

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΨΗΦΙΑΚΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΣΤΗ ΔΙΟΙΚΗΣΗ ΤΗΣ ΦΙΛΟΞΕΝΙΑΣ ΚΑΙ ΤΟΥ
ΤΟΥΡΙΣΜΟΥ



ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Sentiment classification σε online κριτικές ξενοδοχείων της
Κέρκυρας με τη χρήση επιβλεπόμενων μεθόδων μηχανικής μάθησης**

ΑΝΔΡΙΑΝΑ ΑΛΕΞΑΚΗ – ΑΜ: trs.dit2201

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ

ΚΕΡΜΑΝΙΔΟΥ ΚΑΤΙΑ-ΛΗΔΙΑ

Αναπληρώτρια Καθηγήτρια Τμήματος Πληροφορικής

ΚΕΡΚΥΡΑ

Σεπτέμβριος 2024

Πίνακας περιεχομένων

1. ABSTRACT	3
2. Εισαγωγή.....	3
3. Ανάλυση Συναισθημάτων σε Διαδικτυακές Κριτικές, Επιδράσεις του eWoM και η Λύση μέσω Μηχανικής Μάθησης	4
3.1. Διερεύνηση της επιρροής του Electronic Word-of-Mouth και η αντιμετώπιση των σχετικών προκλήσεων με καινοτόμες λύσεις.....	4
3.2. Η Επίδραση των Διαδικτυακών Κριτικών στην Εμπειρία των Ταξιδιωτών	5
3.3. Η Κρισιμότητα των Διαδικτυακών Αξιολογήσεων για τις Επιχειρήσεις	7
3.4. Travel Blogs και η Εξέλιξη του TripAdvisor	7
3.5. Τεχνητή Νοημοσύνη, Μηχανική Μάθηση και Επεξεργασία Φυσικής Γλώσσας (NLP) 9	
4. Θεωρητική Επισκόπηση και παραδείγματα	10
5. Βασικές Έννοιες και Υπόβαθρο για το sentiment classification	11
5.1. Waikato Environment for Knowledge Analysis	11
5.2. Επιβλεπόμενοι και Μη-επιβλεπόμενοι αλγόριθμοι μηχανικής μάθησης	12
5.3 Αλγόριθμοι Μηχανικής Μάθησης	13
5.3.1 Δένδρα αποφάσεων.....	13
5.3.2 Instance-Based K-nearest neighbors (IBK).....	15
5.3.3 Naive Bayes	16
5.4 Κατηγοριοποίηση Συναισθήματος (Sentiment Classification)	17
5.4.1 Τύποι Sentiment Analysis	17
5.4.2 Άλλοι Μέθοδοι Εξαγωγής Feedback.....	18
6. Μεθοδολογία	19
6.1 Συλλογή Δεδομένων.....	20
6.2 Περιγραφή και ανάλυση των δεδομένων.	21
6.3 Προ επεξεργασία δεδομένων.....	22
6.4 Εκπαίδευση των Μοντέλων	27
6.5 Σετ πειραμάτων 1.....	27
6.6 Σετ πειραμάτων 2	30
6.7 Σετ πειραμάτων 3	34
6.8 Σετ πειραμάτων 4.....	37
7. Αποτελέσματα.....	41

7.1 Σετ πειραμάτων 1	41
7.2 Σετ πειραμάτων 2	44
7.3 Σετ πειραμάτων 3	47
7.4 Σετ πειραμάτων 4	50
7.5 Σύγκριση Αλγορίθμων	54
7.6 Σύγκριση με προηγούμενες μελέτες	58
8. Δυσκολίες.....	59
8.1 Λήψη Δεδομένων	59
8.2 Επισημάνση Δεδομένων (Labeling).....	59
8.3 Δημιουργία Διανύσματος Χαρακτηριστικών (Feature Vector)	59
8.4 Ανισορροπία Δείγματος	60
9. Συμπεράσματα.....	60
8.1. Επιλογή Χαρακτηριστικών (Feature Selection).....	61
8. 2. Διαχείριση Ανισορροπημένων Δεδομένων	61
8. 3. Προεπεξεργασία Δεδομένων	61
8. 4. Χρήση Περισσότερων Δεδομένων	61
8.5 Συμπεράσματα	62
10. Προτάσεις για Μελλοντική Έρευνα.....	62
9.1 Εφαρμογή Μετα-Μάθησης (Meta-Learning):.....	62
9.2 Χρήση Προηγμένων Νευρωνικών Δικτύων:.....	62
9.3 Ανάλυση Συναισθημάτων σε Πραγματικό Χρόνο:	63
9.4 Ενσωμάτωση Πολυτροπικών Δεδομένων:.....	63
9.5 Εξέταση Νέων Μεθόδων Προεπεξεργασίας Δεδομένων:	63
9.6 Ενσωμάτωση Κοινωνικών Δικτύων και Big Data:.....	63
9.7 Εφαρμογές στην Ιατρική και την Υγεία:.....	63
9.8 Βελτίωση Εξυπηρέτησης Πελατών:.....	64
9.9 Τεχνητή Νοημοσύνη και Αυτοματισμοί:.....	64
11. Βιβλιογραφία	64
12. Παράρτημα	65
11.1 Κώδικας εξόρυξης Δεδομένων	65

1. ABSTRACT

Η παρούσα διπλωματική εργασία επικεντρώνεται στην κατηγοριοποίηση συναισθημάτων σε διαδικτυακές κριτικές ξενοδοχείων της Κέρκυρας, χρησιμοποιώντας επιβλεπόμενες μεθόδους μηχανικής μάθησης. **Ο στόχος είναι η ανάπτυξη μοντέλων που μπορούν να κατηγοριοποιήσουν με ακρίβεια τις κριτικές ως θετικές, αρνητικές ή ουδέτερες, εφαρμόζοντας διάφορους αλγόριθμους επιβλεπόμενης μηχανικής μάθησης, όπως Random Forest, Naive Bayes, και k-Nearest Neighbors (k-NN), στο περιβάλλον Waikato Environment for Knowledge Analysis (WEKA).** Η μελέτη αντιμετωπίζει προκλήσεις όπως η ανισορροπία των κλάσεων και η προεπεξεργασία δεδομένων, χρησιμοποιώντας τεχνικές όπως η Synthetic Minority Over-sampling Technique (SMOTE) για τη βελτίωση της απόδοσης των μοντέλων. **Τα πειραματικά αποτελέσματα δείχνουν ότι ο αλγόριθμος J48 παρέχει την υψηλότερη ακρίβεια στην κατηγοριοποίηση συναισθημάτων, ιδιαίτερα όταν τα δεδομένα είναι ισορροπημένα.** Αυτή η έρευνα συμβάλλει στη βιομηχανία του τουρισμού προσφέροντας πληροφορίες σχετικά με τα συναισθήματα των πελατών, επιτρέποντας στις επιχειρήσεις να λαμβάνουν αποφάσεις βασισμένες σε δεδομένα για τη βελτίωση της ποιότητας των υπηρεσιών και της ικανοποίησης των πελατών.

Key Words: Sentiment analysis, Sentiment Classification, Text Mining, Machine Learning, Deep Learning, Online Reviews, Natural Language Processing, Word of Mouth

2. Εισαγωγή

Τα ταξίδια και ο τουρισμός είναι ένας από τους πιο υποσχόμενους τομείς στον κόσμο και η πληροφόρηση είναι εξαιρετικά σημαντική στον κλάδο αυτό. Στις μέρες μας, ο τρόπος με τον οποίο οι ταξιδιώτες χρησιμοποιούν αυτές τις πληροφορίες σε όλη τη διαδικασία λήψης αποφάσεων για το ταξίδι τους έχει αλλάξει. Η ευρεία ανάπτυξη της τεχνολογίας πληροφοριών και επικοινωνιών, ιδίως στον τουριστικό κλάδο, είναι ένας από τους παράγοντες που έκαναν αυτή την αλλαγή να συμβεί. Επίσης, συνδέεται κυρίως με την εμφάνιση των travel blogs ως μια από τις πλατφόρμες Web 2.0, τα οποία πρόσφεραν διαδραστικότητα στους χρήστες.

Στην σημερινή ψηφιακή εποχή, **οι διαδικτυακές κριτικές έχουν αναδειχθεί σε καθοριστικό παράγοντα που διαμορφώνει τις αποφάσεις των καταναλωτών-ταξιδιωτών** (Gretzel & Yoo, 2008). Η σημασία των διαδικτυακών κριτικών εκτείνεται πολύ πέρα από τις απλές εκφράσεις ικανοποίησης ή δυσαρέσκειας- χρησιμεύουν ως ‘πυξίδα’ για τα άτομα που περιηγούνται σε μια ‘θάλασσα’ επιλογών, καθοδηγώντας τους προς τις βέλτιστες αποφάσεις αγοράς, γαστρονομικές εμπειρίες, ταξιδιωτικούς προορισμούς και

πολλά άλλα. Οι διαδικτυακές κριτικές αποτελούν ισχυρό εργαλείο διάδοσης πληροφοριών, καθώς όχι μόνο επηρεάζουν τις επιλογές των καταναλωτών-ταξιδιωτών, αλλά διαδραματίζουν σημαντικό ρόλο στη διαμόρφωση της φήμης και της επιτυχίας των επιχειρήσεων. Τελικά, η κατανόηση της σπουδαιότητας των διαδικτυακών κριτικών αναδεικνύεται ζωτικής σημασίας τόσο για τους καταναλωτές-ταξιδιώτες όσο και για τις επιχειρήσεις, καθώς ο τομέας του τουρισμού αναμετράται σε ένα περιβάλλον υψηλού ανταγωνισμού.

3.Ανάλυση Συναισθημάτων σε Διαδικτυακές Κριτικές, Επιδράσεις του eWoM και η Λύση μέσω Μηχανικής Μάθησης

3.1. Διερεύνηση της επιρροής του Electronic Word-of-Mouth και η αντιμετώπιση των σχετικών προκλήσεων με καινοτόμες λύσεις

Το eWOM ή αλλιώς Electronic Word-of-Mouth αναφέρεται στην ψηφιακή ή διαδικτυακή εκδοχή της παραδοσιακής επικοινωνίας από στόμα σε στόμα. Το eWOM περιλαμβάνει τις απόψεις, τις συστάσεις και τις εμπειρίες των ατόμων σχετικά με προϊόντα, υπηρεσίες ή εμπορικά σήματα τις οποίες μοιράζονται σε διάφορες διαδικτυακές πλατφόρμες, όπως τα μέσα κοινωνικής δικτύωσης, ιστοσελίδες κριτικών, φόρουμ και άλλα ψηφιακά κανάλια. Έχει καταστεί σημαντικός παράγοντας στη διαμόρφωση των αντιλήψεων των καταναλωτών και επηρεάζει τις αγοραστικές αποφάσεις στο διαδικτυακό περιβάλλον καθώς θεωρείται πιο αξιόπιστη πηγή από τη διαφήμιση (Cantallops & Salvi, 2014). Το eWOM μπορεί να είναι τόσο θετικό όσο και αρνητικό, και ο αντίκτυπός του συχνά μεγεθύνεται λόγω της ευρείας εμβέλειας και προσβασιμότητας των ψηφιακών blog. Οι επιχειρήσεις και οι έμποροι πρέπει να παρακολουθούν στενά το eWOM με σκοπό να κατανοήσουν τα συναισθήματα των καταναλωτών, να διαχειριστούν τη διαδικτυακή φήμη τους και να αξιοποιήσουν τις θετικές κριτικές για σκοπούς μάρκετινγκ, αλλά και τις αρνητικές ως ευκαιρία για βελτίωση.

Η κατανόηση, δηλαδή, των συναισθημάτων των ταξιδιωτών κρίνεται αναγκαία για τις τουριστικές επιχειρήσεις. Ωστόσο, με τον τεράστιο όγκο κριτικών που προκύπτει από την πολυάριθμη προσφορά ταξιδιωτικών blog, η ανάγνωση και αξιολόγησή τους γίνεται ανέφικτη για τις επιχειρήσεις, αλλά και για τους ταξιδιώτες. Μια λύση σε αυτό το πρόβλημα παρέχεται μέσω της ανάλυσης συναισθημάτων. Αυτή η τεχνική NLP¹, γνωστή και ως sentiment analysis, επιτρέπει στις επιχειρήσεις και στους ταξιδιώτες να αξιολογήσουν αυτόματα τον τόνο των κριτικών,

¹ NLP: Natural Language Processing

κατηγοριοποιώντας τα ως θετικά, αρνητικά ή ουδέτερα. Η διαδικασία αυτή περιλαμβάνει την προεπεξεργασία κειμένου, την εξαγωγή χαρακτηριστικών και την εκπαίδευση μοντέλων μηχανικής μάθησης. **Η ανάλυση συναισθήματος μας δίνει τη δυνατότητα να αναλύσουμε μεγάλους όγκους δεδομένων κειμένου και να αποκτήσουμε πολύτιμες πληροφορίες για τα σχόλια των πελατών,** επιτρέποντας μας την πιο τεκμηριωμένη λήψη αποφάσεων και την ανταποκρινόμενη εμπλοκή με τα συναισθήματα που εκφράζονται στο διαδίκτυο.

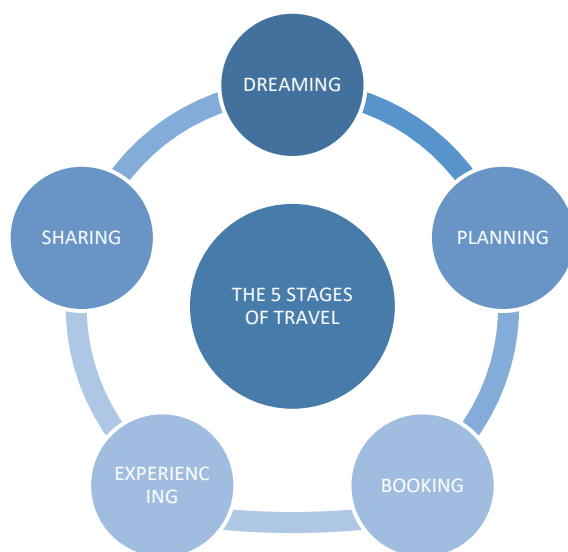
Η ανάλυση συναισθήματος όμως αποτελεί μια πολύπλοκη τεχνική NLP. Γεγονός το οποίο προκύπτει από τη διαφοροποιημένη φύση της ανθρώπινης γλώσσας και των συναισθημάτων. Η γλώσσα είναι εγγενώς διαφορούμενη και τα συναισθήματα μπορούν να εκφραστούν με ποικίλους τρόπους, όπως σαρκασμός και ειρωνεία. Επιπλέον, η ανάλυση συναισθήματος αντιμετωπίζει προκλήσεις κατά το χειρισμό εκφράσεων που εξαρτώνται από τα συμφραζόμενα και τις εξελισσόμενες γλωσσικές τάσεις. Το συναίσθημα, επίσης ποικίλλει σε διάφορους τομείς, καθιστώντας δύσκολη τη δημιουργία καθολικά εφαρμόσιμων μοντέλων. Η ενσωμάτωση, τελικά αλγορίθμων μηχανικής μάθησης απαιτεί μεγάλα όγκου δεδομένων (Big Data²) και συνεχή βελτίωση του μοντέλου ώστε να προσαρμόζεται στις εξελισσόμενες γλωσσικές τάσεις.

3.2. Η Επίδραση των Διαδικτυακών Κριτικών στην Εμπειρία των Ταξιδιωτών

Οι διαδικτυακές κριτικές έχουν εξελιχθεί σε καθοριστικό παράγοντα για τη λήψη αποφάσεων των ταξιδιωτών. Αναδεικνύονται ως αναπόσπαστη πηγή πληροφοριών και συμβουλών, επηρεάζοντας τις επιλογές τους σε κάθε στάδιο του ταξιδιού, από τον προγραμματισμό μέχρι την αξιολόγηση της εμπειρίας τους (Dwityas & Briandana, 2017). Τα στάδια του ταξιδιού είναι πέντε (SmartArt 1) και διαμορφώνονται σημαντικά από τον αντίκτυπο των διαδικτυακών κριτικών. Αναλυτικότερα, από την αρχική σπίθα έμπνευσης (**Dreaming**), όπου οι κριτικές επηρεάζουν τις επιλογές προορισμού, μέχρι το στάδιο του σχεδιασμού (**Planning**), όπου οι ταξιδιώτες βασίζονται σε σχόλια για τεκμηριωμένες αποφάσεις. Καθώς τα άτομα προχωρούν στο στάδιο της κράτησης (**Booking**), οι θετικές κριτικές επικυρώνουν τις επιλογές τους, ενώ οι αρνητικές προκαλούν επανεξέταση. Ακόμη και κατά τη διάρκεια του “πραγματικού” ταξιδιού (**Experiencing**), οι διαδικτυακές κριτικές παίζουν καθοριστικό ρόλο στις επιτόπιες αποφάσεις. Στο τελικό στάδιο, οι ταξιδιώτες μοιράζονται τις εμπειρίες τους μέσω κριτικών (**Sharing**), συμβάλλοντας στο **eWOM**. Τέλος, η έρευνα των Gretzel και Yoo (Gretzel & Yoo, 2008) έχει αποδείξει ότι οι ταξιδιώτες βασίζονται στις διαδικτυακές κριτικές κυρίως για τις αποφάσεις τους σχετικά με τη διαμονή (Booking), παρά στον προγραμματισμό των ταξιδιών κατά τη

² Big Data: Τα μεγάλου όγκου δεδομένα αναφέρονται σε εξαιρετικά μεγάλα και πολύπλοκα σύνολα δεδομένων που υπερβαίνουν τις δυνατότητες των παραδοσιακών μεθόδων επεξεργασίας δεδομένων, απαιτώντας προηγμένες τεχνολογίες για την ανάλυση, την αποθήκευση και την εξαγωγή πολύτιμων αποτελεσμάτων.

διάρκεια της διαμονής τους (Experiencing). Το συγκεκριμένο, όμως αποτέλεσμα μπορεί να διαφέρει με βάση την ηλικιακή ομάδα και το φύλλο.



SMARTART 1 ΤΑ ΠΕΝΤΕ ΣΤΑΔΙΑ ΤΟΥ ΤΑΞΙΔΙΟΥ

Οι διαδικτυακές κριτικές μπορούν, επίσης, να καθησυχάσουν τους ταξιδιώτες οι οποίοι μπορεί να βιώνουν pre-travel anxiety³. Η αυθεντικότητα που εμπεριέχεται στις κριτικές ενισχύει την αίσθηση εμπιστοσύνης και διαφάνειας, στοιχεία ζωτικής σημασίας για όσους αναζητούν εμπιστοσύνη στις ταξιδιωτικές τους αποφάσεις. Πέρα από απλές αξιολογήσεις, αυτές οι μαρτυρίες αποτελούν έναν θησαυρό από συμβουλές εκ των έσω, παρέχοντας ανεκτίμητες συμβουλές για κρυμμένα διαμάντια, τοπικές γνώσεις και πρακτικές προτάσεις. Εμβαθύνοντας στη συλλογική σοφία της ταξιδιωτικής κοινότητας, τα άτομα μπορούν να μετριάσουν τους πιθανούς κινδύνους, να αντιμετωπίσουν προληπτικά τις ανησυχίες τους και να λάβουν τεκμηριωμένες αποφάσεις που ευθυγραμμίζονται με τις προσδοκίες τους. Στην ουσία, οι διαδικτυακές κριτικές χρησιμεύουν όχι μόνο ως πρακτικός οδηγός αλλά και ως υποστηρικτική ταξιδιωτική κοινότητα, ενδυναμώνοντας τα άτομα να ξεκινήσουν τα ταξίδια τους με αίσθημα ετοιμότητας και αυτοπεποίθησης, περιορίζοντας τελικά το άγχος πριν από το ταξίδι.

Η σημασία των διαδικτυακών κριτικών έγκειται στην ικανότητά τους να χρησιμεύουν ως eWOM, προσφέροντας πολύτιμες πληροφορίες, προοπτικές και συστάσεις. Οι διαδικτυακές αξιολογήσεις συμβάλλουν στη διαφάνεια, βοηθούν στον μετριασμό του κινδύνου και ενισχύουν τους καταναλωτές με τις γνώσεις που χρειάζονται για τη λήψη τεκμηριωμένων αποφάσεων. Η κατανόηση της σημασίας των διαδικτυακών κριτικών εκτείνεται πέρα από τις ατομικές προτιμήσεις, διαμορφώνει τις τάσεις της αγοράς, επηρεάζει τις επιχειρήσεις και προωθεί ένα δυναμικό ψηφιακό

³ Σύμφωνα με τους Gudykunst και Hammer (Gudykunst & Hammer, 1988), το άγχος είναι το συναίσθημα που "αναφέρεται στο φόβο των αρνητικών συνεπειών". Όταν ένα άτομο σχεδιάζει μια αγορά (ή στην προκειμένη περίπτωση ένα ταξίδι) που είναι ριψοκίνδυνη, αυτό οδηγεί σε φόβο για άγνωστες συνέπειες και αίσθημα "άγχους".

οικοσύστημα όπου η συλλογική φωνή των καταναλωτών ασκεί ουσιαστική επιρροή.

3.3. Η Κρισιμότητα των Διαδικτυακών Αξιολογήσεων για τις Επιχειρήσεις

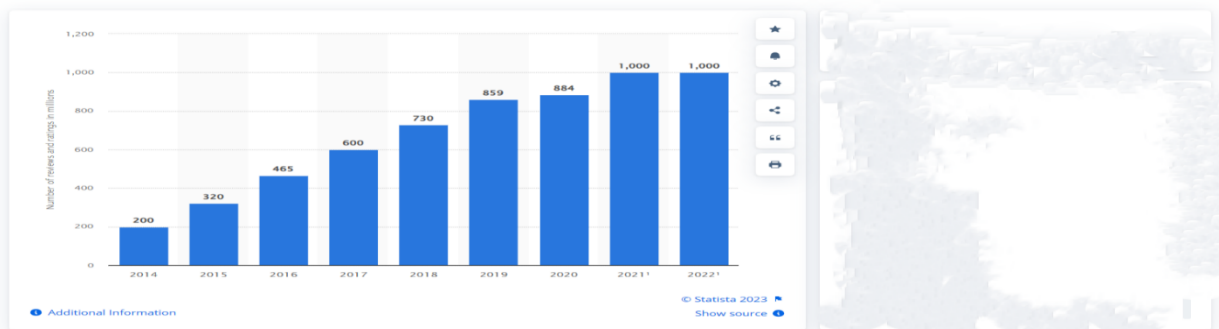
Ο τουρισμός αποτελεί μια από τις σημαντικότερες οικονομικές δραστηριότητες στην Ελλάδα. Είναι άλλωστε εμπειρικά γνωστό σε όλους πως αρκετές περιοχές της Ελλάδας, κυρίως νησιωτικές και παραθαλάσσιες, ζουν σχεδόν αποκλειστικά από τον τουρισμό. Με βάση τα παραπάνω, φαίνεται ότι η τουριστική βιομηχανία στην Ελλάδα και ευρύτερα βρίσκεται σε ένα άκρος ανταγωνιστικό περιβάλλον. Οι επιχειρήσεις στον τομέα αυτό πρέπει να ανταγωνιστούν για την προσοχή και την επιλογή των ταξιδιωτών, καθώς, οι ταξιδιώτες μέσω του διαδικτύου έχουν τη δυνατότητα να λάβουν αποφάσεις βασιζόμενοι σε ποίκιλα κριτήρια, όπως για παράδειγμα τη τιμή, τη τοποθεσία, τις παροχές, τα σχόλια κ.α. **Ένας αποτελεσματικός τρόπος για τις τουριστικές επιχειρήσεις να κερδίσουν όμως ανταγωνιστικό πλεονέκτημα και να ξεχωρίσουν είναι μέσω των διαδικτυακών κριτικών.** Πέρα από τις απλές μαρτυρίες, οι διαδικτυακές κριτικές συμβάλλουν στο ψηφιακό αποτύπωμα μιας τουριστικής επιχείρησης, επηρεάζοντας την ορατότητα στις μηχανές αναζήτησης. Σε αυτή την εποχή του ψηφιακού περιεχομένου που δημιουργείται από τους χρήστες του διαδικτύου, η σημασία της διαχείρισης και της ανταπόκρισης στις διαδικτυακές κριτικές δεν πρέπει να υποτιμηθεί, καθώς δεν επηρεάζουν μόνο το παρόν, αλλά συμβάλλουν στη μακροπρόθεσμη επιτυχία μιας επιχείρησης στο διαρκώς εξελισσόμενο πεδίο των ταξιδιών. Τέλος, **οι κριτικές προσφέρουν σημαντικά πλεονεκτήματα στον ανταγωνιστικό αγώνα για την επιβίωση και την επιτυχία των επιχειρήσεων στον τομέα του τουρισμού.**

3.4. Travel Blogs και η Εξέλιξη του TripAdvisor

Όλο και περισσότεροι ταξιδιώτες στρέφονται σε πλατφόρμες κοινωνικής δικτύωσης, όπως το Facebook, το You Tube, το Instagram και τα travel blogs (TripAdvisor, Google κ.α), για να λάβουν πληροφορίες από πρώτο χέρι, να μάθουν για τις εμπειρίες, τις κριτικές και τις συστάσεις άλλων για διάφορους προορισμούς, υπηρεσίες κ.α. Λόγω αυτού του γεγονότος, **τα τελευταία χρόνια παρατηρείται έντονη άνοδος στον αριθμό των κριτικών που κοινοποιούνται στις διαδικτυακές πλατφόρμες που προαναφέρθηκαν.** Ανάμεσα σε αυτές, ξεχωρίζει η δημοφιλής πλατφόρμα TripAdvisor, όπου το 2022 είχε 1 δισεκατομμύριο συνολικό αριθμό σχολίων και βαθμολογιών (Εικόνα 1). Το TripAdvisor είναι μια διαδικτυακή πλατφόρμα που έχει μεταμορφώσει τον τρόπο με τον οποίο οι ταξιδιώτες σχεδιάζουν τα ταξίδια τους και μοιράζονται τις εμπειρίες τους (About us: Tripadvisor, 2024). Η πλατφόρμα ιδρύθηκε το 2000 από τον Stephen Kaufer και αρχικά επικεντρώθηκε στην παροχή αξιόπιστων κριτικών

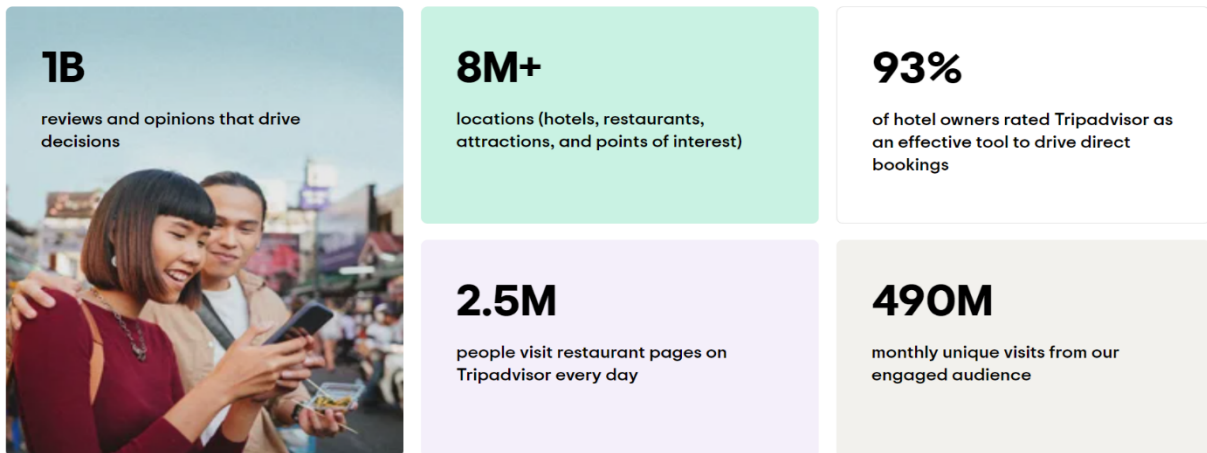
για ξενοδοχεία. Με την πάροδο των ετών, το TripAdvisor επέκτεινε το πεδίο εφαρμογής του για να συμπεριλάβει κριτικές και πληροφορίες για εστιατόρια, αξιοθέατα και διάφορες υπηρεσίες που σχετίζονται με τα ταξίδια. **Το καινοτόμο μοντέλο του βασίζεται στο περιεχόμενο που δημιουργείται από τους χρήστες, επιτρέποντας στους ταξιδιώτες να μοιράζονται τις απόψεις και τις γνώσεις τους, δημιουργώντας μια τεράστια βάση δεδομένων με κριτικές.** Η επιτυχία του TripAdvisor αποδίδεται στη δέσμευσή του για διαφάνεια και εκδημοκρατισμό των ταξιδιωτικών πληροφοριών. Παρά τις προκλήσεις που αντιμετωπίζει, όπως η διαχείριση των ψεύτικων κριτικών, το TripAdvisor παρέμεινε πηγή πληροφόρησης για εκατομμύρια ταξιδιώτες παγκοσμίως, επηρεάζοντας τις αποφάσεις και διαμορφώνοντας το τοπίο της τουριστικής βιομηχανίας (Εικόνα 2). Σήμερα, η πλατφόρμα συνεχίζει να εξελίσσεται, ενσωματώνοντας χαρακτηριστικά όπως δυνατότητες κρατήσεων και εξατομικευμένες συστάσεις, διατηρώντας την ιδιότητά της ως κομβικού παίκτη στην ταξιδιωτική κοινότητα.

Total number of user reviews and ratings on Tripadvisor worldwide from 2014 to 2022
(in millions)



ΕΙΚΟΝΑ 1 ΣΥΝΟΛΙΚΟΣ ΑΡΙΘΜΟΣ ΣΧΟΛΙΩΝ ΚΑΙ ΒΑΘΜΟΛΟΓΙΩΝ ΠΑΓΚΟΣΜΙΩΣ ΣΤΟ TRIPADVISOR ΑΠΟ ΤΟ 2014 ΕΩΣ ΤΟ 2022 (ΣΕ ΕΚΑΤΟΜΜΥΡΙΑ). ΠΗΓΗ: STATISTA⁴

⁴ [Statista - The Statistics Portal for Market Data, Market Research and Market Studies](#)



ΕΙΚΟΝΑ 2 ΔΕΔΟΜΕΝΑ ΓΙΑ ΤΟ TRIPADVISOR ΠΗΓΗ: TRIPADVISOR⁵

3.5. Τεχνητή Νοημοσύνη, Μηχανική Μάθηση και Επεξεργασία Φυσικής Γλώσσας (NLP)

Η Τεχνητή Νοημοσύνη είναι μια ευρεία και, καμιά φορά, ασαφής έννοια που απασχολεί πολλές ερμηνείες και συζητήσεις, προκαλώντας σύγχυση με τους διάφορους ορισμούς που προτείνονται. Ένας κοινός ορισμός της ΤΝ αναφέρεται στη δημιουργία υπολογιστικών συστημάτων που επιτρέπουν στις μηχανές να μιμούνται διάφορες σύνθετες ανθρώπινες δεξιότητες (Sheikh, Prins, & Schrijvers, 2021). Αυτά τα υπολογιστικά συστήματα περιλαμβάνουν εργασίες που συνδέονται με τη μάθηση, τη σκέψη και την αυτοδιόρθωση. Απώτερος στόχος τους είναι να αναπαράγουν και να επεκτείνουν την ανθρώπινη νοημοσύνη, χρησιμοποιώντας υπολογιστικά συστήματα. Η ΤΝ στοχεύει επίσης στη βελτίωση των υπολογιστικών ικανοτήτων μέσω της ανάπτυξης αλγορίθμων και τεχνικών προγραμματισμού, με σκοπό την αποτελεσματική χρήση των υπολογιστικών πόρων. (Kok, Boers, Kusters, & Putten, 2009).

Η Τεχνητή Νοημοσύνη περιλαμβάνει διάφορες τεχνικές, συμπεριλαμβανομένης της μηχανικής μάθησης. Η μηχανική μάθηση είναι ένα υποσύνολο της τεχνητής νοημοσύνης που εστιάζει στην ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στους υπολογιστές να μαθαίνουν αυτόνομα από τα δεδομένα χωρίς ρητό προγραμματισμό. Αυτό σημαίνει ότι τα συστήματα μηχανικής μάθησης μπορούν να αναγνωρίζουν πρότυπα, να προσαρμόζονται σε νέες πληροφορίες, και να βελτιώνουν την επίδοσή τους με τον χρόνο. Αυτή η προσέγγιση χρησιμοποιείται ευρέως σε πολλούς τομείς, όπως η αναγνώριση προτύπων, η αναλυτική πρόβλεψη, και η αυτόματη διακριτοποίηση δεδομένων. Η μηχανική μάθηση χωρίζεται σε τρεις κύριες κατηγορίες (Sheikh, Prins, & Schrijvers, 2021). Σε επιβλεπόμενη μηχανική μάθηση, όπου περιλαμβάνει την εκπαίδευση αλγορίθμων σε σύνολα δεδομένων με

⁵ [TripAdvisor for Business | Reach Millions of Travelers](#)

ετικέτες, όπου το μοντέλο μαθαίνει να αντιστοιχίζει τα δεδομένα εισόδου σε αντίστοιχες ετικέτες εξόδου, καθιστώντας το κατάλληλο για εργασίες όπως η ταξινόμηση και η παλινδρόμηση. Η μη-επιβλεπόμενη μηχανική μάθηση, από την άλλη πλευρά, εκπαιδεύει αλγορίθμους σε σύνολα δεδομένων χωρίς ετικέτες, επιτρέποντάς τους να βρίσκουν μοτίβα και σχέσεις μέσα στα δεδομένα χωρίς ρητές ετικέτες-στόχους, που χρησιμοποιούνται συνήθως για εργασίες ομαδοποίησης και συσχέτισης. Τέλος, reinforcement μηχανική μάθηση (RL) είναι ένας τύπος μηχανικής μάθησης όπου ένας πράκτορας μαθαίνει αλληλεπιδρώντας με ένα περιβάλλον, λαμβάνοντας ανατροφοδότηση με τη μορφή ανταμοιβών ή ποινών. Στόχος του πράκτορα είναι να λαμβάνει αποφάσεις που μεγιστοποιούν τις αθροιστικές ανταμοιβές με την πάροδο του χρόνου. Η RL χρησιμοποιείται σε εφαρμογές όπως τα παιχνίδια, ο ρομποτικός έλεγχος και τα αυτόνομα συστήματα, όπου η προσαρμοστική λήψη αποφάσεων είναι απαραίτητη. Αυτές οι κατηγορίες παρέχουν μια θεμελιώδη δομή για την κατανόηση της ποικιλομορφίας των τεχνικών μηχανικής μάθησης.

Η Επεξεργασία Φυσικής Γλώσσας ή NLP, αναφέρεται στον τομέα της τεχνητής νοημοσύνης και της υπολογιστικής γλωσσολογίας που επικεντρώνεται στο να επιτρέπει στους υπολογιστές να κατανοούν, να ερμηνεύουν και να παράγουν ανθρώπινη γλώσσα. Η NLP περιλαμβάνει την ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στις μηχανές να επεξεργάζονται και να αναλύουν μεγάλο όγκο δεδομένων φυσικής γλώσσας, εξάγοντας νόημα, εντοπίζοντας πρότυπα και εκτελώντας εργασίες όπως η γλωσσική μετάφραση, η ανάλυση συναισθήματος, η αναγνώριση ομιλίας και η παραγωγή κειμένου (Chowdhary, 2020). Στόχος της NLP είναι να γεφυρώσει το χάσμα μεταξύ της ανθρώπινης επικοινωνίας και της κατανόησης από τους υπολογιστές, διευκολύνοντας την απρόσκοπτη αλληλεπίδραση μεταξύ ατόμων και μηχανών με τη χρήση φυσικής γλώσσας.

4. Θεωρητική Επισκόπηση και παραδείγματα

Η ανάλυση συναισθημάτων σε διαδικτυακές κριτικές έχει προσελκύσει το ενδιαφέρον πολλών ερευνητών τα τελευταία χρόνια. Σύμφωνα με τους Cantallops και Salvi (Cantallops & Salvi, 2014), το ηλεκτρονικό στόμα με στόμα (eWOM) έχει σημαντική επίδραση στις αγοραστικές αποφάσεις των καταναλωτών και θεωρείται πιο αξιόπιστη πηγή από τη διαφήμιση. Επιπλέον, η μελέτη των Gretzel και Yoo (Gretzel & Yoo, 2008) δείχνει ότι οι ταξιδιώτες βασίζονται κυρίως στις διαδικτυακές κριτικές για να λάβουν αποφάσεις σχετικά με τη διαμονή τους.

Στο πλαίσιο αυτό, η μηχανική μάθηση και η ανάλυση συναισθημάτων χρησιμοποιούνται ευρέως για την εξαγωγή χρήσιμων πληροφοριών από τις κριτικές. Αλγόριθμοι όπως οι

Random Forest, J48, Naive Bayes και IBK (NN = 3) έχουν αποδειχθεί αποτελεσματικοί στην κατηγοριοποίηση των συναισθημάτων. Η χρήση τεχνικών **όπως η προεπεξεργασία δεδομένων, η επιλογή χαρακτηριστικών και η εξισορρόπηση των κλάσεων μέσω SMOTE συμβάλλει στη βελτίωση της απόδοσης αυτών των μοντέλων.**

Μια σχετική μελέτη των Zhang και Law (Ye, Zhang, & Law, 2009) χρησιμοποιώντας δεδομένα από το TripAdvisor, εξέτασε την ανάλυση συναισθημάτων με τη χρήση του αλγορίθμου Random Forest. Τα αποτελέσματα έδειξαν ότι ο Random Forest είχε υψηλή ακρίβεια στην πρόβλεψη των συναισθημάτων και ήταν ιδιαίτερα αποτελεσματικός στην κατηγοριοποίηση θετικών και αρνητικών κριτικών.

5. Βασικές Έννοιες και Υπόβαθρο για το sentiment classification

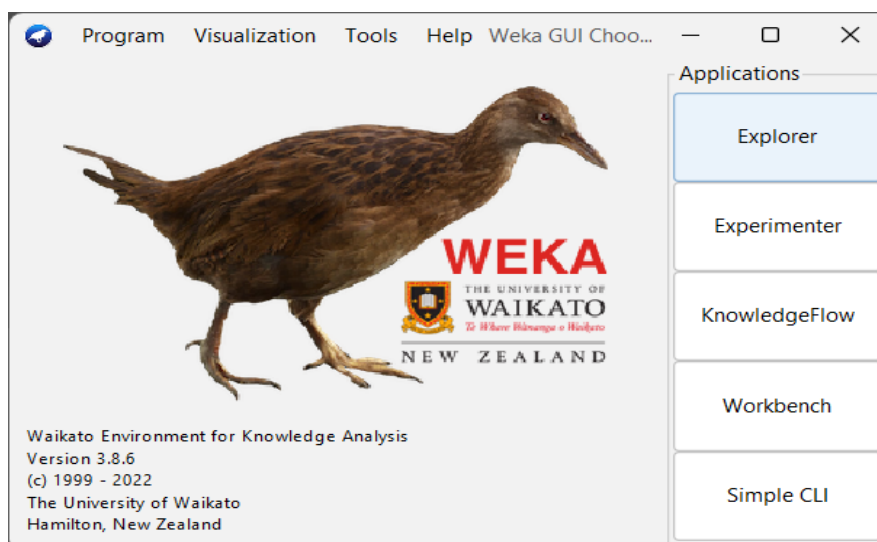
5.1. Waikato Environment for Knowledge Analysis

Το Waikato Environment for Knowledge Analysis (Hall, et al., 2009) ή εν συντομία WEKA δημιουργήθηκε από την αντιληπτή ανάγκη για ένα φιλικό προς το χρήστη εργαλείο που θα επέτρεπε στους ερευνητές εύκολη πρόσβαση σε σύγχρονες τεχνικές μηχανικής μάθησης. Κατά την έναρξη του το 1992, οι αλγόριθμοι μάθησης ήταν διαθέσιμοι σε διάφορες γλώσσες, για χρήση σε διαφορετικές πλατφόρμες και λειτουργούσαν σε διάφορες μορφές δεδομένων. **Σήμερα, το WEKA είναι ένα δημοφιλές εργαλείο ανοικτού κώδικα⁶ για μηχανική μάθηση, που επιτρέπει στους χρήστες να εκτελούν ποικίλες εργασίες όπως προεπεξεργασία δεδομένων, ταξινόμηση, παλινδρόμηση, ομαδοποίηση, εξόρυξη κανόνων συσχέτισης και οπτικοποίηση.** Το WEKA είναι αναπτυγμένο σε Java, καθώς προσφέρει και μια API σε Java για ενσωμάτωση σε προσαρμοσμένες εφαρμογές. Υποστηρίζει μια ευρεία γκάμα αλγορίθμων μηχανικής μάθησης όπως δέντρα αποφάσεων, support vectors machine, neural network, μεθόδους ensemble και πολλά άλλα, καθιστώντας το κατάλληλο τόσο για αρχάριους όσο και για έμπειρους χρήστες στον τομέα της μηχανικής μάθησης και της εξόρυξης δεδομένων. Η ευελιξία του, η ευκολία χρήσης και η εκτενής τεκμηρίωση έχουν συμβάλει στην ευρεία υιοθέτησή του τόσο σε ακαδημαϊκό όσο και σε βιομηχανικό περιβάλλον για την ανάλυση και τον πειραματισμό δεδομένων.

Στην παρούσα εργασία, θα χρησιμοποιήσουμε το Weka 3.8.6 για τους προαναφερθέντες λόγους. **Αναλυτικότερα, ο στόχος της παρούσας εργασίας είναι να παρουσιάσει τα βήματα που απαιτούνται για την εκτέλεση μιας ανάλυσης συναισθήματος διαδικτυακών κριτικών, χρησιμοποιώντας το WEKA ως μοναδικό εργαλείο-**

⁶ [Weka 3 - Data Mining with Open Source Machine Learning Software in Java \(waikato.ac.nz\)](http://weka.ac.nz)

αφετηρία μας είναι το άρθρο των (Ye, Zhang, & Law, 2009). Πιστεύουμε ότι, εκτός από ένα εξαιρετικό εργαλείο ανάλυσης, το WEKA είναι πολύ χρήσιμο σε εκπαιδευτικό πλαίσιο, διότι επιτρέπει στους φοιτητές να εξοικειωθούν με τους σημαντικότερους αλγορίθμους μηχανικής μάθησης και να εμβαθύνουν στις θεωρητικές πτυχές τους.



ΕΙΚΟΝΑ 3 WEKA MAIN MENU

5.2. Επιβλεπόμενοι και Μη-επιβλεπόμενοι αλγόριθμοι μηχανικής μάθησης

Το WEKA προσφέρει διάφορους αλγόριθμους μηχανικής μάθησης, τόσο επιβλεπόμενης (supervised) όσο και μη επιβλεπόμενης (unsupervised), που μπορούν να εφαρμοστούν σε σύνολα δεδομένων μετά τη φάση της προεπεξεργασίας. **Στην παρούσα μελέτη, θα επικεντρωθούμε κυρίως σε επιβλεπόμενες τεχνικές ταξινόμησης κειμένου για την εξόρυξη γνώμης, δηλαδή στη "μάθηση με επίβλεψη".** Αντίθετα, η προσέγγιση σημασιολογικού προσανατολισμού για την εξόρυξη γνώμης αποτελεί παράδειγμα "μάθησης χωρίς επίβλεψη", καθώς δεν απαιτεί προηγούμενη εκπαίδευση για την ανάλυση των δεδομένων. Σε αυτήν την προσέγγιση, λαμβάνεται υπόψη η σημασιολογική πλευρά των λέξεων και η κλίση τους προς το θετικό ή το αρνητικό φάσμα. **Η επικέντρωση στη επιβλεπόμενη μηχανική μάθηση επέρχεται από τη μελέτη των (Chaovalit & Zhou, 2005, January), οι οποίοι διαπίστωσαν ότι η προσέγγιση επιβλεπόμενης μηχανικής μάθησης ήταν πιο αξιόπιστη από την προσέγγιση μη επιβλεπόμενου σημασιολογικού προσανατολισμού.** Γενικότερα, στη μηχανική μάθηση με επίβλεψη, τα μοντέλα μαθαίνουν από ένα σύνολο δεδομένων όπου για κάθε είσοδο υπάρχει μια γνωστή επιθυμητή έξοδος. Αυτό επιτρέπει στα μοντέλα να προβλέπουν με μεγαλύτερη ακρίβεια νέα δεδομένα, καθώς μαθαίνουν τη σχέση μεταξύ των εισόδων και των επιθυμητών εξόδων σε σύγκριση με τη μηχανική μάθηση χωρίς επίβλεψη.

Μερικοί από τους αλγόριθμους επιβλεπόμενης μάθησης που είναι διαθέσιμοι στο Weka περιλαμβάνουν δέντρα αποφάσεων (όπως J48 και Random Forest), Support Vector Machines (SMO), k-nearest neighbors (IBK (NN = 3)), Naive Bayes και πολλούς άλλους.

5.3 Αλγόριθμοι Μηχανικής Μάθησης

5.3.1 Δένδρα αποφάσεων

Τα δέντρα απόφασης είναι μια μέθοδος μηχανικής μάθησης που χρησιμοποιείται για την ταξινόμηση ή την πρόβλεψη, κατασκευάζοντας ένα μοντέλο που λαμβάνει αποφάσεις βασισμένες σε μια σειρά από κανόνες που προκύπτουν από τα δεδομένα.

5.3.1.1 J48

Ο **J48** είναι ένας αλγόριθμος απόφασης που χρησιμοποιείται στην ανάλυση δεδομένων και την εξόρυξη γνώσης από δεδομένα. Είναι μια εφαρμογή του αλγορίθμου C4.5 που αναπτύχθηκε από τον Ross Quinlan (Quinlan, 1993). Ο **J48** χρησιμοποιείται συχνά σε μηχανική μάθηση και αναλύσεις δεδομένων για την κατασκευή δέντρων απόφασης που μπορούν να χρησιμοποιηθούν για την ταξινόμηση δεδομένων.

Ο αλγόριθμος J48 κατασκευάζει δέντρα απόφασης χρησιμοποιώντας μια μέθοδο επαναλαμβανόμενης διαίρεσης των δεδομένων σε υποσύνολα. Η διαίρεση βασίζεται στην επιλογή της ιδιότητας που προσφέρει την καλύτερη διαχωριστική ικανότητα, συχνά με βάση την μέτρηση του "Information Gain". Το δέντρο κατασκευάζεται με την επαναλαμβανόμενη εφαρμογή αυτής της διαδικασίας μέχρι να επιτευχθεί ένα συγκεκριμένο κριτήριο διακοπής, όπως το μέγεθος των φύλλων ή η καθαρότητα των υποσυνόλων.

Τα κύρια παραμετρικά στοιχεία που επηρεάζουν την απόδοσή του είναι:

- **Confidence Factor (CF):** Καθορίζει το επίπεδο εμπιστοσύνης για το κλάδεμα του δέντρου. Ένα μικρότερο CF οδηγεί σε πιο έντονο κλάδεμα, αποφεύγοντας την υπερεκπαίδευση.
- **MinNumObj:** Ο ελάχιστος αριθμός των αντικειμένων που απαιτούνται για τη δημιουργία ενός κόμβου. Ένας μικρός αριθμός μπορεί να οδηγήσει σε πιο σύνθετα δέντρα, ενώ ένας μεγαλύτερος αριθμός μπορεί να οδηγήσει σε πιο απλά δέντρα.
- **Binary Splits:** Καθορίζει αν οι διαχωρισμοί θα είναι δυαδικοί, απλοποιώντας τη δομή του δέντρου και διευκολύνοντας την ερμηνεία του αποτελέσματος.

Ο J48 επιλέχθηκε για την ικανότητά του να δημιουργεί εύκολα ερμηνεύσιμα μοντέλα και να χειρίζεται τόσο διακριτές όσο και συνεχιζόμενες μεταβλητές. Η δυνατότητα του να πραγματοποιεί κλάδεμα (pruning) βοηθά στην αποφυγή της υπερεκπαίδευσης, καθιστώντας τον κατάλληλο για προβλήματα με υψηλή διαστατικότητα και πολυπλοκότητα. Επιπλέον, ο J48 παρέχει σαφή και οπτικά ευανάγνωστα δέντρα αποφάσεων, που διευκολύνουν την κατανόηση των διαδικασιών λήψης αποφάσεων.

5.3.1.2 Random Forest

Ο αλγόριθμος Random Forest (Breiman, 2001) είναι μια ισχυρή και ευέλικτη μέθοδος μηχανικής μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Δημιουργεί ένα σύνολο από δέντρα απόφασης (forest) κατά την εκπαίδευση και παράγει την τελική πρόβλεψη μέσω της πλειοψηφικής ψήφου (στην ταξινόμηση) ή του μέσου όρου (στην παλινδρόμηση) των προβλέψεων αυτών των δέντρων. Η τυχαιοποίηση εισάγεται τόσο κατά την επιλογή των δειγμάτων εκπαίδευσης για κάθε δέντρο όσο και κατά την επιλογή των χαρακτηριστικών που χρησιμοποιούνται σε κάθε διαχωρισμό, γεγονός που συμβάλλει στη μείωση της υπερεκπαίδευσης (overfitting).

Στο WEKA, ο Random Forest υλοποιείται με τη δημιουργία πολλαπλών δέντρων απόφασης από διαφορετικά τυχαία δείγματα των δεδομένων και τη συνδυαστική χρήση των προβλέψεών τους για να επιτευχθεί μια πιο ακριβής και σταθερή πρόβλεψη. Αυτή η μέθοδος είναι ιδιαίτερα αποτελεσματική σε μεγάλα και πολύπλοκα σύνολα δεδομένων, καθιστώντας την μια δημοφιλή επιλογή για πολλές εφαρμογές μηχανικής μάθησης.

Τα κύρια παραμετρικά στοιχεία που επηρεάζουν την απόδοσή του είναι:

- **Number of Trees (n_estimators):** Καθορίζει τον αριθμό των δέντρων στο δάσος. Η αύξηση του αριθμού των δέντρων συνήθως βελτιώνει την ακρίβεια, αλλά αυξάνει τον υπολογιστικό χρόνο και τους πόρους που απαιτούνται. Ένας συνήθης αριθμός είναι μεταξύ 100 και 500 δέντρων, αλλά αυτό μπορεί να προσαρμοστεί ανάλογα με το μέγεθος και την πολυπλοκότητα των δεδομένων.
- **Maximum Depth (max_depth):** Το μέγιστο βάθος των δέντρων επηρεάζει την ικανότητα του μοντέλου να απομνημονεύει τα δεδομένα εκπαίδευσης. Ένα μεγαλύτερο βάθος μπορεί να οδηγήσει σε υπερεκπαίδευση (overfitting), ενώ ένα μικρότερο βάθος μπορεί να οδηγήσει σε υποεκπαίδευση (underfitting). Η βέλτιστη τιμή για το max_depth συνήθως προσδιορίζεται μέσω δοκιμών και σφαλμάτων ή διασταυρούμενης επικύρωσης (cross-validation).
- **Minimum Samples Split (min_samples_split):** Καθορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτούνται για να διαχωριστεί ένας κόμβος. Μικρότεροι αριθμοί μπορούν να δημιουργήσουν πιο πολύπλοκα δέντρα, ενώ μεγαλύτεροι αριθμοί βοηθούν στη μείωση της πολυπλοκότητας και της υπερεκπαίδευσης.
- **Minimum Samples Leaf (min_samples_leaf):** Καθορίζει τον ελάχιστο αριθμό δειγμάτων που πρέπει να έχει ένα φύλλο κόμβου. Αυτό βοηθά στη μείωση της πιθανότητας υπερεκπαίδευσης και εξασφαλίζει ότι τα φύλλα κόμβου περιέχουν αρκετά δείγματα για να είναι αντιπροσωπευτικά.
- **Bootstrap:** Καθορίζει αν τα υποσύνολα των δεδομένων εκπαίδευσης που χρησιμοποιούνται για την εκπαίδευση κάθε δέντρου θα δημιουργούνται με δειγματοληψία με αντικατάσταση (bootstrap sampling). Αυτή η τεχνική βελτιώνει τη γενίκευση του μοντέλου και την ακρίβειά του.

Ο Random Forest επιλέχθηκε λόγω της ικανότητάς του να αντιμετωπίζει προβλήματα υπερεκπαίδευσης και της ανθεκτικότητάς του σε δεδομένα με υψηλή διαστατικότητα. Η χρήση πολλαπλών δέντρων αποφάσεων βοηθά στην εξομάλυνση των σφαλμάτων που μπορεί να προκύψουν από μεμονωμένα δέντρα και παρέχει υψηλή ακρίβεια και σταθερότητα. Επιπλέον, η μέθοδος bootstrap και η τυχαία επιλογή χαρακτηριστικών κατά την κατασκευή των δέντρων συμβάλλουν στη βελτίωση της απόδοσης και της γενίκευσης του μοντέλου. Ο Random Forest είναι επίσης ευέλικτος και μπορεί να χρησιμοποιηθεί για μια ποικιλία προβλημάτων ταξινόμησης και παλινδρόμησης.

5.3.2 Instance-Based K-nearest neighbors (IBK)

Ο αλγόριθμος Instance-Based k-nearest neighbors (k-NN) (Cover, 1967) είναι μια απλή και διαισθητική μέθοδος μηχανικής μάθησης για ταξινόμηση και παλινδρόμηση, όπου η κατηγορία ή η τιμή μιας νέας παρατήρησης προβλέπεται με βάση τις κατηγορίες ή τιμές των k πλησιέστερων γειτόνων της στο σύνολο εκπαίδευσης. Η απόσταση μεταξύ των παρατηρήσεων μετريέται συνήθως με την Ευκλείδεια απόσταση, αν και μπορούν να χρησιμοποιηθούν και άλλες μετρικές αποστάσεων.

Ο IBK είναι η υλοποίηση του αλγορίθμου k-NN στο λογισμικό WEKA. Ο IBK προσφέρει επιπλέον δυνατότητες, όπως η επιλογή του βέλτιστου αριθμού γειτόνων k μέσω διασταυρούμενης επικύρωσης και η χρήση διαφορετικών μεθόδων υπολογισμού αποστάσεων, όπως η Ευκλείδεια ή η Manhattan απόσταση. Αυτό καθιστά τον IBK έναν ευέλικτο και ισχυρό αλγόριθμο για την ταξινόμηση και την παλινδρόμηση σε διάφορα είδη δεδομένων.

Τα κύρια παραμετρικά στοιχεία που επηρεάζουν την απόδοσή του είναι:

- **K (Number of Neighbors):** Καθορίζει τον αριθμό των γειτόνων που χρησιμοποιούνται για την ταξινόμηση. Μικρότερο K μπορεί να οδηγήσει σε υπερεκπαίδευση, ενώ μεγαλύτερο K μπορεί να μειώσει την ακρίβεια.
- **Distance Weighting:** Καθορίζει αν οι γείτονες θα έχουν όλοι την ίδια βαρύτητα ή αν οι πιο κοντινοί γείτονες θα έχουν μεγαλύτερη επιρροή.
- **Cross-Validation:** Χρησιμοποιείται για την επιλογή του καλύτερου K μέσω διασταυρούμενης επικύρωσης.

Ο IBK επιλέχθηκε λόγω της ικανότητάς του να χειρίζεται πολύπλοκα μοτίβα δεδομένων και της απλότητάς του στην κατανόηση και εφαρμογή. Είναι ιδιαίτερα χρήσιμος για δεδομένα με πολλές διαστάσεις και μπορεί να προσαρμοστεί εύκολα σε διαφορετικά είδη προβλημάτων ταξινόμησης. Επιπλέον, προσφέρει ευελιξία στην επιλογή των παραμέτρων, όπως ο αριθμός των γειτόνων και η μέθοδος ζύγισης των αποστάσεων.

5.3.3 Naive Bayes

Ο αλγόριθμος Naive Bayes (McCallum, 1998) είναι μια απλή, αλλά ισχυρή μέθοδος ταξινόμησης που βασίζεται στο Θεώρημα του Bayes με την υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών. Χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης κειμένου, όπως η φιλτράρισμα ανεπιθύμητων μηνυμάτων (spam filtering) και η ανάλυση συναισθήματος.

Στο WEKA, ο Naive Bayes υλοποιείται με έναν αλγόριθμο που υποθέτει ότι η παρουσία ή απουσία ενός συγκεκριμένου χαρακτηριστικού δεν σχετίζεται με την παρουσία ή απουσία οποιουδήποτε άλλου χαρακτηριστικού, δεδομένης της κατηγορίας. Ο Naive Bayes στο WEKA είναι ευέλικτος και μπορεί να προσαρμοστεί για διάφορα είδη δεδομένων, καθιστώντας τον κατάλληλο για πολλά διαφορετικά προβλήματα ταξινόμησης.

Τα κύρια παραμετρικά στοιχεία που επηρεάζουν την απόδοσή του είναι:

- UseKernelEstimator: Αν είναι true, χρησιμοποιείται εκτιμητής πυρήνα για συνεχείς μεταβλητές.
- UseSupervisedDiscretization: Αν είναι true, χρησιμοποιείται επιβλεπόμενη διακριτοποίηση για τη μετατροπή των συνεχών μεταβλητών σε διακριτές κατηγορίες.

Ο Naive Bayes επιλέχθηκε λόγω της ταχύτητας και της αποδοτικότητάς του, ειδικά για μεγάλα σύνολα δεδομένων. Παρά την απλότητα της υπόθεσης ανεξαρτησίας, συχνά παρέχει αξιόπιστα αποτελέσματα σε πολλά πρακτικά προβλήματα. Είναι επίσης ανθεκτικός στον θόρυβο και μπορεί να διαχειριστεί τόσο διακριτά όσο και συνεχή δεδομένα, καθιστώντας τον κατάλληλο για μια ευρεία ποικιλία εφαρμογών.

Στον παρακάτω πίνακα παρατίθενται μερικά από τα πλεονεκτήματα και μειονεκτήματα για καθέναν από τους παραπάνω αλγορίθμους.

Αλγόριθμος	Πλεονεκτήματα	Μειονεκτήματα
k-Nearest Neighbors (k-NN)	1. Απλότητα 2. Δεν χρειάζεται εκπαίδευση 3. Χρήση Διάφορων Μετρικών	1. Υπολογιστικά Ακριβός 2. Ευαίσθητος στον Θόρυβο 3. Απαιτεί Κανονικοποίηση δεδομένων
Naïve Bayes	1. Απλότητα και Ταχύτητα 2. Αποδοτικό σε μικρά Σετ δεδομένων 3. Διαχειρίζεται καλά τα δεδομένα υψηλής διάστασης	1. Υπόθεση Ανεξαρτησίας 2. Κακή Απόδοση σε σπάνιες κατηγορίες
Random Forest	1. Υψηλή ακρίβεια 2. Ανθεκτικό στην υπερεκπαίδευση 3. Διαχειρίζεται ετερογενή δεδομένα	1. Υπολογιστική πολυπλοκότητα 2. Δύσκολο να ερμηνευτεί 3. Απαιτεί προσεκτική ρύθμιση παραμέτρων
J48	1. Ερμηνευσιμότητα	1. Υπερεκπαίδευση

	2. Γρήγορη εκπαίδευση 3. Διαχειρίζεται μη γραμμικά δεδομένα 4. Διαχειρίζεται διαφορετικούς τύπους χαρακτηριστικών	2. Ευαισθησία στον θόρυβο 3. Πολυπλοκότητα με πολλά χαρακτηριστικά 4. Μπορεί να γίνει περίπλοκο με μεγάλα σύνολα δεδομένων
--	---	--

5.4 Κατηγοριοποίηση Συναισθήματος (Sentiment Classification)

Η κατηγοριοποίηση συναισθήματος αναφέρεται στη διαδικασία κατηγοριοποίησης του συναισθηματικού περιεχομένου ενός κειμένου, όπως μιας κριτικής, ενός σχολίου ή μιας αναφοράς, βάσει του είδους του συναισθήματος που εκφράζεται σε αυτό. Ο στόχος είναι να καθοριστεί αν η γενική τάση της γνώμης που εκφράζεται στο κείμενο είναι θετική, αρνητική ή ουδέτερη. Η ανάλυση συναισθήματος χρησιμοποιεί τεχνικές μηχανικής μάθησης και επεξεργασίας φυσικής γλώσσας (NLP) για να αυτοματοποιήσει την κατάταξη του συναισθηματικού περιεχομένου του κειμένου. Αυτό επιτρέπει την εξαγωγή χρήσιμων πληροφοριών και συμπερασμάτων, όπως η κατανόηση της αντίδρασης του κοινού σε ένα προϊόν, μια υπηρεσία ή ένα γεγονός.

Η κατηγοριοποίηση συναισθήματος μπορεί να εφαρμοστεί σε διάφορους τομείς, όπως το μάρκετινγκ, την εξυπηρέτηση πελατών, τη δημόσια υγεία και τα μέσα κοινωνικής δικτύωσης. Μέσω αυτής της ανάλυσης, οι επιχειρήσεις μπορούν να κατανοήσουν καλύτερα τις ανάγκες και τις αντιδράσεις των πελατών τους, να βελτιώσουν τις στρατηγικές τους και να ενισχύσουν την ικανοποίηση των πελατών. Επιπλέον, μπορεί να χρησιμοποιηθεί για την παρακολούθηση των τάσεων και την ανίχνευση κρίσεων, επιτρέποντας την άμεση αντίδραση σε αρνητικές τάσεις ή γεγονότα. Με την αυτοματοποίηση της διαδικασίας κατηγοριοποίησης συναισθημάτων, οι οργανισμοί μπορούν να επεξεργαστούν μεγάλα σύνολα δεδομένων γρήγορα και αποτελεσματικά, εξάγοντας πολύτιμα insights που θα ήταν δύσκολο να αποκτηθούν με χειροκίνητες μεθόδους.

5.4.1 Τύποι Sentiment Analysis

- **Lexicon-Based Sentiment Analysis**

Η Lexicon-Based μέθοδος ανάλυσης συναισθημάτων βασίζεται στη χρήση ενός προκαθορισμένου λεξικού συναισθημάτων, το οποίο περιέχει λίστες λέξεων και φράσεων που έχουν ανατεθεί σε συναισθηματικές κατηγορίες (π.χ., θετικά, αρνητικά, ουδέτερα). Αυτή η μέθοδος αξιολογεί το συναισθήμα ενός κειμένου με βάση την παρουσία αυτών των λέξεων. Ένα παράδειγμα είναι το SentiWordNet, ένα επεκτάσιμο λεξικό που αποδίδει συναισθηματικές βαθμολογίες σε λέξεις της αγγλικής γλώσσας.

- **Dictionary-Based Sentiment Analysis**

Η Dictionary-Based μέθοδος είναι παρόμοια με τη Lexicon-Based, αλλά βασίζεται σε πιο εκτεταμένα λεξικά συναισθημάτων, όπως το Harvard General Inquirer ή το AFINN. Αυτά τα λεξικά περιέχουν λέξεις που έχουν ταξινομηθεί με βάση τη συναισθηματική τους πολικότητα και δύναμη. Η μέθοδος αυτή συχνά συνδυάζεται με τεχνικές προεπεξεργασίας κειμένου, όπως η αφαίρεση σταματημένων λέξεων (stop words) και η μετατροπή των λέξεων σε βασικές μορφές (stemming).

- **Corpus-Based Sentiment Analysis**

Η Corpus-Based μέθοδος χρησιμοποιεί μεγάλες συλλογές κειμένων (corpora) για την εκπαίδευση μοντέλων μηχανικής μάθησης ή την εξαγωγή συναισθηματικών λεξικών. Αυτή η προσέγγιση είναι πιο δυναμική και μπορεί να προσαρμοστεί σε συγκεκριμένα πλαίσια ή γλώσσες. Για παράδειγμα, ένα corpus των Twitter αναρτήσεων μπορεί να χρησιμοποιηθεί για την εκπαίδευση ενός μοντέλου που ανιχνεύει συναισθήματα στα κοινωνικά δίκτυα.

5.4.2 Άλλοι Μέθοδοι Εξαγωγής Feedback

- **Ερωτηματολόγια**

Τα ερωτηματολόγια είναι μια κλασική μέθοδος συλλογής feedback από τους χρήστες. Μπορούν να είναι δομημένα με κλειστές ερωτήσεις για εύκολη ανάλυση ή ανοικτές για πιο ποιοτικές απαντήσεις. Είναι χρήσιμα για τη συλλογή συγκεκριμένων πληροφοριών και μπορούν να διανεμηθούν μέσω email, ιστοσελίδων, ή ακόμα και σε έντυπη μορφή.

- **Ιστοσελίδες - Travel Blogs**

Οι ιστοσελίδες και τα travel blogs αποτελούν πλούσιες πηγές feedback, ειδικά στον τομέα του τουρισμού. Οι ταξιδιώτες συχνά μοιράζονται τις εμπειρίες τους μέσω λεπτομερών άρθρων, τα οποία μπορούν να αναλυθούν για την εξαγωγή συναισθημάτων και την κατανόηση των προτιμήσεων και των αντιλήψεων τους για διάφορους προορισμούς και υπηρεσίες.

- **Mobile Applications**

Οι εφαρμογές κινητών προσφέρουν άμεση και εύκολη πρόσβαση για τη συλλογή feedback από τους χρήστες. Μέσω push notifications, in-app surveys, και αξιολογήσεων, οι επιχειρήσεις μπορούν να συλλέξουν δεδομένα σε πραγματικό χρόνο. Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη για τη συνεχή βελτίωση των υπηρεσιών.

- **Social Media**

Τα κοινωνικά δίκτυα όπως το Facebook, το Twitter, και το Instagram αποτελούν σημαντικές πηγές feedback. Οι χρήστες συχνά εκφράζουν τις απόψεις και τα συναισθήματά τους δημόσια, επιτρέποντας στις επιχειρήσεις να αναλύσουν αυτά τα δεδομένα για να εντοπίσουν τάσεις, να κατανοήσουν τις αντιδράσεις των καταναλωτών, και να διαχειριστούν τη φήμη τους. Τεχνικές όπως το web scraping και η ανάλυση συναισθημάτων μπορούν να εφαρμοστούν για την εξαγωγή χρήσιμων πληροφοριών από αυτές τις πλατφόρμες.

6. Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε στη παρούσα εργασία για την ανάλυση συναισθήματος σε διαδικτυακές κριτικές πεντάστερων ξενοδοχείων της Κέρκυρας από το TripAdvisor περιλαμβάνει τη συλλογή και προεπεξεργασία των δεδομένων, τον καθαρισμό τους μέσω τεχνικών όπως tokenization, αφαίρεση stop words, καθώς και τη χειροκίνητη επισήμανση των κριτικών σε αρνητικά, θετικά και ουδέτερα. Ακολούθησε ο διαχωρισμός των δεδομένων σε σύνολα εκπαίδευσης και δοκιμής, και στη συνέχεια εφαρμογή της μεθόδου SMOTE για εξισορρόπηση των δεδομένων σε κάποια από τα σετ δεδομένων μας. Τέλος, εκπαιδεύτηκαν διάφοροι αλγόριθμοι, όπως Random Forest και Naive Bayes, ενώ η απόδοση τους αξιολογήθηκε βάσει μετρικών όπως η ακρίβεια και η ευστοχία. Αναλυτικότερα:



6.1 Σύλλογή Δεδομένων

Ένας θεμελιώδης παράγοντας για την ακρίβεια των μοντέλων μηχανικής μάθησης είναι η απόκτηση ενός σημαντικού όγκου δεδομένων. Η ιστοσελίδα TripAdvisor επελέγη ως πηγή συλλογής των κριτικών, δεδομένης της ευρείας χρήσης της από τους ταξιδιώτες για την αξιολόγηση των ξενοδοχειακών υπηρεσιών.

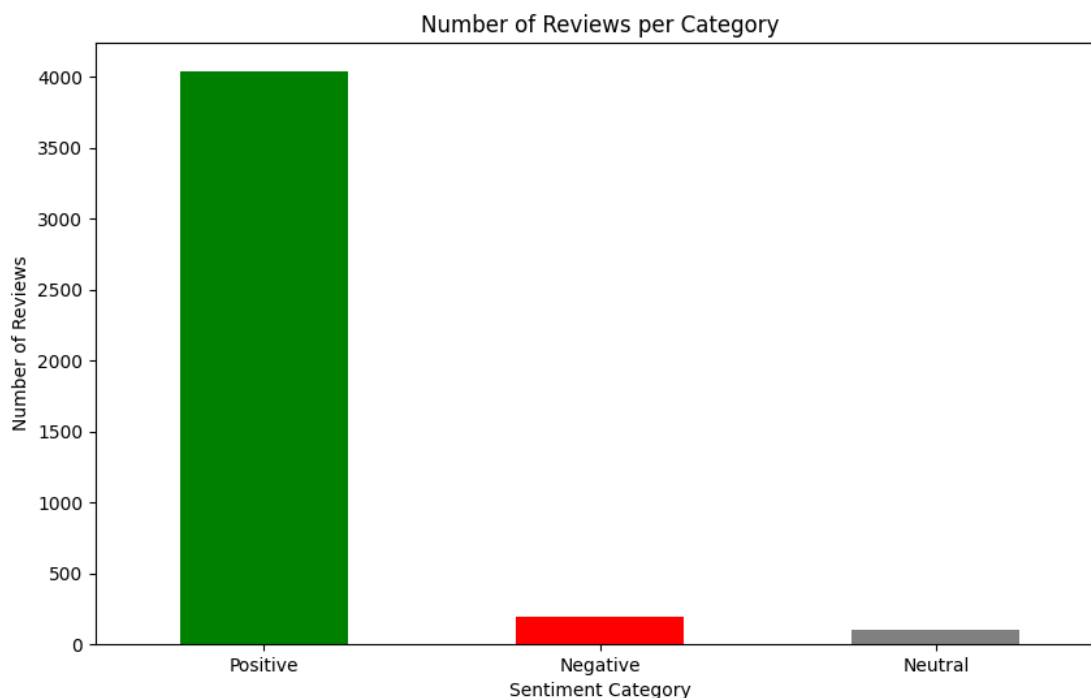
Λόγω της έλλειψης ενός δημόσιου API για την άμεση πρόσβαση στα δεδομένα της ιστοσελίδας, τροποποιήθηκε ένα πρόγραμμα αυτοματοποίησης (web scraping) στη γλώσσα προγραμματισμού Python (Van Rossum, 2009). Το εν λόγω πρόγραμμα χρησιμοποιεί τη βιβλιοθήκη Selenium, η οποία διευκολύνει τον έλεγχο διαφόρων λειτουργιών του web browser, όπως η διαχείριση πολλαπλών παραθύρων, η πλοήγηση σε διάφορες σελίδες, και η

εκτέλεση ενεργειών σε JavaScript, κρίσιμα για την επιτυχία του scraping σε περίπλοκες ιστοσελίδες.

Η διαδικασία scraping περιλαμβάνει το επαναληπτικό άνοιγμα μιας ιστοσελίδας από μια προκαθορισμένη λίστα, την ταυτοποίηση και επέκταση των κριτικών σε πλήρη μορφή, και τη συλλογή πληροφοριών όπως η βαθμολογία, ο τίτλος, το κείμενο, καθώς και η ημερομηνία δημοσίευσης της κριτικής. Το παραπάνω script χρησιμοποιήθηκε για την απόκτηση 4333 κριτικών από ξενοδοχεία στην Κέρκυρα, προσδίδοντας στην έρευνα μια εκτενή βάση δεδομένων για ανάλυση. Αναλυτικότερες πληροφορίες για την τεχνική υλοποίηση του web scraper παρουσιάζονται στην ενότητα [11.1](#) της παρούσας εργασίας.

6.2 Περιγραφή και ανάλυση των δεδομένων.

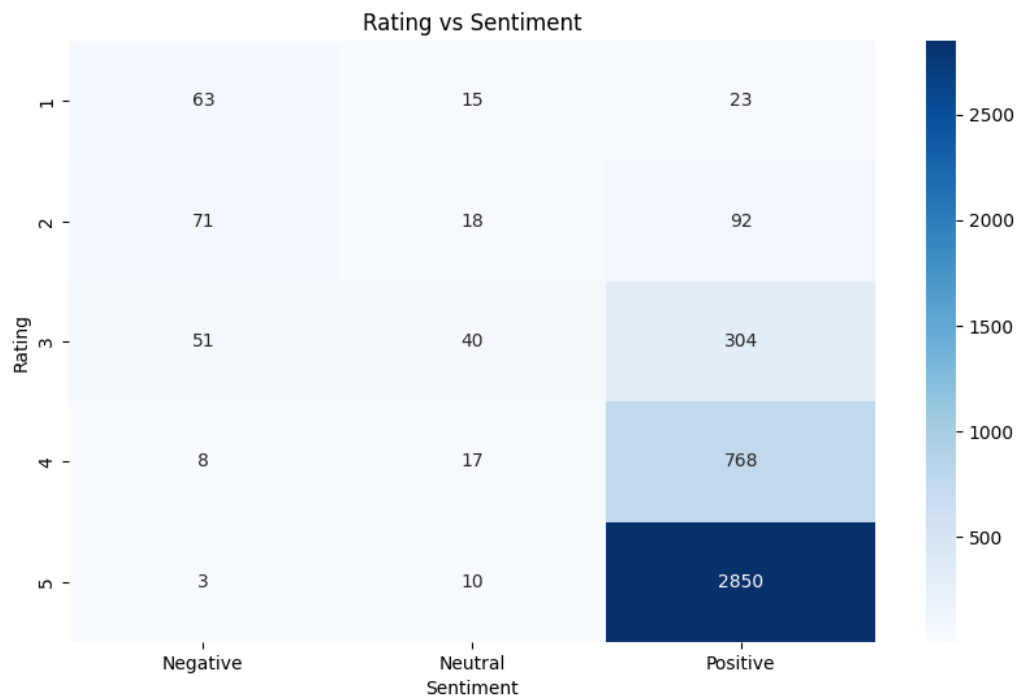
Το τελικό σετ δεδομένων περιλαμβάνει 4.333 κριτικές ξενοδοχείων. Κάθε κριτική αποτελείται από τρία βασικά στοιχεία: τον βαθμό (Rating), τον τίτλο της κριτικής όπως εμφανίζεται στο TripAdvisor (Review_Title), και το κυρίως κείμενο (Review_Text). Οι κριτικές αναλύθηκαν μεμονωμένα και κατηγοριοποιήθηκαν ως Θετικές (Positive), Αρνητικές (Negative), ή Ουδέτερες (Neutral), βάσει του περιεχομένου του κειμένου και του δοθέντα βαθμού. Η τελική κατανομή των κατηγοριών απεικονίζεται στην Εικόνα 4 Κατανομή κριτικών.



ΕΙΚΟΝΑ 4 ΚΑΤΑΝΟΜΗ ΚΡΙΤΙΚΩΝ

Όπως αναφέρεται στην Εικόνα 4, οι Θετικές κριτικές αποτελούν την κυρίαρχη κατηγορία στο σετ δεδομένων μας, **πράγμα που οδηγεί σε ανισορροπία μεταξύ των κλάσεων.**

Αυτή η ανισορροπία απαιτεί ειδική προσοχή στην επεξεργασία των δεδομένων, ώστε τα τελικά μοντέλα μας να μην εμφανίζουν προκατάληψη υπέρ της επικρατούσας κατηγορίας. Όπως διαπιστώνεται και από την Εικόνα 5, τα χαρακτηριστικά μας εμφανίζουν υψηλή συσχέτιση με την επικρατούσα κλάση.



ΕΙΚΟΝΑ 5 ΒΑΘΜΟΣ ΚΑΙ ΣΥΝΑΙΣΘΗΜΑ

6.3 Προ επεξεργασία δεδομένων

Στο πλαίσιο της ανάπτυξης των μοντέλων, είναι απαραίτητο να διασφαλίσουμε ότι τα δεδομένα μας προσαρμόζονται στην κατάλληλη μορφή για επεξεργασία. **Η διαδικασία προετοιμασίας και προ-επεξεργασίας των δεδομένων ακολουθεί τα καθορισμένα βήματα που περιγράφονται παρακάτω και βασίζονται στην δημοσίευση των (Rossini, 2021).** Το αρχικό σετ δεδομένων εμφανίζει σημαντική ανισορροπία στις κλάσεις του, γεγονός που μας οδηγεί στην υλοποίηση τεσσάρων διαφορετικών σετ πειραμάτων με σκοπό την εξέταση των επιπτώσεων αυτής της ανισορροπίας.

Το πρώτο σετ εκπαιδεύει τους αλγορίθμους με την αρχική μορφή του σετ δεδομένων, προβλέποντας πιθανώς μη θετικά αποτελέσματα λόγω της ανισορροπίας. Το δεύτερο σετ χρησιμοποιεί την τεχνική SMOTE για την εξισορρόπηση των κλάσεων και εκπαιδεύει τους ίδιους αλγορίθμους στα νέα δεδομένα. Το τρίτο σετ, σε αντίθεση με το δεύτερο, αντλεί πραγματικά δεδομένα από ένα διαθέσιμο σετ στο Kaggle με τίτλο "[Trip Advisor Hotel Reviews](#)", τα οποία παρέχουν παρόμοια χαρακτηριστικά με τα δικά μας. Τέλος το τέταρτο

σετ χρησιμοποιεί την τεχνική SMOTE, πλέον στο πιο ισορροπημένο σετ, με σκοπό την εξισορρόπηση με λιγότερα τεχνητά δεδομένα.

Οι τεχνικές προ-επεξεργασίας που εφαρμόζονται σε όλα τα σετ δεδομένων, τόσο στο εκπαιδευτικό όσο και στο δοκιμαστικό σετ (εκτός του SMOTE), περιλαμβάνουν τη μετατροπή των κριτικών σε αλφαριθμητικά με τη χρήση του φίλτρου **NominalToString** του Weka. Στη συνέχεια, εφαρμόζεται το φίλτρο **StringToWordVector**, το οποίο μετατρέπει τα κείμενα σε διανύσματα χαρακτηριστικών. Αυτό το φίλτρο υλοποιείται με δύο τεχνικές: Bag-of-Words και **TF-IDF**, όπου η δεύτερη τεχνική χρησιμοποιήθηκε για την παρούσα έρευνα. **Το φίλτρο επιτρέπει επίσης τη ρύθμιση παραμέτρων** όπως η αφαίρεση συχνών λέξεων, η μετατροπή σε πεζούς ή κεφαλαίους χαρακτήρες, και η εφαρμογή stemming για τη μείωση των λέξεων στη βασική τους μορφή.

Πιο αναλυτικά:

- **IDFTransform** (Inverse Document Frequency Transform) εάν είναι ενεργοποιημένο (True), υπολογίζει την αντίστροφη συχνότητας εγγράφου (IDF) για κάθε όρο. Το IDF είναι ένα μέτρο που βοηθά στην αποτίμηση της σημαντικότητας ενός όρου σε ένα συγκεκριμένο έγγραφο σε σύγκριση με ένα σύνολο εγγράφων. Όσο λιγότερο συχνά εμφανίζεται ένας όρος, τόσο μεγαλύτερη είναι η IDF τιμή του, κάτι που τον κάνει πιο σημαντικό.
- **TFTransform** (Term Frequency Transform) όταν ορίζεται σε True, υπολογίζει τη συχνότητα εμφάνισης κάθε όρου μέσα σε ένα έγγραφο. Αυτό βοηθά να κατανοηθεί πόσο συχνά εμφανίζεται ένας όρος σε σχέση με το συνολικό αριθμό όρων μέσα στο έγγραφο.
- **LowerCaseTokens** εάν είναι True, μετατρέπει όλα τα αλφαριθμητικά δεδομένα σε πεζά γράμματα, πράγμα που βοηθά στην αποφυγή διακρίσεων μεταξύ λέξεων που αρχίζουν από κεφαλαία ή πεζά γράμματα (π.χ., "Apple" και "apple" θεωρούνται ίδια).
- **MinTermFreq** ορίζει την ελάχιστη συχνότητα εμφάνισης ενός όρου για να συμπεριληφθεί στο τελικό σύνολο δεδομένων. Για παράδειγμα, αν οριστεί σε 10, τότε ένας όρος πρέπει να εμφανίζεται τουλάχιστον 10 φορές για να ληφθεί υπόψη.
- **NormalizeDocLengths** αν είναι True, κανονικοποιεί τα δεδομένα βάσει του μήκους του εγγράφου, προσπαθώντας να μειώσει τις διακυμάνσεις που οφείλονται στα μεγάλα ή μικρά μήκη των εγγράφων.
- **OutputWordCounts** αν οριστεί σε True, εξάγει την ακριβή συχνότητα των λέξεων αντί μόνο της δυαδικής παρουσίας ή απουσίας τους.
- **Stemmer** ορίζει τον αλγόριθμο βελτιστοποίησης των λέξεων που θα χρησιμοποιηθεί. Ο SnowballStemmer είναι ένας δημοφιλής βελτιστοποιητής που μειώνει τις λέξεις στη βασική τους μορφή (π.χ., "running" σε "run").
- **StopwordsHandler** αναφέρεται στη διαχείριση λέξεων που δεν έχουν σημαντική συμβολή στην ανάλυση του κειμένου (π.χ., "and", "the"). Το WordsFromFile σημαίνει ότι θα φορτώσει μια λίστα λέξεων-στοπ από ένα αρχείο. Όσον αφορά το αρχείο, επιλέχθηκε μια εκτενής λίστα λέξεων η οποία μπορεί να βρεθεί στο [Github](#).

- **Tokenizer** ορίζει τον τρόπο διαχωρισμού του κειμένου σε λέξεις ή φράσεις. Ο `WordTokenizer` (default) διαχωρίζει το κείμενο σε λέξεις βασισμένο στους χαρακτήρες κενού, κόμμα κτλ.

Κατά την προετοιμασία για την εκπαίδευση των αλγορίθμων, **οργανώσαμε τα δεδομένα σε δύο διακριτά σετ**: ένα για την εκπαίδευση των μοντέλων (train) και ένα για την επαλήθευση τους (test). Αυτός ο διαχωρισμός εξασφαλίζει την ακεραιότητα των αποτελεσμάτων μας, καθώς τα μοντέλα εκπαιδεύονται και δοκιμάζονται σε **ανεξάρτητα σετ δεδομένων**. Δεδομένου ότι το Weka δεν παρέχει εγγενή λειτουργία διαχωρισμού των δεδομένων, χρησιμοποιήθηκε η συνάρτηση `train_test_split` της βιβλιοθήκης `scikit-learn` της Python, η οποία επιτρέπει **τον τυχαίο διαχωρισμό των δεδομένων σε δύο σετ, με το σετ δοκιμών να αποτελεί το 20% του συνολικού όγκου δεδομένων**.

Στο δεύτερο και τέταρτο σετ πειραμάτων θα εφαρμοστεί επίσης η **τεχνική SMOTE** για την εξισορρόπηση των κλάσεων στο σετ εκπαίδευσης. Ο αλγόριθμος SMOTE χρησιμοποιείται για να αντιμετωπίσει το πρόβλημα ανισορροπίας των κλάσεων, δημιουργώντας συνθετικά δείγματα για τις υποεκπροσωπούμενες κλάσεις. Οι παράμετροι που χρειάζονται για την εφαρμογή του αλγορίθμου προσδιορίζονται με βάση τον υπολογισμό της αναλογίας των συνθετικών δεδομένων που απαιτούνται για κάθε κλάση, ως εξής (Εξίσωση 1):

$$N_{synthetic} = N_{majority} - N_{minority}$$

$$oversampling_ratio = \frac{N_{synthetic}}{N_{minority}} \times 100$$

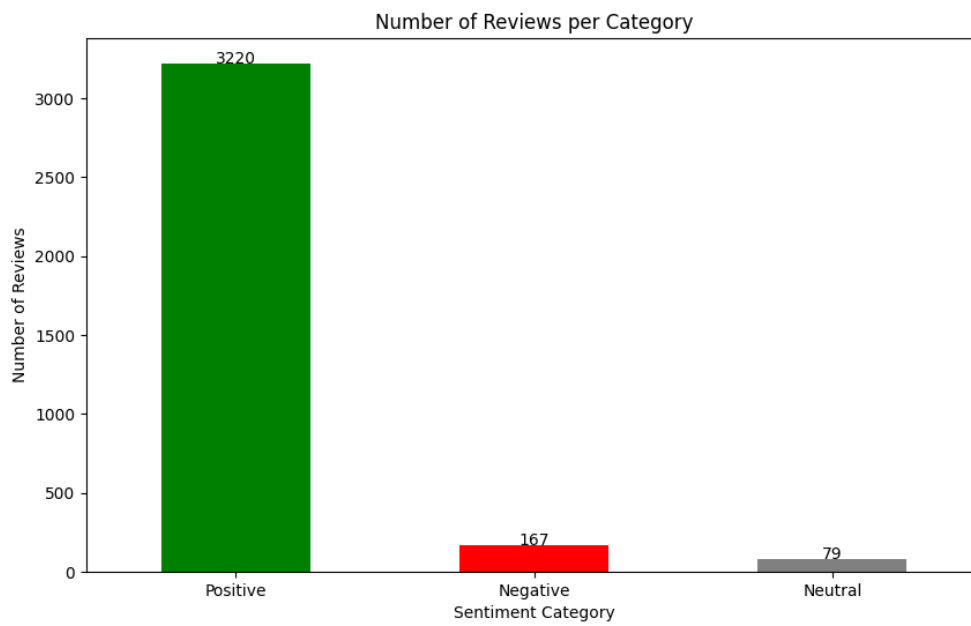
ΕΞΙΣΩΣΗ 1 ΥΠΟΛΟΓΙΣΜΟΣ ΠΟΣΟΣΤΟΥ OVERSAMPLING

Στα πλαίσια των πειραμάτων, εκτός από τα προδιαγραφόμενα attributes, εφαρμόζονται και μέθοδοι επιλογής χαρακτηριστικών όπως η *GainRatioAttributeEval* και η *InfoGainAttributeEval*, με σκοπό τη **μείωση των χαρακτηριστικών στα 100 καλύτερα**.

Η *InfoGainAttributeEval* αξιολογεί τα χαρακτηριστικά με βάση τη μείωση εντροπίας που επιφέρουν, ενώ η *GainRatioAttributeEval* λαμβάνει υπόψη τη διακύμανση στις τιμές των χαρακτηριστικών. Η αξιολόγηση αυτή χρησιμοποιείται για την αποφυγή υπερεκτίμησης χαρακτηριστικών με πολλές διακριτές τιμές. Ο *Ranker* χρησιμοποιείται για την ταξινόμηση και επιλογή των 100 κορυφαίων χαρακτηριστικών βάσει του κέρδους πληροφορίας.

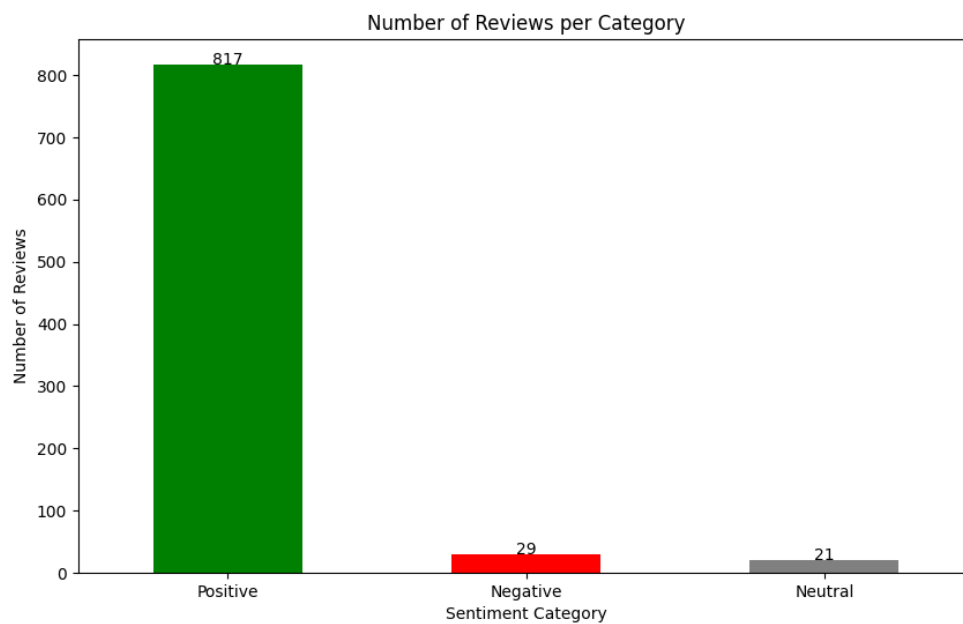
Μετά τον διαχωρισμό σε train και test τα δυο σετ έχουν κατανομή σε Positive, Negative, Neutral την εξής:

Train: Positive: 3220, Negative: 167, Neutral: 79



EIKONA 6

Test: Positive: 817, Negative: 29, Neutral: 21

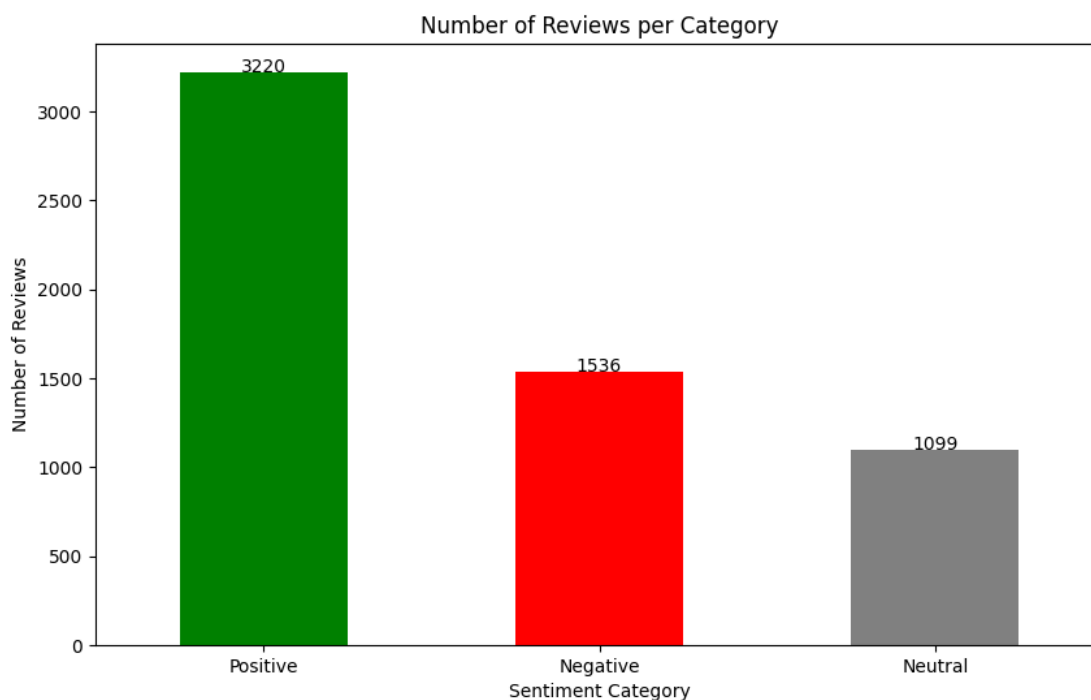


EIKONA 7

Για την εξισορρόπηση των κλάσεων στο φίλτρο SMOTE του Weka, έχουμε ορίσει ως παραμέτρους τα ποσοστά 3950% για την κλάση Neutral και 1816% για την κλάση Negative. Ακολουθώντας την εφαρμογή του φίλτρου, η κατανομή των δεδομένων στο σετ εκπαίδευσης είναι ως εξής: Positive: 3220, Negative: 3340, Neutral: 3239.

Στο τρίτο σετ πειραμάτων, **αντί για τη δημιουργία συνθετικών δεδομένων, αξιοποιούμε πραγματικά δεδομένα από ένα δεύτερο σετ**, τα οποία έχουν επισημανθεί με τον ίδιο τρόπο όπως και το αρχικό μας σετ. Η ένωση των δύο σετ οδηγεί στην τελική κατανομή των δεδομένων: Positive: 3220, Negative: 1536, Neutral: 1099. Μετά το διαχωρισμό σε σετ εκπαίδευσης και δοκιμής, η κατανομή του σετ εκπαίδευσης παρουσιάζει ανισορροπία, αλλά σε σημαντικά μικρότερο βαθμό σε σύγκριση με το πρώτο σετ πειραμάτων: Positive: 3231, Negative: 1235, Neutral: 911.

Τέλος, στο τέταρτο σετ πειραμάτων **εφαρμόζουμε και πάλι την τεχνική SMOTE στο σετ που δημιουργήθηκε στο τρίτο πείραμα, με στόχο τη δημιουργία λιγότερων συνθετικών δεδομένων**. Θα ορίσουμε στο φίλτρο SMOTE τα ποσοστά 254% για την κλάση Neutral και 161% για την κλάση Negative. Μετά την εφαρμογή του φίλτρου, η κατανομή στο σετ εκπαίδευσης είναι: Positive: 3231, Negative: 3211, Neutral: 3188.



ΕΙΚΟΝΑ 8

6.4 Εκπαίδευση των Μοντέλων

Στην παρούσα ενότητα αναφερόμαστε στα αποτελέσματα της εκπαίδευσης των επιλεγμένων μοντέλων, συζητούμε τα χαρακτηριστικά που κρίθηκαν κατάλληλα για την υλοποίηση κάθε μοντέλου, και παρουσιάζουμε αναλυτικά τα αποτελέσματα ανά σετ πειραμάτων. Τα μοντέλα που εκπαιδεύτηκαν περιλαμβάνουν τα *Random Forest*, *J48*, *Naive Bayes* και *IBK (NN = 3)*, τα οποία έχουν περιγραφεί προηγουμένως.

Η διαδικασία της εκπαίδευσης για κάθε μοντέλο πραγματοποιήθηκε στο σετ εκπαίδευσης δεδομένων, ενώ η απόδοσή τους αξιολογήθηκε μέσω της μεθόδου 10-fold cross-validation και στη συνέχεια επιβεβαιώθηκε στο αντίστοιχο σετ δοκιμής δεδομένων.

Η μέθοδος 10-fold cross-validation, η οποία αποτελεί στάνταρ τεχνική στη μηχανική μάθηση για την εκτίμηση της απόδοσης ενός μοντέλου, συνίσταται στον εξής αλγόριθμο:

1. **Διάρθρωση των Δεδομένων:** Το συνολικό σετ δεδομένων χωρίζεται τυχαία σε 10 ισομεγέθη υποσύνολα.
2. **Εκπαίδευση και Αξιολόγηση:** Κάθε μοντέλο εκπαιδεύεται σε 9 από τα 10 υποσύνολα, ενώ το υπολειπόμενο υποσύνολο χρησιμοποιείται για αξιολόγηση. Αυτή η διαδικασία επαναλαμβάνεται 10 φορές, με κάθε υποσύνολο να χρησιμοποιείται ως σετ δοκιμής μια φορά.
3. **Υπολογισμός Μέσου Αποτελέσματος:** Το μέσο αποτέλεσμα των δεκαπλών αξιολογήσεων χρησιμοποιείται ως δείκτης της συνολικής απόδοσης του μοντέλου.

6.5 Σετ πειραμάτων 1

Στο **πρώτο σετ** πειραμάτων, χρησιμοποιήθηκαν τέσσερις αλγόριθμοι: *Random Forest*, *J48*, *Naive Bayes* και *IBK (NN = 3)* χωρίς καμία επιλογή χαρακτηριστικών. **Τα αποτελέσματα έδειξαν ότι ο IBK (NN = 3) είχε την υψηλότερη ευστοχία με 93.59%**, ενώ ο αλγόριθμος *Naive Bayes* εμφάνισε τη χαμηλότερη απόδοση με ακρίβεια 60.55%. Ο *Random Forest* είχε καλή απόδοση με ακρίβεια 93.04% και ο *J48* με 93.21%. Η περιοχή ROC ήταν υψηλή για τον *Random Forest* (95.2%) και τον *J48* (75.7%), αλλά αρκετά χαμηλή για τον *Naive Bayes* (80.7%) και τον *IBK* (81.6%). **Συνολικά, το πρώτο σετ έδειξε ότι οι δένδρο-βασισμένοι αλγόριθμοι (*Random Forest* και *J48*) είχαν σταθερά καλή απόδοση, ενώ ο *Naive Bayes* δεν έχει ικανοποιητικά αποτελέσματα χωρίς επιλογή χαρακτηριστικών.** Τα συνολικά αποτελέσματα φαίνονται στους Πίνακας 1 Πίνακας 2 Πίνακας 3.

	No Feature Selection				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	93.5%	93.0%	89.8%	95.2%	93.04%
J48	92.0%	93.2%	92.6%	75.7%	93.21%
NaiveBayes	93.1%	60.6%	72.3%	80.7%	60.55%
IBK (NN = 3)	91.1%	93.6%	92.2%	81.6%	93.59%

classified as	Positive	Negative	Neutral
Positive	3220	0	0
Negative	164	3	0
Neutral	77	0	2

Random Forest

classified as	Positive	Negative	Neutral
Positive	3151	48	21
Negative	82	72	13
Neutral	54	17	8

J48

classified as	Positive	Negative	Neutral
Positive	1981	88	1151
Negative	27	67	73
Neutral	12	16	51

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	3180	33	7
Negative	98	64	5
Neutral	67	12	0

IBK (NN=3)

ΠΙΝΑΚΑΣ 1 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 1

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	91.4%	93.0%	92.1%	80.4%	92.98%
J48	91.4%	93.0%	92.1%	80.4%	92.98%
NaiveBayes	94.4%	71.4%	79.9%	93.4%	71.35%
IBK (NN = 3)	89.9%	92.8%	90.3%	74.0%	92.75%

classified as	Positive	Negative	Neutral
Positive	3151	59	10
Negative	88	69	10
Neutral	52	24	3

Random Forest

classified as	Positive	Negative	Neutral
Positive	3151	59	10
Negative	88	69	10
Neutral	52	24	3

J48

classified as	Positive	Negative	Neutral
Positive	2301	184	735
Negative	4	112	51
Neutral	5	14	60

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	3200	7	13
Negative	146	14	6
Neutral	76	2	1

IBK (NN=3)

ΠΙΝΑΚΑΣ 2 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 1 (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	92.7	94.5%	93.0%	95.1%	94.51%
J48	92.4%	93.6%	92.9%	77.8%	93.59%
NaiveBayes	94.3%	72.3%	80.8%	92.4%	72.30%
IBK (NN = 3)	90.6%	92.8%	90.7%	75.0%	92.81%

classified as	Positive	Negative	Neutral
Positive	3210	9	1
Negative	100	64	3
Neutral	64	13	2

Random Forest

classified as	Positive	Negative	Neutral
Positive	3156	51	13
Negative	74	81	12
Neutral	55	17	7

J48

classified as	Positive	Negative	Neutral
Positive	2353	138	729
Negative	6	100	61
Neutral	9	17	53

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	3193	7	20
Negative	142	17	8
Neutral	69	3	7

IBK (NN=3)

ΠΙΝΑΚΑΣ 3 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 1 (INFOGAINATTRIBUTEVAL)

6.6 Σετ πειραμάτων 2

Στο **δεύτερο σετ** πειραμάτων εφαρμόστηκε η τεχνική SMOTE για την εξισορρόπηση των κλάσεων στο σετ εκπαίδευσης, πάλι χωρίς επιλογή χαρακτηριστικών. Τα αποτελέσματα έδειξαν σημαντική βελτίωση της απόδοσης των μοντέλων σε σχέση με το πρώτο σετ. Ο Random Forest πέτυχε εντυπωσιακή ακρίβεια 99.18% και περιοχή ROC

100%. Ο J48 και πάλι είχε καλή απόδοση με 96.59% ακρίβεια. Ο Naive Bayes και ο IBK παρουσίασαν καλύτερα αποτελέσματα από το πρώτο σετ, με ακρίβειες 71.42% και 71.69% αντίστοιχα. **Η εφαρμογή της τεχνικής SMOTE φάνηκε να βοηθά στην εξισορρόπηση των δεδομένων και στην αύξηση της ακρίβειας των μοντέλων.** Τα συνολικά αποτελέσματα φαίνονται στους Πίνακες 4, Πίνακας 5, Πίνακας 6.

	No Feature Selection				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	99.2%	99.2%	99.2%	100.0%	99.18%
J48	96.6%	96.6%	96.6%	97.2%	96.59%
NaiveBayes	86.7%	80.9%	81.2%	93.1%	71.42%
IBK (NN = 3)	82.2%	71.7%	64.3%	83.9%	71.69%

classified as	Positive	Negative	Neutral
Positive	3172	48	0
Negative	18	3322	0
Neutral	14	0	3225

Random Forest

classified as	Positive	Negative	Neutral
Positive	3052	74	94
Negative	54	3260	26
Neutral	50	36	3153

J48

classified as	Positive	Negative	Neutral
Positive	1958	28	1233
Negative	153	2899	231
Neutral	7	0	2142

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	446	446	2328
Negative	0	3340	0
Neutral	0	0	3239

IBK (NN=3)

ΠΙΝΑΚΑΣ 4 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 2

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	90.7%	88.8%	88.9%	97.5%	88.79%
J48	85.1%	85.1%	85.1%	94.3%	85.06%
NaiveBayes	83.4%	74.9%	74.4%	90.2%	74.90%
IBK (NN = 3)	89.5%	87.2%	87.3%	96.2%	87.22%

classified as	Positive	Negative	Neutral
Positive	2899	78	243
Negative	18	2456	699
Neutral	10	22	3128

Random Forest

classified as	Positive	Negative	Neutral
Positive	2894	140	186
Negative	22	2735	416
Neutral	279	384	2497

J48

classified as	Positive	Negative	Neutral
Positive	1592	224	1404
Negative	8	2423	742
Neutral	0	19	3141

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	2817	134	269
Negative	16	2378	779
Neutral	8	14	3138

IBK (NN=3)

ΠΙΝΑΚΑΣ 5 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 2 (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	98.5%	98.5%	98.5%	99.9%	98.49%
J48	95.0%	95.0%	95.0%	96.3%	94.98%
NaiveBayes	82.7%	76.2%	76.4%	93.7%	76.21%
IBK (NN = 3)	95.4%	95.0%	94.9%	97.9%	94.98%

classified as	Positive	Negative	Neutral
Positive	3114	89	17
Negative	17	3154	2
Neutral	16	3	3141

Random Forest

classified as	Positive	Negative	Neutral
Positive	2982	95	143
Negative	47	3053	73
Neutral	75	46	3039

J48

classified as	Positive	Negative	Neutral
Positive	2219	128	873
Negative	95	1974	1104
Neutral	54	18	3088

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	2745	164	311
Negative	3	3169	1
Neutral	0	0	3160

IBK (NN=3)

ΠΙΝΑΚΑΣ 6 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 2 (INFOGAINATTRIBUTEVAL)

6.7 Σετ πειραμάτων 3

Στο **τρίτο σετ** πειραμάτων, αντί για τη δημιουργία συνθετικών δεδομένων, χρησιμοποιήθηκαν επιπλέον πραγματικά δεδομένα για την εξισορρόπηση των κλάσεων. **Τα αποτελέσματα έδειξαν ότι η ακρίβεια των μοντέλων μειώθηκε συγκριτικά με το δεύτερο σετ.** Ο Random Forest είχε ακρίβεια 81.96%, ενώ ο J48 έφτασε το 77.44%. Ο Naive Bayes και ο IBK είχαν παρόμοια απόδοση με 78.68% και 64.47% ακρίβεια αντίστοιχα. Η περιοχή ROC παρέμεινε υψηλή για όλους τους αλγόριθμους, ιδίως για τον Random Forest (95.9%). **Η χρήση πραγματικών δεδομένων φαίνεται να προσέφερε καλύτερη ισορροπία στις κλάσεις, αλλά όχι την ίδια απόδοση με τη χρήση συνθετικών δεδομένων.** Τα συνολικά αποτελέσματα φαίνονται στους Πίνακες 7, Πίνακας 8, Πίνακας 9.

	No Feature Selection				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	81.0%	82.0%	79.0%	95.9%	81.96%
J48	76.9%	77.4%	77.2%	83.6%	77.44%
NaiveBayes	78.6%	78.7%	78.5%	90.8%	78.68%
IBK (NN = 3)	77.8%	64.5%	67.5%	85.9%	64.47%

classified as	Positive	Negative	Neutral
Positive	3109	18	4
Negative	166	992	77
Neutral	219	486	206

Random Forest

classified as	Positive	Negative	Neutral
Positive	3045	88	98
Negative	96	748	391
Neutral	153	387	371

J48

classified as	Positive	Negative	Neutral
Positive	3087	75	69
Negative	91	679	474
Neutral	123	314	474

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	2359	27	845
Negative	61	438	736
Neutral	80	161	670

IBK (NN=3)

ΠΙΝΑΚΑΣ 7 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 3

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	78.5%	79.9%	79.0%	94.1%	79.85%
J48	78.4%	80.4%	78.5%	91.2%	80.37%
NaiveBayes	69.8%	72.0%	70.4%	81.5%	71.95%
IBK (NN = 3)	77.9%	79.4%	78.5%	91.4%	79.44%

classified as	Positive	Negative	Neutral
Positive	3105	49	77
Negative	94	860	281
Neutral	157	425	329

Random Forest

classified as	Positive	Negative	Neutral
Positive	3084	55	92
Negative	86	1008	141
Neutral	195	486	230

J48

classified as	Positive	Negative	Neutral
Positive	2982	73	176
Negative	342	594	299
Neutral	358	260	293

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	3121	42	68
Negative	122	815	298
Neutral	172	403	336

IBK (NN=3)

ΠΙΝΑΚΑΣ 8 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 3 (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	82.0%	83.2%	82.1%	95.7%	83.22%
J48	79.8%	80.2%	80.0%	89.1%	80.24%
NaiveBayes	78.1%	78.8%	78.3%	91.7%	78.78%
IBK (NN = 3)	78.2%	75.4%	76.6%	89.4%	75.41%

classified as	Positive	Negative	Neutral
Positive	3179	34	18
Negative	90	946	199
Neutral	137	424	350

Random Forest

classified as	Positive	Negative	Neutral
Positive	3075	75	81
Negative	89	818	328
Neutral	118	371	422

J48

classified as	Positive	Negative	Neutral
Positive	3043	56	132
Negative	100	863	272
Neutral	104	477	330

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	2836	103	292
Negative	34	803	398
Neutral	74	421	416

IBK (NN=3)

ΠΙΝΑΚΑΣ 9 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 3 (INFOGAINATTRIBUTEVAL)

6.8 Σετ πειραμάτων 4

Στο **τέταρτο σετ** πειραμάτων, συνδυάστηκε η χρήση πραγματικών δεδομένων και η τεχνική SMOTE για την εξισορρόπηση των κλάσεων. Ο Random Forest είχε ακρίβεια 95.80%, ο J48 82.55%, ο Naive Bayes 75.79%, και ο IBK (NN = 3) 68.43%. **Η περιοχή ROC ήταν εξαιρετικά υψηλή για τον Random Forest (99.4%), γεγονός που δείχνει τη σημαντική βελτίωση της ικανότητας διάκρισης των κλάσεων. Η συνδυαστική προσέγγιση φαίνεται να προσφέρει την καλύτερη ισορροπία μεταξύ ακρίβειας και εξισορρόπησης δεδομένων, βελτιώνοντας τη συνολική απόδοση των μοντέλων.** Τα συνολικά αποτελέσματα φαίνονται στους Πίνακας 10, Πίνακας 11, Πίνακας 12.

	No Feature Selection				
Αλγοριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	95.8%	95.8%	95.8%	99.4%	95.80%
J48	82.5%	82.6%	82.5%	87.8%	82.55%
NaiveBayes	75.1%	75.8%	75.2%	88.2%	75.79%
IBK (NN = 3)	83.2%	68.4%	60.2%	81.2%	68.43%

classified as	Positive	Negative	Neutral
Positive	3180	35	16
Negative	71	3022	118
Neutral	86	78	3024

Random Forest

classified as	Positive	Negative	Neutral
Positive	2991	86	154
Negative	110	2532	569
Neutral	194	567	2427

J48

classified as	Positive	Negative	Neutral
Positive	3152	37	42
Negative	240	2177	794
Neutral	345	873	1970

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	262	83	2886
Negative	0	3140	71
Neutral	0	0	3188

IBK (NN=3)

ΠΙΝΑΚΑΣ 10 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 4

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	78.5%	78.4%	78.5%	90.7%	78.44%
J48	77.1%	77.4%	77.2%	88.0%	77.37%
NaiveBayes	70.2%	70.2%	69.7%	85.8%	70.19%
IBK (NN = 3)	76.6%	75.4%	75.9%	87.2%	75.41%

classified as	Positive	Negative	Neutral
Positive	3041	46	144
Negative	54	2304	853
Neutral	113	866	2209

Random Forest

classified as	Positive	Negative	Neutral
Positive	3035	61	135
Negative	76	2318	817
Neutral	215	875	2098

J48

classified as	Positive	Negative	Neutral
Positive	2888	66	277
Negative	193	2341	677
Neutral	189	1468	1531

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	2762	99	370
Negative	34	2278	899
Neutral	73	893	2222

IBK (NN=3)

ΠΙΝΑΚΑΣ 11 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 4 (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F-Measure	ROC Area	Accuracy
Random Forest	87.6%	87.7%	87.6%	96.8%	87.71%
J48	78.2%	78.2%	78.2%	85.9%	78.21%
NaiveBayes	73.2%	73.4%	72.6%	87.7%	73.36%
IBK (NN = 3)	88.8%	87.7%	87.9%	95.4%	87.74%

classified as	Positive	Negative	Neutral
Positive	3144	32	55
Negative	60	2737	414
Neutral	118	504	2566

Random Forest

classified as	Positive	Negative	Neutral
Positive	2994	87	150
Negative	80	2309	822
Neutral	168	791	2229

J48

classified as	Positive	Negative	Neutral
Positive	3037	61	133
Negative	194	2418	599
Neutral	262	1316	1610

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	2676	108	447
Negative	20	2857	334
Neutral	18	253	2917

IBK (NN=3)

ΠΙΝΑΚΑΣ 12 ΑΠΟΤΕΛΕΣΜΑΤΑ TRAINING ΣΕΤ 4 (INFOGAINATTRIBUTEVAL)

Τέλος η χρήση των τεχνικών GainRatioAttributeEval και InfoGainAttributeEval για την επιλογή των καλύτερων 100 χαρακτηριστικών βελτίωσε τις επιδόσεις των μοντέλων σε όλες τις περιπτώσεις. Ο Random Forest εμφάνισε βελτιωμένα αποτελέσματα με ακρίβεια 94.51% και 98.49% αντίστοιχα στις δύο μεθόδους. Ο J48, ο Naïve Bayes και ο IBK επίσης παρουσίασαν καλύτερες επιδόσεις με την επιλογή χαρακτηριστικών. Η περιοχή ROC και η ακρίβεια αυξήθηκαν σημαντικά, υποδεικνύοντας ότι η επιλογή χαρακτηριστικών μπορεί να βελτιώσει την απόδοση των μοντέλων και να μειώσει την πολυπλοκότητα των δεδομένων.

7. Αποτελέσματα

Τα αποτελέσματα των μοντέλων θα αξιολογηθούν και συγκριθούν σε κάθε σετ πειραμάτων βάσει των εξής μετρικών απόδοσης:

- **Accuracy (Ακρίβεια):** Μετρά το ποσοστό των σωστών προβλέψεων από το σύνολο των παρατηρήσεων.
- **Precision (Ευστοχία):** Καταγράφει το ποσοστό των πραγματικών θετικών προβλέψεων από το σύνολο των προβλεπόμενων θετικών.
- **Recall (Ανάκληση):** Μετρά το ποσοστό των πραγματικών θετικών προβλέψεων από το σύνολο των πραγματικών θετικών δεδομένων.
- **F1-Score:** Αποτελεί έναν συνδυαστικό δείκτη της ακρίβειας και της ανάκλησης, υπολογιζόμενος ως ο αρμονικός μέσος των δύο.
- **ROC Curve (ROC-AUC):** Μετρά την επίδοση ενός ταξινομητή στο να διακρίνει μεταξύ δύο κλάσεων, με βάση το εμβαδόν κάτω από την καμπύλη ROC (Receiver Operating Characteristic).

Η Ακρίβεια (Accuracy) υπολογίζει την αναλογία των σωστών προβλέψεων στο συνολικό αριθμό των παρατηρήσεων, ενώ η Ακρίβεια (Precision) καταγράφει το ποσοστό των σωστών θετικών προβλέψεων μεταξύ όλων των προβλεπόμενων θετικών. Η Ανάκληση (Recall) εστιάζει στην ικανότητα του μοντέλου να εντοπίζει όλες τις πραγματικές θετικές περιπτώσεις. Το F1-Score παρέχει μια ισορροπημένη μέτρηση των δύο προηγούμενων μετρήσεων, ιδανική για περιπτώσεις ανισορροπημένων δεδομένων. Το ROC-AUC δίνει μια γενική εικόνα της ικανότητας του μοντέλου να διαχωρίζει τις κλάσεις, όπου τιμές κοντά στο 1 υποδηλώνουν υψηλή αποτελεσματικότητα. Οι υπολογισμοί των παραπάνω μετρικών γίνεται με τις παρακάτω εξισώσεις.

$$Accuracy = \frac{Correctly\ Classified\ Instances}{Total\ Instances}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

7.1 Σετ πειραμάτων 1

Όπως αναφέρθηκε προηγουμένως, στο πρώτο σετ πειραμάτων χρησιμοποιήσαμε τα δεδομένα που συλλέξαμε από το TripAdvisor χωρίς επεξεργασία, για την εκπαίδευση και επαλήθευση των αλγορίθμων μας. Τα αποτελέσματα φαίνονται στους Πίνακες 13, Πίνακας

14, Πίνακας 15. Χωρίς χρήση feature selection, οι αλγόριθμοι Random Forest και IBK (NN = 3) είχαν την καλύτερη απόδοση με ακρίβεια 94.34% και 95.04% αντίστοιχα. Ωστόσο, ο αλγόριθμος Naive Bayes παρουσίασε χαμηλότερη απόδοση με ακρίβεια 58.36%, λόγω της χαμηλής ανάκλησης (Recall) που ήταν 58.4%. **Η χρήση feature selection βελτίωσε και στις δυο περιπτώσεις (GainRatioAttributeEval, InfoGainAttributeEval) την συνολική εικόνα όλων των αλγορίθμων εκτός του IBK ο οποίος παρουσίασε μείωση στην απόδοση του με την χρήση λιγότερων χαρακτηριστικών.**

	No Feature Selection				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	?	94.3%	?	95.5%	94.34%
J48	92.4%	93.3%	92.8%	68.8%	93.31%
Naive Bayes	93.0%	58.4%	70.7%	76.5%	58.36%
IBK (NN = 3)	93.6%	95.0%	94.0%	85.2%	95.04%

classified as	Positive	Negative	Neutral
Positive	817	0	0
Negative	28	1	0
Neutral	21	0	0

Random Forest

classified as	Positive	Negative	Neutral
Positive	797	15	5
Negative	16	10	3
Neutral	14	5	2

J48

classified as	Positive	Negative	Neutral
Positive	487	29	301
Negative	6	7	16
Neutral	4	5	12

NaïveBayes

classified as	Positive	Negative	Neutral
Positive	809	8	0
Negative	14	14	1
Neutral	16	4	1

IBK (NN = 3)

ΠΙΝΑΚΑΣ 13 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 1 ΧΩΡΙΣ FEATURE SELECTION

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	92.7%	93.7%	93.2%	89.2%	93.65%
J48	92.7%	93.7%	93.2%	89.2%	93.65%
Naive Bayes	94.0%	66.9%	77.3%	89.2%	66.89%
IBK (NN = 3)	89.6%	93.9%	91.16%	74.1%	93.88%

classified as	Positive	Negative	Neutral
Positive	797	11	9
Negative	14	14	1
Neutral	15	5	1

Random Forest

classified as	Positive	Negative	Neutral
Positive	797	11	9
Negative	14	14	1
Neutral	15	5	1

J48

classified as	Positive	Negative	Neutral
Positive	560	49	208
Negative	1	10	18
Neutral	4	7	10

NaiveBayes

classified as	Positive	Negative	Neutral
Positive	813	1	3
Negative	26	0	3
Neutral	19	1	1

IBK (NN = 3)

ΠΙΝΑΚΑΣ 14 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 1 ΜΕ FEATURE SELECTION (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	?	94.7%	?	96.6%	94.69%
J48	92.9%	93.5%	93.1%	80.0%	93.54%
Naive Bayes	94.2%	72.9%	81.4%	88.5%	72.89%
IBK (NN = 3)	92.5%	94.3%	92.6%	82.3%	94.34%

classified as	Positive	Negative	Neutral
Positive	811	6	0
Negative	19	10	0
Neutral	18	3	0

Random Forest

classified as	Positive	Negative	Neutral
Positive	609	30	178
Negative	1	12	16
Neutral	6	4	11

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	795	18	4
Negative	13	15	1
Neutral	13	7	1

J48

classified as	Positive	Negative	Neutral
Positive	810	3	4
Negative	24	4	1
Neutral	16	1	4

IBK (NN = 3)

ΠΙΝΑΚΑΣ 15 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 1 ΜΕ FEATURE SELECTION (INFOGAINATTRIBUTEVAL)

7.2 Σετ πειραμάτων 2

Στο δεύτερο σετ πειραμάτων εφαρμόστηκε η τεχνική SMOTE (Synthetic Minority Over-sampling Technique) για την αντιμετώπιση της ανισορροπίας στα δεδομένα. Χωρίς feature selection, ο αλγόριθμος Random Forest είχε και πάλι την μεγαλύτερη ακρίβεια 94.57%, ενώ ο IBK (NN = 3) παρουσίασε χαμηλή ανάκληση (Recall) 15.6%, επηρεάζοντας αρνητικά το F1 Score του που ήταν 23.0%. Αυτό δείχνει ότι η τεχνική SMOTE βελτιώνει την απόδοση σε

ορισμένους αλγορίθμους αλλά όχι σε όλους. Η χρήση **feature selection** με **GainRatioAttributeEval** και **InfoGainAttributeEval** βελτίωσε σημαντικά την απόδοση, ειδικά για τον αλγόριθμο **J48** που έφτασε σε ακρίβεια **90.88%** και **91.3%** αντίστοιχα, βελτιώνοντας την ισορροπία ανάμεσα στην ακρίβεια και την ανάκληση. Τα αποτελέσματα φαίνονται στους Πίνακες 16,17 και 18.

	No Feature Selection				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	?	94.6%	?	92.7%	94.57%
J48	92.9%	90.9%	91.8%	63.2%	90.88%
Naive Bayes	93.1%	57.5%	69.9%	71.9%	57.43%
IBK (NN = 3)	94.6%	15.6%	23.0%	62.2%	15.57%

classified as	Positive	Negative	Neutral
Positive	810	7	0
Negative	19	10	0
Neutral	20	1	0

Random Forest

classified as	Positive	Negative	Neutral
Positive	773	22	22
Negative	12	11	6
Neutral	9	8	4

J48

classified as	Positive	Negative	Neutral
Positive	477	10	329
Negative	9	5	15
Neutral	2	3	16

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	110	121	586
Negative	0	16	13
Neutral	0	12	9

IBK (NN=3)

ΠΙΝΑΚΑΣ 16 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 2 ΧΩΡΙΣ FEATURE SELECTION

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	93.5%	87.2%	90.0%	80.4%	87.19%
J48	95.0%	88.5%	91.3%	91.9%	88.46%
Naive Bayes	93.9%	48.9%	62.7%	79.9%	48.90%
IBK (NN = 3)	94.2%	86.0%	89.6%	86.9%	86.04%

classified as	Positive	Negative	Neutral
Positive	740	21	56
Negative	5	11	13
Neutral	8	6	7

Random Forest

classified as	Positive	Negative	Neutral
Positive	727	30	60
Negative	2	13	14
Neutral	6	9	6

J48

classified as	Positive	Negative	Neutral
Positive	401	55	361
Negative	2	12	15
Neutral	2	8	11

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	744	27	46
Negative	1	13	15
Neutral	3	8	10

IBK (NN=3)

ΠΙΝΑΚΑΣ 17 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 2 (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	92.4%	88.2%	90.1%	71.4%	88.23%
J48	94.2%	90.3%	92.1%	75.1%	90.31%
Naive Bayes	93.9%	66.0%	76.5%	89.0%	65.97%
IBK (NN = 3)	94.4%	83.3%	87.9%	90.9%	83.27%

classified as	Positive	Negative	Neutral
Positive	789	20	8
Negative	7	20	2
Neutral	11	7	3

Random Forest

classified as	Positive	Negative	Neutral
Positive	760	20	37
Negative	4	18	7
Neutral	9	7	5

J48

classified as	Positive	Negative	Neutral
Positive	550	28	229
Negative	3	9	17
Neutral	3	4	13

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	696	38	83
Negative	3	18	8
Neutral	4	9	8

IBK (NN=3)

ΠΙΝΑΚΑΣ 18 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 2 (INFOGAINATTRIBUTEVAL)

7.3 Σετ πειραμάτων 3

Στο τρίτο σετ πειραμάτων συνδυάστηκαν το αρχικό σετ δεδομένων με ένα δεύτερο σετ το οποίο επισημάνθηκε με τον ίδιο τρόπο με το αρχικό μας. Χωρίς feature selection, ο Random Forest είχε ακρίβεια 83.42% και ο J48 76.65%, ενώ η χρήση feature selection με GainRatioAttributeEval και InfoGainAttributeEval βελτίωσε την απόδοση του Random

Forest σε 80.52% και 84.16% αντίστοιχα. Ο Naive Bayes βελτιώθηκε επίσης σημαντικά με την ακρίβεια να φτάνει το 80.07% και 79.7%. **Αυτό δείχνει ότι η προσθήκη περισσότερων δεδομένων μπορεί να βελτιώσει την απόδοση, αλλά η χρήση feature selection είναι απαραίτητη για την καλύτερη αξιοποίηση των δεδομένων.** Τα συνολικά αποτελέσματα φαίνονται στους Πίνακες 19, Πίνακας 20, Πίνακας 21.

	No Feature Selection				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	82.7%	83.4%	80.8%	96.3%	83.42%
J48	75.9%	76.7%	76.2%	82.2%	76.65%
Naive Bayes	80.7%	80.1%	80.1%	91.7%	80.07%
IBK (NN = 3)	78.9%	61.8%	65.9%	86.1%	61.78%

classified as	Positive	Negative	Neutral
Positive	804	2	0
Negative	46	267	15
Neutral	42	116	51

Random Forest

classified as	Positive	Negative	Neutral
Positive	761	21	24
Negative	36	193	101
Neutral	38	94	77

J48

classified as	Positive	Negative	Neutral
Positive	772	17	17
Negative	27	181	122
Neutral	19	66	124

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	556	5	245
Negative	13	121	196
Neutral	13	42	154

IBK (NN=3)

ΠΙΝΑΚΑΣ 19 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 3 ΧΩΡΙΣ FEATURE SELECTION

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	79.1%	80.5%	79.6%	94.0%	80.52%
J48	80.2%	82.2%	80.3%	91.4%	82.23%
Naive Bayes	69.1%	71.0%	69.6%	81.8%	71.0%
IBK (NN = 3)	79.2%	80.6%	79.7%	91.6%	80.59%

classified as	Positive	Negative	Neutral
Positive	769	12	25
Negative	25	246	59
Neutral	35	106	68

Random Forest

classified as	Positive	Negative	Neutral
Positive	773	11	22
Negative	22	281	27
Neutral	44	113	52

J48

classified as	Positive	Negative	Neutral
Positive	739	19	48
Negative	90	159	81
Neutral	83	69	57

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	775	11	20
Negative	31	224	75
Neutral	41	91	77

IBK (NN=3)

ΠΙΝΑΚΑΣ 20 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 3 (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	83.0%	84.2%	83.0%	95.7%	84.16%
J48	79.3%	79.9%	79.6%	89.3%	79.92%
Naive Bayes	79.5%	80.1%	79.7%	91.8%	80.07%
IBK (NN = 3)	79.2%	76.7%	77.8%	90.1%	76.72%

classified as	Positive	Negative	Neutral
Positive	790	7	9
Negative	31	261	38
Neutral	26	102	81

Random Forest

classified as	Positive	Negative	Neutral
Positive	760	20	26
Negative	29	230	71
Neutral	30	94	85

J48

classified as	Positive	Negative	Neutral
Positive	759	13	34
Negative	31	240	59
Neutral	18	113	78

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	721	20	65
Negative	16	213	101
Neutral	11	100	98

IBK (NN=3)

ΠΙΝΑΚΑΣ 21 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 3 (INFOGAINATTRIBUTEVAL)

7.4 Σετ πειραμάτων 4

Στο τέταρτο σετ πειραμάτων εφαρμόστηκε η τεχνική SMOTE στο συνδυασμένο σετ δεδομένων από το τρίτο σετ. Χωρίς feature selection, η απόδοση του Random Forest έφτασε το 84.23% και η χρήση feature selection με GainRatioAttributeEval και InfoGainAttributeEval έφερε την ακρίβεια του στο 80.81% και 84.90% αντίστοιχα. Ο αλγόριθμος Naive Bayes παρουσίασε σημαντική βελτίωση με ακρίβεια 82.60%

χρησιμοποιώντας InfoGainAttributeEval. Αυτό υπογραμμίζει τη σημασία της αντιμετώπισης της ανισορροπίας στα δεδομένα μέσω τεχνικών όπως το SMOTE και την επιλογή χαρακτηριστικών για τη βελτίωση της απόδοσης των μοντέλων. Τα συνολικά αποτελέσματα φαίνονται στους Πίνακες 22, Πίνακας 23, Πίνακας 24

	No Feature Selection				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	83.2%	84.2%	83.5%	96.3%	84.23%
J48	76.9%	76.4%	76.6%	83.1%	76.43%
Naive Bayes	79.7%	81.0%	80.2%	89.8%	81.04%
IBK (NN = 3)	80.0%	29.7%	28.4%	63.9%	29.73%

classified as	Positive	Negative	Neutral
Positive	793	11	2
Negative	31	244	55
Neutral	24	89	96

Random Forest

classified as	Positive	Negative	Neutral
Positive	737	24	45
Negative	24	212	94
Neutral	34	96	79

J48

classified as	Positive	Negative	Neutral
Positive	795	5	6
Negative	44	199	87
Neutral	37	76	96

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	69	13	724
Negative	0	177	153
Neutral	0	55	154

IBK (NN=3)

ΠΙΝΑΚΑΣ 22 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 4 ΧΩΡΙΣ FEATURE SELECTION

	Feature Selection with GainRatioAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	82.2%	80.8%	81.2%	94.3%	80.81%
J48	79.7%	77.2%	77.9%	91.1%	77.24%
Naive Bayes	79.2%	77.2%	78.1%	91.0%	77.24%
IBK (NN = 3)	80.0%	74.4%	76.3%	89.0%	74.42%

classified as	Positive	Negative	Neutral
Positive	764	8	34
Negative	19	185	126
Neutral	13	58	138

Random Forest

classified as	Positive	Negative	Neutral
Positive	745	13	48
Negative	20	173	137
Neutral	31	57	121

J48

classified as	Positive	Negative	Neutral
Positive	704	14	88
Negative	35	232	63
Neutral	14	92	103

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	689	23	94
Negative	13	178	139
Neutral	13	62	134

IBK (NN=3)

ΠΙΝΑΚΑΣ 23 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 4 (GAINRATIOATTRIBUTEVAL)

	Feature Selection with InfoGainAttributeEval (100 Best)				
Αλγόριθμος	Precision	Recall	F1 Score	ROC Area	Accuracy
Random Forest	84.5%	84.9%	84.7%	95.9%	84.90%
J48	79.5%	77.0%	77.9%	87.9%	77.02%
Naive Bayes	82.3%	82.6%	82.4%	93.1%	82.60%
IBK (NN = 3)	79.5%	73.1%	75.4%	88.8%	73.08%

classified as	Positive	Negative	Neutral
Positive	785	6	15
Negative	21	240	69
Neutral	20	72	117

Random Forest

classified as	Positive	Negative	Neutral
Positive	733	17	56
Negative	18	188	124
Neutral	22	72	115

J48

classified as	Positive	Negative	Neutral
Positive	768	8	30
Negative	31	236	63
Neutral	20	82	107

Naïve Bayes

classified as	Positive	Negative	Neutral
Positive	668	23	115
Negative	8	197	125
Neutral	10	81	118

IBK (NN=3)

ΠΙΝΑΚΑΣ 24 ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΕΤ 4 (INFOGAINATTRIBUTEVAL)

7.5 Σύγκριση Αλγορίθμων

1. Random Forest

- a. Σε όλα τα σετ πειραμάτων, **ο Random Forest καταφέρνει να έχει σταθερά υψηλές τιμές** σε όλες τις μετρικές, όπως ακρίβεια και F1-score (Γράφημα 3). Ιδιαίτερα με την επιλογή χαρακτηριστικών (Feature Selection), καταφέρνει να διατηρήσει ή να βελτιώσει τις επιδόσεις του (Γράφημα 6&7), ενώ το ROC-AUC παραμένει κοντά στο 1, υποδεικνύοντας εξαιρετική ικανότητα διαχωρισμού των κλάσεων. Το μόνο μειονέκτημα αυτού του αλγορίθμου είναι ότι, λόγω της πολυπλοκότητάς του, σε ορισμένες περιπτώσεις αποτυγχάνει να παράγει αποτελέσματα και εμφανίζει ερωτηματικά (Βλέπε σετ πειραμάτων 1 και 2).
- b. Στο σετ πειραμάτων 1 και με την επιλογή χαρακτηριστικών (InfoGainAttributeEval), **ο Random Forest πετυχαίνει ακρίβεια 94.69% και F1-score 94.7% (Βλέπε Γράφημα 6), που είναι από τις καλύτερες επιδόσεις συνολικά.**

2. J48

- a. Ο αλγόριθμος **J48 παρουσιάζει επίσης υψηλές τιμές (Γράφημα 1), ιδιαίτερα σε ό,τι αφορά το F1-score και την ακρίβεια και επιτυγχάνει να παράγει αποτελέσματα για όλα τα σύνολα δεδομένων.** Σε κάποιες περιπτώσεις όμως δεν ξεπερνά τον Random Forest, ειδικά όταν αξιολογούμε την καμπύλη ROC-AUC και τη συνολική ακρίβεια.
- b. Στο σετ πειραμάτων 3 με την επιλογή χαρακτηριστικών (InfoGainAttributeEval), ο J48 φτάνει ακρίβεια 79.92%, αρκετά χαμηλότερη από τον Random Forest. Στα σετ πειράματων 1 και 2, το F1-score ήταν σημαντικά υψηλότερο από αυτό του Random Forest (Βλέπε Γράφημα 5,6 &7).

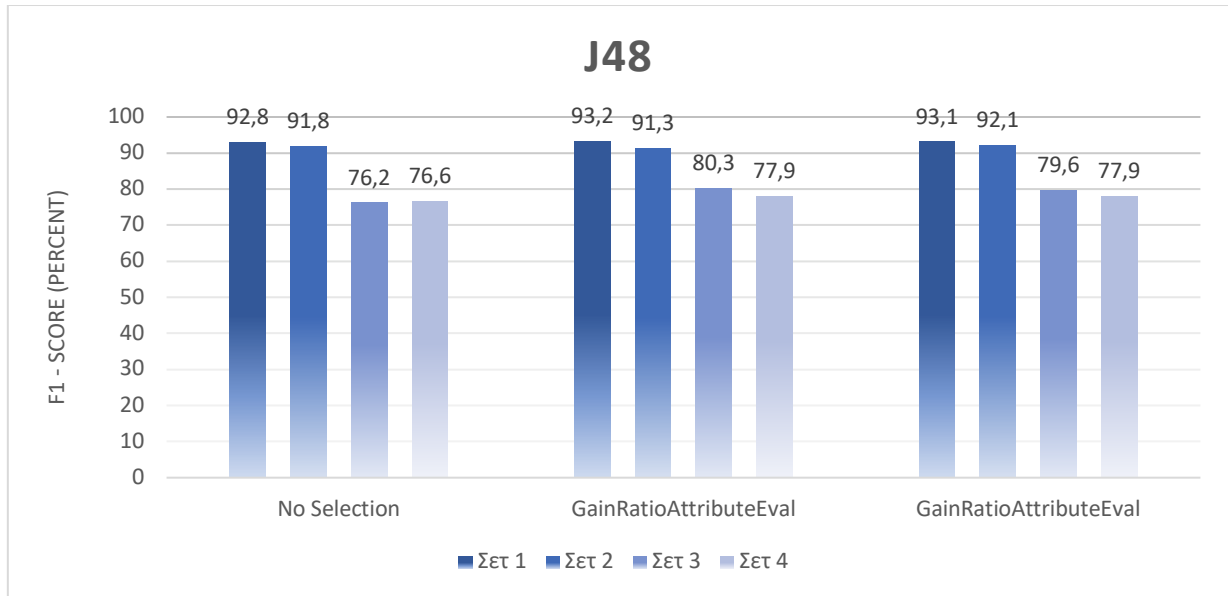
3. Naïve Bayes

- a. Ο **Naïve Bayes παρουσιάζει σχετικά καλές επιδόσεις σε μερικά από τα πειράματα (Γράφημα 4), ειδικά στην ακρίβεια προβλέψεων (precision).** Ωστόσο, η ανάκληση (recall) και το F1-score συχνά είναι χαμηλότερα σε σύγκριση με άλλους αλγορίθμους.
- b. Η **ακρίβειά του σε αρκετά από τα πειράματα παραμένει σημαντικά χαμηλότερη**, όπως στο σετ πειραμάτων 1 χωρίς επιλογή χαρακτηριστικών, όπου φτάνει μόνο το 58.36% (Βλέπε Γράφημα 5).

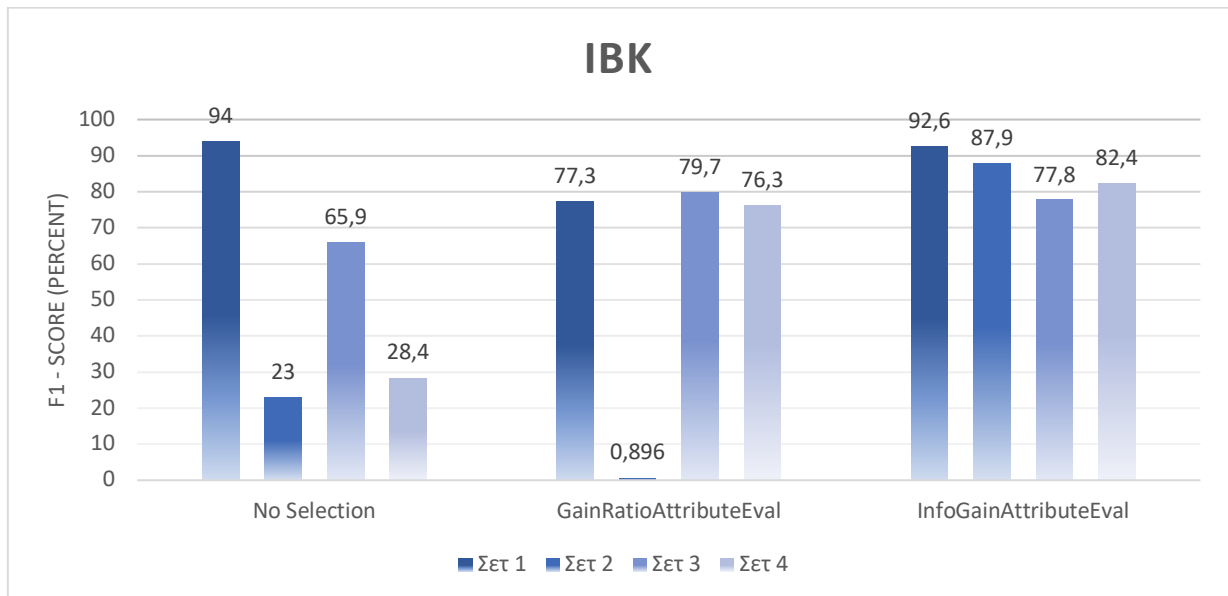
4. IBK (NN = 3)

- a. Ο **IBK (NN = 3) παρουσιάζει αστάθεια στις επιδόσεις του ανάμεσα στα πειράματα (Βλέπε Γράφημα 2).** Ενώ μπορεί να έχει καλή ακρίβεια προβλέψεων, η συνολική του ακρίβεια και το F1-score είναι συχνά χαμηλότερα από τα αντίστοιχα του Random Forest και του J48.
- b. Σε συγκεκριμένα πειράματα, όπως στο σετ πειραμάτων 2 χωρίς επιλογή χαρακτηριστικών, η ακρίβεια πέφτει στο 15.57%, γεγονός **που δείχνει**

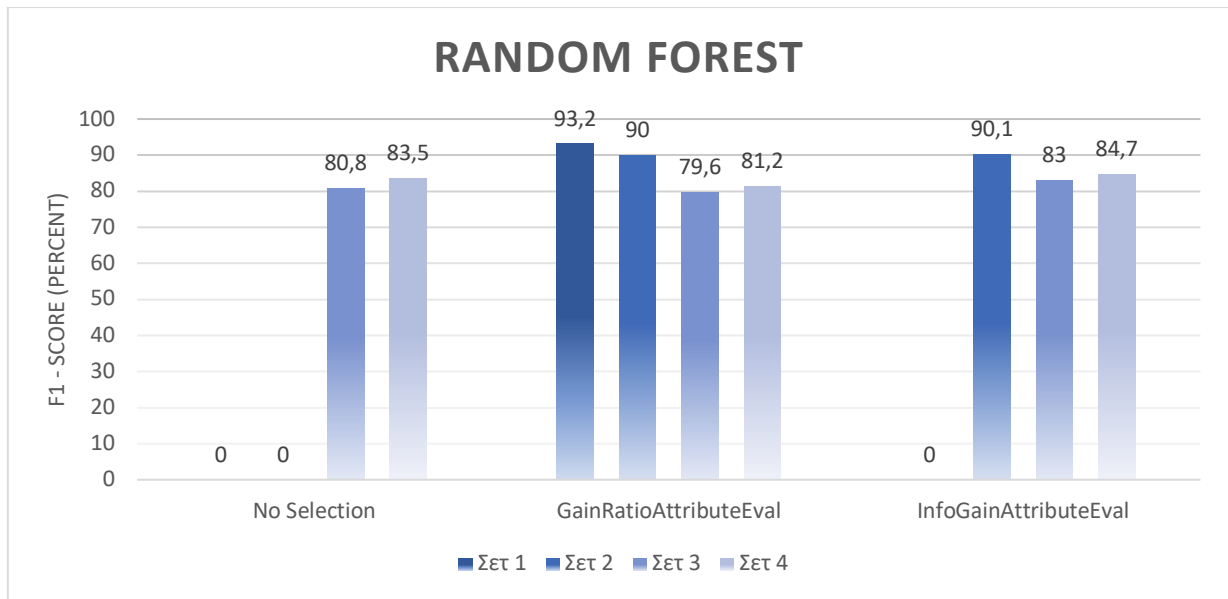
σημαντικά προβλήματα στην απόδοση του μοντέλου (Βλέπε Γράφημα 5).



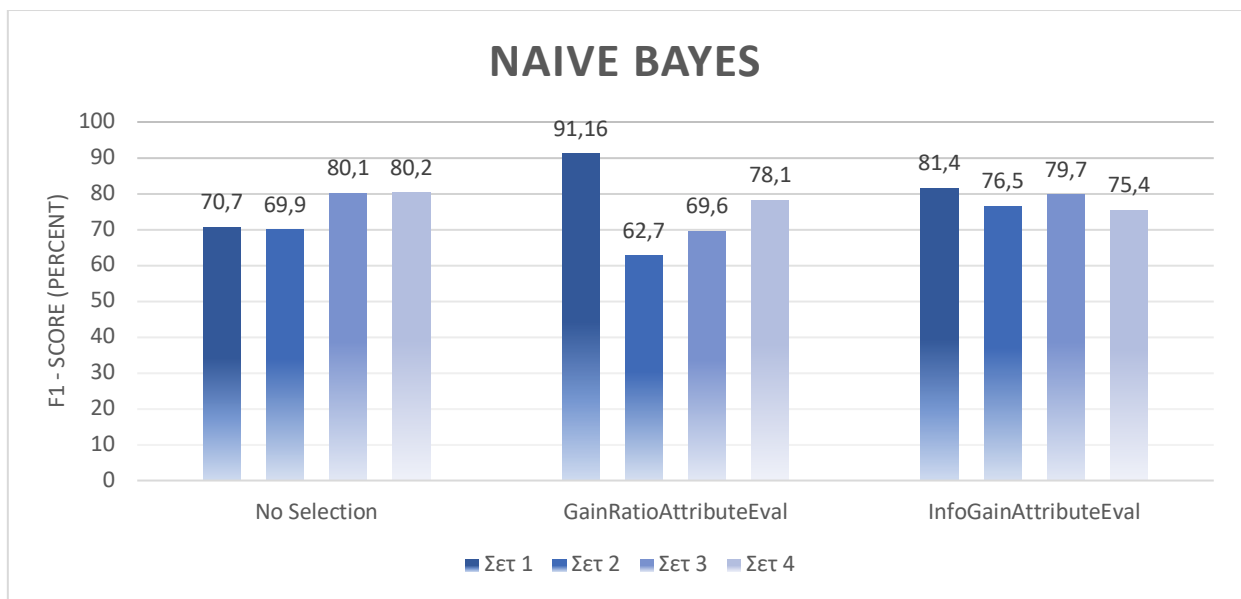
ΓΡΑΦΗΜΑ 1 F1-SCORE ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ J48 ΣΕ ΟΛΑ ΤΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ



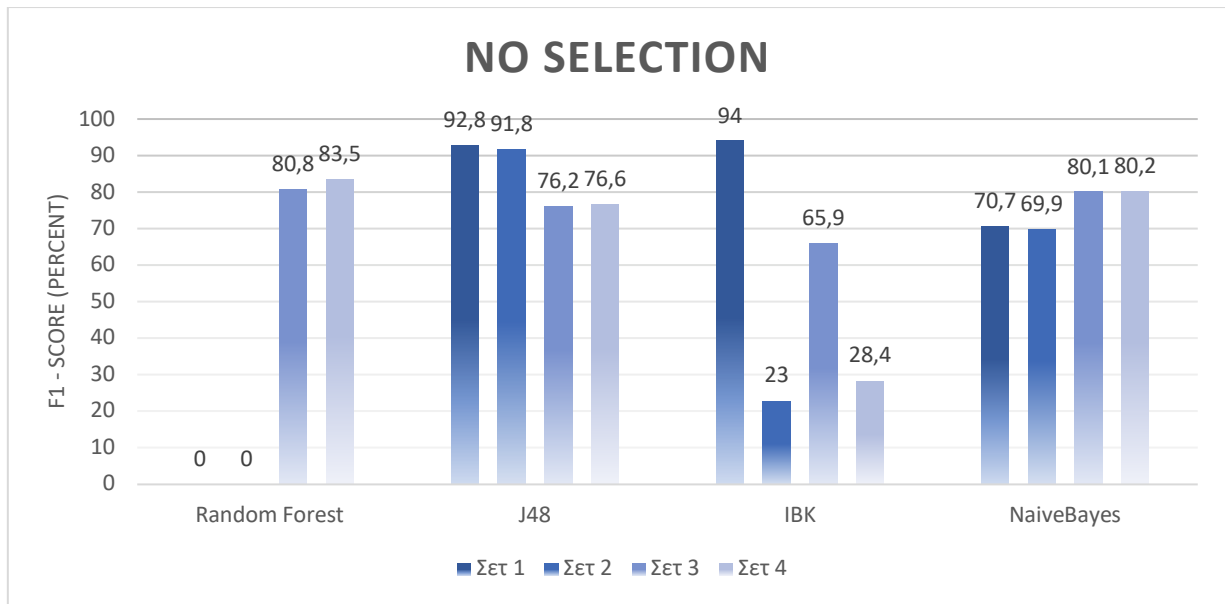
ΓΡΑΦΗΜΑ 2 F1-SCORE ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ IBK ΣΕ ΟΛΑ ΤΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ



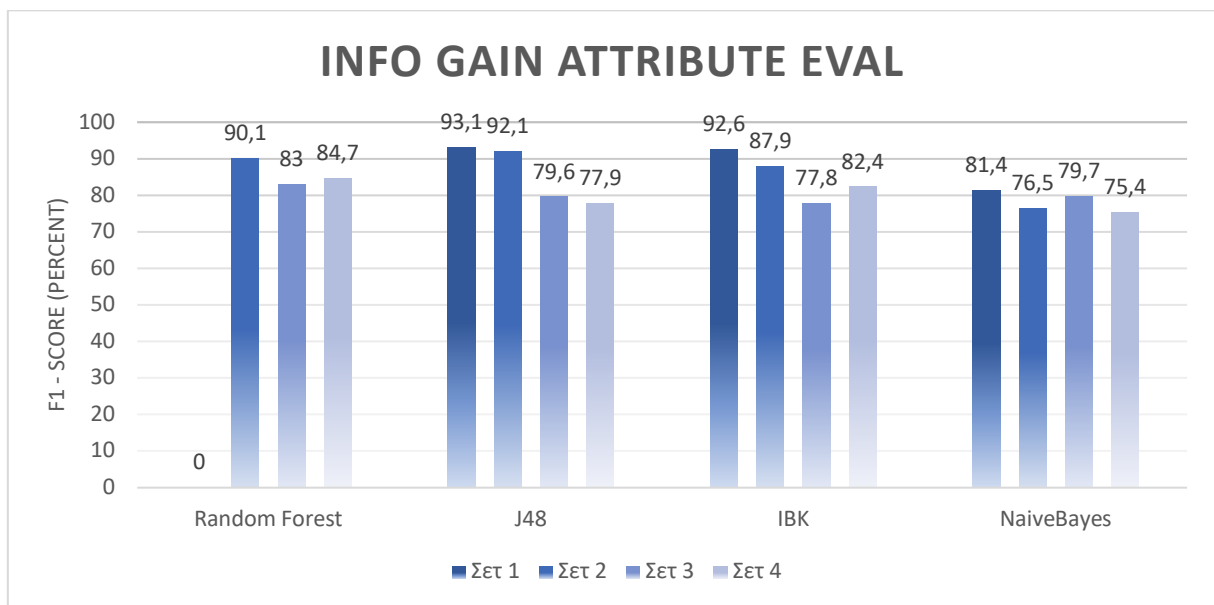
ΓΡΑΦΗΜΑ 3 F1-SCORE ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ RANDOM FOREST ΣΕ ΟΛΑ ΤΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ



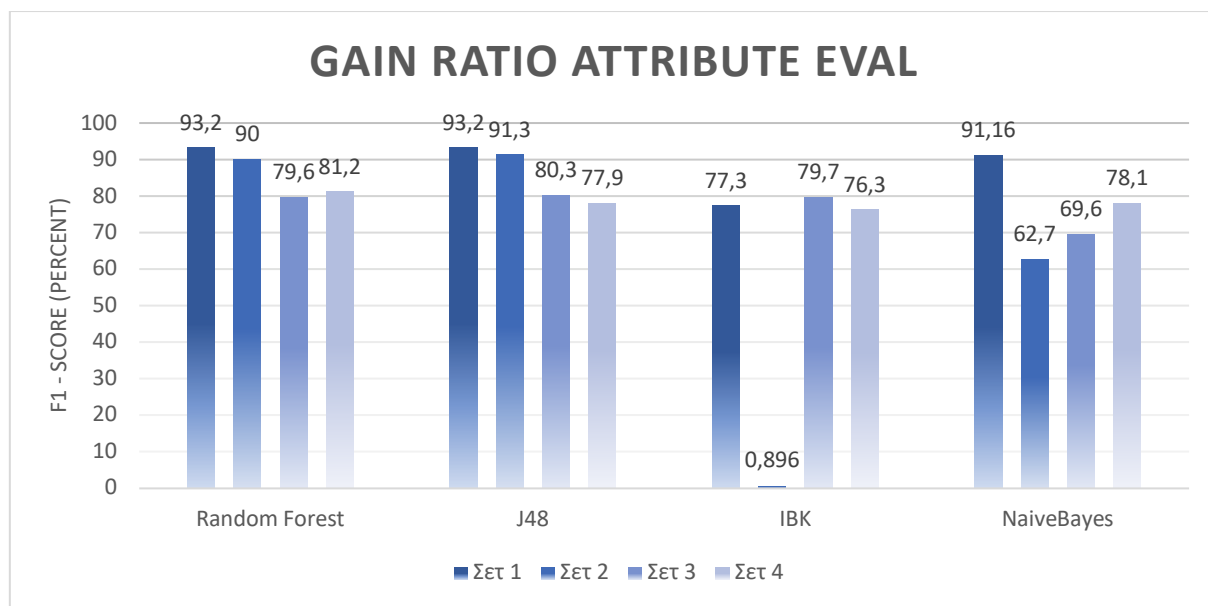
ΓΡΑΦΗΜΑ 4 F1-SCORE ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ NAIVE BAYES ΣΕ ΟΛΑ ΤΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ



ΓΡΑΦΗΜΑ 5 F1-SCORE ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΣΕ ΟΛΑ ΤΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ ΧΩΡΙΣ FEATURE SELECTION



ΓΡΑΦΗΜΑ 6 F1-SCORE ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΣΕ ΟΛΑ ΤΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ ΜΕ FEATURE SELECTION ΚΑΙ INFO GAIN ATTRIBUTE EVAL



ΓΡΑΦΗΜΑ 7 F1-SCORE ΤΩΝ ΑΛΓΟΡΙΘΜΩΝ ΣΕ ΟΛΑ ΤΑ ΣΕΤ ΔΕΔΟΜΕΝΩΝ ΜΕ FEATURE SELECTION ΚΑΙ GAIN RATIO ATTRIBUTE EVAL

7.6 Σύγκριση με προηγούμενες μελέτες

Η παρούσα διπλωματική εργασία επικεντρώνεται στην ανάλυση συναισθημάτων σε διαδικτυακές κριτικές πεντάστερων ξενοδοχείων της Κέρκυρας, εφαρμόζοντας επιβλεπόμενες μεθόδους μηχανικής μάθησης για την κατηγοριοποίηση των κριτικών ως θετικές, αρνητικές ή ουδέτερες. **Σε σύγκριση με προηγούμενες μελέτες, αυτή η εργασία διακρίνεται τόσο ως προς το πεδίο εφαρμογής όσο και ως προς τη μεθοδολογική προσέγγιση που ακολουθεί.** Συγκεκριμένα, ενώ η μελέτη των (Fazzolari, 2018) επικεντρώνεται στην εφαρμογή τεχνικών εξόρυξης δεδομένων σε μεγάλο όγκο ταξιδιωτικών κριτικών για την εξαγωγή γενικών πληροφοριών και προτάσεων.

Αναφορικά με την απόδοση των μοντέλων, η εργασία αυτή ανέδειξε ότι οι δένδροκεντρικοί αλγόριθμοι (Random Forest και J48) είχαν την καλύτερη απόδοση, επιτυγχάνοντας υψηλότερη ακρίβεια στην κατηγοριοποίηση των συναισθημάτων σε σύγκριση με άλλους αλγόριθμους, όπως ο Naive Bayes και ο IBK (k-nearest neighbors). Αυτό έρχεται σε αντίθεση με τη μελέτη "Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques" (Anis Sarah, 2021), όπου ο Support Vector Machine (SVM) αναδείχθηκε ως ο αλγόριθμος με την καλύτερη απόδοση, επιτυγχάνοντας ακρίβεια 86.3% σε ένα γενικότερο σύνολο δεδομένων κριτικών ξενοδοχείων. Ενώ ο SVM έδειξε εξαιρετική απόδοση στη μελέτη τους, η παρούσα εργασία αποδεικνύει ότι ο J48 και ο Random Forest μπορεί να είναι πιο αποδοτικοί σε συγκεκριμένα τοπικά πλαίσια, όπως η

Κέρκυρα, ειδικά όταν οι κλάσεις δεδομένων είναι ανισομερείς και απαιτούν ειδικές τεχνικές εξισορρόπησης όπως το SMOTE.

Σε σχέση με τη μελέτη "Opinion mining from online hotel reviews – A text summarization approach" (Chou, 2017), η οποία επικεντρώνεται στη σύνοψη κειμένου από διαδικτυακές κριτικές, η παρούσα εργασία προσφέρει έναν πιο προσαρμοσμένο και εξειδικευμένο τρόπο αντιμετώπισης των προκλήσεων που αφορούν την ανάλυση συναισθημάτων σε τοπικό επίπεδο. Τα μοντέλα που αναπτύχθηκαν εδώ, παρόλο που μπορεί να μην επιτυγχάνουν τις ίδιες ακρίβειες με τα γενικότερα μοντέλα άλλων μελετών, είναι καλύτερα προσαρμοσμένα στις συγκεκριμένες ανάγκες και τα χαρακτηριστικά των δεδομένων της Κέρκυρας, προσφέροντας ακριβέστερα και πιο αξιόπιστα αποτελέσματα για αυτό το πλαίσιο.

Εν κατακλείδι, η παρούσα εργασία αναδεικνύει τη σημασία της προσαρμογής των μοντέλων μηχανικής μάθησης στις ιδιαιτερότητες του εκάστοτε συνόλου δεδομένων, και παρέχει πρακτικές κατευθύνσεις για την τουριστική βιομηχανία όσον αφορά τη βελτιστοποίηση της ανάλυσης συναισθημάτων σε τοπικό επίπεδο. Αν και τα αποτελέσματα μπορεί να διαφέρουν από εκείνα που αναφέρονται σε άλλες μελέτες, η προσέγγιση αυτή προσφέρει μια πιο εξειδικευμένη και συνεπή ανάλυση των δεδομένων, καθιστώντας τα μοντέλα πιο αποδοτικά για τις ανάγκες της συγκεκριμένης περιοχής.

8. Δυσκολίες

8.1 Λήψη Δεδομένων

Κατά τη διαδικασία λήψης δεδομένων μέσω scraping, αντιμετωπίσαμε δυσκολίες καθώς οι κλάσεις (HTML class) που χρησιμοποιούσαμε για τον εντοπισμό του κουμπιού της επόμενης σελίδας και της πλήρους προβολής των κριτικών άλλαζαν ονομασία. Αυτό είχε ως αποτέλεσμα την ανάγκη συνεχών τροποποιήσεων στον κώδικα για να μπορέσουμε να κατεβάσουμε απαραίτητα δεδομένα.

8.2 Επισήμανση Δεδομένων (Labeling)

Επειδή ο στόχος της εργασίας ήταν η ανάλυση συναισθήματος και τα δεδομένα που χρησιμοποιήθηκαν και στα δύο σύνολα δεδομένων αφορούσαν πραγματικά σχόλια επισκεπτών και βαθμολόγηση με τη μορφή αστεριών, ήταν απαραίτητο να χαρακτηριστούν ως αρνητικά, θετικά ή ουδέτερα. Καθώς δεν υπήρχε διαθέσιμη αυτοματοποιημένη μέθοδος, η διαδικασία έπρεπε να γίνει χειροκίνητα.

8.3 Δημιουργία Διανύσματος Χαρακτηριστικών (Feature Vector)

Κατά τη διαδικασία της προεπεξεργασίας και της δημιουργίας διανυσμάτων χαρακτηριστικών, προέκυψε η εξής δυσκολία: το φίλτρο "string to word vector" έπρεπε να εφαρμοστεί πριν από τον διαχωρισμό των δεδομένων σε σύνολο εκπαίδευσης και δοκιμής, ώστε και τα δύο να έχουν τα ίδια χαρακτηριστικά. Αν η διαδικασία γινόταν ξεχωριστά για τα δύο σύνολα μετά τον διαχωρισμό, αυτά δεν θα είχαν τον ίδιο αριθμό και τύπο

χαρακτηριστικών, με αποτέλεσμα οι ταξινομητές να χρησιμοποιούν μόνο τα κοινά χαρακτηριστικά.

8.4 Ανισορροπία Δείγματος

Τα σύνολα δεδομένων του πραγματικού κόσμου σε πολλούς τομείς όπως ο τουρισμός, η ιατρική, η ανίχνευση εισβολών, οι συναλλαγές απάτης και η βιοπληροφορική είναι εξαιρετικά ανισόρροπα. **Στα προβλήματα ταξινόμησης, τα μη ισορροπημένα σύνολα δεδομένων επηρεάζουν αρνητικά την ακρίβεια των προβλέψεων κλάσεων.** Το πρόβλημα των μη ισορροπημένων συνόλων δεδομένων στη μηχανική μάθηση είναι ένα σημαντικό ερευνητικό ζήτημα.

Τα μη ισορροπημένα σύνολα δεδομένων είναι πιο σημαντικά όταν ορίζετε μια εποπτευόμενη Μηχανική Εκμάθηση με δύο ή περισσότερες κλάσεις. **Τα μη ισορροπημένα δεδομένα υποδεικνύουν ότι ο αριθμός των δεδομένων που είναι προσβάσιμα για κάθε τάξη ποικίλλει.** Εάν υπάρχουν για παράδειγμα δύο κατηγορίες, τα ισορροπημένα δεδομένα σημαίνουν 50 ποσοστιαίες μονάδες για κάθε κατηγορία. Η ελαφρά ανισορροπία δεν αποτελεί ανησυχία για τις περισσότερες προσεγγίσεις μηχανικής μάθησης. Ως αποτέλεσμα, εάν η μία κατηγορία έχει το 60% των δεδομένων ενώ η άλλη το 40%, δεν θα πρέπει να υπάρχει αισθητή μείωση της απόδοσης. Μόνο όταν τα μη ισορροπημένα σύνολα δεδομένων της μηχανικής μάθησης είναι ακραία (δηλαδή 90% για τη μία τάξη και 10% για την άλλη), οι τυπικές παράμετροι βελτιστοποίησης ή οι μετρήσεις απόδοσης θα είναι αναποτελεσματικές και θα απαιτούν προσαρμογή. Υπάρχουν πολλές στρατηγικές που μπορούν να χρησιμοποιηθούν για να εξισορροπηθούν τα δεδομένα, κάποιες εξ' αυτών είναι οι εξής:

- Συλλογή Περισσότερων Δεδομένων
- Υποδειγματοληψία
- Υπερδειγματοληψία (τεχνική SMOTE).

9. Συμπεράσματα

Συνολικά, αν χρειαζόταν να επιλεγεί ένας αλγόριθμος, ο J48 θα θεωρούνταν το καλύτερο μοντέλο, καθώς πέτυχε υψηλότερες τιμές ακρίβειας, F1-score και ROC-AUC στις περισσότερες συνθήκες, ανεξάρτητα από την επιλογή χαρακτηριστικών. Ειδικά όταν εφαρμόζεται η επιλογή χαρακτηριστικών με την τεχνική InfoGainAttributeEval (Βλέπε Γράφημα 6), ο J48 επιδεικνύει άριστες επιδόσεις, καθιστώντας το τον πιο αξιόπιστο αλγόριθμο για την κατηγοριοποίηση συναισθημάτων στις διαδικτυακές κριτικές.

Η ανάλυση των τεσσάρων σετ πειραμάτων καταδεικνύει όχι μόνο την υπεροχή του J48, αλλά και τη σημασία της προσεκτικής επιλογής και εφαρμογής κατάλληλων τεχνικών, όπως η επιλογή χαρακτηριστικών, για τη βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης. Αυτή η στρατηγική μπορεί να είναι καθοριστική για την επιτυχία στην

κατηγοριοποίηση συναισθημάτων, επισημαίνοντας τη σύνδεση μεταξύ της τεχνολογικής προσέγγισης και των τελικών αποτελεσμάτων.

8.1. Επιλογή Χαρακτηριστικών (Feature Selection)

Η διαδικασία της επιλογής χαρακτηριστικών είναι κρίσιμη για τη μείωση της διάστασης του προβλήματος και την εξάλειψη άχρηστων ή μη σχετικών δεδομένων. Η χρήση τεχνικών επιλογής χαρακτηριστικών και οι μέθοδοι βασισμένες σε δέντρα αποφάσεων, μπορεί να ενισχύσει την ακρίβεια των μοντέλων ταξινόμησης μειώνοντας τον θόρυβο στα δεδομένα. Επιπλέον, η μείωση των χαρακτηριστικών συμβάλλει στη μείωση της υπερπροσαρμογής (overfitting), καθιστώντας το μοντέλο πιο γενικεύσιμο σε νέα δεδομένα.

8.2. Διαχείριση Ανισορροπημένων Δεδομένων

Σε πολλές περιπτώσεις, τα δεδομένα είναι ανισορροπημένα, με αποτέλεσμα ορισμένες κλάσεις να υπερεκπροσωπούνται και άλλες να είναι υποεκπροσωπούνται. Η χρήση τεχνικών όπως το SMOTE (Synthetic Minority Over-sampling Technique) μπορεί να δημιουργήσει συνθετικά δεδομένα της μειοψηφικής κλάσης, εξισορροπώντας έτσι την κατανομή των δεδομένων και βελτιώνοντας τις μετρικές απόδοσης του μοντέλου. Τεχνικές όπως η υποδειγματοληψία της πλειοψηφικής κλάσης ή η χρήση πιο προηγμένων αλγορίθμων ταξινόμησης μπορούν επίσης να συμβάλουν στη διαχείριση αυτής της πρόκλησης.

8.3. Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων είναι ένα βήμα που δεν μπορεί να παραβλεφθεί. Η κανονικοποίηση, η απομάκρυνση των ελλিপών τιμών, η μετατροπή κατηγοριοποιημένων δεδομένων σε αριθμητικές μορφές και η εξομάλυνση των χαρακτηριστικών είναι μερικές από τις τεχνικές που μπορούν να χρησιμοποιηθούν για την βελτίωση της ποιότητας των δεδομένων. Η κατάλληλη προεπεξεργασία των δεδομένων οδηγεί σε πιο αξιόπιστα και αποτελεσματικά μοντέλα.

8.4. Χρήση Περισσότερων Δεδομένων

Η ποσότητα των δεδομένων που χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου μπορεί να επηρεάσει σημαντικά την απόδοσή του. Περισσότερα δεδομένα μπορούν να προσφέρουν μια πιο αντιπροσωπευτική εικόνα του προβλήματος και να βελτιώσουν την ικανότητα του μοντέλου να γενικεύει. Ωστόσο, η συλλογή και η διαχείριση μεγάλων όγκων δεδομένων απαιτεί επίσης κατάλληλη υποδομή και τεχνικές διαχείρισης δεδομένων για να διασφαλιστεί η αποτελεσματικότητα και η αποδοτικότητα της διαδικασίας.

8.5 Συμπεράσματα

Η ανάλυση των τεσσάρων σετ πειραμάτων αποδεικνύει ότι η βελτίωση των μοντέλων ταξινόμησης συναισθημάτων απαιτεί μια ολοκληρωμένη προσέγγιση που περιλαμβάνει την επιλογή κατάλληλων τεχνικών προεπεξεργασίας, την εξισορρόπηση των δεδομένων, την επιλογή χαρακτηριστικών και την χρήση μεγαλύτερων σετ δεδομένων. Η συνεργασία αυτών των στοιχείων μπορεί να οδηγήσει στην ανάπτυξη πιο ακριβών, αποδοτικών και αξιόπιστων μοντέλων, τα οποία είναι κρίσιμα για την επιτυχή ανάλυση και κατανόηση των συναισθημάτων.

10. Προτάσεις για Μελλοντική Έρευνα

Η παρούσα μελέτη έδειξε ότι οι τεχνικές μηχανικής μάθησης μπορούν να βελτιώσουν σημαντικά την ανάλυση συναισθημάτων σε διαδικτυακές κριτικές. Η χρήση αλγορίθμων όπως οι Random Forest, J48, Naïve Bayes και IBK (NN = 3) απέδειξε ότι μπορούν να επιτευχθούν υψηλά επίπεδα ακρίβειας στην κατηγοριοποίηση των συναισθημάτων, με τη βοήθεια προεπεξεργασίας δεδομένων, επιλογής χαρακτηριστικών και εξισορρόπησης κλάσεων μέσω της τεχνικής SMOTE. **Παρόλα αυτά, η ανισορροπία των δεδομένων παραμένει μια πρόκληση που απαιτεί περαιτέρω διερεύνηση.**

Επιπλέον, οι παραδοσιακές τεχνικές μηχανικής μάθησης μπορεί να υστερούν στην αναγνώριση πολύπλοκων συναισθημάτων όπως ο σαρκασμός και η ειρωνεία. Σε αυτό το πλαίσιο, η χρήση νευρωνικών δικτύων, όπως τα Long Short-Term Memory (LSTM), έδειξε σημαντική βελτίωση στην ακρίβεια και την αξιοπιστία των προβλέψεων.

9.1 Εφαρμογή Μετα-Μάθησης (Meta-Learning):

Η εφαρμογή τεχνικών μετα-μάθησης μπορεί να βοηθήσει τα μοντέλα να προσαρμόζονται καλύτερα σε νέα δεδομένα και να βελτιώνουν την απόδοσή τους σε διαφορετικά περιβάλλοντα. Αυτό μπορεί να επιτευχθεί μέσω της εκπαίδευσης των μοντέλων σε διάφορα σύνολα δεδομένων από διαφορετικούς τομείς.

9.2 Χρήση Προηγμένων Νευρωνικών Δικτύων:

Η ενσωμάτωση νευρωνικών δικτύων όπως τα μοντέλα BERT (Bidirectional Encoder Representations from Transformers) και GPT (Generative Pre-trained Transformer) μπορεί να βελτιώσει την ακρίβεια της ανάλυσης συναισθημάτων. Αυτά τα μοντέλα έχουν αποδειχθεί εξαιρετικά αποτελεσματικά στην κατανόηση και επεξεργασία της φυσικής γλώσσας, επιτρέποντας την ανάλυση πιο περίπλοκων και υπονοούμενων συναισθημάτων σε κείμενα.

9.3 Ανάλυση Συναισθημάτων σε Πραγματικό Χρόνο:

Η ανάπτυξη συστημάτων που μπορούν να αναλύουν συναισθήματα σε πραγματικό χρόνο θα μπορούσε να προσφέρει σημαντικά πλεονεκτήματα σε τομείς όπως η εξυπηρέτηση πελατών και τα μέσα κοινωνικής δικτύωσης. Αυτό θα επιτρέψει την άμεση ανταπόκριση σε αρνητικά σχόλια ή κρίσεις και τη βελτίωση της αλληλεπίδρασης με τους χρήστες.

9.4 Ενσωμάτωση Πολυτροπικών Δεδομένων:

Η μελλοντική έρευνα μπορεί να επικεντρωθεί στην ανάλυση συναισθημάτων μέσω πολλαπλών μορφών δεδομένων, όπως κείμενο, εικόνες, βίντεο και ήχος. Για παράδειγμα, η ανάλυση βίντεο και φωνής από πλατφόρμες όπως το YouTube μπορεί να προσφέρει μια πιο ολοκληρωμένη εικόνα των συναισθημάτων των χρηστών.

9.5 Εξέταση Νέων Μεθόδων Προεπεξεργασίας Δεδομένων:

Η διερεύνηση νέων μεθόδων προεπεξεργασίας δεδομένων, όπως η χρήση προηγμένων τεχνικών κανονικοποίησης και επιλογής χαρακτηριστικών, θα μπορούσε να βελτιώσει την απόδοση των μοντέλων. Επίσης, η εφαρμογή τεχνικών ενίσχυσης δεδομένων (data augmentation) μπορεί να συμβάλλει στη μείωση της ανισορροπίας των δεδομένων.

9.6 Ενσωμάτωση Κοινωνικών Δικτύων και Big Data:

Η ανάλυση δεδομένων από κοινωνικά δίκτυα, όπως το Facebook, το Twitter και το Instagram, μπορεί να προσφέρει χρήσιμες πληροφορίες για τις τάσεις και τα συναισθήματα του κοινού. Η ενσωμάτωση αυτών των δεδομένων με άλλες πηγές big data μπορεί να παρέχει μια πιο ολοκληρωμένη εικόνα. Αυτές οι προτάσεις παρέχουν κατευθυντήριες γραμμές για την περαιτέρω ανάπτυξη και βελτίωση της ανάλυσης συναισθημάτων σε online κριτικές, ανοίγοντας νέους δρόμους για την αξιοποίηση των δεδομένων που προέρχονται από τις απόψεις των καταναλωτών.

9.7 Εφαρμογές στην Ιατρική και την Υγεία:

Η ανάλυση συναισθημάτων μπορεί να βρει εφαρμογές στην ιατρική και την υγεία, όπως η ανίχνευση της ψυχολογικής κατάστασης των ασθενών μέσω των κοινωνικών δικτύων ή των ημερολογίων τους. Αυτή η προσέγγιση μπορεί να βοηθήσει στην έγκαιρη ανίχνευση και παρέμβαση σε περιπτώσεις ψυχικής υγείας.

9.8 Βελτίωση Εξυπηρέτησης Πελατών:

Η ανάλυση των συναισθημάτων των πελατών μέσω των κριτικών τους και της αλληλεπίδρασής τους με την επιχείρηση μπορεί να βοηθήσει στη βελτίωση της ποιότητας των υπηρεσιών και της ικανοποίησης των πελατών. Ειδικά τα συστήματα που χρησιμοποιούν real-time sentiment analysis μπορούν να εντοπίζουν και να επιλύουν προβλήματα άμεσα.

9.9 Τεχνητή Νοημοσύνη και Αυτοματισμοί:

Η ενσωμάτωση των τεχνικών ανάλυσης συναισθημάτων με άλλα συστήματα τεχνητής νοημοσύνης, όπως chatbots και ψηφιακοί βοηθοί, μπορεί να βελτιώσει τη διαδραστικότητα και την προσαρμοστικότητα αυτών των συστημάτων στις ανάγκες των χρηστών.

11. Βιβλιογραφία

- Anis Sarah, S. S. (2021). Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques. *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2020* (σσ. 227-234). Springer International Publishing.
- Breiman, L. (2001). Random forests. *Machine learning*, 5--32.
- Cantalalops, A. S., & Salvi, F. (2014). New consumer behavior: A review of research on eWOM and hotels. *International Journal of Hospitality Management*, σσ. 41-51.
- Chaovalit, P., & Zhou, L. (2005, January). Movie review mining: A comparison between supervised and unsupervised classification approaches. *In Proceedings of the 38th annual Hawaii international conference on system sciences* (σσ. 112c-112c). IEEE.
- Chawla, N. V. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, σσ. 321–357.
- Chou, Y.-H. H.-L.-L. (2017). Opinion mining from online hotel reviews – A text summarization approach. *Information Processing & Management*, 436-449.
- Chowdhary, K. R. (2020). *Natural Language Processing*. India: Springer.
- Cover, T. a. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 21--27.
- Dwityas, N. A., & Briandana, R. (2017). Social media in travel decision making process. *International Journal of Humanities and Social Science*, σσ. 193-201.
- Fazzolari, M. P. (2018). A study on online travel reviews through intelligent data analysis. *Information Technology & Tourism*.

- Gretzel, U., & Yoo, K. H. (2008). Use and impact of online travel reviews. *In Information and communication technologies in tourism 2008*, Springer, Vienna, σσ. 35-46.
- Gudykunst, W. B., & Hammer, M. R. (1988). Strangers and hosts: An uncertainty reduction based theory of intercultural adaptation. *Cross-cultural adaptation: Current approaches*, σσ. 106-139.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 10-18.
- Hall, M., Holmes, G., Frank, E., Pfahringer, B., Reutuma, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 10-18.
- Kok, J. N., Boers, E. J., Kusters, W. A., & Putten, V. d. (2009). Artificial intelligence: definition, trends, techniques, and cases. *Artificial intelligence*, σσ. 270-299.
- McCallum, A. a. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*, (σσ. 41--48).
- Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.
- Rossini, D. M. (2021). Text categorization with WEKA: A survey. *Machine Learning with Applications*.
- Sheikh, H., Prins, S., & Schrijvers, E. (2021). *Mission AI The New System Technology*. The Netherlands: Springer.
- Van Rossum, G. a. (2009). *Python 3 Reference Manual*. CreateSpace.
- Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, σσ. 6527-6535.
- About us: Tripadvisor. (2024, Ιανουάριος). Ανάκτηση από TrpAdvisor: <https://tripadvisor.mediaroom.com/US-about-us>

12. Παράρτημα

11.1 Κώδικας εξόρυξης Δεδομένων

```
import csv
import os
import time

from selenium import webdriver
```

```

from selenium.webdriver.common.action_chains import ActionChains
from selenium.webdriver.common.by import By
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.support.ui import WebDriverWait

path_to_file = "./Results.csv"
num_pages = 45
urls = [
    "https://www.tripadvisor.com/Hotel_Review-g662629-d12831759-Reviews-Rodostamo_Hotel_Spa-Corfu_Town_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g8323529-d228887-Reviews-Corfu_Imperial_Grecotel_Beach_Luxe_Resort-Limni_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g662629-d228902-Reviews-Corfu_Palace_Hotel-Corfu_Town_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g776025-d23182681-Reviews-Angsana_Corfu-Benitses_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g4283440-d240844-Reviews-MarBella_Mar_Bella_Collection-Agios_Ioannis_Peristeron_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g1191733-d276098-Reviews-Kontokali_Bay_Resort_and_Spa-Kontokali_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g1191902-d270531-Reviews-La_Grotta_Verde-Agios_Gordios_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g662629-d289555-Reviews-Mon_Repos_Palace-Corfu_Town_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g189458-d478750-Reviews-Kerkyra_Blue_Hotel_N_Spa-Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g775848-d483122-Reviews-Atlantica_Nissaki_Beach-Nissaki_Corfu_Ionian_Islands.html",
    "https://www.tripadvisor.com/Hotel_Review-g775844-d653313-Reviews-Philoxenia_Hotel-Ermones_Corfu_Ionian_Islands.html",
]
names = [
    "Rodostamo_Hotel_Spa",
    "Corfu_Imperial_Grecotel_Beach_Luxe_Resort",
    "Corfu_Palace",
    "Angsana_Corfu",
    "MarBella_Corfu",
    "Kontokali_Bay",
    "La_Grotta_Verde",
    "Mon_Repos_Palace",
    "Kerkyra_Blue",
    "Atlantica_Nissaki",
    "Philoxenia_Hotel",
]

# Check if the file exists
if os.path.exists(path_to_file):
    # If the file exists, delete it

```

```

os.remove(path_to_file)
print(f"Deleted existing file: {path_to_file}")

# Open the file to save the review
csvFile = open(path_to_file, "a", encoding="utf-8")
csvWriter = csv.writer(csvFile)
csvWriter.writerow(
    ["Hotel_Name", "Date_Of_Stay", "Rating", "Review_title", "Review_text"]
)
for url, name in zip(urls, names):
    # Create a ChromeOptions object and set options
    chrome_options = webdriver.ChromeOptions()
    chrome_options.headless = (
        False # Set to True for headless mode, False for GUI mode
    )
    driver = webdriver.Chrome(options=chrome_options)
    driver.get(url)

    # change the value inside the range to save more or less reviews
    for i in range(0, num_pages):
        time.sleep(2)

        try:
            # Check for the presence of the privacy statement or "I Accept" button
            privacy_statement = driver.find_element(
                By.ID, "onetrust-button-group-parent"
            ) # Replace with the actual element ID
            accept_button = driver.find_element(
                By.ID, "onetrust-accept-btn-handler"
            ) # Replace with the actual element ID

            # If the privacy statement element is found, click the "I Accept" button
            if privacy_statement:
                accept_button.click()
        except:
            pass

        container = driver.find_elements(By.XPATH, ".*//div[@data-reviewid]")
        # Wait for the overlay to become invisible
        # wait = WebDriverWait(driver, 10)
        # overlay = wait.until(EC.invisibility_of_element_located((By.CLASS_NAME,
"onetrust-pc-dark-filter"))))
        try:
            read_more_button = driver.find_element(
                By.XPATH, ".*//div[contains(@data-test-target,'expand-review')]"
            )
            print(driver.execute_script("arguments[0].click();", read_more_button))
            driver.execute_script("arguments[0].click();", read_more_button)

```

```

except:
    pass
for j in range(len(container)):
    try:
        title = (
            container[j]
            .find_element(
                By.XPATH, ".*div[contains(@data-test-target, 'review-title')]"
            )
            .text
        )
        print("title = {}".format(title))
        date = (
            container[j]
            .find_element(
                By.XPATH, ".*span[contains(@class, 'teHYY _R Me S4 H3')]"
            )
            .text[14:]
        )
        print("date = {}".format(date))
        rating = (
            int(
                container[j]
                .find_element(By.XPATH, ".*span['ui_bubble_rating bubble_']")
                .get_attribute("class")
                .split("_")[3]
            )
            / 10.0
        )
        print("Rating = {}".format(rating))
        review = (
            container[j]
            .find_element(By.XPATH, ".*span[@class='QewHA H4 _a']")
            .text.replace("\n", " ")
        )
        print("Review = {}".format(review))

        csvWriter.writerow(
            [
                name,
                date,
                rating,
                title,
                review,
            ]
        )
    except:
        print("problem with the comment!")

```

```
    pass

try:
    # change the page
    next_button = driver.find_element(
        By.XPATH, '//*[@class="ui_button nav next primary "]'
    )
    next_button.click()
except:
    print("No next page button!")
    break

driver.close()
```