεψη των κλάσεων μελλοντικών δεδομένων. Ένα παράδειγμα γραμμικού ορίου φαίνεται στην Εικόνα 3.2.



Εικόνα 3.2: Λογιστική παλινδρόμηση - Γραμμικό όριο απόφασης (Sekhar, 2019)

Η λογιστική παλινδρόμηση αποτελεί δημοφιλή μέθοδο σε προβλήματα ταξινόμησης διότι διαθέτει πολλά πλεονεκτήματα. Κάποια από αυτά είναι η ταχύτητα, η απλότητα της λόγω γραμμικών ορίων απόφασης και η ικανότητα της να είναι λιγότερη επιρρεπής στην υπερπροσαρμογή λόγω αυτών των ορίων.

Όλα τα παραπάνω κάνουν τη λογιστική παλινδρόμηση πολύ διαδεδομένη σε προβλήματα ταξινόμησης, ωστόσο έχει και τα αρνητικά της σημεία. Κάποιες φορές, οι απλές υποθέσεις μοντελοποίησης μπορεί να οδηγήσουν σε υποπροσαρμογή σε περιπτώσεις μεγάλων και σύνθετων συνόλων δεδομένων (Rao et al., 2021).

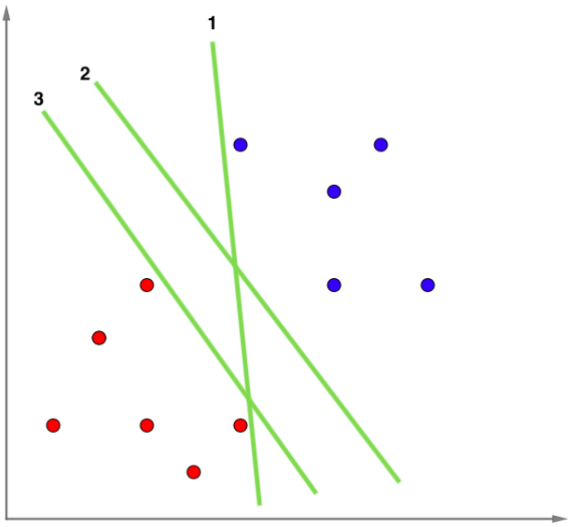
Η υπερπροσαρμογή και υποπροσαρμογή θα αναλυθούν στο υποκεφάλαιο 2.2.6.

3.2.4.2 Μηχανή διανυσμάτων υποστήριξης (Support Vector Machines)

Η μηχανή διανυσμάτων υποστήριξης ή αλλιώς SVM πρόκειται για έναν αλγόριθμο επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιείται σε προβλήματα ταξινόμησης. Αξιοποιείται για την ταξινόμηση των σημείων δεδομένων με τη βοήθεια του υπερεπιπέδου, το οποίο βασίζεται στον αριθμό των διαστάσεων των δεδομένων (π.χ.

για τον δυσδιάστατο χώρο είναι μια γραμμή και για τον τρισδιάστατο είναι ένα επίπεδο) (Kazi, 2020). Όπως και τα υπόλοιπα μοντέλα μηχανικής μάθησης, έτσι και το SVM, παίρνει ως είσοδο δεδομένα εκπαίδευσης ήδη χωρισμένα σε κατηγορίες και έπειτα είναι σε θέση να κατηγοριοποιήσει νέα κείμενα. Ωστόσο, είναι σημαντικό να σημειωθεί πως μπορεί να χρησιμοποιηθεί και για πολυγραμμική ταξινόμηση και όχι μόνο για δυαδική (διαχωρισμός δεδομένων σε δύο μονάχα κατηγορίες).

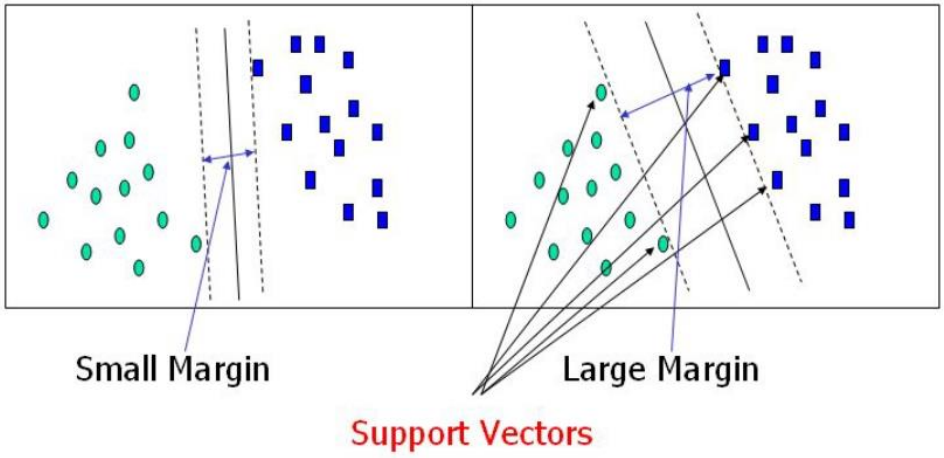
Για να γίνει πιο κατανοητή η έννοια του υπερεπιπέδου, έστω ότι πρέπει να διαχωριστούν τα μπλε σημεία από τα κόκκινα του διαγράμματος στην Εικόνα 3.3.



Εικόνα 3.3: Υπερεπίπεδο – SVM (Kazi, 2020)

Υπάρχουν πολλοί τρόποι να διαχωριστούν αυτά τα σημεία από μια γραμμή, όμως ας πούμε ότι έχουμε τρεις από αυτούς, όπως φαίνεται στην Εικόνα 3.3. Διαισθητικά, η γραμμή 2 φαίνεται και ο καλύτερος τρόπος διαχωρισμού των σημείων, καθώς δεν είναι πιο κοντά σε μια από τις δύο κατηγορίες σημείων, όπως συμβαίνει με τη γραμμή 3 αλλά ούτε είναι κοντά σε κάποια σημεία όπως η γραμμή 1. Με λίγα λόγια, η γραμμή 2 διαισθητικά βρίσκεται σε μέγιστη απόσταση τόσο από τα κόκκινα, όσο και από τα μπλε σημεία. Άρα, αυτό που χρειάζεται ο αλγόριθμος για να έχει την καλύτερη απόδοση είναι να βρεθεί το βέλτιστο υπερεπίπεδο που θα χωρίζει τα μπλε σημεία από τα κόκκινα, δηλαδή αυτό που θα έχει το μέγιστο περιθώριο (margin) και από τις δύο κατηγορίες σημείων και θα είναι εύκολο στον υπολογισμό (Kazi, 2020). Για να μεγιστοποιηθεί το περιθώριο του ταξινομητή χρησιμοποιούνται τα διανύσματα υποστήριξης, τα οποία επηρεάζουν τον προσανατολισμό του επιπέδου και η διαγραφή

τους αλλάζει τη θέση του υπερεπιπέδου (Gandhi, 2018). Η οπτική αναπαράσταση τους φαίνεται στην Εικόνα 3.4.



Εικόνα 3.4: Παράδειγμα περιθωρίων και διανυσμάτων στήριξης του μοντέλου SVM (Gandhi, 2018)

Στο μοντέλο SVM ισχύει ότι αν η τιμή της γραμμικής συνάρτησης είναι μεγαλύτερη από το 1, την ταυτίζουμε με τη μια κλάση και αντίστοιχα αν έχει την τιμή -1 τότε την ταυτίζουμε με τη δεύτερη κλάση. Οπότε οι τιμές της εξόδου της γραμμικής συνάρτησης βρίσκονται στο διάστημα [-1,1]. Η συνάρτηση απώλειας (loss function) του μοντέλου, η οποία βοηθά στη μεγιστοποίηση του περιθωρίου (margin) δίνεται από τον παρακάτω μαθηματικό τύπο (Gandhi, 2018):

$$c(x, y, f(x)) = \begin{cases} 0, & \alpha\nu\ y * f(x) \geq 1 \\ 1\ -\ y * f(x), & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Ο παραπάνω τύπος μεταφράζεται ως εξής: αν η πραγματική τιμή και η τιμή πρόβλεψης έχουν το ίδιο πρόσημο, το κόστος είναι 0, αλλιώς υπολογίζεται από την πράξη $1\ -\ y * f(x)$.

Αν προστεθεί και η παράμετρος κανονικοποίησης, η συνάρτηση κόστους θα πάρει την ακόλουθη μορφή:

$$min_w \|w\|^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)$$

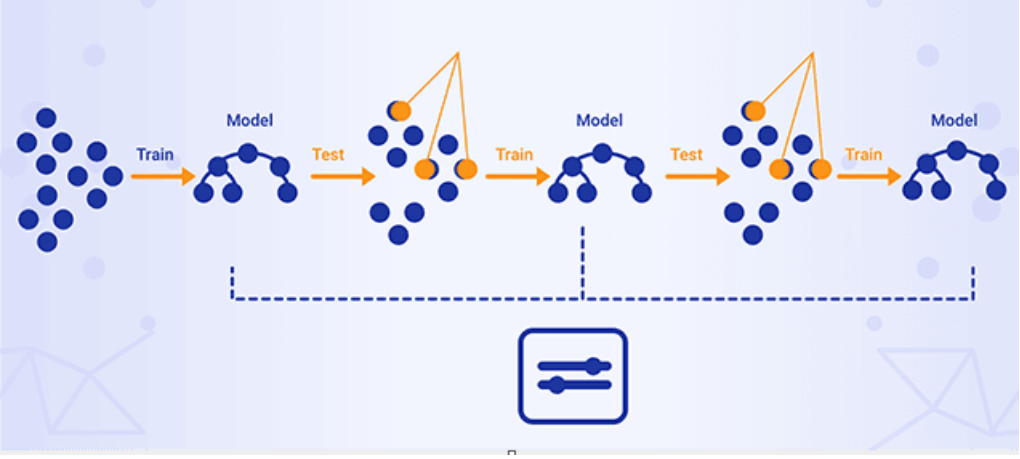
Σε σύγκριση με τους νεότερους αλγορίθμους (π.χ. νευρωνικά δίκτυα), έχει δύο κύρια πλεονεκτήματα. Αυτά είναι η υψηλότερη ταχύτητα και η μεγαλύτερη απόδοση σε περιορισμένο αριθμό δειγμάτων, πράγμα που κάνει το SVM κατάλληλο για προβλήματα ταξινόμησης κειμένου, αφού εκεί συναντάμε συνήθως σύνολα δεδομένων μικρότερου μεγέθους (το πολύ μερικές χιλιάδες δείγματα) (MonkeyLearn, 2017). Όμως, δε λειτουργεί τόσο καλά σε μεγαλύτερα σύνολα δεδομένων αφού απαιτεί αρκετό χρόνο για την εκπαίδευση του και δεν είναι αποτελεσματικό σε δεδομένα με «θόρυβο» (Bambrick, 2016).

Gradient Boosting

Ο αλγόριθμος Gradient Boosting συνδυάζει πολλά αδύναμα μοντέλα πρόβλεψης, με σκοπό τη δημιουργία ενός ισχυρού τελικού μοντέλου, δηλαδή χρησιμοποιεί τις αδύναμες υποθέσεις για να βελτιστοποιεί με επαναληπτικό τρόπο τον αλγόριθμο. Πρόκειται στην ουσία για έναν συνδυασμό μεμονωμένων δέντρων απόφασης (decision trees), όπου κάθε δέντρο προσπαθεί να ελαχιστοποιήσει τα σφάλματα του προηγούμενου δέντρου. Ο βασικός στόχος του αλγόριθμου Gradient Boosting είναι να βελτιστοποιήσει τη συνάρτηση απώλειας (loss function) (Wallstreetmojo Editorial Team, 2022). Τα τρία βασικά στοιχεία του αλγορίθμου είναι τα παρακάτω:

- **Η συνάρτηση απώλειας (loss function):** Πρόκειται για μια συνάρτηση που διαφοροποιείται ανάλογα με το πρόβλημα. Ως παράδειγμα συνάρτησης απώλειας σε ένα πρόβλημα ταξινόμησης μπορεί να οριστεί η λογαριθμική απώλεια (Wallstreetmojo Editorial Team, 2022).
- **Οι weak learners:** Πρόκειται για τους «αδύναμους μαθητές» που συνδυάζονται για την εύρεση του ισχυρού «μαθητευόμενου». Στην περίπτωση μας οι weak learners είναι τα δέντρα απόφασης, τα οποία συνδέονται και κάθε ένα από αυτά προσπαθεί να ελαχιστοποιήσει το σφάλμα του τελευταίου. Πρόκειται για μια αργή αλλά ακριβής διαδικασία (Gaurav, 2021).

- **Το προσθετικό μοντέλο (additive model):** Πρόκειται για την επαναληπτική προσθήκη των δέντρων (weak learners) ένα τη φορά. Με κάθε επανάληψη μειώνεται η τιμή της συνάρτησης απώλειας, συνεπώς με αυτόν τον τρόπο η διαδικασία καταλήγει στο ισχυρό τελικό μοντέλο (ο ζητούμενος «μαθητευόμενος») (Kurama, 2020). Αυτή η διαδικασία αναπαριστάται γραφικά στην Εικόνα 3.5.



Εικόνα 3.5: Οπτικοποίηση αλγορίθμου Gradient Boosting (Akira.AI, 2020)

Τα προτερήματα του αλγορίθμου Gradient Boosting είναι η ακρίβεια των προβλέψεων του και η μεγάλη ευελιξία, αφού μπορεί να χρησιμοποιηθεί σε διαφορετικές συναρτήσεις απώλειας. Επίσης, δεν απαιτεί προ-επεξεργασία των δεδομένων και λειτουργεί επιτυχημένα με κατηγορικές αλλά και αριθμητικές τιμές. Από την άλλη μεριά, φαίνεται πως πολύ συχνά υπερτονίζει τις ακραίες τιμές με αποτέλεσμα να οδηγεί σε υπερπροσαρμογή, απαιτεί αρκετό χρόνο και μνήμη λόγω των πολλών μοντέλων και χρειάζεται ρύθμιση υπερπαραμέτρων λόγω της ευέλικτης φύσης του (Kurama, 2020). Η υπερπροσαρμογή θα αναλυθεί στο υποκεφάλαιο 2.2.6.

3.2.4.3 XGBoost (eXtreme Gradient Boosting)

Πρόκειται για μια παραλλαγή του αλγορίθμου Gradient Boosting. Ο αλγόριθμος είναι βασισμένος σε δέντρα αποφάσεων και χρησιμοποιεί τη διαδικασία boosting για να επιλύει πολλά προβλήματα της επιστήμης των δεδομένων με ακριβή και γρήγορο τρόπο (Brownlee, 2016). Κάποια από τα χαρακτηριστικά του XGBoost είναι τα εξής (Pedamkar, 2022):

- Δημιουργεί παράλληλα δέντρα αποφάσεων.
- Μπορεί να εφαρμοστεί σε πολύ μεγάλα σύνολα δεδομένων αφού χρησιμοποιεί τη δύναμη της παράλληλης επεξεργασίας και την ισχύ υπολογιστών πολλαπλών πυρήνων.
- Μπορεί να ανιχνεύει και να μαθαίνει από μη γραμμικά μοτίβα δεδομένων.
- Περιλαμβάνει τεχνικές κανονικοποίησης ώστε να μπορεί να γενικεύσει επαρκώς και συνεπώς, να αποφευχθεί η υπερπροσαρμογή του μοντέλου.
- Έχει ενσωματωμένο χειρισμό ελλιπών τιμών.

3.2.5 Αξιολόγηση μοντέλων στην Μηχανική Μάθηση

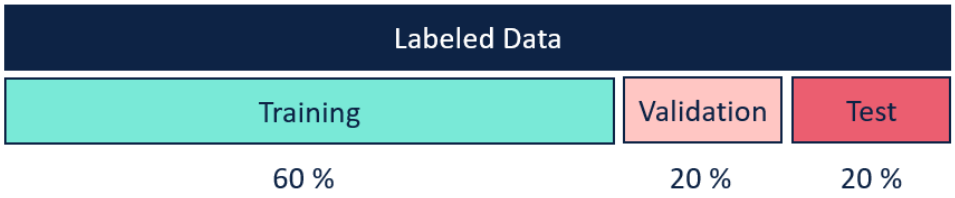
Η αξιολόγηση των μοντέλων Μηχανικής Μάθησης αποτελεί πολύ σημαντικό μέρος της διαδικασίας κατασκευής τους. Επικεντρώνεται στη σύγκριση της ορθότητας των μοντέλων και την επιλογή εκείνου με την υψηλότερη επίδοση. Οι μέθοδοι που χρησιμοποιούνται για την αξιολόγηση των αλγορίθμων είναι οι εξής: η μέθοδος Hold-Out και η Cross-Validation.

Μέθοδος Hold-Out

Στη μέθοδο Hold-Out το σύνολο δεδομένων χωρίζεται σε τρία διαφορετικά σύνολα: τα δεδομένα εκπαίδευσης (training set), τα δεδομένα επαλήθευσης (validation set) και τα δεδομένα δοκιμής (test set) (SydneyF, 2019).

- **Δεδομένα εκπαίδευσης (training set):** Χρησιμοποιούνται για την εκπαίδευση του μοντέλου.
- **Δεδομένα επαλήθευσης (validation set):** Χρησιμοποιούνται για την ρύθμιση υπερπαραμέτρων των μοντέλων.
- **Δεδομένα δοκιμής (test set):** Χρησιμοποιούνται για την αξιολόγηση των επιδόσεων των μοντέλων σε νέα δεδομένα.

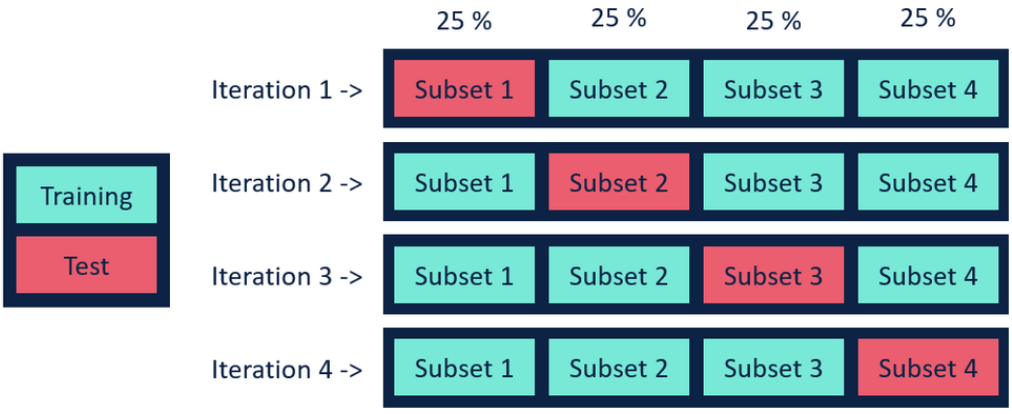
Στην Εικόνα 3.6 δίνεται ένα παράδειγμα χωρισμού των δεδομένων για την εφαρμογή της μεθόδου Hold-Out.



Εικόνα 3.6: Παράδειγμα χωρισμού δεδομένων στην μέθοδο Hold-out (SydneyF, 2019)

Μέθοδος Cross-Validation

Αυτή η μέθοδος στηρίζεται στην διαδικασία της αναδειγματοληψίας. Υπάρχουν διάφορες παραλλαγές της μεθόδου αλλά η k-Fold Cross-Validation είναι η πιο συνήθης. Η μέθοδος αυτή έχει μονό μια παράμετρο, k (συνήθως παίρνει την τιμή 5 ή 10), το οποίο αναφέρεται στο πλήθος των ομάδων στις οποίες πρέπει να χωριστεί το σύνολο δεδομένων. Αφού χωριστεί το σύνολο σε k υποσύνολα (σχεδόν του ίδιου μεγέθους), ένα από τα υποσύνολα κρατείται και η εκπαίδευση του μοντέλου γίνεται με όλα τα υπόλοιπα. Έπειτα, για k σε πλήθος φορές, επιλέγεται ένα υποσύνολο των δεδομένων για την αξιολόγηση του μοντέλου και τα υπόλοιπα χρησιμοποιούνται για την εκπαίδευση του μοντέλου (Brownlee, 2018). Η διαδικασία αυτή περιγράφεται και γραφικά στην Εικόνα 3.7.



Εικόνα 3.7: Διαδικασία μεθόδου k-Fold Cross-Validation (SydneyF, 2019)

Μετρικές αξιολόγησης μοντέλων ταξινόμησης (evaluation metrics)

Οι μετρικές αξιολόγησης βοηθούν στην επιλογή του κατάλληλου αλγορίθμου για το δεδομένο σύνολο δεδομένων. Παρακάτω αναφέρονται οι κύριες μετρικές αξιολόγησης για προβλήματα ταξινόμησης (Garlapati, 2021), (Jordan, 2017):

- **Classification Accuracy (ορθότητα)**

  Η ορθότητα είναι το κυριότερο μέτρο αξιολόγησης. Δίνεται από τον λόγο των σωστών προβλέψεων προς τον συνολικό αριθμό των προβλέψεων που έκανε το μοντέλο. Δηλαδή από τον τύπο:

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Total\ number\ of\ predictions\ made}$$

  Εάν η ορθότητα είναι υψηλή τότε το μοντέλο μπορεί να αξιοποιηθεί και σε άλλες εφαρμογές. Αν δεν είναι, το μοντέλο δεν κατατάσσει σωστά τα δεδομένα και δεν ταιριάζει στο συγκεκριμένο πρόβλημα ταξινόμησης.

  Πολλές φορές η ορθότητα δεν αποτελεί την καλύτερη μέτρηση της επιτυχίας ενός αλγορίθμου. Αυτό μπορεί να γίνει κατανοητό με το επόμενο παράδειγμα. Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων με email και θέλουμε να τα κατατάξουμε σε spam / not spam. Αν το 99% των email είναι not spam και μόνο 1% είναι spam και εμείς χρησιμοποιήσουμε την ορθότητα για την πρόβλεψη ενός spam email, η τιμή της θα είναι πάντα κοντά στο 99% (Wadhawan, 2019). Οι άνισα κατανεμημένες κλάσεις είναι πολύ συχνό φαινόμενο στα προβλήματα μηχανικής μάθησης και καθιστούν την ορθότητα αναξιόπιστο μέτρο.

- **Confusion matrix (Πίνακας σύγχυσης)**

  Ο πίνακας σύγχυσης είναι ένας πίνακας $N \times N$, όπου N είναι ο επιθυμητός αριθμός κλάσεων (για παράδειγμα, σε προβλήματα δυαδικής ταξινόμησης θα έχουμε μια μήτρα 2×2) και χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός αλγορίθμου. Οι όροι του πίνακα ορίζουν τον αριθμό των λανθασμένων αλλά και των σωστών προβλέψεων (βλ. Εικόνα 3.6). Οι όροι αυτοί είναι οι παρακάτω:

		PREDICTED	
		0 (Negative)	1 (Positive)
ACTUAL	0 (Negative)	TN	FP
	1 (Positive)	FN	TP

Εικόνα 3.8: Παράδειγμα πίνακα σύγχυσης (confusion matrix) (Wadhawan, 2019)

TP: ο αριθμός των σωστών ταξινομήσεων στη θετική κλάση (true positive)
TN: ο αριθμός των σωστών ταξινομήσεων στην αρνητική κλάση (true negative)
FP: ο αριθμός των λανθασμένων ταξινομήσεων στη θετική κλάση (false positive)
FN: ο αριθμός των λανθασμένων ταξινομήσεων στην αρνητική κλάση (false negative)

Μέσω αυτών των όρων του πίνακα σύγχυσης γίνεται ο υπολογισμός πολλών μετρικών, χρήσιμων για την αξιολόγηση του αλγορίθμου. Αυτές οι μετρικές είναι οι εξής:

- Η **ορθότητα (ή accuracy)**, η οποία έχει αναφερθεί παραπάνω, δίνεται από τον μέσο όρο των τιμών TP(True Positive) και TN(True Negative). Δηλαδή,

$$accuracy = \frac{true\ positive + true\ negative}{total\ of\ predictions}$$

- Η **λανθασμένη ταξινόμηση (ή misclassification)**, η οποία ισούται με τις προβλέψεις που δεν ήταν σωστές, δηλαδή δίνεται από τη σχέση:

$$misclassification = \frac{false\ positive + false\ negative}{total\ of\ predictions}$$

Ή αλλιώς,

$$misclassification = 1 - accuracy$$

- Η **ακρίβεια (ή precision)**, η οποία ορίζεται ως το κλάσμα των αληθώς θετικών προβλέψεων από το σύνολο όλων των δειγμάτων που προβλέφθηκε ότι ανήκουν σε μια συγκεκριμένη κλάση:

$$precision = \frac{false\ positive}{true\ positive + false\ positive}$$

- Η **ανάκληση (ή recall)**, που ορίζεται ως το κλάσμα των δειγμάτων που προβλέφθηκε ότι ανήκουν σε μια συγκεκριμένη κλάση από το σύνολο

των δειγμάτων που πραγματικά ανήκουν στην κλάση αυτή. Δηλαδή, ισούται με:

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

- Το **F1 score**, είναι ο αρμονικός μέσος όρος των **recall** και **precision** και δίνεται από την παρακάτω σχέση:

$$F1\ score = \frac{2 * recall * precision}{recall + precision}$$

Ή αλλιώς,

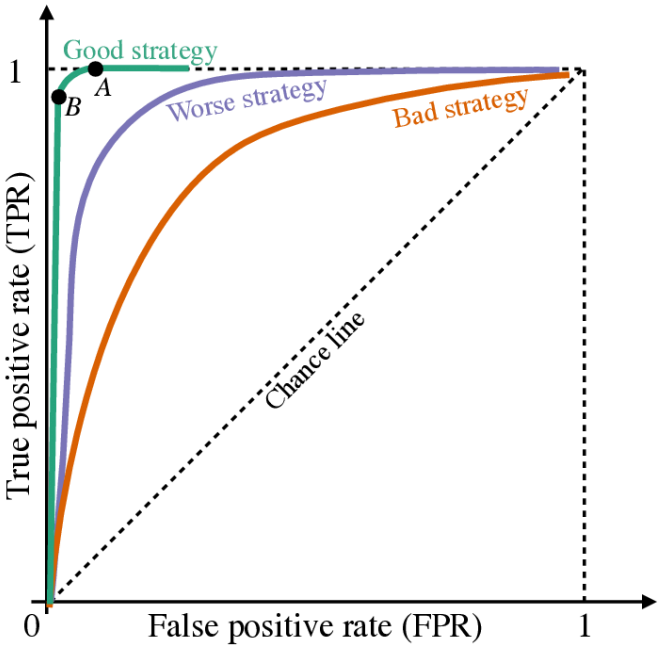
$$F1\ score = \frac{2 * true\ positive}{2 * true\ positive + false\ negative + false\ positive}$$

Όσο μεγαλύτερο είναι το F1 score τόσο καλύτερη η απόδοση του μοντέλου. Πρόκειται για το καλύτερο μέτρο αυτής της απόδοσης και συνεπώς ο υπολογισμός του μπορεί να γίνει χωρίς να υπολογιστεί η ακρίβεια και η ανάκληση ξεχωριστά.

- Τέλος, μια ακόμη τεχνική μέτρησης είναι το εμβαδόν της περιοχής από την καμπύλη ROC και ονομάζεται AUC (Area under curve). Χρησιμοποιεί τα κατώτατα όρια για τον προσδιορισμό του ποσοστού ψευδώς θετικών (False Positive Rate) αποτελεσμάτων και του ποσοστού αληθώς θετικών (True Positive Rate) αποτελεσμάτων. Το TPR στην ουσία είναι η ανάκληση (recall) και το FPR ορίζεται από τον παρακάτω τύπο:

$$FPR = \frac{false\ positive}{false\ posotive + true\ negative}$$

Στην Εικόνα 3.7 το TPR απεικονίζεται στον άξονα y και το FPR στον άξονα x. Το ιδανικό σημείο είναι η πάνω αριστερή γωνία όπου το TPR είναι 1 και το FPR είναι 0. Αυτό σημαίνει ότι υπάρχουν false negative και false positive.



Εικόνα 3.9: Καμπύλη ROC (Gwirtz et al., 2020)

Αν το μοντέλο είναι άριστο, τότε η τιμή του AUC είναι κοντά στο 1 και συνεπώς η ταξινόμηση είναι καλή. Αντίθετα, αν έχει τιμή κοντά στο 0, το μοντέλο δεν είναι επιτυχημένο και η ταξινόμηση δεν είναι η επιθυμητή, δηλαδή ταξινομεί τα δεδομένα σε λάθος κλάση. Τέλος, αν το AUC είναι ίσο με 0.5, τότε ο αλγόριθμος δεν έχει την ικανότητα διαχωρισμού των κλάσεων (Narkhede, 2018).

3.2.6 Σφάλματα

Σφάλμα διακύμανσης (Variance error)

Το σφάλμα διακύμανσης είναι η μεταβλητότητα της πρόβλεψης του μοντέλου για μια δοθείσα παρατήρηση και μας δείχνει τη διασπορά των δεδομένων. Ένα μοντέλο με υψηλή διακύμανση δίνει μεγάλη προσοχή στα δεδομένα εκπαίδευσης και δε γενικεύει σε νέα δεδομένα (Sahani, 2020).

Σφάλμα μεροληψίας (Bias error)

Πρόκειται για τη διαφορά της μέσης πρόβλεψης του αλγορίθμου και της πραγματικής τιμής. Ένα μοντέλο με υψηλό bias (μεροληψία) δίνει λίγη προσοχή στα δεδομένα εκπαίδευσης και υπεραπλουστεύει το μοντέλο (Sahani, 2020).

Υποπροσαρμογή (Underfitting)

Η υποπροσαρμογή είναι η δυσκολία του μοντέλου να αντιληφθεί την υποκειμενική τάση των δεδομένων. Έτσι, σε αυτήν την περίπτωση, ο αλγόριθμος αποδίδει πολύ καλά στα δεδομένα εκπαίδευσης αλλά κάνει ανεπιτυχείς προβλέψεις στα δεδομένα δοκιμής. Τα μοντέλα που έχουν την τάση να υποπροσαρμόζονται, παρουσιάζουν υψηλό bias και χαμηλό variance (Sahani, 2020).

Κάποιες τεχνικές αντιμετώπισης της υποπροσαρμογής είναι οι παρακάτω (Geeksforgeeks, 2022):

1. Αύξηση της πολυπλοκότητας του μοντέλου
2. Χρήση τεχνικών εξαγωγής χαρακτηριστικών
3. Αφαίρεση του «θορύβου» στα δεδομένα
4. Αύξηση του χρόνου εκπαίδευσης του μοντέλου
5. Ρύθμιση των παραμέτρων του μοντέλου

Υπερπροσαρμογή (Overfitting)

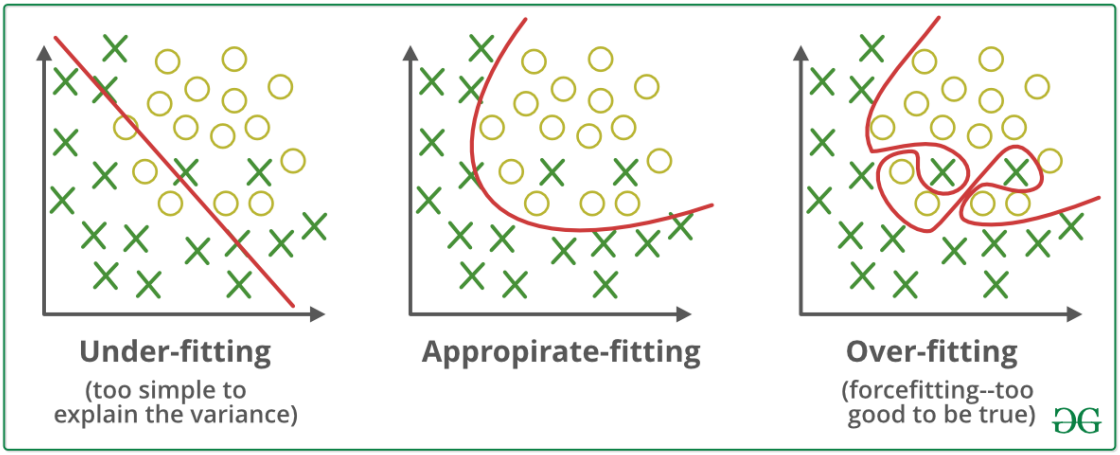
Η υπερπροσαρμογή ορίζεται ως η υπερβολική προσαρμογή των μοντέλων μηχανικής μάθησης στα δεδομένα εκπαίδευσης (training set), η οποία οδηγεί σε ανακριβείς προβλέψεις στα δεδομένα δοκιμής (test set). Τα μοντέλα που έχουν την τάση να υπερπροσαρμόζονται, παρουσιάζουν χαμηλό bias και υψηλό variance (Sahani, 2020).

Κάποιες τεχνικές αντιμετώπισης της υποπροσαρμογής είναι οι παρακάτω (Geeksforgeeks, 2022):

1. Αύξηση των δεδομένων εκπαίδευσης
2. Μείωση της πολυπλοκότητας του μοντέλου
3. Κανονικοποίηση των δεδομένων
4. Διακοπή της διαδικασίας της εκπαίδευσης σε περίπτωση αύξησης των απωλειών

5. Ρύθμιση των παραμέτρων του μοντέλου

Στην Εικόνα 3.8 αναπαρίστανται γραφικά τα σφάλματα της υπερπροσαρμογής και υποπροσαρμογής.



Εικόνα 3.10: Οπτική παρουσίαση των overfitting και underfitting

3.3 Ανάλυση συναισθήματος

3.3.1 Ορισμός

Η ανάλυση συναισθήματος ή αλλιώς εξόρυξη γνώμης, πρόκειται για μια μέθοδο επεξεργασία φυσικής γλώσσας (Natural Language Processing ή αλλιώς NLP) που αναγνωρίζει αν το συναίσθημα στα δεδομένα είναι θετικό, αρνητικό ή ουδέτερο. Χρησιμοποιείται κυρίως σε δεδομένα κειμένων και αποτελεί χρήσιμο εργαλείο για εταιρίες που επιθυμούν να ενημερώνονται για τη γνώμη και το συναίσθημα των καταναλωτών για τα προϊόντα τους.

Ωστόσο, στις μέρες μας, η μέθοδος αυτή χρησιμοποιείται ακόμη και σε προβλήματα Τεχνητής νοημοσύνης (Artificial Intelligence). Οι Yakaew, Dailey και Racharak (2021), δείχνουν με τη μελέτη τους πως είναι δυνατή ακόμα και η δημιουργία ενός μοντέλου αναγνώρισης των συναισθημάτων σε πραγματικό χρόνο μέσω της φωνής και των κινήσεων του προσώπου.

Η ανάλυση συναισθήματος φαίνεται να λειτουργεί με μεγαλύτερη αποτελεσματικότητα σε κείμενα με υποκειμενικό χαρακτήρα και όχι αντικειμενικό. Αυτό συμβαίνει διότι

κατά βάση, ο αντικειμενικός λόγος δεν εκδηλώνει τα συναισθήματα και τη διάθεση του ατόμου. Η ερεύνα των υποκειμενικών προτάσεων είναι πολύ δημοφιλής στην εξόρυξη δεδομένων, αφού φιλοξενεί πολλές προκλήσεις (Nguyen et al., 2018).

3.3.2 Κατηγορίες ταξινόμησης

Στη σύγχρονη βιβλιογραφία, η ταξινόμηση συναισθήματος από ένα κείμενο μπορεί να γίνει με τρείς διαφορετικούς τρόπους, ανάλογα με την προσέγγιση που επιλέγεται (βλ. Εικόνα 3.9). Οι τρόποι αυτοί περιγράφονται παρακάτω:

1. **Ανάλυση σε επίπεδο κειμένου**

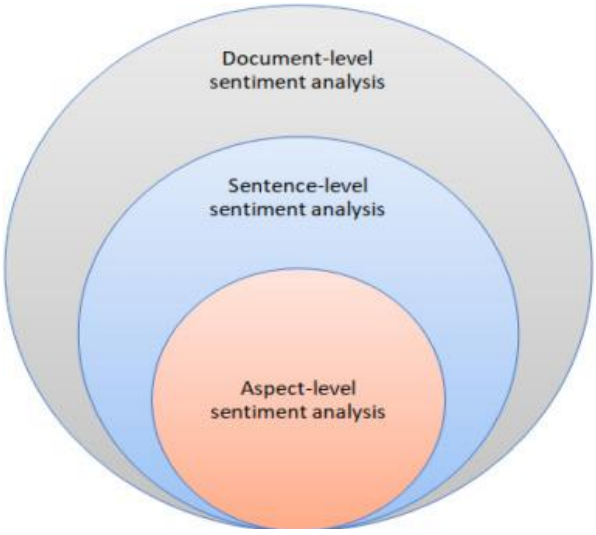
   Πρόκειται για την πιο απλή μέθοδο ανάλυσης αφού εξάγει ένα μονάχα συναίσθημα για ένα ολόκληρο κείμενο. Δηλαδή, αποδίδει στην ουσία τη γνώμη του συγγραφέα για το αντικείμενο για το οποίο μιλάει ή προβληματίζεται.

2. **Ανάλυση σε επίπεδο πρότασης**

   Η μέθοδος αυτή δεν αντιμετωπίζει το κείμενο σαν μια ολότητα αλλά βρίσκει το συναίσθημα του συγγραφέα αναλύοντας κάθε πρόταση ξεχωριστά. Πρόκειται για μια πιο λεπτομερή μέθοδο από την προηγούμενη με καλύτερα αποτελέσματα, αφού ένα κείμενο σπάνια εκφράζει ένα μόνο συναίσθημα.

3. **Ανάλυση σε επίπεδο οντότητας και άποψης**

   Οι προηγούμενες δύο μέθοδοι λειτουργούν καλά όταν σε ένα κείμενο ή μια πρόταση, υπάρχει ένα θέμα συζήτησης. Πολλές φορές όμως οι άνθρωποι μπορεί να εκφέρουν περισσότερες από μια απόψεις για ένα θέμα. Η ανάλυση αυτή αποσκοπεί στην άντληση όλων των συναισθημάτων από ένα κείμενο για το θέμα για το οποίο αναπτύσσεται (Feldman, 2013).



Εικόνα 3.11: Επίπεδα ανάλυσης κειμένου (Birjali et al., 2021)

3.3.3 Τεχνική ανάλυσης με βάση λεξικά

Η τεχνική ανάλυσης με βάση λεξικά, ή αλλιώς Lexicon-based approach, υπολογίζει την πολικότητα του κειμένου βασιζόμενη σε λεξικά, που δημιουργούνται χειροκίνητα και αποτελούνται από λέξεις με το αντίστοιχο βαθμό συναισθήματος.

Η διαδικασία εύρεσης της ζητούμενης πολικότητας ξεκινά με τη διάταξη των λέξεων και των βαθμών συναισθήματος σε ένα λεξικό. Ακολουθεί ο υπολογισμός του βαθμού πολικότητας του δοθέντος κειμένου με τη χρήση του λεξικού και τέλος ο υπολογισμός του τελικού βαθμού συναισθήματος (Taboada et al., 2011). Η τεχνική αυτή έχει το πλεονέκτημα ότι δε χρειάζεται εκπαίδευση, σε αντίθεση με τα μοντέλα Μηχανικής Μάθησης.

Το πρόβλημα όμως προκύπτει στο ότι η μέθοδος αυτή στηρίζεται στο συναισθηματικό υπόβαθρο μιας λέξης και παραβλέπει το συνολικό νόημα μιας πρότασης ή το ειρωνικό ύφος που μπορεί να έχει. Για παράδειγμα, η κριτική «Περίμενα απολαυστικό το φαγητό, αλλά τελικά απογοητεύτηκα», θα κατηγοριοποιηθεί με λάθος τρόπο από την εν λόγω τεχνική μιας και περιέχει τη λέξη «απολαυστικό», που είναι θετική συναισθηματικά αλλά στην πραγματικότητα η πρόταση δεν αποτελεί καλή κριτική.

3.3.3.1 Με χρήση του λεξικού TextBlob

Το λεξικό TextBlob αποτελεί μια βιβλιοθήκη της Python και χρησιμοποιείται για εφαρμογές επεξεργασίας φυσικής γλώσσας (NLP) σε δεδομένα κειμένου. Πρόκειται για ένα απλό API που βασίζεται στο Natural Language Toolkit, ή αλλιώς NLTK, το οποίο αποτελεί μια συλλογή βιβλιοθηκών και εφαρμογών επεξεργασίας φυσικής γλώσσας για αγγλικά κείμενα.

Το λεξικό αυτό λαμβάνει ως είσοδο μια πρόταση και επιστρέφει δύο τιμές, την πολικότητα (polarity) και την υποκειμενικότητα (subjectivity) (Barai, 2021). Οι τιμές που παίρνει η πολικότητα βρίσκονται στο διάστημα μεταξύ του 1 και του -1, όπου το -1 προσδιορίζει αρνητικές λέξεις (π.χ. «απογοητευτικό», «απαίσιο») και το 1 προσδιορίζει θετικές λέξεις (π.χ. «υπέροχο», «καταπληκτικό»). Αντίστοιχα, οι τιμές της υποκειμενικότητας κυμαίνονται στο διάστημα [0,1], όπου οι πιο υποκειμενικές προτάσεις παίρνουν τιμές κοντά στο 1 και οι λιγότερο αντικειμενικές παίρνουν τιμές κοντά στο 0. Στην πρώτη περίπτωση δηλαδή, φαίνεται ότι η πρόταση εμπεριέχει την προσωπική γνώμη του συγγραφέα και όχι πραγματικές πληροφορίες.

Πέρα από την εξαγωγή αποτελεσμάτων για την πολικότητα και την αντικειμενικότητα, η βιβλιοθήκη TextBlob διαθέτει και άλλες συναρτήσεις που βοηθούν στην προετοιμασία των δεδομένων. Η προετοιμασία των δεδομένων μπορεί να περιλαμβάνει τις παρακάτω ενέργειες (Sharma, 2020):

- Διαχωρισμό του κειμένου σε λέξεις ή προτάσεις.
- Μετατροπή λέξεων από το κείμενό σε ενικό ή πληθυντικό αριθμό.
- Εύρεση του λήμματος των λέξεων.
- Διόρθωση τυχών ορθογραφικών στο κείμενο.
- Εύρεση πλειάδων n διαδοχικών λέξεων (N-grams).

3.3.3.2 Με χρήση του λεξικού Vader

Το λεξικό Vader (Valence Aware Dictionary and Sentiment Reasoner) είναι ένας αναλυτής συναισθήματος που βασίζεται σε κανόνες (Geeksforgeeks, 2021). Το λεξικό αυτό χρησιμοποιεί τη συναισθηματική φόρτιση των λέξεων του κειμένου σύμφωνα με τον σημασιολογικό τους προσανατολισμό, δηλαδή το αν είναι θετικά ή αρνητικά

φορτισμένες και επιστρέφει μια τιμή που προσδιορίζει αν μια πρόταση εκφράζει θετικό, αρνητικό ή και ουδέτερο συναίσθημα. Η τιμή αυτή λέγεται compound score και υπολογίζεται αθροίζοντας τις 3 βαθμολογίες neg, neu, και pos, οι οποίες προσδιορίζουν το πόσο θετική, ουδέτερη και αρνητική είναι η πρόταση αντίστοιχα, αφού πρώτα κανονικοποιηθούν μεταξύ -1 και +1.Τα κριτήρια απόφασης είναι παρόμοια με το TextBlob, αφού -1 είναι η πιο αρνητική τιμή και το 1 η πιο θετική. Πιο συγκεκριμένα, τα όρια των βαθμολογιών είναι τα εξής:

Θετικό συναίσθημα: $compound\ score >= 0.05$

Ουδέτερο συναίσθημα: $compound\ score > -0.05\ και\ compound\ score < 0.05$

Αρνητικό συναίσθημα: $compound\ score <= -0.05$

Και οι δύο μέθοδοι, TextBlob και Vader, προσφέρουν πολλές δυνατότητες και βοηθούν στην εξόρυξη σημαντικών πληροφοριών που αφορούν τα συναισθήματα που κατακλύζουν ένα σώμα κειμένου. Η εφαρμογή και των δύο αυτών μεθόδων σε ένα δείγμα από σύνολο δεδομένων και η σύγκριση τους, οδηγεί στην επιλογή του καλύτερου δυνατού αποτελέσματος ανάλογα με τις απαιτήσεις του προβλήματος και το σύνολο δεδομένων.

3.3.4 Τεχνική ανάλυσης με βάση τη μηχανική μάθηση

Η τεχνική αυτή έχει να κάνει με την κατασκευή ενός μοντέλου μηχανικής μάθησης για την κατάταξη των δεδομένων κειμένου συνήθως σε δύο κατηγορίες συναισθήματος, θετικό και αρνητικό. Στην τεχνική αυτή αποτελεί σημαντικό βήμα η προετοιμασία των δεδομένων, η οποία αποτελείται από την εκκαθάριση και την κανονικοποίηση τους ώστε να υπάρχει αμεροληψία. Στη συνέχεια ακολουθεί η δημιουργία του μοντέλου και η εκτίμηση των αποτελεσμάτων του.

Το κυριότερο μειονέκτημα της μηχανικής μάθησης είναι ότι τα μοντέλα απαιτούν μεγάλα σύνολα δεδομένων για την εκπαίδευση τους (training set) που θα πρέπει να είναι αμερόληπτα και καλής ποιότητας. Αν αυτό δε συμβεί, το μοντέλο δε θα είναι

αρκετά νοήμων για να αποφέρει τα σωστά αποτελέσματα στο δοκιμαστικό σύνολο δεδομένων (test set) (Shayaa et al., 2018).

3.4 Θεματική κατηγοριοποίηση κειμένου

3.4.1 Ορισμός

Η θεματική κατηγοριοποίηση κειμένου (topic modeling) είναι μια τεχνική επεξεργασίας φυσικής γλώσσας (Natural Language Processing ή αλλιώς NLP), η οποία έχει σκοπό τον καθορισμό θεμάτων (topics) σε ένα σώμα κειμένου, ή την εύρεση λεξικών μοτίβων. Αυτό το πετυχαίνει αξιοποιώντας τη συχνότητα εμφάνισης των λέξεων ή των φράσεων στο κείμενο με σκοπό τον υπολογισμό της αντίστοιχης πιθανότητας εμφάνισης τους σε ένα συγκεκριμένο θέμα και την κατηγοριοποίηση του κειμένου με βάση την ομοιότητα τους. Επίσης, μελετά το πόσο συχνά εμφανίζονται κάποιες λέξεις μαζί, πράγμα που βοηθά αρκετά στην κατανόηση των topics (Zvornicanin, 2022).

Το topic modeling αποτελεί μια μέθοδο μη επιβλεπόμενης μάθησης και οι βασικοί αλγόριθμοι που χρησιμοποιεί είναι η Latent Semantic Analysis (LSA) και η Latent Dirichlet Allocation (LDA). Οι παραπάνω μέθοδοι θα αναλυθούν πιο διεξοδικά παρακάτω.

3.4.2 Latent Semantic Analysis (LSA)

Η Latent Semantic Analysis αποτελεί μια από τις πιο βασικές μεθόδους θεματικής κατηγοριοποίησης κειμένου. Η βασική ιδέα είναι η κατασκευή ενός πίνακα από τα υπάρχοντα δεδομένα, document-term matrix (μήτρα κείμενο -όρος), για τη δημιουργία δύο νέων πινάκων, document-topic matrix (μήτρα κείμενο-θέμα) και topic-term matrix (μήτρα θέμα-όρος).

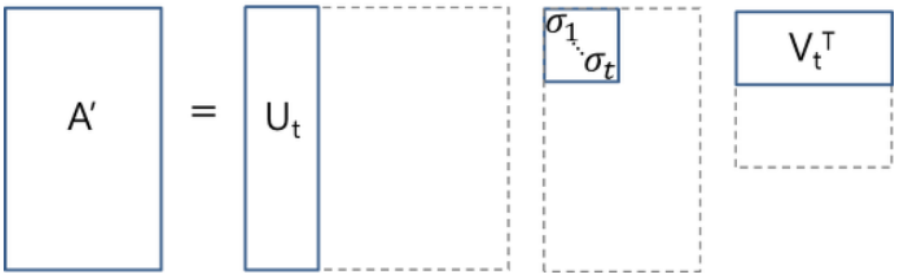
Για το πρώτο βήμα, δηλαδή τη δημιουργία του document-term matrix, λαμβάνονται m έγγραφα και n λέξεις από το λεξιλόγιο και βάση αυτών κατασκευάζεται ένας πίνακας A διαστάσεων $m \times n$, στον οποίο κάθε σειρά αντιπροσωπεύει ένα έγγραφο και κάθε στήλη αφορά μια λέξη.

Στην πιο απλή έκδοση του LSA, κάθε καταχώρηση αφορά τον αριθμό των φορών που η λέξη i εμφανίστηκε στο έγγραφο j. Η πρακτική αυτή όμως, δεν αποτελεί τόσο αξιόπιστο τρόπο εύρεσης των topics, αφού δε λαμβάνεται υπόψη η σημαντικότητα της κάθε λέξης. Για αυτόν τον λόγο, στα LSA μοντέλα πολύ συχνά αυτές οι καταχωρίσεις αντικαθίστανται με το αντίστοιχο TF-IDF σκορ.

Αφού κατασκευαστεί ο πίνακας document-term (πίνακας A), η διαδικασία προχωρά με την εύρεση των topics. Όμως το πρόβλημα είναι ότι κατά πάσα πιθανότητα, ο πίνακας Α θα είναι αραιός και πολύ θορυβώδης σε πολλές διαστάσεις του. Για τον λόγο αυτό, είναι απαραίτητη η εφαρμογή της τεχνικής μείωσης των διαστάσεων του πίνακα, η λεγόμενη τεχνική Singular Value Decomposition (SVD).

Η SVD είναι μια μέθοδος γραμμικής άλγεβρας που παραγοντοποιεί οποιαδήποτε πίνακα M σε ένα γινόμενο τριών ξεχωριστών πινάκων: $M = U * S * V$, όπου S είναι ένας διαγώνιος πίνακας των μοναδικών τιμών του M. Οπότε, αυτό που κάνει η εν λόγω τεχνική είναι να μειώνει τη διάσταση του πίνακα επιλέγοντας μόνο τις t μεγαλύτερες μοναδικές τιμές και κρατώντας μόνο τις πρώτες στήλες t των U και V. Άρα εδώ, το t είναι μια υπερπαράμετρος που επιλέγεται ανάλογα με τον επιθυμητό αριθμό των topics. Οπότε, ο τύπος που δίνει τον Α, θα είναι ο παρακάτω και δίνεται υπό την μορφή πινάκων στην Εικόνα 3.10.

$$A \approx U_t S_t V_t^T$$



Εικόνα 3.12: Εξίσωση πίνακα όρων-εγγράφων A (document-term matrix A) (Xu, 2018)

Σε αυτήν την περίπτωση, το $U \in \mathbb{R} \wedge (m \times t)$ είναι η μήτρα document-term και το $V \in \mathbb{R} \wedge (n \times t)$ είναι η μήτρα topic-term. Οι στήλες αυτών των δύο πινάκων αντιστοιχούν σε ένα topic. Οι σειρές του πίνακα U αντιστοιχούν στους όρους του κειμένου εκφραζόμενους με όρους των θεμάτων και οι σειρές του πίνακα V

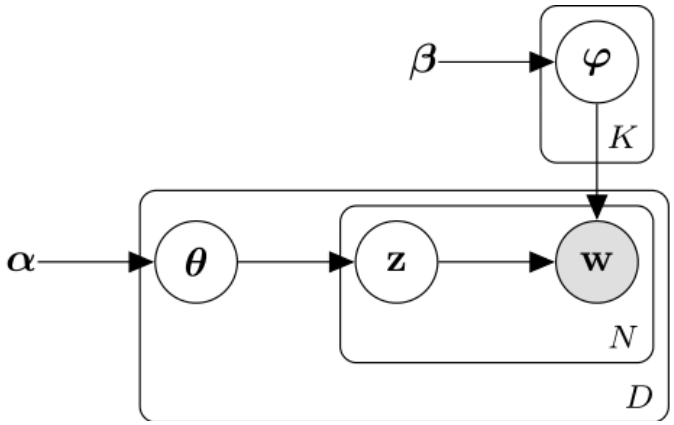
αντιπροσωπεύουν τους διανυσματικούς όρους εκφραζόμενους με όρους των θεμάτων (Xu, 2018).

Η μέθοδος LSA είναι γρήγορη και αποδοτική, αλλά έχει κάποια μειονεκτήματα (Goyal, 2021):

- Αφού πρόκειται για ένα γραμμικό μοντέλο, δεν αποδίδει σωστά σε σύνολα δεδομένων με μη-γραμμικές σχέσεις.
- Προϋποθέτει την τεχνική Singular Value Decomposition (SVD) για την απόδοση σωστών αποτελεσμάτων, κάτι που είναι πιο απαιτητικό υπολογιστικά και δύσκολο να εκσυγχρονιστεί για νέα δεδομένα.
- Δεν μπορούμε να γνωρίζουμε τα topics και το συναίσθημα (θετικό ή αρνητικό) των όρων είναι αυθαίρετο.
- Χρειάζεται μεγάλο σύνολο δεδομένων για να αποδοθεί το μέγιστο δυνατό, με μεγαλύτερη ακρίβεια.
- Προσφέρει λιγότερο αποδοτική αναπαράσταση.

3.4.3 Latent Dirichlet Allocation (LDA)

Η Latent Dirichlet Allocation (LDA) αποτελεί μια πολύ γνωστή μέθοδο ομαδοποίησης χωρίς επίβλεψη η οποία χρησιμοποιείται συνήθως για την ανάλυση κειμένου και χρησιμοποιείται για εργασίες θεματικής κατηγοριοποίησης. Αποτελεί ένα ιεραρχικό μοντέλο Bayes τριών επιπέδων το οποίο περιγράφει μεγάλες συλλογές κειμένων ανά λέξη. Το LDA μοντέλο παρουσιάζεται γραφικά στην Εικόνα 3.11.



Εικόνα 3.13: Γραφική αναπαράσταση του μοντέλου LDA (Mathworks, 2017)

Στην Εικόνα 3.11 ο σκιασμένος κόμβος είναι οι παρατηρούμενες μεταβλητές, οι μη σκιασμένοι κόμβοι είναι η λανθάνουσες μεταβλητές, οι κόμβοι χωρίς περίγραμμα είναι οι παράμετροι, τα βέλη προσδιορίζουν τις εξαρτήσεις μεταξύ των μεταβλητών και οι πλάκες προσδιορίζουν τους επαναλαμβανόμενους κόμβους (Mathworks, 2017).

Κάθε κείμενο περιέχει ένα σύνολο θεμάτων, τα οποία προσδιορίζονται από λέξεις συνδεδεμένες με κάποια πιθανότητα. Ο τρόπος που αναγνωρίζει η LDA τα θέματα του κειμένου στηρίζεται στην εύρεση της δομής των κειμένων. Συνεπώς, για μια συλλογή κειμένων D, ακολουθεί την παρακάτω διαδικασία (Ζαραφέτα, 2019):

1. Για k=1,…,K:
   $\varphi^{(k)} \sim Dirichlet(\beta)$
2. Για κάθε κείμενο $d \in D$:
   A. $\theta_d \sim Dirichlet(a)$
   B. Επιλέγει μια λέξη $w_i \in$ D:
      i. $z_i \sim Discrete(\theta_d)$
      ii. $w_i \sim Discrete\left(\varphi^{(z_i)}\right)$

όπου K ο αριθμός των θεμάτων, $\varphi^{(k)}$ η διακριτή κατανομή πιθανότητας σε ένα σταθερό λεξιλόγιο, θ_d μια κατανομή συγκεκριμένων κειμένων πάνω στα διαθέσιμα θέματα, z_i ένα θέμα που περιέχει τη λέξη w_i και α, β οι υπερπαράμετροι των κατανομών Dirichlet.

Δοθέντων των παραμέτρων α και β, η κοινή κατανομή ενός μείγματος θεμάτων θ, ένα σετ από K θέματα z και ένα σετ K λέξεων w δίνεται από τον τύπο (Blei et al., 2003):

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{K} p(z_n|\theta)p(w_n|z_n, \beta)$$

,όπου $p(z_n\,|\theta)$ είναι τα θ_i για κάθε i τέτοιο ώστε, $z_n^i = 1$. Έπειτα, γίνεται ολοκλήρωση ως προς θ και άθροιση ως προς z για τον υπολογισμό της οριακής κατανομής του εγγράφου (Blei et al., 2003).:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta)\right)d\theta$$

Η LDA επιτυγχάνει υψηλή ποιότητα συνοχής μεταξύ των θεμάτων αλλά αγνοεί κάθε δευτερεύουσα πληροφορία σχετικά με την ομοιότητα των λέξεων (Petterson et al., 2003). Αυτή της η αδυναμία την κάνει να υπολείπεται σε διάφορες πτυχές. Ένα πρόβλημα που εμφανίζεται είναι ότι θέματα που εκτιμώνται για σπάνια εμφανιζόμενες λέξεις μπορεί να είναι αναξιόπιστα. Για παράδειγμα, θα έπρεπε τα θέματα που σχετίζονται με τα συνώνυμα να είναι παρόμοια, έτσι ώστε σε περίπτωση που μια από τις λέξεις είναι σπάνια αλλά η άλλη είναι κοινή, να μπορούν να βελτιωθούν οι εκτιμήσεις με τη σπάνια λέξη. Επίσης, θα αποτελούσε πρόοδο αν θα υπήρχε μεγαλύτερη συνοχή μεταξύ γλωσσών. Για παράδειγμα, λέξεις όπως, “democracy” και “democracia” που είναι διαφορετικές αλλά έχουν την ίδια σημασία και θα έπρεπε να ανήκουν στο ίδιο θέμα και με αυτόν τον τρόπο να ενισχυθεί η ισχύς του μοντέλου.

4 ΜΕΘΟΔΟΛΟΓΙΑ

Στο κεφάλαιο αυτό θα αναλυθούν διεξοδικά τα βήματα που ακολουθούνται στην επίλυση ενός προβλήματος στην ανάλυση δεδομένων. Η διαδικασία ξεκινά με τον προσδιορισμό του προβλήματος και ακολουθεί η συλλογή των δεδομένων, ο καθαρισμός τους και η ανάλυση τους με χρήση της περιγραφικής στατιστικής και οπτικοποιήσεων. Έπειτα, εφαρμόζονται οι απαραίτητοι αλγόριθμου και η αξιολόγηση αυτών. Τέλος, εφαρμόζεται και βελτιστοποιείται το καλύτερο μοντέλο (βλ. Εικόνα 4.1).

Εικόνα 4.1: Μεθοδολογία επίλυσης ενός προβλήματος ανάλυσης δεδομένων (Lateef, 2022)

4.1 Προσδιορισμός του προβλήματος

Το στάδιο κατανόησης του προβλήματος της επιχείρησης αποτελεί το πιο σημαντικό βήμα στην Επιστήμη των Δεδομένων, αφού προκειμένου να λυθεί ένα πρόβλημα, θα πρέπει να κατανοηθεί εις βάθος. Η επιχείρηση θα πρέπει να θέσει με σαφήνεια το πρόβλημα της στον επιστήμονα, έτσι ώστε να βρεθεί το σωστό μοντέλο για την επίλυση του, χωρίς να χρειαστούν πολλές προσεγγίσεις και υποθέσεις που θα κάνουν τη λύση να απέχει πολύ από τον στόχο της. Από την άλλη μεριά, παίζει σημαντικό ρόλο, η εμπειρία, οι γνώσεις και η δημιουργικότητα του επιστήμονα για τη σωστή διατύπωση του ζητούμενου της επιχείρησης (Szabłowski, 2020).

Ο στόχος της ερευνητικής ομάδας είναι να σχεδιάσουν έναν τρόπο επίλυσης του προβλήματος με χρήση εργαλείων εξόρυξης δεδομένων. Οι πιο συνήθεις μέθοδοι επίλυσης περιλαμβάνουν μοντέλα κατηγοριοποίησης, παλινδρόμησης κ.α. Οι επιστήμονες, σε αυτήν τη φάση, θα πρέπει να είναι σε θέση να απαντήσουν τα παρακάτω ερωτήματα: Τι ακριβώς θέλουμε να κάνουμε; Πώς θα το πετύχουμε; Τι εργαλεία πρέπει να χρησιμοποιήσουμε; Μετά την απάντηση αυτών των ερωτημάτων, ξεκινά η διαδικασία επίλυσης, με την κατασκευή ενός πιο απλού πρωταρχικού μοντέλου, το οποίο στη συνέχεια εμπλουτίζεται και προσαρμόζεται ανάλογα με απώτερο

σκοπό να αντικατοπτρίζει με μεγαλύτερη σαφήνεια τις επιχειρηματικές ανάγκες (Provost & Fawcett, 2013).

4.2 Συλλογή δεδομένων

Ο τρόπος συλλογής των δεδομένων διαφέρει ανάλογα με τον τύπο τους, ο οποίος μπορεί να ανήκει σε μια από τις δύο κατηγορίες (Amadebai, 2022):

1. **Ποιοτικά δεδομένα:** Πρόκειται για τα δεδομένα που δεν είναι αριθμητικά και αποτελούνται από λέξεις, φράσεις ή προτάσεις. Αυτά εκδηλώνουν συνήθως απόψεις και συναισθήματα των ανθρώπων για κάποιο θέμα. Συλλέγονται κυρίως μέσω συνεντεύξεως, εγγράφων κλπ.
2. **Ποσοτικά δεδομένα:** Πρόκειται για δεδομένα που είναι αριθμητικά και μπορούν να αναλυθούν με μαθηματικές μεθόδους. Τα δεδομένα αυτά είναι συνήθως πιο αξιόπιστα και η συλλογή τους γίνεται μέσω ερωτηματολογίων, παρατηρήσεων, πειραμάτων κλπ.

Ένας δεύτερος τρόπος διάκρισης των δεδομένων είναι ο διαχωρισμός με βάση τη διαδικασία συλλογής τους. Ανάλογα με αυτή, χωρίζονται σε δύο κατηγορίες, τα πρωτογενή και τα δευτερογενή δεδομένα (βλ. Εικόνα 4.2).

Εικόνα 4.2: Κατηγορίες μεθόδων συλλογής δεδομένων (Singh, 2022)

Τα **πρωτογενή δεδομένα** είναι αυτά που συλλέγονται άμεσα από τα άτομα και αποτελούν ένα καινούριο σύνολο δεδομένων αφού δεν έχει αξιοποιηθεί ή τροποποιηθεί

από άλλο άτομο. Αυτό το χαρακτηριστικό τα καθιστά μοναδικά αλλά χρονοβόρα ως προς τη συλλογή και πολλές φορές πλαστά, εφόσον μπορούν εύκολα να τροποποιηθούν (Amadebai, 2022). Τα είδη των ερευνών που χρησιμοποιούνται για την απόκτηση των πρωτογενών δεδομένων είναι τα εξής:

- **Έρευνές ή ερωτηματολόγια:** Οι έρευνές και τα ερωτηματολόγια αποτελούν τους πιο συνήθεις τρόπους συλλογής πρωτογενών δεδομένων. Πρόκειται για ερωτήσεις που συντάσσονται και αποστέλλονται στο δείγμα της έρευνας για να απαντηθούν. Για την επίτευξη μιας ορθής μελέτης, συνίσταται η αξιολόγηση των ερωτηματολογίων από ειδικούς για την επισήμανση τυχών ανεπαρκειών στις ερωτήσεις ή τυχών λαθών στις διάφορες τεχνικές που χρησιμοποιούνται. Παλιότερα ο διαμοιρασμός αυτών των ερευνών ή των ερωτηματολογίων γινόταν αποκλειστικά σε έγγραφα που θα συμπλήρωναν τα άτομα του δείγματος. Πλέον, υπάρχουν πιο σύγχρονοι τρόποι μέσω διαδικτύου και πιο συγκεκριμένα, μέσω ηλεκτρονικού ταχυδρομείου, μέσων κοινωνικών δικτύων κ.α. (Formplus, 2022)
- **Πειράματα:** Αυτή η μέθοδος ξεκινά με την επιλογή του αντικειμένου ανάλυσης. Στη συνέχεια διενεργούνται διάφορες ενέργειες στο αντικείμενο αυτό, έτσι ώστε να καταγραφούν τα πρωτογενή δεδομένα από τον ερευνητή. Οι ενέργειες αυτές είναι απαραίτητο να επαναληφθούν όσες φορές απαιτηθεί, έως ότου τα αποτελέσματα να είναι έγκυρα και αξιόπιστα. Τέλος, οι καταγραφές αυτές αναλύονται και εξάγεται ένα συμπέρασμα (Formplus, 2022).
- **Συνεντεύξεις ή ομάδες συζητήσεων:** Οι συνεντεύξεις αποτελούν μια συνήθη μέθοδο συλλογής πληροφοριών. Οι ερωτήσεις οι οποίες επιλέγονται από τον ερευνητή μπορεί να διατυπώνονται άμεσα ή έμμεσα, να είναι εστιασμένες ή όχι και οι απαντήσεις αυτών καταγράφονται με σκοπό την ανάλυση τους. Μια συνέντευξη μπορεί να πραγματοποιηθεί και με προσωπική επαφή αλλά και τηλεφωνικά (φωνητικές κλήσεις ή βιντεοκλήσεις). Από την άλλη πλευρά, οι ομάδες συζητήσεων δεν εμπεριέχουν τη διαδικασία ερωτήσεων και απαντήσεων, αλλά περιλαμβάνει συζητήσεις και αλληλεπιδράσεων μεταξύ διαφόρων μελών. Εδώ, οι ερευνητές έχουν παθητικό ρόλο, εφόσον δεν είναι αυτοί που καθοδηγούν τη συζήτηση αλλά βρίσκονται στον χώρο με σκοπό την επίβλεψη της διαδικασίας (Formplus, 2022).

Τα **δευτερογενή δεδομένα** αποτελούνται από παρελθοντικά δεδομένα, δηλαδή πληροφορίες που δε λαμβάνονται για πρώτη φορά. Δε θεωρούνται τόσο αξιόπιστα όσο

τα πρωτογενή αλλά είναι χρήσιμα για οποιονδήποτε επιθυμεί να τα αξιοποιήσει στην ερευνά του αφού είναι εύκολα στην πρόσβαση. Επίσης, σε αντίθεση με τα πρωτογενή δεδομένα, η συλλογή των δευτερογενών δεδομένων δεν απαιτεί πολύ χρόνο και κόπο (Amadebai, 2022). Οι πηγές από τις οποίες αντλούνται είναι οι εξής:

- **Βιβλία:** Αποτελούν την πιο αξιόπιστη πηγή άντλησης πληροφοριών για τη διεξαγωγή μιας έρευνας και η πληθώρα των ειδών τους, καλύπτει τις ανάγκες των ερευνητών.
- **Περιοδικά:** Σε αντίθεση με τα βιβλία που εκτυπώνονται μονάχα μια φορά, τα περιοδικά δημοσιεύονται σε μεγαλύτερη συχνότητα. Αυτό έχει αντίκτυπο στην αρχαιότητα των πληροφοριών και καθιστά τα περιοδικά πιο σύγχρονα και ενημερωμένα με βάση την επικαιρότητα (Olson, 2014).
- **Κυβερνητικά αρχεία:** Τα κυβερνητικά αρχεία αποτελούν μια αξιόπιστη πηγή δεδομένων και πολύ συχνά, εμπεριέχουν πληροφορίες που φαίνονται χρήσιμες στο μάρκετινγκ, τις κοινωνικές επιστήμες, τις επιστήμες υγείας κ.α. (Formplus, 2022)
- **Βάσεις δεδομένων:** Οι βάσεις δεδομένων μπορεί να αποτελούνται από πολλών διαφόρων ειδών δεδομένα, όπως κυβερνητικά δεδομένα, δεδομένα πελατών και πωλήσεων κ.α. Η επιλογή της κατάλληλης βάσης δεδομένων από τον ερευνητή, εξαρτάται από τις ανάγκες της έρευνας που επιχειρεί.

4.3 Προετοιμασία δεδομένων (pre-processing)

Μετά τη συλλογή των δεδομένων, ξεκινά η διαδικασία του pre-processing, η οποία περιλαμβάνει την αφαίρεση του «θορύβου» και συνάμα την μετατροπή των δεδομένων, με σκοπό την διευκόλυνση του αλγόριθμου ταξινόμησης και συνεπώς την επίτευξη των καλύτερων δυνατών επιδόσεων. Πρόκειται για την πιο χρονοβόρα αλλά και την πιο σημαντική διαδικασία της έρευνας, αφού επηρεάζει άμεσα τα αποτελέσματα του αλγορίθμου ταξινόμησης αλλά και τη θεματική κατηγοριοποίηση του κειμένου. Οι εργασίες αυτές θα αναλυθούν παρακάτω όπως και οι υπόλοιπες που εφαρμόστηκαν στην παρούσα εργασία.

4.3.1 Punctuation, digits, lower casing

Η πρώτη επεξεργασία που εφαρμόζουμε στα δεδομένα είναι η αφαίρεση των σημείων στίξης και των αριθμητικών ψηφίων. Είναι μια πολύ σημαντική διαδικασία, αφού τα παραπάνω στοιχεία δεν εμπεριέχουν σημαντική πληροφορία και θεωρούνται «θόρυβος» στο σώμα κειμένου. Αυτό συνεπάγεται στο ότι δε βοηθούν τον αλγόριθμο στην κατανόηση και στην ταξινόμηση τους. Εφόσον στην παρούσα εργασία χρησιμοποιούνται μοντέλα όπως οι μηχανές SVM και η Λογιστική Παλινδρόμηση κ.α., είναι απαραίτητο να πραγματοποιηθεί αυτή η επεξεργασία πριν την εκτέλεση των αλγορίθμων. Ακόμη, σε αυτό το βήμα, όλες οι αξιολογήσεις θα πρέπει να μετατραπούν σε πεζά γράμματα. Αυτό συμβαίνει γιατί αν ακόμη και αν οι λέξεις είναι ίδιες, αν η μια διαφέρει από την άλλη στον τρόπο γραφής, δηλαδή η μια έχει κάποια πεζά ή κεφαλαία γράμματα διαφορετικά από την άλλη, θεωρούνται ανόμοιες από τον αλγόριθμο.

4.3.2 Tokenization

Μετά τον αρχικό καθαρισμό των κειμένων των αξιολογήσεων, προχωράμε στη διαδικασία του tokenization. Το tokenization είναι η κατάτμηση του κειμένου σε επιμέρους τμήματα που ονομάζονται tokens. Στην παρούσα εργασία εφαρμόστηκε διαίρεση των κριτικών κατά λέξη, η οποία είναι και η πιο συνήθης τεχνική. Για παράδειγμα, η πρόταση «η οθόνη έχει πολύ καλή ανάλυση», θα μετατραπεί στις λέξεις: «η», «οθόνη», «έχει», «πολύ», «καλή», «ανάλυση».

4.3.3 Αφαίρεση stop words

Το βήμα αυτό περιλαμβάνει την αφαίρεση των λέξεων που χρησιμοποιούνται πολύ συχνά και δεν προσφέρουν χρήσιμες πληροφορίες για την εκπαίδευση του αλγορίθμου και συνεπώς, επηρεάζουν αρνητικά την ικανότητα να ταξινομεί με επιτυχία τα δεδομένα και θα πρέπει να τις αγνοήσουμε. Για παράδειγμα, ένα μέρος αυτών των λέξεων είναι το εξής:

[‘is’,’are’,’an’,’a’,’as’,‘when’,’where’,’then’,’what’,’which’,’he’,’she’,’it’,’has’,’you’,’these’,’with’,’by’,‘it’,‘how’,’between’,’than’,’from’,’some’,’most’,’these’,’should’]

Όπως είναι κατανοητό, οι λέξεις αυτές δε θα πρέπει να καταλαμβάνουν χώρο στα δεδομένα και κατά συνέπεια, να αυξάνουν τον χρόνο εκπαίδευσης του μοντέλου. Όμως, η αφαίρεση τους δεν είναι πάντα η σωστή λύση, αφού πρέπει να αποφασίζεται ανάλογα με τις ανάγκες του προβλήματος. Για παράδειγμα, η ανάλυση συναισθήματος, είναι μια εργασία κατά την οποία δεν προτιμάται να αφαιρούνται τα stop words. Αυτό συμβαίνει γιατί σε μια κριτική όπως η παρακάτω: «Η οθόνη δεν ήταν καλής ποιότητας», αν αφαιρεθούν τα stopwords, θα γίνει: «οθόνη καλής ποιότητας». Άρα, το νόημα αλλοιώνεται και τελικά, η ανάλυση καταλήγει σε λανθασμένα συμπεράσματα. Αντίθετα, για τη θεματική κατηγοριοποίηση, είναι απαραίτητη η διαδικασία αφαίρεσης αυτών των λέξεων αφού με αυτόν τον τρόπο αφαιρείται μεγάλο μέρος ασήμαντων λέξεων και παραμένει όλη η ουσία. Έτσι, τα τελικά topics που δημιουργούνται, είναι πιο καθαρά και ευανάγνωστα.

4.3.4 Lemmatization

Το επόμενο βήμα στην επεξεργασία των κειμένων είναι η διαδικασία εύρεσης του λήμματος των λέξεων ανάλογα με τη σημασία τους. Δηλαδή, η λημματοποίηση αποσκοπεί στην αφαίρεση των καταλήξεων των λέξεων και επιστρέφει τη βασική τους μορφή. Ακόμη, προσδίδει περιεχόμενο στις λέξεις και επομένως, τις συνδέει με άλλες παρόμοιες σε σημασία.

Για παράδειγμα, κάποιες μετατροπές λέξεων φαίνονται στον Πίνακα 4.1.

Πίνακας 4.1: Μετατροπή λέξεων με τη μέθοδο Lemmatization

Original	Lemmatization
phones	phone
charging	charge
better	good

Αυτή η διαδικασία είναι απαραίτητη, αφού το λεξικό που θα δημιουργήσουμε δε θα έχει πολλές λέξεις με το ίδιο νόημα αλλά γραμμένες αλλιώς, οπότε θα είναι και πιο καθαρό και θα μειωθεί ο χρόνος διαδικασίας. Στο παρών πρόβλημα γίνεται χρήση της βιβλιοθήκης spacy για την εφαρμογή της λημματοποίησης.

Πολλοί συγχέουν το lemmatization με το stemming, αφού και τα δύο αξιοποιούνται για την κατασκευή των λημμάτων των λέξεων. Ωστόσο, πρόκειται για δύο πολύ διαφορετικές μεθόδους. Πιο συγκεκριμένα, το lemmatization, όπως αναφέρθηκε παραπάνω, οδηγεί στον εντοπισμό των λημμάτων των λέξεων, δηλαδή πραγματικές λέξεις. Αυτό δε συμβαίνει στην περίπτωση του stemming, γιατί αφαιρεί τις καταλήξεις των λέξεων και πολλές φορές τις αλλοιώνει, αλλάζει το νόημα τους και δημιουργεί ορθογραφικά λάθη. Για παράδειγμα, η λέξη «earrings», με τη μέθοδο του lemmatization θα μετατραπεί σε «earring», ενώ με τη μέθοδο του stemming σε «ear». Στην πρώτη περίπτωση, δίνεται το λήμμα της λέξης, το οποίο στην προκειμένη περίπτωση είναι ο ενικός αριθμός, ενώ στη δεύτερη περίπτωση αλλάζει εντελώς η σημασία της λέξης. Όπως είναι φυσικό, το stemming θεωρείται μια πιο ανακριβής μέθοδος συγκριτικά με το lemmatization, όποτε η τελευταία χρησιμοποιείται συχνότερα σε προβλήματα επεξεργασίας φυσικής γλώσσας.

4.3.5 Κανονικοποίηση δεδομένων

Η κανονικοποίηση των δεδομένων αποτελεί ένα σημαντικό βήμα για την ταξινόμηση τους. Όπως έχει αναφερθεί και νωρίτερα, η ταξινόμηση αποσκοπεί στην πρόβλεψη της κλάσης των δεδομένων που έχουν δοθεί ως είσοδος. Με άλλα λόγια είναι η προσέγγιση μιας συνάρτησης απεικόνισης (f) των τιμών εισόδου (x) σε διακριτές μεταβλητές εξόδου (y). Το πρόβλημα που εμφανίζεται πολύ συχνά σε προβλήματα ταξινόμησης είναι η ύπαρξη ανισοβαρών κλάσεων στα δεδομένα. Δηλαδή, μια από τις κλάσεις να έχει σημαντικά υψηλότερο αριθμό παρατηρήσεων από τις άλλες. Για παράδειγμα, έστω ότι μια τράπεζα επιθυμεί να ελέγξει αν κάποιες συναλλαγές είναι δόλιες. Αν στα δεδομένα που διαθέτει, 40 μόνο από τις 3000 είναι δόλιες (κάτω από το 2%), τότε πάνω από το 98% των συνολικών συναλλαγών δεν είναι δόλιες (Mazumder, 2021). Αυτή η ανισότητα στις κλάσεις, οδηγεί το μοντέλο ταξινόμησης σε μεροληπτικά συμπεράσματα και πρέπει να αντιμετωπιστεί πριν την είσοδο των δεδομένων στον αλγόριθμο. Κάποιοι τρόποι αντιμετώπισης του εν λόγω προβλήματος είναι οι εξής:

1. Η επιλογή του **f1 score** για την αξιολόγηση του μοντέλου ταξινόμησης. Αν και η **ορθότητα (accuracy)** αποτελεί μια πολύ σημαντική μετρική για την εκτίμηση της ποιότητας των αποτελεσμάτων ενός μοντέλου, δεν είναι ιδανική για προβλήματα ανισόρροπων κλάσεων. Αντίθετα, το f1 score είναι το

καταλληλότερο μέτρο αξιολόγησης αυτών των προβλημάτων, αφού σε περίπτωση αύξησης των ψευδώς θετικών ή των ψευδώς αρνητικών αποτελεσμάτων, η τιμή του θα είναι χαμηλή. Αυτό σημαίνει ότι το f1 score θα αυξηθεί μονάχα αν βελτιωθεί η ποιότητα των προβλέψεων, συνεπώς μονάχα αν ο αλγόριθμος αναγνωρίζει σωστά όλες τις κλάσεις (Mazumder, 2021).

2. Με την εφαρμογή της τεχνικής της **επαναδειγματοληψίας**. Η τεχνική αυτή χρησιμοποιεί το μη ισορροπημένο σύνολο δεδομένων και το επαναδειγματοληπτεί με σκοπό την κατασκευή ισοβαρών κλάσεων. Αυτή η διαδικασία μπορεί να γίνει με δύο τρόπους. Ο πρώτος τρόπος είναι η υπερδειγματοληψία της κλάσης που μειονεκτεί με τη χρήση της αντικατάστασης. Ο δεύτερος τρόπος εξισορρόπησης των κλάσεων είναι η διαγραφή τυχαίων γραμμών από την κλάση που περιέχει περισσότερες παρατηρήσεις για να ισοφαρίσει τον αριθμό των παρατηρήσεων της κλάσης που μειονεκτεί. Με την εξισορρόπηση των δεδομένων, ο αλγόριθμος θα αντιμετωπίζει με τον ίδιο τρόπο όλες τις κλάσεις και τα αποτελέσματα του θα είναι αμερόληπτα (Mazumder, 2021).
3. Με τη μέθοδο **SMOTE** ή αλλιώς **Synthetic Minority Oversampling Technique**. Εφαρμόζει μια τεχνική υπερδειγματοληψίας της κλάσης που περιέχει μικρό αριθμό παρατηρήσεων. Πιο συγκεκριμένα, χρησιμοποιεί τα υπάρχοντα δεδομένα και τα χαρακτηριστικά τους, για την κατασκευή νέων στη μειονεκτική κλάση. Συνεπώς, τα νέα δεδομένα που δημιουργούνται δεν αποτελούν αντίγραφα των ήδη υπαρχόντων. Η διαδικασία κατασκευή τους ξεκινά με τη συλλογή δειγμάτων των χαρακτηριστικών των υπαρχόντων γειτονικών δεδομένων από τον αλγόριθμο και τον συνδυασμό τους, έτσι ώστε να κατασκευαστούν νέα παραδείγματα (Microsoft, 2021).
4. Με χρήση **ορίων πιθανοτήτων**. Όπως είναι γνωστό, οι αλγόριθμοι ταξινόμησης κάνουν χρήση των πιθανοτήτων για την πρόβλεψη της κλάσης μιας παρατήρησης. Μπορεί να γίνει ανάθεση των πιθανοτήτων αυτών σε μια κλάση με βάση ένα όριο πιθανότητας. Συνήθως αυτό το όριο είναι το 0.5, δηλαδή αν η πιθανότητα είναι < 0.5 ανήκει σε μια συγκεκριμένη κλάση, αλλιώς ανήκει σε άλλη κλάση. Πολύ συχνά, για να βρεθεί το κατάλληλο όριο, το οποίο θα χωρίσει αποτελεσματικά τις κλάσεις, αξιοποιούνται οι καμπύλες ROC. Έτσι, είμαστε σίγουροι ότι το όριο που επιλέγεται είναι σωστό (Mazumder, 2021).

Στο πρόβλημα μας, θα γίνει εφαρμογή της επαναδειγματοληψίας για την κανονικοποίηση των κλάσεων, έτσι ώστε να έχουμε πιο ακριβή και αμερόληπτα συμπεράσματα. Επιπλέον, θα αξιολογήσουμε τα αποτελέσματα και με τη χρήση του f1 score για να είμαστε σίγουροι για την αμεροληψία του μοντέλου ταξινόμησης.

4.4 Εξερεύνηση των δεδομένων

Η εξερεύνηση αποτελεί το τέταρτο βήμα στην επίλυση ενός προβλήματος στην Επιστήμη των Δεδομένων. Στο βήμα αυτό γίνεται χρήση της στατιστικής και μέσων οπτικοποίησης με σκοπό την κατανόηση και την περιγραφή των χαρακτηριστικών των δεδομένων. Κάποια από τα χαρακτηριστικά από αυτά μπορεί να είναι το μέγεθος του δείγματος, η ακρίβεια κ.α. Επίσης, εντοπίζονται σχέσεις μεταξύ των μεταβλητών, τυχών ακραίες ή ελλείπουσες τιμές, αποκαλύπτονται μοτίβα και άλλες σημαντικές πληροφορίες για τους αναλυτές, με τις οποίες αποκομίζουν μια μεγαλύτερη εικόνα για αυτά (Heavy, 2022).

Συγκριτικά με τη μελέτη αριθμητικών δεδομένων, η οπτικοποίηση τους, βοηθά σε πολύ μεγαλύτερο βαθμό την κατανόηση τους από τον άνθρωπο. Αυτό αποδίδεται στο γεγονός ότι η απόδοση νοήματος και η επεξήγηση χιλιάδων γραμμών είναι εξαιρετικά δύσκολη. Συνεπώς, τα οπτικά στοιχεία (όπως γραμμές, διαγράμματα, σχήματα, σημεία) αποτελούν σημαντική βοήθεια για τους επιστήμονες για την εξερεύνηση και έπειτα για τον καθαρισμό του δείγματος, αφού πολλές ανωμαλίες ή σχέσεις μεταξύ αυτών είναι αδύνατον να εντοπιστούν χωρίς τη βοήθεια οπτικοποιήσεων (Heavy, 2022).

Η διερευνητική ανάλυση των δεδομένων χωρίζεται σε τέσσερεις κατηγορίες ανάλογα με το πλήθος των μεταβλητών που ο αναλυτής επιθυμεί να ερευνήσει αλλά και τον τρόπο που θα επιλέξει για αυτό, δηλαδή αν τις αναπαραστήσει γραφικά ή όχι. Οι κατηγορίες αυτές είναι οι παρακάτω (IBM Cloud Education, 2020):

1. **Μη γραφική αναπαράσταση μιας μεταβλητής**: Πρόκειται για τον πιο απλό τρόπο ανάλυσης, αφού αφορά μονάχα μια μεταβλητή. Στην περίπτωση αυτή αναζητούνται τυχών μοτίβα στα δεδομένα αλλά ταυτόχρονα γίνεται και η περιγραφή τους.
2. **Μη γραφική αναπαράσταση πολλαπλών μεταβλητών**: Πρόκειται για την ανάλυση πολλών μεταβλητών ταυτοχρόνως. Σε αυτήν την ανάλυση δε γίνεται

μονάχα η περιγραφή των μεταβλητών αλλά αναζητούνται τυχών σχέσεις μεταξύ δύο ή περισσοτέρων μεταβλητών, των οποίων η εύρεση γίνεται μέσω πινάκων και χρήση στατιστικής.

3. **Γραφική αναπαράσταση μιας μεταβλητής**: Οι μη γραφικές μέθοδοι δε δείχνουν πάντα ολοκληρωμένα την εικόνα των δεδομένων, συνεπώς πολύ συχνά χρειάζεται η αναπαράσταση τους γραφικά. Κάποιοι τύποι αναπαραστάσεων μιας μεταβλητής είναι οι παρακάτω:
    - Το **ιστόγραμμα (histogram)**: Πρόκειται για ένα ραβδόγραμμα όπου οι ράβδοι αντιπροσωπεύουν τη συχνότητα ή την αναλογία των παρατηρήσεων για ένα εύρος τιμών.
    - Το **φυλλόγραμμα (stem-and-leaf plot)**: Δείχνει όλες τις τιμές της μεταβλητής και το σχήμα της κατανομής.
    - Το **θηκόγραμμα (box plot)**: Απεικονίζει γραφικά πέντε αριθμητικά δεδομένα: τη μικρότερη παρατήρηση, το πρώτο τεταρτημόριο (Q1), τη διάμεσο (δ), το τρίτο τεταρτημόριο (Q3) και τη μέγιστη παρατήρηση.
4. **Γραφική αναπαράσταση πολλαπλών μεταβλητών**: Πρόκειται για τη γραφική αναπαράσταση παραπάνω από μιας μεταβλητής, με σκοπό την εύρεση των σχέσεων μεταξύ των μεταβλητών. Κάποιοι τύποι τέτοιων γραφικών αναπαραστάσεων είναι (IBM Cloud Education, 2020):
    - Το **ομαδοποιημένο ραβδόγραμμα (grouped bar plot)**: Είναι το πιο σύνηθες διάγραμμα αναπαράστασης πολλών μεταβλητών. Σε αυτό το διάγραμμα κάθε ομάδα αντιπροσωπεύει μια μεταβλητή και κάθε κομμάτι εντός της ράβδου αντιπροσωπεύει τις τιμές μιας άλλης μεταβλητής.
    - Το **διάγραμμα διασποράς (scatter plot)**: Απεικονίζει τα σημεία των τιμών των μεταβλητών σε δύο άξονές για να γίνει κατανοητό πόσο επηρεάζεται η μια από την άλλη.
    - Το **διάγραμμα φυσαλίδων (bubble chart)**: Στο διάγραμμα αυτό γίνεται αναπαράσταση των δεδομένων με φυσαλίδες (κύκλους) σε δύο άξονες. Είναι στην πραγματικότητα ένας τύπος διαγράμματος διασποράς. Η διαφορά αυτών των δύο γραφικών αναπαραστάσεων είναι ότι το διάγραμμα διασποράς δείχνει τις τιμές των μεταβλητών και τη σύγκριση αυτών, ενώ στο διάγραμμα φυσαλίδων, οι φυσαλίδες αντικαθιστούν τα σημεία αυτά για να δείξουν τη σύγκριση τους.

4.5 Μοντελοποίηση και αξιολόγηση

Η μοντελοποίηση των δεδομένων είναι η διαδικασία κατασκευής αλγορίθμων πρόβλεψης με σκοπό την εξόρυξη δεδομένων. Οι αλγόριθμοι αυτοί εφαρμόζουν τεχνικές (όπως η ταξινόμηση, η ομαδοποίηση, η συσχέτιση) και αναδεικνύουν μοτίβα σε περίπλοκα σύνολα δεδομένων. Στο Κεφάλαιο 2 έχει γίνει αναφορά στις κατηγορίες αυτών των μοντέλων αλλά και περιγραφή ορισμένων εξ αυτών.

Αφού επιτευχθεί η κατασκευή του μοντέλου, ακολουθεί η αυστηρή αξιολόγηση του. Σκοπός αυτής της αξιολόγησης είναι η επιβεβαίωση ότι τα αποτελέσματα του αλγορίθμου που κατασκευάστηκε είναι έγκυρα και αξιόπιστα. Πιο συγκεκριμένα, τα μοτίβα που εξάγονται από το σύνολο δεδομένων θα πρέπει να είναι αξιόπιστα και να μην οφείλονται σε δειγματοληπτικές ανωμαλίες. Επιπλέον, εξίσου σημαντική είναι η ικανοποίηση των επιχειρηματικών στόχων. Όσο αξιόπιστο και αν είναι το μοντέλο, αν δεν οδηγεί στη λύση του επιχειρηματικού προβλήματος θα πρέπει να αναπροσαρμοστεί ανάλογα με τις ανάγκες της επιχείρησης ή να αντικατασταθεί από ένα νέο μοντέλο. Ακόμη, ένας επιστήμονας θα πρέπει να λάβει υπόψιν του πως, πολλές φορές, παρόλο που ο αλγόριθμός καθίσταται εξαιρετικά ακριβής (>99%), επιστρέφει πολλά ψευδή αποτελέσματα. Πράγμα που είναι πιθανό να συμβεί σε περιπτώσεις ανισορροπίας των δεδομένων. Παρόλα αυτά, αυτό το λάθος του αλγορίθμου μπορεί να εντοπιστεί στο στάδιο της εξερεύνησης του συνόλου δεδομένων, όπως έχει αναφερθεί παραπάνω (Provost & Fawcett, 2013).

4.6 Ανάπτυξη και βελτιστοποίηση

Η ανάπτυξη είναι η διαδικασία κατά την οποία τα αποτελέσματα του αλγορίθμου που δημιουργήθηκε και αξιολογήθηκε από τον αναλυτή, χρησιμοποιούνται στο πραγματικό κόσμο με σκοπό τη συνεισφορά στις επιχειρήσεις. Για παράδειγμα, σε ένα πρόβλημα πρόβλεψης απομάκρυνσης πελατών από μια τράπεζα, η ανάπτυξη αφορά την εφαρμογή του μοντέλου πρόβλεψης στα πραγματικά δεδομένα της τράπεζας με σκοπό την εύρεση των απογοητευμένων πελατών. Έπειτα, η επιχείρηση θα μπορεί να κάνει προσπάθειες διατήρησης αυτών των πελατών μέσω κάποιων ενεργειών (Provost & Fawcett, 2013). Εδώ είναι σημαντικό να σημειωθεί ότι οι αλγόριθμοι που δημιουργούνται, δεν προορίζονται για την επίλυση μονάχα ενός προβλήματος αλλά απαιτείται να είναι

ευέλικτα και να ανταποκρίνονται επιτυχώς σε νέα δεδομένα. Συνεπώς, η επιτυχία του επιχειρηματικού στόχου προϋποθέτει πολύ καλό προγραμματισμό και προετοιμασία. Όπως είναι λογικό, ελάχιστα από τα μοντέλα που αναπτύσσονται, ανταποκρίνονται στους επιθυμητούς επιχειρηματικούς στόχους και πολλές φορές απαιτούν μια σημαντική επένδυση πόρων (Weedmark, 2021).

Το τελικό βήμα στην επίλυση ενός προβλήματος στην Επιστήμη των Δεδομένων είναι η βελτιστοποίηση της υπάρχουσας λύσης. Σε αυτό το στάδιο θα πρέπει να ερμηνευτεί το αποτέλεσμα και να διαπιστωθεί αν είναι αυτό που είχε ζητηθεί. Αν δεν είναι ακριβώς αυτό που ήθελε να πετύχει ο επιστήμονας και η επιχείρηση, τότε αφού κατανοηθεί εις βάθος ο λόγος για τον οποίο ο αλγόριθμος κατέληξε σε αυτό το αποτέλεσμα, μπορεί να αναπτυχθεί ξανά ένας νέος αλγόριθμος που θα ανταποκρίνεται στις ανάγκες του αρχικού επιχειρηματικού στόχου (Geeksforgeeks, 2020).

5 CASE STUDY

5.1 Προσδιορισμός του προβλήματος

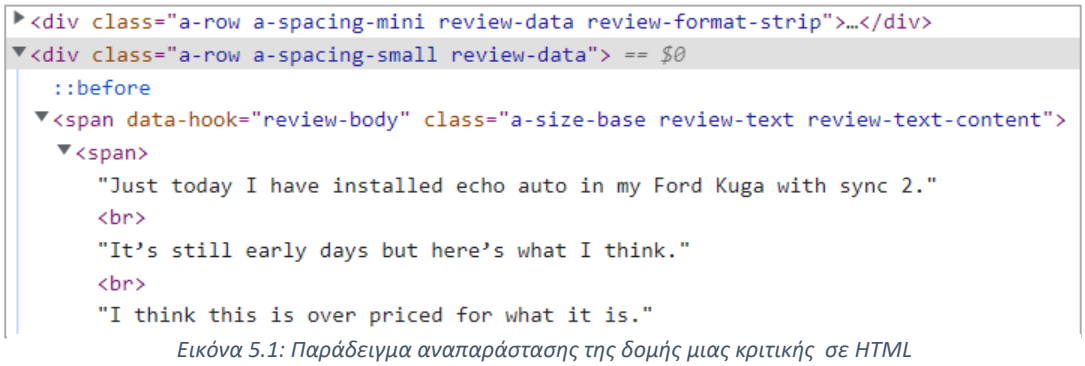
Ο στόχος κάθε επιχείρησης λιανικού εμπορίου είναι η προτίμηση των προϊόντων της από τους πελάτες και η ικανοποίηση τους από την ποιότητα αυτών, έτσι ώστε να επιτευχθεί αύξηση του κέρδους και εδραίωση της καλής φήμης της εταιρίας. Για αυτόν τον λόγο, είναι σημαντικό, η επιχείρηση να ενημερώνεται για τυχών παράπονα των πελατών για κάποιο προϊόν και να προχωρά σε μετατροπές ή στη χειρότερη περίπτωση, απόσυρση του προϊόντος. Επιπλέον, εδώ είναι σημαντικό να αναφερθεί ότι οι άνθρωποι τείνουν να επηρεάζονται σε μεγάλο βαθμό από τις κριτικές που πραγματοποιούν οι αγοραστές στο διαδίκτυο γιατί τους θεωρούν μια αξιόπιστη πηγή πληροφοριών λόγω ότι βρίσκονται στην ίδια «ομάδα». Ένας τρόπος για να αναλυθεί η γνώμη του καταναλωτικού κοινού είναι η ανάλυση των κριτικών που οι πελάτες γράφουν στην ιστοσελίδα της εταιρίας. Η ανάλυση αυτή ξεκινά με την κατηγοριοποίηση των αξιολογήσεων σε θετικές και αρνητικές (sentiment analysis), αξιοποιώντας κατάλληλα μοντέλα με σκοπό να βρεθούν οι κριτικές που ενδιαφέρουν την εταιρία. Ο διαχωρισμός τους χειροκίνητα είναι μια πολύ χρονοβόρα διαδικασία, η οποία είναι πρακτικά αδύνατο να πραγματοποιηθεί αν ο όγκος των πληροφοριών είναι μεγάλος. Αφού προβλεφθεί η κατηγορία των αξιολογήσεων, θα απομονωθούν οι αρνητικές κριτικές και θα βρεθούν τα θέματα στα οποία αναφέρονται οι καταναλωτές σε αυτές. Με αυτόν τον τρόπο, θα μπορεί η επιχείρηση να μάθει τι ακριβώς απογοήτευσε τους αγοραστές και θα προχωρήσει σε ενέργειες βελτιστοποίησης του προϊόντος. Όλες αυτές οι πολύτιμες πληροφορίες θα αποκτηθούν χωρίς να χρειαστεί κάποιος να διαβάσει μία μία τις κριτικές και να κρατά σημειώσεις, αλλά με τη βοήθεια των μοντέλων Θεματικής Κατηγοριοποίησης (topic modeling).

5.2 Συλλογή και περιγραφή του συνόλου δεδομένων

Το σύνολο δεδομένων αποτελείται από 3720 κριτικές αγοραστών σε αγγλική γλώσσα για το προϊόν Alexa Echo Auto της εταιρίας Amazon. Αυτές οι αξιολογήσεις αντλήθηκαν με τη μέθοδο του **web scraping** από την ιστοσελίδα της Amazon. Το **web scraping** είναι μια αυτόματη μέθοδος άντλησης μεγάλου όγκου δεδομένα από

ιστότοπους. Παίρνει μη δομημένα δεδομένα σε μορφή HTML και τα μετατρέπει σε δομημένα, τοποθετώντας τα σε υπολογιστικά φύλλα ή βάσεις δεδομένων.

Σε αυτήν την εργασία δημιουργήθηκε ένας αλγόριθμος web scraping με χρήση της Python. Ο αλγόριθμος αυτός χρησιμοποιεί τρεις βιβλιοθήκες: την bs4 από την οποία γίνεται προσθήκη της BeautifulSoup, την requests και την pandas. Η BeautifulSoup είναι υπεύθυνη για την εξαγωγή των δεδομένων από αρχεία HTML και XML και με τη συνεργασία με τη βιβλιοθήκη requests, η οποία κάνει τη «αίτηση» για την άντληση δεδομένων από την ιστοσελίδα, αντλούνται όσες πληροφορίες χρειάζονται. Για την παρούσα εργασία, οι πληροφορίες αυτές είναι ο τίτλος, ο αριθμός των αστεριών με τα οποία έχει αξιολογηθεί η αγοραστική εμπειρία των καταναλωτών και φυσικά, το σώμα κειμένου των αξιολογήσεων. Οι πληροφορίες αυτές αποτελούν περιεχόμενα των στοιχείων του HTML αρχείου της ιστοσελίδας. Τα στοιχεία αυτά και συνεπώς το περιεχόμενο τους, βρίσκονται μέσα από μονοπάτια (paths) που αποτελούνται από html μορφοποίηση (div, class) (Mozilla, 2022). Στην Εικόνα 5.1 δίνεται ένα παράδειγμα μονοπατιού για να βρεθεί το σώμα κειμένου μιας κριτικής. Με την κατασκευή ενός βρόχου επανάληψης, αφού πρώτα γίνει η «αίτηση» στην ιστοσελίδα του προϊόντος (στην οποία υπάρχει ο αριθμός των σελίδων των κριτικών), ξεκινά η διαδικασία του web scraping σε κάθε σελίδα αξιολογήσεων. Τέλος, με τη βοήθεια της βιβλιοθήκης pandas οι πληροφορίες αυτές αποθηκεύονται σε ένα πλαίσιο δεδομένων (data frame) και εξάγονται σε ένα υπολογιστικό φύλλο (excel).



Εικόνα 5.1: Παράδειγμα αναπαράστασης της δομής μιας κριτικής σε HTML

5.3 Προετοιμασία των δεδομένων

Μετά τη συλλογή των δεδομένων και την εξερεύνηση τους, ακολουθεί ο καθαρισμός τους και η επιλογή των κριτικών που θα ταξινομηθούν. Αφού πραγματοποιηθεί η δυαδική ταξινόμηση, πρέπει να κατασκευαστούν δύο κλάσεις αξιολογήσεων. Για αυτόν τον λόγο, οι κριτικές που έχουν τρία αστέρια, δηλαδή αυτές που είναι ουδέτερες, αφαιρούνται από τα δεδομένα και πλέον το σύνολο δεδομένων αποτελείται από 3.312 αξιολογήσεις. Ένας ακόμη λόγος της αφαίρεσης των ουδέτερων κριτικών είναι η επίδραση τους στα μοντέλα ταξινόμησης, αφού λόγω της ουδετερότητας τους είναι πιο δύσκολα να ταξινομηθούν με αποτέλεσμα να εξάγουν ανακριβή αποτελέσματα. Συνεπώς, δημιουργείται μια νέα μεταβλητή, η "rate class", όπου χωρίζει τα δεδομένα σε δύο κλάσεις ως εξής: οι κριτικές που έχουν 2 ή λιγότερα αστέρια κατατάσσονται στην κλάση 0 που συμβολίζει τις αρνητικές αξιολογήσεις και αντίστοιχα αυτές που έχουν από 4 αστέρια και πάνω κατατάσσονται στην κλάση θετικών αξιολογήσεων, η οποία συμβολίζεται με την τιμή 1.

Επιπλέον, διαπιστώθηκε πως ο τίτλος των κριτικών, εμπεριέχει σημαντική πληροφορία, εξίσου σημαντική με αυτήν του κύριου σώματος κειμένου. Για αυτόν τον λόγο, έγινε προσθήκη του κειμένου των τίτλων στο κύριο σώμα κειμένου, όπως φαίνεται στην Εικόνα 5.2.

	title	rating	body
0	Unlocking the potential of Alexa	5	Unlocking the potential of Alexa Your browser does not support HTML5 video.\n\n\n The success ...
1	Will not navigate in the UK	1	Will not navigate in the UK WARNING: useless for navigation! Do not buy if you expect to be able...
2	Works brilliantly with older cars and stereos via AUX, too!	5	Works brilliantly with older cars and stereos via AUX, too! This really is an excellent little b...
3	Small but Perfectly Formed and with a Great Sense of Hearing	5	Small but Perfectly Formed and with a Great Sense of Hearing Configured and installed in less th...
4	Meh - at best	2	Meh - at best .... Once the novelty wears off (should not take more than 5 minutes), you begin t...
...	...	...	...
3307	Excellent	5	Excellent Great little kit worth the money
3308	Don't think this is one of those I'd queue up in a hurry	1	Don't think this is one of those I'd queue up in a hurry To be honest, from what I've read the ...
3309	Its a good product but would be better to be able to charge up	5	Its a good product but would be better to be able to charge up Only thing thats not great is whe...
3310	Handy for hands free communications and music selection.	4	Handy for hands free communications and music selection. Works well. Handy for hands free calls,...
3311	Easy to use	5	Easy to use Nice and compact easy to use.

Εικόνα 5.2: Δείγμα από το σύνολο δεδομένων μετά την συγχώνευση των τίτλων με το κύριο σώμα των κριτικών

Στο επόμενο στάδιο, γίνεται ο καθαρισμός του συνόλου δεδομένων ακολουθώντας τα βήματα που διευκρινίστηκαν λεπτομερώς στο Κεφάλαιο 3. Παρακάτω, ανακεφαλαιώνονται τα βήματα αυτά συνοπτικά:

1. Αφαίρεση των σημείων στίξης και των αριθμών. Επίσης, σε αυτό το σημείο μετατρέπονται όλα τα γράμματα σε πεζά. Οι εργασίες αυτές πραγματοποιήθηκαν με τη βοήθεια των regular expressions.
2. Διαίρεση του κειμένου κατά λέξη (tokenization) με χρήση των regular expressions.
3. Αφαίρεση των ασήμαντων λέξεων (stopwords). Για τη διαδικασία αυτή αξιοποιήθηκε το πακέτο nltk.corpus της NLTK (Natural Language Toolkit), η οποία έχει αποθηκευμένη μια λίστα λέξεων χαμηλής σημασίας σε 16 διαφορετικές γλώσσες (Geeksforgeeks, 2022). Στην εργασία αυτή χρησιμοποιήθηκαν μονάχα η λίστα των stopwords σε αγγλική γλώσσα. Εδώ είναι σημαντικό να αναφερθεί ότι από τη λίστα αυτή, αφαιρέθηκαν οι λέξεις “no” και “not”, αφού αποτελούν πολύ σημαντικές λέξεις στα προβλήματα ανάλυσης συναισθήματος και προστέθηκαν οι λέξεις [“alexa”, “auto”, “car”, “echo”, “amazon”] που είναι εξαιρετικά συχνές στα δεδομένα και δε δίνουν πολύτιμη πληροφορία στα μοντέλα ταξινόμησης.
4. Δημιουργία λημμάτων με χρήση της βιβλιοθήκης spacy. Χρησιμοποιήθηκε η τεχνική του lemmatization και όχι του stemming γιατί το stemming «έκοβε» τις καταλήξεις των λέξεων και χανόταν το νόημα τους.

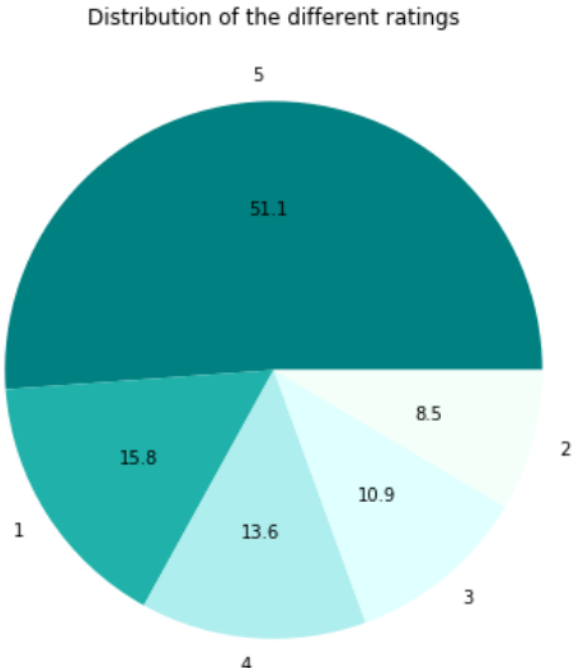
Μετά από τον καθαρισμό του συνόλου δεδομένων, το σώμα κειμένου των κριτικών θα έχει την μορφή που φαίνεται στην Εικόνα 5.3.

body
unlock potential browser support html video success failure largely go depend deeply find ecosys...
navigate warn useless navigation buy expect able say get direction postcode street name favorite...
disappoint sorry big fan station house want love first disappointment come air vent mount includ...
work brilliantly old car stereo aux really excellent little box trick get year old skoda octavia...
small perfectly form great sense hearing configure instal less minute use vent attachment rather...
...
excellent great little kit worth money
think queue hurry honest read seem little pointless let explain pay equivalent box microphone co...
good product well able charge thing s great answer customer hear person hear switch loud speaker
handy hand free communication music selection work well handy hand free call texte read message ...
easy use nice compact easy use

Εικόνα 5.3: Δείγμα από το κύριο σώμα κειμένου των κριτικών μετά τον καθαρισμό

5.4 Εξερεύνηση των δεδομένων

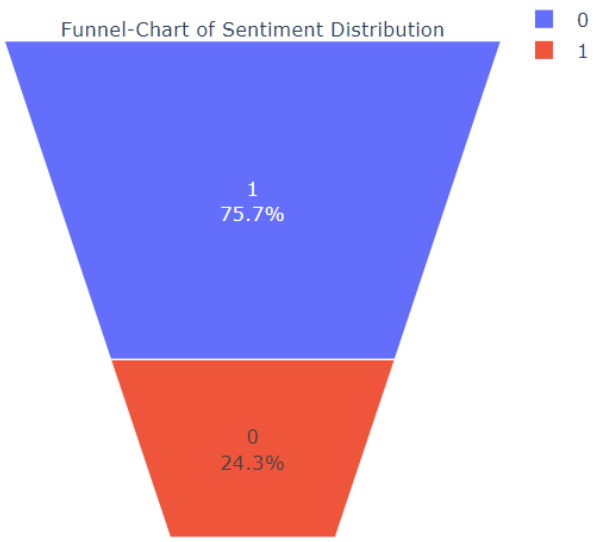
Η πρώτη μεταβλητή που είναι απαραίτητο να ερευνηθεί είναι η μεταβλητή “rating”. Για τον λόγο αυτό κατασκευάστηκε η πίτα της κατανομής των αξιολογήσεων (βλ. Εικόνα 5.4) ανά ποσοστό πλήθους αστερίων με τη βοήθεια της βιβλιοθήκης **matplotlib** της python.



Εικόνα 5.4: Ποσοστά κριτικών ανά πλήθος αστεριών

Όπως είναι φανερό, η πλειοψηφία των αξιολογήσεων έχουν 5 αστέρια, πράγμα που σημαίνει πως λίγοι παραπάνω από τους μισούς αγοραστές έμειναν ευχαριστημένοι. Όμως, το αμέσως επόμενο μεγαλύτερο ποσοστό κατέχουν οι αξιολογήσεις με 1 αστέρι, οι οποίες κατακτούν το 15% επί του συνόλου. Τέλος, το 13.6%, το 10.9% και το 8.5% αποτελούν κριτικές με 4, 3 και 2 αστέρια αντίστοιχα. Αυτό σημαίνει ότι το 24.3% από το σύνολο των αγοραστών που άφησαν κριτική, δεν έμειναν ευχαριστημένοι από την αγορά τους.

Έπειτα, μας ενδιαφέρει η εύρεση των ποσοστών των θετικών και αρνητικών κριτικών στο σύνολο των δεδομένων. Έτσι, στην Εικόνα 5.5, φαίνεται το διάγραμμα που κατασκευάστηκε με τη χρήση της βιβλιοθήκης **plotly** και αναπαριστά την κατανομή του συναισθήματος στο δείγμα, μετά τη συγχώνευση των ουδέτερων κριτικών στην κατηγορία των θετικών (όπου 1: «Θετικό», 0: «Αρνητικό»).



Εικόνα 5.5: Ποσοστά κριτικών ανά συναίσθημα

Παρατηρούμε ότι οι κλάσεις βαθμολογίας είναι εξαιρετικά ανισοβαρείς, κάτι που θα πρέπει να αντιμετωπιστεί πριν χρησιμοποιηθούν τα δεδομένα για ταξινόμηση. Ο τρόπος που θα γίνει αυτό, θα αναλυθεί αργότερα στην κατασκευή των μοντέλων.

5.5 Ταξινόμηση αξιολογήσεων με χρήση των λεξικών VADER, TextBlob

Στο στάδιο αυτό χρησιμοποιούνται οι βιβλιοθήκες TextBlob και VADER για να βρεθεί η βαθμολογία πολικότητας των κριτικών και να καταταχθούν ανάλογα με τη βαθμολογία αυτή, σε θετικές ή αρνητικές.

Με την εφαρμογή της βιβλιοθήκης TextBlob έχουμε ως έξοδο, μια καινούρια στήλη με όνομα "SA_tb" που εμπεριέχει την τιμή 1 σε κριτικές που έχουν πολικότητα μεγαλύτερη του μηδενός ($polarity > 0$) και την τιμή 0 σε αυτές που έχουν πολικότητα μικρότερη του μηδενός ($polarity < 0$). Αντίστοιχα, με την εφαρμογή της βιβλιοθήκης VADER, έχουμε ως έξοδο μια νέα στήλη που εισάγει την τιμή 1 σε κριτικές με συνολική βαθμολογία μεγαλύτερη του 0 ($compound > 0$) και την τιμή 0 σε κριτικές με συνολική βαθμολογία μικρότερη του 0 ($compound < 0$). Ένα δείγμα του συνόλου δεδομένων μετά την εφαρμογή των TextBlob και Vader φαίνεται στην Εικόνα 5.6.

body	rate_class	SA_tb	SA_vd
unlock potential browser support html video success failure largely go depend deeply find ecosys...	1	1	1
navigate warn useless navigation buy expect able say get direction postcode street name favorite...	0	1	1
disappoint sorry big fan station house want love first disappointment come air vent mount includ...	1	1	1
work brilliantly old car stereo aux really excellent little box trick get year old skoda octavia...	1	1	1
small perfectly form great sense hearing configure instal less minute use vent attachment rather...	1	1	1
navigation awful music control work feel good mount great mic pick voice loud music roar wind st...	1	1	1
meh good novelty wear take minute begin wonder money well spend else regard hardware build quali...	0	1	1

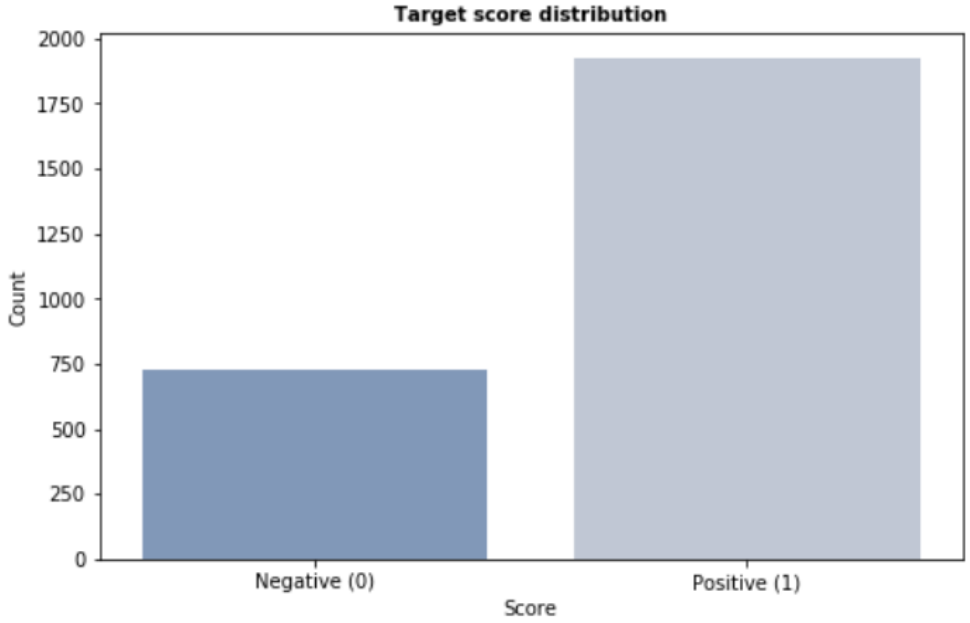
Εικόνα 5.6: Δείγμα του συνόλου δεδομένων μετά την εφαρμογή TextBlob και VADER

5.6 Ταξινόμηση αξιολογήσεων με χρήση μοντέλων Μηχανικής Μάθησης

Το επόμενο βήμα της ανάλυσης αποτελείται από την κατασκευή μοντέλων ταξινόμησης Μηχανικής Μάθησης για την ταξινόμηση της μεταβλητής "rate_class", δηλαδή την ταξινόμηση των κριτικών σε θετικές και αρνητικές.

Όμως, πριν την εφαρμογή των μοντέλων ταξινόμησης είναι απαραίτητο να γίνουν οι παρακάτω ενέργειες:

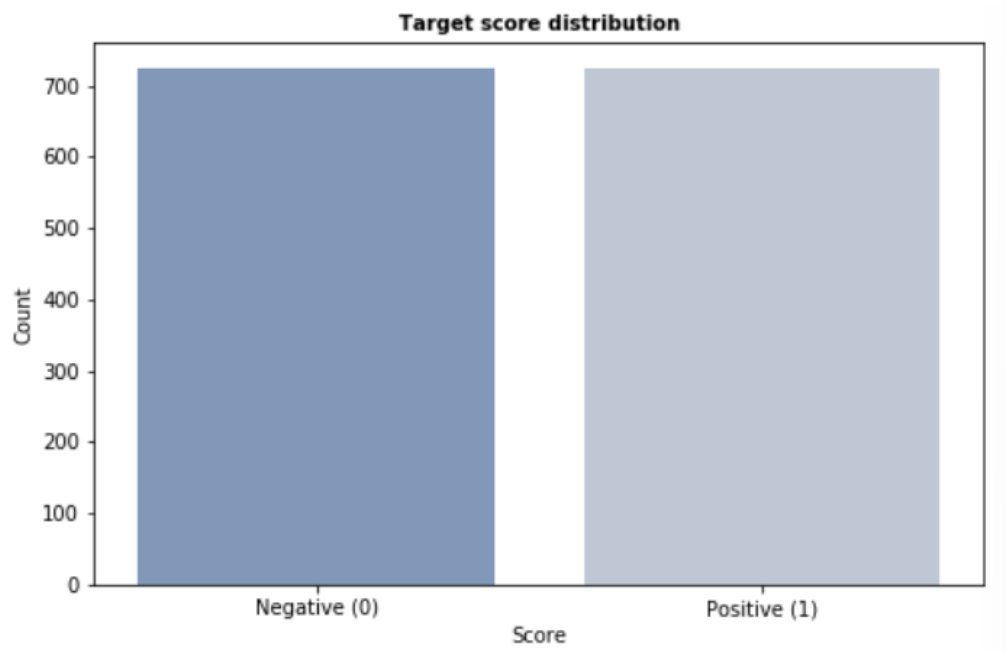
1. Αρχικά, θα πρέπει τα δεδομένα να χωριστούν σε δεδομένα εκπαίδευσης (training) και δοκιμής (test). Πιο συγκεκριμένα, η εκπαίδευση του μοντέλου έγινε με το 80% του συνόλου δεδομένων, ενώ το υπόλοιπο 20% χρησιμοποιήθηκε για την αξιολόγηση των μοντέλων. Επιπλέον, πρέπει να ελεγχθεί το πρόβλημα της κλάσης βαθμολογίας, η οποία -όπως έχει αναφερθεί παραπάνω- είναι εξαιρετικά ανισοβαρής (βλ. Εικόνα 5.7).



Εικόνα 5.7: Ραβδόγραμμα θετικής και αρνητικής κλάσης

Αφού, η θετική κλάση έχει πολλά παραπάνω δεδομένα από την αρνητική, θα εφαρμοστεί μείωση της κλάσεως των θετικών κριτικών, ώστε να έχει το μέγεθος της αρνητικής κλάσης. Η τεχνική αυτή ονομάζεται **downsampling.**

Το διάγραμμα της Εικόνας 5.7, μετά τη μείωση της θετικής κλάσης θα λάβει την μορφή του διαγράμματος της Εικόνας 5.8.



Εικόνα 5.8: Ραβδόγραμμα θετικής και αρνητικής κλάσης μετά την τεχνική του downsampling

Οι τεχνικές εξισορρόπησης των δεδομένων θα εφαρμοστεί μονάχα στα δεδομένα εκπαίδευσης (training set), αφού είναι σημαντικό να διατηρηθεί η ανισορροπία στα δεδομένα δοκιμής (test set), γιατί με αυτόν τον τρόπο τα μοντέλα θα αποδώσουν αποτελέσματα που ανταποκρίνονται σε μια πραγματική κατάσταση.

2. Εφαρμόζουμε την τεχνική Bag-of-words (BOW) στα κείμενα των αξιολογήσεων. Με αυτόν τον τρόπο μετατρέπουμε τα κείμενα από αδόμητα και ακατάστατα, σε δομημένα και σταθερού μήκους, έτσι ώστε να είναι κατανοητά και πιο εύκολα διαχειρίσιμα από τα μοντέλα. Πιο συγκεκριμένα, εφαρμόζονται τα μοντέλο CountVectorizer και TfidfVectorizer στο ίδιο σετ εκπαίδευσης και δοκιμής και στη συνέχεια εισάγονται διαδοχικά σε κάθε ένα μοντέλο για να συγκριθούν οι βαθμολογίες που επιτυγχάνονται με αυτές τις δύο τεχνικές αναπαράστασης.
3. Τέλος, χρησιμοποιείται η μέθοδος SVD, η οποία εφαρμόζεται για τη μείωση των διαστάσεων. Ο λόγος που κάνει σημαντική αυτήν διαδικασία είναι ότι τα CountVectorizer και TfidfVectorizer παράγουν ένα πολύ μεγάλο αριθμό χαρακτηριστικών, γεγονός που απαιτεί υψηλούς πόρους και αρκετό χρόνο για την ταξινόμηση των δεδομένων. Με αυτόν τον τρόπο μεγιστοποιούμε τον

αριθμό των χαρακτηριστικών διατηρώντας κοντά στο 75% της **εξηγούμενης διακύμανσης** (explained variance). Η εξηγούμενη διακύμανση μετρά την απόκλιση μεταξύ ενός μοντέλου και των πραγματικών δεδομένων. Συνεπώς, υψηλότερα ποσοστά εξηγούμενης διακύμανσης υποδηλώνουν ισχυρότερη ισχύ της συσχέτισης.

Logistic Regression & BOW

Το πρώτο μοντέλο κατασκευάστηκε με τη χρήση του αλγορίθμου Logistic Regression και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του CountVectorizer. Επιπλέον, πραγματοποιήθηκε ρύθμιση υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.1.

Πίνακας 5.1: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο Logistic Regression-BOW

C	0.1
max_iter	100
penalty	l2

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου Logistic Regression με χρήση των παραμέτρων του Πίνακα 5.1 και της τεχνικής Cross Validation με 5 folds φαίνονται στον Πίνακα 5.2.

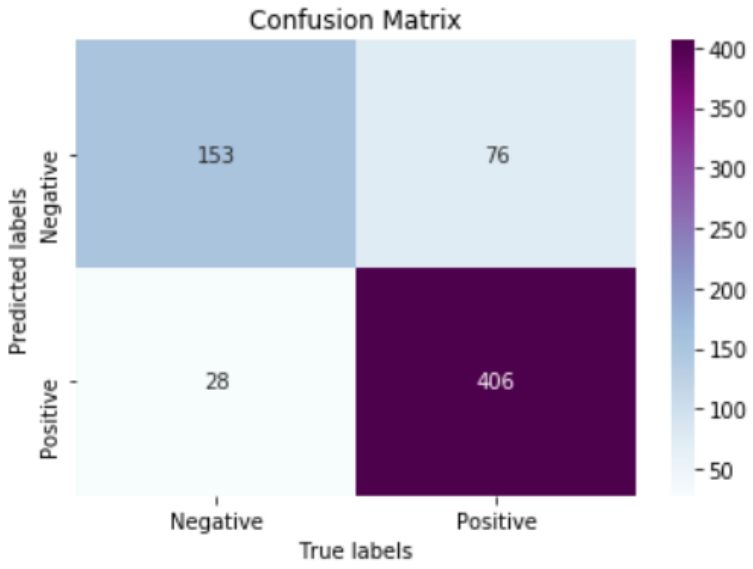
Πίνακας 5.2: Αποτελέσματα αξιολόγησης για το μοντέλο Logistic Regression-BOW

Accuracy on train data	88.5%
Accuracy on test data	84.3%
Precision on test data	84.2%
Recall on test data	93.5%
F1 on test data	88.6%
AUC test data	80.2%

Έπειτα, στον Πίνακα 5.3 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου Logistic Regression στα test data με χρήση του CountVectorizer και στην Εικόνα 5.9 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

Πίνακας 5.3: Λεπτομερής αναφορά ταξινόμησης του μοντέλου Logistic Regression-BOW (test data)

	Precision	Recall	F1-score	Support
Negative	0.85	0.67	0.75	229
Positive	0.84	0.94	0.89	434
Accuracy			0.84	663
Macro avg	0.84	0.80	0.82	663
Weight avg	0.84	0.84	0.84	663



Εικόνα 5.9: Πίνακας σύγχυσης του μοντέλου Logistic Regression-BOW (test data)

Logistic Regression & TF-IDF

Κατασκευάστηκε μοντέλο με τη χρήση του αλγορίθμου Logistic Regression και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του TfidfVectorizer. Επιπλέον, πραγματοποιήθηκε ρύθμιση

υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.4.

Πίνακας 5.4: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο Logistic Regression-TFIDF

C	1.5
max_iter	100
penalty	l2

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου Logistic Regression με χρήση των παραμέτρων του Πίνακα 5.4 και της τεχνικής Cross Validation με 5 folds φαίνονται στον Πίνακα 5.5.

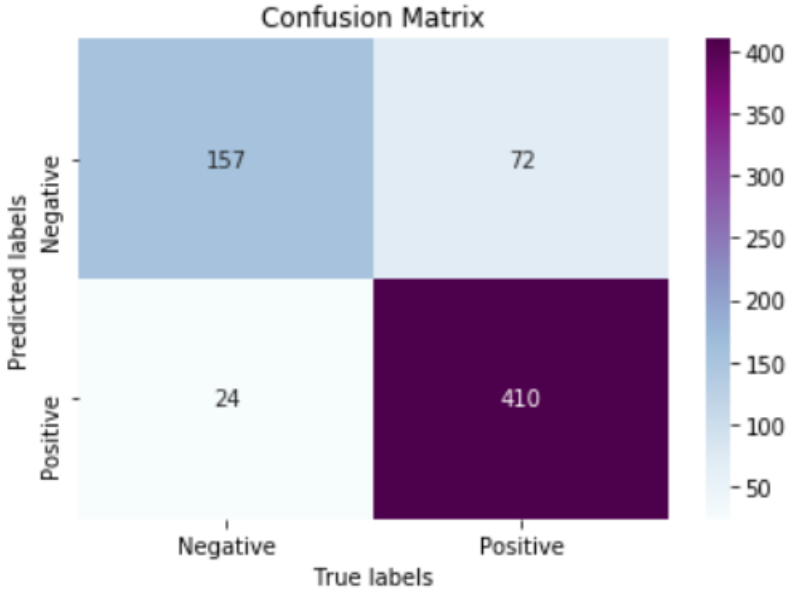
Πίνακας 5.5: Αποτελέσματα αξιολόγησης για το μοντέλο Logistic Regression-TFIDF

Accuracy on train data	91.9%
Accuracy on test data	85.5%
Precision on test data	85.1%
Recall on test data	94.5%
F1 on test data	89.5%
AUC test data	81.5%

Έπειτα, στον Πίνακα 5.6 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου Logistic Regression στα test data με χρήση του TfidfVectorizer και στην Εικόνα 5.10 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

Πίνακας 5.6: Λεπτομερής αναφορά ταξινόμησης του μοντέλου Logistic Regression-TFIDF (test data)

	Precision	Recall	F1-score	Support
Negative	0.87	0.69	0.77	229
Positive	0.85	0.94	0.90	434
Accuracy			0.86	663
Macro avg	0.86	0.82	0.83	663
Weight avg	0.86	0.86	0.85	663



Εικόνα 5.10: Πίνακας σύγχυσης του μοντέλου Logistic Regression-TFIDF (test data)

Support Vector Machine & BOW

Κατασκευάστηκε μοντέλο με τη χρήση του αλγορίθμου Support Vector Machine (SVM) και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του CountVectorizer. Επιπλέον, πραγματοποιήθηκε ρύθμιση υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.7.

Πίνακας 5.7: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο SVM-BOW

C	10
gamma	0.001
kernel	rbf

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του SVM μοντέλου με χρήση των παραμέτρων του Πίνακα 5.7 και της μεθόδου Cross Validation με 5 folds φαίνονται στον Πίνακα 5.8.

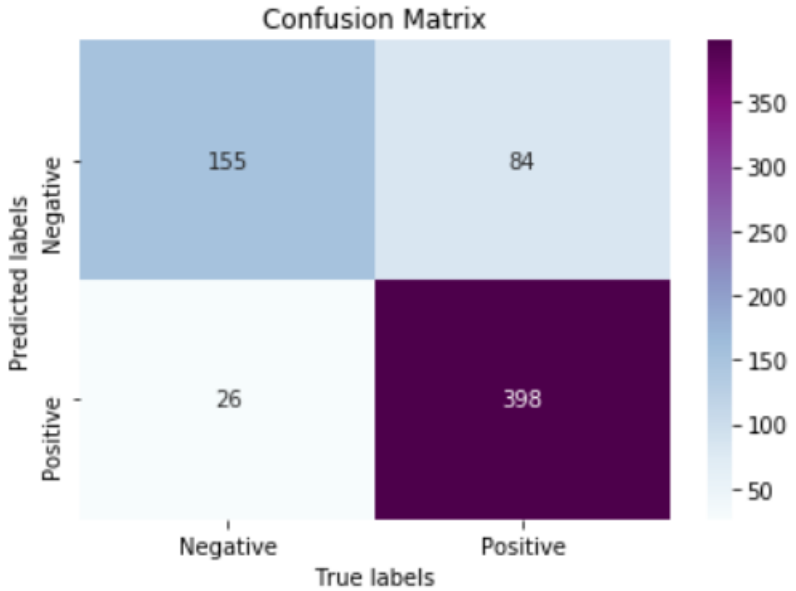
Πίνακας 5.8: Αποτελέσματα αξιολόγησης για το μοντέλο SVM-BOW

Accuracy on train data	87.3%
Accuracy on test data	83.4%
Precision on test data	82.6%
Recall on test data	93.9%
F1 on test data	87.9%
AUC test data	79.4%

Έπειτα, στον Πίνακα 5.9 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου SVM στα test data με χρήση του CountVectorizer και στην Εικόνα 5.11 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

Πίνακας 5.9: Λεπτομερής αναφορά ταξινόμησης του μοντέλου SVM-BOW (test data)

	Precision	Recall	F1-score	Support
Negative	0.86	0.65	0.74	239
Positive	0.83	0.94	0.88	424
Accuracy			0.83	663
Macro avg	0.84	0.79	0.81	663
Weight avg	0.84	0.83	0.83	663



Εικόνα 5.11:Πίνακας σύγχυσης του μοντέλου SVM-BOW (test data)

Support Vector Machine & TF-IDF

Κατασκευάστηκε μοντέλο με τη χρήση του αλγορίθμου Support Vector Machine (SVM) και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του TfidfVectorizer. Επιπλέον, πραγματοποιήθηκε ρύθμιση υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.10.

Πίνακας 5.10: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο SVM-TFIDF

C	1
gamma	1
kernel	rbf

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου SVM με χρήση των παραμέτρων του Πίνακα 5.10 και της μεθόδου Cross Validation με 5 folds φαίνονται στον Πίνακα 5.11.

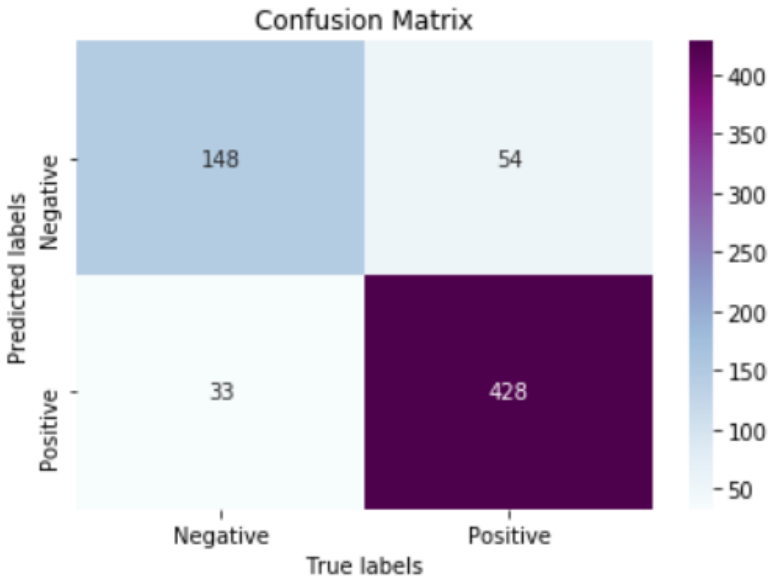
Πίνακας 5.11: Αποτελέσματα αξιολόγησης για το μοντέλο SVM- TFIDF

Accuracy on train data	96.7%
Accuracy on test data	86.9%
Precision on test data	88.8%
Recall on test data	92.8%
F1 on test data	90.8%
AUC test data	83.1%

Έπειτα, στον Πίνακα 5.12 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου SVM στα test data με χρήση του TfidfVectorizer και στην Εικόνα 5.12 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

Πίνακας 5.12: Λεπτομερής αναφορά ταξινόμησης του μοντέλου SVM-TFIDF (test data)

	Precision	Recall	F1-score	Support
Negative	0.82	0.73	0.77	204
Positive	0.89	0.93	0.91	461
Accuracy			0.87	663
Macro avg	0.85	0.83	0.84	663
Weight avg	0.87	0.87	0.87	663



Εικόνα 5.12: Πίνακας σύγχυσης του μοντέλου SVM-TFIDF (test data)

Gradient Boosting & BOW

Κατασκευάστηκε μοντέλο με τη χρήση του αλγορίθμου Gradient Boosting και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του CountVectorizer. Επιπλέον, πραγματοποιήθηκε

ρύθμιση υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.13.

Πίνακας 5.13: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο Gradient Boosting-BOW

n_estimators	200
min_samples_split	100
min_samples_leaf	50
max_depth	10
learning_rate	0.25

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου Gradient Boosting με χρήση των παραμέτρων του Πίνακα 5.13 και της μεθόδου Cross Validation με 5 folds φαίνονται στον Πίνακα 5.14.

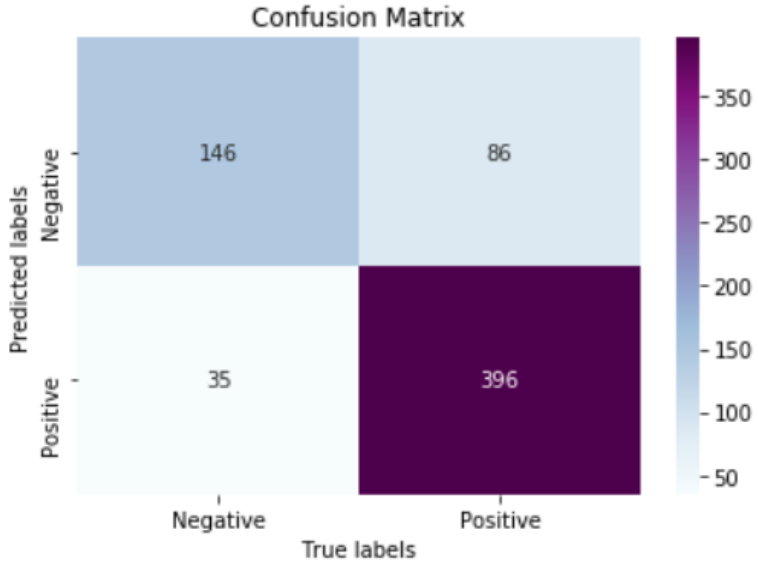
Πίνακας 5.14: Αποτελέσματα αξιολόγησης για το μοντέλο Gradient Boosting-BOW

Accuracy on train data	99.9%
Accuracy on test data	81.7%
Precision on test data	82.2%
Recall on test data	91.9%
F1 on test data	86.7%
AUC test data	77.4%

Έπειτα, στον Πίνακα 5.15 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου Gradient Boosting στα test data με χρήση του CountVectorizer και στην Εικόνα 5.13 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

Πίνακας 5.15: Λεπτομερής αναφορά ταξινόμησης του μοντέλου Gradient Boosting-BOW (test data)

	Precision	Recall	F1-score	Support
Negative	0.81	0.63	0.71	232
Positive	0.82	0.92	0.87	431
Accuracy			0.82	663
Macro avg	0.81	0.77	0.79	663
Weight avg	0.82	0.82	0.81	663



Εικόνα 5.13: Πίνακας σύγχυσης του μοντέλου Gradient Boosting-BOW (test data)

Gradient Boosting & TFIDF

Κατασκευάστηκε μοντέλο με τη χρήση του αλγορίθμου Gradient Boosting και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του TfidfVectorizer. Επιπλέον, πραγματοποιήθηκε ρύθμιση υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.16.

Πίνακας 5.16: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο Gradient Boosting-TFIDF

n_estimators	150
min_samples_split	150
min_samples_leaf	40
max_depth	30
learning_rate	0.25

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου Gradient Boosting με χρήση των παραμέτρων του Πίνακα 5.16 και της μεθόδου Cross Validation με 5 folds φαίνονται στον Πίνακα 5.17.

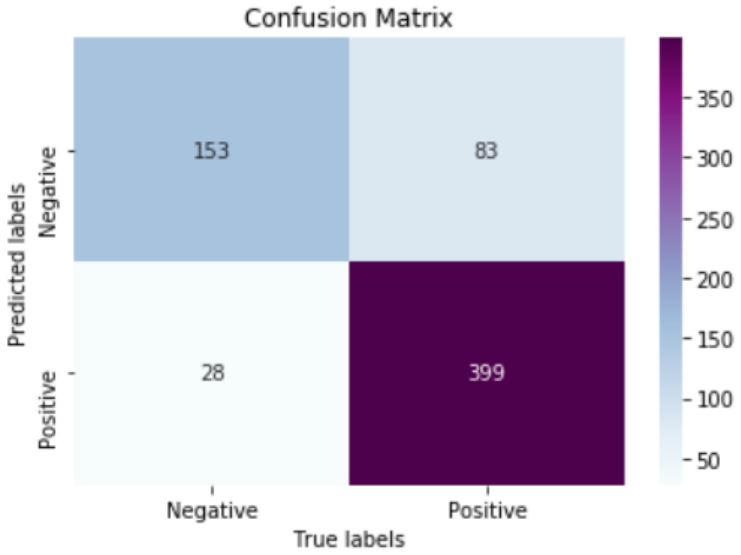
Πίνακας 5.17: Αποτελέσματα αξιολόγησης για το μοντέλο Gradient Boosting-TFIDF

Accuracy on train data	99.9%
Accuracy on test data	83.3%
Precision on test data	82.8%
Recall on test data	93.4%
F1 on test data	87.8%
AUC test data	79.1%

Έπειτα, στον Πίνακα 5.18 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου Gradient Boosting στα test data με χρήση του TfidfVectorizer και στην Εικόνα 5.14 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

Πίνακας 5.18: Λεπτομερής αναφορά ταξινόμησης του μοντέλου Gradient Boosting-TFIDF (test data)

	Precision	Recall	F1-score	Support
Negative	0.85	0.65	0.73	236
Positive	0.83	0.93	0.88	427
Accuracy			0.83	663
Macro avg	0.84	0.79	0.81	663
Weight avg	0.83	0.83	0.83	663



Εικόνα 5.14: Πίνακας σύγχυσης του μοντέλου Gradient Boosting-TFIDF (test data)

XG Boosting & BOW

Κατασκευάστηκε μοντέλο με τη χρήση του αλγορίθμου XG Boosting και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του CountVectorizer. Επιπλέον, πραγματοποιήθηκε ρύθμιση υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.19.

Πίνακας 5.19: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο XG Boosting-BOW

n_estimators	200
min_samples_leaf	50
min_child_weight	3
max_depth	15
learning_rate	0.15
gamma	0.2

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου XG Boosting με χρήση των παραμέτρων του Πίνακα 5.19 και της μεθόδου Cross Validation με 5 folds φαίνονται στον Πίνακα 5.20.

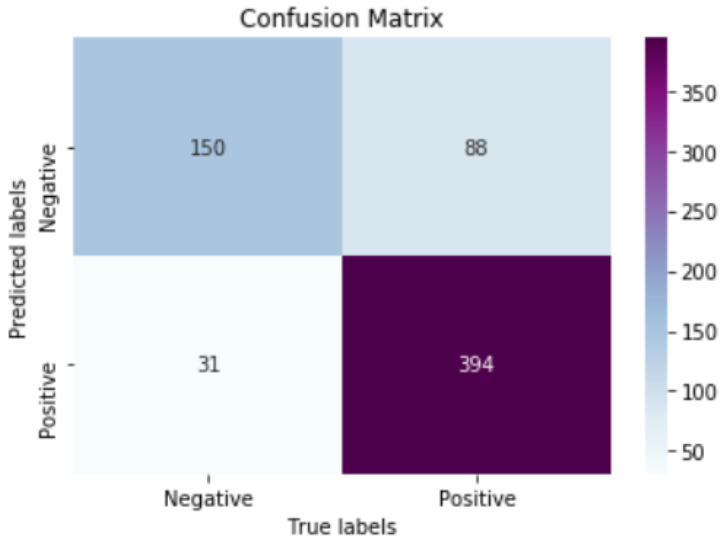
Πίνακας 5.20: Αποτελέσματα αξιολόγησης για το μοντέλο XG Boosting-BOW

Accuracy on train data	99.9%
Accuracy on test data	82.1%
Precision on test data	81.7%
Recall on test data	92.7%
F1 on test data	86.9%
AUC test data	77.9%

Έπειτα, στον Πίνακα 5.21 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου XG Boosting στα test data με χρήση του CountVectorizer και στην Εικόνα 5.15 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

Πίνακας 5.21: Λεπτομερής αναφορά ταξινόμησης του μοντέλου XG Boosting-BOW (test data)

	Precision	Recall	F1-score	Support
Negative	0.83	0.63	0.72	238
Positive	0.82	0.93	0.87	425
Accuracy			0.82	663
Macro avg	0.82	0.78	0.79	663
Weight avg	0.82	0.82	0.81	663



Εικόνα 5.15: Πίνακας σύγχυσης του μοντέλου XG Boosting-BOW (test data)

XG Boosting & TFIDF

Κατασκευάστηκε μοντέλο με τη χρήση του αλγορίθμου XG Boosting και έχει ως είσοδο τα δεδομένα στα οποία εφαρμόστηκε η τεχνική BOW για την εξαγωγή χαρακτηριστικών με χρήση του TfidfVectorizer. Επιπλέον, πραγματοποιήθηκε ρύθμιση υπερπαραμέτρων του μοντέλου και οι τελικές τιμές που δόθηκαν φαίνονται στον Πίνακα 5.22.

Πίνακας 5.22: Υπερπαράμετροι που χρησιμοποιήθηκαν έπειτα από tuning για το μοντέλο XG Boosting-TFIDF

n_estimators	200
min_samples_leaf	50
min_child_weight	3
max_depth	20
learning_rate	0.1
gamma	0.2

Τα αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου XG Boosting με χρήση των παραμέτρων του Πίνακα 5.22 και της μεθόδου Cross Validation με 5 folds φαίνονται στον Πίνακα 5.23.

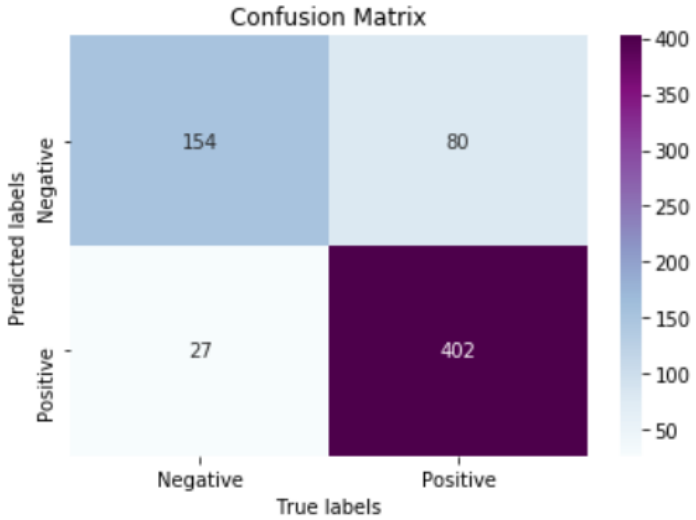
Πίνακας 5.23: Αποτελέσματα αξιολόγησης για το μοντέλο XG Boosting-TFIDF

Accuracy on train data	99.9%
Accuracy on test data	83.9%
Precision on test data	83.4%
Recall on test data	93.7%
F1 on test data	88.3%
AUC test data	79.8%

Έπειτα, στον Πίνακα 5.24 γίνεται λεπτομερής αναφορά της επίδοσης του μοντέλου XG Boosting στα test data με χρήση του TfidfVectorizer και στην Εικόνα 5.16 δίνεται ο αντίστοιχος πίνακας σύγχυσης.

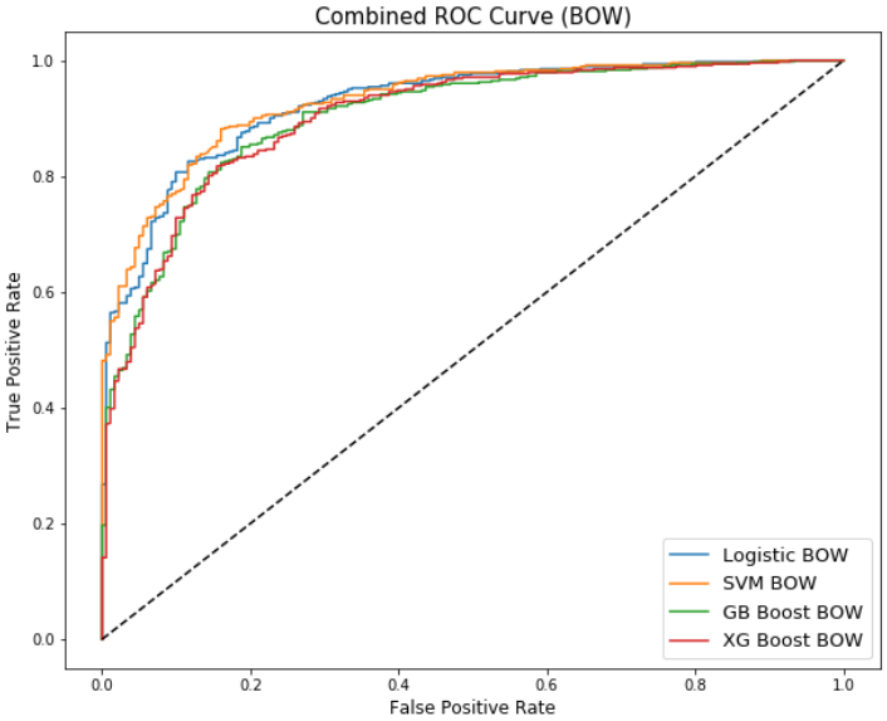
Πίνακας 5.24: Λεπτομερής αναφορά ταξινόμησης του μοντέλου XG Boosting-TFIDF (test data)

	Precision	Recall	F1-score	Support
Negative	0.85	0.66	0.74	234
Positive	0.83	0.94	0.88	429
Accuracy			0.84	663
Macro avg	0.84	0.80	0.81	663
Weight avg	0.84	0.84	0.83	663

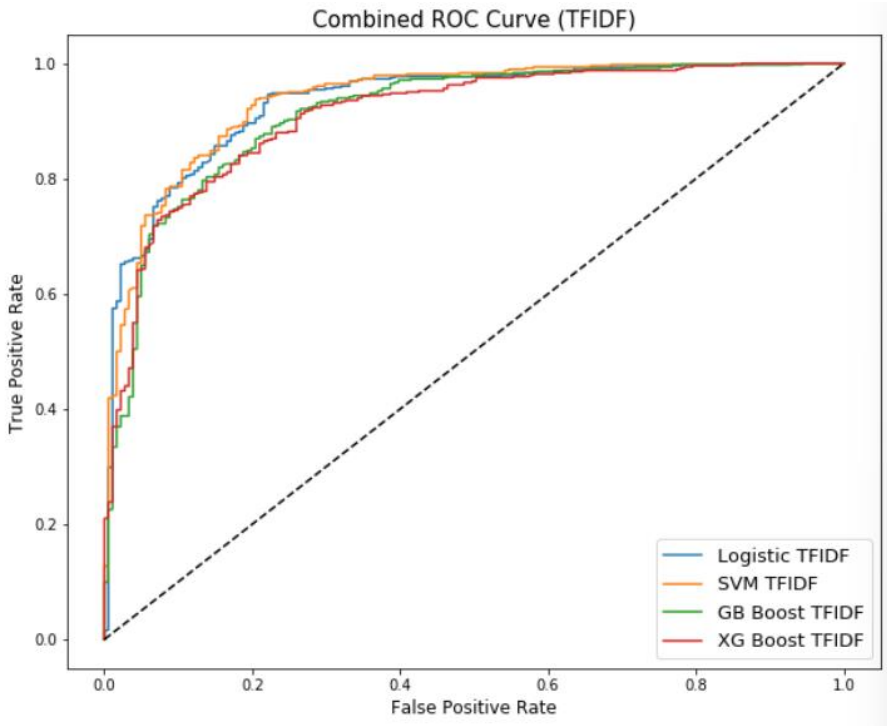


Εικόνα 5.16: Πίνακας σύγχυσης του μοντέλου XG Boosting- TFIDF (test data)

Τέλος, κατασκευάστηκαν δύο συνδυαστικά διαγράμματα καμπυλών ROC (Receiver Operator Characteristic), όπου το πρώτο αναπαριστά τις καμπύλες των μοντέλων στα οποία έχει εφαρμοστεί η τεχνική CountVectorizer (βλ. Εικόνα 5.17) και το δεύτερο τις καμπύλες των μοντέλων στα οποία έχει εφαρμοστεί η τεχνική TfidfVectorizer (βλ. Εικόνα 5.18).



Εικόνα 5.17: Καμπύλες ROC όλων των μοντέλων με χαρακτηριστικά BOW



Εικόνα 5.18: : Καμπύλες ROC όλων των μοντέλων με χαρακτηριστικά TF-IDF

Όπως αναφερθεί στην υποενότητα 3.2.5, για να είναι άριστο το μοντέλο, θα πρέπει η τιμή του AUC (Area under curve), δηλαδή του εμβαδόν της περιοχής κάτω από την καμπύλη ROC, να είναι κοντά στο 1. Αντίθετα, αν η τιμή του εμβαδόν κοντά στο 0, το μοντέλο δεν είναι επιτυχημένο και η ταξινόμηση δεν είναι η επιθυμητή. Στις Εικόνες 5.17, 5.18 παρατηρείται πως τα μοντέλα έχουν αρκετά καλή απόδοση με κάποια από αυτά να πλησιάζουν κοντά στην τιμή 1.

5.7 Προετοιμασία Θεματικής Κατηγοριοποίησης αξιολογήσεων

Το τελευταίο βήμα στην ανάλυση του συνόλου δεδομένων είναι η κατασκευή και αξιολόγηση μοντέλων θεματικής κατηγοριοποίησης με σκοπό την εξαγωγή θεμάτων «συζήτησης» των καταναλωτών για το δεδομένο προϊόν. Όπως έχει αναφερθεί παραπάνω, στην παρούσα εργασία θα κατασκευαστούν δύο μοντέλα θεματικής κατηγοριοποίησης, LDA και LSA. Το σύνολο δεδομένων το οποίο θα δεχθούν ως είσοδο τα μοντέλα, θα είναι μονάχα οι αρνητικές κριτικές που κατέταξε το βέλτιστο μοντέλο στη φάση της ανάλυσης συναισθήματος. Στην Εικόνα 5.19 φαίνονται οι συχνότερες λέξεις που εμφανίζονται στις αρνητικές αξιολογήσεις, έτσι ώστε να δημιουργηθεί μια πρώτη εικόνα των αρνητικών αξιολογήσεων.

Εικόνα 5.19: Wordcloud των πιο συχνών λέξεων στις αρνητικές κριτικές

Εφόσον το σύνολο δεδομένων έχει ήδη καθαριστεί για την ανάλυση συναισθήματος, δε θα χρειαστεί να γίνει ξανά για τη θεματική κατηγοριοποίηση. Η μόνη επιπλέον επεξεργασία που θα πρέπει να εφαρμοστεί στα δεδομένα είναι η κατασκευή χαρακτηριστικών. Δηλαδή, το σώμα κειμένου θα αναπαρασταθεί από αριθμητικά διανύσματα. Αυτό θα πραγματοποιηθεί με τη χρήση του CountVectorizer.

Έπειτα, δημιουργούνται τα δύο μοντέλα με είσοδο τον ίδιο αριθμό topics (έγινε η επιλογή των 10 topics) με σκοπό τη σύγκριση τους και την επιλογή του βέλτιστου για την τελική ανάλυση. Εδώ αξίζει να σημειωθεί ότι τα μοντέλα δεν είχαν ως είσοδο επιπλέον παραμέτρους. Η σύγκριση τους θα πραγματοποιηθεί με τη βοήθεια της τεχνικής μείωσης διαστάσεων **T-SNE**. Πρόκειται για μια μέθοδο που είναι χρήσιμη στην οπτικοποίηση των αποτελεσμάτων της ομαδοποίησης που επιτυγχάνει το κάθε μοντέλο. Η εμφάνιση των αποτελεσμάτων ενός μοντέλου θεματικής κατηγοριοποίησης αποτελεί μια δύσκολη διαδικασία αφού έπειτα την εξαγωγή των χαρακτηριστικών τα δεδομένα απέκτησαν πολλές διαστάσεις. Η τεχνική T-SNE λύνει αυτό το πρόβλημα αφού επιτυγχάνει μείωση των διαστάσεων και καθιστά δυνατή την απεικόνιση των θεμάτων των μοντέλων σε ένα δισδιάστατο επίπεδο.

Αφού επιλεχθεί το βέλτιστο μοντέλο, το τελικό βήμα της θεματικής κατηγοριοποίησης είναι η εύρεση του κατάλληλου αριθμού θεμάτων για την καλύτερη δυνατή απόδοση του. Αυτό θα γίνει με τη βοήθεια του **βαθμού συνοχής (coherence score)** των topics της μεθόδου, ο οποίος προσδιορίζει το πόσο ερμηνεύσιμα είναι τα θέματα των μοντέλων από τους ανθρώπους, παρατηρώντας την ομοιότητα των πιο συχνών λέξεων. Συνεπώς, δημιουργούνται πολλαπλά μοντέλα με διαφορετικές τιμές θεμάτων και βρίσκεται το coherence score σε κάθε περίπτωση. Τέλος, επιλέγεται ο αριθμός θεμάτων με τον υψηλότερο βαθμό συνοχής και γίνεται η ερμηνεία των topics.

Πριν την εύρεση του υψηλότερου βαθμού συνοχής είναι απαραίτητο να πραγματοποιηθούν κάποιες μετατροπές στο σώμα κειμένου. Πιο συγκεκριμένα:

1. Κατακερματισμός του σώματος κειμένου των κριτικών ανά λέξη με χρήση των regular expressions.
2. Αφαίρεση κάποιων λέξεων που εμφανίζονται πολύ συχνά, έχουν μικρή σημασία και δε προσδίδουν πολύτιμη πληροφορία καθώς και όσων έχουν πλήθος γραμμάτων μικρότερο ή ίσο με 3.

3. Δημιουργία bigrams και trigrams, ώστε να βρεθούν ζευγάρια λέξεων που εμφανίζονται πολύ συχνά μαζί με χρήση του πακέτου models της βιβλιοθήκης gensim.

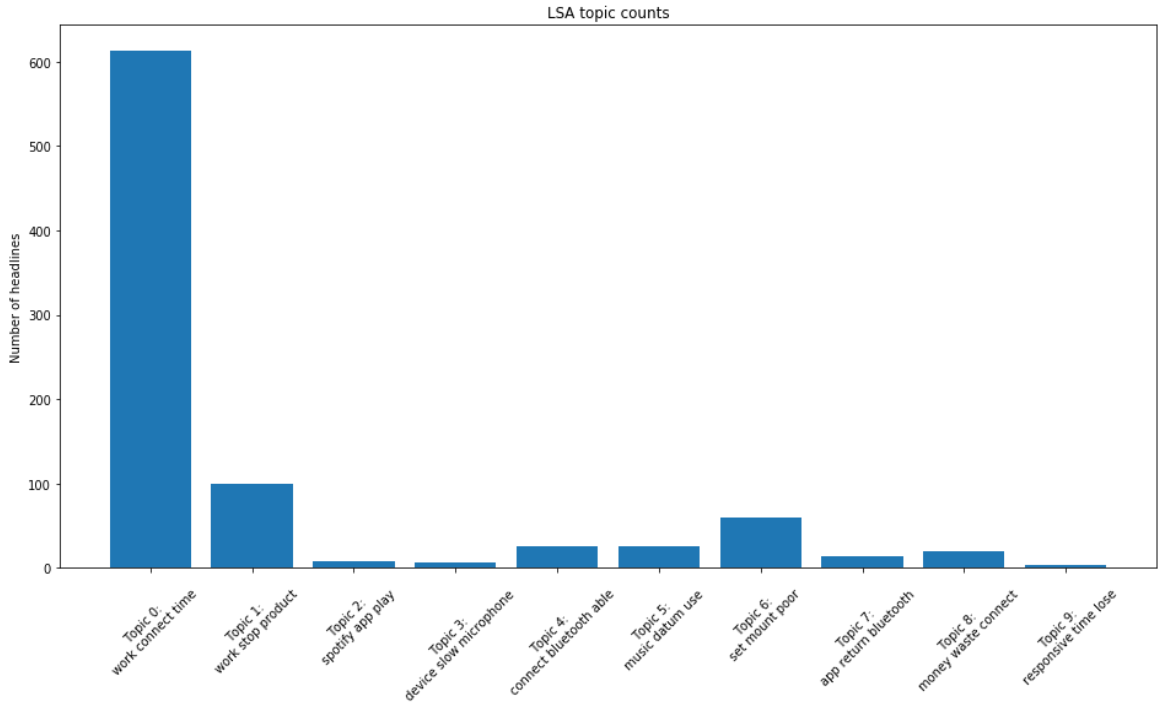
4. Δημιουργία του λεξικού που απαιτείται για την εφαρμογή της θεματικής κατηγοριοποίησης και αφορά την συχνότητα που εμφανίζεται μια λέξη στο σύνολο δεδομένων, με χρήση του corpora.Dictionary της βιβλιοθήκης genism. Στη φάση αυτή, για τη διατήρηση των υψηλής σημασίας λέξεων, έγινε φιλτράρισμα των λέξεων που εμφανίζονται σε λιγότερες από 15 κριτικές και σε περισσότερες από 0.5 (στο συνολικό μέγεθος του σώματος κειμένου, όχι ως απόλυτο αριθμό).

5. Εφαρμογή της μεθόδου doc2bow για τη δημιουργία χαρακτηριστικών του κειμένου. Με αυτόν τον τρόπο βρίσκεται ο αριθμός τον λέξεων και η συχνότητα που εμφανίζονται στο κείμενο. Έτσι η κάθε λέξη αναπαρίσταται με ένα διάνυσμα (word_id, word_frequency).

5.8 Σύγκριση των μοντέλων LDA, LSA

Για να γίνει η σύγκριση των μοντέλων LDA και LSA κατασκευάστηκαν ραβδογράμματα (βλ. Εικόνες 5.20, 5.22) σχετικά με τα μεγέθη των θεμάτων στα οποία καταλήγουν τα δύο μοντέλα ως προς τις πιο συχνές λέξεις και τα διαγράμματα T-SNE για κάθε μοντέλο (βλ. Εικόνες 5.21, 5.23).

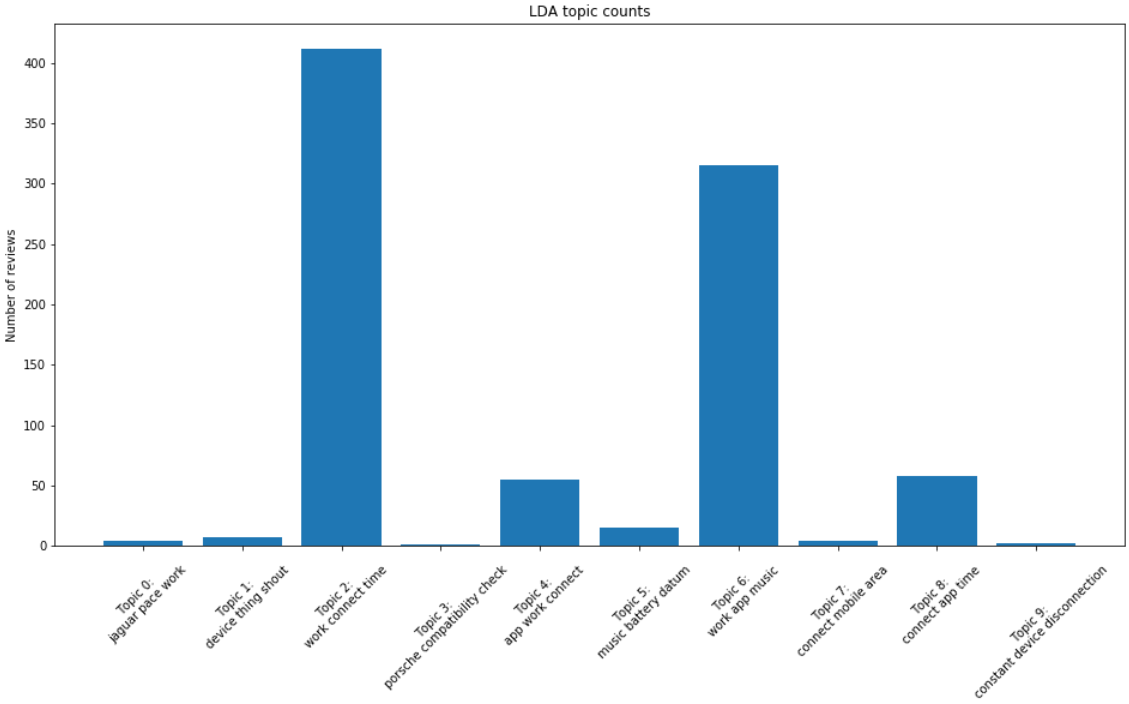
Μέθοδος LSA για 10 topics



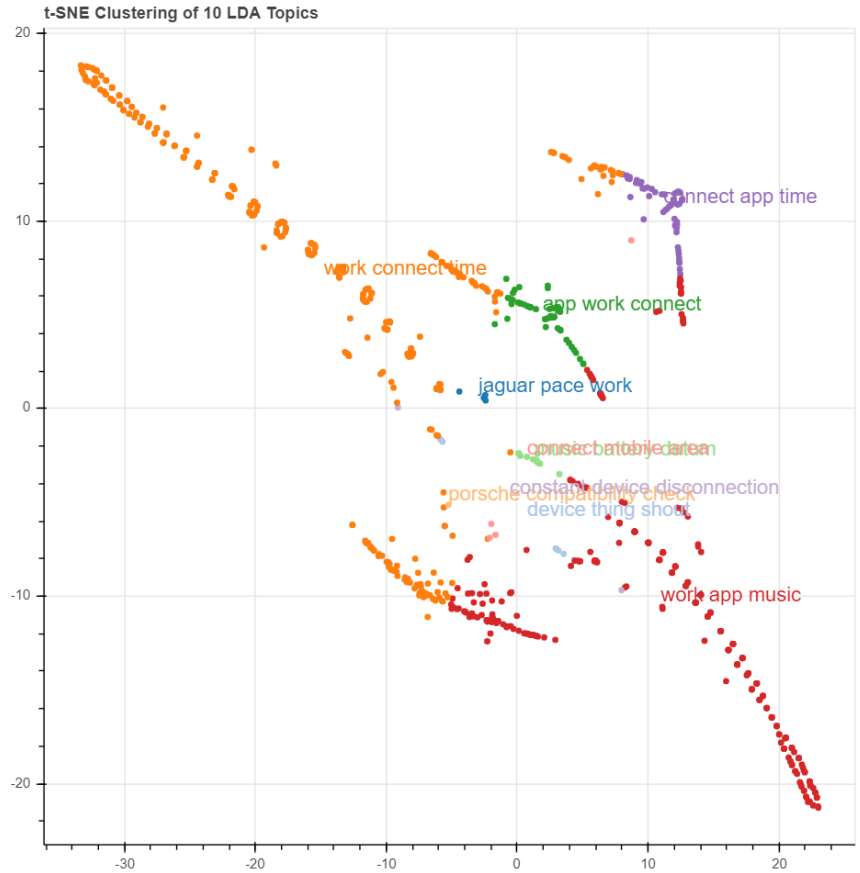
Εικόνα 5.20: Ραβδόγραμμα του πλήθους των πιο συχνών λέξεων ανά topic της μεθόδου

Εικόνα 5.21: Διάγραμμα T-SNE της μεθόδου LSA για 10 topics

Μέθοδος LDA για 10 topics



Εικόνα 5.22: Ραβδόγραμμα του πλήθους των πιο συχνών λέξεων ανά topic της μεθόδου LDA



Εικόνα 5.22: Διάγραμμα T-SNE της μεθόδου LDA για 10 topics

Γενικά, το διάγραμμα T-SNE ενός καλού μοντέλου αναπαρίσταται με σημεία (αξιολογήσεις) του ίδιου χρώματος (δηλαδή του ίδιου topic) συνωστισμένα μεταξύ τους και όσο το δυνατό περισσότερο απομονωμένα από τα σημεία (αξιολογήσεις) άλλου χρώματος (διαφορετικού topic). Αντίθετα, αν το διάγραμμα είναι χαοτικό και τα topics μπερδεύονται, τότε το μοντέλο δεν διαχωρίζει σωστά τα θέματα και συνεπώς τα αποτελέσματα του δεν αντιπροσωπεύουν την γνώμη των καταναλωτών στις αξιολογήσεις. Στο παρών πρόβλημα, τα διαγράμματα T-SNE (βλ. Εικόνες 5.21, 5.23) δείχνουν ότι η μέθοδος LDA κατηγοριοποιεί και ομαδοποιεί πολύ καλύτερα τα δεδομένα από ότι η LSA.

Στα ραβδογράμματα των δύο μοντέλων (βλ. Εικόνες 5.20, 5.22) φαίνεται πως και τα δύο μεροληπτούν ως προς κάποια θέματα, αφού κατατάσσουν μεγάλο αριθμό των αξιολογήσεων σε ένα ή δύο topics. Πιο συγκεκριμένα, η μέθοδος LSA κατατάσσει πάνω από 600 κριτικές στο θέμα 0 (topic 0), γεγονός που δείχνει μεροληψία του μοντέλου ως προς το συγκεκριμένο θέμα. Από την άλλη μεριά, η LDA φαίνεται να μεροληπτεί ως προς τα θέματα 2 και 6 (topics 2, 6) αφού πάνω από 700 αξιολογήσεις κατατάσσονται σε αυτά τα δύο θέματα. Παρόλα αυτά, η μέθοδος LDA υπερτερεί της LSA αφού μεροληπτεί σε μικρότερο βαθμό. Για τους παραπάνω λόγους, θα γίνει χρήση της LDA για τη διαδικασία του Topic Modeling.

5.9 Κατασκευή βέλτιστου μοντέλου Θεματικής Κατηγοριοποίησης

Έτσι, η διαδικασία προχωρά με τη δημιουργία του μοντέλου της LDA. Οι παράμετροι που χρησιμοποιούνται για την εκπαίδευση του μοντέλου φαίνονται στον Πίνακα 5.25.

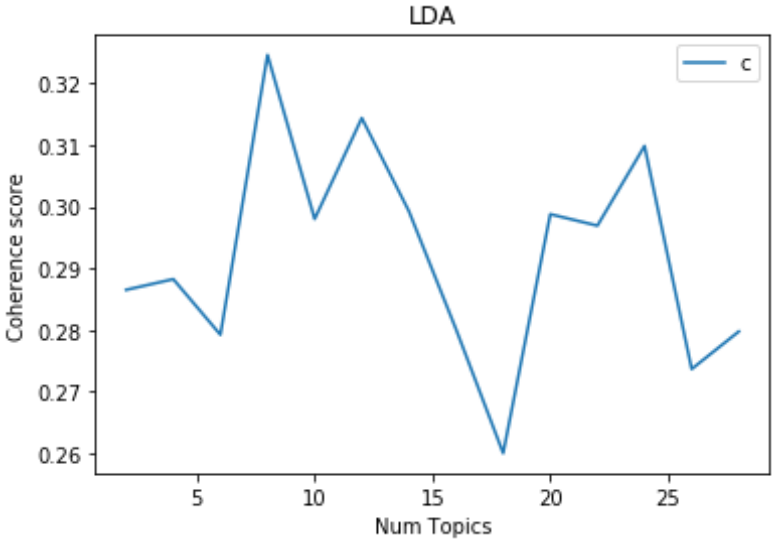
Πίνακας 5.25: Παράμετροι που επιλέχθηκαν για τη μέθοδο LDA

num_topics	10
chunksize	100
passes	10
alpha	auto

Όπου, num_topics: ο αριθμός των θεμάτων, chucksize: ο αριθμός των εγγράφων που χρησιμοποιούνται σε κάθε τμήμα εκπαίδευσης, passes: ο συνολικός αριθμός των

περασμάτων εκπαίδευσης και, alpha: η υπερπαράμετρος που επηρεάζει την αραιότητα των θεμάτων.

Στη συνέχεια, κατασκευάζεται το διάγραμμα βαθμού συνοχής (βλ. Εικόνα 5.24) ανά αριθμό θεμάτων και τα αποτελέσματα φαίνονται λεπτομερώς στον Πίνακα 5.26.



Εικόνα 5.23: Διάγραμμα βαθμού συνοχής ανά αριθμό topics της μεθόδου LDA

Πίνακας 5.26: Πίνακας τιμών του βαθμού συνοχής ανά αριθμό topics της μεθόδου LDA

Number of topics	Coherence value
2	0.2865
4	0.2882
6	0.2792
8	0.3246
10	0.2980
12	0.3144
14	0.2993
16	0.2801
18	0.2600
20	0.2988
22	0.2969
24	0.3099
26	0.2736
28	0.2797

Όπως φαίνεται στον Πίνακα 5.26, η τιμή που είναι υπογραμμισμένη με κίτρινο χρώμα είναι η καλύτερη τιμή του βαθμού συνοχής και επιτυγχάνεται με την κατασκευή του μοντέλου LDA για 8 θέματα.

Παρακάτω φαίνεται κάθε topic με τα αντίστοιχα βάρη των λέξεων που περιέχονται σε αυτό. Τα βάρη ορίζουν τη σημαντικότητα της κάθε λέξης στο κάθε θέμα, οπότε όσο μεγαλύτερο είναι το βάρος, τόσο αυτή η λέξη κλειδί προσδιορίζει το νόημα του topic.

Topic 1

'0.059*"call" + 0.053*"problem" + 0.051*"worth" + 0.041*"price" + '

'0.040*"mobile" + 0.040*"week" + 0.040*"purchase" + 0.040*"iphone" + '

'0.035*"expect" + 0.032*"turn"

Topic 2

'0.130*"bluetooth" + 0.115*"poor" + 0.081*"voice" + 0.057*"sound" + '

'0.039*"quality" + 0.035*"recognition" + 0.034*"audio" + 0.034*"issue" + '

'0.034*"frustrating" + 0.031*"different

Topic 3

'0.107*"compatible" + 0.072*"fail" + 0.071*"avoid" + 0.060*"refund" + '

'0.053*"support" + 0.042*"hear" + 0.041*"month" + 0.036*"unable" + '

'0.035*"service" + 0.035*"shout"

Topic 4

'0.093*"lose" + 0.058*"signal" + 0.055*"unit" + 0.054*"stop" + '

'0.047*"internet" + 0.045*"first" + 0.043*"actually" + 0.043*"hard" + '

'0.043*"difficult" + 0.041*"song"

Topic 5

'0.122*"useless" + 0.109*"disappoint" + 0.060*"able" + 0.056*"spend" + '

'0.051*"drive" + 0.041*"journey" + 0.038*"properly" + 0.038*"setup" + '

'0.036*"trouble" + 0.036*"useful"

Topic 6

'0.164*"spotify" + 0.121*"radio" + 0.098*"bother" + 0.087*"disconnect" + '

'0.051*"reconnect" + 0.047*"show" + 0.042*"vehicle" + 0.041*"drive" + '

'0.039*"maybe" + 0.038*"instead"

Topic 7

'0.094*"pointless" + 0.092*"wire" + 0.078*"mount" + 0.069*"vent" + '

'0.057*"terrible" + 0.054*"ever" + 0.053*"stereo" + 0.045*"bluetooth" + '

'0.045*"save" + 0.043*"option”

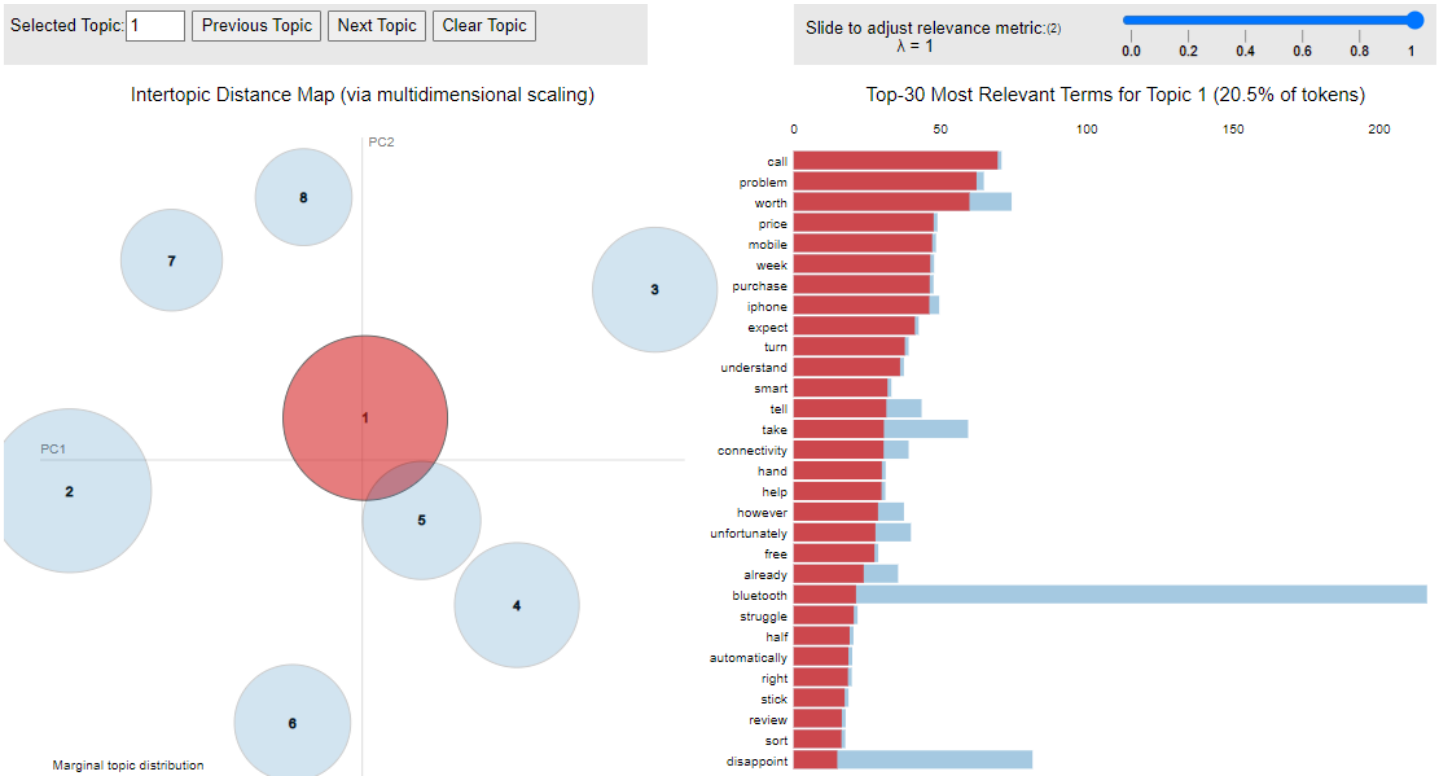
Topic 8

'0.175*"speaker" + 0.116*"cable" + 0.098*"hope" + 0.068*"battery" + '

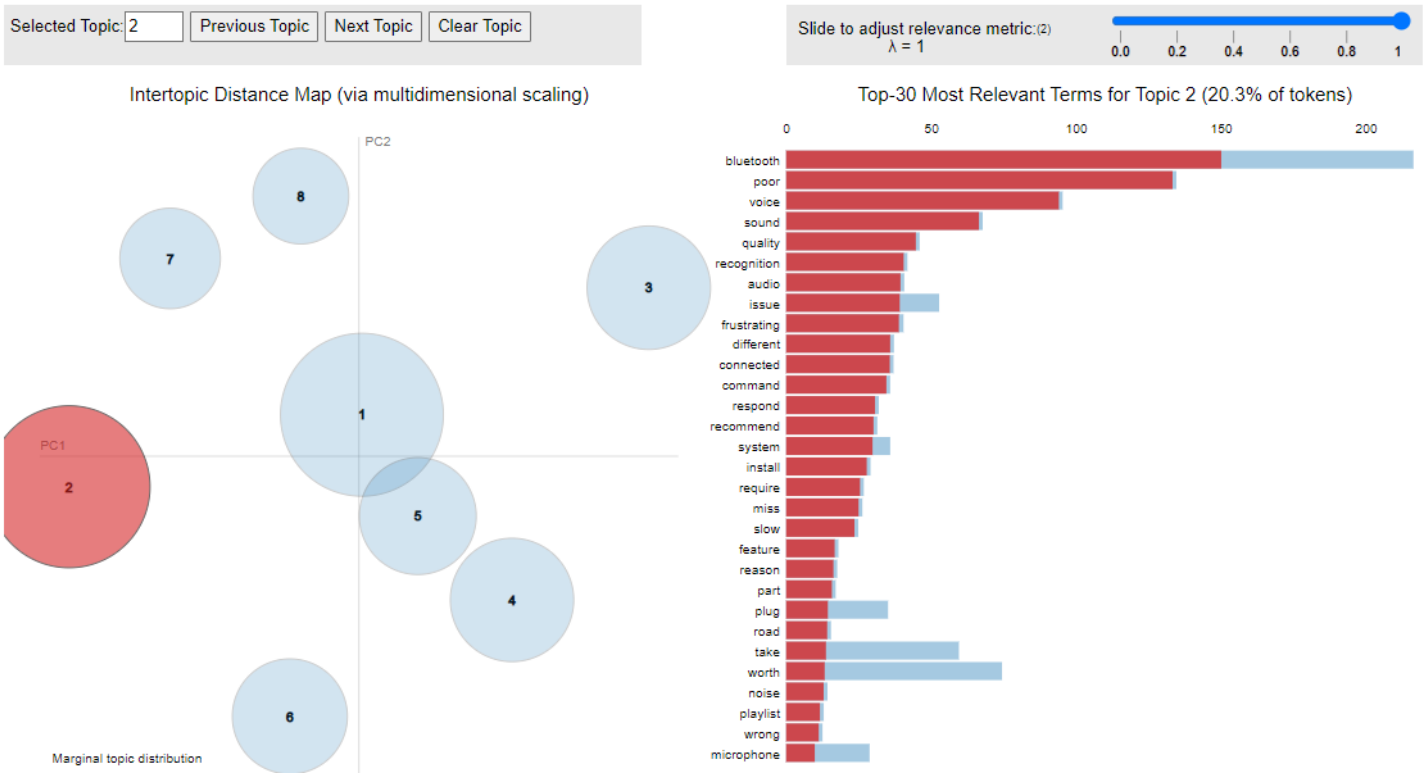
'0.056*"leave" + 0.055*"high" + 0.044*"enough" + 0.039*"drain" + '

'0.038*"shame" + 0.038*"happen"

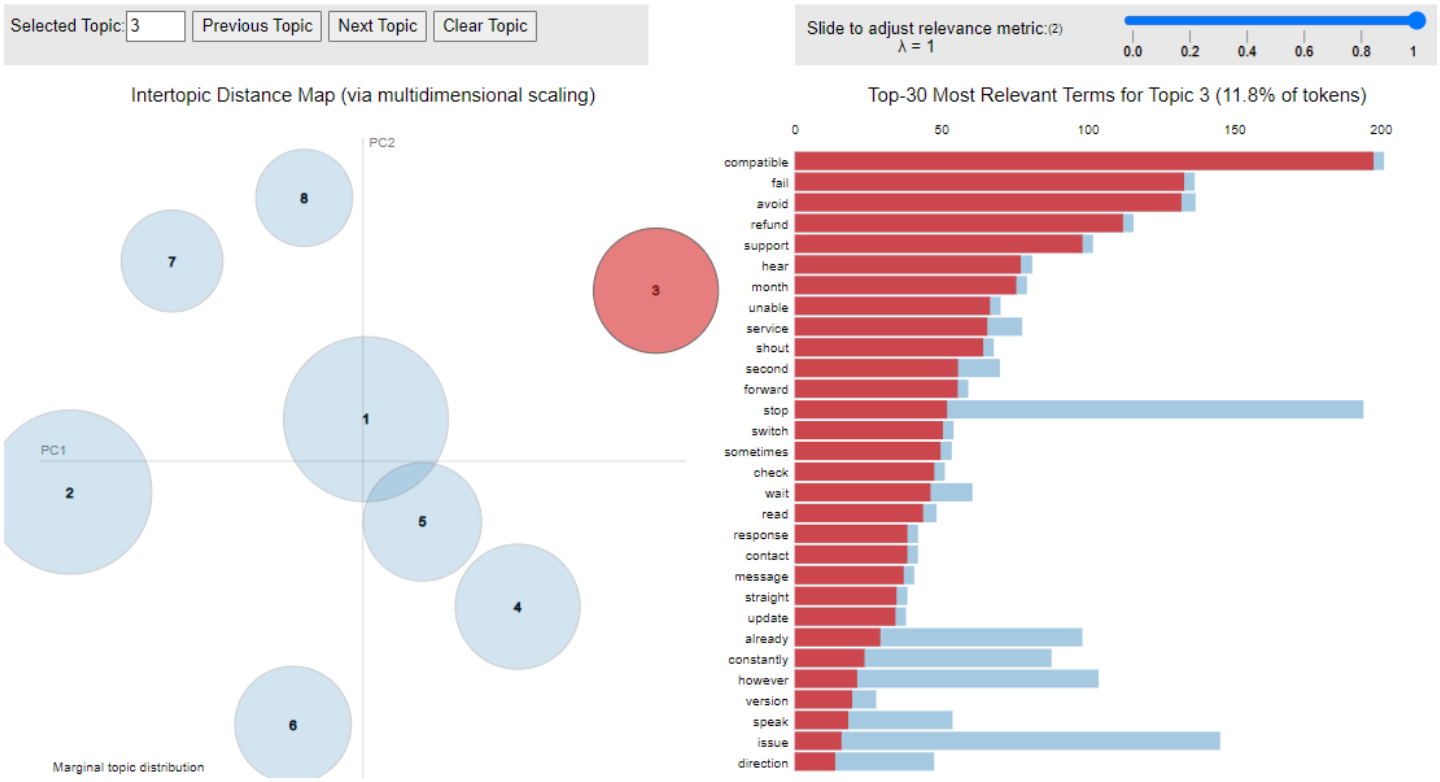
Στις επόμενες σελίδες παρουσιάζονται οπτικά τα topics καθένα ξεχωριστά, με χρήση του πακέτου pyLDAvis (βλ. Εικόνες 5.25-6.32).



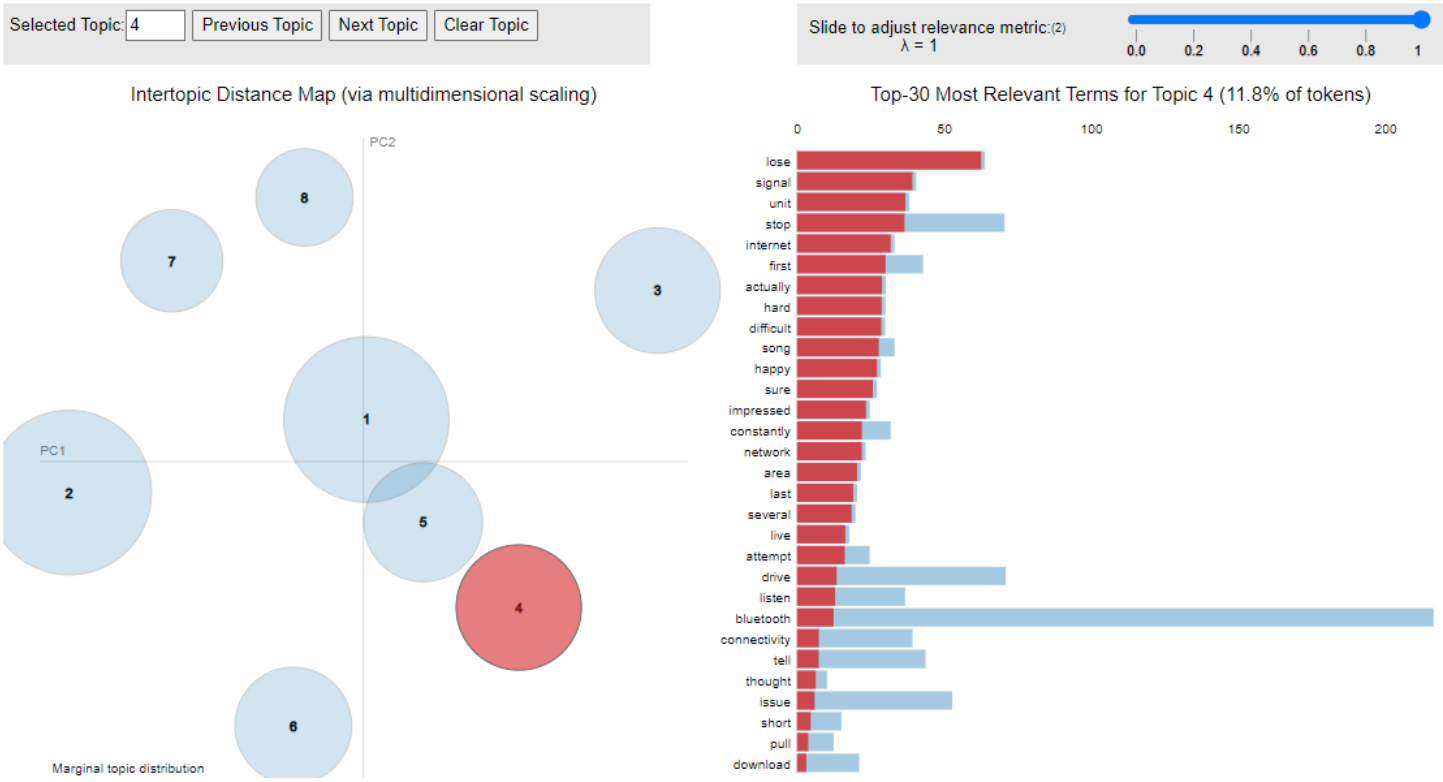
Εικόνα 5.25: Αναπαράσταση του topic 1 στο διάγραμμα pyLDAvis



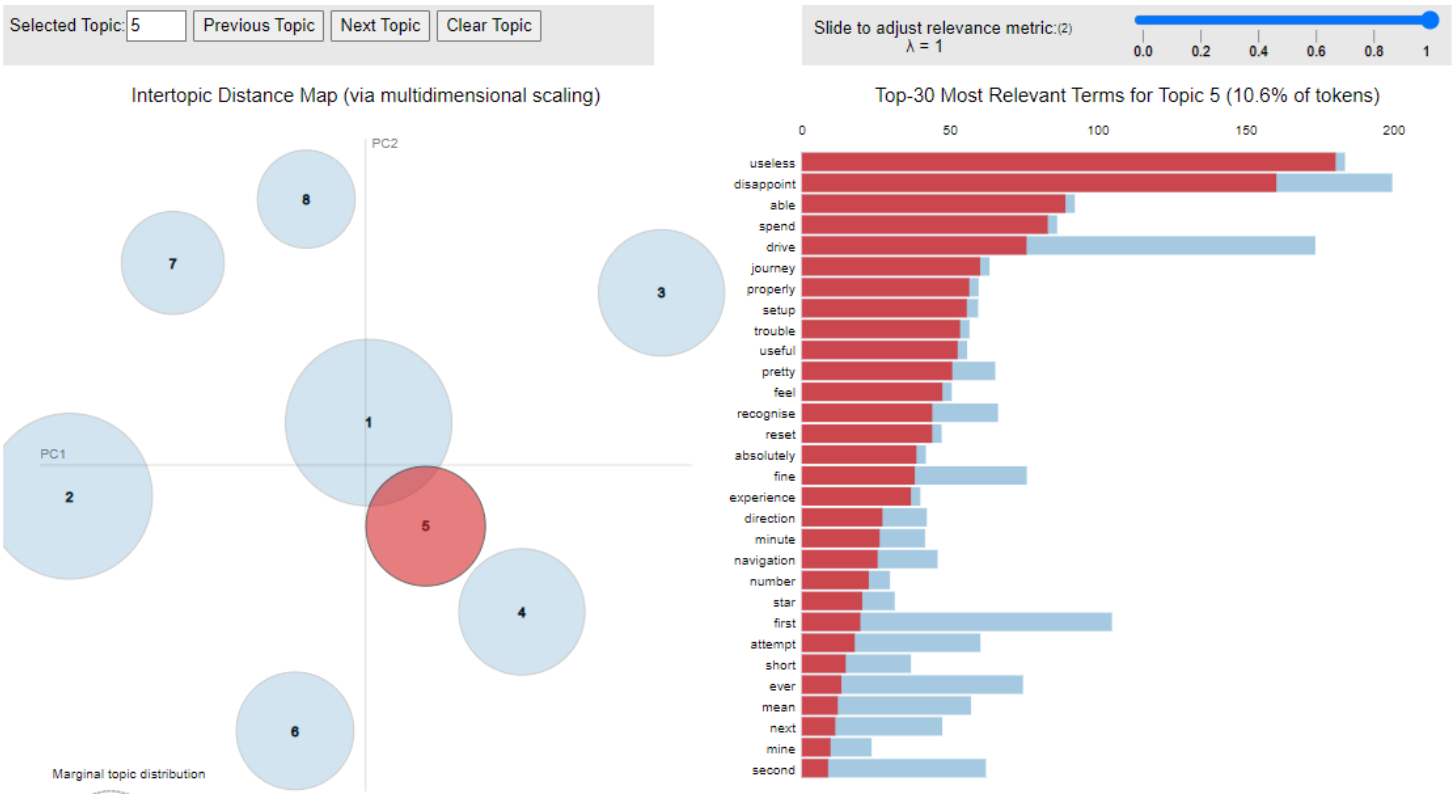
Εικόνα 5.24: Αναπαράσταση του topic 2 στο διάγραμμα pyLDAvis



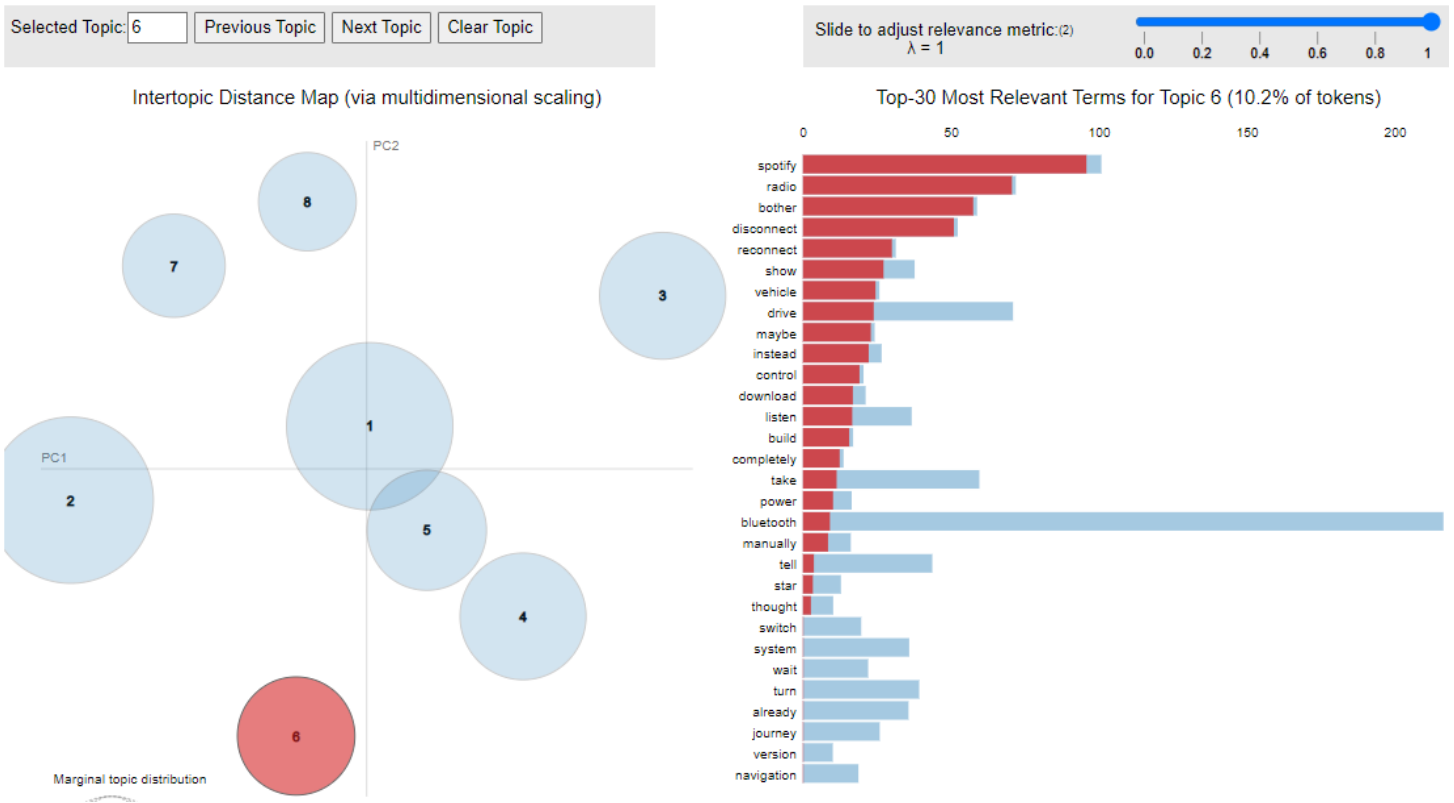
Εικόνα 5.27: Αναπαράσταση του topic 3 στο διάγραμμα pyLDAvis



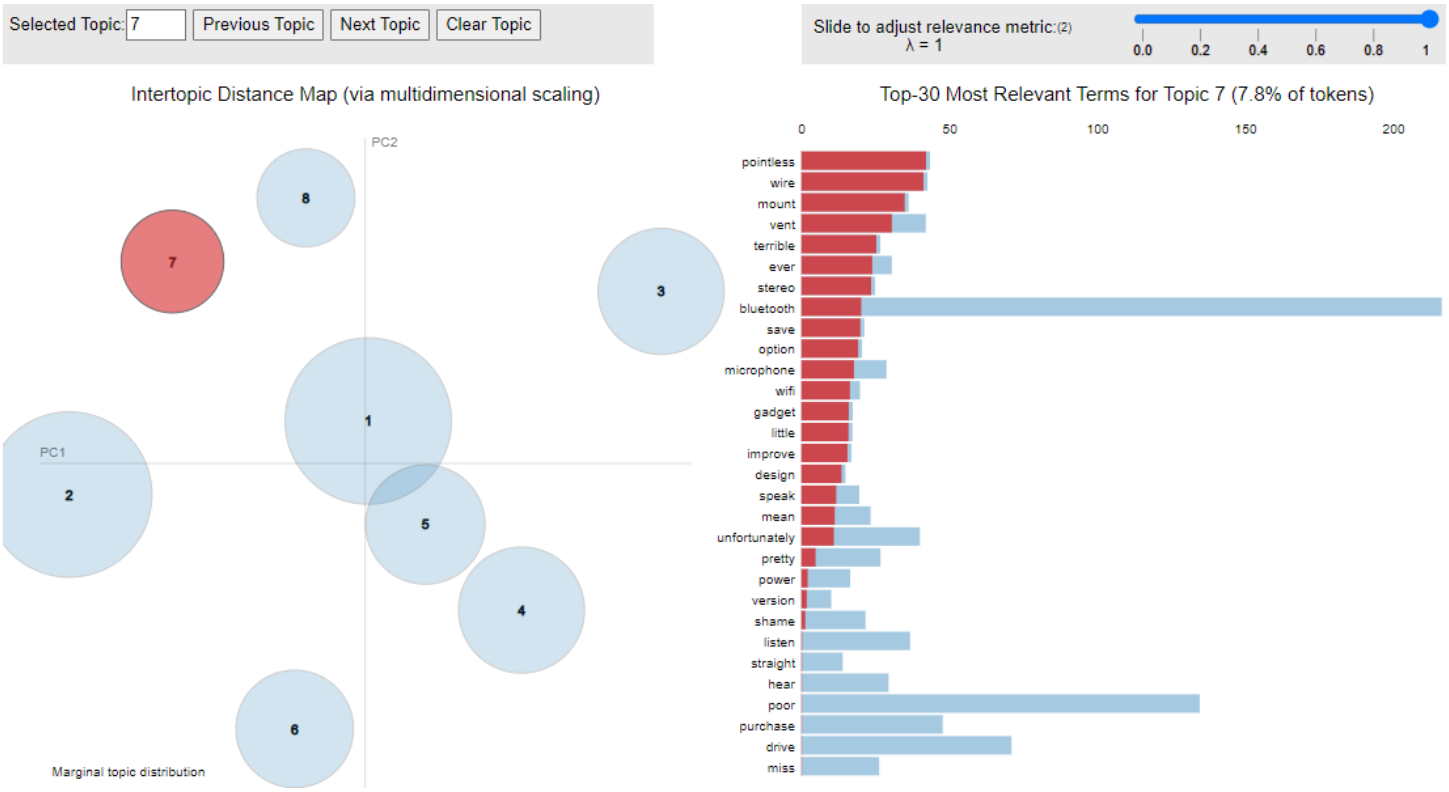
Εικόνα 5.26: Αναπαράσταση του topic 4 στο διάγραμμα pyLDAvis



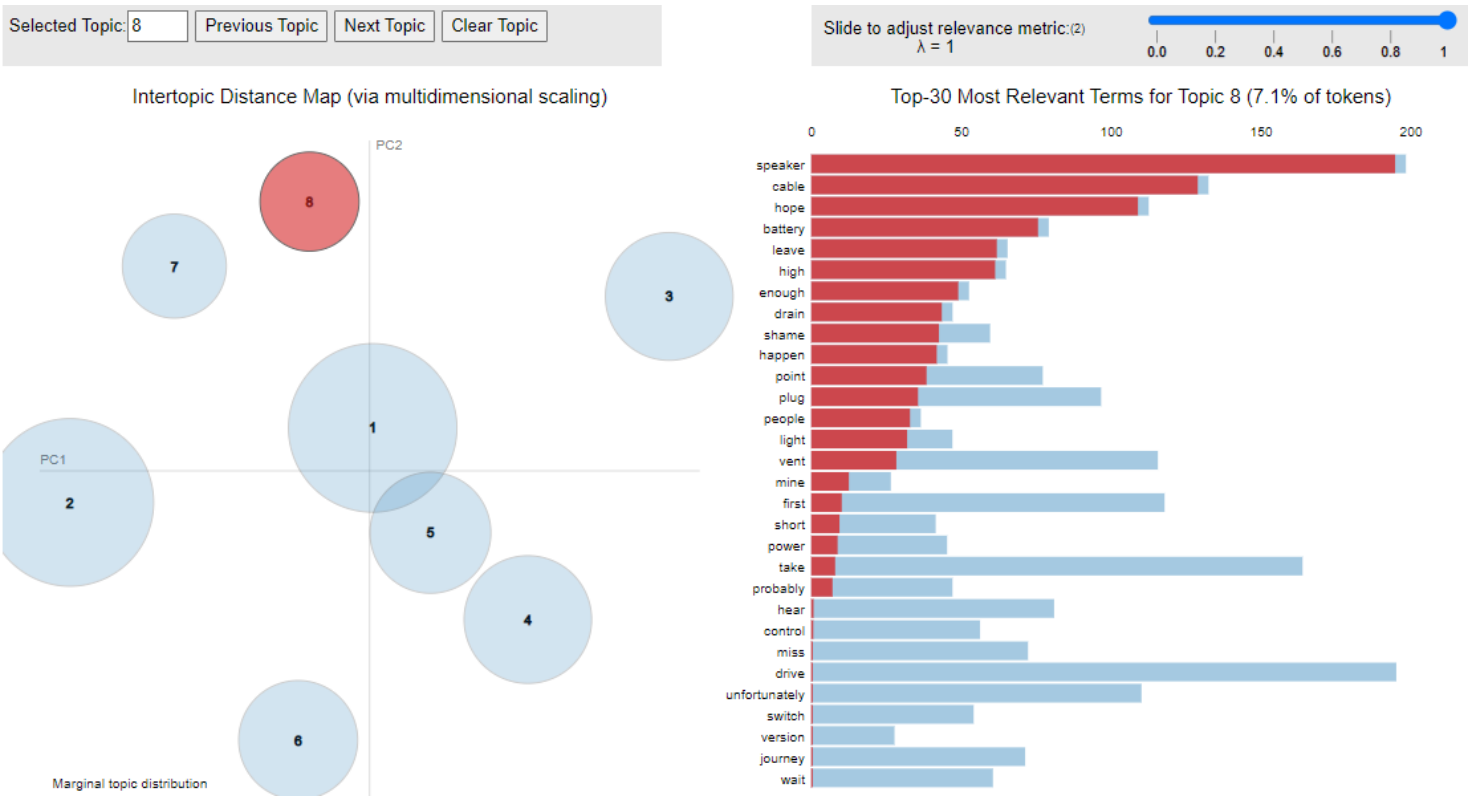
Εικόνα 5.28: Αναπαράσταση του topic 5 στο διάγραμμα pyLDAvis



Εικόνα 5.29: Αναπαράσταση του topic 6 στο διάγραμμα pyLDAvis



Εικόνα 5.31: Αναπαράσταση του topic 7 στο διάγραμμα pyLDAvis



Εικόνα 5.30: Αναπαράσταση του topic 8 στο διάγραμμα pyLDAvis

Στο διάγραμμα pyLDAvis, κάθε κύκλος αναπαριστά ένα θέμα. Το μέγεθος του κάθε κύκλου είναι ανάλογο με τη συχνότητα του θέματος που του αντιστοιχεί. Αυτό σημαίνει ότι όσο πιο μεγάλος είναι ο κύκλος τόσο πιο διαδεδομένο είναι το θέμα που αναπαριστά. Όσο πιο διάσπαρτοι και μεγάλοι είναι οι κύκλοι του διαγράμματος, τόσο καλύτερη και η απόδοση της θεματικής κατηγοριοποίησης. Επιπλέον, στο κάθε θέμα, οι γκρίζες μπάρες στα δεξιά υποδεικνύουν τη συνολική συχνότητα της κάθε λέξης, ενώ οι κόκκινες προσδιορίζουν τη συχνότητα της λέξης στο επιλεγμένο θέμα.

Στα διαγράμματα pyLDAvis που κατασκευάστηκαν σε αυτήν την εργασία (βλ. Εικόνες 5.25-6.32), οι κύκλοι είναι αρκετά ομοιόμορφοι, παρόλο που οι κύκλοι του topic 1 και το topic 2 φαίνεται να έχουν λίγο μεγαλύτερη διάμετρο. Αυτό σημαίνει ότι τα topics 1 και 2 είναι τα περισσότερο διαδεδομένα. Ακόμα, οι κύκλοι είναι αρκετά καλά διαχωρισμένοι, με τα topics 1 και 5 να έχουν μια μικρή τομή. Άρα, φαίνεται ότι έχουν κάποιες κοινές λέξεις. Παρόλα αυτά, η κατηγοριοποίηση των θεμάτων από την LDA είναι αρκετά ικανοποιητική.

6 ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Σύγκριση αποτελεσμάτων ταξινόμησης

Αρχικά, γίνεται η αξιολόγηση και σύγκριση των μεθόδων TextBlob και Vader. Για τον σκοπό αυτό, έγινε σύγκριση των αποτελεσμάτων της καθεμίας μεθόδου με τη στήλη "rate_class" του συνόλου δεδομένων. Τα αποτελέσματα φαίνονται στον Πίνακα 6.1.

Lexicon	Accuracy
TextBlob	78.8%
VADER	80.9%

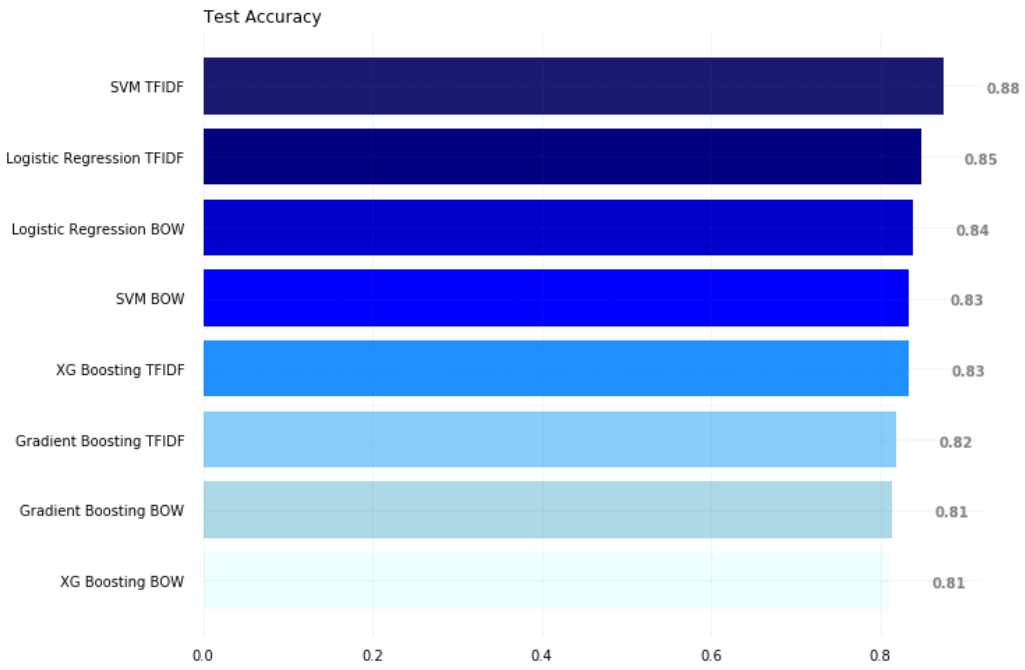
Πίνακας 6.1: Accuracy των μεθόδων ταξινόμησης TextBlob, VADER

Στον Πίνακα 6.2 φαίνονται τα αποτελέσματα της αξιολόγησης των μοντέλων στα test data, ταξινομημένα κατά φθίνουσα σειρά των τιμών του accuracy. Οι καλύτερες επιδόσεις για κάθε μετρική είναι υπογραμμισμένες με κίτρινο χρώμα.

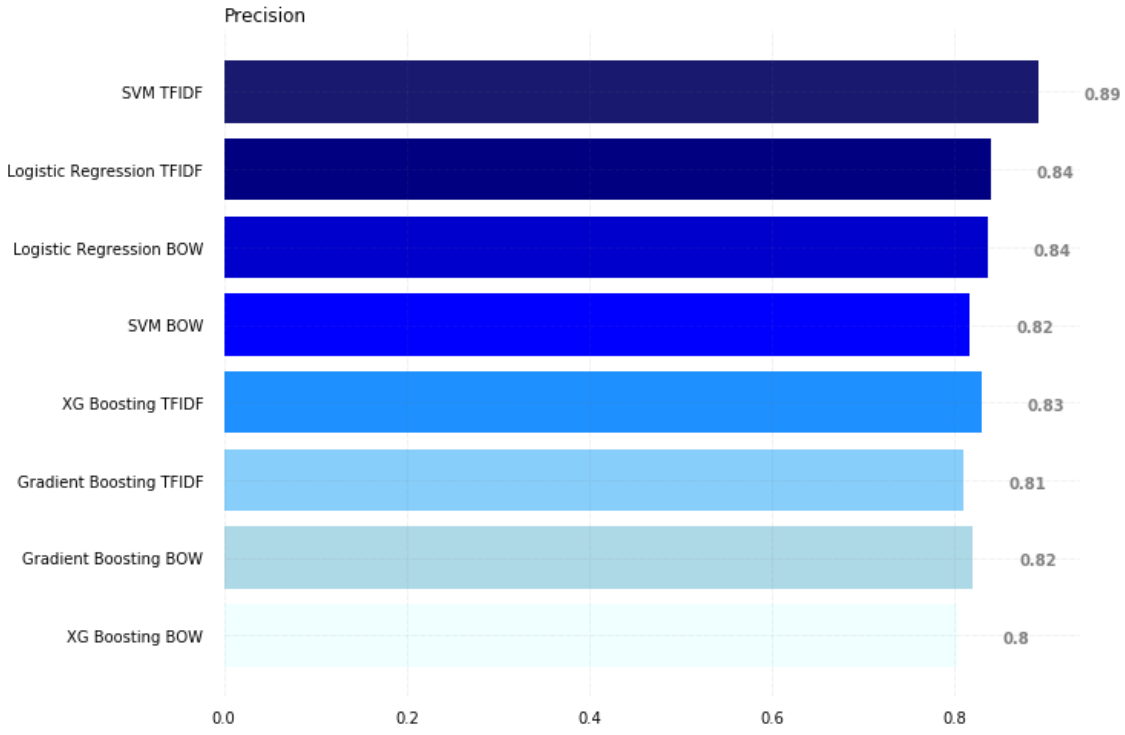
Classifier	Accuracy	Precision	Recall	F1 Score	AUC
SVM TFIDF	0.869	0.888	0.928	0.908	0.831
Logistic Regression_ TFIDF	0.855	0.851	0.945	0.895	0.815
Logistic Regression BOW	0.843	0.842	0.935	0.886	0.802
XG Boosting TFIDF	0.839	0.834	0.937	0.883	0.798
SVM BOW	0.834	0.826	0.939	0.879	0.794
Gradient Boosting TFIDF	0.833	0.828	0.934	0.878	0.791
XG Boosting TFIDF	0.821	0.817	0.927	0.869	0.779
Gradient Boosting BOW	0.817	0.822	0.919	0.867	0.774

Πίνακας 6.2: Συγκεντρωτικός πίνακας αποτελεσμάτων μοντέλων Μηχανικής Μάθησης

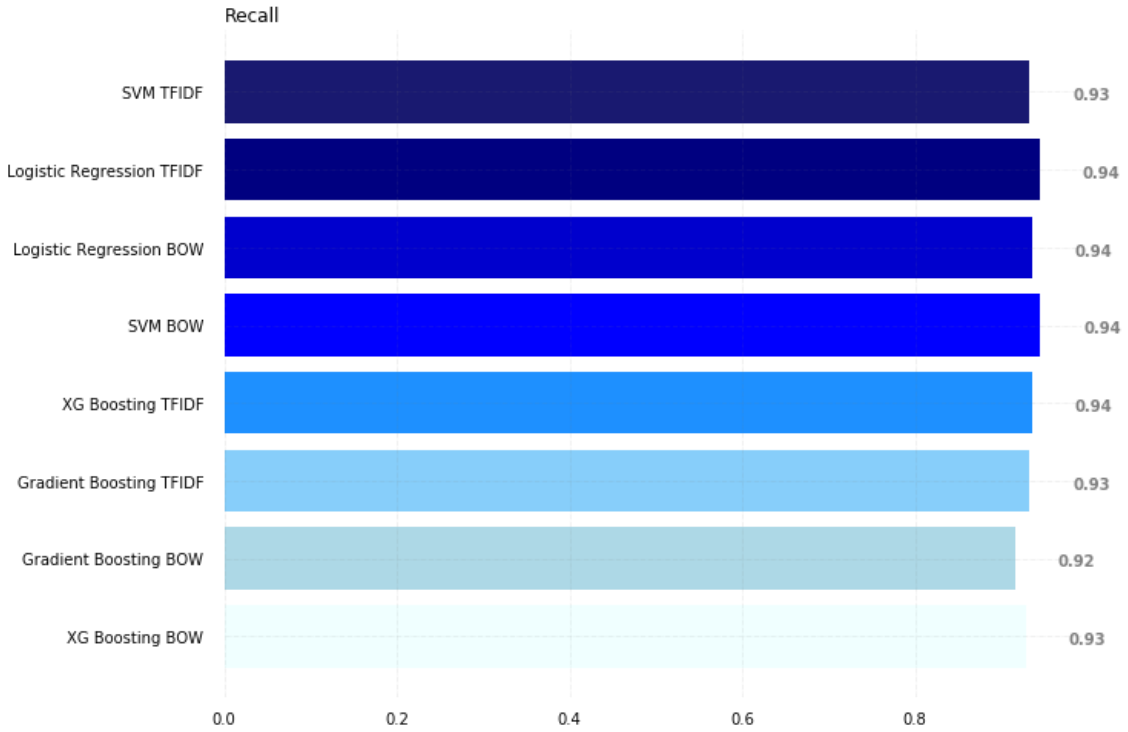
Επιπλέον, κατασκευάστηκαν τα γραφήματα των επιδόσεων των μοντέλων ανά μετρική (βλ. Εικόνες 6.1-6.5).



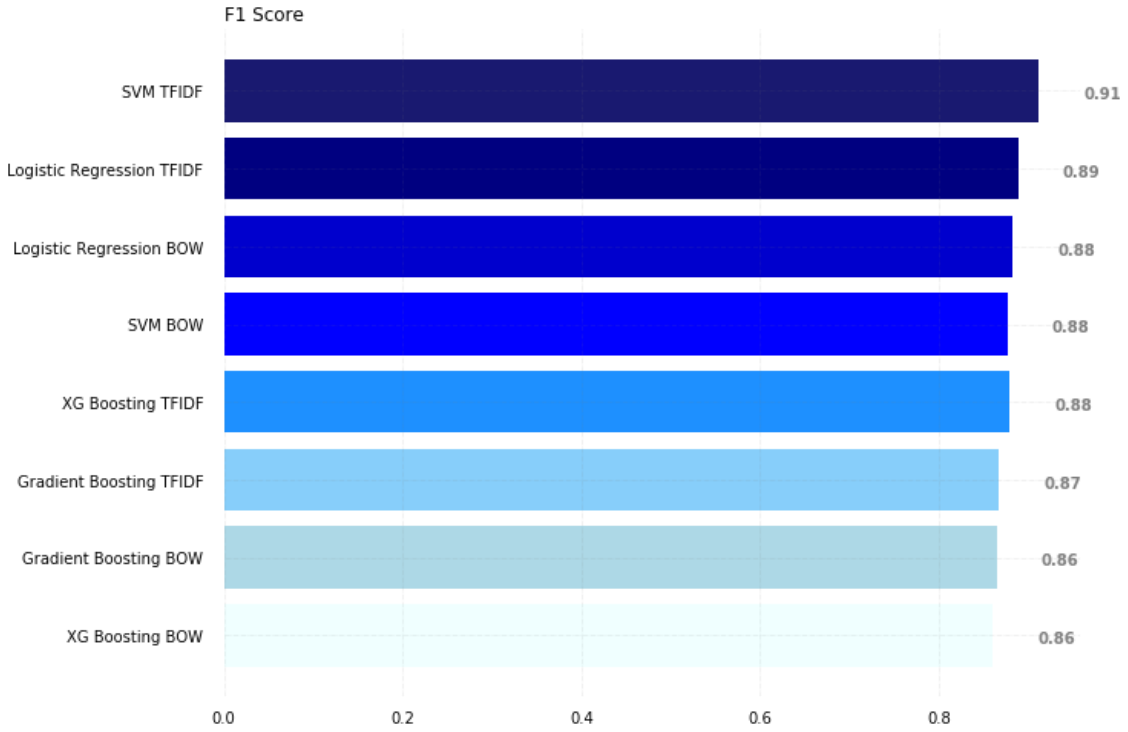
Εικόνα 6.1: Ραβδόγραμμα των επιδόσεων της μετρικής accuracy ανά μοντέλο



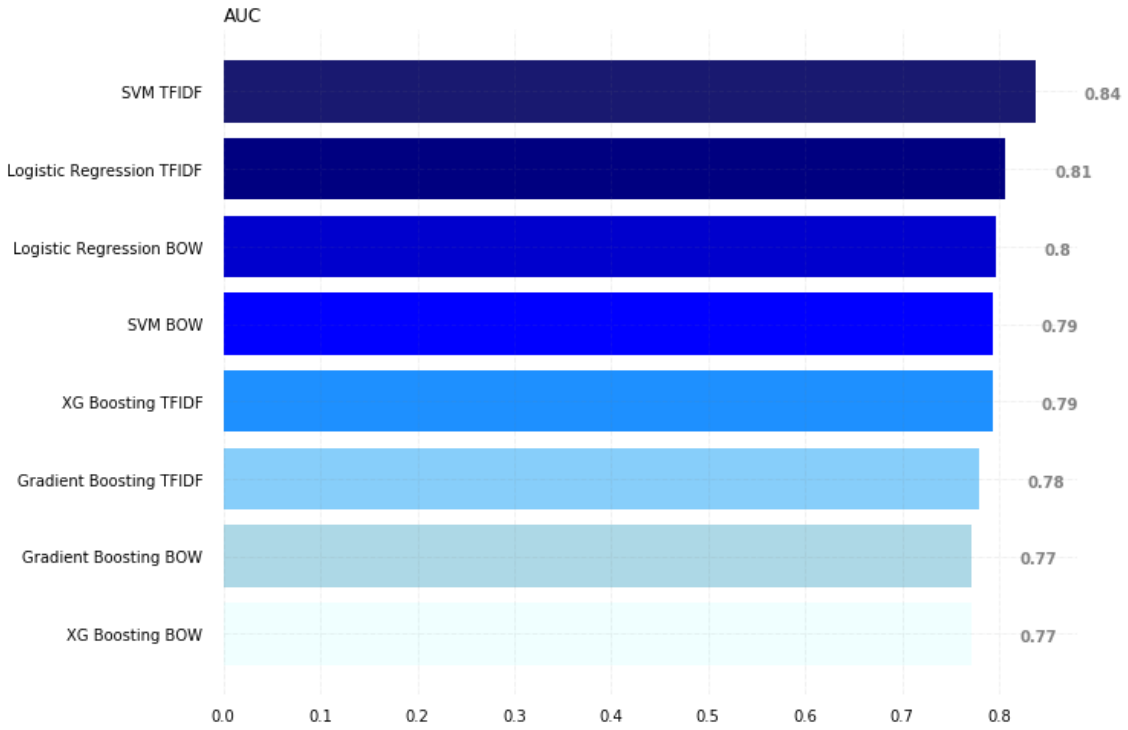
Εικόνα 6.2: Ραβδόγραμμα των επιδόσεων της μετρικής precision ανά μοντέλο



Εικόνα 6.3: Ραβδόγραμμα των επιδόσεων της μετρικής recall ανά μοντέλο



Εικόνα 6.4: Ραβδόγραμμα των επιδόσεων της μετρικής F1 Score ανά μοντέλο



Εικόνα 6.5: Ραβδόγραμμα των επιδόσεων της μετρικής AUC ανά μοντέλο

Συγκρίνοντας τα παραπάνω αποτελέσματα, την καλύτερη επίδοση σε όλες τις μετρικές, εκτός από το Recall, παρουσιάζει το μοντέλο SVM στο οποίο είχε εφαρμοστεί η τεχνική εξαγωγής χαρακτηριστικών TF-IDF. Όπως έχει αναφερθεί στην υποενότητα 4.3.5, το F1 score είναι το καταλληλότερο μέτρο αξιολόγησης των προβλημάτων με ανισόρροπες κλάσεις, συνεπώς αξίζει να αναφερθεί πως το μοντέλο SVM με εξαγωγή χαρακτηριστικών TF-IDF έχει πολύ καλή επίδοση στην συγκεκριμένη μετρική (0.908). Τα δύο επόμενα υψηλότερης απόδοσης μοντέλα είναι τα Logistic Regression με χρήση των CountVectorizer και TF-IDF αντίστοιχα, με το δεύτερο να έχει την υψηλότερη τιμή της μετρικής Recall. Αντίθετα, τα Gradient Boosting μοντέλα και το XG Boosting (Tf-idf) έχουν τις πιο αδύναμες επιδόσεις.

6.2 Ανάλυση των topics της θεματικής κατηγοριοποίησης

Σε αυτήν την υποενότητα θα αναλυθούν τα topics που βρέθηκαν από το LDA μοντέλο με σκοπό να εξαχθούν συμπεράσματα για την γνώμη του καταναλωτικού κοινού για το προϊόν Alexa Echo Auto. Για αυτόν τον λόγο, δίνεται ένας τίτλος για κάθε θέμα, ώστε να γίνουν κατανοητά τα χαρακτηριστικά ή οι λειτουργίες που δεν ικανοποίησαν τους καταναλωτές και μια μικρή επεξήγηση για το καθένα:

Topic 1: "Calling problems"

Σε αυτό το θέμα, οι καταναλωτές αναφέρουν ότι δεν μπορούν να πραγματοποιήσουν μια εξερχόμενη κλήση. Το προϊόν αυτό υπόσχεται ότι μπορεί να συνδεθεί με το κινητό τηλέφωνο και να πραγματοποιεί κλήσεις, οπότε αυτό σημαίνει ότι μπορεί να μη συνδέεται επιτυχώς με όλα τα κινητά τηλέφωνα. Ο ισχυρισμός αυτός έχει βάση αφού στο topic εμφανίζονται λέξεις όπως "mobile","iphone", "smart", "connectivity".

Topic 2: "Bluetooth, sound and voice recognition quality"

Στο topic 2 οι αγοραστές φαίνεται να αναφέρονται στην ποιότητα του ήχου, στην ικανότητα σύνδεσης με το Bluetooth και στην αναγνώριση φωνής. Στο θέμα υπάρχουν πολλές λέξεις που αναφέρονται στην αναγνώριση φωνής και πιο συγκεκριμένα, στην αναγνώριση εντολών.

Topic 3: "Compatibility and support service"

Οι λέξεις που προσδιορίζουν το topic 3 αναφέρονται κυρίως στη συμβατότητα της συσκευής και στην ποιότητα της εξυπηρέτησης των πελατών από το support της εταιρίας.

Topic 4: "Internet and signal connectivity"

Το 4^{ο} θέμα αναφέρεται σχεδόν αποκλειστικά στο πρόβλημα της σύνδεσης του διαδικτύου και του σήματος της συσκευής.

Topic 5: "Journey navigation problems"

Σε αυτό το θέμα, οι αγοραστές αναφέρονται στην ποιότητα της πλοήγησης της και των εντολών πλοήγησης της συσκευής.

Topic 6: "Spotify and radio connection"

Το topic 5 είναι το θέμα που αναφέρεται στη μουσική. Συμφώνα με τους αγοραστές, η εφαρμογή spotify και το ραδιόφωνο φαίνεται να αποσυνδέονται και συνεπώς, να δημιουργούν ενόχληση κατά την οδήγηση.

Topic 7: "Design and microphone"

Στο 7ο θέμα οι επικρατέστερες λέξεις έχουν να κάνουν με την κατασκευή της συσκευής. Πιο συγκεκριμένα, γίνεται αναφορά στο καλώδιο του προϊόντος και στη βάση του. Μια υπόθεση για τη σύνδεση αυτών των δύο λέξεων είναι ότι λόγω κατασκευής της συσκευής, τα δύο αυτά εξαρτήματα δεν είναι πρακτικά. Επίσης, όπως φαίνεται, υπάρχουν παράπονα για τη λειτουργία του μικροφώνου. Αν λάβουμε υπόψιν τις λέξεις, "microphone, "power", "listen", "poor", "hear", μάλλον υπάρχει πρόβλημα στην ποιότητα του ήχου του μικροφώνου.

Topic 8: "Speaker and battery quality"

Τε τελευταίο θέμα αναφέρεται στην ποιότητα του ηχείου της συσκευής και της μπαταρίας, αφού ου λέξεις, "battery", "drain", μας προϊδεάζουν για ύπαρξη προβλήματος στην αντοχή της μπαταρίας.

Σύμφωνα με την παραπάνω ανάλυση των topics, φαίνεται πως τα κύρια και πιο πολυσυζητημένα προβλήματα του δεδομένου προϊόντος είναι η ποιότητα του σήματος, η συμβατότητα του με άλλες συσκευές ή εφαρμογές, η αναγνώριση φωνής, η πλοήγηση και ο ήχος.

6.3 Συμπεράσματα

Η κατηγοριοποίηση των αξιολογήσεων και η θεματική κατηγοριοποίησή τους φαίνεται να παρέχουν πολύτιμες πληροφορίες για την εταιρεία. Καταρχάς, μέσω των αλγορίθμων ταξινόμησης, οι αρνητικές κριτικές μπορούν να διαχωριστούν από τις θετικές, καθιστώντας την ανάλυσή τους πιο πρακτική. Σαφώς, η μη αυτόματη συλλογή

και κατηγοριοποίηση 3.312 αξιολογήσεων είναι δύο πολύ χρονοβόρες διαδικασίες και είναι πρακτικά αδύνατον να εκτελεσθούν από τους υπαλλήλους της εταιρείας. Επομένως, η κατασκευή αλγορίθμων web scraping και μοντέλων ταξινόμησης είναι εξαιρετικά χρήσιμη για την συλλογή και την ταξινόμηση των δεδομένων. Αυτές οι αναλύσεις εξοικονομούν πολύτιμο χρόνο και πόρους για την επιχείρηση. Βέβαια, οι πληροφορίες που παρέχονται στην εταιρεία θα πρέπει να είναι ακριβείς, διαφορετικά δεν θα μπορέσουν να λύσουν το πρόβλημα και να καλύψουν τις ανάγκες της. Στην παρούσα εργασία, το βέλτιστο μοντέλο ταξινόμησης, ύστερα από την ρύθμιση παραμέτρων και την εξαγωγή χαρακτηριστικών, κατάφερε να επιτύχει 88% επιτυχία, γεγονός που καθιστά τα αποτελέσματα ικανοποιητικά ακριβή. Η αναλυτική αξιολόγηση του βέλτιστου μοντέλου γίνεται στην υποενότητα 6.1.

Στη συνέχεια, μέσω της κατηγοριοποίησης των θεμάτων, μπορούν εύκολα να εξαχθούν πληροφορίες σχετικά με τις αδυναμίες των προϊόντων, ώστε η επιχείρηση να μπορεί εύκολα να τις κατανοήσει και να εφαρμόσει κάποιες βελτιώσεις. Χωρίς τη δημιουργία ενός μοντέλου θεματικής κατηγοριοποίησης, είναι σχεδόν αδύνατο να εξαχθούν τα θέματα «συζήτησης» του καταναλωτικού κοινού σχετικά με ένα προϊόν, αφού θα χρειαζόταν πολύς χρόνος και ανθρώπινο δυναμικό για να μελετηθούν όλες οι αξιολογήσεις και στη συνέχεια να εξαχθούν τα topics. Ο βέλτιστος αλγόριθμος που κατασκευάστηκε στην αυτήν την μελέτη, επιτυγχάνει καλό διαχωρισμό των θεμάτων, γεγονός που τα καθιστά μοναδικά και ευανάγνωστα. Αυτές οι πληροφορίες είναι πολύ σημαντικές για την εταιρία, αφού βοηθούν στην κατασκευή ενός προϊόντος υψηλότερης ποιότητας και δημοφιλίας. Επιπλέον, οδηγούν στην ικανοποίηση των πελατών και συνεπώς, στην αύξηση των εσόδων της.
Τα ευρήματα αυτής της μελέτης καταδεικνύουν την συμβολή του Data Science στο ηλεκτρονικό εμπόριο. Οι εταιρίες που χρησιμοποιούν αυτές τις μεθόδους για να αναπτύξουν λειτουργικά προϊόντα και να καλύψουν τις ανάγκες των πελατών τους, έχουν σημαντικό πλεονέκτημα έναντι άλλων ανταγωνιστικών εταιριών, αφού έχουν σαν επίκεντρο τον πελάτη και επιτυγχάνουν όχι μονό την αύξηση των κερδών τους αλλά και την ανάπτυξη μιας σχέσης εμπιστοσύνης με το καταναλωτικό κοινό.

Ως επέκταση αυτής της μελέτης, οι κριτικές θα μπορούσαν να κατηγοριοποιηθούν σε τρεις κατηγορίες συναισθήματος: «θετικό», «αρνητικό» και «ουδέτερο», ώστε να αποδοθούν πιο συγκεκριμένες κλάσεις και να παραχθούν πιο «καθαρά» topics με τη

μέθοδο της θεματικής κατηγοριοποίησης. Επιπλέον, θα είχε πολύ ενδιαφέρον να γίνει ο καθαρισμός των δεδομένων με τέτοιον τρόπο, έτσι ώστε να είναι δυνατή η ανάλυση των συναισθημάτων των κριτικών με χρήση των emoticons (π.χ. 😊, ☹).

ΒΙΒΛΙΟΓΡΑΦΙΑ

Byeongki Jeonga, Janghyeok Yoona, Jae-Min Leeb. (2019) Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis, International Journal of Information Management (48), p.p 280-290.

Heidi Nguyen, Aravind Veluchamy, Mamadou Diop et al. (2018) Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches, SMU Data Science Review: Vol. 1: No. 4, Article 7.

Atitaya Yakaew, Matthew N. Dailey, Teeradaj Racharak. (2021) Multimodal Sentiment Analysis on Video Streams using Lightweight Deep Neural Networks, 10th International Conference on Pattern Recognition Applications and Methods.

Ronen Feldman. (2013) Techniques and applications for sentiment analysis, Communications of the ACM, Volume 56, Issue 4, p.p. 82-89.

Maite Taboada, Julian Brooke, Milan Tofiloski et al. (2011), Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics, 37 (2): 267–307.

SAS. (2022) Machine Learning. What it is and why it matters. [Online] Available from: https://www.sas.com/en_us/insights/analytics/machine-learning [Accessed: 22 June 2022].

Shahid Shayaa, Noor Ismawati Jaafar, Shamshul Bahri et al. (2018) Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges, IEEE Access, Volume 6, p.p. 37807-37827.

Enes Zvornicanin. (2022) Topic Modeling and Latent Dirichlet Allocation (LDA). [Online] Available from: https://datascienceplus.com/topic-modeling-and-latent-dirichlet-allocation-lda/ [Accessed: 27 June 2022].

Joyce Xu. (2018) Topic Modeling with LSA, PLSA, LDA & lda2Vec. [Online] Available from: https://medium.com/nanonets/topic-modeling-with-lsa-psla-lda-and-lda2vec-555ff65b0b05 [Accessed: 27 June 2022].

Chirag Goyal. (2021) Part 16 : Step by Step Guide to Master NLP – Topic Modelling using LSA. [Online] Available from: https://www.analyticsvidhya.com/blog/2021/06/part-16-step-by-step-guide-to-master-nlp-topic-modelling-using-lsa/ [Accessed: 28 June 2022].

David M. Blei, Andrew Y. Ng, Michael I. Jordan. (2003) Latent Dirichlet Allocation, Journal of Machine Learning Research 3, p.p. 993-1022.

James Petterson , Alex Smola , Tiberio Caetano et al. (2010) Word Features for Latent Dirichlet Allocation, Advances in Neural Information Processing Systems 23 (NIPS 2010).

Ben Lutkevich. (2021) Natural language processing (NLP). [Online] Available from: https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP [Accessed: 1 July 2022].

Monkeylearn (2022) Text Classification: What it is And Why it Matters. https://monkeylearn.com/text-classification/ [Accessed: 1 July 2022].

Maheshwar Nareddy,Dr. Goutam Chakraborty. (2012) Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 1st ed., Oxford: Elsevier.

Engati (2021) Bag of words Model. [Online] Available from: https://www.engati.com/glossary/bag-of-words [Accessed: 2 July 2022].

Δημήτρης Ε. Πιζάνιας. (2018) Εξόρυξη άποψης και ανάλυση συναισθήματος σε μέσα κοινωνικής δικτύωσης, Διπλωματική εργασία, Τμήμα Στατιστικής και Αφαλιστικής Επιστήμης.

Dan Jurafsky, James H. Martin. (2009) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 3rd ed., London: Prentice Hall.

Jacob Eisenstein. (2019) Introduction to Natural Language Processing, Illustrated ed., London: MIT Press.

Mohit Kumar Barai. (2021) Sentiment Analysis with TextBlob and Vader. https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/ [Accessed: 9 July 2022].

Himanshu Sharma. (2020) Let's Learn TextBlob Quickstart – A Python Library For Processing Textual Data. https://analyticsindiamag.com/lets-learn-textblob-quickstart-a-python-library-for-processing-textual-data/ [Accessed: 9 July 2022].

Geeksforgeeks. (2021) https://www.geeksforgeeks.org/python-sentiment-analysis-using-vader/ [Accessed: 10 July 2022].

Arni S.R. Srinivasa Rao, C.R. Rao (2021) Data Science: Theory and Applications, Handbook of Statistics, Volume 44, p.p. 2-331.

Sidharth Sekhar. (2019) Math Behind Logistic Regression Algorithm. https://medium.com/analytics-vidhya/logistic-regression-b35d2801a29c [Accessed: 12 July 2022].

MonkeyLearn. (2017) Support Vector Machines (SVM) Algorithm Explained. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/ [Accessed: 14 July 2022].

Arshad Kazi (2020) Mathematics behind SVM (Support Vector Machine). https://www.arshad-kazi.com/mathematics-behind-svmsupport-vector-machine/ [Accessed: 14 July 2022].

Weizhang Liang, Suizhi Luo, Guoyan Zhao et al. (2020) Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms, Mathematics 2020, Volume 8, Issue 5, 765.

Harsha Vardhan Garlapati. (2021) Machine Learning Model Evaluation. https://www.knowledgehut.com/blog/data-science/machine-learning-model-evaluation [Accessed: 21 July 2022].

Neha Wadhawan. (2019) Machine Learning Model Evaluation Methods: which one to use? https://medium.com/@wadhawan.neha06/machine-learning-model-evaluation-methods-which-one-to-use-f659cd20d759 [Accessed: 21 July 2022].

Jeremy Jordan. (2017) Evaluating a machine learning model. https://www.jeremyjordan.me/evaluating-a-machine-learning-model/ [Accessed: 21 July 2022].

Sarang Narkhede. (2018) Understanding AUC - ROC Curve. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [Accessed: 22 July 2022].

Mika V. Mantyla, Daniel Graziotin, Miikka Kuutila. (2018) The evolution of sentiment analysis—A review of research topics, venues, and top cited papers, Computer Science Review, Volume 27, p.p. 16-32.

Mozilla. (2022) HTML elements reference. https://developer.mozilla.org/en-US/docs/Web/HTML/Element [Accessed: 25 July].

Geeksforgeeks. (2022) Removing stop words with NLTK in Python. https://www.geeksforgeeks.org/removing-stop-words-nltk-python/ [Accessed: 27 July].

Emidio Amadebai. (2022) The 5 Methods of Collecting Data Explained. https://www.analyticsfordecisions.com/methods-of-collecting-data/ [Accessed: 14 August].

Formplus. (2022) What is Primary Data? + [Examples & Collection Methods]. https://www.formpl.us/blog/primary-data [Accessed: 16 August].

Mike Olson. (2014) What are the Differences Between Books, Magazines and Electronic Media? https://content.spendit.com/2014/01/what-are-the-differences-between-books-magazines-and-electronic-media/ [Accessed: 18 August].

Formplus. (2022) What is Secondary Data? + [Examples, Sources, & Analysis]. https://www.formpl.us/blog/secondary-data [Accessed: 18 August].

Bartosz Szabłowski. (2020) Complete Data Science project: Business Understanding. https://towardsdatascience.com/complete-data-science-project-part-1-business-understanding-b8456bb14bd4 [Accessed: 22 August].

Foster Provost, Tom Fawcett. (2013) Data Science for Business, 1st ed., O'Reilly Media Publication, Inc.

Saikat Mazumder. (2021) 5 Techniques to Handle Imbalanced Data For a Classification Problem. https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/ [Accessed: 25 August].

Microsoft. (2021) SMOTE. https://docs.microsoft.com/en-us/azure/machine-learning/component-reference/smote [Accessed: 25 August].

Heavy. (2022) Data Exploration - A Complete Introduction. https://www.heavy.ai/learn/data-exploration [Accessed: 26 August].

IBM Cloud Education. (2020) Exploratory Data Analysis. https://www.ibm.com/cloud/learn/exploratory-data-analysis#toc-types-of-e-64hsTW2A [Accessed: 26 August].

David Weedmark, (2021) A Guide to Machine Learning Model Deployment. https://www.dominodatalab.com/blog/machine-learning-model-deployment [Accessed: 27 August].

Geeksforgeeks. (2020) Optimization for Data Science. https://www.geeksforgeeks.org/optimization-for-data-science/ [Accessed: 27 August].

K. Gwirtz, M. Morzfeld, A. Fourniero et al. (2020) Can one use Earth's magnetic axial dipole field intensity to predict reversals?, Geophysical Journal International, Volume 225, Issue 1, p.p. 277–297.

Marouane Birjali, Mohammed Kasri, Abderrahim Beni-Hssane. (2021) A comprehensive survey on sentiment analysis: Approaches, challenges and trends, Knowledge-Based Systems, Volume 226, p.p. 107-134.

Zulaikha Lateef. (2022) 5 Data Science Projects – Data Science Projects For Practice. https://www.edureka.co/blog/data-science-projects/ [Accessed: 19 August].

Priya Singh. (2022) Data Collection: Important Methods. https://www.embibe.com/exams/data-collection/ [Accessed: 18 August].

Wallstreetmojo Editorial Team. (2022) What is Gradient Boosting? https://www.wallstreetmojo.com/gradient-boosting/ [Accessed: 11 September].

Gaurav. (2021) An Introduction to Gradient Boosting Decision Trees. https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/ [Accessed: 11 September].

Vihar Kurama. (2020) Gradient Boosting In Classification: Not a Black Box Anymore! https://blog.paperspace.com/gradient-boosting-for-classification/ [Accessed: 11 September].

Akira.AI (2020) Gradient Boosting. http://www.akira.ai/glossary/gradient-boosting [Accessed: 11 September].

Jason Brownlee. (2016) A Gentle Introduction to XGBoost for Applied Machine Learning. https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/ [Accessed: 11 September].

Priya Pedamkar. (2022) What is XGBoost Algorithm? https://www.educba.com/xgboost-algorithm/ [Accessed: 11 September].

Ζαραφέτα Κυριακή Ηλέκτρα. (2019) Υλοποίηση των μοντέλων LDA και Word2Vec και σύγκριση των εργαλείων TMG και Text Analytics, Διπλωματική εργασία, Πανεπιστήμιο Πατρών.

Sonia Jessica. (2022) How Does Logistic Regression Work? https://www.kdnuggets.com/2022/07/logistic-regression-work.html [Accessed: 15 September].

Ayush Pant. (2019) Introduction to Logistic Regression, https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148 [Accessed: 15 September].

Rohith Gandhi. (2018) Support Vector Machine — Introduction to Machine Learning Algorithms. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47 [Accessed: 16 September].

Noel Bambrick. (2016) Support Vector Machines: A Simple Explanation. https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html [Accessed: 16 September].

Geeksforgeeks. (2022) ML | Underfitting and Overfitting. https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/ [Accessed: 16 September].

Gaurav Rajesh Sahani. (2020) Elucidating Bias, Variance, Under-fitting, and Over-fitting. https://medium.com/analytics-vidhya/elucidating-bias-variance-under-fitting-and-over-fitting-273846621622 [Accessed: 16 September].

Mathworks. (2017) Latent Dirichlet allocation (LDA) model. https://www.mathworks.com/help/textanalytics/ref/ldamodel.html [Accessed: 17 September].

Ben Lutkevich, Wesley Chai, Brian Holak. (2022) E-commerce. https://www.techtarget.com/searchcio/definition/e-commerce [Accessed: 1 October].

Matthew Woodward. (2022) Ecommerce Statistics 2022 – Everything You Need To Know. https://www.matthewwoodward.co.uk/seo/ecommerce-statistics/ [Accessed: 1 October].

Think-plus. (2020) Τι είναι το ecommerce (ηλεκτρονικό εμπόριο); https://think-plus.gr/ti-einai-to-ecommerce/ [Accessed: 1 October].

FortuneGreece. (2022) Ηλεκτρονικό εμπόριο και πανδημία: Τεράστια η ανάπτυξη του 2021, μεγαλύτεροι όμως οι κίνδυνοι κυβερνο-απατεώνων. https://www.fortunegreece.com/article/ilektroniko-emporio-ke-pandimia-terastia-i-anaptixi-tou-2021-megaliteri-omos-i-kindini-kiverno-apateonon/ [Accessed: 1 October].

April Berthene. (2022) Coronavirus pandemic adds $219 billion to US ecommerce sales in 2020-2021. https://www.digitalcommerce360.com/article/coronavirus-impact-online-retail/ [Accessed: 1 October].

Meghna Pant. (2019) 9 Interesting Applications of Data Science in the E-commerce industry. https://datascience.foundation/datatalk/9-interesting-applications-of-data-science-in-the-e-commerce-industry [Accessed: 2 October].

Great Learning. (2020) How Applying Data Science in E-Commerce Will Boost Online Sales. https://medium.com/@mygreatlearning/how-applying-data-science-in-e-commerce-will-boost-online-sales-ac42239afa91 [Accessed: 3 October].

JavaPoint. (2022) Machine Learning Tutorial. https://www.javatpoint.com/machine-learning [Accessed: 29 October].

SydneyF. (2019) Holdouts and Cross Validation: Why the Data Used to Evaluate your Model Matters. https://community.alteryx.com/t5/Data-Science/Holdouts-and-Cross-Validation-Why-the-Data-Used-to-Evaluate-your/ba-p/448982 [Accessed: 29 October].

Jason Brownlee. (2018) A Gentle Introduction to k-fold Cross-Validation. https://machinelearningmastery.com/k-fold-cross-validation/ [Accessed: 29 October].