ΑΝΑΓΝΩΡΙΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΑΠΟ ΚΕΙΜΕΝΟ

Διλμπέρης Εμμανουήλ, ME2104
Τμήμα Μεταπτυχιακών Σπουδών
"Πληροφοριακά Συστήματα και Υπηρεσίες"
Ειδίκευση: Μεγάλα Δεδομένα και Αναλυτική
Πανεπιστήμιο Πειραιά
E-mail: e.dilmperis@gmail.com

Κατσιούλας Κοσμάς, ME2109
Τμήμα Μεταπτυχιακών Σπουδών
"Πληροφοριακά Συστήματα και Υπηρεσίες"
Ειδίκευση: Μεγάλα Δεδομένα και Αναλυτική
Πανεπιστήμιο Πειραιά
E-mail: kosmas.katsioulas96@gmail.com

Μοβσεσιάν Άρμεν, ME2117
Τμήμα Μεταπτυχιακών Σπουδών
"Πληροφοριακά Συστήματα και Υπηρεσίες"
Ειδίκευση: Μεγάλα Δεδομένα και Αναλυτική
Πανεπιστήμιο Πειραιά
E-mail:armenmov95@gmail.com

Περίληψη **Στη συγκεκριμένη εργασία γίνεται ανάλυση δεδομένων που αφορούν κριτικές χρηστών για ταινίες. Η λήψη των δεδομένων έγινε από το Kaggle και αναλύονται με τη χρήση της γλώσσας προγραμματισμού Python και σχετικών βιβλιοθηκών της. Χρησιμοποιήθηκαν διάφορες μέθοδοι για την κατηγοριοποίηση των κειμένων κριτικής , με σκοπό το διαχωρισμό αυτών σε θετικών και αρνητικών. Εν συνεχεία έγινε αξιολόγηση της απόδοσης των μοντέλων, με σκοπό την επιλογή του βέλτιστου. Με τη χρήση του προτεινόμενου μοντέλου, αναπτύχθηκε εφαρμογή κατηγοριοποίησης νέων κριτικών. Τέλος, έγινε εφαρμογή κανόνων συσχέτισης στο αρχικό dataset, με σκοπό την εξαγωγή συσχετίσεων μεταξύ λέξεων σε θετικές και αντίστοιχα σε αρνητικές κριτικές.**

I. ΕΙΣΑΓΩΓΗ

Στη σύγχρονη εποχή ο όγκος των δεδομένων που παράγεται καθημερινά είναι τεράστιος , αλλά και η επιτακτική ανάγκη για άντληση περισσότερης γνώσης , κάνει την ανάλυση των μεγάλων δεδομένων αναγκαία για κάθε οργανισμό. Για την εξόρυξη δεδομένων και την άντληση γνώσης, χρησιμοποιούνται μέθοδοι μηχανικής μάθησης. Ένας κλάδος της εξόρυξης δεδομένων , είναι η εξόρυξη πληροφορίας από κείμενα. (text mining). Στη συγκεκριμένη εργασία γίνεται εξόρυξη γνώμης, η οποία είναι ταυτόσημη με την ανάλυση συναισθήματος. Η ανάλυση συναισθήματος (sentiment analysis), είναι ένα συνεχώς αναπτυσσόμενο πεδίο της επιστήμης των υπολογιστών. Συνδυάζει πολλούς ερευνητικούς τομείς όπως η επεξεργασία φυσικής γλώσσας, η εξόρυξη δεδομένων και η εξόρυξη κειμένου. Είναι η μελέτη που αναλύει την γνώμη, το συναίσθημα, την αξιολόγηση, τη συμπεριφορά των ανθρώπων απέναντι σε οντότητες, όπως προϊόντα, υπηρεσίες, οργανισμούς, γεγονότα και τις ιδιότητές τους. Για αυτόν τον λόγο η ανάλυση συναισθήματος αποκτά γρήγορα μείζονα σημασία σε οργανισμούς που προσπαθούν να ενσωματώσουν μεθόδους υπολογιστικής ευφυίας στις λειτουργίες τους.

Η εφαρμογή της ανάλυσης συναισθήματος γίνεται σε πολλαπλούς τομείς. Ένας από αυτούς είναι ο χώρος των Social Media, όπου οι χρήστες κάνουν χιλιάδες αναρτήσεις σε πολύ μικρό χρονικό διάστημα . Για παράδειγμα στο Twitter γίνονται 6.000 tweets ανά δευτερόλεπτο και ένα μέρος αυτών σίγουρα θα αφορά επιχειρήσεις και οργανισμούς, οι οποίες μπορούν με τη σειρά τους να λάβουν σημαντικό feedback από αυτή τη διαδικασία. Μια άλλη εφαρμογή της ανάλυσης συναισθήματος είναι στη διαχείριση εξυπηρέτησης πελατών, όπου μεγάλος όγκος εισερχόμενων κλήσεων μπορεί να ταξινομηθεί σε κλάσεις, όπως για παράδειγμα επείγουσες και μη επείγουσες, και να δημιουργηθεί η κατάλληλη σειρά προτεραιότητας εξυπηρέτησης.

Σημαντική επίδραση παρατηρείται και στον τομέα της ανάλυσης κριτικών σε διάφορες ταινίες. Πιο συγκεκριμένα, η ανάλυση συναισθήματος επιδιώκει την εξόρυξη γνώμης του κοινού, ανάλογα με την κριτική τους σε μια συγκεκριμένη ταινία. Οι κριτικές των χρηστών ταξινομούνται με βάση τη διάθεση που εκφράζουν, ως θετικές, αρνητικές στη συγκεκριμένη εργασία, ενώ σε άλλα project μπορεί να υπάρξει και η κλάση της ουδέτερης κριτικής.

II. ΣΤΟΧΟΣ ΠΡΟΒΛΗΜΑΤΟΣ

Στη παρούσα εργασία ,χρησιμοποιήθηκε ένα dataset που ανακτήθηκε από το Kaggle και τα δεδομένα αυτού προήλθαν από τις κριτικές χρηστών στον ιστότοπο του imdb. Στόχος του προβλήματος είναι η εξαγωγή ενός μοντέλου ταξινόμησης, το οποίο θα κατατάσσει τις κριτικές των χρηστών σε δύο κλάσεις, ως αρνητικές και θετικές, σύμφωνα με το περιεχόμενο.

Προϋποθέσεις για την επίτευξη του στόχου είναι:

- η ανάλυση των δεδομένων
- ο καθαρισμός των δεδομένων
- η μετατροπή των κειμένων σε διανύσματα
- η επιλογή των κατάλληλων ταξινομητών
- η αξιολόγηση των τελικών μοντέλων

Στη συνέχεια υλοποιήθηκε μια εφαρμογή η οποία δέχεται σαν είσοδο ένα κείμενο κριτικής από τον χρήστη και επιστρέφει τη κατηγορία στην οποία ανήκει (θετική, αρνητική) .Η ταξινόμηση αυτή επιτυγχάνεται με τη βοήθεια του μοντέλου που αναπτύχθηκε προηγουμένως. Τέλος αξιοποιήθηκαν οι αξιολογήσεις των χρηστών για την εξαγωγή κανόνων συσχέτισης στην κάθε μια από τις δύο κλάσεις ξεχωριστά.

III. ΜΕΘΟΔΟΛΟΓΙΑ

i)Ανάλυση και προεπεξεργασία δεδομένων

Αρχικά με τη χρήση της γλώσσας Python έγινε φόρτωση των δεδομένων σε Dataframe, με τη βοήθεια της βιβλιοθήκης pandas. Παρατηρήθηκε ότι το dataset περιείχε πέντε στήλες από τις οποίες κρατήθηκαν μόνο δυο. Η μία ήταν η στήλη που περιείχε την αξιολόγηση (review) και η άλλη ήταν η στήλη με τις ετικέτες των κλάσεων(label). Οι υπόλοιπες αφαιρέθηκαν από το Dataframe. Επίσης έγινε έλεγχος για ελλείπουσες τιμές (missing values) αλλά δεν εντοπίστηκε καμία. Προβάλλοντας ένα μέρος των δεδομένων βρέθηκαν γραμμές οι οποίες, σαν ετικέτα, δεν περιείχαν μια από τις δύο κλάσεις, αλλά την τιμή unsup. Εν συνεχεία αυτές αφαιρέθηκαν.

Μέσω της βιβλιοθήκης re δημιουργήθηκαν οι κατάλληλες εντολές για να αφαιρεθούν από τις αξιολογήσεις:

- οι διευθύνσεις url
- τα tags της γλώσσας σήμανσης html (πχ \<br\s)
- ειδικοί χαρακτήρες όπως '@,#'
- σημεία στίξης
- περιττά κενά διαστήματα(extra spaces)

Αφού έγινε αφαίρεση του θορύβου από τα σχόλια, στο επόμενο στάδιο της επεξεργασίας πραγματοποιήθηκε stemming στις αξιολογήσεις των χρηστών. Οι δύο συναρτήσεις που δοκιμάστηκαν ήταν οι WordNetLemmatizer και PorterStemmer ,οι οποίες προέρχονται από τη βιβλιοθήκη nltk, και λειτουργούν ως εξής :

Original Word	WordNetLemmatizer	PorterStemmer
loves	love	love
killing	killing	kill
connections	connection	connect
children	child	children
frequency	frequency	frequenc

Εικόνα 1

Το PorterStemmer και το WordNetLemmatizer, παίρνουν και τα δυο τη ρίζα των λέξεων που δέχονται σαν είσοδο. Η κύρια διαφορά τους είναι ότι το αποτέλεσμα του PorterStemmer ίσως να μην είναι πραγματική λέξη, όπως φαίνεται στην *Εικόνα 1* στη λέξη frequency. Η μέθοδος αυτή μας δίνει ως αποτέλεσμα τη λέξη frequenc , η οποία δεν αντιστοιχεί σε κάποια πραγματική. Για τον προαναφερθέντα λόγο, αλλά και εξαιτίας της τελικής απόδοσης των μοντέλων, από αυτές τις δύο μεθόδους, επιλέχθηκε η WordNetLemmatizer.

Στο επόμενο βήμα, έγινε αφαίρεση των stopwords από τις κριτικές των χρηστών . Ως stopwords ορίζονται:

- αντωνυμίες (πχ 'me', 'her', 'each' κ.α.)
- χρονικοί ,αιτιολογικοί και τοπικοί σύνδεσμοι (πχ 'when', 'then', 'while' , 'because' κ.α.)

Η υλοποίηση αυτού του σταδίου έγινε με τη χρήση της έτοιμης λίστας stopwords, για την αγγλική γλώσσα, η οποία βρίσκεται μέσα στο πακέτο nltk.corpus.

Ύστερα από την πραγματοποίηση των παραπάνω βημάτων, ολοκληρώθηκε η απαιτούμενη επεξεργασία των δεδομένων για τον καθαρισμό τους. Προτού γίνει η μετατροπή των κριτικών σε διάνυσμα, για την καλύτερη αντίληψη των δεδομένων, έγινε η οπτικοποίηση τους με τη χρήση της βιβλιοθήκης WordCloud. Μέσω της WordCloud, γίνεται αναπαράσταση των συχνότερα εμφανιζόμενων λέξεων στις κριτικές και το μέγεθος τους στο γράφημα είναι ανάλογο με τη συχνότητα τους . Δηλαδή , όσο μεγαλύτερη είναι η συχνότητα εμφάνισης της λέξης , τόσο μεγαλύτερο είναι το μέγεθος της στο γράφημα(Εικόνα 2). Για την αντικειμενικότητα του γραφήματος, έχουν αφαιρεθεί οι λέξεις 'movie' και 'film' διότι ήταν πιο συχνές, κάτι το οποίο είναι λογικό, διότι το dataset αναφέρεται σε κριτικές ταινιών.

Εικόνα 2

ii) Μετατροπή κειμένου σε διάνυσμα

Οι αλγόριθμοι ταξινόμησης δε δέχονται γράμματα ή λέξεις για εκπαίδευση . Για το λόγο αυτό, είναι απαραίτητο να γίνει μετατροπή των κειμένων σε διανύσματα. Οι τεχνικές που δοκιμάστηκαν ήταν :

- Bag of Words (BoW)
- συχνότητας όρου – αντίστροφης συχνότητας όρου (TF-IDF)
- Word2Vec

Η τεχνική Bag of Words(BoW) δημιουργεί έναν αραιό πίνακα , όπου η κάθε λέξη αντιστοιχεί σε μία στήλη του και κάθε αξιολόγηση αντιστοιχεί σε μία γραμμή. Οι τιμές του πίνακα αντιστοιχούν στην συχνότητα της λέξης στο συγκεκριμένο σχόλιο.

Η τεχνική TF-IDF δεν είναι τόσο αφελής όσο η Bag of Words(BoW), παρόλο που και αυτή βασίζεται στη συχνότητα. Είναι μία αριθμητική στατιστική που προορίζεται να αντικατοπτρίζει τη σημασία μιας λέξης για ένα έγγραφο ή μια συλλογή εγγράφων. Στο Bag of Words η διανυσματοποίηση αφορούσε μόνο τη συχνότητα των λέξεων σε μία κριτική. Αυτό έχει ως αποτέλεσμα, λέξεις όπως για παράδειγμα 'movies', 'film' ,οι οποίες δε συνεισφέρουν στην εξαγωγή συμπεράσματος για τη διάθεση των χρηστών απέναντι στην ταινία, να αποκτούν το ίδιο επίπεδο σημαντικότητας με λέξεις που συνεισφέρουν.

Ο αλγόριθμος του TF-IDF βοηθάει στην επίλυση αυτού του θέματος, διότι οι λέξεις που επαναλαμβάνονται πολύ συχνά ,δεν υπερέχουν αυτών που έχουν μικρότερη συχνότητα αλλά είναι πιο σημαντικές. Διακρίνεται σε δύο σκέλη. Το tf, το οποίο υπολογίζει τη συχνότητα εμφάνισης της λέξης στο έγγραφο όπως ακολούθως:

$$tf = \frac{Συχνότητα\ λέξης\ στο\ έγγραφο}{Συνολικός\ αριθμός\ λέξεων\ στο\ έγγραφο}$$

και το idf, το οποίο υπολογίζει την αντίστροφη αναλογία των εγγράφων που περιέχουν τη συγκεκριμένη λέξη, με σκοπό, μια κοινή λέξη που βρίσκεται σε πολύ μεγάλο ποσοστό των εγγράφων, να είναι μικρότερης σημασίας για το τρέχον έγγραφο.

$$idf = log\frac{Συνολικός\ αριθμός\ των\ κειμένων}{Τα\ κείμενα\ που\ περιέχουν\ τη\ λέξη}$$

Το τελικό TF-IDF score προκύπτει από:

$$TF - IDF = tf * idf$$

Η τεχνική Word2Vec ,σε αντίθεση με τις δύο προηγούμενες τεχνικές , δεν αντιμετωπίζει την κάθε λέξη σαν «ατομική» οντότητα . Ο σκοπός και η χρησιμότητα του Word2Vec είναι να ομαδοποιήσει τα διανύσματα παρόμοιων λέξεων μαζί στο διανυσματικό χώρο. Στην παρακάτω εικόνα , φαίνεται το πως δουλεύει ο αλγόριθμος Word2Vec για τις λέξεις "good" και "bad" .

good	bad
(decent, 0.7512123584747314)	(terrible, 0.6581383943557739)
(alright, 0.6433106064796448)	(awful, 0.6545064449310303)
(great, 0.6401292681694031)	(horrible, 0.6514533758163452)
(ok, 0.6113848090171814)	(lousy, 0.5819767713546753)
(okay, 0.5912384986877441)	(sucks, 0.574227511882782)
(fine, 0.5896051526069641)	(crappy, 0.5432813167572021)
(excellent, 0.5836659669876099)	(atrocious, 0.5348544716835022)
(cool, 0.5834649205207825)	(dreadful, 0.5229642987251282)
(nice, 0.5804026126861572)	(lame, 0.5168747901916504)

Εικόνα 3

Σε κάθε γραμμή , εμφανίζεται η λέξη που θεώρησε ο αλγόριθμος ως όμοια με την αρχική μαζί με το βαθμό ομοιότητας , ο οποίος έχει ως βέλτιστη τιμή τη μονάδα.

Από τις τρείς τεχνικές , επιλέχθηκε ο αλγόριθμος του TF-IDF λόγω της καλύτερης τελικής απόδοσης του μοντέλου. Από την Εικόνα 4 παρατηρείται ότι με την τεχνική Bag of Words το μοντέλο είναι επιρρεπές στο φαινόμενο του overfitting, διότι υπάρχει μεγάλη απόκλιση στο ποσοστό της ακρίβειας ανάμεσα στα test και train sets.

Vectorization Technique	Accuracy on Test Set	Accuracy on Train Set
Bag of Words	88.70%	99.52%
TF-IDF	90.33%	93.46%
Word2Vec	73.42%	74.82%

Εικόνα 4

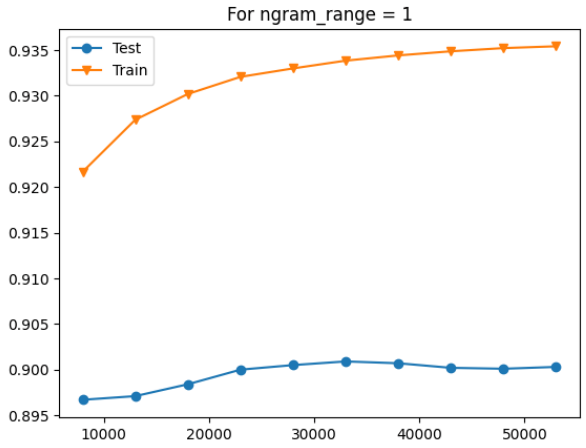
Χρησιμοποιώντας τον αλγόριθμο TF-IDF για το μετασχηματισμό των κειμένων σε διανύσματα , έγινε επεξεργασία κάποιων παραμέτρων. Οι τρείς παράμετροι αυτοί είναι η max_feature, η ngram_range και η max_df.

Η max_feature, παίρνει σαν τιμή έναν ακέραιο αριθμό που δηλώνει το πλήθος των λέξεων που θα κρατήσει σαν χαρακτηριστικά(features), με κριτήριο το term frequency.

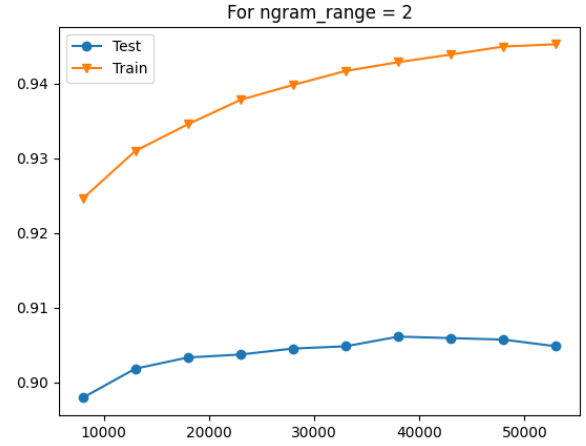
Η ngram_range παίρνει ως είσοδο μια πλειάδα από δυο ακεραίους ,οι οποίοι αναπαριστούν το κατώτερο και ανώτερο όριο του εύρους τιμών n, για διαφορετικά ngram προς εξαγωγή. Όλες οι τιμές του n για τις οποίες ισχύει min_n ≤ n ≤max_n θα χρησιμοποιηθούν. Για παράδειγμα για ngram=(1,2) ο αλγόριθμος θα επιστρέψει όλες τις λέξεις μόνες τους , αλλά και ανά δυο σαν features. Για ngram_range = (1,1) , ο TFIDF επιστρέφει 136874 features και για ngram_range = (1,2) επιστρέφει 2664143 features. Αυτό σημαίνει ότι αυξάνονται πάρα πολύ οι διαστάσεις των δεδομένων , κάτι το οποίο μπορεί να οδηγήσει σε overfitting , σε αύξηση της πολυπλοκότητας αλλά και σε αύξηση του χρόνου εκπαίδευσης των μοντέλων ταξινόμησης.

Η max_df παίρνει ως είσοδο έναν ακέραιο αριθμό ο οποίος λειτουργεί σα μέγιστο όριο εμφάνισης μιας λέξης μέσα σε μια συλλογή κειμένων. Αν κάποια λέξη υπερβεί αυτό το όριο, τότε ο αλγόριθμος την αποκλείει σαν επιλογή χαρακτηριστικού(feature).

Στα γραφήματα που ακολουθούν παρατηρείται, για διάφορες τιμές των max_feature και ngram_range , η απόδοση του τελικού μοντέλου σε train και test set.



Εικόνα 5



Εικόνα 6

Από τις Εικόνες 5 και 6, γίνεται αντιληπτό ότι όσο αυξάνονται τα features η διαφορά των αποδόσεων σε train και test set μεγαλώνει, ενώ από μια τιμή και έπειτα η απόδοση στο test set αρχίζει να φθίνει.

Σαν τιμές των παραμέτρων max_feature και ngram_range δόθηκαν οι 18000 και (1,2) αντίστοιχα, με τις οποίες εξάγεται το βέλτιστο αποτέλεσμα. Επίσης, η παράμετρος max_df πήρε την τιμή 24000, που σημαίνει ότι αν μια λέξη εμφανίζεται τουλάχιστον σε 24000 κριτικές από τις 50000 που βρίσκονται στο dataset, δε θα επιλεχθεί .

iii) Δημιουργία μοντέλου ταξινόμησης

Για τη δημιουργία του τελικού μοντέλου ταξινόμησης χρησιμοποιήθηκαν οι αλγόριθμοι ταξινόμησης Logistic Regression και Support Vector Machine(SVM).

Η Λογιστική Παλινδρόμηση (Logistic Regression) είναι μια μέθοδος στατιστικής ανάλυσης για την πρόβλεψη ενός δυαδικού αποτελέσματος, όπως αυτό στην παρούσα εργασία, με βάση προηγούμενες παρατηρήσεις ενός συνόλου δεδομένων. Ένα μοντέλο λογιστικής παλινδρόμησης προβλέπει μια εξαρτημένη μεταβλητή δεδομένων αναλύοντας τη σχέση ανάμεσα σε μία ή περισσότερες υπάρχουσες ανεξάρτητες μεταβλητές.

Το SVM ή Support Vector Machine είναι ένα μοντέλο για προβλήματα ταξινόμησης και παλινδρόμησης. Μπορεί να λύσει γραμμικά και μη γραμμικά προβλήματα και να λειτουργήσει καλά για πολλά πρακτικά προβλήματα. Η ιδέα του SVM είναι απλή, ο αλγόριθμος δημιουργεί μια γραμμή ή ένα υπερεπίπεδο που διαχωρίζει τα δεδομένα σε κλάσεις.

Οι δύο παραπάνω αλγόριθμοι , χρησιμοποιήθηκαν για την ταξινόμηση των κριτικών ταινιών σε θετικές και αρνητικές.

iv) Αξιολόγηση μοντέλων ταξινόμησης

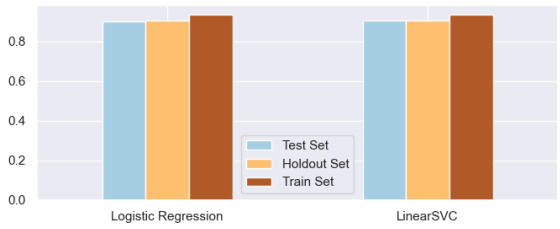
Η αξιολόγηση των μοντέλων έγινε βάσει της ακρίβειας ταξινόμησης (Accuracy score), του χρόνου εκπαίδευσης αλλά και του χρόνου πρόβλεψης.

Αρχικά γίνεται διαχωρισμός του dataset σε train, test και holdout set. Το train set αποτελείται από το 80% του αρχικού dataset, το test set από το 18% και το holdout set από το 2%(1000 δείγματα). Το holdout set δεν χρησιμοποιείται στη διαδικασία της εκπαίδευσης, με σκοπό να έχουμε μια αμερόληπτη εκτίμηση της απόδοσης του μοντέλου. Το set αυτό παίζει σημαντικό ρόλο καθώς διασφαλίζει ότι το μοντέλο μπορεί να γενικευτεί καλά σε δεδομένα που δεν έχει δει.. Οι αλγόριθμοι εκπαιδεύονται με το train set και προβλέπουν τις τιμές για τα test και holdout set. Για τον αλγόριθμο της Λογιστικής Παλινδρόμησης δε χρειάστηκε κάποια μετατροπή των αρχικών παραμέτρων ενώ στον αλγόριθμο SVM η καλύτερη απόδοση προέκυψε όταν η παράμετρος C πήρε την τιμή 0.1.

Παρακάτω παρουσιάζονται τα αποτελέσματα για τη μετρική accuracy score κάθε μέθοδο που δοκιμάστηκε για την ανάπτυξη των μοντέλων και στα τρία σετ:

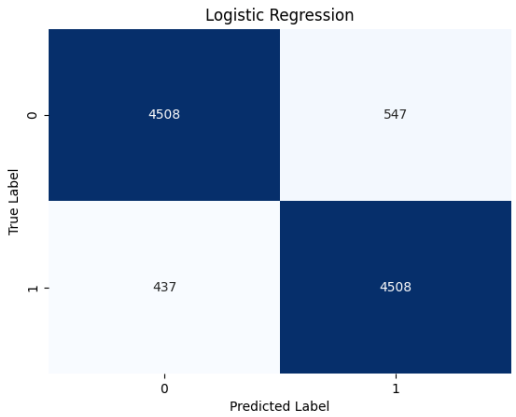
	Test Set	Holdout Set	Train Set
Logistic Regression	90.13%	90.4%	93.25%
LinearSVC	90.31%	90.5%	93.45%

Εικόνα 7

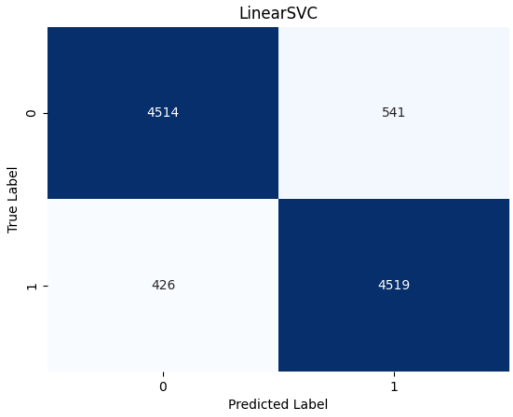


Εικόνα 8

Πέρα από την παραπάνω μετρική έγινε έλεγχος με τη χρήση confusion matrix για να διαπιστωθεί ότι τα μοντέλα ταξινομούν με την ίδια ακρίβεια και τις δύο κλάσεις.



Εικόνα 9 (0 = negative reviews, 1 = positive reviews)



Εικόνα 10 (0 = negative reviews, 1 = positive reviews)

Στη συνέχεια χρονομετρήθηκαν τα μοντέλα ως προς τις διαδικασίες της εκπαίδευσης και της πρόβλεψης. Για την εκτίμηση του χρόνου που χρειάζεται για να εκτελεστεί το μοντέλο κατηγοριοποίησης και να δώσει αποτελέσματα, πάρθηκε ο μέσος χρόνος εκτέλεσης από 10 επαναλήψεις , όπου στην κάθε μια επιλέγονται τυχαία 1000 κριτικές από το dataset. Η μονάδα μέτρησης χρόνου είναι σε δευτερόλεπτα.

Models	Train Time	Mean Prediction Time
Logistic Regression	107.54	2.264
SVM	102.05	2.239

v) Προτεινόμενο μοντέλο ταξινόμησης

Από την προηγούμενη ενότητα , γίνεται αντιληπτό ότι και τα δύο μοντέλα αποδίδουν εξίσου καλά ως προς την ακρίβεια ταξινόμησης. Ακόμη παρατηρήθηκε ότι οι χρόνοι εκπαίδευσης και πρόβλεψης των μοντέλων δεν είχαν μεγάλες διαφορές. Ως τελικό μοντέλο ταξινόμησης επιλέχθηκε το μοντέλο που αναπτύχθηκε με τη χρήση του αλγορίθμου ταξινόμησης LinearSVC καθώς ήταν αυτό που υπερτερούσε στις λεπτομέρειες.

IV. ΑΝΑΠΤΥΞΗ ΕΦΑΡΜΟΓΗΣ

Με τη βοήθεια του προτεινόμενου μοντέλου δημιουργήθηκε μια εφαρμογή, με τη βοήθεια της γλώσσας προγραμματισμού Python, η οποία δίνει τη δυνατότητα στο χρήστη να γράφει μια κριτική και το σύστημα με τη σειρά του αναγνωρίζει και εκτυπώνει το συναίσθημα του χρήστη, δηλαδή θετικό ή αρνητικό.

Πιο αναλυτικά, με το που γίνεται εισαγωγή ενός σχολίου από το χρήστη ,αυτόματα γίνεται μια προεπεξεργασία σε αυτό, όμοια με αυτή που υλοποιήθηκε για την ανάπτυξη του κατηγοροποιητή (αφαίρεση ειδικών χαρακτήρων, lemmatization, αφαίρεση stopwords κ.α.). Στη συνέχεια μετατρέπεται σε διάνυσμα και εισάγεται στο μοντέλο , το οποίο με τη σειρά του θα προβλέψει την κλάση στην οποία ανήκει.

V. ΑΝΑΛΥΣΗ ΚΑΙ ΕΞΑΓΩΓΗ ΣΥΣΧΕΤΙΣΕΩΝ

Για την εκτέλεση αυτής της διαδικασίας δημιουργήθηκε εφαρμογή στη γλώσσα προγραμματισμού Python στην οποία με τη βοήθεια σχετικών βιβλιοθηκών υλοποιείται ο Apriori αλγόριθμος. Οι μετρικές που χρησιμοποιήθηκαν είναι:

- το support,
- το confidence,
- και το lift

Προκειμένου να γίνει εξαγωγή συσχετίσεων πρέπει να ικανοποιούνται κάποια κατώτατα όρια τόσο για το support όσο και για το confidence. Για παράδειγμα, ορίζοντας το support με τιμή 0.1, δηλαδή να περιέχεται τουλάχιστον στο 10% των κριτικών, και το confidence με τιμή 0.5 , η εφαρμογή δίνει τα ακόλουθα αποτελέσματα.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
(time)	(movie)	0.33708	0.65312	0.23216	0.688739	1.054536
(time)	(film)	0.33708	0.53632	0.19328	0.573395	1.069129
(time)	(like)	0.33708	0.49444	0.18688	0.554408	1.121286
(time, movie)	(film)	0.23216	0.53632	0.12460	0.536699	1.000706
(time, film)	(movie)	0.19328	0.65312	0.12460	0.644661	0.987048

Εικόνα 11(Δείγμα από το συνολικό πίνακα που δίνει η εφαρμογή)

Η παραπάνω διαδικασία εφαρμόστηκε σε δύο υποσύνολα του αρχικού dataset, τα οποία είναι οι θετικές και οι αρνητικές κριτικές των χρηστών. Ύστερα από πολλές δοκιμές, παρατηρήθηκε ότι υπάρχουν λέξεις όπως movies, watch, like κ.α., οι οποίες σε συνδυασμό μεταξύ τους παρουσιάζουν υψηλό confidence, όμως το support τους ήταν εξίσου υψηλό , καθιστώντας τη συσχέτιση μη σημαντική. Επίσης υπήρχαν συνδυασμοί λέξεων που ενώ παρουσίαζαν υψηλό confidence δεν μπορούσαν να θεωρηθούν κανόνες συσχέτισης με κάποιο ενδιαφέρον διότι ο συνδυασμός τους είναι στα πλαίσια της λογικής. Ένα τέτοιο παράδειγμα είναι, όταν παρουσιάζονται οι λέξεις <worst,movie> είναι λογικό να ακολουθεί μέσα στη πρόταση η λέξη <ever> τις περισσότερες φορές.

Για μεγάλες τιμές του support δεν υπήρξαν κάποιοι κανόνες συσχέτισης που παρουσίασαν ιδιαίτερο ενδιαφέρον. Έτσι, έγινε έλεγχος για χαμηλότερες τιμές του support, θέτοντας ως κατώτατο όριο εμφάνισης του συνδυασμού των λέξεων τις 500 κριτικές.

Για τις θετικές κριτικές έγινε η εξαγωγή των παρακάτω κανόνων συσχέτισης:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
white	black	0.04176	0.05812	0.02216	0.530651	9.130271
supporting	cast	0.04000	0.13780	0.02244	0.561000	4.071118

Εικόνα 12

Από τον παραπάνω πίνακα , φαίνεται ότι όταν υπάρχει η λέξη <white> συνυπάρχει η λέξη <black> με confidence 0.5306 και lift 9.13, κάτι το οποίο δηλώνει την υψηλή σημαντικότητα της συσχέτισης τους. Διαβάζοντας τις αξιολογήσεις που περιέχουν και τις δύο λέξεις παρατηρήθηκε ότι οι κριτικές απευθύνονται ως επί το πλείστον σε ασπρόμαυρες ταινίες. Ομοίως, όταν εμφανίζεται η λέξη <supporting> ακολουθεί και η λέξη <cast> με confidence 0.5610 και lift 4.07.

Αντίστοιχα, για τις αρνητικές κριτικές:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift
film, low	budget	0.03704	0.06608	0.02172	0.586393	8.873987
bad, special	effects	0.03648	0.08964	0.02608	0.714912	7.975371

Εικόνα 13

Όταν παρουσιάζεται ο συνδυασμός των λέξεων <film,low> τότε εμφανίζεται και η λέξη <budget> με confidence 0.5863 και lift 8.87. Τέλος, με το συνδυασμό των λέξεων <bad,special> ακολουθεί η λέξη <effects> με confidence 0.7149 και lift 7.97.

VI. ΠΑΡΑΤΗΡΗΣΕΙΣ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Για το μοντέλο ταξινόμησης που προτάθηκε, έκπληξη προκάλεσε το γεγονός ότι από τα 2.664.143 συνολικά χαρακτηριστικά που δόθηκαν από τον μετασχηματιστή TF-IDF, 18.000 από αυτά ήταν αρκετά για τη βελτιστοποίηση του μοντέλου, δηλαδή την αύξηση της απόδοσής του ως προς την ακρίβεια και τη μείωση του χρόνου εκπαίδευσης.

Κάτι ακόμα που παρατηρήθηκε ήταν ότι, οι ensemble μέθοδοι ταξινόμησης όπως Random Forest, Gradient Boosting και AdaBoosting Classifier δεν ήταν εξίσου αποδοτικοί με το προτεινόμενο μοντέλο, ακόμα και μετά τη ρύθμιση των υπερπαραμέτρων τους. Δύο άλλα μειονεκτήματά τους ήταν ο χρόνος εκπαίδευσης των μοντέλων που δημιουργήθηκαν με αυτές τις μεθόδους και το overfit.

Model	Accuracy on Test	Accuracy on Train	Train Time
RandomForest	85.42%	95.21%	271.05
GradientBoosting	84.67%	97.01%	851.44
AdaBoosting	77.54%	89.94%	739.38

Εικόνα 14

Όσον αφορά την εφαρμογή που αναπτύχθηκε, καταφέρνει σε μεγάλο ποσοστό να κατατάξει στη σωστή κλάση τις νέες κριτικές που εισάγονται από το χρήστη. Εκεί που υστερεί όμως είναι στην ειρωνεία αλλά και όταν το σχόλιο περιέχει την ονομασία της ταινίας ή κάποια λόγια από την ταινία τα οποία έχουν χροιά αντίθετη από το συναίσθημα του χρήστη.

Παραδείγματος χάριν η κριτική «The Magnificent Seven was not magnificent at all.» ταξινομήθηκε λανθασμένα ως θετική ενώ η κριτική «The Bad boys it's a good choice for Saturday Night» ταξινομήθηκε λανθασμένα ως αρνητική.

Ακόμη διαπιστώθηκε ότι οι αξιολογήσεις χρηστών που συγκρίνουν μια ταινία με ένα prequel της μπορεί να μπερδέψει τον αλγόριθμο.

Ένα τέτοιο παράδειγμα είναι η κριτική «The first chapter of the trilogy was the best movie I've ever seen, but chapter two seems to be a scrappy work, I am disappointed.».

Τέλος, αναφορικά με τη διαδικασία εξαγωγής κανόνων συσχέτισης, ύστερα από πολλές δοκιμές, διαπιστώθηκε ότι δεν ήταν εφικτή η εξαγωγή ενδιαφέρουσας πληροφορίας. Αυτό υποδηλώνει ότι τα dataset που αφορούν κριτικές ταινιών δεν είναι κατάλληλα για την ανακάλυψη σημαντικών σχέσεων και ισχυρών κανόνων.

VII. ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Mining of Massive Datasets. Cambridge University Press. 2014 (2nd Edition), Jure Leskovec, Anand Rajaraman, Jeff Ullman.

[2] Data Mining: Concepts and Techniques, , Han and M. Kamber, Morgan Kaufmann, 2006.

[3] Sentiment Analysis Using Support Vector Machine Ms. Gaurangi Patil , Ms. Varsha Galande , Mr. Vedant Kekan , Ms. Kalpana Dange

[4] Μιχάλης Βαζιργιάννης, Μαρία Χαλκίδη, Εξόρυξη Γνώσης από Βάσεις Δεδομένων και τον Παγκόσμιο Ιστό, Εκδ. Gutenberg.

[5]Sotiris Kotsiantis, Dimitris Kanellopoulos, Association Rules Mining: A Recent Overview,2006.

[6] Association Rule Mining: A Survey , Qiankun Zhao.,Sourav S. Bhowmick.,(2003).

[7] Introduction to Information Retrieval, C.P. Manning, P. Raghavan, and H. Schutze, Cambridge Univ. Press, 2008.

[8] "The influence of preprocessing on text classification using a bag-of-words representation", Yaakov HaCohen-Kerner, Daniel Miller,Yair Yigal (2020).

[9] Vectorization of text documents for identifying unifiable news articles, AK Singh, M Shashi - Int. J. Adv. Comput. Sci. Appl, 2019.

[10] "An Introduction to Logistic Regression Analysis and Reporting", Chao-Ying Joanne Peng,Kuk Lida Lee & Gary M. Ingersoll (2011).

[11] Text Mining for Sentiment Analysis of Twitter Data Shruti Wakade, Chandra Shekar, Kathy J. Liszka and Chien-Chung Chan

[12] Aspect-based sentiment analysis of movie reviews on discussion boards, Tun Thura Thet, Jin-Cheon Na and Christopher S.G. Khoo

[13] New Avenues in Opinion Mining and Sentiment Analysis, Erik Cambria; Björn Schuller; Yunqing Xia; Catherine Havasi

[14] Sentiment analysis using product review data, 2015, Xing Fang and Justin Zhan

[15] An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation, Bijoyan Das and Sarit Chakraborty

[16] A Text Mining Technique Using Association Rules Extraction, 2007, Hany Mahgoub, Dietmar Rösner, Nabil Ismail, Fawzy Torkey