Sentiment Analysis On Movie Reviews

Vasileios Theiou, Vlachos Evgenios

Department of Electrical and Computer Engineering, University of Thessaly, Volos
Email: {evgvlachos, theiou}@uth.gr

Στην παρούσα εργασία μελετάται το θέμα του sentiment analysis σε κριτικές ταινιών για την εξαγωγή του συναισθήματος του ατόμου που γράφει την κριτική με τη χρήση πολλαπλών μοντέλων μηχανικής μάθησης. Ειδικότερα, κατόπιν μελέτης 2 ερευνών και επιλογής ενός dataset παρουσιάζονται και συγκρίνονται 5 διαφορετικά μοντέλα με στόχο την εύρεση του βέλτιστου αλγορίθμου για την ανάλυση της φυσικής γλώσσας και πρόβλεψη του ποσοστού του συναισθήματος κάθε κριτικής. Επιπλέον, δημιουργείται ένα γραφικό περιβάλλον πιο φιλικό στο χρήστη το οποίο δίνει τη δυνατότητα καταχώρησης μιας κρικτής και επιλογή συγκεκριμένων αλγορίθμων για εξαγωγή αποτελέσματος. Τέλος, γίνεται αναφορά στα συμπεράσματα κάθως και σε μελλοντικά σχέδια για βελτιστοποιήση.

Keywords: sentiment analysis; review; tokenize; classification

I. INTRODUCTION

Στην περίοδο που ζούμε, η τεχνολογία έχει μεγάλο αντίκτυπο στη ζωή των ανθρώπων. Τα δεδομένα που προκύπτουν από τους χρήστες είναι πολυάριθμα και δημιουργούνται καθημερινά στα μέσα κοινωνικής δικτύωσης,σε σελίδες για κριτικές προιόντων ή ταινιών, σε forums κ.α. Για αυτούς τους λόγους δεν πρέπει να μείνουν ανεκμετάλλευτα. Έτσι για να κατανοηθούν τα συναισθήματα και οι απόψεις των πελατών για τις προσφερόμενες υπηρεσίες σε αυτούς, κρίνεται απαραίτητη η εύρεση ενός μηχανισμού ο οποίος θα έχει την ικανότητα να προβλέπει είτε τα θετικά είτε τα αρνητικά συναισθήματα των ατόμων. Ο τομέας αυτός ονομάζεται sentiment analysis και έχει σημαντική επιρροή στις διάφορες εταιρίες, καθώς τους δίνεται η δυνατότητα να κατανοήσουν τις απόψεις των πελατών τους με αποτέλεσμα εάν κάτι δεν είναι αρεστό να μπορούν να το αναβαθμίσουν και να το βελτιώσουν. Η ανάλυση συναισθήματος, αναφέρεται στη χρήση της φυσικής γλώσσας(natural language processing) και την ανάλυση κειμένου ώστε να αξιολογηθούν και να εξαχθούν οι απόψεις που προκύπτουν από το κείμενο. Επιπλέον, η ανάλυση συναισθήματος ονομάζεται και αλλίως ως opinion mining καθώς, εκτός από την αναγνώριση της γνώμης και του συναισθήματός,εξάγει και χαρακτηριστικά όπως:εάν η άποψη που εξέφρασε ο ομιλητής είναι θετική ή αρνητική,ποιος εκφράζει την άποψη(άτομο,οντότητα) αλλά και ποιο είναι το γεγονός για το οποίο γίνεται λόγος. Η ανάλυση συναισθήματος έχει πολλές πρακτικές εφαρμογές για αυτό το λόγο,υπήρξε τεράστια αύξηση του ενδιαφέροντος για την έρευνα και ανάπτυξη διαφόρων αναλύσεων και προβλεπτικών τεχνικών ώστε να δημιουργηθούν νέα μοντέλα που θα είναι ακριβέστερα. Μεγάλη είναι η συνεισφορά της ανάλυσης συναισθήματος στα συστήματα συστάσεων, καθώς αναλύοντας και κατηγοριοποιώντας τη γνώμη των ανθρώπων σύμφωνα με τις δικές τους προτιμήσεις και ενδιαφέροντα, το σύστημα μπορεί να προβλέψει ποιο αντικείμενο πρέπει να προταθεί και ποιο όχι. Σχετικά με τη λήψη αποφάσεων , τα συναισθήματα και οι ιδέες των ανθρώπων είναι πολύ σημαντικοί παράγοντες για τη λήψη μιας απόφασης. Κατά την αγορά οποιουδήποτε αντικειμένου είτε είναι βιβλίο ,είτε ρούχα ,είτε ηλεκτρονικό, ο χρήστης διαβάζει τις πρώτες κριτικές για το συγκεκριμένο προϊόν,δηλαδή αυτές που έχουν κατηγοριοποιηθεί ως θετικές. Επιπλέον, η παρακολούθηση ομάδων συζήτησης όπως blogs στα μέσα κοινωνικής δικτύωσης είναι δυνατή με την ανάλυση συναισθημάτων. Μέσα από αυτήν, μπορεί να ανιχνευτούν υβριστικές ή προσβλητικές λέξεις που χρησιμοποιούνται σε tweets, αναρτήσεις ή φόρουμ και ιστολόγια στο Διαδίκτυο. Τα δεδομένα μας προήλθαν από ένα dataset με 50000 κριτικές του IMDB, για το οποίο έχουν εφαρμοστεί πολλές τεχνικές και έχουν ληφθεί πολυάριθμα αποτελέσματα. Στην συγκεκριμένη εργασία, στόχος είναι αρχικά η προεπεξεργασία του συνόλου δεδομένων που έχουμε ώστε στη συνέχεια να μπορέσουμε να εφαρμόσουμε τα διάφορα προβλεπτικά μοντέλα. Αφου γίνει η εφαρμογή θα προχωρήσουμε σε μία σύγκριση των αποτελεσμάτων με τη χρήση κατάλληλων διαγραμμάτων και πάνω στο πιο αποδοτικό μοντέλο θα κάνουμε δοκιμές ξεχωριστών προτάσεων ώστε να διαπιστώσουμε αν η πρόταση έχει θετικό ή αρνητικό συναίσθημα.

II. REASONS FOR CHOOSING THE SOURCES

Προκειμένου να υλοποιήσουμε τα μοντέλα μας στο σύνολο δεδομένων που έχουμε, συνδυάσαμε 2 διαφορετικές έρευνες οι οποίες εφάρμοσαν τεχνικές μηχανικής μάθησης για πρόβλεψη συναισθήματος σε κριτικές ταινιών. Αφού συνδυάσαμε το περιεχόμενο των 2 ερευνών και συγκρίναμε τα αποτελέσματά τους στο δικό μας σύνολο δεδομένων, προχωρήσαμε στην εφαρμογή ενός αναδρομικού νευρωνικού δικτύου LSTM με στόχο την μέγιστη ακρίβεια πρόβλεψης. Οι 2 αυτές πηγές έχουν τους παρακάτω τίτλους:Sentiment Analysis of Movie Reviews using Machine Learning Classifiers και Sentiment Analysis of Movie Reviews using Machine Learning Techniques και προέρχονται από την πλατφόρμα του Google Scholar.

III. PREPROCESSING OF THE DATASET

Πριν ξεκινήσουμε την εφαρμογή των διαφόρων μοντέλων, πρωταρχικής σημασίας είναι η κατανόηση των δεδομένων που διαθέτουμε ώστε να έχουμε μια πλήρη εικόνα και να μπορέσουμε να επιλέξουμε τις κατάλληλες παραμέτρους για τους αλγορίθμους που εφαρμόζουμε.

Μέσω κατάλληλων διαγραμμάτων και με επιλογή ngram = (1,2) βλέπουμε την συχνότητα των διπλότυπων φράσεων του συνόλου δεδομένων μας τόσο για τις θετικές όσο και για τις αρνητικές κριτικές.

	index	frequency
0	br br	4996
1	of the	4140
2	in the	2577
3	the film	1456
4	and the	1431
5	to the	1332
6	this movie	1321
7	this film	1140
8	it is	1115
9	the movie	1082

FIGURE 1. Positive phrases for ngram=(1,2)

	index	frequency
0	br br	5306
1	of the	3669
2	in the	2647
3	this movie	1727
4	to be	1370
5	the film	1360
6	the movie	1327
7	and the	1272
8	to the	1140
9	this film	1080

FIGURE 2. Negative phrases for ngram=(1,2)

Παρακάτω βλέπουμε ένα διάγραμμα που αναπαριστά τον αριθμό των θετικών συναισθημάτων και τον αριθμό των αρνητικών συναισθημάτων κάθε κριτικής.Διαπιστώνουμε ότι ο αριθμός αυτός είναι περίπου ίδιος και περίπου ίσος με 2500 θετικές και 2500 αρνητικές κριτικές.

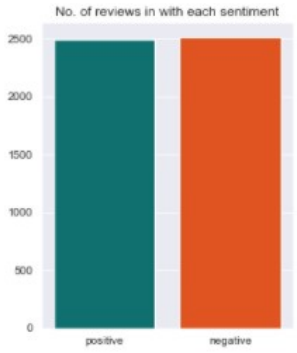

FIGURE 3. number of reviews in each sentiment

Στη συνέχεια, εντοπίσαμε κάποια στατιστικά στοιχεία για τα δεδομένα μας που αφορούν την μέση τιμή του αριθμού των λέξεων κάθε κριτικής, τον αριθμό των stop words κάθε κριτικής που θα αναλυθεί παρακάτω αλλά και την μέση τιμή των γραμμάτων κάθε λέξης στην κριτική. Η ίδια διαδιασία ακολουθήθηκε και για την εύρεση των μεγίστων τιμών για κάθε χαρακτηριστικό που αναφέρθηκε παραπάνω. Έτσι, καταλήξαμε στις εξής παρατηρήσεις:

- Average Word Length: Μέση τιμή = 4.64, Μέγιστη τιμή = 12.29
- Stop Word Count: Μέση τιμή = 105.36, Μέγιστη τιμή = 521
- Word Count: Μέση τιμή = 234.75, Μέγιστη τιμή = 1186

Έτσι, μπορούμε να δούμε ότι υπάρχουν κριτικές με πολυάριθμες λέξεις οι οποίες δεν προσφέρουν κάποιο συναίσθημα και δεν συνεισφέρουν στην πρόβλεψη του(stop words), αλλά και κριτικές με μικρό αριθμό λέξεων με αποτέλεσμα να γίνεται δυσκλότερη η κατανόηση του νοήματος της κάθε κριτικής.

Προχωρώντας, παρέχονται δύο διαγράμματα που δείχνουν το πλήθος των κριτικών που έχουν συγκεκριμένο αριθμό stop words και αντίστοιχα το πλήθος των κριτικών που έχουν συγκεκριμένο αριθμό πλήθους λέξεων. Άρα, μπορούμε να συμπεράνουμε ότι, βλέποντας τα διαγάμματα, κατανοούμε ότι, κατά μέσο όρό υπάρχουν περίπου 70 κριρικές με αριθμό stop words ίσο με 100 λέξεις και ότι υπάρχουν 40 κριτικές με αριθμό συνολικών λέξεων περίπου 200.

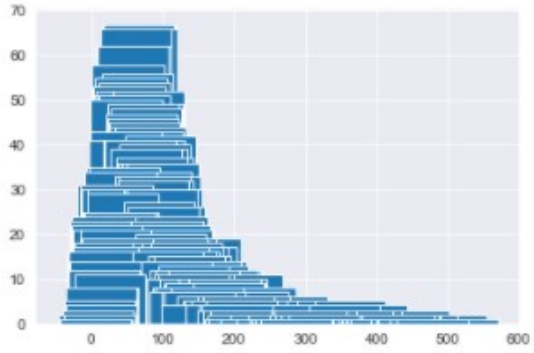

FIGURE 4. Stop Word count diagram

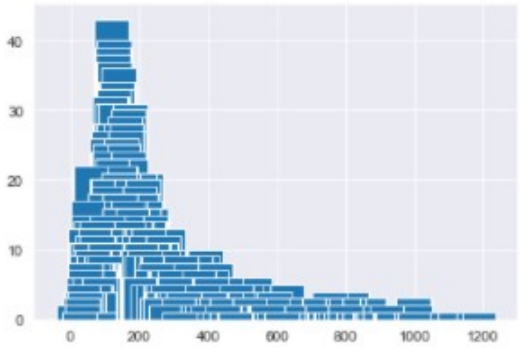

FIGURE 5. Word Count diagram

Η προεπεξεργασία κειμένου είναι μια μέθοδος για τον καθαρισμό των δεδομένων κειμένου και την προετοιμασία για την τροφοδοσία τους στο μοντέλο. Τα δεδομένα κειμένου περιέχουν θόρυβο σε διάφορες μορφές όπως συναισθήματα, σημεία στίξης. Υπάρχουν πολλές βιβλιοθήκες που χρησιμοποιούνται για την αντιμετώπιση προβλημάτων φυσικής γλώσσας. Το NLTK (Natural language toolkit) και το spacy είναι βιβλιοθήκες που χρησιμοποιούνται για την εκτέλεση εργασιών φυσικής γλώσσας, όπως η αφαίρεση stop words κ.λπ.

Πρώτο βήμα είναι η μετατροπή όλων των γραμμάτων σε μικρά καθώς είναι εύκολο για ένα μηχάνημα να ερμηνεύσει τις λέξεις επειδή το πεζό και το κεφαλαίο αντιμετωπίζονται διαφορετικά.

Στην συνέχεια ακολουθεί η αφαίρεση των σημείων στίξης. Αυτή συμβαίνει καθώς τα σημεία στίξης δεν προσδίδουν κάποιο συναίσθημα και μπορούν να μπερδέψουν το μοντελο που εφαρμόζουμε με αποτέλεσμα να έχουμε ανακριβή αποτελέσματα.

Αφού τελειώσει η μεταροπή όλων των γραμμάτων σε πεζά και η αφαίρεση των σημείων στίξης, προχωράμε στην αφαίρεση των stop words. Οι λέξεις αυτές, είναι οι πιο συνηθισμένες λέξεις σε ένα κείμενο που δεν παρέχουν καμία πολύτιμη πληροφορία όπως οι λέξεις αυτές, εκεί, αυτό, πού, κ.λπ. Χρησιμοποιώντας τη βιβλιοθήκη NLTK βλέπουμε ότι υπάρχουν περίπου 180 τέτοιες λέξεις τις οποίες αφαιρούμε από όλες τις προτάσεις του συνόλου δεδομένων μας. Εάν θέλουμε να προσθέσουμε οποιαδήποτε νέα λέξη σε ένα σύνολο λέξεων, τότε είναι εύκολο να χρησιμοποιήσουμε τη μέθοδο προσθήκης. Έτσι μετά από αυτήν την εφαρμογή, κάθε κριτική θα έχει μειωθεί και θα είναι πιο εύκολο να επεξεργαστεί από τα διάφορα μοντέλα που θα εφαρμόσουμε.

Επιπλέον, για να κάνουμε τον υπολογιστή μας να κατανοήσει οποιοδήποτε κείμενο, πρέπει να αναλύσουμε κάθε λέξη με τρόπο που να μπορεί να κατανοήσει το μηχάνημά μας. Εκεί μπαίνει η έννοια του tokenization στην Επεξεργασία Φυσικής Γλώσσας (NLP). Το Tokenization είναι ουσιαστικά ο διαχωρισμός μιας φράσης, μιας πρότασης, μιας παραγράφου ή ενός ολόκληρου εγγράφου κειμένου σε μικρότερες ενότητες, όπως μεμονωμένες λέξεις ή όρους. Κάθε μία από αυτές τις μικρότερες μονάδες λέγεται token. Έτσι θα μπορέσουμε να αναλύσουμε ευκολότερα το κείμενο μέσω της ανάλυσης μεμονομένων λέξεων κατανοώντας το νόημα του κειμένου και αναλύοντας τη σειρά των λέξεων.

Το προελευταίο κομμάτι της διαδικασίας του text processing που ονομάζεται stemming, είναι η μείωση της λέξης στο ριζικό στέλεχος, για παράδειγμα run, running, runs, runed που προέρχεται από την ίδια λέξη με το run. Ουσιαστικά η βασική αρχή είναι να αφαιρέσει το πρόθεμα ή το επίθημα από λέξεις όπως ing, s, es, κ.λπ. Η βιβλιοθήκη NLTK χρησιμοποιείται πάλι σε αυήτν την περίπτωση για να κάνει stemming τις λέξεις. Υπάρχουν διάφοροι τύποι αλγορίθμων στεμμινγ όπως porter stemmer, snowball stemmer, το Porter Stemmer χρησιμοποιείται ευρέως στη βιβλιοθήκη NLTK και είναι ο αλγόριθμος που εφαρμόσαμε στα δεδομένα μας.

Το τελευταίο βήμα για την ολοκλήρωση του preprocessing των δεδομένων μας είναι η εφαρμογή του countvectorizer. Το CountVetorizer είναι ένα εργαλείο που παρέχεται από τη βιβλιοθήκη scikit-learn στην Python. Χρησιμοποιείται για τη μετατροπή ενός δεδομένου κειμένου σε διάνυσμα με βάση τη συχνότητα (πλήθος) κάθε λέξης που εμφανίζεται σε ολόκληρο το κείμενο. Αυτό είναι χρήσιμο όταν έχουμε πολλά τέτοια κείμενα και θέλουμε να μετατρέψουμε κάθε λέξη σε κάθε κείμενο σε διανύσματα (για χρήση σε περαιτέρω ανάλυση κειμένου). Το CountVetorizer δημιουργεί έναν πίνακα στον οποίο κάθε μοναδική λέξη αντιπροσωπεύεται από μια στήλη του πίνακα και κάθε δείγμα κειμένου από το έγγραφο είναι μια γραμμή στον πίνακα. Η τιμή κάθε κελιού δεν

είναι παρά η μέτρηση της λέξης στο συγκεκριμένο δείγμα κειμένου. Η χρήση του countvectorizer μπορεί να βοηθήσει στην αποτελεσματική ελαχιστοποίηση του χρόνου εκτέλεσης του κώδικα.

Τέλος, παρακάτω παραθέτουμε 2 εικόνες των πιο συχνών λέξεων των δεδομένων μας πριν και μετά την προεπεξεργασία.

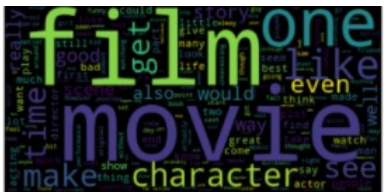

FIGURE 6. Word Cloud diagram

FIGURE 7. Word Cloud diagram

IV. ALGORITHMS

Μετά την προεπεξεργασία των δεδομένων μας, κάναμε χρήση των παρακάτω αλγορίθμων ώστε να μπορέσουμε να προβλέψουμε είτε τα θετικά, είτε τα αρνητικά συναισθήματα της κάθε κριτικής:

- Logistic Regression
- SGD Classifier
- Multinomial Naive Bayes
- Random Forest
- Recurrent Neural Network(LSTM)

Παρακάτω θα γίνει μία συνοπτική επεξήγηση κάθε μοντέλου που εφαρμόσαμε.

IV.I. Logistic Regression

Στην επεξεργασία φυσικής γλώσσας, ο logistic regression είναι ένας αλγόριθμος μηχανικής μάθησης για ταξινόμηση και έχει επίσης πολύ στενή σχέση με τα νευρωνικά δίκτυα. Κυρίως, χρησιμοποιείται για την πρόβλεψη ενός δυαδικού αποτελέσματος (όπως 0 / 1, Λάθος / Σωστό, Όχι / Ναι) όταν δίνεται ένα σύνολο από ανεξάρτητες μεταβλητές. Γενικά περιγράφει την πιθανότητα ενός αποτελέσματος 0 ή 1 με μια logistic συνάρτηση S-shaped. Η Logistic Regression είναι πολύ αποτελεσματική σε δεδομένα κειμένου και ο υποκείμενος αλγόριθμος είναι επίσης αρκετά εύκολο να κατανοηθεί.

IV.II. Naive Bayes

Ο πολυωνυμικός αλγόριθμος Naive Bayes είναι μια πιθανολογική μέθοδος εκμάθησης που χρησιμοποιείται κυρίως στην Επεξεργασία Φυσικής Γλώσσας (NLP). Ο αλγόριθμος βασίζεται στο θεώρημα του Bayes και προβλέπει την ετικέτα ενός κειμένου. Υπολογίζει την πιθανότητα κάθε ετικέτας για ένα δεδομένο δείγμα και στη συνέχεια δίνει την ετικέτα με την υψηλότερη πιθανότητα ως έξοδο. Αυτό το μοντέλο είναι εύκολο στην κατασκευή και ιδιαίτερα χρήσιμο για πολύ μεγάλα σύνολα δεδομένων.

IV.III. Random Forest

Είναι μία μέθοδος για ταξινόμηση και παλινδρόμιση. Κατασκευάζει μια σειρά από δέντρα απόφασης κατά την διάρκεια της εκπαίδευσης. Για να ταξινομήσει μία νέα περίπτωση, στέλνει τη νέα υπόθεση σε καθένα από τα δέντρα που έχουν δημιουργηθεί. Κάθε δέντρο εκτελεί ταξινόμηση και βγάζει μια κλάση. Η κλάση εξόδου επιλέγεται με βάση την πλειοψηφία που είναι ο μέγιστος αριθμός που δημιουργείται από διάφορα δέντρα. Ένα ισχυρό πλεονέκτημα του Random Forest είναι ότι μειώνει το overfitting στα δέντρα αποφάσεων και συμβάλλει στη βελτίωση της ακρίβειας.

IV.IV. SGD Classifier

Αυτό το μοντέλο είναι ένας γραμμικός ταξινομητής που χρησιμοποιεί τη stochastic gradient descent κατά τη διάρκεια της εκπαίδευσης. Ο ταξινομητής SGD εφαρμόζει βασικά μια απλή διαδικασία εκμάθησης SGD που υποστηρίζει διάφορες loss functions και penalties για ταξινόμηση. Το συγκεκριμένο μοντέλο είναι υπολογιστικά γρήγορο καθώς μόνο ένα δείγμα υποβάλλεται σε επεξεργασία κάθε φορά. Για μεγαλύτερα σύνολα δεδομένων, μπορεί να συγκλίνει πιο γρήγορα, καθώς προκαλεί συχνότερες ενημερώσεις στις παραμέτρους.

IV.V. LSTM

Ένας καλός λόγος για να χρησιμοποιήσουμε το LSTM είναι ότι είναι αποτελεσματικό στην απομνημόνευση σημαντικών πληροφοριών. Αν δούμε και άλλες τεχνικές ταξινόμησης μη νευρωνικών δικτύων, αυτές εκπαιδεύονται σε πολλές λέξεις ως ξεχωριστές εισόδους που είναι απλώς μια λέξη που δεν έχει πραγματικό νόημα ως πρόταση, και ενώ προβλέπει την κλάση θα δώσει την έξοδο σύμφωνα με στατιστικά στοιχεία και όχι σύμφωνα με το νόημα. Αυτό σημαίνει ότι κάθε λέξη ταξινομείται σε μία από τις κατηγορίες. Αυτό δεν είναι το ίδιο στο LSTM. Στο LSTM μπορούμε να χρησιμοποιήσουμε μια συμβολοσειρά πολλαπλών λέξεων για να μάθουμε την κλάση στην οποία ανήκει. Αυτό είναι πολύ χρήσιμο όταν εφαρμόζουμε επεξεργασία φυσικής γλώσσας. Εάν χρησιμοποιήσουμε κατάλληλα επίπεδα embedding και encoding στο LSTM, το μοντέλο θα μπορεί να ανακαλύψει την πραγματική σημασία στη συμβολοσειρά εισόδου και να δώσει την πιο ακριβή κλάση εξόδου.

V. MODEL EVALUATION

Κάθε μοντέλο παρέχει πολυάριθμες παραμέτρους, ώστε να έχουμε όσο δυνατόν καλύτερο αποτέλεσμα πρόβλεψης. Για αυτόν τον λόγο μετά από πολλές δοκιμές καταλήξαμε στις παρακάτω παραμέτρους για κάθε μοντέλο.

- Logistic Regression: i)Solver = lbfgs, ii)C = 0.1, iii)random state=42, iv)max iter=1000, v)overbose = 1
- SGD Classifier: i)loss = perceptron, ii)max iter = 1000, iii)random state = 42
- Random Forest: i)n estimators = 200, ii)criterion = entropy
- LSTM: Χρησιμοποιήθηκαν τα εξής layers:i)Embedding, ii)Dense. Ως output layer χρησιμοποιήθηκε ένα dense layer ενός νευρώνα με activation function την sigmoid.

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_2 (Embedding)     (None, 120, 100)          117379800

 bidirectional_2 (Bidirectio  (None, 128)              84480
 nal)

 dense_4 (Dense)             (None, 24)                3096

 dense_5 (Dense)             (None, 1)                 25

=================================================================
Total params: 117,467,401
Trainable params: 117,467,401
Non-trainable params: 0
_________________________________________________________________
```

FIGURE 8. LSTM Summary Report

Μετά την εφαρμογή των διαφόρων προβλεπτικών μοντέλων παρατηρούμε την ακρίβεια τους και μπορούμε να κάνουμε μια σύγκριση αυτών. Βλέπουμε ότι καλύτερη ακρίβεια επιτυγχάνεται με την εφαρμογή του Recurrent neural network LSTM, η οποία πλησίασε το 79%. Στη συνέχεια, ακολουθούν τα μοντέλα Multinomial Naive bayes και logistic regression με ακρίβεια 71% και 70%. Ενώ τέλος, οι αλγόριθμοι SGD Classifier και Random Forest με ακρίβεια 67% και 62%. Έτσ, μπορούμε να διαπιστώσουμε με μεγάλη ευκολία ότι το LSTM δίνει με διαφορά την καλύτερη ακρίβεια πρόβλεψης.

Παρακάτω παρουσιάζουμε αναύτικά το classification report κάθε μοντέλου που εφαρμόσαμε.

	precision	recall	f1-score	support
-1	0.65	0.81	0.72	528
1	0.76	0.58	0.66	541
accuracy			0.69	1069
macro avg	0.71	0.70	0.69	1069
weighted avg	0.71	0.69	0.69	1069

FIGURE 9. Logistic Regression Classification Report

	precision	recall	f1-score	support
-1	0.68	0.73	0.70	528
1	0.71	0.66	0.69	541
accuracy			0.69	1069
macro avg	0.69	0.69	0.69	1069
weighted avg	0.69	0.69	0.69	1069

FIGURE 10. SGD Classification Report

	precision	recall	f1-score	support
-1	0.68	0.73	0.70	528
1	0.72	0.66	0.69	541
accuracy			0.70	1069
macro avg	0.70	0.70	0.70	1069
weighted avg	0.70	0.70	0.70	1069

FIGURE 11. Multinomial Naive Bayes Report

	precision	recall	f1-score	support
-1	0.75	0.46	0.57	528
1	0.62	0.85	0.71	541
accuracy			0.66	1069
macro avg	0.68	0.65	0.64	1069
weighted avg	0.68	0.66	0.64	1069

FIGURE 12. Random Forest Classification Report

Ένα ιδανικό σύστημα έχει υψηλό precision και υψηλό recall καθώς θα επιστρέψει αποτελέσματα, τα οποία κατηγοριοποιούνται ως σωστά. Χαμηλό recall σημαίνει ότι οι περισσότερες από τις θετικές τιμές δεν προβλέπονται ποτέ. Ενώ υψηλό precision σημαίνει ότι οι θετικές κριτικές κατηγοριοποιούνται σωστά ως θετικές στο σύνολο των θετικών κριτικών.

Παρακάτω παρατίθεται ένα συγκριτικό διάγραμμα των ακριβειών πρόβλεψης κάθε μεθόδου.

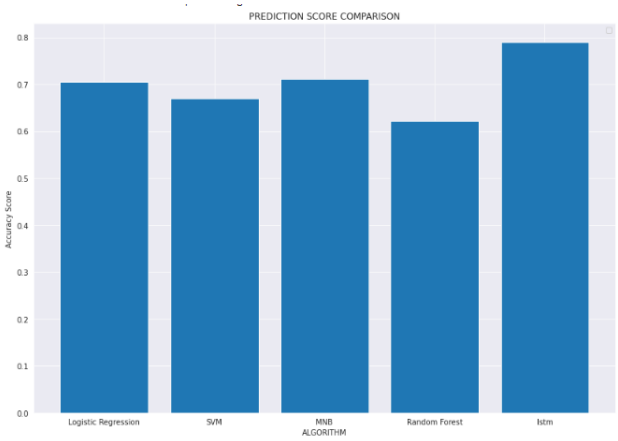

FIGURE 13. Accuracy Comparison

VI. PROBLEMS WE FACED

Κατά τη διάρκεια εκτέλεσης των αλγορίθμων μας, διαπιστώσαμε ότι ο χρόνος που απαιτούνταν για να λάβουμε τα αποτελέσματα που θέλαμε ήταν μεγάλος. Για αυτόν τον λόγο,επιλέξαμε να εφαρμόσουμε τόσο την πρεπεξεργασία των δεδομένων μας, όσο και τα διάφορα προβλεπτικά μοντέλα που επιλέξαμε σε ένα δείγμα του dataset μας ώστε να εξοικονομήσουμε τόσο χρόνο όσο και πόρους από το σύστημα μας. Σημαντικό ρόλο έπαιξε η αλλαγή της παραμέτρου ngram σε (1,2) στον count vectorizer καθώς αρχικά είχαμε επιλέξει ngram = (1,3). Τα μοντέλα μας ήταν οριακά πιο αποδοτικά, όμως απαιτούνταν αρκετός χρόνος εκτέλεσης.Έτσι, μετά απο αυτήν την μετατροπή και τον περιορισμό του dataset, λάβαμε τα αποτελέσματα που παρουσιάστηκαν παραπάνω.

VII. CONSCLUSIONS AND FUTURE WORK

Αρχικά, συγκρίνοντας τις 2 έρευνες, παρατηρούμε ότι και οι 2 εφαρμόζουν sentiment analysis με την χρήση μοντέλων κατηγοριοποίησης για την πρόβλεψη του συναισθήματος μιας κριτικής.Στο δικό μας σύνολο δεδομένων, οι αλγόριθμοι κατηγοριοποίησης των ερευνών πετύχαιναν ακρίβεια που δεν ξεπέρασε το 71% οπότε, δημιουργήσαμε ένα αναδρομικό νευρωνικό δίκτυο για να την βελτιστοποιήσουμε.Η ακρίβεια που επιτεύχθηκε στο συγκεκριμένο μοντέλο άγγιξε το 79%. Τέλος, στο γραφικό περιβάλλον που υλοποιήσαμε, ο χρήστης έχει την δυνατότητα να επιλέξει ποιον αλγόριθμο θα εφαρμόσει σε κάποια νέα κριτική και να εξάγει το συναίσθημα της, ενώ ταυτόχρονα παρέχονται και συγκριτικά διαγράμματα της απόδοσης των αλγορίθμων πάνω στο δικό μας dataset καθώς και classification reports. Ιδανικά, ως μελλοντικό σχέδιο προτείνουμε την ανάπτυξη ενός γραφικού περιβάλλοντος το οποίο θα δίνει τη δυνατότητα στο χρήστη να επιλέγει αυτός τις παραμέτρους για κάθε αλγόριθμο σε δικό του dataset και να υλοποιεί δικά του πειράματα λαμβάνοντας συγκριτικά διαγράμματα ώστε να μπορεί να επιλέξει το βέλτιστο.

REFERENCES

[1] Palak Baid, Neelam Chaplot, Apoorva Gupta, "Sentiment Analysis of Movie Reviews using Machine Learning Techniques",Department of Computer Science Engineering Jaipur Engineering College and Research Center Jaipur, Rajasthan, India, 2017.

[2] Mamtesh, Seema Mehla, "Sentiment Analysis of Movie Reviews using Machine Learning Classifiers",National Institute of Technology Kurukshetra, India,2019.