ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση συναισθήματος προτιμήσεων χρηστών στον ταξιδιωτικό κλάδο με αλγορίθμους μηχανικής μάθησης

Φοιτήτρια : Ακριβή Γρηγοροπούλου

ΑΜ:18390184

Επιβλέπουσα Καθηγήτρια

Παναγιώτα Τσελέντη

ΑΘΗΝΑ, ΟΚΤΩΒΡΙΟΣ 2024

(Κενό φύλλο)

UNIVERSITY OF WEST ATTICA

FACULTY OF ENGINEERING

DEPARTMENT OF INFORMATICS & COMPUTER ENGINEERING

DIPLOMA THESIS

Sentiment analysis of user preferences in the tourism sector using machine learning algorithms

Student: Akrivi Grigoropoulou

Registration Number:18390184

Supervisor

Panagiota Tselenti

ATHENS, OCTOBER 2024

(Κενό φύλλο)

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Συναισθηματική ανάλυση προτιμήσεων χρηστών στον ταξιδιωτικό κλάδο με αλγορίθμους μηχανικής μάθησης

Ακριβή Γρηγοροπούλου
Α.Μ. 18390184

Εισηγητής:

Τσελέντη Παναγιώτα

Panagiota Tselenti
Digitally signed by Panagiota Tselenti
Date: 2024.10.08 12:37:26 +03'00'

Εξεταστική Επιτροπή:

Κρούσκα Ακριβή

Akrivi Krouska
Digitally signed by Akrivi Krouska
Date: 2024.10.08 13:16:45 +03'00'

Τρούσσας Χρήστος

Christos Troussas
Digitally signed by Christos Troussas
Date: 2024.10.08 13:02:13 +03'00'

Ημερομηνία εξέτασης:

04/10/24

(Κενό φύλλο)

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Η κάτωθι υπογεγραμμένη Ακριβή Γρηγοροπούλου του Σπυρίδωνος , με αριθμό μητρώου 18390184 φοιτήτρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

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

Η δηλούσα

Ακριβή Γρηγοροπούλου

(Κενό φύλλο)

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία ολοκληρώθηκε μετά από επίμονες προσπάθειες, σε ένα ενδιαφέρον γνωστικό αντικείμενο, όπως αυτό της ανάλυση συναισθήματος με αλγορίθμους μηχανικής μάθησης . Την προσπάθειά μου αυτή υποστήριξε η επιβλέπουσα καθηγήτρια μου κ. Παναγιώτα Τσελέντη, την οποία θα ήθελα να ευχαριστήσω για την αμέριστη υποστήριξη, την ανταπόκριση και το γνήσιο ενδιαφέρον της για την παρούσα διπλωματική εργασία.

Ακόμη, θέλω να εκφράσω την βαθύτατη ευγνωμοσύνη και αγάπη μου στην οικογένειά μου, που με στήριξε σε όλες τις δύσκολες στιγμές και ήταν δίπλα μου στηρίζοντάς με, σε κάθε μου βήμα.

Τέλος, θα ήθελα να ευχαριστήσω τους φίλους μου και το αγόρι μου, που ήταν οι «συνταξιδιώτες» μου σε αυτό το ταξίδι, στις δύσκολες στιγμές και στις χαρούμενες, και μια συνεχής πηγή δύναμης, υποστήριξης και ενθάρρυνσης σε κάθε εμπόδιο.

(Κενό φύλλο)

Περίληψη

Η παρούσα διπλωματική εργασία έχει ως στόχο την ανάπτυξη μιας πλήρους ανάλυσης της αγοράς στον τομέα του ταξιδιωτικού κλάδου. Αρχικά, επιλέχτηκε ένα κατάλληλο σύνολο δεδομένων από την πλατφόρμα Kaggle το οποίο περιέχει διάφορες τουριστικές πληροφορίες οι οποίες αφορούν όχι μόνο τα ίδια τα ξενοδοχεία, όπως το όνομα του, την τοποθεσία του, την κατηγορία του ξενοδοχείου κ.α αλλά και πληροφορίες για τους ίδιους τους χρήστες που γράφουν την κριτική, όπως το username τους, το κειμένο της κριτικής και την βαθμολογία. Το σύνολο δεδομένων αρχικά αναλύθηκε, οπτικοποιήθηκε και έπειτα επεξεργάστηκε ώστε να μπορέσουμε να κατανοήσουμε ολοκληρωτικά τα δεδομένα μας και να εξάγουμε διάφορα συμπεράσματα , που θα μας επιτρέπουν να αποκτήσουμε μια πιο ολοκληρωμένη εικόνα και να αναδείξουμε σημαντικούς τομείς της αγοράς ώστε να ανιχνευθούν προτιμήσεις πελατών και να αναλυθούν οι τάσεις του ταξιδιωτικού κλάδου. Εξετάστηκαν οι προτιμήσεις των χρηστών λαμβάνοντας υπόψη την τοποθεσία που έχουν επισκεφθεί και τις παροχές των ξενοδοχείων (όπως η προσφορά πρωινού) κ.λπ. Επιπλέον, μελετήθηκε η ανίχνευση συναισθηματικών προτιμήσεων και αντιδράσεων των χρηστών βάσει των σχολίων και των αναφορών τους. Στο δεύτερο μέρος της διπλωματικής εργασίας, εφαρμόστηκαν διάφοροι αλγόριθμοι μηχανικής και βαθιάς μάθησης με σκοπό την εκπαίδευσή τους πάνω στα δεδομένα , ώστε να είναι ικανοί να προβλέπουν τις βαθμολογίες από τα κείμενα κριτικών. Τα αποτελέσματα αυτά συλλέχτηκαν και σχολιάστηκαν, ενώ αξιολογήθηκαν και οι αλγόριθμοι που χρησιμοποιήθηκαν. Μέσω αυτής της έρευνας επιδιώκεται η καλύτερη κατανόηση των προτιμήσεων των χρηστών προκειμένου να προσδιοριστούν μοτίβα και τάσεις στον τουριστικό τομέα αλλά και η εύρεση του καταλληλότερου αλγορίθμου για τα συγκεκριμένα δεδομένα.

.

ΛΕΞΕΙΣ – ΚΛΕΙΔΙΑ

Συναισθηματική ανάλυση , ταξιδιωτικός κλάδος ,εξόρυξη δεδομένων, μηχανική μάθηση,

Συναισθηματικές Προτιμήσεις

Abstract

This thesis aims to develop a complete market analysis in the field of the travel industry. Initially, a suitable data set was selected from the Kaggle platform which contains various tourist information concerning not only the hotels themselves, such as the name, location, category of the hotel, etc. but also information about the hotels themselves. users writing the review, such as their username, review rating and rating. The data set will first be analyzed, visualized and then processed so that we can fully understand our data and draw various conclusions, which will allow us to have a more complete picture and highlight important aspects of the market to detect customer preferences and analyze travel industry trends. User preferences will be considered taking into account the location they have visited and hotel amenities (such as breakfast offer) etc. In addition, the detection of emotional preferences and reactions of users based on their comments and reports will be studied. In the second part of the thesis, various machine learning and deep learning algorithms will be applied in order to train them on the data to be able to predict the scores from the review texts. These results will be collected and commented, while the algorithms used will be evaluated. Through this research you seek a better understanding of user preferences in order to identify patterns and trends in the tourism sector but also to find the most appropriate algorithm for the specific data.

Keywords

Emotional analysis, travel sector, data mining, machine learning, emotional preferences.

Περιεχόμενα

Εικόνες

Πίνακες

1. Εισαγωγή

Ο τομέας του τουρισμού αποτελεί έναν από τους πιο αναπτυσσόμενους αλλά και δυναμικούς τομείς στη σημερινή κοινωνία, καθώς επηρεάζει άμεσα την παγκόσμια οικονομία αλλά και επιδρά σημαντικά στην πολιτιστική ανάπτυξη των περιοχών που τον φιλοξενούν.

Ζούμε στην εποχή της ψηφιοποίησης , με αποτέλεσμα η τεχνολογία να έχει καταφέρει να επεκταθεί αν όχι σε όλους, σε αρκετούς τομείς. Ο τουρισμός είναι ένας από αυτούς , αφού κατάφερε να εξελιχθεί από μια απλή διαδικασία κράτησης ξενοδοχείων σε μια πιο πολύπλοκη και διασκεδαστική εμπειρία με πολλές περισσότερες δυνατότητες και επιλογές.

Πιο συγκεκριμένα, μετά την διάδοση του διαδικτύου από τα τέλη της δεκαετίας του 1990 κατάφερε να φέρει επανάσταση στον ταξιδιωτικό και τουριστικό κλάδο. Δημιουργώντας έτσι τη βάση για την ανάπτυξη του λεγόμενου ηλεκτρονικού τουρισμού (e-Tourism) (Xiang, 2018).

Η ψηφιοποίηση των διαδικασιών με τις οποίες κάποιος πλέον μπορεί να οργανώσει ένα ταξίδι, δημιουργεί άπειρα δεδομένα περιέχοντας πολύτιμες πληροφορίες για τις προτιμήσεις, τις ανάγκες και τις συναισθηματικές αντιδράσεις των ταξιδιωτών. Έτσι οι εταιρείες αξιοποιώντας τα δεδομένα, αποκτούν τη δυνατότητα, βελτίωσης της εμπειρίας των χρηστών , κατανόησης των τάσεων της αγοράς αλλά και πιθανής ανάπτυξης καινοτόμων τουριστικών υπηρεσιών. Πλέον με τον ηλεκτρονικό τουρισμό, έχουν ολοκληρωτικά διαφοροποιηθεί οι διαδικασίες που αξιοποιούνται και καταναλώνονται από τις τουριστικές υπηρεσίες, καθώς και οι στρατηγικές που θα ακολουθήσουν οι εταιρείες του τουριστικού κλάδου για τη μεγιστοποίηση των κερδών τους. Επομένως, διάφορες λειτουργίες, όπως το εμπόριο και το μάρκετινγκ επωφελούνται από τον πλούτο ελεύθερης πληροφορίας και αλλάζουν ολοκληρωτικά τις στρατηγικές τους.

Σήμερα, ο τρόπος με τον οποίο θα επιλεχθούν καταλύματα, αξιοθέατα και μεταφορά έχει αλλάξει οριστικά. Υπάρχει λοιπόν, η ανάγκη δημιουργίας πλατφορμών, όπου ο χρήστης θα μπορεί να ασκεί κριτική για το ξενοδοχείο που επισκέφτηκε, το μέρος που είδε ή την πτήση που πραγματοποίησε και να συμμετέχει άμεσα. Οι πλατφόρμες αυτές που επιτρέπουν στον χρήστη να αναλάβει τον ρόλο του κύριου οργανωτή της ταξιδιωτικής του εμπειρίας ποικίλουν στο είδος τους.

Υπάρχουν ιστοσελίδες που ασχολούνται με κριτικές καταστημάτων, εστιατορίων και άλλων επιχειρήσεων από χρήστες, όπως το GoogleMaps**,** ενώ άλλες αναφέρονται σε κρατήσεις

καταλυμάτων αλλά και σε αξιολογήσεις από πραγματικούς ταξιδιώτες όπως το Booking.com και η Airbnb. Μια ακόμα τέτοια κατηγορία αποτελεί η πλατφόρμα κριτικών σε ταξιδιωτικές κρατήσεις όπως το Trip.com και το TripAdvisor.

Το TripAdvisor έχει καταφέρει να συγκεντρώσει εκατομμύρια κριτικές που αφορούν τον ταξιδιωτικό κλάδο, όπως ξενοδοχεία, αξιοθέατα και μαγαζιά εστίασης σε όλον τον κόσμο και αποτελεί πλέον μια πλατφόρμα όπου οι χρήστες αξιοποιούν και εμπιστεύονται. Μετρά 350 εκατομμύρια μοναδικούς επισκέπτες ανά μήνα και περιέχει πάνω από 320 εκατομμύρια κριτικές που καλύπτουν καταλύματα, εστιατόρια και αξιοθέατα (TripAdvisor, 2016).

Για να μπορέσουν οι εταιρείες του ταξιδιωτικού κλάδου να αξιοποιήσουν όλα αυτά τα δεδομένα τα οποία προσφέρονται δωρεάν πρέπει να χρησιμοποιήσουν ορισμένες τεχνικές. Μια από αυτές είναι αυτή της ανάλυσης συναισθήματος. Η διαδικασία αυτή περιέχει διάφορα εργαλεία και τεχνικές όπου δίνουν τη δυνατότητα διαβάζοντας μια πρόταση ή ένα κείμενο να προσδιορίζεται η στάση του συγγραφέα. Υπάρχουν διάφοροι τρόποι που προσδιορίζονται τα συναισθήματα ανάλογα με τους αλγορίθμους και τις τεχνικές που θα χρησιμοποιηθούν (*What Is Sentiment Analysis? - Sentiment Analysis Explained - AWS*)

Από την άλλη πλευρά, η μηχανική μάθηση έχει σημειώσει ραγδαία εξέλιξη και χρησιμοποιείται πλέον σχεδόν σε όλους τους τομείς. Συλλέγονται δεδομένα από τους χρήστες των πλατφορμών και χρησιμοποιούνται ώστε να εκπαιδευτούν τα μοντέλα της μηχανικής μάθησης και να καταφέρουν να εξάγουν όλες αυτές τις πληροφορίες που χρειάζονται για την καλύτερη εξυπηρέτηση των χρηστών.

Οι (Roh et al., 2018) διερευνούν τις τεχνικές αυτές με τις οποίες γίνεται η συλλογή δεδομένων αλλά και τη δημιουργία ασθενών ετικετών στις περιπτώσεις όπου τα δεδομένα δεν επαρκούν ή είναι ελάχιστα. Οι συγγραφείς προτείνουν διάφορες μεθόδους όπως η εξαγωγή γεγονότων από γνωστικές βάσεις και η επανετικέλαση, ενώ μέσω ενός παραδείγματος κάνουν τον αναγνώστη να κατανοήσει τα στάδια τα οποία πρέπει να ακολουθηθούν για να εκπαιδευτεί ένα μοντέλο μηχανικής μάθησης. Η μελέτη τονίζει πιο έντονα κάποιες τεχνικές όπως, την κοινοποιητική ετικετοποίηση όπου παρουσιάζεται ως μια πρακτική τεχνική για μεγάλους όγκους δεδομένων αλλά δίνουν και ιδιαίτερη σημασία στην τεχνική της αυτό-ετικετοποίησης (self-labeling) και της επαναληπτικής ετικετοποίησης (self-training) ως λύσεις που είναι εξαιρετικά χρήσιμες όταν δεν υπάρχουν αρκετά ετικετοποιημένα δεδομένα για εκπαίδευση μοντέλων μηχανικής μάθησης.

Η συγκεκριμένη διπλωματική εργασία συνδυάζει τη μηχανική μάθηση και την ανάλυση συναισθήματος από διάφορα δεδομένα. Με την εύρεση ενός πλήρους και ποικιλόμορφου συνόλου δεδομένων, την ανάλυση των συναισθηματικών αντιδράσεων των χρηστών, μέχρι την εφαρμογή μοντέλων μηχανικής και βαθιάς μάθησης, η εργασία αυτή αποσκοπεί στην διεξαγωγή συμπερασμάτων μέσω προηγμένων αλγορίθμων για την πλήρη κατανόηση των αναγκών της αγοράς και τη βελτίωση πτυχών του τουριστικού τομέα. Πιο συγκεκριμένα, με την χρήση ενός έτοιμου συνόλου ταξιδιωτικών δεδομένων συμπεριλαμβάνοντας πληροφορίες όπως ονόματα και τοποθεσίες για τα ίδια τα ξενοδοχεία αλλά και διάφορες πληροφορίες για τους ίδιους τους χρήστες όπως το όνομα χρήστη, τις κριτικές για κάθε ξενοδοχείο αλλά και τις βαθμολογίες, θα αναλυθούν, θα επεξεργαστούν και θα οπτικοποιηθούν με στόχο την καλύτερη κατανόηση. Έπειτα θα εφαρμοστούν αλγόριθμοι μηχανικής και βαθιάς μάθησης οι οποίοι θα εκπαιδευτούν ώστε να προβλέπουν την βαθμολογία, με βάση το κείμενο της κριτικής. Τα αποτελέσματα των αλγορίθμων θα σχολιαστούν και θα συγκριθούν ώστε να συμπεράνουμε πιο μοντέλο εφαρμόζεται καλύτερα στα συγκεκριμένα δεδομένα συμβάλλοντας με αυτόν τον τρόπο στην καλύτερη αντίληψη της χρησιμότητας των μοντέλων για την ανάλυση και την κατανόηση των προτιμήσεων των χρηστών αλλά και την πιθανή πρόβλεψη των τουριστικών τάσεων.

1.1 Τουριστική Συμπεριφορά

Ο τρόπος με τον οποίο κάθε ταξιδιώτης θα βιώσει και έπειτα θα σχολιάσει την ταξιδιωτική του εμπειρία εξαρτάται από πολλούς παράγοντες. Σύμφωνα με τον (Parrinello, 1993) σημαντικό παράγοντα της τουριστικής συμπεριφοράς έχει η γνωστική και αισθητική προσωπικότητα του καθενός. Μετά από μελέτες παρατηρήθηκε λοιπόν πως για τα ίδια μέρη , παραδείγματος χάρη για ένα ξενοδοχείο, υπάρχουν διαφορετικές αξιολογήσεις (Parrinello,1993 , Tahir&Meltem, 2018)

Επίσης σημαντικό ρόλο έχει ο τύπος της πλατφόρμας όπου δημοσιεύονται τα σχόλια (Gretzelandyoo 2018). Για παράδειγμα, στα μέσα κοινωνικής δικτύωσης συνηθίζεται να δίνεται περισσότερη έμφαση στο κομμάτι της ψυχαγωγίας σε σύγκριση με τις ιστοσελίδες τουρισμού. Ο νέος τρόπος αποτύπωσης των εντυπώσεών τους ονομάζεται eWOM (electronic word-of-mouth) και είναι ολοένα και πιο σημαντικό κριτήριο στην βιομηχανία του τουρισμού. Περιέχει δηλαδή, τις αξιολογήσεις από τις εμπειρίες των ταξιδιωτών ή των καταναλωτών αλλά αποτελεί και έναν τρόπο ανταλλαγής απόψεων σχολίων και συστάσεων. Μαθαίνοντας λοιπόν για το eWOM των ταξιδιωτών στις διάφορες πλατφόρμες, ενισχύεται η

κατανόηση της ψυχολογίας και έπειτα η διεξαγωγή συμπερασμάτων από τους ειδικούς. (Zhou*etal.*, 2020)

Κατά την διάρκεια ενός ταξιδιού, κάθε ταξιδιώτης, ανάλογα την εμπειρία που βιώνει και τον χαρακτήρα που έχει, αναπτύσσει διάφορα συναισθήματα. Μερικά από αυτά είναι η λύπη, η χαρά, η αγωνία, ο ενθουσιασμός και η δυσαρέσκεια. Αυτά τα συναισθήματα επηρεάζουν την άποψή και την γνώμη που αναπτύσσει για την εμπειρία του.Διάφορες μελέτες όπως του Kim Uysal (2013), αναζητούν και εξηγούν το πώς επηρεάζεται η ικανοποίηση των ταξιδιωτών από τα συναισθήματα που έχουν αναπτύξει κατά την διάρκεια των ταξιδιών τους. Η καλύτερη κατανόηση των παραπάνω, βοηθάει στην πιο σωστή ανάλυση συναισθημάτων βελτιώνοντας τις στρατηγικές μάρκετινγκ και ενισχύοντας την ταξιδιωτική εμπειρία.

1.2 Ανάλυση συναισθήματος σε ταξιδιωτικά δεδομένα

Οι τεχνολογικές αλλαγές του διαδικτύου, και η ψηφιοποίηση έχουν επιφέρει ραγδαίες αλλαγές στην τουριστική βιομηχανία. Οι ταξιδιώτες έχουν πλέον την δυνατότητα να διατυπώσουν και να αναρτήσουν τα δικά τους σχόλια για την ταξιδιωτική τους εμπειρία (Neidhardt et al., 2017; Yang et al., 2017; Ye et al., 2009) που αποτελούν πιο αξιόπιστες πηγές συγκριτικά με τους ιστότοπους των εταιρειών (Akehurst, 2009).

Μετά από αρκετές μελέτες αποδεικνύεται ότι τώρα πια οι περισσότεροι αναλυτές τουρισμού αντλούν τα δεδομένα τους είτε από επαγγελματικές εφαρμογές τουρισμού, για παράδειγμα το TripAdvisor, είτε από μέσα κοινωνικής δικτύωσης όπως το Twitter (Alaei et al., 2019). Στις δύο παραπάνω περιπτώσεις το κείμενο που αποτελεί πηγή δεδομένων είναι αρκετά μικρό. Σύμφωνα με διάφορα άρθρα, το όριο των 140 χαρακτήρων στα tweets έχει αυξηθεί στους 280 χαρακτήρες για απλούς χρήστες.

Έτσι λοιπόν δημιουργείται η ανάγκη ύπαρξης πιο αποτελεσματικής ανάλυσης, αφού πλέον οι αναλυτές έχουν να αντιμετωπίσουν δεδομένα μικρά σε όγκο και μη δομημένα. Οι (Shimada et al., 2011) βασιζόμενοι στον αλγόριθμο μηχανικής μάθησης NaïveBayes, κατάφεραν να τον εκπαιδεύσουν χρησιμοποιώντας δεδομένα χωρίς ετικέτες. Αντικατέστησαν τις λέξεις με συναισθήματα « χαμογελαστό πρόσωπο» για τα θετικά και « λυπημένο πρόσωπο» για τα αρνητικά ταξινομώντας έτσι τις κριτικές στις δύο παραπάνω κατηγορίες.

Μια διαφορετική προσέγγιση υιοθέτησαν οι (Misopoulos et al., 2014) με τη μέθοδο λεξικού, αξιολογώντας κριτικές Twitter μιας αεροπορικής εταιρείας. Αντί να χρησιμοποιήσουν την

χρήση συμβόλων, η ανάλυσή τους βασίστηκε στην πολικότητα κατηγοριοποιώντας την σε αρνητικά, θετικά και ουδέτερα συναισθήματα. Παρόλα αυτά, η παραπάνω μέθοδος βασίστηκε σε 20 λέξεις – κλειδιά κάτι που μπορεί να μην καλύπτει πλήρως τα συναισθήματα των χρηστών.

Σύμφωνα με έρευνες (Buhalis & Law, 2008) αποδεικνύεται ότι πλέον ο τουρισμός επηρεάζεται όχι μόνο από τις αποφάσεις των καταναλωτών αλλά και από την εξέλιξη της τεχνολογίας. Η ταχεία ανάπτυξη του διαδικτυακού τουρισμού καθιστά αναγκαία τη δημιουργία νέων τεχνολογιών ώστε να υπάρχει καλύτερη διαχείρισή των νέων δεδομένων που προκύπτουν .

Ένα από αυτά τα πολλά εργαλεία που έχουν αναπτυχθεί είναι η ανάλυση συναισθήματος. Η ανάλυση αυτή παρέχει ακρίβεια και ταχύτητα και με την χρήση διαφόρων μοτίβων κάνει το κείμενο κατανοητό στους ειδικούς (Thelwall, 2019). Έτσι πλέον η τεχνολογία έχει την δυνατότητα να επεξεργαστεί μεγάλο όγκο δεδομένων αλλά και να τον κατανοήσει, βοηθώντας έτσι αποτελεσματικά στην λήψη αποφάσεων.

2. Θεωρητικό υπόβαθρο

2.1 Ανάλυση Συναισθήματος

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing – NLP) αποτελεί τη βασική προσέγγιση ώστε να γίνει αντιληπτή η ανθρώπινη επικοινωνία βασιζόμενη στον κλάδο της ψυχολογίας και της ψυχιατρικής (Brockopp, 1983).

Η ανάλυση συναισθήματος, αποτελεί ένα πεδίο της Επεξεργασίας Φυσικής Γλώσσας και έχει ως σκοπό να μπορέσει να κατανοήσει τη συναισθηματική κατάσταση ενός χρήστη από τα κείμενά του (Akhtar et al., 2020). Συγκεντρώνει και αναλύει τις απόψεις και τις εντυπώσεις των χρηστών (Roh, Heo and Whang, 2018) και παράλληλα προβλέπει τον συναισθηματικό τόνο και την κατεύθυνση του κειμένου. Η τομή μεταξύ του ανθρώπινου συναισθήματος και του υπολογιστή ονομάστηκε Συναισθηματική Υπολογιστική (Affective Computing) και ερευνήθηκε την δεκαετία του 1990 από την Rosalind Picard (Picard, R.W,2003). Σκοπός της Συναισθηματικής Υπολογιστικής είναι να μπορέσει να υλοποιήσει ένα τέτοιο σύστημα το οποίο θα διευκολύνει την αλληλεπίδραση ανθρώπου-υπολογιστή δημιουργώντας έτσι υπολογιστικά συστήματα ικανά να αναγνωρίζουν και να εκφράζουν συναισθήματα (What Is Interaction Design (IxD)— Updated 2024 | IxDF) Πρωταρχικός ρόλος αυτών των υπολογιστικών συστημάτων είναι να μπορέσουν να αντιληφθούν και να ερμηνεύσουν συναισθήματα.

Σε αυτό βοήθησε το μοντέλο συναισθημάτων του Ekman όπου κατηγοριοποίησε τα ανθρώπινα συναισθήματα σε έξι βασικές κατηγορίες. Θυμός, αηδία, φόβος, χαρά, λύπη, έκπληξη (Ekman.P, 1999). Το μοντέλο αυτό, μεταγενέστερα χρησιμοποιήθηκε σε πολλές μελέτες ώστε να αναγνωριστεί η συναισθηματική κατάσταση από συστήματα.

Τα αποτελέσματα μετά την ανάλυση συναισθήματος ενός κειμένου εξαρτώνται από το είδος αυτού. Η ανάλυση συναισθήματος έχει τη δυνατότητα να εφαρμοστεί σε όλα τα είδη κειμένων (Mohammad, 2015). Διάφοροι ερευνητές σύμφωνα με τον Saif Mohammad ερεύνησαν την ανάλυση συναισθήματος σε διαφορετικά είδη κειμένων.

Παρόλα αυτά, τα δεδομένα από το twitter είναι από τα πιο διαδεδομένα κείμενα για την εφαρμογή συναισθηματικής ανάλυσης (Chikersal et al., 2015).Ο τεράστιος αριθμός πληροφοριών που έχουν αναρτηθεί στο διαδίκτυο, καθιστά την συναισθηματική ανάλυση απαραίτητη.

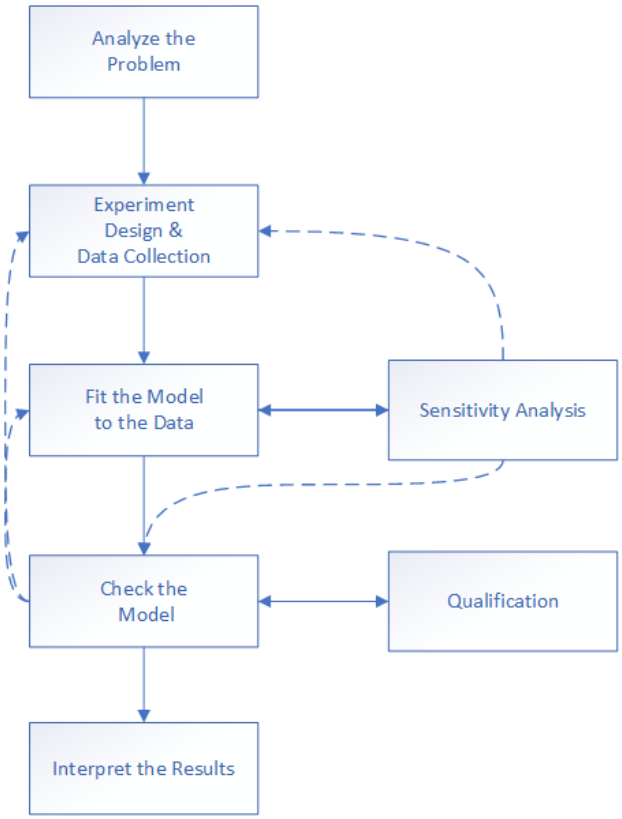
Τα μοντέλα συναισθηματικής ανάλυσης δεν περιορίζονται μόνο στην αναγνώριση μίας γλώσσας, αλλά σε πληθώρα αυτών. Σε αυτές συμπεριλαμβάνονται και τα ελληνικά εφόσον διάφορες ερευνητικές μελέτες έχουν γίνει πάνω σε κείμενα της ελληνικής γλώσσας (Spatiotis et al., 2017).

2.2 Εισαγωγή στην ανάλυση δεδομένων και διαδικασίες ανάλυσης κειμένου

Η ανάλυση δεδομένων (Data analysis) αποτελεί μία διαδικασία στην οποία για την διεξαγωγή χρήσιμων πληροφοριών, πρέπει πρώτα τα δεδομένα που έχουμε να επεξεργαστούν και να τροποποιηθούν κατάλληλα. Ο στόχος της ανάλυσης δεδομένων είναι να αποβεί εφικτό να ληφθούν αποφάσεις από την διεξαγωγή των συμπερασμάτων. Ο Daniel Johnson, για να κάνει πιο κατανοητή την έννοια δίνει ένα παράδειγμα αναφέροντας, πως στην καθημερινή ζωή οι άνθρωποι για να αποφασίσουν κάτι που αφορά το μέλλον, επεξεργάζονται μια παρόμοια κατάσταση που έχει γίνει στο παρελθόν. Για να γίνει αυτό, απαραίτητη προϋπόθεση είναι να έχουν τις αναμνήσεις τους. Το ίδιο ακριβώς κάνει και ο αναλυτής με τα δεδομένα και αυτό είναι ουσιαστικά η ανάλυση των δεδομένων (Daniel Johnson, 2024)

Η ανάλυση των δεδομένων αποτελείται από μια διαδικασία πολλαπλών βημάτων. Αρχικά περιλαμβάνει την ανάλυση του προβλήματος ώστε να είναι ξεκάθαρο το πρόβλημα που αντιμετωπίζεται και ποιος είναι ο σκοπός . Έπειτα ακολουθεί συλλογή δεδομένων που με βάση την παραπάνω ανάλυση διαλέγονται τα απαραίτητα δεδομένα. Ακολουθεί η προσαρμογή του μοντέλου στα δεδομένα που σε αυτό το στάδιο τα παραπάνω δεδομένα χρησιμοποιούνται για την ανάπτυξη του μοντέλου. Γίνεται επίσης η ανάλυση ευαισθησίας και η ποιοτική αξιολόγηση του μοντέλου και στο τέλος ακολουθεί ο έλεγχος του. Η διαδικασία ολοκληρώνεται με την ερμηνεία των αποτελεσμάτων που έχουν προκύψει.

Ακολουθεί η εικόνα με τα βήματα της διαδικασίας ανάλυσης δεδομένων που περιγράψαμε παραπάνω:



Εικόνα 1 Διαδικασία Ανάλυσης Δεδομένων

Για να μπορέσει η ανάλυση των δεδομένων να γίνει σωστά, είτε είναι σε κείμενο είτε σε φράσεις είτε σε μεμονωμένες λέξεις είναι αναγκαίο όλα τα δεδομένα να έρθουν στην κατάλληλη μορφή. Για αυτό το λόγο όπου χρειάζεται εφαρμόζεται το στάδιο της προ-επεξεργασίας κειμένου, ώστε τα μοντέλα να μπορούν να επεξεργαστούν κατάλληλα τα δεδομένα .

2.2.1 Συλλογή δεδομένων

Αποτελεί το πρώτο και ένα από τα πιο σημαντικά βήματα ώστε να ακολουθήσει η ανάλυση των δεδομένων. Θα πρέπει να συλλεχθούν κατάλληλα σύνολα δεδομένων ώστε να χρησιμοποιηθούν αργότερα για την επίτευξη των στόχων. Τα δεδομένα αυτά είτε είναι δεδομένα ερευνών, ιστότοπων κοινωνικής δικτύωσης είτε κείμενα από βιβλία, άρθρα κ.α

2.2.2 Καθαρισμός και προ επεξεργασία κειμένου

Τα περισσότερα κείμενα που επιλέγονται για την εκπαίδευση των μοντέλων περιέχουν μέσα θόρυβο. Αυτό σημαίνει ότι είτε περιέχουν λέξεις όπως προθέματα (stop words), ορθογραφικά λάθη, σημεία στίξης είτε μπορεί να λείπουν τελείως λέξεις ή να είναι ανεπιθύμητες αλλοιώνοντας το νόημα. Έτσι επιτυγχάνεται η μείωση του θορύβου και δεν

επιβαρύνεται η διαδικασία της ταξινόμησης με δεδομένα τα οποία δεν συνεισφέρουν στην επίτευξη του επιθυμητού αποτελέσματος .

2.2.2.1 Καθαρισμός δεδομένων

Αποτελεί ένα κρίσιμο βήμα για την προ-επεξεργασία. Σε αυτό το σημείο γίνεται η αφαίρεση ελλιπών τιμών (stop words), η αφαίρεση διπλότυπων εγγραφών και η κανονικοποίηση των δεδομένων. Επιπλέον, αν τα ίδιου τύπου δεδομένα είναι καταχωρισμένα με διαφορετικούς τρόπους, ενοποιούνται.

2.2.2.2 Tokenization και μετατροπή σε πεζά.

Το tokenization (διαχωρισμός σε λέξεις) επιτρέπει την διάσπαση του κειμένου σε μικρότερες μονάδες, τα tokens. Έτσι, επιτρέπεται στα μοντέλα να χειριστούν και να επεξεργαστούν το κείμενο με ευκολία και μικρότερη πιθανότητα λαθών. Στη διάσπαση των tokens συμπεριλαμβάνονται και τα σημεία στίξης. Από την άλλη η διαδικασία της μετατροπής σε πεζούς χαρακτήρες βοηθάει ώστε να μην υπάρξει κατηγοριοποίηση της ίδιας λέξης διπλές φορές.

2.2.2.3 Stemming και Lemmatization

Σε ένα κείμενο η πιθανότητα να εντοπιστεί η ίδια λέξη με πολλές μορφές είναι αρκετά μεγάλη. Η διαδικασία του Stemming εντοπίζει τις λέξεις αυτές και βρίσκει την ρίζα τους. Σκοπός του δηλαδή είναι η μετατροπή της κάθε λέξης στη βασική της μορφή. Πχ “reports” “report ή “τα good, better, best” λημματοποιούνται σε “good, good, good”). Η διαδικασία του Lemmatization είναι αρκετά παρόμοια με την παραπάνω, αλλά σε αυτή την περίπτωση ψάχνει την σωστή λεξική μορφή της κάθε λέξης, σε αντίθεση με την διαδικασία stemming που απλά κόβει καταλήξεις .

2.2.2.4 Term Frequency Inverse Document Frequency (TF/IDF)

Το IDF αποτελεί μια αρκετά γνωστή τεχνική, στατιστικής ανάλυσης που αποκαλύπτει την σημασιολογική βαρύτητα μιας λέξης σε μια συλλογή εγγράφων (V. Mohan,2015). Ουσιαστικά εντοπίζει πόσο συχνά εμφανίζεται μια λέξη μέσα σε ένα κείμενο με σκοπό να αξιολογηθεί η σημασία της στο κείμενο αυτό.

Υπολογίζεται με το εξής τύπο :

$$IDF(w) = log\ log\ \left(\frac{N}{n_w}\right) \qquad (1)$$

2.2.2.5 Μοντέλο Bag of Words

Σε αυτό το μοντέλο (BOW) παραβλέπεται η γραμματική και η σειρά των λέξεων, αλλά διατηρείται η πολλαπλότητα τους. Σκοπός είναι η μετατροπή των κειμένων σε αριθμητικές μορφές ώστε να τοποθετηθούν σε διανύσματα. Η καταχώριση στο διάνυσμα αντιπροσωπεύει την εμφάνιση μιας λέξης σε ένα κείμενο. Όταν η λέξη βρίσκεται στο κείμενο τότε εντοπίζεται με 1 αλλιώς με 0. Το σύνολο των λέξεων που διατηρήθηκε ονομάζεται vocabulary, αφού σε κάθε λέξη αντιστοιχεί ένας μοναδικός αριθμός.

3. Μηχανική Μάθηση και Deep Learning

3.1 Μηχανική Μάθηση

Η ανάλυση συναισθήματος που αφορά την διαδικασία αναγνώρισης και κατηγοριοποίησης συναισθημάτων από κείμενο ή ήχο, έχει αποκτήσει πλέον σημαντική θέση εξαιτίας των πλεονεκτημάτων εφαρμογής αλγορίθμων μηχανικής μάθησης.

Η Μηχανική μάθηση (Machine Learning) αποτελεί ένα πεδίο της τεχνιτής νοημοσύνης η οποία βασίζεται στην άποψη ότι τα συστήματα έχουν την δυνατότητα να μαθαίνουν από δεδομένα, να ορίζουν με ακρίβεια πρότυπα και να λαμβάνουν αποφάσεις με ελάχιστη ανθρώπινη εμπλοκή. Σκοπός της είναι να μπορέσουν να δημιουργηθούν μηχανές ικανές να «μαθαίνουν» από δεδομένα χωρίς να είναι ρητά προγραμματισμένες , αλλά μέσω της αξιοποίησης προηγούμενης γνώσης και εμπειρίας.

Ένας σχετικός γενικός ορισμός Μηχανικής Μάθησης δίνεται από τον Mitchell (1997):

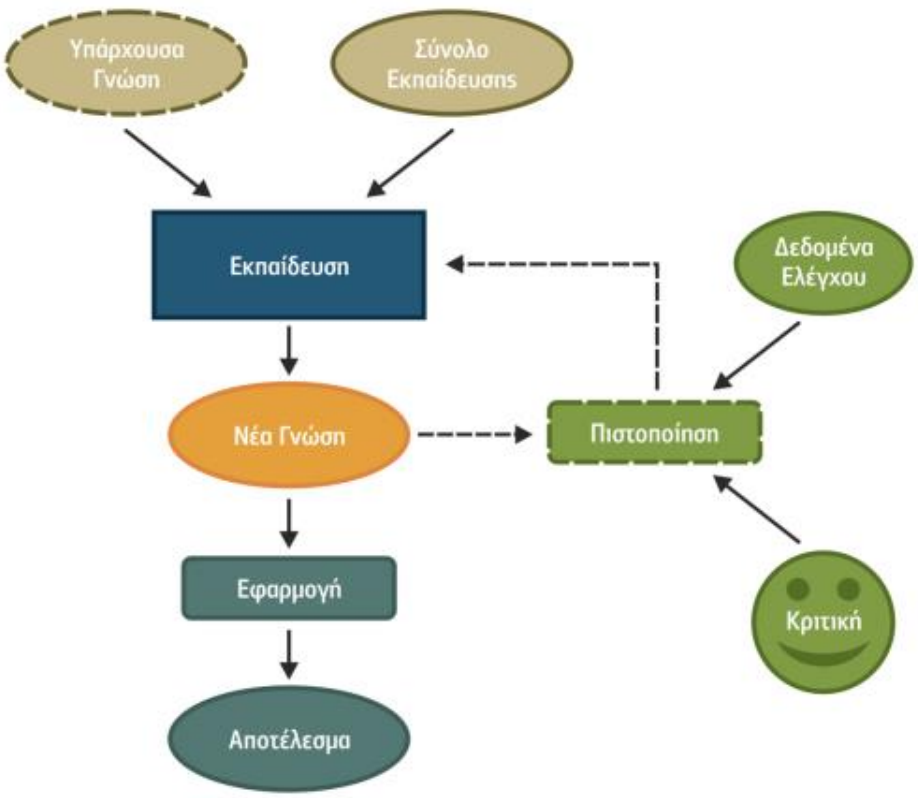
> «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες από το T, όπως μετριέται από το P, βελτιώνεται μέσω της εμπειρίας E.»

Η μηχανική μάθηση ενσωματώνεται ολοένα και περισσότερο στις βιομηχανίες αυτοματοποιώντας τις χειροκίνητες εργασίες και κάνοντας τον τρόπο ζωής όλο και πιο εύκολο. Διακρίνεται σε τρεις βασικές κατηγορίες: Εποπτευόμενη μάθηση (Supervised Learning), μη επιβλεπόμενη μάθηση (Unsupervised Learning) και ενισχυτική μάθηση (Reinforcement Learning).

Όσον αφορά την ανάλυση κειμένου χρήσιμη είναι η εφαρμογή τεχνικών όπως το TF-IDF (Term Frequency-Inverse Document Frequency), το οποίο εφαρμόζεται σε κειμενικά δεδομένα με σκοπό την μετατροπή τους σε αριθμητική μορφή ώστε να είναι αναγνωρίσιμα από αλγορίθμους μηχανικής μάθησης. Ουσιαστικά η παραπάνω τεχνική μετράει τις λέξεις δίνοντας αξία στην συχνότητα εμφάνισής τους. Επομένως οι πιο σπάνιες λέξεις και πιθανόν οι πιο σημαντικές αποκτούν μεγαλύτερη βαρύτητα σε αντίθεση με τις λέξεις που εμφανίζονται πιο συχνά, ο οποίες μπορεί να αποτελούν άρθρα ή συνδετικές λέξεις. Με αυτόν τον τρόπο οι αλγόριθμοι μηχανικής μάθησης βελτιώνουν την απόδοσή τους καθώς δίνουν έμφαση σε σημαντικά μοτίβα.

Η εποπτευόμενη μάθηση χρησιμοποιεί ως είσοδο δεδομένα που έχουν επισημανθεί με ετικέτες (σύνολο εκπαίδευσης) για την εκπαίδευση ενός μοντέλου στοχεύοντας στην ακριβή πρόβλεψη και κατηγοριοποίηση νέων δεδομένων. Από την άλλη, η μη-επιβλεπόμενη μάθηση χρησιμοποιεί αλγορίθμους που κατασκευάζουν ένα μοντέλο με κάποιο σύνολο δεδομένων χωρίς να γνωρίζουν τις επιθυμητές εξόδους. Η ενισχυτική μάθηση χρησιμοποιεί την εκπαίδευση μοντέλων, μαθαίνοντας μέσα από την αλληλεπίδραση με το περιβάλλον μια στρατηγική ενεργειών.

Στην παρακάτω εικόνα απεικονίζεται η γενική διαδικασία που ακολουθεί ένας αλγόριθμος μηχανικής μάθησης χρησιμοποιώντας ως είσοδο ένα σύνολο εκπαίδευσης ώστε να επιτευχθεί ο σκοπός του και να εξάγει νέα γνώση.



Εικόνα 2 Γενικός τρόπος λειτουργίας των αλγορίθμων Μηχανικής Μάθησης.Γεωργούλη, Α. (2015). Μηχανική Μάθηση

Εμβαθύνοντας, οι αλγόριθμοι πραγματεύονται τη μείωση της διάστασης, την ομαδοποίηση, την πρόβλεψη και την ενισχυτική μάθηση. Η μείωση της διάστασης αποτελεί μια από τις βασικές τεχνικές μη-επιβλεπόμενης μάθησης, συμπιέζοντας τα δεδομένα, μειώνοντας τις

διαστάσεις τους και διατηρώντας όσο το δυνατόν περισσότερες από τις σημαντικές πληροφορίες. (Tatsat, Puri, & Lookabaugh, 2020).Γνωρίζοντας πάντα, το πιθανό ενδεχόμενο απώλειας πληροφοριών επιδιώκεται να επιλέγονται οι σωστές τεχνικές ώστε να υπάρχει ισορροπία.

Στην ίδια κατηγορία βρίσκεται και η ομαδοποίηση (clustering). Σε αυτή την περίπτωση ο αλγόριθμος προσπαθεί να ομαδοποιήσει τα δεδομένα, με βάση τις ομοιότητες μεταξύ τους, εντοπίζοντας κρυφές δομές οι οποίες μπορεί να μην ήταν εμφανείς στην αρχή .

Επόμενη εφαρμογή της μηχανικής μάθησης πάνω σε δεδομένα αποτελεί η πρόβλεψη. Σε αυτή την περίπτωση γίνεται χρήση μοντέλων πάνω στα υπάρχοντα δεδομένα που είναι ήδη εκπαιδευμένα με σκοπό είτε να προβλέψουν μελλοντικά γεγονότα είτε να αξιολογήσουν τα δεδομένα προβλέποντας τη συναισθηματική κατάσταση που ακολουθεί. (Liu, B., Liu, J., Shen, J., & Li, Z. 2020).

Τέλος, η ενισχυτική μάθηση, βασίζεται σε έναν αλγόριθμο βοηθό (agent) που μαθαίνει να παίρνει αποφάσεις με βάση το περιβάλλον του αλληλοεπιδρώντας με αυτό. Πιο αναλυτικά το μοντέλο αξιολογεί με βάση ανταμοιβές και τιμωρίες (εκτελώντας ενέργειες, αξιολογεί τα αποτελέσματά του και ανάλογα λαμβάνει θετική ή αρνητική αξιολόγηση την οποία θα αξιοποιήσει σε μελλοντική πρόβλεψη)

3.1.1 Αλγόριθμοι Μηχανικής Μάθησης

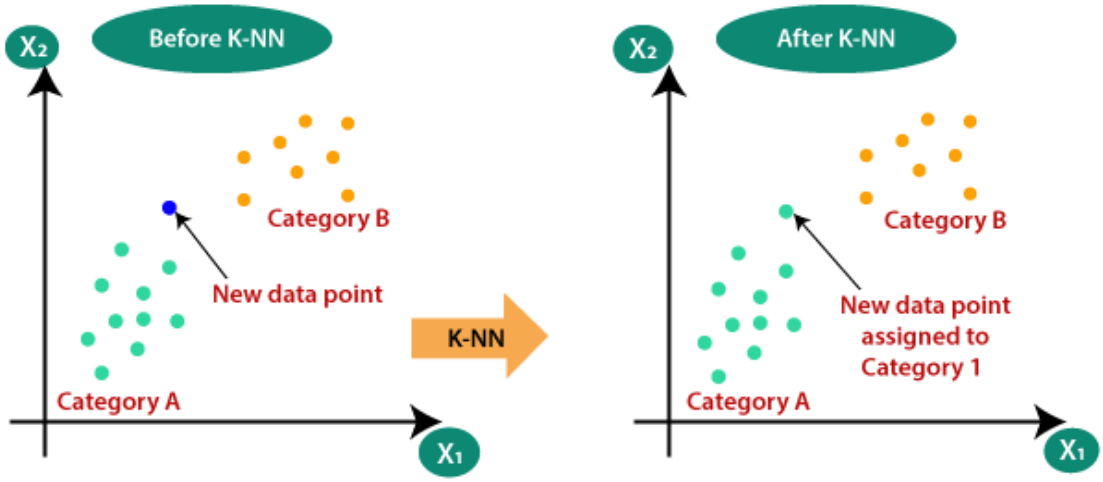
K-Nearest Neighbors (KNN)

Ο αλγόριθμος K-Nearest Neighbors χαρακτηρίζεται από τους πιο γνωστούς αλλά και απλούς αλγόριθμους ταξινόμησης και παλινδρόμησης. Βασίζεται στη μνήμη ή στα στιγμιότυπα ενώ η ταξινόμηση των νέων αντικειμένων γίνεται με βάση τα κοινά χαρακτηριστικά τους με τα ήδη γνωστά δεδομένα εκπαίδευσης. Αποτελεί μη παραμετρικό μοντέλο και ένα από τα μεγαλύτερα πλεονεκτήματά του είναι η ικανότητα αποτελεσματικής διαχείρισης μεγάλων συνόλων δεδομένων αλλά και η απλότητά του. Βέβαια το γεγονός πως απαιτεί σάρωση όλων των δεδομένων εκπαίδευσης κάθε φορά που χρειάζεται να ταξινομηθεί μπορεί να γίνει αρκετά χρονοβόρο ειδικά όταν έχει να κάνει με ένα μεγάλο σύνολο δεδομένων (Acito, F. 2023).

Ο KNN αρχικά προτάθηκε το 1951 από του Fix και Hodges (Evelyn F, Hodges JL Jr .1951)ενώ στην πορεία τροποποιήθηκε από τους Cover και Hart (Cover TM, Hart P (1967).

Η λογική του αλγορίθμου εξαρτάται ουσιαστικά από τον υπολογισμό των αποστάσεων ανάμεσα στα δύο σύνολα δεδομένων (του ήδη ελεγμένου και του ελεγχόμενου του δείγματος).Στο τέλος το δείγμα τοποθετείται στην κλάση του πλησιέστερου γείτονα. (Ali, Neagu, & Trundle, 2019)

Η χρήση του αλγορίθμου KNN χρήζει την καλή προ επεξεργασία δεδομένων και την χρήση σωστών παραμέτρων για τη βελτιστοποίηση της απόδοσής του ενώ έτσι καθίσταται ιδανικός αλγόριθμος για προβλήματα κατηγοριοποίησης όπου υπάρχει άμεση σύγκριση με άλλα δεδομένα.



Εικόνα 3 Πριν και μετά την εφαρμογή του KNN (Javapoint)

Naive Bayes

Ο Naive Bayes είναι ένας αλγόριθμος ταξινόμησης που βασίζεται στη θεωρία των πιθανοτήτων. Παρά την σημασία του ονόματός του (αθώος) αποτελεί έναν πολύ αποτελεσματικό και χρήσιμο αλγόριθμο καθώς χρησιμοποιείται για μεγάλη ποικιλία πραγματικών εφαρμογών. Συνηθίζεται να εφαρμόζεται στην κατηγοριοποίηση κειμένων, και στη συναισθηματική ανάλυση, προσδιορίζοντας αν το κείμενο έχει θετική, αρνητική ή ουδέτερη διάθεση ακόμα και διάγνωση ασθενειών με βάση τα διαγνωσμένα χαρακτηριστικά (WICKRAMASINGHE, I. and KALUTARAGE, H. 2021)

Ο Άγγλος Thomas Bayes ήταν ο πρώτος που ανέπτυξε τη θεωρία το 1701-1761 , η οποία περιγράφει τη διαδικασία υπολογισμού της πιθανότητας μιας υπόθεσης δεδομένων νέων στοιχείων(Price, R. 2007)

Η βασική ιδέα πίσω από τον αλγόριθμο είναι ο υπολογισμός της πιθανότητας ενός δεδομένου να ανήκει σε μια συγκεκριμένη κλάση χρησιμοποιώντας την πιθανότητα της κλάσης με δεδομένα με βάση τα χαρακτηριστικά και την περίπτωση. (Atmadja, Uriawan, Pritisen, Maylawati, & Arbain, 2019). Υπολογίζει στην συνέχεια τις πιθανότερες της κάθε κλάσης και διαλέγει αυτή με την υψηλότερη.

Ο αλγόριθμος βασίζεται στον εξής τύπο:

$$P(H\backslash E) = \frac{\mathrm{P(H\backslash E)*P(H)}}{\mathrm{P(E)}} \quad (2)$$

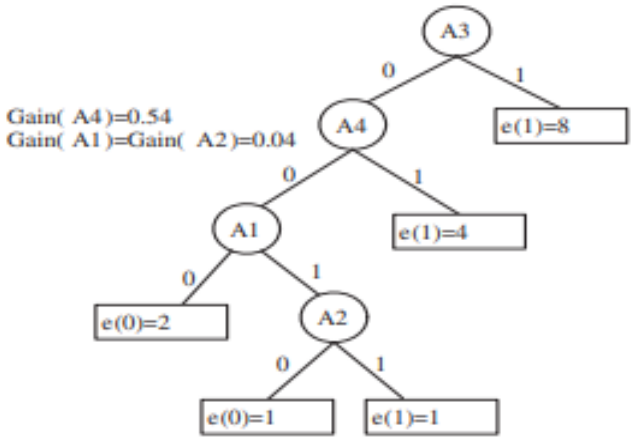
Decision Tree

Τα Δέντρα απόφασης είναι ένας αλγόριθμος μηχανικής μάθησης τα οποία χρησιμοποιούνται για τη δημιουργία συστημάτων ταξινόμησης που στηρίζονται σε πολλές μεταβλητές ή για τη δημιουργία αλγορίθμων πρόβλεψης για μια μεταβλητή στόχο. Αποτελείται από την Ρίζα (root) που αντιπροσωπεύει το σύνολο των δεδομένων, τους εσωτερικούς κόμβους (Internal Nodes) που αφορούν τις ερωτήσεις για τα χαρακτηριστικά των δεδομένων και τέλος τα φύλλα (Leaves) που αποτυπώνουν τις τελικές αποφάσεις ή κατηγορίες. Η δυνατότητα να χειρίζεται δεδομένα μικτού τύπου και η απλότητα του, αποτελούν μερικά από τα βασικά του πλεονεκτήματα.(J Su, H Zhang - Aaai, 2006).

Βασική ιδέα του αλγορίθμου, είναι να ταξινομεί τα δεδομένα σε τμήματα που μοιάζουν με κλάδους που κατασκευάζουν ένα ανεστραμμένο δέντρο. Ξεκινάει από την ρίζα του δέντρου και επαναλαμβάνει την διαδικασία έως ότου βρεθούν κριτήρια τερματισμού. Όταν το μέγεθος του δείγματος είναι αρκετά μεγάλο, τα δεδομένα της μελέτης μπορούν να χωριστούν σε σύνολα δεδομένων εκπαίδευσης και επικύρωσης.(Loh, W.-Y. 2015). Έχουν τη δυνατότητα είτε να κατηγοριοποιήσουν τα δεδομένα, παίρνοντας αποφάσεις σε κάθε εσωτερικό κόμβο κάνοντας τη διαδρομή από τη ρίζα προς τα φύλλα, είτε προβλέποντας αριθμητικές τιμές, με τα φύλλα του δέντρου να περιέχουν τις μέσες τιμές των στόχων .

Τα δέντρα αποφάσεων πρωτοεμφανίστηκαν το 1960 (Song & Lu, 2015)και αποτελούσαν τους πρώτους αλγορίθμους σε ηλεκτρονική μορφή κατά την υιοθέτηση ψηφιακών κυκλωμάτων στους ηλεκτρονικούς υπολογισμούς, ενώ αποτελούν πλέον από τους πιο βασικούς αλγορίθμους εξόρυξης δεδομένων. (De Ville, 2013).Αρχικά, πρωτοεμφανίστηκαν στον τομέα της στατιστικής ενώ έχουν επεκταθεί σε πολλές εφαρμογές ακόμα και στην ιατρική. Παράδειγμα ιατρικής χρήσης των δέντρων απόφασης αποτελεί η διάγνωση μιας ιατρικής κατάστασης, που οι κατηγορίες που το δέντρο απόφασης ορίζει θα μπορούσαν είτε να είναι

διαφορετικές κλινικές κατηγορίες ή μια πάθηση, είτε διάφοροι ασθενείς με μια πάθηση που θα έπρεπε να λάβουν διαφορετικές θεραπείες ο καθένας (Fallon et al., 2013)



Εικόνα 4 Δέντρο απόφασης που δημιουργήθηκε για Boolean σύστημα (Su, J., & Zhang, H)

3.2 Βαθιά μάθηση

Μεταξύ των διαφόρων αλγορίθμων μηχανικής μάθησης, η βαθιά μάθηση (deep learning,DL) αποτελεί ένα βασικό υποπεδίο. Έχει τις ρίζες της στα συμβατικά νευρωνικά δίκτυα και καταφέρνει να ξεπερνά τους προκατόχους της.(Archana & Jeevaraj, 2024).Με την χρήση των νευρωνικών δικτύων η βαθιά μάθηση επιτρέπει να αποκτήσουν πολύπλοκα δεδομένα ώστε να μπορούν να προβλέψουν και να λάβουν αποφάσεις χωρίς να υπάρχει απαραίτητα ανθρώπινη παρέμβαση. Αποτελεί πλέον ευρέως γνωστή τεχνική και τίθεται σε εφαρμογή σε διάφορους τομείς τεχνητής νοημοσύνης, όπως π.χ σημασιολογική ανάλυση , μεταφορά μάθησης , φυσική γλώσσα, επεξεργασία και όραση υπολογιστή.(Guo et al., 2016)

Η αρχική έννοια των τεχνητών νευρωνικών δικτύων (Artificial Neural Network, ANN) το 1943 ήταν ως μαθηματικό μοντέλο ενός τεχνητού νευρώνα. Το 2006 η βαθιά μάθηση προτάθηκε ως ANN με πολλά επίπεδα, η οποία έχει σημαντική ικανότητα μάθησης.(Talaei Khoei et al., 2023).Εκεί διαφοροποιείται με τα παραδοσιακά νευρωνικά δίκτυα ,αφού τα πολλά επίπεδα δίνουν την δυνατότητα να αναγνωρίζουν σύνθετα μοτίβα που δεν ήταν εφικτό παλαιότερα. Υπάρχουν διάφορες κατηγορίες νευρωνικών δικτύων που εκμεταλλεύεται η βαθιά μάθηση ώστε να γίνει η ανάλυση και η επεξεργασία πολύπλοκων δεδομένων ανάλογα με τις ανάγκες κάθε φορά .

Κάποιες από τις πιο βασικές αρχιτεκτονικές είναι ο Long Short-Term Memory (LSTM) και τα Recurrent Neural Networks (RNNs) που αποτελούν τους κατάλληλους αλγόριθμους επεξεργασίας σειρών δεδομένων και φυσικής γλώσσας και η αρχιτεκτονική των Transformers που παρόλο που είναι αρκετά πρόσφατη, χρησιμοποιώντας μηχανισμούς προσοχής έχει καταφέρει να καινοτομήσει.

Σε αυτό το πλαίσιο, για την εφαρμογή των αλγορίθμων βαθιάς μάθησης, κρίνεται αναγκαία η χρήση διαφόρων τεχνικών όπως το Spatial Dropout. Σε αντίθεση με το παραδοσιακό Dropout όπου κατα την εκπαίδευση του μοντέλου απενεργοποιούνται επιλεκτικά ολόκληροι νευρώνες με σκοπό την υπερπροσαρμογή του μοντέλου στα δεδομένα, η πιο εξελιγμένη αυτή τεχνική επιλέγει να αποβάλει ολόκληρα χαρακτηριστικά χάρτες (feature maps) ώστε το μοντέλο να γίνει ακόμα πιο ανθεκτικό.

Επιπλέον, τα μοντέλα βαθιάς μάθησης λειτουργούν με την επιλογή του κατάλληλου Batch Size το οποίο κατα την εκπαίδευση του μοντέλου ρυθμίζει την ποσότητα δειγμάτων εκπαίδευσης σε κάθε επανάληψη. Επομένως, όσο πιο σωστή είναι η ποσότητα τον δειγμάτων για κάθε μοντέλο, θα αυξάνεται η ακρίβεια και η ταχύτητα της εκπαίδευσης.

3.2.2 Αλγόριθμοι Βαθιάς μάθησης

Recurrent Neural Networks (RNN)

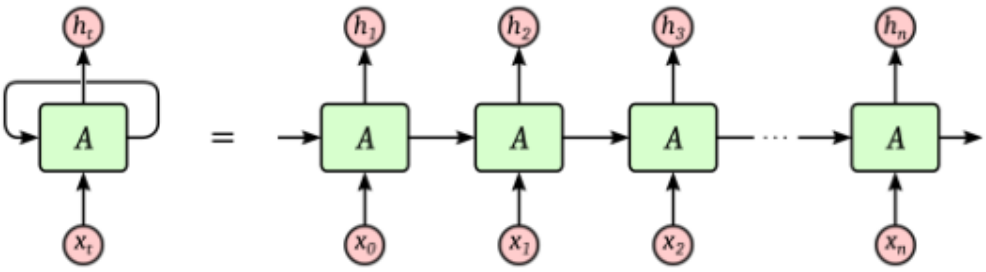
Τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNNs) είναι ένας τύπος νευρωνικού δικτύου το οποίο έχει την δυνατότητα να επεξεργάζεται δεδομένα που δεν έχουν σταθερό μήκος, είναι δηλαδή μεταβλητό είτε στην είσοδο, είτε στην έξοδο. Έτσι αποδίδουν πολύ καλά σε διαδοχικά δεδομένα, όπως δεδομένα χρονοσειρών ή την επίλυση εργασιών του NLP (Banerjee et al., 2019). Βασικό προτέρημα του παραπάνω δικτύου αποτελεί η αναδρομική του φύση, καθώς έχοντας την δυνατότητα να διατηρεί μνήμη, σε αντίθεση με τα παραδοσιακά νευρωνικά δίκτυα, χρησιμοποιούν την τελευταία κατάσταση ως είσοδο στην επόμενη κατάσταση για τον ίδιο ή άλλους νευρώνες σε επόμενα χρονικά βήματα.(Li et al., 2018). Επίσης, RNN που λαμβάνουν ως είσοδο μεγάλο αριθμό δεδομένων και δύνανται να αντιμετωπίσουν μεγαλύτερες ακολουθίες σε τιμές από τα παραδοσιακά νευρωνικά δίκτυα, είναι δυνατό να κατασκευαστούν (Goodfellow, 2016).

Η έννοια του RNN πρωτοεμφανίστηκε σε μια επιστολή που δημοσιεύτηκε από το Nature το 1986 από τους (Rumelhart et al, 1986) με σκοπό να περιγράψει μια νέα διαδικασία μάθησης .Ενώ ο 1997, προτάθηκε μια από τις πιο δημοφιλείς αρχιτεκτονικές RNN, το δίκτυο μακράς

βραχυπρόθεσμης μνήμης (LSTM) όπου αποτελεί μια από τις πιο γνωστές παραλλαγές του και μπορεί να επεξεργαστεί μεγάλες ακολουθίες.

Ωστόσο, η εκπαίδευσή του RNN μπορεί να γίνει αρκετά περίπλοκη λόγω του φαινομένου της "εξασθένησης" (vanishing gradient) ή "έκρηξης" του βαθμίδιου (exploding gradient)(Li et al., 2018).Το φαινόμενο αυτό παρατηρείται στις περιπτώσεις όπου τα δεδομένα στην διαδικασία της εκμάθησης παίρνουν ακραίες τιμές. Έτσι το μοντέλο δεν είναι εφικτό να αναγνωρίσει μακροχρόνιες εξαρτήσεις, με αποτέλεσμα να μην είναι δυνατή η εκπαίδευση του. Για την αντιμετώπιση του προβλήματος κλίσης έχουν προταθεί αρκετές παραλλαγές.

Υπάρχουν διάφοροι τύποι RNN, ένας από αυτούς είναι ο SimpleRNN που είναι μια από τις πρώτες και πιο απλές υλοποιήσεις RNNs και περιλαμβάνει ένα απλό νευρωνικό δίκτυο με συνδέσεις ανάδρασης. Χρησιμοποιείται για την επεξεργασία σειριακών δεδομένων αφού η προηγούμενη πληροφορία επηρεάζει τη μέθοδο με την οποία επεξεργάζεται η επόμενη. Ακόμα, το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM) αποτελεί μια από τις πιο γνωστές παραλλαγές αφού έχει την ικανότητα να αντιμετωπίζει το πρόβλημα των εξαφανιζόμενων και εκρηγνυόμενων κλίσεων. Τέλος, παρόμοια λογική με αυτή του LSTM αλλά με τροποποιημένη αρχιτεκτονική αποστέλλει μια τρίτη παραλλαγή ο Gated Recurrent Unit (GRU) RNN. Καταφέρνει να μειώσει τον χρόνο υπολογισμού όταν υπάρχουν αρκετοί παράμετροι αλλά και επιλύει το πρόβλημα της κλίσης καταγράφοντας τις μακροπρόθεσμες εξαρτήσεις.(Editor Wolfgang Walz, 2023.)



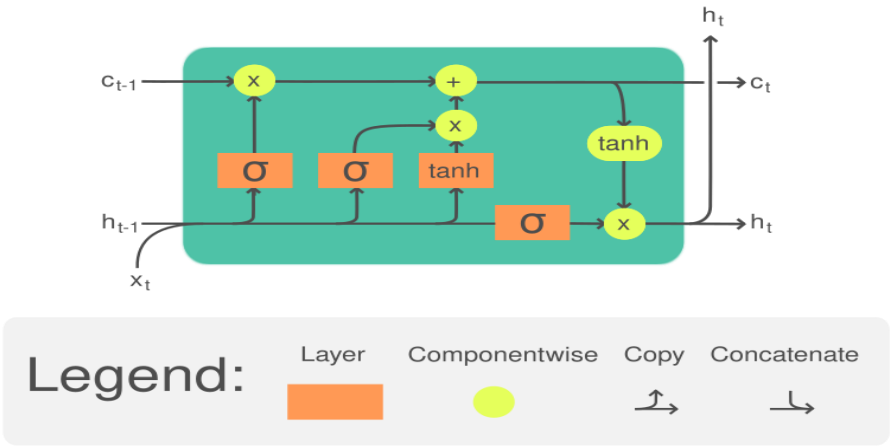
Εικόνα 5 Απεικόνιση της αρχιτεκτονικής ενός RNN (Kvit, J.,2016)

Long Short-Term Memory (LSTM)

Τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM) αποτελούν το επίκεντρο της βαθιάς μάθησης, αφού έχουν την δυνατότητα να αναπαράγουν όλα τα πολύπλοκα αποτελέσματα που βασίζονται στον RNN αλλά και να μάθουν τα δεδομένα εισόδου ακόμα και όταν αυτά είναι πολύ μεγάλα καταφέρνοντας έτσι να ξεπεράσουν τα RNNs (Sherstinsky, 2018).Η δομή τους είναι πιο σύνθετη από τα παραδοσιακά δίκτυα. Αποτελούνται από τρεις πύλες: την πύλη εισόδου (input gate), την πύλη διανύσματος απώλειας μνήμης (forget gates) και μία πύλη στο διάνυσμα της εξόδου. Οι πύλες αυτές έχουν την δυνατότητα να ελέγχουν τη ροή των πληροφοριών και με αυτό τον τρόπο το δίκτυο έχει την δυνατότητα να επιλέξει ποιες πληροφορίες θα διατηρήσει και ποιες όχι ανάλογα με την σημασία τους(output gate) (Zaki & Meira, 2020).

Η επιτυχία των LSTM οφείλεται στη πύλη απώλειας μνήμης και από τη συνάρτηση στην στοιβάδα εξόδου, ενώ έχει αναφερθεί σε έρευνά πως καμία άλλη παραλλαγή των LSTM δε δίνει σημαντική διαφορά στα αποτελέσματα (Greff et al ,2016).Σημαντική αναφορά αποτελεί, ότι τα δίκτυα LSTM έχουν μια κρυφή κατάσταση που μπορεί να αποθηκεύει πληροφορίες για μεγάλα χρονικά διαστήματα. Αυτό τους επιτρέπει να μαθαίνουν και να κάνουν εκτιμήσεις με βάση μακροπρόθεσμα πρότυπα στα δεδομένα (Shewalkar, Nyavanandi, & Ludwig, 2019).

Η πρώτη επίσημη εμφάνιση των δικτύων έγινε το 1997, ενώ από τότε έχουν δημοσιευθεί πολυάριθμες θεωρητικές και πειραματικές εργασίες. Τα δίκτυα LSTM, έχουν χρησιμοποιηθεί για μοντελοποίηση γλώσσας, ομιλία σε κείμενο, μεταγραφή, αυτόματη μετάφραση και άλλες πολλές εφαρμογές (Lin & Tegmark, 2017).



Εικόνα 6 Κύτταρο Μνήμης Μακροχρόνιας και Βραχυπρόθεσμης Διάρκειας(LSTM) (Guillaume Chevalie, Wikipedia)

Transformers

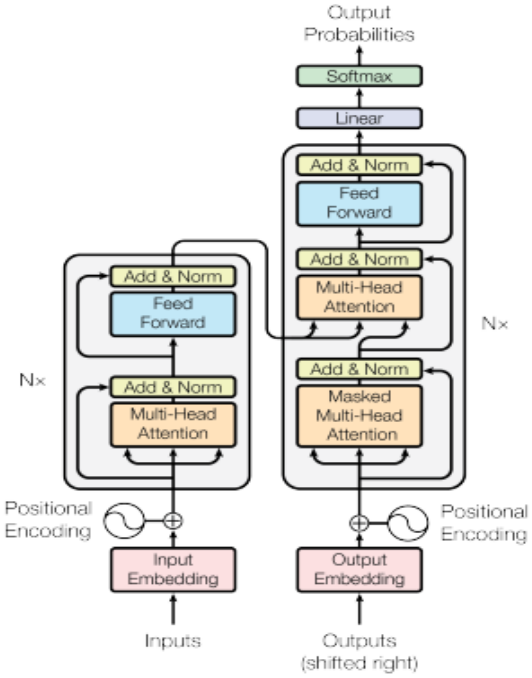
Οι μετασχηματιστές (Transformers) είναι ένας τύπος βαθύ νευρωνικού δικτύου όπου χάρη στα εκπληκτικά αποτελέσματα που έφεραν, ξεκίνησε το ενδιαφέρον για προβλήματα όρασης υπολογιστή. Πρωτοεμφανίστηκαν το 2017 από την εργασία "Attention is All You Need" και πλέον αποτελούν το βασικό παράγοντα για σύγχρονα μοντέλα γλωσσικής επεξεργασίας (Vaswani et al., 2017)

Ο σχεδιασμός τους είναι αρκετά απλοϊκός και τους επιτρέπει την επεξεργασία και τον συνδυασμό δεδομένων με διαφορετικούς τρόπους. Ένα βασικό προτέρημα είναι η δυνατότητα να υπάρχει η μοντελοποίηση μεγάλων εξαρτήσεων μεταξύ των στοιχείων της ακολουθίας εισόδου. Επίσης, υποστηρίζουν την παράλληλη επεξεργασία της ακολουθίας σε σύγκριση με επαναλαμβανόμενα δίκτυα, π.χ., τη μακροπρόθεσμη βραχυπρόθεσμη μνήμη (Khan et al., 2022)

Σε αντίθεση με τις παραδοσιακές μεθόδους επανάληψης, οι μετασχηματιστές χρησιμοποιούν την προσοχή για να μάθουν από ένα ολόκληρο τμήμα μιας ακολουθίας, με την βοήθεια του κωδικοποιητή (encoder) και του αποκωδικοποιητή (decoder).Ο πρώτος αφορά την είσοδο και ενσωματώνει την πληροφορία, ενώ ο δεύτερος χρησιμοποιεί την πληροφορία που έχει ενσωματωθεί από τον κωδικοποιητή για να παράγει την έξοδο (Vaswani et al., 2017)

Οι μετασχηματιστές έφεραν την επανάσταση στην ανάπτυξη προηγμένων μοντέλων φυσικής γλώσσας αφού χρησιμοποιήθηκαν σε έργα όπως το BERT (Bidirectional Encoder

Representations from Transformers) και το GPT (Generative Pre-trained Transformer).Τα μοντέλα αυτά έχουν βελτιώσει διάφορες εφαρμογές και έχουν καταφέρει να δημιουργήσουν νέες δυνατότητες στην επιστήμη αλλά και στον κόσμο της τεχνητής νοημοσύνης (Khan et al., 2022).



Εικόνα 7 Αρχιτεκτονική του Transformer (Vaswani et al., 2017)

4. Μεθοδολογία και Αποτελέσματα

4.1 Εισαγωγή

Στο πρώτο μέρος της συγκεκριμένης διπλωματικής εργασίας θα ασχοληθούμε με την ανάλυση δεδομένων από κριτικές σε διάφορα ξενοδοχεία. Σκοπός είναι να εξάγουμε συμπεράσματα σχετικά με τις υπηρεσίες και την ικανοποίηση των πελατών. Συγκεκριμένα, τα δεδομένα που θα χρησιμοποιηθούν, θα επεξεργαστούν ώστε να μπορεί να γίνει σωστά η ανάλυση και έπειτα θα προχωρήσουμε στην οπτικοποίηση και εξερεύνηση των δεδομένων για την διεξαγωγή συμπερασμάτων.

Στο δεύτερο μέρος, θα ακολουθήσει η εφαρμογή αλγορίθμων μηχανικής και βαθιάς μάθησης πάνω στα επεξεργασμένα δεδομένα, με στόχο τα μοντέλα αυτά να εκπαιδευτούν κατάλληλα ώστε να είναι ικανά να προβλέψουν τις πιθανές βαθμολογίες με βάση τις κριτικές των χρηστών. Τα αποτελέσματα των παραπάνω εφαρμογών θα ελεγχθούν και θα σχολιαστούν κατάλληλα έχοντας ως στόχο την διεξαγωγή συμπερασμάτων όχι μόνο για τα ίδια τα μοντέλα αλλά και το πόσο κατάλληλο είναι το κάθε μοντέλο όσον αφορά το συγκεκριμένο σύνολο δεδομένων. Τέλος τα μοντέλα θα συγκριθούν μεταξύ τους ώστε να διαπιστωθεί πιο μοντέλο παράγει καλύτερα αποτελέσματα.

4.2 Εργαλεία

Η επιλογή της γλώσσας python για τη συγκεκριμένη διπλωματική εργασία έγινε χάρη στις πολλές δυνατότητες που προσφέρει αλλά και στην ικανότητα ευελιξίας της, αφού χρησιμοποιεί εξειδικευμένες βιβλιοθήκες κατάλληλες για ανάλυση.

Οι πιο σημαντικές βιβλιοθήκες που χρησιμοποιήθηκαν στην συγκεκριμένη εργασία είναι οι Pandas , NumPy , Matplotlib.

Η Pandas, αποτελεί από τις πιο γνωστές βιβλιοθήκες δεδομένων της Python. Παρέχει τεράστιες δυνατότητες στη διαδικασία της ανάλυσης και επεξεργασίας των κειμένων, μέσω των DataFrames , τα οποία δίνουν τη δυνατότητα αποθήκευσης και επεξεργασίας μεγάλων συνόλων δεδομένων ώστε αργότερα να υποστούν την επεξεργασία που χρειάζεται.

Η NumPy χρησιμοποιείται για αριθμητικές πράξεις. Παρέχει εργαλεία τα οποία είναι απαραίτητα για την ερμηνεία των αποτελεσμάτων

Η Matplotlib αφορά την οπτικοποίηση των δεδομένων. Μέσω της δημιουργίας διαφόρων διαγραμμάτων (διαγράμματα γραμμών, ράβδων, διασποράς, ιστογραμμάτων κ.α) τα δεδομένα οπτικοποιήθηκαν και ερευνήθηκαν οι σχέσεις μεταξύ τους.

Οι βιβλιοθήκες αυτές είναι σχεδόν αλληλένδετες αφού η μία συμπληρώνει την άλλη δημιουργώντας την πλήρη επεξεργασία, ανάλυση και οπτικοποίηση δεδομένων φέρνοντας ένα επιθυμητό αποτέλεσμα.

4.3 Σύνολα δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε προέρχεται από την πλατφόρμα Kaggle και η συλλογή των δεδομένων έγινε από 3 διαφορετικά σύνολα τα οποία περιέχουν κριτικές για διάφορα ξενοδοχεία όπου βρίσκονται στη χώρα της Αμερικής. Οι εγγραφές αφορούν διάφορες χρονικές περιόδους, πιο συγκεκριμένα περιλαμβάνουν κριτικές μέχρι και τον Ιούνιο του 2019.

Το 7282_1.csv , το Datafiniti_Hotel_Reviews.csv και το Datafiniti_Hotel_Reviews_Jun19.csv αποτελούν σύνολα δεδομένων με 55,912 εγγραφές συνολικά και 28 στήλες. Το καθένα αναφέρεται σε διαφορετική ημερομηνία ενώ και τα τρία περιλαμβάνουν διάφορες πληροφορίες που αφορούν το ξενοδοχείο όπως το όνομα, την τοποθεσία (γεωγραφικό μήκος και πλάτος) τη διεύθυνση αλλά και στοιχεία που αφορούν τις κριτικές που έχουν κάνει οι πελάτες όπως το πλήρες κείμενο, την ημερομηνία που έγινε η κριτική και τη βαθμολογία της. Όλες αυτές οι πληροφορίες είναι αρκετά βοηθητικές ώστε να γίνει η συναισθηματική ανάλυση και να εξαχθούν διάφορα μοτίβα και συμπεράσματα για τις προτιμήσεις των χρηστών.

Για την καλύτερη κατανόηση του κάθε συνόλου δεδομένων πριν την επεξεργασία και την ανάλυσή τους ακολουθούν τρεις πίνακες οι οποίοι συνοψίζουν τις στήλες των συνόλων δεδομένων, αναλύοντας τις πληροφορίες που περιλαμβάνει κάθε μία..

Πιο συγκεκριμένα το πρώτο dataset με όνομα **7282_1.csv** αποτελείται από 35.313 γραμμές και 19 στήλες.

Χαρακτηριστικά	Περιγραφή
address	Διεύθυνση του ξενοδοχείου.
categories	Τύποι ξενοδοχείων
city	Πόλη όπου βρίσκεται το ξενοδοχείο
country	Χώρα στην οποία βρίσκεται το ξενοδοχείο
latitude	Γεωγραφικό πλάτος
longitude	Γεωγραφικό μήκος
name	Όνομα του ξενοδοχείου
postalCode	Ταχυδρομικός κώδικας του ξενοδοχείου
province	Πολιτεία όπου βρίσκεται το ξενοδοχείο
reviews.date	Ημερομηνία της κριτικής(διάστημα διαμονής)
reviews.dateAdded	Ημερομηνία που γράφτηκε η κριτική
reviews.doRecommend	Εάν ο πελάτης προτείνει το συγκεκριμένο ξενοδοχείο (Ναι/Όχι)
reviews.id	κωδικός για κάθε κριτική
reviews.rating	Βαθμολογία που δόθηκε από τον πελάτη
reviews.text	Το πλήρες κείμενο της κριτικής
reviews.title	Τίτλος της κριτικής
reviews.userCity	Πόλη από την οποία προέρχεται ο χρήστης
reviews.username	Όνομα χρήστη που έγραψε την κριτική.
reviews.userProvince	Πολιτεία που προέρχεται ο κάθε χρήστης που έγραψε την κριτική

Πίνακας 1 Χαρακτηριστικά του συνόλου δεδομένων 7282_1.csv

Το δεύτερο και το τρίτο σύνολο δεδομένων που ακολουθεί με όνομα **Datafiniti_Hotel_Reviews.csv** και **Datafiniti_Hotel_Reviews_Jun19** αποτελούνται από 10.000 γραμμές και 25 στήλες το κάθε ένα. Παρατηρήθηκε ότι περιλαμβάνουν κάποιες επιπλέον στήλες οι οποίες δεν υπάρχουν στο προηγούμενο σύνολο δεδομένων. Στον παρακάτω πίνακα παρουσιάζονται οι στήλες που υπάρχουν μόνο στο δεύτερο και στο τρίτο αρχείο και όχι στο πρώτο αλλά και η περιγραφή τους όπως προηγουμένως.

Χαρακτηριστικά	Περιγραφή
dateUpdated	Ημερομηνία της τελευταίας ενημέρωσης της κριτικής
keys	Κλειδιά που σχετίζονται με την εγγραφή
reviews.dateSeen	Ημερομηνίες που η κριτική έγινε αντιληπτή
reviews.sourceURLs	URL όπου προέρχονται οι κριτικές
sourceURLs	URL των πηγών του ξενοδοχείου
websites	Ιστότοποι που σχετίζονται με την εγγραφή ή το ξενοδοχείο

Πίνακας 2 Χαρακτηριστικά τών συνόλών δεδομένων Datafiniti_Hotel_Reviews.csv και Datafiniti_Hotel_Reviews_Jun19

4.4 Προ-επεξεργασία δεδομένων

Όπως έχουμε αναφέρει αρκετές φορές παραπάνω, στα σύνολα των δεδομένων μας χρειάζεται να γίνει επεξεργασία. Επομένως αφού φορτώσουμε τα σύνολα δεδομένων σε ένα Dataframe ξεκινάει η επεξεργασία. Παρατηρήσαμε ότι στο σύνολο δεδομένων «7282_1.csv» η βαθμολογία των ξενοδοχείων ξεκινάει από το 0 και καταλήγει στο 10.Σε αντίθεση τα άλλα δύο σύνολα έχουν βαθμολογίες από το 1 έως και το 5. Έτσι πριν ενώσουμε τα σύνολα δεδομένων, φέρνουμε όλες τις βαθμολογήσεις των κριτικών στην ίδια κλίμακα αλλά και τις ημερομηνίες στις οποίες γράφτηκαν οι κριτικές στο ίδιο format.

Έπειτα ενώνουμε τα σύνολα δεδομένων ώστε να δημιουργήσουμε ένα ενιαίο dataset με όλες τις πληροφορίες συγκεντρωμένες ώστε να προχωρήσουμε στην συνολική ανάλυση αυτών

Χαρακτηριστικά	count	unique
address	55912	3934
categories	55912	2038
city	55912	1818
country	55912	1
latitude	55826	NaN
longitude	55826	NaN
name	55912	3600
postalCode	55857	2826
province	55912	287
reviews.date	55653	NaN
reviews.dateAdded	35912	1029
reviews.doRecommend	0	NaN
reviews.id	0	NaN
reviews.rating	55050	NaN
reviews.text	55887	53955
reviews.title	54284	36814
reviews.userCity	30427	5485
reviews.username	55869	29232
reviews.userProvince	30221	835
Id	20000	2975
dateAdded	20000	2754
dateUpdated	20000	3029
primaryCategories	20000	8
Keys	20000	2976
Reviews.dateSeen	20000	2416
Reviews.sourceURLs	20000	14258
sourceURLs	20000	3271
websites	20000	2985

Πίνακας 3 Σύνολο εγγραφών ανά χαρακτηριστικό στο ενοποιημένο dataset

Όπως είναι εμφανές το καινούργιο σύνολο δεδομένων περιέχει αρκετές πληροφορίες οι οποίες δεν είναι απαραίτητες στην έρευνά μας. Επιπλέον στον παραπάνω πίνακα φαίνεται

πόσες διακριτές τιμές έχει η κάθε κατηγορία, για παράδειγμα στην κατηγορία με το όνομα country εμφανίζεται μόνο μία διακριτή τιμή και αυτή είναι η χώρα της Αμερικής, αφού εκεί βρίσκονται τα ξενοδοχεία μας. Τέλος παρατηρούμε ότι στον πίνακα εμφανίζονται κάποιες ελλιπείς τιμές NaN. Αυτό σημαίνει ότι το συγκεκριμένο πεδίο είναι κενό. Επομένως για να γίνει σωστά η επεξεργασία κρατάμε μόνο τις κοινές στήλες που είναι χρήσιμες στην ανάλυση μας αλλά και αντιμετωπίζουμε τις ελλιπείς τιμές.

Χαρακτηριστικά	
address	0
categories	0
city	0
latitude	86
longitude	86
name	0
province	0
reviews.date	259
reviews.rating	862
reviews.text	25
reviews.title	1628
reviews.userCity	25485
reviews.username	43
reviews.userProvince	25691

Πίνακας 4 Χαρακτηριστικά με ελλιπείς τιμές

Στις κατηγορίες οι οποίες είναι σημαντικές για την ανάλυσή μας (για παράδειγμα η κατηγορία rating και text) υπάρχουν κάποιες μηδενικές τιμές των οποίων σβήνουμε τις γραμμές. Το ποσοστό είναι αρκετά μικρό, ώστε να μην δημιουργηθούν ψευδή αποτελέσματα. Επιπλέον, στην κατηγορία της βαθμολογίας το σύνολο δεδομένων έχει τιμές σε μορφή πραγματικών αριθμών τους οποίους στρογγυλοποιούμε σε μορφή ακεραίων ώστε να υπάρχει αργότερα καλύτερη οπτικοποίηση των δεδομένων.

Μετά από αυτές τις αλλαγές οι τελικές διαστάσεις των δεδομένων είναι οι εξής:

Γραμμές	Στήλες
55025	14

Πίνακας 5 Τελικές διαστάσεις του συνόλου δεδομένων

Όπως καταλαβαίνουμε το συνολικό dataset αποτελείται από 55025 γραμμές και 14 στήλες.

Ενώ τα τελικά δεδομένα μας είναι τα εξής:

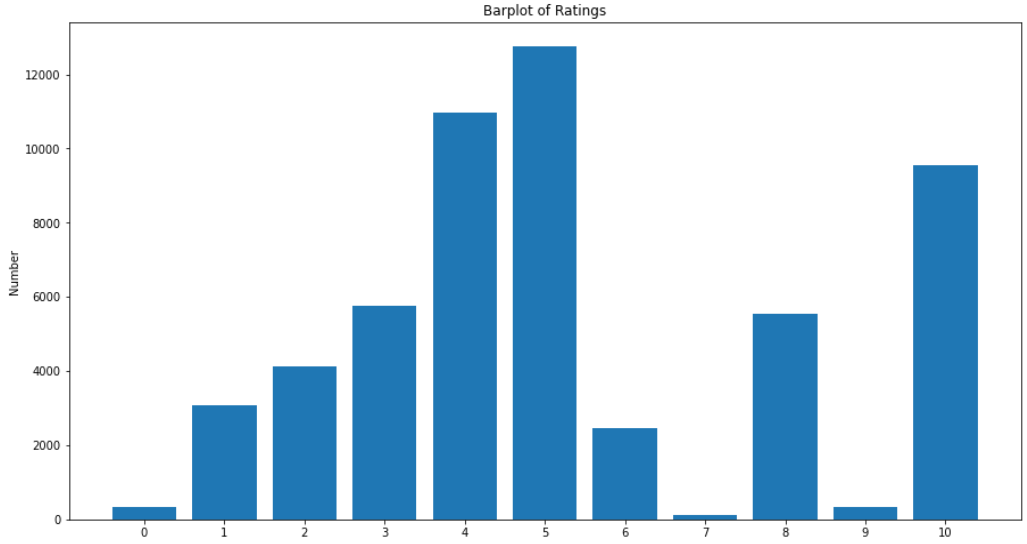
Χαρακτηριστικά	count	unique
address	55025	3816
categories	55025	1982
city	55025	1769
latitude	54949	NaN
longitude	54949	NaN
name	55025	3517
province	55025	256
reviews.date	54768	NaN
reviews.rating	55025	NaN
reviews.text	55025	53189
reviews.title	54146	36745
reviews.userCity	30298	5466
reviews.username	54982	28699
reviews.userProvince	30108	835

Πίνακας 6 Τελικό σύνολο δεδομένων

1.5 Εξερεύνηση των δεδομένων

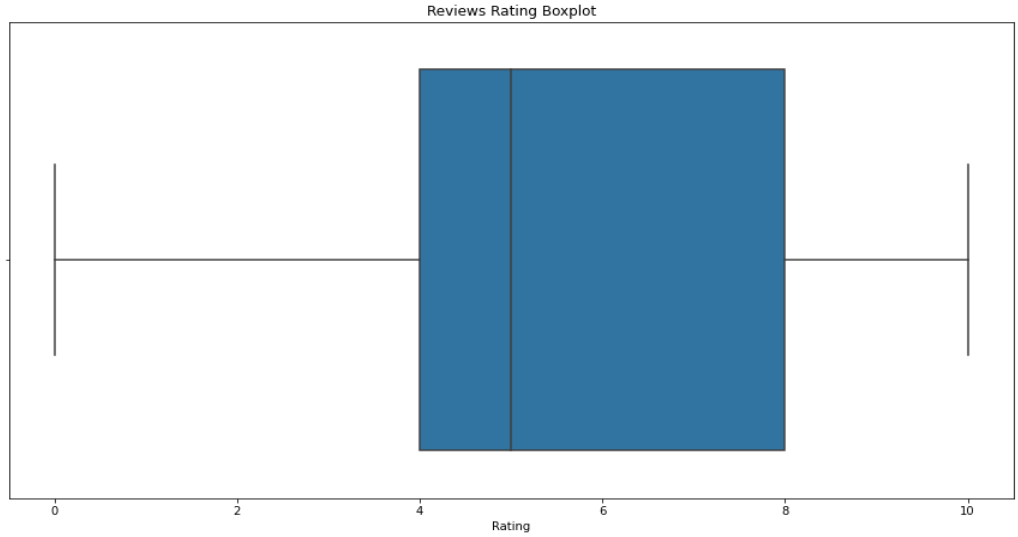
Για την καλύτερη κατανόηση των δεδομένων μας ώστε να έχουμε μία ολοκληρωμένη εικόνα, ακολουθούν διάφορα διαγράμματα και πίνακες που δίνουν μορφή σε αυτά τα δεδομένα. Ένας από τους πιο γνωστούς τρόπος αναπαράστασης, αποτελεί το διάγραμμα Barplot. Στη συγκεκριμένη περίπτωση (*Εικόνα 8*),στον κάθετο άξονα καταγράφεται η ποσότητα, ενώ στον οριζόντιο η εμβέλεια της βαθμολογίας για κάθε κριτική, καθώς είναι προφανές ότι διαβαθμίζεται από το 0 έως το 10 έχοντας μόνο ακέραιους αριθμούς. Βλέπουμε ότι η

βαθμολογία με τον αριθμό 5 αντιστοιχεί στον αριθμό 12.000. Αυτό δείχνει ότι οι περισσότερες από το σύνολο των κριτικών μας έχουν την βαθμολογία αυτή. Ταυτόχρονα, εξίσου υψηλά είναι και η βαθμολογία 4 και 10 με περίπου 10.000 η κάθε μία. Τέλος παρατηρούμε ότι οι υπόλοιπες βαθμολογίες είναι αρκετά χαμηλά ιδιαίτερα οι κριτικές με αριθμό 0,7, και 9.



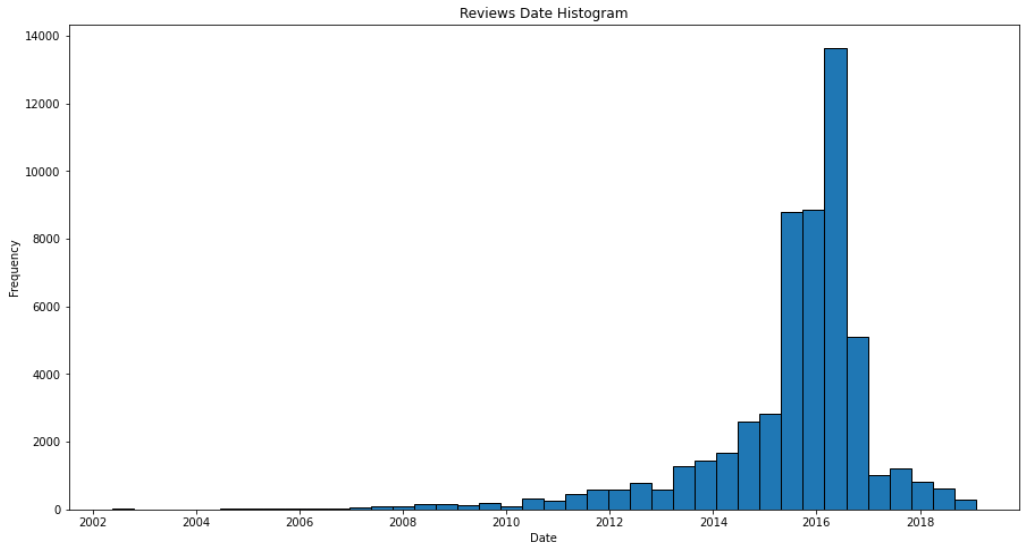
Εικόνα 8 Διάγραμμα Barplot από τις κριτικές

Ένας επιπλέον τρόπος οπτικοποίησης δεδομένων, επίσης γνωστός, είναι μέσω του διαγράμματος Boxplot. Παρακάτω(*Εικόνα 9*),έχοντας τις ίδιες συνιστώσες όπως προηγουμένως, επιβεβαιώνεται ότι το μέγιστο πλήθος των κριτικών έχουν βαθμολογία με αριθμό 5. Μπορούμε επίσης να εντοπίσουμε και να αναλύσουμε καλύτερα τη διασπορά των δεδομένων. Είναι φανερό ότι το 50% των κριτικών αναφέρεται σε βαθμολογίες από 4 έως 8.Επιπλέον μπορούμε να καταλάβουμε ότι οι ακραίες τιμές σε αυτήν την περίπτωση συμπεριλαμβάνονται στα δεδομένα, αλλά και ότι δεν αποτελούν τιμές outlier, τιμές δηλαδή που βρίσκονται μακριά από τα υπόλοιπα δεδομένα.



Εικόνα 9 Διάγραμμα Boxplot από τις κριτικές

Παρακάτω(*Εικόνα 10*) οπτικοποιούμε τα δεδομένα με διάγραμμα ιστογράμματος. Στον οριζόντιο άξονα απεικονίζονται οι κριτικές ανά έτος, ξεκινώντας από το 2002 φτάνοντας μέχρι το 2018, ενώ στον κάθετο άξονα φαίνεται το πλήθος των κριτικών που έγιναν την κάθε χρονολογία. Από αυτό το διάγραμμα καταλαβαίνουμε ότι η περίοδος κοντά στο 2016 αποτελεί την χρονιά με τις περισσότερες κριτικές, ενώ μέχρι το 2012 παρατηρούμε ότι ο αριθμός που γίνονταν κριτικές είναι αρκετά μικρός. Τέλος, εντοπίζεται ότι μετά τη μεγάλη αύξηση των κριτικών εν έτη 2016 ακολουθεί μια σταδιακή πτώση αυτών, φτάνοντας στο 2018 με αρκετά χαμηλή δραστηριότητα κριτικών.



Εικόνα 10 Ιστόγραμμα με τη χρονολογία των κριτικών

Η μέθοδος οπτικοποίησης με wold cloud περιλαμβάνει τις λέξεις που εμφανίζονται πιο συχνά είτε σε ένα κείμενο είτε σε ένα σύνολο δεδομένων, με σκοπό να έχουμε μια γρήγορη εικόνα των δεδομένων μας.

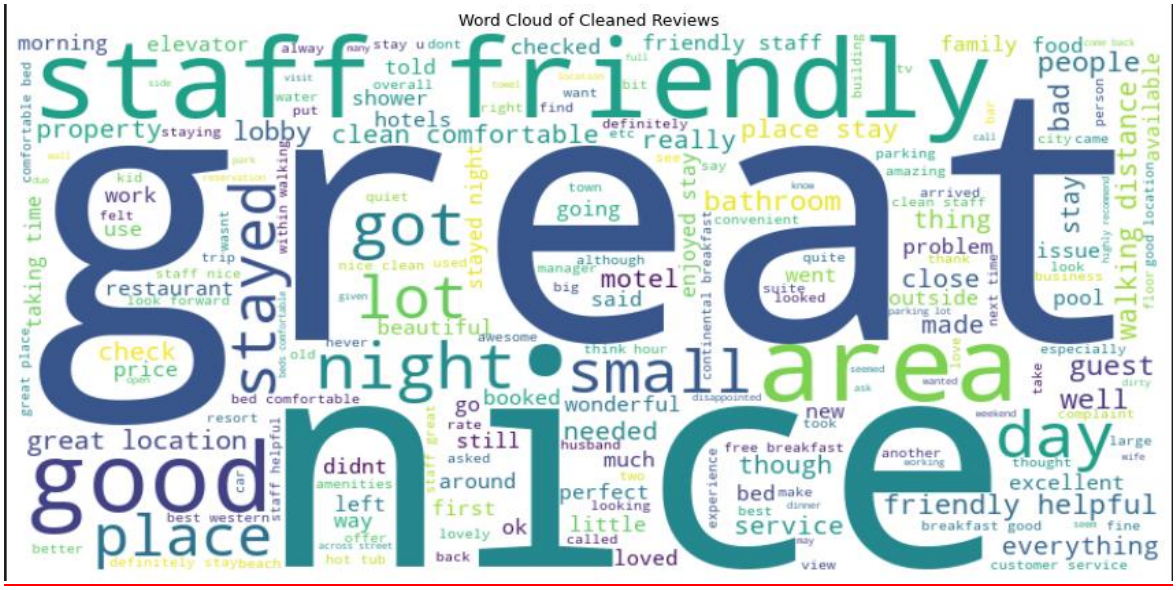
Για να μπορέσουμε να δημιουργήσουμε το word cloud από τα δεδομένα μας, είναι απαραίτητο να κάνουμε κάποιες τροποποιήσεις στο σύνολο δεδομένων που αναφέραμε και παραπάνω.

Αρχικά, κατεβάζουμε την βιβλιοθήκη με τα stopwords και επιπλέον φτιάχνουμε μια δική μας λίστα με κάποια “extra StropWords” τα οποία δε χρειαζόμαστε στην απεικόνιση. Έτσι σε μια συνάρτηση κάνουμε τις απαραίτητες αλλαγές, αφαιρώντας δηλαδή τα stopwords και τα extra stopwords αλλά και τα προθέματα της κάθε λέξης ενώ ταυτόχρονα μετατρέπουμε όλα τα κεφαλαία σε πεζά.

Μετά από τις παραπάνω τροποποιήσεις έχουμε σαν αποτέλεσμα το παρακάτω(*Εικόνα 11*) διάγραμμα. Το διάγραμμα αυτό δείχνει τις πιο συχνές λέξεις στις κριτικές των χρηστών. Όσο πιο μεγάλη είναι η απεικόνιση της λέξης, τόσο μεγαλύτερη είναι και η συχνότητα εμφάνισής της, ενώ όσο πιο μικρή εμφανίζεται μια λέξη τόσο μικρότερη είναι και η συχνότητα εμφάνισης της αντίστοιχα.

Στο διάγραμμά μας είναι εμφανές, ότι λέξεις όπως “great," "nice," και "friendly" έχουν μεγαλύτερη συχνότητα χρήσης κάτι που σημαίνει ότι το μεγαλύτερο ποσοστό των χρηστών έχει μείνει ικανοποιημένο αποκτώντας μια ευχάριστη εμπειρία.

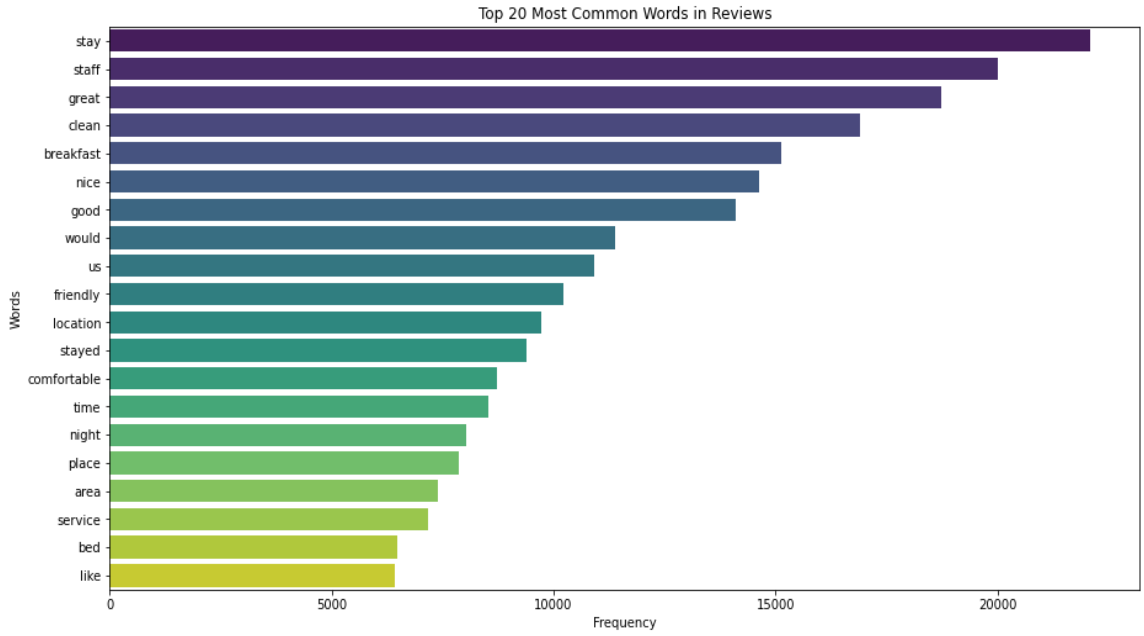
Τέλος, μεγάλη συχνότητα έχουν και οι λέξεις “staff” “area”. Μπορούμε να συμπεράνουμε λοιπόν ότι το μεγαλύτερο πλήθος των κριτικών μας προέρχονται από μία ευχάριστη εμπειρία, δίνοντας μεγάλη έμφαση στην περιοχή αλλά και στο προσωπικό του κάθε ξενοδοχείου.



Εικόνα 11 Worldcloud απο τις κριτικές

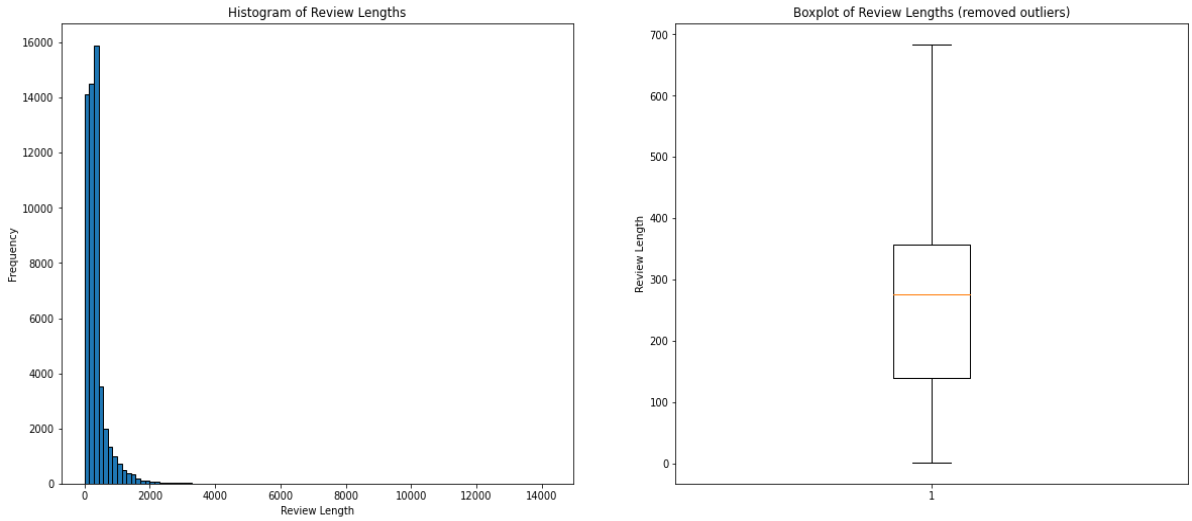
Ένας ακόμα τρόπος για να έχουμε πρόσβαση στα δεδομένα, αποκτώντας μια «γρήγορη» εικόνα για το τι προτιμούν περισσότερο οι πελάτες και ποια θέματα αφορούν συχνά οι κριτικές τους, αποτελούν τα διαγράμματα Bar Chart.

Στο παρακάτω διάγραμμα(*Εικόνα 12*), ο κάθετος άξονας αφορά τις λέξεις που έχουν χρησιμοποιηθεί στις περισσότερες κριτικές και ο οριζόντιος τη συχνότητά τους. Παρατηρώντας το επιβεβαιώνεται το διάγραμμα του word cloud καθώς βλέπουμε να εμφανίζονται αρκετά ψηλά οι λέξεις “great," "nice," και "friendly" αλλά και οι λέξεις “staff”, ”breakfast”, “clean” και “stay”, κατανοώντας με αυτό ότι οι παροχές του κάθε ξενοδοχείου αλλά και η τοποθεσία του παίζει καθοριστικό ρόλο για την εμπειρία των χρηστών.



Εικόνα 12 Bar Chart με τις 20 πιο κοινές λέξεις

Παρακάτω ακολουθούν δύο είδη διαγραμμάτων τα οποία έχουμε δει και πιο πάνω. Στο ιστόγραμμα και στο Boxplot που ακολουθούν(*Εικόνα 13*), φαίνεται το πλήθος των λέξεων ή χαρακτήρων που έχει η κάθε κριτική. Παρατηρούμε επομένως ότι η κάθε κριτική αποτελείται από 0 έως και περίπου 2.000 χαρακτήρες και από αυτό καταλαβαίνουμε ότι οι περισσότερες κριτικές είναι σύντομες. Από την άλλη το Boxplot δείχνει ότι οι περισσότερες κριτικές περιλαμβάνουν γύρω στους 300 χαρακτήρες, ενώ οι ακραίες τιμές έχουν αφαιρεθεί σε αυτή την περίπτωση για την καλύτερη κατανόηση.



Εικόνα 13 Ιστόγραμμα και boxplot του μήκους των κριτικών

4.6 Εξερεύνηση των δεδομένων σε σχέση με το συναίσθημα

Για αρχή επιλέχτηκαν, οι 10 κριτικές με το μεγαλύτερο polarity. Παρατηρούμε λοιπόν, βλέποντας τον πίνακα που ακολουθεί ότι το polarity όσων κριτικών εμφανίζονται αντιστοιχεί στο **1** που υποδηλώνει ότι η κριτική είναι θετική. Από την άλλη το subjectivity σε όλες τις κριτικές που παρουσιάζονται στον πίνακα είναι και αυτό ίσο με το 1, που υποδηλώνει το πόσο υποκειμενική είναι η κάθε κριτική. Σε αυτήν την περίπτωση το 1 υποδηλώνει ότι είναι πολύ υποκειμενική. Επομένως συμπεραίνουμε ότι ισχύει πως στον πίνακα εμφανίζονται θετικές κριτικές έχοντας polarity ίσο με 1. Βλέπουμε επίσης πως η βαθμολογία των κριτικών κυμαίνεται από το 4 έως το 10.

	reviews.text	polarity	subjectivity	reviews.rating
26865	Staff were awesome!	1.0	1.0	4
25900	Excellent htel au look sympa, surtout les cham...	1.0	1.0	5
6424	Excellent	1.0	1.0	5
33754	Excellent all the way around.	1.0	1.0	5
31973	Very nice hotel!!	1.0	1.0	5
5182	I will be returning the service was excellent	1.0	1.0	10
19076	Two night stay for a getaway with our service ...	1.0	1.0	5
28609	Awesome managers!!	1.0	1.0	4
5145	I could go on and on..... I won't. Totally awe...	1.0	1.0	5
753	Wonderful stay, Staff was so helpful	1.0	1.0	4

Πίνακας 7 Κριτικές με τα 10 καλύτερα polarity

Από την άλλη επιλέξαμε να εμφανίσουμε τις 10 κριτικές με τα μικρότερα polarity. Όπως βλέπουμε παρακάτω, σε αυτήν την περίπτωση το polarity αντιστοιχεί στο **-1** υποδηλώνοντας ότι οι κριτικές είναι αρνητικές. Το subjectivity εναλλάσσεται από το **0.6** στο **1.0** με αποτέλεσμα στις περιπτώσεις όπου η τιμή είναι **0.6** να μην είναι ξεκάθαρα υποκειμενική όπως στις περιπτώσεις όπου η τιμή ισούται με **1.** Επίσης βλέπουμε ότι οι κριτικές κυμαίνονται από το 1 έως το 3, δηλαδή είναι αρκετά χαμηλές.

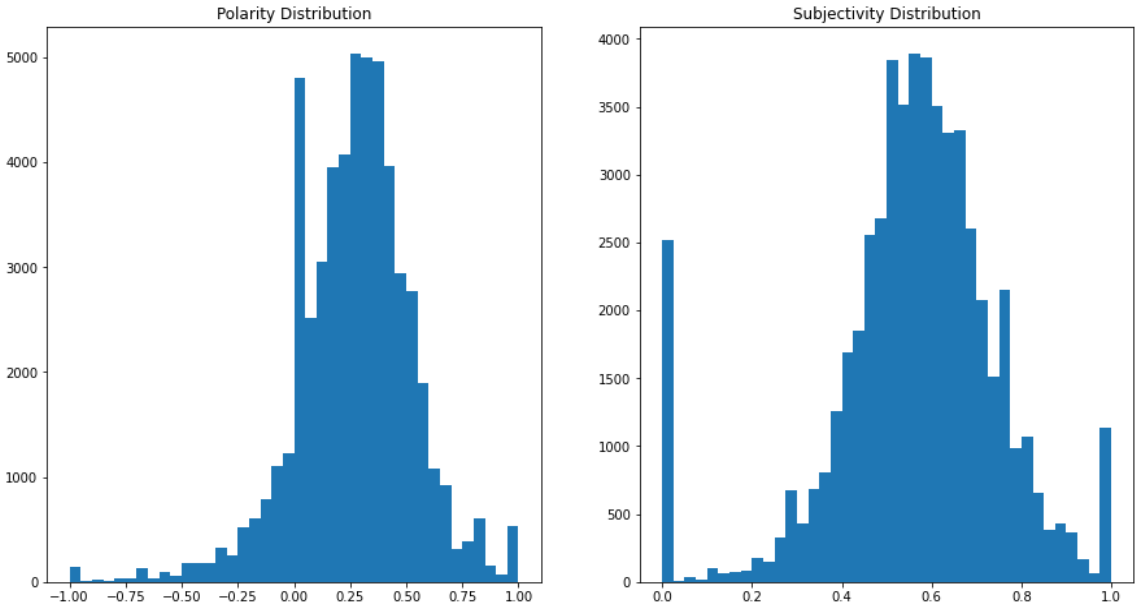
	reviews.text	polarity	subjectivity	reviews.rating
22276	It was horrible I got in room and the staff wa...	-1.0	1.000000	1
28114	Room was horrible	-1.0	1.000000	2
27318	Worst place I have ever been. That hotel shoul...	-1.0	1.000000	1
13532	place is overpriced and awful	-1.0	1.000000	1
14738	Horrible, some employee went into my room whil...	-1.0	1.000000	1
34215	la limpieza de la habitacion estuvo terrible, ...	-1.0	1.000000	3
3491	Found a sock and a glove in our room (NOT OUR...	-1.0	1.000000	1
7368	The condition of the carpet and tubs were in b...	-1.0	1.000000	1
27655	Absolutely Horrible!	-1.0	1.000000	1
14735	BED BUGS, ANTS. Bits all over my body! In a wo...	-1.0	1.000000	1

Πίνακας 8 Κριτικές με τα 10 μικρότερα polarity

Όσον αφορά το πρώτο διάγραμμα που ακολουθεί(*Εικόνα 14*) ,παρατηρούμε ότι το μεγαλύτερο πλήθος του polarity κυμαίνεται μεταξύ των τιμών 0.25 έως 0.75. Συμπεραίνουμε ότι η πλειοψηφία των κριτικών είναι θετικές, με ελάχιστες να είναι αρνητικές.

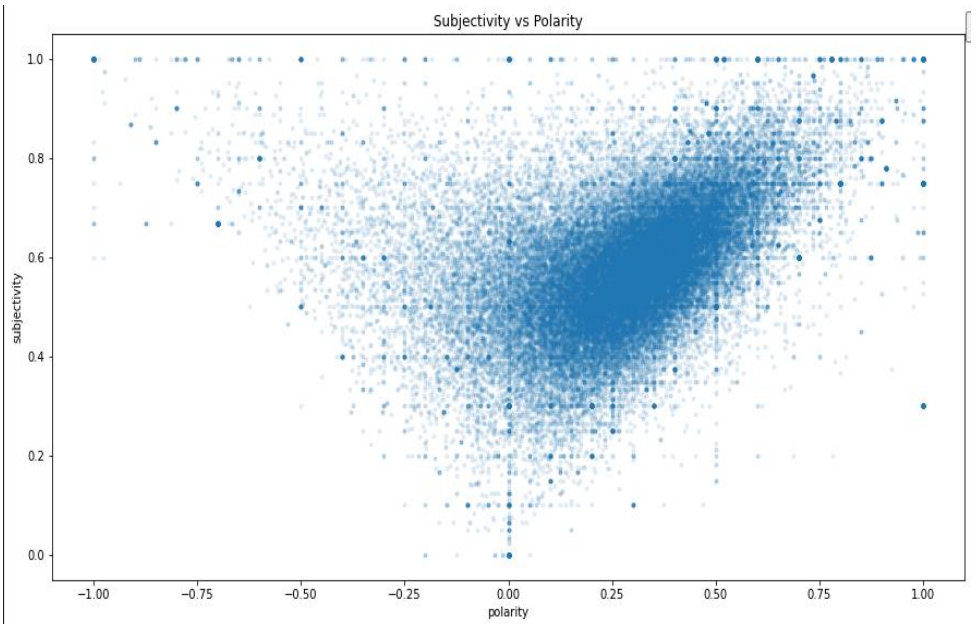
Από την άλλη, το δεύτερο διάγραμμα στα δεξιά(*Εικόνα 14*), δηλώνει την υποκειμενικότητα των κριτικών και κυμαίνεται από το 0 στο 1.Διαπιστώνουμε, ότι η πλειοψηφία των κριτικών αντιστοιχεί σε μέτρια επίπεδα υποκειμενικότητας αφού εντοπίζουμε μεγάλη αύξηση στο διάστημα 0.3 με 0.8 . Επομένως, συμπεραίνουμε ότι οι κριτικές αφορούν μέτρια έως υψηλή υποκειμενικότητα, κάτι το οποίο περιμέναμε αφού οι κριτικές συνήθως αφορούν προσωπικές απόψεις και εμπειρίες.

Παρακάτω έχουμε την κατανομή των Polarity και του subjectivity σε άξονες



Εικόνα 14 Κατανομή των Polarity και subjectivity

Ακολουθεί διάγραμμα scatter plot(*Εικόνα 15*)που απεικονίζει τη σχέση του Subjectivity με το Polarity. Στο συγκεκριμένο διάγραμμα, όταν το polarity είναι από 0 έως και 0.75 και το subjectivity από 0.3 έως 0.9 εντοπίζεται μια αρκετά μεγάλη συγκέντρωση σημείων σε εκείνη την περιοχή. Από αυτό συμπεραίνουμε, ότι όσο πιο θετική είναι η πόλωση, τόσο πιο υποκειμενική τείνει να είναι η κριτική.



Εικόνα 15 scatter plot μεταξύ Subjectivity και Polarity

Στον παρακάτω πίνακα(*Πίνακας 9*) αποτυπώνεται η συσχέτιση μεταξύ διαφόρων μεταβλητών. Όταν δύο μεταβλητές συσχετίζονται εννοείται ότι οι δύο μεταβλητές αυτές αυξάνονται και μειώνονται ανάλογα. Την μεγαλύτερη συσχέτιση που βλέπουμε με βάση το παραπάνω πίνακα την έχει το subjectivity και το polarity με 0.39 που δείχνει ότι όσο πιο θετική είναι η κριτική τόσο πιο υποκειμενική θα είναι και το αντίστροφο.

Η αμέσως επόμενη συσχέτιση που παρατηρείται είναι ανάμεσα στο reviews.rating με το polarity με 0.35 καθώς δηλώνει ότι όταν το polarity είναι αρνητικό θα είναι χαμηλή η βαθμολογία ενώ όταν είναι θετικό το polarity, η βαθμολογία θα είναι υψηλή.

Πίνακας 9 Πίνακας συσχετίσεων

4.7 Ερευνητικά ερωτήματα

Όπως έχουμε αναφέρει και παραπάνω με την ανάλυση των δεδομένων μας μπορούμε να μάθουμε χρήσιμες πληροφορίες για τα θέματα που αφορούν, εξάγοντας έτσι διάφορα συμπεράσματα, ώστε να προσαρμοστούν στρατηγικές βελτίωσης και προώθησης ανάλογα με τα ευρήματα. Καθώς παραπάνω έχουμε εφαρμόσει τις κατάλληλες μεθόδους ακολουθούν τα ερευνητικά ερωτήματα τα οποία θα εξετάσουμε.

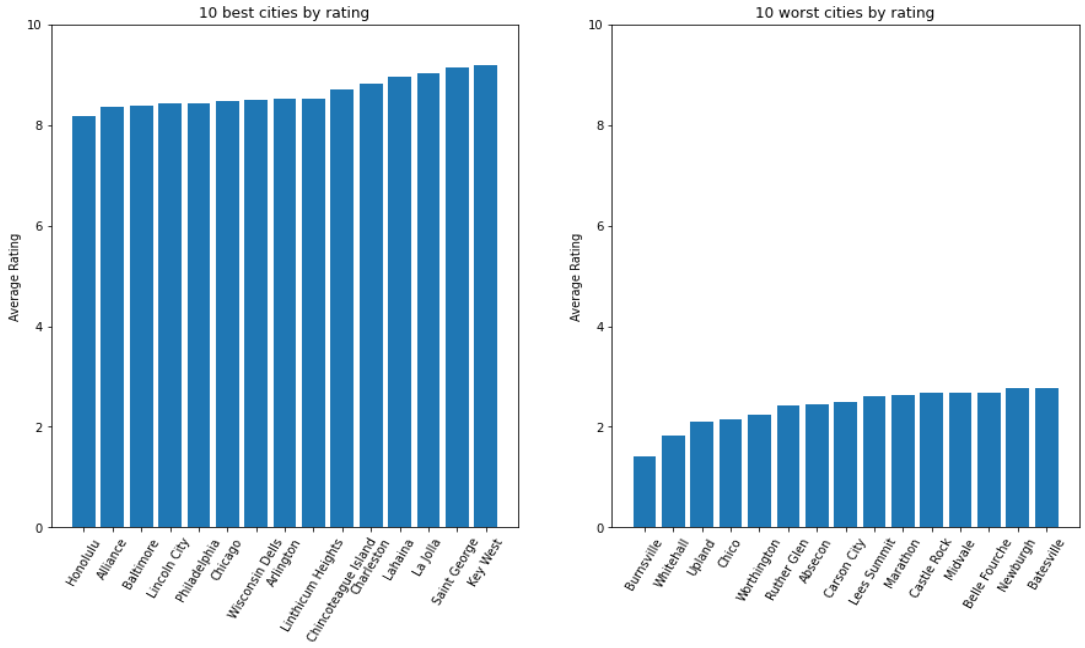
1. Ποιες πόλεις (city) έχουν τις υψηλότερες και τις χαμηλότερες μέσες βαθμολογίες;
2. Συχνότητα λέξεων σε καλές/κακές κριτικές
3. Πώς σχετίζεται η αναφορά σε ειδικά χαρακτηριστικά ή ανέσεις (όπως αναφέρονται στο κείμενο της κριτικής reviews.text) με τις βαθμολογίες;
4. Πώς σχετίζεται η αναφορά σε ειδικά χαρακτηριστικά ή ανέσεις (όπως αναφέρονται στο κείμενο της κριτικής reviews.text) με το polarity της κριτικής;

Παρακάτω ακολουθεί η περαιτέρω ανάλυση των ερωτημάτων αυτόν καθώς και η διεξαγωγή των συμπερασμάτων που προκύπτουν με βάση τα διαγράμματα που ακολουθούν.

- *Ποιες πόλεις (city) έχουν τις υψηλότερες και τις χαμηλότερες μέσες βαθμολογίες;*

Με την βοήθεια του διαγράμματος Barplot(*Εικόνα 16*), καταφέραμε να δείξουμε με βάση τις κριτικές ποιες είναι οι 10 καλύτερες και ποιες οι 10 χειρότερες περιοχές, ανάλογα με τις κριτικές των ξενοδοχείων που περιείχαν. Στον οριζόντιο άξονα και στους δύο πίνακες βρίσκονται τα ονόματα των περιοχών ενώ στον κάθετο ο βαθμός της κριτικής.

Βλέποντας αναλυτικά και τους δύο πίνακες, παρατηρούμε ότι οι πόλεις που είναι στο αριστερό διάγραμμα, αποτελούν μερικές από τις πιο δημοφιλείς πόλεις της Αμερικής, επομένως είναι λογικό να έχουν περισσότερα και καλύτερα ξενοδοχεία καθώς και περισσότερες κριτικές εφόσον ελκύουν περισσότερο τουρισμό

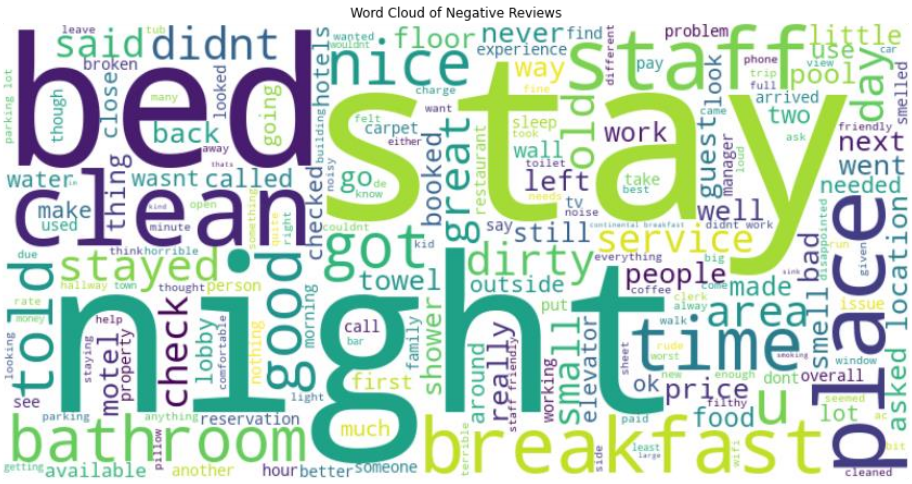


Εικόνα 16 Πόλεις με τις υψηλότερες και τις χαμηλότερες βαθμολογίες

.

- *Συχνότητα λέξεων σε καλές/κακές κριτικές*

Παρακάτω φαίνονται δύο διαγράμματα world cloud(*Εικόνα 17*)(*Εικόνα 18*). Το πρώτο αφορά τις λέξεις που εμφανίζονται πιο συχνά στις αρνητικές κριτικές ενώ το δεύτερο αυτές στις θετικές κριτικές. Πιο συγκεκριμένα, όταν η βαθμολογία των κριτικών είναι μικρότερη ή ίση του 3 αφορά αρνητική βαθμολογία και επομένως αρνητικές κριτικές ενώ τα υπόλοιπα θετικές . Λέξεις που εμφανίζονται στο πρώτο διάγραμμα όπως bed", "clean", "staff", "breakfast", "room", και "bathroom" αφορούν τα ζητήματα που αντιμετωπίζουν οι πελάτες όσον αφορά τη διαμονή τους στο ξενοδοχείο και πιθανότατα τα προβλήματα που τους δυσαρεστούν. Από την άλλη στο διάγραμμα με τις θετικές κριτικές εμφανίζονται κυρίως λέξεις όπως "great", "location", "nice", "good", "clean" δηλώνοντας ότι οι πελάτες έχουν μείνει αρκετά ικανοποιημένοι από την συνολική διαμονή τους. Επιπλέον, η λέξη clean είναι αρκετά εμφανής στο διάγραμμα των αρνητικών κριτικών, δείχνοντας έτσι ότι η καθαριότητα είναι συχνό πρόβλημα που αντιμετωπίζουν οι πελάτες και πως είναι κάτι το οποίο χρήζει βελτίωσης. Από την άλλη με βάση τις θετικές κριτικές, οι λέξεις location" και η "area", έχουν μεγάλη εμφάνιση στο διάγραμμα και φαίνεται ότι προσφέρουν θετικά στην αξιολόγηση

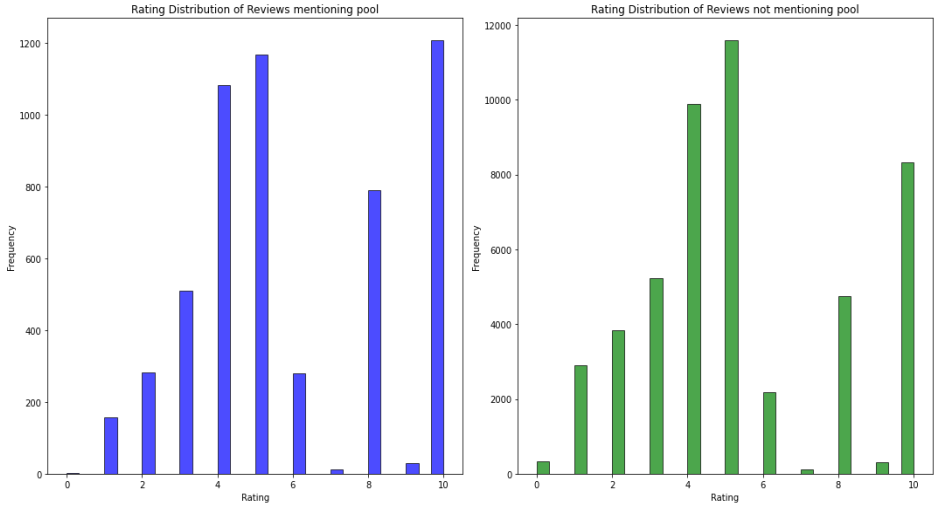


Εικόνα 17 Worldcloud με λέξεις στις αρνητικές κριτικές

Εικόνα 18 Worldcloud με λέξεις στις θετικές κριτικές

- *Πώς σχετίζεται η αναφορά σε ειδικά χαρακτηριστικά ή ανέσεις (όπως αναφέρονται στο κείμενο της κριτικής reviews.text) με τις βαθμολογίες;*

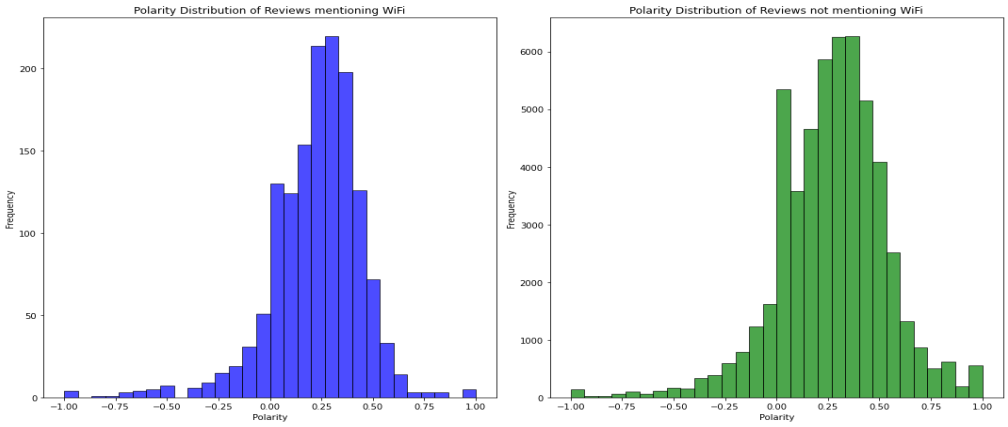
Παρακάτω απεικονίζονται τα διαγράμματα(*Εικόνα 19*) , για την κατανομή των βαθμολογιών στις κριτικές που αναφέρουν ή παραλείπουν τη λέξη “πισίνα” θέλοντας να δούμε πόσο επηρεάζει την βαθμολογία των χρηστών. Στον οριζόντιο άξονα φαίνεται η βαθμολογία των κριτικών ενώ στον κάθετο εμφανίζεται πόσες κριτικές έλαβε κάθε βαθμολογία για τα διαφορετικά σύνολα κριτικών (με ή χωρίς αναφορά στην πισίνα).Αυτό που παρατηρούμε είναι ότι υπάρχουν περισσότερες κριτικές που δεν αναφέρουν την συγκεκριμένη λέξη παρά από αυτές που την συμπεριλαμβάνουν.



Εικόνα 19 Συσχέτιση της λέξης “ pool” με τις βαθμολογίες

- *Πώς σχετίζεται η αναφορά σε ειδικά χαρακτηριστικά ή ανέσεις (όπως αναφέρονται στο κείμενο της κριτικής reviews.text) με το polarity της κριτικής;*

Παρατηρώντας τα δύο διαγράμματα παρακάτω (*Εικόνα 20*), παρατηρούμε ότι και σε αυτή την περίπτωση οι κριτικές που δεν αναφέρουν την λέξη “wifi” είναι πολύ περισσότερες από αυτές που την αναφέρουν. Παρόλα αυτά φαίνεται να ακολουθεί μια κανονική κατανομή με την πολικότητα να κυμαίνεται κυρίως από το 0.25 έως το 0.75 .Δείχνοντας με αυτό τον τρόπο ότι οι περισσότεροι χρήστες που αναφέρουν την λέξη wifi έχουν συνολικά μια θετική εμπειρία.



Εικόνα 20 Συσχέτιση της λέξης " Wi-Fi " με τις βαθμολογίες

5. Πειράματα και αποτελέσματα

Στην παρούσα εργασία, σκοπός μας είναι να εφαρμοστούν αλγόριθμοι βαθιάς μάθησης πάνω στα δεδομένα μας, να εκπαιδευτούν και να μπορέσουν να προβλέψουν τις βαθμολογίες των χρηστών. Στόχος μας είναι να αξιολογήσουμε τα μοντέλα για την απόδοσή τους αλλά και να τα συγκρίνουμε μεταξύ τους. Για την αξιολόγηση των μοντέλων θα χρησιμοποιηθούν μετρικές όπως η ακρίβεια (accuracy), η απώλεια (loss), και ο πίνακας σύγχυσης (confusion matrix), τα οποία μας βοηθάνε να εξάγουμε συμπεράσματα σχετικά με την αποτελεσματικότητα του κάθε μοντέλου. Για την εφαρμογή των αλγορίθμων, το σύνολο τον δεδομένων διαχωρίζεται σε σύνολο εκπαίδευσης (train set), το οποίο περιλαμβάνει ένα αρκετά μεγάλο ποσοστό από τα δεδομένα ώστε να εκπαιδεύσουν το μοντέλο αλλά και στο σύνολο δοκιμής (test set), όπου τα δεδομένα χρησιμοποιούνται με σκοπό την αξιολόγησή των μοντέλων.

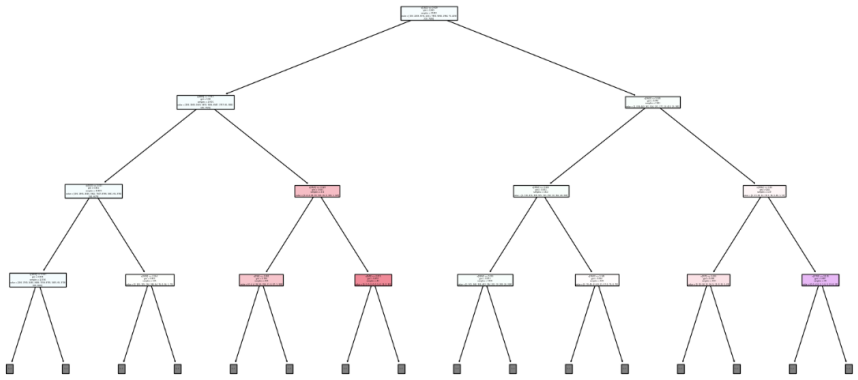
Χρήση του μοντέλου Random Forest

Ο Random Forest(RF) αποτελεί έναν αλγόριθμο επιβλεπόμενης μάθησης και σκοπός του είναι να συνδυάσει τις προβλέψεις των δεδομένων μειώνοντας τον κίνδυνο για υπερπροσαρμογή δημιουργώντας δέντρα σε πολλά και διάφορα σύνολα δεδομένων. Τα χαρακτηριστικά των δεδομένων αποτελούν και τους κόμβους απόφασης και με αυτό τον τρόπο χωρίζονται τα δεδομένα σε κλάδους. Το κάθε δέντρο αποφασίζει ανεξάρτητα και στο τέλος γίνεται ο συνδυασμός των αποτελεσμάτων.

Στην συγκεκριμένη εφαρμογή έχει επιλεγεί ο **RandomForestClassifier** ενώ χρησιμοποιείται ο αλγόριθμος **TF-IDF** ώστε να τροποποιηθούν κατάλληλα τα δεδομένα για είσοδο , καθώς οι αλγόριθμοι μηχανικής μάθησης μπορούν να επεξεργαστούν αριθμητικά διανύσματα και όχι απευθείας κείμενα.

Ο αλγόριθμος, δημιουργεί πολλά δέντρα απόφασης χρησιμοποιώντας τυχαία χαρακτηριστικά, το κάθε δέντρο παίρνει την δική του απόφαση και η τελική πρόβλεψη προκύπτει από την πλειοψηφία αυτών. Η τυχαία επιλογή δεδομένων έχει ως στόχο το μοντέλο να μην υπερπροσαρμοστεί στα δεδομένα, κάνοντας το έτσι πιο αποτελεσματικό.

Ακολουθεί το διάγραμμα (*Εικόνα 21*) ενός από τα πολλά δέντρα που χρησιμοποιούνται για την παραγωγή του Random forest.



Εικόνα 21 Αρχιτεκτονική ενός δέντρου που συμπεριλαμβάνεται στον Random Forest

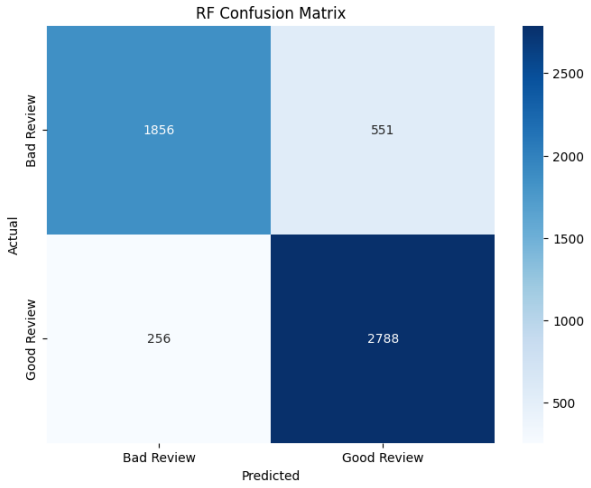
Ακολουθεί ο πίνακας δεδομένων (*Πίνακας 10*) από το σύνολο εκπαίδευσης(training set) του μοντέλου. Αξιολογώντας τα αποτελέσματα παρατηρούμε ότι η ακρίβεια είναι στο 1.00 όπως και οι τιμές των υπολοίπων recall, f1-score, support. Από τα παραπάνω συμπεραίνουμε ότι το μοντέλο κάνει πολύ καλές προβλέψεις τόσο στις θετικές όσο και στις αρνητικές κριτικές καθώς έχει εκπαιδευτεί αρκετά καλά πάνω στο σύνολο εκπαίδευσης. Αυτό βέβαια προφυλάσσει και από τον κίνδυνο **υπερπροσαρμογής** (overfitting), καθώς το ότι έχει πολύ καλά αποτελέσματα στο training set μπορεί να επηρεάσει την δυνατότητά του να γενικεύει σε νέα δεδομένα.

	precision	recall	f1-score	support
Bad	1.00	1.00	1.00	9600
Good	1.00	1.00	1.00	12203
Accuracy			1.00	21803
macro avg	1.00	1.00	1.00	21803
weighted avg	1.00	1.00	1.00	21803

Πίνακας 10 Σύνολο εκπέδευσης του μοντέλου RF

Ακολουθεί ο πίνακας σύγχυσης ο οποίος αφορά το σετ ελέγχου (*Πίνακας 11*). Στον οριζόντιο άξονα απεικονίζονται οι προβλεπόμενες κλάσεις , ενώ στον κάθετο οι πραγματικές και είναι χωρισμένες σε δύο κατηγορίες θετικές και αρνητικές. Ταυτόχρονα, κάθε κελί απεικονίζει πόσα δείγματα ταξινομήθηκαν στην κάθε κλάση. Πιο συγκεκριμένα, παρατηρούμε ότι από τις αρνητικές κριτικές έχουν γίνει 1.856 σωστές προβλέψεις και στις θετικές 2.788. Λάθος

ταξινομήθηκαν 551 κακές κριτικές ,οπού προβλέφθηκαν ως καλές, και 256 καλές κριτικές όπου προβλέφθηκαν ως κακές, αντίστοιχα. Συμπερασματικά, το μοντέλο επιτυγχάνει καλύτερη ταξινόμηση στις θετικές από ότι τις αρνητικές κριτικές, χωρίς όμως να παρουσιάζει μεγάλη απόκλιση.



Πίνακας 11 Πίνακας σύγχυσης του μοντέλου RF στο σετ αξιολόγησης

Ακολουθεί ο πίνακας δεδομένων(*Πίνακας 12*) από το σύνολο δοκιμών (test set) του μοντέλου. Το μοντέλο πέτυχε συνολική ακρίβεια 0.85 ενώ παρατηρούμε ότι οι μετρικές απόδοσης για τις κακές κριτικές είναι ελαφρώς μειωμένες σε σχέση με τις καλές με το F1 score να ισούται με 0.82 και 0.87 αντίστοιχα. Το μοντέλο επομένως δεν χειρίζεται τόσο καλά τις αρνητικές όσο τις θετικές κριτικές, παρόλα αυτά εμφανίζει αρκετά καλή ικανότητα στο να αναγνωρίζει καινούργια δεδομένα.

	precision	recall	f1-score	support
Bad	0.88	0.77	0.82	2407
Good	0.83	0.92	0.87	2044
Accuracy			0.85	5451
macro avg	0.86	0.84	0.85	5451
weighted avg	0.85	0.85	0.85	5451

Πίνακας 12 Σύνολο αξιολόγησης του μοντέλου RF

Χρήση του μοντέλου SVM

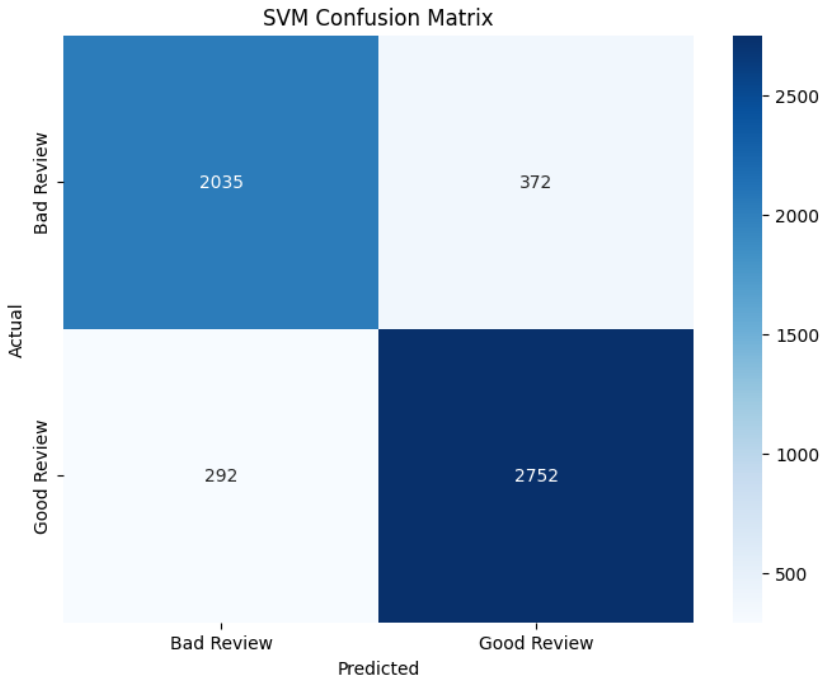
Το Support Vector Machine (SVM) μοντέλο αποτελεί αλγόριθμο μηχανικής μάθησης και πιο συγκεκριμένα επιβλεπόμενης μάθησης. Λειτουργεί διαχωρίζοντας τα δεδομένα σε διάφορες κατηγορίες ,στα υπερεπίπεδα , αναζητώντας αυτό με τη μεγαλύτερη απόσταση μεταξύ του υπερεπίπεδου και των πιο κοντινών σημείων. Προσπαθώντας με αυτό τον τρόπο να μειώσει την πιθανότητα λάθους.

Παρακάτω υπάρχει ο πίνακας δεδομένων*(Πίνακας 13)*από το σύνολο εκπαίδευσης του SVM μοντέλου. Παρατηρούμε ότι η συνολική ακρίβεια του μοντέλου βρίσκεται στο 92% , το οποίο σημαίνει ότι το μοντέλο έχει εκπαιδευτεί πολύ καλά και η ικανότητα του να προβλέπει έχει μεγάλη ακρίβεια. Παρατηρώντας τις μετρικές είναι εξίσου καλές και για τις θετικές αλλά και για τις αρνητικές κριτικές, με το precision να είναι 0.91 για τις κακές αξιολογήσεις και 0.92 για τις καλές, ενώ οι αντίστοιχοι δείκτες recall φτάνουν και στα δύο το 0.93.

	precision	recall	f1-score	support
Bad	0.91	0.89	0.90	9600
Good	0.92	0.93	0.93	12203
Accuracy			0.92	21803
macro avg	0.92	0.91	0.91	21803
weighted avg	0.92	0.92	0.92	21803

Πίνακας 13 Σύνολο εκπαίδευσης του μοντέλου SVM

Ακολουθεί ο πίνακας σύγχυσης *(Πίνακας 14*) μετά την εφαρμογή του παραπάνω αλγορίθμου καθώς μας βοηθάει να διακρίνουμε τις σωστές και λάθος προβλέψεις του μοντέλου. Πιο συγκεκριμένα, οι αξιολογήσεις έχουν ταξινομηθεί αρκετά καλά αφού το μοντέλο έκανε σωστά 2035 προβλέψεις για αρνητικές κριτικές και 2752 προβλέψεις για θετικές, ενώ 372 κακές κριτικές ταξινομήθηκαν λανθασμένα ως καλές και 292 καλές κριτικές ταξινομήθηκαν ως κακές, αντίστοιχα. Από αυτό συμπεραίνουμε ότι το μοντέλο μπορεί να διαχωρίσει αρκετά καλά τις κατηγορίες, αλλά συνεχίζει να υπάρχει και η πιθανότητα λάθους.



Πίνακας 14 Πίνακας σύγχυσης του μοντέλου SVM στο σετ αξιολόγησης

Τα αποτελέσματα του παρακάτω πίνακα δεδομένων(*Πίνακας 15*) από το σύνολο δοκιμών (test set) του μοντέλου SVM, με συνολική ακρίβεια 88% δείχνοντας ότι το μοντέλο είναι αρκετά αποδοτικό σε καινούργια δεδομένα. Οι μετρικές τόσο για τις καλές όσο και για τις κακές κριτικές είναι αρκετά υψηλές, ενώ πιο συγκεκριμένα το F1-score είναι ίσο με το 0.88 υποδηλώνοντας ισορροπία μεταξύ των μετρικών. Συμπερασματικά, το μοντέλο αποδίδει πολύ καλά και αξιόπιστα στον σύνολο δοκιμών.

	precision	recall	f1-score	support
Bad	0.87	0.85	0.86	2407
Good	0.88	0.90	0.89	3044
Accuracy			0.88	5451
macro avg	0.88	0.87	0.88	5451
weighted avg	0.88	0.88	0.88	5451

Πίνακας 15 Σύνολο αξιολόγησης του μοντέλου SVM

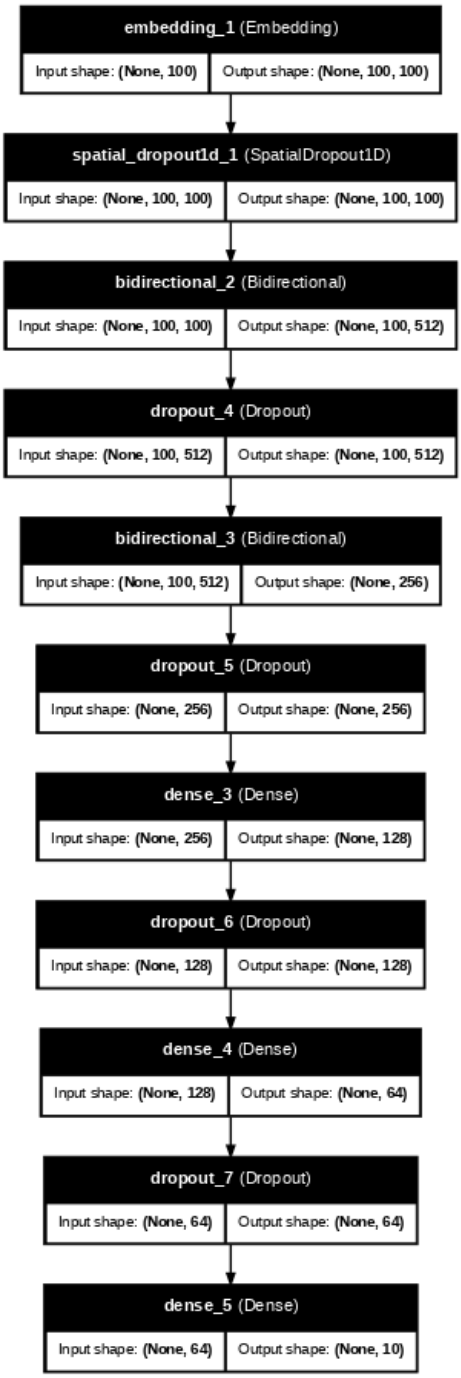
Χρήση του μοντέλου LSTM

Η εκπαίδευση του LSTM μοντέλου έγινε με την χρήση εξελιγμένων τεχνικών, με στόχο όσο την το δυνατόν καλύτερη απόδοση του αλγορίθμου και την αποφυγή της υπερπροσαρμογής, θέλοντας δηλαδή να αποφύγουμε το μοντέλο να μάθει καλά τα δεδομένα ώστε αργότερα να δυσκολεύεται να τα γενικεύσει σε καινούργια.

Τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης και επικύρωσης ενώ προσθέτοντας την τεχνική EarlyStopping, με την οποία η εκπαίδευση σταματάει όταν η απώλεια στα δεδομένα επικύρωσης δεν βελτιώνεται για 10 εποχές, επιτυγχάνουμε την αποθήκευση των καλύτερων βαρών. Για να μπορέσουμε να διακόψουμε την εκτέλεση χειροκίνητα όταν χρειάζεται προσθέσαμε να custom callback. Το μοντέλο εκπαιδεύτηκε για μέγιστο αριθμό 150 εποχών, με μέγεθος batch 128, και τα καλύτερα βάρη αποθηκεύονται αυτόματα.

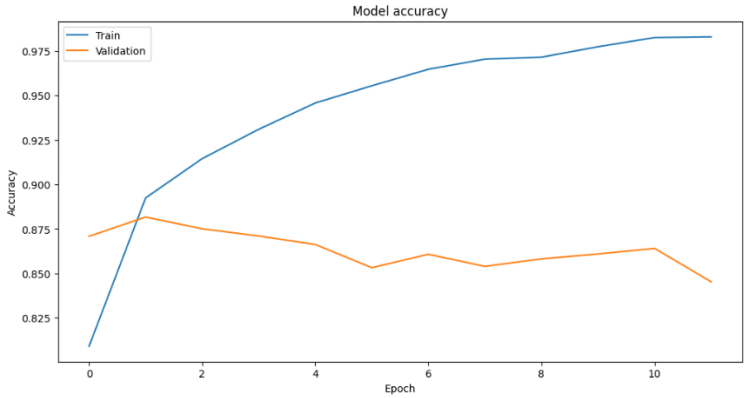
Το μοντέλο που αναπτύχθηκε βασίστηκε στην Bidirectional LSTM αρχιτεκτονική. Το πρώτο επίπεδο αφορά ένα επίπεδο ενσωμάτωσης , ώστε να μετατραπούν οι είσοδοι σε διανύσματα και ακολουθείται από ένα επίπεδο SpatialDropout με στόχο να αποφύγει το μοντέλο την υπερπροσαρμογή. Το μοντέλο αποτελείται από δύο LSTM επίπεδα με 256 και 128 μονάδες ενώ μετά από κάθε επίπεδο ακολουθεί Dropout με σκοπό την κανονικοποίηση. Τέλος χρησιμοποιείται το softmax για τη λειτουργία της ταξινόμησης, ενώ περιλαμβάνονται και τα επίπεδα dense τα οποία μειώνουν την διάσταση.

Ακολουθεί η αρχιτεκτονική του μοντέλου(*Εικόνα 22*) σε διάγραμμα για την καλύτερη κατανόηση.



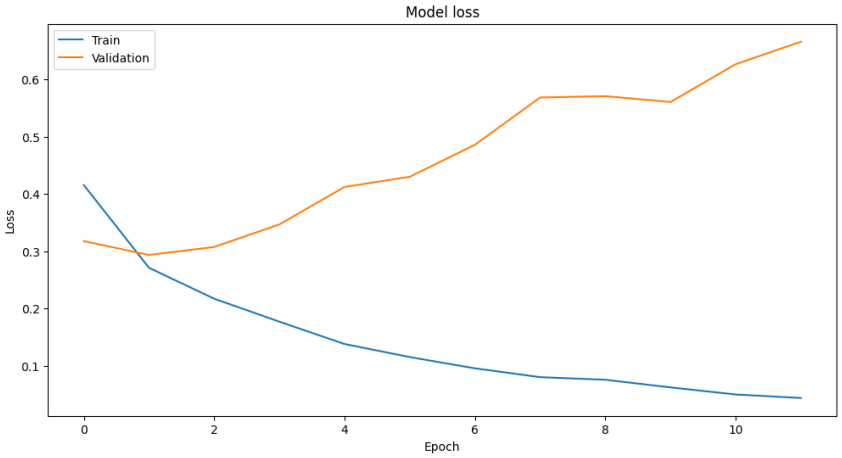
Εικόνα 22 Αρχιτεκτονική του μοντέλου LSTM

Στο παρακάτω διάγραμμα που ακολουθεί(*Εικόνα 23*), καταγράφεται η ακρίβεια (accuracy) του μοντέλου κατά τη εκπαίδευση του. Το μοντέλο εκτέλεσε 10 εποχές και στην 11η διακόπηκε. Η μπλε γραμμή που αφορά την εκπαίδευσή (train) του παρατηρούμε ότι αυξάνεται ραγδαία, υποδηλώνοντας ότι μαθαίνει πολύ καλά τα δεδομένα όσο αλλάζουν οι «εποχές» με την ακρίβεια του σετ εκπαίδευσης να φτάνει στο 97%. Παρόλα αυτά, η πορτοκαλί γραμμή, (Validation) που αφορά την αξιολόγηση της απόδοσης , παρατηρούμε ότι σιγά σιγά μειώνεται με την ακρίβεια να παραμένει γύρω στο 88%. Η διαφορά αυτή ανάμεσα στα σύνολα δηλώνει πιθανή υπερπροσαρμογή, και υποδεικνύει ότι το μοντέλο μπορεί να μην έχει την δυνατότητα να γενικεύει σε καινούργια δεδομένα.



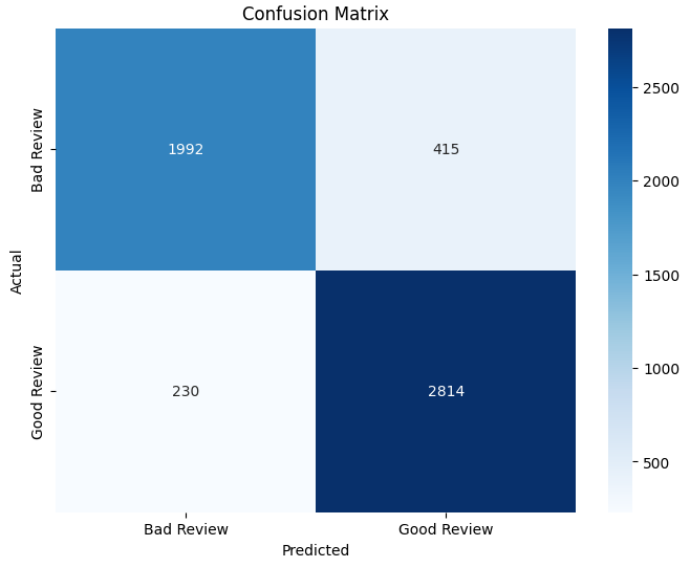
Εικόνα 23 Διάγραμμα ακρίβειας κατά την εκπαίδευση του μοντέλου LSTM

Στο επόμενο διάγραμμα(*Εικόνα 24*), παρουσιάζεται η απώλεια κατά την διάρκεια εκπαίδευσης. Αυτή τη φορά η μπλε γραμμή που αφορά την εκπαίδευση δείχνει την μείωση της απώλειας, δηλαδή το μοντέλο μαθαίνει καλά τα δεδομένα, από την άλλη, η πορτοκαλί γραμμή, αυξάνεται μετά από μερικές εποχές, επιβεβαιώνοντας το φαινόμενο υπερπροσαρμογής μιας και το μοντέλο βελτιώνει την απόδοσή του στα δεδομένα εκπαίδευσης αλλά χειροτερεύει στα δεδομένα επικύρωσης.



Εικόνα 24 Διάγραμμα απώλειας κατά την εκπαίδευση του μοντέλου LSTM

Παρακάτω ο πίνακας σύγχυσης (*Πίνακας 16*) μας δίνει την δυνατότητα να εντοπίσουμε για ποιες βαθμολογίες έγιναν σωστές και λάθος προβλέψεις. Έτσι παρατηρούμε ότι το μοντέλο έχει προβλέψει σωστά 1992 αρνητικές κριτικές και 230 λανθασμένα. Από την άλλη, από τις θετικές κριτικές έγινε σωστά πρόβλεψή για τις 2814 ενώ 415 ταξινομήθηκαν λανθασμένα στις αρνητικές.



Πίνακας 16 Πίνακας σύγχυσης του μοντέλου LSTM στο σετ αξιολόγησης

Ο πίνακας με τις μετρικές απόδοσης(*Πίνακας 17*), μας βοηθάει να αποκτήσουμε μια συνολική εικόνα για την επίδοση του μοντέλου. Η ικανότητα του μοντέλου να διακρίνει σωστά μεταξύ των κατηγοριών εξετάζεται με τις μετρικές απόδοσης. Παρακάτω βλέπουμε την ακρίβεια να είναι ίση με το 88% δείχνοντάς μας ότι το μοντέλο κάνει πολύ καλές προβλέψεις.

Συγκεκριμένα, η μετρική recall είναι ίση με 0.83, η οποία δείχνει ότι το μοντέλο κατάφερε να ταξινομήσει σωστά 83% από τις αρνητικές κριτικές. Το recall για τις θετικές είναι στο 0.92, όπου είναι υψηλότερο. Αντιλαμβανόμαστε ότι το μοντέλο διαχειρίζεται καλύτερα τις θετικές από τις αρνητικές κριτικές, προσφέρει όμως μια ισορροπημένη εικόνα απόδοσης.

	precision	recall	f1-score	support
Bad (0)	0.90	0.83	0.86	2407
Good(1)	0.87	0.92	0.90	3044
Accuracy			0.88	5451
macro avg	0.88	0.88	0.88	5451
weighted avg	0.88	0.88	0.88	5451

Πίνακας 17 Σύνολο αξιολόγησης του μοντέλου LSTM

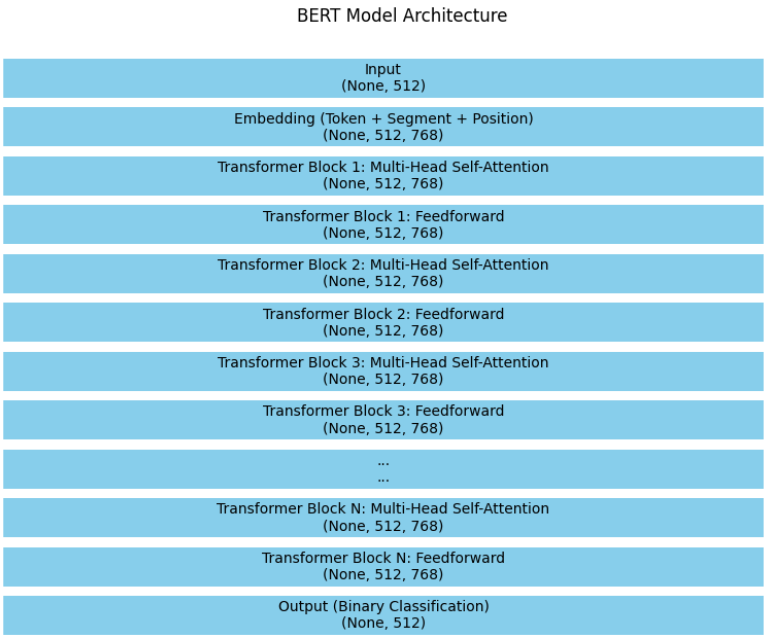
Χρήση του μοντέλου BERT

Το BERT (Bidirectional Encoder Representations from Transformers) είναι ένα μοντέλο όπου η αρχιτεκτονική του βασίζεται σε αυτή των transformers.

Όπως και προηγουμένως, η εκπαίδευση του αλγορίθμου, πραγματοποιήθηκε με τέτοιο τρόπο ώστε να αποφευχθεί η υπερπροσαρμογή του μοντέλου ενώ στοχεύει στην καλύτερη και βέλτιστη απόδοσή του.

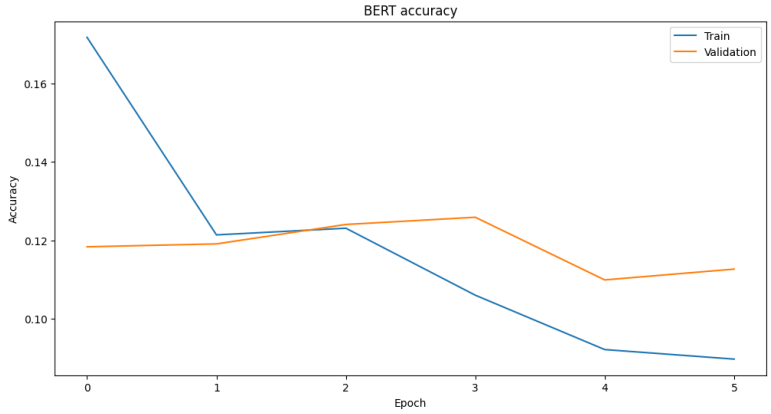
Το μοντέλο εκπαιδεύεται για μέγιστο αριθμό 100 εποχών, με batch size ίσο με 128 αλλά η εκπαίδευση τερματίζεται νωρίτερα λόγω της χειροκίνητης παρέμβασης. Σε αυτή την περίπτωση το μοντέλο μας εκπαιδεύτηκε για έξι εποχές μια και η διαδικασία εκπαίδευσης ήταν αρκετά χρονοβόρα.

Ακολουθεί η αρχιτεκτονική του μοντέλου(*Εικόνα 25*)



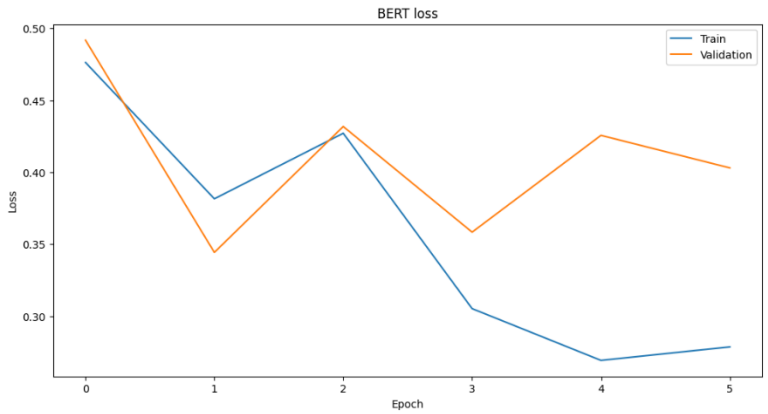
Εικόνα 25 Αρχιτεκτονική του μοντέλου BERT

Ακολουθεί το διάγραμμα(*Εικόνα 26*) στο οποίο καταγράφεται η ακρίβεια (accuracy) του μοντέλου BERT κατά τη εκπαίδευση του. Σε αυτή την περίπτωση παρατηρείται ότι και στο σετ εκπαίδευσης (μπλε γραμμή) όσο και στο σετ επικύρωσης (πορτοκαλί γραμμή) ακολουθεί μια σταδιακή πτώση όσο αυξάνονται οι εποχές. Το μοντέλο καθώς εκπαιδεύεται μειώνει την αποτελεσματικότητά του καταλήγοντας σε πιθανή υπερπροσαρμογή και αδυναμία γενίκευσης.



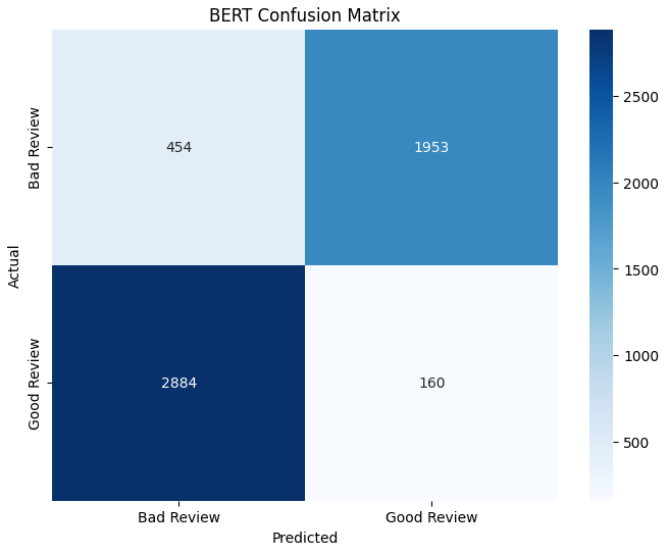
Εικόνα 26 Διάγραμμα ακρίβειας κατά την εκπαίδευση του μοντέλου BERT

Ακολουθεί το διάγραμμα απώλειας του μοντέλου(*Εικόνα 27*). Σε αυτή την περίπτωση αρχικά παρατηρούμε ότι στο σετ εκπαίδευσης (μπλε γραμμή) όπως και στο σετ επικύρωσης (πορτοκαλί γραμμή) η απώλεια μειώνεται . Από την Τρίτη εποχή και μετά παρατηρείται απόκλιση ανάμεσα στα δύο σετ υποδηλώνοντας την μη ικανότητα του μοντέλου να γενικεύει και καινούργια δεδομένα, ενώ ταυτόχρονα συνδέεται με υπερπροσαρμογή.



Εικόνα 27 Διάγραμμα απώλειας κατά την εκπαίδευση του μοντέλου BERT

Παρακάτω, παρατηρώντας τον πίνακα σύγχυσης του μοντέλου (*Πίνακας 18*) φανερώνονται αρκετά προβλήματα στην ταξινόμηση του. Παρουσιάζεται πρόβλημα στην προσπάθεια να διακρίνει σωστά τις κατηγορίες, με τα περισσότερα λάθη να γίνονται στην κατηγορία με τις θετικές κριτικές. Το μοντέλο είναι εμφανές ότι δεν μπορεί να ταξινομήσει σωστά τις θετικές και τις αρνητικές κριτικές ,αφού οι σωστές προβλέψεις για τα αρνητικά είναι 454 ενώ για τα θετικά σχόλια μόλις 160. Η ανισορροπία αυτή δηλώνει ότι πιθανά ο αλγόριθμος να μην προλαβαίνει να κάνει επαρκή εκπαίδευση μέχρι το πέρας των έξι εποχών και πιθανόν να χρειαζόταν περισσότερο χρόνο ώστε να εκπαιδευτεί πλήρως .



Πίνακας 18 Πίνακας σύγχυσης του μοντέλου BEPT στο σετ αξιολόγησης

Τα αποτελέσματα αξιολόγησης του μοντέλου BERT (*Πίνακας 19*) μέσω των μετρικών precision, recall και f1-score παρουσιάζουν αρκετά χαμηλές επιδόσεις, με την συνολική ακρίβεια να φτάνει το 0,11. Συγκεκριμένα, το precision είναι μόλις 0.08 για τα θετικά σχόλια ακόμα πιο χαμηλό από τα αρνητικά. Συμπερασματικά, το μοντέλο αδυνατεί να ταξινομήσει σωστά την πλειονότητα των δεδομένων και οι αποδόσεις του μοντέλου είναι πολύ κάτω από το αποδεκτό επίπεδο.

	precision	recall	f1-score	support
Bad (0)	0.14	0.19	0.16	2407
Good(1)	0.08	0.05	0.06	3044
Accuracy			0.11	5451
macro avg	0.11	0.12	0.11	5451
weighted avg	0.10	0.11	0.10	5451

Πίνακας 19 Σύνολο αξιολόγησης του μοντέλου BERT

6. Συμπεράσματα

Η ανάλυση διάφορων χαρακτηριστικών στα παραπάνω δεδομένα, όπως ημερομηνίες, προορισμοί κ.α περιλαμβάνεται στην διαδικασία ταξινόμησης. Στην συγκεκριμένη περίπτωση από τα αποτελέσματα εύκολα καταλαβαίνουμε ότι δεν παρουσιάζεται ιδανική κατανομή των δεδομένων, αφού ορισμένες κατηγορίες περιέχουν περισσότερα δείγματα από άλλες, επηρεάζοντας τα αποτελέσματα του μοντέλου. Η αξιολόγηση της ακρίβειας με τη χρήση διαφόρων αλγορίθμων κατέληξε σε μια σχετικά ικανοποιητική ακρίβεια. Πιο συγκεκριμένα η ακρίβεια για κάθε αλγόριθμο, Random forest (0,85), SVM(0,88) LSTM(0,88),BERT(0,11) με τον μέσο όρο ακρίβειας να υπολογίζεται:

$$\text{Μέσος όρος ακρίβειας} = \frac{0,85+0,88+0,88+0,11}{4} = 0,68(\text{ή } 68\%) \qquad (3)$$

Παρατηρώντας τα αποτελέσματα του αλγόριθμου **Random Forest(RF)** που παρουσιάζονται παραπάνω μέσω του συνόλου εκπαίδευσης, συνόλου δοκιμών και τον πίνακα σύγχυσης, παρατηρούμε ότι η συνολική ακρίβεια του μοντέλου είναι αρκετά ικανοποιητική αφού ισούται με το 85%. Ταυτόχρονα, αρκετά καλές είναι οι και υπόλοιπες μετρικές precision, recall και F1-score με αυτές από τις θετικές κριτικές να υπερτερούν από τις αρνητικές. Αυτό σημαίνει ότι το μοντέλο καταφέρνει να ταξινομήσει ελαφρός καλύτερα τις θετικές από τις αρνητικές κριτικές και αυτό επιβεβαιώνεται και από τον πίνακα σύγχυσης. Συμπερασματικά το μοντέλο έχει ικανοποιητική απόδοση στο σετ δοκιμών ενώ ταυτόχρονα ισορροπεί την απόδοσή του, αφήνοντας βέβαια μικρά περιθώρια βελτίωσης.

Μετά την εφαρμογή στα δεδομένα του αλγορίθμου **Support Vector Machine (SVM),** παρουσιάστηκε μια αρκετά ικανοποιητική απόδοση με βάση το σύνολο δοκιμών, με την ακρίβεια να είναι ίση με 88%. Παρατηρούμε, από τις υπόλοιπες μετρικές απόδοσης ότι το μοντέλο αποδίδει παρόμοια και στις θετικές αλλά και αρνητικές κριτικές χωρίς μεγάλες αποκλείσεις με το recall να είναι 0.90 και 0.85 αντίστοιχα. Το μοντέλο έχει την δυνατότητα να γενικεύει σε καινούργια δεδομένα διακρίνοντας σωστά τις κλάσεις χωρίς να εμφανίζει μεγάλες δυσκολίες στην ταξινόμηση των κριτικών. Παρόλο που οι μετρικές απόδοσης στο σύνολο δοκιμών είναι μειωμένες σε σχέση με αυτές στο σύνολο εκπαίδευσης το μοντέλο παρουσιάζει αρκετά ικανοποιητικά αποτελέσματα και θα μπορούσε να θεωρηθεί κατάλληλο για την ταξινόμηση των κριτικών.

Ακολουθεί η εφαρμογή του μοντέλου LSTM, όπου έχει καλή απόδοση με ακρίβεια ίση με 0,88 ωστόσο υπάρχει σημαντικό περιθώριο βελτίωσης. Το μοντέλο εμφάνισε μεγάλη διαφορά στα αποτελέσματα μεταξύ του σετ εκπαίδευσης και επικύρωσης αφήνοντας ενδείξεις υπερπροσαρμογής. Παρατηρώντας τις μετρικές απόδοσης, συμπεραίνουμε ότι ταξινομεί καλύτερα τις θετικές από τις αρνητικές κριτικές αφού η ανάκληση (recall) των θετικών είναι υψηλότερη από αυτή των αρνητικών. Συμπερασματικά, το μοντέλο εμφάνισε μια ικανοποιητική απόδοση, όμως παρουσίασε σημάδια υπερεκπαίδευσης ενώ είναι εμφανές πως δεν γενικεύει τόσο καλά σε νέα δεδομένα ,κυρίως όταν αφορούν τις αρνητικές κριτικές.

Συνοψίζοντας, τα αποτελέσματα του BERT μοντέλου, εμφανίζουν αρκετά χαμηλές προβλέψεις με τη συνολική ακρίβεια να βρίσκεται στην τελευταία θέση από τις προηγούμενες και να είναι ίση με 0,11. Οι περισσότερες μετρικές είναι πολύ χαμηλές ειδικά για τις θετικές κριτικές με το recall να είναι μόλις 0,05. Από το διάγραμμα απώλειας και πίνακα σύγχυσης είναι εμφανές ότι το μοντέλο έχει υπερεκπαιδευτεί στα δεδομένα και ότι το μοντέλο δυσκολεύεται στην σωστή ταξινόμηση ανάμεσα στις κλάσεις, αντίστοιχα. Τέλος, είναι πιθανό η χαμηλή απόδοση του μοντέλου να οφείλεται στο γεγονός ότι εκπαιδεύτηκε μόλις σε έξι εποχές .

Μετά την εφαρμογή των τεσσάρων αλγορίθμων (BERT, LSTM, Random Forest, SVM) συμπεραίνουμε ότι το μοντέλο SVM έχει την καλύτερη ακρίβεια (88%),επομένως διακρίνεται ως το καλύτερο μοντέλο από τα παραπάνω. Με βάση τα αποτελέσματά του αποτελεί το πιο ισορροπημένο μοντέλο έχοντας καλή ισορροπία μεταξύ του precision και recal ταξινομώντας καλά τα δεδομένα ανάμεσα στις δύο κλάσεις. Δεύτερο τοποθετείται το Random Forest με την ακρίβεια (85%) να βρίσκεται αρκετά κοντά σε αυτή του SVM αλλά με ελάχιστα πιο χαμηλή ισορροπία ανάμεσα στις μετρικές .Το μοντέλο LSTM παρόλο που η ακρίβειά του είναι ίση με του SVM (88%) παρουσιάζει προβλήματα υπερπροσαρμογής κάνοντάς το μη αποδοτικό στην ταξινόμηση των δύο κλάσεων. Τέλος, τελευταίο στην κατάταξη βρίσκεται το μοντέλο BERT με μόλις 11% ακρίβεια ενώ όλες οι υπόλοιπες μετρικές είναι και αυτές αρκετά χαμηλές. Αξίζει βέβαια να σημειωθεί ότι ο BERT εκτελέστηκε μόνο για τέσσερις εποχές, με αποτέλεσμα να μην έχει αρκετό χρόνο ώστε να εκπαιδευτεί πλήρως.

Η συναισθηματική ανάλυση αποτελεί μια πολύπλοκη διαδικασία και οι παράμετροι από τους οποίους εξαρτάται ποικίλουν. Η χαμηλή απόδοση του του BERT και του LSTM μπορεί να οφείλεται σε διάφορες παραμέτρους. Η ποιότητα και η ποσότητα των δεδομένων αποτελούν δύο από τις πιο βασικές, αφού όταν τα χαρακτηριστικά είναι παρόμοια τα μοντέλα

δυσκολεύονται να τα διαχωρίσουν στις σωστές κλάσεις με αποτέλεσμα να υπάρχει σύγχυση μεταξύ τους. Επιπλέον τα δεδομένα μπορεί να μην είναι ιδανικά, οδηγώντας σε κακή απόδοση. Για αυτό το λόγο έχει τεράστια σημασία η επιλογή των δεδομένων αλλά και η επεξεργασία τους, πριν τα τροφοδοτήσουμε σε κάποιο μοντέλο. Τέλος, είναι πιθανή η ύπαρξη υπερεκπαίδευσης των μοντέλων, με αποτέλεσμα να δυσκολεύεται να γενικεύσει σε καινούργια δεδομένα. Τα μοντέλα SVM και Random Forest παρουσίασαν καλύτερη ακρίβεια, επηρεάστηκαν όμως από την κατανομή των δεδομένων αφού στην περίπτωση που τα δεδομένα είναι ανισόρροπα παραπλανούν την εκπαίδευση του μοντέλου. Συνολικά, πρωταρχικό ρόλο για την καλή απόδοση του κάθε μοντέλου έχει η σωστή επεξεργασία των δεδομένων και η προσαρμογή των μοντέλων στις ιδιαιτερότητες των δεδομένων ώστε να παράγονται άρτια αποτελέσματα.

7. Βιβλιογραφία

1. Akehurst, G. (2009). User generated content: The use of blogs for tourism organisations and tourism consumers. *Service Business*, *3*(1), 51–61. https://doi.org/10.1007/S11628-008-0054-2/METRICS
2. Akhtar, M. S., Ekbal, A., & Cambria, E. (2020). How Intense Are You? Predicting Intensities of Emotions and Sentiments using Stacked Ensemble [Application Notes]. *IEEE Computational Intelligence Magazine*, *15*(1), 64–75. https://doi.org/10.1109/MCI.2019.2954667
3. Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, *58*(2), 175–191. https://doi.org/10.1177/0047287517747753
4. Archana, R., & Jeevaraj, P. S. E. (2024). Deep learning models for digital image processing: a review. *Artificial Intelligence Review*, *57*(1). https://doi.org/10.1007/s10462-023-10631-z
5. Banerjee, I., Ling, Y., Chen, M. C., Hasan, S. A., Langlotz, C. P., Moradzadeh, N., Chapman, B., Amrhein, T., Mong, D., Rubin, D. L., Farri, O., & Lungren, M. P. (2019). Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial Intelligence in Medicine*, *97*, 79–88. https://doi.org/10.1016/j.artmed.2018.11.004
6. Brockopp, D. Y. (1983). What Is NLP? *The American Journal of Nursing*, *83*(7), 1012. https://doi.org/10.2307/3463336
7. Buhalis, D., & Law, R. (2008). Progress in information technology and tourism management: 20 years on and 10 years after the Internet—The state of eTourism research. *Tourism Management*, *29*(4), 609–623. https://doi.org/10.1016/J.TOURMAN.2008.01.005
8. Chikersal, P., Poria, S., & Cambria, E. (2015). *SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning*. http://URL.com.
9. Daniel Johnson. (2024). *Τι είναι η Ανάλυση Δεδομένων; Έρευνα, Τύποι & Παράδειγμα*.
10. Editor Wolfgang Walz, S. (n.d.). *Machine Learning for Brain Disorders*. http://www.springer.com/series/7657
11. Fallon, B., Ma, J., Allan, K., Pillhofer, M., Trocmé, N., & Jud, A. (2013). Opportunities for prevention and intervention with young children: Lessons from the Canadian incidence study of reported child abuse and neglect. *Child and Adolescent Psychiatry and Mental Health*, *7*(1). https://doi.org/10.1186/1753-2000-7-4

12. Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48. https://doi.org/10.1016/j.neucom.2015.09.116
13. Hippner, H., & Rentzmann, R. (2006). Text mining. *Informatik-Spektrum*, *29*(4), 287–290. https://doi.org/10.1007/S00287-006-0091-Y/METRICS
14. Li, S., Li, W., Cook, C., Zhu, C., & Gao, Y. (n.d.). *Independently Recurrent Neural Network (IndRNN): Building A Longer and Deeper RNN*.
15. Misopoulos, F., Mitic, M., Kapoulas, A., & Karapiperis, C. (2014). Uncovering customer service experiences with Twitter: The case of airline industry. *Management Decision*, *52*(4), 705–723. https://doi.org/10.1108/MD-03-2012-0235/FULL/XML
16. Mohammad, S. M. (n.d.). *#Emotional Tweets*. http://www.ark.cs.cmu.edu/GeoText
17. Parrinello, G. L. (1993). Motivation and anticipation in post-industrial tourism. *Annals of Tourism Research*, *20*(2), 233–249. https://doi.org/10.1016/0160-7383(93)90052-5
18. Roh, Y., Heo, G., & Whang, S. E. (2018). A Survey on Data Collection for Machine Learning: a Big Data -- AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162
19. Sherstinsky, A. (2018). *Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network*. https://doi.org/10.1016/j.physd.2019.132306
20. Shimada, K., Inoue, S., Maeda, H., & Endo, T. (2011). Analyzing tourism information on twitter for a local city. *Proceedings - 1st ACIS International Symposium on Software and Network Engineering, SSNE 2011*, 61–66. https://doi.org/10.1109/SSNE.2011.27
21. Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry*, *27*(2), 130–135. https://doi.org/10.11919/j.issn.1002-0829.215044
22. Spatiotis, N., Paraskevas, M., Perikos, I., & Mporas, I. (2017). Examining the impact of feature selection on sentiment analysis for the greek language. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10458 LNAI*, 353–361. https://doi.org/10.1007/978-3-319-66429-3_34/COVER
23. Talaei Khoei, T., Ould Slimane, H., & Kaabouch, N. (2023). Deep learning: systematic review, models, challenges, and research directions. In *Neural Computing and Applications* (Vol. 35, Issue 31, pp. 23103–23124). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00521-023-08957-4

24. Thelwall, M. (2019). Sentiment analysis for tourism. *Big Data and Innovation in Tourism, Travel, and Hospitality: Managerial Approaches, Techniques, and Applications*, 87–104. https://doi.org/10.1007/978-981-13-6339-9_6/COVER
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762
26. *What is Interaction Design (IxD)? — updated 2024 | IxDF*. (n.d.). Retrieved April 22, 2024, from https://www.interaction-design.org/literature/topics/interaction-design
27. *What is Sentiment Analysis? - Sentiment Analysis Explained - AWS*. (n.d.). Retrieved March 5, 2024, from https://aws.amazon.com/what-is/sentiment-analysis/
28. Acito, F. (2023). k Nearest Neighbors. In: Predictive Analytics with KNIME. Springer, Cham.
29. Evelyn F, Hodges JL Jr (1951) Discriminatory analysis-nonparametric discrimination: consistency properties. Technical report, California University, Berkeley
30. Cover TM, Hart P (1967) Nearest neighbor pattern classifcation. IEEE Trans Inf Theory 13(1):21–27
31. WICKRAMASINGHE, I. and KALUTARAGE, H. 2021. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft computing [online], 25(3), pages 2277-2293
32. Price, R. (2007). Bayes' Theorem: A Historical Perspective and Modern Application. *Philosophical Transactions of the Royal Society*, 53(2), 123-145.
33. Loh, W.-Y. (2015). Regression trees with unbiased variable selection and interaction detection. *Statistical Science, 30*(1), 89–106. https://doi.org/10.1214/14-STS511
34. Su, J., & Zhang, H. (2006). A fast decision tree learning algorithm. Faculty of Computer Science, University of New Brunswic
35. De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics, 5*(6), 448-455.
36. Rumelhart DE, Hinton GE, Williams RJ (1986) Learning representations by back-propagating errors. Nature 323(6088):533–536
37. Henry W. Lin and Max Tegmark. Criticality in formal languages and statistical physics. Entropy, 19(7):299, Aug 2017
38. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR), 54*(10s), 1-41

39. V. Mohan, “Preprocessing Techniques for Text Mining-An Overview Privacy Preserving 68 Data Mining View project,” 2015. Accessed: May 11, 2021. [Online]. Available: https://www.researchgate.net/publication/339529230