ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Διπλωματική Εργασία

Ανάλυση και Πρόβλεψη Συναισθήματος από Διαδικτυακές Κριτικές Πελατών

Φοιτητής: Παπαδόπουλος Ευστάθιος

ΑΜ: 711141048(cs141048)

Επιβλέπων Καθηγητής

Αλέξανδρος Μπουσδέκης

Συν-επιβλέπων Καθηγητής

Γεώργιος Μιαούλης

ΑΘΗΝΑ-ΑΙΓΑΛΕΩ, Μάρτιος/2022

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάλυση και Πρόβλεψη Συναισθήματος από Διαδικτυακές Κριτικές Πελατών

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
1	ΓΕΩΡΓΙΟΣ ΜΙΑΟΥΛΗΣ	Ομότιμος Καθηγητής	Georgios Miaoulis Digitally signed by Georgios Miaoulis DN: c=GR, l=Athens, o=University of West Attica, ou=Class B - Private Key created and stored in software CSP, sn=Miaoulis, givenName=Georgios, serialNumber=3179983822, cn=Georgios Miaoulis, email=gmiaoul@uniwa.gr Date: 2022.03.09 12:53:56 +02'00'
2	ΠΑΡΙΣ ΜΑΣΤΟΡΟΚΩΣΤΑΣ	Καθηγητής	PARIS MASTOROKOSTAS Digitally signed by PARIS MASTOROKOSTAS Date: 2022.03.09 09:21:02 +02'00'
3	ΓΕΩΡΓΙΟΣ ΜΠΑΡΔΗΣ	Επίκουρος Καθηγητής	Georgios Bardis Digitally signed by Georgios Bardis Date: 2022.03.09 17:05:01 +02'00'

Δήλωση Συγγραφέα Διπλωματικής Εργασίας

Ο κάτωθι υπογεγραμμένος Παπαδόπουλος Ευστάθιος του Κωνσταντίνου, με αριθμό μητρώου 141048 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

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

Ο Δηλών

Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω την οικογένεια μου που με στήριξε όλα αυτά τα χρόνια της φοίτησής μου, καθώς και τον επιβλέπων καθηγητή μου για την καθοδήγηση που μου παρείχε καθ' όλη την διάρκεια της εκπόνησης της διπλωματικής εργασίας.

Περίληψη

Η διπλωματική αφορά την ανάλυση και πρόβλεψη συναισθήματος σε διαδικτυακές κριτικές πελατών με τη χρήση μεθόδων αναλυτικής δεδομένων και μηχανικής μάθησης. Η προσέγγιση της διπλωματικής επεξεργάζεται το κείμενο των σχολίων, προβλέπει την γνώμη των επόμενων πελατών και εξετάζει την επίδραση διαφόρων κριτηρίων στην συνολική γνώμη του πελάτη. Η προσέγγιση επικυρώνεται στον τομέα του τουρισμού, χρησιμοποιώντας ένα dataset που προέρχεται από το TripAdvisor. Τα κριτήρια αφορούν διαφορετικές υπηρεσίες και διαδικασίες των ξενοδοχείων. Η γλώσσα υλοποίησης της διπλωματικής είναι η Python.

Λέξεις κλειδία: Μηχανική μάθηση, Ανάλυση Συναισθήματος, Προβλεπτική ανάλυση, Online Κριτικές Ξενοδοχείων, Υπηρεσίες Ξενοδοχείων

Abstract

The thesis is about analyzing and predicting sentiment in online customer reviews using data analytics and machine learning methods. The thesis' approach processes the text of the comments, predicts the opinion of subsequent customers and examines the impact of various criteria on the overall customer opinion. The approach is validated in the tourism domain using a dataset derived from TripAdvisor. The criteria relate to different hotel services and processes. The implementation language of the thesis is Python.

Keywords: Machine learning, Sentiment analysis, Predictive analytics, Online Hotel Reviews, Hotel Services

ΚΑΤΑΛΟΓΟΣ ΕΙΚΟΝΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ

1.1 Γενικά

Οι κριτικές επισκεπτών είναι ένας σημαντικός παράγοντας που επηρεάζει τις κρατήσεις/αγορές των ανθρώπων. Οι περισσότεροι άνθρωποι όταν αναζητούν ένα μέρος για διακοπές στο Expedia/Booking/TripAdvisor, τι κάνουν; Πηγαίνουν συνήθως να ελέγξουν τις κριτικές.

Οι κριτικές των επισκεπτών έχουν σημαντικό αντίκτυπο σε μία επιχείρηση. Με άλλα λόγια, επηρεάζουν σαφώς την απόφαση των ανθρώπων για κράτηση, πράγμα που σημαίνει ότι, πρέπει να δίνεται ιδιαίτερη προσοχή στο τι λένε οι άνθρωποι για κάποιο ξενοδοχείο.

Είναι απαραίτητο εκτός από τις καλές κριτικές, να δίνεται έμφαση και στις αρνητικές, διότι αυτές μπορούν να αξιοποιηθούν με τρόπο που μπορεί να βοηθήσει να μάθει μια επιχείρηση τα μέγιστα για τους πελάτες της. Επίσης μπορούν να πουν αν μία επιχείρηση συμβαδίζει με τις προσδοκίες των πελατών της, κάτι που είναι ζωτικής σημασίας για την ανάπτυξη στρατηγικών μάρκετινγκ με βάση τις προσωποποιήσεις των πελατών.

Οι κριτικές είναι σημαντικές και, πρέπει να αρχίσουν να αξιοποιούνται.

1.2 Αντικείμενο και στόχοι της Διπλωματικής

Η εκπόνηση της συγκεκριμένης μελέτης αποσκοπεί στην ανάλυση της γνώμης των πελατών ξενοδοχείων για διάφορες υπηρεσίες των ξενοδοχείων. Τα αποτελέσματα της ανάλυσης μπορούν να βοηθήσουν τα ξενοδοχεία να βελτιώσουν τις υπηρεσίες και τις διαδικασίες τους. Η διπλωματική έχει 2 βασικά βήματα: (1) επεξεργασία των online reviews, τα οποία είναι κείμενο και (2) ανάλυση της επίδρασης της γνώμης στο συνολικό review rating που δίνει ο πελάτης σε ένα ξενοδοχείο. Το dataset είναι πραγματικό και προέρχεται από το TripAdvisor.

1.3 Διάρθρωση της Διπλωματικής Εργασίας.

Η παρούσα Διπλωματική Εργασία θα αναπτυχθεί σε 4 συνολικά κεφάλαια. Η διάρθρωση της εργασίας είναι η εξής:

1. Στο Κεφάλαιο 2, δίνεται ο ορισμός της ανάλυσης συναισθήματος(2.1), αναλύεται το θεωρητικό της υπόβαθρο(2.1.1), και παρουσιάζονται αλγόριθμοι ανάλυσης συναισθήματος(2.1.2). Στην συνέχεια δίνεται ο ορισμός της προβλεπτικής αναλυτικής(2.2), αναλύεται και εδώ το θεωρητικό της υπόβαθρο(2.2.1), περιγράφονται μέθοδοι προβλεπτικής ανάλυσης και το κεφάλαιο κλείνει με το τί ρόλο παίζει η προβλεπτική ανάλυση όταν την χρησιμοποιούμε σε Online κριτικές.
2. Στο Κεφάλαιο 3, πραγματοποιείται μία επισκόπηση της προτεινόμενης προσέγγισης(3.1), επεξηγούνται το σύνολο των δεδομένων καθώς και η προ-επεξεργασία τους(3.2), και κατόπιν ορίζονται οι κατηγορίες προς αξιολόγηση(3.3). Στην συνέχεια παρουσιάζεται ο τρόπος με τον οποίο θα βρεθούν οι κατηγορίες σε διαδικτυακά σχόλια(3.4), και παρουσιάζεται ο αλγόριθμος συσταδοποίησης K-means(3.5). Το κεφάλαιο κλείνει με την περιγραφή μίας μεθόδου Ensemble και οι αλγόριθμοι ταξινόμησης από την οποία θα απαρτίζεται.
3. Το Κεφάλαιο 4, πραγματεύεται τη σημασία της Ανάλυσης Συναισθήματος στις Ξενοδοχειακές επιχειρήσεις(4.1), καθώς και το σύνολο των δεδομένων που θα χρησιμοποιηθεί(4.2). Μετέπειτα πραγματοποιείται λεπτομερής περιγραφή της υλοποίησης που προτάθηκε παραπάνω(4.3) και αξιολόγησή της(4.4). Το κεφάλαιο κλείνει με την σύγκριση των μέχρι τώρα αποτελεσμάτων(4.5).
4. Τέλος, στο Κεφάλαιο 5 περιγράφονται τα συμπεράσματα που προκύπτουν από τη μελέτη καθώς και τα αποτελέσματα της σύγκρισης από τη χρήση αυτών των μεθόδων.

Κεφάλαιο 2

Βιβλιογραφική Επισκόπηση

2.1 Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος είναι η διαδικασία προσδιορισμού του κατά πόσον ένα κείμενο είναι θετικό, αρνητικό ή ουδέτερο. Ένα σύστημα ανάλυσης συναισθήματος για την ανάλυση κειμένου συνδυάζει τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) και μηχανικής μάθησης για να αποδίδει σταθμισμένες βαθμολογίες συναισθήματος στις οντότητες, τα θέματα και τις κατηγορίες μέσα σε μια πρόταση ή φράση.

Η ανάλυση συναισθήματος βοηθά τους αναλυτές δεδομένων των μεγάλων επιχειρήσεων να μετρήσουν την κοινή γνώμη, να παρακολουθούν τη φήμη των εμπορικών σημάτων και των προϊόντων και να κατανοούν τις εμπειρίες των πελατών. Επιπλέον, οι εταιρείες ανάλυσης δεδομένων συχνά ενσωματώνουν API ανάλυση συναισθήματος τρίτων στη δική τους πλατφόρμα διαχείρισης εμπειρίας πελατών, παρακολούθησης κοινωνικών μέσων ή ανάλυσης εργατικού δυναμικού, προκειμένου να παρέχουν χρήσιμες πληροφορίες στους δικούς τους πελάτες.

Το κεφάλαιο εξετάζει πώς λειτουργεί η βασική ανάλυση συναισθήματος, τα πλεονεκτήματα και τα μειονεκτήματα της ανάλυσης συναισθήματος και το ρόλο της μηχανικής μάθησης στην ανάλυση συναισθήματος. Τέλος, θα δούμε μερικούς αλγόριθμους ανάλυσης συναισθήματος.

2.1.1 Θεωρητικό Υπόβαθρο

Πώς λειτουργεί η ανάλυση συναισθήματος:

Η ανάλυση συναισθήματος λειτουργεί με το σπάσιμο ενός μηνύματος σε θεματικές ενότητες και στη συνέχεια με την ανάθεση μιας βαθμολογίας συναισθήματος σε κάθε θέμα.

Για παράδειγμα ας δούμε την ακόλουθη κοινωνική ανάρτηση:

Δοκίμασα τον νέο υπολογιστή Dell G5 Gaming. Εντυπωσιάστηκα πραγματικά. Η κάρτα γραφικών ήταν λίγο απογοητευτική, αλλά είναι δύσκολο να νικήσεις τον G5 σε αυτή την τιμή.

Ένα εργαλείο ανάλυσης συναισθήματος θα το χώριζε αυτό σε θεματικές ενότητες και στη συνέχεια θα απέδιδε μια βαθμολογία συναισθήματος σε κάθε θέμα, ανάλογα με μια προκαθορισμένη κλίμακα:

- Dell 5 Gaming. Εντυπωσιάστηκα πραγματικά = +4
- Η κάρτα γραφικών…απογοητευτική = -2
- Δύσκολο να νικήσεις…την τιμή = +3

Στη συνέχεια, το ρομπότ θα αθροίσει τις βαθμολογίες ή θα χρησιμοποιήσει κάθε βαθμολογία ξεχωριστά για να αξιολογήσει τα στοιχεία της δήλωσης. Σε αυτή την περίπτωση, υπήρχε ένα συνολικά θετικό συναίσθημα, αλλά ένα αρνητικό συναίσθημα για την κάρτα γραφικών.

Ωστόσο, η ανάλυση συναισθήματος δεν είναι εύκολη! Σκεφτείτε την ακόλουθη πρόταση:

Είμαι ΤΟΣΟ χαρούμενος που η πτήση μου καθυστέρησε.

Ενώ οι περισσότεροι από εμάς μπορούμε να το αντιληφθούμε αυτό ως σαρκασμό, μπορεί να είναι δύσκολο για ένα ρομπότ. Ωστόσο, η ανάλυση συναισθήματος βελτιώνει διαρκώς τις δυνατότητές της με την πάροδο του χρόνου χάρη στην πρόοδο της τεχνητής νοημοσύνης και τη συμβολή των ανθρώπινων κριτών.[1]

Πλεονεκτήματα της ανάλυσης συναισθήματος:

1. Βελτίωση της εξυπηρέτησης πελατών

Ένα από τα οφέλη της ανάλυσης συναισθήματος είναι η δυνατότητα εντοπισμού των βασικών μηνυμάτων από τις απόψεις και τις σκέψεις των πελατών για μια μάρκα.

Αυτό βοηθά το τμήμα εξυπηρέτησης πελατών να γνωρίζει τυχόν σχετικά ζητήματα ή προβλήματα.

Καθώς η μέθοδος επιτρέπει στους οργανισμούς να κατανοήσουν καλύτερα τους πελάτες τους, η ανάλυση συναισθήματος παρέχει μια σαφή εικόνα των προβλημάτων και πείθει τον οργανισμό να αναζητήσει μια λύση.

Επιπλέον, έχοντας μια γρήγορη ανίχνευση της ανάλυσης συναισθήματος πάνω στα δυσμενή σχόλια των πελατών, ο οργανισμός μπορεί να δράσει γρήγορα διερευνώντας τη βασική αιτία και παρέχοντας στο τμήμα εξυπηρέτησης πελατών μια αποτελεσματική λύση.

Τίποτα δεν είναι καλύτερο από μια άμεση ανταπόκριση στην αντιμετώπιση ενός ζητήματος από την ίδια την εταιρεία.

2. Ανάπτυξη ποιοτικών προϊόντων:

Το να κάνετε τους πελάτες ευτυχισμένους και να παραμείνουν πιστοί σε μια μάρκα είναι μια επίπονη δουλειά. Ως εκ τούτου, ένα ακόμα από τα οφέλη της ανάλυσης συναισθήματος είναι ότι καθιστά την όλη διαδικασία ευκολότερη και ταυτόχρονα παρέχει ευκαιρίες βελτίωσης.

Αυτό επιτρέπει στην ομάδα μάρκετινγκ να ερευνά καλύτερα τις τρέχουσες τάσεις και τις προτιμήσεις των πελατών.

Οι απαντήσεις των πελατών μπορούν να χρησιμοποιηθούν ως κατευθυντήριες γραμμές για τη βελτίωση της ποιότητας των υπηρεσιών, την καλύτερη μελλοντική ανάπτυξη προϊόντων, τη μείωση της απομάκρυνσης πελατών ή τη βελτίωση του τρόπου παρουσίασης του προϊόντος.

‘Όποιος καταλαβαίνει καλύτερα τον πελάτη, κερδίζει’ Mike Gospe

3. Ανακάλυψη νέων στρατηγικών μάρκετινγκ:

Με περισσότερα δεδομένα και πληροφορίες που συλλέγονται μέσω της ανάλυσης συναισθήματος, οι οργανισμοί θα μπορούσαν να αναπτύξουν μια αποτελεσματική στρατηγική μάρκετινγκ.

Το αποτέλεσμα από τις στρατηγικές μπορεί να μετρηθεί από τα θετικά ή αρνητικά βασικά μηνύματα των πελατών.

Παρατηρώντας τις συζητήσεις των πελατών στα μέσα κοινωνικής δικτύωσης και εντοπίζοντας τα συγκεκριμένα μηνύματα-κλειδιά που σχετίζονται με το εμπορικό σήμα, μπορούν να σχεδιαστούν συγκεκριμένες εκστρατείες μάρκετινγκ για τους καταναλωτές-στόχους.

4. Βελτίωση της αντίληψης των μέσων ενημέρωσης:

Ένα άλλο πλεονέκτημα της ανάλυσης συναισθήματος είναι η δυνατότητα παρακολούθησης της αντίληψης των δημοσιογράφων, των συγγραφέων, των αρθρογράφων, των αναλυτών αγοράς, των ερευνητών των μέσων ενημέρωσης ή των ανεξάρτητων συνεργατών για την εταιρεία, είτε πρόκειται για το προϊόν, την υπηρεσία, τις αξίες της εταιρείας, το ανθρώπινο δυναμικό κ.λπ.

Αυτό είναι ζωτικής σημασίας, καθώς οποιαδήποτε παρερμηνεία ή αρνητική χροιά μπορεί να οδηγήσει σε αρνητικά βασικά μηνύματα που διαμορφώνουν μια ανεπιθύμητη αντίληψη.

Η γνώση του ποιος γράφει τι ιστορικά και ποιο είναι το ενδιαφέρον του και πόσο επικριτικός είναι σε ορισμένα θέματα βοηθά το τμήμα σχέσεων με τα μέσα ενημέρωσης να συσκευάσει ένα κατάλληλο και ελκυστικό περιεχόμενο για αυτούς.[2]

5. Ανάλυση μεγάλων δεδομένων:

Η ανάλυση συναισθήματος μπορεί να καταστήσει δυνατή την επεξεργασία τεράστιων ποσοτήτων δεδομένων σε πραγματικό χρόνο. Επίσης, όταν επεξεργάζεστε και αναλύετε μεγαλύτερο όγκο δεδομένων, τόσο πιο αξιόπιστες, χρήσιμες και έγκαιρες είναι οι πληροφορίες. Επίσης μπορούμε να αντλήσουμε πληροφορίες σχετικά με το πόσο ευχαριστημένοι είναι οι επισκέπτες από την εμπειρία του ξενοδοχείου σας.

6. Βελτίωση της απόδοσης του προσωπικού:

Τα ξενοδοχεία πρέπει να κάνουν ήδη πολλά για να βελτιώσουν την απόδοση του προσωπικού, αλλά η διαδικασία βελτίωσης θα μπορούσε να γίνει απλά καλύτερη και αποτελεσματικότερη αν τα ξενοδοχεία αρχίσουν να αναγνωρίζουν πού ακριβώς το προσωπικό αποδίδει αποτελεσματικά και πού όχι.

Για παράδειγμα, οι επισκέπτες γράφουν κριτικές στο διαδίκτυο για κάποιο από το προσωπικό που δεν προσέφερε μια εξαιρετική εμπειρία. Μπορούμε να αναλύσουμε τις κριτικές και να καλύψουμε το κενό βελτιώνοντας τους πιο απαραίτητους τομείς. Επίσης, ένα άλλο παράδειγμα θα ήταν - ένας επισκέπτης γράφει μια κριτική όπου αναφέρει την εξαιρετική εξυπηρέτηση που παρείχε το προσωπικό. Δεν θα ήταν υπέροχο να μοιραστόμασταν αυτό το σχόλιο με την ομάδα μας και να επαινούσαμε το προσωπικό; Προφανώς, ναι! Αυτό θα κρατήσει το προσωπικό σε εγρήγορση και θα του δώσει κίνητρο να διαπρέψει και να προσφέρει εξαιρετική εμπειρία στους επισκέπτες.[3]

Μειονεκτήματα της ανάλυσης συναισθήματος:

Όσον αφορά τις προκλήσεις της ανάλυσης συναισθήματος, υπάρχουν αρκετά πράγματα με τα οποία παλεύουν οι εταιρείες προκειμένου να επιτύχουν ακρίβεια στην ανάλυση συναισθήματος. Η ανάλυση συναισθήματος μπορεί να είναι δύσκολη στην επεξεργασία φυσικής γλώσσας, απλώς και μόνο επειδή οι μηχανές πρέπει να εκπαιδευτούν να αναλύουν και να κατανοούν τα συναισθήματα όπως κάνει ο ανθρώπινος εγκέφαλος. Καθώς η επιστήμη των δεδομένων συνεχίζει να εξελίσσεται, το λογισμικό ανάλυσης συναισθήματος είναι σε θέση να αντιμετωπίσει καλύτερα αυτά τα ζητήματα. Ακολουθούν τα κύρια εμπόδια στην ανάλυση συναισθήματος.

1. Τόνος:

Ο τόνος μπορεί να είναι δύσκολο να ερμηνευτεί προφορικά και ακόμη πιο δύσκολο να κατανοηθεί στον γραπτό λόγο. Τα πράγματα περιπλέκονται ακόμη περισσότερο όταν κάποιος προσπαθεί να αναλύσει έναν τεράστιο όγκο δεδομένων που μπορεί να εμπεριέχονται τόσο υποκειμενικές όσο και αντικειμενικές απαντήσεις. Οι μάρκες μπορεί να αντιμετωπίσουν δυσκολίες στην εύρεση υποκειμενικών συναισθημάτων και στη σωστή ανάλυσή τους για τον επιδιωκόμενο τόνο τους.

2. Πολικότητα:

Λέξεις όπως "αγάπη" και "μίσος" έχουν υψηλή θετική (+1) και αρνητική (-1) βαθμολογία στην πολικότητα. Αυτές είναι εύκολα κατανοητές. Υπάρχουν όμως ενδιάμεσες συζυγίες λέξεων όπως "όχι και τόσο κακός" που μπορεί να σημαίνουν "μέτριος" και επομένως βρίσκονται στη μέση πολικότητα (-75). Μερικές φορές φράσεις όπως αυτές παραλείπονται, γεγονός που αραιώνει τη βαθμολογία συναισθήματος.

3. Σαρκασμός:

Οι άνθρωποι χρησιμοποιούν την ειρωνεία και τον σαρκασμό σε περιστασιακές συζητήσεις και μιμίδια(memes) στα μέσα κοινωνικής δικτύωσης. Η πράξη της έκφρασης αρνητικών συναισθημάτων με τη χρήση πισώπλατων φιλοφρονήσεων μπορεί να καταστήσει δύσκολο για τα εργαλεία ανάλυσης συναισθήματος να ανιχνεύσουν το πραγματικό πλαίσιο του τι πραγματικά υπονοεί η απάντηση. Αυτό μπορεί συχνά να οδηγήσει σε μεγαλύτερο όγκο "θετικών" ανατροφοδοτήσεων που στην πραγματικότητα είναι αρνητικές.

4. Emojis:

Το πρόβλημα με το περιεχόμενο των μέσων κοινωνικής δικτύωσης που βασίζεται σε κείμενο, όπως το Twitter, είναι ότι κατακλύζονται από emojis. Οι εργασίες NLP εκπαιδεύονται ώστε να είναι εξειδικευμένες στη γλώσσα. Ενώ μπορούν να εξάγουν κείμενο ακόμη και από εικόνες, τα emojis είναι μια γλώσσα από μόνα τους. Οι περισσότερες λύσεις ανάλυσης συναισθημάτων αντιμετωπίζουν τα emojis σαν ειδικούς χαρακτήρες που αφαιρούνται από τα δεδομένα κατά τη διαδικασία εξόρυξης κειμένου. Με αυτόν τον τρόπο όμως οι εταιρείες δεν θα λάβουν ολιστικές γνώσεις από τα δεδομένα.

5. Ιδιώματα:

Τα προγράμματα μηχανικής μάθησης δεν κατανοούν απαραίτητα ένα σχήμα λόγου. Για παράδειγμα, ένας ιδιωματισμός όπως το "not my cup of tea" θα μπερδέψει τον αλγόριθμο επειδή κατανοεί τα πράγματα με την κυριολεκτική τους έννοια. Ως εκ τούτου, όταν ένας ιδιωματισμός χρησιμοποιείται σε ένα σχόλιο ή μια κριτική, η πρόταση μπορεί να παρερμηνευτεί από τον αλγόριθμο ή ακόμη και να αγνοηθεί. Για να ξεπεραστεί αυτό το πρόβλημα, μια πλατφόρμα ανάλυσης συναισθήματος πρέπει να εκπαιδευτεί στην κατανόηση ιδιωματισμών. Όταν πρόκειται για πολλαπλές γλώσσες, το πρόβλημα αυτό γίνεται πολλαπλό.

6. Αρνήσεις:

Οι αρνήσεις, που δίνονται από λέξεις όπως not, never, cannot, were not, κ.λπ. μπορούν να προκαλέσουν σύγχυση στο μοντέλο ML. Για παράδειγμα, ένας μηχανικός αλγόριθμος πρέπει να καταλάβει ότι μια φράση που λέει: "Δεν μπορώ να μην πάω στην επανένωση της τάξης μου", σημαίνει ότι το άτομο σκοπεύει να πάει στην επανένωση της τάξης.[4]

Πώς χρησιμοποιείται η μηχανική μάθηση για την ανάλυση συναισθήματος:

Ο πρωταρχικός ρόλος της μηχανικής μάθησης στην ανάλυση συναισθήματος είναι να βελτιώσει και να αυτοματοποιήσει τις λειτουργίες ανάλυσης κειμένου χαμηλού επιπέδου στις οποίες βασίζεται η ανάλυση συναισθήματος, συμπεριλαμβανομένης της επισήμανσης μέρους του λόγου. Για παράδειγμα, οι επιστήμονες δεδομένων μπορούν να εκπαιδεύσουν ένα μοντέλο μηχανικής μάθησης για τον εντοπισμό ουσιαστικών τροφοδοτώντας το με έναν μεγάλο όγκο εγγράφων κειμένου που περιέχουν παραδείγματα με προ-ετικέτες. Χρησιμοποιώντας εποπτευόμενες και μη εποπτευόμενες τεχνικές μηχανικής μάθησης,

όπως νευρωνικά δίκτυα και βαθιά μάθηση, το μοντέλο θα μάθει πώς μοιάζουν τα ουσιαστικά.

Μόλις το μοντέλο είναι έτοιμο, ο ίδιος επιστήμονας δεδομένων μπορεί να εφαρμόσει αυτές τις μεθόδους εκπαίδευσης για τη δημιουργία νέων μοντέλων για τον εντοπισμό άλλων μερών του λόγου. Το αποτέλεσμα είναι η γρήγορη και αξιόπιστη επισήμανση των μερών του λόγου που βοηθά το ευρύτερο σύστημα ανάλυσης κειμένου να εντοπίζει αποτελεσματικότερα τις φράσεις που περιέχουν συναισθήματα.

Η μηχανική μάθηση βοηθά επίσης τους αναλυτές δεδομένων να επιλύσουν δύσκολα προβλήματα που προκαλούνται από την εξέλιξη της γλώσσας. Για παράδειγμα, η φράση "sick burn" μπορεί να έχει πολλές ριζικά διαφορετικές έννοιες. Η δημιουργία ενός συνόλου κανόνων ανάλυσης συναισθήματος για να ληφθεί υπόψη κάθε πιθανή σημασία είναι αδύνατη. Αλλά αν τροφοδοτήσετε ένα μοντέλο μηχανικής μάθησης με μερικές χιλιάδες παραδείγματα με προ-σημειώσεις, μπορεί να μάθει να καταλαβαίνει τι σημαίνει "sick burn" στο πλαίσιο των βιντεοπαιχνιδιών, σε σχέση με το πλαίσιο της υγειονομικής περίθαλψης.[5]

2.1.2 Αλγόριθμοι Ανάλυσης Συναισθήματος

Όλα ξεκινούν με τη δημιουργία μιας βιβλιοθήκης συναισθημάτων

Οι βιβλιοθήκες συναισθήματος αποτελούνται από πολλαπλά λεξικά που έχουν έναν εξαντλητικό κατάλογο φράσεων και επιθέτων που έχουν βαθμολογηθεί προηγουμένως χειροκίνητα. Αυτός είναι ο ίδιος τρόπος με τον οποίο αντιλαμβανόμαστε τις φράσεις. Την πρώτη φορά που ακούμε μια φράση, μπορεί να μην την καταλαβαίνουμε, αλλά με βάση το πλαίσιο στο οποίο χρησιμοποιείται, αρχειοθετούμε στον εγκέφαλό μας αν έχει θετική, αρνητική ή ουδέτερη χροιά.

Οι σημασιολογικές βιβλιοθήκες το κάνουν με τον ίδιο τρόπο, αλλά οι ανθρώπινοι κωδικοποιητές θα βαθμολογούν με το χέρι κάθε μία από αυτές τις φράσεις. Αυτό μπορεί να είναι αρκετά δύσκολο, επειδή όλοι πρέπει να συμφωνήσουν σχετικά με τη βαθμολογία που πρέπει να δοθεί. Για παράδειγμα, αν ένα άτομο δώσει στη λέξη "awful" βαθμολογία -0,5 και ένα άλλο άτομο δώσει στη λέξη "dislike" την ίδια βαθμολογία, τότε η ανάλυση συναισθήματος θα θεωρήσει ότι και οι δύο λέξεις έχουν την ίδια αρνητική ένταση. Γνωρίζουμε όμως ότι το 'awful' θα πρέπει να υπερτερεί του 'dislike'.

Εάν χρειαζόμαστε μια πολυγλωσσική μηχανή συναισθήματος, τότε θα χρειαστούμε μοναδικές βιβλιοθήκες για κάθε γλώσσα. Και κάθε μία από αυτές τις βιβλιοθήκες πρέπει να συντηρείται, να τροποποιούνται οι βαθμολογίες και να προστίθενται ή να αφαιρούνται νέες φράσεις.

Μόλις η βιβλιοθήκη συναισθήματος είναι έτοιμη, το επόμενο βήμα είναι να αποφασίσουμε για το μοντέλο αλγορίθμου που θα προσδιορίσει το συναίσθημα πίσω από το κείμενο. Πρόκειται συνήθως για μια επιλογή από 3 σημαντικά μοντέλα αλγορίθμων ανάλυσης συναισθήματος. Το μοντέλο που κάποιος μπορεί να επιλέξει εξαρτάται από τον όγκο των δεδομένων που αναμένετε να επεξεργαστούμε και την ακρίβεια που χρειαζόμαστε.

1. Προσέγγιση βασισμένη σε κανόνες ή λεξικό

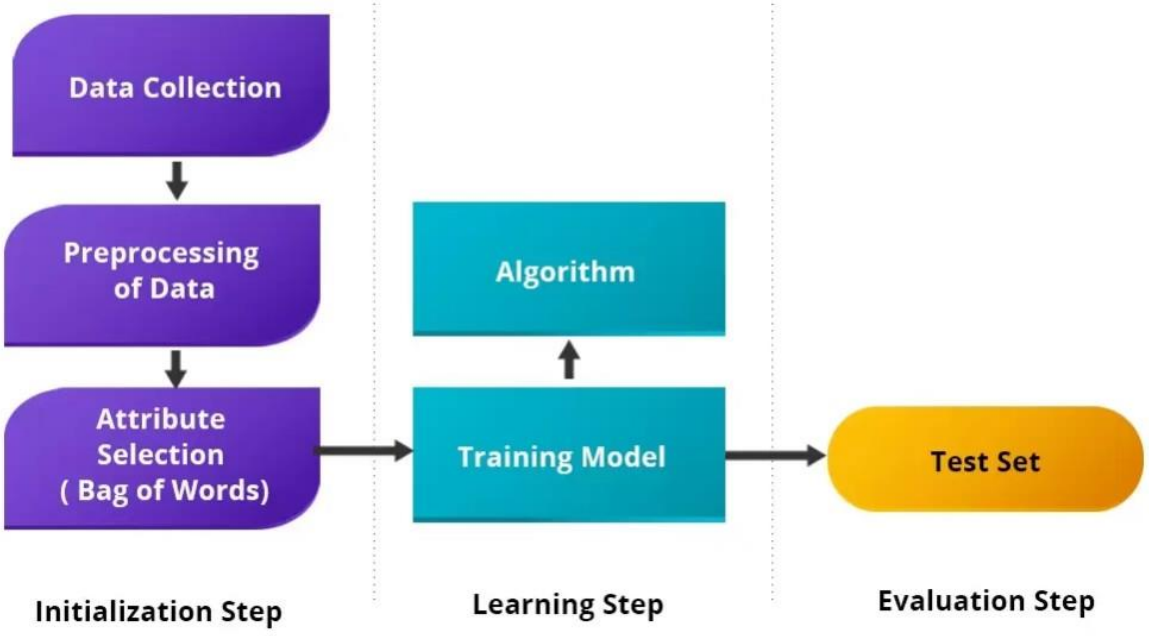
Αυτή η προσέγγιση βασίζεται σε χειροκίνητα διαμορφωμένους κανόνες για την ταξινόμηση των δεδομένων για τον προσδιορισμό του συναισθήματος. Αυτή η προσέγγιση χρησιμοποιεί λεξικά λέξεων με θετικές ή αρνητικές τιμές για να δηλώσει την πολικότητα και την ισχύ του συναισθήματος για τον υπολογισμό μιας βαθμολογίας. Μπορεί επίσης να προστεθεί πρόσθετη λειτουργικότητα με τη συμπερίληψη εκφράσεων. Οι αλγόριθμοι ανάλυσης συναισθήματος που βασίζονται σε κανόνες μπορούν να προσαρμοστούν με βάση το πλαίσιο αναπτύσσοντας ακόμη πιο έξυπνους κανόνες.

Πώς λειτουργεί: Μετράει τον αριθμό των θετικών και αρνητικών λέξεων στο συγκεκριμένο κείμενο. Εάν ο αριθμός των θετικών είναι μεγαλύτερος από τον αριθμό των αρνητικών, θα επιστρέψει ένα θετικό συναίσθημα. Εάν και τα δύο είναι ίσα, θα επιστρέψει ένα ουδέτερο συναίσθημα.

Μειονεκτήματα:

- Το μειονέκτημα αυτής της προσέγγισης είναι ότι δεν λαμβάνει υπόψη τον τρόπο με τον οποίο οι λέξεις συνδυάζονται σε μια πρόταση, εξετάζει μόνο τις εμφανίσεις.
- Εφαρμόζεται γρήγορα, αλλά το μοντέλο συνεπάγεται με μακροπρόθεσμη δαπάνη, καθώς απαιτεί τακτική συντήρηση, ώστε να έχουμε συνεπή και βελτιωμένα αποτελέσματα.

Βήματα που εμπλέκονται στην εκπαίδευση ενός ταξινομητή στην ανάλυση συναισθήματος:



Εικόνα 2.1: Βήματα εκπαίδευσης ενός ταξινομητή.[6]

2. Αυτοματοποιημένη προσέγγιση ή προσέγγιση μηχανικής μάθησης:

Αντί για σαφώς καθορισμένους κανόνες, αυτό το μοντέλο ανάλυσης συναισθήματος χρησιμοποιεί τη μηχανική μάθηση για να καταλάβει την ουσία της δήλωσης. Αυτό διασφαλίζει ότι η ακρίβεια της ανάλυσης βελτιώνεται και οι πληροφορίες μπορούν να επεξεργαστούν με πολλά κριτήρια χωρίς να είναι πολύ περίπλοκες. Η προσέγγιση αυτή περιλαμβάνει τη χρήση αλγορίθμων μηχανικής μάθησης υπό επίβλεψη. Ένας αλγόριθμος εκπαιδεύεται με πολλά δείγματα αποσπασμάτων μέχρι να μπορεί να προβλέψει με ακρίβεια το συναίσθημα του κειμένου. Στη συνέχεια, μεγάλα κομμάτια κειμένου τροφοδοτούνται στον ταξινομητή και αυτός προβλέπει το συναίσθημα ως αρνητικό, ουδέτερο ή θετικό.

Τα μοντέλα μηχανικής μάθησης μπορεί να είναι δύο ειδών:

α. Παραδοσιακά μοντέλα - Αυτή η μέθοδος απαιτεί τη συγκέντρωση ενός συνόλου δεδομένων με παραδείγματα για θετικές, αρνητικές και ουδέτερες κλάσεις, στη συνέχεια την επεξεργασία αυτών των δεδομένων και, τέλος, την εκπαίδευση του αλγορίθμου με βάση τα παραδείγματα. Αυτές οι μέθοδοι χρησιμοποιούνται κυρίως για τον προσδιορισμό της πολικότητας του κειμένου.

Οι παραδοσιακές μέθοδοι μηχανικής μάθησης, όπως οι μέθοδοι Naïve Bayes, Logistic Regression και Support Vector Machines (SVM), χρησιμοποιούνται ευρέως για την ανάλυση συναισθήματος μεγάλης κλίμακας, επειδή έχουν δυνατότητα κλιμάκωσης.

β. Μοντέλα βαθιάς μάθησης- Αυτό παρέχει πιο ακριβή αποτελέσματα από τα παραδοσιακά μοντέλα και περιλαμβάνει μοντέλα νευρωνικών δικτύων όπως το CNN (Convoluted Neural Network), το RNN (Recurrent Neural Network) και το DNN (Deep Neural Network).

Τα κύρια μοντέλα που χρησιμοποιούνται για τους αλγορίθμους ταξινόμησης ανάλυσης συναισθήματος είναι το Naïve Bayes και το Deep Learning:

Ανάλυση συναισθήματος Naive Bayes:

Ονομάζεται "Naïve" επειδή χρησιμοποιεί την υπόθεση ότι η εμφάνιση ενός χαρακτηριστικού είναι ανεξάρτητη από άλλα χαρακτηριστικά. Για παράδειγμα, αναγνωρίζει το πορτοκάλι με βάση το χρώμα, το σχήμα και τη γεύση, με κάθε χαρακτηριστικό να αξιολογείται ανεξάρτητα για να καταλήξει στο συμπέρασμα. Το 'Bayes' είναι επειδή βασίζεται στην αρχή του θεωρήματος Bayes.

Το θεώρημα Bayes βασίζεται στην έννοια της υπό συνθήκη πιθανότητας ή της πιθανότητας να συμβεί το γεγονός A όταν συμβεί το γεγονός B. Το θεώρημα στην πραγματικότητα δηλώνει ότι η πιθανότητα του A αν το B είναι αληθές = η πιθανότητα του B αν το A είναι αληθές, πολλαπλασιασμένη με το επί τοις εκατό της πιθανότητας να είναι αληθές το A και το σύνολο διαιρούμενο με την πιθανότητα να είναι αληθές το B:

$$P(A|B)= \frac{P(B|A)P(A)}{P(B)}$$

Στην ανάλυση συναισθήματος Naïve Bayes, ο ταξινομητής Bayes ταξινομεί έγγραφα, κείμενα ή προϊόντα ως θετικά ή αρνητικά.

Για παράδειγμα, στην πρόταση "Μου αρέσει πολύ αυτό το προϊόν", υπάρχει μια σαφής αίσθηση του θετικού συναισθήματος. Ο ταξινομητής υπολογίζει κάθε τιμή πιθανότητας και η κλάση επιλέγεται ως θετική επειδή η θετική τιμή υπερτερεί.

Βαθιά μάθηση:

Η ανάλυση συναισθήματος που χρησιμοποιεί βαθιά μάθηση NLP είναι σε θέση να μάθει μοτίβα μέσω πολλαπλών επιπέδων από μη δομημένα και μη επισημασμένα δεδομένα για να εκτελέσει ανάλυση συναισθήματος. Δύο τεχνικές νευρωνικών δικτύων είναι κοινές - CNN ή Convolutional Neural Networks για την επεξεργασία εικόνων και RNN ή Recurrent Neural Networks για εργασίες NLP.

3. Υβριδική προσέγγιση:

Τα υβριδικά μοντέλα ανάλυσης συναισθήματος είναι η πιο σύγχρονη, αποτελεσματική και ευρέως χρησιμοποιούμενη προσέγγιση για την ανάλυση συναισθήματος. Με την προϋπόθεση ότι έχετε καλά σχεδιασμένα υβριδικά συστήματα, μπορείτε στην πραγματικότητα να επωφεληθείτε από τα οφέλη τόσο των αυτόματων όσο και των συστημάτων που βασίζονται σε κανόνες. Τα υβριδικά μοντέλα μπορούν να προσφέρουν τη δύναμη της μηχανικής μάθησης σε συνδυασμό με την ευελιξία της προσαρμογής.[7]

2.2 Προβλεπτική Αναλυτική (Predictive Analytics)

Ο όρος predictive analytics αναφέρεται στη χρήση στατιστικών και τεχνικών μοντελοποίησης για την πραγματοποίηση προβλέψεων σχετικά με μελλοντικά αποτελέσματα και επιδόσεις. Η προβλεπτική ανάλυση εξετάζει τα τρέχοντα και ιστορικά πρότυπα δεδομένων για να καθορίσει εάν αυτά τα πρότυπα είναι πιθανό να εμφανιστούν ξανά. Αυτό επιτρέπει στις επιχειρήσεις και τους επενδυτές να προσαρμόσουν το πού χρησιμοποιούν τους πόρους τους για να επωφεληθούν από πιθανά μελλοντικά γεγονότα. Η προβλεπτική ανάλυση μπορεί επίσης να χρησιμοποιηθεί για τη βελτίωση της λειτουργικής αποτελεσματικότητας και τη μείωση του κινδύνου.

2.2.1 Θεωρητικό Υπόβαθρο

Αν και η προβλεπτική ανάλυση υπάρχει εδώ και δεκαετίες, είναι μια τεχνολογία της οποίας η ώρα έφτασε. Όλο και περισσότεροι οργανισμοί στρέφονται στην προγνωστική ανάλυση για να αυξήσουν το τελικό αποτέλεσμα και το ανταγωνιστικό τους πλεονέκτημα. Γιατί τώρα;

- Ο αυξανόμενος όγκος και οι τύποι δεδομένων και το μεγαλύτερο ενδιαφέρον για τη χρήση των δεδομένων για την παραγωγή πολύτιμων πληροφοριών.

- Ταχύτεροι, φθηνότεροι υπολογιστές.
- Ευκολότερο στη χρήση λογισμικό.
- Σκληρότερες οικονομικές συνθήκες και ανάγκη για ανταγωνιστική διαφοροποίηση.

Με το διαδραστικό και εύχρηστο λογισμικό να γίνεται όλο και πιο διαδεδομένο, η προβλεπτική ανάλυση δεν είναι πλέον μόνο ο τομέας των μαθηματικών και των στατιστικολόγων. Οι επιχειρηματικοί αναλυτές και οι ειδικοί των επιχειρήσεων χρησιμοποιούν επίσης αυτές τις τεχνολογίες.[8]

Γιατί η προβλεπτική ανάλυση είναι σημαντική:

Η ανάγκη για προβλεπτική ανάλυση είναι αναμφισβήτητα πιο κρίσιμη από ποτέ. "Η παραδοσιακή έννοια της μάθησης από τα λάθη δεν ισχύει πλέον- η πραγματικότητα στις μέρες μας μοιάζει περισσότερο με το "ένα χτύπημα και είσαι εκτός"", έγραψε ο Delen, καθηγητής διοικητικής επιστήμης και πληροφοριακών συστημάτων στο Πολιτειακό Πανεπιστήμιο της Οκλαχόμα, στην εισαγωγή του στο Predictive Analytics, Second Edition. "Οι οργανισμοί που χρησιμοποιούν την επιχειρηματική ανάλυση όχι μόνο μπορούν να επιβιώσουν, αλλά συχνά ευδοκιμούν σε τέτοιου είδους συνθήκες".

Τα δεδομένα είναι η ζωογόνος δύναμη της επιχειρηματικής ανάλυσης και, ολοένα και περισσότερο, το καύσιμο των επιχειρήσεων. Οι εταιρείες, μεγάλες και μικρές, λειτουργούν με δεδομένα που παράγονται και συλλέγονται από τις δραστηριότητές τους και από εξωτερικές πηγές. Για παράδειγμα, οι εταιρείες συλλέγουν δεδομένα για κάθε βήμα του ταξιδιού του αγοραστή, παρακολουθώντας πότε, τι, πόσο και πόσο συχνά αγοράζουν οι πελάτες. Παρακολουθούν επίσης τις αποτυχίες των πελατών, τα παράπονα, τις καθυστερήσεις πληρωμών, τις πιστωτικές αθετήσεις και την απάτη.

Όμως, ο τεράστιος όγκος δεδομένων που συσσωρεύουν οι επιχειρήσεις σχετικά με τους πελάτες τους, τις επιχειρηματικές λειτουργίες, τους προμηθευτές, την απόδοση των εργαζομένων κ.ο.κ. δεν είναι χρήσιμος αν δεν αξιοποιηθεί. "Τα δεδομένα έχουν γίνει τόσο διαδεδομένα στις επιχειρηματικές λειτουργίες που η απλή πρόσβαση σε περισσότερα ή καλύτερα δεδομένα δεν αποτελεί από μόνη της βασική διαφορά", σημειώνει ο ειδικός στα αναλυτικά συστήματα Donald Farmer, διευθυντής της συμβουλευτικής εταιρείας TreeHive Strategy. "Αυτό που αλλάζει τα επιχειρηματικά αποτελέσματα σήμερα είναι ο τρόπος με

τον οποίο κατανοούμε και ενεργούμε με βάση τα δεδομένα μας. Αυτή η κατανόηση απαιτεί αναλυτικά στοιχεία".

Η προβλεπτική ανάλυση δίνει στις επιχειρήσεις ένα πλεονέκτημα, αναζητώντας σημαντικά μοτίβα σε αυτά τα συσσωρευμένα δεδομένα, και στη συνέχεια δημιουργώντας μοντέλα που προβλέπουν τι θα συμβεί πιθανότατα στο μέλλον. Για παράδειγμα, με βάση τη συμπεριφορά ενός πελάτη στο παρελθόν και τη συμπεριφορά άλλων πελατών με παρόμοια χαρακτηριστικά, πόσο πιθανό είναι ο πελάτης να ανταποκριθεί σε έναν συγκεκριμένο τύπο προσφοράς μάρκετινγκ, να αθετήσει μια πληρωμή ή να βιδώσει;

Τα έμπειρα τμήματα πωλήσεων και μάρκετινγκ έχουν εδώ και καιρό αξιοποιήσει την προβλεπτική μοντελοποίηση, αλλά η χρήση της προβλεπτικής ανάλυσης μπορεί πλέον να βρεθεί σε όλες τις επιχειρηματικές λειτουργίες και κλάδους. Χρησιμοποιείται τακτικά από τους οργανισμούς για τη βελτίωση βασικών μετρήσεων απόδοσης με τη μείωση του κινδύνου, τη βελτιστοποίηση των λειτουργιών και την αύξηση της αποδοτικότητας, καθώς και για τον καθορισμό στρατηγικών που τελικά προσδίδουν ανταγωνιστικό πλεονέκτημα.[9]

Λειτουργία Προβλεπτικής Ανάλυσης:

Οι προβλεπτικές αναλύσεις βασίζονται σε μεγάλο βαθμό στη μηχανική μάθηση (ML). Η ML είναι ένας συνδυασμός στατιστικής και επιστήμης των υπολογιστών που χρησιμοποιείται για τη δημιουργία μοντέλων μέσω της επεξεργασίας δεδομένων με αλγορίθμους. Αυτά τα μοντέλα μπορούν να αναγνωρίσουν τάσεις και μοτίβα στα δεδομένα που είναι γενικά βαθύτερα σε πολυπλοκότητα από ό,τι μόνο οι οπτικές μέθοδοι ανακάλυψης δεδομένων από μόνες τους. Χρησιμοποιώντας δεδομένα από διάφορες πηγές (για παράδειγμα, το Διαδίκτυο των πραγμάτων (IoT), αισθητήρες, μέσα κοινωνικής δικτύωσης και μια σειρά συσκευών), η μηχανική μάθηση επεξεργάζεται τα δεδομένα αυτά μέσω εξελιγμένων αλγορίθμων και δημιουργεί μοντέλα για τον εντοπισμό και την επίλυση ενός προβλήματος και την πραγματοποίηση προβλέψεων.

Οι προβλεπτικές αναλύσεις βασίζονται επίσης στην επιστήμη των δεδομένων, η οποία είναι μια πιο περιεκτική έννοια από την απλή ML. Η επιστήμη των δεδομένων συνδυάζει τη στατιστική, την επιστήμη των υπολογιστών και τη γνώση συγκεκριμένων εφαρμογών για την επίλυση ενός προβλήματος. Σε ένα επιχειρηματικό περιβάλλον, συνδυάζει μεθόδους μηχανικής μάθησης με επιχειρηματικά δεδομένα, διαδικασίες και τεχνογνωσία τομέα για

την επίλυση ενός επιχειρηματικού προβλήματος. Βασικά, παρέχει προγνωστικές γνώσεις στους υπεύθυνους λήψης αποφάσεων.

Μπορούμε να ενσωματώσουμε ένα μοντέλο για να προβλέψουμε ένα πιθανό αποτέλεσμα ή να παρέχουμε μια βελτιστοποιημένη λύση για αλλαγές στις παραμέτρους της διαδικασίας απευθείας μέσα στις επιχειρηματικές διαδικασίες. Ένα μοντέλο παρέχει ανταγωνιστικό πλεονέκτημα επειδή κάνει τα εξής:

- Ενισχύει τις δυνατότητες
- Επιταχύνει τη λήψη αποφάσεων
- Επεξεργάζεται μεγάλες ποσότητες διαφορετικών τύπων δεδομένων
- Μειώνει γενικά το κόστος λειτουργίας
- Δημιουργεί νέες ροές εσόδων
- Οδηγεί σε διαφοροποιημένα προϊόντα και προσφορές υπηρεσιών[10]

2.2.2 Κατηγορίες Αλγορίθμων Προβλεπτικής Αναλυτικής

Εάν εργάζεστε ή διαβάζετε για την ανάλυση, τότε η προγνωστική ανάλυση είναι ένας όρος που έχετε ξανακούσει. Επί του παρόντος, το πιο περιζήτητο μοντέλο στον κλάδο, τα μοντέλα προβλεπτικής ανάλυσης έχουν σχεδιαστεί για να αξιολογούν ιστορικά δεδομένα, να ανακαλύπτουν μοτίβα, να παρατηρούν τάσεις και να χρησιμοποιούν αυτές τις πληροφορίες για να συντάσσουν προβλέψεις σχετικά με τις μελλοντικές τάσεις.

Οι δυνητικές εφαρμογές της προβλεπτικής ανάλυσης ποικίλλουν ευρέως, όπως και οι τύποι των μοντέλων που χρησιμοποιούνται για την τροφοδότηση των συμπερασμάτων που προκύπτουν. Ο προσδιορισμός των τύπων τεχνικών προβλεπτικής ανάλυσης που είναι οι καλύτερες για τον οργανισμό σας ξεκινά με έναν σαφώς καθορισμένο στόχο. Μόλις γνωρίζετε ποιο ερώτημα θέλετε να απαντήσετε, μπορείτε να επιλέξετε το μοντέλο που σας εξυπηρετεί καλύτερα. Τα μοντέλα προγνωστικής ανάλυσης μπορούν να ομαδοποιηθούν χονδρικά σε αυτούς τους τέσσερις τύπους:

1. *Μοντέλα παλινδρόμησης (Regression Models):*

Τα μοντέλα παλινδρόμησης εκτιμούν τη δύναμη μιας σχέσης μεταξύ μεταβλητών. Το μοντέλο παρακολουθεί τον τρόπο με τον οποίο οι ενέργειες (ανεξάρτητες μεταβλητές) επηρεάζουν τα αποτελέσματα (εξαρτημένες μεταβλητές) και χρησιμοποιεί αυτές τις

πληροφορίες για να προβλέψει τις μελλοντικές επιπτώσεις. Αυτά τα στατιστικά μοντέλα μπορεί να είναι απλά, με μία ανεξάρτητη μεταβλητή και μία εξαρτημένη μεταβλητή ή μια πολλαπλή γραμμική παλινδρόμηση με δύο ή περισσότερες ανεξάρτητες μεταβλητές. Υπάρχουν διάφορες τεχνικές παλινδρόμησης και μπορούν να χρησιμοποιηθούν ανάλογα με την εφαρμογή και τους τύπους των εμπλεκόμενων μεταβλητών. Καθορίζοντας τη σχέση μεταξύ των μεταβλητών, οι οργανισμοί μπορούν να εκτελέσουν ανάλυση σεναρίων, γνωστή και ως ανάλυση "τι θα γίνει αν", για να προσθέσουν νέες ανεξάρτητες μεταβλητές και να δουν πώς επηρεάζουν το αποτέλεσμα. Οι οργανισμοί θα μπορούσαν να χρησιμοποιήσουν ένα μοντέλο παλινδρόμησης για να προσδιορίσουν πώς οι ιδιότητες ενός προϊόντος επηρεάζουν την πιθανότητα αγοράς. Αναλύοντας τη σχέση μεταξύ του χρώματος του προϊόντος και της πιθανότητας αγοράς, ένας οργανισμός μπορεί να δει μια συσχέτιση μεταξύ μπλε πουκάμισων και περισσότερων πωλήσεων. Επειδή η συσχέτιση δεν ισοδυναμεί με αιτιώδη συνάφεια, ο οργανισμός μπορεί να διερευνήσει πώς άλλοι παράγοντες επηρεάζουν την πιθανότητα αγοράς, όπως το μέγεθος, η εποχικότητα ή η τοποθέτηση του προϊόντος. Μπορεί να χρησιμοποιήσει αυτές τις γνώσεις για να βοηθήσει στις προσπάθειες μάρκετινγκ ή στην ανάπτυξη προϊόντων, ώστε να καθορίσει ποια προϊόντα θα μπορούσαν να έχουν καλή απόδοση στο μέλλον.

2. *Μοντέλα ταξινόμησης (Classification Models):*

Τα μοντέλα ταξινόμησης τοποθετούν τα δεδομένα σε κατηγορίες με βάση την ιστορική γνώση. Η ταξινόμηση ξεκινά με ένα σύνολο δεδομένων εκπαίδευσης όπου κάθε δεδομένο έχει ήδη επισημανθεί. Ο αλγόριθμος ταξινόμησης μαθαίνει τις συσχετίσεις μεταξύ των δεδομένων και των ετικετών και κατηγοριοποιεί κάθε νέο δεδομένο. Ορισμένες δημοφιλείς τεχνικές μοντέλων ταξινόμησης περιλαμβάνουν τα δέντρα αποφάσεων, τα τυχαία δάση και την ανάλυση κειμένου. Επειδή τα μοντέλα ταξινόμησης μπορούν εύκολα να επανεκπαιδευτούν με νέα δεδομένα, χρησιμοποιούνται σε πολλούς κλάδους. Οι τράπεζες χρησιμοποιούν συχνά μοντέλα ταξινόμησης για τον εντοπισμό δόλιων συναλλαγών. Ο αλγόριθμος μπορεί να αναλύσει εκατομμύρια προηγούμενες συναλλαγές για να μάθει πώς μπορεί να μοιάζουν οι μελλοντικές δόλιες συναλλαγές και να ειδοποιήσει τους πελάτες όταν η δραστηριότητα στον λογαριασμό τους φαίνεται ύποπτη.

3. *Μοντέλα ομαδοποίησης (Clustering Models):*

Τα μοντέλα ομαδοποίησης τοποθετούν τα δεδομένα σε ομάδες με βάση παρόμοια χαρακτηριστικά. Ένα μοντέλο ομαδοποίησης χρησιμοποιεί έναν πίνακα δεδομένων, ο οποίος συσχετίζει κάθε στοιχείο με σχετικά χαρακτηριστικά. Με αυτόν τον πίνακα, ο αλγόριθμος θα ομαδοποιήσει τα στοιχεία που έχουν τα ίδια χαρακτηριστικά, εντοπίζοντας μοτίβα στα δεδομένα που μπορεί προηγουμένως να ήταν κρυμμένα. Οι οργανισμοί μπορούν να χρησιμοποιούν μοντέλα ομαδοποίησης για να ομαδοποιούν τους πελάτες και να δημιουργούν πιο εξατομικευμένες στρατηγικές στόχευσης. Για παράδειγμα, ένα εστιατόριο μπορεί να ομαδοποιήσει τους πελάτες του με βάση την τοποθεσία και να στείλει φυλλάδια μόνο σε πελάτες που ζουν σε μια ορισμένη απόσταση οδήγησης από τη νεότερη τοποθεσία του.[11]

2.2.3 Προβλεπτική Αναλυτική σε Διαδικτυακές Κριτικές Πελατών

Είναι δύσκολο να φανταστεί κανείς έναν κόσμο πριν από τις διαδικτυακές κριτικές. Το 2021, δεν θα αγοράζατε ποτέ μια νέα φωτογραφική μηχανή χωρίς να ξέρετε πρώτα τι πιστεύουν όλοι οι άλλοι γι' αυτήν. Πώς αλλιώς μπορείτε να επιβεβαιώσετε ότι η ποιότητα της εικόνας, η μορφή της οθόνης, οι τεχνικές δυνατότητες, η απόδοση της μπαταρίας και η ταχύτητα λήψης είναι όλες στο ίδιο επίπεδο; Οι καταναλωτές εμπιστεύονται τις διαδικτυακές κριτικές - ακόμη και αν δεν έχουν ιδέα ποιος τις έγραψε.

Για τις μάρκες, τα οφέλη των διαδικτυακών κριτικών είναι διπλά.

1. Οι καλές κριτικές είναι μια μεγάλη πηγή οργανικής έκθεσης και είναι πιθανό να οδηγήσουν σε αύξηση των πωλήσεων.
2. Η ανάλυση κειμένου των κριτικών μπορεί να βοηθήσει τους οργανισμούς να κατανοήσουν καλύτερα τους πελάτες τους - συμπεριλαμβανομένων των προτιμήσεων, των προσδοκιών, των παραπόνων και των απόψεών τους.

Γιατί οι κριτικές πελατών έχουν σημασία;

Οι κριτικές και οι διαδικτυακές αξιολογήσεις επηρεάζουν όλο και περισσότερο την αγοραστική συμπεριφορά. Παρακάτω παραθέτονται ορισμένοι αριθμοί που αποτυπώνουν τη σημασία της καλλιέργειας θετικών δημόσιων επαίνων από τους πελάτες.

- Το 97% των καταναλωτών αναζητούν τοπικά προϊόντα ή υπηρεσίες στο διαδίκτυο.
- Το 91% των ατόμων ηλικίας 18-34 ετών εμπιστεύεται τις διαδικτυακές κριτικές.

- Το 93% των καταναλωτών ισχυρίζονται ότι οι κριτικές επηρέασαν την απόφαση αγοράς τους.
- Οι πελάτες είναι διατεθειμένοι να δαπανήσουν 31% περισσότερο για ένα προϊόν με καλές κριτικές.
- Μόνο το 13% των πελατών θα εξετάσει μια επιχείρηση με 1 ή 2 αστέρια.
- 4 στους 5 καταναλωτές άλλαξαν γνώμη για μια αγορά αφού διάβασαν κακές κριτικές.

Για να τα συνοψίσουμε όλα αυτά: οι διαδικτυακές αξιολογήσεις είναι ένα δίκοπο μαχαίρι. Οι θετικές κριτικές δεν είναι μόνο ένας πολύ καλός τρόπος για να αυξήσετε τις πωλήσεις, οι αρνητικές κριτικές είναι ένας σίγουρος τρόπος για να χάσετε πιθανούς πελάτες.

Κατανόηση των προφίλ των κριτών:

Οι πληροφορίες πίσω από τον κριτικό είναι εξίσου σημαντικές με την ίδια την κριτική. Η καταγραφή των χαρακτηριστικών του κριτικού δημιουργεί μια πηγή πληροφοριών πλούσια σε μεταδεδομένα. Αυτό το πρόσθετο επίπεδο δεδομένων σας επιτρέπει να εντοπίζετε τάσεις και συσχετίσεις με βάση τα προφίλ πελατών.

Χρήσιμες πληροφορίες:

- Ηλικία
- Τοποθεσία
- Φύλο
- Ημερομηνία αγοράς
- Είδος(-α) που αγοράστηκε(-ονται)
- Συνολική βαθμολογία
- Εύρος εισοδήματος
- Επίπεδο εκπαίδευσης
- Κανάλι αναθεώρησης[12]

Σημασία Ανάλυσης Κριτικών:

1. Περισσότερη επιχειρηματική ανάπτυξη

- Αναλύοντας τα σχόλια και ακούγοντας τους πελάτες, υπάρχει πιθανότητα αύξησης στα ποσοστά επιτυχίας των upselling και cross-selling κατά 15% έως 20%.

- Αυτό σημαίνει βελτιωμένη διατήρηση των πελατών, λιγότερες αποχωρήσεις και υψηλότερη αξία ζωής των πελατών, καθώς και μικρότερο κόστος για τη διατήρηση των αγοραστών.

2. Καλύτερη εμπειρία πελάτη

- Η ακρόαση των πελατών σας σημαίνει τελικά τη δημιουργία μιας καλύτερης εμπειρίας πελατών με την πάροδο του χρόνου, οδηγώντας σε θετικές αλλαγές στα σωστά σημεία.
- Θα ξέρετε πού πονάει, και πού πρέπει να βελτιωθείτε.

3. Καλύτερα προϊόντα και υπηρεσίες

Η σωστή ανάλυση των ανατροφοδοτήσεων των πελατών σας θα βοηθήσει τις διαδικασίες ανάπτυξης των προϊόντων, ώστε να μπορείτε να αναπτύξετε καλύτερα προϊόντα που οι πελάτες σας θα εκτιμήσουν περισσότερο, καθώς θα ανταποκρίνονται καλύτερα στις απαιτήσεις τους.

Πώς να αναλύετε τα σχόλια των πελατών:

Υπάρχουν 4 τύποι λύσεων για την ανάλυση των ανατροφοδοτήσεων των πελατών:

- Εξωτερικός αναλυτής δεδομένων
- Πρόσληψη αναλυτή δεδομένων
- Ανάθεση σε εξωτερικό οργανισμό
- Προϊόντα SaaS

Συμπέρασμα:

Η προβλεπτική ανάλυση είναι ένας έξυπνος τρόπος για να προσθέσετε περισσότερη διορατικότητα και σαφήνεια στις επιχειρηματικές σας αποφάσεις. Παρόλο που μπορεί να χρειαστεί πολύς χρόνος για να συλλέξετε χρήσιμα δεδομένα και να καταστρώσετε ένα σχέδιο για την ταξινόμησή τους, όταν δείτε τα αποτελέσματα από αυτά που μπορούν να κάνουν, θα αξίζει τον κόπο.

ΚΕΦΑΛΑΙΟ 3

3.1 Επισκόπηση της Προτεινόμενης Προσέγγισης

Παρακάτω αναλύεται μία προτεινόμενη προσέγγιση για το πώς μπορούμε χρησιμοποιώντας sentiment analysis να αναλύσουμε την γνώμη πελατών ξενοδοχείων για διάφορες υπηρεσίες των ξενοδοχείων:

Αρχικά επιλέγουμε την πηγή από την οποία θα πάρουμε το dataset που θα αφορά κριτικές ξενοδοχείων. Αφού κατέληξα στο TripAdvisor, ακολουθεί η προ-επεξεργασία δεδομένων η οποία αποτελεί το πρώτο βήμα μηχανικής μάθησης στο οποίο μετατρέπουμε τα ακατέργαστα δεδομένα σε μία μορφή χρήσιμη για τα μοντέλα μηχανικής μάθησης.

Αυτά τα **Δεδομένα Εισόδου** στην συνέχεια θα τα συγκρίνω με τα index set της κάθε κατηγορίας χρησιμοποιώντας την τεχνική **Fuzzy String Matching**, έτσι ώστε να δούμε πόσο ταιριάζουν τα δεδομένα μεταξύ τους. Ο **ορισμός των κατηγοριών** πραγματοποιήθηκε έπειτα από προσωπική επιλογή.

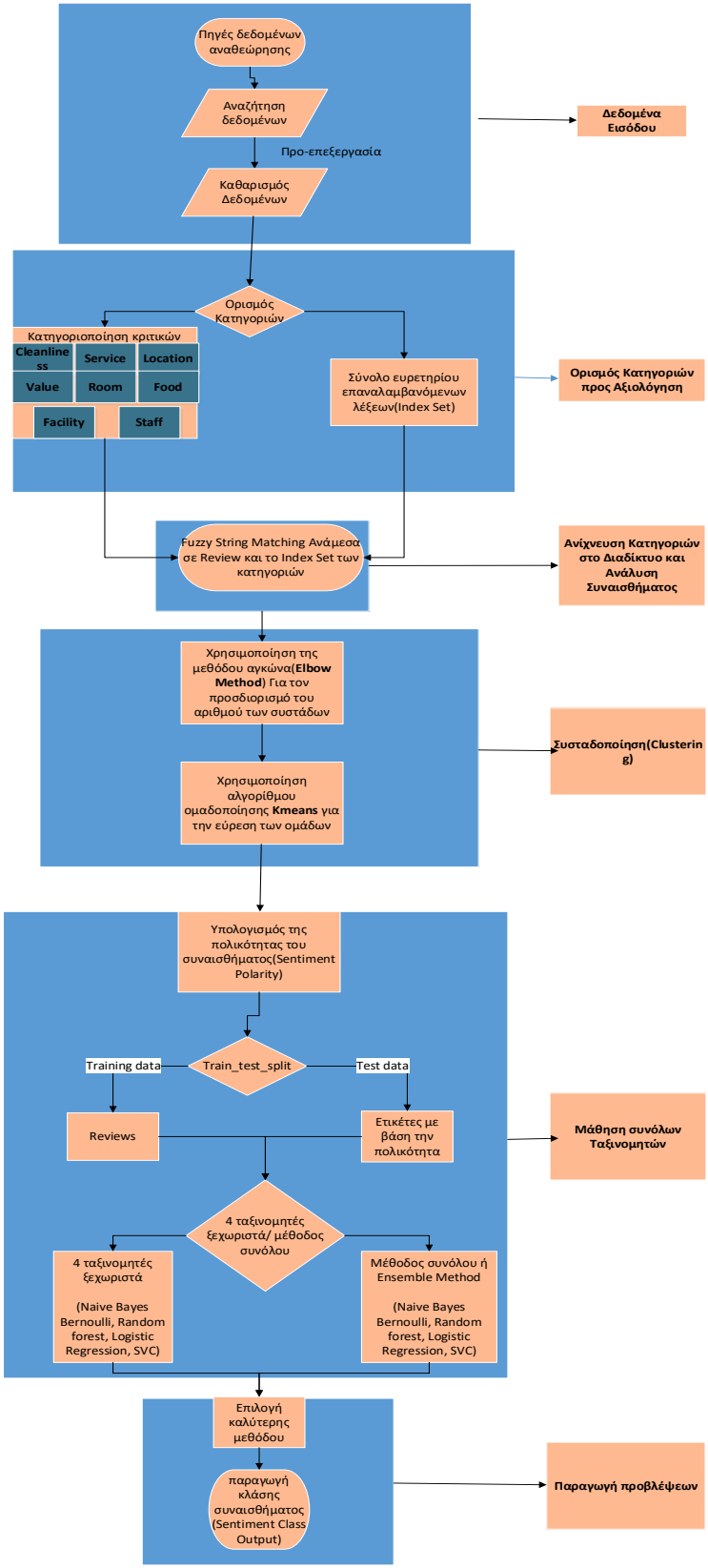
Στην συνέχεια αποφάσισα να κανω **Συσταδοποίηση (Clustering)** χρησιμοποιώντας τον αλγόριθμο K-means, με σκοπό την δημιουργία συστάδων, όπου κάθε συστάδα θα αποτελείται από ξενοδοχεία τα οποία ‘ταιριάζουν’ περισσότερο σε μία ή περισσότερες κατηγορίες. Για την εύρεση κατάλληλης τιμής για τον αριθμό των συστάδων χρησιμοποιήθηκε η μέθοδος Elbow.

Μετέπειτα κατηγοριοποίησα τις κριτικές ανάλογα με το συναίσθημα χρησιμοποιώντας 2 τρόπους:

- 4 μεμονωμένους αλγορίθμους ταξινόμησης (Naïve Bayes, Random Foret, Logistic Regression, SVM linear)
- Μέθοδο Ensemble (Συνδυασμός των παραπάνω 4 αλγορίθμων σε έναν ενιαίο)

με σκοπό την **Μάθηση 2 ταξινομητών** ικανών να προβλέψουν σωστά το συναίσθημα της κάθε κριτικής.

Τέλος, αφού πραγματοποιηθούν συγκρίσεις χρησιμοποιώντας μετρικές όπως (accuracy, precision, recall, f1-score), αλλά και πίνακες σύγχησης και classification reports, θα επιλέξουμε ποια είναι η αποδοτικότερη μέθοδος με την οποία θα μπορέσουμε να πραγματοποιήσουμε την **Παραγωγή Προβλέψεων**.



Εικόνα 3.1: Προτεινόμενη προσέγγιση για ανάλυση γνώμης κριτικών ξενοδοχείων

3.2 Δεδομένα Εισόδου

Σχετικά με το Tripadvisor:

Η Tripadvisor, Inc. είναι μια αμερικανική διαδικτυακή ταξιδιωτική εταιρεία που λειτουργεί έναν ιστότοπο και μια εφαρμογή για κινητά με περιεχόμενο που δημιουργείται από τους χρήστες και έναν ιστότοπο συγκριτικών αγορών. Προσφέρει επίσης online κρατήσεις ξενοδοχείων και κρατήσεις για μεταφορές, καταλύματα, ταξιδιωτικές εμπειρίες και εστιατόρια. Εμείς θα χρησιμοποιήσουμε ένα dataset που αφορά κριτικές πελατών πάνω στα ξενοδοχεία που επισκέφθηκαν.

Σχετικά με το σύνολο δεδομένων:

Τα ξενοδοχεία διαδραματίζουν καθοριστικό ρόλο στα ταξίδια και με την αυξημένη πρόσβαση στις πληροφορίες αναδύθηκαν νέοι τρόποι επιλογής των καλύτερων. Με αυτό το σύνολο δεδομένων, το οποίο αποτελείται από περίπου 10k κριτικές που συλλέγονται από το Tripadvisor, μπορούμε να εξερευνήσουμε τι κάνει ένα καλό ξενοδοχείο.

Προεπεξεργασία δεδομένων:

Τα δεδομένα που συλλέγονται από ιστότοπους περιέχουν συνήθως πολύ θόρυβο: λάθη, πληροφορίες χωρίς νόημα, ασυνεπή μορφοποίηση ή ελλιπείς προτάσεις. Αυτό δυσκολεύει την επεξεργασία τους από τις μηχανές και επηρεάζει τα αποτελέσματα της ανάλυσης.

Για να το αποφύγουμε αυτό, θα πρέπει να προεπεξεργαστούμε ή να καθαρίσουμε τα δεδομένα μας πριν εκτελέσουμε οποιοδήποτε είδος ανάλυσης κειμένου. Ακολουθούν ορισμένοι τρόποι με τους οποίους μπορούμε να προετοιμάσουμε τα δεδομένα μας και να βελτιώσουμε τη συνολική ποιότητά τους:

- Αφαίρεση Stop Words. Οι λέξεις στάσεις (όπως a, at, is, from, there, κ.λπ.) εμφανίζονται συχνά στα κείμενα αλλά δεν προσθέτουν σχετικές πληροφορίες.
- Αφαίρεση των emojis, τους ειδικούς χαρακτήρες, τα στοιχεία HTML, τα σημεία στίξης κ.λπ.
- Μετατροπή όλων των δεδομένων του κειμένου σε πεζά γράμματα.

- Αναγωγή των λέξεων στη μορφή της ρίζας τους (lemmatization).

3.3 Ορισμός Κατηγοριών προς Αξιολόγηση

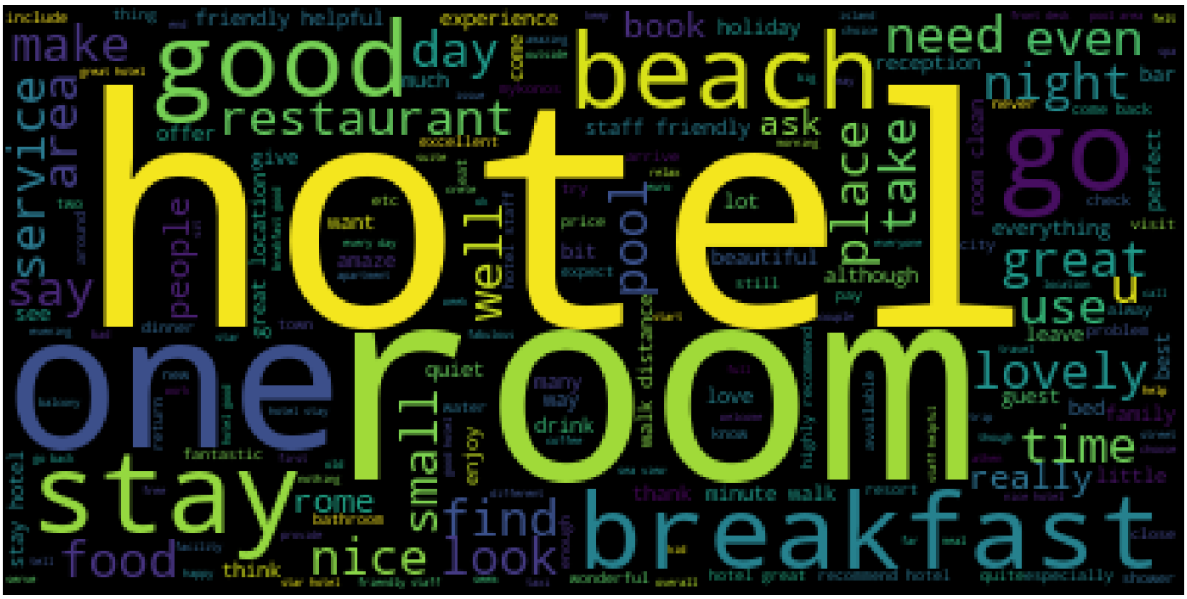
Είναι απαραίτητο να κατηγοριοποιήσουμε τις κριτικές μας που ανέφεραν οι πελάτες. Αυτό μπορεί να επιτευχθεί χρησιμοποιώντας το Word Cloud:

Τα νέφη λέξεων ή νέφη ετικετών είναι γραφικές αναπαραστάσεις της συχνότητας των λέξεων που δίνουν μεγαλύτερη έμφαση στις λέξεις που εμφανίζονται συχνότερα σε ένα πηγαίο κείμενο. Όσο μεγαλύτερη είναι η λέξη στην απεικόνιση, τόσο πιο συχνή ήταν η λέξη στο έγγραφο (ή στα έγγραφα). Αυτός ο τύπος οπτικοποίησης μπορεί να βοηθήσει τους αξιολογητές στη διερευνητική ανάλυση κειμένου, εντοπίζοντας λέξεις που εμφανίζονται συχνά σε ένα σύνολο συνεντεύξεων, εγγράφων ή άλλου κειμένου. Μπορεί επίσης να χρησιμοποιηθεί για την επικοινωνία των πιο σημαντικών σημείων ή θεμάτων στο στάδιο της υποβολής εκθέσεων.

Μια ποικιλία από γεννήτριες σύννεφων λέξεων και ετικετών είναι ελεύθερα διαθέσιμες στο διαδίκτυο και η διαδικασία δημιουργίας τους είναι απλή. Οι αξιολογητές μπορούν απλώς να εισάγουν κείμενο (για παράδειγμα, ένα σύνολο συνεντεύξεων) σε ένα πλαίσιο κειμένου και το εργαλείο δημιουργεί μια γραφική αναπαράσταση των λέξεων. Οι περισσότερες γεννήτριες σύννεφων λέξεων διαθέτουν χαρακτηριστικά που επιτρέπουν στους χρήστες να αλλάζουν τα χρώματα, τη γραμματοσειρά και να αποκλείουν κοινές ή παρόμοιες λέξεις.[13]

Παράδειγμα:

Ακολουθεί ένα παράδειγμα σύννεφου λέξεων που δημιουργήθηκε από κριτικές ξενοδοχείων:



Εικόνα 3.2: Σύννεφο λέξεων από κριτικές ξενοδοχείων.

Όσον αφορά τα Index set μας ή Index terms κάθε κατηγορίας μπορούμε να πάρουμε τις πιο γνωστές λέξεις που αντιπροσωπεύουν κάθε κατηγορία χωρίς την χρησιμοποίηση κάποιας συγκεκριμένης μεθόδου.

3.4 Ανίχνευση Κατηγοριών στα Διαδικτυακά Σχόλια και Ανάλυση Συναισθήματος

Το fuzzy string matching είναι η τεχνική εύρεσης συμβολοσειρών που ταιριάζουν με μια δεδομένη συμβολοσειρά μερικώς και όχι ακριβώς. Όταν ένας χρήστης γράφει λάθος μια λέξη ή εισάγει μια λέξη μερικώς, το ασαφές ταίριασμα συμβολοσειρών βοηθά στην εύρεση της σωστής λέξης - όπως βλέπουμε στις μηχανές αναζήτησης.

Ο αλγόριθμος πίσω από το fuzzy string matching δεν εξετάζει απλώς την ισοδυναμία δύο συμβολοσειρών, αλλά ποσοτικοποιεί το πόσο κοντά είναι δύο συμβολοσειρές μεταξύ τους. Αυτό γίνεται συνήθως με τη χρήση μιας μετρικής απόστασης που είναι γνωστή ως "απόσταση επεξεργασίας". Αυτή προσδιορίζει την εγγύτητα δύο συμβολοσειρών προσδιορίζοντας τις ελάχιστες αλλαγές που πρέπει να γίνουν για να μετατραπεί η μία συμβολοσειρά σε άλλη. Υπάρχουν διάφοροι τύποι αποστάσεων επεξεργασίας που μπορούν να χρησιμοποιηθούν, όπως η απόσταση Levenshtein, η απόσταση Hamming, η απόσταση Jaro κ.λπ.

Εμείς θα δούμε την απόσταση Levenshtein, επίσης γνωστή και ως απόσταση επεξεργασίας όπως είδαμε παραπάνω.

Η απόσταση Levenshtein:

Η απόσταση Levenshtein είναι μια μετρική που μετρά πόσο απέχουν δύο ακολουθίες λέξεων. Με άλλα λόγια, μετράει τον ελάχιστο αριθμό επεξεργασιών που πρέπει να κάνετε για να αλλάξετε μια ακολουθία λέξεων σε μια άλλη. Αυτές οι επεξεργασίες μπορεί να είναι εισαγωγές, διαγραφές ή αντικαταστάσεις. Η μετρική αυτή πήρε το όνομά της από τον Vladimir Levenshtein, ο οποίος την εξέτασε αρχικά το 1965.

Ο επίσημος ορισμός της απόστασης Levenshtein μεταξύ δύο συμβολοσειρών a και b μπορεί να θεωρηθεί ως εξής:

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Όπου 1(ai≠bj) δηλώνει 0 όταν a=b και 1 διαφορετικά. Είναι σημαντικό να σημειωθεί ότι οι σειρές στο παραπάνω ελάχιστο αντιστοιχούν σε μια διαγραφή, μια εισαγωγή και μια αντικατάσταση με αυτή τη σειρά.

Είναι επίσης δυνατό να υπολογιστεί ο λόγος ομοιότητας Levenshtein με βάση την απόσταση Levenshtein. Αυτό μπορεί να γίνει χρησιμοποιώντας τον ακόλουθο τύπο:

(|a|+|b|)-leva,b(i,j)|a|+|b|

όπου |a| και |b| είναι τα μήκη της ακολουθίας a και της ακολουθίας b αντίστοιχα.[14]

Τώρα ας ρίξουμε μια ματιά στην πιο χρησιμοποιούμενη βιβλιοθήκη για το ταίριασμα συμβολοσειρών - το πακέτο FuzzyWuzzy.

Επισκόπηση βιβλιοθήκης FuzzyWuzzy:

Πρόκειται για μια βιβλιοθήκη Python που αναπτύχθηκε αρχικά από την SeatGeek. Η βασική μέθοδος που χρησιμοποιείται εδώ είναι ο υπολογισμός της απόστασης Levenshtein μεταξύ δύο συμβολοσειρών.

Η βιβλιοθήκη μπορεί να εγκατασταθεί με τη χρήση του pip:

- pip install fuzzywuzzy
- pip-install python-Levenshtein

Αποκτώντας μια βαθύτερη κατανόηση σχετικά με τη μέθοδο υπολογισμού των ποσοστών ομοιότητας, ας δούμε τους διαφορετικούς τύπους αναλογιών Fuzz που εκτελούν την όλη διαδικασία.

Τύποι αναλογιών Fuzz και οι μηχανισμοί τους:

1. Ratio (Simple ratio)

Όταν έχετε ένα πολύ απλό σύνολο συμβολοσειρών που μοιάζουν σχεδόν μεταξύ, μπορείτε να χρησιμοποιήσετε την απλή αναλογία από το πακέτο FuzzyWuzzy.

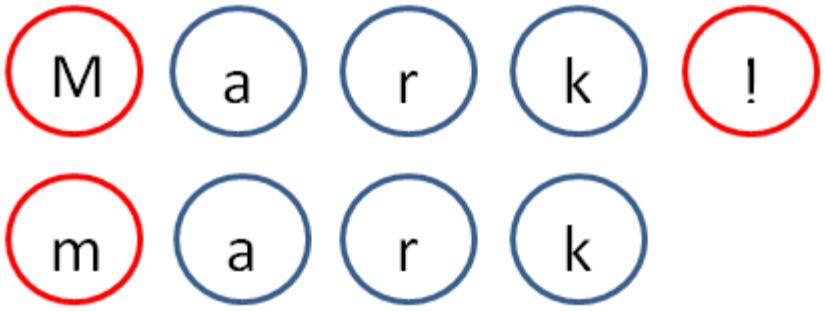
Ακολουθεί ο κώδικας για την κατανόηση της αντιστοίχισης με τη χρήση της απλής αναλογίας:

```
String1 = "Humpty Dumpty sat on a wall"
String2 = "Humpty Dumpty Sat on a Wall!

fuzz.ratio("Humpty Dumpty sat on a wall", "Humpty
Dumpty Sat on a Wall!")
>>> 91
```

Όπως φαίνεται στον παραπάνω κώδικα, η πρώτη συμβολοσειρά ταιριάζει με τη δεύτερη με ποσοστό 91%. Η διαφορά έγκειται στο θαυμαστικό που λείπει '!'.

Αυτή η αναλογία χρησιμοποιεί μια απλή τεχνική που περιλαμβάνει τον υπολογισμό της απόστασης επεξεργασίας (απόσταση Levenshtein) μεταξύ δύο συμβολοσειρών. Το μοναδικό χαρακτηριστικό του "fuzz.ratio" έγκειται στο γεγονός ότι λαμβάνει υπόψη τις ελάχιστες διαφορές που υπάρχουν μεταξύ των δύο συμβολοσειρών. Για παράδειγμα, αναγνωρίζει τα σημεία στίξης που λείπουν, τις λέξεις με διαφορετική πεζότητα, τις ανορθόγραφες λέξεις κ.λπ.



Εικόνα 3.3: Παράδειγμα Ratio

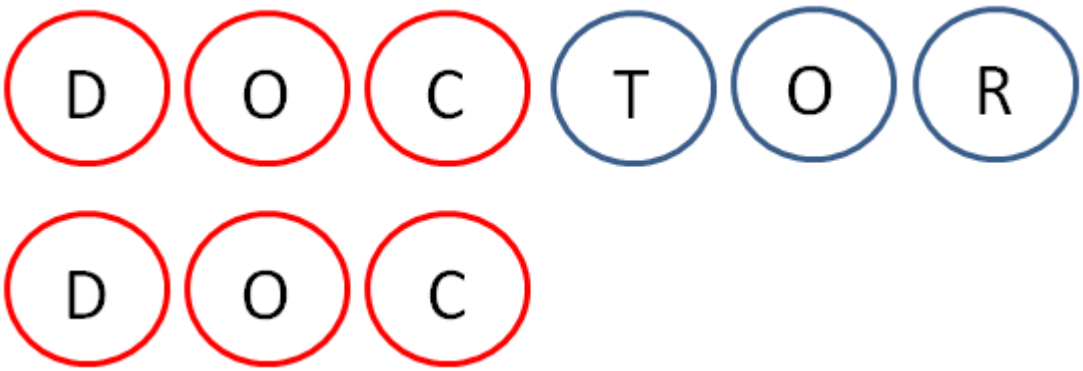
2. Partial Ratio

Τις περισσότερες φορές η απλή αναλογία δεν θα λειτουργήσει, καθώς είναι πολύ άκαμπτη στον εντοπισμό των αντιστοιχιών. Για παράδειγμα, όταν δεν θα θέλατε να λάβετε υπόψη όλες τις μικρές λεπτομέρειες, όπως τις λέξεις που σταματούν, τα σημεία στίξης, τα κεφαλαία γράμματα κ.λπ., είναι προτιμότερο να χρησιμοποιήσετε τη Μερική Αναλογία.

```
String1 = "Humpty Dumpty sat on a wall"
String2 = "Humpty"
fuzz.partial_ratio("Humpty Dumpty sat on a wall",
"Humpty")
>>> 100
```

Όπως παρατηρήθηκε, η μόνη κοινή λέξη μεταξύ των δύο σειρών ήταν το "Humpty", αλλά έδωσε 100% ταύτιση.

Όταν έχουμε να κάνουμε με υποσύνολα, δηλαδή με μια σύντομη συμβολοσειρά που αποτελεί μέρος κάποιας άλλης μακράς συμβολοσειράς, χρησιμοποιούμε τη συνάρτηση Partial Ratio. Ο μηχανισμός αυτής της αναλογίας ασχολείται με κάτι που είναι γνωστό ως "βέλτιστη μερική λογική". Για παράδειγμα, έστω ότι το μικρότερο μήκος συμβολοσειράς είναι 'm' και το μεγαλύτερο μήκος συμβολοσειράς είναι 'n'. Τότε, η μερική αναλογία βρίσκει ένα υποσύνολο μήκους 'm' που ταιριάζει καλύτερα.



Εικόνα 3.4: Παράδειγμα Partial Ratio

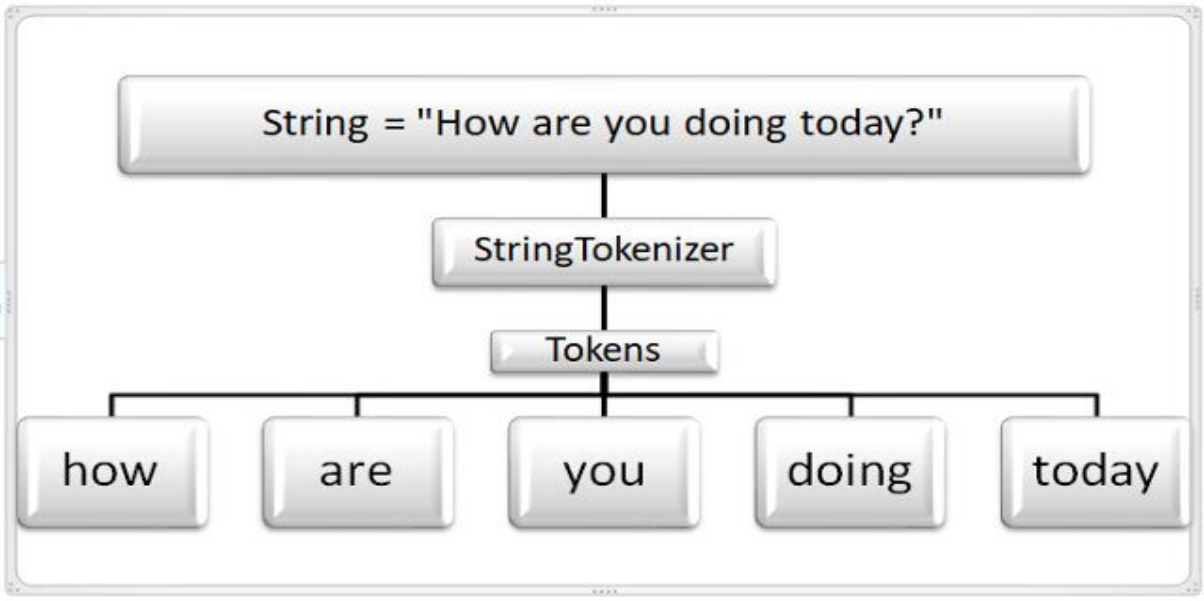
3. Token Set Ratio

Όταν δεν σας ενδιαφέρει ο αριθμός των φορών που επαναλαμβάνεται μια λέξη στη συμβολοσειρά, τότε είναι προτιμότερο να χρησιμοποιήσετε την αναλογία Token Set Ratio από το πακέτο.

```
String1 = "Humpty Dumpty sat on a wall"
String2 = "Humpty Humpty Dumpty sat on a wall"

fuzz.token_set_ratio("Humpty Dumpty sat on a wall",
"Humpty Humpty Dumpty sat on a wall")
>>> 100
```

Όπως φαίνεται από το παραπάνω παράδειγμα, οι συμβολοσειρές διαφέρουν απλώς ως προς τον αριθμό των φορών που χρησιμοποιείται η λέξη "Humpty". Επομένως, αν δεν σκοπεύουμε να λάβουμε υπόψη την επανάληψη, τότε το Token Set Ratio δίνει 100% ταύτιση.



Εικόνα 3.5: Παράδειγμα Token Set Ratio

Η ομοιότητα μεταξύ συγκεκριμένων συμβολοσειρών είναι ένα ακέραιο μέτρο (int) που κυμαίνεται από [0 100]. Η διαδικασία λήψης του ποσοστού ομοιότητας, περιλαμβάνει πρώτα τη διάσπαση των συμβολοσειρών σε μάρκες (ή λέξεις). Στη συνέχεια πραγματοποιείται η ταξινόμηση αυτών των tokens.

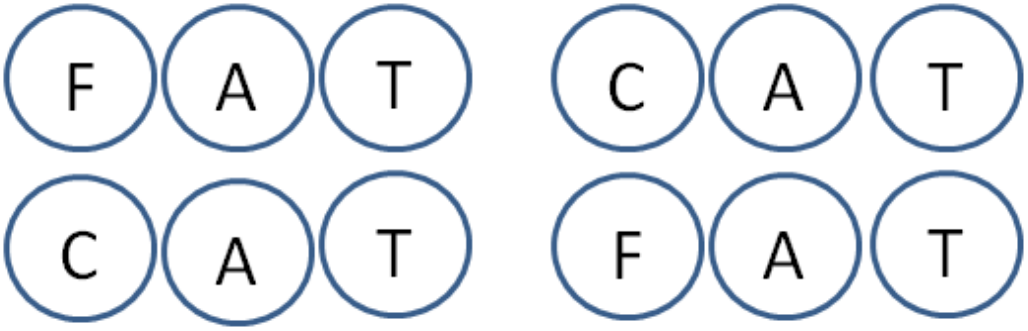
Το Token Set Ratio χωρίζεται σε δύο διαφορετικά σύνολα: Σύνολο τομής και σύνολο υπολειμμάτων. Επιπλέον, η συνάρτηση token αφαιρεί όλα τα σημεία στίξης εξαλείφοντας όλους τους μη αλφαβητικούς, μη αριθμητικούς χαρακτήρες. Τα πάντα μετατρέπονται σε πεζά. Η τελική αντιστοίχιση είναι το αποτέλεσμα των ομοιοτήτων που υπάρχουν μεταξύ των μετασχηματισμένων συμβολοσειρών (με τη μορφή συμβόλων).

4. Token Sort Ratio

Εάν η σειρά με την οποία οι λέξεις τοποθετούνται σε μια συγκεκριμένη πρόταση δεν έχει σημασία, τότε ο καλύτερος τρόπος για να ταιριάξετε δύο συμβολοσειρές είναι η χρήση της αναλογίας Token Sort Ratio από το πακέτο.

```
String1 = "Humpty Dumpty sat on a wall"
String2 = "Dumpty Humpty wall on sat a"
fuzz.token_sort_ratio("Humpty Dumpty sat on a
wall","Dumpty Humpty wall on sat a")
>>> 100
```

Όπως παρατηρείται από το παραπάνω παράδειγμα, οι συμβολοσειρές διέφεραν μόνο ως προς τη διάταξη των λέξεων. Ως εκ τούτου, η αναλογία ταξινόμησης Token έδωσε 100% ταύτιση.



Εικόνα 3.6: Παράδειγμα Token Sort Ratio

Αντί να συγκρίνει απευθείας τις συμβολοσειρές, το Token Sort Ratio χωρίζει επίσης τις συμβολοσειρές σε συμβολικά στοιχεία. Οι μάρκες (λέξεις) συγκρίνονται στη συνέχεια χρησιμοποιώντας τον απλό μηχανισμό αναλογίας. Ένα ενδιαφέρον χαρακτηριστικό αυτής της αναλογίας είναι ότι δεν λαμβάνει υπόψη τη σειρά των tokens που εμφανίζονται στις συμβολοσειρές. Ωστόσο, η αναλογία ταξινόμησης συμβόλων δεν είναι τόσο ευέλικτη όσο η αναλογία συνόλου συμβόλων όσον αφορά τις επαναλήψεις λέξεων.

Οι παραπάνω συμβολοσειρές "Fat Cat" και "Cat Fat" δίνουν 100% ταύτιση με το Token Sort Ratio.[15]

Συμπέρασμα:

Παρόλο που οι παραπάνω τέσσερις τύποι αναλογιών fuzz κάνουν τη ζωή μας ευκολότερη με την ανάκτηση των συμβολοσειρών που μοιάζουν μεταξύ τους, η διαδικασία περιπλέκεται όταν αποφασίζεται ο συγκεκριμένος τύπος των αναλογιών που πρέπει να χρησιμοποιηθεί για ένα δεδομένο σύνολο συμβολοσειρών.

Η χρήση ενός συγκεκριμένου τύπου fuzz ratio εξαρτάται από διάφορους παράγοντες όπως:

- Σύνολο τιμών που δίνονται για την αντιστοίχιση μιας συμβολοσειράς

- Υπόβαθρο του συνόλου δεδομένων που παρέχεται
- Χρήση των αποτελεσμάτων εξόδου
- Τύπος των συμβολοσειρών που αντιστοιχίζονται

3.5 Συσταδοποίηση (Clustering)

Η συσταδοποίηση είναι μια τεχνική μηχανικής μάθησης χωρίς επίβλεψη. Είναι η διαδικασία διαχωρισμού του συνόλου δεδομένων σε ομάδες στις οποίες τα μέλη της ίδιας ομάδας έχουν ομοιότητες στα χαρακτηριστικά. Οι αλγόριθμοι συσταδοποίησης που χρησιμοποιούνται συνήθως είναι οι K-Means, η ιεραρχική, η συσταδοποίηση με βάση την πυκνότητα, η συσταδοποίηση με βάση το μοντέλο κ.λπ.

Στην ανάλυση συστάδων, η μέθοδος του αγκώνα είναι μια ευρετική μέθοδος που χρησιμοποιείται για τον προσδιορισμό του αριθμού των συστάδων σε ένα σύνολο δεδομένων. Η μέθοδος συνίσταται στην απεικόνιση της εξηγούμενης διακύμανσης ως συνάρτηση του αριθμού των συστάδων και στην επιλογή του αγκώνα της καμπύλης ως τον αριθμό των συστάδων που πρέπει να χρησιμοποιηθούν.

K-means αλγόριθμος:

Ο αλγόριθμος K-Means δεν χρειάζεται συστάσεις. Είναι απλός και ίσως ο πιο συχνά χρησιμοποιούμενος αλγόριθμος ομαδοποίησης.

Η βασική ιδέα πίσω από τον αλγόριθμο k-means συνίσταται στον ορισμό k συστάδων έτσι ώστε η συνολική διακύμανση εντός της συστάδας (ή το σφάλμα) να είναι ελάχιστη.

Ένα κέντρο συστάδας είναι ο εκπρόσωπος της συστάδας του. Η τετραγωνική απόσταση μεταξύ κάθε σημείου και του κέντρου της συστάδας του είναι η απαιτούμενη διακύμανση. Στόχος της συσταδοποίησης k-means είναι η εύρεση αυτών των k συστάδων και των κέντρων τους με ταυτόχρονη μείωση του συνολικού σφάλματος.[16]

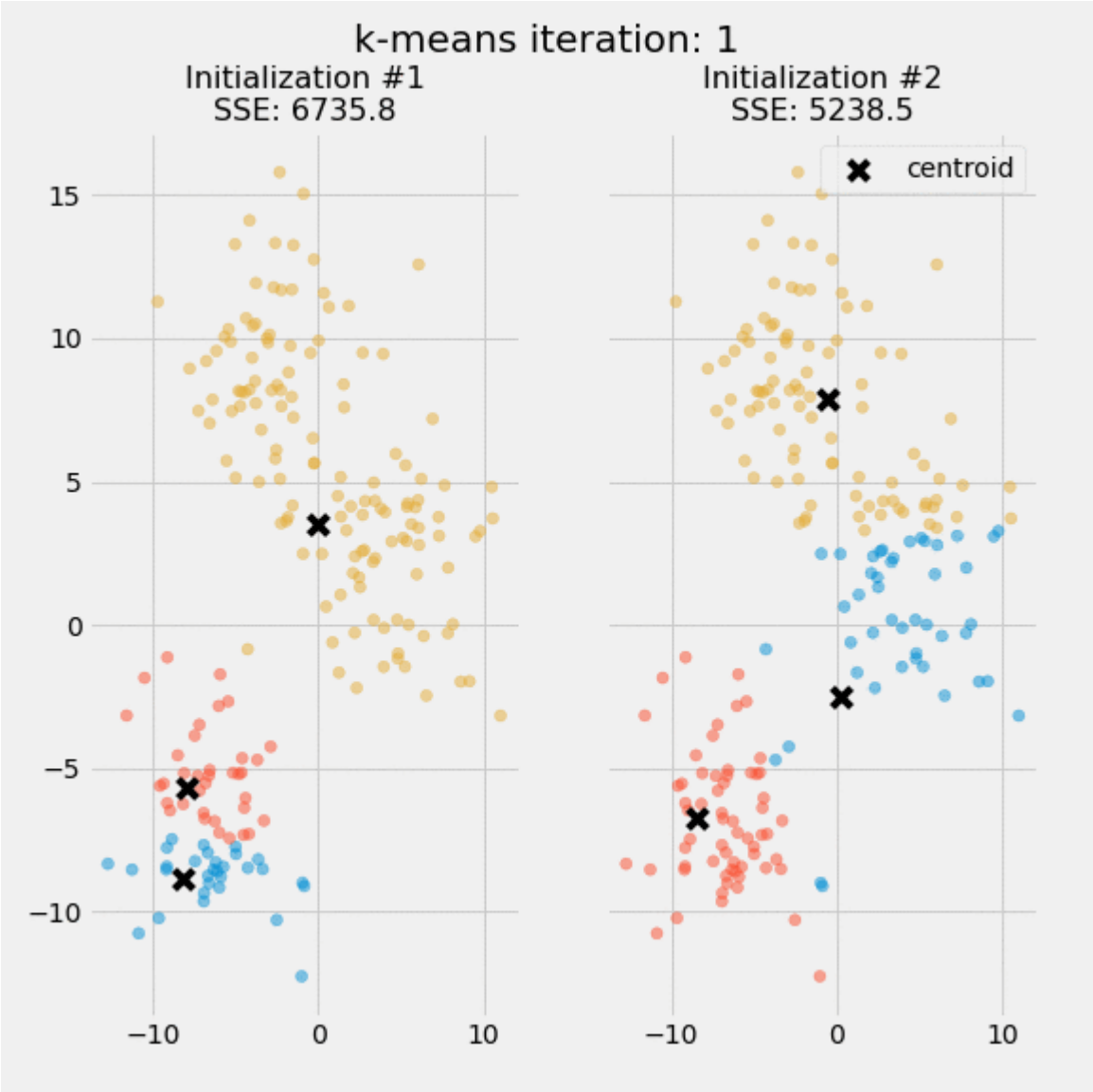
Το κύριο στοιχείο του αλγορίθμου λειτουργεί με μια διαδικασία δύο βημάτων που ονομάζεται μεγιστοποίηση της προσδοκίας. Το βήμα της προσδοκίας αναθέτει κάθε σημείο δεδομένων στο πλησιέστερο κεντροειδές του. Στη συνέχεια, το βήμα μεγιστοποίησης υπολογίζει τον μέσο όρο όλων των σημείων για κάθε συστάδα και ορίζει το νέο κεντροειδές. Ακολουθεί η συμβατική έκδοση του αλγορίθμου k-means:

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

Η ποιότητα των αναθέσεων των συστάδων καθορίζεται με τον υπολογισμό του αθροίσματος του τετραγωνικού σφάλματος (SSE) μετά τη σύγκλιση των κεντροειδών ή την αντιστοίχιση με την ανάθεση της προηγούμενης επανάληψης. Το SSE ορίζεται ως το άθροισμα των τετραγωνικών ευκλείδειων αποστάσεων κάθε σημείου από το πλησιέστερο κεντροειδές του. Δεδομένου ότι πρόκειται για ένα μέτρο σφάλματος, ο στόχος του k-means είναι να προσπαθήσει να ελαχιστοποιήσει αυτή την τιμή.

Το παρακάτω σχήμα δείχνει τα κεντροειδή και την ενημέρωση του SSE κατά τις πρώτες πέντε επαναλήψεις από δύο διαφορετικές εκτελέσεις του αλγορίθμου k-means στο ίδιο σύνολο δεδομένων:



Εικόνα 3.7: Αρχικοποίηση των κεντροειδών και η χρήση της SSE

Σκοπός αυτού του σχήματος είναι να δείξει ότι η αρχικοποίηση των κεντροειδών είναι ένα σημαντικό βήμα. Επισημαίνει επίσης τη χρήση του SSE ως μέτρο της απόδοσης της ομαδοποίησης. Μετά την επιλογή ενός αριθμού συστάδων και των αρχικών κεντροειδών, το βήμα μεγιστοποίησης προσδοκίας επαναλαμβάνεται μέχρι οι θέσεις των κεντροειδών να φτάσουν σε σύγκλιση και να παραμείνουν αμετάβλητες.[17]

Πολύ κομψός αλγόριθμος. Αλλά υπάρχει μια παγίδα. Πώς αποφασίζετε τον αριθμό των συστάδων;

Θα εξηγήσω λεπτομερώς μία από τις δύο μεθόδους που μπορούν να είναι χρήσιμες για την εύρεση αυτού του μυστηριώδους k στο k-Means.

Αυτές οι μέθοδοι είναι:

- Η μέθοδος του αγκώνα
- Η μέθοδος της σιλουέτας

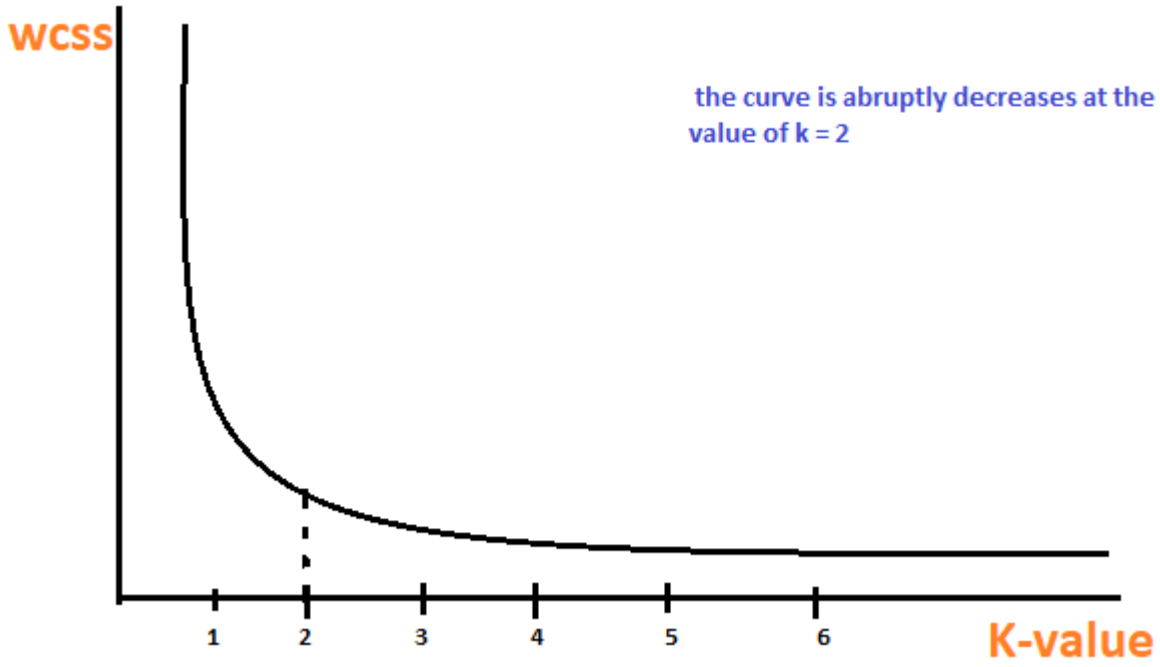
Η μέθοδος του αγκώνα:

Η μέθοδος του αγκώνα είναι μια από τις πιο διάσημες μεθόδους με τις οποίες μπορείτε να επιλέξετε τη σωστή τιμή του k και να ενισχύσετε την απόδοση του μοντέλου σας. Ας δούμε πώς λειτουργεί αυτή η μέθοδος του αγκώνα.

Πρόκειται για μια εμπειρική μέθοδο για την εύρεση της καλύτερης τιμής του k. Διαλέγει το εύρος των τιμών και παίρνει την καλύτερη από αυτές. Υπολογίζει το άθροισμα των τετραγώνων των σημείων και υπολογίζει τη μέση απόσταση.

Όταν η τιμή του k είναι 1, το άθροισμα του τετραγώνου εντός της συστάδας(wcss) θα είναι υψηλό. Καθώς η τιμή του k αυξάνεται, το άθροισμα του τετραγώνου εντός της συστάδας θα μειώνεται.

Τέλος, δούμε ένα γράφημα μεταξύ των τιμών k και του αθροίσματος των τετραγώνων εντός της συστάδας για να πάρουμε την τιμή k. Θα εξετάσουμε προσεκτικά το γράφημα. Σε κάποιο σημείο, το γράφημά μας θα μειωθεί απότομα. Αυτό το σημείο θα θεωρηθεί ως τιμή του k.



Εικόνα 3.8: Καμπύλη εύρεσης της τιμής του k.

Στην παραπάνω εικόνα φαίνεται πολύ καθαρά ότι το σημείο στο οποίο η καμπύλη μειώνεται απότομα είναι στην τιμή 2.

Πλεονεκτήματα του K-means:

- Είναι πολύ απλό στην εφαρμογή του.
- Είναι επεκτάσιμο σε ένα τεράστιο σύνολο δεδομένων και επίσης ταχύτερο σε μεγάλα σύνολα δεδομένων.
- Προσαρμόζει τα νέα παραδείγματα πολύ συχνά.
- Γενίκευση των συστάδων για διαφορετικά σχήματα και μεγέθη.

Μειονεκτήματα του K-means:

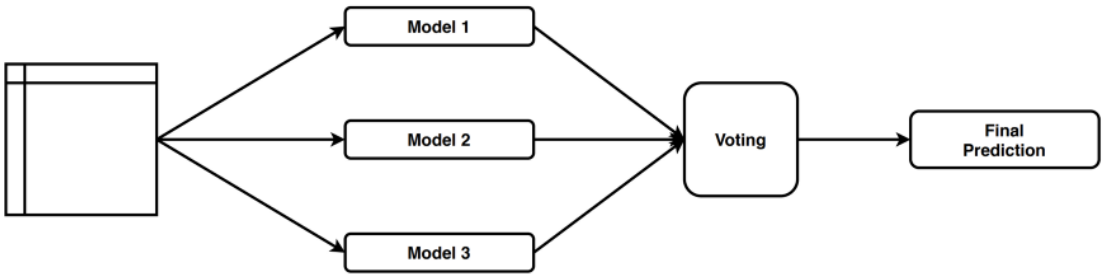
- Είναι ευαίσθητο στις ακραίες τιμές.
- Η επιλογή των τιμών k με το χέρι είναι μια δύσκολη εργασία.
- Καθώς αυξάνεται ο αριθμός των διαστάσεων μειώνεται η επεκτασιμότητά του.[18]

3.6 Μάθηση Συνόλων Ταξινομητών (Ensemble of classifiers)

Γενικά, τα μοντέλα συνόλου συνδυάζουν πολλαπλά μοντέλα βάσης για να βελτιώσουν την απόδοση πρόβλεψης. Το πιο γνωστό παράδειγμα ενός ensemble μοντέλου είναι το Random Forest, το οποίο - απλουστεύοντας σημαντικά τη λογική του αλγορίθμου - συνδυάζει πολλαπλά δέντρα απόφασης και συγκεντρώνει τις προβλέψεις τους με τη χρήση πλειοψηφικής ψήφου σε περίπτωση προβλήματος ταξινόμησης ή με τη λήψη του μέσου όρου για εργασίες παλινδρόμησης.

Ομοίως με το Random Forest, το Voting Ensemble εκτιμά πολλαπλά μοντέλα βάσης και χρησιμοποιεί ψηφοφορία για να συνδυάσει τις επιμέρους προβλέψεις ώστε να καταλήξει στις τελικές. Ωστόσο, η βασική διαφορά έγκειται στους εκτιμητές βάσης. Μοντέλα όπως το Voting Ensemble (και το Stacking Ensemble) δεν απαιτούν τα μοντέλα βάσης να είναι ομοιογενή. Με άλλα λόγια, μπορούμε να εκπαιδεύσουμε διαφορετικούς εκτιμητές βάσης, για παράδειγμα, ένα Δέντρο Αποφάσεων και μια Λογιστική Παλινδρόμηση, και στη συνέχεια να χρησιμοποιήσουμε το Voting Ensemble για να συνδυάσουμε τα αποτελέσματα.

Το ακόλουθο διάγραμμα παρουσιάζει τη ροή εργασιών του Συνόλου ψηφοφορίας:



Εικόνα 3.9: Ροή εργασιών ενός Voting Ensemble.

Ο ταξινομητής ψηφοφορίας υποστηρίζει δύο τύπους ψηφοφορίας:

- Hard: Η τελική πρόβλεψη κλάσης γίνεται με ψηφοφορία πλειοψηφίας - ο εκτιμητής επιλέγει την πρόβλεψη κλάσης που εμφανίζεται συχνότερα μεταξύ των βασικών μοντέλων.

- Soft: Η τελική πρόβλεψη κλάσης γίνεται με βάση τη μέση πιθανότητα που υπολογίζεται με τη χρήση όλων των προβλέψεων των βασικών μοντέλων. Για παράδειγμα, εάν το μοντέλο 1 προβλέπει τη θετική κλάση με πιθανότητα 70%, το μοντέλο 2 με πιθανότητα 90%, τότε το Voting Ensemble θα υπολογίσει ότι υπάρχει 80% πιθανότητα η παρατήρηση να ανήκει στη θετική κλάση και θα επιλέξει τη θετική κλάση ως πρόβλεψη. Επιπλέον, μπορούμε να χρησιμοποιήσουμε προσαρμοσμένα βάρη για τον υπολογισμό του σταθμισμένου μέσου όρου. Αυτό ενδείκνυται για περιπτώσεις στις οποίες εμπιστευόμαστε περισσότερο ορισμένα μοντέλα, αλλά εξακολουθούμε να θέλουμε να λάβουμε υπόψη αυτά που εμπιστευόμαστε λιγότερο.

Ένα πράγμα που πρέπει να έχουμε υπόψη μας είναι ότι για να χρησιμοποιήσουμε τη μέθοδο soft voting, όλα τα βασικά μοντέλα πρέπει να διαθέτουν τη μέθοδο predict_proba. Η soft ψηφοφορία μπορεί να οδηγήσει σε καλύτερες επιδόσεις από τη hard ψηφοφορία (αλλά όχι απαραίτητα), καθώς με τον μέσο όρο των πιθανοτήτων "δίνει μεγαλύτερη βαρύτητα" στις σίγουρες ψήφους.

Το Voting Ensemble είναι μια χρήσιμη τεχνική, η οποία είναι ιδιαίτερα χρήσιμη όταν ένα μεμονωμένο μοντέλο παρουσιάζει κάποια προκατάληψη. Είναι επίσης πιθανό το Voting Ensemble να καταλήξει σε καλύτερη συνολική βαθμολογία από την καλύτερη από τις βασικές εκτιμήτριες, καθώς συγκεντρώνει τις προβλέψεις πολλαπλών μοντέλων και προσπαθεί να καλύψει τις πιθανές αδυναμίες των μεμονωμένων μοντέλων. Ένας τρόπος για να βελτιωθεί η απόδοση του συνόλου είναι να γίνουν οι εκτιμητές βάσης όσο το δυνατόν πιο διαφορετικοί.[19]

Οι πιο συχνοί ταξινομητές που χρησιμοποιούνται για την δημιουργία ενός Ensembler είναι:

1. Naïve Bayes Bernoulli
2. Random Forest
3. Logistic Regression
4. SVC(Support Vector Classifier)

1. Naïve Bayes Bernoulli:

Το Bernoulli Naive Bayes είναι μέρος της οικογένειας Naive Bayes. Δέχεται μόνο δυαδικές τιμές. Το πιο γενικό παράδειγμα είναι όπου ελέγχουμε αν κάθε τιμή θα είναι ή όχι μια λέξη

που εμφανίζεται σε ένα έγγραφο. Αυτό είναι ένα πολύ απλουστευμένο μοντέλο. Σε περιπτώσεις όπου η καταμέτρηση της συχνότητας των λέξεων είναι λιγότερο σημαντική, το Bernoulli μπορεί να δώσει καλύτερα αποτελέσματα. Με απλά λόγια, πρέπει να μετράμε σε κάθε τιμή δυαδικά χαρακτηριστικά εμφάνισης όρων, δηλαδή μια λέξη εμφανίζεται σε ένα έγγραφο ή όχι. Αυτά τα χαρακτηριστικά χρησιμοποιούνται αντί για την εύρεση της συχνότητας μιας λέξης στο έγγραφο.

Για να το κατανοήσουμε με απλά λόγια, η κατανομή Bernoulli έχει δύο αμοιβαία αποκλειόμενα αποτελέσματα: P(X=1)=p ή P(X=0)=1-p. Στο θεώρημα BernoulliNB, μπορούμε να έχουμε πολλαπλά χαρακτηριστικά, αλλά κάθε ένα θεωρείται ότι είναι μεταβλητή με δυαδική τιμή, δηλαδή boolean. Επομένως, αυτή η κατηγορία απαιτεί τα δείγματα να αναπαριστώνται ως διανύσματα χαρακτηριστικών με δυαδική τιμή. Σε περίπτωση που παρέχεται οποιοδήποτε άλλο είδος δεδομένων, τότε ένα παράδειγμα BernoulliNB μπορεί να δυαδικοποιήσει την είσοδό του.

Ο κανόνας απόφασης για το Naive Bernoulli Bayes βασίζεται σε:

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

Σύμφωνα με τον τύπο του κανόνα απόφασης, το x πρέπει να είναι δυαδικό. Σκεφτείτε τον τύπο στην περίπτωση όπου xi=1 και στην περίπτωση όπου xi=0. Έτσι i είναι το γεγονός όπου xi=1 ή το γεγονός όπου xi=0.

Τα πλεονεκτήματα του BernoulliNB περιλαμβάνουν καλύτερες επιδόσεις σε ορισμένα σύνολα δεδομένων, ιδίως σε εκείνα με μικρότερα έγγραφα.

Είναι καλύτερο όταν εφαρμόζεται σε σενάρια πραγματικής ζωής όπου απαιτούνται άμεσα (γρήγορα) αποτελέσματα. Επίσης, χρησιμοποιείται συχνότερα για να επιτύχει καλύτερα αποτελέσματα σε προβλήματα πολλαπλών κλάσεων και κανόνων ανεξαρτησίας. Έτσι, έχει υψηλότερο ποσοστό επιτυχίας από άλλους αλγορίθμους. Σε περίπτωση μικρού όγκου δεδομένων ή μικρών εγγράφων(για παράδειγμα στην ταξινόμηση κειμένου), ο Bernoulli Naive Bayes δίνει πιο ακριβή και ακριβή αποτελέσματα σε σύγκριση με άλλα μοντέλα.[20]

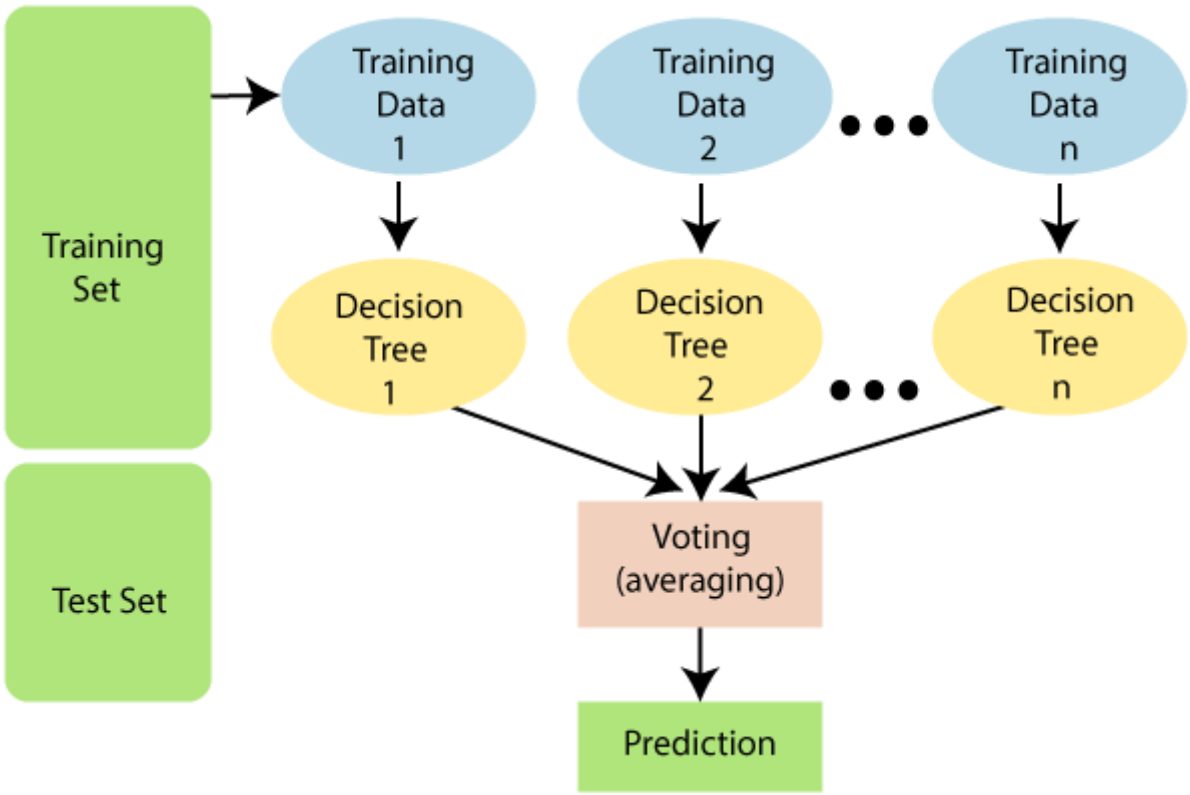
2. Random Forest:

Το Random Forest είναι ένας δημοφιλής αλγόριθμος μηχανικής μάθησης που ανήκει στην τεχνική μάθησης με επίβλεψη. Μπορεί να χρησιμοποιηθεί τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα παλινδρόμησης στην ML. Βασίζεται στην έννοια της μάθησης συνόλου, η οποία είναι μια διαδικασία συνδυασμού πολλαπλών ταξινομητών για την επίλυση ενός σύνθετου προβλήματος και τη βελτίωση της απόδοσης του μοντέλου.

Όπως υποδηλώνει το όνομα, "το Random Forest είναι ένας ταξινομητής που περιέχει έναν αριθμό δέντρων απόφασης σε διάφορα υποσύνολα του δεδομένου συνόλου δεδομένων και λαμβάνει το μέσο όρο για να βελτιώσει την ακρίβεια πρόβλεψης του εν λόγω συνόλου δεδομένων". Αντί να βασίζεται σε ένα δέντρο απόφασης, το random forest λαμβάνει την πρόβλεψη από κάθε δέντρο και με βάση τις πλειοψηφικές ψήφους των προβλέψεων, και προβλέπει την τελική έξοδο.

Ο μεγαλύτερος αριθμός δέντρων στο forest οδηγεί σε μεγαλύτερη ακρίβεια και αποτρέπει το πρόβλημα της υπερπροσαρμογής.

Το παρακάτω διάγραμμα εξηγεί τη λειτουργία του αλγορίθμου Random Forest:



Εικόνα 3.10: Λειτουργία αλγορίθμου Random Forest.

Πώς λειτουργεί ο αλγόριθμος Random Forest;

Το random forest λειτουργεί σε δύο φάσεις: η πρώτη είναι η δημιουργία του τυχαίου δάσους συνδυάζοντας Ν δέντρα αποφάσεων και η δεύτερη είναι η πραγματοποίηση προβλέψεων για κάθε δέντρο που δημιουργήθηκε στην πρώτη φάση.

Η διαδικασία εργασίας μπορεί να εξηγηθεί στα παρακάτω βήματα:

Βήμα-1: Επιλέξτε τυχαία Κ σημεία δεδομένων από το σύνολο εκπαίδευσης.

Βήμα-2: Δημιουργία των δέντρων απόφασης που σχετίζονται με τα επιλεγμένα σημεία δεδομένων (υποσύνολα).

Βήμα-3: Επιλέξτε τον αριθμό Ν για τα δέντρα απόφασης που θέλετε να δημιουργήσετε.

Βήμα-4: Επαναλάβετε τα βήματα 1 & 2.

Βήμα-5: Για νέα σημεία δεδομένων, βρείτε τις προβλέψεις κάθε δέντρου απόφασης και αναθέστε τα νέα σημεία δεδομένων στην κατηγορία που κερδίζει τις περισσότερες ψήφους.

Πλεονεκτήματα του Random Forest:

- Το Random Forest είναι ικανό να εκτελεί τόσο εργασίες ταξινόμησης όσο και παλινδρόμησης.
- Είναι ικανό να χειρίζεται μεγάλα σύνολα δεδομένων με υψηλή διαστατικότητα.
- Ενισχύει την ακρίβεια του μοντέλου και αποτρέπει το ζήτημα της υπερπροσαρμογής.

Μειονεκτήματα του Random Forest:

- Παρόλο που το random forest μπορεί να χρησιμοποιηθεί τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης, δεν είναι και τόσο κατάλληλο για εργασίες παλινδρόμησης.[21]

3. Logistic Regression:

Η λογιστική παλινδρόμηση είναι ένας αλγόριθμος "επιβλεπόμενης μηχανικής μάθησης" που μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση της πιθανότητας μιας συγκεκριμένης κατηγορίας ή ενός γεγονότος. Χρησιμοποιείται όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα και το αποτέλεσμα είναι δυαδικό ή διχοτομικό στη φύση του.

Αυτό σημαίνει ότι η λογιστική παλινδρόμηση χρησιμοποιείται συνήθως για προβλήματα δυαδικής ταξινόμησης.

Η δυαδική ταξινόμηση αναφέρεται στην πρόβλεψη της μεταβλητής εξόδου που είναι διακριτή σε δύο κλάσεις.

Μερικά παραδείγματα δυαδικής ταξινόμησης είναι τα Ναι/Όχι, Πέρασμα/Αποτυχία, Κέρδος/Χάσιμο, Καρκινογόνος/Μη καρκινογόνος κ.λπ.

Εξίσωση γραμμικής παλινδρόμησης:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

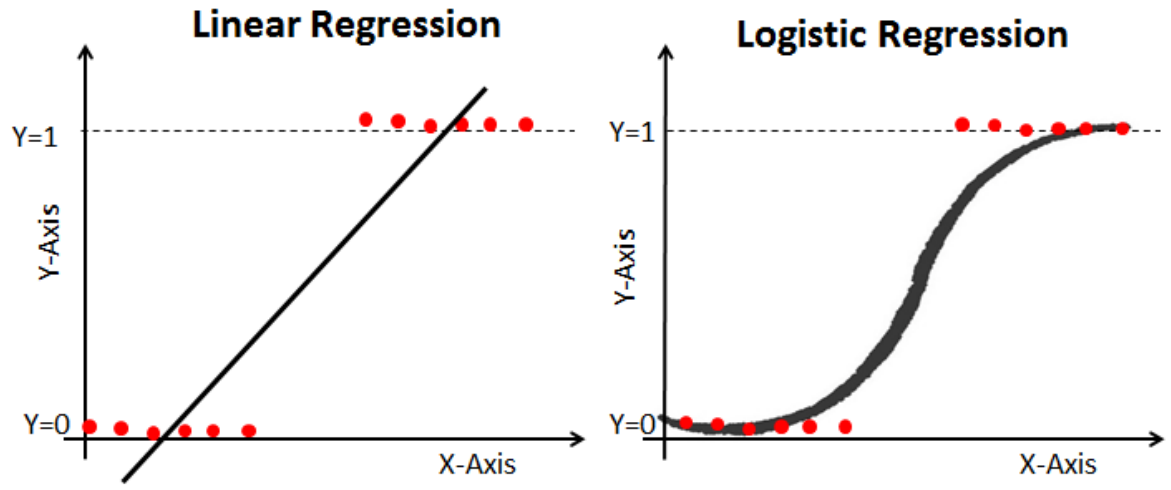
Όπου, y είναι η εξαρτημένη μεταβλητή και x1, x2 ... και Xn είναι επεξηγηματικές μεταβλητές.

Ιδιότητες της λογιστικής παλινδρόμησης:

- Η εξαρτημένη μεταβλητή στη λογιστική παλινδρόμηση ακολουθεί την κατανομή Bernoulli.
- Η εκτίμηση γίνεται μέσω της μέγιστης πιθανοφάνειας.
- Δεν υπάρχει τετράγωνο R, η καταλληλότητα του μοντέλου υπολογίζεται μέσω των Concordance, KS-Statistics.

Γραμμική παλινδρόμηση έναντι Λογιστικής παλινδρόμησης:

Η γραμμική παλινδρόμηση μας δίνει μια συνεχή έξοδο, αλλά η λογιστική παλινδρόμηση παρέχει μια σταθερή έξοδο. Ένα παράδειγμα συνεχούς εξόδου είναι η τιμή του σπιτιού και η τιμή της μετοχής. Παράδειγμα της διακριτής εξόδου είναι η πρόβλεψη του αν ένας ασθενής έχει καρκίνο ή όχι. Η γραμμική παλινδρόμηση εκτιμάται με τη χρήση της μεθόδου Ordinary Least Squares (OLS), ενώ η λογιστική παλινδρόμηση εκτιμάται με τη χρήση της Maximum Likelihood Estimation (MLE).



Εικόνα 3.11: OLS και MLE μεθόδους.

Maximum Likelihood Estimation - Least Square Method:

Η MLE είναι μια μέθοδος μεγιστοποίησης της "πιθανότητας", ενώ η OLS είναι μια μέθοδος προσέγγισης που ελαχιστοποιεί την απόσταση. Η μεγιστοποίηση της συνάρτησης πιθανότητας προσδιορίζει τις παραμέτρους που είναι πιθανότερο να παράγουν τα παρατηρούμενα δεδομένα. Από στατιστική άποψη, η MLE θέτει τη μέση τιμή και τη διακύμανση ως παραμέτρους στον προσδιορισμό των συγκεκριμένων παραμετρικών τιμών για ένα δεδομένο μοντέλο. Αυτό το σύνολο παραμέτρων μπορεί να χρησιμοποιηθεί για την πρόβλεψη των δεδομένων που απαιτούνται σε μια κανονική κατανομή.

Οι OLS υπολογίζονται με την προσαρμογή μιας γραμμής παλινδρόμησης στα δεδομένα που έχει το ελάχιστο άθροισμα των τετραγωνικών αποκλίσεων (ελάχιστο τετραγωνικό σφάλμα). Και οι δύο χρησιμοποιούνται για την εκτίμηση των παραμέτρων ενός μοντέλου γραμμικής παλινδρόμησης. Η MLE προϋποθέτει μια κοινή συνάρτηση μάζας πιθανότητας, ενώ η OLS δεν απαιτεί στοχαστικές υποθέσεις για την ελαχιστοποίηση της απόστασης.[22]

4. SVC(Support Vector Classifier):

Σκεφτείτε τους αλγορίθμους μηχανικής μάθησης ως ένα οπλοστάσιο γεμάτο με τσεκούρια, σπαθιά, λεπίδες, τόξα, στιλέτα κ.λπ. Έχετε διάφορα εργαλεία, αλλά οφείλετε να μάθετε να τα χρησιμοποιείτε τη σωστή στιγμή. Ως αναλογία, σκεφτείτε την "παλινδρόμηση" ως ένα σπαθί ικανό να τεμαχίζει και να κόβει δεδομένα αποτελεσματικά, αλλά ανίκανο να χειριστεί εξαιρετικά πολύπλοκα δεδομένα. Αντίθετα, οι "Μηχανές διανυσμάτων

υποστήριξης" είναι σαν ένα κοφτερό μαχαίρι - λειτουργεί σε μικρότερα σύνολα δεδομένων, αλλά σε πολύπλοκα, μπορεί να είναι πολύ ισχυρότερη και ισχυρότερη στην κατασκευή μοντέλων μηχανικής μάθησης.

Τι είναι η μηχανή διανυσμάτων υποστήριξης(SVM);

Το "Support Vector Machine" (SVM) είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη που μπορεί να χρησιμοποιηθεί για προκλήσεις ταξινόμησης ή παλινδρόμησης. Ωστόσο, χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης. Στον αλγόριθμο SVM, απεικονίζουμε κάθε στοιχείο δεδομένων ως ένα σημείο στον n-διάστατο χώρο (όπου n είναι ο αριθμός των χαρακτηριστικών που έχετε) με την τιμή κάθε χαρακτηριστικού να είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Στη συνέχεια, πραγματοποιούμε ταξινόμηση βρίσκοντας το υπερεπίπεδο που διαφοροποιεί πολύ καλά τις δύο κλάσεις.

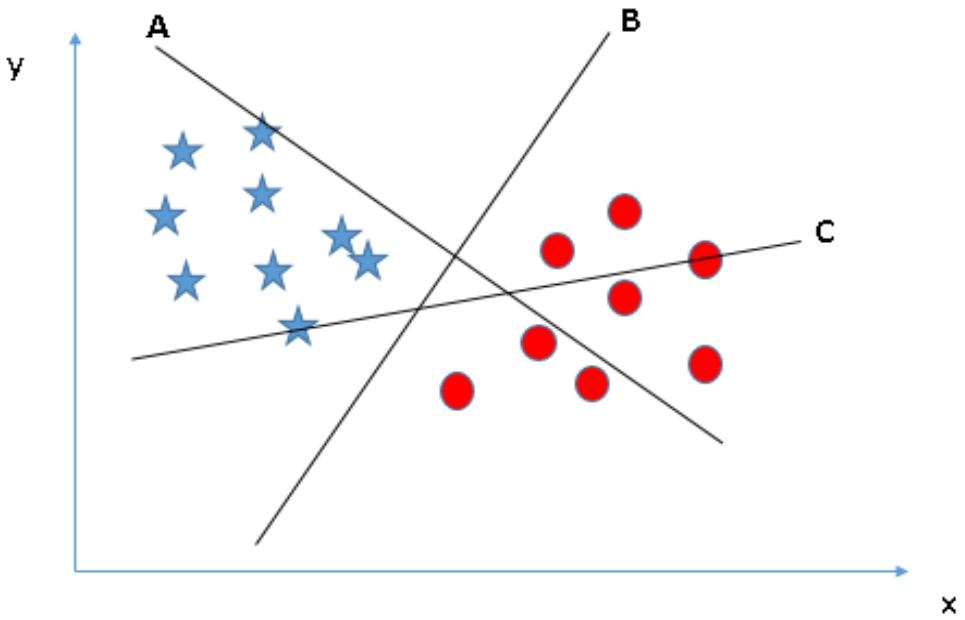
Τα Support Vectors είναι απλώς οι συντεταγμένες των μεμονωμένων παρατηρήσεων. Ο ταξινομητής SVM είναι ένα σύνορο που διαχωρίζει καλύτερα τις δύο κλάσεις (υπερεπίπεδο/γραμμή).

Λειτουργία SVM:

Παραπάνω, συνηθίσαμε στη διαδικασία διαχωρισμού των δύο κλάσεων με ένα υπερεπίπεδο. Τώρα το ερώτημα είναι "Πώς μπορούμε να προσδιορίσουμε το σωστό υπερεπίπεδο;".

Ας καταλάβουμε:

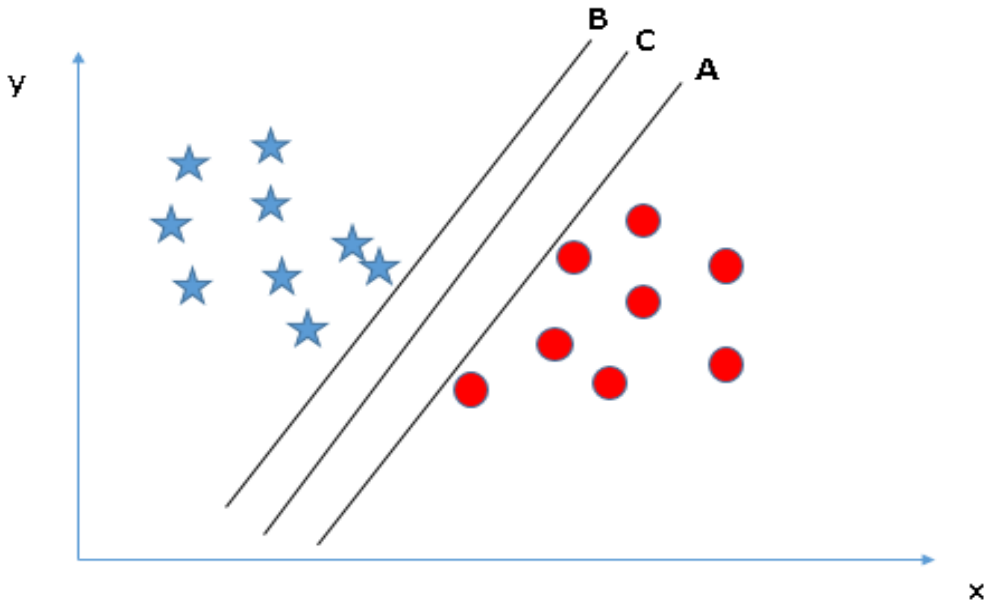
- Προσδιορισμός του σωστού υπερεπιπέδου (Σενάριο-1): Εδώ, έχουμε τρία υπερεπίπεδα (A, B και C). Τώρα, προσδιορίστε το σωστό υπερεπίπεδο για να ταξινομήσετε τα αστέρια και τους κύκλους.



Εικόνα 3.12: Προσδιορισμός σωστού υπερεπιπέδου. Παράδειγμα 1.

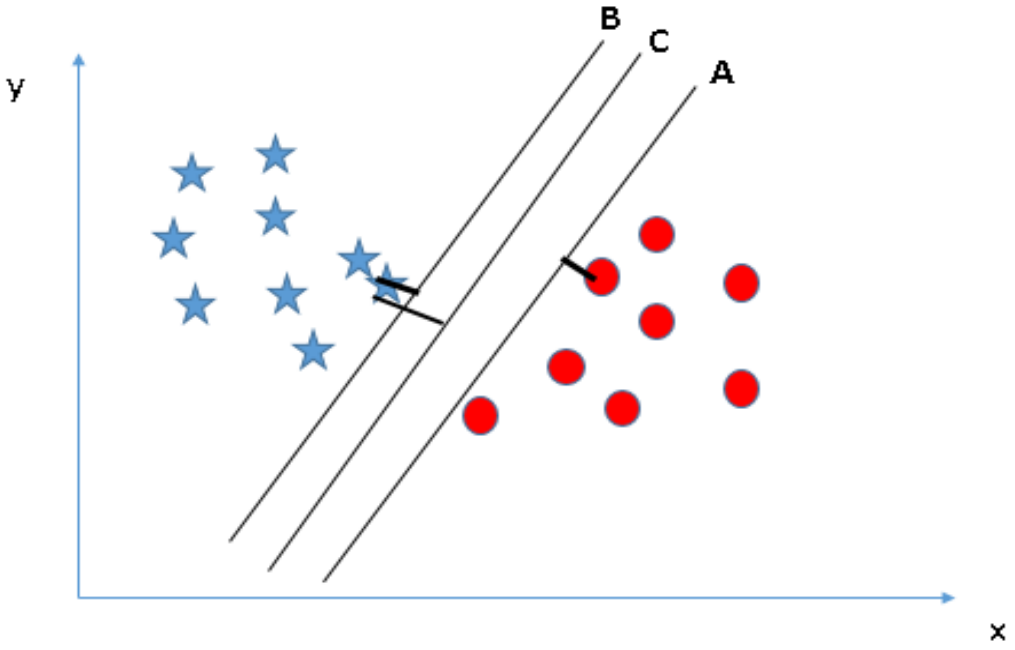
Πρέπει να θυμάστε έναν κανόνα για να εντοπίσετε το σωστό υπερεπίπεδο: "Επιλέξτε το υπερεπίπεδο που διαχωρίζει καλύτερα τις δύο κλάσεις". Σε αυτό το σενάριο, το υπερεπίπεδο "B" έχει επιτελέσει άριστα αυτή τη δουλειά.

- Προσδιορίστε το σωστό υπερεπίπεδο (Σενάριο-2): Εδώ, έχουμε τρία υπερεπίπεδα (A, B και C) και όλα διαχωρίζουν καλά τις κλάσεις. Τώρα, πώς μπορούμε να προσδιορίσουμε το σωστό υπερεπίπεδο;



Εικόνα 3.13: Προσδιορισμός σωστού υπερεπιπέδου. Παράδειγμα 2-1.

Εδώ, η μεγιστοποίηση των αποστάσεων μεταξύ του πλησιέστερου σημείου δεδομένων (είτε κλάση) και του υπερεπιπέδου θα μας βοηθήσει να αποφασίσουμε το σωστό υπερεπίπεδο. Αυτή η απόσταση ονομάζεται περιθώριο. Ας δούμε το παρακάτω στιγμιότυπο:



Εικόνα 3.14: Προσδιορισμός σωστού υπερεπιπέδου. Παράδειγμα 2-2.

Παραπάνω, μπορείτε να δείτε ότι το περιθώριο για το υπερεπίπεδο C είναι υψηλό σε σύγκριση με τα δύο A και B. Ως εκ τούτου, ονομάζουμε το σωστό υπερεπίπεδο ως C. Ένας άλλος λόγος για την επιλογή του υπερεπιπέδου με υψηλότερο περιθώριο είναι η ανθεκτικότητα. Εάν επιλέξουμε ένα υπερεπίπεδο με χαμηλό περιθώριο, τότε υπάρχει μεγάλη πιθανότητα λανθασμένης ταξινόμησης.[23]

ΚΕΦΑΛΑΙΟ 4

4.1 Η σημασία της Ανάλυσης Συναισθήματος στις Ξενοδοχειακές Επιχειρήσεις

Κάθε κριτική που δημοσιεύουν οι επισκέπτες σε οποιαδήποτε πλατφόρμα εξετάζεται για ανάλυση συναισθήματος. Αυτό βοηθά στην εξαγωγή των θετικών και αρνητικών στοιχείων κάθε ξενοδοχείου.

Σε αντίθεση με το NPS (Net Promoter Score) που χρησιμοποιεί βαθμολογίες για την αξιολόγηση κάθε κριτικής, η ανάλυση συναισθήματος επικεντρώνεται στο περιεχόμενό τους.

Με τη βοήθεια της μηχανικής μάθησης και της βαθιάς μάθησης, το σύστημα ενημερώνεται για διάφορες ορολογίες και φράσεις για την ταξινόμηση των δηλώσεων. Για την επεξεργασία των ανατροφοδοτήσεων, οι τέσσερις κύριοι παράγοντες που λαμβάνονται υπόψη είναι οι εξής:

1. **Precision**
2. **Recall**
3. **F1 Score**
4. **Accuracy**

Οι τρεις αρχικές πτυχές χρησιμοποιούν έναν συγκεκριμένο αλγόριθμο για την αξιολόγηση του συνόλου δεδομένων των ξενοδοχειακών κριτικών για την ανάλυση συναισθήματος. Εν τω μεταξύ, η ακρίβεια βοηθά στη μέτρηση του κατά πόσον τα δεδομένα που αναλύονται είναι σωστά ή όχι.

- Ποια είναι όμως τα οφέλη της ανάλυσης συναισθήματος στις ξενοδοχειακές επιχειρήσεις;

α. Μέτρηση της φήμης των ξενοδοχείων στην αγορά

Ένα από τα κύρια οφέλη από τη χρήση της ανάλυσης συναισθήματος είναι η γνώση της φήμης του ξενοδοχείου. Τα περισσότερα λογισμικά διαχείρισης φήμης χρησιμοποιούν έναν

συνδυασμό NLP, NPS και μηχανικής μάθησης για την αξιολόγηση της φήμης ενός ξενοδοχείου.

Οι κριτικές από όλες τις πλατφόρμες κρατήσεων στις οποίες ένα άτομο είναι καταχωρημένο συνεργάζονται για να σχηματίσουν ένα σύνολο δεδομένων. Αυτά τα δεδομένα στη συνέχεια περιορίζονται περαιτέρω σε τρία διαφορετικά τμήματα χρησιμοποιώντας ανάλυση συναισθήματος ως θετικά, αρνητικά και ουδέτερα.

Αφού ταξινομηθούν τα δεδομένα, χρησιμοποιούνται ορισμένες μετρήσεις για τον υπολογισμό της βαθμολογίας φήμης του ξενοδοχείου.

β. Κατανόηση των ελαττωμάτων ενός ξενοδοχείου

Κάθε δήλωση που εμπίπτει στην αρνητική κατηγορία έχει μια φράση που σχετίζεται με υπηρεσίες που δεν αρέσουν στους επισκέπτες. Ας εξετάσουμε ένα παράδειγμα αρνητικής κριτικής:

‘Το προσωπικό του ξενοδοχείου ήταν αρκετά αγενές απέναντί μου όταν παραπονέθηκα για το ακάθαρτο δωμάτιο’

Αν αναλύσουμε αυτή την κριτική, μπορούμε να καταλάβουμε ξεκάθαρα ότι ο επισκέπτης δεν είναι ευχαριστημένος με δύο πράγματα:

- Τη συμπεριφορά του προσωπικού
- Την καθαριότητα του ξενοδοχείου

Έτσι, το ξενοδοχείο μπορεί να λάβει διορθωτικά μέτρα για να διασφαλίσει ότι άλλοι επισκέπτες δεν θα αντιμετωπίσουν το ίδιο πρόβλημα.

γ. Ανάλυση των ανταγωνιστών

Δεν είναι εφικτό κάποιο ξενοδοχείο να λάβει πληροφορίες περί των οικονομικών στοιχείων ενός ανταγωνιστή για ανάλυση. Ωστόσο, με την ανάλυση συναισθήματος των κριτικών του ξενοδοχείου, μπορούμε να αξιολογήσουμε το επίπεδο φήμης στο οποίο βρίσκονται στην αγορά.

Πολλά λογισμικά διαχείρισης φήμης διαθέτουν τη δυνατότητα ανάλυσης ανταγωνιστών. Μόλις ενημερώσουμε το σύνολο των ανταγωνιστών μας στο σύστημα, αυτό αντλεί σχόλια από όλες τις πηγές και μας παρέχει μια έκθεση ανάλυσης κριτικών ξενοδοχείων.

Αυτή η έκθεση αποτελείται από τις επαινετικές και επικριτικές πτυχές όλων των ανταγωνιστών-ξενοδοχείων.

δ. Επανεξέταση όλων των δεδομένων με μία μόνο κίνηση

Χρειάζεται να διαθέτουμε μεγάλο όγκο δεδομένων από κριτικές ξενοδοχείων για την ανάλυση συναισθήματος. Όπως ανέφερα παραπάνω, τα σχόλια μπορούν να αξιοποιηθούν από πολλαπλές πλατφόρμες αξιολόγησης.

Εξοικονομούμε χρόνο αλλά και προσπάθεια για να εξετάσθεί κάθε κριτική ξεχωριστά. Επιπλέον, η μαζική ανάλυση της ανατροφοδότησης αναδεικνύει τους πιο κρίσιμους και αξιολογητικούς παράγοντες.[24]

4.2 Το σύνολο δεδομένων από το TripAdvisor

Στον σημερινό κόσμο η ανάλυση συναισθήματος μπορεί να διαδραματίσει ζωτικό ρόλο σε κάθε κλάδο. Η ταξινόμηση των tweets, των σχολίων στο Facebook ή των κριτικών προϊόντων με τη χρήση ενός αυτοματοποιημένου συστήματος μπορεί να εξοικονομήσει πολύ χρόνο και χρήμα. Ταυτόχρονα, η πιθανότητα σφάλματος είναι μικρότερη. Παρακάτω, θα εξηγήσω μια εργασία ανάλυσης συναισθήματος με τη χρήση του συνόλου δεδομένων κριτικών ξενοδοχείων.

Θα χρησιμοποιήσω την python και βιβλιοθήκες της python.

Εργαλεία που χρησιμοποιήθηκαν:

1. Python
2. Pandas Library
3. NLTK corpus
4. Re
5. PyCharm ως IDE

Σύνολο δεδομένων

Θα χρησιμοποιήσω ένα σύνολο δεδομένων για την προεπεξεργασία, όπως ανέφερα προηγουμένως. Το σύνολο δεδομένων περιέχει κριτικές ξενοδοχείων από το Trip Advisor.

Έχει 3 φύλλα εργασίας(General and Traveler Ranking, Prices και Reviews), ωστόσο εμείς θέλουμε μόνο το 3ο φύλλο, δηλαδή αυτό που περιέχει και τις κριτικές των ξενοδοχείων(Reviews). Το συγκεκριμένο φύλλο αποτελείται από 9 στήλες: Review's Title, Review's Date, Reviewer's Username, Full Review, Rating, Hotel's Name, Hotel's Location, Hotel's Class. Από αυτές τις στήλες θα χρησιμοποιήσουμε τις Review's Title, Full Review και Rating. Οι στήλες Review's Title και Full Review είναι δεδομένα κειμένου ενώ η στήλη Rating περιέχει αριθμούς από το 1 έως το 5 που δείχνει το πόσο καλή είναι μία κριτική ή όχι.

Η δουλειά μας είναι να αναλύσουμε τις κριτικές ως θετικές και αρνητικές. Ας ρίξουμε μια ματιά στο σύνολο δεδομένων. Παρακάτω βλέπουμε ένα μείγμα από το σύνολο των δεδομένων μας.

	A	B	C	D	E	F	G	H	I	J
1	Review's Title	Review's Date	Reviewer's Username	Reviewer's Location	Full Review	Rating	Hotel's Name	Hotel's Location	Hotel's Class	
2	Great location, cc	September 9, 2017	themisb	Messery, France	Nice. Brilliant loc	5 of 5 bubl	The Zillers Bou	Athens	4 Stars	
3	Great service and	September 9, 2017	hlk613	Cincinnati, Ohio	The upscale hote	4 of 5 bubl	Daios Luxury L	Thessaloniki	5 Stars	
4	Perfect location f	September 9, 2017	Somebodyaround		Nice hotel with fr	4 of 5 bubl	The Bristol Hot	Thessaloniki	5 Stars	
5	Best breakfast &	September 9, 2017	caronaf1	New York City, New Yo	I love this hotel, s	5 of 5 bubl	Archipelagos	Mykonos	5 Stars	
6	Best Hotel in myk	September 9, 2017	Zaid A	Stockholm, Sweden	Good Hospitality	5 of 5 bubl	Kirini - My Myk	Mykonos	5 Stars	
7	Fairly nice hotel,	September 9, 2017	bcmlawer	Melville, New York	If you want a hot	3 of 5 bubl	Apanema Resc	Mykonos	4 Stars	
8	Not worth it!!	September 9, 2017	caebayer	California	We stayed at Sa	2 of 5 bubl	San Antonio St	Mykonos	4 Stars	
9	Spoiled our Wed	September 9, 2017	Theresa K		We stayed in a F	3 of 5 bubl	Aphrodite Beac	Mykonos	4 Stars	
10	Fantastic experie	September 9, 2017	Ahmed N		It was unbelievat	5 of 5 bubl	Tharroe of Myk	Mykonos	5 Stars	
11	Amazing team	September 9, 2017	SachaLondon	London	We've just spent	5 of 5 bubl	Tharroe of Myk	Mykonos	5 Stars	
12	Three days wasn	September 9, 2017	Kirsty W	Auburn, Alabama	Wow.....what ca	5 of 5 bubl	Petinos Hotel	Mykonos	4 Stars	
13	A place to relax	September 9, 2017	T-lavoyageur		A nice hotel with	4 of 5 bubl	San Marco Hot	Mykonos	4 Stars	
14	Unreal Experienc	September 9, 2017	karagiannopoulosc	Philadelphia, Pennsylva	Mykonos should	5 of 5 bubl	Kouros Hotel &	Mykonos	5 Stars	
15	Mr Bex	September 9, 2017	luc b	Leopoldsburg, Belgium	One world on its	5 of 5 bubl	Kouros Hotel &	Mykonos	5 Stars	
16	Five-star by all m	September 9, 2017	rihamasri	Amman, Jordan	I wouldn't say it's	5 of 5 bubl	Myconian Impe	Mykonos	5 Stars	
17	Outstanding!	September 9, 2017	tonydsydney	Sydney, Australia	Just enjoyed a w	5 of 5 bubl	Grace Mykono	Mykonos	3 Stars	
18	Great location an	September 9, 2017	Hiro H	Sydney	We stayed for 4	5 of 5 bubl	Petasos Beach	Mykonos	4 Stars	
19	Nice location but	September 9, 2017	Chantelle J		We stayed for 3	3 of 5 bubl	Petasos Beach	Mykonos	4 Stars	
20	Wonderful holida	September 9, 2017	Claudiamatthews		My husband and	5 of 5 bubl	Mykonos Esse	Mykonos	4 Stars	

Προεπεξεργασία δεδομένων

Στην πραγματική ζωή, οι επιστήμονες δεδομένων σπάνια λαμβάνουν δεδομένα που είναι πολύ καθαρά και ήδη προετοιμασμένα για τα μοντέλα μηχανικής μάθησης. Σχεδόν για κάθε

έργο, πρέπει να αφιερώσουμε χρόνο για τον καθαρισμό και την επεξεργασία των δεδομένων. Ας καθαρίσουμε λοιπόν το σύνολο δεδομένων.

Μερικά από τα βήματα προεπεξεργασίας είναι τα εξής:

- Αφαίρεση των μηδενικών τιμών.
- Μετατροπή όλου του κειμένου σε πεζά γράμματα.
- Αφαίρεση μη αλφαριθμητικών λέξεων/χαρακτήρων (όπως αριθμοί και σημεία στίξης) με χρήση regex.
- Λημματοποίηση(Lemmatizing) / Tokenization
- Αφαίρεση λέξεων διακοπής(Stop words)

Αφαίρεση των μηδενικών τιμών:

Μια σημαντική διαδικασία καθαρισμού δεδομένων είναι η εξάλειψη των μηδενικών τιμών. Η συνάρτηση Pandas DataFrame dropna() χρησιμοποιείται για την αφαίρεση γραμμών και στηλών με τιμές Null/NaN. Από προεπιλογή, αυτή η συνάρτηση επιστρέφει ένα νέο DataFrame και το αρχικό DataFrame παραμένει αμετάβλητο.

Ας ελέγξουμε πόσες μηδενικές τιμές έχουμε στο σύνολο δεδομένων.

```
# drop the null values
reviews_df = reviews_df.dropna()
print(len(reviews_df) - len(reviews_df.dropna()))
```

Έχουμε μηδενικές τιμές σε 0 γραμμές. Πράγμα το οποίο είναι σημαντικό διότι θα έχουμε αρκετά δεδομένα για να εκπαιδεύσουμε τον μελλοντικό μας αλγόριθμο όπως θα δούμε σε επόμενη ενότητα.

Πρέπει να έχουμε όλα τα δεδομένα συμβολοσειράς στη στήλη Full_Review. Εάν υπάρχουν δεδομένα που έχουν άλλους τύπους, θα προκαλέσουν προβλήματα στα επόμενα βήματα.

Τώρα, θα ελέγξουμε τον τύπο δεδομένων όλων των κριτικών. Εάν υπάρχει κάποια γραμμή που έχει δεδομένα σε οποιονδήποτε άλλο τύπο εκτός από συμβολοσειρά, θα την αλλάξουμε σε συμβολοσειρά.

```
# if there is any row having data in any other type than string we will change
that to a string.
for i in range(0,len(reviews_df)-1):
    if type(reviews_df.iloc[i]['Full Review']) != str:
        reviews_df.iloc[i]['Full Review'] = str(reviews_df.iloc[i]['Full
Review'])
```

Καθώς κάνουμε ανάλυση συναισθήματος, είναι σημαντικό να πούμε στο μοντέλο μας τι είναι θετικό συναίσθημα και τι είναι αρνητικό συναίσθημα.

Στη στήλη αξιολόγησης(Rating), έχουμε αξιολογήσεις από το 1 έως το 5. Μπορούμε να ορίσουμε το 1 και το 2 ως κακές κριτικές και το 4 και το 5 ως καλές κριτικές.

Τι γίνεται όμως με το 3;

Το 3 βρίσκεται στη μέση. Δεν είναι ούτε καλό ούτε κακό. Απλά μέτριο. Αλλά θέλουμε να ταξινομήσουμε τις καλές ή κακές κριτικές. Έτσι, αποφάσισα να ξεφορτωθώ όλα τα 3.

```
# drop all reviews with rating equal to 3
reviews_df = reviews_df[reviews_df['Rating'] != '3 of 5 bubbles']
reviews_df.reset_index(drop=True, inplace=True)
```

Στην συνέχεια θα συμβολίσουμε τα θετικά συναισθήματα ως 1 και τα αρνητικά συναισθήματα ως 0. Η συνάρτηση 'sentiment' επιστρέφει 1 αν η βαθμολογία είναι 4 ή περισσότερο, διαφορετικά επιστρέφει 0. Στη συνέχεια, θα εφαρμόσουμε τη συνάρτηση sentiment και θα δημιουργήσουμε μια νέα στήλη που θα αναπαριστά το θετικό και το αρνητικό συναίσθημα ως 1 ή 0.

```
# denote positive sentiments as 1 and negative sentiments as 0
def sentiment(n):
```

```
    pattern = '(4|5)'
    return 1 if re.match(pattern, n) else 0

reviews_df['Sentiment'] = reviews_df['Rating'].apply(sentiment)
print(reviews_df['Sentiment'].head())
```

Αρχικά, πρέπει να προετοιμάσουμε τα χαρακτηριστικά εκπαίδευσης(training set). Συνδυάζουμε τις δύο στήλες 'Review's Title' και 'Full Review' και δημιουργούμε μια ενιαία στήλη. Γράφουμε μια συνάρτηση 'combined_features' που θα συνδυάζει και τις δύο στήλες. Στη συνέχεια, εφαρμόζουμε τη συνάρτηση και δημιουργούμε μια νέα στήλη 'all_features' που θα περιέχει τις συμβολοσειρές από τις στήλες Review's Title και Full Review.

```
# Combine both Title and Review columns and make one single column
def combined_features(row):
    return row["Review's Title"] + ' '+ row['Full Review']
reviews_df['all_features'] = reviews_df.apply(combined_features, axis=1)
print(reviews_df.head())
```

Μετατροπή όλου του κειμένου σε πεζά γράμματα:

Πρόκειται για ένα από τα πιο συνηθισμένα βήματα προεπεξεργασίας, όπου το κείμενο μετατρέπεται κατά προτίμηση σε πεζά γράμματα.

Αυτή την μετατροπή μπορούμε να την επιτύχουμε χρησιμοποιώντας την συνάρτηση lower() όπως βλέπουμε παρακάτω:

```
# converting all text to lower case.
reviews_df['all_features'] = reviews_df['all_features'].str.lower()
```

Αφαίρεση μη αλφαριθμητικών λέξεων/χαρακτήρων (όπως αριθμοί και σημεία στίξης):

Μια απλή λύση είναι η χρήση κανονικών εκφράσεων για την αφαίρεση μη αλφαριθμητικών χαρακτήρων από μια συμβολοσειρά. Η ιδέα είναι να χρησιμοποιηθεί η έκρφαση [r'[^a-zA-

Z]\s?] , η οποία ταιριάζει με κάθε χαρακτήρα που δεν είναι χαρακτήρας λέξης. Αυτή την αφαίρεση μπορούμε να την επιτύχουμε χρησιμοποιώντας την παρακάτω γραμμή κώδικα:

```
# striping out non alphanumeric words/characters (such as numbers and
punctuation) using regex
reviews_df['all_features'] = reviews_df['all_features'].str.replace(r'[^a-zA-Z
]\s?',r' ',regex=True)
```

Λημματοποίηση(Lemmatizing) / Tokenization:

Στις περισσότερες περιπτώσεις τα ρήματα αντιμετωπίζονται σαν ουσιαστικά όταν προσπαθούμε να χρησιμοποιήσουμε μόνο το Wordnet Lemmatizer. Για να το ξεπεράσουμε αυτό, χρησιμοποιούμε ετικέτες POS (Part of Speech).

Προσθέτουμε μια ετικέτα με μια συγκεκριμένη λέξη που καθορίζει τον τύπο της (ρήμα, ουσιαστικό, επίθετο κ.λπ.).

Για παράδειγμα,

Word + Type (POS tag) —> Lemmatized Word

driving + verb ‘v’ —> drive

dogs + noun ‘n’ —> dog

Αφαίρεση λέξεων διακοπής(Stop words):

Οι λέξεις διακοπής είναι οι λέξεις που χρησιμοποιούνται αρκετά συχνά και αφαιρούνται από το κείμενο, καθώς δεν προσθέτουν καμία αξία στην ανάλυση. Οι λέξεις αυτές έχουν λιγότερο ή καθόλου νόημα.

Η βιβλιοθήκη του NLTK αποτελείται από έναν κατάλογο λέξεων που θεωρούνται stop words για την αγγλική γλώσσα. Μερικές από αυτές είναι: [i, me, my, myself, we, our, ours, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself, yourselves, he, most,

other, some, such, no, nor, not, only, own, same, so, then, too, very, can, will, just, don, don't, should, should've, aren't, could, couldn't, didn't, didn't]

Εμείς θέλουμε να αφαιρέσουμε όλες τις λέξεις διακοπής και συγκεκριμένα στην Αγγλική γλώσσα. Αυτό θα επιτευχθεί με τον παρακάτω κώδικα:

```
# Removing stop words
stops = set(stopwords.words("english"))

def remove_stops(row):
    my_list = row['all_features']
    meaningful_words = [w for w in my_list if not w in stops]
    return (meaningful_words)

reviews_df['all_features'] = reviews_df.apply(remove_stops, axis=1)
```

4.3 Υλοποίηση

Ορισμός Κατηγοριών και εύρεση των index set (EDA)

Η διερευνητική ανάλυση δεδομένων ή (EDA) είναι η κατανόηση των συνόλων δεδομένων συνοψίζοντας τα κύρια χαρακτηριστικά τους και συχνά απεικονίζοντάς τα οπτικά. Αυτό το βήμα είναι πολύ σημαντικό, ιδίως όταν φτάνουμε στη μοντελοποίηση των δεδομένων προκειμένου να εφαρμόσουμε τη μηχανική μάθηση. Η απεικόνιση στην EDA αποτελείται από ιστογράμματα, Box plot, Scatter plot και πολλά άλλα. Συχνά απαιτείται πολύς χρόνος για τη διερεύνηση των δεδομένων.

Μέσω της διαδικασίας της EDA, καταφέραμε να βρούμε τις πτυχές(aspects) ή διαφορετικά τα index set των κατηγοριών(categories) τα οποία στην συνέχεια θα συγκριθούν χρησιμοποιώντας την μέθοδο Fuzzy String Matching. Εδώ θα ήθελα να προσθέσω ότι τις κατηγορίες τις όρισα εγώ ο ίδιος ποιες θα είναι και οι οποίες τελικά θα αποτελέσουν τις υπηρεσίες(services) των ξενοδοχείων:

1. Cleanliness
2. Service
3. Location
4. Value

5. Room
6. Food
7. Facility
8. Staff

Όσον αφορά τα index set χρησιμοποίησα EDA, και πιο συγκεκριμένα χρησιμοποιώντας την βιβλιοθήκη της python wordcloud, με την οποία μπορούμε να βρούμε τις πιο συνηθισμένες λέξεις που χρησιμοποιούνται για την κριτική ενός ξενοδοχείου. Παρακάτω παρουσιάζω σε μορφή κώδικα το index set για κάθε κατηγορία:

```
# Declaration of our index terms of the category Cleanliness
aspects = ['satisfactory', 'ample', 'hygienic', 'proper', 'spotless', 'odor',
'dirty', 'clean', 'smell']

# Declaration of our index terms of the category Service
aspects = ['desk', 'check in', 'check out', 'reliable', 'fast', 'convenient',
'service']

# Declaration of our index terms of the category Location
aspects = [ 'railway', 'view', 'station', 'airport', 'distance', 'far',
'close', 'train', 'metro', 'transport', 'market', 'mall', 'surrounding',
'areas', 'highway', 'traffic', 'out']

# Declaration of our index terms of the category Value
aspects = ['price', 'amount', 'rate', 'cheap', 'worth', 'low', 'money',
'economical', 'reasonable', 'fee', 'expensive', 'charge']

# Declaration of our index terms of the category Room
aspects = ['bed', 'bunk-beds', 'toilet', 'bathroom', 'shower', 'dryer',
'fridge', 'space', 'spacious', 'outdated', 'noisy']

# Declaration of our index terms of the category Food
aspects = ['drink', 'breakfast', 'spicy', 'food', 'tasty', 'tea', 'coffee',
'buffet',
'bar', 'restaurant', 'dinner', 'lunch', 'brunch', 'delicious']
```

```
# Declaration of our index terms of the category Facility
aspects = ['front', 'pool', 'gym', 'wifi', 'spa', 'internet', 'wireless',
'broken', 'parking', 'ventilation']

# Declaration of our index terms of the category Staff
aspects = ['friendly', 'helpful', 'reliable', 'quick', 'good', 'polite',
'staff']
```

Fuzzy String Matching

Υπάρχουν πολλοί τρόποι να συγκρίνουμε κείμενο στην python. Όμως, συχνά αναζητούμε έναν εύκολο τρόπο σύγκρισης κειμένου. Η σύγκριση κειμένου είναι απαραίτητη για διάφορους σκοπούς ανάλυσης κειμένου και επεξεργασίας φυσικής γλώσσας.

Ένας από τους ευκολότερους τρόπους σύγκρισης κειμένου στην python είναι η χρήση της βιβλιοθήκης fuzzy-wuzzy. Εδώ, η μέγιστη βαθμολογία είναι το 100, με βάση την ομοιότητα των συμβολοσειρών. Βασικά, μας δίνεται ο δείκτης ομοιότητας. Η βιβλιοθήκη χρησιμοποιεί την απόσταση Levenshtein για τον υπολογισμό της διαφοράς μεταξύ δύο συμβολοσειρών.

Στην περίπτωσή μας για να συγκρίνω το πόσο όμοια είναι μια κριτική σε σχέση με τις πτυχές(aspects) κάθε κατηγορίας χρησιμοποίησα την τεχνική fuzz.token_set_ratio, διότι σε περίπτωση που οι δύο συμβολοσειρές έχουν διαφορετικό μήκος, άλλα ήδη συναρτήσεων ταξινόμησης ενδέχεται να μην είναι σε θέση να αποδώσουν καλά σε αυτή την κατάσταση.

Τέλος αποφάσισα να κρατήσω μόνο τις κριτικές των οποίων ο λόγος(ratio) είναι μεγαλύτερος από 40 και στην συνέχεια να φτιάξω 8 Dataframes τα οποία θα περιέχουν τιμές 1 ή 0 που θα δηλώνουν εάν μία κριτική πληροί ή όχι την παραπάνω προϋπόθεση. Το κάθε Dataframe αντιστοιχεί σε μία κατηγορία, δηλαδή θα έχουν μορφή σαν το παρακάτω παράδειγμα:

Cleanlines
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
1
0

Clustering χρησιμοποιώντας Kmeans – Elbow Method

Στην συνέχεια συνδύασα τις στήλες των 8 προηγούμενων Dataframe και όσες κριτικές ήταν ίσες με 1 τις αποθήκευσα σε μία νέα στήλη 'Comb'. Αυτή η νέα στήλη δείχνει σε ποιες κατηγορίες αναφέρεται η κάθε κριτική. Δηλαδή θα είναι της μορφής:

	Comb
0	Food Staff
1	Service
3	Location Food
4	Location Room Staff

Η παραπάνω στήλη θα αποτελέσει την είσοδο στον αλγόριθμο ομαδοποίησης Kmeans ο οποίος θα χρησιμοποιηθεί για την εύρεση ομάδων που δεν έχουν επισημανθεί ρητά στα

δεδομένα, δηλαδή στην περίπτωσή μας η κάθε συστάδα θα αποτελείται από ξενοδοχεία τα οποία ‘ταιριάζουν’ περισσότερο σε μία ή περισσότερες υπηρεσίες. Πιο συγκεκριμένα επειδή ο αλγόριθμος ομαδοποίησης Kmeans δέχεται μόνο ακέραιες τιμές θα χρειαστεί να χρησιμοποιήσουμε την βιβλιοθήκη TfidfVectorizer η οποία θα μετατρέψει τις υπηρεσίες μας-κατηγορίες σε διανύσματα χαρακτηριστικών που μπορούν να χρησιμοποιηθούν ως είσοδο στον Kmeans.

Πρωτού όμως δούμε πώς λειτουργεί ο αλγόριθμος συστασοποίησης Kmeans θα χρειαστεί να βρούμε τον αριθμό των συστάδων. Η τεχνική που μπορούμε να χρησιμοποιήσουμε για να βρούμε αυτόν τον αριθμό ονομάζεται μέθοδος αγκώνα(Elbow Method) και η ιδέα είναι να εκτελέσουμε την ομαδοποίηση k-means για ένα εύρος συστάδων k (ας πούμε από 1 έως 15) και για κάθε τιμή, υπολογίζουμε το άθροισμα των τετραγωνικών αποστάσεων από κάθε σημείο προς το κέντρο που του έχει ανατεθεί(παραμορφώσεις).

Όταν οι παραμορφώσεις σχεδιαστούν και το διάγραμμα μοιάζει με βραχίονα τότε ο "αγκώνας"(το σημείο καμπής της καμπύλης) είναι η καλύτερη τιμή του k. Παρακάτω βλέπουμε τον κώδικα για να επιτύχουμε το αποτέλεσμα που επιθυμούμε:

```
distortions = []
K = range(1,15)
for k in K:
    kmeanModel = KMeans(n_clusters=k)
    kmeanModel.fit(new_X)
    distortions.append(kmeanModel.inertia_)
```

Το σημείο στο οποίο δημιουργείται το σχήμα του αγκώνα είναι 6, δηλαδή για την τιμή μας Κ ένας βέλτιστος αριθμός συστάδων είναι 6. Τώρα ας εκπαιδεύσουμε το μοντέλο στο σύνολο δεδομένων με αριθμό συστάδων 6.

```
kmeanModel = KMeans(n_clusters=6)
y_kmeans = kmeanModel.fit_predict(X)
print(y_kmeans)
```

Ταξινομητές ξεχωριστά / Μέθοδος Συνόλου (Ensemble Method)

Αφού βρήκαμε το πώς συσταδοποιούνται τα ξενοδοχεία με βάση τις κατηγορίες-υπηρεσίες, πάμε τώρα να δούμε εάν <<μαντέψουν>> επιτυχώς οι ταξινομητές που θα χρησιμοποιήσω το σωστό sentiment(συναίσθημα) για κάθε κριτική.

Θα δούμε 2 διαφορετικούς τρόπους:

1. 4 ταξινομητές ξεχωριστά.
2. Μοντέλο Ensemble(Συνδυασμός των 4 ταξινομητών σε ένα).

4 ταξινομητές ξεχωριστά:

Αρχικά μπαίνουμε με το σκεπτικό ότι δεν έχουμε labels, καθώς η βαθμολογία αστέρων δεν εκφράζει την ακριβή εμπειρία του πελάτη. Επομένως, χρησιμοποίησα την βιβλιοθήκη TextBlob, η οποία αποτελείτα από 2 στήλες:

1. Polarity: κυμαίνεται ε τιμές -1 με +1.
2. Subjectivity: κυμαίνεται σε τιμές 0 με +1.

Εάν το Polarity μίας κριτικής σύμφωνα με το TextBlob είναι > 1 τότε θα αποθηκεύεται σε μία νέα στήλη <<label>> το καινούργιο μας sentiment (0 ή 1).

Στην συνέχεια προσθέτω τις τιμές TF-IDF (Term Frequency – inverse Document Frequency) για κάθε λέξη.

- Το TF υπολογίζει τον κλασικό αριθμό φορών που η λέξη εμφανίζεται στο κείμενο
- Το IDF υπολογίζει τη σχετική σημασία αυτής της λέξης που εξαρτάται από το πόσες φορές μπορεί να εμφανιστεί μία λέξη σε πολλά κείμενα.

Τώρα ήρθε η ώρα να χρησιμοποιήσουμε τους 4 ταξινομητές μας ξεχωριστά:

Bernoulli Naive Bayes Classifier:

Ο ταξινομητής Naive Bayes του Bernoulli υποθέτει ότι όλα τα χαρακτηριστικά μας είναι δυαδικά, ώστε να λαμβάνουν μόνο δύο τιμές. Σε εμάς είναι όντως δυαδικά και μπορεί να υλοποιηθεί με τον παρακάτω τρόπο:

```
# naive bayes bernoulii
from sklearn.naive_bayes import BernoulliNB
```

```
Bernouli_NB = BernoulliNB()
Bernouli_NB.fit(features_train, labels_train)
```

Παραπάνω βλέπουμε το πώς με κώδικα γραμμένο σε python μπορεί να εκπαιδευτεί ένας ταξινομητής Bernoulli. Δεν έχουμε χρησιμοποιήσει καμία από τις παραμέτρους του.

Random Forest Classifier:

Ο ταξινομητής Random Forest είναι ένας μετα-εκτιμητής που προσαρμόζει έναν αριθμό ταξινομητών δέντρων απόφασης σε διάφορα υποδείγματα του συνόλου δεδομένων και χρησιμοποιεί τον μέσο όρο για να βελτιώσει την ακρίβεια πρόβλεψης και να ελέγξει την υπερβολική προσαρμογή. Το μέγεθος του υποδείγματος ελέγχεται με την παράμετρο max_samples αν bootstrap=True (προεπιλογή). Εμείς θα χρησιμοποιήσουμε ολόκληρο το σύνολο δεδομένων μας. Μπορεί να υλοποιηθεί με τον παρακάτω τρόπο:

```
# Random Forest
from sklearn.ensemble import RandomForestClassifier

Random_F = RandomForestClassifier(n_estimators=200, random_state=0)
Random_F.fit(features_train, labels_train)
```

Όπου n_estimator = 200 συμβολίζει τον αριθμό των δέντρων στο δάσος(forest).

Logistic Regression:

Ο Logistic Regression είναι μια τεχνική ταξινόμησης που χρησιμοποιείται στη μηχανική μάθηση. Χρησιμοποιεί μια λογιστική συνάρτηση για τη μοντελοποίηση της εξαρτημένης μεταβλητής. Η εξαρτημένη μεταβλητή έχει διχοτομικό χαρακτήρα, δηλαδή θα μπορούσαν να υπάρχουν μόνο δύο πιθανές κλάσεις (π.χ.: είτε ο καρκίνος είναι κακοήθης είτε όχι). Κατά συνέπεια, η τεχνική αυτή χρησιμοποιείται κατά την επεξεργασία δυαδικών δεδομένων. Μπορεί να υλοποιηθεί με τον παρακάτω τρόπο:

```
# Logistic Regression
from sklearn.linear_model import LogisticRegression
```

```
Logistic_Reg = LogisticRegression()
Logistic_Reg.fit(features_train, labels_train)
```

SVM Linear:

Το SVM ή Support Vector Machine είναι ένα γραμμικό μοντέλο για προβλήματα ταξινόμησης και παλινδρόμησης. Μπορεί να επιλύσει γραμμικά και μη γραμμικά προβλήματα και λειτουργεί καλά για πολλά πρακτικά προβλήματα. Η ιδέα του SVM είναι απλή: Ο αλγόριθμος δημιουργεί μια γραμμή ή ένα υπερεπίπεδο που διαχωρίζει τα δεδομένα σε κλάσεις. Αυτό μπορεί να υλοποιηθεί με τον παρακάτω τρόπο:

```
from sklearn.svm import SVC

model = svm.SVC(kernel='linear', C=1)
model.fit(features_train, labels_train)
```

Όπου θα χρησιμοποιήσουμε την παράμετρο kernel(πυρήνα), η οποία μετασχηματίζει έναν χώρο δεδομένων εισόδου στην απαιτούμενη μορφή. Εμείς θα χρησιμοποιήσουμε linear kernel μπορεί να χρησιμοποιηθεί ως κανονικό γινόμενο τελείας δύο δεδομένων παρατηρήσεων. Επιπλέον η C: είναι η παράμετρος κανονικοποίησης, του όρου σφάλματος.

Μοντέλο Ensemble(Συνδυασμός των 4 ταξινομητώ σε ένα):

Όπως και παραπάνω έτσι και εδώ χρησιμοποίησα την βιβλιοθήκη TextBlob και TF-IDF.

Εδώ όμως εφάρμοσα ένα μοντέλο Ensemble με βάση την ψηφοφορία (Voting Based), στο οποίο συνδύασα τους ταξινομητές Naïve Bayes, Random Forest, Logistic Regression, και Support Vector Machine με σκοπό να τους συγκρίνω όπως θα δούμε σε παρακάτω ενότητα.

Αρχικά ένωσα και τους 4 ταξινομητές μέσα σε μία λίστα ‘estimators’

```
# create the sub models
estimators = []
model1 = BernoulliNB()
estimators.append(('bnb', model1))
model2 = RandomForestClassifier(n_estimators=200, random_state=0)
```

```
estimators.append(('rfc', model2))
model3 = LogisticRegression()
estimators.append(('logistic', model3))
model4 = svm.SVC(kernel='linear', C=1, gamma=1)
estimators.append(('svm', model4))
```

και στην συνέχεια δημιούργησα το Ensemble μοντέλο και το εκπαίδευσα.

```
# create the ensemble model
ensemble = VotingClassifier(estimators)
ensemble.fit(features_train, labels_train)
```

4.4 Αξιολόγηση της προτεινόμενης προσέγγισης

Είναι σημαντικό να αξιολογηθεί η απόδοση του μοντέλου ταξινόμησης προκειμένου να χρησιμοποιηθούν αυτά τα μοντέλα στην παραγωγή για την επίλυση προβλημάτων του πραγματικού κόσμου. Τα μέτρα απόδοσης στα μοντέλα ταξινόμησης μηχανικής μάθησης χρησιμοποιούνται για να αξιολογηθεί πόσο καλά αποδίδουν οι αλγόριθμοι ταξινόμησης μηχανικής μάθησης σε ένα δεδομένο πλαίσιο. Αυτά τα μέτρα απόδοσης περιλαμβάνουν το accuracy, το precision, το recall και το F1-score. Επειδή μας βοηθά να κατανοήσουμε τα πλεονεκτήματα και τους περιορισμούς αυτών των μοντέλων όταν κάνουμε προβλέψεις σε νέες καταστάσεις, η απόδοση των μοντέλων είναι απαραίτητη για τη μηχανική μάθηση.

Παρακάτω θα δούμε το πόσο καλά αποδώσανε τα μοντέλα ταξινόμησης τόσο ξεχωριστά όσο και ενωμένα.

Οι ετικέτες-στόχοι στο σύνολο δεδομένων για τις κριτικές ξενοδοχείων που πήραμε από το TripAdvisor είναι Θετικό (1) και Αρνητικο (0). Δημιούργησα ένα διαχωρισμό εκπαίδευσης και δοκιμής(train_test_split), όπου το 25% του συνόλου δεδομένων προορίζεται για σκοπούς δοκιμής.

```
# Dividing Data into Training and Test Sets
from sklearn.model_selection import train_test_split
```

```
features_train, features_test, labels_train, labels_test = train_test_split(t,
labels, test_size=0.25, random_state=0)
```

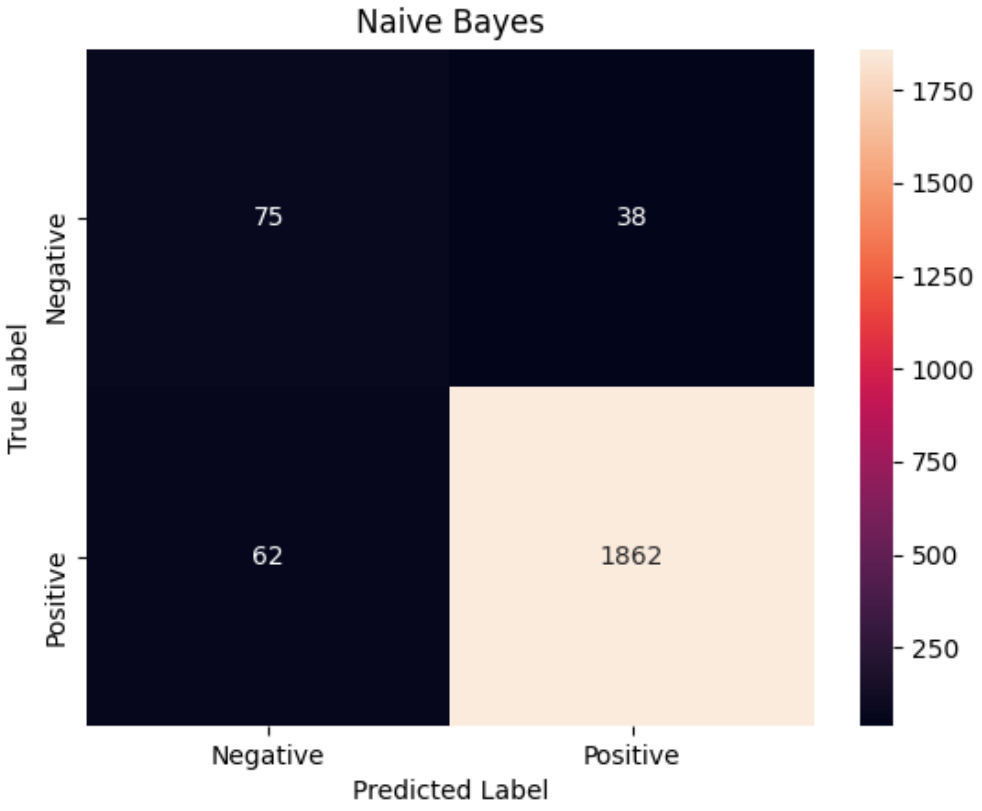
Μετρικές για 4 ξεχωριστά μοντέλα:

Στην συνέχεια εκπαίδευσα και τα 4 μοντέλα ξεχωριστά και εκτύπωσα τον πίνακα σύγχησης(Confusion Matrix) τους. Ακολουθεί ο κώδικας για την εκτύπωση του πίνακα σύγχυσης(Naïve Bayes Bernouli):

```
# create the heatmap
from sklearn.metrics import confusion_matrix

print('Naive_Bayes')
print(confusion_matrix(labels_test, labels_pred))
```

Εκτυπώθηκε ο ακόλουθος πίνακας σύγχυσης για τον αλγόριθμο Naïve Bayes Bernoulli:



Σχήμα 4.1: Πίνακας σύγχυσης για το Naïve Bayes Bernoulli που αντιπροσωπεύει τις προβλέψεις(predictions) έναντι των πραγματικών δεδομένων(actual data) σε δεδομένα δοκιμής(test data).

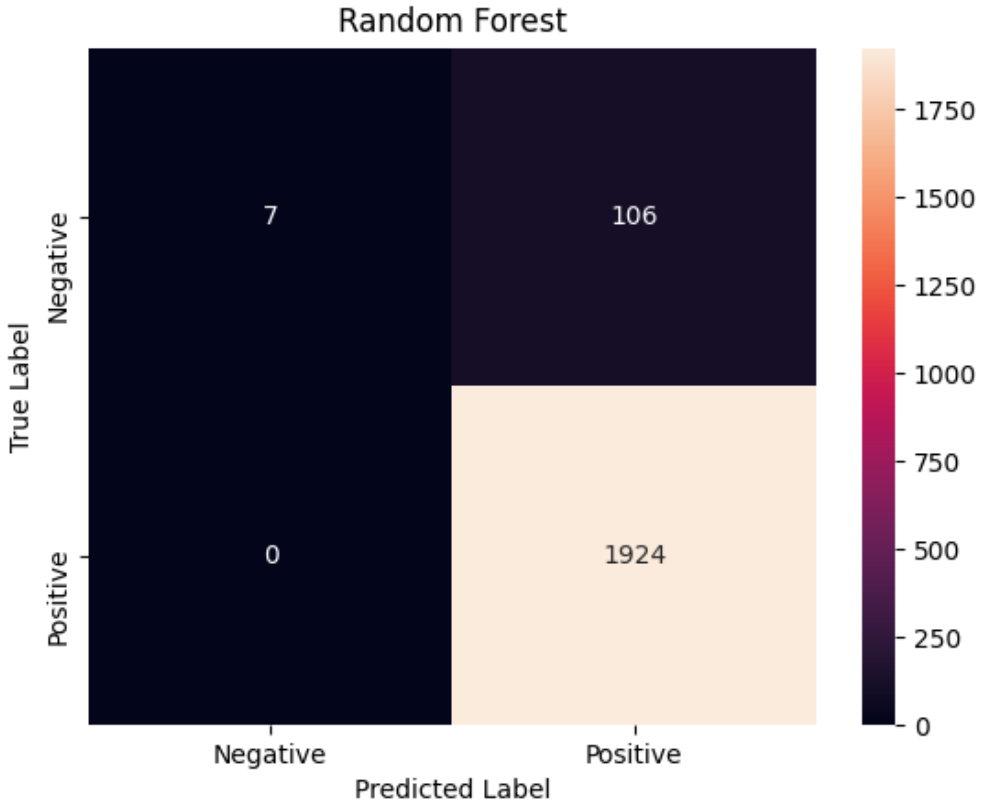
Τα προβλεπόμενα αποτελέσματα των δεδομένων στο παραπάνω διάγραμμα θα μπορούσαν να διαβαστούν με τον ακόλουθο τρόπο δεδομένου ότι το 1 αντιπροσωπεύει θετικό σχόλιο:

- Αληθές θετικό (TP): Το αληθές θετικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των θετικών από τις πραγματικές θετικές περιπτώσεις. Από τις 1924 πραγματικές θετικές περιπτώσεις, οι 1862 είναι σωστά προβλεπόμενες θετικές. Συνεπώς, η τιμή του True Positive είναι 1862.
- Ψευδώς θετικό (FP): Το ψευδώς θετικό αντιπροσωπεύει την τιμή των εσφαλμένων θετικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των αρνητικών (από τις 113) που προβλέπονται εσφαλμένα ως θετικές. Από τα 113 πραγματικά αρνητικά, 38 προβλέπονται εσφαλμένα ως θετικά. Συνεπώς, η τιμή του ψευδούς θετικού είναι 38.
- Αληθές αρνητικό (TN): Το αληθές αρνητικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των αρνητικών από τις πραγματικές αρνητικές περιπτώσεις. Από τις 113 πραγματικές αρνητικές περιπτώσεις, οι 75 προβλέπονται σωστά ως αρνητικές. Συνεπώς, η τιμή του True Negative είναι 75.
- Ψευδές αρνητικό (FN): Το ψευδές αρνητικό αντιπροσωπεύει την τιμή των εσφαλμένων αρνητικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των θετικών (από τις 1924) που προβλέπονται εσφαλμένα ως αρνητικές. Από τα 1924 πραγματικά θετικά, 62 προβλέπονται εσφαλμένα ως αρνητικά. Συνεπώς, η τιμή του ψευδούς αρνητικού είναι 62.

Ακολουθεί ο κώδικας για την εκτύπωση του πίνακα σύγχυσης(Random Forest Classifier):

```
# create the heatmap
print('Random Forest ')
print(confusion_matrix(labels_test, labels_pred))
```

Εκτυπώθηκε ο ακόλουθος πίνακας σύγχυσης για τον αλγόριθμο Random Forest Classifier:



Σχήμα 4.2: Πίνακας σύγχυσης για το Random Forest Classifier που αντιπροσωπεύει τις προβλέψεις(predictions) έναντι των πραγματικών δεδομένων(actual data) σε δεδομένα δοκιμής(test data).

Τα προβλεπόμενα αποτελέσματα των δεδομένων στο παραπάνω διάγραμμα θα μπορούσαν να διαβαστούν με τον ακόλουθο τρόπο δεδομένου ότι το 1 αντιπροσωπεύει θετικό σχόλιο:

- Αληθές θετικό (TP): Το αληθές θετικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των θετικών από τις πραγματικές θετικές περιπτώσεις. Από τις 1924 πραγματικές θετικές περιπτώσεις, οι 1924 είναι σωστά προβλεπόμενες θετικές. Συνεπώς, η τιμή του True Positive είναι 1924.
- Ψευδώς θετικό (FP): Το ψευδώς θετικό αντιπροσωπεύει την τιμή των εσφαλμένων θετικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των αρνητικών (από τις 113) που προβλέπονται εσφαλμένα ως θετικές. Από τα 113 πραγματικά αρνητικά, 106 προβλέπονται εσφαλμένα ως θετικά. Συνεπώς, η τιμή του ψευδούς θετικού είναι 106.
- Αληθές αρνητικό (TN): Το αληθές αρνητικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των αρνητικών από τις πραγματικές αρνητικές περιπτώσεις. Από τις

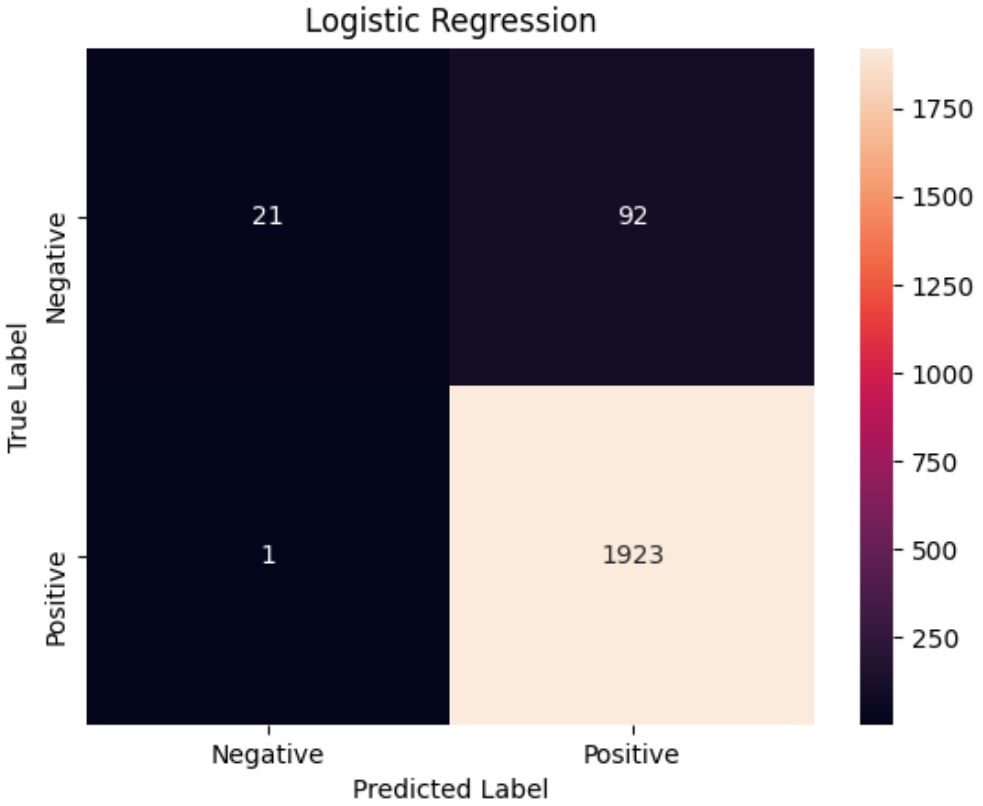
113 πραγματικές αρνητικές περιπτώσεις, οι 7 προβλέπονται σωστά ως αρνητικές. Συνεπώς, η τιμή του True Negative είναι 7.

- Ψευδές αρνητικό (FN): Το ψευδές αρνητικό αντιπροσωπεύει την τιμή των εσφαλμένων αρνητικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των θετικών (από τις 1924) που προβλέπονται εσφαλμένα ως αρνητικές. Από τα 1924 πραγματικά θετικά, 0 προβλέπονται εσφαλμένα ως αρνητικά. Συνεπώς, η τιμή του ψευδούς αρνητικού είναι 0.

Ακολουθεί ο κώδικας για την εκτύπωση του πίνακα σύγχυσης(Logistic Regression):

```
# create the heatmap
print('Logistic Regression ')
print(confusion_matrix(labels_test, labels_pred))
```

Εκτυπώθηκε ο ακόλουθος πίνακας σύγχυσης για τον αλγόριθμο Logistic Regression:



Σχήμα 4.3: Πίνακας σύγχυσης για το Logistic Regression που αντιπροσωπεύει τις προβλέψεις(predictions) έναντι των πραγματικών δεδομένων(actual data) σε δεδομένα δοκιμής(test data).

Τα προβλεπόμενα αποτελέσματα των δεδομένων στο παραπάνω διάγραμμα θα μπορούσαν να διαβαστούν με τον ακόλουθο τρόπο δεδομένου ότι το 1 αντιπροσωπεύει θετικό σχόλιο:

- Αληθές θετικό (TP): Το αληθές θετικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των θετικών από τις πραγματικές θετικές περιπτώσεις. Από τις 1924 πραγματικές θετικές περιπτώσεις, οι 1923 είναι σωστά προβλεπόμενες θετικές. Συνεπώς, η τιμή του True Positive είναι 1923.
- Ψευδώς θετικό (FP): Το ψευδώς θετικό αντιπροσωπεύει την τιμή των εσφαλμένων θετικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των αρνητικών (από τις 113) που προβλέπονται εσφαλμένα ως θετικές. Από τα 113 πραγματικά αρνητικά, 92 προβλέπονται εσφαλμένα ως θετικά. Συνεπώς, η τιμή του ψευδούς θετικού είναι 92.
- Αληθές αρνητικό (TN): Το αληθές αρνητικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των αρνητικών από τις πραγματικές αρνητικές περιπτώσεις. Από τις

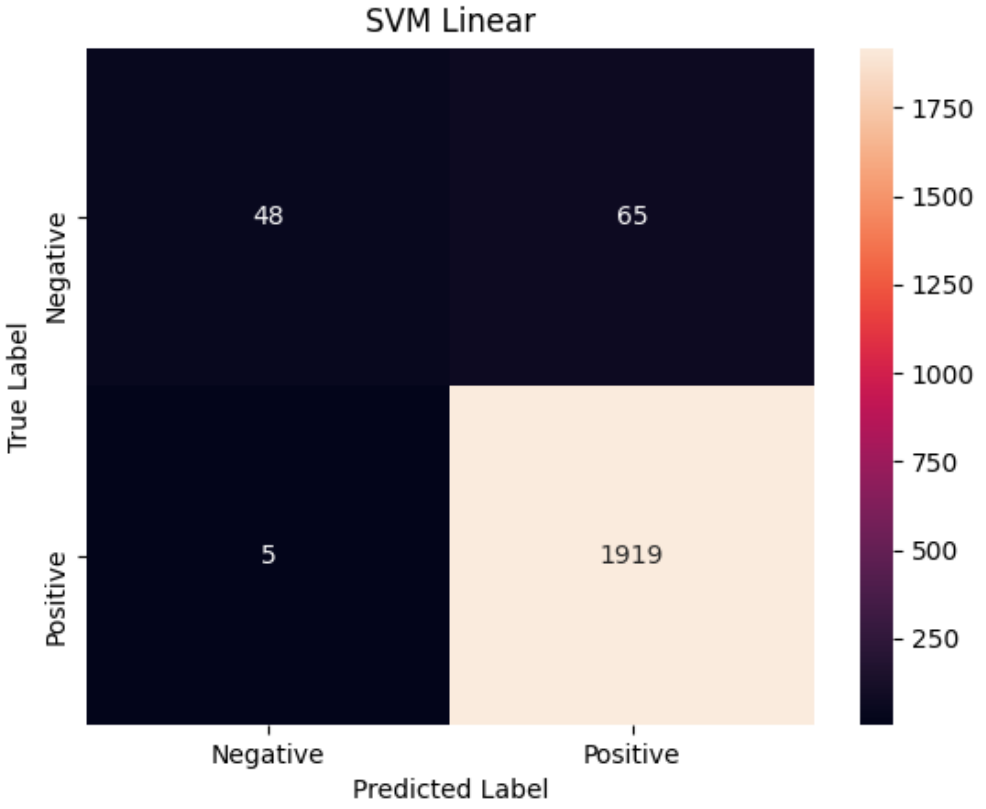
113 πραγματικές αρνητικές περιπτώσεις, οι 21 προβλέπονται σωστά ως αρνητικές. Συνεπώς, η τιμή του True Negative είναι 21.

- Ψευδές αρνητικό (FN): Το ψευδές αρνητικό αντιπροσωπεύει την τιμή των εσφαλμένων αρνητικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των θετικών (από τις 1924) που προβλέπονται εσφαλμένα ως αρνητικές. Από τα 1924 πραγματικά θετικά, 1 προβλέπεται εσφαλμένο ως αρνητικό. Συνεπώς, η τιμή του ψευδούς αρνητικού είναι 1.

Ακολουθεί ο κώδικας για την εκτύπωση του πίνακα σύγχυσης(SVM linear):

```
# create the heatmap
print('SVM Linear ')
print(confusion_matrix(labels_test, labels_pred))
```

Εκτυπώθηκε ο ακόλουθος πίνακας σύγχυσης για τον αλγόριθμο SVM linear:



Σχήμα 4.4: Πίνακας σύγχυσης για το SVM linear ταξινομητή που αντιπροσωπεύει τις προβλέψεις(predictions) έναντι των πραγματικών δεδομένων(actual data) σε δεδομένα δοκιμής(test data).

Τα προβλεπόμενα αποτελέσματα των δεδομένων στο παραπάνω διάγραμμα θα μπορούσαν να διαβαστούν με τον ακόλουθο τρόπο δεδομένου ότι το 1 αντιπροσωπεύει θετικό σχόλιο:

- Αληθές θετικό (TP): Το αληθές θετικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των θετικών από τις πραγματικές θετικές περιπτώσεις. Από τις 1924 πραγματικές θετικές περιπτώσεις, οι 1919 είναι σωστά προβλεπόμενες θετικές. Συνεπώς, η τιμή του True Positive είναι 1919.
- Ψευδώς θετικό (FP): Το ψευδώς θετικό αντιπροσωπεύει την τιμή των εσφαλμένων θετικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των αρνητικών (από τις 113) που προβλέπονται εσφαλμένα ως θετικές. Από τα 113 πραγματικά αρνητικά, 65 προβλέπονται εσφαλμένα ως θετικά. Συνεπώς, η τιμή του ψευδούς θετικού είναι 65.

- Αληθές αρνητικό (TN): Το αληθές αρνητικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των αρνητικών από τις πραγματικές αρνητικές περιπτώσεις. Από τις 113 πραγματικές αρνητικές περιπτώσεις, οι 48 προβλέπονται σωστά ως αρνητικές. Συνεπώς, η τιμή του True Negative είναι 48.
- Ψευδές αρνητικό (FN): Το ψευδές αρνητικό αντιπροσωπεύει την τιμή των εσφαλμένων αρνητικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των θετικών (από τις 1924) που προβλέπονται εσφαλμένα ως αρνητικές. Από τα 1924 πραγματικά θετικά, 5 προβλέπεται εσφαλμένο ως αρνητικό. Συνεπώς, η τιμή του ψευδούς αρνητικού είναι 5.

Στην συνέχεια βρήκα accuracy, precision, recall, and f1-score για όλους τους ταξινομητές ξεχωριστά:

Μοντέλο Naïve Bayes Bernoulli

Accuracy	Precision	Recall	F1 Score
0.95	0.98	0.97	0.97

Μοντέλο Random Forest

Accuracy	Precision	Recall	F1 Score
0.95	0.95	1.00	1.00

Μοντέλο Logistic Regression

Accuracy	Precision	Recall	F1 Score
0.95	0.95	1.00	0.98

Μοντέλο SVM Linear

Accuracy	Precision	Recall	F1 Score
0.97	0.97	1.00	0.98

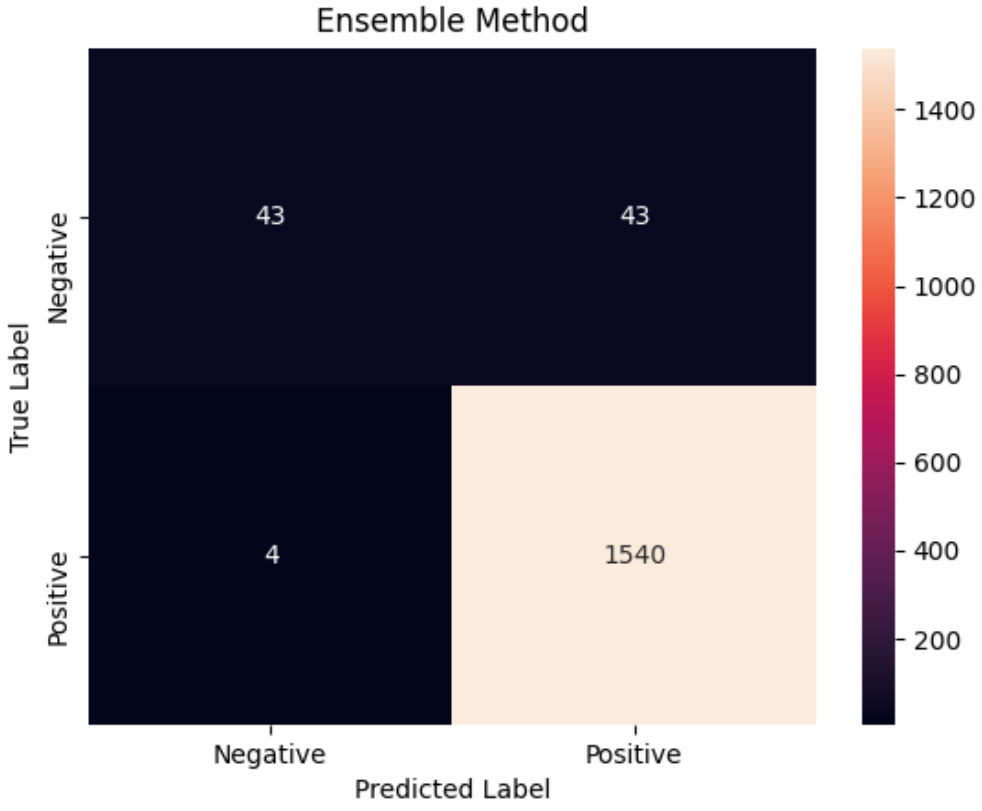
Τις παραπάνω μετρικές τις βρήκα χρησιμοποιώντας τους παρακάτω τύπους:

- Accuracy = (TP +TN) / (P + N)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 Score = (2 * Precision * Recall) / (Precision + Recall)

Παρατηρούμε ότι και τα 4 μοντέλα ξεχωριστά έχουν υψηλές τιμές σε όλες τις μετρικές. Αυτό πιθανόν σημαίνει ότι τα μοντέλα είναι κατά κάποιο τρόπο ισορροπημένα, δηλαδή έχουν καταφέρει να ταξινομήσουν σωστά τόσο τις θετικές κριτικές όσο και τις αρνητικές.

Μετρικές για το Μοντέλο Ensemble(Συνδυασμός των 4 ταξινομητών σε ένα):

Στην συνέχεια εκπαίδευσα και τα 4 μοντέλα μαζί δημιουργώντας ένα νέο μοντέλο που ονομάζεται μέθοδος Ensemble και εκτύπωσα τον πίνακα σύγχησης(Confusion Matrix) του.



Σχήμα 4.5: Πίνακας σύγχυσης για το μοντέλο Ensemble που αντιπροσωπεύει τις προβλέψεις(predictions) έναντι των πραγματικών δεδομένων(actual data) σε δεδομένα δοκιμής(test data).

Τα προβλεπόμενα αποτελέσματα των δεδομένων στο παραπάνω διάγραμμα θα μπορούσαν να διαβαστούν με τον ακόλουθο τρόπο δεδομένου ότι το 1 αντιπροσωπεύει θετικό σχόλιο:

- Αληθές θετικό (TP): Το αληθές θετικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των θετικών από τις πραγματικές θετικές περιπτώσεις. Από τις 1544 πραγματικές θετικές περιπτώσεις, οι 1540 είναι σωστά προβλεπόμενες θετικές. Συνεπώς, η τιμή του True Positive είναι 1540.
- Ψευδώς θετικό (FP): Το ψευδώς θετικό αντιπροσωπεύει την τιμή των εσφαλμένων θετικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των αρνητικών (από τις 86) που προβλέπονται εσφαλμένα ως θετικές. Από τα 86 πραγματικά αρνητικά, 43 προβλέπονται εσφαλμένα ως θετικά. Συνεπώς, η τιμή του ψευδούς θετικού είναι 43.
- Αληθές αρνητικό (TN): Το αληθές αρνητικό αντιπροσωπεύει την τιμή των σωστών προβλέψεων των αρνητικών από τις πραγματικές αρνητικές περιπτώσεις. Από τις 86 πραγματικές αρνητικές περιπτώσεις, οι 43 προβλέπονται σωστά ως αρνητικές. Συνεπώς, η τιμή του True Negative είναι 43.
- Ψευδές αρνητικό (FN): Το ψευδές αρνητικό αντιπροσωπεύει την τιμή των εσφαλμένων αρνητικών προβλέψεων. Η τιμή αυτή αντιπροσωπεύει τον αριθμό των θετικών (από τις 1544) που προβλέπονται εσφαλμένα ως αρνητικές. Από τα 1924 πραγματικά θετικά, 4 προβλέπεται εσφαλμένο ως αρνητικό. Συνεπώς, η τιμή του ψευδούς αρνητικού είναι 4.

Στην συνέχεια βρήκα accuracy, precision, recall, and f1-score για την μέθοδο Ensemble:

Μοντέλο Ensemble Method(Συνδυασμός 4 αλγοριθμών ταξινόμησης)

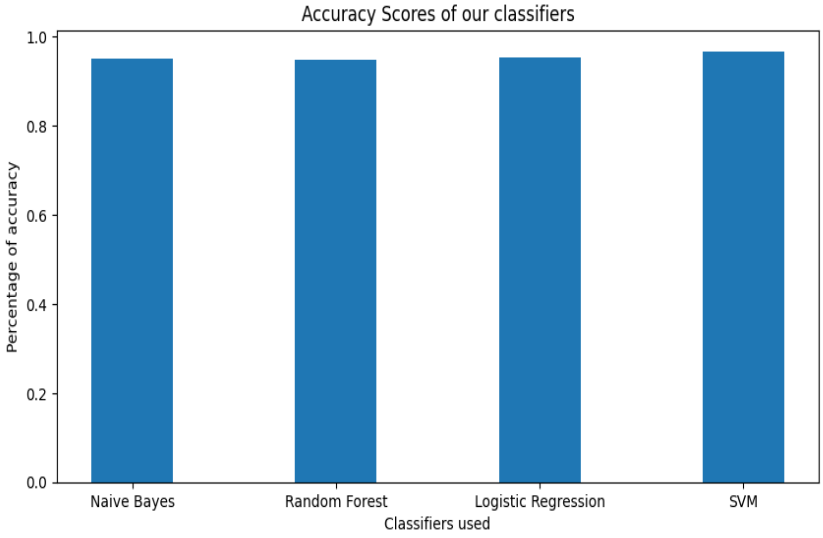
Accuracy	Precision	Recall	F1 Score
0.95	0.98	0.97	0.97

Παρατηρούμε ότι και σε αυτήν την περίπτωση οι μετρικές είναι υψηλές, δηλαδή πάλι έχουν ταξινομηθεί σωστά τόσο οι θετικές όσο και οι αρνητικές κριτικές.

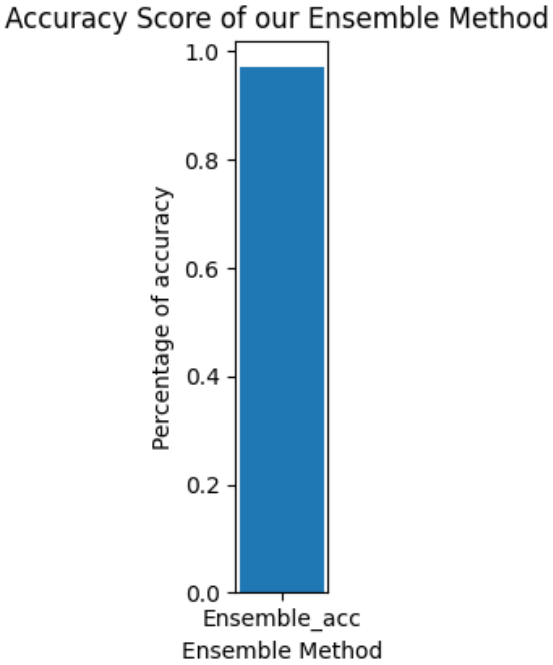
4.5 Συγκριτική Ανάλυση

Αφού αναλύθηκε παραπάνω η απόδοση των μοντέλων μας παρατηρήσαμε ότι όλα τα μοντέλα φαίνονται να είναι αρκετά ισορροπημένα. Ωστόσο το ισορροπημένο δεν σημαίνει απαραίτητα ότι είναι και καλό. Στην ενότητα αυτή παρουσιάζονται κα συγκρίνονται, με χρήση γραφημάτων και πινάκων τα αποτελέσματα αυτών.

Αφού παρατηρήσαμε ότι όλες οι μετρικές των μοντέλων μας είναι αρκετά υψηλές, στην συνέχεια βρήκα τις μετρικές για την κάθε κλάση ξεχωριστά χρησιμοποιώντας classification report. Οι κλάσεις που έχουμε είναι οι αρνητικές κριτικές και οι θετικές. Με αυτόν τον τρόπο θα δούμε τα accuracy, precision, recall και F1 score της κάθε κλάσεως ξεχωριστά.



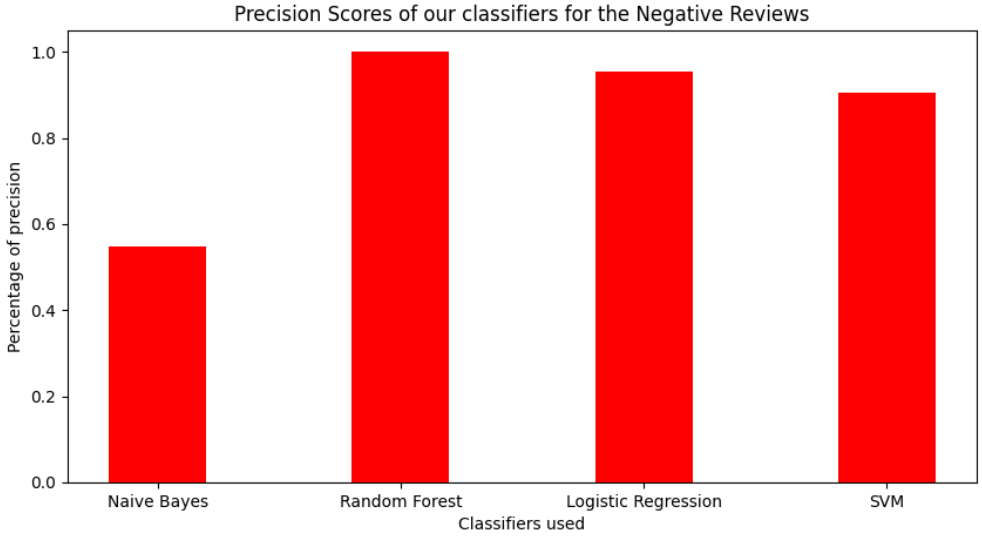
Εικόνα 4.6: Accuracy score των ταξινομητών



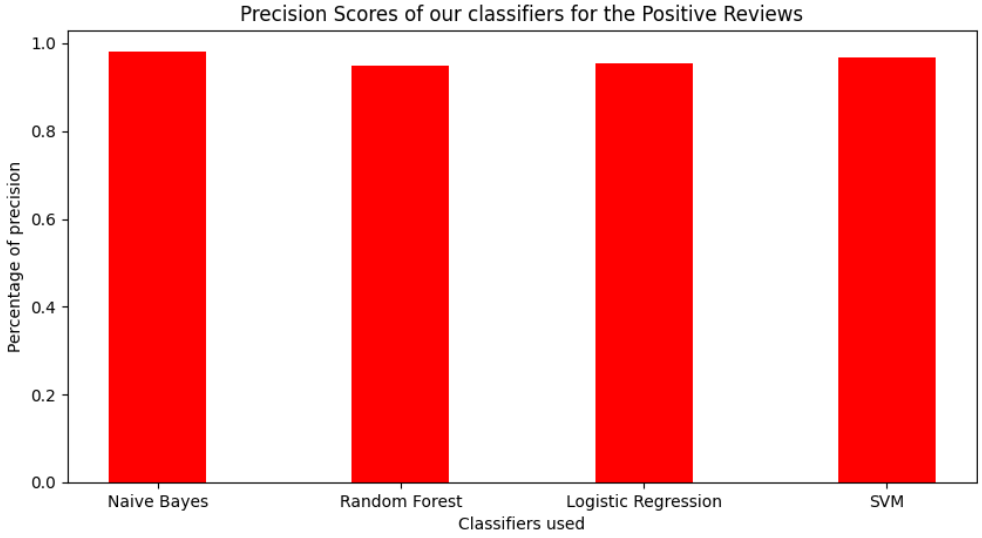
Εικόνα 4.7: Accuracy score της μεθόδου Ensemble

Παραπάνω βλέπουμε ότι το accuracy ανεξαρτήτως ποια μέθοδο επιλέξουμε είναι σχεδόν τέλειο. Η ακρίβεια, χωρίς αμφιβολία, είναι μια σημαντική μετρική που πρέπει να λαμβάνεται υπόψη, αλλά δεν δίνει πάντα την πλήρη εικόνα.

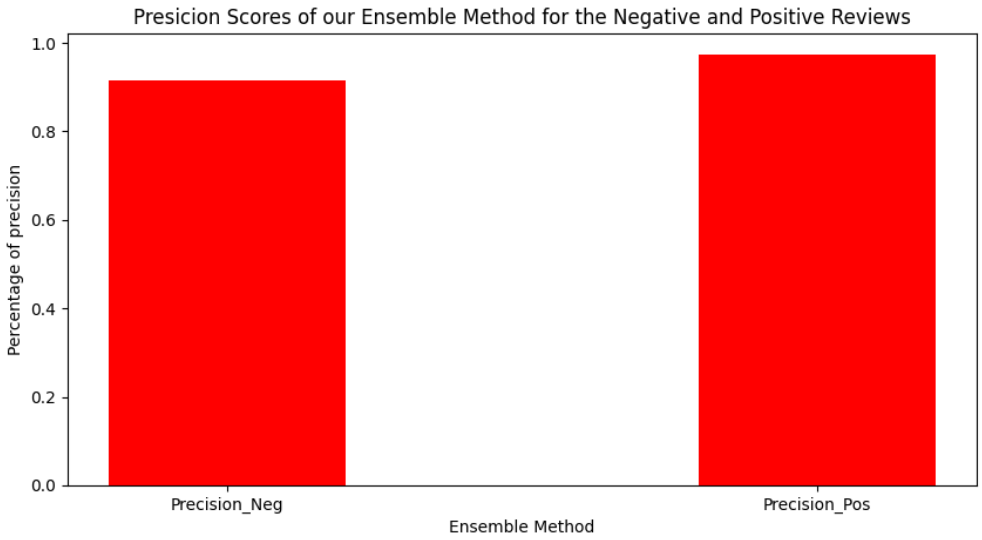
Για αυτό θα δούμε και τις άλλες μετρικές μας ξεκινώντας από το precision.



Εικόνα 4.8: Precision score των ταξινομητών (Αρνητικές Κριτικές)



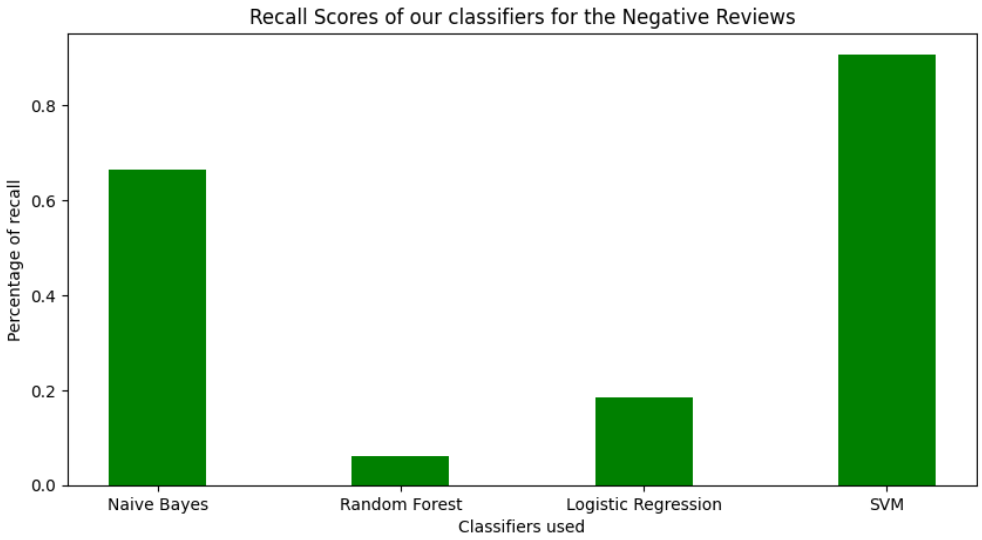
Εικόνα 4.9: Precision score των ταξινομητών (Θετικές Κριτικές)



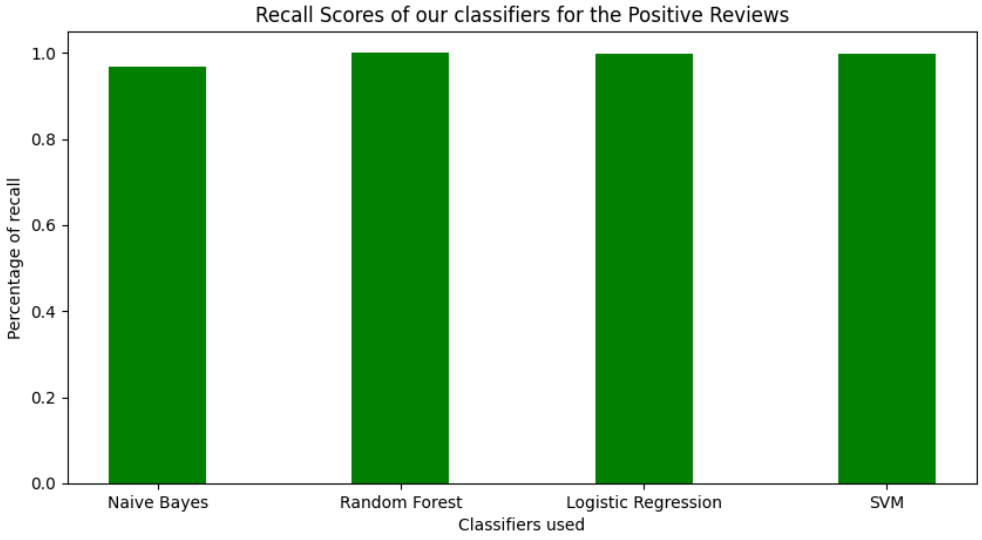
Εικόνα 4.10: Precision score της μεθόδου Ensemble (Αρνητικές-Θετικές Κριτικές)

Εδώ παρατηρούμε ότι και πάλι όλα τα μοντέλα έχουν τιμή μεγαλύτερη του 90% τόσο για τις αρνητικές κριτικές όσο και τις θετικές, εκτός από τον Naïve Bayes, ο οποίος για τις αρνητικές κριτικές έχει χαμηλή τιμή στο precision του. Αυτό σημαίνει ότι αν θέλαμε να διαλέξουμε τον «καλύτερο» αλγόριθμο ταξινόμησης με βάση το πόσο σωστά ταξινομούνται οι αρνητικές κριτικές θα ήταν πιο συνετό να αποφύγουμε τον Naïve Bayes και να επιλέξουμε έναν από τους υπόλοιπους.

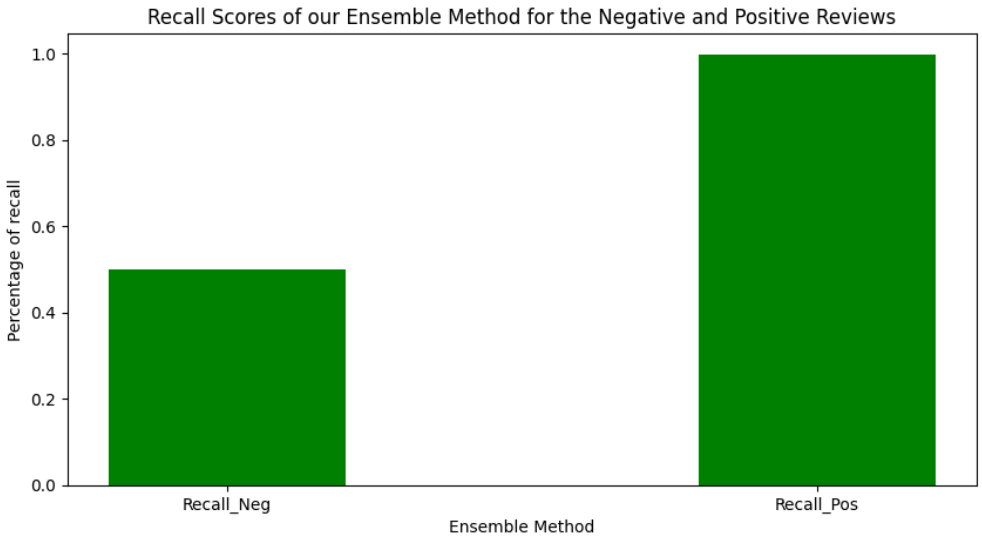
Αξίζει να σημειωθεί πριν δούμε το recall της κάθε κλάσης, ότι το precision που είδαμε προηγουμένως μας δείχνει το πόσες κριτικές έχουν προβλεφθεί σωστά σε σχέση με τις εκπαιδευμένες, ενώ το recall θα μας δείξει το πόσα αποτελέσματα έχουν επιστραφεί.



Εικόνα 4.11: Recall score των ταξινομητών (Αρνητικές Κριτικές)



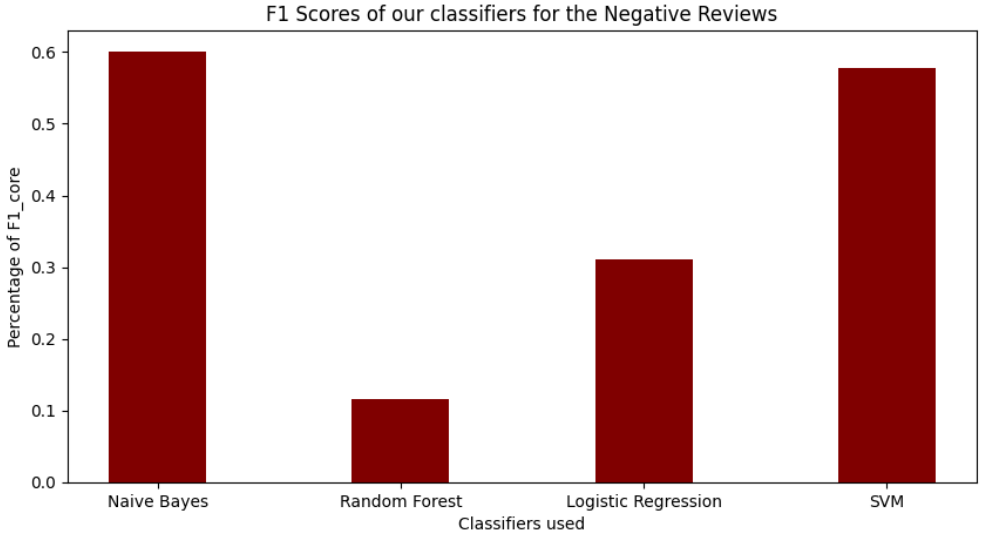
Εικόνα 4.12: Recall score των ταξινομητών (Θετικές Κριτικές)



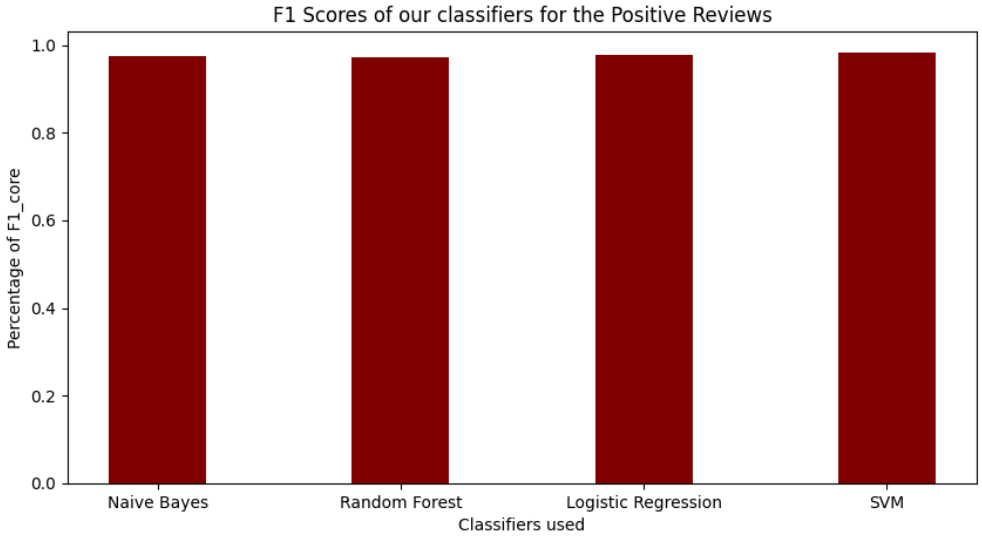
Εικόνα 4.13: Recall score της μεθόδου Ensemble (Αρνητικές-Θετικές Κριτικές)

Όπως σωστά περιμέναμε το recall για την θετική κλάση είναι και πάλι υψηλό, ωστόσο για την αρνητική κλάση κάποιοι ταξινομητές έχουν επιστρέψει λιγότερα αποτελέσματα σε σχέση με άλλους. Επομένως αν δεν μας νοιάζει που έχουν επιστραφεί λίγα αποτελέσματα και πιστεύουμε ότι είμαστε ικανοποιημένοι, τότε διαλέγουμε τον SVM. Ωστόσο βλέπουμε ότι η μέθοδος Ensemble επίστρεφει έναν μέσο όρο αυτών των τιμών με τιμή γύρω στο 50%.

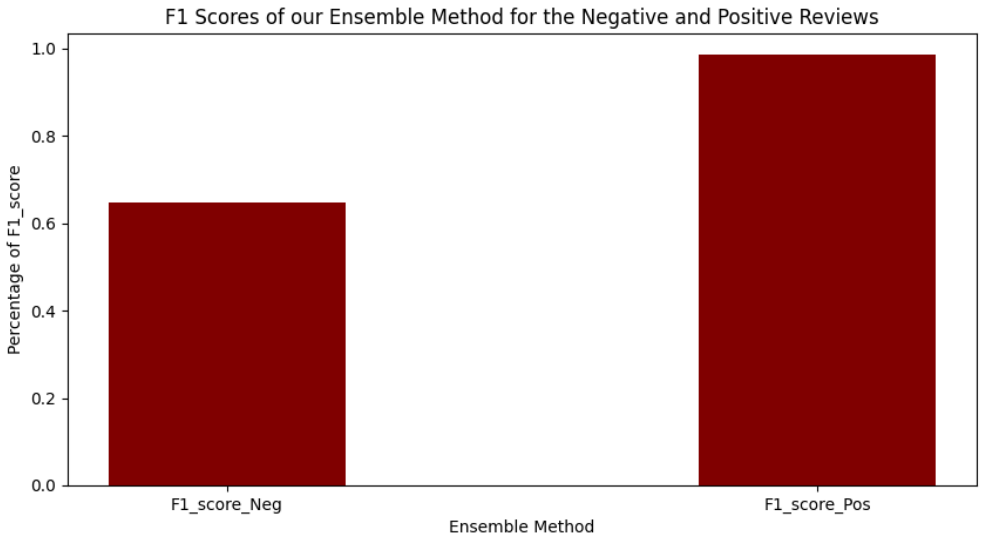
Από τις σημαντικά διαφορετικές τιμές μεταξύ θετικής και αρνητικές κλάσης είναι ασφαλές να πούμε ότι το dataset μας είναι αρκετά imbalanced(μη ισορροπημένο). Η καλύτερη μετρική που μας απομένει όταν ένα dataset είναι μη ισορροπημένο είναι το f1-score, το οποίο είναι ο αρμονικός μέσος όρος μεταξύ precision & recall.



Εικόνα 4.14: F1 score των ταξινομητών (Αρνητικές Κριτικές)



Εικόνα 4.15: F1 score των ταξινομητών (Θετικές Κριτικές)



Εικόνα 4.16: F1 score της μεθόδου Ensemble (Αρνητικές-Θετικές Κριτικές)

Το f1-score θα αποτελέσει τον τελικό μας «κριτή» με βάση τον οποίο θα διαλέξουμε ποιος ταξινομητής είναι ο πιο αποδοτικός. Επειδή καταλήξαμε ότι το dataset είναι imbalanced, θα δώσουμε βάση στα αποτελέσματα της αρνητικής κλάσης, διότι σε ένα imbalanced dataset δίνεται βαρύτητα στην κλάση με τις περισσότερες παρατηρήσεις. Αν κοιτάξουμε τους 4 ταξινομητές ξεχωριστά παρατηρούμε ότι ο πιο αποδοτικός θα είναι ο είτε ο Naïve Bayes Bernoulli είτε ο SVM Linear. Ωστόσο επειδή δεν παρατηρείται ακριβώς ποιος είναι αυτός με την μεγαλύτερη τιμή δημιούργησα τα παρακάτω πινακάκια ύστερα από τα αποτελέσματα που μας έδωσε το classification report.

Μοντέλο Naïve Bayes Bernoulli

Classes	*Precision*	*Recall*	*Accuracy*	*F1-score*
Negative	0.55	0.66	0.95	0.60
Positive	0.98	0.97	0.95	0.97

Μοντέλο Random Forest

Classes	*Precision*	*Recall*	*Accuracy*	*F1-score*
Negative	1.00	0.06	0.95	0.12
Positive	0.95	1.00	0.95	0.97

Μοντέλο Logistic Regression

Classes	*Precision*	*Recall*	*Accuracy*	*F1-score*
Negative	0.95	0.19	0.95	0.31
Positive	0.95	1.00	0.95	0.98

Μοντέλο SVM Linear

Classes	*Precision*	*Recall*	*Accuracy*	*F1-score*
Negative	0.91	0.42	0.95	0.58
Positive	0.97	1.00	0.95	0.98

Βλέπουμε ότι ο Naïve Bayes έχει F1-score 0.60, ενώ ο SVM Linear έχει 0.58. Επομένως σε αυτήν την περίπτωση αποδοτικότερος είναι αυτός του Naïve Bayes Bernoulli για το dataset που πήραμε από το TripAdvisor.

Εάν όμως κάποιος επιθυμούσε έναν αποδοτικότερο αλγόριθμο ταξινόμησης;

Παρακάτω βλέπουμε και το f1-score της Ensemble μεθόδου που χρησιμοποίησα, ο οποίος είναι συνδυασμός όλων των παραπάνω ταξινομητών

Μοντέλο Ensemble

Classes	*Precision*	*Recall*	*Accuracy*	*F1-score*
Negative	0.91	0.50	0.95	0.64
Positive	0.97	0.99	0.95	0.98

Βλέπουμε ότι το f1-score της Ensemble μεθόδου μας έχει τιμή 0.64, δηλαδή παρατηρείται αν και μικρή μια βελτίωση στην αποδοτικότητα του αλγορίθμου μας.

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

5.1 Συμπεράσματα

Στην παρούσα Διπλωματική Εργασία προτάθηκαν 2 διαφορετικοί μεθόδοι που θα μπορούσε κάποιος να ακολουθήσει για να κατηγοριοποιήσει κριτικές ξενοδοχείων οι οποίες συλλέχθηκαν από το TripAdvisor σε θετικές ή αρνητικές με επιβλεπόμενη μάθηση(Supervised Learning), αφού όμως πρώτα χρησιμοποιώντας συσταδοποίηση(Clustering) με μη-επιβλεπόμενη μάθηση(Unsupervised Learning) βρήκαμε συστάδες οι οποίες αποτελούνται από ξενοδοχεία τα οποία 'ταιριάζουν' περισσότερο σε μία ή περισσότερες υπηρεσίες.

Σύμφωνα με τα αποτελέσματα της ομαδοποίησης των κριτικών σε συστάδες με βάση τις υπηρεσίες των ξενοδοχείων και με τη χρήση του αλγορίθμου K-Means ο προσδιορισμός του καλύτερου αριθμού συστάδων με τη μέθοδο Elbow είναι 6. Ωστόσο ο αριθμός των συνδυασμών των υπηρεσιών με τις οποίες μπορεί ένα ξενοδοχείο να ταιριάζει είναι πάρα πολύ μεγάλος. Οπότε, η μέθοδος με την οποία προσπάθησα να ομαδοποιήσω τις τα ξενοδοχεία-υπηρεσίες δεν είναι και ο πιο επιθυμητός.

Στην συνέχεια κατηγοριοποίησα τις κριτικές ανάλογα με το συναίσθημα χρησιμοποιώντας αρχικά 4 διαφορετικούς αλγόριθμους ταξινόμησης(Naïve Bayes Bernoulli, Random Forest, Logistic Regression, SVM Linear) και στο τέλος έναν συνδυασμό αυτών(Ensemble Method). Αφού κατάφερα να βρω προγραμματιστικά τους πίνακες σύγχησης τόσο των 4 αλγορίθμων ξεχωριστά όσο και τον συνδυασμό αυτών βρήκα τις μετρικές accuracy, precision, recall και f1-score και με βάση αυτές είδαμε ότι ο πιο αποδοτικός αλγόριθμος ήταν αυτός του Random Forest Classifier, ο οποίος είχε βαθμολογία f1-score 1.00, δηλαδή άριστη. Ωστόσο είναι σχεδόν αδύνατο να υπάρξει τέλεια βαθμολογία, οπότε αποφάσισα να δω τις μετρικές της κάθες κλάσης ξεχωριστά.

Αφού λοιπόν χρησιμοποίησα classification report για βρω τις μετρικές των κλάσεων κατέληξα στα εξής συμπεράσματα:

- Η κατανομή των κριτικών στην κάθε τάξη δεν είναι ισορροπημένη.

- Τα precision και recall έχουν αρκετές διακυμάνσεις στις τιμές τους.
- Το dataset για τις κριτικές των ξενοδοχείων από το Trip Advisor είναι αρκετά imbalanced.
- Όταν ένα dataset είναι imbalanced η καλύτερη μετρική είναι αυτή της f1-score.

Με βάση τα παραπάνω συμπεράσματα είδαμε ότι ατομικά ο καλύτερος αλγόριθμος ταξινόμησης είναι ο Naïve Bayes Bernoulli με τιμή 0,60. Όταν όμως είδαμε και το f1-score της Ensemble μεθόδου συνειδητοποιήσαμε ότι είναι αποδοτικότερη με τιμή 0,64. Για την επιλογή μου βασίστηκα στις μετρικές που έδειξαν για την αρνητική κλάση διότι το μη-ισορροπημένο dataset ευνοεί περισσότερο την κλάση με τις περισσότερες παρατηρήσεις.

5.2 Μελλοντική Εργασία

Η μελέτη αυτή μπορεί να χρησιμοποιηθεί ως βάση για περαιτέρω έρευνα και εργασίες. Μια προσθήκη που θα μπορούσε να γίνει είναι η χρήση περισσότερων αλγορίθμων συσταδοποίησης όπως π.χ ο K-RMS και να συγκρίνουμε τα αποτελέσματα με τον Kmeans που χρησιμοποίησα. Επίσης, επειδή το dataset μας είναι imbalanced θα μπορούσαμε να χρησιμοποιήσουμε την τεχνική τυχαίας υπερδειγματοληψίας(Random Over-sampling) δηλαδή να προσθεθούν αυτόματα περισσότερες αντιγραφές στην τάξη μειονότητας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] https://keyhole.co/blog/how-does-sentiment-analysis-work/#:~:text=Sentiment%20analysis%20works%20by%20breaking,sentiment%20score%20to%20each%20topic.&text=A%20sentiment%20analysis%20tool%20would,really%20impressed%20%3D%20%2B4

[2] https://www.commsights.com/benefits-of-sentiment-analysis-for-businesses/

[3] https://revenue-hub.com/how-sentiment-analysis-can-help-your-hotel-build-a-strong-online-reputation/

[4] https://www.repustate.com/blog/sentiment-analysis-challenges-with-solutions/

[5] https://www.lexalytics.com/technology/sentiment-analysis#basics

[6] https://itechindia.co/wp-content/uploads/2021/11/inner11.jpeg

[7] https://itechindia.co/blog/which-of-the-3-algorithms-models-should-you-choose-for-sentiment-analysis-2/

[8] https://www.sas.com/en_us/insights/analytics/predictive-analytics.html#:~:text=Predictive%20analytics%20is%20the%20use,will%20happen%20in%20the%20future.

[9] https://searchbusinessanalytics.techtarget.com/definition/predictive-analytics

[10] https://www.tibco.com/reference-center/what-is-predictive-analytics

[11] https://www.tableau.com/learn/articles/what-is-predictive-analytics

[12] https://relativeinsight.com/how-to-analyze-customer-reviews/

[13] https://www.betterevaluation.org/en/evaluation-options/wordcloud#:~:text=Word%20clouds%20or%20tag%20clouds,frequently%20in%20a%20source%20text.&text=Most%20word%20cloud%20generators%20have,exclude%20common%20or%20similar%20words.

[14] https://www.datacamp.com/community/tutorials/fuzzy-string-python

[15] https://www.jashds.com/blog/2019/05/13/fuzzy-stringmatching-python

[16] https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

[17] https://realpython.com/k-means-clustering-python/

[18] https://www.analyticsvidhya.com/blog/2020/10/a-simple-explanation-of-k-means-clustering/

[19] https://levelup.gitconnected.com/ensemble-learning-using-the-voting-classifier-a28d450be64d

[20] https://medium.com/@nansha3120/bernoulli-naive-bayes-and-its-implementation-cca33ccb8d2e

[21] https://www.javatpoint.com/machine-learning-random-forest-algorithm

[22] https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

[23] https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[24] https://www.ezeeabsolute.com/blog/sentiment-analysis-for-hotel-reviews/