University of Thessaly Department of Electrical and Computer Engineering

Μηχανική Μάθηση – Ανάλυση Συναισθήματος

Machine Learning - Sentiment Analysis

Διπλωματική Εργασία

Φερφέλης Θωμάς

Επιβλέπων

Σταμούλης Γεώργιος

Καθηγητής

Βόλος, Οκτώμβριος 2019

University of Thessaly Department of Electrical and Computer Engineering

Μηχανική Μάθηση – Ανάλυση Συναισθήματος

Machine Learning - Sentiment Analysis

Διπλωματική Εργασία

Φερφέλης Θωμάς

Επιβλέποντες:

Σταμούλης Γεώργιος

Παπαδάκης Νικόλαος

Τσουκαλάς Ελευθέριος

Βόλος, Οκτώμβριος 2019

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ. Γεώργιο Σταμούλη για την στήριξη, τις γνώσεις και την καθοδήγηση που μου παρείχε καθόλη την διάρκεια της πτυχιακής εργασίας βοηθώντας με να ξεπεράσω όλα τα εμπόδια που δημιουργούνταν.

Επιπλέον θα ήθελα να ευχαριστήσω όλους τους καθηγητές της σχολής μας για την προσφορά τους στο τμήμα αλλά και στην εκπαίδευση γενικότερα.

Τέλος, δεν θα μπορούσα να ξεχάσω την οικογένεια μου για την στήριξη που μου παρείχε καθόλη την διάρκεια των σπουδών μου.

To my Family and Friends

Περίληψη

Σε αυτήν την εργασία θα ερευνήσουμε την χρήση της μηχανικής μάθησης για την ανάλυση συναισθήματος σε ένα dataset που περιέχει κριτικές προϊόντων του Amazon. Αυτό το πετυχαίνουμε με την διαμόρφωση του κειμένου εισαγωγής κάθε κριτικής αξιολογώντας τους λεξικούς πόρους που έχουμε στην διάθεσή μας και τα χαρακτηριστικά τους.

Έπειτα χρησιμοποιούμε classification για την κατηγοριοποιήση, την αναγνώριση και την διαφοροποίση των δεδομένων ανάλογα με το συναίσθημα που παράγουν. Σε αυτό το σημείο κάνουμε χρήση τεσσάρων αλγορίθμων.

Τέλος τους συγκρίνουμε με βάση την ακρίβεια και παρουσιάζουμε τον καλύτερο για την προσσέγγιση του συγκεκριμένου dataset.

Abstract

In this paper we will explore the use of machine learning for sentiment analysis in a dataset containing Amazon product reviews. We do this by formulating the introduction text of each review by evaluating the lexical resources at our disposal and their features. Then we use the classification method to categorize, identify and differentiate the data according to the feeling they produce. At this point we use four algorithms. Finally, we compare them based on accuracy and present the best one for approximating this dataset.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Η εξόρυξη δεδομένων είναι μια ευρέως διαδεδομένη διαδικάσία στις μέρες μας. Μέσω αυτής βρίσκουμε πληροφορίες από ποικίλες βάσεις δεδομένων με την βοήθεια αλγορίθμων ομαδοποίησης ή κατηγοριοποίησης και τις αρχές της στατιστικής, της τεχνητής νοημοσύνης και της μηχανικής μάθησης. Στόχος της είναι η πληροφορία που θα εξαχθεί να είναι κατανοητή από τον άνθρωπο ώστε να πάρει τις σωστές αποφάσεις ανάλογα με την έρευνα που διεξάγει. Γενικότερα χρησιμοποιείται από πολλούς επιστημονικούς τομείς όπως η ιατρική, η οικονομία και οι τηλεπικοινωνίες.

Τα τελευταία χρόνια τα ηλεκτρονικά καταστήματα έχουν ασκήσει μεγάλη επιρροή στις επιλογές των καταναλωτών. Έτσι πολλές επιχειρήσεις αξιοποιώντας αυτό το φαινόμενο προσπαθούν να προωθήσουν ενεργά τους καταναλωτές να ‘διαδώσουν την λέξη’ για τα προϊόντα τους. Η επιρροή που ασκείται λόγω των κριτικών όσον αφορά τις πωλήσεις είναι πολύ μεγάλη γεγονός που οδήγησε πολλούς να ασχοληθούν με το πεδίο εκμετάλλευσής τους.

Το αντικείμενο της συγκεκριμένης πτυχιακής εργασίας αναφέρεται σε δεδομένα κειμένου που αποτελούνται απο κριτικές προϊόντων του Amazon. Για να πετύχουμε την ανάλυσή τους χρησιμοποιήσαμε αλγορίθμους μηχανικής μάθησης. Πιο συγκεκριμένα χρησιμοποιήσαμε τους Gradient Boosting, Multinomial Logistic Regression, Multinomial Naïve Bayes και Random Forest. Έπειτα βάσει της ακρίβειας του

καθενός αποφασίσαμε τον αποτελεσματικότερο για τα συγκεκριμένα δεδομένα.

Αναλυτικά τα κεφάλαια της εργασίας έχουν ως εξής:

Κεφάλαιο 1: Αρχικά περιγράφει την διαδικασία και σημασία της εξόρυξης δεδομένων στο πλαίσιο της εισαγωγής. Έπειτα αναφέρεται στην επιρροή των κριτικών στα ηλεκτρονικά καταστήματα όσον αφορά την προώθηση προϊόντων. Τέλος, συγκεκριμενοποιεί το αντικείμενο της έρευνας (Amazon Reviews) παραθέτοντας τους αλγορίθμους κατηγοριοποίησης που χρησιμοποιήσαμε για να πετύχουμε την αποτελεσματικότερη ανάλυση.

Κεφάλαιο 2: Στο συγκεκριμένο κεφάλαιο θα αναλυθούν τρία πολύ σημαντικά βήματα για την διαδικασία της εξόρυξης δεδομένων σε θεωρητική μορφή. Αρχικά θα αναφέρθούμε στο dataset που χρησιμοποιήσαμε και στην περιγραφή του αναλυτικά, ώστε να περάσουμε στο επόμενο βήμα το οποίο είναι η προεπεξεργασία του. Τέλος θα εξηγηθούν και οι μετασχηματισμοί με στόχο να το φέρουμε σε επιθυμητή μορφή για να γίνει η κατηγοριοποίηση.

Κεφάλαιο 3: Στο κεφάλαιο 3 θα γίνει θεωρητική αναφορά στην διαδικασία της προεπεξεργασίας. Έπειτα θα δοθεί η υλοποίησης αυτής από εμάς μέχρι να έρθει στην κατάλληλη μορφή για την κατηγοριοποίηση.

Κεφάλαιο 4: Σε αυτό το σημείο θα αναλυθούν θεωρητικά οι τέσσερις αλγόριθμοι κατηγοριοποίησης αλλά και η υλοποίηση τους. Έπειτα θα παρουσιαστεί η απόδοση τους ως προς το test και train accuracy αλλά και ως προς τον confusion matrix.

Κεφάλαιο 5: Στο τελευταίο αυτό μέρος θα εμφανιστεί το επικρατέστερο μοντέλο για την ανάλυση των δεδομένων μας αλλά και οι στόχοι μας για μελλοντική έρευνα πάνω στο sentiment analysis.

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Ανάλυση Σελίδων Πωλήσεων

Amazon

Είναι μια αμερικανική πολυεθνική εταιρεία τεχνολογίας με έδρα το Σιάτλ της Ουάσιγκτον, η οποία επικεντρώνεται στο ηλεκτρονικό εμπόριο, στο cloud computing, στην ψηφιακή ροή και στην τεχνητή νοημοσύνη. Θεωρείται μία από τις εταιρείες τεχνολογίας Big Four μαζί με το Google, την Apple και το Facebook. Η Amazon είναι γνωστή για τη διαταραχή των καθιερωμένων βιομηχανιών μέσω της τεχνολογικής καινοτομίας και της μαζικής κλίμακας. Είναι η μεγαλύτερη αγορά ηλεκτρονικού εμπορίου παγκοσμίως, AI assistant provider και πλατφόρμα υπολογιστικού cloud, όπως μετράται από τα έσοδα και την κεφαλαιοποίηση της αγοράς. Η Amazon είναι η μεγαλύτερη εταιρεία Διαδικτύου από έσοδα στον κόσμο. Είναι ο δεύτερος μεγαλύτερος ιδιωτικός εργοδότης στις Ηνωμένες Πολιτείες και μία από τις πιο πολύτιμες εταιρείες στον κόσμο. Είναι η δεύτερη μεγαλύτερη εταιρεία τεχνολογίας από τα έσοδα. Όλα αυτά τα χρόνια, στα πλαίσια της αναζήτησης ενός ανοιχτού, ευέλικτου και δομημένου συστήματος, πρόσθεσε λειτουργίες όπως Verified Purchase, ψήφους και κριτικές.

Alibaba

Είναι μια κινεζική πολυεθνική εταιρεία χαρτοφυλακίου που ειδικεύεται στο ηλεκτρονικό εμπόριο, το λιανικό εμπόριο, το Διαδίκτυο και την τεχνολογία. Η εταιρεία, που ιδρύθηκε στις 4 Απριλίου 1999 στο Hangzhou, Zhejiang, παρέχει υπηρεσίες διαδικτυακών πωλήσεων από καταναλωτές προς καταναλωτές, επιχειρήσεις προς καταναλωτές και επιχειρηματικές συναλλαγές, καθώς και ηλεκτρονικές υπηρεσίες πληρωμών, μηχανές αναζήτησης αγορών και υπηρεσίες cloud computing. Διαθέτει και εκμεταλλεύεται μια ποικιλία επιχειρήσεων σε όλο τον κόσμο σε πολλούς τομείς και ονομάζεται ως μία από τις πιο θαυμαστές εταιρείες του κόσμου από την Fortune.

eBay

Είναι μια αμερικανική πολυεθνική εταιρία ηλεκτρονικού εμπορίου που εδρεύει στο San Jose της Καλιφόρνια και διευκολύνει τις πωλήσεις μεταξύ καταναλωτών μέσω των ιστοσελίδων τους. Το eBay ιδρύθηκε από τον Pierre Omidyar το φθινόπωρο του 1995 και έγινε μια αξιοσημείωτη ιστορία επιτυχίας της φούσκας dot-com. Το eBay είναι μια επιχείρηση πολλών δισεκατομμυρίων δολαρίων με δραστηριότητες σε περίπου 30 χώρες, από το 2011. Η εταιρεία διαχειρίζεται τον ιστότοπο του eBay, έναν ηλεκτρονικό ιστότοπο δημοπρασιών και αγορών, στον οποίο οι άνθρωποι και οι επιχειρήσεις αγοράζουν και πωλούν μια μεγάλη ποικιλία αγαθών και υπηρεσιών παγκοσμίως. Ο ιστότοπος είναι ελεύθερος να χρησιμοποιηθεί για τους αγοραστές, αλλά οι πωλητές χρεώνουν τα τέλη για την καταχώριση στοιχείων.

Walmart

Είναι μια αμερικανική πολυεθνική εταιρία λιανικής που λειτουργεί μια αλυσίδα υπερκαταστημάτων, εκπτωτικών πολυκαταστημάτων και παντοπωλείων με έδρα το Bentonville του Αρκάνσας. Η εταιρεία ιδρύθηκε από τον Sam Walton το 1962 και ενσωματώθηκε στις 31 Οκτωβρίου 1969. Διαθέτει επίσης και λειτουργεί τις λιανικές αποθήκες του Sam's Club. Από τις 31 Ιουλίου 2019, η Walmart διαθέτει 11.389 καταστήματα και κλαμπ σε 27 χώρες, που λειτουργούν με 55 διαφορετικά ονόματα. Η εταιρεία λειτουργεί με την επωνυμία Walmart στις Ηνωμένες Πολιτείες και τον Καναδά, ως Walmart de México y Centroamérica στο Μεξικό και την Κεντρική Αμερική, ως Asda στο Ηνωμένο Βασίλειο και ως ο όμιλος Seiyu στην Ιαπωνία. Έχει επιχειρήσεις που ανήκουν εξ ολοκλήρου στην Αργεντινή, τη Χιλή, τον Καναδά και τη Νότια Αφρική. Από τον Αύγουστο του 2018, η Walmart κατέχει μόνο μειοψηφική συμμετοχή στην Walmart Brasil, η οποία μετονομάστηκε σε Grupo Big τον Αύγουστο του 2019, με το 20% των μετοχών της εταιρείας και η εταιρεία ιδιωτικών μετοχών Advent International που κατέχει το 80% της ιδιοκτησίας της.

2.2 Amazon Marketplace

Ιστορικά

Το 1994, ο Jeff Bezos ενσωμάτωσε την Amazon. Επιλέγει τη θέση του Σιάτλ λόγω τεχνικού ταλέντου, καθώς η Microsoft βρίσκεται εκεί. Τον Μάιο του 1997, η οργάνωση έγινε δημόσια. Η εταιρεία άρχισε να πωλεί μουσική και βίντεο το 1998, οπότε άρχισε να λειτουργεί διεθνώς με την εξαγορά online πωλητών βιβλίων στο Ηνωμένο Βασίλειο και τη

Γερμανία. Την επόμενη χρονιά, ο οργανισμός πούλησε και άλλα είδη βίντεο, ηλεκτρονικά είδη ευρείας κατανάλωσης, είδη οικιακής βελτίωσης, λογισμικό, παιχνίδια και παιχνίδια. Το 2002, η εταιρεία ξεκίνησε την υπηρεσία Amazon Web Services (AWS), η οποία παρείχε στοιχεία σχετικά με τη δημοτικότητα του ιστοτόπου, τα πρότυπα κίνησης στο Internet και άλλα στατιστικά στοιχεία για τους εμπόρους και τους προγραμματιστές. Το 2006, ο οργανισμός αύξησε το χαρτοφυλάκιο AWS, όταν διατέθηκε Elastic Compute Cloud (EC2), η οποία ενοικιάζει την εξουσία επεξεργασίας υπολογιστών, καθώς και η απλή υπηρεσία αποθήκευσης (S3), η οποία μισθώνει την αποθήκευση δεδομένων μέσω του Διαδικτύου. Την ίδια χρονιά, η εταιρεία ξεκίνησε την εκπλήρωση από την Amazon, η οποία διαχειρίστηκε την απογραφή των ατόμων και των μικρών εταιρειών που πουλούσαν τα υπάρχοντά τους μέσω της ιστοσελίδας της. Το 2012, η Amazon αγόρασε την Kiva Systems για να αυτοματοποιήσει τις δραστηριότητές της διαχείρισης αποθεμάτων, αγοράζοντας την αλυσίδα σουπερμάρκετ της Whole Foods Market πέντε χρόνια αργότερα το 2017.

Προϊόντα και Υπηρεσίες

Οι σειρές προϊόντων της Amazon.com που διατίθενται στην ιστοσελίδα της περιλαμβάνουν διάφορα μέσα (βιβλία, DVD, μουσικά CD, βιντεοκασέτες και λογισμικό), είδη ένδυσης, προϊόντα για βρέφη, ηλεκτρονικά είδη ευρείας κατανάλωσης, προϊόντα ομορφιάς, γκουρμέ φαγητά και είδη παντοπωλείου. Η Amazon διαθέτει ξεχωριστές ιστοσελίδες λιανικής για ορισμένες χώρες και προσφέρει επίσης τη διεθνή ναυτιλία ορισμένων προϊόντων της σε ορισμένες άλλες χώρες.

Κριτικές

Το Amazon επιτρέπει στους χρήστες να υποβάλλουν κριτικές στην ιστοσελίδα κάθε προϊόντος. Οι αξιολογητές πρέπει να βαθμολογούν το προϊόν σε μια κλίμακα αξιολόγησης από ένα έως πέντε αστέρια. Το Amazon παρέχει μια επιλογή αποδείξεων για τους αναθεωρητές που υποδεικνύουν το πραγματικό όνομα του κριτικού (βάσει επιβεβαίωσης λογαριασμού πιστωτικής κάρτας) ή που δείχνουν ότι ο κριτικός είναι ένας από τους κορυφαίους κριτές κατά δημοτικότητα. Οι πελάτες μπορούν να σχολιάσουν ή να ψηφίσουν σχετικά με τις αναθεωρήσεις, υποδεικνύοντας εάν τους βοήθησε μια κριτική. Αν μια κριτική δίνεται αρκετές "helpful" hits, εμφανίζεται στην πρώτη σελίδα του προϊόντος. Το 2010, η Amazon αναφέρθηκε ως η μεγαλύτερη μοναδική πηγή των αναθεωρήσεων των καταναλωτών στο Διαδίκτυο. Υπήρξαν περιπτώσεις θετικών αναθεωρήσεων που γράφονται και δημοσιεύονται από εταιρείες δημοσίων σχέσεων για λογαριασμό των πελατών τους και περιπτώσεις συγγραφέων που χρησιμοποιούν ψευδώνυμα για να αφήσουν αρνητικές κριτικές για τα έργα των αντιπάλων τους.

2.3 Εξόρυξη Δεδομένων

Η εξόρυξη δεδομένων είναι η διαδικασία της ανεύρεσης προτύπων σε μεγάλα σύνολα δεδομένων που περιλαμβάνουν μεθόδους στη διασταύρωση της μηχανικής μάθησης, των στατιστικών και των συστημάτων βάσης δεδομένων (Fig 2.1). Είναι ένας διεπιστημονικός υποτομέας της επιστήμης των υπολογιστών και των στατιστικών με γενικό στόχο την εξαγωγή πληροφοριών (με έξυπνες μεθόδους) από ένα σύνολο δεδομένων και τη μετατροπή των πληροφοριών σε κατανοητή

δομή για περαιτέρω χρήση. Εκτός από το στάδιο της πρώτης ανάλυσης, περιλαμβάνει επίσης πτυχές διαχείρισης βάσεων δεδομένων και δεδομένων, προεπεξεργασία δεδομένων, εκτιμήσεις μοντέλων και συμπερασμάτων, μετρήσεις ενδιαφερόντων, εκτιμήσεις πολυπλοκότητας, μετα-επεξεργασία ανακαλυφθέντων δομών, οπτικοποίηση και ηλεκτρονική ενημέρωση.

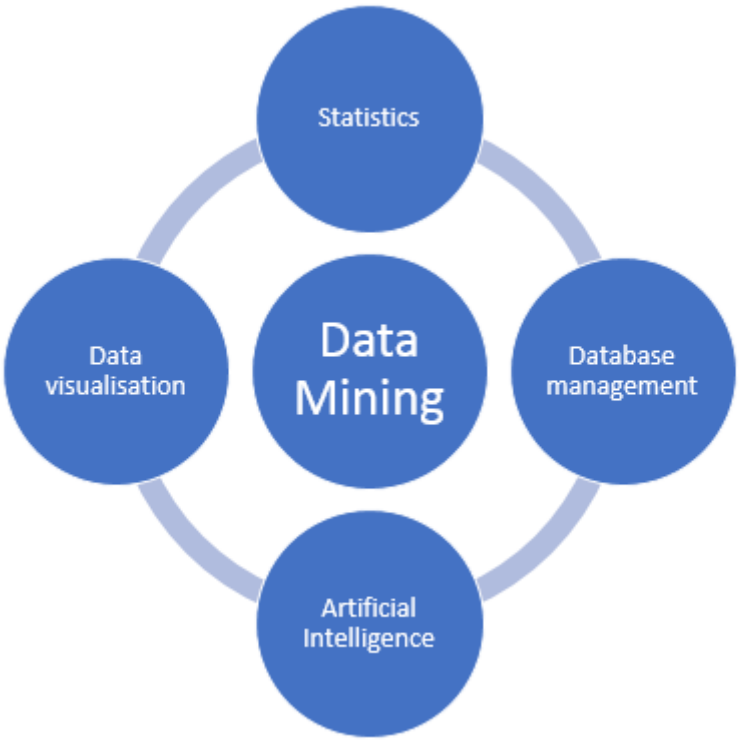

Fig 2.1

2.3.1 KDD Process

Ο KDD [1], αναφέρεται στην ευρεία διαδικασία εξεύρεσης γνώσης στα δεδομένα και τονίζει την εφαρμογή υψηλού επιπέδου συγκεκριμένων μεθόδων εξόρυξης δεδομένων (Fig 2.2). Έχει ενδιαφέρον για τους ερευνητές στη μηχανική μάθηση, την αναγνώριση προτύπων, τις βάσεις δεδομένων, τις στατιστικές, την τεχνητή νοημοσύνη, την απόκτηση γνώσεων για συστήματα εμπειρογνωμόνων και την οπτικοποίηση δεδομένων.

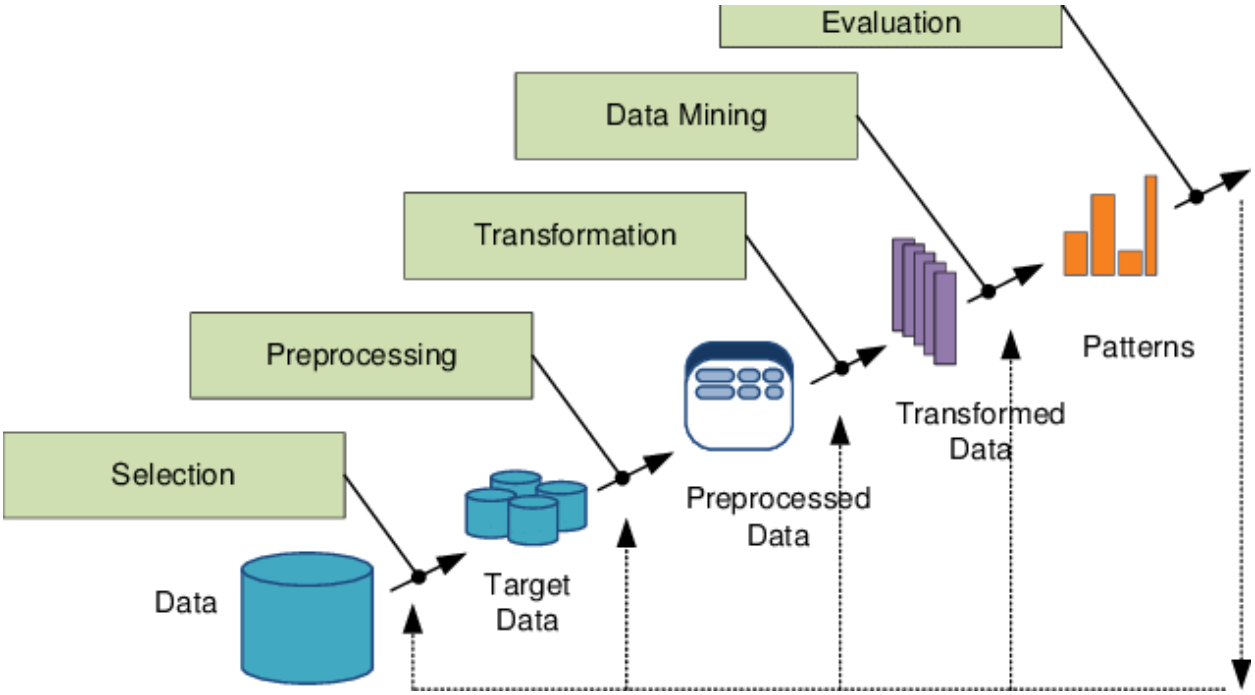

Fig 2.2

Αναλυτικότερα τα στάδια είναι τα εξής:

1) Data Cleaning και Preprocessing: Είναι η απαλειφή των θορυβώδων και άσχετων δεδομένων.
    - Missing Values
    - Θορυβώδη δεδομένα με random ή variant error
    - Με εργαλεία μετασχηματισμού δεδομένων ή με ανίχνευση

2) Data Selection: Ορίζεται ως η διαδικασία όπου δεδομένα σχετικά με την ανάλυση επιλέγονται και ανακτώνται από την συλλογή δεδομένων.

3) Data Transformation: Είναι η διαδικασία όπου μετατρέπουμε τα δεδομένα σε κατάλληλη μορφή για την εξόρυξη.

4) Data Mining: Έξυπνες τεχνικές που εφαρμόζονται για την εξαγωγή χρήσιμων προτύπων.
    - Classification
    - Characterization

5) Pattern Evaluation: Ο εντοπισμός των αυξανόμενων προτύπων που αντιπροσωπεύουν τις γνώσεις βάσει συγκεκριμένων μέτρων.
    - Score για κάθε pattern
    - Σύνοψη με σκοπό την κατανόηση από τον χρήστη

6) **Use of Knowledge**

2.3.2 Εξόρυξη δεδομένων σε Supervised Data

Η supervised μάθηση είναι ένα task μηχανικής μάθησης για την εκμάθηση μιας συνάρτησης που χαρτογραφεί μια είσοδο σε μια έξοδο βασισμένη σε παραδείγματα ζευγών εισόδου-εξόδου. Συγκεντρώνει μια συνάρτηση από τα supervised data κατάρτισης που αποτελούνται από ένα σύνολο εκπαιδευτικών παραδειγμάτων. Ένας supervised αλγόριθμος εκμάθησης αναλύει τα δεδομένα εκπαίδευσης και παράγει μια συναγόμενη συνάρτηση, η οποία μπορεί να χρησιμοποιηθεί για χαρτογράφηση νέων παραδειγμάτων. Ένα βέλτιστο σενάριο θα επιτρέψει στον αλγόριθμο να προσδιορίσει σωστά τις ετικέτες της κλάσης για αθέατες περιπτώσεις. Αυτό απαιτεί τον αλγόριθμο μάθησης να γενικεύεται από τα δεδομένα εκπαίδευσης σε αόρατες καταστάσεις με έναν "λογικό" τρόπο.

Supervised Technics and Algorithms

Κατηγοριοποίηση (Classification) : είναι μια διαδικασία στην οποία οι ιδέες και τα αντικείμενα αναγνωρίζονται, διαφοροποιούνται και κατανοούνται. (Random Forest, Naïve Bayes, Logistic Regression) (Fig 2.3)

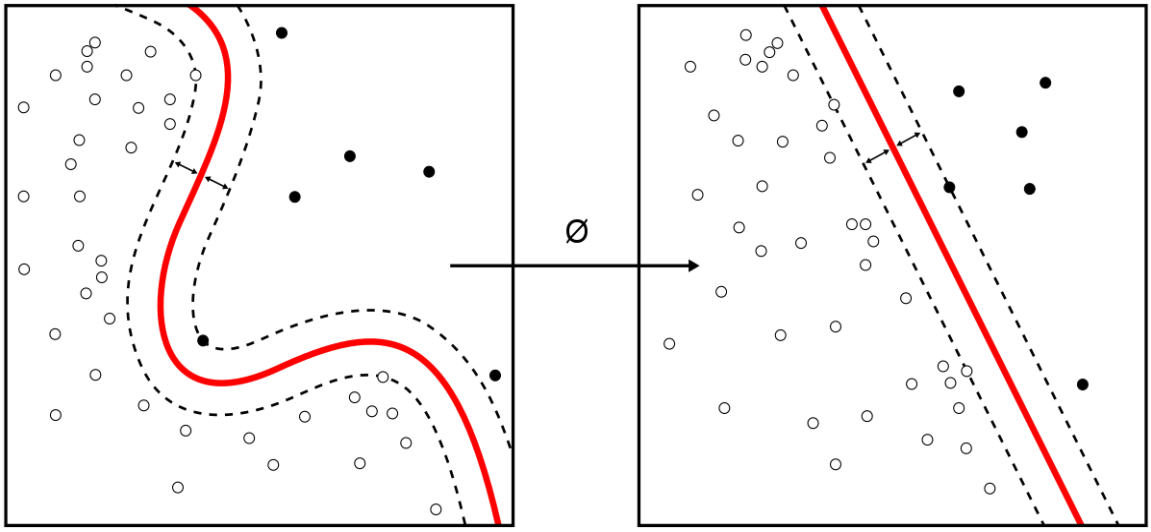

Fig 2.3

Οπισθοδρόμηση (Regression) : Υπολογίζει την εξαρτώμενη προσδοκία της εξαρτημένης μεταβλητής λαμβάνοντας υπόψη τις ανεξάρτητες μεταβλητές – δηλαδή τη μέση τιμή της εξαρτημένης μεταβλητής όταν οι ανεξάρτητες μεταβλητές είναι σταθερές. (Linear-Polynomial Regression) (Fig 2.4)

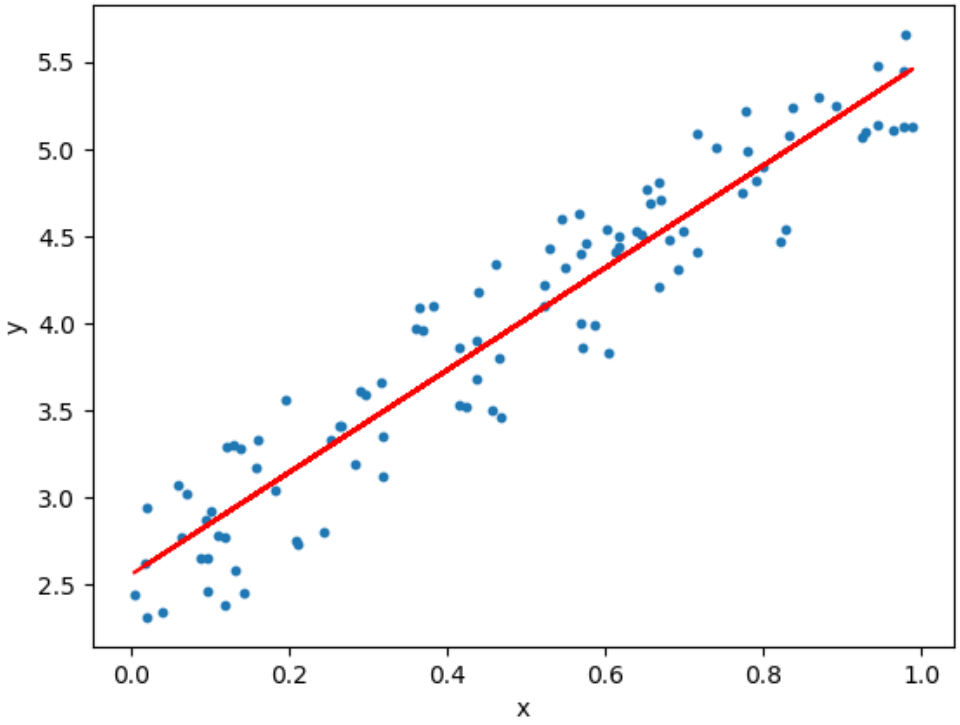

Fig 2.4

2.3.3 Εξόρυξη δεδομένων σε unsupervised data

Η unsupervised μάθηση είναι ένας τύπος αυτο-οργανωμένης εκμάθησης Hebbian που βοηθάει να βρούμε παλιότερα άγνωστα μοτίβα σε σύνολο δεδομένων χωρίς προϋπάρχουσες ετικέτες. Είναι επίσης γνωστή ως αυτο-οργάνωση και επιτρέπει πυκνότητες πιθανότητας μοντελοποίησης συγκεκριμένων εισροών.

Unsupervised Technics and Algorithms

Συσσωμάτωση(Clustering) : Είναι η διαδικασία ομαδοποίησης ενός συνόλου αντικειμένων με τέτοιο τρόπο ώστε τα αντικείμενα της ίδιας ομάδας (που ονομάζεται σύμπλεγμα) να είναι πιο παρόμοια μεταξύ τους (σε κάποια έννοια) με αυτά που ανήκουν σε άλλες ομάδες (συστάδες). (K-means, DBSCAN) (Fig 2.5)

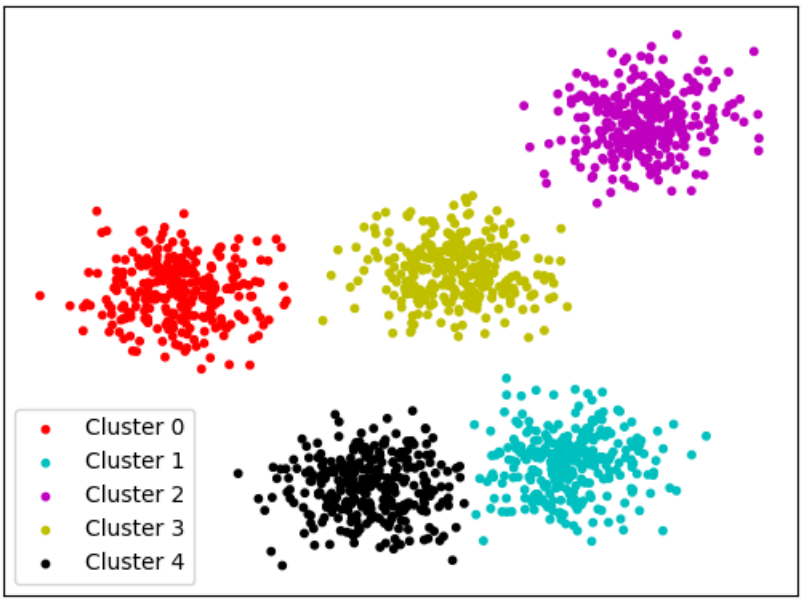

Fig 2.5

Νευρωνικά Δίκτυα(Neural Networks) : Είναι συστήματα υπολογιστών που εμπνέονται από, αλλά όχι ταυτόσημα, βιολογικά νευρωνικά δίκτυα που αποτελούν ζωικό εγκέφαλο. Αυτά τα συστήματα «μαθαίνουν» να εκτελούν εργασίες εξετάζοντας παραδείγματα, γενικά χωρίς να προγραμματίζονται με συγκεκριμένους κανόνες. (Hebbian Learning, Autoencoders) (Fig 2.6)

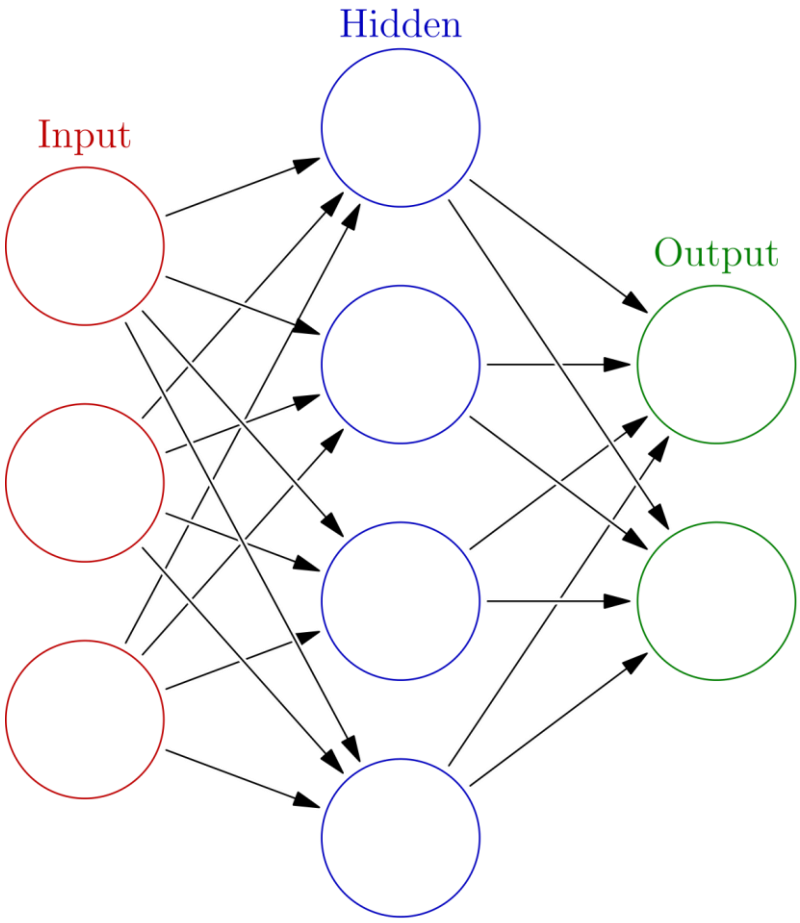

Fig 2.6

2.4 Sentiment Analysis

Το sentiment analysis [2] είναι ένας τομέας της Επεξεργασίας Φυσικής Γλώσσας (NLP) που αναπτύσσει συστήματα που προσπαθούν να εντοπίσουν και να εξάγουν απόψεις μέσα στο κείμενο. Συνήθως, εκτός από την αναγνώριση της γνώμης, αυτά τα συστήματα εξάγουν χαρακτηριστικά της έκφρασης.

Επί του παρόντος, η ανάλυση συναισθημάτων είναι ένα θέμα ιδιαίτερου ενδιαφέροντος και ανάπτυξης, καθώς έχει πολλές πρακτικές εφαρμογές. Δεδομένου ότι οι πληροφορίες που παρέχονται δημόσια και ιδιωτικά μέσω του διαδικτύου αυξάνονται διαρκώς, μεγάλος αριθμός κειμένων που εκφράζουν απόψεις σε review sites, forum, blogs και social media.

Με τη βοήθεια συστημάτων sentiment analysis, αυτή η αδόμητη πληροφορία θα μπορούσε να μετατραπεί αυτόματα σε δομημένα

δεδομένα των απόψεων για προϊόντα, υπηρεσίες, brands, πολιτική ή οποιοδήποτε θέμα που οι άνθρωποι μπορούν να εκφράσουν τις απόψεις τους. Αυτά τα δεδομένα μπορούν να είναι πολύ χρήσιμα για εμπορικές εφαρμογές, όπως η ανάλυση μάρκετινγκ, οι δημόσιες σχέσεις, οι αναθεωρήσεις προϊόντων, η βαθμολόγηση του καθαρού υποκινητή, η ανατροφοδότηση προϊόντων και η εξυπηρέτηση των πελατών.

2.4.1 Τύποι Sentiment Analysis

Ακριβής Ανάλυση Συναισθήματος

Πολλές φορές ενδιαφερόμαστε για μία πιο συγκεκριμένη ανάλυση, οπότε αντί για θετικό,αρνητικό και ουδέτερο μπορούμε να δείξουμε περισσότερες κατηγορίες

1. Πολύ θετικό
2. Θετικό
3. Ουδέτερο
4. Αρνητικό
5. Πολύ αρνητικό

Μερικά συστήματα αναγνωρίζουν και άλλα συναισθήματα όπως θυμός, λύπη, χαρά, ενθουσιασμό και άλλα.

Aspect-Based Sentiment Analysis

Όταν αναλύεις το συναίσθημα για κάποια προϊόντα, μπορεί να ενδιαφερθείς για συγκεκριμένα χαρακτηριστικά του καθενός και όχι για την γενική θετική ή αρνητική εικόνα του. Για παράδειγμα, σε ένα κινητό μπορεί κάποιος να ενδιαφέρεται για την ποιότητα και διάρκεια της

μπαταρίας του και όχι για το σύνολο του σαν συσκευη. Με βάση αυτό το ενδιαφέρον ορίζεται η Aspect-Based Sentiment Analysis.

Multilingual Sentiment Analysis

Απαιτεί πολλή προεπεξεργασία και καταναλώνει αρκετούς πόρους. Οι πόροι είτε ειναι διαθέσιμοι Online (συναισθηματικά λεξικά) είτε χρειάζεται να τους δημιουργήσουμε εμείς (πχ αλγόριθμοι εντοπισμού θορύβου).

2.4.2 Αυτόματες Προσεγγίσεις Αλγορίθμων

Οι αυτόματες προσεγγίσεις βασίζονται σε αλγορίθμους μηχανικής μάθησης. Το κάθε task μοντελοποιείται σε ένα πρόβλημα κατηγοριοποίησης όπου ο classifier λαμβάνει ένα text και ως output εξάγει συναίσθημα.

Εξαγωγή Χαρακτηριστικών από κείμενο

Αναφέρεται στην μετατροπή το κειμένου σε αριθμητική αναπαράσταση, συνήθως διανύσματα. Κάθε διάνυσμα αναπαριστά την συχνότητα μίας λέξης. Αυτή η διαδικασία καθιστά δυνατό για τις λέξεις με ίδια αναπαράσταση να έχουν και ίδιο συναίσθημα πράγμα το οποίο μπορεί να αυξήσει την απόδοση των classifiers.

Training-Prediction Processes

Στη διαδικασία εκπαίδευσης, το μοντέλο μας μαθαίνει να συσχετίζει μια συγκεκριμένη είσοδο (δηλαδή ένα κείμενο) με την αντίστοιχη έξοδο με

βάση τα test samples που χρησιμοποιούνται για την εκπαίδευση. Ο extractor χαρακτηριστικών μεταφέρει την εισαγωγή κειμένου σε ένα διάνυσμα χαρακτηριστικών. Ζεύγη φορέων χαρακτηριστικών και ετικετών (π.χ. θετικών, αρνητικών ή ουδέτερων) τροφοδοτούνται στον αλγόριθμο μηχανικής μάθησης για τη δημιουργία ενός μοντέλου. Στη διαδικασία πρόβλεψης, ο extractor χρησιμοποιείται για να μετασχηματίσει τα text inputs σε διανύσματα χαρακτηριστικών. Αυτά τα διανύσματα χαρακτηριστικών τροφοδοτούνται στη συνέχεια στο μοντέλο, το οποίο δημιουργεί προβλεπόμενες ετικέτες (πάλι, θετικές, αρνητικές ή ουδέτερες).

Sentiment Analysis Metrics and Evaluation

Υπάρχουν πολλοι τρόποι για την λήψη απόδοσης και κατανόηση του πόσο ακριβές ένα μοντέλο είναι. Ο πιο διαδεδομένος είναι το cross-validation.

Η διαδικασία του είναι αρκετά απλή. Αυτό που κάνει είναι να χωρίζει τα δεδομένα σε 2 κατηγορίες (train,test). Το train αποτελείται από το 75-80% των δεδομένων και χρησιμοποιείται για να εξασκήσει τον classifier. Το test αποτελείται από το 20-25% των ίδιων δεδομένων και χρησιμοποιείται για να λάβουμε μετρικές απόδοσης. Αυτή η διαδικασία γίνεται πολλές φορές και υπολογίζεται ο μέσος όρος κάθε μετρικής.

Αν το testing set είναι συνέχεια το ίδιο μπορεί η ανάλυση σου να έχει ρυθμιστεί τόσο πολύ σε ένα data set ώστε είναι πιθανό να μην μπορείς να αναλύσεις διαφορετικό σετ. Το Cross-validation βοηθάει συγκεκριμένα τέτοια περίπτωση.

Precision, Recall, and Accuracy

Είναι μετρικές [3] για να υπολογίσουμε την απόδοση ενός classifier.

- Το Precision μετράει πόσα κείμενα είχαν προβλεφθεί σωστά ως ανήκοντα σε μια δεδομένη κατηγορία από όλα τα κείμενα που είχαν προβλεφθεί (σωστά και εσφαλμένα) ως μέλη της κατηγορίας.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Το Recall μετράει πόσα κείμενα είχαν προβλεφθεί σωστά ως ανήκοντα σε μια δεδομένη κατηγορία από όλα τα κείμενα που θα έπρεπε να είχαν προβλεφθεί ως ανήκοντα στην κατηγορία. Γνωρίζουμε επίσης ότι όσο περισσότερα δεδομένα θα διανείμουμε με τους ταξινομητές μας, τόσο καλύτερη θα είναι η ανάκληση.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- Το Accuracy μετράει πόσα κείμενα έχουν προβλεφθεί σωστά (που ανήκουν ή όχι σε μια δεδομένη κατηγορία) από όλα τα κείμενα του corpus.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- F1-score: αυτό είναι μόνο ο αρμονικός μέσος όρος ακρίβειας και ανάκλησης

$$f_1\text{-score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Confusion Matrix

Ο confusion matrix είναι μια σύνοψη των αποτελεσμάτων πρόβλεψης σε ένα πρόβλημα ταξινόμησης. Ο αριθμός σωστών και εσφαλμένων προβλέψεων συνοψίζεται με τις τιμές μετρήσεων και αναλύεται ανά κατηγορία. Παρουσιάζει τους τρόπους με τους οποίους το μοντέλο ταξινόμησης συγχέεται όταν κάνει προβλέψεις. Μας δίνει τη δυνατότητα να δούμε όχι μόνο τα σφάλματα που δημιουργούνται από έναν ταξινομητή αλλά κυρίως τα είδη των σφαλμάτων που γίνονται.

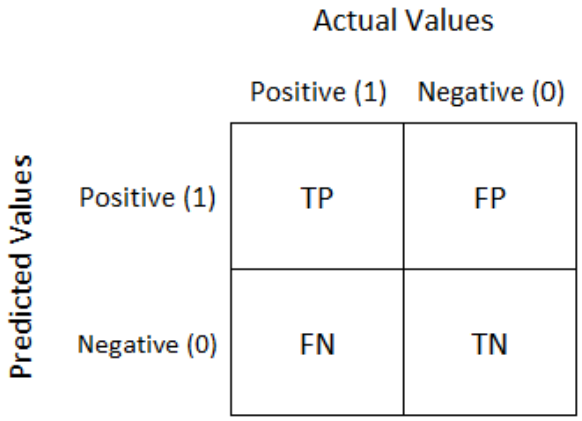

Fig 2.7

TP = True Positive (παρατήρηση θετική/πρόβλεψη θετική)

FP = False Positive (παρατήρηση αρνητική/πρόβλεψη θετική)

FN = False Negative (παρατήρηση θετική/πρόβλεψη αρνητική)

TN = True Negative (παρατήρηση αρνητική/πρόβλεψη αρνητική)

2.4.3 Sentiment Analysis σε Αγορές

Χρησιμότητα

- Ανάλυση κριτικών προϊόντων
- Σύγκριση συναισθήματος με παγκόσμιες αγορές
- Ανάλυση σε social media για real-time γενονότα
- Ανάλυση κριτικών για ανεπιθύμητα σχόλια πελατών

Πλεονεκτήματα

- Νέες πηγές πληροφοριών
- Πρόσθεση της ποιοτικής διάστασης σε ήδη συγκεντρωμένες ποσοτικές γνώσεις.
- Παροχή πληροφοριών σε πραγματικό χρόνο και όχι εκ των υστέρων.
- Συμπλήρωση κενών όπου τα δημόσια δεδομένα είναι σπάνια στις αναδυόμενες αγορές, για παράδειγμα.

Κεφάλαιο 3

Προεπεξεργασία και Υλοποίηση

Σε αυτήν την διπλωματική εργασία θα αναλύσουμε κριτικές πελατών του Amazon για ποικίλα προϊόντα. Στόχος μας είναι να ξεχωρίσουμε τις positive, neutral και negative κριτικές για το κάθε προϊόν. Σαν αποτέλεσμα θα γίνει classification με την χρήση αλγορίθμων ώστε να βρεθεί το καλύτερο μοντέλο αναπαράστασης των αποτελεσμάτων. Για να γίνει όμως αυτό θα πρέπει να ακολουθήσουμε κάποια διαδικασία η οποία περιγράφεται αναλυτικά παρακάτω.

3.1 Exploratory Data Analysis

Το dataset προέρχεται από το Kaggle έχοντας λάβει άδεια για χρήση και επεξεργασία. Η αρχική της μορφής περιλαμβάνει πολλές πληροφορίες όπως φαίνεται παρακάτω (Fig 3.1).

	id	asins	brand	categories	colors	dateAdded	dateUpdated	dimension	ean	keys	...	Score
0	AVpe7AsMilAPnD_xQ78G	B00QJDU3KY	Amazon	Amazon Devices,mazon.co.uk	NaN	2016-03-08T20:21:53Z	2017-07-18T23:52:58Z	169 mm x 117 mm x 9.1 mm	NaN	kindlepaperwhite/b00qjdu3ky	...	5.0
1	AVpe7AsMilAPnD_xQ78G	B00QJDU3KY	Amazon	Amazon Devices,mazon.co.uk	NaN	2016-03-08T20:21:53Z	2017-07-18T23:52:58Z	169 mm x 117 mm x 9.1 mm	NaN	kindlepaperwhite/b00qjdu3ky	...	5.0
2	AVpe7AsMilAPnD_xQ78G	B00QJDU3KY	Amazon	Amazon Devices,mazon.co.uk	NaN	2016-03-08T20:21:53Z	2017-07-18T23:52:58Z	169 mm x 117 mm x 9.1 mm	NaN	kindlepaperwhite/b00qjdu3ky	...	4.0
3	AVpe7AsMilAPnD_xQ78G	B00QJDU3KY	Amazon	Amazon Devices,mazon.co.uk	NaN	2016-03-08T20:21:53Z	2017-07-18T23:52:58Z	169 mm x 117 mm x 9.1 mm	NaN	kindlepaperwhite/b00qjdu3ky	...	5.0

reviews.sourceURLs	reviews.text	reviews.title	reviews.userCity	reviews.userProvince	reviews.username	sizes	upc	weight
https://www.amazon.com/Kindle-Paperwhite-High-...	I initially had trouble deciding between the p...	Paperwhite voyage, no regrets!	NaN	NaN	Cristina M	NaN	NaN	205 grams
https://www.amazon.com/Kindle-Paperwhite-High-...	Allow me to preface this with a little history...	One Simply Could Not Ask For More	NaN	NaN	Ricky	NaN	NaN	205 grams
https://www.amazon.com/Kindle-Paperwhite-High-...	I am enjoying it so far. Great for reading. Ha...	Great for those that just want an e-reader	NaN	NaN	Tedd Gardiner	NaN	NaN	205 grams
https://www.amazon.com/Kindle-Paperwhite-High-...	I bought one of the first Paperwhites	Love / Hate relationship	NaN	NaN	Dougal	NaN	NaN	205 grams

Fig 3.1

Για την εξόρυξη αυτής, κρίθηκε αναγκαία η επεξεργασία της ώστε να κρατήσουμε τις πιο σημαντικές στήλες. Όπως μπορούμε να διαπιστώσουμε οι στήλες που θα μας οδηγήσουν σε μια ακριβή ανάλυση ειναι οι reviews.text και Score.

	reviews.text	Score
0	I initially had trouble deciding between the p...	5.0
1	Allow me to preface this with a little history...	5.0
2	I am enjoying it so far. Great for reading. Ha...	4.0
3	I bought one of the first Paperwhites and have...	5.0
4	I have to say upfront - I don't like coroporat...	5.0

Fig 3.2

Παρατηρήθηκε μια διαφοροποίηση στα Score. Παρόλο που σε κάθε περίπτωση υπάρχει μία κριτική, δεν συμβαίνει το ίδιο και για την στήλη Score (ορισμένα εμφανίζονται ως NaN). Για αυτόν τον λόγο ‘καθαρίσαμε’ σε πρώτη φάση το dataset ώστε σε κάθε κριτική να αντιστοιχεί και μία βαθμολογία και έπειτα διορθώσαμε τα Index. Μετά καταχωρήσαμε ένα συναίσθημα για κάθε βαθμολογία.

Positive = 5,4

Neutral = 3

Negative = 2,1

3.1.1 Rebalancing και Oversampling

Σε αυτό το σημείο παρατηρήσαμε ότι τα θετικά σχόλια υπερτερούσαν κατά πάρα πολύ από τα υπόλοιπα δύο οπότε για λόγους ισορροπίας καταφύγαμε σε μία πολύ σημαντική μέθοδο που ονομάζεται oversampling [4].

Ας υποθέσουμε ότι, για να αντιμετωπιστεί το ζήτημα των διακρίσεων λόγω φύλου, έχουμε δεδομένα της έρευνας σχετικά με τους μισθούς σε ένα συγκεκριμένο τομέα, π.χ. λογισμικό ηλεκτρονικών υπολογιστών. Είναι γνωστό ότι οι γυναίκες υποεκπροσωπούνται σημαντικά σε ένα τυχαίο δείγμα μηχανικών λογισμικού. Ας υποθέσουμε ότι μόνο το 20% των μηχανικών λογισμικού είναι γυναίκες, δηλαδή οι άντρες είναι 4 φορές συχνότεροι από τις γυναίκες. Εάν σχεδιάζαμε μια έρευνα για τη συλλογή δεδομένων, θα εξετάσαμε 4 φορές περισσότερα θηλυκά από τα αρσενικά, έτσι ώστε στο τελικό δείγμα και τα δύο φύλα να εκπροσωπούνται εξίσου.

Ένας τρόπος να καταπολεμηθεί αυτό το ζήτημα είναι να δημιουργηθούν νέα δείγματα στις κατηγορίες που υποεκπροσωπούνται (Fig 3.3).

Random Oversampling

Η τυχαία υπερδειγματοληψία περιλαμβάνει τη συμπλήρωση των δεδομένων εκπαίδευσης με πολλαπλά αντίγραφα ορισμένων από τις τάξεις των μειονοτήτων. Η υπερδειγματοληψία μπορεί να γίνει

περισσότερες από μία φορές (2x, 3x, 5x, 10x, κλπ.). Αυτή είναι μία από τις πιο πρόωρες προτεινόμενες μεθόδους, που αποδείχθηκε επίσης ισχυρή. Αντί να αντιγράψουμε κάθε δείγμα στην τάξη των μειονοτήτων, ορισμένοι από αυτούς μπορεί να επιλέγονται τυχαία με αντικατάσταση.

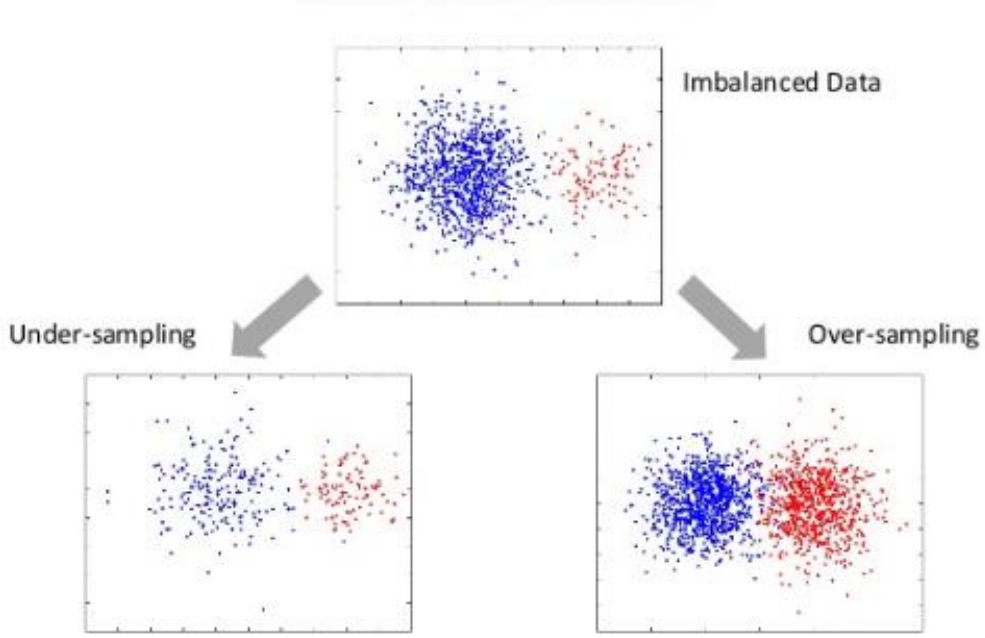

Fig 3.3

Μετά το balance, τα data μετασχηματίστηκαν ως εξης:

Positive	977		Positive	977
Neutral	124	→→→	Neutral	977
Negative	76		Negative	977

3.1.2 Length Κριτικών

Στους παρακάτω γράφους φαίνεται το μέσο μέγεθος των κριτικών ανάλογα με την κατηγορία που κατατάσσονται.

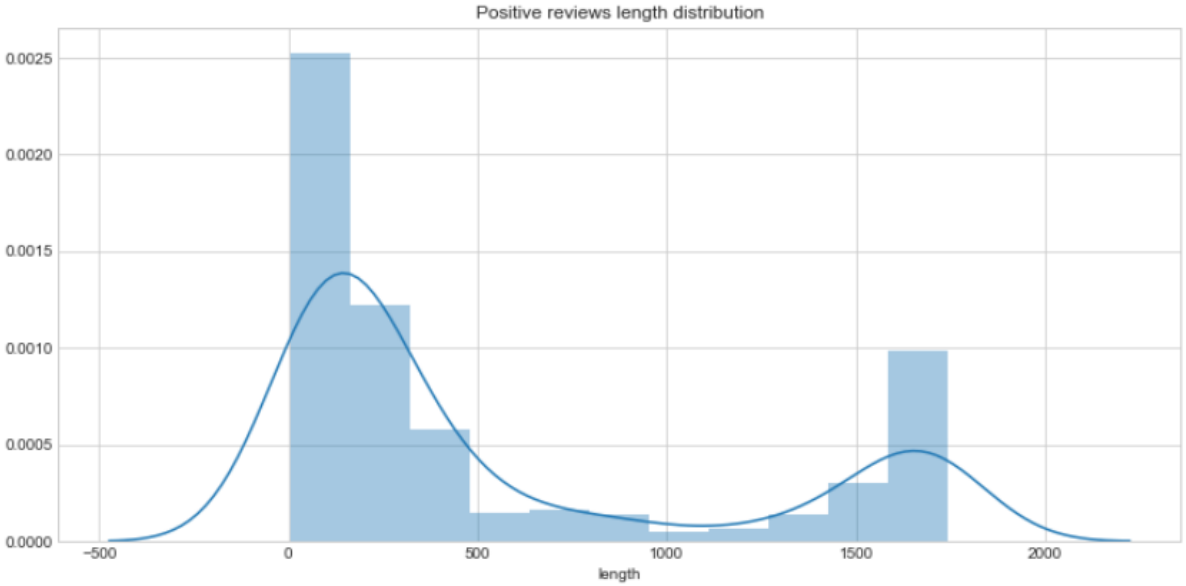

Fig 3.4

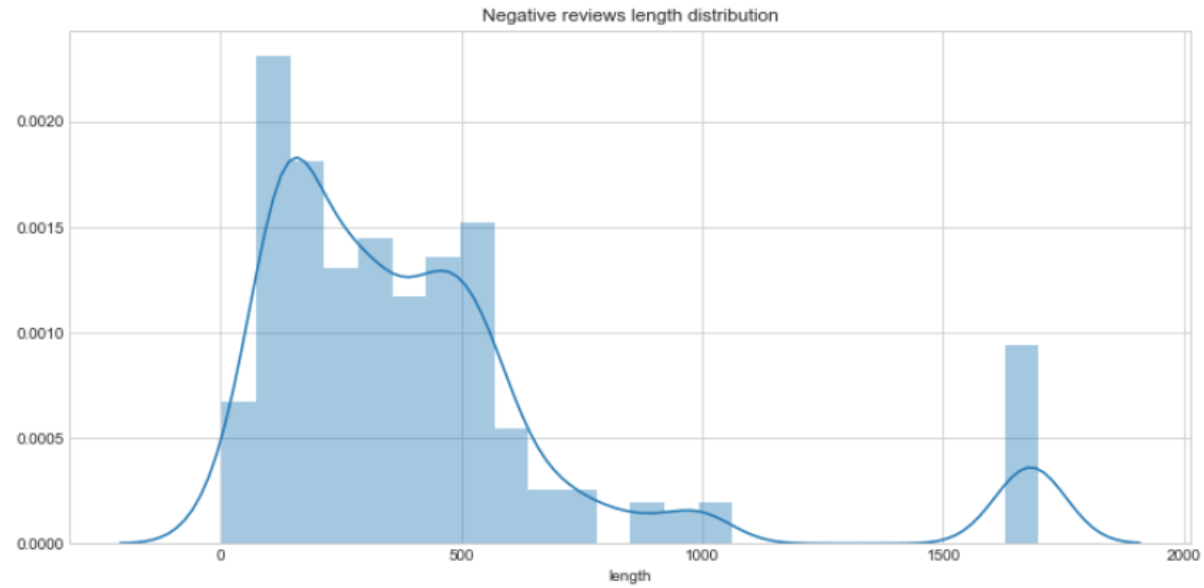

Fig 3.5

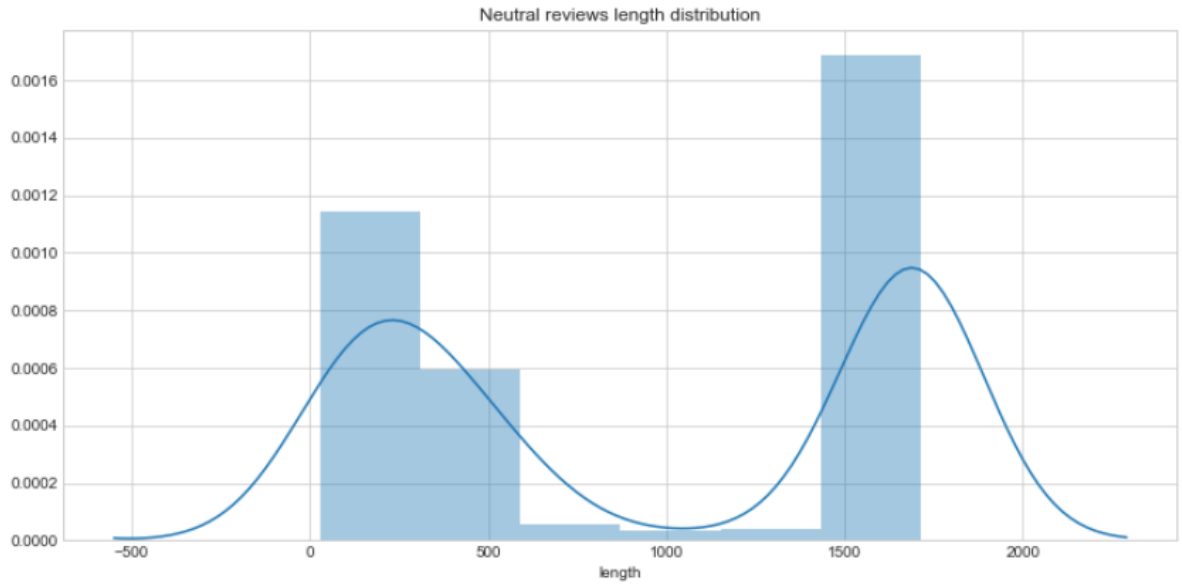

Fig 3.6

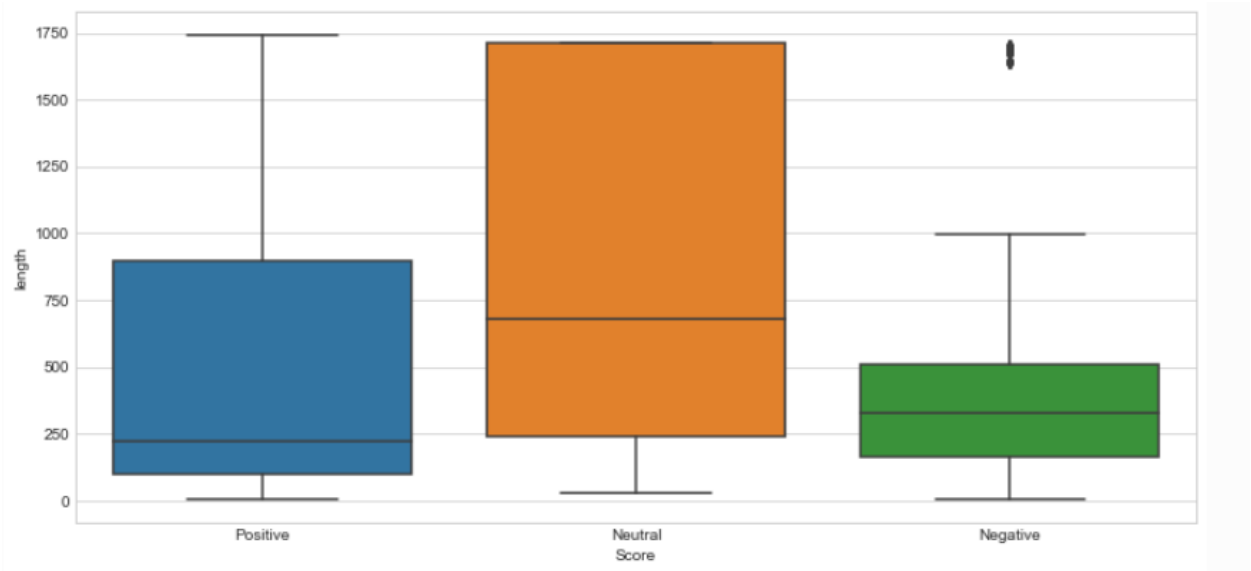

Fig 3.7

Παρατηρείται μια διαφοροποίηση στα μεγέθη των κριτικών ανάλογα με την κατηγορία. Γιαυτό το λόγο θα κανονικοποιήσουμε τα features στην συνέχεια.

3.2 Feature Engineering

Η μέθοδος αναπαράστασης κειμένου είναι γνωστή ως πλαίσιο BOW. Σε αυτό το πλαίσιο, ένα έγγραφο θεωρείται ως μια σακούλα λέξεων και αντιπροσωπεύεται από ένα διάνυσμα χαρακτηριστικό που περιέχει όλες τις λέξεις που εμφανίζονται στο corpus. Αν και το BOW είναι απλή και αρκετά αποτελεσματική μέθοδος στην ταξινόμηση κειμένου, πολλές από τις πληροφορίες από το πρωτότυπο έγγραφο απορρίπτονται, η σειρά λέξεων διαταράσσεται και οι συντακτικές δομές σπάνε. Επομένως, απαιτούνται εξελιγμένες μεθόδοι εξαγωγής χαρακτηριστικών με μια βαθύτερη κατανόηση των εγγράφων για εργασίες sentiment classification.

Η προεπεξεργασία των δεδομένων είναι η διαδικασία καθαρισμού και προετοιμασίας του κειμένου για ταξινόμηση. Τα ηλεκτρονικά κείμενα περιέχουν συνήθως πολύ θόρυβο και μη ενημερωτικά μέρη, όπως ετικέτες HTML, σενάρια και διαφημίσεις. Επιπλέον, σε επίπεδο λέξεων, πολλές λέξεις στο κείμενο δεν επηρεάζουν τον γενικό προσανατολισμό του. Η διατήρηση αυτών των λέξεων καθιστά το dimensionality του προβλήματος υψηλό και ως εκ τούτου την ταξινόμηση πιο δύσκολη, καθώς κάθε λέξη στο κείμενο αντιμετωπίζεται ως μία διάσταση. Εδώ είναι η υπόθεση της σωστής προεπεξεργασίας των δεδομένων: για να μειωθεί ο θόρυβος στο κείμενο, θα πρέπει να βελτιωθεί η απόδοση του ταξινομητή και να επιταχυνθεί η διαδικασία ταξινόμησης, βοηθώντας έτσι την ανάλυση του συναισθηματικού χρόνου. Η όλη διαδικασία περιλαμβάνει διάφορα βήματα: καθαρισμό online κειμένου, απομάκρυνση λευκού χώρου, επέκταση συντομογραφίας, απομάκρυνση λέξεων, άρνηση χειρισμού και τέλος επιλογή χαρακτηριστικών. Όλα τα βήματα αλλά και τα τελευταία ονομάζονται μετασχηματισμοί.

Χαρακτηριστικά στο opinion mining είναι οι λέξεις, οι όροι ή οι φράσεις που εκφράζουν έντονα τη γνώμη ως θετική ή αρνητική [5].

3.2.1 Text Cleaning

Καθαρισμός Ειδικών Χαρακτήρων

Οι ειδικοί χαρακτήρες όπως '\n' πρέπει να αφαιρεθούν από το κείμενο εφόσον δεν περιέχουν καμία δύναμη πρόβλεψης.

```
df['reviews'] = df['reviews.text'].str.replace("\r", " ")
df['reviews'] = df['reviews'].str.replace("\n", " ")
df['reviews'] = df['reviews'].str.replace("    ", " ")
```

executed in 15ms, finished 10:42:59 2019-09-22

```
df['reviews'] = df['reviews'].str.replace('"', '')
```

Upcase/downcase

Περιμένουμε ότι η λέξη 'voyage' και 'Voyage' να έχουν ίδια δύναμη πρόβλεψης. Γιαυτό το λόγο τα μετατρέπουμε όλα σε μικρά.

```
#we downcase cause Thomas and thomas is the same word
df['reviews'] = df['reviews'].str.lower()
```

Σημεία Στίξης

Αφαιρούμε όλα τα σημεία στίξης. ("?", "!", ";")

```
punctuation_signs = list("?:!.,;")
df['reviews'] = df['reviews']

for punct_sign in punctuation_signs:
    df['reviews'] = df['reviews'].str.replace(punct_sign, '')
```

Κτητικά Επίθετα

Περιμένουμε ότι το ‘Thomas’ και το ‘Thomas’s’ να έχουν την ίδια δύναμη πρόβλεψης. Οπότε αφαιρούμε τα ‘s.

```
#we remove possessive pronouns
df['reviews'] = df['reviews'].str.replace("'s", "")
```

Stemming ή Lemmatization

Stemming είναι η διαδικασία της μείωσης των παράγωγων λέξεων στη ρίζα τους. Λεμματοποίηση είναι η διαδικασία μείωσης μιας λέξης στο λέμμα της. Η κύρια διαφορά μεταξύ των δύο μεθόδων είναι ότι η λεμματοποίηση παρέχει υπάρχουσες λέξεις, ενώ το stemming παρέχει τη ρίζα, η οποία μπορεί να μην είναι μια υπάρχουσα λέξη.

Η λεμματοποίηση αναφέρεται συνήθως στη σωστή λειτουργία με τη χρήση λεξιλογίου και μορφολογικής ανάλυσης λέξεων, που συνήθως αποσκοπεί στην απομάκρυνση των τελειωτικών τελειών μόνο και στην επιστροφή της βασικής ή λεξικογραφικής μορφής μιας λέξης, η οποία είναι γνωστή ως λέμμα. Αν αντιμετωπιστεί το token ‘saw’, το stemming θα μπορούσε να επιστρέψει μόνο ‘s’, ενώ η λεμματοποίηση θα επιχειρούσε να επιστρέψει είτε ‘saw’ είτε ‘see’ ανάλογα με το αν η χρήση του συμβόλου ήταν ως ρήμα ή ουσιαστικό. Χρησιμοποιήσαμε ένα Lemmatizer βασισμένο στο WordNet.

Χρησιμοποιήσαμε την πρώτη κριτική για να ελέγξουμε τον μέχρι τώρα καθαρισμό ο οποίος φαίνεται παρακάτω (Fig 3.8).

```
I initially had trouble deciding between the paperwhite and the voyage because reviews more or less said the same thing: the pa
perwhite is great, but if you have spending money, go for the voyage.Fortunately, I had friends who owned each, so I ended up b
uying the paperwhite on this basis: both models now have 300 ppi, so the 80 dollar jump turns out pricey the voyage's page pres
s isn't always sensitive, and if you are fine with a specific setting, you don't need auto light adjustment).It's been a week a
nd I am loving my paperwhite, no regrets! The touch screen is receptive and easy to use, and I keep the light at a specific set
ting regardless of the time of day. (In any case, it's not hard to change the setting either, as you'll only be changing the li
ght level at a certain time of day, not every now and then while reading).Also glad that I went for the international shipping
option with Amazon. Extra expense, but delivery was on time, with tracking, and I didnt need to worry about customs, which I ma
y have if I used a third party shipping service.


Transformed to:


i initially have trouble decide between the paperwhite and the voyage because review more or less say the same thing the paperw
hite be great but if you have spend money go for the voyagefortunately i have friends who own each so i end up buy the paperwhi
te on this basis both model now have 300 ppi so the 80 dollar jump turn out pricey the voyage page press isn't always sensitive
and if you be fine with a specific set you don't need auto light adjustment)it be a week and i be love my paperwhite no regret
the touch screen be receptive and easy to use and i keep the light at a specific set regardless of the time of day (in any case
it not hard to change the set either as you'll only be change the light level at a certain time of day not every now and then w
hile reading)also glad that i go for the international ship option with amazon extra expense but delivery be on time with track
and i didnt need to worry about customs which i may have if i use a third party ship service
```

Fig 3.8

Stop Words

Αυτό το βήμα καθαρισμού εξαρτάται επίσης από το τι τελικά θα κάνουμε με τα δεδομένα μας μετά την προεπεξεργασία. Τα stop words είναι οι λέξεις που χρησιμοποιούνται πολύ συχνά που χάνουν λίγο το σημασιολογικό τους νόημα. Λέξεις όπως “of, are, the, it, is” είναι μερικά παραδείγματα stop words. Σε εφαρμογές όπως μηχανές αναζήτησης εγγράφων και ταξινόμηση εγγράφων, όπου οι λέξεις-κλειδιά είναι πιο σημαντικές από τους γενικούς όρους, η κατάργηση των stop words μπορεί να είναι μια καλή ιδέα. Έχουμε κατεβάσει μια λίστα αγγλικών stop words από το πακέτο nltk και στη συνέχεια τους διαγράψαμε από το corpus.

```
stop_words = list(stopwords.words('english'))
#for example
#i love cooking = example, word = i
#so after the following code in the end we will have output = love cooking
```

executed in 7ms, finished 10:43:27 2019-09-22

```
df['reviews'] = df['reviews']

for stop_word in stop_words:

    regex_stopword = r"\b" + stop_word + r"\b"
    df['reviews'] = df['reviews'].str.replace(regex_stopword, '')
```

```
I initially had trouble deciding between the paperwhite and the voyage because reviews more or less said the same thing: the pa
perwhite is great, but if you have spending money, go for the voyage.Fortunately, I had friends who owned each, so I ended up b
uying the paperwhite on this basis: both models now have 300 ppi, so the 80 dollar jump turns out pricey the voyage's page pres
s isn't always sensitive, and if you are fine with a specific setting, you don't need auto light adjustment).It's been a week a
nd I am loving my paperwhite, no regrets! The touch screen is receptive and easy to use, and I keep the light at a specific set
ting regardless of the time of day. (In any case, it's not hard to change the setting either, as you'll only be changing the li
ght level at a certain time of day, not every now and then while reading).Also glad that I went for the international shipping
option with Amazon. Extra expense, but delivery was on time, with tracking, and I didnt need to worry about customs, which I ma
y have if I used a third party shipping service.


Transformed to:


 initially  trouble decide   paperwhite   voyage  review   less say   thing  paperwhite  great     spend money go   voyagefortu
nately   friends      end  buy  paperwhite   basis  model   300 ppi   80 dollar jump turn  pricey  voyage page press ' always s
ensitive     fine   specific set  ' need auto light adjustment)   week    love  paperwhite  regret  touch screen  receptive  ea
sy  use   keep  light   specific set regardless   time  day (  case   hard  change  set either  '   change  light level   certa
in time  day  every     reading)also glad   go   international ship option  amazon extra expense  delivery   time  track   didn
t need  worry  customs   may    use  third party ship service
```

Fig 3.9

Label Coding

Τα μοντέλα μηχανικής μάθησης απαιτούν αριθμητικά χαρακτηριστικά και ετικέτες για την πρόβλεψη. Για το λόγο αυτό, δημιουργήσαμε ένα λεξικό για τη χαρτογράφηση κάθε ετικέτας σε ένα αριθμητικό αναγνωριστικό.

1→ Positive

2→ Neutral

0→ Negative

```
from sklearn.preprocessing import LabelEncoder
```
executed in 2ms, finished 10:43:37 2019-09-22

```
encoder = LabelEncoder()

df["sentiment"] = encoder.fit_transform(df["Score"])
```

Train-Test Split

Πρέπει να διαχωρίσουμε ένα test set προκειμένου να αποδείξουμε την ποιότητα των μοντέλων μας όταν προβλέπουμε δεδομένα που δεν έχουν δει. Επιλέξαμε έναν τυχαίο διαχωρισμό με το 80% των παρατηρήσεων που συνθέτουν τη δοκιμασία εκπαίδευσης και το 20% των παρατηρήσεων που συνθέτουν το σετ δοκιμών. Θα εκτελέσουμε τη διαδικασία ρύθμισης υπερπαραμέτρων με cross validation στα δεδομένα εκπαίδευσης, θα την προσαρμόσουμε στο τελικό μοντέλο και στη συνέχεια θα την αξιολογήσουμε με εντελώς αόρατα δεδομένα ώστε να αποκτήσουμε μια μέτρηση αξιολόγησης όσο το δυνατόν λιγότερο μεροληπτική.

```
X_train, X_test, y_train, y_test = train_test_split(df['reviews'],
                                                    df['sentiment'],
                                                    test_size=0.20,
                                                    random_state=8,
                                                    shuffle=True)
```

Αντιπροσώπευση κειμένου

- TF-IDF Vectors (Term Frequency-Inverse Document Frequency)

  Είναι μια βαθμολογία που αντιπροσωπεύει τη σχετική σημασία ενός όρου στο έγγραφο και ολόκληρο το corpus [6].

$$TFIDF(t,d) = TF(t,d) \times \log\left(\frac{N}{DF(t)}\right)$$

  - t: term (i.e. a word in a document)
  - d: document
  - $TF(t)$: term frequency (i.e. how many times the term t appears in the document d)
  - N: number of documents in the corpus
  - $DF(t)$: number of documents in the corpus containing the term t

Η τιμή TF-IDF αυξάνεται αναλογικά με τον αριθμό των φορών που εμφανίζεται μια λέξη στο έγγραφο και αντισταθμίζεται από τον αριθμό των εγγράφων στο corpus που περιέχουν τη λέξη, γεγονός που βοηθά να προσαρμόζεται στο γεγονός ότι ορισμένες λέξεις εμφανίζονται πιο συχνά γενικά.

Λαμβάνει επίσης υπόψη το γεγονός ότι ορισμένα έγγραφα μπορεί να είναι μεγαλύτερα από άλλα, κανονικοποιώντας τον όρο TF.

Η μέθοδος TF-IDF Vectors ονομάζεται συχνά μέθοδος Bag of Words, καθώς η σειρά των λέξεων σε μια πρόταση αγνοείται.

Επιλέξαμε την μέθοδο TF-IDF Vectors για να αναπαραστήσουμε τα έγγραφα στο corpus.

- Το TF-IDF είναι ένα απλό μοντέλο που αποδίδει εξαιρετικά αποτελέσματα σε αυτόν τον συγκεκριμένο τομέα.
- Η δημιουργία χαρακτηριστικών TF-IDF είναι μια γρήγορη διαδικασία, η οποία θα μας οδηγήσει σε μικρότερο χρόνο αναμονής για το χρήστη όταν χρησιμοποιούμε την εφαρμογή web.
- Μπορούμε να συντονίσουμε τη διαδικασία δημιουργίας χαρακτηριστικών για να αποφύγουμε ζητήματα όπως η υπερφόρτωση.

Κατά τη δημιουργία των λειτουργιών με αυτήν τη μέθοδο, μπορούμε να επιλέξουμε κάποιες παραμέτρους:

- Εύρος N-gram: είμαστε σε θέση να εξετάσουμε unigrams, bigrams, trigrams
- Max/Min συχνότητα εγγράφου: όταν δημιουργούμε το λεξιλόγιο, μπορούμε να αγνοήσουμε όρους που έχουν μια συχνότητα εγγράφων αυστηρά υψηλότερη / χαμηλότερη από το δεδομένο όριο.
- Μέγιστες λειτουργίες: μπορούμε να επιλέξουμε τις κορυφαίες λειτουργίες N που ταξινομούνται βάσει της συχνότητας όρων σε ολόκληρο το σώμα.

Parameter	Value
N-Gram	(1,2)
Max DF	1
Min DF	10
Max Features	4.500

Min DF: Για να αφαιρέσουμε τις λέξεις που δεν εμφανίζονται σε περισσότερα απο 10 έγγραφα.

Max DF: Για να μην αγνοήσουμε άλλες λέξεις.

Max Features: Για να αποφύγουμε το overfitting λόγω μεγάλου αριθμού features αναλογικά με training παρατηρήσεις.

```
ngram_range = (1,2)
min_df = 10
max_df = 1.
max_features = 4500
```

executed in 4ms, finished 11:01:02 2019-09-22

```
tfidf = TfidfVectorizer(encoding='utf-8',
                        ngram_range=ngram_range,
                        stop_words=None,
                        lowercase=False,
                        max_df=max_df,
                        min_df=min_df,
                        max_features=max_features,
                        norm='l2',
                        sublinear_tf=True)

features_train = tfidf.fit_transform(X_train).toarray()
labels_train = y_train
print(features_train.shape)

features_test = tfidf.transform(X_test).toarray()
labels_test = y_test
print(features_test.shape)
# we transform every word into a vector
```

executed in 436ms, finished 11:01:03 2019-09-22

Τελικό Dataset

	reviews.text	Score	length	reviews	sentiment
0	I initially had trouble deciding between the p...	Positive	1064	initially trouble decide paperwhite voya...	2
1	Allow me to preface this with a little history...	Positive	1424	allow preface little history () casual...	2
2	I am enjoying it so far. Great for reading. Ha...	Positive	182	enjoy far great read original fire sinc...	2
3	I bought one of the first Paperwhites and have...	Positive	1671	buy one first paperwhites please c...	2
4	I have to say upfront - I don't like coroporat...	Positive	1916	say upfront - ' like coroporate hermetical...	2
5	Had older model, that you could text to speech...	Positive	223	older model could text speech one ' like ...	2
6	This is a review of the Kindle Paperwhite laun...	Positive	10670	review kindle paperwhite launch july 2015...	2
7	I love my kindle! I got one for my fiance on h...	Positive	859	love kindle get one fiance birthday lo...	2
8	Vraiment bon petit appareil , lger et facile d...	Positive	157	vraiment bon petit appareil lger et facile e...	2
9	Exactly what it is supposed to be. Works great...	Positive	118	exactly suppose work great love built-...	2
10	Trs heureux que les livres soient sur Icloud. ...	Positive	158	trs heureux que les livres soient sur icloud a...	2
11	Only 4 stars because I found it very confusing...	Positive	470	4 star find confuse first kindle fin...	2
12	Almost like reading a real book. Don't is cris...	Positive	281	almost like read real book ' crisp sharp ea...	2
13	I already had an original Kindle Paperwhite an...	Positive	716	already original kindle paperwhite love b...	2
14	Received my new kindle in perfect condition an...	Positive	205	receive new kindle perfect condition hap...	2
15	I'm a first-time Kindle owner, so I have nothi...	Positive	1661	' first-time kindle owner nothing compare...	2

Fig 3.10

Το τελικό dataset θα έχει αυτήν την μορφή

Index: 0-2930 rows

Reviews.text: Χωρίς Text Cleaning

Score: Βαθμολογία

Length: Μήκος κάθε κριτικής

Reviews: Μετά το Text Cleaning

Sentiment: Μετά το Label encoding

Κεφάλαιο 4

Αλγόριθμοι Κατηγοριοποίησης και Εκτίμηση

4.1 Multinomial Naïve Bayes

Ο MNB κάνει μια απλοποιημένη (naive) υπόθεση για το πώς αλληλεπιδρούν τα χαρακτηριστικά. Ο Naive Bayes είναι ένας πιθανοτικός ταξινομητής , που σημαίνει ότι για ένα έγγραφο d, από όλες τις κατηγορίες c ∈ C , επιστρέφει την κλάση c η οποία έχει το μέγιστο posterior [7].

$$\hat{c} = \underset{c \in C}{arg\,max} P(c|d)$$

Έπειτα, εξετάζουμε τον νόμο του Bayes,

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}$$

P(d): ανεξάρτητη πιθανότητα

P(d|c): πιθανότητα του d δεδομένου ότι το c είναι αληθής

P(c|d): πιθανότητα του c δεδομένου ότι το d είναι αληθή

$$P_{NB} := \frac{P(c)(\prod_{i=1}^{m})P(f_i|c)^{n_i(d)}}{P(d)}$$

- Η μέθοδος train σχετίζεται με την σχετική συχνότητα του P(c) και του P(fi|c)

4.1.1 Απόδοση MNB

- Κάναμε χρήση του MNM καλώντας τον από την βιβλιοθήκη της Python, sklearn.
- Υπολογίσαμε το train και το test accuracy.

```
The training accuracy is:
0.887372013652

The test accuracy is:
0.860306643952
```

- Εμφανίσαμε τον Confusion Matrix με όλες του τις μετρικές.

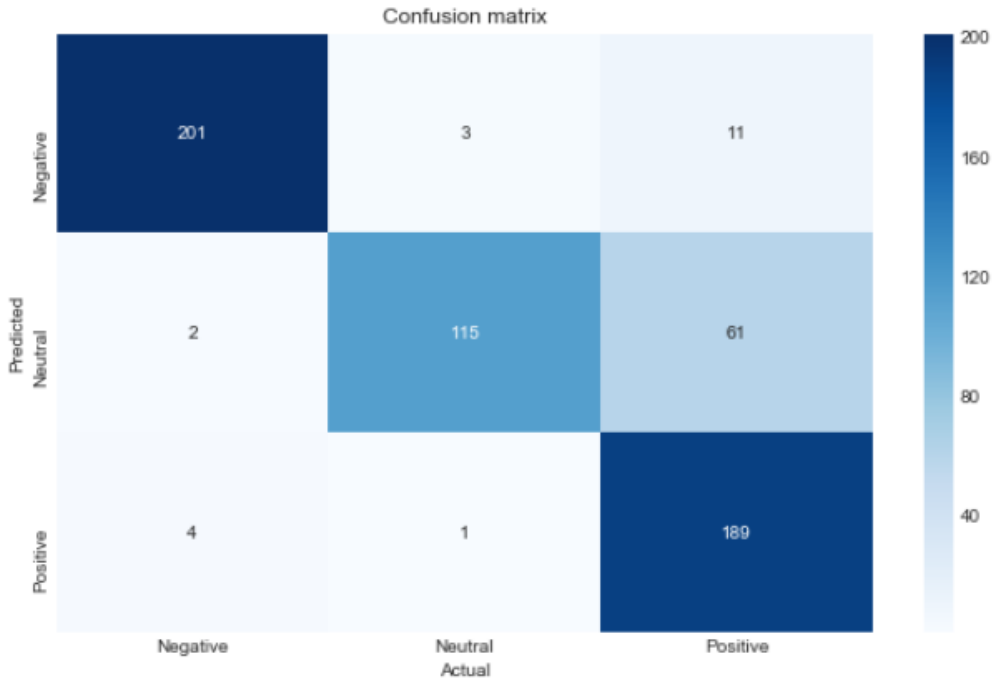

Fig 4.1

4.2 Multinomial Logistic Regression

Χρησιμοποιείται για την πρόβλεψη της κατηγορικής τοποθέτησης ή την πιθανότητα κατηγορίας μέλους σε μια εξαρτημένη μεταβλητή που βασίζεται σε πολλαπλές ανεξάρτητες μεταβλητές. Οι ανεξάρτητες μεταβλητές μπορούν να είναι είτε διχοτομικές είτε συνεχείς. Ο multinomial logistic regression χρησιμοποιεί εκτίμηση μέγιστης πιθανότητας για να αξιολογήσει την πιθανότητα κατηγορικής συμμετοχής [8].

Μοντέλο

Υιοθετούμε την κοινή τεχνική εκπροσώπησης της αναπαράστασης των ετικετών κλάσης χρησιμοποιώντας $\{1, \dots, m\}$ και φορέα κωδικοποίησης $y = [y^{(1)}, \dots, y^{(m)}]^T$. Τα δείγματα εκπαίδευσης n μπορεί να αναπαρασταθούν ως σύνολο δεδομένα εκπαίδευσης
$D = \{(x_1, y_1), \dots\dots, (x_n, y_n)$

Πιθανότητα x να ανήκει σε μία κλάση i

$$P\left(y^{(i)} = 1 | \boldsymbol{x}, \boldsymbol{w}\right) = \frac{\exp\left(\boldsymbol{w}^{(i)^T} \boldsymbol{x}\right)}{\sum_{j=1}^{m} \exp\left(\boldsymbol{w}^{(j)^T} \boldsymbol{x}\right)}$$

Για $i \in \{1, .., m\}$ όπου $w^{(i)}$ είναι ο φορέας βάρους που αντιστοιχεί στην κλάση iκαι ο δείκτης T υποδηλώνει τον vector/matrix transpose.

Για $m > 2$ ενδείκνυται ο MLR

$$\sum_{i=1}^{m} P\left(y^{(i)} = 1 | \boldsymbol{x}, \boldsymbol{w}\right) = 1$$

Λόγω της συνθήκης εξομάλυνσης ο φορέας βάρους για μία από τις κατηγορίες δεν χρειάζεται να υπολογιστεί.

Θέτουμε $w^{(m)} = 0$ και οι μόνες παράμετροι που πρέπει να μάθουμε είναι το βάρος φορείς $w^{(i)}$ για $i \in \{1, \dots, m-1\}$.

Χρησιμοποιούμε το w για να μειώσουμε το $(d(m-1))$ -dimensional διάνυσμα των παραμέτρων που πρέπει να μάθουν.

4.2.1 Απόδοση MLR

- Κάναμε χρήση του MLR καλώντας τον από την βιβλιοθήκη της Python, sklearn.
- Υπολογίσαμε το train και το test accuracy.

```
The training accuracy is:
0.991894197952

The test accuracy is:
0.979557069847
```

- Εμφανίσαμε τον Confusion Matrix με όλες του τις μετρικές.

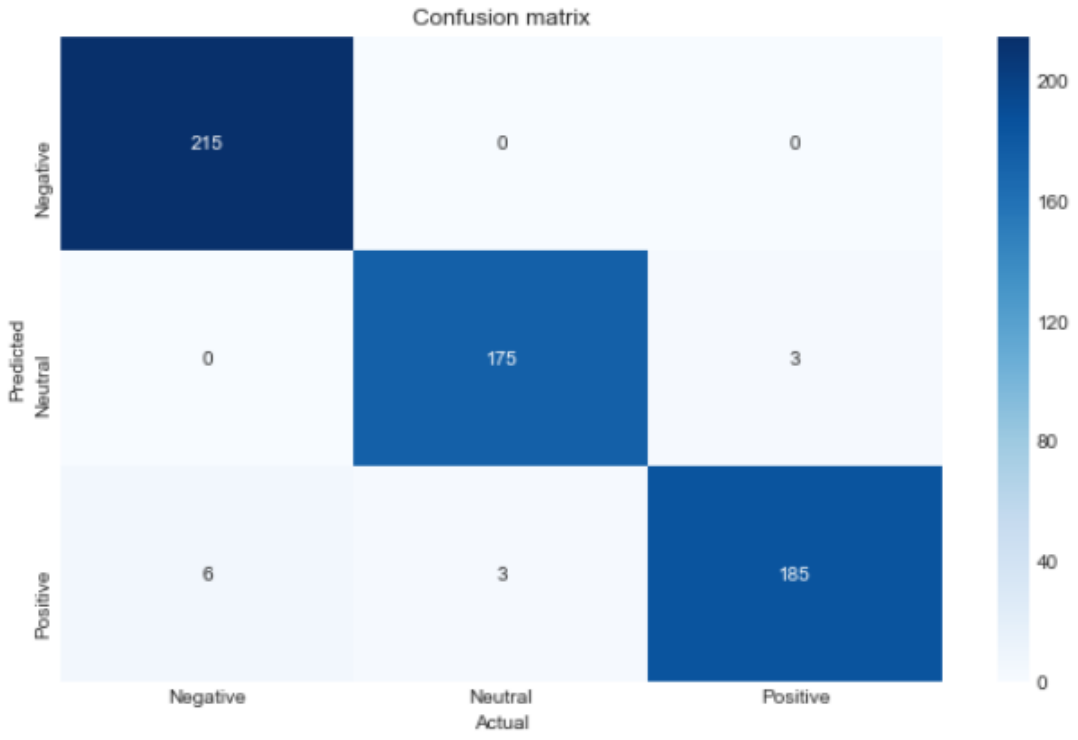

Fig 4.2

4.3 Random Forest

Ο RF αποτελείται από ένα συνδυασμό ταξινομητών δέντρων όπου κάθε ταξινομητής δημιουργείται χρησιμοποιώντας έναν τυχαίο φορέα δειγματοληπτημένο ανεξάρτητα από την είσοδο φορέα και κάθε δέντρο εκπέμπει μια ψηφοφορία μονάδας για την πιο δημοφιλή κατηγορία για να ταξινομήσει μια εισροή διάνυσμα.

Ο RF που χρησιμοποιήθηκε για τη μελέτη αυτή χρησιμοποιεί τυχαία επιλεγμένα χαρακτηριστικά ή συνδυασμό χαρακτηριστικών σε κάθε κόμβο για ανάπτυξη ενός δέντρου. Μία μέθοδος για τη δημιουργία ενός συνόλου δεδομένων εκπαίδευσης με τυχαία σχεδίαση με παραδείγματα αντικατάστασης N, όπου N είναι το μέγεθος του αρχικού σετ εκπαίδευσης, χρησιμοποιήθηκε για κάθε συνδυασμό χαρακτηριστικών που επιλέχθηκε. Ο σχεδιασμός ενός δέντρου απόφασης απαιτεί την επιλογή ενός μέτρου επιλογής χαρακτηριστικών και μια μέθοδο pruning.

Για ένα dataset T το Gini Index μπορεί να γραφτεί:

$$\sum \sum_{j \neq i} (f(C_i, T)/|T|)(f(C_j, T)/|T|)$$

Όπου $f(C_i, T)/|T|$ είναι η πιθανότητα μίας περίπτωσης να ανήκει στην κλάση C_i.

Κάθε φορά που ένα δέντρο καλλιεργείται στο μέγιστο βάθος σε νέα δεδομένα εκπαίδευσης χρησιμοποιώντας ένα συνδυασμός χαρακτηριστικών. Αυτά τα πλήρως αναπτυγμένα δέντρα δεν γίνονται prune.
Έτσι, ο RF αποτελείται από N δέντρα, όπου N είναι ο αριθμός των δέντρων που καλλιεργούνται, τα οποία μπορούν να είναι οποιαδήποτε τιμή ορίζεται από το χρήστη. Για να ταξινομήσετε ένα νέο σύνολο

δεδομένων, κάθε περίπτωση των συνόλων δεδομένων μεταφέρεται σε καθένα από τα N δέντρα. Το δάσος επιλέγει μια κλάση που έχει τα μέγιστα των ψήφων N, για την περίπτωση αυτή [9].

4.3.1 Απόδοση RF

- Κάναμε χρήση του RF καλώντας τον από την βιβλιοθήκη της Python, sklearn.
- Υπολογίσαμε το train και το test accuracy.

```
The training accuracy is:
0.999146757679

The test accuracy is:
0.984667802385
```

- Εμφανίσαμε τον Confusion Matrix με όλες του τις μετρικές.

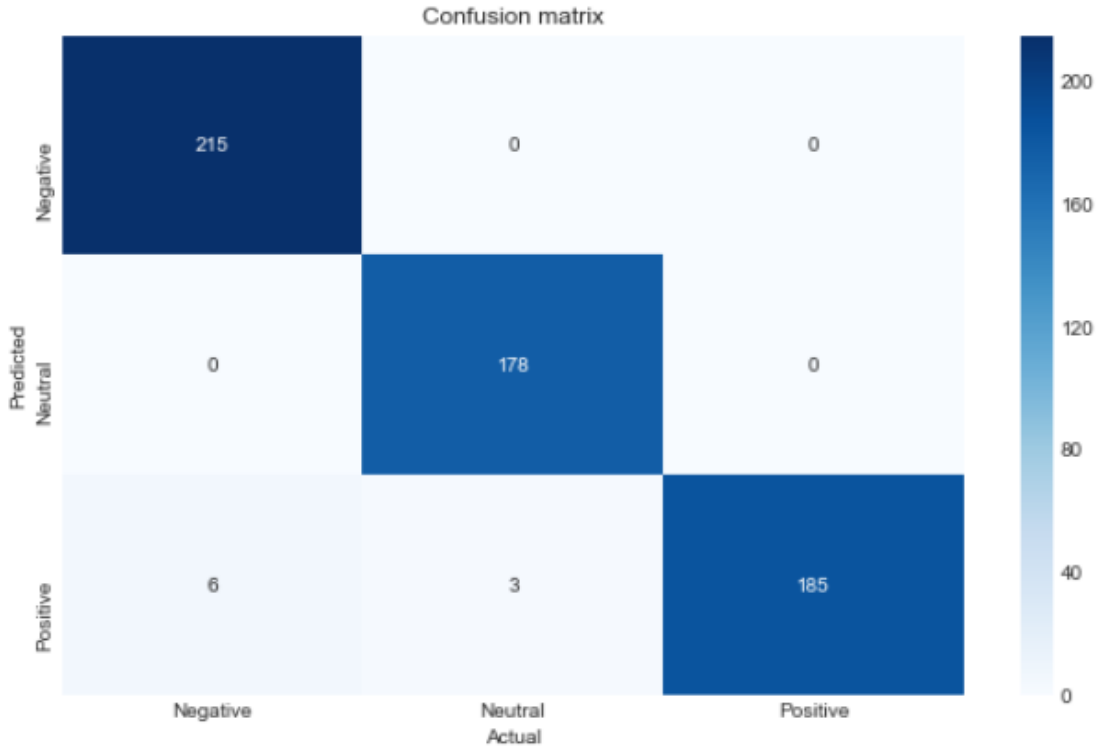

Fig 4.3

4.4 Gradient Boosting Machine

Το Boosting είναι μια μέθοδος μετατροπής των αδύναμων μαθητών σε ισχυρούς μαθητές. Στην ενίσχυση, κάθε νέο δέντρο είναι προσαρμοσμένο σε μια τροποποιημένη έκδοση του αρχικού συνόλου δεδομένων. Ο GBM μπορεί να εξηγηθεί εύκολα με την χρήση του αλγόριθμου AdaBoost. Ο αλγόριθμος AdaBoost ξεκινά με την εκπαίδευση ενός δέντρου αποφάσεων στο οποίο σε κάθε παρατήρηση αποδίδεται ίσο βάρος. Μετά την αξιολόγηση του πρώτου δένδρου, αυξάνουμε τα βάρη εκείνων των παρατηρήσεων που είναι δύσκολο να ταξινομηθούν και να μειώνουμε τα βάρη για εκείνων που είναι εύκολο να ταξινομηθούν. Κατά συνέπεια, το δεύτερο δέντρο αναπτύσσεται σε αυτά τα σταθμισμένα δεδομένα. Εδώ, η ιδέα είναι να βελτιωθούν οι προβλέψεις του πρώτου δέντρου. Το νέο μας μοντέλο είναι ως εκ τούτου Δέντρο 1 + Δέντρο 2. Στη συνέχεια υπολογίζουμε το σφάλμα ταξινόμησης από αυτό το νέο μοντέλο 2-δένδρου και δημιουργούμε ένα τρίτο δέντρο για να προβλέψουμε τα αναθεωρημένα υπολείμματα. Επαναλαμβάνουμε αυτή τη διαδικασία για έναν συγκεκριμένο αριθμό επαναλήψεων. Τα επόμενα δέντρα μας βοηθούν να ταξινομήσουμε τις παρατηρήσεις που δεν είναι καλά ταξινομημένες από τα προηγούμενα δέντρα. Οι προβλέψεις του τελικού μοντέλου του συνόλου είναι επομένως το σταθμισμένο άθροισμα των προβλέψεων που προέκυψαν από τα προηγούμενα μοντέλα δέντρων.
Το Gradient Boosting εκπαιδεύει πολλά μοντέλα με βαθμιαίο, προσθετικό και διαδοχικό τρόπο. Εντοπίζει τις ελλείψεις χρησιμοποιώντας Gradient loss function. Η loss function είναι ένα μέτρο που υποδεικνύει πόσο καλά είναι οι συντελεστές του μοντέλου για την τοποθέτηση των υποκείμενων δεδομένων [10]. Ένα από τα μεγαλύτερα κίνητρα για τη χρήση του GBM είναι ότι επιτρέπει σε κάποιον να βελτιστοποιήσει μια συνάρτηση κόστους που καθορίζεται από το

χρήστη, αντί για μια λειτουργία απώλειας που συνήθως προσφέρει λιγότερο έλεγχο και δεν ανταποκρίνεται ουσιαστικά στις εφαρμογές του πραγματικού κόσμου.

4.4.1 Απόδοση GBM

- Κάναμε χρήση του GBM καλώντας τον από την βιβλιοθήκη της Python, sklearn.
- Υπολογίσαμε το train και το test accuracy.

```
The training accuracy is:
0.992320819113

The test accuracy is:
0.962521294719
```

- Εμφανίσαμε τον Confusion Matrix με όλες του τις μετρικές.

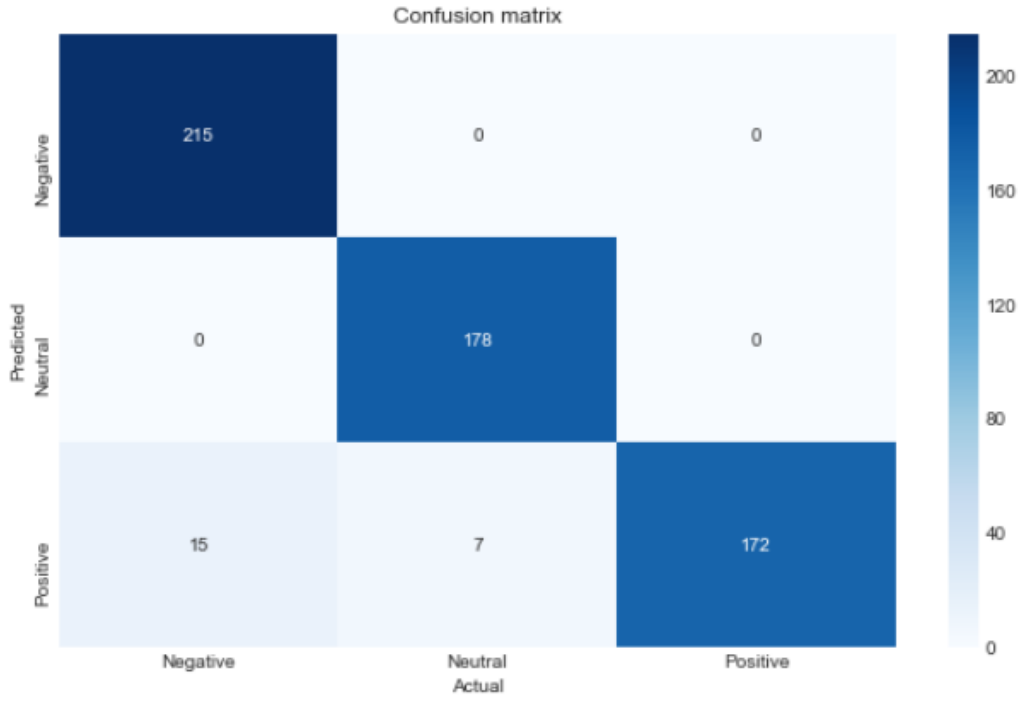

Fig 4.4

Κεφάλαιο 5

Αποτελέσματα και Μελλοντική Έρευνα

Έχοντας εξετάσει πλήρως και τους τέσσερις αλγόριθμους (GBM,MNB,RF,MLR) προχωρήσαμε και δημιουργήσαμε έναν πίνακα που περιλαμβάνει πλήρως την απόδοσή τους.

	Model	Test Set Accuracy	Training Set Accuracy
0	Gradient Boosting	0.962521	0.992321
1	Logistic Regression	0.979557	0.991894
2	Multinomial Naïve Bayes	0.860307	0.887372
3	Random Forest	0.984668	0.999147

Παρατηρώντας και ταξινομώντας τα Test/Train accuracy καταλήξαμε στον επικρατέστερο.

	Model	Test Set Accuracy	Training Set Accuracy
3	Random Forest	0.984668	0.999147
1	Logistic Regression	0.979557	0.991894
0	Gradient Boosting	0.962521	0.992321
2	Multinomial Naïve Bayes	0.860307	0.887372

Ο **Random Forest** φαίνεται να έχει το καλύτερο test accuracy από τους υπόλοιπους τρεις για το συγκεκριμένο dataset. Έκτος απο το accuracy φαίνεται ότι είναι ο επικρατέστερος και από άποψη overfitting αφού παρουσιάζει το λιγότερο από όλους [11]. (Overfitting=Train_acc-Test_acc)

Σαν μελλοντική έρευνα θα μπορούσαμε να ασχοληθούμε και με άλλες ιστοσελίδες ηλεκτρονικών πωλήσεων όπως για παράδειγμα το eBay. Επίσης θα μπορούσαμε να συγκεκριμενοποιήσουμε την έρευνα μας σε συγκεκριμένα προϊόντα από τα οποία θα καταλήγαμε σε αυτά που φέρνουν μεγαλύτερη δημοσιότητα και αξιοπιστία στο εκάστοτε κατάστημα. Τέλος, με έναν γράφο σύγκρισης, θα μπορούσαμε να συγκρίνουμε τα καταστήματα μεταξύ τους καταλήγοντας στο επικρατέστερο.

Βιβλιογραφία

[1] GeeksforGeeks, «KDD Process in Data Mining,» [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/kdd-process-in-data-mining/.

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

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

[4] imbalanced-learn, «Over-sampling,» [Ηλεκτρονικό]. Available: https://imbalanced-learn.readthedocs.io/en/stable/over_sampling.html.

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

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

[7] A. M. Kibriya, E. Frank, B. Pfahringer και G. Holmes, «Multinomial Naive Bayes for Text Categorization Revisited,» σε *AI 2004: Advances in Artificial Intelligence*, 2004, pp. 488-499.

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

[9] P. M., «Random forest classifier for remote sensing classification,» *Taylor and Francis Online,* τόμ. 26, αρ. 1, pp. 217-222, 2005.

[10] A. JAIN, «Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python,» 2016.

[11] C. W. Kim, «Overfitting (What They Are & Train, Validation, Test & Regularization),» *Medium,* 15 April 2018.

[12] M. Thapliyal, «Document Classification Using Machine Learning,» *Medium,* 8 December 2018.

[13] 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.

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

[15] E. K. Ikonomakis, V. Tampakas και S. Kotsiantis, «Text Classification Using Machine Learning Techniques,» *Research Gate,* τόμ. 4, αρ. 8, pp. 966-974, 2005.

[16] C. D. Manning, P. Raghavan και H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[17] E. Haddi, X. Liu και Y. Shi, «The Role of Text Pre-processing in Sentiment Analysis,» *Science Direct,* τόμ. 17, pp. 26-32, 2013.

[18] R. Xia, C. Zong και S. Li, «Ensemble of feature sets and classification algorithms for sentiment classification,» *Science Direct,* τόμ. 181, αρ. 6, pp. 1138-1152, 2011.

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

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

[21] «GeeksforGeeks,» [Ηλεκτρονικό]. Available: https://www.geeksforgeeks.org/confusion-matrix-machine-learning/.

[22] . J. Starkweather και A. K. Moske , «Multinomial Logistic Regression,» 2011.

[23] B. Krishnapuram, L. Carin, M. A. Figueiredo και A. J. Hartemink, «Sparse Multinomial Logistic Regression:,» *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, τόμ. 27, αρ. 6, 2005.