ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αξιολόγηση της Ποιότητας Ξενοδοχειακών Υπηρεσιών με Αλγορίθμους Ανάλυσης Συναισθήματος

ΦΟΙΤΗΤΕΣ:

Μερεντίτης Δημήτριος cs 141042

Νικολακέας Θεόδωρος cs 141017

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

Μπουσδέκης Αλέξανδρος

Εγκρίθηκε από την κάτωθι τριμελή επιτροπή

Αθανάσιος Βουλόδημος

Athanasios Voulodimos
Digitally signed by Athanasios Voulodimos
Date: 2021.11.23 18:37:51 +02'00'

Γεώργιος Μπαρδής

Georgios Bardis
Digitally signed by Georgios Bardis
Date: 2021.11.23 19:12:38 +02'00'

Αλέξανδρος Μπουσδέκης

Alexandros Bousdekis
Digitally signed by Alexandros Bousdekis
Date: 2021.11.22 14:20:27 +02'00'

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της διπλωματικής μας εργασίας, θα θέλαμε να εκφράσουμε τις θερμές μαςευχαριστίες σε όλους όσους συνέβαλλαν στην εκπόνησή της.

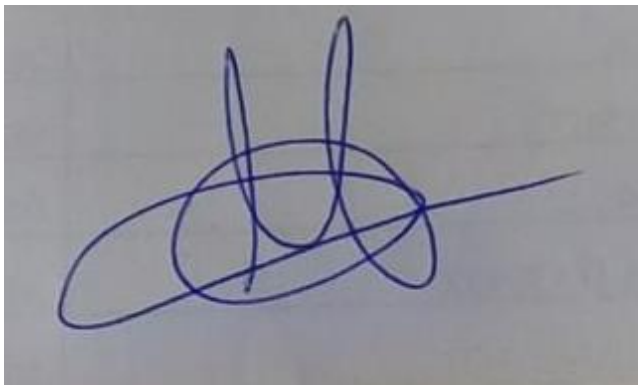
Ευχαριστούμε θερμά τον επιβλέπων καθηγητή μας, κύριο Αλέξανδρο Μπουσδέκη, για την εμπιστοσύνη που μας έδειξε εξ' αρχής, αναθέτοντάς μας το συγκεκριμένο θέμα,την επιστημονική του καθοδήγηση, τις υποδείξεις του, την επιμονή του, το αμείωτο ενδιαφέρον του, τη συμπαράστασή του, τη συνεχή του υποστήριξη από την αρχή μέχρι το τέλος.

Επιπλέον,θα θέλαμε να εκφράσουμε την ευγνωμοσύνη μας στις οικογενείες μας για όλη τη στήριξη, τη συμπαράσταση και την κατανόησή τους, καθ' όλη τη διάρκεια των σπουδών μας.

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η **ΜΕΡΕΝΤΙΤΗΣ ΔΗΜΗΤΡΙΟΣ** του **ΑΝΔΡΕΑ** , με αριθμό μητρώου **141042** φοιτητής/τρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής **ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ**. του Τμήματος **ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ** , δηλώνω υπεύθυνα ότι: «Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

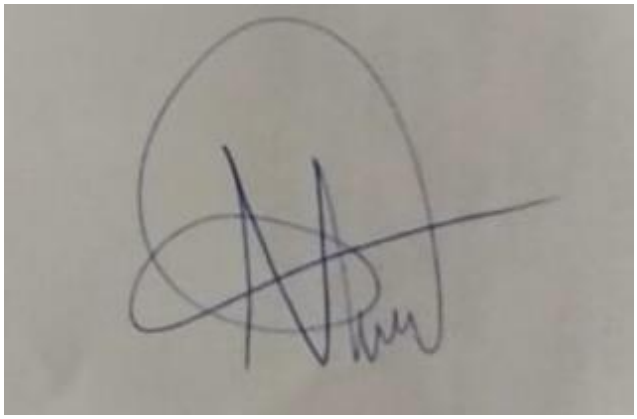
Ο/Η Δηλών



ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένος/η **ΝΙΚΟΛΑΚΕΑΣ ΘΕΟΔΩΡΟΣ** του **ΠΑΝΑΓΙΩΤΗ** , με αριθμό μητρώου **141017** φοιτητής/τρια του Πανεπιστημίου Δυτικής Αττικής της Σχολής **ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ**. του Τμήματος **ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ** , δηλώνω υπεύθυνα ότι: «Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος. Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου».

Ο/Η Δηλών



Περίληψη

Τα σημερινά ξενοδοχεία δέχονται τις κρατήσεις τους μέσα από διάφορες ιστοσελίδες όπου οι πελάτες παραθέτουν τη γνώμη τους για τις υπηρεσίες που προσφέρουν και έτσι κάνουν τις ιστοσελίδες να λειτουργούν επίσης ως πολύτιμή πηγή πληροφοριών. Μεταξύ αυτών των ιστοσελίδων, το **Trivago** είναι ένας από τους πιο δημοφιλείς , όπου εκατόμμμύρια χρήστες παραθέτουν τη γνώμη τους και αλληλεπιδρούνκαθημερινά. Τα σχόλια καθορίζουν την κοινή γνώμη / το συναίσθημά όσον αφορά μία υπηρεσία, μία τοποθεσία κ.λπ.

Η παρακολούθηση και η ανάλυση αυτών των σχολίων δίνουν πολύτιμή ανατροφοδότηση (feedback) στους χρήστες. Λόγω του μμεγάλου μμεγέθους αυτών των δεδομένων, η ανάλυση συναισθήματος επιλέγεται ως μια τεχνική για την ανάλυση αυτών των δεδομένων, λόγω της ευκολίας στον προσδιορισμό των απόψεων των χρηστών χωρίς να χρειαστεί να ελεγχθούνεκατόμμμύριασχόλια με το χέρι.

Σε αυτή τη διπλωματική , ασχοληθήκαμε με την ανάλυση συναισθήματος με τη μέθοδο του λεξικού.

Για τη διεξαγωγή της , χρησιμοποιήθηκε έναDatasetμε 10276 σχόλια από αναρτήσεις που έχουν γίνει στον ισότοπο **Trivago** καθώς επίσης και κάποιες εντολές και βιβλιοθήκες της γλώσσας Python.

Την ανάλυση συναισθήματος σαν πρόβλημα μπορούμε να τη προσεγγίσουμε με διαφορετικούς τρόπους ωστόσο στη παρούσα εργασία αφού ταξινομήσαμε τα σχόλια σε θετικά και αρνητικά , με τη βοήθεια της μεθόδου του λεξικού τα χωρίσαμε σε βαθμίδες από το 1 έως το 5 ανάλογα το πόσο θετικά η αρνητικά είναι.

Η ανάλυση συναισθημάτων με βάση το λεξικό είναι μια υπολογιστική προσέγγιση για τη μέτρηση του συναισθήματος που μεταφέρει ένα κείμενο στον αναγνώστη μέσα από διάφορες λέξεις κλειδιά.

Abstract

Today's hotels accept their reservations through various websites where customers quote their opinion about the services they offer and that make the websites also function as a valuable source of information. Among these websites, Trivago is one of he most popular, where millions of users comment and interact on a daily basis.

Comments determine public opinion / awareness regarding a service, a location, etc..Monitoring and analyzing these comments provides valuable feedback to users. Because of the staff of this data, emotion analysis was chosen as a technique for analyzing this data, because of the ease in determining users' views without having to check millions of comments in hand.

In this dissertation, we dealt with the analysis of emotion using the dictionary method.

To do this, we used a Dataset with 10,276 comments from posts made on the Trivago site as well as some Python commands and libraries.

The analysis of emotion as a problem can be approached in different ways, however in the present work, after we classified the comments into positive and negative, with the help of the dictionary method we divided them into grades from 1 to 5 depending on how positive or negative they are.

Dictionary-based emotion analysis is a computational approach to measuring the emotion that a text conveys to the reader through various keywords

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1 - Εισαγωγή

1.1 Γενικά

Η **Ανάλυση Συναισθήματος (Sentiment Analysis)** είναι μια από τις τεχνικές Επεξεργασίας της Φυσικής Γλώσσας (Natural Language Processing) που καθορίζει αν ένα κομμάτι γραφής (κείμενο) είναι θετικό, αρνητικό ή ουδέτερο.

Βοηθά επίσης στην άντληση της γνώμη ή της στάσης ενός ατόμου και ως εκ τούτου είναι επίσης γνωστή ως εξόρυξη γνώμης (opinion mining).

Το συναίσθημά (sentiment) χρησιμοποιείται συνήθως για να εκφράσει πώς οι άνθρωποι αισθάνονται για ένα θέμα.

Η κατηγοριοποίηση ενός κειμένου σε θετικό ή αρνητικό είναι μια φυσική διαδικασία για τον άνθρωπο που μπορεί να πραγματοποιηθεί χειροκίνητα, όμως δεν είναι ικανή να δουλέψει αποτελεσματικά για τεράστιο όγκο δεδομένων που αντλούνται από τον παγκόσμιο ιστό. Έτσι, η ανάπτυξη συστημάτων αυτόματης κατηγοριοποίησης των δεδομένων και ακόμα περισσότερο η αναγνώριση του σημασιολογικούπεριεχομένου των δεδομένων, είναι πιο αναγκαία από ποτέ.

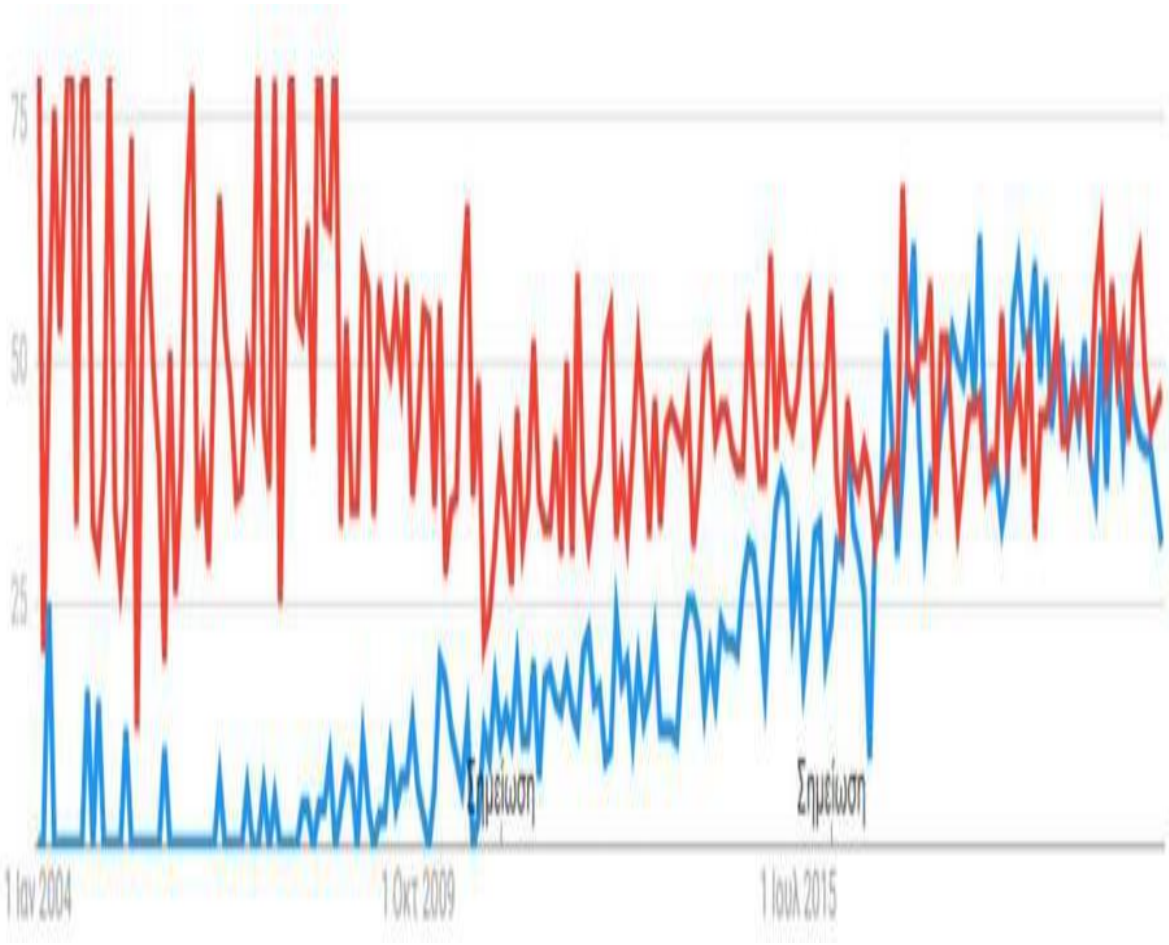
Για αυτόν τον λόγο έχουν αναπτυχθεί πολλές προσεγγίσεις στο πρόβλημα της αυτόματης εξαγωγής συναισθημάτων, όπως είναι η προσέγγιση βασισμένη στη μμηχανική μάθηση (machine learning-based), η προσέγγιση βασισμένη σε λεξικό (lexicon-based) και η υβριδική (hybrid) προσέγγιση.

1.2 Ιστορική Αναδρομή

Οι ρίζες της ανάλυσης συναισθημάτων στις μελέτες σχετικά με την ανάλυση της κοινής γνώμης εμφανίζονται στις αρχές του 20ου αιώνα και στην ανάλυση υποκειμενικότητας του κειμένου που διενεργήθηκε από την Κοινότητα της Γλωσσολογίας στης δεκαετία του 1990. Ωστόσο, το ξέσπασμα της ανάλυσης συναισθήματος που βασίζεται σε υπολογιστή προέκυψε μόνο με τη διαθεσιμότητα υποκειμενικών κειμένων στο διαδίκτυο. Κατά συνέπεια, το 99% των δημοσιεύσεων δημοσιεύθηκε μετά το 2004. Οι δημοσιεύσεις που αφορούν ανάλυση συναισθήματος είναι διάσπαρτες σε πολλαπλούς τόπους δημοσιεύσεων, και ο συνολικός αριθμός δημοσιεύσεων στους 15 κορυφαίους τόπους αντιπροσωπεύει μόνο περίπου το 30% των δημοσιεύσεων συνολικά. Τα τελευταία χρόνια, η ανάλυση συναισθήματος έχει μετατοπιστεί από την ανάλυση των διαδικτυακών κριτικών προϊόντων σε κείμενα κοινωνικής δικτύωσης από το Twitter και το Facebook. Πολλά θέματα πέρα από τις χρηματιστηριακές αγορές, οι εκλογές, οι καταστροφές, η ιατρική, η μηχανική λογισμικού και η διαδικτυακής παρενόχλησης (cyberbullying) επεκτείνουν την αξιολόγηση του συναισθήματος. Έχει παρατηρηθεί μια τεράστια αύξηση του αριθμού των δημοσιεύσεων που επικεντρώνονται στην ανάλυση συναισθημάτων και στην εξόρυξη απόψεων κατά τα τελευταία έτη.

Σύμφωνα με τα δεδομένα, σχεδόν 7.000 έγγραφα αυτού του θέματος έχουν καταχωρηθεί και, το πιο ενδιαφέρον, το 99% των άρθρων έχουν εμφανιστεί μετά το 2004. Η ανάλυση συναισθήματος αποτελεί έναν από τους ταχύτερα αναπτυσσόμενους τομείς έρευνας. Ο Πίνακας 5 δείχνει την αύξηση των αναζητήσεων που έγιναν με μια συμβολοσειρά αναζήτησης "ανάλυση συναισθήματος" και "σχόλια πελατών" στις Ηνωμένες Πολιτείες Αμερικής στη μηχανή Google Search. Παρατηρήθηκε ότι οι πρώτες ακαδημαϊκές μελέτες που μετρούν τις δημόσιες δηλώσεις είναι κατά τη διάρκεια και μετά το 2ο Παγκόσμιο Πόλεμο και τα κίνητρά τους είναι άκρως πολιτικής φύσεως. Η σύγχρονης ανάλυση συναισθήματος συνέβη μόνο στα μέσα της δεκαετίας του 2000, και επικεντρώθηκε στις κριτικές προϊόντων που διατίθενται στο διαδίκτυο, π.χ. IMdb, Yelp. Έκτοτε, η χρήση της ανάλυσης συναισθημάτων έχει φθάσει σε πολλούς άλλους τομείς, όπως η πρόβλεψη των χρηματοπιστωτικών αγορών και οι αντιδράσεις σε τρομοκρατικές επιθέσεις. Επιπλέον, η έρευνα για την ανάλυση συναισθήματος και η επεξεργασία της φυσικής γλώσσας έχει αντιμετωπίσει πολλά προβλήματα που συμβάλλουν στη δυνατότητα εφαρμογής της ανάλυσης συναισθημάτων για την ανίχνευση ειρωνείας και την πολύγλωσση υποστήριξη. Επιπροσθέτως, όσον αφορά τα συναισθήματα, οι προσπάθειες προχωρούν από την απλή ανίχνευση της Πόλωσης σε πιο σύνθετες αποχρώσεις των συναισθημάτων και τη διαφοροποίηση των αρνητικών συναισθημάτων όπως ο θυμός και η θλίψη (Mäntylä, etal., 2016) . Μελέτη Δεδομένων Κοινωνικών Δικτύων

Δεδομένα που δείχνουν τη σχετική δημοτικότητα των ερωτημάτων αναζήτησης "ανάλυση συναισθήματος" και "σχόλια πελατών" στις Ηνωμένες Πολιτείες Αμερικής



1.3 Τι είναι το Sentiment Analysis

Το διαδίκτυο αλλά και η πραγματική ζωή βρίθουν από όλων των ειδών δεδομένα, ασύντακτα και συντεταγμένα, δεδομένα που δεν έχουν μετρηθεί η συνδυαστεί και που μπορεί να ανήκουν ή και μην ανήκουν σε μια δομή. Ας δώσουμε όμως μερικά παραδείγματα χρήσης μεγάλου όγκου δεδομένων και μετα-δεδομένων. Υπάρχουν δεδομένα που αν συνδυαστούν περιγράφουν συστήματα όπως λόγου χάρη οι δείκτες των μετρήσεων λειτουργίας και βλάβης κάποιων μηχανών σε ένα εργοστάσιο. Τα στοιχεία που διαβάζουμε στο log των αυτών των μηχανών μπορεί να μην μας λένε πολλά όμως, αν γνωρίζουμε όμως με ποιο τρόπο να τα συνδυάσουμε και ταυτόχρονα πως να μελετήσουμε παλαιότερα, μπορούμε να δημιουργήσουμε μια εφαρμογή που να προβλέπει τον πιθανό χρόνο μια μελλοντικής βλάβης ή την ενεργειακή κατανάλωση σε βάθος χρόνου και ούτω καθεξής. Ας μεταβούμε σε ένα άλλο πεδίο όπου η χρήση μοντέλων πρόβλεψης είναι απαραίτητη.

Στο διαδίκτυο υπάρχει μια τεράστια ζήτηση για πληροφορίες και κατά συνέπεια μεγάλη προσφορά. Οι ειδήσεις και ειδικότερα το γραπτό κείμενο ή φυσική γλώσσα είναι ένα μεγάλο κομμάτι αυτών των πληροφοριών. Συλλέγοντας μεγάλο όγκο αυτών των δεδομένων μπορούμε να κατασκευάσουμε ένα προγνωστικό μοντέλο που δύναται να αποφασίσει αν ένα κείμενο είναι θετικό, ευχάριστο, ενθουσιώδες ή αντίθετα, αρνητικό, απαισιόδοξο ή λυπητερό. Έχουμε λοιπόν τη δυνατότητα να εξάγουμε το συναίσθημα ενός κειμένου. Η ικανότητα να προβλέπουμε αυτόματα με σχετικά μεγάλο βαθμό αξιοπιστίας το συναίσθημα έχει προεκτάσεις και χρήση σε πολλούς κλάδους της οικονομικής και πολιτικής ζωής. Στα επόμενα κεφάλαια θα ασχοληθούμε με τα επιμέρους κομμάτια ενός τέτοιου μοντέλου πρόβλεψης το οποίο εν τέλει θα κατασκευάσουμε ως μέρος ενός προγράμματος που θα ασχολείται με την ανάλυση συναισθήματος συγκεκριμένων κριτικών που γίνονται σε ξενοδοχεία.

1.4 Αντικείμενο και Στόχοι της Διπλωματικής Εργασίας

Η συγκεκριμένη διπλωματική εργασία ασχολείται με τη συλλογή συγκεκριμένων δεδομένων από ένα μεγάλο Dataset - το οποίο περιέχει κάτι παραπάνω από 10.000 αξιολογήσεις- την εξόρυξη γνώσης από αυτό και ειδικότερα την κατάλληλη επεξεργασία του για την εξαγωγή στοχευμένων αποτελεσμάτων και συμπερασμάτων. Στο πλαίσιο της εκπόνησης της Διπλωματικής εργασίας, χρησιμοποιήθηκαν κάποιες τεχνικές συλλογής δεδομένων από το Dataset , κάποιες συγκεκριμένες βιβλιοθήκες της Pythonαλλά και βασικοί αλγόριθμοι.

Πιο συγκεκριμένα ο κύριος στόχος της έρευνας είναι να γίνει:

- Σωστή αξιολόγηση των χρηστών που λαμβάνουμε από το datasetγια τα διάφορα ξενοδοχεία
- Ορθή ταξινόμηση των δεδομένων σε δικό μας Dataset με βάση της πληροφορίες που λαμβάνουμε στο
- Εξαγωγή αποτελεσμάτων και παρουσίαση συμπερασμάτων ανάλογα το συναίσθημα.

1.5 Διάρθρωση της Διπλωματικής Εργασίας

Στο πρώτο κεφάλαιο προσδιορίσαμε την έννοια του Sentiment Analysis , και το αντικείμενο της διπλωματικής εργασίας .Υπάρχει επίσης η περίληψη και μια ιστορική αναδρομή του Sentiment Analysis.

Στη συνέχεια, στο πρώτο μέρος του Δεύτερου κεφαλαίου γίνεται ανάλυση των πλεονεκτημάτων του Sentiment Analysis γενικά αλλά και για τους οργανισμούς.Ύστερα αναλύεται η χρήση του Sentiment Analysis στους οργανισμούς αλλά και γενικά.Στη συνέχεια , βλέπουμε τις τεχνικές κατηγοριοποίησης συναισθήματος και τις μεθόδους.Τέλος , δίνεται έμφαση στην ανάλυση συναισθήματος με τεχνικές βασισμένες σε λεξικό αλλά και με τεχνικές βασισμένες στη μηχανική μάθηση.

Στο κεφάλαιο Τρία γίνεται η αξιολόγηση της ποιότητας των ξενοδοχειακών υπηρεσιών.Ξεκινάει με τη προ επεξεργασία των δεδομένων και ύστερα την εξαγωγή δεδομένων.Ύστερα αναλύεται το πως γίνεται η ανάλυση συναισθήματος με υπηρεσίες Fuzzy String Matching.

Στο Τέταρτο κεφάλαιο υπάρχει η τεχνική υλοποίση αλλά και η ανάλυση των βιβλιοθηκών Python που χρησιμοποιήθηκαν.Αρχικά , εξηγείται γιατί και που χρησιμοποιήθηκαν οι συγκεκριμένες βιβλιοθήκες και ύστερα βλέπουμε κομμάτια κώδικα που βοηθαέι στη κατανόηση της τεχνικής υλοποίησης του προγράμματος μας.

Στο Πέμπτο κεφάλαιο προσδιορίσαμε το συμπέρασμα της διπλωματικής εργασίας και τη μελλοντική εργασία.Τέλος παρουσιαζεται η βιβλιογραφία απο όπου αντλήσαμε τις πληροφορίες για τη διπλωματική μας εργασία.

ΚΕΦΑΛΑΙΟ 2 - Ανάλυση Συναισθήματος

2.1 Ανάλυση Συναισθήματος

Η ανάλυση συναισθημάτων ή αλλιώς εξόρυξη γνώμης, αναφέρεται στη χρήση και επεξεργασίας της φυσικής γλώσσας, της ανάλυσης κειμένων, της αξιοποίησης της υπολογιστικής γλωσσολογίας και της βιομετρίας για τον συστηματικό εντοπισμό, την εξαγωγή, την ποσοτικοποίηση και τη μελέτη των συναισθηματικών καταστάσεων και των υποκειμενικών πληροφοριών που χαρακτηρίζουν ένα άτομο ή μια ομάδα ατόμων. Η ανάλυση συναισθήματος εφαρμόζεται ευρέως με την αξιοποίηση της στάσης ενός πελάτη, όπως αποτυπώνεται σε κριτικές και σε απαντήσεις ερευνών, με την αξιοποίηση των onlineκαι των κοινωνικών μέσων ενημέρωσης και δικτύωσης.

Σε γενικές γραμμές, η ανάλυση συναισθημάτων έχει ως στόχο να καθορίσει τη στάση ενός ομιλητή ή άλλου υποκειμένου σε σχέση με κάποιο θέμα ή να καθορίσει τη συνολική συναισθηματική αντίδραση ενός ατόμου σε ένα ερέθισμα με το οποίο βρίσκεται ή μπορεί να βρεθεί σε αλληλεπίδραση. Η ανάλυση συναισθήματος ερευνά την απλή συμπεριφορά ενός ατόμου και την κρίση του έως και την επιδιωκόμενη από μια επιχείρηση συμπεριφορά του ατόμου.

Η τεχνική της ανάλυσης συναισθημάτων βρίσκει ευρεία εφαρμογή στα μέσα κοινωνικής δικτύωσης και σε πλατφόρμες όπως είναι η booking και η trivagoστις οποίες εκφράζονται κριτικές και βαθμολογίες βάση των εμπειριών των ατόμων σχετικά με τις διακοπές τους. Τα σχόλια των πελατών της booking αποτελούν μια τεράστια βάση δεδομένων στην οποία υπάρχει αχανής πληροφόρηση και πληθώρα δεδομένων που δύναται να αξιοποιηθούν μέσω της ανάλυσης συναισθημάτων. Η ποιότητα της εξυπηρέτησης των πελατών μπορεί με αυτόν τον τρόπο να βελτιωθεί σε μεγάλο ποσοστό προς όφελος τόσο αυτών που διεξάγουν την έρευνα όσο και των πελάτων. Η ανάλυση συναισθημάτων έχει εφαρμογή σε έγγραφα και άρθρα στα οποία εξετάζεται η πολικότητα, δηλαδή σε τι βαθμό υπάρχει θετική, αρνητική ή πιο ουδέτερη εκφρασμένη γνώμη που μπορεί να επηρεάσει τον αναγνώστη.

Πέραν του εντοπισμού της πολικότητας ενός άρθρου και εγγράφου, άλλες τεχνικές ανάλυσης συναισθημάτων επικεντρώνουν το ενδιαφέρον τους στην ανάλυση συναισθηματικών καταστάσεων όπως είναι η στενοχώρια, το άγχος, η χαρά, ο θυμός κλπ. Ωστόσο επισημαίνεται ότι η ανάλυση συναισθημάτων έχει κατά βάση στατικό χαρακτήρα και δεν αποτελεί δυναμική εκτίμηση, αφού σε περίπτωση που ληφθούν υπόψη σχόλια και στάσεις καταναλωτών για μια συγκεκριμένη υπηρεσία που έλαβαν, τότε θα διαμορφωθούν συμπεράσματα που θα αφορούν τη δεδομένη υπηρεσία και χρονική στιγμή. Για το λόγο αυτό πρέπει η ανάλυση συναισθημάτων να συνδυάζει ένα εύρος σχολίων, συναισθηματικών καταστάσεων και στάσεων διαφόρων υποκειμένων για να εξαχθεί πιο ασφαλές συμπέρασμα, ιδιαίτερα για μια επιχείρηση που επιζητάς την κατανόηση της συναισθηματικής κατάστασης πελάτων και υποψηφίων πελάτων της.

Όσον αφορά τις τεχνικές που αφορούν την ανάλυση συναισθημάτων με επιστροφή αποτελεσμάτων υπό την μορφή κλίμακας και όχι με την μορφή Ναι / Όχι, κρίνεται ότι η διαδικασία συσχετίσεων ενδείκνυται για να καταστεί εφικτή η ακριβής μέτρηση, αφού εξετάζεται σε τι βαθμό οι προβλεπόμενες τιμές βρίσκονται κοντά με τα επιδιωκόμενα αποτελέσματα. Η ανάπτυξη των κοινωνικών μέσων δικτύωσης, η δυνατότητα ελεύθερης έκφρασης στο διαδίκτυο και οπουδήποτε στον Παγκόσμιο Ιστό σχολίων πελάτων επιχειρήσεων, καταναλωτών, απλών πολιτών για οποιοδήποτε ζήτημα αφορά άμεσα η έμμεσα την καθημερινότητα τους, την επαγγελματική τους ζωή κλπ. Το γεγονός ότι δεν απαιτούνται ιδιαίτερες γνώσεις για την αλληλεπίδραση χρηστών, αλλά και για την έκφραση σχολίων σε sitesή σε blogs, διευκολύνει όχι μόνο τους χρήστες αλλά και τις επιχειρήσεις και τους ερευνητές, οι οποίοι μπορούν να χρησιμοποιήσουν τις γνώμες που εκφράζονται σε ιστοτόπους και εν γένει στο διαδίκτυο, για να τοποθετήσουν με πιο αποτελεσματικό τρόπο τα προϊόντα τους και να τροποποιήσουν τις ενέργειες προβολής και προώθησης των προϊόντων τους.

Η τάση για ανάλυση των συναισθημάτων είναι αναντίρρητη στον κόσμο των επιχειρήσεων, αφού πλέον θεωρείται πολύτιμη η πληροφορία που αποσπάται από τα γενικά σχόλια απλών χρηστών του διαδικτύου. Απαντώνται όμως αρκετές δυσκολίες για όσους επιδιώκουν να εξορύξουν δεδομένα και γνώση από σχόλια. Ο υπολογισμός που πραγματοποιείται μετά την σωστή παραμετροποίηση και μετασχηματισμό των δεδομένων από ένα υπολογιστικό σύστημα, ενέχει τον κίνδυνο να εξάγει συμπεράσματα που δεν μπορούν να συμπεριλάβουν ιδιαιτερότητες που διακρίνουν την ανθρώπινη φύση και τους ανθρώπους άλλου φύλλου, άλλης εθνικότητας και άλλης κουλτούρας. Εμπόδια προκύπτουν από την πληθώρα γλωσσών και με τον τρόπο που μεταφράζεται ένα σχόλιο από μια γλώσσα σε μια άλλη, πολιτιστικές και πολιτισμικές ιδιαιτερότητες μπορούν να επηρεάσουν σε σημαντικό βαθμό τον τρόπο γραφής ενός κειμένου και εγγράφου, άρα η εξόρυξη αξιόπιστης γνώμης γίνεται ακόμα πιο δυσχερής. Προς επίρρωση των ανωτέρω, αξίζει να αναφερθεί ότι υπάρχουν περιπτώσεις κατά τις οποίες ακόμα και άνθρωποι έρχονται σε διαφωνία για το ύφος ενός κειμένου, επομένως πρέπει να εντοπίζονται και να χρησιμοποιούνται κείμενα και σχόλια μεγάλου μήκους και με καθαρά εκφρασμένη άποψη για την βελτίωση της αξιοπιστίας των ευρημάτων.

2.2 Πλεονεκτήματα της ανάλυσης συναισθήματος

Τα συστήματα ανάλυσης συναισθημάτων επιτρέπουν στις εταιρείες να κάνουν αίσθηση αυτής της θάλασσας μη δομημένου κειμένου αυτοματοποιώντας τις επιχειρηματικές διαδικασίες, δημιουργώντας χρήσιμες πληροφορίες και εξοικονομώντας ώρες μη αυτόματης επεξεργασίας δεδομένων, με άλλα λόγια, καθιστώντας τις ομάδες πιο αποτελεσματικές. Μερικά από τα πλεονεκτήματα της ανάλυσης συναισθήματος είναι τα εξής:

1) **Δυνατότητα κλιμάκωσης:** Η ανάλυση συναισθημάτων επιτρέπει την επεξεργασία δεδομένων σε κλίμακα με αποδοτικό και οικονομικά αποδοτικό τρόπο.

2) **Ανάλυση σε πραγματικό χρόνο:** Η ανάλυση συναισθημάτων μπορεί να χρησιμοποιηθεί για τον εντοπισμό κρίσιμων πληροφοριών που επιτρέπουν την ευαισθητοποίηση της κατάστασης κατά τη διάρκεια συγκεκριμένων σεναρίων σε πραγματικό χρόνο. Υπάρχει κρίση δημοσίων σχέσεων στα μέσα κοινωνικής δικτύωσης που θα εκραγεί; Ένας θυμωμένος πελάτης που πρόκειται να δημιουργήσει προβλήματα; Ένα σύστημα ανάλυσηςσυναισθήματος μπορεί να μας βοηθήσει να εντοπίσουμε αμέσως αυτά τα είδη καταστάσεων και να αναλάβουν δράση.

3) **Σαφή κριτήρια:** Οι άνθρωποι δεν τηρούν σαφή κριτήρια για την αξιολόγηση του αισθήματος ενός κειμένου. Εκτιμάται ότι διαφορετικοί άνθρωποι συμφωνούν μόνο σε ποσοστό 60- 65% όταν κρίνουν το συναίσθημα για ένα συγκεκριμένο κομμάτι κειμένου. Είναι ένα υποκειμενικό έργο που επηρεάζεται σε μεγάλο βαθμό από προσωπικές εμπειρίες, σκέψεις και πεποιθήσεις.

Χρησιμοποιώντας ένα συγκεντρωτικό σύστημα ανάλυσης συναισθημάτων, οι εταιρείες μπορούν να εφαρμόσουν τα ίδια κριτήρια σε όλα τα δεδομένα τους. Αυτό συμβάλλει στη μείωση των σφαλμάτων και στη βελτίωση της συνέπειας των δεδομένων. Επιπλέον, η ανάλυση συναισθήματος είναι χρήσιμη στην παρακολούθηση των μέσων κοινωνικής δικτύωσης, επειδή βοηθά να πραγματοποιηθούν τα παρακάτω: Προτεραιότητα δράσης: Η ανάλυση συναισθήματος επιτρέπει το εύκολο φιλτράρισμα στις μη αναγνωσμένες αναφορές με θετικά ή και αρνητικά σχόλια, δίνοντας την ευχέρεια για την ιεράρχηση των δράσεων που πρέπει να γίνουν. Συντονισμός σε μια συγκεκριμένη χρονική στιγμή – δηλαδή εάν υπάρχει προβάδισμα στην καμπάνια ενός προϊόντος ή την ημέρα που το αντίστοιχο προϊόν ενός ανταγωνιστή είχε πτώση σε επίπεδο δημόσιων αναφορών. Ανάλυση του ανταγωνιστή. Η χρήση του

εργαλείου της συναισθηματικής ανάλυσης βοηθά στο να παρατηρηθεί εάν υπάρχει μια αρνητική απόκριση σε μια συγκεκριμένη λειτουργία του νέου προϊόντος του ανταγωνιστή.

Όσον αφορά την παρακολούθηση επωνυμίας, η ανάλυση συναισθημάτων είναι χρήσιμη επειδή βοηθά στην κατανόηση του πώς η φήμη της επωνυμίας σας εξελίσσεται με την πάροδο του χρόνου.Στην εξερεύνηση του ανταγωνισμού και πώς η φήμη του εξελίσσεται με την πάροδο του χρόνου.Στον εντοπισμό πιθανών κρίσεων δημοσίων σχέσεων. Δίνοντας προτεραιότητα στις πυρκαγιές που πρέπει να τεθούν αμέσως σε έλεγχο και τις αναφορές που μπορούν να περιμένουν. Στον συντονισμό σε μια συγκεκριμένη χρονική στιγμή. Ελέγχοντας τις αναφορές μια συγκεκριμένης ημέρας ή ενός μόνο προϊόντος.

2.2.2 Πλεονεκτήματα της ανάλυσης συναισθήματος στους οργανισμούς

Τα οφέλη της ανάλυσης συναισθημάτων και η χρήση της από τους ιδιοκτήτες επιχειρήσεων τους βοηθούν να αποκτήσουν πλεονέκτημα έναντι των ανταγωνιστών τους.Όροι όπως «εξόρυξη απόψεων» και «ταυτοποίηση κειμένου» περιγράφουν συχνά το νόημα της ανάλυσης συναισθημάτων ως μια κατάλληλη μέθοδο που χρησιμοποιείται από τους εμπόρους για την αναγνώριση των προτιμήσεων των πελατών.Τα δεδομένα που συλλέγονται από τις απαντήσεις των πελατών, όπως tweets, σχόλια, σχόλια και οποιαδήποτε γραφή που σχετίζεται με προϊόντα ή υπηρεσίες μελετώνται και αυτή η διαδικασία ονομάζεται ανάλυση συναισθημάτων.Οι έμποροι και οι οργανώσεις συνεχίζουν αυτήν τη διαδικασία για να παραμείνουν σχετικοί στον ανταγωνιστικό τομέα και να βρουν έναν κατάλληλο τρόπο για να προωθήσουν την επιχείρησή τους.Καθώς τα μέσα κοινωνικής δικτύωσης παίζουν σημαντικό ρόλο στη ζωή σχεδόν κάθε ανθρώπου, οι περισσότεροι οργανισμοί το χρησιμοποιούν ως μέσο επικοινωνίας με τους πελάτες τους ή άλλες εταιρείες.Ενώ οι εταιρείες χρησιμοποιούν τα μέσα κοινωνικής δικτύωσης ως εργαλείο για να αλληλεπιδρούν με τους πελάτες τους, τα σχόλια και τα σχόλια που προέρχονται από τους πελάτες αναλύονται και χρησιμοποιούνται από τις εταιρείες ως οδηγό για τη βελτίωση των προϊόντων και των υπηρεσιών τους.

Παρακάτω παρατίθενται μερικά από τα πλεονεκτήματα της ανάλυσης συναισθημάτων και μια σύντομη εξήγηση για το πώς η ανάλυση συναισθημάτων βοηθά σε έναν οργανισμό.

1)Βελτίωση της εξυπηρέτησης πελατών

Ένα από τα οφέλη της ανάλυσης συναισθημάτων είναι η δυνατότητα παρακολούθησης των βασικών μηνυμάτων από τις απόψεις και τις σκέψεις των πελατών.Αυτό βοηθά το τμήμα εξυπηρέτησης πελατών να έχει επίγνωση τυχόν σχετικών θεμάτων ή προβλημάτων.Καθώς η μέθοδος επιτρέπει στους οργανισμούς να κατανοούν καλύτερα τους πελάτες τους, η ανάλυση συναισθημάτων παρέχει μια σαφή εικόνα των προβλημάτων και πείθει τον οργανισμό να αναζητήσει μια λύση.Επιπλέον, έχοντας μια γρήγορη ανίχνευση ανάλυσης συναισθημάτων για δυσμενείς παρατηρήσεις πελατών, ο οργανισμός μπορεί να ενεργήσει γρήγορα διερευνώντας τη βασική αιτία και παρέχοντας στο τμήμα εξυπηρέτησης πελατών αποτελεσματική επίλυση. Τίποτα δεν αντιστοιχεί στην άμεση απάντηση στην αντιμετώπιση ενός ζητήματος από την ίδια την εταιρεία.

2. Ανάπτυξη ποιοτικών προιόντων

Το να κάνετε τους πελάτες ευτυχισμένους και να παραμείνετε πιστοί σε ένα εμπορικό σήμα είναι μια φορολογική δουλειά. Ως εκ τούτου, ένα άλλο από τα οφέλη της ανάλυσης συναισθημάτων διευκολύνει την όλη διαδικασία και ταυτόχρονα παρέχει ευκαιρίες βελτίωσης.Αυτό επιτρέπει στην ομάδα μάρκετινγκ να ερευνήσει καλύτερα τις τρέχουσες τάσεις και τις προτιμήσεις των πελατών.Οι απαντήσεις από τους πελάτες μπορούν να χρησιμοποιηθούν ως κατευθυντήρια γραμμή για τη βελτίωση της ποιότητας των υπηρεσιών, την καλύτερη μελλοντική ανάπτυξη του προϊόντος, τη μείωση της αναταραχής πελατών ή τη βελτίωση του τρόπου παρουσίασης του προϊόντος.

3. Ανακάλυψη νέων στρατηγικών marketing

Με περισσότερα δεδομένα και πληροφορίες που συλλέγονται μέσω ανάλυσης συναισθημάτων, οι οργανισμοί θα μπορούσαν να αναπτύξουν μια αποτελεσματική στρατηγική μάρκετινγκ.Το αποτέλεσμα των στρατηγικών μπορεί να μετρηθεί από τα θετικά ή αρνητικά βασικά μηνύματα των πελατών.Παρατηρώντας τις συνομιλίες των πελατών στα κοινωνικά τους μέσα και εντοπίζοντας τα συγκεκριμένα βασικά μηνύματα που σχετίζονται με την επωνυμία σας, μπορούν να σχεδιαστούν συγκεκριμένες καμπάνιες μάρκετινγκ για τους καταναλωτές

4. Βελτιώση των αντιλήψεων απο τα μέσα ενημέρωσης.

Ένα άλλο πλεονέκτημα της ανάλυσης συναισθημάτων είναι να μπορείς να παρακολουθείς την κατανόηση των δημοσιογράφων, συγγραφέων, αρθρογράφων, αναλυτών αγοράς, ερευνητών μέσων ενημέρωσης ή ανεξάρτητων συνεργατών προς την εταιρεία, είτε πρόκειται για το προϊόν, την υπηρεσία, τις αξίες της εταιρείας, το ανθρώπινο δυναμικό κ.λπ.Αυτό είναι ζωτικής σημασίας καθώς κάθε παρερμηνεία ή αρνητική χροιά μπορεί να οδηγήσει σε αρνητικά βασικά μηνύματα που σχηματίζουν μια ανεπιθύμητη αντίληψη.Γνωρίζοντας ποιος γράφει τι ιστορικά και ποιο είναι το ενδιαφέρον τους και πόσο κρίσιμο είναι για ορισμένα θέματα βοηθά το τμήμα σχέσεων με τα μέσα ενημέρωσης να συγκεντρώσει ένα κατάλληλο και ελκυστικό περιεχόμενο για αυτούς.

5. Αύξηση εσόδων από πωλήσεις

Η ανάλυση συναισθημάτων αποτυπώνει τις εντυπώσεις και τις διαθέσεις των πελατών και αυτός είναι σίγουρα ένας πολύ καλός τρόπος για να βελτιώσετε τα κέρδη από τις πωλήσεις!Καθώς εντοπίζονται αρνητικά βασικά μηνύματα και η ομάδα μάρκετινγκ κάνει τη μαγεία της για να λύσει τα προβλήματα και να βελτιστοποιήσει την ποιότητα του προϊόντος, οι οργανισμοί μπορούν να εκτιμήσουν υψηλότερη χρηματική απόδοση.Αυτό επιτυγχάνεται με τη χρήση της ανάλυσης συναισθημάτων από τη διοίκηση για τη βελτίωση των προϊόντων και των υπηρεσιών. Επιπλέον, οι πελάτες αισθάνονται ότι ακούγονται και οι ανάγκες τους φροντίζονται, βελτιώνοντας έτσι και την εικόνα μιας εταιρείας.

6. Βελτίωση της Διαχείρισης Κρίσεων

Η συχνή παρακολούθηση των απαντήσεων ή των απόψεων των πελατών απέναντι σε ένα εμπορικό σήμα θα βοηθούσε στον γρήγορο εντοπισμό τυχόν ζητημάτων. ένα από τα οφέλη της ανάλυσης συναισθημάτων.Η αποφυγή κάθε κλιμακούμενης καταγγελίας είναι ένας από τους σκοπούς της ανάλυσης συναισθημάτων, η οποία επιτρέπει την αποτελεσματική και ταχεία διαχείριση κρίσεων.Οι έγκαιρες προληπτικές ενέργειες είναι πολύ σημαντικές, καθώς βοηθούν στην εξάλειψη της κρίσης επικοινωνίας στο διαδίκτυο, η οποία θα μπορούσε εύκολα να εξαπλωθεί σε όλο το Διαδίκτυο σε λίγα λεπτά.Καθώς η ανάλυση συναισθημάτων επιτρέπει στους οργανισμούς να παρακολουθούν στενά κάθε αρνητικό νήμα ή σχόλιο στο διαδίκτυο, πιθανά ζητήματα ή κρίσεις μπορούν να αντιμετωπιστούν νωρίς

πριν από την κλιμάκωση.Όσο ενδιαφέροντα κι αν είναι αυτά τα οφέλη των αναλύσεων συναισθημάτων, οι εταιρείες θα πρέπει πρώτα να κατανοήσουν τα είδη της ανάλυσης συναισθημάτων και πού να τα εφαρμόσουν.

2.3 Χρήση της ανάλυση συναισθήματος

- Ανάλυση των δημοσιεύσεων tweets η ενός Dataset για ένα χρονικό διάστημα, για να δείτε το συναίσθημα ενός συγκεκριμένου ακροατηρίου.

- Εκτέλεση της ανάλυσης συναισθημάτων σε όλα τα μέσα κοινωνικής δικτύωσης που αναφέρονται στο brand που ενδιαφέρει και κατηγοριοποιήστε αυτόματα με επείγοντα χαρακτήρα.

- Αυτόματη ειδοποίηση της αντίστοιχης ομάδας της εταιρείας ή του οργανισμού με τις αναφορέςπου γίνονται στο διαδίκτυο γύρω από το πεδίο ενδιαφέροντος της.

- Χρήση αναλύσεων για την απόκτηση καλύτερης εικόνας για το τι συμβαίνει στα κανάλια των μέσων κοινωνικής δικτύωσης.

- Εκπαίδευση Chat bots.Μπορούμε εύκολα να αν έχουμε ένα ChatBot στον ισότοπό μας να επωφεληθεί από την ανάλυση συναισθημάτων.Αυτό γίνεται γιατί μπορούμε να εκπαιδεύσουμε να αναγνωρίζει και να ανταποκρίνεται στα συναισθήματα των πελατών.

2.3.1 Χρήση της ανάλυσης συναισθήματος στους οργανισμούς

Η χρήση της ανάλυσης συναισθημάτων ήταν μια μεγάλη βοήθεια για τους οργανισμούς να παρακολουθούν τη φήμη της μάρκας τους και να παραμένουν σε εγρήγορση για τυχόν ζητήματα που προκύπτουν.Αυτό επιτυγχάνεται με την ανάλυση των κριτικών και του τόνου που είναι ενσωματωμένα στο κείμενο και μπορούν να ανακτηθούν από διαφορετικές πλατφόρμες.

Παρακάτω παρατίθενται μερικές περιπτώσεις χρήσης της ανάλυσης συναισθημάτων και μια σύντομη εξήγηση για το πώς η ανάλυση συναισθημάτων βοηθά σε έναν οργανισμό.

1)Παρακολούθηση κοινωνικών μέσων

Ζώντας στη σύγχρονη εποχή με περισσότερα από 500.000 Tweets και 510.000 σχόλια στο Facebook που γράφονται καθημερινά, η παρακολούθηση των μέσων κοινωνικής δικτύωσης είναι η πιο συχνά χρησιμοποιούμενη μέθοδος από τις εταιρείες.Τα δεδομένα που μοιράζονται στο διαδίκτυο είναι το τεράστιο ορυχείο πληροφοριών που περιέχει τις απόψεις, τις προτιμήσεις και τα συναισθήματα των ανθρώπων σχεδόν σε κάθε πτυχή της ζωής τους.Είτε θα αγοράσετε το νέο μοντέλο smartphone είτε θα αποφασίσετε ποιο ηλεκτρονικό κατάστημα θα επισκεφθείτε, οι λεπτομέρειες της διαδικτυακής συνομιλίας παρέχουν επιπλέον εικόνα για την επιχείρηση.Σύμφωνα με στατιστικά στοιχεία, το 46% των ανθρώπων επέλεξαν να χρησιμοποιήσουν τα μέσα κοινωνικής δικτύωσης για να επεκτείνουν τα παράπονά τους στην εταιρεία που προορίζονται. Χάρη στην ανάλυση συναισθημάτων, τα παράπονα που λαμβάνονται μπορούν να εντοπιστούν, να κατηγοριοποιηθούν ανάλογα με τη σοβαρότητα της καταγγελίας και να επιλυθούν με τακτοποιημένο τρόπο.Λίγοι οργανισμοί έχουν χρησιμοποιήσει μοντέλα ανάλυσης συναισθημάτων στη σελίδα τους στα μέσα κοινωνικής δικτύωσης, σύμφωνα με τα οποία τα σχόλια επισημαίνονται αμέσως ως θετικά, αρνητικά ή ουδέτερα.

2)Έρευνα αγοράς

Η χρήση της ανάλυσης συναισθημάτων στη διεξαγωγή έρευνας αγοράς είναι απαραίτητη από την εποχή των παραδοσιακών μεθόδων.Ενώ τα προηγούμενα εργαλεία, όπως οι κάρτες σχολίων, οι έντυπες έρευνες, οι συνεντεύξεις και οι ομάδες εστίασης που υιοθετήθηκαν για τον προσδιορισμό των προτιμήσεων και των ανατροφοδοτήσεων των πελατών προς μια μάρκα, η συλλογή διαδικτυακών συναισθημάτων με συστηματική σειρά παρέχουν επίσης το ίδιο αποτέλεσμα.Η ανάλυση συναισθημάτων που πραγματοποιήθηκε κατά την ταξινόμηση των απαντήσεων των πελατών επιτρέπει στους οργανισμούς να κατανοήσουν τις σκέψεις ή τα συναισθήματά τους για το εμπορικό σήμα.

3) Υποστήριξη πελατών

Οι οργανισμοί σε όλο τον κόσμο προλαβαίνουν το γεγονός ότι το να παραμένεις ανταγωνιστικός στην απαιτητική αγορά σημαίνει να χρησιμοποιείς αναλύσεις επιχειρηματικών δεδομένων.Η βελτίωση των λειτουργιών εξυπηρέτησης πελατών είναι ζωτικής σημασίας καθώς παρέχει μια καλή εμπειρία στον πελάτη και θα παραμείνει πιστός.Οι συνομιλίες υποστήριξης και η εποικοδομητική ανατροφοδότηση

σχετικά με τις συζητήσεις στο φόρουμ είναι λίγα εργαλεία που χρησιμοποιούνται στην ανάλυση συναισθημάτων για τον εντοπισμό του τόνου του κειμένου των πελατών.Οποιαδήποτε επιχείρηση εξαρτάται σε μεγάλο βαθμό από το επίπεδο ικανοποίησης των πελατών τους, επομένως οι λεπτομέρειες που εξάγονται από την ανάλυση θεωρούνται πολύτιμες για προτάσεις και λύσεις σε οποιοδήποτε σχετικό ζήτημα.Ταυτόχρονα, η ανάλυση συναισθημάτων βοηθά επίσης στον προσδιορισμό της εμπειρίας των πελατών από το τμήμα υποστήριξης πελατών και αποφασίζει εάν τα ζητήματά τους έχουν επιλυθεί επιτυχώς ή όχι.

4)Βελτίωση της ποιότητας των προιόντων

Αλλά δεν είναι μόνο οι προσπάθειές σας μάρκετινγκ που πρέπει να προσαρμοστούν για να προσεγγίσουν το κοινό σας πιο αποτελεσματικά, τα προϊόντα και οι υπηρεσίες σας θα πρέπει επίσης να εξελιχθούν για να ικανοποιήσουν τις προτιμήσεις των πελατών σας. Θα μπορούσατε να τοποθετήσετε τους πελάτες επιτόπου με μια έρευνα, αλλά πιθανότατα θα λάβετε μόνο αρνητικές απαντήσεις από απογοητευμένους πελάτες. Όπως μια μύγα στον τοίχο, η ανάλυση συναισθημάτων σάς επιτρέπει να δείτε πώς οι καταναλωτές συζητούν ειλικρινά εσάς και τα προϊόντα σας στο διαδίκτυο, επισημαίνοντας συχνά πολύ ακριβή σημεία που μπορούν να βελτιωθούν. Και γιατί να μην είναι; Ο πελάτης έχει πάντα δίκιο, σωστά. Βελτιώστε την ποιότητα των προϊόντων σας

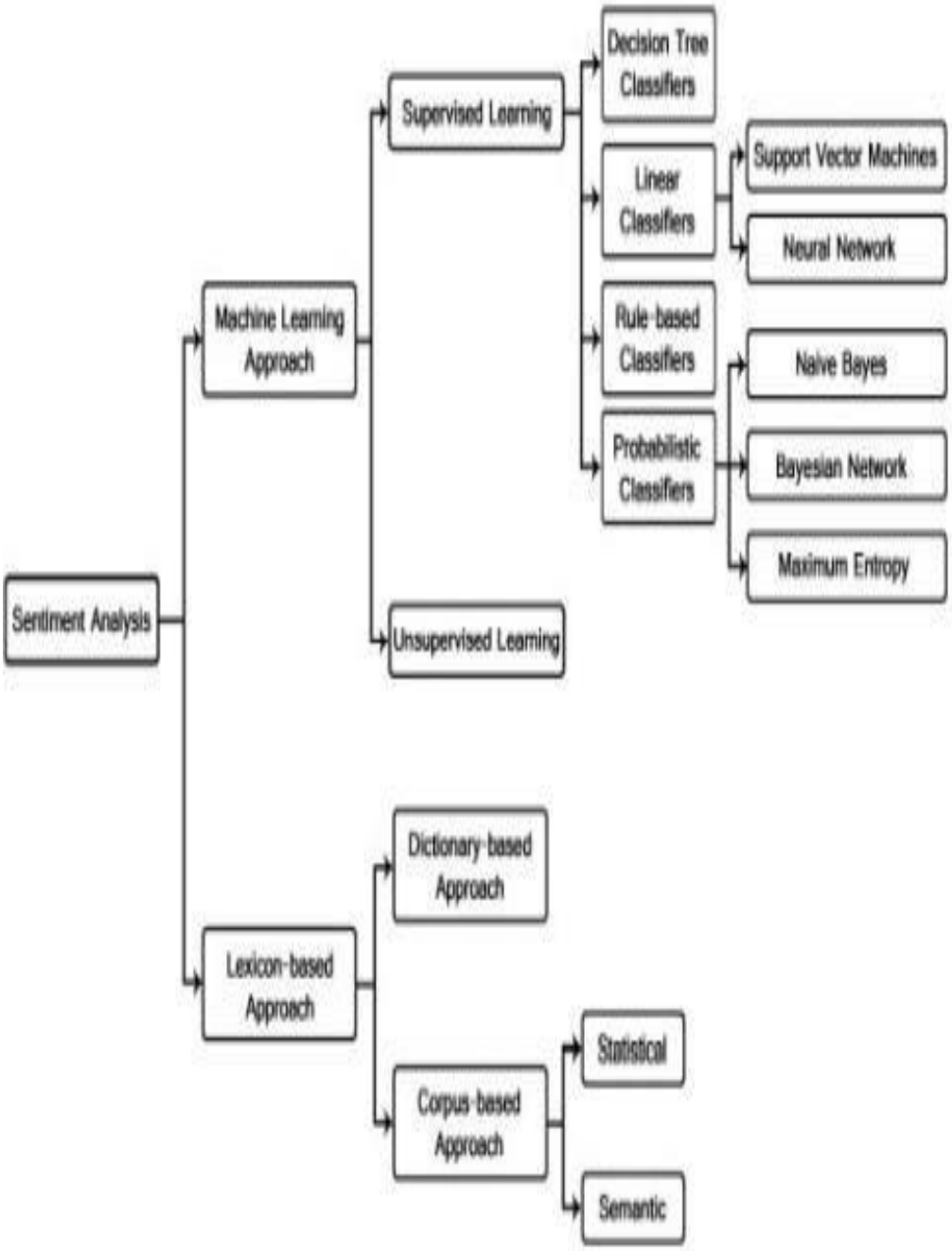
Αλλά δεν είναι μόνο οι προσπάθειές σας μάρκετινγκ που πρέπει να προσαρμοστούν για να προσεγγίσουν το κοινό σας πιο αποτελεσματικά, τα προϊόντα και οι υπηρεσίες σας θα πρέπει επίσης να εξελιχθούν για να ικανοποιήσουν τις προτιμήσεις των πελατών σας. Θα μπορούσατε να τοποθετήσετε τους πελάτες επιτόπου με μια έρευνα, αλλά πιθανότατα θα λάβετε μόνο αρνητικές απαντήσεις από απογοητευμένους πελάτες. Όπως μια μύγα στον τοίχο, η ανάλυση συναισθημάτων σάς επιτρέπει να δείτε πώς οι καταναλωτές συζητούν ειλικρινά εσάς και τα προϊόντα σας στο διαδίκτυο, επισημαίνοντας συχνά πολύ ακριβή σημεία που μπορούν να βελτιωθούν. Και γιατί να μην είναι; Ο πελάτης έχει πάντα δίκιο, σωστά.

2.4 Τεχνικές Κατηγοριοποίησης Συναισθήματος

Στο πεδίο του Sentiment Analysis χρησιμοποιούνται διάφορες τεχνικές όπου κάθε μία χρησιμοποιεί και διαφορετικό είδος αλγορίθμων ή συνδυασμό αυτών. Οι δύο μεγάλες κατηγορίες στις οποίες μπορούμε να τις χωρίσουμε είναι οι εξής:

- **Τεχνικές βασισμένες σε μηχανική μάθηση (Machine Learning based techniques)**

- **Τεχνικές βασισμένες σε λεξικό (Lexicon based techniques)**


Sentiment Analysis
Machine Learning Approach
Lexicon-based Approach
Supervised Learning
Unsupervised Learning
Decision Tree Classifiers
Linear Classifiers
Rule-based Classifiers
Probabilistic Classifiers
Support Vector Machines
Neural Network
Naive Bayes
Bayesian Network
Maximum Entropy
Dictionary-based Approach
Corpus-based Approach
Statistical
Semantic

2.4.1 Μέθοδοι ανάλυσης συναισθήματος

Οι μέθοδοι για την ανάλυση συναισθήματος κατατάσσονται σε μία απο τις παρακάτω κατηγορίες:

1)Keyword spotting

Οι μέθοδοι που ανήκουν σε αυτή τη κατηγορία χρησιμοποιούν κάποιες ενδεικτικές λέξεις για την αναγνώριση συναισθήματος που ονομάζονται affect words.Η προσέγγιση αυτή θεωρείτε απλοική γιατί βασίζεται μόνο σε λέξεις που προσδίδουν ξεκάθαρα κάποιο συναίσθημα με αποτέλεσμα να μην μπορεί να γίνει σωστή ανάλυση συναισθήματος όταν για παράδειγμα υπάρχει άρνηση μέσα στο κείμενο ή όταν το συναίσθημα εκφράζεται μέσα απο λέξεις ή φράσεις που δεν κάνουν χρήση λέξεων.

2)Lexical affinity

Στη κατηγορία αυτή γίνεται χρήση των affect words όπως και στη παραπάνω κατηγορία με τη διαφορά πως αυτή τη φορά προσδίδεται σε κάθε λέξη μεγάλη πιθανότητα να σχετίζεται με κάποιο συναίσθημα.Για παράδειγμα, η λέξη “accident” έχει 75% πιθανότητα να έχει αρνητική επίδραση.Οι μέθοδοι που ανήκουν σε αυτή τη κατηγορία έχουν καλύτερη επίδοση απο τις μεθόδους που ανήκουν στην κατηγορία keyword spotting , αλλά υπάρχουν και περιπτώσεις που αποτυγχάνουν.

3)Statistical Methods

Αυτή η προσέγγιση κάνει χρήση αλγόριθμων μηχανικής μάθησης σε συνδιασμό με κείμενα που έχουν καταταχθεί χειροκίνητα σε κάποιο συναίσθημα με σκοπό να δημιουργηθεί μία μηχανή αναγνώρισης εγγράφων.Οι μέθοδοι αυτοί λειτουργούν καλά σε μεγάλα κείμενα.

4)Concept–based approaches

Οι μέθοδοι αυτές εστιάζουν στη σημασιολογική ανάλυση του κειμένου μέσω της χρήσης σημασιολογικών δικτύων που επιτρέπουν την ομαδοποίηση νοητικών και συγκινησιακών πληροφοριών που σχετίζονται με τις απόψεις της φυσικής γλώσσας.Στόχος των μεθόδων αυτών είναι να συνταχθεί η σημασιολογική και συναι σθηματική πληροφορία που σχετίζεται με τις απόψεις της φυσικής γλώσσας.

2.5 Τεχνικές βασισμένες σε Λεξικό

Η προσέγγιση βασισμένη σε λεξικό (lexicon-based method) υποθέτει ότι ο συναισθηματικός προσανατολισμός ενός κειμένου μπορεί να συναχθεί από το συναισθηματικό προσανατολισμό των επιμέρους λέξεων και φράσεων του. Σε αντίθεση με τις μεθόδους βασισμένες σε μηχανική μάθηση, η προσέγγιση βάση λεξικού δεν απαιτεί την εκπαίδευση ενός ταξινομητή. Αντιθέτως, χρησιμοποιεί λεξικά συναισθήματος για να αποδώσει το συναίσθημα των συναισθηματικά φορτισμένων λέξεων στο κείμενο. Η απόδοση μιας μεθόδου Ανάλυσης Συναισθήματος βασισμένη σε λεξικό συνήθως καθορίζεται από τον τύπο του λεξικού συναισθήματος και από τον αλγόριθμο ανίχνευσης συναισθήματος (sentiment detection algorithms), δηλαδή από τον αλγόριθμο που χρησιμοποιείται για τον εντοπισμό των συναισθηματικά φορτισμένων λέξεων του κειμένου και τον υπολογισμό του συνολικού συναισθήματος.Οι βασισμένες σε λεξικό μέθοδοι έχουν το πλεονέκτημα της απλότητας και της ταχύτητας, στοιχεία που αποτελούν απαραίτητη προϋπόθεση ιδίως όταν απαιτείται η ανάλυση τεράστιων όγκων δεδομένων, όπως για παράδειγμα σε συλλογές που αποτελούνται από δεκάδες χιλιάδες tweets είτε ενός μεγάλου Datasetσε Excel. Από την άλλη, οι μέθοδοι αυτοί έρχονται συχνά αντιμέτωπες με δύο βασικούς περιορισμούς:

1) Οι συμβατικές μέθοδοι αδυνατούν να ανιχνεύσουν σύνθετους τύπους συναισθήματος σε tweets και Dataset , όπως για παράδειγμα την άρνηση ή το ενισχυμένο συναίσθημα. Επιπλέον εξαιτίας του γεγονότος ότι στα λεξικά συναισθήματος οι λέξεις επισημαίνονται με μια σταθερή πολικότητα, δεν λαμβάνεται υπόψιν το ευρύτερο σημασιολογικό πλαίσιο (context).

2) Τα παραδοσιακά λεξικά, όπως τα MPQA και Sentiment Word Net για την αγγλική γλώσσα, είναι σχεδιασμένα για να χειρίζονται επίσημα και καλογραμμένα κείμενα. Στο Twitter όμως, ο ιδιαίτερος και ανεπίσημος τρόπος γραφής εξαιτίας του περιορισμού των 140 χαρακτήρων, οδηγεί στον εντοπισμό ανάκτηση ελάχιστων μη-συμβατικών ή ασυνήθιστων λέξεων, αφού αυτές σπάνια είναι καταχωρημένες στα παραδοσιακά λεξικά Κατά συνέπεια, οι μέθοδοι βασισμένες σε λεξικό μπορούν να εφαρμοστούν άμεσα σε δεδομένα από το Twitter, χωρίς να χάνεται πολύτιμος χρόνος για την εκπαίδευση ταξινομητών, με το βασικό περιορισμό ότι η απόδοση και αποτελεσματικότητα όλου του συστήματος συνδέεται αναπόσπαστα με τη λεξική πηγή στην οποία βασίζεται· Μια μέθοδος βασισμένη σε λεξικό είναι τόσο καλή, όσο και το λεξικό που χρησιμοποιεί.

Τέλος, η απόδοση τέτοιων συστημάτων, σε επίπεδο ακρίβειας και χρονικής πολυπλοκότητας, επιδεινώνεται δραστικά με την εκθετική αύξηση του μεγέθους του λεξικού.Στο σημείο αυτό αξίζει να σημειωθεί ότι σε οποιαδήποτε εφαρμογή ανάλυσης κειμένου υπάρχει η δυνατότητα επιλογής μεταξύ στατιστικών ή συντακτικών τεχνικών.

Οι συντακτικές τεχνικές (syntactic techniques) μπορούν να οδηγήσουν σε καλύτερη ακρίβεια γιατί κάνουν χρήση των συντακτικών κανόνων μιας γλώσσας με σκοπό να ανιχνεύσουν τα ρήματα, τα επίθετα και τα ουσιαστικά.

Δυστυχώς αυτού του είδους οι τεχνικές είναι άρρηκτα συνδεδεμένες με τη γλώσσα του κειμένου και συνεπώς δεν μπορούν να εφαρμοστούν σε άλλες γλώσσες. Από την άλλη οι στατιστικές τεχνικές (statistical techniques) έχουν πιθανό υπόβαθρο και εστιάζουν στις σχέσεις μεταξύ των λέξεων.Οι τεχνικές αυτές έχουν δύο σημαντικά πλεονεκτήματα σε σχέση με τις συντακτικές: μπορούμε να τις χρησιμοποιήσουμε σε διάφορες γλώσσες τροποποιώντας τες ελάχιστα ή και καθόλου και μπορούμε να χρησιμοποιήσουμε μετάφραση μηχανής (machine translation) του αρχικού σετ δεδομένων και να έχουμε και πάλι ικανοποιητικά αποτελέσματα. Οι δύο προαναφερθείσες μέθοδοι μπορούν να συνδυαστούν τόσο με τεχνικές βασισμένες σε λεξικό, όσο και με τεχνικές μηχανικής μάθησης.

Οι κυριότερες μέθοδοι για την σύνταξη των λεξικών είναι οι εξής:

***2.5.1 Dictionary based techniques*:** Σε αυτές τις τεχνικές ξεκινάμε με την συλλογή, με το χέρι, ενός συνόλου από γνωστές λέξεις που εκφράζουν συναίσθημα. Στη συνέχεια ψάχνουμε στο λεξικό WordNet για συνώνυμα και αντώνυμα και τα προσθέτουμε στο λεξικό μας. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να μην υπάρχουν άλλες λέξεις να προστεθούν. Η προσέγγιση αυτή έχει μειονέκτημα στο ότι δεν μπορεί να βρει opinion words για ειδικότερα θέματα.

2.5.2 Corpus based techniques: Αυτές οι τεχνικές βασίζονται σε συντακτικά μοτίβα σε μεγάλα corpora. Μπορούν να παράγουν opinion words με σχετικά μεγάλη ακρίβεια, χρειάζονται όμως ένα πολύ μεγάλο δεδομένων για training. Έχουν πλεονέκτημα ότι μπορούν να παράγουν opinion words για ειδικότερα θέματα, ανάλογα και με τα δεδομένα που παρέχουν για το training. Γίνεται προφανές ότι οι τεχνικές που χρησιμοποιούνε machine learning μπορούν να παρέχουν πολύ καλύτερα αποτελέσματα από τις τεχνικές με λεξικό. Οι τεχνικές με λεξικό έχουν σαφείς περιορισμούς καθώς λαμβάνουν κυρίως υπ' όψη κάθε λέξη ξεχωριστά, πράγμα που οδηγεί πολύ εύκολα σε αστοχίες. Οι τεχνικές machine learning, μπορεί να απαιτούν μεγάλο όγκο δεδομένων για να εκπαιδευτούν, όμως εστιάζουν πέρα από κάθε λέξη και στη σύνδεση με τις υπόλοιπες λέξεις για να κάνουν την κατηγοριοποίηση, γεγονός που προσεγγίζει περισσότερο την ανθρώπινη κρίση.. Στην Εικόνα 1 παραθέτουμε έναν πίνακα με τα καλύτερα ποσοστά απόδοσης που έχουμε εντοπίσει στην βιβλιογραφία για Sentiment classification

ΕΙΚΟΝΑ 1

Authors	Classifier	Features	Domain	Accuracy
Turney	Pointwise Mutual Information	bigrams	movies, cars, banks	66-84%
Pang & Lee	NB, ME, SVMs	unigrams, bigrams, feature presence or frequence, negation	IMDb[1]	82.9%
Pang & Lee	NB, SVMs	sentence level subjectivity based on minimum cuts	IMDb	87.2%
Mullen & Collier	Hybrid SVM	Turney values, Osgood values, lemma models	IMDb[1], record reviews	87%
Bai et al.	two-stage Markov Blanket Classifier	dependence among words, minimal vocabulary	IMDb[1]	87.5%
Whitelaw et al.	SMO with linear kernel	appraisal groups	movie reviews	90.2%
Kennedy & Inkpen	SVMs, term counting, combination	term frequencies	IMDb	86.2%
Annett & Kondrak	WordNet and SVM	num of pos/neg adj/adv, min distance from pivot words in WordNet	movie reviews, blog posts	65.4-77.5%
Zhou & Chaovalit	ontology-supported polarity mining	n-grams, words, word senses	movie reviews	72.2%
Abbasi et al.	Genetic Algorithm(GA), Information Gain(IG), GA+IG	stylistic and syntactic features	movie reviews, web forum postings	91.7%

2.5.3 Τεχνικές βασισμένες σε Λεξικό – Αδυναμίες

Η μέθοδος αυτή, παρά την απλότητά της, έχει κάποιες αδυναμίες που περιορίζουν την απόδοσή της στην ταξινόμηση κειμένου. Οι βασικοί περιορισμοί με τους οποίους έρχεται αντιμέτωπη η μέθοδος αυτή είναι:

1)Άρνηση (Negation): Η διαδικασία που περιγράφτηκε παραπάνω, αγνοεί την άρνηση, πουστα αγγλικά ανιχνεύεται με τις λέξεις not, never, nothing, nobody κτλ. Η ύπαρξη μίαςάρνησης λέξης, επηρεάζει σαφώς το νόημα της λέξης ή των λέξεων που ακολουθούν 48 αντιστρέφοντας το polarity. Για παράδειγμα, ενώ η πρόταση “The movie was good” αντιστοιχεί σε μία θετική κριτική, η πρόταση “The movie was not good” αντιστοιχεί σε μία αρνητική κριτική. Η προφανής επιλογή αντιμετώπισης του φαινομένου αυτού και αυτή που χρησιμοποιείται συνήθως είναι η αντιστροφή του polarity των λέξεων που ακολουθούν την λέξη άρνησης μέχρι το επόμενο σημείο στίξης ή κάποιον αντιθετικό σύνδεσμο, πχ. but, however κ.ά. Όμως η επιλογή αυτή έχει αδυναμίες. Για παράδειγμα, αν υποθέσουμε ότι η λέξη excellent έχει βαθμολογία +5 και η λέξη good έχει βαθμολογία +3, τότε προκύπτει ότι η άρνηση not good έχει μεγαλύτερη βαθμολογία από την άρνηση not excellent, στοιχείο που δεν συνάδει με τη διαίσθησή μας.

2)Μετατόπιση Έντασης/Σθένους (Valence Shifters): Πέραν της άρνησης, που σαφώς επηρεάζει το νόημα των λέξεων που ακολουθούν, υπάρχουν και οι λέξεις μεταβολής έντασης που αυξάνουν (intensifiers) ή μειώνουν (downtoners) την ένταση της επόμενης λέξης. Παραδείγματα τέτοιων λέξεων στα αγγλικά είναι very, truly, really, slightly, more, less κ.α. Η εξέταση των λέξεων αυτών είναι σημαντική, ιδιαίτερα στην περίπτωση που επιδιώκουμε πολυεπίπεδη συναισθηματική κατάταξη. Κάποιοι ερευνητές, όπως οι Kennedy και Inkpen και οι Polanyi and Zaenen , αντιμετώπισαν το φαινόμενο με απλή πρόσθεση και αφαίρεση: Αν βρεθεί κάπου μέσα στο κείμενο μία λέξη intensifier, το polarity της επόμενης λέξης αυξάνεται κατά μία σταθερή ποσότητα, ενώ αντίθετα, αν βρεθεί μία λέξη downtoner, το polarity της επόμενης λέξης μειώνεται κατά την ίδια σταθερή ποσότητα. Το πρόβλημα αυτής της προσέγγισης είναι ότι δεν λαμβάνει υπόψη τις διαφορές μεταξύ των valence shifters. Πχ. η λέξη extraordinarily είναι για παράδειγμα πολύ πιο δυνατός «ενισχυτής» από τη λέξη rather.

3)Σειρά Λέξεων: Η lexicon-based μέθοδος αγνοεί τη σειρά των λέξεων που εμφανίζονται στο κείμενο. Αυτή η BoW (Bag of Words) μοντελοποίηση του κειμένου εμφανίζει αδυναμίες αφού η σειρά των λέξεων μπορεί και να αντιστρέψει το polarity μίας πρότασης. Αν χρησιμοποιήσουμε και πάλι το παράδειγμα:

- That's not true, I'm a fan of this movie.
- That's true, I'm not a fan of this movie.

βλέπουμε ότι οι δύο παραπάνω προτάσεις χρησιμοποιούν το ίδιο σύνολο λέξεων αλλά έχουν εντελώς διαφορετικό (αντίθετο) νόημα. Το παραπάνω πρόβλημα αντιμετωπίζεται με την επισήμανση της άρνησης (λέξη not) και τον προσδιορισμό της εμβέλειάς της όπως εξηγήθηκε νωρίτερα.

4)Ύπαρξη Αντιθετικών/Εναντιωματικών Συνδέσμων (Adversative conjunctions):

Οι αντιθετικοί σύνδεσμοι σε μία πρόταση, όπως υποδηλώνει το όνομά τους, συνδέουν δύο φράσεις αντίθετης πολικότητας. Παραδείγματα αντιθετικών συνδέσμων στα αγγλικά είναι 49 οι λέξεις but, although, however κ.ά. Συνήθως, το polarity της συνολικής πρότασης καθορίζεται από το δεύτερο συστατικό της πρότασης.

Για παράδειγμα, στην πρόταση "The car is nice but expensive" η προδιάθεση του συγγραφέα ή ομιλητή είναι εναντίον της αγοράς του αυτοκινήτου, ενώ στην πρόταση "The car is expensive but nice", η προδιάθεση του συγγραφέα ή ομιλητή ως προς την αγορά του αυτοκινήτου είναι θετική.

5)Ιδιωματισμοί: Μία άλλη αστοχία της μεθόδου είναι ότι μελετώντας κάθε λέξη ξεχωριστά,αγνοούμε την ύπαρξη φράσεων των οποίων οι επιμέρους λέξεις προσδίδουν ένα ιδιαίτερο συνολικό νόημα. Παράδειγμα αποτελεί η φράση "once in a blue moon", που σημαίνει πολύ σπάνια. Για την αντιμετώπισή τους απαιτείται η χρήση ειδικού λεξικού που θα περιλαμβάνει τέτοιους ιδιωματισμούς και η ανάλυση του κειμένου όχι σε επίπεδο λέξεων αλλά σε επίπεδο φράσεων.

6)Αμφισημία: Το φαινόμενο κατά το οποίο μία λέξη ή φράση έχει διαφορετικό νόημα ανάλογα με το ευρύτερο νοηματικό πλαίσιο στο οποίο χρησιμοποιείται ονομάζεται αμφισημία. Για παράδειγμα, η λέξη “unpredictable” έχει θετική έννοια όταν αναφέρεται σε μία ταινία και συνήθως αρνητική όταν αναφέρεται στον καιρό. Ακόμη, η φράση “go read the book” είναι θετική για μία κριτική βιβλίου αλλά αρνητική για μία κριτική ταινίας. Ο χειρισμός της αμφισημίας είναι ένα από τα δυσκολότερα εμπόδια που καλείται να λύσει η ανάλυση συναισθήματος και έχει συγκεντρώσει το ενδιαφέρον πολλών ερευνητών.

7)Ειρωνεία: Πολλές φορές οι άνθρωποι επιστρατεύουν το χιούμορ και τον σαρκασμό για να εκφράσουν την συνήθως αρνητική άποψή τους. Η ανίχνευση της ειρωνείας είναι πολλές φορές δύσκολη από τον άνθρωπο, πόσο μάλλον από μία μηχανή.

Για παράδειγμα, η πρόταση “The restaurant was great in that it will make all future meals seem more delicious” εμπεριέχει ειρωνεία που μπορεί να μη γίνει αντιληπτή από κάποιο αναγνώστη που την «προσπερνά» γρήγορα. Το πρόβλημα λοιπόν της ανίχνευσης ειρωνείας κειμένου με αυτοματοποιημένο τρόπο είναι δύσκολο και η επιστημονική έρευνα έχει επικεντρωθεί σε μεθόδους εντοπισμού της με διάφορες μεθόδους μελετώντας το δυαδικό πρόβλημα της ταξινόμησης μίας πρότασης ως σαρκαστική ή όχι.

***8)Πολλαπλοί στόχοι*:** Πολλές φορές σε μία κριτική γίνεται αναφορά σε παραπάνω από μία οντότητες (πρόσωπα, προϊόντα, γεγονότα) ή ακόμα και σε διαφορετικά χαρακτηριστικά (aspects) της ίδιας οντότητας. Στην περίπτωση αυτή συνήθως δεν ενδιαφέρει η ταξινόμηση του κειμένου συνολικά ως μία θετική ή αρνητική άποψη αλλά εξετάζεται το κείμενο σε επίπεδο χαρακτηριστικών (aspect-level). Η ανάλυση συναισθήματος σε επίπεδο χαρακτηριστικών (Aspect-level SA) στοχεύει στον προσδιορισμό του συναισθήματος ως προς συγκεκριμένα aspects των οντοτήτων και διαφέρει από την ανάλυση συναισθήματος σε επίπεδο κειμένου (Document-level SA). Το πρώτο βήμα είναι η αναγνώριση των οντοτήτων και των χαρακτηριστικών τους. Στη συνέχεια εξάγονται οι γλωσσικές 50 εκφράσεις που αναφέρονται στα ζεύγη (Entity, Aspect) και με χρήση συνωνύμων και λεξικού αποδίδεται το αντίστοιχο συναίσθημα.

Για παράδειγμα, στην πρόταση:

‘‘The voice quality of this phone is not good, but the battery life is long”, η Aspect-level SA θα έδινε: {(phone, voice quality), negative}

{(phone, battery), positive}

Μία ενδιαφέρουσα εφαρμογή του sentiment analysis είναι η ανάλυση άποψης, σχετικά με μία οντότητα, από κείμενα χρηστών ενός κοινωνικού μέσου, για παράδειγμα του twitter. Στο πρόβλημα όμως αυτό εισέρχονται ακόμη περισσότερες δυσκολίες (πέραν αυτών που αναφέρθηκαν προηγουμένως) που σχετίζονται με την ιδιαίτερη μορφή των tweets. Συγκεκριμένα, το μικρό τους μέγεθος (έως 140 χαρακτήρες) υποχρεώνει τον χρήστη να εκφράζει την άποψή του με λίγες

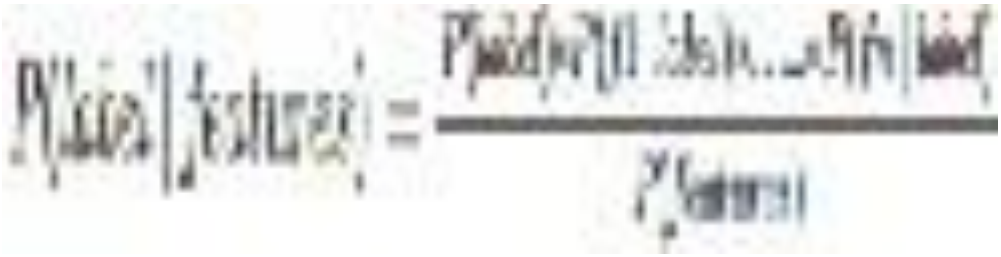
χρωματισμένες λέξεις οι οποίες μπορεί να μην βρίσκονται στο λεξικό. Τότε, η εξαγωγή συμπεράσματος ως προς την πολικότητα της άποψης είναι εξαιρετικά δύσκολη αν όχι αδύνατη. Επίσης, στα tweets, είναι συχνή η χρήση συντομογραφιών (λόγω του περιορισμού μεγέθους), όπως και συχνά είναι και τα ορθογραφικά λάθη λόγω επιπολαιότητας. Ακόμη πολλές φορές τα μηνύματα περιέχουν λέξεις από περισσότερες της μίας γλώσσας, ενώ ειδικότερα για τους Έλληνες χρήστες η χρήση greeklish είναι ευρέως διαδεδομένη. Όλα αυτά, απαιτούν ένα λεξικό το οποίο θα αφομοιώνει τις τάσεις των χρηστών του διαδικτύου, ενσωματώνοντας συχνά misspellings,συντομογραφίες και λέξεις και άλλων γλωσσών κάτι, λεξικό που όμως είναι δύσκολο να κατασκευαστεί.

Σε αντίθεση με τη μηχανική μάθηση, η βασισμένη σε λεξικό μέθοδος δεν απαιτεί την εκπαίδευση ενός ταξινομητή πάνω σε επισημειωμένα δεδομένα εξοικονομώντας έτσι σημαντικό χρόνο. Ωστόσο, απαιτεί ένα συναισθηματικό λεξικό, περιορίζοντας την εφαρμογή της μεθόδου στην ανάλυση κειμένων γραμμένων στην γλώσσα του λεξικού και την απόδοσή της στην ποιότητα ή πληρότητα του λεξικού. Από την άλλη πλευρά, οι μέθοδοι μηχανικής μάθησης πετυχαίνουν συνήθως καλά αποτελέσματα με αντίτιμο την ανάγκη εκπαίδευσης που μπορεί να είναι χρονοβόρα. Γι' αυτό, η επιστημονική έρευνα τα τελευταία χρόνια προσανατολίζεται στη χρήση υβριδικών μεθόδων που συνδυάζουν λεξικό με μηχανική μάθηση ώστε να επωφεληθούν από τα πλεονεκτήματα των επιμέρους μεθόδων, δηλαδή της ταχύτητας της lexicon-based προσέγγισης και της ακρίβειας της machine learning προσέγγισης.

2.6 Τεχνικές βασισμένες στη Μηχανική Μάθηση

Οι προσεγγίσεις μηχανικής μάθησης όπου είναι εφαρμόσιμες στο Sentiment Analysis ανήκουν στην κατηγορία της εποπτευόμενης (supervised) εκπαίδευσης. Για να εφαρμοστούν οι τεχνικές μηχανικής μάθησης χρειάζονται δύο διαφορετικές κατηγορίες κειμένων: τα κείμενα για εκπαίδευση και τα κείμενα όπου θα γίνουν τα test. Ένας αυτόματος ταξινομητής (classifier) χρησιμοποιεί τα κείμενα εκπαίδευσης για να μάθει τα διαφοροποιητικά χαρακτηριστικά των κειμένων και στη συνέχεια να δοκιμάσει την ακρίβεια και την αποδοτικότητά του στα κείμενα για test. Ένας μεγάλος αριθμός από τεχνικές machine learning έχει χρησιμοποιηθεί για να την κατηγοριοποίηση γνωμών. 35 Οι αλγόριθμοι που έχουν συναντήσει περισσότερο στην βιβλιογραφία και έχουν χρησιμοποιηθεί περισσότερο για κατηγοριοποίηση γνώμης είναι:

2.6.1 Naïve Bayes (NB): Το μοντέλο κατηγοριοποίησης Naïve Bayes υπολογίζει την aposteriori πιθανότητα μίας κλάσης, βασισμένο στην κατανομή των λέξεων της κλάσης στο κείμενο. Αυτό το μοντέλο χρησιμοποιεί για το feature extraction την τεχνική Bagof Words (BOW), μη λαμβάνοντας υπ'' όψη τη θέση της λέξης στο κείμενο. Χρησιμοποιεί το Θεώρημα Bayes για να υπολογίσει την πιθανότητα ένα δεδομένο σύνολο από feature (feature set) να ανήκει σε ένα συγκεκριμένο label. Θεωρώντας ότι όλα τα features είναι ανεξάρτητα μεταξύ τους, προκύπτει ότι



όπου P(label) είναι η apriori πιθανότητα για ένα τυχαίο feature να ορίσει ένα label, P(f|label) είναι η apriori πιθανότητα για κάθε σύνολο από feature να κατηγοριοποιηθεί ως label, P(features) είναι η apriori πιθανότητα ότι ένα δεδομένο σύνολο από features εμφανίζεται.

Ο αλγόριθμος Naive Bayes είναι από τις βασικές τεχνικές ταξινόμησης κειμένου και παρά τηναπλότητά του και τις υποθέσεις ανεξαρτησίας που κάνει, αποδίδει καλά σε πολλά προβλήματα.Η καλή απόδοση συνδυάζεται μάλιστα με χαλαρές απαιτήσεις ως προς τη CPU και τη μνήμη, ενώκαι ο χρόνος εκπαίδευσης είναι σημαντικά μικρότερος σε σχέση με άλλες μεθόδους. Όμως, ο Naive Bayes είναι ένας κακός εκτιμητής, καθώς συχνά υπερεκτιμά τις πιθανότητες εξόδου. Στο γενικότερο πρόβλημα της αναγνώρισης προτύπων, όπως αναφέραμε και προηγουμένως,καλούμαστε να επιλέξουμε μία κλάση cj στην οποία θεωρούμε ότι ανήκει ένα πρότυπο με βάσητο διάνυσμα χαρακτηριστικών του, έστω $\boldsymbol{x}$. Η επιλογή μας γίνεται μέσα από N πιθανές κλάσεις $c1, c2, \dots, cN$. Αν ορίσουμε τη δεσμευμένη πιθανότητα: $P(cj|\boldsymbol{x}), j = 1, \dots, N$, ως την πιθανότητα το πρότυπο $\boldsymbol{x}$ να ανήκει στην κλάση cj, γνωστή και ως εκ των υστέρων πιθανότητα (a posteriori probability), τότε η διαίσθησή μας λέει να επιλέξουμε για το $\boldsymbol{x}$, την κλάση που μεγιστοποιεί την παραπάνω a posteriori πιθανότητα , έστω την κλάση k. Δηλαδή θεωρούμε τον ακόλουθο κανόνα απόφασης:

Το πρότυπο $\boldsymbol{x}$ αντιστοιχίζεται στην κλάση ck, όπου: $k = arg \max P(cj\ j|\boldsymbol{x}), j = 1, \dots, N$ Αυτός ακριβώς είναι ο κανόνας απόφασης στον ταξινομητή Naive Bayes, και γι' αυτό ονομάζεται και Maximum A Posteriori (MAP) ταξινομητής. Η πιθανότητα $P(cj|\boldsymbol{x})$ εφαρμόζοντας το θεώρημα του Bayes υπολογίζεται ως εξής: $P(cj|\boldsymbol{x}) = P(cj, \boldsymbol{x})P(\boldsymbol{x}) = P(\boldsymbol{x}|cj)P(cj)P(\boldsymbol{x})$ Όπου: $P(cj)$ είναι η πρότερη πιθανότητα (prior probability) της κλάσης j $P(\boldsymbol{x}|cj)$ είναι η πιθανότητα του χαρακτηριστικού $\boldsymbol{x}$ δεδομένης της κλάσης cj (class conditional probability density function) Παρατηρούμε πως η πιθανότητα $P(x)$ δεν χρειάζεται να υπολογιστεί διότι στην εφαρμογή του κανόνα Naive Bayes εμφανίζεται ως σταθερή

ποσότητα, ανεξάρτητη του j και δεν επηρεάζει τη μεγιστοποίηση. Στο σημείο αυτό έρχεται να εφαρμοστεί και η υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών δεδομένης της κλάσης cj, οπότε η πιθανότητα $P(\boldsymbol{x}|cj)$ υπολογίζεται ως το γινόμενο των επιμέρους $P(xi|cj)$.

Δηλαδή: $P(\boldsymbol{x}|cj) = \prod P(xi|cj) ni=1$ Για την απόφαση του ταξινομητή λοιπόν, αρκεί ο υπολογισμός των πιθανοτήτων $P(cj)$ και $P(xi|cj)$. Οι πιθανότητες αυτές εκτιμώνται κάνοντας χρήση της εκτίμησης μέγιστης πιθανοφάνειας (Maximum Likelihood Estimation – MLE) πάνω στο training set. Σύμφωνα με τη MLE, οι παράμετροι ενός στατιστικού μοντέλου επιλέγονται έτσι ώστε να συμφωνούν με ταδεδομένα που έχουμε στη διάθεσή μας. Έτσι, η πιθανότητα $P(cj)$ υπολογίζεται ως το ποσοστότων προτύπων στο training set που ανήκουν στη κλάση cj και η πιθανότητα $P(\boldsymbol{x}|cj)$ υπολογίζεταιαπό τις επιμέρους πιθανότητες $P(xi|cj)$ οι οποίες εκτιμώνται ίσες με τις αντίστοιχες συχνότητεςτων χαρακτηριστικών στο ίδιο training set.

2.6.2 Maximum Entropy (ME): Η μέθοδος αυτή κωδικοποιεί feature sets, που έχουν ήδη κάποιο label, σε διανύσματα. Αυτό το κωδικοποιημένο διάνυσμα χρησιμοποιείται για να υπολογιστεί ένα βάρος για κάθε feature, τα οποία στην συνέχεια συνδυάζονται για να καθοριστεί το πιο πιθανό label για το feature set. Ο classifier παραμετροποιείται από ένα σύνολο από X{weights}, το οποίο χρησιμοποιείται για να συνδυάσει τα κοινά features που παράγονται από ένα feature set με ένα X{encoding}. Η πιθανότητα για κάθε label υπολογίζεται έτσι:

$$P(fs|label) = \frac{dotprod(weights, encode(fs, label))}{sum(dotprod(weights, encode(fs, l))\ forlinlabels)}$$

Ο αλγόριθμος μέγιστης εντροπίας (Maximum Entropy) που ονομάζεται και αλγόριθμος λογιστικής παλινδρόμησης (Logistic Regression) υλοποιεί παρά την ονομασία του ένα γραμμικό μοντέλο με σκοπό την ταξινόμηση και όχι την παλινδρόμηση. Πρόκειται για έναν πιθανοτικό ταξινομητή του οποίου οι πιθανότητες εξόδου μοντελοποιούνται κάνοντας χρήση μιας λογιστικής συνάρτησης, δηλαδή συνάρτησης της μορφής:

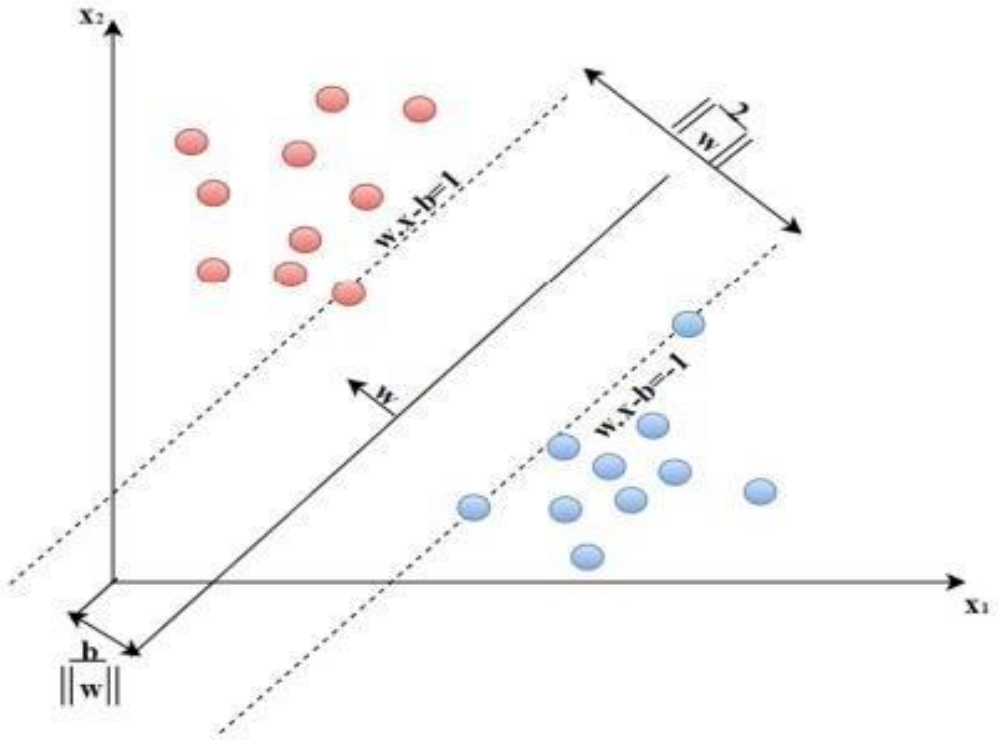
$f(x) = L / 1 + e^{-k(x-x0)}$

Η γενίκευση της λογιστικής συνάρτησης στην περίπτωση πολλών εισόδων ονομάζεται softmaxfunction και ορίζεται παρακάτω. Όπως υποδηλώνει το όνομά του, ο Max Entropy βασίζεται στην αρχή της μέγιστης εντροπίας,σύμφωνα με την οποία μεταξύ όλων των μοντέλων που ταιριάζουν με τα δεδομένα επιλέγεταιεκείνο που δεν κάνει καμία άλλη υπόθεση πέρα των περιορισμών που επιβάλλονται από totraining set και συνεπώς η κατανομή είναι όσο το δυνατόν ομοιόμορφη. Ο ταξινομητής MaxEntropy

χρησιμοποιείται σε πολλά προβλήματα ταξινόμησης κειμένου, όπως ανίχνευση γλώσσας,ταξινόμηση με βάση το θέμα του κειμένου, sentiment analysis και άλλα.Ανόμοια με τον επίσης πιθανοτικό ταξινομητή Naive Bayes, που αναπτύχθηκε στηνπροηγούμενη ενότητα, ο Max Entropy δεν υποθέτει ότι τα χαρακτηριστικά είναι υπό συνθήκηανεξάρτητα μεταξύ τους. Το γεγονός ότι ο αλγόριθμος μέγιστης εντροπίας δεν κάνει κάποιαυπόθεση ανεξαρτησίας μεταξύ των χαρακτηριστικών τον καθιστά ιδιαίτερα αποδοτικό σεπροβλήματα ταξινόμησης κειμένου, όπου τα χαρακτηριστικά-λέξεις δεν είναι προφανώςανεξάρτητα μεταξύ τους. Το μειονέκτημά του σε σχέση με τον Naive Bayes είναι ο μεγαλύτεροςχρόνος εκπαίδευσης λόγω του προβλήματος

2.6.3 Support Vector Machines (SVM): Εκτός από τους κατηγοριοποιητές που χρησιμοποιούν πιθανολογικά μοντέλα, υπάρχουν και εκείνοι που προσεγγίζουν τα δεδομένα γραμμικά. Ένας από αυτούς είναι οι μηχανές υποστήριξης διανυσμάτων (Support Vector Machines / SVM), όπου γίνεται για πρώτη φορά λόγος από τους Vapnik και Chermovenkis το 1963.Πρόκειται για έναν αλγόριθμο μηχανικής μάθησης που βασίζεται σε διανύσματα και επιχειρεί να βρει τη διαχωριστική γραμμή μεταξύ κλάσεων, με τέτοιο τρόπο ώστε νακατηγοριοποιούνται καλύτερα τα νέα δεδομένα. Ο SVM είναι γνωστός για την καλή απόδοση γενίκευσης που οφείλεται στην ιδιότητα του μέγιστου περιθωρίου. Γι' αυτό πολλές φορές τον αποκαλούν και κατηγοριοποιητή μέγιστου περιθωρίου.Πρόκειται για ένα μοντέλο μάθησης υπό επίβλεψη (supervised) το οποίο χρησιμοποιείται τόσο στην κατηγοριοποίηση (classification) όσο και στην παλινδρόμηση (regression).Εχοντας ως βάση ενα αρχικό σύνολο κατηγοριοποιημένων δεδομένων, ο στόχος είναι η δημιουργία ενός βέλτιστου υπερεπιπέδου (hyper-plane,) το οποίο θα διαχωρίζει τις ομάδες μεταξύ τους. Βέλτιστο υπερεπίπεδο είναι αυτό του οποίο το περιθώριο από τις ομάδες είναι το μεγαλύτερο δυνατό.Έστω ένα σύνολο δεδομένων εκπαίδευσης, καθένα από τα οποία ανήκει σε μία από τις δύο κατηγορίες (Εικόνα 2) ο αλγορίθμος SVM δημιουργεί ένα μοντέλο που εκχωρεί νέα δεδομένα σε μία από τις δύο κατηγορίες.

ΕΙΚΟΝΑ 2



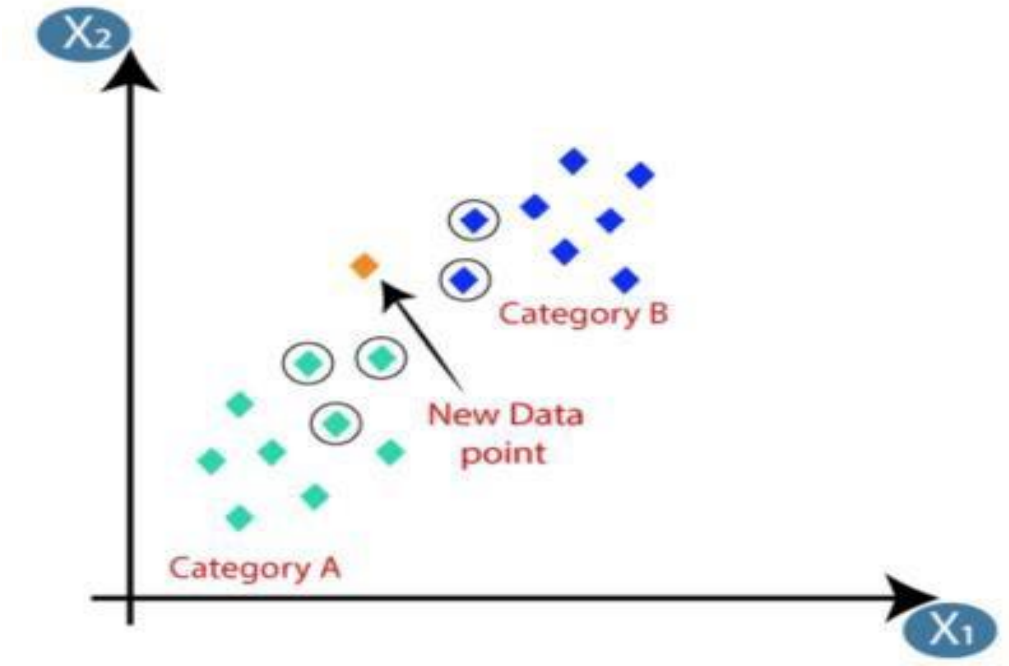
2.6.4 K-Nearest Neighbors (K-NN):

Ο αλγόριθμος κοντινότερων γειτόνων (K-Nearest Neighbors / K-NN) είναι ένας από τους πιο γνωστούς ταξινομητές στον τομέα του Machine Learning. Ουσιαστικά αυτό που κάνει είναι να διαχωρίζει τα δεδομένα σε κλάσεις, με σημείο αναφοράς και σύγκρισης τους κοντινότερους γείτονες.Ας υποθέσουμε ότι έχουμε ένα training set με κατηγοριοποιημένα δεδομένα. Κάθε νεοδεδομένο θα πρέπει να κατηγοριοποιείται σύμφωνα με τα κοντινά του δεδομένα. Αν επομένως η κατηγοριοποίηση ενός παραδείγματος δεν είναι γνωστή, τότε μπορεί να γίνει ελέγχοντας την κλάση των κοντινών του γειτόνων. Ο K-NN υπολογίζει την απόσταση μεταξύ του testing set και του training set. Τελικά, η απόσταση με τη μικρότερη τιμή,δηλαδή η κοντινότερη, χρησιμοποιείται για την κατηγοριοποίηση νέων δεδομένων.

Ένα παράδειγμα που παρουσιάζει τη λειτουργία του κατηγοριοποιητή K-NN φαίνεται στο παρακάτω σχήμα(ΕΙΚΟΝΑ 3). Το πρόβλημα αποτελείται από δύο κλάσεις, πράσινους ρόμβους(κατηγορία Α) και μπλε ρόμβους (κατηγορία Β). Για να ταξινομήσουμε το νέο δείγμα, πορτοκαλί ρόμβος, σε μια από τις δύο κλάσεις θα χρειαστεί να λάβουμε υπόψη τουςκοντινότερους γείτονές του, όπως φαίνονται στο

σχήμα. Αν ορίσουμε τη μεταβλητή k = 5, παρατηρούμε ότι από τα πέντε κοντινότερα δείγματα, τα τρία ανήκουν στην κατηγορία Ακαι υπόλοιπα δύο στην κατηγορία Β. Επομένως, λαμβάνοντας την πλειοψηφία μπορούμε να κατατάξουμε το άγνωστο δείγμα στην κατηγορία Α.

ΕΙΚΟΝΑ 3

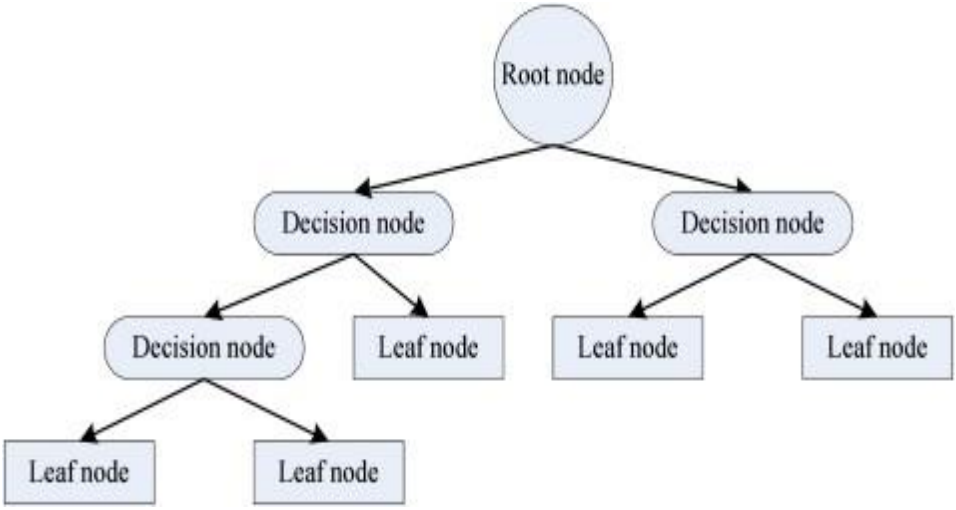


Η μετρική της απόστασης παίζει πολύ σημαντικό ρόλο στην απόδοση του αλγορίθμου.Εξίσου σημαντικός παράγοντας είναι η τιμή του k, που αποτελεί τη βασική παράμετρο του αλγορίθμου. Αν το k είναι πολύ μεγάλο, οι κλάσεις με πολλά κατηγοριοποιημένα δεδομέναμπορεί να υπερισχύσουν και τα αποτελέσματα να είναι μεροληπτικά. Το αντίθετο συμβαίνει αν το k είναι πολύ μεγάλο.

2.6.5 Decision Trees *:* Τα Δέντρα Αποφάσεων (Decision Trees / DT) μπορούν να προσαρμοστούν σε σχεδόν οποιοδήποτε τύπο δεδομένων, επομένως είναι ένας από τους πιο ευρέως χρησιμοποιούμενους αλγόριθμους μηχανικής μάθησης. Πρόκειται για έναν αλγόριθμομάθησης με επίβλεψη (supervised) που χωρίζει τα δεδομένα εκπαίδευσης σε όλο και μικρότερα τμήματα, προκειμένου να προσδιορίσει τα πρότυπα που μπορούν ναχρησιμοποιηθούν για την κατηγοριοποίηση. Στη συνέχεια, τα δεδομένα παρουσιάζονται με μορφή λογικής δομής παρόμοιας με εκείνη της εικόνας 4 που μπορεί εύκολα να γίνεικατανοητή χωρίς στατιστική γνώση. Όταν χτίζεται ένα Decision Tree, χωρίζεται ο χώρος εισόδου με έναν ιεραρχικό τρόπο αναδρομικά. Αυτή η μέθοδος είναι γνωστή και ως «διαίρεικαι βασίλευε».Τα δέντρα αποφάσεων είναι χτισμένα χρησιμοποιώντας ευριστικό διαχωρισμό (heuristic partitioning). Η δομή ενός δέντρου αποφάσεων αποτελείται από έναν κόμβο πουονομάζεται

ρίζα (root), ο οποίος αντιπροσωπεύει ολόκληρο το σύνολο δεδομένων, τους κόμβους απόφασης (decision nodes), που εκτελούν υπολογισμούς και τους κόμβους φύλλα ή αλλιώς τερματικούς κόμβους (leaf nodes / terminal nodes), που αντιπροσωπεύουν μία κατηγορία. Στη φάση της εκπαίδευσης ο αλγόριθμος μαθαίνει ποιες αποφάσεις πρέπει να ληφθούν προκειμένου να χωριστούν τα δεδομένα εκπαίδευσης σε κατηγορίες.

ΕΙΚΟΝΑ 4

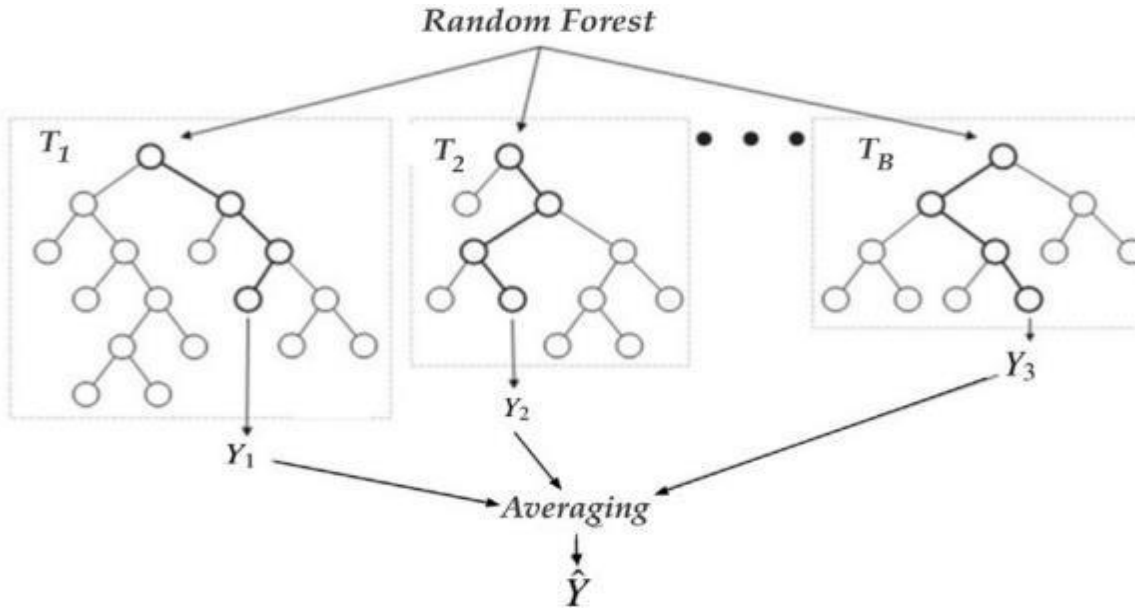


Κατα την κατηγοριοποίηση τα δεδομένα εισόδου διαβιβάζονται μέσω του δέντρου. Σε κάθε κόμβο απόφασης (decision node) ένα συγκεκριμένο χαρακτηριστικό από τα δεδομένα εισόδου συγκρίνεται με μια σταθερά που καθορίστηκε στη φάση της εκπαίδευσης. Ο υπολογισμός που λαμβάνει χώρα σε κάθε κόμβο απόφασης συνήθως συγκρίνει το επιλεγμένο χαρακτηριστικό με αυτήν την προκαθορισμένη σταθερά. Η απόφαση βασίζεται στο αν το χαρακτηριστικό είναι μεγαλύτερο ή μικρότερο από τη σταθερά, δημιουργώντας διακλαδώσεις στο δέντρο. Τα δεδομένα τελικά θα περάσουν από τους κόμβους απόφασης μέχρι να φτάσουν σε έναν τερματικό κόμβο (leaf node / terminal node) που αντιπροσωπεύει μία καθορισμένη κατηγορία

2.6.6 Random Forest: Το Τυχαίο Δάσος (Random Forest / RF) είναι ένας αλγόριθμος μάθησης με επίβλεψη (supervised) που δημιουργεί τυχαία και συγχωνεύει πολλαπλά δέντρα αποφάσεων (decision trees) σε ένα "δάσος". Πιο συγκεκριμένα, το τυχαίο δάσος χτίζει πολλαπλά δέντρα απόφασης και τα συγχωνεύει για να πάρει μια πιο ακριβή και σταθερή πρόβλεψη. Κάθε δέντρο εκπαιδεύεται χρησιμοποιώντας ένα υποσύνολο των δεδομένων εκπαίδευσης. Σε κάθε κόμβο απόφασης (decision node) επιλέγεται ένα τυχαίο υποσύνολο των χαρακτηριστικών και ο αλγόριθμος λαμβάνει υπόψη μόνο τις διαφορές σε αυτά τα χαρακτηριστικά.Είναι ένας ευέλικτος και εύχρηστος αλγόριθμος, ο οποίος

παράγει εξαιρετικά αποτελέσματα τις περισσότερες φορές. Χρησιμοποιείται ευρέως λόγω της απλότητας του και επειδή μπορεί να χρησιμοποιηθεί τόσο για προβλήματα κατηγοριοποίησης (classification) όσο και για προβλήματα παλινδρόμησης (regression).Η βασική δομή του random forest απεικονίζεται στην εικόνα 5.

ΕΙΚΟΝΑ 5



2.6.7 Artificial Neural Networks: Τα Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks / ANN) ή απλά Νευρωνικά Δίκτυα (Neural Networks / NN) είναι μια κατηγορία μοντέλων που χρησιμοποιούνται στη μηχανική μάθηση με επιτυχία σε πληθώρα πεδίων. Ονομάζονται έτσι γιατί είναι εμπνευσμένα από τους νευρώνες του εγκεφάλου, αλλά όσον αφορά τη μηχανική μάθηση,είναι περισσότερο ένα απλοποιημένο μοντέλο που βρίσκει εφαρμογή σε συγκεκριμένες εργασίες παρά σε κάτι που έχει σχέση με την πραγματική λειτουργία του ανθρώπινουεγκεφάλου. Παρόλα αυτά μπορούμε να πούμε ότι αποτελεί την πλέον κατάλληλη μέθοδο για ανάπτυξη ευφυών αλγορίθμων και γενικότερα διαδικασιών σχετιζόμενων με τη νοημοσύνη, όπως η μάθηση, η μνήμη κ.λ.π.Η επιτυχία του μοντέλου αποδίδεται στην ικανότητά του να ελαχιστοποιεί μια συνάρτηση κόστους ύστερα από την διαδικασία της εκπαίδευσης η οποία έχει απώτερο σκοπό τη ρύθμιση εσωτερικών παραμέτρων, των συναπτικών βαρών (synaptic weights).Το πιο απλό νευρωνικό δίκτυο αποτελείται από ένα μόνο νευρώνα και είναι το μοντέλο perceptron. Οι μόνες συνδέσεις που υπάρχουν είναι αυτές μεταξύ των εισόδων χ1,χ2...,χν και του νευρώνα.Τα βασικά συστατικά του μοντέλου perceptron είναι τα εξής:

- Είσοδος (input): διάνυσμα x= [x1,x2,...,xn], όπου αντιπροσωπεύει το δείγμα και το n αντιστοιχεί στον συνολικό αριθμό των δειγμάτων.

- Συναπτικά Βάρη (Synaptic Weights): w = [wk1,wk2,...,wkn].

- Το κατώφλι Bk πρόκειται για δευτερεύουσα παράμετρο του συστήματος, η οποία συνήθως επιλέγεται με στόχο την καλύτερη ευελιξία του.

- αθροιστής (Summing Function): ο οποίος στην έξοδό του δίνει το άθροισμα των σταθμισμένων εισόδων.

- Η συνάρτηση ενεργοποίησης (Activation Function): από αυτήν περνά η έξοδος του αθροιστή και δίνει αποτέλεσμα στο διάστημα [0, 1] ή [-1, 1], ανάλογα με τον τύπο της συνάρτησης που επιλέχθηκε.

Σημαντικό κομμάτι για την supervised τεχνική κατηγοριοποίησης που διαλέγουμε αποτελεί η επιλογή των features. Τα features ουσιαστικά λένε στον classifier πως αναπαρίστανται τα κείμενα. Τα features που χρησιμοποιούνται περισσότερο στην κατηγοριοποίηση συναισθήματος είναι:

Terms presence and frequency: Τα features αυτά περιλαμβάνουν κυρίως unigrams ή n-grams και τη συχνότητα τους στο κείμενο, όπου n-gram αποτελεί μία ακολουθία από tokens, που εδώ είναι οι λέξεις. Είτε δίνεται σε κάθε λέξη μία δυαδική τιμή, ανάλογα με το αν εμφανίζεται ή όχι, είτε χρησιμοποιείται η συχνότητα εμφάνισης όρων για την σχετική σημασία τους στο κείμενο. Έχουν χρησιμοποιηθεί ευρέως και με επιτυχία στην κατηγοριοποίηση συναισθήματος.

Part of Speech (POS): Το POS χρησιμοποιείται για την αποσαφήνιση εννοιών που στη συνέχεια χρησιμοποιούνται για την επιλογή των features. Με το POS, κάθε όρος σε κάθε όρο μίας πρότασης θα μπαίνει μία ταμπέλα που δείχνει τη θέση και το ρόλο της στην γραμματική της πρότασης. Για παράδειγμα, μπορούμε να εντοπίζουμε επίθετα και επιρρήματα που συνήθως δείχνουν συναίσθημα.

Negations: Οι αρνήσεις είναι σημαντικό feature που πρέπει να λαμβάνεται υπ' όψη καθώς μπορεί να αλλάξει το νόημα μιας πρότασης.

Opinion words and phrases: Σε αυτήν την κατηγορία λέξεων ή φράσεων ανήκουν λέξεις ή φράσεις που χρησιμοποιούνται συχνά για να εκφράσουν θετική ή αρνητική γνώμη.

Οι τεχνικές μηχανικής μάθησης και οι αλγόριθμοι που χρησιμοποιούν ταξινομούνται σε τέσσερις κύριες κατηγορίες και χρησιμοποιούνται ανάλογα με την φύση του προβλήματος:

Επιτηρούμενη μάθηση (Supervised Learning): Συχνά αναφέρεται και ως επιβλεπόμενη μάθηση ή μάθηση με επίβλεψη. Για την τεχνική αυτή και την εκπαίδευση των αλγορίθμων απαιτείται η ύπαρξη ενός συνόλου δεδομένων με ταξινομημένη πολικότητα (Σύνολο Εκπαίδευσης – Training Set). Χρησιμοποιώντας το σύνολο εκπαίδευσης ως βάση (μοντέλο) λαμβάνουν μεγάλο όγκο δεδομένων, το

αναλύουν, εξάγουν πρότυπα και συμπεράσματα και οδηγούν σε προβλέψεις. Δημιουργείται ένα σύνολο δεδομένων εισόδου (Σύνολο Εκπαίδευσης – Training Set) στο οποίο είναι γνωστή η έξοδος. Δηλαδή για κάθε ένα από τα δεδομένα εισόδου είναι γνωστός ο συναισθηματικός προσανατολισμός (Ταξινομημένη Πολικότητα, κατηγορία, κλάση ή ετικέτα). Ο αλγόριθμος λαμβάνει ως βάση το Σύνολο Εκπαίδευσης, τροφοδοτείται με δεδομένα εισόδου άγνωστης πολικότητας, τα οποία είναι κείμενα σε φυσική γλώσσα, τα επεξεργάζεται, προσπαθεί να εξαγάγει πρότυπα που να συμφωνούν με το σύνολο εκπαίδευσης, κατηγοριοποιεί–χαρακτηρίζει τα δεδομένα εισόδου και εξάγει συμπεράσματα με χαρακτηριστικά πολικότητας που δύνανται να οδηγήσουν σε προβλέψεις. Η επιβλεπόμενη μάθηση χρησιμοποιείται σε προβλήματα Ταξινόμησης (Classification), Πρόγνωσης (Prediction) και Διερμηνείας (Interpretation). Εκπρόσωποι των αλγορίθμων επιβλεπόμενης μάθησης είναι οι αλγόριθμοι Naive Bayes και Support Vector Machine (SVM).

Ημι-επιτηρούμενη μάθηση (Semi-supervised Learning): Την συναντάμε και ως ημι-επιβλεπόμενη μάθηση. Σε αυτή την κατηγορία το Σύνολο Εκπαίδευσης περιέχει γνωστές και άγνωστες εξόδους. Ο Αλγόριθμος καλείται να εκπαιδευτεί και να αποφασίσει έχοντας μέρος της πληροφορίας ως δεδομένη. Οι αλγόριθμοι που ανήκουν σε αυτή την κατηγορίας ταξινόμησης είναι οι γράφοι.

Μη-επιτηρούμενη μάθηση (Un – supervised Learning): Αλλιώς Μη-επιβλεπόμενη μάθηση χωρίς επίβλεψη. Στον αλγόριθμο δεν παρέχεται καμία πληροφορία και πρέπει να ανακαλύψει την δομή, τα μοτίβα ή τις συσχετίσεις των δεδομένων μόνος του. Χρησιμοποιείται σε προβλήματα Ανάλυσης Συσχετισμών (Association Analysis) και Ομαδοποίησης (Clustering). Οι πιο γνωστοί αλγόριθμοι μη επιτηρούμενης μάθησης και ταξινόμησης είναι οι αλγόριθμοι συσταδοποίησης (clustering).

Ενισχυτική μάθηση (Reinforcement Learning): Στην ενισχυτική μάθηση ο αλγόριθμος μαθαίνει μέσα από την άμεση αλληλεπίδραση με το περιβάλλον. Χρησιμοποιείται σε προβλήματα Σχεδιασμού (Planning).Χρησιμοποιεί το σύστημα μάθησης της ανταμοιβής ή της τιμωρίας ανάλογα με την επιτυχή ή μη εκτέλεση της αποστολής του. Εκπρόσωποι των αλγορίθμων ενισχυτικής μάθησης είναι οι αλγόριθμοι δυναμικού προγραμματισμού που χρησιμοποιούνται συχνά σε επίλυση προβλημάτων βελτιστοποίησης

2.7 NLP Tools

Για να μπορούμε να εφαρμόσουμε το Sentiment Analysis στα Reviews που συλλέγουμε, μπορούμε να χρησιμοποιήσουμε ένα NLPtoolkit.Τα Natural Language Processing tool kits προσφέρουνε επιλογές για όλο το φάσμα των πεδίων που περιλαμβάνει το NLP και επομένως περιλαμβάνουν και τα εργαλεία εκείνα που μας επιτρέπου να κάνουμε SentimentAnalysis.Επίσης πολλά συνδυάζουν το NLP μαζί με machine learning, εστιάζοντας ιδιαίτερα στο κομμάτι αυτό.Για την εργασία μας, θα πρέπει να βρούμε κάποια εργαλεία τα οποία να μπορούμε να τα ενοποιήσουμε με το project μας και όχι κάποια τα οποία δουλεύουν μόνα τους χωρίς να μπορούμε να τα χρησιμοποιήσουμε μέσα στο project μας. Κάνοντας

μια έρευνα στο internet, τα εργαλεία όπου χρησιμοποιούνται περισσότερο για το SentimentAnalysis και μπορούν να τρέξουν σε Python είναι:

Stanford NLP: Έχει αναπτυχθεί από το NLP Group στο πανεπιστήμιο του Stanford και περιλαμβάνει ένα μεγάλο εύρος από εργαλεία ανάλυσης γλώσσας , όπως Part-Of-Speech tagging (POS),Named Entity Recognition (NER) καθώς και Sentiment Analysis (SA).Για τα περισσότερα από αυτά υποστηρίζει πολλές γλώσσες όμως για το SA υποστηρίζει μόνο Αγγλικά. Για το Sentiment Analysis χρησιμοποιείται ένα καινούργιο είδοςαπό Recursive Neural Network , όπου έχουν αναπτύξει και έχει εκπαιδευτεί σε ένα συγκεκριμένο dataset. Η άδεια χρήσης του είναι GNUGPLv3.

Lingpipe: Έχει αναπτυχθεί από την εταιρία Alias και αποτελεί μία πλατφόρμα για διάφορες εργασίες πάνω στην ανάλυση γλώσσας.Και αυτό παρέχει εργαλεία για όλα τα είδη NLP και υποστηρίζει αρκετές γλώσσες ενώ μπορούμε να του παρέχουμε συνεχώς νέα δεδομένα για εκμάθηση.Για το SentimentAnalysis, χρησιμοποιεί language model classifiers για την κατηγοριοποίηση των κειμένων.Για το featureselection χρησιμοποιούνται κυρίως τα n-gram. Η άδεια χρήσης του είναι GNUAGPL.

2.7.1 Στοιχεία Επεξεργασίας Φυσικής Γλώσσας (NLP)

Πριν κάνουμε επισκόπηση στο πώς μπορούμε να επιτύχουμε την επεξεργασία φυσικής γλώσσας (natural language processing-NLP) και την κατανόηση φυσικής γλώσσας (natural language understanding-NLU), θα παρουσιάσουμε τα οφέλη που μπορεί να επιφέρει η χρήση φυσικής γλώσσας κατά την επικοινωνία χρήστηυπολογιστή. Πλήθος τομέων μπορούν να επωφεληθούν από τη χρήση της, με κυριότερο, καταρχάς, την επικοινωνία ανθρώπου-μηχανής (human-computer interaction). Στο χώρο αυτό, η χρήση φυσικής γλώσσας επιτρέπει στους χρήστες να χρησιμοποιούν το φυσικό τρόπο επικοινωνίας τους και όχι τεχνητές γλώσσες (προγραμματισμού, μηχανής, κ.ά.) ή δομημένα μενού. Μια τέτοια προσέγγιση έχει και προτερήματα και μειονεκτήματα. Ναι μεν δεν απαιτείται εκπαίδευση στη χρήση της γλώσσας, αλλά αυτό διευκολύνει περισσότερο τους περιστασιακούς χρήστες και λιγότερο τους εξειδικευμένους, όπως είναι για παράδειγμα οι προγραμματιστές ή οι υπάλληλοι γραφείου που εισάγουν στοιχεία σε φόρμες.Μια δεύτερη περιοχή είναι αυτή της διαχείρισης πληροφορίας (information management), όπου η NLP θα μπορούσε να ενεργοποιήσει διαδικασίες αυτόματης διαχείρισης και επεξεργασίας της πληροφορίας με βάση τη διερμηνεία της. Αν, για παράδειγμα, ένα σύστημα μπορούσε να κατανοήσει το νόημα ενός εγγράφου, θα μπορούσε να το αρχειοθετήσει μαζί με τα άλλα αντίστοιχα έγγραφα.Τρίτη περιοχή είναι αυτή της αναζήτησης σε βάσεις δεδομένων (database searching). Οι συνήθεις τρόποι έκφρασης μιας επιθυμητής πληροφορίας είναι μέσω επιλογής από λίστες, συμπλήρωσης μενού ή σύνταξης του αιτήματος σε τεχνητή γλώσσα (special query language-SQL). Η χρήση τεχνητής γλώσσας επιτρέπει μεν την ανάπτυξη απλών μηχανισμών αναζήτησης, αλλά και πάλι ο χρήστης πρέπει να έχει κάποια γνώση σχετικά με τη δομή της βάσης. Από την άλλη πλευρά, ο χρήστης είναι πιο εξοικειωμένος με το περιεχόμενο ή την περιοχή ενδιαφέροντος της βάσης παρά με τη δομή της. Με τη χρήση φυσικής γλώσσας,τα αιτήματα μπορεί να περιοριστούν σε όρους σχετικούς με

το περιεχόμενο και την περιοχή ενδιαφέροντος.Είτε πρόκειται για Alexa, Siri, Google Assistant, Bixby ή Cortana, όλοι όσοι διαθέτουν smartphone ή έξυπνο ηχείο έχουν έναν βοηθό ενεργοποιημένο με φωνή. Κάθε χρόνο, αυτοί οι βοηθοί φωνής φαίνεται να γίνονται καλύτεροι στην αναγνώριση και εκτέλεση των πραγμάτων που τους λέμε να κάνουν. Αλλά αναρωτηθήκατε ποτέ πώς αυτοί οι βοηθοί επεξεργάζονται τα πράγματα που λέμε; Το καταφέρνουν να το κάνουν αυτό χάρη στην Επεξεργασία Φυσικής Γλώσσας ή στο NLP.Ιστορικά, τα περισσότερα λογισμικά μπόρεσαν να ανταποκριθούν μόνο σε ένα καθορισμένο σύνολο συγκεκριμένων εντολών. Θα ανοίξει ένα αρχείο επειδή κάνατε κλικ στο Άνοιγμα ή ένα υπολογιστικό φύλλο θα υπολογίσει έναν τύπο βασισμένο σε συγκεκριμένα σύμβολα και ονόματα τύπων. Ένα πρόγραμμα επικοινωνεί χρησιμοποιώντας τη γλώσσα προγραμματισμού στην οποία κωδικοποιήθηκε, και έτσι θα παράγει έξοδο όταν του δοθεί είσοδος που αναγνωρίζει. Σε αυτό το πλαίσιο, οι λέξεις είναι σαν ένα σύνολο διαφορετικών μηχανικών μοχλών που παρέχουν πάντα την επιθυμητή έξοδο.Αυτό έρχεται σε αντίθεση με τις ανθρώπινες γλώσσες, οι οποίες είναι περίπλοκες, μη δομημένες και έχουν πολλές έννοιες που βασίζονται στη δομή των προτάσεων, στον τόνο, στον τόνο, στο χρόνο, στα σημεία στίξης και στο πλαίσιο. Η Επεξεργασία Φυσικής Γλώσσας είναι ένας κλάδος της τεχνητής νοημοσύνης που προσπαθεί να γεφυρώσει αυτό το χάσμα μεταξύ αυτού που μια μηχανή αναγνωρίζει ως είσοδος και της ανθρώπινης γλώσσας. Αυτό συμβαίνει ώστε όταν μιλάμε ή πληκτρολογούμε φυσικά, το μηχάνημα παράγει έξοδο σύμφωνα με όσα είπαμε.Αυτό γίνεται λαμβάνοντας τεράστιες ποσότητες σημείων δεδομένων για να αντλήσουμε νόημα από τα διάφορα στοιχεία της ανθρώπινης γλώσσας, πάνω από τις έννοιες των πραγματικών λέξεων. Αυτή η διαδικασία συνδέεται στενά με την έννοια που είναι γνωστή ως μηχανική μάθηση, η οποία επιτρέπει στους υπολογιστές να μάθουν περισσότερα καθώς αποκτούν περισσότερα σημεία δεδομένων. Αυτός είναι ο λόγος για τον οποίο τα περισσότερα μηχανήματα επεξεργασίας φυσικών γλωσσών με τα οποία αλληλεπιδρούμε συχνά φαίνονται να βελτιώνονται με την πάροδο του χρόνου.

Για να φωτίσουμε καλύτερα την ιδέα, ας ρίξουμε μια ματιά στις τεχνικές που χρησιμοποιούνται στο NLP για την επεξεργασία γλώσσας και πληροφοριών.

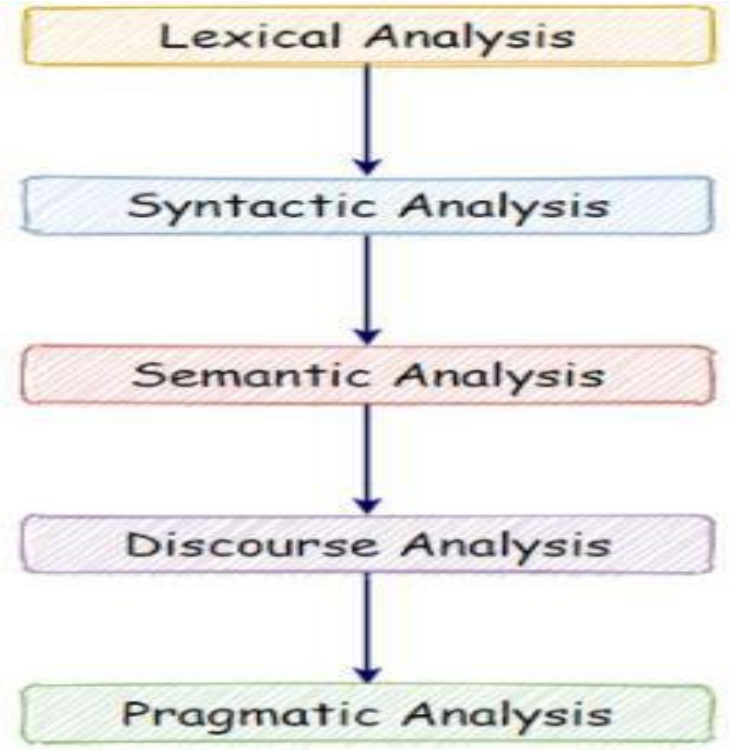
***1)Lexical Analysis** : Με τη λεξιλογική ανάλυση, χωρίζουμε ένα ολόκληρο κομμάτι κειμένου σε παραγράφους, προτάσεις και λέξεις. Περιλαμβάνει τον εντοπισμό και την ανάλυση της δομής των λέξεων.*

***2)Syntactic Analysis:** Η συντακτική ανάλυση περιλαμβάνει την ανάλυση λέξεων σε μια πρόταση για γραμματική, και την τακτοποίηση λέξεων με τρόπο που δείχνει τη σχέση μεταξύ των λέξεων*

*3)**Semantic Analysis**: Η σημασιολογικήανάλυση αντλεί την ακριβή έννοια των λέξεων και αναλύει τη σημασία του κειμένου.*

***4) Disclosure Integration:** Λαμβάνει υπόψιν το πλαίσιο του κειμένου καιεξετάζειτονόηματης πρότασης πριν τελειώσει.*

***5) Pragmatic Analysis**: Ασχολείται με τη συνολική επικοινωνία και ερμηνεία της γλώσσας.Ασχολείται επίσης και με την εκπόνηση ουσιαστικής χρήσης της γλώσσας σε διάφορες καταστάσεις.*



2.7.3 Τεχνικές Επεξεργασίας Φυσικής Γλώσσας(NLP)

Η Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing ή NLP) είναι μια ευρεία έννοια που καλύπτει οποιαδήποτε επεξεργασία ή χειρισμό της φυσικής γλώσσας, μέσω Η/Υ.Από το πιο απλό, όπως το άθροισμα των συχνοτήτων των λέξεων για τον διαχωρισμό διαφορετικών στυλ συγγραφής, έως το πολύ περίπλοκο, όπως είναι το να γίνουν "κατανοητές" ανθρώπινες εκφράσεις, τουλάχιστον στο βαθμό που να είναι δυνατόν να δοθεί κάποια λογική απάντηση σε αυτές.Ορισμένες κυρίαρχες τεχνικές στην επεξεργασία της Φυσικής Γλώσσας είναι οι παρακάτω:

1)Bag-of-Words (BoW): Το μοντέλο σάκου από λέξεις (bag-of-words model) είναι μία απλή αναπαράσταση κειμένου που χρησιμοποιείται συχνά σε διάφορες εφαρμογές Επεξεργασίας Φυσικής Γλώσσας και όχι μόνο. Σε αυτή την αναπαράσταση το κείμενο είναι ένας “σάκος” ο οποίος περιέχει όλες τις λέξεις του κειμένου χωρίς να ενδιαφέρεται για γραμματική, συντακτικό,στίξη και τη σειρά των λέξεων στο κείμενο (μόνο στη γενική του μορφή διατηρεί τις επαναλαμβανόμενες λέξεις).

2) Bag-of-N-grams: Τα n-grams είναι μία τεχνική μοντελοποίησης ακολουθιών λέξεων. Ένα μοντέλο Bag-of-Ngrams έχει την απλότητα του μοντέλου bag-of-words, αλλά επιτρέπει τη διατήρηση περισσότερων πληροφοριών, επειδή συλλαμβάνει περισσότερες πληροφορίες σχετικά με κάθε λέξη.Τα n-grams είναι μία τεχνική μοντελοποίησης ακολουθιών λέξεων. Ένα n-gram περιλαμβάνει n συνεχόμενες λέξεις. Για παράδειγμα, στη φράση «Στον Γιώργο αρέσουν οι ταινίες.», τα bi-grams (n = 2) είναι τα εξής: «Στον Γιώργο», «Γιώργο αρέσουν», « αρέσουν οι», «οι ταινίες». Αυτή η μέθοδος είναι ιδιαίτερα χρήσιμη όταν ένα κείμενο έχει θόρυβο,όταν το σύνολο των κειμένων είναι μικρό ή όταν το κείμενο έχει μεταφορικό λόγο.

3) Tokenization: Το βασικό βήμα για την αναπαράσταση ενός κειμένου είναι ο διαχωρισμός του σε διακριτές λέξεις, που μπορούν να ονομαστούν tokens. Tokenization είναι η διαδικασία διαχωρισμού του κειμένου σε λογικά κομμάτια, όπως για παράδειγμα σε λέξεις. Σε ό,τι αφορά τα κείμενα, ο διαχωρισμός γίνεται ως επί το πλείστον στα κενά και στα σημεία στίξης. Το ποια από αυτά τα tokens θα χρησιμοποιηθούν αργότερα για την αναπαράσταση του κειμένου, είναι στην ευχέρεια του ερευνητή και υπάρχουν πολλές μέθοδοι που βοηθούν σε αυτό το κομμάτι.

4)Part-Of-Speech Tagging (POS Tagging): Η σύνταξη είναι ίσως το πιο σημαντικό στοιχείο ενός κειμένου, πέρα από το λεξιλόγιο και τη φρασεολογία και μπορεί να δώσει πολύ ενδιαφέρουσες πληροφορίες για το περιεχόμενό του. Η διαδικασία κατά την οποία γίνεται κατηγοριοποίηση των
λέξεων σε μέρη του λόγου και επισημαίνονται κατάλληλα, ονομάζεται part-of-speech tagging, POS-tagging, ή απλά tagging, ή πιο απλά η επισήμανση ενός κειμένου με τα αντίστοιχα μέρη του λόγου που το απαρτίζουν.Τα μέρη του λόγου (Parts of speech) είναι γνωστά επίσης και ως κατηγορίες λέξεων (word classes) ή λεκτικές κατηγορίες (lexical categories). Το σύνολο των tags που χρησιμοποιείται για τη συγκεκριμένη διεργασία είναι γνωστό ώς tagset. Στόχος του NLP είναι η αξιοποίηση των tags αλλά και το αυτόματο tagging.Στον τομέα της ανάλυσης κειμένων αποτελεί κοινή πρακτική η αναγνώριση μερών του λόγου για κάθε λέξη. Αυτή η μέθοδος μπορεί να βοηθήσει στον εντοπισμό των πιο σημαντικών λέξεων, αλλά και σε άλλες μεθόδους όπως το stemming και η λημματοποίηση

5)Αφαίρεση Λέξεων χωρίς Αξία (Stop-words): Η αφαίρεση των λέξεων που δεν προσδίδουν κάποια αξία είναι ένα από τα πιο συνηθισμέναβήματα επεξεργασίας κειμένου. Η ιδέα είναι να αφαιρεθούν όλες οι κοινές λέξεις, οι οποίες εμφανίζονται συχνά σε μια συλλογή κειμένων. Οι λέξεις αυτές δεν έχουν καμία αξία σε κάποιες NLP εφαρμογές, κάτι που σημαίνει ότι οι λέξεις αυτές δεν είναι ιδιαίτερα σημαντικές. Υπάρχουν και περιπτώσεις βέβαια, στις οποίες οι stop-words παίζουν πάρα πολύ μικρό ρόλο και η επιρροή τους σε μια NLP εφαρμογή είναι αμελητέα. Εκτός από τις ήδη υπάρχουσες και έτοιμες λίστες από stop-words, ένας απλός τρόπος για να παραχθεί μια τέτοια λίστα, είναι με βάση τη συχνότητα μιας λέξης στα κείμενα που εξετάζονται, όπου εάν η λέξη είναι παρούσα σε όλα τα κείμενα, τότε μπορεί να χαρακτηριστεί ως stop-word. Έχει γίνει αρκετή έρευνα σχετικά με τη βέλτιστη λύση στο θέμα και κάποιες βιβλιοθήκες, όπως η NLTK [21], παρέχουν πρόσβαση σε λίστες με 2,400 stopwords για 16 γλώσσες.

6)Stemming: Το Stemming είναι η διαδικασία κατά την οποία εντοπίζεται η ρίζα της λέξης ή η λέξη μετασχηματίζεται σε μια πιο απλή μορφή [19]. Αυτή η μέθοδος είναι πολύ χρήσιμη στον τομέα της ανάλυσης συναισθήματος, ώστε να μειώνονται οι διαστάσεις του training set,δηλαδή να μειώνονται οι λέξεις. Ο λόγος είναι ότι οι διαφορετικές γραμματικές μορφές μιας λέξης πολλές φορές δεν αλλάζουν το νόημά της. Για παράδειγμα η λέξη “γράφω” και η λέξη “γράφοντας” έχουν ουσιαστικά την ίδια βαρύτητα, αλλά χωρίς τη διαδικασία stemming θα υπήρχαν δύο φορές μέσα στο training set

2.7.4 Δυσκολίες στην Κατανόηση Φυσικής Γλώσσας(NLP)

Η μεγαλύτερη δυσκολία στην επεξεργασία φυσικής γλώσσας είναι η διφορούμενη ερμηνεία που προκαλεί ασάφεια στη γλώσσα (ambiguity of language) σε πολλά επίπεδα:Καταρχάς, ασάφεια σε επίπεδο σύνταξης (ambiguity at syntactic level) της γλώσσας.Κάποιες συντακτικά ορθά προτάσεις επιδέχονται πάνω από μια διερμηνεία, ανάλογα με τοπώς θα αναλυθούν συντακτικά, καθιστάμενες συντακτικά ασαφείς.

Για παράδειγμα:Χτύπησα τον κλέφτη με το τσεκούρι. Το τσεκούρι ήταν το όπλο με το οποίο χτύπησα τονκλέφτη ή χτύπησα τον κλέφτη που κρατούσε το τσεκούρι;

Δευτερευόντως, ασάφεια σε επίπεδο λεξιλογικό (ambiguity at lexical level), όταν το νόημαμιας λέξης είναι διφορούμενο.

Για παράδειγμα:Το πρώτο γράμμα του Γιώργου. Εννοεί το πρώτο γράμμα που έγραψε ο Γιώργος ή το γράμμα του αλφαβήτου από το οποίο αρχίζει το όνομα «Γιώργος»; Η λέξη γράμμα έχει δυο έννοιες,της επιστολής και του γράμματος του αλφαβήτου.

Τρίτον, ασάφεια σε αναφορικό επίπεδο (ambiguity at referential level), όταν δεν είναι ευκρινές το σε ποιον, πού ή σε τι η πρόταση αναφέρεται.

Για παράδειγμα:Ο Γιάννης χτύπησε τον Πέτρο, γιατί του αρέσει η Μαίρη. Σε ποιον αρέσει η Μαίρη, στο Γιάννη ή στον Πέτρο;

Τέταρτον, ασάφεια σε σημασιολογικό επίπεδο (ambiguity at semantic level), όταν, με διατήρηση της ίδιας συντακτικής ανάλυσης, η πρόταση επιδέχεται τουλάχιστον δυο διαφορετικές ερμηνείες.

Για παράδειγμα:Τον άφησε στα κρύα του λουτρού. Η πρόταση κυριολεκτεί ότι κάποιος άφησε κάποιον άλλον στα κρύα ενός λουτρού ή παρουσιάζει μεταφορικά ότι τον παράτησε και έφυγε στη μέση κάποιας συνεργασίας; Τέλος, ασάφεια σε πραγματολογικό επίπεδο (pragmatic level), κατά τη διερμηνεία μιας πρότασης, όταν λαμβάνουμε υπόψη το πλαίσιο του κειμένου που την περιέχει.

Στην παρακάτω πρόταση δεν είναι εύκολο να λυθεί η πραγματολογική ασάφεια: Οι δεινόσαυροι έχουν εξαφανιστεί πολλά χρόνια. Πόσα χρόνια είναι τα πολλά χρόνια; Οι επιμέρους ασάφειες μπορεί να συνδυαστούν σε μια μόνη πρόταση.

Για παράδειγμα, η παρακάτω πρόταση περιέχει ασάφεια σε λεξιλογικό, αναφορικό και πραγματολογικό επίπεδο:Ο Νίκος ζήτησε από τον Ηλία να τον αντικαταστήσει στη δουλειά σήμερα, πριν φύγει ταξίδι.Υπάρχει αμφιβολία σχετικά με την απόδοση του χρονικού προσδιορισμού "σήμερα" σε ένα από τα δύο ρήματα, όπως και για το ποιος θα φύγει ταξίδι (πραγματολογικό και αναφορικό επίπεδο).Ο Γιάννης χτύπησε εχθές τον Πέτρο, που είναι γάιδαρος, επειδή του αρέσει η Μαίρη. Σε επίπεδο σημασιολογικής ασάφειας δεν ξέρουμε αν ο Πέτρος είναι ζώο γάιδαρος ή άνθρωπος που συμπεριφέρεται ως γάιδαρος. Από πλευράς αναφορικής δεν ξέρουμε σε ποιον αρέσει η Μαίρη και από πλευράς πραγματολογικής δεν ξέρουμε το πότε είναι το «χθες».

2.8 Αξιολόγηση Υπηρεσιών από Σχόλια πελάτων

2.8.1 Αξιολόγηση Ποιότητας Υπηρεσιών

Η αξιολόγηση της ποιότητας των υπηρεσιών είναι μια υποκειμενική, κατά κύριο λόγο, διαδικασία, αφού ο κάθε πελάτης, καταναλωτής και χρήστης υπηρεσιών έχει διαφορετικά κριτήρια για το τί θεωρείται ποιοτικό και τί όχι. Σχετικά με τον ξενοδοχειακό κλάδο, η ποιότητα στις υπηρεσίες που προσφέρουν τα ξενοδοχεία αποτελεί ανταγωνιστικό και σε πολλές περιπτώσεις στρατηγικό ανταγωνιστικό πλεονέκτημα. Η σε βάθος κατανόηση των αναγκών των πελατών των ξενοδοχείων από την διοίκηση, ο ορισμός ελαχίστων προδιαγραφών ποιότητας και η σημασία που αποδίδεται σε κάθε υπηρεσία ανά τμήμα, συνθέτουν το πλαίσιο μέσα στο οποίο κινούνται οι ενέργειες για αξιολόγηση και προσπάθεια μέτρησης της ποιότητας των υπηρεσιών.

Λόγω του ανταγωνισμού στον ξενοδοχειακό κλάδο τόσο σε εθνικό όσο και σε διεθνές επίπεδο, η τακτική μέτρηση της ποιότητας των υπηρεσιών με συγκεκριμένα εργαλεία, η διόρθωση των αποκλίσεων και προβληματικών υπηρεσιών, η εξασφάλιση θετικής εικόνας από τους πελάτες και ο έλεγχος της αντίληψης των πελατών για το τι είναι ποιοτικό κατά την κρίση τους, είναι σημαντικοί παράμετροι που καθιστούν αναγκαία την ύπαρξη ενός συστήματος μέτρησης και αξιολόγησης της ποιότητας που αντιλαμβάνονται οι πελάτες ότι τους προσφέρεται.

Ο αντικειμενικός χαρακτήρας της αξιολόγησης της ποιότητας των υπηρεσιών σε ένα ξενοδοχείο αφορά στοιχεία που δεν σχετίζονται με την άποψη και τη στάση του πελάτη και είναι εντελώς ανεξάρτητα από τις προσωπικές προσδοκίες του κάθε πελάτη. Παρατηρούνται στον ξενοδοχειακό κλάδο γνωρίσματα τεχνικής φύσεως όπως είναι η διακόσμηση και το μέγεθος των δωματίων όπως επίσης και η πληθώρα των μέσων εκγύμνασης που περιλαμβάνει το τμήμα γυμναστηρίου ενός ξενοδοχείου, γνωρίσματα χρονικά που σχετίζονται με τον χρόνο παραμονής στο ξενοδοχείο των πελατών, υπάρχουν τυπικά γνωρίσματα αξιολόγησης της ποιότητας όπως είναι η ανταπόκριση των όσων αναφέρονται σε διαφημιστικές μπροσούρες και φυλλάδια με τα όσα παρέχονται στο ξενοδοχείο εν τέλει και τέλος γνωρίσματα που αφορούν την παροχή υπηρεσιών και μπορούν να είναι από την λειτουργία υποδοχής έως και την λειτουργία της καθαριότητας των δωματίων κλπ.

Η υποκειμενική φύση της αξιολόγησης της ποιότητας των υπηρεσιών άπτεται της ιδιοσυγκρασίας, της προσωπικότητας και των επιθυμιών που εκφράζει ένας πελάτης. Για αυτό το λόγο δημιουργούνται προβλήματα σχετικά με την δημιουργία ενός αξιόπιστου μετρήσιμου αποτελέσματος αξιολόγησης ποιότητας υπηρεσιών, το οποίο θα μπορεί να αξιοποιηθεί οποιαδήποτε χρονική στιγμή από διάφορους εμπλεκόμενους φορείς στον ξενοδοχειακό κλάδο και πιο συγκεκριμένα από το ξενοδοχείο στο οποίο πραγματοποιείται η εκάστοτε μέτρηση.

2.8.2 ρόλος των Online Reviews

Οι ενδιάμεσες τουριστικές ιστοσελίδες μέσω των οποίων ο καταναλωτής μπορεί να κάνει κράτηση σε ένα ξενοδοχείο, διευκολύνουν τον χρήστη λόγω του περιεχομένου το οποίο μπορεί να διαμοιράζεται με άλλους χρήστες καθώς και με την μορφή σχολίων να αξιολογεί θετικά ή αρνητικά κάποιο ξενοδοχείο που έχει επισκεφθεί. Ενώ οι αξιολογήσεις αυτές ποικίλλουν σε μορφή, η πλειοψηφία τους βασίζεται στο user-generatedπαραδοσιακό σύστημα της αξιολόγησης των αστεριών. Ωστόσο, μπορούν επίσης να βασιστούν μόνο στις αντιλήψεις των ταξιδιωτών αντί να χρησιμοποιούν σαφή κριτήρια που χρησιμοποιούνται στο σύστημα παραδοσιακής αξιολόγησης. Το user-generated περιεχόμενο μπορεί να θεωρηθεί ως μια μορφή ηλεκτρονικού WOM(Word of mouth).

Σε όρους μάρκετινγκ, το user-generated περιεχόμενο στις ιστοσελίδες είναι μια αποτελεσματική μέθοδος e-marketingαπό καταναλωτή σε καταναλωτή. Οι ιστοσελίδες και τα μέσα κοινωνικής δικτύωσης έχουν αλλάξει το πεδίο της επικοινωνίας “WOM”. Ενώ στο παρελθόν αυτή η τακτική Word of Mouth (από στόμα σε στόμα) βασίστηκε σε ανθρώπους που μιλούν μεταξύ τους ή αποτελούν καθοδηγητές γνώμης για κάποιες μικρές ομάδες ατόμων, σήμερα, το Διαδίκτυο έχει επεκταθεί και ουσιαστικά μεταλλάξει το “Word -of-mouth” σε ένα τεράστιο μέσο επικοινωνίας μέσα σε προκαθορισμένες ομάδες, φίλους, ή χιλιάδες ξένους συνδεδεμένους όμως μεταξύ τους σε διαδικτυακές κοινότητες.

Οι καταναλωτές δίνουν μεγαλύτερη εμπιστοσύνη σε σχόλια από τους ταξιδιώτες για τα ταξίδια τους και αυτά έχουν μεγαλύτερο αντίκτυπο στις πωλήσεις των ξενοδοχειακών επιχειρήσεων, ενώ εξετάζοντας τις εμπειρίες των άλλων καταναλωτών από τα σχόλια τους και από άλλο υλικό είναι η πιο δημοφιλής πηγή πληροφόρησης. Μέσα από διάφορες έρευνες που έχουν γίνει, έχει διαπιστωθεί ότι τα δημοσιευμένα σχόλια των χρηστών φαίνεται να λειτουργούν ως μια πρόσθετη πηγή των πληροφοριών που οι τουρίστες που αναζητούν ένα ξενοδοχείο το θεωρούν μέρος της διαδικασίας συλλογής πληροφοριών και όχι απλά μια πηγή πληροφοριών. Τα κίνητρα για τους χρήστες να δημοσιεύουν στα social media τα διάφορα σχόλια τους και τις κριτικές τους έχουν διερευνηθεί από διάφορους ερευνητές και τα αποτελέσματα δείχνουν ότι οι αρνητικές εμπειρίες είναι πιο πιθανό να παρακινήσουν τους δυσαρεστημένους καταναλωτές να δημοσιεύσουν κάτι σχετικά στο Διαδίκτυο. Άλλες παρεμφερείς έρευνες επικεντρώθηκαν στο πώς τα χαρακτηριστικά των ταξιδιών (η εξοικείωση με τους προορισμούς, το μήκος του ταξιδιού, η θέση των προορισμών) επηρεάζουν την επιλογή του περιεχομένου που

δημιουργείται από τους χρήστες. Τα αποτελέσματα έδειξαν ότι τα χαρακτηριστικά των ταξιδιών διαδραματίζουν θεμελιώδη ρόλο στην επιλογή του περιεχομένου που δημιουργείται από χρήστες.

Η έννοια της ποιότητας γενικότερα αφορά ένα ευρύ φάσμα διαδικασιών, λειτουργιών, υπηρεσιών, προϊόντων τα οποία έχουν ως απώτερο στόχο την καλύτερη και πληρέστερη ικανοποίηση των πελατών – χρηστών, οι οποίοι με την σειρά τους εκφράζουν ανάγκες και επιθυμίες σε διάφορες χρονικές καταστάσεις της ζωής τους. Οι αυξημένες απαιτήσεις των καταναλωτών και ο οξυμένος ανταγωνισμός μεταξύ των επιχειρήσεων που προσφέρουν προϊόντα και υπηρεσίες, οδηγούν στην δημιουργία ποιοτικότερων προϊόντων και υπηρεσιών, δηλαδή σε προϊόντα και υπηρεσίες με χαρακτηριστικά που ικανοποιούν τις επιθυμίες των καταναλωτών.

Λαμβάνοντας υπόψη ότι η ποιότητα είναι από τις πιο σημαντικές συνιστώσες για επιτυχία μιας επιχειρηματικής δραστηριότητας, γίνεται κατανοητό ότι η κατανόηση της από την πλευρά των επιχειρήσεων, η πολύπλευρη ανάγνωση της και η τακτική μέτρηση της, είναι στοιχεία που πρέπει να εξετάζονται από την διοίκηση μιας επιχείρησης, για να υπάρχει αποδοτική και ικανοποιητική αξιοποίηση της έννοιας.

Ωστόσο για να προσεγγισθεί με πιο πλήρη τρόπο η έννοια της ποιότητας, έχουν αναπτυχθεί διάφορες θεωρίες που αφορούν τον τρόπο ανάλυσης της ποιότητας. Μία εξ αυτών αναλύει την ποιότητα κάτω από πέντε διαστάσεις. Αρχικά, η ποιότητα έχει υπερβατική αξία καθώς ο καθένας έχει προσωπική και υποκειμενική άποψη για την ποιότητα σε ένα προϊόν και υπηρεσία η οποία από χρονική στιγμή σε χρονική στιγμή ενδέχεται να αλλάξει. Εν συνέχεια η ποιότητα μπορεί να εξετασθεί με βάση το προϊόν, αφού διαφορές στην ποιότητα σημαίνουν διαφορές στην ποσότητα κάποιου συστατικού ή κάποιας ιδιότητας που απαντάται σε ένα προϊόν ή υπηρεσία. Ακολούθως, η έννοια της ποιότητας έχει υποκειμενική χροιά και ενδέχεται κατά την μέτρηση της να υπάρξουν μεροληπτικές ενέργειες ενώ ως τέταρτη διάσταση θεωρείται η ποιότητα που προσδίδεται κατά την κατασκευή σε ένα προϊόν και τέλος ως Πέμπτη διάσταση θεωρείται ότι είναι η ψυχολογική διάσταση της αξίας ενός προϊόντος, καθώς ένα προϊόν όπως πχ ένα smart phone αξίας 250 ευρώ, μπορεί για έναν καταναλωτή να είναι ποιοτικό βάση των αναγκών του, όχι όμως εξίσου ποιοτικό για έναν άλλο καταναλωτή με διαφορετικές ανάγκες.

Τέλος, η ποιότητα τόσο στον ξενοδοχειακό κλάδο που εξετάζουμε, όσο και γενικά στον τρόπο γίνεται αντιληπτή από τους καταναλωτές, επηρεάζεται από μια σειρά παραγόντων και διαστάσεων όπως είναι οι κάτωθι επιδόσεις, χαρακτηριστικά, γνωρίσματα, αξιοπιστία, συμμόρφωση με προδιαγραφές, ανθεκτικότητα, συντηρησιμότητα, αισθητική.

ΚΕΦΑΛΑΙΟ 3 - Αξιολόγηση της Ποιότητας Ξενοδοχειακών Υπηρεσιών

3.1 Προ Επεξεργασία Δεδομένων

Για τη παρούσα διπλωματική εργασία , μας δόθηκε ένα Dataset σε μορφή Excel το οποίο αποτελείται από 10276 σχόλια πελατών.ToDataset αναφέρεται σε αληθινά σχόλια πελατών των ξενοδοχειακών μονάδων και η πηγή είναι το Trivago.

Αποτελείται από 3 σελίδες.

Στη Πρώτη σελίδα αναφέρονται τα εξής:

Hotel Name, Hotel Class, Hotel Location, Traveler Ranking , Hotel Rating, Reviews

Hotel Name	Hotel Class	Hotel Location	Traveler Ranking	Hotel Rating	Reviews
Palazzo Manfredi - Relais & Chateaux	5.0 Stars Hotel	Rome	#36 of 1,274 Hotels in Rome	4.5 of 5 bubbles	489 reviews
Hotel Lord Byron	5.0 Stars Hotel	Rome	#45 of 1,274 Hotels in Rome	4.5 of 5 bubbles	681 reviews
Hotel Splendide Royal	5.0 Stars Hotel	Rome	#50 of 1,274 Hotels in Rome	4.5 of 5 bubbles	1,451 reviews
Hotel De Russie	5.0 Stars Hotel	Rome	#55 of 1,274 Hotels in Rome	4.5 of 5 bubbles	1,421 reviews
Aldrovandi Villa Borghese	5.0 Stars Hotel	Rome	#58 of 1,274 Hotels in Rome	4.5 of 5 bubbles	1,412 reviews
NH Collection Roma Palazzo Cinquecen	5.0 Stars Hotel	Rome	#62 of 1,274 Hotels in Rome	4.5 of 5 bubbles	748 reviews
Rome Cavalieri, A Waldorf Astoria Reso	5.0 Stars Hotel	Rome	#66 of 1,274 Hotels in Rome	4.5 of 5 bubbles	6,060 reviews
Boscolo Exedra Roma, Autograph Colle	5.0 Stars Hotel	Rome	#78 of 1,274 Hotels in Rome	4.5 of 5 bubbles	2,962 reviews
Sofitel Rome Villa Borghese	5.0 Stars Hotel	Rome	#88 of 1,274 Hotels in Rome	4.5 of 5 bubbles	2,380 reviews
Gran Melia Rome	5.0 Stars Hotel	Rome	#89 of 1,274 Hotels in Rome	4.5 of 5 bubbles	1,504 reviews

Η Δεύτερη σελίδα αποτελείται από:

Hotel's Name, Room Type, Check-in, Check out, Best Value Ranking, First Deal Provider, First Price ,Second Deal Provider , Second Price , Third Deal Provider, Third price, Hotel's Class , Hotel's Location

Hotel's Name	Room Type	Check-in	Check-out	Best Value Ranking	First Deal Provider	First Price	Second Deal Provider	Second Price	Third Deal Provider	Third Price	Hotel's Class	Hotel's Location
Aphrodite Hotel	1 Room, 2 Adults, 0 Children	15 March	21 March	#27 Best Value of 1,271 hotels in Rome	Agoda.com	€207	Hotels.com	€288	Expedia.com	€288	4 Stars	Rome
Residenza Argentina	1 Room, 2 Adults, 0 Children	15 March	21 March	#28 Best Value of 1,271 hotels in Rome	Agoda.com	€571	Agoda.com	€571	TripAdvisor	€481	4 Stars	Rome
Hotel Memphis	1 Room, 2 Adults, 0 Children	15 March	21 March	#29 Best Value of 1,271 hotels in Rome	Agoda.com	€364	Agoda.com	€364	TripAdvisor	€364	4 Stars	Rome
Hotel Diana Roof Garden	1 Room, 2 Adults, 0 Children	15 March	21 March	#30 Best Value of 1,271 hotels in Rome	Agoda.com	€339	Expedia.com	€435	Hotels.com	€435	4 Stars	Rome
Hotel Romanico Palace	1 Room, 2 Adults, 0 Children	15 March	21 March	#31 Best Value of 1,271 hotels in Rome							4 Stars	Rome
Hilton Garden Inn Rome C	1 Room, 2 Adults, 0 Children	15 March	21 March	#32 Best Value of 1,271 hotels in Rome	Booking.com	€104	Roomdi.com	€140	TripAdvisor	€104	4 Stars	Rome
Hotel Mondial	1 Room, 2 Adults, 0 Children	15 March	21 March	#33 Best Value of 1,271 hotels in Rome	Booking.com	€88	Hotels.com	€88	Expedia.com	€88	4 Stars	Rome
Leon's Place Hotel	1 Room, 2 Adults, 0 Children	15 March	21 March	#34 Best Value of 1,271 hotels in Rome	Booking.com	€146	Hotels.com	€146	TripAdvisor	€146	4 Stars	Rome
Albergo del Sole Al Panth	1 Room, 2 Adults, 0 Children	15 March	21 March	#35 Best Value of 1,271 hotels in Rome	Booking.com	€193	Hotels.com	€193	TripAdvisor	€193	4 Stars	Rome
Gambrinus Hotel	1 Room, 2 Adults, 0 Children	15 March	21 March	#36 Best Value of 1,271 hotels in Rome	Booking.com	€100	Expedia.com	€100	Hotels.com	€100	4 Stars	Rome

Τέλος η Τρίτη σελίδα περιέχει:

Review's Title , Reviewer's Location , Full Review , Rating , Hotel's Name , Hotel's Location , Hotel's Class

Review's Title	Reviewer's Location	Full Review	Rating	Hotel's Name	Hotel's Location	Hotel's Class
Family of four 10 and 13 year old	Adelaide, Australia	We have travelled a fair bit and this is an ok place to stay. it is	3 of 5 bubbles	Adelmar Hotel and Suites	Mykonos	4 Stars
Beautiful Hotel with Superb Staff	London, United Kingdom	I was in a group of three girls staying for 8 nights (sharing 1 roor	5 of 5 bubbles	Adelmar Hotel and Suites	Mykonos	4 Stars
Great hotel	Berlin, Germany	We have stayed there for 2 nights in September 2014. The hot	4 of 5 bubbles	Hotel degli Imperatori	Rome	4 Stars
Helpful staff in a precarious situation	Stockholm, Sweden	I had booked another hotel outside Rome by Expedia.com. Howe	3 of 5 bubbles	Park Hotel Roma Cassia	Rome	4 Stars
Avoid like the plague	Lausanne, Switzerland	I stayed a week at this hotel, and it was the worst experience of	1 of 5 bubbles	Minos Mare Royal	Crete	5 Stars
Dreadful location!	England	The location of this hotel is in the middle of a rubbish filled, dis-u	1 of 5 bubbles	Ardeatina Park Hotel	Rome	4 Stars
Amazing location and great staff	Pretoria, South Africa	Could not have asked for a better location - it's so central to all t	5 of 5 bubbles	Emporikon Athens Hotel	Athens	4 Stars
Very nice staff	barcelona	I have to say that the hotel is not a luxury one and it's not locate	4 of 5 bubbles	Art Hotel	Athens	4 Stars
Basic, but clean	Cheltenham, United Kingd	Spent one evening here before heading off island hopping. Roon	3 of 5 bubbles	Novus City Hotel	Athens	4 Stars
Great location and service	Busselton, Australia	I stayed at this hotel on 3 separate occasions while touring Gree	4 of 5 bubbles	Golden Age Hotel Athens	Athens	4 Stars
Ancient Greek Tour - Cosmos Holidays	Gloucester, United Kingdo	We found this hotel to be really good. Friendly helpful staff, goo	4 of 5 bubbles	Golden Age Hotel Athens	Athens	4 Stars
Perfect as a base for touring Greece	Ithaca, New York	A modestly priced hotel with a million-dollar view of the Acropolis	5 of 5 bubbles	Acropolis View Hotel	Athens	4 Stars
Nice location, hotel and staff	Sheffield	Clean, modern and central. The quiet room did not have a windov	4 of 5 bubbles	Coco-Mat Hotel Athens	Athens	4 Stars
Excellent rooms,wonderful staff,superb rest	Afula, Israel	Spacious and comfortable rooms and bathrooms, a lot of space	5 of 5 bubbles	Daios Luxury Living	Thessaloniki	5 Stars
Perfect position, in the heart of the city	Paphos, Cyprus	We were there for a four night city break. The hotel was clean, c	4 of 5 bubbles	City Hotel Thessaloniki	Thessaloniki	4 Stars
Phenomenal restaurant!	Riyadh, Saudi Arabia	We were amazed and delighted to find world-class cuisine in this	5 of 5 bubbles	Archipelagos	Mykonos	5 Stars

Για την αξιολόγηση της ποιότητας των ξενοδοχειακών υπηρεσιών χρησιμοποιήσαμε τα Rating και τα Full Review από τη σελίδα 3.Στη παραπάνω εικόνα φαίνονται μέσα απο το excel τα δεδομένα του προγράμματος.

3.2 Εξαγωγή δεδομένων

Χρησιμοποιώντας βασικές συναρτήσεις και , ίσως την βασικότερη βιβλιοθήκη της Python για επεξεργασία αρχείων, **pandas** , καταφέραμε να εξάγουμε συγκεκριμένα δεδομένα (tabs) από το αρχείο που μας δόθηκε από τον επιβλέπων καθηγητή, ώστε να τα επεξεργαστούμε και να βγάλουμε χρήσιμα συμπεράσματα που σας αναλύονται παρακάτω.

	Review	Score	Bubbles	Result
0	Nice. Brilliant location opposite the cathedral. Bed and linen ideal for a good nights sleep. Good combination of design in nec	0,987	5	ABSOLUTELY PERFECT
1	The upscale hotel Daios has much to offer its guests. It is located on the waterfront not far from the White Tower and nearby	0,6124	4	PERFECT
2	Nice hotel with friendly staff and free parking near the see shore. Central location. Hotel is more like a boutique hotel of 2 flo	0,7933	4	ABSOLUTELY PERFECT
3	I love this hotel, stayed here last year and repeated this year for the same trip after having such an amazing experience. It is f	0,987	5	ABSOLUTELY PERFECT
4	Good Hospitality & Friendly Recepction Front desk athina antoniou was so friendly and helpfull the hotel has an amazing view	0,9608	5	ABSOLUTELY PERFECT
5	If you want a hotel walking distance from town but don't want to be in the center of town, then this place is good. Not much c	0,8504	3	ABSOLUTELY PERFECT
6	We stayed at San Antonio Summerland 4 nights. We had arrange with the hotel shuttle to pick up us. It was confirmed twice. /	-0,5471	2	DISASTER
7	We stayed in a Panoramic Double Room only to find that we had two singles beds pushed together and a wooden boarder on	0,9743	3	ABSOLUTELY PERFECT
8	It was unbelievable experience!! Very smooth & fast check in and out, super friendly environment and great ppl! Great view	0,9969	5	ABSOLUTELY PERFECT
9	We've just spent a week here and can't agree more with the reviews that highlight the quality of the staff at the Tharroe. Iren	0,9858	5	ABSOLUTELY PERFECT
10	Wow.....what can we say to give you a real insight into petinos hotel. Well in the last five weeks we have travelled through I	0,9865	5	ABSOLUTELY PERFECT
11	A nice hotel with lovely interior and helpful staff, located 15 minutes drive from the Mykonos town. In my opinion the only p	0,7906	4	ABSOLUTELY PERFECT
12	Mykonos should be so proud for having Kouros as on of their hotels!!! We are a family of 4 from USA with kids aged 7 and 9 an	0,9985	5	ABSOLUTELY PERFECT
13	One world on its own ..amazing ! Not possible to explain how relaxing how fantastic looking at a perfect sunset get the very b	0,9768	5	ABSOLUTELY PERFECT
14	I wouldn't say it's luxurious but it definitely pays attention to details. Little things matter, amenities at the room take into cor	0,8962	5	ABSOLUTELY PERFECT
15	Just enjoyed a week at this beautiful spot. One of the very best places we have stayed...great location, quality and service. T	0,9844	5	ABSOLUTELY PERFECT
16	We stayed for 4 nights and the first impression we had was "wow". The pool and bar area with an overlooking view of the oce	0,9884	5	ABSOLUTELY PERFECT
17	We stayed for 3 nights in the deluxe sea view room, paid €540 per night. 1. Our view was a sea view but the view is not what	0,9746	3	ABSOLUTELY PERFECT
18	My husband and I have just returned from a lovely week away at Mykonos Essence. We had booked a deluxe sea view room a	0,9951	5	ABSOLUTELY PERFECT
19	We had a lovely time at this hotel. The staff are all pleasant and helpful. Anastasia was a delight to deal with and couldn't do	0,9915	4	ABSOLUTELY PERFECT
20	Nice hotel, great location. The concierge was incredibly helpful and planned all of our reservations and transfers. When you	0,897	4	ABSOLUTELY PERFECT
21	As everything else in Mykonos, the hotel is convenient but too expensive for what it offers. The room we stayed in was small	0,997	3	ABSOLUTELY PERFECT
22	Marios, the owner, was amazing in providing guidance about places to see and restaurants to eat at. Spotless hotel with nice a	0,9299	5	ABSOLUTELY PERFECT

Η παραπάνω φωτογραφία είναι μια ενδεικτική, καθώς το .xsl αρχείο είναι αρκετά μεγάλο για να παρουσιαστεί εδώ (10.000+ εγγραφές), ώστε να μπορέσουμε να σας αναλύσουμε τι ακριβώς είναι η κάθε στήλη του παραπάνω αρχείου αλλά και πως ακριβώς προέκυψε και ποια η χρησιμότητα της στην διπλωματική μας εργασία.

Η Π**ρώτη στήλη** είναι ένας αύξων αριθμός οπού μετράει τα σχόλια και δεν κάνει τίποτα παραπάνω, μάλιστα το συγκεκριμένο είναι ένα add-on του excel καθώς, την συγκεκριμένη στήλη δεν την δημιουργήσαμε εμείς, έγινε auto-generated μόνη της.

Η **Δεύτερη στήλη** περιέχει το σχόλιο το οποίο έχουμε πάρει από το dataset που μας δόθηκε και ο λόγος που το πήραμε ήταν επειδή το dataset που μας δόθηκε είναι αρκετά μεγάλο και υπήρχαν αρκετά tabs τα οποία ξέραμε εξ' αρχής ότι δεν θα τα χρησιμοποιήσουμε, σκεφτήκαμε ότι θα ήταν ιδανικότερο για εμάς να κρατήσουμε όλα αυτά που χρειαζόμαστε σε ένα δικό μας excel αρχείο ώστε να μπορούμε να το επεξεργαστούμε με μεγαλύτερη ευκολία και άνεση.

Η **Τρίτη στήλη** περιέχει το Sentiment Score του κάθε σχολίου, δηλαδή πόσο θετικό ή αρνητικό είναι σύμφωνα με μια βιβλιοθήκη της Python, την Vader Sentiment, πιο αναλυτικά παραδείγματα για τη συγκεκριμένη στήλη θα παρατεθούν παρακάτω.

Η Τ**έταρτη στήλη** περιέχει τα αστερία που έχουν δοθεί από τον χρήστη,τα αστερία είναι ένα είδος γρήγορης αξιολόγησης. Στο dataset που μας δόθηκε, η συγκεκριμένη στήλη ήταν της μορφής

'1of 5 bubbles', '2 of 5 bubbles','3 of 5 bubbles', '4 of 5 bubbles', '5 of 5 bubbles'

Εμείς αυτό που κάναμε ήταν να απομονώσουμε το πρώτο χαρακτήρα κάθε γραμμής από την συγκεκριμένη στήλη και να τον βάλουμε στο δικό μας,στην ουσία το κάναμε έτσι γιατίχρειαζόμασταν μόνο το πρώτο γράμμα, το οποίο υποδεικνύει τον αριθμό με αστερία που βαθμολογεί ο κάθε χρήστης το ξενοδοχείο.

Η **Πέμπτη και τελευταία στήλη** μας υποδεικνύει ένα δικό μας σχόλιο, βασισμένο στο Sentiment Score που βρίσκεται στην δεύτερη στήλη, χωρισμένο σε 5 διαφορετικά επίπεδα. Πιο αναλυτικά, θα μιλήσουμε για το εύρος διαστήματος κάθε επιπέδου παρακάτω.

Χρησιμοποιώντας συγκεκριμένες μεθόδους και τεχνικές της Python, τις οποίες θα γράψουμε αναλυτικά στο κεφάλαιο 4, καταλήξαμε στο παραπάνω αποτέλεσμα που αποτυπώνεται στον πίνακα.

CRITERIA/ SENTIMENT	LOCATION	STAFF	BREAKFAST	QUIET	BED	CLEANLINESS	ROOM SPACE	PARKING	INTERIOR DESIGN
REVIEWS COUNT	2.822	5.590	3.877	1.074	2.435	4.096	101	307	911
MATCHING OVER 90%	2.783	5.547	3.748	1005	1.244	3.555	80	300	404

Θέσαμε εξ ‘αρχής κάποια κριτήρια, ταοποία μετά από έρευνα - τουλάχιστον ένα από αυτά – απασχολούν τους χρήστες σε κάθε τους αξιολόγηση. Συνεπώς, πολλές φορές μπορεί να υπάρξει συνδυασμός των παραπάνω κριτηρίων από κάθε πελάτη στην κάθε αξιολόγηση που κάνει.

Τα κριτήρια που θέσαμε, είναι τα εξής:

- **Τοποθεσία (Location)**

  Αναφέρεται στην τοποθεσία που βρίσκεται το κάθε κατάλυμα.

- **Προσωπικό (Staff)**

  Αναφέρεται στο προσωπικό που εργάζεται στο κατάλυμα και ο χρήστης αλληλοεπιδρά έμμεσα ή άμεσα μαζί του.

- **Πρωϊνό (Breakfast)**

  Αναφέρεται στην υπηρεσία του πρωϊνού, εάν προσφέρεται, από το κατάλυμα.

- **Ησυχία (Quiet)**

  Αναφέρεται στο κατά πόσο το δωμάτιο και γενικότερα το ξενοδοχείο, οι εγκαταστάσεις, ‘παρέχουν’ αρκετή ησυχία στους καταναλωτές.

- **Κρεβάτι (Bed)**

  Αναφέρεται στην ποιότητα των κρεβατιών έτσι όπως τα έχει αξιολογήσει ο χρήστης.

- **Καθαριότητα (Cleanliness)**

  Αναφέρεται, ίσως σε ένα από τα πιο σημαντικά κριτήρια, στην καθαριότητα του δωματίου και γενικότερα στο κατά πόσο καθαρό είναι το κατάλυμα σαν σύνολο.

- **Χώρος δωματίου (Room Space)**

  Αναφέρεται στο μέγεθος του δωματίου και το κατά πόσο ικανοποίησε τον χρήστη. Επίσης, γίνεται αναφορά και στην αναλογία Τιμής/Μέγεθος δωματίου σε αρκετά σχόλια χρηστών.

- **Χώρος στάθμευσης (Parking)**

  Αναφέρεται στην υπηρεσία στάθμευσης οχημάτων, ή αλλιώς χώρος parking, μια υπηρεσία η οποία δεν παρέχεται σε αρκετά ξενοδοχεία.

- **Εσωτερική Διακόσμηση (Interior Design)**

  Αναφέρεται στην εσωτερική διακόσμηση του δωματίου και γενικότερα του καταλύματος.

Γιατί όμως μπήκαμε στην παραπάνω διαδικασία έρευνας τέτοιων κριτηρίων και πως τα συγκεκριμένα κριτήρια επηρεάζουν τους χρήστες στην επιλογή τους;

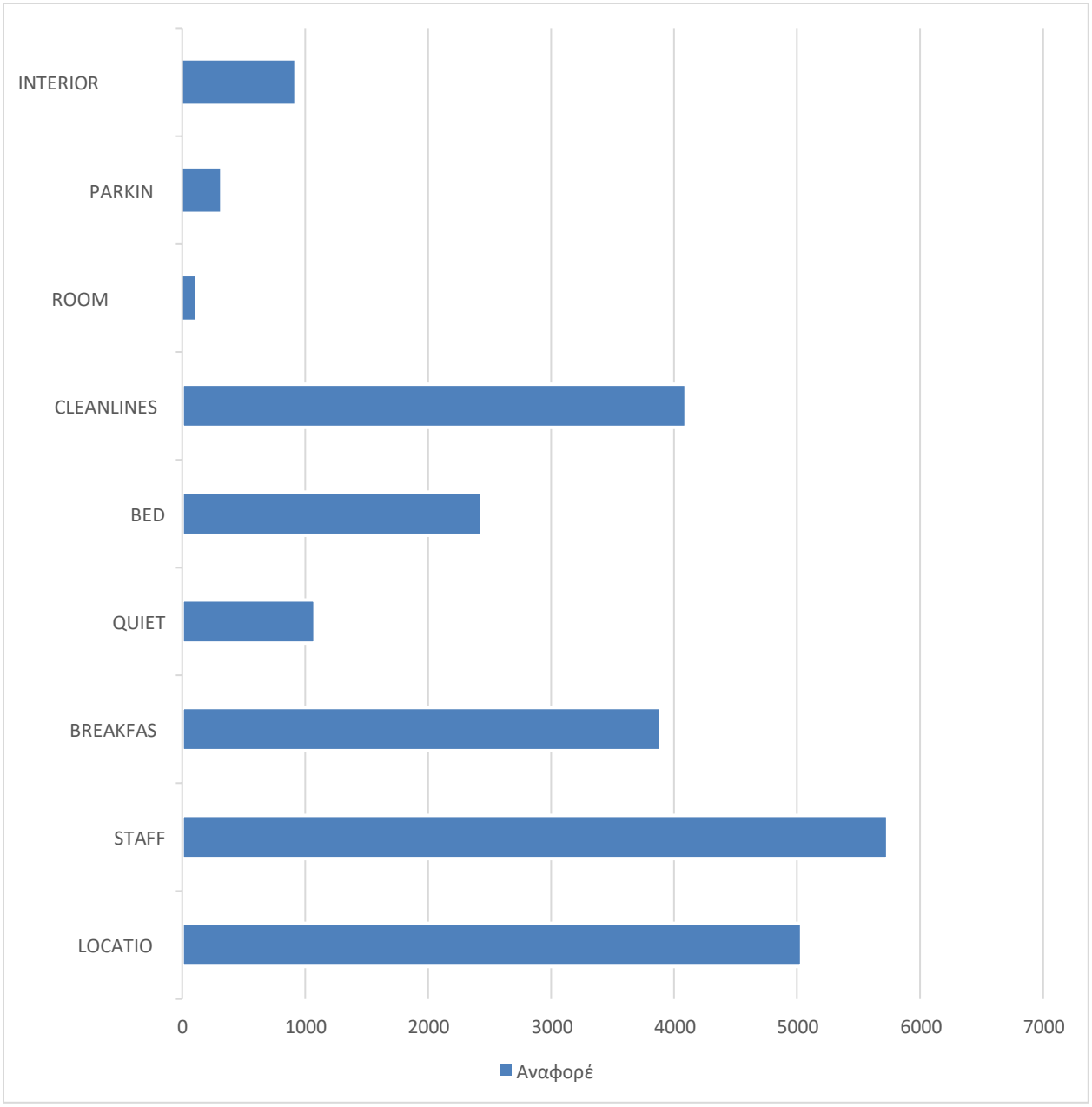
Τα συμπεράσματα που εξάγουμε από αυτό το κομμάτι της έρευνας, είναι αρκετά και χρήσιμα. Για αρχή,μαθαίνουμε, μέσα από ένα αρκετά μεγάλο δείγμα, ποια κριτήρια έχει ο χρήστης ως υψηλή προτεραιότητα. Η διαδικασία αναζήτησης του σύγχρονου ταξιδιώτη έχει αλλάξει αρκετά τα τελευταία χρόνια, και θα είναι μια διαδικασία που θα είναι μονίμως ανάλογη με την τεχνολογία, καθώς όσο προχωράει η τεχνολογία τόσο προχωράει και η εξέλιξη της αναζήτησης. Η αναζήτηση γίνεται αρκετά ευκολότερη για τον χρήστη, καθώς πλέον υπάρχουν ιστοσελίδες που μπορούν να βρουν αυτό που ακριβώς χρειάζονται έχοντας πρόσβαση σε αναζήτηση με διάφορα φίλτρα (Τιμή, τοποθεσία, παροχές κλπ.) αλλά παράλληλα γίνεται πιο δύσκολη για τους ιδιοκτήτες, καθώς πρέπει να προσθέσουν αρκετά στοιχεία του ξενοδοχείου που παλαιότερα είναι πιθανό να μην χρειαζόταν (Wi-Fi, Parking κλπ.).

Συνεπώς η σχέση που υπάρχει ανάμεσα στον χρήστη και στο ξενοδοχείο, είναι αντιστρόφως ανάλογη, καθώς όσο πιο εύκολη γίνεται η αναζήτηση για τον χρήστη όσο εξελίσσεται η τεχνολογία, τόσο πιο δύσκολη γίνεται για τα ίδια τα ξενοδοχεία, καθώς χρειάζεται συνέχεια να προσθέτουν και να παρέχουν παροχές στον χρήστη, πολλές φορές χωρίς επιπλέον κόστος, καθώς πια ο χρήστης έχει να επιλέξει ανάμεσα από μια μεγάλη γκάμα επιλογών και ο ανταγωνισμός είναι αρκετά μεγάλος.

3.2.1 Η διαδικασία αναζήτησης του σύγχρονου ταξιδιώτη

Το 47% των ταξιδιωτών χρησιμοποιεί ιστοσελίδες meta search, για να συγκρίνει τις τιμές ξενοδοχείων (στοιχεία της Research Now μεταξύ 2013-2015). Μόλις επιλεχτεί ένας προορισμός και οριστεί ένα εύρος τιμών, **το επόμενο βήμα** στην αναζήτηση εύρεσης του ιδανικού ξενοδοχείου είναι οι επιλογές να γίνουν πιο συγκεκριμένες. Διαπιστώσαμε ότι οι χρήστες έχουν μια συγκεκριμένη «εικόνα» στο μυαλό τους για τα ξενοδοχεία που είναι σύμφωνη με συγκεκριμένες προτιμήσεις, π.χ. τύπος καταλύματος, στιλ, παροχές και υπηρεσίες. Πιο συγκεκριμένα, τα κριτήρια που επιλέξαμε και αναλύσαμε παραπάνω, δεν μπήκαν τυχαία στο πρόγραμμα μας, καθώς μετά από αναζήτηση σε αρκετές έρευνες καταλήξαμε στο γεγονός ότι αυτά τα 9 παίζουν ένα αρκετά σημαντικό ρόλο, άλλα περισσότερο άλλα λιγότερο, στην επιλογή καταλύματος από τον χρήστη.Συνεπώς,Τα κριτήρια ξενοδοχείου έγιναν ολοένα και πιο σημαντικά με την εξέλιξη των **μηχανών αποτελεσμάτων αναζήτησης**. Σε μια εποχή όπου η επιλογή αποτελεί σημαντικό παράγοντα για τους πελάτες, αυτές οι ιστοσελίδες δίνουν τη δυνατότητα στους χρήστες να πραγματοποιούν αναζήτηση ανάμεσα σε μια τεράστια γκάμα ξενοδοχείων, φιλτράροντας και προσαρμόζοντας τα αποτελέσματα, ώστε να βρουν το ξενοδοχείο που ικανοποιεί τις ανάγκες τους.τα ξενοδοχεία είναι σημαντικό να προβάλλουν σωστά στοιχεία ξενοδοχείου. **Τα ξενοδοχεία που δεν περιλαμβάνουν αυτές τις πληροφορίες στο σύστημα δε θα προβάλλονται στα αποτελέσματα αναζήτησης.** Ως συνέπεια, χάνουν αυτόματα και την ευκαιρία να προβληθούν και να κερδίσουν μια κράτηση.

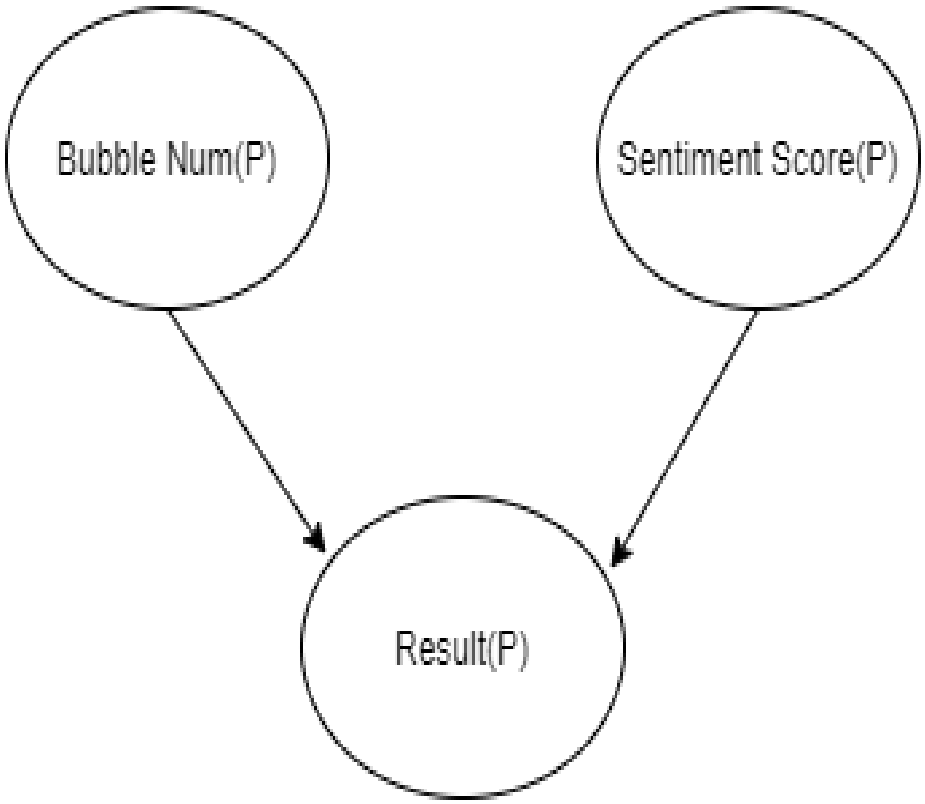
Αποτελέσματα Έρευνας – Γράφικη Αναπαράσταση



Όπως είναι εύκολα αντιληπτό, το κριτήριο του **Προσωπικού** και της **Τοποθεσίας** παίρνουν το μεγαλύτερο μερίδιο στις αναφορές των χρηστών, έχοντας μικρή διαφορά μεταξύ τους, η ακριβής διαφορά τους είναι 700.Στη συνέχεια, λίγο πιο πίσω βρίσκονται τα κριτήριο της **Καθαριότητας** και του Πρωινού, τα οποία και αυτά έχουν μια αρκετά μικρή διαφορά μεταξύ τους, η ακριβής διαφορά τους είναι 209.Έπειτα, αρκετά πιο πίσω από τα υπόλοιπα βρίσκονται τα κριτήρια του **Κρεβατιού**, της **Ησυχίας**, της **Εσωτερικής Διακόσμησης** αλλά και του Χώρου **Στάθμευσης**. Στην τελευταία θέση, με μεγάλη διαφορά, βρίσκεται το κριτήριο **του Χώρου του Δωματίου**, που από ότι φαίνεται οι χρήστες το αγνοούνε τελείως, καθώς οι αναφορές για αυτό το κριτήριο είναι μόλις 101.

3.2.3 Χρήση Bayesian Network

Στην συνέχεια, χρησιμοποιώντας Bayesian Networks 2 επιπέδων, προσπαθούμε να καταλάβουμε ανάλογα με το Sentiment Score που έχει συγκεντρώσει κάθε σχόλιο, ποια είναι η πιθανότητα ο χρήστης να έχει δώσει ένα αντίστοιχο αριθμό από φυσαλίδες (Bubbles rating–Είναι ένας αριθμός από το 1 έως το 5 και στην ουσία είναι ένας γρήγορος τρόπος βαθμολόγησης των υπηρεσιών του ξενοδοχείου χωρίς να γίνεται κάποια αναφορά σε κάτι).



Έχουμε 5 διαφορετικές περιπτώσεις Sentiment Score, μετά από αρκετά πειράματα επάνω στο συγκεκριμένοdataset από πλευράς μας και με την βοήθεια του επιβλέπων καθηγητή καταλήξαμε σε αυτό το εύρος:

- Για Sentiment Score από **(-0.4,-1]** το αποτέλεσμα είναι: **DISASTER**
- Για Sentiment Score **(0.1,-0.4]**το αποτέλεσμαείναι: **MANY THINGS NEEDS TO GET BETTER**
- Για Sentiment Score**(0.4,0.1]** το αποτέλεσμα είναι: **FAIR ENOUGH**
- Για Sentiment Score**(0.7,0.4]** το αποτέλεσμα είναι: **PERFECT**
- Για Sentiment Score **[0.7 , 1]**τοαποτέλεσμαείναι: **ABSOLUTELY PERFECT**

Στην συνέχεια θα παραθέσουμε αναλυτικά αποτελέσματα για κάθε μια από τις περιπτώσεις που μπορούν να προκύψουν για όλους τους πιθανούς συνδυασμούς που προκύπτουν.

Ξεκινάμε λοιπόν με την **πρώτη περίπτωση**, στην οποία βρίσκουμε την πιθανότητα να δώσει ο χρήστης Χ αριθμό από φυσαλίδες (1 έως 5) δεδομένου ότι το Sentiment Score του σχολίου είναι **ABSOLUTELYPERFECT** (δηλ. [0.7 , 1])

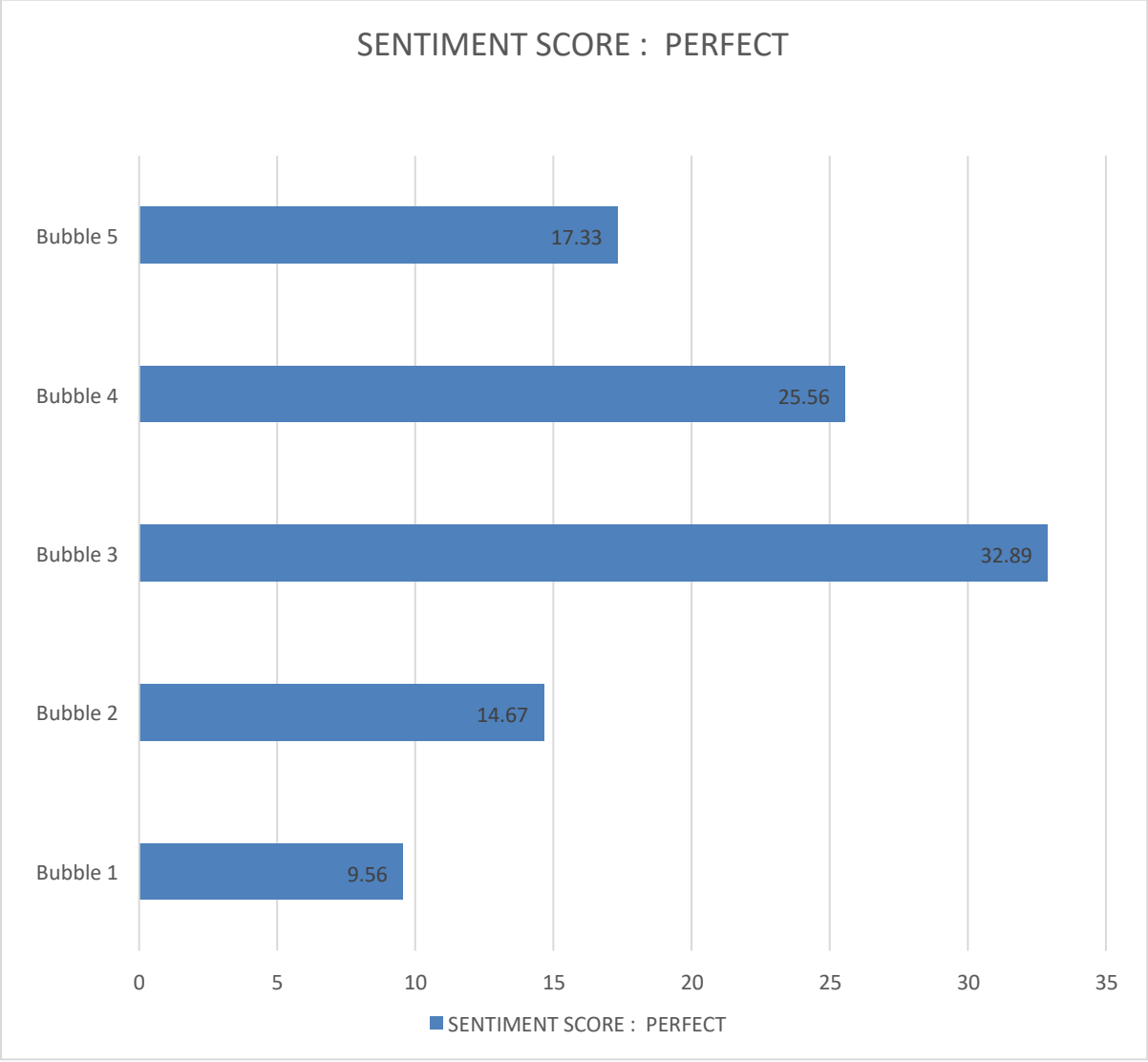
ΠΕΡΙΠΤΩΣΗ 1

BUBBLES BY USER	RATIO (%)
1	1,14
2	2,17
3	9,04
4	28,07
5	59,57

Το **συμπέρασμα** που προκύπτει είναι το περίπου το 60% των αξιολογήσεων συμφωνούν με το Sentiment Score που έβγαλε το πρόγραμμα μας και αν υπολογίσουμε και το 28% που παραπέμπει σε 4 αστέρια, έχουμε ένα ποσοστό επιτυχίας κοντά στο 88%, αν αναλογιστούμε τον αριθμό των σχολίων (κάτι περισσότερο από 10.000) έχουμε βγάλει ένα αρκετά καλό Sentiment Score με πολύ μικρή απόκλιση.

Συνεχίζουμε με την **δεύτερη περίπτωση**, στην οποία βρίσκουμε την πιθανότητα να δώσει ο χρήστης Χ αριθμό από φυσαλίδες (1 έως 5) δεδομένου ότι το Sentiment Score του σχολίου είναι **PERFECT** (δηλ. (0.7 ,0.4])

ΠΕΡΙΠΤΩΣΗ 2

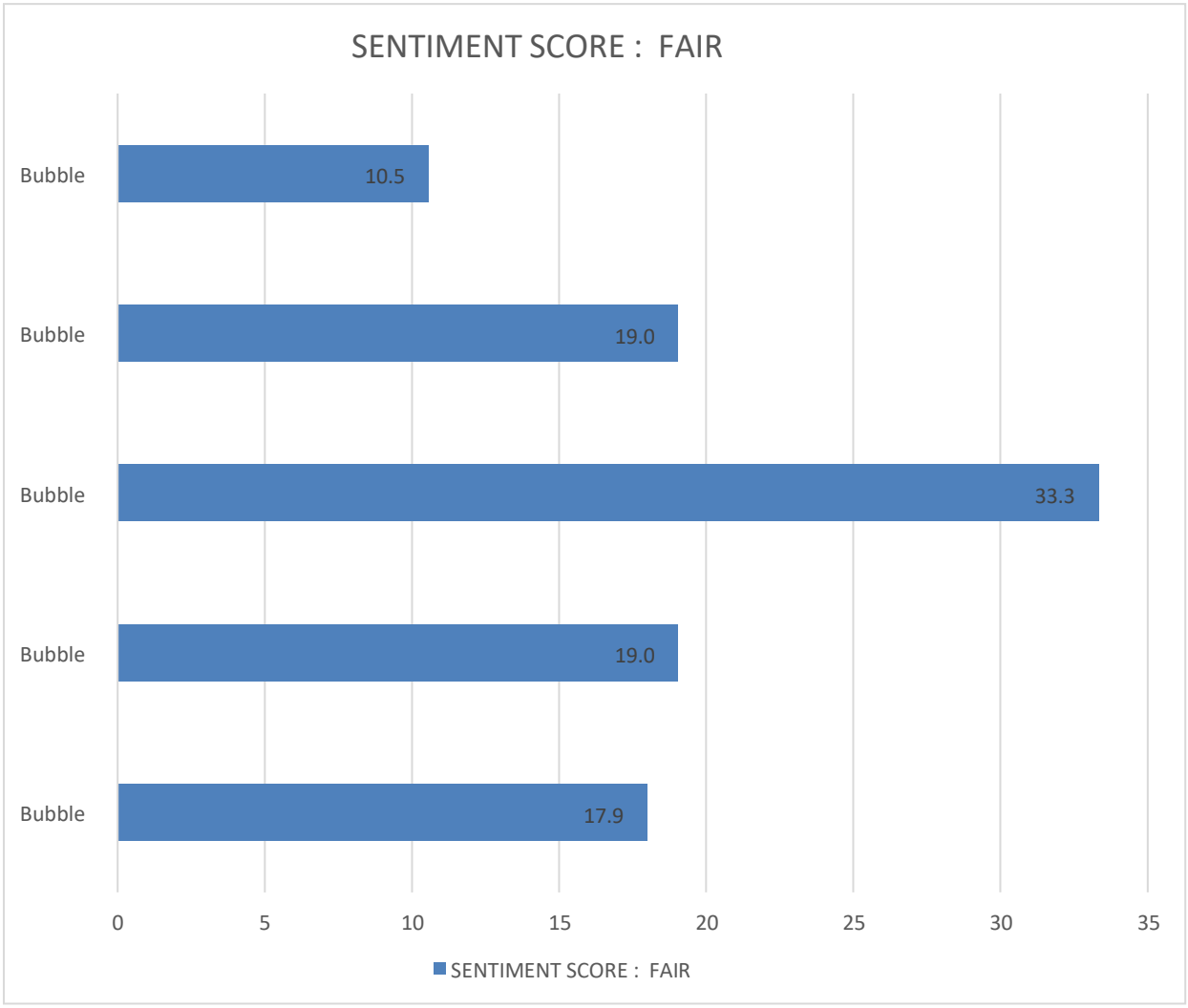


BUBBLES BY USER	RATIO (%)
1	9,56
2	14,67
3	32,89
4	25,56
5	17,33

Το **συμπέρασμα** που προκύπτει είναι το περίπου το 58% των αξιολογήσεων συμφωνούν με το Sentiment Score που έβγαλε το πρόγραμμα μας(αναφερόμαστε στις αξιολογήσεις με 3 και 4 αστέρια που μπορούν κάλλιστα να παραπέμψουν σε ένα sentiment score “PERFECT”).Επίσης, σίγουρα έχουμε και ένα μέρος από τα 5 αστέρια, καθώς πολλοί είναι οι χρήστες που πιθανών να έβαλα 5 αστερία και η αξιολόγηση τους να είχε κάποια αρνητικά, τα οποία όμως πιθανών να μην ήταν στα βασικά κριτήρια. Συνεπώς, και σε αυτή την περίπτωση φτάνουμε κοντά στο 65%-68%, αποτέλεσμα το οποία είναι αρκετά ικανοποιητικό.

Έπειτα πηγαίνουμε στην **τρίτη περίπτωση**, στην οποία βρίσκουμε την πιθανότητα να δώσει ο χρήστης X αριθμό από φυσαλίδες (1 έως 5) δεδομένου ότι το Sentiment Score του σχολίου είναι **FAIRENOUGH** (δηλ. (0.4 , 0.1])

ΠΕΡΙΠΤΩΣΗ 3

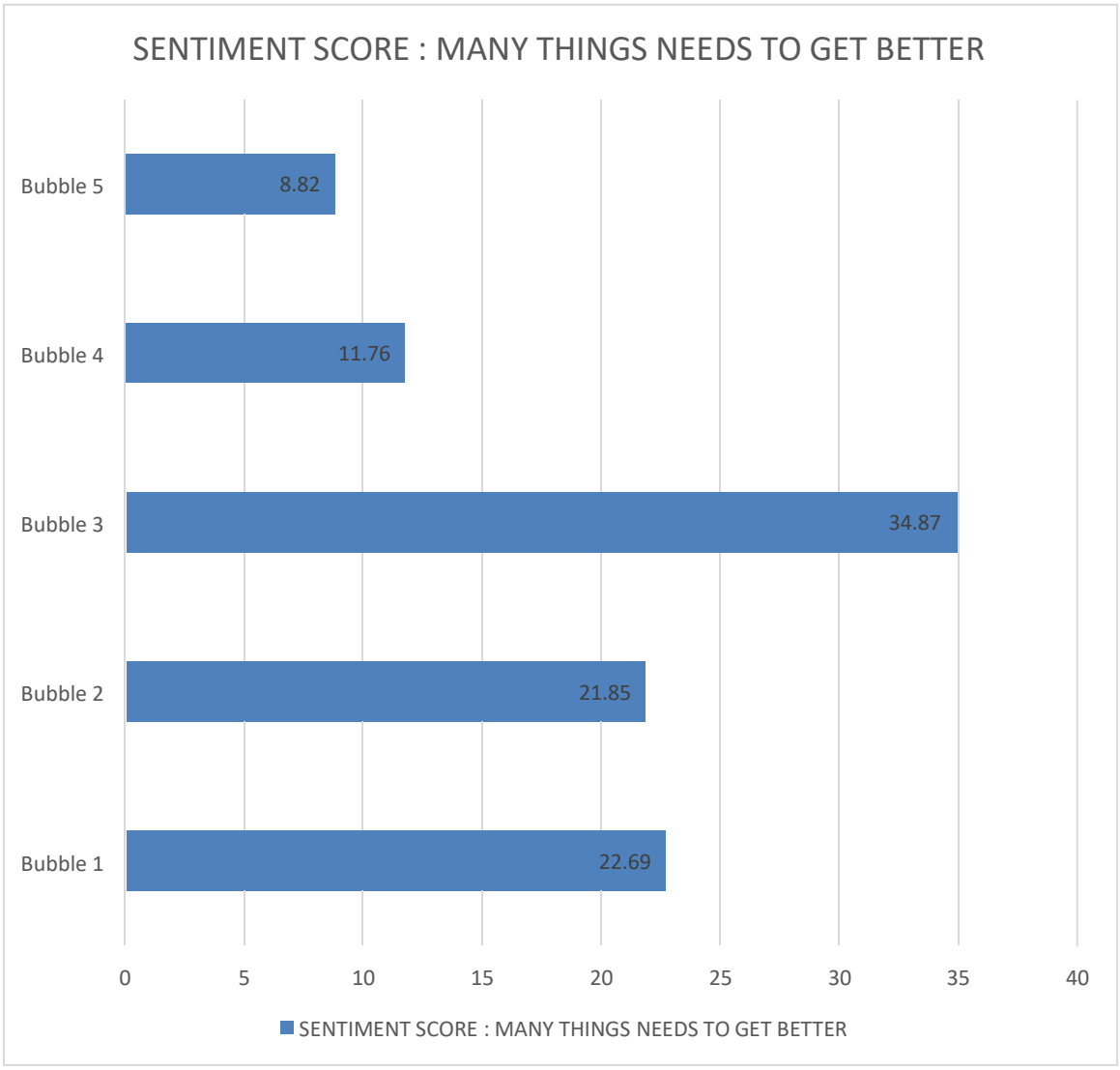


BUBBLES BY USER	RATIO (%)
1	17,99
2	19,05
3	33,33
4	19,05
5	10,58

Όσο κατεβαίνουμε προς τα κάτω σε κλίμακα αξιολόγησης, καταλαβαίνουμε ότι είναι πιο σύνθετο να εξάγουμε ασφαλή συμπεράσματα καθώς το είδος των αποτελεσμάτων είναι αρκετά πιο ουδέτερο και δεν είναι τόσο ξεκάθαρο. Με μια ματιά, μπορούμε να πούμε ότι το **Sentiment Score FAIR ENOUGH** ανήκει στην κατηγορία αστεριών 2 και 3, και σίγουρα παίρνει και ένα ποσοστό από την κατηγορία με τα 4 αστερία. Εμπειρικά, μπορούμε να πούμε ότι πάλι ήμαστε σε ένα ποσοστό κοντά στο 65%, το οποίο μας αφήνει αρκετά ικανοποιημένους για ένα τέτοιο 'ουδέτερο' σχόλιο.

Στη συνέχεια, πηγαίνουμε στην **τέταρτη περίπτωση**, στην οποία βρίσκουμε την πιθανότητα να δώσει ο χρήστης X αριθμό από φυσαλίδες (1 έως 5) δεδομένου ότι το Sentiment Score του σχολίου είναι**MANYTHINGSNEEDSTOGETBETTER** (δηλ. (0.1 , -0.4])

ΠΕΡΙΠΤΩΣΗ 4

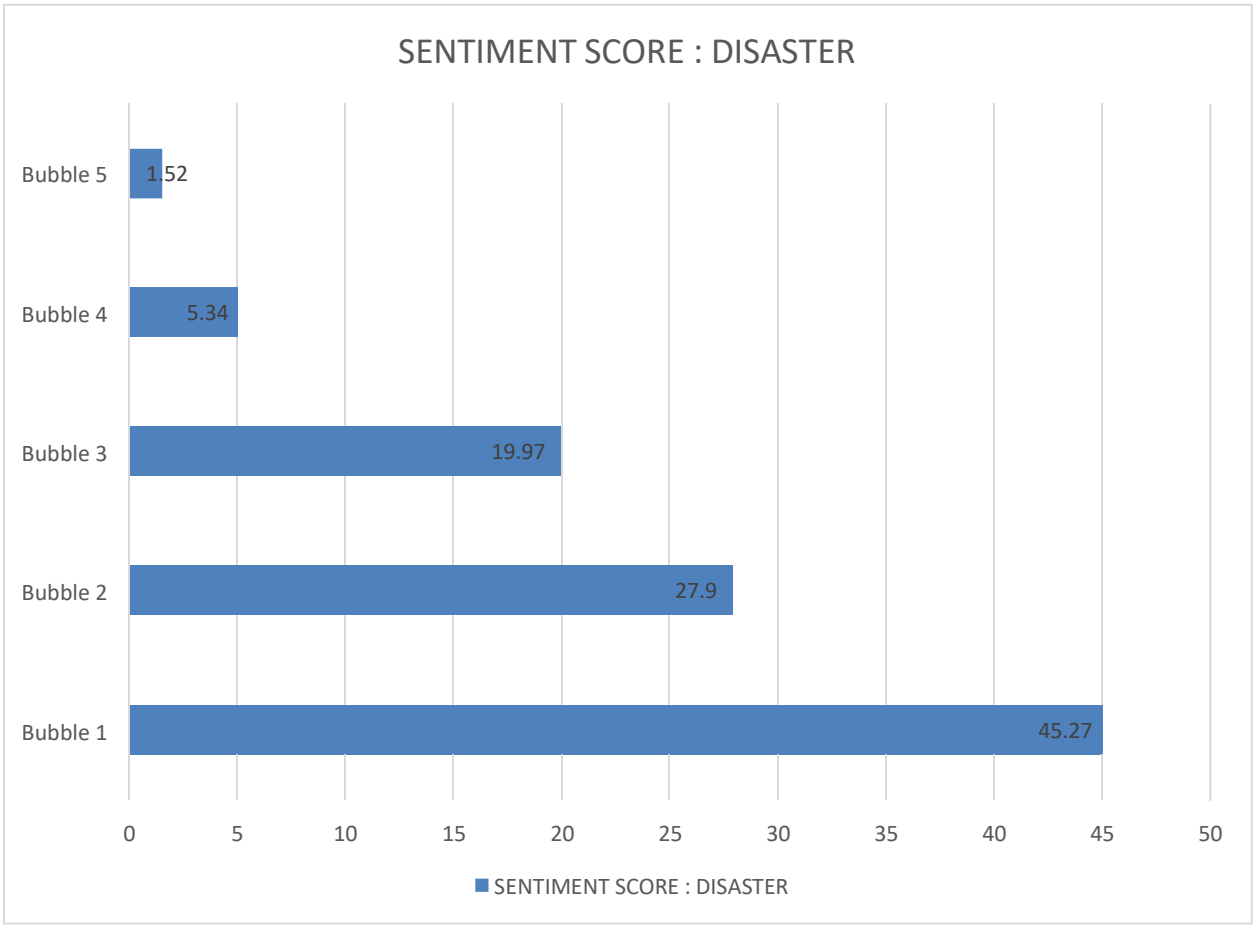


BUBBLES BY USER	RATIO (%)
1	22,69
2	21,85
3	34,87
4	11,76
5	8,82

Πλέον, τα πράγματα γίνονται πιο ξεκάθαρα, σε σχέση με το προηγούμενο στάδιο, καθώς αυτό το sentiment score παραπέμπει ξεκάθαρα σε κατηγορία αστερίων από 3 και κάτω.Μπορούμε να πούμε ότι το **Sentiment Score MANY THINGS NEEDS TO GET BETTER** ανήκει στην κατηγορία αστεριών 1,2 αλλά παίρνει ένα αρκετά μεγάλο ποσοστό και από την κατηγορία 3. Εμπειρικά, και πάντα σύμφωνα με τα προηγούμενα συμπεράσματα, μπορούμε να πούμε ότι πάλι ήμαστε σε ένα ποσοστό κοντά στο 70%

Τέλος,πηγαίνουμε στην **πέμπτη και τελευταία μας περίπτωση**, στην οποία βρίσκουμε την πιθανότητα να δώσει ο χρήστης Χ αριθμό από φυσαλίδες (1 έως 5) δεδομένου ότι το Sentiment Score του σχολίου είναι **DISASTER** (δηλ. (-0.4, -1])

ΠΕΡΙΠΤΩΣΗ 5



BUBBLES BY USER	RATIO (%)
1	45,27
2	27,9
3	19,97
4	5,34
5	1,52

Φτάσαμε στο άλλο άκρο των αξιολογήσεων, ξεκινώντας από **SENTIMENTSCORE : ABSOLUTELY PERFECT**, φτάσαμε σε ένα **SENTIMENTSCORE : DISASTER**. Τα συμπεράσματα που μπορούμε να εξάγουμε από εδώ, είναι παρόμοια με αυτά του **ABSOLUTELY PERFECT**, πιο αναλυτικά, σε αυτή την κατηγορία ανήκουν σίγουρα οι κατηγορίες αστεριών 1 και 2, και ίσως και ένα ελάχιστο ποσοστό από το 3,συνεπώς ήμαστε πάλι σε ένα ποσοστό άνω του 75%.

3.3 Ανάλυση Συναισθήματος για Υπηρεσίες με Fuzzy String Matching

Το **Fuzzy String Matching** είναι μια κατά προσέγγιση τεχνική αντιστοίχισης συμβολοσειράς για προγραμματική αντιστοίχιση παρόμοιων δεδομένων. Αντί να εξετάσουμε απλώς την ισοδυναμία μεταξύ δύο συμβολοσειρών για να καθορίσουμε εάν είναι οι ίδιες, οι αλγόριθμοι ασαφούς αντιστοίχισης λειτουργούν για να ποσοτικοποιήσουν ακριβώς πόσο κοντά είναι δύο συμβολοσειρές μεταξύ τους. Με αυτόν τον τρόπο, μπορούν να βοηθήσουν στον προσδιορισμό της πιθανότητας δύο διαφορετικές συμβολοσειρές να είναι πράγματι ισοδύναμες. Αυτό γίνεται συχνά με την ενσωμάτωση της απόστασης επεξεργασίας.

Υπάρχουν πολλές καταστάσεις όπου οι τεχνικές **Fuzzy String Matching** να είναι χρήσιμες όπως:

Single Customer View: Δημιουργία μεμονωμένης προβολής πελάτη (SCV): Η προβολή ενός πελάτη (SCV) αναφέρεται στη συλλογή όλων των δεδομένων σχετικά με τους πελάτες και τη συγχώνευση τους σε μια ενιαία εγγραφή.

DataAccuracy: Σύμφωνα με μια πρόσφατη μελέτη, πάνω από το 60% των εταιρειών έχουν εφαρμόσει λύσεις βασισμένες στην Μηχανική Μάθηση. Καθώς οι εταιρείες βασίζονται στην τεχνητή νοημοσύνη και τη μηχανική μάθηση, η ακρίβεια των δεδομένων γίνεται εξαιρετικά κρίσιμη.Η πρωτοποριακή έρευνα πραγματοποιείται συχνά για τη βελτίωση της ακρίβειας των νευρωνικών δικτύων και των τεχνολογιών μηχανικής μάθησης, ωστόσο, λίγα γίνονται για να διασφαλιστεί ότι τα δεδομένα αυτά καλής ποιότητας θα τροφοδοτηθούν σε αυτά τα μοντέλα. Ένας μεγάλος αλγόριθμος μηχανικής μάθησης χωρίς ακριβή δεδομένα είναι ανάλογος με την εκτόξευση πυραύλου στον Άρη χρησιμοποιώντας συμπιεσμένο φυσικό αέριο.Η ασαφής αντιστοίχιση συμβολοσειρών μπορεί να συμβάλει στη βελτίωση της ποιότητας και της ακρίβειας των δεδομένων με την αφαίρεση πολλαπλασιασμού δεδομένων, τον εντοπισμό ψευδώς θετικών κλπ.

FraudDetection: Ένας καλός αλγόριθμος αντιστοίχισης ασαφών συμβολοσειρών μπορεί να βοηθήσει στην ανίχνευση απάτης εντός ενός οργανισμού. Αργότερα σε αυτήν την ανάρτηση, θα δούμε πώς η FAA χρησιμοποίησε ασαφή συμβολοσειρά συμβολοσειρών για να ξεχωρίσει αρκετούς πιλότους για επιδείξεις δόλιας συμπεριφοράς.

3.3.1 Αλγόριθμοι Fuzzy Match String

Υπάρχουν πολλοί δημοφιλείς αλγόριθμοι που μπορούν να χρησιμοποιηθούν για την εκτέλεση Fuzzy String Matching.Μερικοί από αυτούς είναι:

3.1.1.1 Απόσταση Levenshtein: Η απόσταση Levenshtein είναι μια μέτρηση που χρησιμοποιείται για τη μέτρηση της διαφοράς μεταξύ 2 ακολουθιών συμβολοσειρών.Μας δίνει ένα μέτρο του αριθμού των εισαγωγών, διαγραφών ή αντικαταστάσεων ενός χαρακτήρα που απαιτούνται για την αλλαγή μιας συμβολοσειράς σε άλλη.Μαθηματικά, μπορούμε να ορίσουμε την απόσταση Levenshtein(Εικόνα 1,4) ως εξής:

Εικόνα 1.4

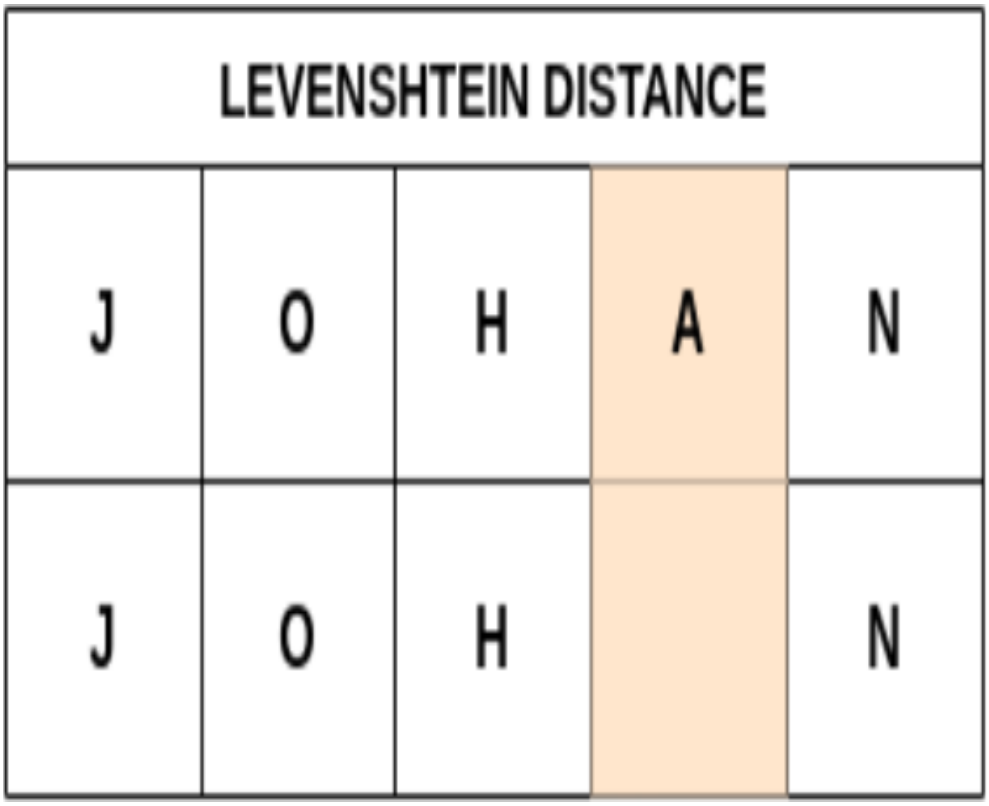
The Levenshtein distance between two strings a, b (of length $|a|$ and $|b|$ respectively) is given by $\operatorname{lev}(a, b)$ where

$$\operatorname{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \operatorname{lev}(\operatorname{tail}(a), b) \\ \operatorname{lev}(a, \operatorname{tail}(b)) \\ \operatorname{lev}(\operatorname{tail}(a), \operatorname{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

Όπου η ουρά (α) είναι μια συμβολοσειρά που περιλαμβάνει όλους τους χαρακτήρες της συμβολοσειράς α, εξαιρώντας τον πρώτο χαρακτήρα. Από τον ορισμό, είναι σαφές ότι η απόσταση **Levenshtein** αποδίδει καλά με ανορθόγραφα ονόματα.

Για παράδειγμα, η απόσταση **Levenshtein** (ο αριθμός εισαγωγών, διαγραφών ή αντικαταστάσεων ενός χαρακτήρα) μεταξύ των συμβολοσειρών JOHN (όπως γράφεται στις ΗΠΑ) και JOHAN (όπως γράφεται στην Ιαπωνία, τη Σουηδία και τη Νορβηγία) 'είναι 1 ως φαίνεται στην Εικόνα 6

Εικόνα 6



3.1.1.3 The Metaphone and Double Metaphone Algorithms:ToMetaphone είναι ένας φωνητικός αλγόριθμος, που δημοσιεύθηκε από τον Lawrence Philips το 1990, για την ευρετηρίαση λέξεων με την αγγλική τους προφορά. Βελτιώνει ουσιαστικά τον αλγόριθμο Soundex χρησιμοποιώντας πληροφορίες σχετικά με παραλλαγές και ασυνέπειες στην αγγλική ορθογραφία και προφορά για να παράγει μια πιο ακριβή κωδικοποίηση. καλύτερη δουλειά στο να ταιριάζουν λέξεις και ονόματα που ακούγονται παρόμοια. Όπως και με το Soundex, οι λέξεις με παρόμοιο ήχο πρέπει να έχουν τα ίδια κλειδιά. Το Metaphone είναι διαθέσιμο ως ενσωματωμένος χειριστής σε πολλά συστήματα.Philips παρήγαγε αργότερα μια νέα έκδοση του αλγορίθμου, την οποία ονόμασε DoubleMetaphone. Σε αντίθεση με τον αρχικό αλγόριθμο του οποίου η εφαρμογή περιορίζεται μόνο στα αγγλικά, αυτή η έκδοση λαμβάνει υπόψη τις ορθογραφικές ιδιαιτερότητες μιας σειράς άλλων γλωσσών. Το 2009 η Philips κυκλοφόρησε μια τρίτη έκδοση, που ονομάζεται Metaphone 3, η οποία επιτυγχάνει ακρίβεια περίπου 99% για αγγλικές λέξεις, μη αγγλικές λέξεις που είναι γνωστές στους Αμερικανούς και τα ονόματα και τα επώνυμα που απαντώνται συνήθως στις Ηνωμένες Πολιτείες, έχουν αναπτυχθεί σύμφωνα με σύγχρονα πρότυπα μηχανικής έναντι δοκιμαστικής πλεξούδας προετοιμασμένων σωστών κωδικοποιήσεων.

Παράδειγμα Double Metaphone (Εικόνα 7)

string1	dblmeta_s1	string2	compare	Notes
My String	["MSTRNK","MSTRNK"]	my string	TRUE	comparison is case-insensitive
judge	["JJ","AJ"]	juge	TRUE	typo
knock	["NK","NK"]	nock	TRUE	silent letters
white	["AT","AT"]	wite	TRUE	missing letters
record	["RKRT","RKRT"]	record	TRUE	two different words in English but match the same
pair	["PR","PR"]	pear	TRUE	these match but are different words.
bookkeeper	["PKPR","PKPR"]	book keeper	FALSE	spaces cause failures in comparison
test1	["TST","TST"]	test123	TRUE	digits are not compared
the end.	["0NT","TNT"]	the endâ€¦.	TRUE	punctuation differences do not matter.
a elephant	["ALFNT","ALFNT"]	an elephant	FALSE	a and an are treated differently.

***3.1.1.4 Cosine Similarity*:** Είναι ένα μέτρο ομοιότητας μεταξύ δύο μη μηδενικών διανυσμάτων ενός εσωτερικού χώρου προϊόντος. Ορίζεται για να ισούται με το συνημίτονο της γωνίας μεταξύ τους, το οποίο είναι επίσης το ίδιο με το εσωτερικό γινόμενο των ίδιων διανυσμάτων που κανονικοποιούνται και στα δύο έχουν μήκος 1. Το συνημίτονο των 0 ° είναι 1 και είναι μικρότερο από 1 για οποιαδήποτε γωνία στο διάστημα (0, π] ακτίνια. Είναι συνεπώς κρίση προσανατολισμού και όχι μεγέθους: δύο διανύσματα με τον ίδιο προσανατολισμό έχουν ομοιότητα συνημίτονο 1, δύο διανύσματα προσανατολισμένα σε 90 ° το ένα μεταξύ τους έχουν ομοιότητα 0, και δύο διανύσματα διαμετρικά αντίθετα έχουν ομοιότητα -1, ανεξάρτητα από το μέγεθός τους. Η ομοιότητα συνημίτονο χρησιμοποιείται ιδιαίτερα σε θετικό χώρο, όπου το αποτέλεσμα περιορίζεται τακτοποιημένα στο {\ display style [0,1]} [0,1]. Το όνομα προέρχεται από τον όρο "κατεύθυνση συνημίτονο": σε αυτή την περίπτωση, τα διανύσματα μονάδων.είναι στο μέγιστο βαθμό "παρόμοια" αν είναι παράλληλα και στο μέγιστο "ανόμοια" αν είναι ορθογώνια (κάθετα). Αυτό είναι ανάλογο με το συνημίτονο, που είναι ενότητα (μέγιστη τιμή) όταν τα τμήματα έχουν μηδενική γωνία και μηδέν (μη συσχετισμένα) όταν το se τα gments είναι κάθετα.Αυτά τα όρια ισχύουν για οποιονδήποτε αριθμό διαστάσεων και η ομοιότητα συνημίτονο χρησιμοποιείται συχνότερα σε θετικούς χώρους υψηλής διάστασης. Για παράδειγμα, στην ανάκτηση πληροφοριών και την εξόρυξη κειμένου, κάθε όρος έχει νοηματικά μια διαφορετική διάσταση και ένα έγγραφο χαρακτηρίζεται από ένα διάνυσμα όπου η τιμή σε κάθε διάσταση αντιστοιχεί στον αριθμό των φορών που εμφανίζεται ο όρος στο έγγραφο. Η ομοιότητα με το συνημίτονο δίνει στη συνέχεια ένα χρήσιμο μέτρο για το πόσο παρόμοια είναι τα δύο έγγραφα ως προς το αντικείμενό τους.Η τεχνική χρησιμοποιείται επίσης για τη μέτρηση της συνοχής εντός συμπλεγμάτων στον τομέα της εξόρυξης δεδομένων.Τύπος Cosine Similarity Εικόνα 8.

Εικόνα 8

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

Given two vectors of attributes, A and B, the cosine similarity, $\cos(\theta)$, is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

where A_i and B_i are components of vector A and B respectively.

3.1.1.2 Ο αλγόριθμος Soundex: *Είναι ένας φωνητικός αλγόριθμος που χρησιμοποιείται για την αναζήτηση ονομάτων που ακούγονται παρόμοια αλλά γράφονται διαφορετικά.Χρησιμοποιείται συχνότερα για γενεαλογικές αναζητήσεις βάσεων δεδομένων.Ο αλγόριθμος Soundex εξάγει έναν τετραψήφιο κωδικό και παίρνει σαν είσοδο ένα όνομα.Ο αλγόριθμος Soundex είναι ο πιο ευρέως γνωστός από όλους τους φωνητικούς αλγόριθμους επειδή είναι ένα τυπικό χαρακτηριστικό του δημοφιλούς λογισμικού βάσεων δεδομένων όπως DB2,PostgreSQL,MySQL,SQLite,Ingres,MS SQLServer,Oracle και SAPASE.Οι βελτιώσεις στο Soundex αποτελούν τη βάση για πολλούς σύγχρονους φωνητικούς αλγόριθμους.Για παράδειγμα, τα ονόματα "Stuart" και "Stewart" παράγουν τον ίδιο κωδικό SoundexS363.*

3.4 Συσχέτιση Υπηρεσιών Συναισθήματος και Βαθμολογίας με Bayesian Networks

Ορισμός

Ένα δίκτυο Bayes είναι ένα πιθανολογικό γραφικό μοντέλο που αντιπροσωπεύει ένα σύνολο μεταβλητών και τις εξαρτήσεις υπό όρους τους μέσω κατευθυνόμενου ακυκλικού γραφήματος (DAG). Τα δίκτυα Bayes είναι ιδανικά για τη λήψη ενός γεγονότος που συνέβη και για την πρόβλεψη της πιθανότητας ότι οποιαδήποτε από τις περισσότερες πιθανές γνωστές αιτίες ήταν ο συντελεστής. Για παράδειγμα, ένα δίκτυο Bayesian θα μπορούσε να αντιπροσωπεύει τις πιθανολογικές σχέσεις μεταξύ ασθενειών και συμπτωμάτων. Δεδομένων των συμπτωμάτων, το δίκτυο μπορεί να χρησιμοποιηθεί για τον υπολογισμό των πιθανοτήτων παρουσίας διαφόρων ασθενειών.

Το Bayesian Network Ν είναι μια γραφική παράσταση μιας κοινής κατανομής πιθανότητας μεταξύ ενός συνόλουτυχαίεςμεταβλητές.Το δίκτυο αποτελείται από δύο στοιχεία: ένα DAGG =(Rn, Mr) που αντιπροσωπεύει τη διαρθρωτική διάταξηενός συνόλου μεταβλητών (κόμβων) Rn = {x1, ..., xn} και ένα αντίστοιχο σύνολο εξάρτησης και ανεξαρτησίας ισχυρισμοί (τόξα) ,Mr μεταξύ των μεταβλητών ένα σύνολο από κατανομές πιθανοτήτων υπό όρους P = {pi, ..., pn} μεταξύ των κόμβων γονέα και παιδιού στο γράφημα.Στο στοιχείο DAG, υποστηρίζει η ύπαρξη ενός κατευθυνόμενου τόξου μεταξύ ενός ζεύγους μεταβλητών xi και xj υπό όρους εξάρτηση μεταξύ των δύο μεταβλητών. Το κατευθυνόμενο τόξο μπορεί επίσης να φαίνεται ότι αντιπροσωπεύει την αιτιότητα μεταξύ μιας μεταβλητής και το άλλο, δηλαδή μεταβλητό.Τοxy είναι μια υπαρξιακή αιτία μεταβλητής xz, επομένως xy →xz Η απουσία κατευθυνόμενου τόξου μεταξύ ζεύγουςτων μεταβλητών, ωστόσο, αντιπροσωπεύει μια υπό όρους ανεξαρτησία, τέτοια ώστε, δεδομένου ενός υποσυνόλου Uμεταβλητώναπό το Rn, ο βαθμός πληροφοριών σχετικά με τη μεταβλητήτοxi δεν αλλάζει γνωρίζοντας το xj, οπότε έχουμε (xi , xj | U).Αυτό συνεπάγεται επίσης ότι το p (xi | xj, U) = p (xi| U). ο γονέας (ες) της μεταβλητής xi∈Rn συμβολίζεται με ένα σύνολο paG (xi) = xj∈Rn | xji∈Mr, και paG (xi) = 0 για ριζικός κόμβος. Το DAGG αντιπροσωπεύεται από το CPT του, το οποίο περιέχει ένα σύνολο αριθμητικών παραμέτρων για κάθε μεταβλητή xiRn Αυτές οι αριθμητικές παράμετροι υπολογίζονται ως το πιθανότητα κάθε μεταβλητής δεδομένου του συνόλου των γονέων, p (xi | paG (Xi)). Πάνω από το σύνολο των μεταβλητών στο Rn.

Συνεπώς, η κοινή πιθανότητα για το BN λαμβάνεται ως(Εικόνα 9):

$$p(x_1, ..., x_n) = \prod_{X_i \in R_n} p(x_i | pa_G(x_i))$$

Εικόνα 9

Έτσι, για μια τυπική εργασία ταξινόμησης, ο ταξινομητής BN θα μάθαινε τις αριθμητικές παραμέτρους ενός CPT από τη δομή DAG G, με εκτίμηση ορισμένων στατιστικών πληροφοριών από τα δεδομένα. Αυτές οι πληροφορίες περιλαμβάνουν, αμοιβαίες πληροφορίες (MI) μεταξύ οι μεταβλητές και η κατανομή χ-τετραγώνου. Το πρώτο βασίζεται στις τοπικές μετρήσεις βαθμολογίας προσέγγιση και το τελευταίο παρουσιάζει προσέγγιση δοκιμών ανεξαρτησίας υπό όρους (CI). Και για τις δύο προσεγγίσεις, χρησιμοποιούνται διαφορετικοί αλγόριθμοι αναζήτησης για τον προσδιορισμό της δομής του δικτύου. Ο στόχος είναι να εξακριβωθεί, σύμφωνα με ένα ή περισσότερα κριτήρια αναζήτησης, το καλύτερο BN που ταιριάζει στο δεδομένα αξιολογώντας το βάρος του τόξου μεταξύ τις μεταβλητές. Τα κριτήρια για την αξιολόγηση της καταλληλόληταστων κόμβων (μεταβλητές), και τα τόξα (παράμετροι) στους αλγόριθμους αναζήτησης BN, εκφράζονται ως κατάλληλοι ή βαθμολογώντας συναρτήσεις εντός του ταξινομητήBN.Στόχος μας είναι να διασφαλίσουμε ότι αυτά τα κριτήρια περιλαμβάνουν πληροφορίες που εξαρτώνται από το συναίσθημα μεταξύ των μεταβλητές. Θα επικεντρωθούμε στην επιβολή κυρώσεων σε υπάρχοντα τοπικά μετρήσεις βαθμολογίας με το συναίσθημά μας να αυξάνει τη βαθμολογία συνάρτηση για τους ταξινομητέςBN, επομένως το SABN που προτείνεται σε αυτό το έγγραφο.Οι τοπικές μετρήσεις βαθμολογίας έχουν ιδιαίτερο ενδιαφέρον επειδή παρουσιάζουν ένα πρακτικό χαρακτηριστικόπου διασφαλίζει ότι η κοινή πιθανότητα του BN είναι αποσυνθέσιμη στο άθροισμα (ή το γινόμενο) της μεμονωμένης πιθανότητας κάθε κόμβου.Από όσο γνωρίζουμε, ελάχιστες ερευνητικές εργασίες έχουν εξετάσει πληροφορίες που εξαρτώνται από το συναίσθημα, ως μέρος των κριτηρίων καταλληλόλητας για τον εντοπισμό της εξάρτησης μεταξύ των μεταβλητές.

ΚΕΦΑΛΑΙΟ 4 - Τεχνική Υλοποίηση - Βιβλιοθήκες της Python

4.1 Ανάλυση βιβλιοθηκών της Python που χρησιμοποιήθηκαν για την υλοποίηση του κώδικα

4.1.1 Pandas

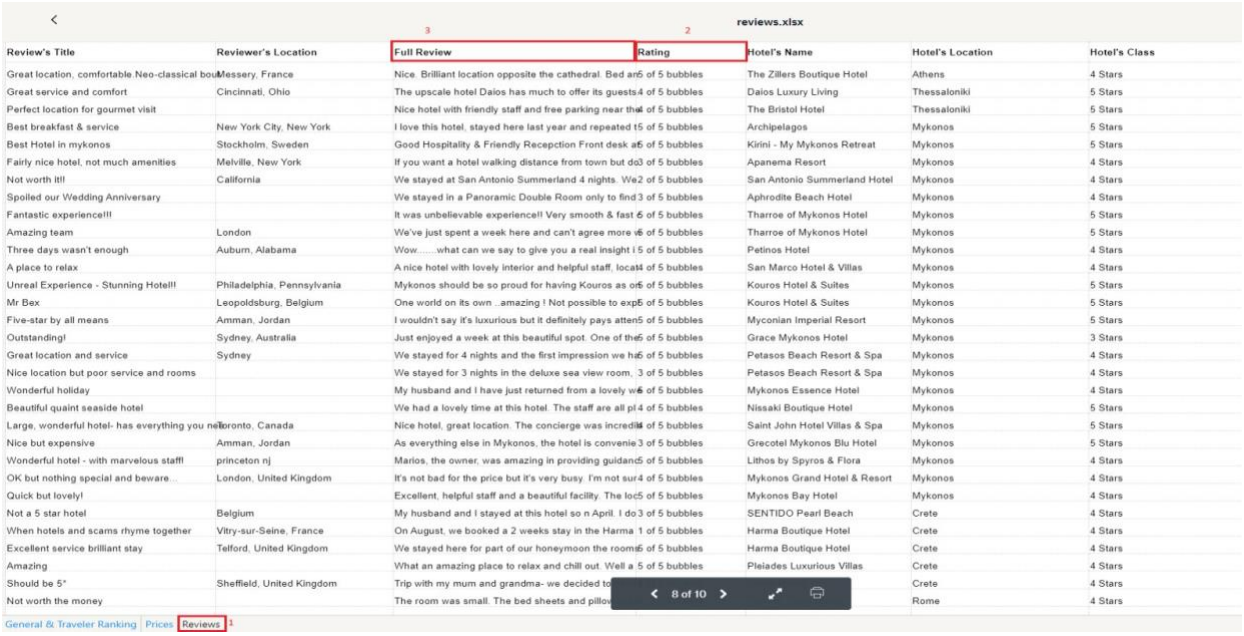
Η Pandas είναι μια βιβλιοθήκη Python για ανάλυση δεδομένων.Ξεκίνησε από τον Wes McKinney το 2008 λόγω ανάγκης για ένα ισχυρό και ευέλικτο εργαλείο ποσοτικής ανάλυσης, η pandas έχει εξελιχθεί σε μία από τις πιο δημοφιλείς βιβλιοθήκες Python. Έχει μια εξαιρετικά ενεργή κοινότητα συνεισφερόντων.Είναι χτισμένη πάνω σε δύο βασικές βιβλιοθήκες Python - mat plot lib για οπτικοποίηση δεδομένων και Num για μαθηματικές πράξεις.Η Pandas λειτουργεί ως περιτύλιγμα σε αυτές τις βιβλιοθήκες, επιτρέποντάς σας να έχετε πρόσβαση σε πολλές μεθόδους lib plot και Num Py με λιγότερο κώδικα.Για παράδειγμα, το pandas .plot () συνδυάζει πολλαπλές μεθόδους lib plot mat σε μία μόνο μέθοδο, επιτρέποντάς σας να σχεδιάσετε ένα γράφημα σε λίγες γραμμές.Πριν από τη pandas,οι περισσότεροι αναλυτές χρησιμοποίησαν την Python για την επεξεργασία δεδομένων και την προετοιμασία και στη συνέχεια άλλαξαν σε ενα πιο συγκεκριμένο τομέα της γλώσσας,όπως η R για το υπόλοιπο της ροής εργασίας τους. Η Pandas εισήγαγε δύο νέους τύπους αντικειμένων για την αποθήκευση δεδομένων που διευκολύνουν τις αναλυτικές εργασίες και εξαλείφουν την ανάγκη εναλλαγής εργαλείων: Σειρές που έχουν δομή που μοιάζει με λίστα και Data Frames, που έχουν δομή πίνακα.

Παρακάτω παραθέτουμε μια φωτογραφία από το excel που μας δόθηκε με ανοιχτή την καρτέλα με το όνομα Reviews και μια φωτογραφία από την κλάση Reviews Initialize.

Η κύρια υλοποίηση της Pandas στο πρόγραμμα μας γίνεται για στην κλάση Reviews Initiliaze.py την οποία δημιουργήσαμε για την συγκεκρίμενη δουλεια, δηλαδή να γίνεται η συλλογή των δεδομένων που χρειάζομαστε από το αρχείο που μας έχει δωθεί από τον καθηγητή. Συγκεκριμένα, εμείς κρατάμε την στήλη με τα σχόλια και την στήλη με τις φυσαλίδες.

Για την εκπόνηση της διπλωματικής εργασίας και για την έρευνα που κάναμε, χρειαστήκαμε μόνο την συγκεκριμένη καρτέλα παίρνοντας στοχευμένες πληροφορίες από κάθε στήλη.

EIKONA 10



reviews.xlsx

Review's Title	Reviewer's Location	Full Review	Rating	Hotel's Name	Hotel's Location	Hotel's Class
Great location, comfortable.Neo-classical bou	Messery, France	Nice. Brilliant location opposite the cathedral. Bed an	5 of 5 bubbles	The Zillers Boutique Hotel	Athens	4 Stars
Great service and comfort	Cincinnati, Ohio	The upscale hotel Daios has much to offer its guests	4 of 5 bubbles	Daios Luxury Living	Thessaloniki	5 Stars
Perfect location for gourmet visit		Nice hotel with friendly staff and free parking near the	4 of 5 bubbles	The Bristol Hotel	Thessaloniki	5 Stars
Best breakfast & service	New York City, New York	I love this hotel, stayed here last year and repeated t	5 of 5 bubbles	Archipelagos	Mykonos	5 Stars
Best Hotel in mykonos	Stockholm, Sweden	Good Hospitality & Friendly Recepction Front desk a	5 of 5 bubbles	Kirini - My Mykonos Retreat	Mykonos	5 Stars
Fairly nice hotel, not much amenities	Melville, New York	If you want a hotel walking distance from town but do	3 of 5 bubbles	Apanema Resort	Mykonos	4 Stars
Not worth it!!	California	We stayed at San Antonio Summerland 4 nights. We	2 of 5 bubbles	San Antonio Summerland Hotel	Mykonos	4 Stars
Spoiled our Wedding Anniversary		We stayed in a Panoramic Double Room only to find	3 of 5 bubbles	Aphrodite Beach Hotel	Mykonos	4 Stars
Fantastic experience!!!		It was unbelievable experience!! Very smooth & fast	5 of 5 bubbles	Tharroe of Mykonos Hotel	Mykonos	5 Stars
Amazing team	London	We've just spent a week here and can't agree more w	5 of 5 bubbles	Tharroe of Mykonos Hotel	Mykonos	5 Stars
Three days wasn't enough	Auburn, Alabama	Wow.....what can we say to give you a real insight i	5 of 5 bubbles	Petinos Hotel	Mykonos	4 Stars
A place to relax		A nice hotel with lovely interior and helpful staff, locat	4 of 5 bubbles	San Marco Hotel & Villas	Mykonos	4 Stars
Unreal Experience - Stunning Hotel!!	Philadelphia, Pennsylvania	Mykonos should be so proud for having Kouros as on	5 of 5 bubbles	Kouros Hotel & Suites	Mykonos	5 Stars
Mr Bex	Leopoldsburg, Belgium	One world on its own ..amazing ! Not possible to exp	5 of 5 bubbles	Kouros Hotel & Suites	Mykonos	5 Stars
Five-star by all means	Amman, Jordan	I wouldn't say it's luxurious but it definitely pays atten	5 of 5 bubbles	Myconian Imperial Resort	Mykonos	5 Stars
Outstanding!	Sydney, Australia	Just enjoyed a week at this beautiful spot. One of the	5 of 5 bubbles	Grace Mykonos Hotel	Mykonos	3 Stars
Great location and service	Sydney	We stayed for 4 nights and the first impression we ha	5 of 5 bubbles	Petasos Beach Resort & Spa	Mykonos	4 Stars
Nice location but poor service and rooms		We stayed for 3 nights in the deluxe sea view room,	3 of 5 bubbles	Petasos Beach Resort & Spa	Mykonos	4 Stars
Wonderful holiday		My husband and I have just returned from a lovely we	4 of 5 bubbles	Mykonos Essence Hotel	Mykonos	4 Stars
Beautiful quaint seaside hotel		We had a lovely time at this hotel. The staff are all pl	4 of 5 bubbles	Nissaki Boutique Hotel	Mykonos	5 Stars
Large, wonderful hotel- has everything you ne	Toronto, Canada	Nice hotel, great location. The concierge was incredi	4 of 5 bubbles	Saint John Hotel Villas & Spa	Mykonos	5 Stars
Nice but expensive	Amman, Jordan	As everything else in Mykonos, the hotel is convenie	3 of 5 bubbles	Grecotel Mykonos Blu Hotel	Mykonos	5 Stars
Wonderful hotel - with marvelous staff!	princeton nj	Marios, the owner, was amazing in providing guidanc	5 of 5 bubbles	Lithos by Spyros & Flora	Mykonos	4 Stars
OK but nothing special and beware...	London, United Kingdom	It's not bad for the price but it's very busy. I'm not sur	4 of 5 bubbles	Mykonos Grand Hotel & Resort	Mykonos	4 Stars
Quick but lovely!		Excellent, helpful staff and a beautiful facility. The loc	5 of 5 bubbles	Mykonos Bay Hotel	Mykonos	4 Stars
Not a 5 star hotel	Belgium	My husband and I stayed at this hotel so n April. I do	3 of 5 bubbles	SENTIDO Pearl Beach	Crete	4 Stars
When hotels and scams rhyme together	Vitry-sur-Seine, France	On August, we booked a 2 weeks stay in the Harma	1 of 5 bubbles	Harma Boutique Hotel	Crete	4 Stars
Excellent service brilliant stay	Telford, United Kingdom	We stayed here for part of our honeymoon the rooms	5 of 5 bubbles	Harma Boutique Hotel	Crete	4 Stars
Amazing		What an amazing place to relax and chill out. Well a	5 of 5 bubbles	Pleiades Luxurious Villas	Crete	4 Stars
Should be 5*	Sheffield, United Kingdom	Trip with my mum and grandma- we decided to			Crete	4 Stars
Not worth the money		The room was small. The bed sheets and pillow			Rome	4 Stars

8 of 10

General & Traveler Ranking | Prices | Reviews

EIKONA 11

```
1 import pandas as pd
2
3 def getReviews():
4     data = pd.read_excel(r'/Users/thodorisnik/Desktop/SentimentAnalysis/reviews.xlsx', sheet_name='Reviews')
5     df = pd.DataFrame(data, columns=['Full Review'])
6     return df.values
7
8 def getBubbles():
9     data = pd.read_excel(r'/Users/thodorisnik/Desktop/SentimentAnalysis/reviews.xlsx', sheet_name='Reviews')
10    rate = pd.DataFrame(data, columns=['Rating'])
11    return rate.values
```

Ανάλυση των συναρτήσεων που δημιουργήσαμε και χρησιμοποιήσαμε στην παραπάνω φωτογραφία (βλ. εικόνα 11)

DefgetReviews () :

Μας επιστρέφει μια λίστα με τα σχόλια όλων των πελατών τα οποία παίρνει από την καρτέλα Reviews (1), την στήλη Full Review (3) (βλ. εικόνα 11) από το αρχείο το οποίο βρίσκεται στο συγκεκριμένο path (4).

DefgetBubbles () :

Μας επιστρέφει μια λίστα με την βαθμολογία που έχουν δώσει οι πελάτες (2) (βλ. εικόνα 4.1)την οποία παίρνει από την καρτέλα Reviews (1), την στήλη Rating από το αρχείο που βρίσκεται στο συγκεκριμένο path (4).

4.1.2 Vader Sentiment Analysis

Η VADER (Valence Aware Dictionary and Sentiment Reasoner) είναι μία βιβλιοθήκη - λεξικό που είναι βασισμένο σε κανόνες εργαλείο ανάλυσης συναισθημάτων που προσαρμόζεται ειδικά στα συναισθήματα που εκφράζονται στα κοινωνικά μέσα. Η VADER χρησιμοποιεί ένα συνδυασμό ενόςλεξικού συναισθημάτων είναι ένας κατάλογος λεξικών χαρακτηριστικών (π.χ. λέξεις) που χαρακτηρίζονται γενικά σύμφωνα με τον σημασιολογικό προσανατολισμό τους ως θετικοί ή αρνητικοί. Το VADER όχι μόνο μιλάει για τη βαθμολογία Θετικότητας και Αρνητικότητας αλλά μας λέει επίσης για το πόσο θετικό ή αρνητικό είναι ένα συναίσθημα.

Παρακάτω παραθέτουμε μια εικόνα η οποία δείχνει πως έχει προσαρμοστεί στον κώδικα μας η παραπάνω βιβλιοθήκη.

Εικόνα 12

```
55 # fill in sentiment_score array
56
57 for s in range(len(reviews)):
58     sent_vader = list(SentimentIntensityAnalyzer().polarity_scores(reviews[s]).values())
59     sentiment_score.append(sent_vader[3])
60
```

Εδώ λοιπόν γίνεται ανάθεση των τιμών του κάθε σχολίου με βάση την συνάρτηση που αναφέραμε παραπάνω, οι τιμές που μπορεί να πάρει το κάθε σχόλιο είναι από -1 (extreme negative) μέχρι και 1 (extreme positive), και με την σειρά τους, αυτές οι τιμές μπαίνουν σε μια λίστα που έχουμε δημιουργήσει οι οποία περιέχει κάθε Sentiment Score για κάθε σχόλιο που υπάρχει, η αντιστοιχία είναι η εξής:

sentiment_score[0] = sentiment score 1^{ου}σχολίου

sentiment_score[1] = sentiment score 2^{ου}σχολίου

sentiment_score[2] = sentiment score 3^{ου}σχολίου

sentiment_score[n] = sentiment score n+1 σχολίου

Στην συνέχεια θα εξηγήσουμε πως γίνεται η κατανομή των σχολίων στα 5 είδη των κατηγοριών που αναλύσαμε στο προηγούμενο κεφάλαιο, τα οποία είναι τα παρακάτω, ξεκινώντας από το πιο αρνητικό και φτάνοντας στο πιο θετικό

- **DISASTER**
- **MANY THINGS NEEDS TO GET BETTER**
- **FAIR ENOUGH**
- **PERFECT**
- **ABSOLUTELY PERFECT**

Εικόνα 13

```
# getting positive or negative depending from sentiment_score[i]

for j in range(len(sentiment_score)):
    if sentiment_score[j] >= 0.70:
        program_feedback.append("ABSOLUTELY PERFECT")
        count_aperfect += 1
    elif 0.70 > sentiment_score[j] >= 0.40:
        program_feedback.append("PERFECT")
        count_perfect += 1
    elif 0.40 > sentiment_score[j] >= 0.1:
        program_feedback.append("FAIR ENOUGH")
        count_fairenough += 1
    elif 0.1 > sentiment_score[j] >= -0.4:
        program_feedback.append("MANY THINGS NEEDS TO GET BETTER")
        count_manythings += 1
    else:
        program_feedback.append("DISASTER")
        count_disaster += 1
    count_total += 1
```

Δημιουργούμε μια forloop η οποία θα γίνει τόσες φορές όσο είναι ο αριθμός των sentimentscore(δηλαδή ο αριθμός των σχολίων) και η οποία θα βλέπει μια – μια τιμή κάθε θέσης της λίστας που δημιουργήσαμε πριν και ανάλογα τι τιμή έχει η κάθε θέση, θα μπαίνει το αντίστοιχο σχόλιο, στην αντίστοιχη θέση μιας άλλης λίστας που αρχικοποιήσαμε και δημιουργήσαμε παραπάνω με όνομα **program_feedback**.

Πιο αναλυτικά:

- Για SentimentScore από **(-1,-0.4]**η τιμή που θα τοποθετηθεί στην αντίστοιχη θέση του πίνακα program_feedback είναι **“DISASTER”**

- Για SentimentScore από **(-0.4,0.1]**η τιμή που θα μπει στην αντίστοιχη θέση του πίνακα program_feedback είναι **“MANY THINGS NEEDS TO GET BETTER”**

- Για SentimentScore από **(0.1,0.4]**η τιμή που θα μπει στην αντίστοιχη θέση του πίνακα program_feedback είναι **“FAIR ENOUGH”**

- Για SentimentScore από **(0.4,0.7]**η τιμή που θα μπει στην αντίστοιχη θέση του πίνακα program_feedback είναι **“PERFECT”**

- Για SentimentScore από **(0.7,1]**η τιμή που θα μπει στην αντίστοιχη θέση του πίνακα program_feedback είναι **“ABSOLUTELY PERFECT”**

4.1.3 FuzzyWuzzy

Η συγκεκριμένη βιβλιοθήκη είναι ίσως η πιο σημαντική βιβλιοθήκη, για την υλοποίηση και εξαγωγή συμπερασμάτων, που χρησιμοποιήσαμε στην έρευνα μας.Παρακάτω σας παραθέτω κάποια ενδεικτικά παραδείγματακώδικα στα οποία χρησιμοποιήθηκε η FuzzyWuzzy

```
264 # CRITERIA A1 : Location
265 word_synonym = ["area","location","district"]
266 count_inits = 0
267 for i in word_synonym:
268     for j in user_reviews:
269         init = fuzz.token_set_ratio(j.lower(),i)
270         if i in j:
271             location_counter += 1
272             if init > 90:
273                 count_inits += 1
274 fw.write(f"Criteria word: {word_synonym[1].upper()}\n")
275 fw.write(f"In how many reviews word {word_synonym[1].upper()} and its synonyms appeared : {location_counter}\n")
276 fw.write(f"In how many reviews the score was over 90% matching with the word {word_synonym[1]} and its synonyms : {count_inits}\n")
277 fw.write("\n")
```

Εικόνα 14

Ο πίνακας word_synonym περιέχει συνώνυμα των λέξεων που αντιστοιχούν σε κάθε ένα κριτήριο. Ο λόγος που έγινε αυτό είναι γιατί μετά από αρκετές δοκιμές καταλήξαμε στο συμπέρασμα ότι υπάρχουν παραπάνω από μια λέξεις που μπορείς να χρησιμοποιήσεις

για να αναφερθείς στην ίδια λέξη, έτσι για κάθε ένα από τα εννιά κριτήρια που έχουμε, σκεφτήκαμε και κάποια συνώνυμα που πιθανώς να χρησιμοποίησε τουλάχιστον ένας χρήστης ώστε να αναφερθεί στο ίδιο κριτήριο.

Συγκεκριμένα, για το κριτήριο A1 : Location, βάλαμε σε ένα πίνακα άλλες δυο λέξεις που όταν χρησιμοποιούνται σε ένα σχόλιο παραπέμπουν στο κριτήριο αυτό.

Η παραπάνω υλοποίηση (βλ. εικόνα 14) μπορεί να φαίνεται μπερδεμένη αλλά στην ουσία είναι αρκετά απλή.

word_synonym: αρχικοποιούμε έναν πίνακα ο οποίος περιέχει συνώνυμα της λέξεις "Location" τα οποία μπορούν να χρησιμοποιηθούν κάλλιστα σε ένα σχόλιο ενός χρήστη.

count_inits:ένας μετρητής, ο οποίος αρχικοποιείται με τιμή 0, και μετράει πόσα σχόλια έχουνscorematchingμεγαλύτερο του 90 τοις εκατό.

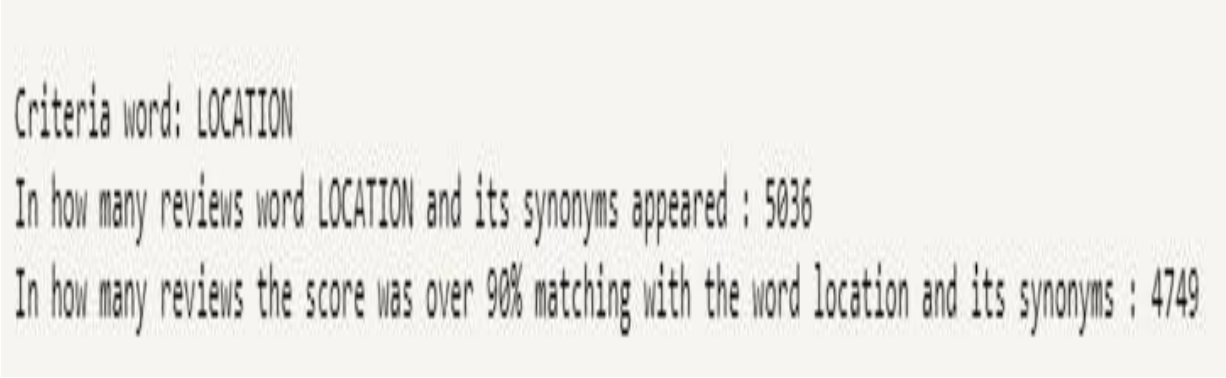
init:είναι η τιμή του matchingscoreπου παίρνει η μεταβλητή για κάθε σχόλιο.Αυτό το καταφέρνουμε χρησιμοποιώντας την συνάρτηση token_set_ratio(stringa,stringb) της βιβλιοθήκης fuzzyWuzzy.

Γιατί όμως χρησιμοποιούμε την συγκεκρίμενη συνάρτηση και όχι κάποια άλλη συνάρτηση της βιβλιοθήκης, όπως π.χ την ratio() ή την partial_ratio() ;

Απάντηση:Η προσέγγιση της **TokenSet** είναι παρόμοια με αυτές που αναφέραμε ακριβώς από πάνω αλλά είναι ακόμα πιο ευέλικτη. Στην συγκεκριμένη συνάρτηση, γίνεται **tokenize** και των δυο συμβολοσειρών αλλά **αντι** να γίνει ταξινόμηση και σύγκριση **αμέσως**, πρώτα τις χωρίζουμε σε δυο ομάδες: το(τα) σημεία τομής και το υπόλοιπο. Με αυτά τα δυο σύνολα που δημιουργούνται, η συνάρτηση τα χρησιμοποιεί ώστε να δημιουργήσει μια συμβολοσειρά σύγκρισης.

location_counter:ένας μετρητής, ο οποίος και αυτός έχει αρχικοποιηθεί με την τιμη 0, και μετράει πόσες φορές περιέχεται κάθε μια λέξη από τον πίνακα word_synonymσε κάθε ένα σχόλιο.

Με αυτόν τον τρόπο εξάγουμε αποτελέσματα για κάθε ένα κριτήριο, ενδεικτικά, το αποτέλεσμα για το συγκεκριμένο κριτήριο είναι το εξής:



Από το οποίο προκύπτει ότι για το κριτήριο αναζήτησης **LOCATION,** γίνεται αναφορά σε αυτό (ή κάποιο από τα συνώνυμα του) 5036 φορές και είχαμε τουλάχιστον 90% matchingσε 4749 σχόλια.

Που σημαίνει ότι το πρόγραμμα μας έχει μια μικρή απόκλισητης τάξεως του **5,7%.**

Στη συνέχεια θα παραθέσουμε με μορφή πίνακα τα αποτελέσματα όλων των κριτηρίων ώστε να εξάγουμε συνολικά συμπεράσματα και να δούμε τι αποκλίσεις υπάρχουν και στα υπόλοιπα.

Criteria Name	***Πίνακας συνώνυμων***	***Αριθμός αναφορών***	***Αριθμός αναφορών με matching > 90%***	***Απόκλιση (%)***
Location	*["area","location","district"]*	*5036*	*4749*	***5,7***
Staff	*["personnel","staff","crew"]*	*5736*	*5681*	***0,98***
Breakfast	*["brunch","breakfast","early meal"]*	*3887*	*3758*	***3,32***
Quiet	*["quietness","quiet"]*	*1079*	*1010*	***6,39***
Bed	*["bed"]*	*2435*	*1244*	***48,91***
Cleanliness	*["clean","cleanliness","purity","tidiness"]*	*4096*	*3555*	***13,21***
Room Space	*["room space","large room","small room"]*	*101*	*80*	***20,79***
Parking	*["parking","parking area","parking space","parking garage","car parking"]*	*307*	*300*	***2,28***
Interior Design	*["interior design","design","interior decoration","decor","decorating"]*	*911*	*404*	***55,65***

CriteriaName: Το όνομα του κριτηρίου

Αριθμός αναφορών: Ο αριθμός των αναφορών χρησιμοποιώντας έναν απλό μετρητη που ελέγχει αν περιέχεται η κάθε λέξη στα σχόλια.

Αριθμός αναφορών με matching> 90%: Ο αριθμός αναφορών που προκύπτει, χρησιμοποιώντας την συνάρτηση της βιβλιοθήκης fuzzyWuzzy, **fuzz.token_set_ratio()**.

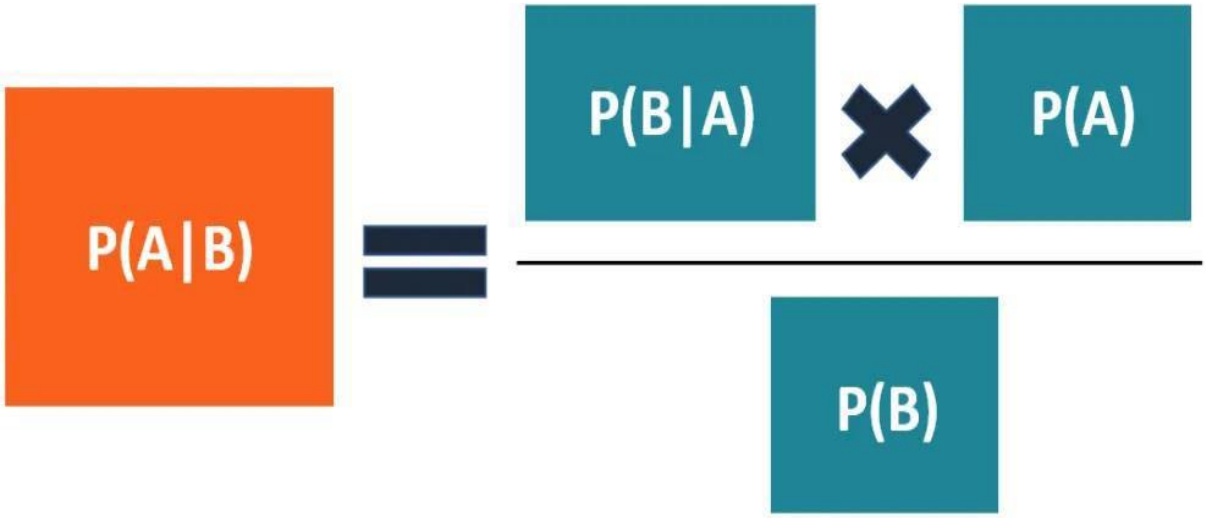
Απόκλιση (%): Πόσο τοις εκάτο απόκλιση υπάρχει μεταξύ των αριθμό αναφορών και αριθμό αναφορών με matchingscoreμεγαλύτερο του 90%.

Η αποκλίσεις που προκύπτουν, είναι αρκετά μικρές σε σχέση με τον όγκο των σχολίων που είχαμε να διαχειριστούμε.

4.2 Η χρήση του Bayes'theorem

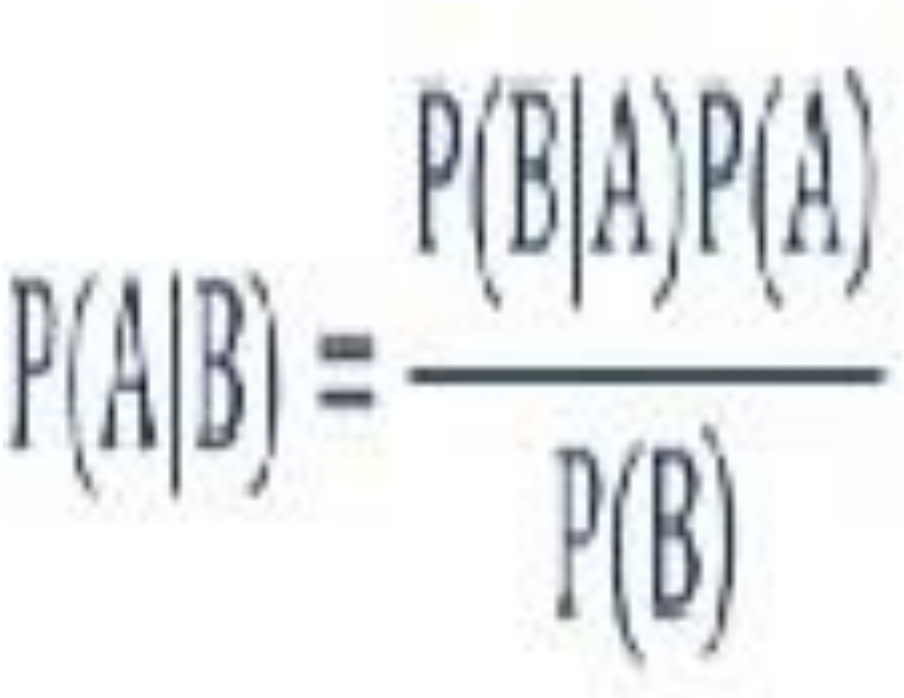
4.2.1 Περίληψη

Στην στατιστική και στη θεωρία πιθανοτήτων, το θεώρημα του Bayes(γνωστό και ως κανόνας του Bayes) είναι ένας μαθηματικός τύπος που χρησιμοποιείται για τον προσδιορισμό της υπό όρους πιθανότητας γεγονότων. Ουσιαστικά, το θεώρημα του Bayesπεριγράφει την πιθανότητα ενός γεγονότος με βάση την προηγούμενη γνώση των συνθηκών που μπορεί να σχετίζονται με το συμβάν.



Εκτός από την στατιστική, το παραπάνω θεώρημα χρησιμοποιείται επίσης σε διάφορους κλάδους, με την ιατρική και τη φαρμακολογία ως τα πιο αξιοσημείωτα παραδείγματα. Επιπλέον, το θεώρημα χρησιμοποιείται συνήθως και σε διάφορους τομείς των οικονομικών. Ορισμένες από τις εφαρμογές περιλαμβάνουν, χωρίς να περιορίζονται σε αυτές, μοντελοποίηση του κινδύνου δανεισμού χρημάτων σε δανειολήπτες ή πρόβλεψη της επιτυχίας μια επένδυσης.

Formula for Bayes' Theorem



Όπου:

- P(A|B) – η πιθανότητα να συμβεί το A, έχοντας ως δεδομένο ότι έχει συμβεί το B
- P(B|A) – η πιθανότητα να συμβεί το B,έχοντας ως δεδομένο ότι έχει συμβεί το A
- P(A) – η πιθανότητα να συμβεί το A
- P(B) – η πιθανότητα να συμβεί το B

4.2.2 Υλοποίηση του Θεωρήματος του Bayes

Σε αυτήν την ενότητα θα υλοποιήσουμε το θεώρημα του Bayesέχοντας ως μεταβλητές , 3 γεγονότα.

Με αυτόν τον τρόπο θέλουμε να δείξουμε πως συνδέονται:

- Το sentiment_score τουσχολίου
- Η βαθμολογία που έδωσε ο χρήστης
- Το κριτήριο αναφοράς

Ξεκινώντας, θα παραθέσουμε το κομμάτι κώδικα το οποίο είναι υπεύθυνο για τον υπολογισμό των παραπάνω πιθανοτήτων αλλά και για την εύρεση των επιμέρους πιθανοτήτων και μεταβλητών που χρειαστήκαμε για να υλοποιήσουμε το θεώρημα του Bayestριών (3) επιπέδων.

Στη συνέχεια θα σας αναφέρουμε παραθέσουμε και τα ενδεικτικά αποτελέσματα του κώδικα,εξηγώντας προς προέκυψαν και τι σημαίνουν στην πραγματικότητα και τέλος θα αναλύσουμε συγκεντρωτικά όλα τα αποτελέσματα που προέκυψαν,για όλες τις περιπτώσεις.

```
401 # Probabilities
402 # After loops are done, we calculate each criteria probability
403 # based on reviews number
404 # p character before every word stands for probability
405
406 plocation = (location_counter / counter_reviews) * 100
407 pstaff = (staff_counter / counter_reviews) * 100
408 pbr = (br_counter / counter_reviews) * 100
409 pquiet = (quiet_counter / counter_reviews) * 100
410 pbed = (bed_counter / counter_reviews) * 100
411 pclean = (clean_counter / counter_reviews) * 100
412 proomspace = (roomspace_counter / counter_reviews) * 100
413 pparking = (parking_counter / counter_reviews) * 100
414 pdesign = (design_counter / counter_reviews) * 100
415
```

Εδώ γίνεται ένας απλός υπολογισμός πιθανότητας του κάθε κριτήριου, καθώς την χρειαζόμαστε γιατί θα μας χρειαστεί για το θεώρημα του Bayes.

```
def
printLocationProbabilities(fopen,string,star_commentreview1,star
_commentreview2,star_commentreview3,star_commentreview4,star_com
mentreview5):
    fw = open("probabilitiesCriteria/location.txt",fopen)
    fw.write("Probability given the review is "+string+" with 5
bubbles and related to location word:
{:.2f}%\n".format((star_commentreview5 / 100) * plocation))
    fw.write("Probability given the review is "+string+" with 4
bubbles and related to location word:
{:.2f}%\n".format((star_commentreview4 / 100) * plocation))
    fw.write("Probability given the review is "+string+" with 3
bubbles and related to location word:
{:.2f}%\n".format((star_commentreview3 / 100) * plocation))
    fw.write("Probability given the review is "+string+" with 2
bubbles and related to location word:
{:.2f}%\n".format((star_commentreview2 / 100) * plocation))
    fw.write("Probability given the review is "+string+" with 1
bubbles and related to location word:
{:.2f}%\n".format((star_commentreview1 / 100) * plocation))
    fw.write("\n")
```

Η συγκεκριμένη συνάρτηση δέχεται ως όρισματα 7 μεταβλητές

- fopen – την πρώτη φορά καλείται με όρισμα “w” ώστε να δημιουργήσει το αρχείο, και έπειτα καλείται με το όρισμα “a” ,το οποίο σημαίνει appendκαι συνεχίζει να γράφει από το τέλος του αρχείου,δεν ξανα δημιουργεί καινούργιο.
- String – Sentiment score [DISASTER,MANY THINGS NEEDS TO GET BETTER,FAIR ENOUGH,PERFECT,ABSOLUTELY PERFECT]
- Star_commentreview3,Star_commentreview2…Star_commentreview5 – Οιπιθανότητεςτωνεκάστοτεφυσαλίδωνπουέχειδώσειοχρήστης

Κλήση της συνάρτησης:

```
# Probabilities of having 'x' bubbles (1 up to 5) given the
review status and is related to location

printLocationProbabilities("w","ABSOLUTELY
PERFECT",pfinalone_aperfect,pfinaltwo_aperfect,pfinalthree_aperf
ect,pfinalfour_aperfect,pfinalfive_aperfect)
printLocationProbabilities("a","PERFECT",pfinalone_perfect,pfina
ltwo_perfect,pfinalthree_perfect,pfinalfour_perfect,pfinalfive_p
erfect)
printLocationProbabilities("a","FAIR
ENOUGH",pfinalone_fe,pfinaltwo_fe,pfinalthree_fe,pfinalfour_fe,p
finalfive_fe)
printLocationProbabilities("a","MANY THINGS NEEDS TO GET
BETTER",pfinalone_many,pfinaltwo_many,pfinalthree_many,pfinalfou
r_many,pfinalfive_many)
printLocationProbabilities("a","DISASTER",pfinalone_ds,pfinaltwo
_ds,pfinalthree_ds,pfinalfour_ds,pfinalfive_ds)
```

Εξαγωγή αποτελέσματος:

location.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to location word: 29.20%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to location word: 13.76%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to location word: 4.43%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to location word: 1.07%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to location word: 0.56%

Probability given the review is PERFECT with 5 bubbles and related to location word: 8.50%
Probability given the review is PERFECT with 4 bubbles and related to location word: 12.53%
Probability given the review is PERFECT with 3 bubbles and related to location word: 16.12%
Probability given the review is PERFECT with 2 bubbles and related to location word: 7.19%
Probability given the review is PERFECT with 1 bubbles and related to location word: 4.68%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to location word: 5.19%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to location word: 9.34%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to location word: 16.34%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to location word: 9.34%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to location word: 8.82%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to location word: 4.32%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to location word: 5.77%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to location word: 17.09%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to location word: 10.71%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to location word: 11.12%

Probability given the review is DISASTER with 5 bubbles and related to location word: 0.75%
Probability given the review is DISASTER with 4 bubbles and related to location word: 2.61%
Probability given the review is DISASTER with 3 bubbles and related to location word: 9.79%
Probability given the review is DISASTER with 2 bubbles and related to location word: 13.67%
Probability given the review is DISASTER with 1 bubbles and related to location word: 22.19%
```

Από το παραπάνω προκύπτει ότι:

Το **29.20%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στην τοποθεσία.

Το **22.19%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στην τοποθεσία.

Εξαγωγή συμπεράσματος:

Ένα εύλογο ποσοστό και ευχαριστήμενων και δυσαρεστημένων πελάτων κάνουν αναφορά στην τοποθεσία, από αυτό το ποσοστό συμπεραίνουμε ότι το κριτήριο τοποθεσίας έρχεται από τα πρώτα πράγματα που βαθμολογεί ένας χρήστης σε ένα κατάλυμα.

bed.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to bed word: 14.12%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to bed word: 6.65%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to bed word: 2.14%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to bed word: 0.52%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to bed word: 0.27%

Probability given the review is PERFECT with 5 bubbles and related to bed word: 4.11%
Probability given the review is PERFECT with 4 bubbles and related to bed word: 6.06%
Probability given the review is PERFECT with 3 bubbles and related to bed word: 7.79%
Probability given the review is PERFECT with 2 bubbles and related to bed word: 3.48%
Probability given the review is PERFECT with 1 bubbles and related to bed word: 2.26%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to bed word: 2.51%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to bed word: 4.51%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to bed word: 7.90%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to bed word: 4.51%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to bed word: 4.26%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to bed word: 2.09%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to bed word: 2.79%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to bed word: 8.26%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to bed word: 5.18%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to bed word: 5.38%

Probability given the review is DISASTER with 5 bubbles and related to bed word: 0.36%
Probability given the review is DISASTER with 4 bubbles and related to bed word: 1.26%
Probability given the review is DISASTER with 3 bubbles and related to bed word: 4.73%
Probability given the review is DISASTER with 2 bubbles and related to bed word: 6.61%
Probability given the review is DISASTER with 1 bubbles and related to bed word: 10.73%
```

Από το παραπάνω προκύπτει ότι:

Το **14.12%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στο κρεβάτι.

Το **10.73%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στην τοποθεσία.

Εξαγωγή συμπεράσματος:

Σε αυτή την περίπτωση βλέπουμε ότι το κρεβάτι δεν παίζει και τόσο σημαντικό ρόλο στην αξιολόγηση των ξενοδοχείων μιας και τα δυο άκρα, δηλαδή ένας απόλυτα ευχαριστήμενος πελάτης και ένας τελείως δυσαρεστημένος πελάτης κάνουν αναφορά σε αυτό με ποσοστά μόλις 14.12% και 10.73%, αντίστοιχα.

staff.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to staff word: 33.26%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to staff word: 15.67%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to staff word: 5.04%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to staff word: 1.21%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to staff word: 0.64%

Probability given the review is PERFECT with 5 bubbles and related to staff word: 9.68%
Probability given the review is PERFECT with 4 bubbles and related to staff word: 14.27%
Probability given the review is PERFECT with 3 bubbles and related to staff word: 18.36%
Probability given the review is PERFECT with 2 bubbles and related to staff word: 8.19%
Probability given the review is PERFECT with 1 bubbles and related to staff word: 5.33%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to staff word: 5.91%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to staff word: 10.63%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to staff word: 18.61%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to staff word: 10.63%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to staff word: 10.04%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to staff word: 4.93%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to staff word: 6.57%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to staff word: 19.47%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to staff word: 12.20%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to staff word: 12.67%

Probability given the review is DISASTER with 5 bubbles and related to staff word: 0.85%
Probability given the review is DISASTER with 4 bubbles and related to staff word: 2.98%
Probability given the review is DISASTER with 3 bubbles and related to staff word: 11.15%
Probability given the review is DISASTER with 2 bubbles and related to staff word: 15.57%
Probability given the review is DISASTER with 1 bubbles and related to staff word: 25.27%
```

Από το παραπάνω προκύπτει ότι:

Το **33.26%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στην προσωπικό.

Το **25.27%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στο προσωπικό.

Εξαγωγή συμπεράσματος:

Σε αυτην την περίπτωση τα ποσοστά παρουσιάζουν μια ελαφρία αύξηση καθώς βλέπουμε ότι το κριτήριο του προσωπικού (“staff”) παίζει σημαντικό ρόλο στην αξιολόγηση που κάνει ο χρήστης,αν σκεφτεί κανείς ότι το 33.26% των ευχαρηστημένων πελατών κάνουν αναφορά στο προσωπικό και το 25.27% των δυσαρεστημένων κάνουν αναφορά επίσης.

breakfast.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to breakfast word: 22.54%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to breakfast word: 10.62%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to breakfast word: 3.42%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to breakfast word: 0.82%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to breakfast word: 0.43%

Probability given the review is PERFECT with 5 bubbles and related to breakfast word: 6.56%
Probability given the review is PERFECT with 4 bubbles and related to breakfast word: 9.67%
Probability given the review is PERFECT with 3 bubbles and related to breakfast word: 12.44%
Probability given the review is PERFECT with 2 bubbles and related to breakfast word: 5.55%
Probability given the review is PERFECT with 1 bubbles and related to breakfast word: 3.61%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to breakfast word: 4.00%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to breakfast word: 7.21%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to breakfast word: 12.61%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to breakfast word: 7.21%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to breakfast word: 6.81%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to breakfast word: 3.34%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to breakfast word: 4.45%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to breakfast word: 13.19%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to breakfast word: 8.27%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to breakfast word: 8.58%

Probability given the review is DISASTER with 5 bubbles and related to breakfast word: 0.58%
Probability given the review is DISASTER with 4 bubbles and related to breakfast word: 2.02%
Probability given the review is DISASTER with 3 bubbles and related to breakfast word: 7.55%
Probability given the review is DISASTER with 2 bubbles and related to breakfast word: 10.55%
Probability given the review is DISASTER with 1 bubbles and related to breakfast word: 17.13%
```

Από το παραπάνω προκύπτει ότι:

Το **22.54%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στο πρωινό.

Το **17.13%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στο πρωινό.

Εξαγωγή συμπεράσματος:

Εδώ τα ποσοστά είναι κάπως μοιρασμένα, καθώς 1 στους 5 πελάτες είτε είναι ευχαριστημένος είτε είναι δυσαρεστημένος θα κάνει κάποια αναφορά στην υπηρεσία του πρωινού που παρέχει το ξενοδοχείο. Το ποσοστό αυτό δεν είναι ιδιαίτερα αξιοσημείωτο , οπότε μπορούμε να καταλάβουμε ότι δεν επηρεάζει σε μεγάλο βαθμό την κριτική που θα κάνει ο κάθε πελάτης.

roomspace.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to roomspace word: 0.59%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to roomspace word: 0.28%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to roomspace word: 0.09%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to roomspace word: 0.02%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to roomspace word: 0.01%

Probability given the review is PERFECT with 5 bubbles and related to roomspace word: 0.17%
Probability given the review is PERFECT with 4 bubbles and related to roomspace word: 0.25%
Probability given the review is PERFECT with 3 bubbles and related to roomspace word: 0.32%
Probability given the review is PERFECT with 2 bubbles and related to roomspace word: 0.14%
Probability given the review is PERFECT with 1 bubbles and related to roomspace word: 0.09%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to roomspace word: 0.10%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to roomspace word: 0.19%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to roomspace word: 0.33%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to roomspace word: 0.19%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to roomspace word: 0.18%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to roomspace word: 0.09%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to roomspace word: 0.12%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to roomspace word: 0.34%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to roomspace word: 0.21%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to roomspace word: 0.22%

Probability given the review is DISASTER with 5 bubbles and related to roomspace word: 0.01%
Probability given the review is DISASTER with 4 bubbles and related to roomspace word: 0.05%
Probability given the review is DISASTER with 3 bubbles and related to roomspace word: 0.20%
Probability given the review is DISASTER with 2 bubbles and related to roomspace word: 0.27%
Probability given the review is DISASTER with 1 bubbles and related to roomspace word: 0.45%
```

Από το παραπάνω προκύπτει ότι:

Το **0.59%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στο χώρο δωματίου.

Το **0.45%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στο χώρο δωματίου.

Εξαγωγή συμπεράσματος:

Σε καμία περίπτωση το ποσοστό αναφοράς δεν ξεπερνάει το 0.59%, είναι με διαφορά το μικρότερο ποσοστό που έχουμε δει ως τώρα και από ότι φαίνεται οι πελάτες δεν ενδιαφέρονται καθόλου για τον χώρο που έχει το δωμάτιο.

quiet.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to quiet word: 6.26%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to quiet word: 2.95%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to quiet word: 0.95%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to quiet word: 0.23%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to quiet word: 0.12%

Probability given the review is PERFECT with 5 bubbles and related to quiet word: 1.82%
Probability given the review is PERFECT with 4 bubbles and related to quiet word: 2.68%
Probability given the review is PERFECT with 3 bubbles and related to quiet word: 3.45%
Probability given the review is PERFECT with 2 bubbles and related to quiet word: 1.54%
Probability given the review is PERFECT with 1 bubbles and related to quiet word: 1.00%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to quiet word: 1.11%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to quiet word: 2.00%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to quiet word: 3.50%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to quiet word: 2.00%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to quiet word: 1.89%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to quiet word: 0.93%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to quiet word: 1.24%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to quiet word: 3.66%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to quiet word: 2.29%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to quiet word: 2.38%

Probability given the review is DISASTER with 5 bubbles and related to quiet word: 0.16%
Probability given the review is DISASTER with 4 bubbles and related to quiet word: 0.56%
Probability given the review is DISASTER with 3 bubbles and related to quiet word: 2.10%
Probability given the review is DISASTER with 2 bubbles and related to quiet word: 2.93%
Probability given the review is DISASTER with 1 bubbles and related to quiet word: 4.75%
```

Από το παραπάνω προκύπτει ότι:

Το **6.26%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στο κριτήριο της ησυχίας.

Το **4.75%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στο κριτήριο της ησυχίας.

Εξαγωγή συμπεράσματος:

Μπορεί τα ποσοστά να μην είναι τόσο μικρά όπως ήταν στο προηγούμενο κριτήριο (βλέπε roomspace), όμως και πάλι προκύπτει ότι περίπου 5% των χρηστών κάνουν αναφορά στο συγκεκριμένο κριτήριο, είτε είναι ευχαριστημένοι είτε δυσαρεστημένοι, ένα ποσοστό το οποίο κατατάσσει το συγκεκριμένο κριτήριο χαμηλά στην ιεραρχία των προτιμήσεων του χρήστη.

parking.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to parking word: 1.78%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to parking word: 0.84%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to parking word: 0.27%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to parking word: 0.06%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to parking word: 0.03%

Probability given the review is PERFECT with 5 bubbles and related to parking word: 0.52%
Probability given the review is PERFECT with 4 bubbles and related to parking word: 0.76%
Probability given the review is PERFECT with 3 bubbles and related to parking word: 0.98%
Probability given the review is PERFECT with 2 bubbles and related to parking word: 0.44%
Probability given the review is PERFECT with 1 bubbles and related to parking word: 0.29%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to parking word: 0.32%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to parking word: 0.57%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to parking word: 1.00%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to parking word: 0.57%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to parking word: 0.54%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to parking word: 0.26%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to parking word: 0.35%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to parking word: 1.04%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to parking word: 0.65%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to parking word: 0.68%

Probability given the review is DISASTER with 5 bubbles and related to parking word: 0.05%
Probability given the review is DISASTER with 4 bubbles and related to parking word: 0.16%
Probability given the review is DISASTER with 3 bubbles and related to parking word: 0.60%
Probability given the review is DISASTER with 2 bubbles and related to parking word: 0.83%
Probability given the review is DISASTER with 1 bubbles and related to parking word: 1.35%
```

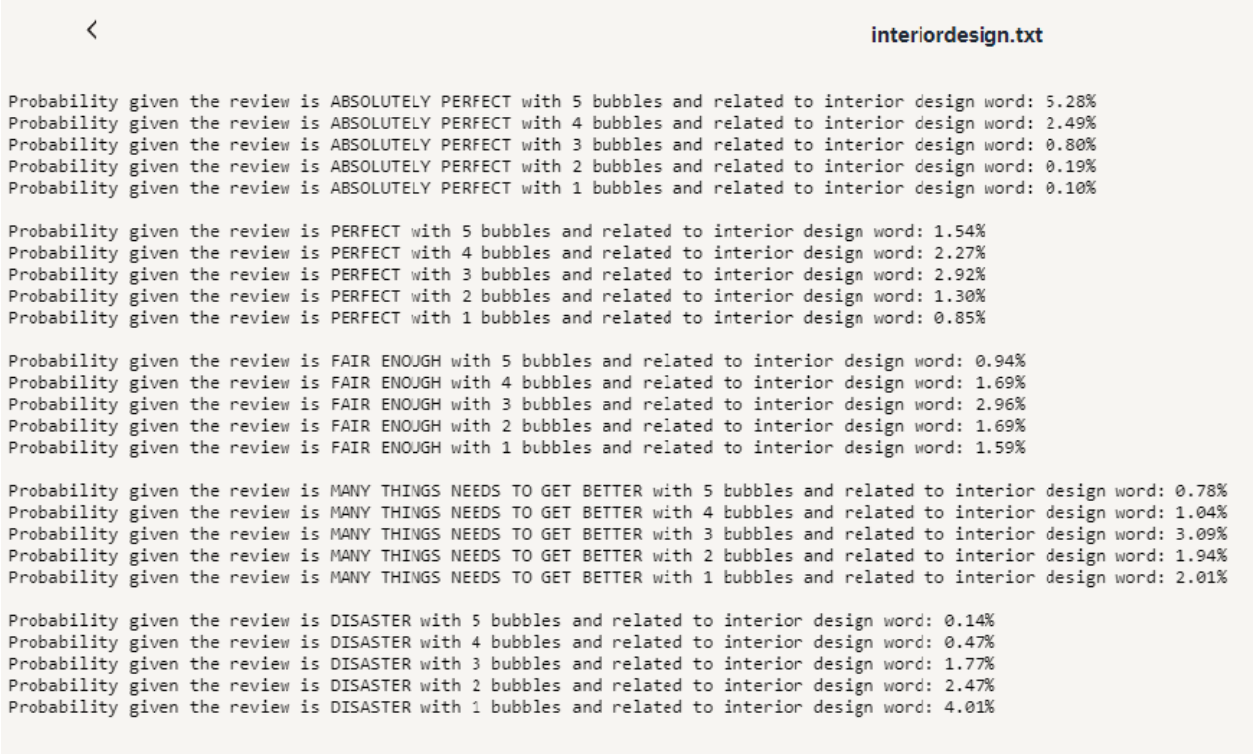
Από το παραπάνω προκύπτει ότι:

Το **1.78%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στο χώρο του parking.

Το **1.35%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στο χώρο του parking.

Εξαγωγή συμπεράσματος:

Αυτή η περίπτωση είναι παρόμοια με αυτή του roomspace, τα ποσοστά είναι πάρα πολύ μικρά,στην πραγματικότητα είναι αμελητέα και καταλαβαίνουμε ότι οι χρήστες δεν νοιάζονται καθόλου για το συγκεκριμένο κριτήριο.



Από το παραπάνω προκύπτει ότι:

Το **5.28%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στην εσωτερική διακόσμιση του δωματίου.

Το **4.01%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στην εσωτερική διακόσμιση του δωματίου.

Εξαγωγή συμπεράσματος:

Με μεγάλη έκπληξη παρατηρήσαμε ότι το συγκεκριμένο κριτήριο παρουσιάζει καλύτερα ποσοστά σε σχέση με τον χώρο στάθμευσης και από τον χώρου δωματίου. Περίπου 1 στους 20 πελάτες κάνει αναφορά στην εσωτερική διακόσμηση που περιβάλλει τον χώρο στον οποίο βρίσκεται, το ποσοστό προφανώς και δεν είναι μεγάλο, αλλά είναι μεγαλύτερο από το ποσοστό κάποιων κριτηρίων που πιστεύαμε ότι θα ήταν πιο ψηλά.

cleanliness.txt

```
Probability given the review is ABSOLUTELY PERFECT with 5 bubbles and related to cleanliness word: 23.75%
Probability given the review is ABSOLUTELY PERFECT with 4 bubbles and related to cleanliness word: 11.19%
Probability given the review is ABSOLUTELY PERFECT with 3 bubbles and related to cleanliness word: 3.60%
Probability given the review is ABSOLUTELY PERFECT with 2 bubbles and related to cleanliness word: 0.87%
Probability given the review is ABSOLUTELY PERFECT with 1 bubbles and related to cleanliness word: 0.46%

Probability given the review is PERFECT with 5 bubbles and related to cleanliness word: 6.91%
Probability given the review is PERFECT with 4 bubbles and related to cleanliness word: 10.19%
Probability given the review is PERFECT with 3 bubbles and related to cleanliness word: 13.11%
Probability given the review is PERFECT with 2 bubbles and related to cleanliness word: 5.85%
Probability given the review is PERFECT with 1 bubbles and related to cleanliness word: 3.81%

Probability given the review is FAIR ENOUGH with 5 bubbles and related to cleanliness word: 4.22%
Probability given the review is FAIR ENOUGH with 4 bubbles and related to cleanliness word: 7.59%
Probability given the review is FAIR ENOUGH with 3 bubbles and related to cleanliness word: 13.29%
Probability given the review is FAIR ENOUGH with 2 bubbles and related to cleanliness word: 7.59%
Probability given the review is FAIR ENOUGH with 1 bubbles and related to cleanliness word: 7.17%

Probability given the review is MANY THINGS NEEDS TO GET BETTER with 5 bubbles and related to cleanliness word: 3.52%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 4 bubbles and related to cleanliness word: 4.69%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 3 bubbles and related to cleanliness word: 13.90%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 2 bubbles and related to cleanliness word: 8.71%
Probability given the review is MANY THINGS NEEDS TO GET BETTER with 1 bubbles and related to cleanliness word: 9.04%

Probability given the review is DISASTER with 5 bubbles and related to cleanliness word: 0.61%
Probability given the review is DISASTER with 4 bubbles and related to cleanliness word: 2.13%
Probability given the review is DISASTER with 3 bubbles and related to cleanliness word: 7.96%
Probability given the review is DISASTER with 2 bubbles and related to cleanliness word: 11.12%
Probability given the review is DISASTER with 1 bubbles and related to cleanliness word: 18.05%
```

Από το παραπάνω προκύπτει ότι:

Το **23.75%** των χρηστών οι οποίοι ήταν απόλυτα ευχαριστημένοι με την διαμονή τους κάνουν αναφορά στην καθαριότητα του δωματίου.

Το **18.05%** των χρηστών οι οποίοι ήταν απόλυτα δυσαραστημένοι με την διαμονή τους, κάνουν αναφορά στην καθαριότητα του δωματίου.

Εξαγωγή συμπεράσματος:

¨Όπως ήταν αναμενόμενο, η συγκεκριμένη κατηγορία απασχολέι αρκετούς χρήστες μιας και ποσοστά σε όλες σχεδόν τις κατηγορίες χρηστών, από τέρμα ευχαριστημένοι μέχρι τέρμα δυσαρεστημένοι, κάνουν αναφορά στην καθαριότητα.

ΚΕΦΑΛΑΙΟ 5 - Συμπεράσματα και Μελλοντική Εργασία

Συμπεράσματα και Μελλοντική Εργασία

Η Ανάλυση Συναισθήματος ορίζεται ως η αυτοματοποιημένη εξαγωγή άποψης από ταξινομητές (ή μηχανές). Η διαδικασία έχει πολλά κοινά χαρακτηριστικά με τον άνθρωπο. Και στις δύο περιπτώσεις το ερώτημα είναι το ίδιο: Έχει τις γνώσεις να το κάνει; Τόσο ο ταξινομητής όσο και ο άνθρωπος πρέπει να γνωρίζουν καλά το αντικείμενο για να μπορέσουν να ταξινομήσουν ένα κείμενο ή ένα σύνολο δεδομένων ως θετικό, αρνητικό ή ουδέτερο. Ακόμη και μεταξύ δύο ανθρώπων που έχουν διαφορετικό γνωστικό επίπεδο, διαφορετικές εμπειρίες και διαφορετικά επίπεδα αναφοράς υπάρχει η πιθανότητα της απόκλισης όταν πρόκειται να ταξινομήσουν ένα κείμενο. Το μεγαλύτερο μειονέκτημα του ανθρώπου είναι ο χρόνος που χρειάζεται για να αξιολογήσει ένα σύνολο δεδομένων. Ειδικά τους τεράστιους όγκους δεδομένων που παράγονται καθημερινά. Αυτό το πρόβλημα λύνεται με την χρήση των μηχανών (ή ταξινομητών). Το μεγαλύτερο μειονέκτημα των ταξινομητών είναι το κατά πόσο επικαιροποιημένη και εξειδικευμένη μπορεί να είναι η γνώση τους για να αξιολογήσουν σωστά τα σύνολα δεδομένων. Βλέπουμε ότι υπάρχει μία αλληλεξάρτηση, τέτοια που να απαιτεί από την μία πλευρά ο άνθρωπος να συντηρεί και να επικαιροποιεί-εκσυγχρονίζει τον ταξινομητή και από την άλλη πλευρά ο ταξινομητής να επεξεργάζεται όγκους δεδομένων σε αποδεκτούς χρόνους και φυσικά να παράγει σωστά αποτελέσματα.

Αξίζει να σημειωθεί πως το σύστημα που υλοποιήθηκε σε αυτή την εργασία θα μπορούσε να μετασχηματιστεί σε ανάλυση συναισθήματος των σχολίων σε πραγματικό χρόνο, αφού δεν χρειάζεται ανθρώπινη επίβλεψη για τη δημιουργία του training set. Επιπλέον, θα μπορούσαν να εξεταστούν τα αποτελέσματα αυτής της μεθόδου σε άλλες γλώσσες πέραν της Αγγλικής.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Agathangelou, P., Katakis, I., Rori, L., Gunopulos, D., & Richards, B. (2017). Understanding online political networks: The case of the far-right and far-left in Greece. In G. Ciampaglia, A. Mashhadi, & T. Yasseri, Social informatics - Part I (pp. 162-177). Springer.

Benedetto, F., & Tedeschi, A. (2016). Big data sentiment analysis for brand monitoring in social media streams by cloud computing. In W. Pedryez, & S.-M. Chen, Sentiment analysis and ontology engineering (pp. 341-377). Springer.

Biagioni, R. (2016). Sentiment analysis. In R. Biagioni, The SenticNet sentiment lexicon: exploring semantic richness in multi-word concepts (pp. 7-16). Springer.

Bogdanovski, A. (2015). Macedonia - back in the global spotlight. Norwegian Institute of International Affairs (23).

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In E. Cambria, D. Das, S. Bandyopadhyay, & A. Feraco, A practical guide to sentiment analysis (pp. 1-8). Springer.

Ceka, B. (2018, 4). Macedonia: A new beginning? Journal of Democracy , 29 (2), pp. 142-157.

Chen, H. (2012). Sentiment analysis. In H. Chen, Darkweb: exploring and data mining the dark side of the web (pp. 171-201). Springer

Dabrowski, M., & Myachenkova, Y. (2018, 2). The Western Balkans on the road to the European Union. Policy Contribution (04), pp. 1-23

Fernandez-Gavilanez, M., Juncal-Martinez, J., Garcia-Mendez, S., Costa-Montenegro, E., & GonzalezCastano, F. J. (2018, 2 26). Creating emoji lexica from unsupervised sentiment analysis of their descriptions. Expert systems with applications , 103, pp. 74-91.

Fidanovski, K. (2018). What's in a name? Possible ways forward in the Macedonian name dispute. School of Slavonic and Eastern Europeran Studies , 31 (1), pp. 18-44.

Guidi, M., Ruiz-Agundez, I., & Canga-Sanchez, I. (2012). Knowledge mining from the twitter social network: the case of Barack Obama. In A. Abraham, Computational social networks: mining and visualization (pp. 211-229). Springer.

Illik, G., & Gjurovski, M. (2014). Modernism as an obstacle: the postmodern nature of the European Union and the Republic of Macedonia. In C. T. Mojanoski, Macedonia and the Balkans, a hundred years after the World War I - Security and Euro-Atlantic integrations (pp. 373-384). Bitola Skopje: University "St. Kliment Ogridski"

Ivanovski, H. (2013). The Macedonia-Greece dispute/ difference over the name issue: mitigating the inherently unsolvable. New Balkan Politics (14), pp

Janev, K., Dimeski, S., & Ristov, I. (2014). Republic of Macedonia - Regional zone of conflict of interests and geopolitics. In A. Pajaziti, J. Jakimovski, H. Jashari, & J. Abdullai, The Balkans in the new millenium: From balkanization to eutopia (pp. 426-436). Skopje: Balkan Sociological Forum - Universitas Europae Orientalis Meridionalis.

Kranjc, J., Smailovic, J., Podpecan, V., Grcar, M., Znidarsic, M., & Lavrac, N. (2015). Active learning for sentiment analysis on data streams: methodology and workflow implementation in the clowdflows platform. Information Processing and Management , 51, pp. 187-203.

Liu, B. (2017). Many facets of sentiment analysis. In E. Cambria, D. Das, S. Bandyopadhyay, & A. Feraco, A practical guide to sentiment analysis (pp. 11-38). Springer

Luo, T., Chen, S., Xu, G., & Zhou, Z. (2013). Sentiment analysis. In T. Luo, S. Chen, G. Xu, & J. Zhou, Trust-based collective view prediction (pp. 53-68). Springer.

Mavromatidis, F. (2010, 3 22). The role of the European Union in the name dispute between Greece and FYR Macedonia. Journal of Contemporary European Studies , 18 (1), pp. 46-62.

McKenna, B., Myers, M. D., & Newman, M. (2017, 3 17). Social media in qualitative research: challenges and recommendations. Information and Organization , 27, pp. 87-99.

Mohammad, S. M. (2017). Challenges in sentiment analysis. In E. Cambria, D. Das, S. Bandyopadhyay, & A. Feraco, A practical guide to sentiment analysis (pp. 61-77). Springer.

Nikodinovska-Stefanovska, S. (2014). EU and regional cooperation in Western Balkans. In C. T. Mojanoski, Macedonia and the Balkans, a hundred years after the World War I - Security and Euro-Atlantic Integrations - Vol. II (pp. 339-353). Bitola Skopje: University "St. Kliment Ohridski".

Ryan, J. (2013). Sentiment analysis and text visualization within Irish politics. Institute of Technology Blanchardstown, Dublin, Ireland.

Sjoeberg, E. (2011). Battlefields of memory: The Macedonian conflict and Greek historical culture. Umea University.

Tanev, I. (1999, 1). Legal aspects of the use of a provisional name for Macedonia in the United Nations system. The American Journal of International Law , 93 (1), pp. 155-160.

Taylor, E. M., Rodriguez, C. O., Velasquez, J. D., Ghosh, G., & Banerjee, S. (2013). Web opinion mining and sentimental analysis. In J. D. Velasquez, V. Palade, & L. Jain, Advanced techniques ing web intelligence - 2 (pp. 105-125). Springer.

Zadrozny, P., & Kodali, R. (2013). Sentiment analysis. In P. Zadrozny, & R. Kodali, Big data analytics using Splunk (pp. 255-282). Springer.

Wang, H., & Zhai, C. (2017). Generative models for sentiment analysis and opinion mining. In E. Cambria, D. Das, S. Bandyopandhyay, & A. Feraco, A practical guide to sentiment analysis (pp. 107-130). Springer.

Tsakalidis, A., Papadopoulos, S., & Kompatsiaris, I. (2014). An enseble model for cross-domain polarity classification on twitter. International Conference on web information systems engineering (pp. 168-177). Springer.

Tziampiris, A. (2005). The name dispute in the Former Yugoslav Republic of Macedonia after the signing of the Interim Accord. In E. Kofos, Athens-Skopje: An uneasy symbiosis (1995-2002) (pp. 225-252). Athens: ELIAMEP

E. Ma, «3 basic approaches in Bag of Words which are better than Word Embeddings,» 3 basic approaches in Bag of Words which are better than Word Embeddings, 22 July 2018.

R. Walimbe, «Handling imbalanced dataset in supervised learning using family of SMOTE algorithm.,» Data Science Central, 24 April 2017.

E. K. Ikonomakis, V. Tampakas θαη S. Kotsiantis, «Text Classification Using Machine Learning Techniques,» Research Gate, ηφκ. 4, αξ. 8, pp. 966-974, 2005.

C. D. Manning, P. Raghavan θαη H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

E. Haddi, X. Liu θαη Y. Shi, «The Role of Text Pre-processing in Sentiment Analysis,» Science Direct, ηφκ. 17, pp. 26-32, 2013.

R. Xia, C. Zong θαη S. Li, «Ensemble of feature sets and classification algorithms for sentiment classification,» Science Direct, ηφκ. 181, αξ. 6, pp. 1138-1152, 2011.

A. Hamza, «Effectively Pre-processing the Text Data Part 1: Text 54 Cleaning,» Towards Data Science, 30 January 2019.

D. Behl, S. Handa θαη A. Arora, «A Bug Mining Tool to Identify and Analyze,» ICROIT, India, 2014.

«GeeksforGeeks,». Available: https://www.geeksforgeeks.org/confusion-matrix-machinelearning/.

J. Starkweather θαη A. K. Moske , «Multinomial Logistic Regression,» 2011.

B. Krishnapuram, L. Carin, M. A. Figueiredo θαη A. J. Hartemink, «Sparse Multinomial Logistic Regression:,» IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, ηφκ. 27, αξ. 6, 2005

GeeksforGeeks, «KDD Process in Data Mining,» [Ζιεθηξνληθφ]. Available: https://www.geeksforgeeks.org/kdd-process-in-datamining/.

«Monkey Learn,» [Ζιεθηξνληθφ]. Available: https://monkeylearn.com/sentiment-analysis/. [Πξφζβαζε 15 September 2019].

H. Ferreira, «Confusion matrix and other metrics in machine learning,» Medium, 5 April 2018.

imbalanced-learn, «Over-sampling,». Available: https://imbalancedlearn.readthedocs.io/en/stable/over_sampling.html.

M. F. Zafra, «Text Classification in Python,» Towards Data Science, 16 June 2019.

W. Scott, «TF-IDF from scratch in python on real world dataset,» Towards Data Science, 15 February 2019.

A. M. Kibriya, E. Frank, B. Pfahringer θα η G. Holmes, «Multinomial Naive Bayes for Text Categorization Revisited,» ζε AI 2004: Advances in Artificial Intelligence, 2004, pp. 488-499.

H. Singh, «Understanding Gradient Boosting Machines,» Towards Data Science, 3 November 2018.

P. M., «Random forest classifier for remote sensing classification