



Πρόγραμμα Μεταπτυχιακών Σπουδών
«Πληροφοριακά Συστήματα & Υπηρεσίες»

Ειδίκευση: Μεγάλα Δεδομένα και Αναλυτική

Εξόρυξη και Προετοιμασία δεδομένων

«Text Mining Amazon Mobile Phone Reviews»

Διδάσκουσα:

Αναπληρώτρια Καθηγήτρια

Μαρία Χαλκίδη

Φοιτητές:

Γαρδέλης Σταύρος – ΜΕ1922

Καπότης Χρήστος – ΜΕ1926

Πλατής Κωνσταντίνος – ΜΕ1940

Τζαβάρας Γιάννης – ΜΕ1946

Αθήνα 2020

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ	2
ΚΕΦΑΛΑΙΟ 1 ^ο : ΕΙΣΑΓΩΓΗ	3
ΚΕΦΑΛΑΙΟ 2 ^ο : ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ	4
ΚΕΦΑΛΑΙΟ 3 ^ο : ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ	7
ΚΕΦΑΛΑΙΟ 4 ^ο : ΑΠΟΤΕΛΕΣΜΑΤΑ	12
4.1 Ανάλυση - συσχετίσεις δεδομένων	12
4.2 Ανάλυση συναισθήματος	14
4.3 Κατηγοριοποίηση.....	16
ΚΕΦΑΛΑΙΟ 5 ^ο : ΣΥΜΠΕΡΑΣΜΑΤΑ	18
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	19

ΚΕΦΑΛΑΙΟ 1^ο: ΕΙΣΑΓΩΓΗ

Στην παρούσα εργασία γίνεται μια αναφορά στην ανάλυση συναισθήματος σε reviews χρηστών που έχουν προχωρήσει στην αγορά ορισμένων μεταχειρισμένων κινητών τηλεφώνων. Σκοπός είναι, αφενός η εξόρυξη πληροφορίας σχετικά με το αν ένα μοντέλο έχει θετικές, αρνητικές ή ουδέτερες κριτικές, αφετέρου, η δημιουργία μοντέλων πρόβλεψης για το αν ένας υποψήφιος αγοραστής θα μείνει ικανοποιημένος από την επιλογή ενός συγκεκριμένου μοντέλου.

Το σύνολο δεδομένων που τέθηκε προς επεξεργασία εντοπίζεται στον ιστότοπο <https://www.kaggle.com/PromptCloudHQ/amazon-reviews-unlocked-mobile-phones/data> και περιλαμβάνει 400.000 reviews μεταχειρισμένων κινητών τηλεφώνων που πουλήθηκαν στο amazon.com. Το σύνολο δεδομένων περιέχει πληροφορίες σχετικά με τα βασικά στοιχεία του κινητού τηλεφώνου, όπως τίτλος προϊόντος, μάρκα, τιμή, βαθμολογία (1-5), κείμενο του review, καθώς και πόσα άτομα βρήκαν χρήσιμη την αξιολόγηση.

Με την τεχνική sentiment analysis έγινε προσπάθεια κατηγοριοποίησης των κριτικών-σχολίων του συνόλου δεδομένων σε τρεις κλάσεις (θετικά, αρνητικά, ουδέτερα), ενώ στη συνέχεια χρησιμοποιήθηκαν διάφοροι αλγόριθμοι κατηγοριοποίησης (δέντρα απόφασης, αλγόριθμος k -Nearest Neighbors, λογιστική παλινδρόμηση, αλγόριθμοι Ταξινόμησης Bayes, νευρωνικά δίκτυα, random forest, SVM) χρησιμοποιώντας σαν κλάση τα αποτελέσματα της sentiment analysis, με σκοπό τη δημιουργία ενός συστήματος πρόβλεψης. Το σύστημα αυτό προβλέπει, δεχόμενο ως είσοδο από τον υποψήφιο πελάτη το μοντέλο, τη μάρκα και το ποσό που διαθέτει για την αγορά της συσκευής, αν θα μείνει ικανοποιημένος ή όχι από την αγορά.

ΚΕΦΑΛΑΙΟ 2^ο: ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Η ανάλυση συναισθήματος αποτελεί μία υποπεριοχή της ταξινόμησης κειμένου (text classification) και αναφέρεται στη διαδικασία εξαγωγής πληροφοριών για τη συναισθηματική κατάσταση ενός χρήστη μέσα από το γραπτό λόγο του. Χρησιμοποιεί τεχνικές επεξεργασίας φυσικού λόγου (natural language processing - NLP), στατιστικές μεθόδους και μεθόδους μηχανικής μάθησης για την ταξινόμηση ενός κειμένου σε κλάσεις που εκφράζουν συναίσθημα.

Ένας πρώτος διαχωρισμός της ανάλυσης συναισθήματος γίνεται με βάση την ακριβή έννοια της συναισθηματικής κατάστασης που επιχειρεί να προσδιορίσει. Αυτή σύμφωνα με τη wikipedia μπορεί να αναφέρεται είτε στη γενικότερη συναισθηματική κατάσταση του συγγραφέα κατά τη συγγραφή του κειμένου (affective state), είτε στο συναίσθημα που μεταδίδεται σκόπιμα από τον συγγραφέα στον αναγνώστη μέσω του κειμένου, είτε στην στάση – άποψη – εκτίμηση του συγγραφέα σχετικά με κάποιο θέμα. Στις πρώτες δύο περιπτώσεις η ταξινόμηση μπορεί να γίνει σε κλάσεις που εκφράζουν συναισθήματα αντιληπτά από τον άνθρωπο. Επιχειρείται δηλαδή η αναγνώριση πραγματικών συναισθημάτων στο κείμενο όπως η χαρά, η λύπη και ο θυμός ή και καταστάσεων όπως η ειρωνεία. Η ταξινόμηση ωστόσο μπορεί να γίνει και σε γενικότερες κλάσεις όπως θετικό, αρνητικό και ουδέτερο συναίσθημα. Στην τελευταία περίπτωση, όπου εξετάζεται η στάση ως προς κάποιο θέμα, η ταξινόμηση γίνεται συνήθως σε δύο (θετική στάση, αρνητική στάση), τρεις (θετική, ουδέτερη, αρνητική) ή πέντε (θετική, μάλλον θετική, ουδέτερη, μάλλον αρνητική, αρνητική) κλάσεις.

Ένας άλλος διαχωρισμός γίνεται με βάση το μέγεθος του κειμένου που εξετάζεται. Έτσι μπορεί να αναζητείται το συναίσθημα ή η πολικότητά του (polarity) σε ένα ολοκληρω κείμενο (document-based sentiment analysis), σε μία πρόταση (sentence-based sentiment analysis) ή ακόμα και σε μεμονωμένες φράσεις (feature/aspect-based sentiment analysis) όταν αυτές αναφέρονται σε χαρακτηριστικά μίας οντότητας (features of an entity) ως προς τα οποία αναζητούμε το συναίσθημα. Μια τέτοια περίπτωση συναντάται στη παρούσα

εργασία, όπου καλούμαστε να εξάγουμε πληροφορία σχετικά με τον αν μια κριτική για την αγορά ενός μεταχειρισμένου κινητού τηλεφώνου (οντότητα) θεωρείται θετική, αρνητική ή ουδέτερη.

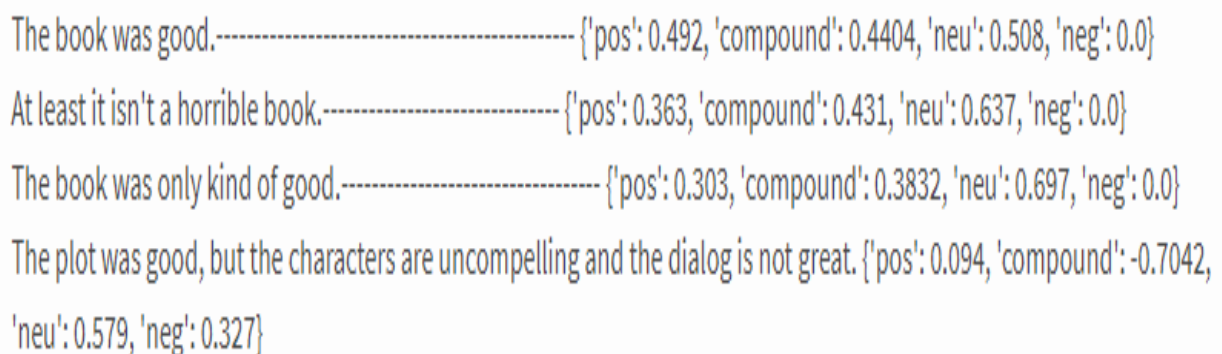
Κατά κύριο λόγο η ανάλυση συναισθήματος προσεγγίζεται με δύο διαφορετικούς τρόπους. Ο πρώτος τρόπος προσέγγισης του προβλήματος της ανάλυσης συναισθήματος είναι η εξαγωγή χαρακτηριστικών και η χρήση μεθόδων μηχανικής μάθησης, ενώ ο δεύτερος είναι η χρήση κατάλληλων λεξικών, μέθοδος που χρησιμοποιήθηκε και στη παρούσα εργασία.

Η μέθοδος «lexicon-based» κάνει χρήση συναισθηματικών λεξικών για να αποδώσει συναισθηματική βαθμολογία σε λέξεις και φράσεις. Στη συνέχεια συνθέτει τις λέξεις και τις φράσεις για να προκύψει συναισθηματική βαθμολογία για όλο το προς εξέταση κείμενο. Μία απλή υλοποίηση της lexicon-based προσέγγισης, στην περίπτωση απλών προτάσεων, είναι η πρόσθεση των συναισθηματικών βαθμολογιών των λέξεων με τον τρόπο που υποδεικνύει το συντακτικό δέντρο της πρότασης. Έτσι ξεκινώντας από τα φύλλα, που αποτελούν λέξεις, γίνονται διαδοχικές αθροίσεις των βαθμολογιών μέχρι το τελικό σκορ στη ρίζα του δέντρου, που χαρακτηρίζει όλη την πρόταση. Ανάλογα με την τελική βαθμολογία η πρόταση χαρακτηρίζεται αρνητική ή θετική. Τέτοιες προσεγγίσεις συχνά περιέχουν και χειρισμό της άρνησης (negation handling). Καθώς η άρνηση αντιστρέφει το συναίσθημα της λέξης (cool - not cool), μπορεί για παράδειγμα η συναισθηματική βαθμολογία λέξεων που συντάσσονται με άρνηση απλά να αλλάζει πρόσημο. Διατηρείται δηλαδή η ένταση του συναισθήματος αλλά αντιστρέφεται η πολικότητα. Ένα ιδιαίτερα δημοφιλές λεξικό συναισθηματικής πολικότητας είναι το VADER (Valence Aware Dictionary and sEntiment Reasoner), η εκδοχή του οποίου αποδίδει συναισθηματική βαθμολογία σε λέξεις σε τρία επίπεδα, θετική, αρνητική ή ουδέτερη.

Το VADER, επικυρωμένο από πολλαπλούς ανεξάρτητους κριτές, ενσωματώνει ένα «χρυσό πρότυπο» λεξικό συναισθημάτων. Το λεξικό VADER είναι ευαίσθητο τόσο στην πολικότητα όσο και στην ένταση των συναισθημάτων που εκφράζονται σε αναρτήσεις κοινωνικών μέσων δικτύωσης. Επίσης βρίσκει εφαρμογή στην ανάλυση συναισθήματος και σε άλλους τομείς. Πάνω από 9.000

χαρακτηριστικά βαθμολογήθηκαν με κλίμακα από το -4 «εξαιρετικά αρνητικό» έως το +4 «εξαιρετικά θετικό», με τα ουδέτερα χαρακτηριστικά να βαθμολογούνται με το 0. Εν συνεχεία συγκρατήθηκαν τα λεξιλογικά χαρακτηριστικά που είχαν μια μη μηδενική μέση τιμή διαβάθμισης, και των οποίων η τυπική απόκλιση ήταν μικρότερη από 2,5. Αυτό είχε ως αποτέλεσμα τη δημιουργία ενός λεξικού που αποτελείται από 7.500 λεξιλογικά χαρακτηριστικά με επικυρωμένες βαθμολογίες που προσδιορίζουν τόσο την πολικότητα ως προς το συναίσθημα (θετικό / αρνητικό), όσο και την ένταση αυτού σε μια κλίμακα από -4 έως 4. Για παράδειγμα, η λέξη «εντάξει» έχει ένα θετικό βάρος +0.9, η λέξη «καλό» +1.9, και η λέξη «μεγάλο» +3.1.

Η συνολική βαθμολογία μιας πρότασης προκύπτει από το άθροισμα της βαθμολογίας των επιμέρους λέξεων από τις οποίες αποτελείται. Στη συνέχεια το αποτέλεσμα κανονικοποιείται με μέγιστη τιμή το +1 (πιο ακραία θετική τιμή) και ελάχιστη το -1 (πιο ακραία αρνητική τιμή). Οι προτάσεις με σκορ (compound) μεγαλύτερο ή ίσο του +0.05 ταξινομούνται ως θετικές, εκείνες με σκορ μικρότερο ή ίσο του -0.05 ως αρνητικές, με τις υπόλοιπες να κατηγοριοποιούνται ως ουδέτερες. Στην παρακάτω εικόνα φαίνεται το αποτέλεσμα της ανάλυσης συναισθήματος με τη χρήση του λεξικού Vader σε ορισμένες προτάσεις.



The book was good.	{ 'pos': 0.492, 'compound': 0.4404, 'neu': 0.508, 'neg': 0.0 }
At least it isn't a horrible book.	{ 'pos': 0.363, 'compound': 0.431, 'neu': 0.637, 'neg': 0.0 }
The book was only kind of good.	{ 'pos': 0.303, 'compound': 0.3832, 'neu': 0.697, 'neg': 0.0 }
The plot was good, but the characters are un compelling and the dialog is not great.	{ 'pos': 0.094, 'compound': -0.7042, 'neu': 0.579, 'neg': 0.327 }

Σχήμα 1. Εφαρμογή του λεξικού Vader σε κριτικές βιβλίου.

ΚΕΦΑΛΑΙΟ 3^ο: ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ

Σκοπός μας στο σημείο αυτό, ήταν η δημιουργία ενός συστήματος που θα προβλέπει αν ένας υποψήφιος αγοραστής θα μείνει ικανοποιημένος από την επικείμενη αγορά, δίνοντας ως είσοδο στον σύστημα το μοντέλο της συσκευής, τη μάρκα και την τιμή στην οποία επιθυμεί να την αγοράσει. Ως κλάση για την κατηγοριοποίηση θέσαμε το αποτέλεσμα της ανάλυσης συναισθήματος που προηγήθηκε στις κριτικές του συνόλου δεδομένων (θετική, αρνητική ή ουδέτερη). Για την υλοποίηση ενός τέτοιου συστήματος χρησιμοποιήθηκαν διάφοροι αλγόριθμοι κατηγοριοποίησης (δέντρα απόφασης, αλγόριθμος k -Nearest Neighbors, λογιστική παλινδρόμηση, αλγόριθμοι ταξινόμησης Bayes, νευρωνικά δίκτυα, random forest, SVM) και εξετάστηκε η ακρίβεια καθενός εξ αυτών. Παρακάτω παρατίθενται ορισμένα βασικά στοιχεία των αλγορίθμων κατηγοριοποίησης που χρησιμοποιήθηκαν.

➤ Δένδρα απόφασης

Δενδροειδής δομή που με γραφικό τρόπο περιγράφει τα δεδομένα. Μια αναπαράσταση σε δένδρο θα μπορούσε να έχει τη μορφή:

- Κάθε κόμβος ορίζει μια συνθήκη ελέγχου της τιμής κάποιου χαρακτηριστικού των περιπτώσεων
- Κάθε κλαδί που φεύγει από ένα κόμβο αντιστοιχεί σε μια διαφορετική διακριτή τιμή του χαρακτηριστικού που σχετίζεται με τον κόμβο.
- Στα κλαδιά φύλλα έχουμε το τι συνέβη.

Τα δένδρα ταξινόμησης χρησιμοποιούνται για να προβλέψουν, με κάποιο βαθμό ακρίβειας, την τιμή της μεταβλητής που μοντελοποιούν με βάση τις τιμές των θεωρούμενων ανεξάρτητων μεταβλητών (χαρακτηριστικών).

➤ K - Κοντινότερος Γείτονας

Ο αλγόριθμος k -Κοντινότερος Γείτονας (k -Nearest Neighbors) είναι η γενίκευση του κανόνα του πλησιέστερου γείτονα στα k πλησιέστερα σημεία στο σημείο δείγμα. Η παράμετρος k καθορίζεται από τον χρήστη και

επιλέγεται συνήθως μέσω πειραμάτων (hyperparameter optimization) αφού η επίδοση διαφορετικών k εξαρτάται από την γενικότερη διάταξη των σημείων εκπαίδευσης και δεν απαντά σε συγκεκριμένους κανόνες. Εν γένει μεγαλύτερες τιμές k μπορούν να βελτιώσουν το αποτέλεσμα της ταξινόμησης αλλά και να συμπεριλάβουν στα σημείαψηφοφόρους απομακρυσμένα δείγματα (outliers) με αρνητικές συνέπειες. Ωστόσο στην περίπτωση δύο κλάσεων αποφεύγονται άρτιες τιμές k για να μην υπάρχει ισοπαλία στην ψήφο της κλάσης. Με την ίδια λογική, στη γενική περίπτωση K κλάσεων αποφεύγονται πολλαπλάσια του K δηλαδή $k = K, k = 2K, k = 3K \dots$ Ως μετρική απόστασης μπορεί να χρησιμοποιηθεί οποιαδήποτε μαθηματικά θεμελιωμένη απόσταση σημείων σε n -διάστατο χώρο. Οι συνηθέστερες επιλογές είναι η Ευκλείδεια απόσταση (Euclidean Distance), η απόσταση Minkowski (Minkowski Distance), η απόσταση Mahalanobis (Mahalanobis Distance) και η απόσταση Hamming (Hamming Distance), που μετρά αποστάσεις ανάμεσα σε strings, σε εφαρμογές επεξεργασίας και ταξινόμησης κειμένου.

➤ Λογιστική παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic regression) αποτελεί στην ουσία ένα μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Y με βάση τη θεωρία των πιθανοτήτων. Στο μοντέλο αυτό όπου η μεταβλητή Y συνήθως έχει δυαδικό χαρακτήρα (λαμβάνει δύο τιμές) στοχεύεται η πρόβλεψη της έκβασης αυτής από ένα πλήθος προβλεπτικών μεταβλητών που μπορεί να είναι ονομαστικές, τακτικές ή ποσοτικές. Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην μεν πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική, στη δε δεύτερη αποκλειστικά ποσοτική. Ενώ κατά την κλασική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων a και b γίνεται με τη μέθοδο των ελαχίστων τετραγώνων, κατά τη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο του λόγου πιθανοφάνειας (μέθοδος συνήθως εφαρμοζόμενη στα γενικευμένα γραμμικά υποδείγματα), δηλαδή επιλέγονται οι πιο πιθανοφανείς τιμές των παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα

αποτελέσματα. Ως επακόλουθο, η πρώτη παραδέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων ενώ στη δεύτερη αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτήν.

➤ *Αλγόριθμοι ταξινόμησης Bayes*

Ο κατηγοριοποιητής Bayes χρησιμοποιεί το γνωστό θεώρημα Bayes για να προβλέψει την κλάση αυτή που μεγιστοποιεί την μεταγενέστερη πιθανότητα. Ο κύριος στόχος είναι να εκτιμηθεί η συνδυαστική συνάρτηση πυκνότητας πιθανότητας για κάθε κλάση που μοντελοποιείται μέσω μιας πολυμεταβλητής κανονικής διανομής. Ο απλουστευμένος κατηγοριοποιητής Bayes υποθέτει ότι τα γνωρίσματα ,προς μελέτη, είναι ανεξάρτητα μεταξύ τους, παρόλα αυτά χρησιμοποιείται πολύ συχνά σε εφαρμογές. Το πιο σημαντικό πρόβλημα που συναντάται με τον κατηγοριοποιητή Bayes είναι η έλλειψη αρκετών δεδομένων ώστε να υπολογιστεί με ακρίβεια η συλλογική πυκνότητα πιθανότητας ειδικά σε μεγάλης διάστασης δεδομένα. Για να λυθεί το πρόβλημα αυτό χρειάζεται να μειωθεί ο αριθμός των παραμέτρων και να χρησιμοποιηθεί ο απλοποιημένος κατηγοριοποιητής Bayes. Ο Naive Bayes ταξινομητής προσφέρει μια απλή πιθανοθεωρητική προσέγγιση στα προβλήματα μάθησης με επίβλεψη, όπου στόχος μας είναι να προβλέψουμε επακριβώς την κατηγορία-κλάση των στιγμιότυπων δοκιμής χρησιμοποιώντας ταξινομημένα στιγμιότυπα εκπαίδευσης που περιλαμβάνουν την πληροφορία της κλάσης που ανήκουν. Βασίζεται σε δυο σημαντικές εκλαϊκευτικές υποθέσεις. Συγκεκριμένα, υποθέτει ότι κάθε χαρακτηριστικό των στιγμιότυπων είναι στοχαστικά ανεξάρτητα των υπολοίπων, δεδομένης της κλάσης και ότι δεν υπάρχουν άλλα κρυφά χαρακτηριστικά που να επηρεάζουν την διαδικασία της πρόβλεψης.

➤ Νευρωνικά δίκτυα

Τα τεχνητά νευρωνικά δίκτυα (Artificial Neural Networks - ANNs) είναι υπολογιστικά μοντέλα που εμπνέονται από βιολογικές διαδικασίες μάθησης. Απαρτίζονται από απλούς υπολογιστικούς κόμβους που καλούνται νευρώνες (neurons) οι οποίοι συνδέονται μεταξύ τους δημιουργώντας ένα δίκτυο. Κάθε διασύνδεση νευρώνων στο δίκτυο χαρακτηρίζεται από κάποιο βάρος w_i που προσαρμόζεται κατά τη φάση εκπαίδευσης του δικτύου. Η εκπαίδευση συνολικά συνιστά τη διαδικασία προσαρμογής όλων των βαρών του δικτύου. Τα παραπάνω απαντώνται γενικά στις διάφορες κατηγορίες νευρωνικών δικτύων (Multilayer perceptron κ.α). Έχουν το μεγάλο πλεονέκτημα της ανοχής που παρουσιάζουν σε δεδομένα εκπαίδευσης με θόρυβο, δηλαδή δεδομένα που περιστασιακά έχουν λανθασμένες τιμές (π.χ. λάθη καταχώρησης). Ωστόσο, αδυνατούν να εξηγήσουν ποιοτικά τη γνώση που μοντελοποιούν.

➤ Random Forest

Ο αλγόριθμος αυτός αναπτύχθηκε από τους Leo Breiman και Adele Cutler το 2001. Αποτελεί μια γενίκευση (επέκταση) του bagging, καθώς μεταφέρει κατά μια έννοια την μεθοδολογία του bootstrap, εκτός από τα instances και στα attributes των δεδομένων. Είναι από τους δημοφιλέστερους αλγόριθμους στην κατηγορία του, κυρίως για την ταχύτητα αλλά και την ακρίβεια που προσφέρει. Σύμφωνα με τους δημιουργούς του, προσφέρει την καλύτερη ακρίβεια μεταξύ των υπαρχόντων αλγορίθμων. Η ταχύτητά του είναι πολύ καλή ακόμα και σε πολύ μεγάλα σύνολα δεδομένων εκπαίδευσης. Μπορεί να χειριστεί αποδοτικά πάρα πολύ μεγάλο αριθμό χαρακτηριστικών (ακόμα και χιλιάδες). Επίσης, δίνει μια εκτίμηση για το ποια χαρακτηριστικά είναι τα πιο σημαντικά στην κατηγοριοποίηση. Δεν χρειάζεται την χρήση διαφορετικού συνόλου δεδομένων για τον έλεγχο ακρίβειας (δεν είναι δηλαδή απαραίτητο το cross-validation), καθώς η εκτίμηση του λάθους γενίκευσης γίνεται από τον ίδιο τον αλγόριθμο κατά την εκτέλεσή του. Τέλος, μπορεί να χειριστεί αποδοτικά ελλιπή δεδομένα. Δεν παρουσιάζει φαινόμενα υπερεκπαίδευσης.

➤ Μηχανές Διανυσμάτων Στήριξης (SVMs)

Η μέθοδος των ΜΔΣ χρησιμοποιείται για κατηγοριοποίηση γραμμικών και μη γραμμικών δεδομένων. Συνοπτικά ο αλγόριθμος δουλεύει ως εξής: χρησιμοποιεί μια μη γραμμική τεχνική αντιστοίχισης για το μετασχηματισμό του πρωτότυπου συνόλου των δεδομένων προς εκπαίδευση σε μια πιο υψηλή διάσταση. Σε αυτή την καινούργια διάσταση αναζητά το βέλτιστο γραμμικό διαχωριστικό υπερεπίπεδο, δηλαδή το όριο που διαχωρίζει τις πλειάδες που ανήκουν σε μια κλάση από αυτές μιας διαφορετικής κλάσης. Με κατάλληλη μη γραμμική αντιστοίχιση σε μια επαρκώς υψηλή διάσταση, τα δεδομένα που ανήκουν σε δύο διαφορετικές κλάσεις μπορούν πάντα να διαχωριστούν με ένα υπερεπίπεδο. Η μέθοδος των ΜΔΣ βρίσκει αυτό το υπερεπίπεδο χρησιμοποιώντας διανύσματα στήριξης, ουσιώδεις δηλαδή πλειάδες προς εκπαίδευση, και περιθώρια, που ορίζονται από τα διανύσματα στήριξης.

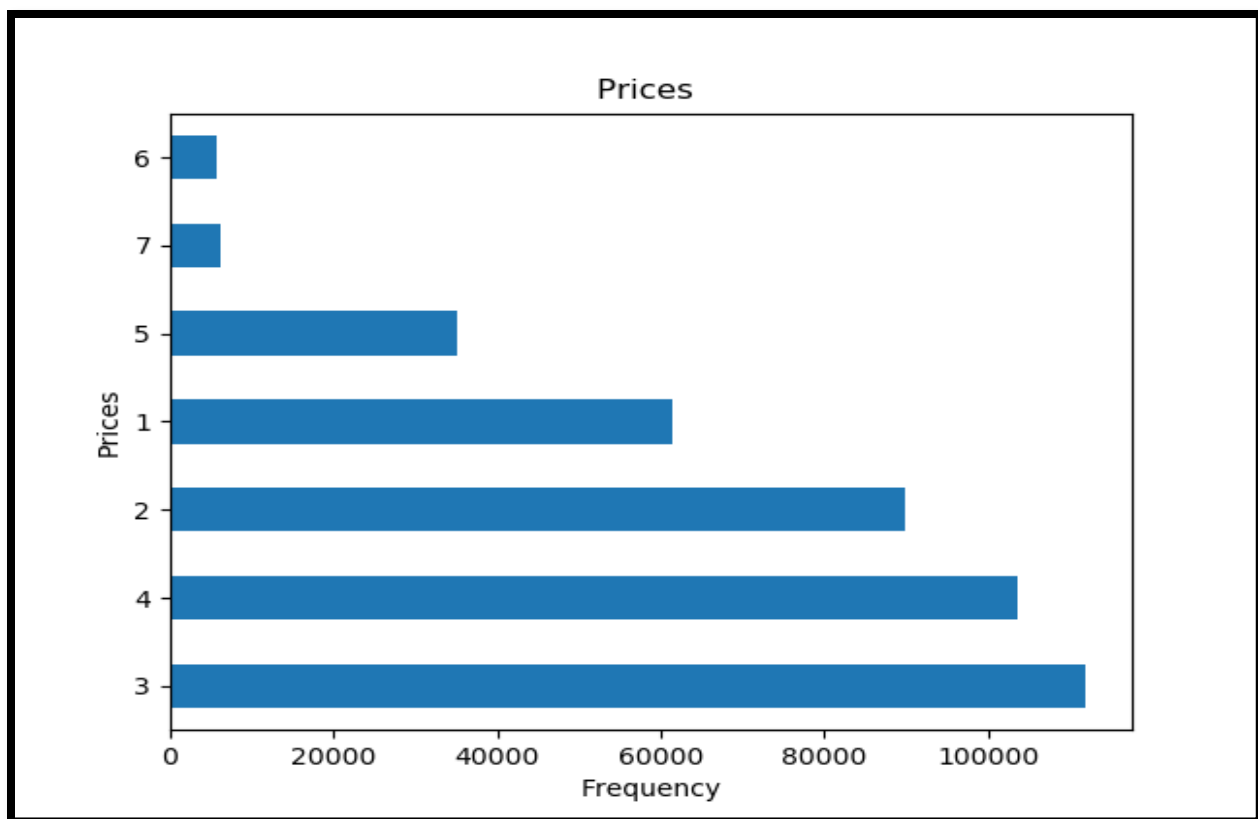
ΚΕΦΑΛΑΙΟ 4^ο : ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Ανάλυση - συσχετίσεις δεδομένων

Σε πρώτη φάση έγινε προσπάθεια ανάλυσης τόσο των δεδομένων, όσο και των συσχετίσεων μεταξύ τους. Αρχικά, εντοπίσαμε ποιό είναι το εύρος τιμών των συσκευών που έχουν τη μεγαλύτερη ζήτηση. Το εύρος των τιμών (σε ευρώ) χωρίστηκε σε επιμέρους 7 ομάδες, όπως παρουσιάζονται παρακάτω.

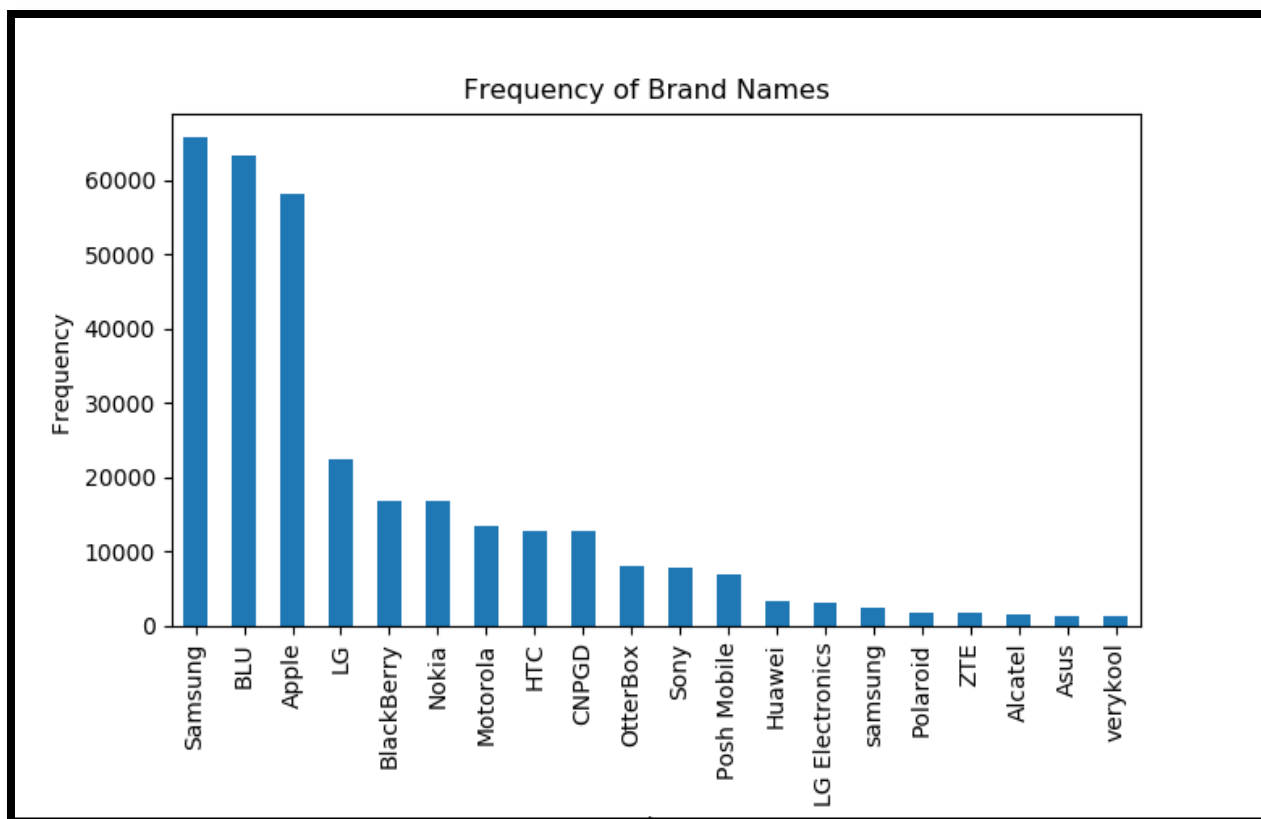
1. ≤ 50
2. > 50 έως ≤ 100
3. > 100 έως ≤ 200
4. > 200 έως ≤ 500
5. > 500 έως ≤ 1000
6. > 1000 έως ≤ 2000
7. διάφορη από τις παραπάνω

Στο ακόλουθο διάγραμμα παρουσιάζεται ο αριθμός των πωλήσεων σε συνάρτηση με το εύρος τιμής που πωλήθηκε κάθε συσκευή. Διαπιστώνουμε, ότι η πλειοψηφία των αγοραστών επέλεξε να διαθέσει ποσό της τάξεων των 100 έως 200 ευρώ για την αγορά μεταχειρισμένου κινητού της επιθυμίας του.



Σχήμα 2. Συχνότητα αγοράς συσκευής συναρτήσει της τιμής της

Ένα άλλο ενδιαφέρον στοιχείο που εξετάσαμε ήταν ο εντοπισμός των εταιριών, των οποίων τα κινητά έρχονται πρώτα σε ζήτηση στο σύνολο δεδομένων μας. Παρατηρήθηκε ότι μοντέλα των εταιρειών Samsung, Blue και Apple εμφάνισαν την υψηλότερη συχνότητα πώλησης, με τις Alcatel, Asus και Very kool να είναι ουραγοί στη σχετική λίστα.



Σχήμα 3. Συχνότητα επιλογής συσκευής εκάστοτε εταιρίας

4.2 Ανάλυση συναισθήματος

Αφού απομονώθηκε η στήλη των reviews από το σύνολο δεδομένων, εφαρμόστηκε σε αυτή, το λεξικό Vader Sentiment. Σε κάθε πρόταση δόθηκε μία συγκεκριμένη βαθμολογία από το +1 (ακραία θετική) έως το -1 (ακραία αρνητική). Οι προτάσεις με μέσο (compound) σκορ μεγαλύτερο του +0.05 ταξινομήθηκαν ως θετικές, αυτές με σκορ μικρότερο του -0.05 ως αρνητικές και αυτές με ενδιάμεσο σκορ ως ουδέτερες. Παρακάτω φαίνονται μερικά στιγμιότυπα της ανάλυσης συναισθήματος σε μία θετική και μία αρνητική κριτική αντίστοιχα.

🌈 "NICE PHONE, NICE UP GRADE FROM MY PANTACH REVUE. VERY CLEAN SET UP AND EASY SET UP. NEVER HAD AN ANDROID PHONE BUT THEY ARE

FANTASTIC TO SAY THE LEAST. PERFECT SIZE FOR SURFING AND SOCIAL MEDIA. GREAT PHONE SAMSUNG"

```
Sentence Overall Rated As Overall sentiment dictionary is : {'neg': 0.072, 'neu': 0.597, 'pos': 0.331, 'compound': 0.9231}  
sentence was rated as 7.199999999999999 % Negative  
sentence was rated as 59.699999999999996 % Neutral  
sentence was rated as 33.1 % Positive
```

🚩 *"I ALREADY HAD A PHONE WITH PROBLEMS... I KNOW IT STATED IT WAS USED, BUT DANG, IT DID NOT STATE THAT IT DID NOT CHARGE. I WISH I WOULD HAVE READ THESE COMMENTS THEN I WOULD HAVE NOT PURCHASED THIS ITEM.... AND ITS CRACKED ON THE SIDE.. DAMAGED GOODS IS WHAT IT IS.... IF TRYING TO CHARGE IT ANOTHER WAY DOES NOT WORK I AM REQUESTING FOR MY MONEY BACK... AND I WILL GET MY MONEY BACK...SIGNED AN UNHAPPY CUSTOMER...."*

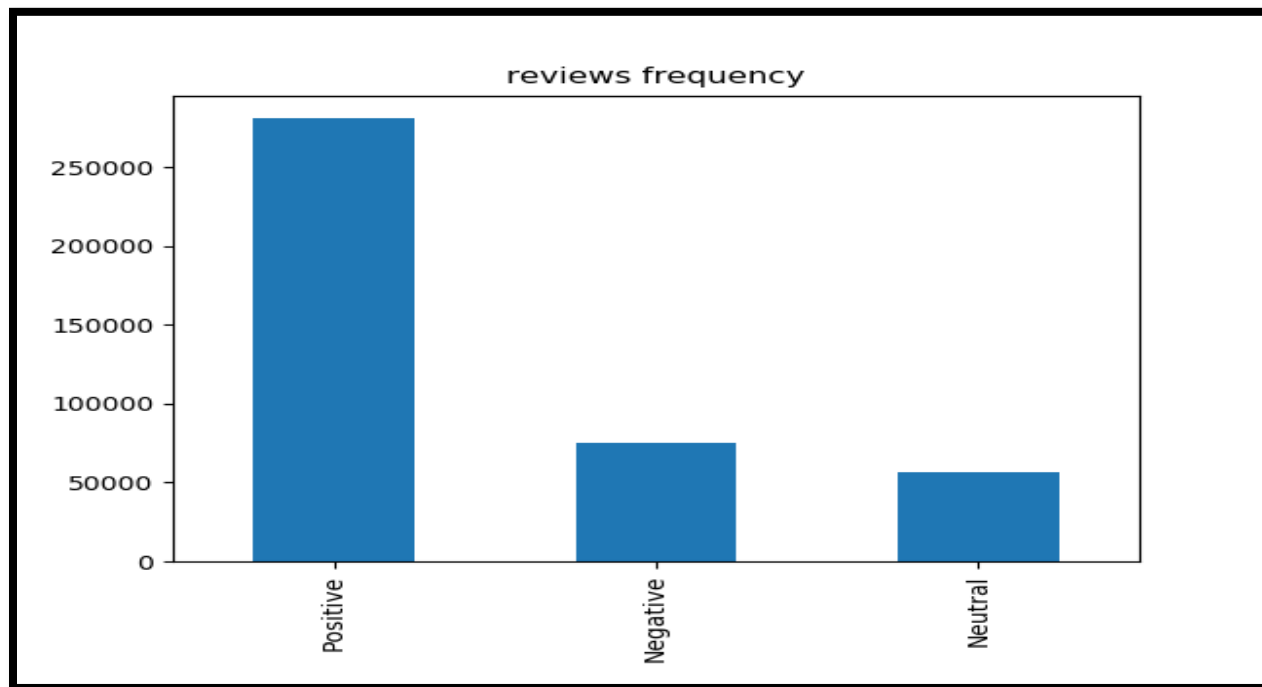
```
Sentence Overall Rated As Overall sentiment dictionary is : {'neg': 0.131, 'neu': 0.869, 'pos': 0.0, 'compound': -0.8869}  
sentence was rated as 13.100000000000001 % Negative  
sentence was rated as 86.9 % Neutral  
sentence was rated as 0.0 % Positive
```

Το αποτέλεσμα της ανάλυσης σε ένα σύνολο 413840 κριτικών, ανέδειξε τα εξής αποτελέσματα:

- 75480 αρνητικές κριτικές
- 56813 ουδέτερες κριτικές
- 281547 θετικές κριτικές

Εν συνεχεία δημιουργήθηκε μία νέα στήλη με την ονομασία «reviews», με τρεις τιμές (positive, neutral, negative), η οποία αντικατέστησε τη στήλη «reviews» του αρχικού συνόλου δεδομένων που περιείχε χαρακτήρες κειμένου (text).

Παρακάτω, ακολουθεί το διάγραμμα που αναπαριστά τη συχνότητα των θετικών, αρνητικών και ουδέτερων κριτικών στο σύνολο δεδομένων μας.



Σχήμα 4. Συχνότητα εμφάνισης κριτικών

4.3 Κατηγοριοποίηση

Σε επόμενη φάση δημιουργήθηκε ένα σύστημα πρόβλεψης, για το εάν κάποιος υποψήφιος αγοραστής θα μείνει ικανοποιημένος από μια επικείμενη αγορά με βάση το μοντέλο, τη μάρκα και τη τιμή της συσκευής που προτίθεται να αγοράσει. Χρησιμοποιήθηκαν διάφοροι αλγόριθμοι κατηγοριοποίησης και εξετάστηκε η ακρίβεια καθενός εξ αυτών.

Ως χαρακτηριστικά (attributes) για το σύνολο εκπαίδευσης χρησιμοποιήθηκαν οι στήλες «product name», «brand name», «price» και ως κλάση (label) η στήλη «reviews» του προηγούμενου ερωτήματος. Χρησιμοποιήθηκαν οι βιβλιοθήκες pandas, numpy καθώς και η sklearn της python, ενώ με τη χρήση του label_encoder δόθηκε μία αναγνωριστική τιμή για κάθε μοναδική εγγραφή των

στηλών «product name» και «brand name». Επίσης το εύρος των τιμών χωρίστηκε σε επιμέρους 7 ομάδες που αναφέραμε παραπάνω. (1^η: ≤ 50 , 2^η : >50 έως ≤ 100 , 3^η : >100 έως ≤ 200 , 4^η : >200 έως ≤ 500 , 5^η : >500 έως ≤ 1000 , 6^η : >1000 έως ≤ 2000 , 7^η : διάφορη από τις παραπάνω) για διευκόλυνση την ταξινόμησης. Τέλος η κλάση «reviews» αποτελούνταν από δύο τιμές: θετική και αρνητική προοπτική.

Η ακρίβεια (accuracy) των αλγορίθμων κατηγοριοποίησης παρουσιάζεται παρακάτω.

- Δένδρα απόφασης : 68%
- Λογιστική παλινδρόμηση : 74.1%
- Απλός ταξινομήτης Bayes : 68.5%
- Νευρωνικό δίκτυο : 68%
- Random forest : 74%

Παρατηρούμε ότι οι μέθοδοι random forest και logistic regression πέτυχαν την υψηλότερη ακρίβεια πρόβλεψης για το μοντέλο που κατασκευάσαμε. Επίσης, αξίζει να τονιστεί ότι ο k-NN καθώς και οι μηχανές διανυσμάτων στήριξης δεν είχαν ικανοποιητική ακρίβεια για το λόγο αυτό και εξαιρέθηκαν από την παρουσίαση.

ΚΕΦΑΛΑΙΟ 5^ο: ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία έγινε προσπάθεια εξόρυξης πληροφορίας, μέσω ανάλυσης συναισθήματος που έλαβε χώρα σε κριτικές χρηστών μεταχειρισμένων κινητών τηλεφώνων. Η ανάλυση των κριτικών πραγματοποιήθηκε με τη χρήση του λεξικού συναισθημάτων VADER (Valence Aware Dictionary and sEntiment Reasoner) της python. Κατά αυτόν τον τρόπο χαρακτηρίστηκε καθεμία κριτική ως θετική, αρνητική ή ουδέτερη ανάλογα το περιεχόμενο της. Εν συνεχεία, σε δεύτερη φάση, δοκιμάστηκε η δημιουργία ενός μοντέλου-συστήματος πρόβλεψης αν ένας υποψήφιος αγοραστής θα μείνει ικανοποιημένος ή όχι από την αγορά ενός συγκεκριμένου κινητού τηλεφώνου, στην τιμή που επιθυμεί να διαθέσει. Εξετάστηκαν διάφοροι αλγόριθμοι ταξινόμησης, με τη μέθοδο της λογιστικής παλινδρόμησης να επιτυγχάνει τη μεγαλύτερη ακρίβεια, της τάξεως του 74.1%. Η ακρίβεια του μοντέλου πρόβλεψης δεν κρίνεται ως ικανοποιητική και αυτό οφείλεται σε μεγάλο βαθμό στην έλλειψη πολλών μεταβλητών-παραγόντων (attributes) που θέσαμε στο σύνολο εκπαίδευσης. Για μελλοντική μελέτη, θα παρουσίαζε ενδιαφέρον η προσθήκη και άλλων γνωρισμάτων στο προς εξέταση σύνολο δεδομένων (πχ: έτος κατασκευή συσκευής, τύπος επεξεργαστή, πιθανές ζημιές κ.α), οπότε και θα αναμένεται υψηλότερη ακρίβεια στο μοντέλο πρόβλεψης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] <https://pypi.org/project/vaderSentiment/>
- [2] <http://www.dbnet.ece.ntua.gr/pubs/uploads/DIPL-2006-15.pdf>
- [3] <http://artemis.cslab.ece.ntua.gr:8080/jspui/handle/123456789/12907>
- [4] <http://www.image.ece.ntua.gr/papers/886.pdf?fbclid=IwAR2VXTNmVNbqXJaA2JeJz0-y2-jHSYhkDJFK9J32Hx76bMNyldlqLty462k>
- [5] <https://evdoxos.ds.unipi.gr/modules/document/index.php?course=DSERV139&openDir=/5dcd05b3AeMm>
- [6] https://books.google.gr/books?id=hSNTDwAAQBAJ&printsec=frontcover&dq=sentiment%20analysis%20data%20mining&hl=el&sa=X&ved=0ahUKEwjytP6x6t_nAhWD-aQKHZMYBVcQ6AEIKTAA&fbclid=IwAR1BV3opjgiQwhoxDUwVzcZtfrsmqxufe6LMLX7yKKgSagBaTfK5EyesGDM#v=onepage&q=sentiment%20analysis%20data%20mining&f=false
- [7] https://books.google.gr/books?id=IimgDQAAQBAJ&printsec=frontcover&dq=sentiment%20analysis%20python&hl=el&sa=X&ved=0ahUKEwiewr3h6t_nAhXG4KQKHUadDQ8Q6AEISDAD&fbclid=IwAR2ILJ0X3yVD2H7yAnHvt0XbcSnlw3nsisTYwy37BOlZo2PUfZQ_JMISjMg#v=onepage&q=sentiment%20analysis%20python&f=false