Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών
Κοινωνικά Πληροφοριακά Συστήματα

Μεταπτυχιακή Διατριβή

Ανάλυση Αξιολογήσεων Προϊόντων στον Ψηφιακό Κόσμο

Αντώνιος Παπαδάκης

Επιβλέπων Καθηγητής
Δημήτρης Αντωνιάδης

Μάιος 2020

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακό Πρόγραμμα Σπουδών

Κοινωνικά Πληροφοριακά Συστήματα

Μεταπτυχιακή Διατριβή

Ανάλυση Αξιολογήσεων Προϊόντων στον Ψηφιακό Κόσμο

Αντώνιος Παπαδάκης

Επιβλέπων Καθηγητής
Δημήτρης Αντωνιάδης

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων για απόκτηση μεταπτυχιακού τίτλου σπουδών στα *Κοινωνικά Πληροφοριακά Συστήματα* από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών του Ανοικτού Πανεπιστημίου Κύπρου.

Μάιος 2020

ΛΕΥΚΗ ΣΕΛΙΔΑ

Περίληψη

Η παρούσα μεταπτυχιακή διατριβή εμπίπτει στην εξόρυξη γνώμης (opinion mining) και στην ανάλυση συναισθήματος (sentiment analysis) μέσω επεξεργασίας της φυσικής γλώσσας (natural language processing). Σκοπός μας ήταν να αναλύσουμε αξιολογήσεις προϊόντων της πλατφόρμας BestPrice.gr. Για τη συλλογή των δεδομένων επιλέχθηκαν δέκα συγκεκριμένες κατηγορίες από τις οποίες συλλέχθηκαν αξιολογήσεις. Στο πρώτο στάδιο της διαδικασίας πραγματοποιήθηκε προεπεξεργασία των δεδομένων (preprocessing), δηλαδή έγινε αφαίρεση των stopwords και καθαρισμός από μη χρήσιμα δεδομένα που δεν είχαν κάποια ουσιαστική αξία. Μετά το στάδιο της προεπεξεργασίας επιλέχθηκαν οι τέσσερις σημαντικότερες κατηγορίες προϊόντων, δηλαδή αυτές με το μεγαλύτερο αριθμό αξιολογήσεων. Για κάθε κατηγορία δημιουργήσαμε word clouds για να βρούμε τις λέξεις με τη μεγαλύτερη συχνότητα και να εντοπίσουμε ομοιότητες και διαφορές των λέξεων αυτών μεταξύ των κατηγοριών. Πραγματοποιήθηκε λεξικογραφική ανάλυση μέσω ελληνικού συναισθηματικού λεξικού όπου υπολογίστηκαν 8 διαφορετικά scores τα οποία αφορούν το sentiment, το subjectivity και 6 affects τα οποία είναι τα anger, disgust, fear, happy, sad και surprise. Με χρήση γραφημάτων τύπου scatter plot και boxplot, υπολογίσαμε το συναίσθημα που εκφράζεται για κάθε κατηγορία καθώς και πώς αυτό επηρεάζεται από το rating. Τα αποτελέσματα έδειξαν ότι όσο ανεβαίνει το rating (1-5), τόσο μεγαλώνει και το score για το συναίσθημα (θετικό), τη χαρά αλλά και για την υποκειμενικότητα. Αντίθετα όσο ανεβαίνει το rating μειώνονται τα scores για θυμό, απέχθεια, φόβο, και λύπη, πράγμα το οποίο είναι λογικό όταν μιλάμε για μια θετική αξιολόγηση. Όσον αφορά το score της έκπληξης είναι ουδέτερο καθώς μπορεί να απεικονίζει και θετικό αλλά και αρνητικό συναίσθημα.

Summary

This postgraduate dissertation falls into the category of opinion mining and sentiment analysis through natural language processing. Our goal was to analyze product reviews of the BestPrice.gr platform. Ten specific categories were selected to collect the data from which reviews were collected. In the first stage of the process, pre-processing of the data took place, ie the stopwords were removed and cleaned of useless data that did not have any real value. After the pre-processing stage, the four most important product categories were selected, ie those with the highest number of ratings. For each category we created word clouds to find the words with the highest frequency and to identify similarities and differences of these words between the categories. A lexicographical analysis was performed through a Greek emotional dictionary where 8 different scores were calculated which concern sentiment, subjectivity and 5 affects which are anger, disgust, fear, happy, sad and surprise. Using scatter plot and boxplot graphs, we calculated the emotion expressed for each category and how this is affected by the rating. The results showed that the higher the rating (1-5), the higher the score for sentiment (positive), happiness and subjectivity. On the contrary, as the rating goes up, scores for anger, disgust, fear, and sadness decrease, which makes sense when it comes to a positive evaluation. As for the surprise score, it is neutral as it can reflect both positive and negative emotion.

Ευχαριστίες

Ευχαριστώ τον επιβλέποντα μου κύριο Δημήτρη Αντωνιάδη για την βοήθεια και την υπομονή του. Επιπλέον, ευχαριστώ την εταιρία BestPrice για τη συνεργασία της στη συλλογή και την παροχή των δεδομένων.

Περιεχόμενα

Κεφάλαιο 1
Εισαγωγή

Με τη ραγδαία εξέλιξη του διαδικτύου αλλά και της τεχνολογίας γενικότερα, το εμπόριο έχει περάσει σε άλλο επίπεδο, το ηλεκτρονικό. Το ηλεκτρονικό εμπόριο αποτελεί πλέον βασικό πυλώνα της αγοράς αγαθών όπως ρούχα, ηλεκτρονικά προϊόντα, οικιακές συσκευές κτλ., αλλά και απαραίτητο στοιχείο μιας επιχείρησης που έχει ως αντικείμενο την πώληση τέτοιων αγαθών. Ολοένα και περισσότερα ηλεκτρονικά καταστήματα κάνουν αισθητή την παρουσία τους με στόχο το καταναλωτικό κοινό προς όφελος και των δύο πλευρών. Αφενός οι καταναλωτές βρίσκουν πολύ εύκολα και άμεσα τα προϊόντα που έχουν ανάγκη και σκοπεύουν να αγοράσουν, αφετέρου οι επιχειρηματίες βρίσκουν εύκολα αγοραστές ή εν δυνάμει αγοραστές των προϊόντων που παράγουν ή απλώς εμπορεύονται. Με αυτό τον τρόπο και ο καταναλωτής είναι ευχαριστημένος που βρίσκει άμεσα το προϊόν που αναζητά χωρίς να πηγαίνει σε φυσικό κατάστημα περιμένοντας στην ουρά, αλλά και ο πωλητής αυξάνει τις πωλήσεις του και γίνεται πιο γνωστός στο καταναλωτικό κοινό μέσω του διαδικτύου.

Προτού προβεί σε αγορά ενός προϊόντος ένας καταναλωτής έχει πλέον τη δυνατότητα να αναζητήσει πληροφορίες για το προϊόν που τον ενδιαφέρει. Οι πληροφορίες αυτές μπορεί να αφορούν είτε τα επίσημα χαρακτηριστικά του, για παράδειγμα στην επίσημη σελίδα του προϊόντος, είτε αξιολογήσεις του προϊόντος αυτού από άλλους καταναλωτές που το έχουν ήδη αγοράσει. Επίσης, υπάρχουν περιπτώσεις όπου ο καταναλωτής θέλει να αγοράσει ένα προϊόν, πχ έναν αφυγραντήρα, αλλά δεν έχει στο μυαλό του κάποιον συγκεκριμένο. Οπότε θα πρέπει να αναζητήσει προϊόντα της συγκεκριμένης κατηγορίας (αφυγραντήρες), να τα συγκρίνει μεταξύ τους ως προς τα χαρακτηριστικά και τις αξιολογήσεις από άλλους χρήστες και να καταλήξει σε ένα από αυτά.

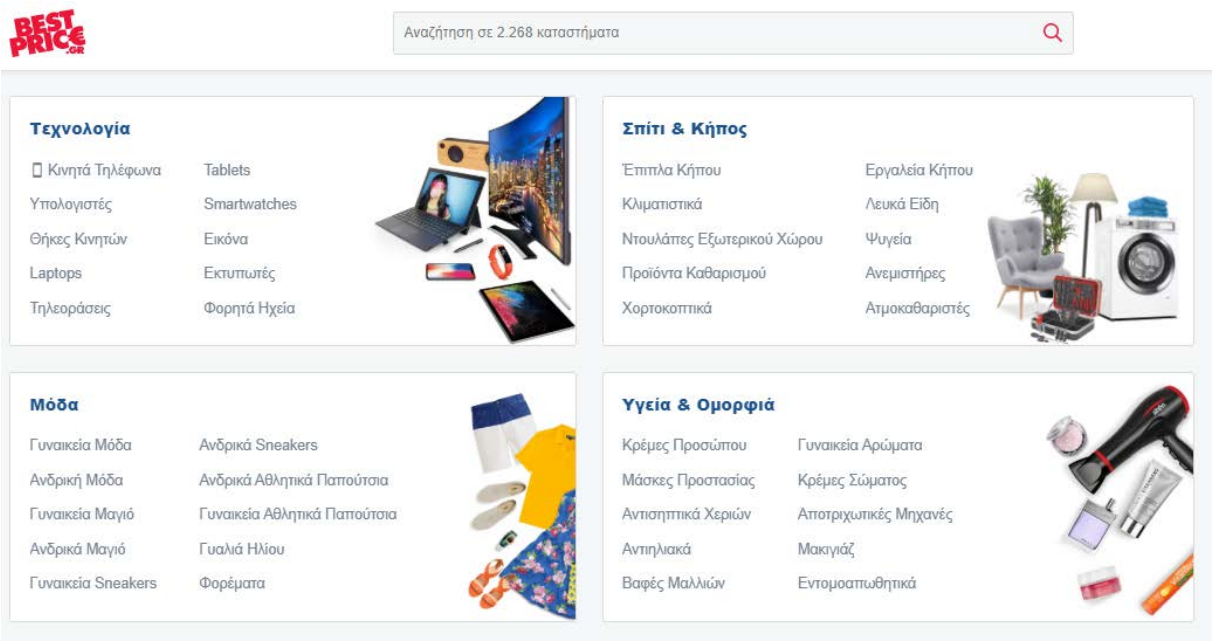
Ένα σημαντικό ζήτημα που προκύπτει είναι το κατά πόσο οι αξιολογήσεις προϊόντων είναι έγκυρες, αντικειμενικές και αληθής και αν γενικά υπάρχει αμεροληψία στο

κομμάτι αυτό, της δημοσίευσης αξιολογήσεων. Στην παρούσα διατριβή θα ασχοληθούμε με τη συλλογή και την ανάλυση αξιολογήσεων στον ψηφιακό κόσμο.

1.1 Σκοπός Διατριβής

Η παρούσα διατριβή έχει σκοπό την ανάλυση αξιολογήσεων προϊόντων που υπάρχουν στον ψηφιακό κόσμο και εμπίπτει στην περιοχή της εξόρυξης γνώμης (Opinion Mining) μέσω της ανάλυσης φυσικής γλώσσας (Natural Language Processing). Πέρα από τα ηλεκτρονικά καταστήματα υπάρχουν και άλλες διαδικτυακές πλατφόρμες οι οποίες παρέχουν δυνατότητες αναζήτησης προϊόντων, αναζήτησης πληροφοριών για αυτά καθώς και εύρεσης καταστημάτων όπου παρέχουν αυτά τα προϊόντα. Είτε μιλάμε για ένα e-shop, είτε για μια άλλη διαδικτυακή πλατφόρμα υπάρχουν αξιολογήσεις των προϊόντων από χρήστες που τα έχουν αγοράσει, δοκιμάσει και γενικά έχουν μια γνώμη για αυτά. Στόχος μας είναι να συλλεχθούν και να αναλυθούν δεδομένα μιας διαδικτυακής πλατφόρμας που θα αφορούν αξιολογήσεις προϊόντων.

Η πλατφόρμα με την οποία δουλέψαμε είναι η **www.bestprice.gr** όπου μπορείς να αναζητήσεις ανάμεσα σε εκατομμύρια προϊόντα (~ 9,5 , 10/2019) τα οποία παρέχονται από μεγάλο πλήθος καταστημάτων (~ 2000, 10/2019) που δηλώνουν την παρουσία τους μέσα στην πλατφόρμα. Ένας χρήστης μπορεί να κάνει εγγραφή, να δημιουργήσει προσωπικό προφίλ και να αναπτύξει «σχέσεις» με άλλους χρήστες (προσθήκη φίλου) και να γράψει αξιολογήσεις.



Εικόνα 1. Αρχική σελίδα του BestPrice.gr (2020)

1.2 Αναγκαιότητα και σπουδαιότητα της έρευνας

Η ανάγκη της έρευνας αυτής έγκειται στο γεγονός ότι δεν υπάρχει απόλυτη εγκυρότητα στον ψηφιακό κόσμο όσον αφορά το περιεχόμενο του. Ιδιαίτερα όταν υπάρχει στο τραπέζι το κέρδος, η εγκυρότητα αυτή μειώνεται δραματικά. Το ηλεκτρονικό εμπόριο έχει αντικαταστήσει μεγάλο μέρος του παραδοσιακού εμπορίου σε συγκεκριμένους τομείς, όπως είναι η τεχνολογία. Πλέον το καταναλωτικό κοινό έχει τη δυνατότητα να αναζητήσει, να ερευνήσει, να κρίνει και να αποφασίσει για ένα προϊόν που τον ενδιαφέρει είτε σε ένα ηλεκτρονικό κατάστημα, είτε σε ένα blog, είτε σε μια διαδικτυακή πλατφόρμα γενικά. Όμως αυτό που θα τον επηρεάσει περισσότερο είναι οι αξιολογήσεις του προϊόντος που αναζητά και όσο περισσότερες αξιολογήσεις βρει για αυτό το προϊόν τόσο πιο κοντά θα είναι στην απόφαση του είτε να το αγοράσει είτε να το απορρίψει. Το ζήτημα εδώ είναι κατά πόσο αυτές οι αξιολογήσεις είναι έγκυρες και αν τελικά βοηθούν κάποιον χρήστη ή αντιθέτως τον παραπλανούν είτε άθελα είτε ηθελημένα. Οι άνθρωποι δύσκολα μπαίνουν στη διαδικασία να αξιολογήσουν ένα προϊόν απλώς για «να κάνουν καλό» ενώ πιο εύκολα όταν ξέρουν ότι θα κερδίσουν κάτι. Για παράδειγμα εάν η πλατφόρμα διεξάγει ένα διαγωνισμό και ζητάει από χρήστες να αξιολογήσουν προϊόντα για να μπουν σε μια κλήρωση θα ήταν ένα καλό κίνητρο για να γράψουν fake αξιολογήσεις.

Τα αποτελέσματα της έρευνας αυτής θα συνεισφέρουν στο να κατανοήσουμε τη σπουδαιότητα των αξιολογήσεων προϊόντων στον ψηφιακό κόσμο καθώς και να σκεφτούμε τρόπους ελέγχου της ποιότητας και της εγκυρότητας αυτών. Αυτό θα έχει θετικό αντίκτυπο τόσο στην πλευρά του πωλητή, ο οποίος θέλει οι πελάτες του να βρίσκουν τις σωστές πληροφορίες όταν βρίσκονται στην πλατφόρμα του, όσο και στην πλευρά των αγοραστών οι οποίοι θέλουν να νιώθουν σιγουριά για τις πληροφορίες που λαμβάνουν μέσα από μια διαδικτυακή πλατφόρμα. Επιπρόσθετα, ο πωλητής ή ο ιδιοκτήτης της πλατφόρμας θα μπορεί να χρησιμοποιήσει τα αποτελέσματα έτσι ώστε να προσφέρει καλύτερες υπηρεσίες στους χρήστες, πχ να υπάρχει μια σημείωση, ένα hint, σε μια αξιολόγηση όπου να γράφει «*αυτή η αξιολόγηση είναι υπερβολική σε σχέση με τις υπόλοιπες*». Με αυτό τον τρόπο θα μπορεί και ο χρήστης να έχει μια πιο πραγματική εικόνα μιας αξιολόγησης.

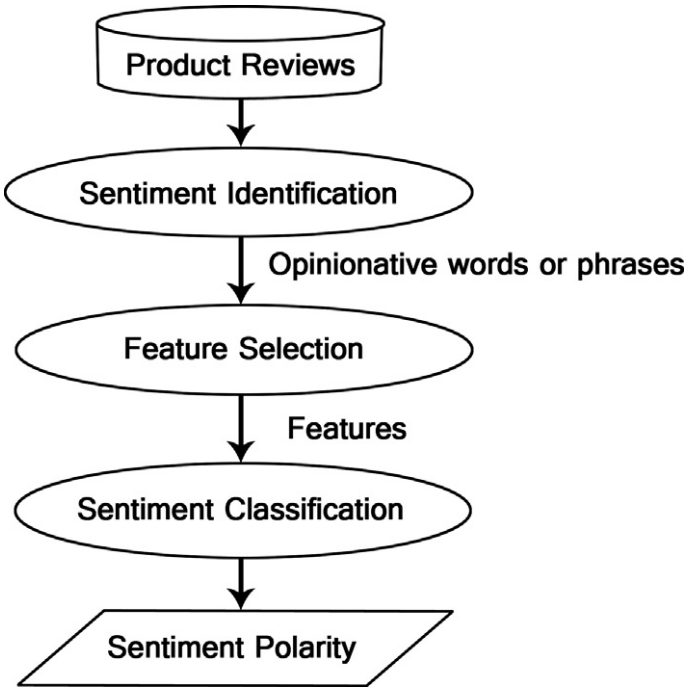
Αξίζει τέλος να αναφέρουμε ότι η παρούσα μεταπτυχιακή διατριβή θα συνεισφέρει στην ανάλυση συναισθήματος καθώς και στην εξόρυξη γνώμης πάνω στην ελληνική γλώσσα, όπου δεν υπάρχουν πάρα πολλές σχετικές μελέτες, σε σύγκριση με την αγγλική.

Κεφάλαιο 2
Βιβλιογραφική Ανασκόπηση

Σε αυτό το κεφάλαιο θα κάνουμε μια βιβλιογραφική ανασκόπηση για τις τεχνικές και τις μεθόδους που αφορούν την επεξεργασία φυσικής γλώσσας και την ανάλυση συναισθήματος καθώς και την εξόρυξη γνώμης.

2.1 Ανάλυση Συναισθήματος

Η ανάλυση συναισθήματος - ΑΣ (sentiment analysis), είναι μια υπολογιστική μελέτη απόψεων, συναισθημάτων και στάσεων που εκφράζονται σε κείμενα έναντι μιας οντότητας. Η ΑΣ συμβάλλει στην επίτευξη διαφόρων στόχων, όπως η παρατήρηση της δημόσιας διάθεσης σχετικά με την πολιτική κίνηση, η ευφυΐα της αγοράς, η μέτρηση της ικανοποίησης των πελατών, η πρόβλεψη πωλήσεων ταινιών και πολλά άλλα. Σύμφωνα με τους Walaa Medhat et al. (2014), ο στόχος της ΑΣ είναι να βρει γνώμες, να προσδιορίσει τα συναισθήματα που εκφράζουν και στη συνέχεια, να ταξινομήσει την πολικότητά τους. Η ΑΣ μπορεί να θεωρηθεί διαδικασία ταξινόμησης όπως φαίνεται στην **Εικόνα 2**. Υπάρχουν τρία βασικά επίπεδα ταξινόμησης στην ΑΣ: **document-level sentence-level** και **aspect-level**. Η ΑΣ σε document-level στοχεύει να ταξινομήσει ένα έγγραφο γνώμης ως έκφραση θετικής ή αρνητικής γνώμης ή συναισθήματος. Θεωρεί ολόκληρο το έγγραφο μια βασική μονάδα πληροφοριών. Η ΑΣ σε sentence-level στοχεύει στην ταξινόμηση των συναισθημάτων που εκφράζονται σε κάθε πρόταση. Το πρώτο βήμα είναι να προσδιορίσουμε εάν η πρόταση είναι υποκειμενική ή αντικειμενική. Εάν η πρόταση είναι υποκειμενική, η ΑΣ θα καθορίσει εάν η πρόταση εκφράζει θετικές ή αρνητικές απόψεις. Δεν υπάρχει ουσιαστική διαφορά μεταξύ document-level και sentence-level, επειδή οι προτάσεις είναι απλώς μικρά documents.



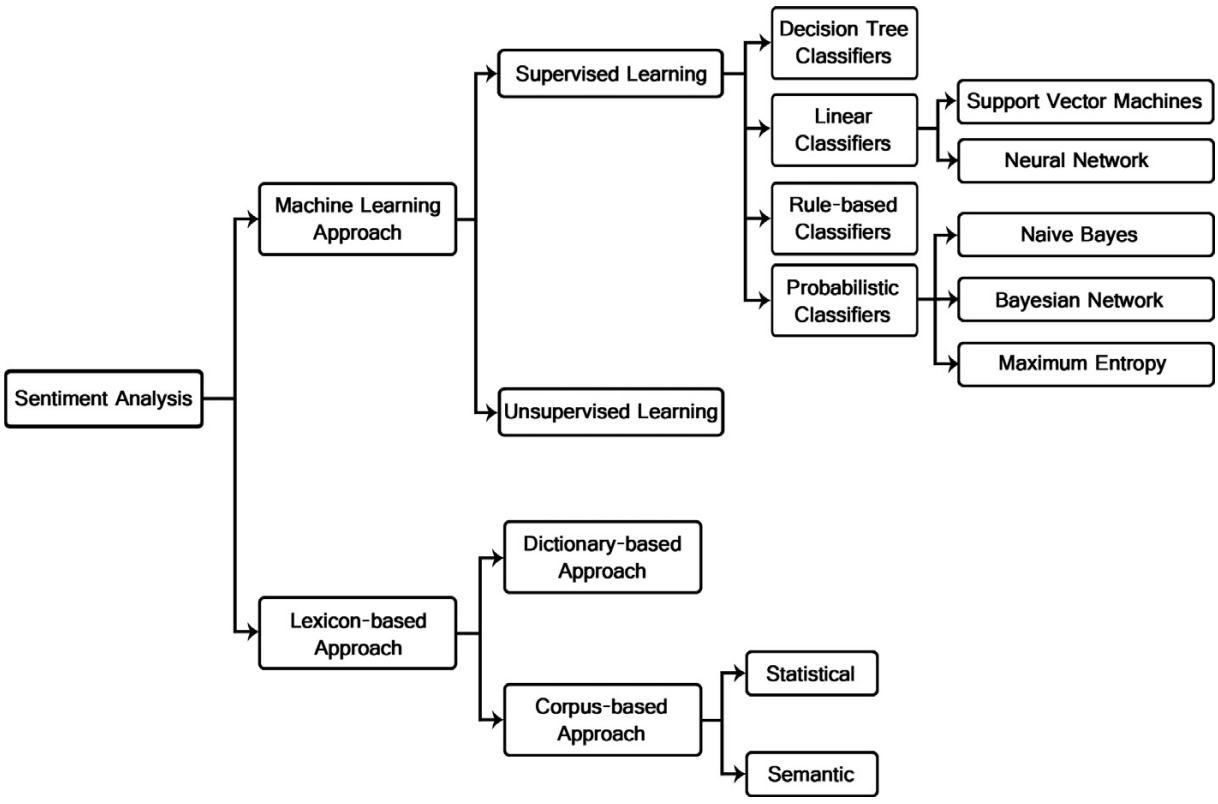
Εικόνα 2. Διαδικασία ανάλυσης συναισθήματος από τους Walaa Medhat et al. (2014)

Η ταξινόμηση σε document-level ή σε sentence-level δεν παρέχει αρκετές λεπτομέρειες σχετικά με τη γνώμη που εκφράζεται για όλες τις πτυχές της οντότητας, για να ληφθούν αυτές οι λεπτομέρειες πρέπει να πάμε σε Aspect-level sentiment analysis. Η ανάλυση σε επίπεδο aspect στοχεύει στην ταξινόμηση του συναισθήματος σε σχέση με τις συγκεκριμένες πτυχές των οντοτήτων. Το πρώτο βήμα είναι ο προσδιορισμός των οντοτήτων και των aspects τους. Κάποιος που εκφράζει μια γνώμη μπορεί να δώσει διαφορετικές απόψεις για διαφορετικές πτυχές της ίδιας οντότητας, όπως για παράδειγμα σε αυτή την πρόταση « *Η ποιότητα φωνής αυτού του τηλεφώνου δεν είναι καλή, αλλά η διάρκεια ζωής της μπαταρίας είναι μεγάλη* ».

Οι τεχνικές ταξινόμησης συναισθημάτων μπορούν να χωριστούν σε προσέγγιση μηχανικής μάθησης (Machine Learning - ML), προσέγγιση με βάση λεξικά (Lexicon-Based) και υβριδική προσέγγιση (Hybrid). Η προσέγγιση μηχανικής εκμάθησης εφαρμόζει τους διάσημους αλγόριθμους ML και χρησιμοποιεί γλωσσικά χαρακτηριστικά. Η προσέγγιση με βάση το λεξικό βασίζεται σε ένα λεξικό συναισθημάτων, δηλαδή μια συλλογή από γνωστούς και προκατασκευασμένους όρους συναισθημάτων. Χωρίζεται σε dictionary-based προσέγγιση και corpus-based προσέγγιση, που χρησιμοποιούν στατιστικές ή σημασιολογικές μεθόδους για να βρουν

την πολικότητα του συναισθήματος. Η υβριδική προσέγγιση συνδυάζει και τις δύο προσεγγίσεις και είναι πολύ συχνή μέθοδος. Οι μέθοδοι ταξινόμησης κειμένου που χρησιμοποιούν την προσέγγιση ML μπορούν να χωριστούν σε **supervised** και **unsupervised** μεθόδους μάθησης. Οι supervised μέθοδοι χρησιμοποιούν μεγάλο αριθμό εγγράφων εκπαίδευσης (training documents). Οι unsupervised μέθοδοι χρησιμοποιούνται όταν είναι δύσκολο να βρεθούν αυτά τα training documents.

Η προσέγγιση που βασίζεται στο λεξικό εξαρτάται από την εύρεση του λεξικού που θα χρησιμοποιηθεί για την ανάλυση του κειμένου. Υπάρχουν δύο μέθοδοι σε αυτήν την προσέγγιση, η **dictionary based**, η οποία εξαρτάται από την εύρεση λέξεων-seeds γνώμης και στη συνέχεια, αναζητά στο λεξικό συνώνυμα και ανώνυμα. Η **corpus-based** προσέγγιση, ξεκινά με μια λίστα λέξεων-seeds γνώμης και ψάχνει αντιστοιχία των λέξεων αυτών μέσα στο κείμενο. Στην παρούσα μελέτη θα χρησιμοποιήσουμε τη dictionary based προσέγγιση.



Εικόνα 3. Τεχνικές ταξινόμησης συναισθήματος, Walaa Medhat et al. (2014)

Σχετική μελέτη έχει πραγματοποιηθεί στο έργο “Low-Quality Product Review Detection in Opinion Summarization” (Jingjing Liu, et al, 2007), όπου η ανάλυση τους βασίζεται σε αξιολογήσεις προϊόντων της πλατφόρμας Amazon. Το κέντρο της ανάλυσης αυτής είναι η ποιότητα των αξιολογήσεων μέσω opinion mining. Η Amazon παρέχει τη δυνατότητα στους χρήστες να ψηφίζουν θετικά ή αρνητικά αξιολογήσεις άλλων χρηστών. Σύμφωνα με την έρευνα έχουν τεθεί τρεις τύποι μεροληψίας των χρηστών ως προς τις αξιολογήσεις. Ο πρώτος είναι «imbalance vote» οποίος αναφέρεται σε αξιολογήσεις όπου η έκταση τους είναι μικρή αλλά παρόλα αυτά έχουν πολλές θετικές ψήφους. Ο δεύτερος τύπος είναι «winner circle» όπου λέει ότι όσες περισσότερες ψήφους έχει μια αξιολόγηση τόσο περισσότερο θα επηρεάζεται και η αντικειμενικότητα του αναγνώστη. Ο τρίτος τύπος είναι ο «early bird». Εδώ αναφέρεται ότι η ημερομηνία δημοσίευσης μιας αξιολόγησης επηρεάζει τη συσσώρευση ψήφων σε αυτήν. Όσο πιο νωρίς γραφτεί μια αξιολόγηση τόσες περισσότερες ψήφους θα μαζέψει, οπότε αξιολογήσεις με ποιοτικό περιεχόμενο που έχουν δημοσιευτεί αργά θα έχουν πολύ λίγες ψήφους. Όσον αφορά την ποιότητα των αξιολογήσεων έχουν τεθεί τέσσερις κατηγορίες: 1) «η καλύτερη αξιολόγηση», όπου είναι μια πλήρης και πολύ λεπτομερής περιγραφή του προϊόντος με αποδεικτικά στοιχεία, 2) «καλή κρητική», όπου είναι μια καλή περιγραφή του προϊόντος αλλά χωρίς τις απαραίτητες αποδείξεις, 3) «τίμια κρητική», όπου είναι μια μικρή περιγραφή του προϊόντος χωρίς πολλές λεπτομέρειες, 4) «κακή κρητική», όπου είναι συνήθως μια λανθασμένη περιγραφή με παραπλανητικά στοιχεία για τον αναγνώστη.

2.1 Εξόρυξη Γνώμης

Σύμφωνα με τους Jesus Serrano-Guerrero et al. (2015), μια γνώμη θα μπορούσε απλώς να οριστεί ως θετικό ή αρνητικό συναίσθημα, άποψη, στάση ή εκτίμηση σχετικά με μια οντότητα (προϊόν, άτομο, γεγονός, οργανισμός ή θέμα) ή μια πτυχή αυτής της οντότητας από έναν χρήστη ή μια ομάδα χρηστών. Οι απόψεις μπορούν να ταξινομηθούν σε διαφορετικές ομάδες, για παράδειγμα, θα μπορούσαν να είναι κανονικές (regular) και συγκριτικές (comparative). Οι περισσότερες απόψεις είναι κανονικές και μπορούν να χωριστούν σε άμεσες ή έμμεσες. Οι άμεσες απόψεις εκφράζουν μια ιδέα για μια οντότητα ή μια πτυχή μιας οντότητας, ενώ οι έμμεσες

απόψεις εκφράζουν μια γνώμη για μια οντότητα ή μια πτυχή μιας οντότητας με βάση τις επιπτώσεις σε άλλες οντότητες. Από την άλλη πλευρά, οι συγκριτικές προτάσεις εκφράζουν την ομοιότητα μεταξύ οντοτήτων λαμβάνοντας υπόψη κοινές πτυχές ή χαρακτηριστικά. Επιπλέον, οι απόψεις μπορούν να ταξινομηθούν σε σαφείς (explicit) ή έμμεσες (implicit), ανάλογα με το αν εκφράζουν υποκειμενικές ή αντικειμενικές ιδέες. Εκτός από το συναίσθημα (sentiment) και τη γνώμη (opinion), υπάρχουν δύο κοντινές έννοιες, το subjectivity και το emotion. Μια υποκειμενική πρόταση μπορεί να εκφράσει κάποια προσωπικά συναισθήματα, απόψεις ή πεποιθήσεις, ωστόσο, δεν συνεπάγεται απαραίτητα κανένα συναίσθημα. Έτσι, η διαφορά μεταξύ αντικειμενικών και υποκειμενικών προτάσεων είναι ότι, μια αντικειμενική πρόταση εκφράζει κάποιες πραγματικές πληροφορίες για τον κόσμο, ενώ μια υποκειμενική πρόταση εκφράζει κάποια προσωπικά συναισθήματα, απόψεις ή πεποιθήσεις, πχ «Νομίζω ότι έχουν φύγει». Ωστόσο, η υποκειμενικότητα μερικές φορές περιλαμβάνει συναισθήματα σε κάποιο βαθμό όταν ασχολείται με επιρροή (affect), κρίση, εκτίμηση, κερδοσκοπία, συμφωνία κ.λπ. Από την άλλη πλευρά, ένα συναίσθημα (emotion) μπορεί να θεωρηθεί ως έκφραση των δικών μας υποκειμενικών συναισθημάτων και σκέψεων.

Σύμφωνα και πάλι με τους Jesus Serrano-Guerrero et al. (2015), το emotion σαν έννοια είναι πολύ κοντά στο sentiment καθώς, ο τρόπος μέτρησης της δύναμης μιας άποψης συνδέεται με την ένταση ορισμένων συναισθημάτων, όπως η αγάπη, η χαρά, η έκπληξη, ο θυμός, η θλίψη ή ο φόβος. Για παράδειγμα στην πρόταση: «Αγαπώ αυτό το αυτοκίνητο», ο ομιλητής εκφράζει την αντικειμενική του αγάπη για το αυτοκίνητό του. Είναι χρήσιμο να αναφερθεί και η έννοια της διάθεσης (mood), η οποία θα μπορούσε να θεωρηθεί ως μείγμα από sentiments, emotions και feelings που ωθούν τον συγγραφέα ενός συγκεκριμένου κειμένου να γράψει αυτό το σχόλιο, την παρατήρηση, την κριτική κ.λπ.

Από μαθηματικής άποψης, η γνώμη μπορεί να οριστεί ως μια πλειάδα αποτελούμενη από 5 μέρη (5-tuple) η οποία είναι:

$$(e_j, a_{jk}, so_{ijkl}, h_i, t_l)$$

Όπου το $\boldsymbol{e_j}$ αντιπροσωπεύει την οντότητα στόχο, το $\boldsymbol{a_{jk}}$ αντιπροσωπεύει το k-th χαρακτηριστικό (feature) αυτής της οντότητας, το $\boldsymbol{so_{ijkl}}$ είναι η τιμή του sentiment της

γνώμης και μπορεί να είναι θετική, αρνητική, ουδέτερη, το h_i αντιπροσωπεύει τον κάτοχο της γνώμης (opinion holder) και το t_l είναι η χρονική στιγμή όπου εκφράστηκε αυτή η γνώμη.

Τεχνικές «επεξεργασίας της φυσικής γλώσσας» περιγράφονται στο έργο "A review of natural language processing techniques for opinion mining systems" (Shiliang Sun, et al, 2016). Όπως αναφέρεται στον εν λόγω έργο, η εξόρυξη γνώμης απαιτεί αρκετά βήματα προ επεξεργασίας του κειμένου για την εξαγωγή χαρακτηριστικών όπως tokenization, τμηματοποίηση λέξεων, προσθήκη ετικετών Part Of Speech (POS), parsing. Η τεχνική tokenization είναι μια διαδικασία στην οποία γίνεται διαχωρισμός μιας πρότασης ή ενός αρχείου σε tokens, δηλαδή λέξεις ή προτάσεις. Κάποιες λέξεις όπως «αυτό», «η», «το» κτλ., αφαιρούνται καθώς δεν δίνουν κάποια σημαντική πληροφορία. Η τμηματοποίηση λέξεων μπορεί να χρησιμοποιηθεί σε γλώσσες όπως τα κινέζικα όπου οι λέξεις δεν έχουν όρια στο διαχωρισμό τους έτσι ώστε να διασπαστούν σε μονάδες όπου θα δίνουν κάποια πληροφορία. Οι ετικέτες POS καθώς η τεχνική parsing αποτελούν τεχνικές ανάλυσης. Οι ετικέτες POS μπορεί να είναι επίθετο ή ουσιαστικό και προστίθενται σε λέξεις μέσα στο κείμενο. Αυτές οι ετικέτες είναι χρήσιμες διότι σε ένα κείμενο όπου εκφράζεται μια γνώμη χρησιμοποιούνται επίθετα, ενώ τα αντικείμενα για τα οποία εκφράζεται η γνώμη είναι ουσιαστικά. Η τεχνική parsing αφορά συντακτικές πληροφορίες και αναπαριστά ένα δέντρο με τη γραμματική των λέξεων μιας πρότασης.

Κεφάλαιο 3
Μεθοδολογία

Η μεθοδολογία που ακολουθήθηκε στην παρούσα μεταπτυχιακή διατριβή ήταν καταρχάς η συλλογή των δεδομένων. Τα δεδομένα, μας τα παραχώρησε η εταιρία BestPrice για τις κατηγορίες που επιλέξαμε. Στο **Κεφάλαιο 4** εξηγούμε τη μορφή και τον όγκο του dataset που λάβαμε. Αφού λάβαμε τα δεδομένα προχωρήσαμε σε μια ομαδοποίηση αυτών, ανά κατηγορία και ανά rating. Έτσι καταλήξαμε στις 4 μεγαλύτερες κατηγορίες σε αριθμό αξιολογήσεων. Πρώτο βήμα πριν την ανάλυση των δεδομένων ήταν να γίνει μια προεπεξεργασία, να αφαιρεθούν stop words και να γίνει «καθάρισμα». Έπειτα χρησιμοποιήσαμε word clouds για να βρούμε τις λέξεις με τη μεγαλύτερη συχνότητα και να εντοπίσουμε ομοιότητες και διαφορές των λέξεων αυτών ανάμεσα στις 4 κατηγορίες. Τα παραπάνω περιγράφονται στο **Κεφάλαιο 5**. Επόμενο στάδιο ήταν να γίνει λεξικογραφική ανάλυση μέσω ελληνικού συναισθηματικού λεξικού. Υπολογίστηκαν 8 διαφορετικά scores τα οποία αφορούν το sentiment, το subjectivity και 6 affects τα οποία αφορούν τα anger, disgust, fear, happy, sad και surprise. Με χρήση γραφημάτων τύπου scatter plot και boxplot, υπολογίσαμε το συναίσθημα που εκφράζεται για κάθε κατηγορία καθώς και πώς αυτό επηρεάζεται από το rating. Τα παραπάνω περιγράφονται στο **Κεφάλαιο 6**. Τα συμπεράσματα αποτυπώνονται στο **Κεφάλαιο 7**.

Εικόνα 4. Τα στάδια της ανάλυσης

3.1 Εργαλεία

Για τις ανάγκες της παρούσας μεταπτυχιακής διατριβής, έγινε χρήση της γλώσσας προγραμματισμού **Python**[1]. Η Python είναι διερμηνευόμενη, γενικού σκοπού και υψηλού επιπέδου γλώσσα και χρησιμοποιείτε κατά κόρων για Data Analysis. Η ανάλυση έγινε με pure Python και δε χρησιμοποιήθηκε κάποιο framework. Ωστόσο, χρησιμοποιήθηκαν κάποιες βιβλιοθήκες και κάποια build-in packages τα οποία είναι το **json** package για να γίνει το parsing των δεδομένων, το **csv** package για να γίνει η χρήση του συναισθηματικού λεξικού, η βιβλιοθήκη **Pandas**[2] για να δημιουργήσουμε data frames (δομή δεδομένων 2 διαστάσεων) των δεδομένων μας έτσι ώστε να αποτυπωθούν στα διαγράμματα (plots) καθώς και η βιβλιοθήκη **Matplotlib**[3] για τη δημιουργία των scatter plots και boxplots. Όσον αφορά την προεπεξεργασία έγινε με regular expressions (RegEx) της Python χωρίς τη χρήση κάποιας βιβλιοθήκης. Για τη δημιουργία των word clouds χρησιμοποιήσαμε το online εργαλείο https://wordart.com.

[1] https://www.python.org/
[2] https://pandas.pydata.org/
[3] https://matplotlib.org/

Κεφάλαιο 4
Συλλογή Δεδομένων

Για τις ανάγκες της παρούσας μεταπτυχιακής διατριβής ζητήθηκαν δεδομένα αξιολογήσεων από την εταιρία BestPrice. Η ομάδα της BestPrice διέθεσε σε εμάς ένα ικανοποιητικό dataset αξιολογήσεων για δέκα συγκεκριμένες κατηγορίες τις οποίες ζητήσαμε. Παρακάτω γίνεται αναφορά στις κατηγορίες αυτές αλλά και στη δομή της κάθε αξιολόγησης.

4.1 Dataset

Για να επιλέξουμε με ποιες κατηγορίες θα ασχοληθούμε λάβαμε υπόψη τη δημοφιλία της κατηγορίας αλλά και τον αριθμό των προϊόντων μέσα στην πλατφόρμα. Επιπλέον, επιλέξαμε κατηγορίες από διαφορετικούς κλάδους όπως τεχνολογία και λευκές συσκευές. Οι κατηγορίες που επιλέχθηκαν φαίνονται στον **Πίνακα 1**.

ΑΑ	Κατηγορία
1	Κινητά Τηλέφωνα
2	Τηλεοράσεις
3	Tablets
4	Laptops
5	Αφυγραντήρες
6	Κλιματιστικά
7	Πλυντήρια Ρούχων
8	Οθόνες Υπολογιστών
9	Αθλητικά Όργανα Μετρήσεων
10	Κουζίνες

Πίνακας 1. Κατηγορίες προϊόντων

Το dataset που λάβαμε ήταν σε μορφή JSON[4] και αποτελείτο από ένα σύνολο **8.337** αξιολογήσεων.

Στο κεφάλαιο 5 κάνουμε μια αξιολόγηση σχετικά με τον αριθμό των αξιολογήσεων της κάθε κατηγορίας και επιλέγουμε ποιες θα χρησιμοποιήσουμε.

4.2 Δομή Αξιολόγησης

Στο BestPrice μια αξιολόγηση έχει συγκεκριμένη δομή, δηλαδή συγκεκριμένες πληροφορίες που σχετίζονται με αυτήν. Η κάθε αξιολόγηση περιλαμβάνει το **κείμενο αξιολόγησης**, τη **βαθμολογία** (1-5 αστέρια), **ημερομηνία και ώρα**, την **κατηγορία** του προϊόντος και τον **τίτλο** του προϊόντος. Στην **Εικόνα 5** φαίνεται η δομή μιας αξιολόγησης σε JSON μορφή.


```
{
  "review": "Άψογο κινητό, μοντέρνο design και αρκετά εύχρηστο",
  "rating": 5,
  "date": "2018.08.25 00:46:05",
  "category": "Κινητά Τηλέφωνα",
  "product": "Apple iPhone 6 16GB"
}
```


Εικόνα 5. Δομή μιας αξιολόγησης προϊόντων

Για την ανάλυση μας θα χρειαστούμε το **κείμενο**, το **rating** καθώς και την **κατηγορία** κάθε αξιολόγησης. Πριν να ξεκινήσουμε την ανάλυση μας, θα κάνουμε μια ομαδοποίηση των δεδομένων μας ανά κατηγορία και ανά rating (1-5). Αυτή η ομαδοποίηση θα μας βοηθήσει να δούμε ποιες κατηγορίες έχουν το μεγαλύτερο πλήθος αξιολογήσεων καθώς και το πλήθος αξιολογήσεων ανά rating.

[4] https://en.wikipedia.org/wiki/JSON

4.3 Ομαδοποίηση Δεδομένων

Αρχικά πραγματοποιήσαμε ομαδοποίηση των αξιολογήσεων ανά κατηγορία, έπειτα ανά αστέρι (rating) για όλες τις κατηγορίες συνολικά και τέλος ανά αστέρι για κάθε κατηγορία ξεχωριστά.

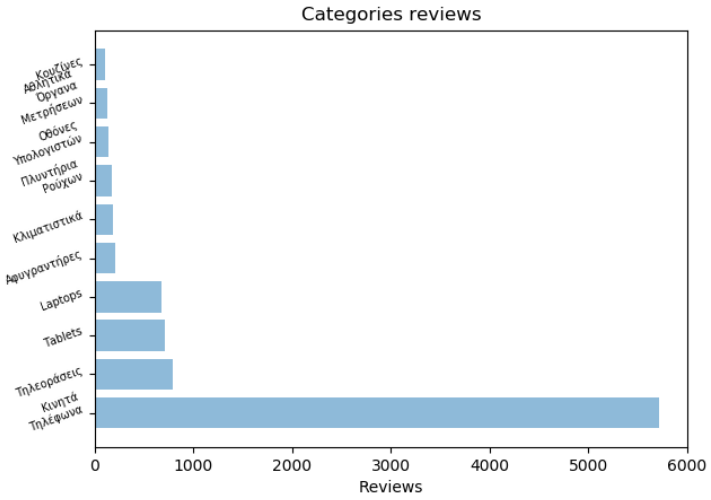
4.3.1 Ομαδοποίηση Δεδομένων ανά Κατηγορία

Υπολογίζοντας το πλήθος αξιολογήσεων ανά κατηγορία βρίσκουμε ότι οι κατηγορίες με τις περισσότερες αξιολογήσεις είναι Κινητά Τηλέφωνα, Τηλεοράσεις, Tablets και Laptops. Στον **Πίνακα 2** παρουσιάζεται αναλυτικά το πλήθος αξιολογήσεων για όλες τις κατηγορίες σε φθίνουσα σειρά.

ΑΑ	Κατηγορία	Πλήθος Αξιολογήσεων
1	Κινητά Τηλέφωνα	5717
2	Τηλεοράσεις	795
3	Tablets	712
4	Laptops	672
5	Αφυγραντήρες	210
6	Κλιματιστικά	183
7	Πλυντήρια Ρούχων	176
8	Οθόνες Υπολογιστών	144
9	Αθλητικά Όργανα Μετρήσεων	123
10	Κουζίνες	105

Πίνακας 2. Πλήθος αξιολογήσεων ανά κατηγορία

Τα αποτελέσματα αποτυπώνονται και στο **Διάγραμμα 1**.



Διάγραμμα 1. Αξιολογήσεις ανά κατηγορία - Ραβδόγραμμα

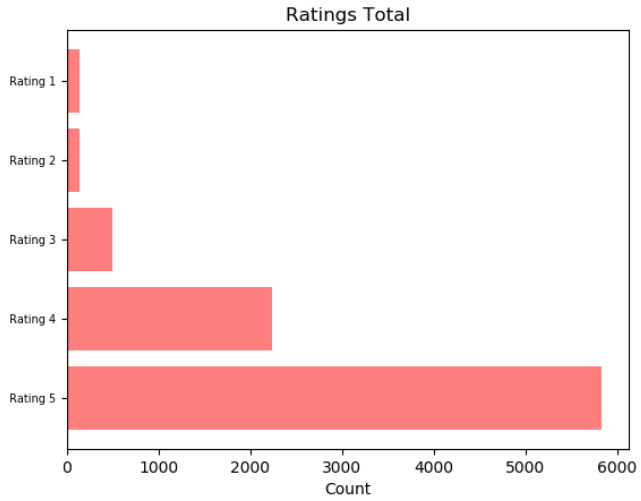
Είναι προφανές ότι η κατηγορία **Κινητά τηλέφωνα** έχει το μεγαλύτερο αριθμό αξιολογήσεων καθώς είναι πολύ δημοφιλής κατηγορία για το καταναλωτικό κοινό και σίγουρα έχει πολύ μεγάλο ενδιαφέρον για την ανάλυση μας.

4.3.2 Ομαδοποίηση Δεδομένων ανά Rating

Ομαδοποιώντας τις αξιολογήσεις ανά rating παρατηρούμε ότι περίπου το 66% του συνόλου αφορά το rating 5, περίπου το 25% αφορά το rating 4 ενώ για rating 3, 2 και 1 οι αξιολογήσεις είναι ελάχιστες. Στον **Πίνακα 3** βλέπουμε τις μετρήσεις καθώς και στο **Διάγραμμα 2** σε μορφή μπάρας.

Rating	Πλήθος Αξιολογήσεων
5	5830
4	2233
3	501
2	138
1	135

Πίνακας 3. Πλήθος αξιολογήσεων ανά rating



Διάγραμμα 2. Αξιολογήσεις ανά rating - Ραβδόγραμμα

4.3.3 Ομαδοποίηση Δεδομένων ανά Rating και ανά Κατηγορία

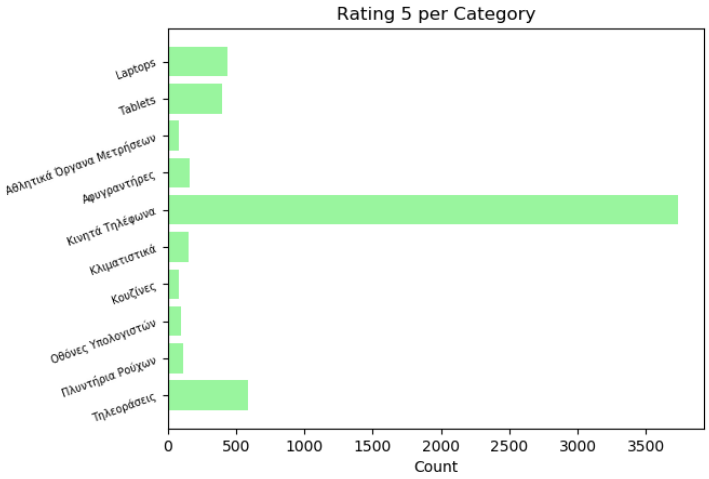
Σε αυτό το στάδιο υπολογίζουμε για κάθε κατηγορία πόσες αξιολογήσεις υπάρχουν για κάθε rating.

Για Rating 5

Στον **Πίνακα 4** παρατηρούμε το πλήθος αξιολογήσεων ανά κατηγορία καθώς και την οπτική απεικόνιση των μετρήσεων στο **Διάγραμμα 3**.

ΑΑ	Κατηγορία	Πλήθος Αξιολογήσεων	Ποσοστό επί του συνόλου
1	Laptops	435	~64.7%
2	Tablets	395	~55.4%
3	Αθλητικά Όργανα Μετρήσεων	83	~67.4%
4	Αφυγραντήρες	157	~74.6%
5	Κινητά Τηλέφωνα	3739	~65.4%
6	Κλιματιστικά	152	~83.0%
7	Κουζίνες	80	~76.2%
8	Οθόνες Υπολογιστών	93	~64.5%
9	Πλυντήρια Ρούχων	108	~61.3%
10	Τηλεοράσεις	588	~74.0%

Πίνακας 4. Πλήθος αξιολογήσεων ανά κατηγορία για rating 5



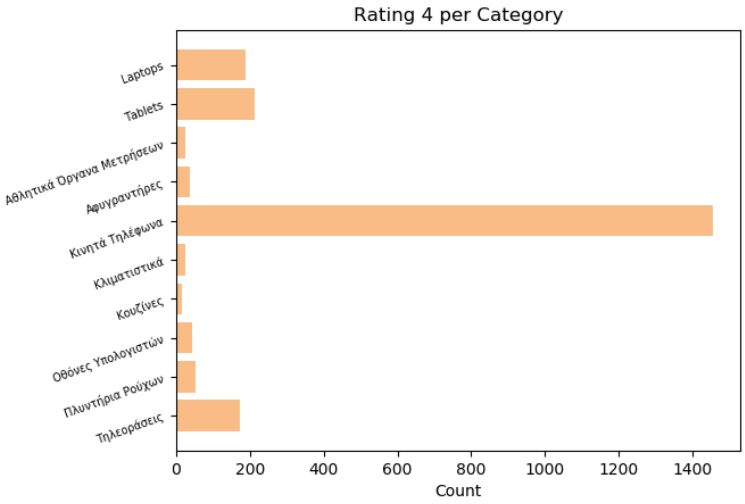
Διάγραμμα 3. Αξιολογήσεις ανά κατηγορία για rating 5 - Ραβδόγραμμα

Για Rating 4

Στον **Πίνακα 5** παρατηρούμε το πλήθος αξιολογήσεων ανά κατηγορία καθώς και την οπτική απεικόνιση των μετρήσεων στο **Διάγραμμα 4**.

ΑΑ	Κατηγορία	Πλήθος Αξιολογήσεων	Ποσοστό επί του συνόλου
1	Laptops	187	~27.8%
2	Tablets	213	~30.0%
3	Αθλητικά Όργανα Μετρήσεων	26	~21.1%
4	Αφυγραντήρες	38	~18.1%
5	Κινητά Τηλέφωνα	1457	~25.5%
6	Κλιματιστικά	25	~13.6%
7	Κουζίνες	16	~15.2%
8	Οθόνες Υπολογιστών	44	~30.5%
9	Πλυντήρια Ρούχων	53	~30.1%
10	Τηλεοράσεις	174	~21.8%

Πίνακας 5. Πλήθος αξιολογήσεων ανά κατηγορία για rating 4



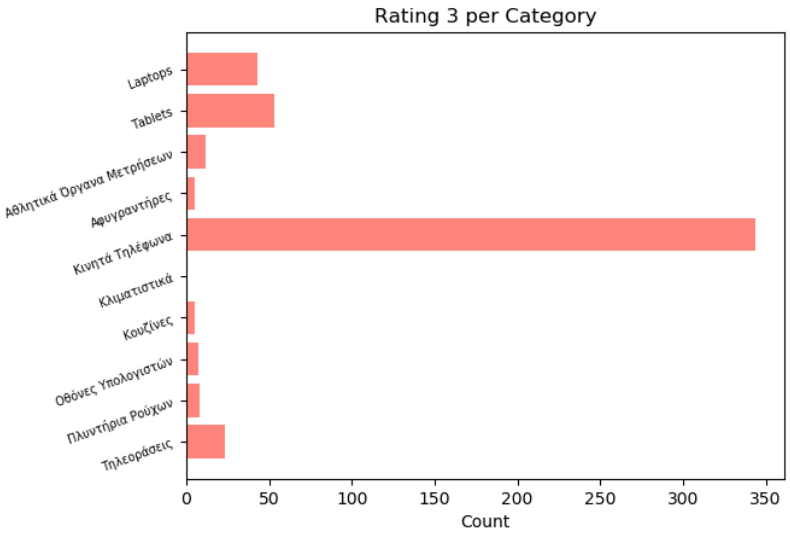
Διάγραμμα 4. Αξιολογήσεις ανά κατηγορία για rating 4 - Ραβδόγραμμα

Για Rating 3

Στον **Πίνακα 6** παρατηρούμε το πλήθος αξιολογήσεων ανά κατηγορία καθώς και την οπτική απεικόνιση των μετρήσεων στο **Διάγραμμα 5**.

ΑΑ	Κατηγορία	Πλήθος Αξιολογήσεων	Ποσοστό επί του συνόλου
1	Laptops	43	~6.4%
2	Tablets	53	~7.5%
3	Αθλητικά Όργανα Μετρήσεων	12	~9.7%
4	Αφυγραντήρες	5	~2.4%
5	Κινητά Τηλέφωνα	344	~6.0%
6	Κλιματιστικά	1	~0.5%
7	Κουζίνες	5	~4.7%
8	Οθόνες Υπολογιστών	7	~4.8%
9	Πλυντήρια Ρούχων	8	~4.5%
10	Τηλεοράσεις	23	~2.9%

Πίνακας 6. Πλήθος αξιολογήσεων ανά κατηγορία για rating 3



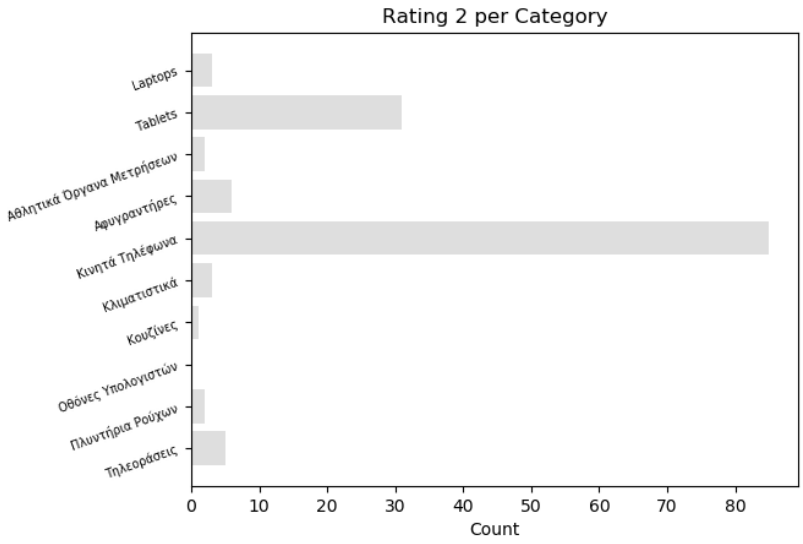
Διάγραμμα 5. Αξιολογήσεις ανά κατηγορία για rating 3 - Ραβδόγραμμα

Για Rating 2

Στον **Πίνακα 7** παρατηρούμε το πλήθος αξιολογήσεων ανά κατηγορία καθώς και την οπτική απεικόνιση των μετρήσεων στο **Διάγραμμα 6**.

ΑΑ	Κατηγορία	Πλήθος Αξιολογήσεων	Ποσοστό επί του συνόλου
1	Laptops	3	~0.4%
2	Tablets	31	~4.3%
3	Αθλητικά Όργανα Μετρήσεων	2	~1.6%
4	Αφυγραντήρες	6	~2.8%
5	Κινητά Τηλέφωνα	85	~1.5%
6	Κλιματιστικά	3	~1.6%
7	Κουζίνες	1	~1.0%
8	Οθόνες Υπολογιστών	0	~0.0%
9	Πλυντήρια Ρούχων	2	~1.1%
10	Τηλεοράσεις	5	~0.6%

Πίνακας 7. Πλήθος αξιολογήσεων ανά κατηγορία για rating 2



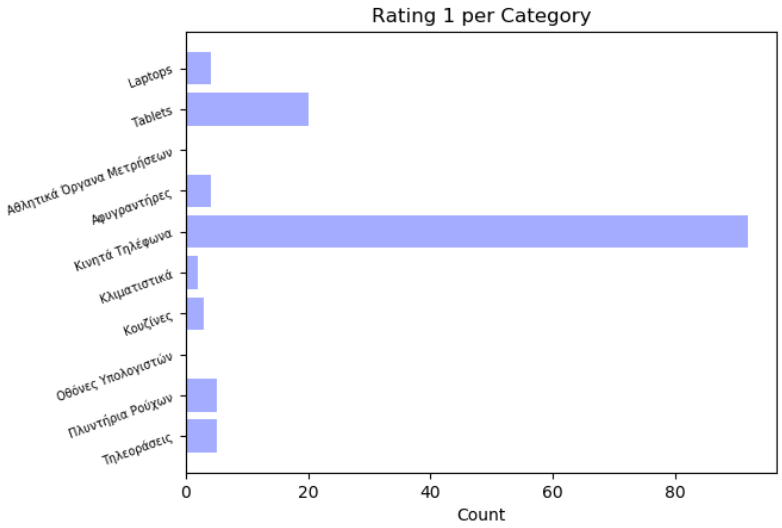
Διάγραμμα 6. Αξιολογήσεις ανά κατηγορία για rating 2 - Ραβδόγραμμα

Για Rating 1

Στον **Πίνακα 8** παρατηρούμε το πλήθος αξιολογήσεων ανά κατηγορία καθώς και την οπτική απεικόνιση των μετρήσεων στο **Διάγραμμα 7**.

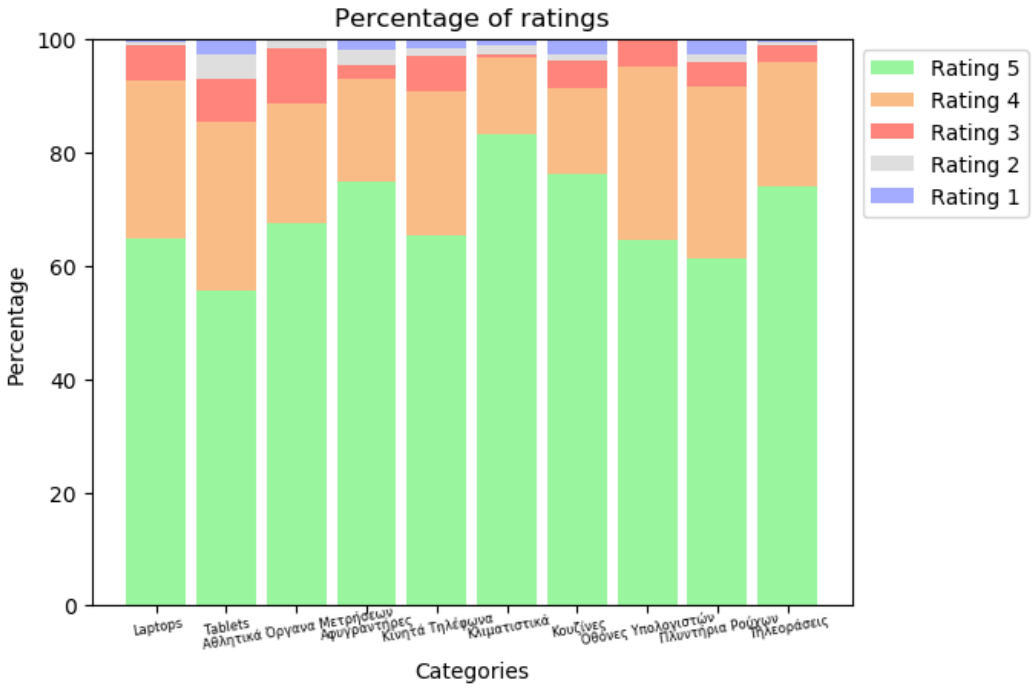
ΑΑ	Κατηγορία	Πλήθος Αξιολογήσεων	Ποσοστό επί του συνόλου
1	Laptops	4	~0.6%
2	Tablets	20	~2.8%
3	Αθλητικά Όργανα Μετρήσεων	0	~0.0%
4	Αφυγραντήρες	4	~1.9%
5	Κινητά Τηλέφωνα	92	~1.6%
6	Κλιματιστικά	2	~1.1%
7	Κουζίνες	3	~2.8%
8	Οθόνες Υπολογιστών	0	~0.0%
9	Πλυντήρια Ρούχων	5	~2.8%
10	Τηλεοράσεις	5	~0.6%

Πίνακας 8. Πλήθος αξιολογήσεων ανά κατηγορία για rating 1



Διάγραμμα 7. Αξιολογήσεις ανά κατηγορία για rating 2 - Ραβδόγραμμα

Η συνολική απεικόνιση των αξιολογήσεων με βάση το rating φαίνεται στο **Διάγραμμα 8**.



Διάγραμμα 8. Αξιολογήσεις ανά κατηγορία με βάση το rating

Κεφάλαιο 5
Προεπεξεργασία Δεδομένων

Η προεπεξεργασία κειμένου (text preprocessing) αποτελεί ένα από τα σημαντικότερα βήματα στην εξόρυξη γνώμης και είναι απαραίτητο για την εξαγωγή χαρακτηριστικών από αυτό. Οι αξιολογήσεις γράφονται από χρήστες, αυτό σημαίνει ότι εκφράζουν τα συναισθήματα τους μέσα από το κείμενο της αξιολόγησης. Συνεπώς μπορεί να χρησιμοποιούν σημεία στίξης (όπως θαυμαστικά) και συντομογραφίες. Επιπλέον μια αξιολόγηση συμπεριλαμβάνει και άρθρα (ο, η το κτλ.) τα οποία δεν δίνουν κάποια χρήσιμη πληροφορία. Αυτά ονομάζονται stop words και θα πρέπει αν αφαιρεθούν πριν την ανάλυση των αξιολογήσεων με βάση τα ελληνικά stop-words[5]. Επειδή το dataset μας είναι σε μορφή JSON μπορεί να περιλαμβάνει και μη χαρακτήρες στα κείμενα των αξιολογήσεων, δηλαδή κάποιες πληροφορίες σε μορφή HTML, για παράδειγμα break lines (
). Συνεπώς είναι απαραίτητη διαδικασία να γίνει «καθαρισμός» των αξιολογήσεων έτσι ώστε να μείνουν μόνο οι χρήσιμες πληροφορίες.

- Παράδειγμα αρχικού κειμένου αξιολόγησης:

  «Πολύ καλό ψήσιμο, γρήγορα και οικονομικά. Πολύ καλή κατασκευή και με λειτουργίες που κάνουν το μαγείρεμα παιχνίδι.»

- Κείμενο μετά την προεπεξεργασία:

  «καλό ψήσιμο οικονομικά καλή κατασκευή λειτουργίες κάνουν μαγείρεμα παιχνίδι»

[5] https://github.com/xtsimpouris/gr-nlp-law/tree/master/Greek%20Stopwords

5.1 Λέξεις ανά Κατηγορία

Εφόσον έχει γίνει η προεπεξεργασία μετρήσαμε πόσες χρήσιμες λέξεις έμειναν για κάθε κατηγορία συνολικά. Στον **Πίνακα 9** παρατηρούμε τον αριθμό λέξεων πριν και μετά την προεπεξεργασία.

ΑΑ	Κατηγορία	Αρχικός Αριθμός Λέξεων	Λέξεις μετά την προεπεξεργασία
1	Κινητά Τηλέφωνα	1.339.949	106.908
2	Τηλεοράσεις	174.643	14.489
3	Tablets	160.059	12.712
4	Laptops	155.022	12.296
5	Αφυγραντήρες	68.624	5.383
6	Κλιματιστικά	41.033	3.195
7	Πλυντήρια Ρούχων	40.409	3.146
8	Οθόνες Υπολογιστών	36.495	2.920
9	Αθλητικά Όργανα Μετρήσεων	38.339	2.975
10	Κουζίνες	22.399	1.800

Πίνακας 9. Αριθμός λέξεων ανά κατηγορία πριν και μετά την προεπεξεργασία

Από τις παραπάνω μετρήσεις καταλαβαίνουμε τη σημαντικότητα αλλά και την αναγκαιότητα της επεξεργασίας των κειμένων των αξιολογήσεων πριν την ανάλυση τους. Για να είναι η ανάλυση μας όσο το δυνατόν πιο έγκυρη, αποφασίσαμε να ασχοληθούμε με τις κατηγορίες που περιέχουν από 10.000 λέξεις και πάνω. Συνεπώς με βάση τα αποτελέσματα επιλέξαμε τις τέσσερις σημαντικότερες κατηγορίες, δηλαδή τα **Κινητά Τηλέφωνα**, τις **Τηλεοράσεις**, τα **Tablets** και τα **Laptops**. Όπως είναι εμφανές οι υπόλοιπες κατηγορίες περιέχουν μικρό αριθμό λέξεων και θεωρήσαμε ότι η ανάλυση πάνω σε αυτές δεν θα ήταν αρκετά έγκυρη.

5.2 Δημιουργία Word Clouds

Ένα word cloud (γνωστό και ως tag cloud) είναι μια οπτική απεικόνιση δεδομένων κειμένου, συνήθως μεμονωμένων λέξεων όπου η σημαντικότητα μιας λέξης καθορίζεται με το μέγεθος του κείμενου καθώς και με το χρώμα του κειμένου. Με απλά λόγια, όσο πιο μεγάλη και πιο έντονη είναι μια λέξη μέσα στο word cloud, τόσο πιο συχνά

εμφανίζεται μέσα στο κείμενο. Ένα word cloud μας βοηθάει να δούμε ποιες λέξεις είναι πιο συχνές αλλά και πιο σημαντικές.

Εικόνα 6. Παράδειγμα word cloud από την Wikipedia

Στη δική μας περίπτωση, τα word clouds θα μας βοηθήσουν να εντοπίσουμε τη συχνότητα των λέξεων στα κείμενα των αξιολογήσεων κι επίσης να δούμε αν υπάρχουν μεταξύ των κατηγοριών αλλά και των ratings κάθε κατηγορίας διαφορές στις λέξεις αυτές ή αν κάποιες συγκεκριμένες λέξεις επαναλαμβάνονται.

Εφόσον λοιπόν «καθαρίσαμε» τα δεδομένα μας δημιουργήσαμε ένα word cloud για κάθε κατηγορία και για κάθε rating. Για να δημιουργήσουμε word clouds πάνω στα δικά μας δεδομένα χρησιμοποιήσαμε το online εργαλείο https://wordart.com, λαμβάνοντας υπόψη μόνο τις χρήσιμες λέξεις μετά την προεπεξεργασία. Επίσης να σημειώσουμε ότι αφαιρέσαμε πρώτα τους τόνους από τις λέξεις, έτσι ώστε μια λέξη που υπήρχε στο αρχικό κείμενο και με τις δύο μορφές (πχ κινητό, κινητο), να υπάρχει τελικά σε μία μορφή.

5.2.1 Word Clouds ανά Κατηγορία

Σε αυτήν την υποενότητα παρουσιάζονται τα word clouds που δημιουργήσαμε για τις τέσσερις κατηγορίες που επιλέξαμε. Για κάθε κατηγορία δημιουργήσαμε ένα word cloud για όλες τις αξιολογήσεις συνολικά καθώς και ένα word cloud για κάθε rating, συνολικά

6 word clouds ανά κατηγορία. Η οπτική αυτή απεικόνιση θα μας βοηθήσει αφενός να εντοπίσουμε ομοιότητες και διαφορές ανάμεσα στα ratings της εκάστοτε κατηγορίας και αφετέρου μεταξύ των τεσσάρων κατηγοριών.

Κινητά Τηλέφωνα

Όπως βλέπουμε παρακάτω, συναντάμε κοινές λέξεις μεταξύ των ratings, κάποιες από αυτές είναι *κινητό, κάμερα, καλή, μπαταρία, οθόνη, γρήγορο* κτλ. Φαίνεται ότι οι χρήστες χρησιμοποιούν τις λέξεις αυτές αρκετά συχνά στις αξιολογήσεις τους. Η λέξη «κινητό» έχει τη μεγαλύτερη συχνότητα από όλες τις λέξεις των αξιολογήσεων, όμως δεν εκφράζει κάποιο συναίσθημα ή γνώμη.

Όλες οι αξιολογήσεις | Rating 5

Εικόνα 7. Word clouds για όλες τις αξιολογήσεις και για rating 5 – Κινητά Τηλέφωνα

Rating 4 | Rating 3

Εικόνα 8. Word clouds για rating 4 και rating 3 – Κινητά Τηλέφωνα

Εικόνα 9. Word clouds για rating 2 και rating 1 – Κινητά Τηλέφωνα

Τηλεοράσεις

Στην κατηγορία τηλεοράσεις παρατηρούμε αρκετές ομοιότητες στα ratings 3, 4 και 5. Συγκεκριμένα κάποιες κοινές λέξεις μεταξύ τους είναι *καλή, τηλεόραση, εικόνα, τιμή, χρώματα* κτλ. Η λέξη «καλή» έχει τη μεγαλύτερη συχνότητα και εκφράζει ένα θετικό συναίσθημα.

Εικόνα 10. Word clouds για όλες τις αξιολογήσεις και για rating 5 – Τηλεοράσεις

Εικόνα 11. Word clouds για rating 4 και rating 3 – Τηλεοράσεις

Στα ratings 1 και 2 όπως φαίνεται και στην **Εικόνα 12**, οι λέξεις που εμφανίζονται πιο συχνά, διαφέρουν από τα υπόλοιπα ratings. Συγκεκριμένα παρατηρούμε λέξεις όπως *πρόβλημα, service, δυστυχώς,* οι οποίες ίσως να δηλώνουν ότι οι αξιολογήσεις εκφράζουν αρνητικά συναισθήματα.

Εικόνα 12. Word clouds για rating 2 και rating 1 – Τηλεοράσεις

Tablets

Στην κατηγορία Tablets παρατηρούμε ανάμεσα στα ratings 3, 4 και 5 κοινές λέξεις όπως *καλό, καλή, οθόνη, γρήγορο, μπαταρία* κτλ. Οι λέξεις «καλό» και «καλή» έχουν χρησιμοποιηθεί περισσότερο από τους χρήστες και δηλώνουν θετικά συναισθήματα.

Εικόνα 13. Word clouds για όλες τις αξιολογήσεις και για rating 5 - Tablets

Εικόνα 14. Word clouds για rating 4 και rating 3 – Tablets

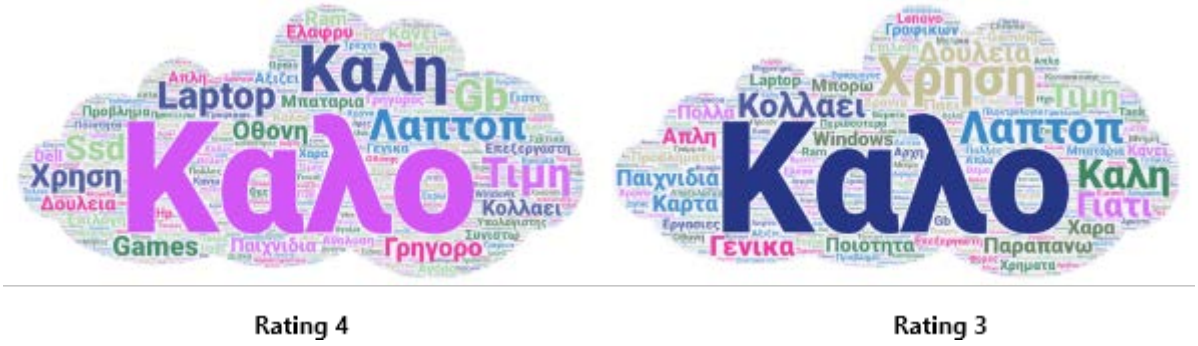
Αντίθετα στα ratings 1 και 2 είναι διαφορετικές οι λέξεις που παρατηρούμε, όπως *χάλασε, εγγύηση, τάμπλετ, κολλάει.*

Εικόνα 15. Word clouds για rating 2 και rating 1 – Tablets

Laptops

Στην κατηγορία αυτή παρατηρούμε ότι οι λέξεις *καλό, χρήση, γρήγορο, τιμή* κτλ. έχουν μεγάλη συχνότητα κι επίσης τις συναντάμε στα ratings 2, 3, 4 και 5.

Εικόνα 16. Word clouds για όλες τις αξιολογήσεις και για rating 5 - Laptops



Εικόνα 17. Word clouds για rating 4 και rating 3 – Laptops

Όσον αφορά το rating 1, παρατηρούμε λέξεις όπως pixel, καμένα, αγορά κτλ. οι οποίες δηλώνουν ότι οι χρήστες έχουν κάνει αρνητικά σχόλια και αυτό επιβεβαιώνεται από το χαμηλό rating.

Εικόνα 18. Word clouds για rating 2 και rating 1 - Laptops

Συγκρίνοντας τα παραπάνω word clouds όλων των κατηγοριών παρατηρούμε ότι κάποιες λέξεις με μεγάλη συχνότητα όπως *καλό, καλή, οθόνη, τιμή* κτλ., είναι κοινές μεταξύ των κατηγοριών.

Κεφάλαιο 6
Ανάλυση Δεδομένων

Στο παρόν κεφάλαιο παρουσιάζεται η ανάλυση των δεδομένων των αξιολογήσεων για τις κατηγορίες που έχουμε επιλέξει μέσω λεξικογραφικής ανάλυσης.

6.1 Λεξικογραφική Ανάλυση

Εφόσον έχει προηγηθεί η προεπεξεργασία των δεδομένων θα προχωρήσουμε σε λεξικογραφική ανάλυση (lexicon based sentiment analysis) με χρήση συναισθηματικού λεξικού, για να υπολογίσουμε 8 διαφορετικά scores τα οποία θα εξηγήσουμε παρακάτω.

6.1.2 GrAFS Lexicon

Το λεξικό που χρησιμοποιήσαμε για την ανάλυση μας είναι το **Greek Affect and Sentiment Lexicon (GrAFS)**[6], το οποίο έχει δημιουργηθεί από τους Tsakalidis et al., και παρουσιάζεται στο άρθρο **Building and evaluating resources for sentiment analysis in the Greek language** (2018). Το άρθρο αυτό είναι αρκετά πρόσφατο και είχε 7 αναφορές στο Google Scholar την περίοδο που το δουλέψαμε. Η δημιουργία του λεξικού αυτού έχει βασιστεί πάνω στην ηλεκτρονική έκδοση του λεξικού του Τριανταφυλλίδη[7] (1998), ένα από τα μεγαλύτερα λεξικά της ελληνικής γλώσσας. Για τη συγκέντρωση των λέξεων χρησιμοποιήθηκαν εργαλεία crawling από τους Tsakalidis et al., και έγινε αναζήτηση στο λεξικό του Τριανταφυλλίδη για λέξεις με πιθανό συναισθηματικό φορτίο, οι οποίες εμπεριείχαν ειρωνικό, υποτιμητικό, υβριστικό, κωμικό ή χυδαίο τόνο.

[6] https://mklab.iti.gr/resources/tsakalidis2017building.zip
[7] http://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/index.html

Επιπλέον, πραγματοποιήθηκε αναζήτηση και στις περιγραφές των λέξεων για συναισθηματικές λέξεις, συγκεκριμένα για τις λέξεις *συναίσθημα, αισθάνομαι, αίσθηση, αίσθημα, συναίσθηση, αισθάνεται* και *νιώθω*.

Η παραπάνω διαδικασία είχε ως αποτέλεσμα να συγκεντρωθούν 2324 λέξεις. Οι λέξεις αυτές σχολιάστηκαν από 4 σχολιαστές, 2 από τον κλάδο της πληροφορικής και 2 από τον κλάδο της γλωσσολογίας. Από κάθε σχολιαστή ζητήθηκε για κάθε λέξη να χαρακτηρίσει την **υποκειμενικότητα** (subjectivity), την **πόλωση** (polarity), δηλαδή αν το συναίσθημα είναι θετικό, αρνητικό ή ουδέτερο, καθώς και κάθε ένα από τα έξι βασικά συναισθήματα (affects) σύμφωνα με τον Ekman (1992), τα οποία είναι ο **θυμός** (anger), η **απέχθεια** (disgust), ο **φόβος** (fear), η **χαρά** (happiness), η **λύπη** (sadness) και η **έκπληξη** (surprise). Η υποκειμενικότητα έρεπε να χαρακτηριστεί ως *αντικειμενική,* ή έντονα ή ασθενώς *υποκειμενική*. Αν μια λέξη ήταν υποκειμενική έπρεπε να ορίσουν την πολικότητα της με βάση τις επιλογές *θετική*, *αρνητική* ή *και τα δύο,* καθώς και να βαθμολογήσουν το κάθε συναίσθημα από 1 (δεν υπάρχει καθόλου) έως 5 (υπάρχει σε μεγάλο βαθμό).

Εν συνεχεία, τα scores την υποκειμενικότητας μετατράπηκαν σε τρεις τιμές, 0 για αντικειμενικό, 0.5 ασθενώς υποκειμενικό και 1 για έντονα υποκειμενικό. Έγινε κανονικοποίηση στα scores των έξι συναισθημάτων έτσι ώστε να έχουν τιμές του εύρους 0-1. Υπολογίστηκε ο μέσος όρος για την πόλωση καθώς και για κάθε score συναισθήματος που όρισαν οι σχολιαστές. Τελικό στάδιο ήταν να παραχθούν όλες οι κλίσεις των λέξεων με τη χρήση εργαλείων επεξεργασίας φυσικής γλώσσας. Η τελική έκδοση του λεξικού περιέχει **32.884** λέξεις. Παρακάτω φαίνεται ένα παράδειγμα με τα scores των λέξεων από το λεξικό.

keyword	subj	positive	negative	anger	disgust	fear	happy	sad	surprise
καλό	0.75	1.0	0.0	0.0	0.0	0.0	0.6875	0.0	0.375
καλή	0.5	0.75	0.0	0.0	0.0	0.0	0.5625	0.0	0.3125
λαχταρώ	0.875	0.75	0.5	0.0	0.0	0.4375	0.5	0.0	0.3125
έπαιζα	0.375	0.5	0.25	0.125	0.0	0.0	0.3125	0.0	0.125

Πίνακας 10. Δείγμα λέξεων από το GrAFS Lexicon

6.1.3 Εφαρμογή του Λεξικού

Για να βρούμε εάν μια λέξη έχει θετικό, αρνητικό ή ουδέτερο συναίσθημα αφαιρούμε από το positive score το negative score. Εάν το αποτέλεσμα είναι αρνητικό , τότε το συναίσθημα είναι αρνητικό, αν είναι θετικό το συναίσθημα είναι κι αυτό θετικό, και αν είναι μηδέν τότε το συναίσθημα είναι ουδέτερο. Για να υπολογίσουμε το κάθε score, για κάθε λέξη της εκάστοτε αξιολόγησης που είχε αντιστοιχία στο λεξικό, βρήκαμε τα 8 scores. Έπειτα υπολογίσαμε το average score για κάθε αξιολόγηση με βάση μόνο τις λέξεις αυτές. Όσες λέξεις δεν βρέθηκαν στο λεξικό αγνοήθηκαν και δεν υπολογίστηκαν στην ανάλυση μας.

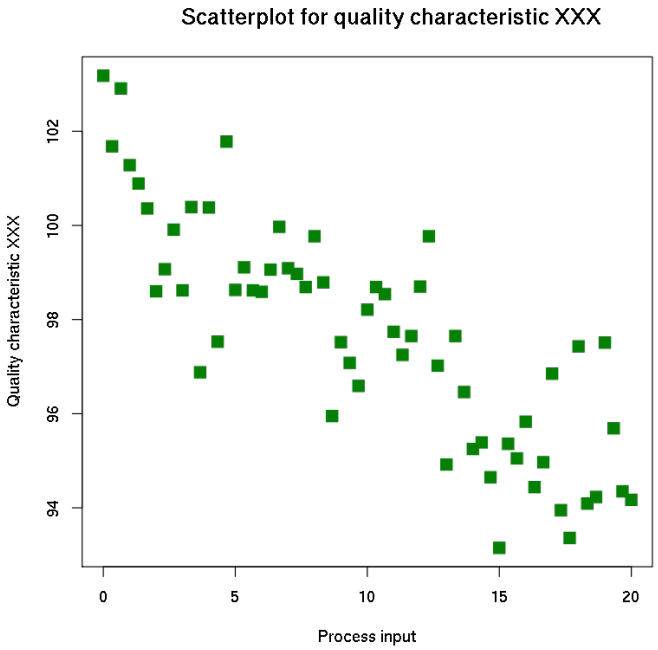
Ένα ζήτημα που προκύπτει στην ανάλυση μας είναι ότι δεν είναι βέβαιο ότι έχουν όλες οι λέξεις των αξιολογήσεων τόνους. Για να αποφύγουμε το πρόβλημα αυτό, αφαιρέσαμε από όλες τις λέξεις του λεξικού τους τόνους καθώς και από όλες τις αξιολογήσεις. Κατά αυτόν τον τρόπο είμαστε βέβαιοι ότι θα γίνει σωστά η αντιστοιχία των λέξεων.

6.2 Τάση Συναισθήματος και Affects με Χρήση Plots

Εφόσον έχουμε βρει τα scores για κάθε αξιολόγηση θα χρησιμοποιήσουμε scatter plots και box plots για να πάρουμε μια οπτική απεικόνιση των δεδομένων. Με τον τρόπο αυτό θα προσπαθήσουμε να καταλάβουμε το συναίσθημα που εκφράζεται (polarity) για κάθε κατηγορία και για κάθε rating. Επιπρόσθετα, θα προσπαθήσουμε να εντοπίσουμε την τάση των affect scores. Στον άξονα x έχουμε το rating κάθε αξιολόγησης και στον y έχουμε το average score, για κάθε ένα από τα παραπάνω scores.

6.2.1 Scatter Plots

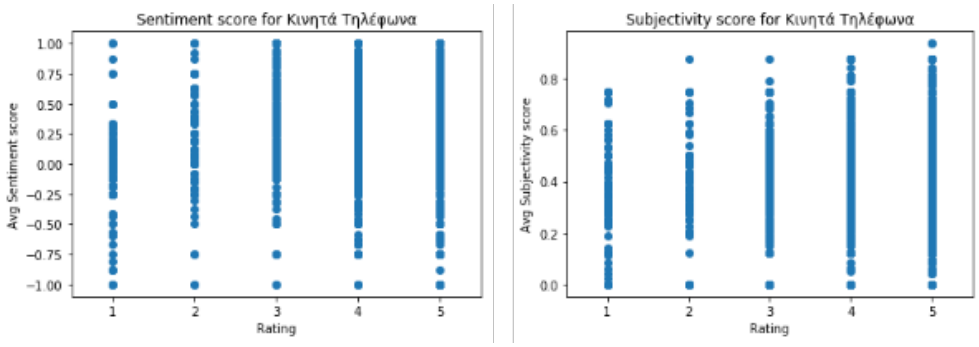
Το scatter plot (γράφημα διασποράς) είναι ένας τύπος γραφήματος ή μαθηματικού διαγράμματος το οποίο χρησιμοποιεί καρτεσιανές συντεταγμένες για την εμφάνιση τιμών, συνήθως για δύο μεταβλητές για ένα σύνολο δεδομένων. Απεικονίζει ζευγάρια τιμών x-y και στόχος του είναι να δείξει κάποια σχέση μεταξύ των μεταβλητών. Στην περίπτωση μας θέλουμε να δούμε τη σχέση που έχει το κάθε rating με το κάθε score, αν για παράδειγμα, όσο ανεβαίνει το rating αυξάνεται ή μειώνεται το sentiment score, το happy score κτλ.



Εικόνα 19. Παράδειγμα scatter plot από τη Wikipedia

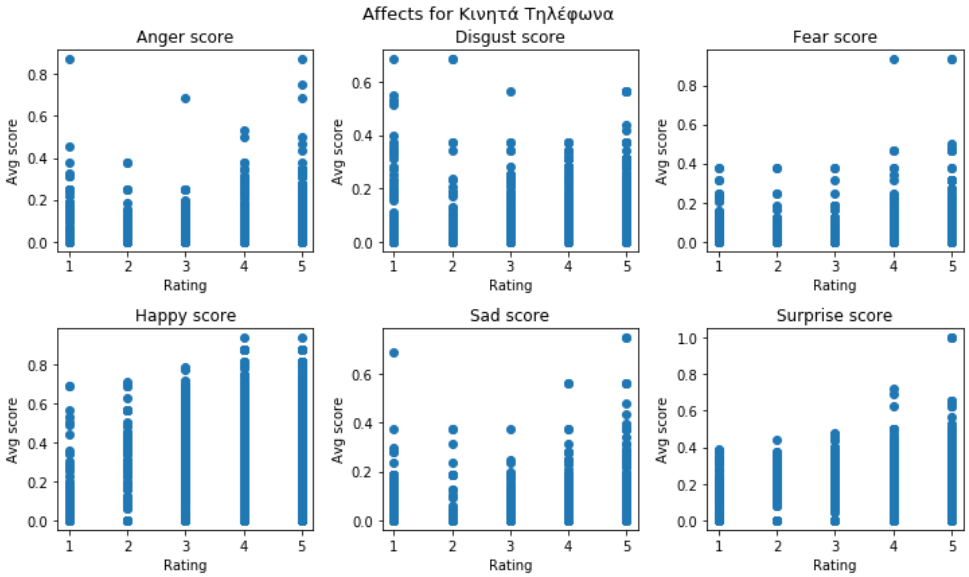
Κινητά Τηλέφωνα

Στα κινητά τηλέφωνα για το sentiment παρατηρούμε ότι για το rating 1 οι περισσότερες τιμές είναι από -0.25 έως 0.25, πράγμα που φαίνεται λογικό για το βαθμό του rating. Για το rating 2 οι τιμές είναι διάσπαρτες, ενώ για τα υπόλοιπα ratings αν και υπάρχουν και τιμές στο -1, παρατηρούμε κυρίως υψηλές τιμές. Το subjectivity δείχνει μια τάση να ανεβαίνει καθώς αυξάνεται το rating.



Εικόνα 20. Sentiment and Subjectivity scores – Κινητά Τηλέφωνα

Παρακάτω παρατηρούμε τα affect scores. Το anger έχει σχετικά χαμηλές τιμές που φτάνουν περίπου στο 0.4 ενώ υπάρχουν και κάποιες τιμές μακριά από το σύνολο (outliers). Στο disgust και στο fear παρατηρούμε επίσης παρόμοιες μετρήσεις με το score να φτάνει κοντά στο 0.4 για όλα τα ratings. Στο γράφημα του happy score φαίνεται μια τάση να ανεβαίνει το score καθώς ανεβαίνει και το rating. Η πρώτη εντύπωση που μας δίνεται είναι ότι εκφράζεται χαρά στα υψηλά ratings. Εν, συνεχεία το sad score φτάνει κοντά στο 0.4 ενώ παρατηρούμε και αρκετές τιμές στο rating 5 πράγματα που μας κάνει εντύπωση. Τέλος στο surprise score δε υπάρχει μεγάλη διαφορά στη διακύμανση του ανάμεσα στα ratings. Να σημειώσουμε εδώ ότι το surprise μπορεί να εκφράζει κα αρνητικό αλλά και θετικό affect.

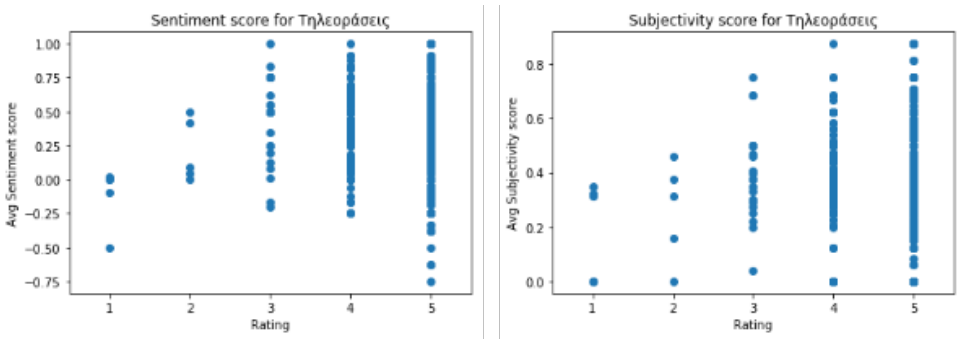


Εικόνα 21. Affect scores – Κινητά Τηλέφωνα

Στην κατηγορία αυτή λόγω του μεγάλου αριθμού αξιολογήσεων είναι δύσκολο να καταλάβουμε ακριβώς που κυμαίνονται οι τιμές. Θα δούμε παρακάτω με τη χρήση των boxplots πιο καθαρές μετρήσεις.

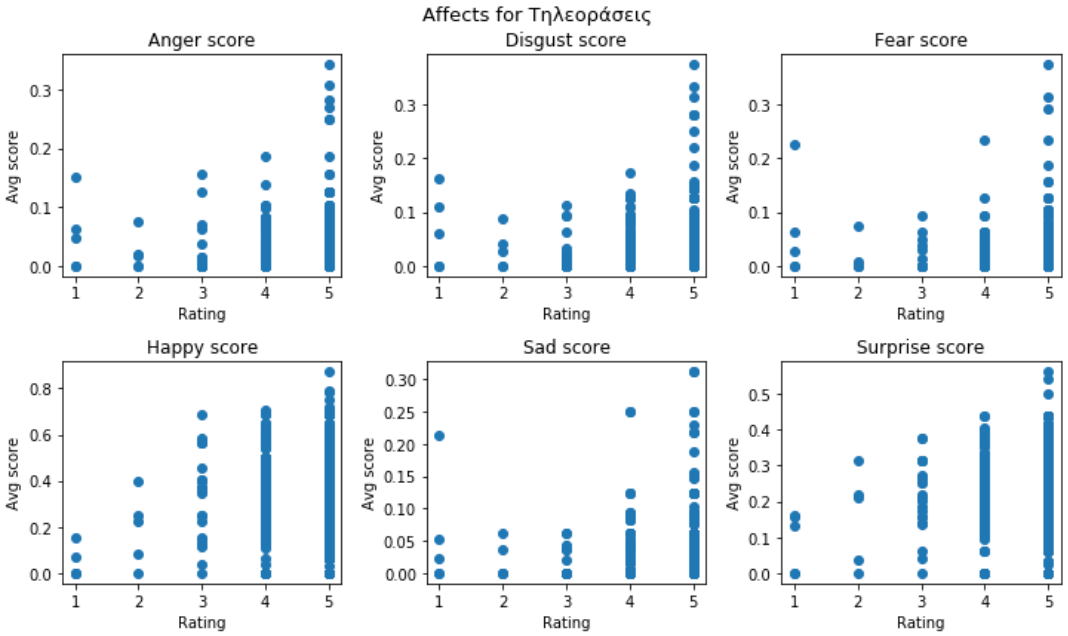
Τηλεοράσεις

Στις τηλεοράσεις όπως φαίνεται παρακάτω, παρατηρούμε γενικά θετικό score του sentiment, αν και οι περισσότερες αξιολογήσεις αφορούν τα ratings 4 και 5. Το subjectivity δείχνει κι αυτό μια τάση να αυξάνεται καθώς αυξάνεται το rating.



Εικόνα 22. Sentiment and Subjectivity scores – Τηλεοράσεις

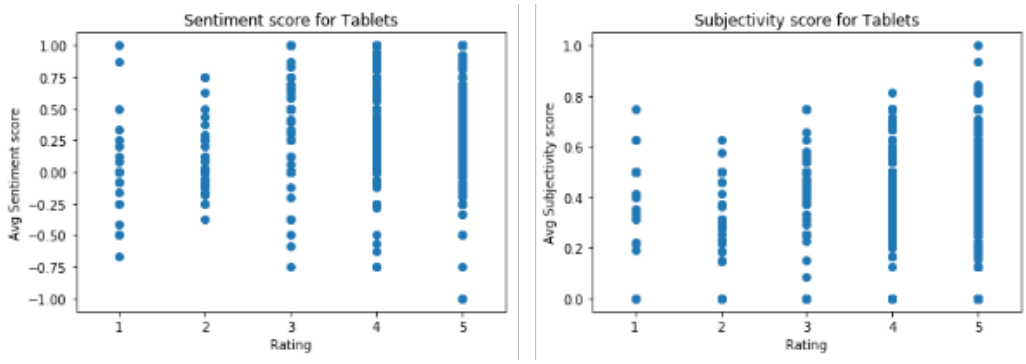
Είναι εντυπωσιακό ότι στο rating 5 το anger score έχει αρκετά υψηλές τιμές καθώς επίσης και το disgust αλλά και το fear, αν και μοιάζουν να είναι outliers. Το happy score φαίνεται κι εδώ ότι αυξάνεται καθώς αυξάνεται το rating. Το sad score φαίνεται να έχει χαμηλές τιμές γενικά ενώ το surprise φαίνεται να αυξάνεται στα υψηλά ratings.



Εικόνα 23. Affect scores – Τηλεοράσεις

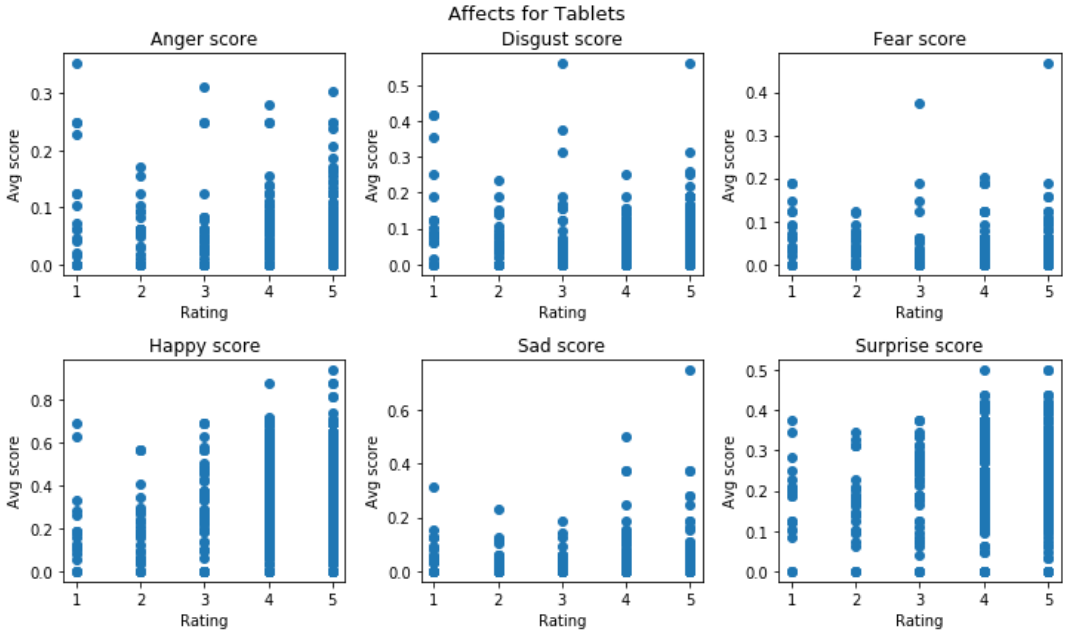
Tablets

Στα tablets στο γράφημα του sentiment score παρατηρούμε ότι στο rating 1 δε συγκεντρώνονται οι τιμές σε κάποιο συγκεκριμένο διάστημα στον άξονα y. Παρομοίως και στα ratings 2 και 3, ενώ στα 4 και 5 παρατηρούμε υψηλά scores γενικά. Στο subjectivity φαίνεται οι τιμές να είναι γύρω από το 0.5 για τα ratings 3, 4 και 5.



Εικόνα 24. Sentiment and Subjectivity scores - Tablets

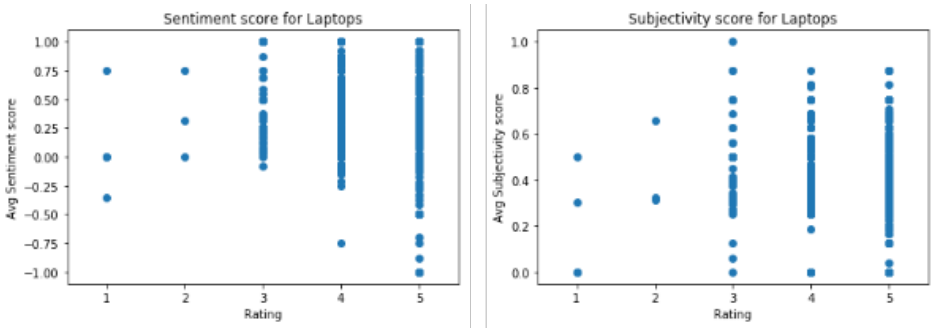
Στα γραφήματα των anger, disgust, fear και sad παρατηρούμε γενικά χαμηλές τιμές σε όλα τα ratings. Το happy score και το surprise δείχνουν να ανεβαίνουν στα υψηλά ratings.



Εικόνα 25. Affect scores – Tablets

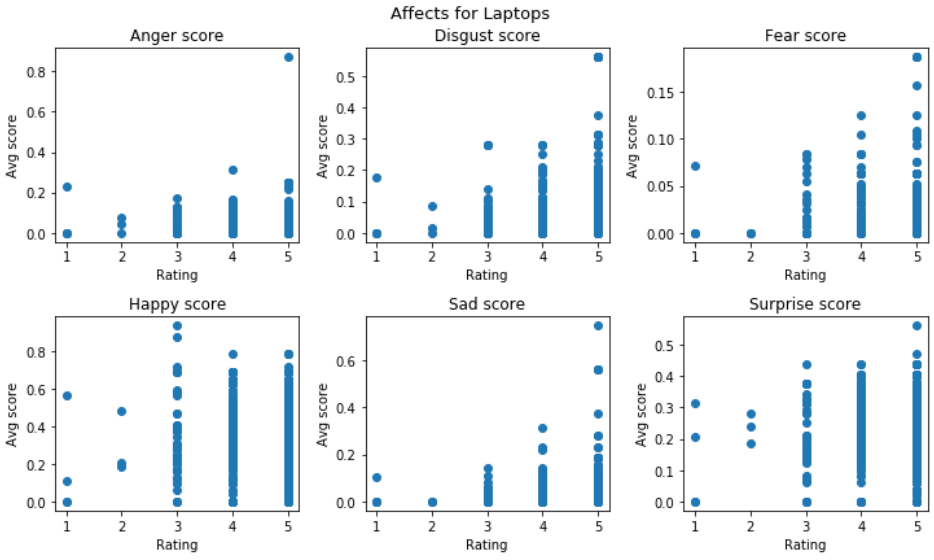
Laptops

Στα laptops λόγω του μικρού αριθμού αξιολογήσεων παρατηρούμε ελάχιστα σημεία πάνω στα γραφήματα για τα ratings 1 και 2. Για τα ratings 3, 4 και 5 βλέπουμε ότι το sentiment score κυμαίνεται περίπου από -0.5 έως 1. Στο subjectivity φαίνεται οι τιμές να είναι γύρω από το 0.5.



Εικόνα 26. Sentiment and Subjectivity scores – Laptops

Όσον αφορά τα affects, παρατηρούμε το anger score, το fear καθώς επίσης και το sad να έχουν αρκετά χαμηλές τιμές. Το disgust δείχνει μια τάση να ανεβαίνει καθώς ανεβαίνει το rating αλλά οι περισσότερες τιμές δεν ξεπερνούν το 0.3. Επίσης το happy score δείχνει να ανεβαίνει και τέλος το surprise για τα ratings 3, 4 και 5 να κυμαίνεται από 0 έως 0.5.

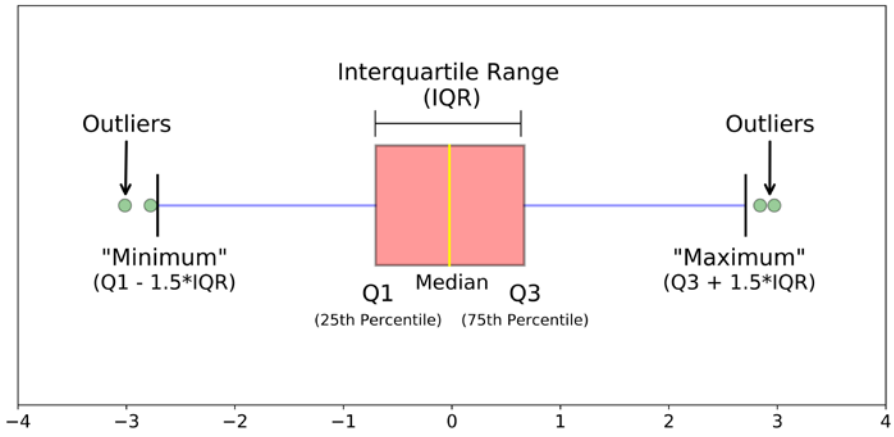


Εικόνα 27. Affect scores - Laptops

Τα παραπάνω scatter plots μας δίνουν μια πρώτη εικόνα για το που κυμαίνονται τα scores, δεν είναι όμως εύκολο να καταλάβουμε που συγκεντρώνονται οι περισσότερες τιμές στον άξονα y. Γι' αυτό το λόγο χρησιμοποιήσαμε και τα box plots τα οποία θα δούμε παρακάτω.

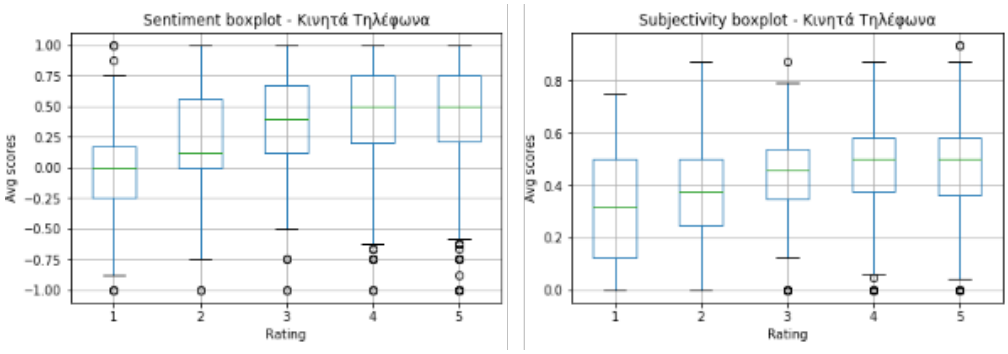
6.2.2 Box Plots

Το box plot είναι ένας τυποποιημένος τρόπος εμφάνισης ενός dataset βάσει μιας περίληψης πέντε αριθμών: το ελάχιστο, το μέγιστο, το διάμεσο και το πρώτο και τρίτο τεταρτημόριο. Ελάχιστο είναι το χαμηλότερο σημείο δεδομένων εξαιρουμένων τυχόν ακραίων τιμών (outliers), μέγιστο είναι το μεγαλύτερο σημείο δεδομένων εξαιρουμένων τυχόν ακραίων τιμών, διάμεσος (Q2 / 50th Percentile) είναι η μέση τιμή (median) του συνόλου δεδομένων, πρώτο τεταρτημόριο (Q1 / 25th Percentile) είναι η διάμεση τιμή του κάτω μισού του συνόλου δεδομένων και τρίτο τεταρτημόριο (Q3 / 75th Percentile), είναι η διάμεση τιμή του άνω μισού του συνόλου δεδομένων. Στην **Εικόνα 28** βλέπουμε τα 5 αυτά μέρη.



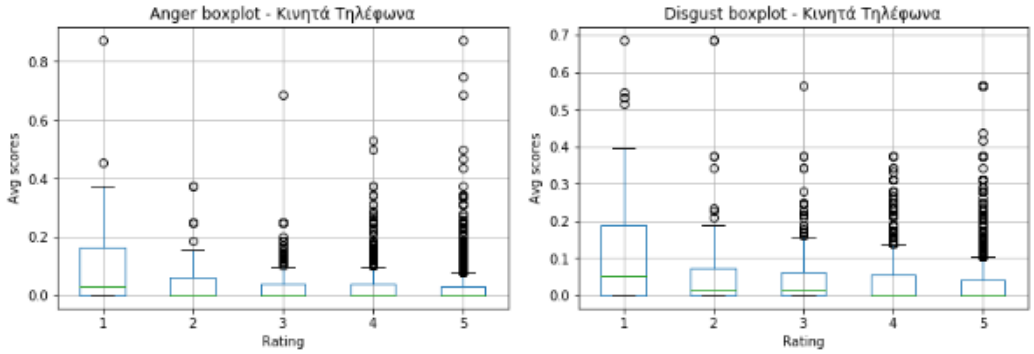
Εικόνα 28. Τα μέρη ενός box plot.

Κινητά Τηλέφωνα

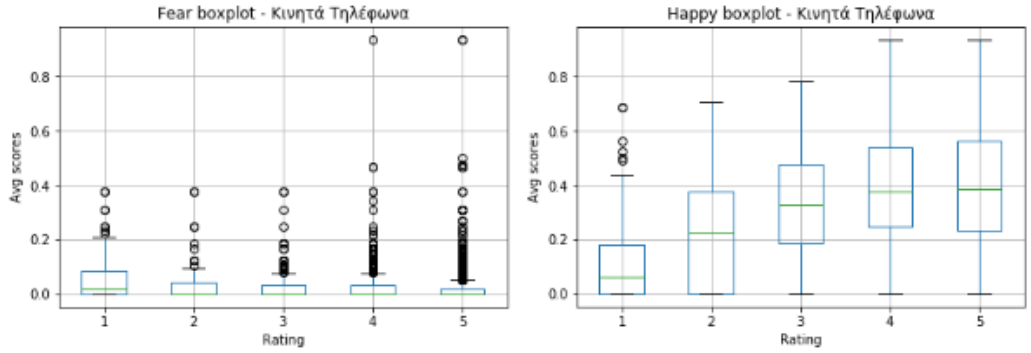


Εικόνα 29. Sentiment and Subjectivity Box Plots – Κινητά Τηλέφωνα

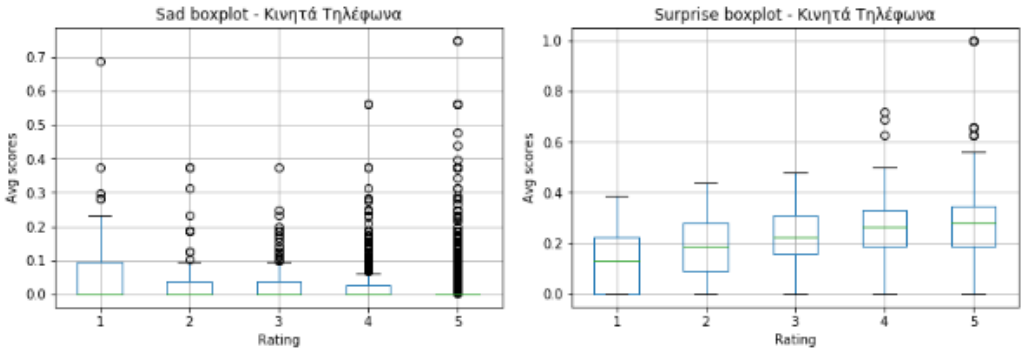
Για το sentiment παρατηρούμε ότι όσο ανεβαίνουν τα αστέρια, ανεβαίνει και η μέση τιμή του average sentiment score. Ενώ η max τιμή για το average score είναι η ίδια για τα αστέρια 2-5. Για το subjectivity παρατηρούμε ότι όσο ανεβαίνουν τα αστέρια ανεβαίνει η μέση τιμή του average subjectivity score.



Εικόνα 30. Anger and Disgust Box Plots – Κινητά Τηλέφωνα



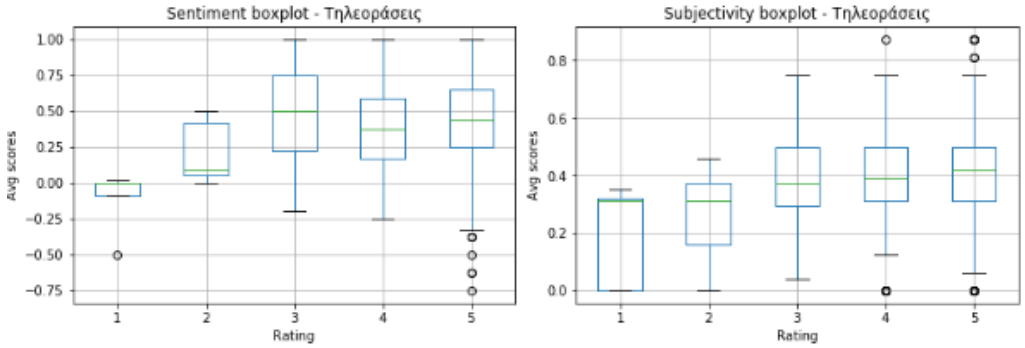
Εικόνα 31. Fear and Happy Box Plots – Κινητά Τηλέφωνα



Εικόνα 32. Sad and Surprise Box Plots – Κινητά Τηλέφωνα

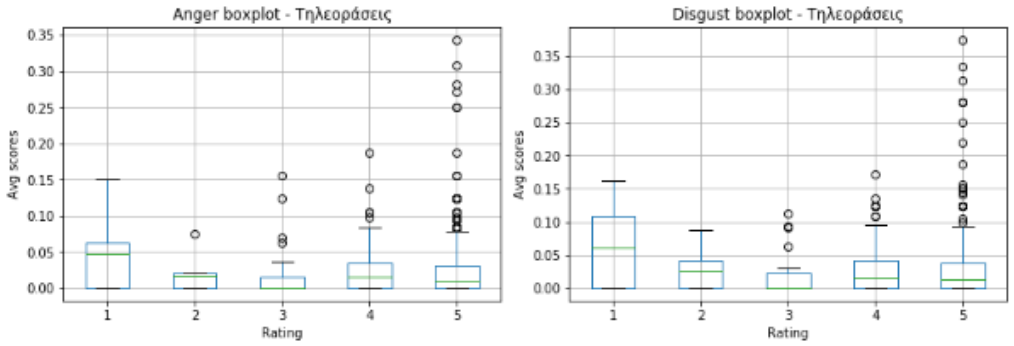
Σχετικά με τα average scores των Anger, Disgust και Fear παρατηρούμε ότι όσο ανεβαίνουν τα αστέρια μικραίνει η μέση τιμή των average scores φτάνοντας κοντά στο 0. Στο Happy box plot παρατηρούμε ότι, όσο ανεβαίνουν τα αστέρια ανεβαίνει η μέση τιμή του average happy score καθώς αυξάνεται και η max τιμή. Στο sad παρατηρούμε ότι Όσο ανεβαίνουν τα αστέρια μειώνεται η μέση τιμή του average sad score καθώς και η max τιμή.

Τηλεοράσεις

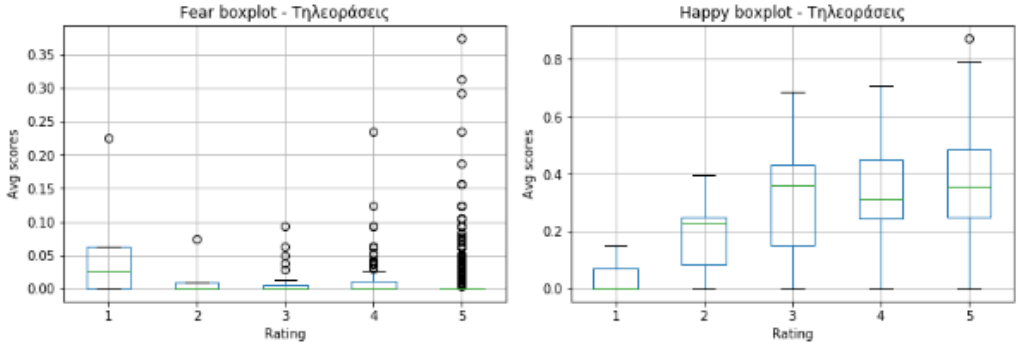


Εικόνα 33. Sentiment and Subjectivity Box Plots – Τηλεοράσεις

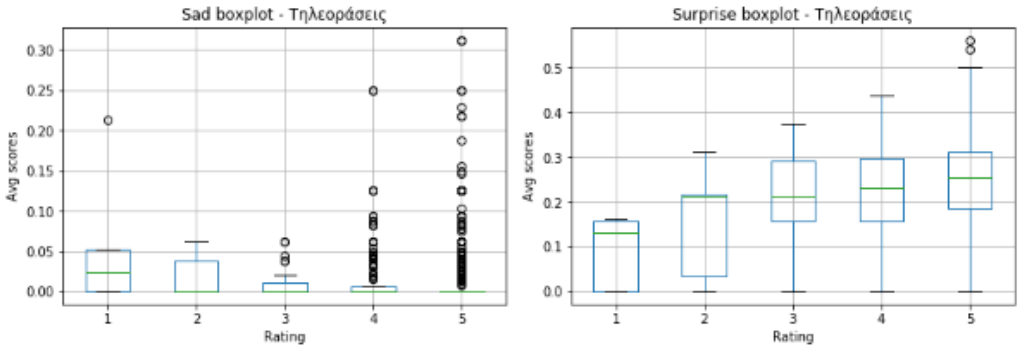
Στο Sentiment φαίνεται ότι στα 3 αστέρια έχουμε τη μεγαλύτερη μέση τιμή του average sentiment score ενώ η max τιμή είναι ίδια για τα αστέρια 3-5. Ενώ στο Subjectivity όσο ανεβαίνουν τα αστέρια ανεβαίνει η μέση τιμή του average subjectivity score ενώ η max τιμή είναι ίδια για τα αστέρια 3-5.



Εικόνα 34. Anger and Disgust Box Plots – Τηλεοράσεις



Εικόνα 35. Fear and Happy Box Plots - Τηλεοράσεις

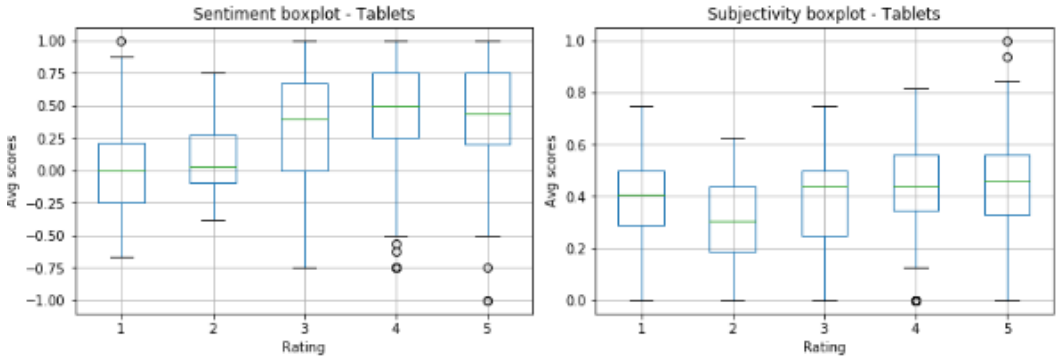


Εικόνα 36. Sad and Surprise Box Plots – Τηλεοράσεις

Παρατηρούμε ότι στο Anger, στο Disgust και στο Fear η μεγαλύτερη μέση τιμή βρίσκεται στο αστέρι 1. Για το Happy score παρατηρούμε ότι όσο μεγαλώνουν τα αστέρια μεγαλώνει η max τιμή ενώ η min τιμή παραμένει σταθερή. Στο Sad βλέπουμε ότι όσο μεγαλώνουν τα αστέρια η μέση τιμή μικραίνει φτάνοντας κοντά στο 0 ενώ ή min τιμή παραμένει σταθερή. Τέλος για το Surprise βλέπουμε ότι όσο μεγαλώνουν τα

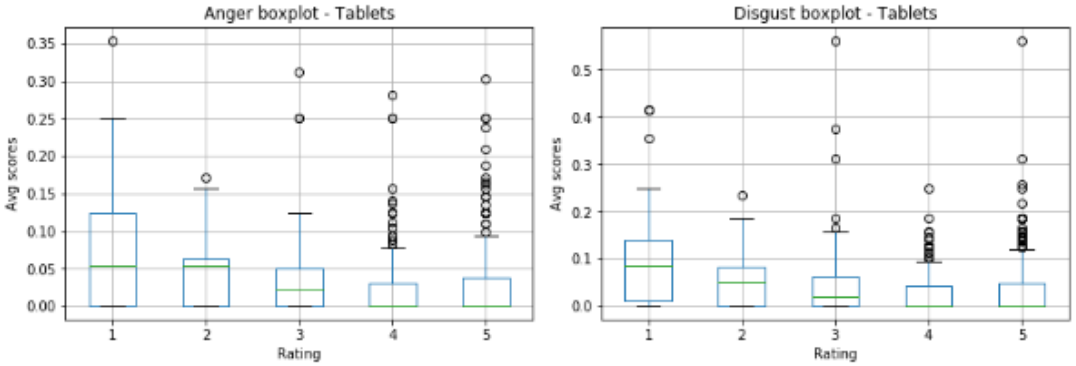
αστέρια μεγαλώνει η μέση τιμή του average score καθώς και η max τιμή, ενώ ή min τιμή παραμένει σταθερή

Tablets

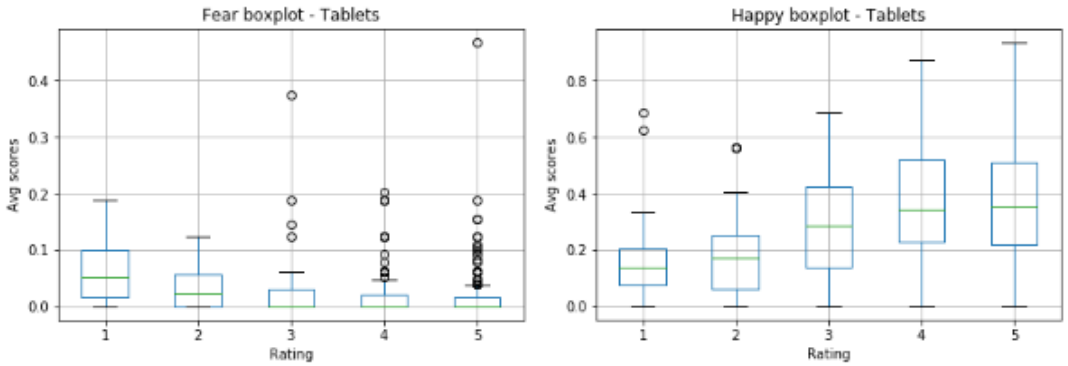


Εικόνα 37. Sentiment and Subjectivity Box Plots - Tablets

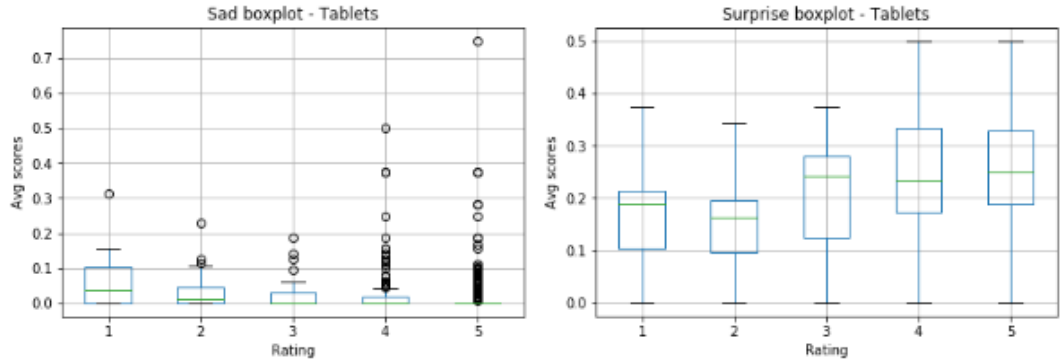
Στο sentiment score βλέπουμε ότι η μεγαλύτερη μέση τιμή είναι στο 4ο αστέρι ενώ τα αστέρια 3-5 έχουν ίδια max τιμή. Ενώ στο Surprise παρατηρούμε μικρές αποκλίσεις στη μέση τιμή (κοντά στο 4) για τα αστέρια 1,3,4,5 ενώ τα αστέρια 1,2,3,5 έχουν ίδια min τιμή.



Εικόνα 38. Anger and Disgust Box Plots – Tablets



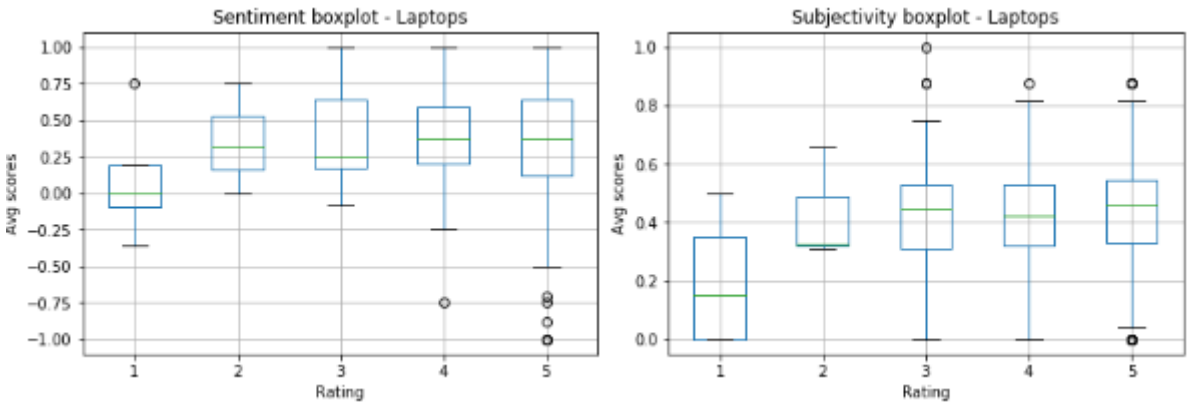
Εικόνα 39. Fear and Happy Box Plots – Tablets



Εικόνα 40. Sad and Surprise Box Plots – Tablets

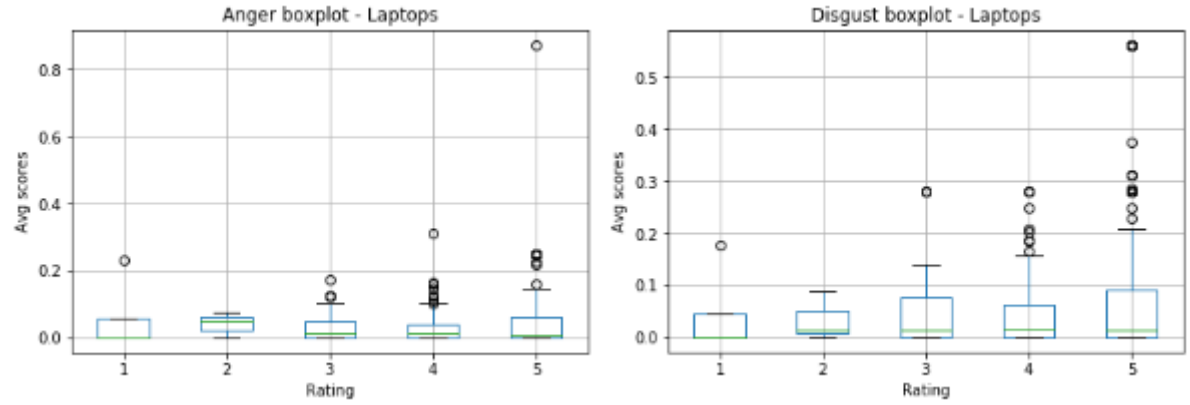
Εδώ παρατηρούμε ότι όσο μεγαλώνουν τα αστέρια η μέση τιμή για τα average scores Anger, Disgust και Fear μικραίνει φτάνοντας κοντά στο 0. Για το Happy βλέπουμε ότι όσο μεγαλώνουν τα αστέρια αυξάνεται η μέση τιμή καθώς και η max τιμή, ενώ η min τιμή μένει σταθερή. Για το Sad score όσο μεγαλώνουν τα αστέρια μικραίνει η μέση τιμή, ενώ στο Surprise έχουμε μικρές αποκλίσεις στη μέση τιμή του average score για τα αστέρια 3-5.

Laptops

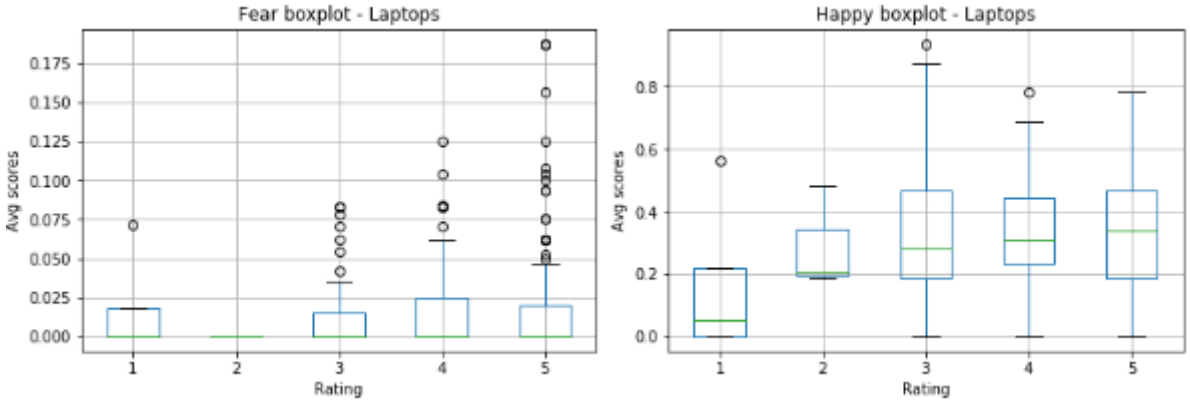


Εικόνα 41. Sentiment and Subjectivity Box Plots – Laptops

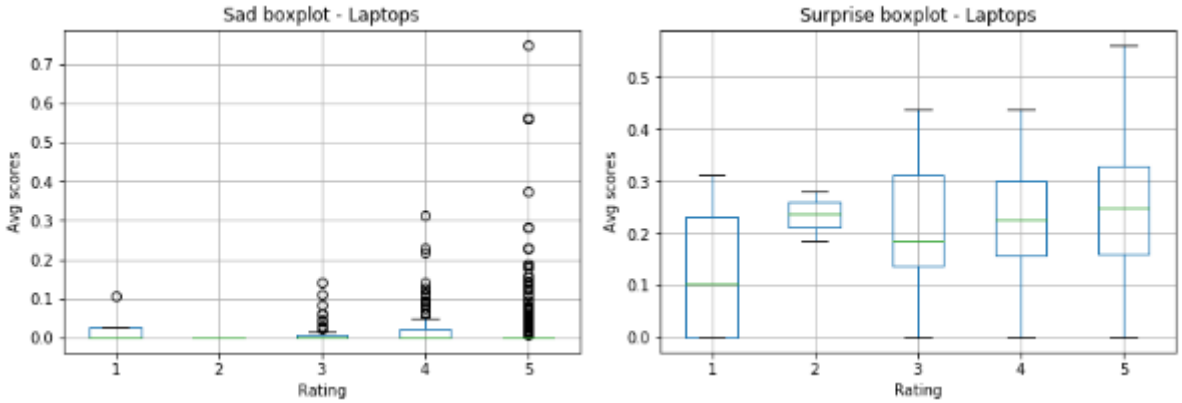
Εδώ όσον αφορά Sentiment, παρατηρούμε σταθερή τιμή για τα αστέρια 3-5 ενώ στο Subjectivity όσο μεγαλώνουν τα αστέρια αυξάνεται και η max τιμή.



Εικόνα 42. Anger and Disgust Box Plots - Laptops



Εικόνα 43. Fear and Happy Box Plots - Laptops



Εικόνα 44. Sad and Surprise Box Plots - Laptops

Για τα scores Anger, Disgust, Fear και Sad παρατηρούμε τη μέση τιμή να βρίσκεται κοντά στο 0. Για το Happy όσο μεγαλώνουν τα αστέρια η μέση τιμή αυξάνεται, ενώ στο Surprise η μεγαλύτερη μέση και max τιμή για το average surprise score φαίνεται να είναι στο αστέρι 5.

Γενικά συμπεράσματα

Με βάση τα παραπάνω γραφήματα, δείξαμε ότι όσο ανεβαίνει το rating, αυξάνεται το score του συναισθήματος (sentiment), αλλά και της χαράς (happy) πράγμα που είναι λογικό καθώς φαίνεται οι χρήστες να εκφράζουν θετικές γνώμες στις αξιολογήσεις τους. Από την άλλη μεριά παρατηρούμε τα αρνητικά συναισθήματα, δηλαδή το φόβο (fear), τη λύπη (sad), την απέχθεια (disgust) και το θυμό (anger) να έχουν υψηλές τιμές σε μικρό rating. Φαίνεται να είναι κι αυτό λογικό καθώς εκφράζουν αρνητικές γνώμες. Τέλος το affect score της έκπληξης (surprise) φαίνεται να ανεβαίνει καθώς ανεβαίνει το rating, πράγμα που σημαίνει ότι πιθανώς να εκπλήσσονται θετικά οι χρήστες, καθώς το συγκεκριμένο affect μπορεί να έχει διττή έννοια, δηλαδή μπορεί να εκφράσει και αρνητικά αλλά και θετικά συναισθήματα.

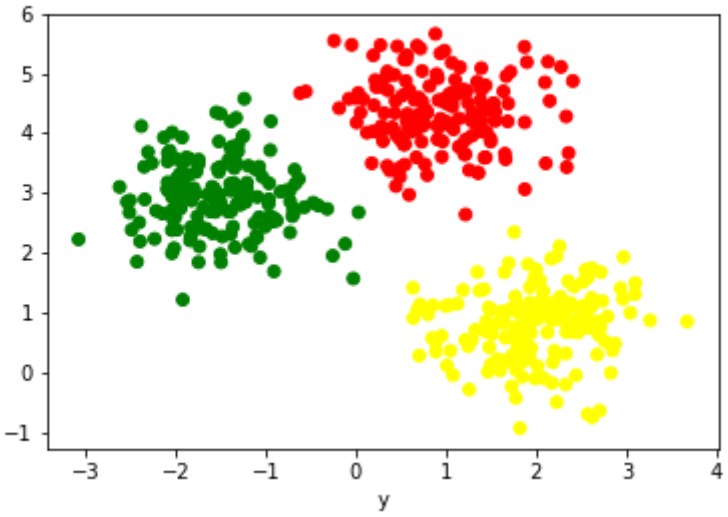
Κεφάλαιο 7
Συμπεράσματα

Όσον αφορά τα κείμενα των αξιολογήσεων, παρατηρήσαμε από τα word clouds ότι κάποιες λέξεις με μεγάλη συχνότητα όπως *καλό, καλή, οθόνη, τιμή κτλ.*, είναι κοινές μεταξύ των κατηγοριών. Αυτό σημαίνει ότι εκφράζονται παρόμοια συναισθήματα και στις 4 κατηγορίες. Από τη λεξικογραφική ανάλυση τα γραφήματα έδειξαν ότι όσο ανεβαίνει το rating (1-5), τόσο μεγαλώνει και το score για το συναίσθημα (θετικό), τη χαρά αλλά και για την υποκειμενικότητα. Αντίθετα όσο ανεβαίνει το rating μειώνονται τα scores για θυμό, απέχθεια, φόβο, και λύπη, πράγμα το οποίο είναι λογικό όταν μιλάμε για μια θετική αξιολόγηση. Όσον αφορά το score της έκπληξης φαίνεται να εκφράζει και θετικό αλλά και αρνητικό συναίσθημα. Επίσης, είδαμε ότι υπάρχουν στα γραφήματα πάρα πολλές τιμές που είναι μακριά από τις υπόλοιπες (outliers). Αυτές οι αξιολογήσεις ίσως είναι υπερβολικές σε σχέση με τις υπόλοιπες ή να μην αναφέρονται στο ίδιο θέμα με βάση το κείμενο τους και θα είχε ενδιαφέρον να εξεταστούν σε κάποια μελλοντική επέκταση της παρούσας μελέτης.

Κεφάλαιο 8
Επεκτάσεις

Η παρούσα μελέτη μπορεί να επεκταθεί στην ανάλυση των αξιολογήσεων. Θα μπορούσε να εφαρμοστεί K-Means clustering για βρούμε πόσα communities σχηματίζονται με βάση τις ομοιότητες των αξιολογήσεων. Αν δηλαδή αξιολογήσεις που έχουν ίδιο rating θα μπουν στο ίδιο community ή θα σχηματιστούν περισσότερα από 5 (όσα είναι τα ratings) communities.

Μια άλλη προσέγγιση θα ήταν να δημιουργήσουμε ακμές μεταξύ των αξιολογήσεων, δηλαδή όλα τα πιθανά ζευγάρια αξιολογήσεων (1-2, 1-3, 1-4 κοκ.) και να υπολογίσουμε το cosine similarity κάθε ζεύγους. Έπειτα και πάλι θα μπορούσαν να δημιουργηθούν communities με βάση το similarity και δούμε στο γράφο που θα σχηματιστεί για κάθε κατηγορία τι ποσοστό έχει το κάθε rating και κατά πόσο τα communities που δημιουργούνται επικεντρώνονται γύρω από ένα θέμα ή όχι.



Εικόνα 45. Παράδειγμα ομαδοποίησης δεδομένων

Βιβλιογραφία

Adam Tsakalidis, S. P. (2018, July 14). Building and evaluating resources for sentiment. *Springer*, σσ. 1021-1044.

Galarnyk, M. (2020, Μάρτιος). *Towards Data Science*. Ανάκτηση από https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Jesus Serrano-Guerrero, J. A.-V. (2015, March 22). Sentiment analysis: A review and comparative analysis of web services. *Elsevier Inc.*, σσ. 18-38.

Liu, J., Cao, Y., Lin, C.-Y., Huang, Y., & Zhou, M. (2007). Low-Quality Product Review Detection in Opinion Summarization. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (σσ. 334-342). Prague.

Sun, S., Luo, C., & Chen, J. (2017, July). A review of natural language processing techniques for opinion mining systems. *Information Fusion, 36*, 10-25.

Walaa Medhat, A. H. (2014, Μάιος 27). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, σσ. 1093–1113.

Wikipedia. (2020, Απρίλιος). Ανάκτηση από https://en.wikipedia.org/wiki/Tag_cloud

Wikipedia. (2020, Μάιος). Ανάκτηση από https://en.wikipedia.org/wiki/Scatter_plot

Wikipedia. (2020, Μάιος). Ανάκτηση από https://en.wikipedia.org/wiki/Box_plot