ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη Συστήματος Ανάλυσης Συναισθήματος με Χρήση Επεξεργασίας Φυσικής Γλώσσας

(Natural Language Processing - NLP)

ΝΑΪΝΤΕΝΟΒΑ ΒΛΑΝΤΙΜΙΡΑ

Α.Μ. 18390186

Εισηγητής: ΧΡΗΣΤΟΣ, ΤΡΟΥΣΣΑΣ, ΕΠ. ΚΑΘΗΓΗΤΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάπτυξη Συστήματος Ανάλυσης Συναισθήματος με Χρήση Επεξεργασίας Φυσικής Γλώσσας

(Natural Language Processing - NLP)

ΝΑΪΝΤΕΝΟΒΑ ΒΛΑΝΤΙΜΙΡΑ

Α.Μ. 18390186

Εισηγητής:

ΧΡΗΣΤΟΣ, ΤΡΟΥΣΣΑΣ, ΕΠ. ΚΑΘΗΓΗΤΗΣ

Εξεταστική Επιτροπή:

ΧΡΗΣΤΟΣ, ΤΡΟΥΣΣΑΣ, ΕΠ. ΚΑΘΗΓΗΤΗΣ

Christos Troussas Digitally signed by Christos Troussas Date: 2024.07.15 11:21:03 +03'00'

ΑΚΡΙΒΗ, ΚΡΟΥΣΚΑ, ΜΕΛΟΣ ΕΔΙΠ

Akrivi Krouska Digitally signed by Akrivi Krouska Date: 2024.07.15 14:01:01 +03'00'

ΠΑΝΑΓΙΩΤΑ, ΤΣΕΛΕΝΤΗ, ΜΕΛΟΣ ΕΔΙΠ

Panagiota Tselenti Digitally signed by Panagiota Tselenti Date: 2024.07.15 22:03:46 +03'00'

Ημερομηνία εξέτασης

ΙΟΥΛΙΟΣ 2024

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΜΕΤΑΠΤΥΧΙΑΚΗΣ ΕΡΓΑΣΙΑΣ

Ο/η κάτωθι υπογεγραμμένη Βλαντιμίρα Ναϊντένοβα του Βλαντιμίροβα, με αριθμό μητρώου 18390186 φοιτήτρια του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών της Σχολής Μηχανικών του Πανεπιστημίου Δυτικής Αττικής, δηλώνω ότι:

«Βεβαιώνω ότι είμαι συγγραφέας της παρούσας διπλωματικής εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτή, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για την συγκεκριμένη διπλωματική εργασία»

Ο/Η Δηλών/ούσα

ΕΥΧΑΡΙΣΤΙΕΣ

Αυτή η διπλωματική εργασία αποτελεί το αποτέλεσμα της αφοσιωμένης μου προσπάθειας και πάθους, αναδεικνύοντας έναν κόσμο γνώσης που με συναρπάζει. Την προσπάθειά μου αυτή υποστήριξε ο επιβλέπων καθηγητής μου, τον οποίο θα ήθελα να ευχαριστήσω θερμά.

Τέλος, θα ήθελα να ευχαριστήσω τους συναδέλφους και φίλους μου από τη σχολή για τις εποικοδομητικές συζητήσεις και τον χρόνο που περάσαμε μαζί, και κυρίως την οικογένειά μου, που είναι πάντα δίπλα μου στηρίζοντας με σε κάθε μου βήμα.

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία ασχολείται με ανάπτυξη ενός συστήματος που αξιοποιεί της τεχνικές της Επεξεργασίας Φυσικής Γλώσσας (NLP) για να επιτευχθεί συναισθηματική ανάλυση σε δεδομένα κειμένου. Διερευνώνται οι προσεγγίσεις της συναισθηματικής ανάλυσης, όπως αυτές της βασισμένης σε λεξικό και σε αλγορίθμους μηχανικής μάθησης. Αναλύονται οι έννοιες της της Επεξεργασίας Φυσικής Γλώσσας και της συναισθηματικής ανάλυσης, όπως επίσης η σημασία και οι χρήσεις τους στο χρονικό πλαίσιο που διανύουμε. Επιπλέον, αναφέρονται οι κυριότερες μέθοδοι που χρησιμοποιούνται για την επίτευξη της ανάλυσης συναισθήματος. Ακόμη, γίνεται επεξηγείται η μεθοδολογία που ακολουθήθηκε αλλά και οι τεχνικές σχεδιασμού του συστήματος. Παρουσιάζεται ο κώδικας και επεξηγείται, ώστε να γίνει κατανοητή οι προσεγγίσεις που χρησιμοποιήθηκαν. Για την ανάλυση συναισθήματος χρησιμοποιηθήκαν δύο προσεγγίσεις, η βασισμένη σε λεξικό που αξιοποίησε το VADER, ένα module του NLTK, και η προσέγγιση μέσω μηχανικής μάθησης και συγκεκριμένα επιβλεπόμενης, με τα μοντέλα του Naive Bayes, Supervised Vector Machine (SVM) και του Decision Tree.

Τα στοιχεία στα οποία επικεντρώνεται η παρούσα διπλωματική εργασία είναι η κατανόηση των τεχνολογιών που χρησιμοποιούνται, ο τρόπος με τον οποίο έχει δομηθεί η εφαρμογή και η μεθοδολογία που ακολουθήθηκε. Τέλος, παρουσιάζεται η αξιολόγηση του συστήματος, τα τελικά συμπεράσματα και ο πηγαίος κώδικας.

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Επεξεργασία Φυσικής Γλώσσας, Ανάλυση Συναισθήματος

ABSTRACT

This thesis deals with the development of a system that utilizes Natural Language Processing Techniques (NLP) to achieve sentiment analysis in text data. Approaches to sentiment analysis, such as those of dictionary-based and machine learning algorithms, are explored. The concepts of natural language processing and sentiment analysis are analyzed, as well as their importance and use in the current time frame. In addition, the main methods used to achieve sentiment analysis are mentioned. The methodology followed and the design techniques of the rectum are also explained. The code is presented and explained to understand the approaches used. Two approaches, a dictionary based with VADER, an NLTK module, and the approach to machine learning, and in particular supervised, with Naive Bayes, Supervised Vector Machine (SVM), and Decision Tree models were used to analyze emotion.

The elements focused on in this thesis are the understanding of the technologies used, how the application and the methodology followed have been structured. Finally, the system evaluation, the final conclusions, and the source code are presented.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΣΥΝΤΟΜΟΓΡΑΦΊΕΣ

SA Sentiment Analysis

NLP Natural Language Processing

ML Machine Learning

AI Artificial Intelligence

SA Sentiment Analysis

JSON JavaScript Object Notation

NLTK Natural Language ToolKit

NB Naive Bayes

LR Logistic Regression

SVM Supervised Vector Machine

DT Decision Tree

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή του αντικειμένου της διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία σηματοδοτεί μια σημαντική εξερεύνηση στον περίπλοκο τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP), με κύρια εστίαση στην ανάπτυξη ενός εξελιγμένου συστήματος ανάλυσης συναισθήματος (sentiment analysis). Ο κεντρικός στόχος της παρούσας διατριβής είναι να εξεταστούν και να εφαρμοστούν τεχνικές που αξιοποιούν τη δύναμη της NLP για τη σχολαστική ανάλυση κειμενικού περιεχομένου, με ιδιαίτερη έμφαση στις περίπλοκες αποχρώσεις της αναγνώρισης συναισθήματος.

Σε μια εποχή κορεσμένη από ροές κειμένου, η διάκριση των συναισθημάτων από τον γραπτό λόγο αποκτά βαθιά σημασία σε διάφορους τομείς, όπως τα σχόλια των πελατών στο ηλεκτρονικό εμπόριο, οι κριτικές των ασθενών στην υγειονομική περίθαλψη και οι δημόσιες γνώμες στα μέσα κοινωνικής δικτύωσης. Το σχεδιαζόμενο σύστημα ανάλυσης συναισθήματος φιλοδοξεί να φωτίσει τα συναισθηματικά περιγράμματα μέσα στα δεδομένα κειμένου, διευκολύνοντας τις διαφοροποιημένες ερμηνείες και τη λήψη τεκμηριωμένων αποφάσεων, που είναι ζωτικής σημασίας όχι μόνο για την κατανόηση των συναισθημάτων των χρηστών αλλά και για τη διαμόρφωση στρατηγικών σε τομείς που κυμαίνονται από τις επιχειρήσεις και την υγειονομική περίθαλψη έως τη συμμετοχή στα μέσα κοινωνικής δικτύωσης.

Το πεδίο εφαρμογής αυτής της μελέτης περιλαμβάνει την ολιστική αντίληψη ενός συστήματος ανάλυσης συναισθήματος, ενορχηστρώνοντας τη συμφωνία των τεχνικών NLP για γλωσσική επεξεργασία. Ιδιαίτερη έμφαση θα δοθεί στην επιμέλεια και την προετοιμασία συνόλων δεδομένων, τη διερεύνηση μοντέλων NLP και την αξιολόγηση της αποτελεσματικότητας του συστήματος έναντι διαφόρων κριτηρίων αναφοράς.

1.2 Σκοπός και στόχοι

Ο πρωταρχικός στόχος αυτής της ακαδημαϊκής προσπάθειας είναι να εμβαθύνει στις πτυχές της γλωσσικής επεξεργασίας, ξεδιπλώνοντας τις ιδιαιτερότητες που διέπουν την κατανόηση της γλώσσας. Επιπλέον, η παρούσα εργασία αποσκοπεί στη εξέταση και αξιολόγηση των ποικίλων μεθοδολογιών που χρησιμοποιούνται στην ανάπτυξη ενός συστήματος ανάλυσης συναισθήματος. Η έρευνα αυτή επιδιώκει να συμβάλει όχι μόνο στη θεωρητική κατανόηση της γλωσσικής επεξεργασίας αλλά και στην

πρακτική εφαρμογή αυτής της γνώσης στη δημιουργία ενός προηγμένου συστήματος ανάλυσης συναισθήματος. Επίσης περιλαμβάνει μια εις βάθος εξέταση των τεχνικών και αλγορίθμων αιχμής που έχουν καθοριστική σημασία για την ισχυρή κατασκευή ενός συστήματος ανάλυσης συναισθήματος.

Παρακάτω περιγράφονται οι θεμελιώδεις στόχοι που καθοδηγούν την προσπάθεια αυτή.

- Ανάπτυξη του συστήματος: Ο πρωταρχικός στόχος είναι η σχολαστική ανάπτυξη ενός συστήματος ανάλυσης συναισθήματος, με τη χρήση τεχνικών NLP για την ενίσχυση της ικανότητας του συστήματος να διακρίνει και να κατηγοριοποιεί συναισθήματα από διάφορες πηγές κειμένου.
- Επιλογή και προεπεξεργασία συνόλου δεδομένων: Επιμέλεια και προετοιμασία συνόλου δεδομένων, εξασφαλίζοντας μια ολοκληρωμένη αξιολόγηση της προσαρμοστικότητας και της αποτελεσματικότητας του συστήματος ανάλυσης συναισθήματος.
- Εξερεύνηση μοντέλων NLP: Μελέτη και ανάπτυξη σύγχρονων μοντέλων NLP για να διαπιστωθεί ο αντίκτυπός τους στην ακρίβεια της ανάλυσης συναισθήματος, προωθώντας τη βαθύτερη κατανόηση της δυνατότητας εφαρμογής και των περιορισμών τους.
- Αξιολόγηση επιδόσεων: Αυστηρή αξιολόγηση του συστήματος ανάλυσης συναισθήματος που αναπτύχθηκε, μετρώντας την αποτελεσματικότητά του σε σύγκριση με καθιερωμένες μεθοδολογίες.

Η έρευνα αυτή ευθυγραμμίζεται όχι μόνο με τους ακαδημαϊκούς στόχους αλλά και με ένα βαθύ προσωπικό ενδιαφέρον για την προώθηση της γνώσης στα δυναμικά πεδία της μηχανικής μάθησης και τους παράγωγους υποκλάδους της, όπως αυτός της NLP.

1.3 Μεθοδολογία και Δομή της Διπλωματικής Εργασίας

Το κείμενο είναι δομημένο σε πέντε κεφάλαια ως εξής:

- Κεφάλαιο 1: Εισαγωγή
  Το εισαγωγικό κεφάλαιο που παρουσιάζει το αντικείμενο της διπλωματικής εργασίας, καθορίζει τον σκοπό και τους στόχους αυτής και παρουσιάζει την δομή της.

- Κεφάλαιο 2: Επισκόπηση του θεωρητικού πεδίου
  Στο επόμενο αυτό κεφάλαιο εμβαθύνει σε θεωρητικό επίπεδο στις εφαρμογές μηχανικής μάθησης, τις προκλήσεις και τις αναδυόμενες τάσεις στην ανάλυση συναισθήματος.

- Κεφάλαιο 3: Μεθοδολογία και εργαλεία
  Παρουσιάζοντας λεπτομερώς την επιλεγμένη μεθοδολογία και τα εργαλεία, το κεφάλαιο αυτό

δίνει έμφαση στη συστηματική προσέγγιση για την επιλογή και την επεξεργασία συνόλων δεδομένων, παρέχοντας τα θεμέλια για τη μετέπειτα τεχνική υλοποίηση.

- Κεφάλαιο 4: Τεχνική υλοποίηση
  Γεφυρώνοντας τη θεωρία και την πράξη, το κεφάλαιο αυτό παρουσιάζει τις πρακτικές πτυχές της ανάπτυξης του συστήματος ανάλυσης συναισθήματος

- Κεφάλαιο 5: Αποτελέσματα και Συμπεράσματα
  Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα του συστήματος που αναπτύχθηκε, και αξιολογείται κριτικά η απόδοσή του σε σχέση με προκαθορισμένα σημεία αναφοράς. Επιπλέον καταλήγει σε συμπεράσματα, παρουσιάζει προβληματισμούς σχετικά με τις επιπτώσεις τους και παρέχει συστάσεις για μελλοντική έρευνα στους τομείς της ανάλυσης συναισθήματος και της NLP.

ΚΕΦΑΛΑΙΟ 2

ΕΠΙΣΚΟΠΗΣΗ ΘΕΩΡΗΤΙΚΟΥ ΠΕΔΙΟΥ

2.1 Εισαγωγή κεφαλαίου

Το δεύτερο κεφάλαιο της διπλωματικής εργασίας επικεντρώνεται στη θεωρητική ανάλυση του πεδίου της ανάλυσης συναισθημάτων. Κατά τη διάρκεια αυτού του κεφαλαίου, θα εξετάσουμε βασικές έννοιες που αφορούν τον χώρο της ανάλυσης συναισθημάτων, όπως ορισμοί, τύποι συναισθημάτων και προκλήσεις που αντιμετωπίζονται κατά την εκτέλεση αυτού του είδους αναλύσεων. Επιπλέον, θα εξετάσουμε τις διαδικασίες προεπεξεργασίας κειμένου που απαιτούνται για την αποτελεσματική ανάλυση συναισθημάτων και θα εισαγάγουμε βασικές έννοιες της μηχανικής μάθησης που χρησιμοποιούνται για την κατηγοριοποίηση κειμένου. Η κατανόηση αυτών των θεμελιωδών αρχών αποτελεί τη βάση για την εξέταση των προτεινόμενων μοντέλων και των αποτελεσμάτων τους στο επόμενο κεφάλαιο.

2.2 Επεξεργασία Φυσικής Γλώσσας - Natural Language Processing (NLP)

2.2.1 Η σημασία της NLP και η χρήση της

Η Επεξεργασία Φυσικής Γλώσσας (NLP) είναι το πεδίο του σχεδιασμού μεθόδων και αλγορίθμων που είτε λαμβάνουν ως είσοδο είτε παράγουν ως έξοδο μη δομημένα δεδομένα φυσικής γλώσσας (Goldberg, 2017). Σύμφωνα με τον Goldberg (2017), ενώ οι άνθρωποι διαπρέπουν στη χρήση της γλώσσας, στην έκφραση, στην κατανόηση και στην ερμηνεία σύνθετων νοημάτων, αγωνίζονται να διατυπώσουν επίσημα τους κανόνες που διέπουν τη γλώσσα.

Η επεξεργασία φυσικής γλώσσας συνδυάζει την υπολογιστική γλωσσολογία - μοντελοποίηση της ανθρώπινης γλώσσας βάσει κανόνων - με στατιστικά μοντέλα, μοντέλα μηχανικής μάθησης και βαθιάς μάθησης. Μαζί, αυτές οι τεχνολογίες επιτρέπουν στους υπολογιστές να επεξεργάζονται την ανθρώπινη γλώσσα με τη μορφή κειμένου ή φωνητικών δεδομένων και να "κατανοούν" το πλήρες νόημά της, μαζί με την πρόθεση και το συναίσθημα του ομιλητή ή του συγγραφέα. (What Is Natural Language Processing? | IBM, n.d.)

Στην ψηφιακή μας εποχή, τα δεδομένα είναι πανταχού παρόντα, με ένα σημαντικό μέρος να είναι η μη δομημένη ανθρώπινη γλώσσα. Το NLP παρέχει ένα μέσο κατανόησης και αξιοποίησης αυτού του τεράστιου όγκου γλωσσικών δεδομένων. Επιτρέπει στις μηχανές να κατανοούν το συναίσθημα πίσω

από τις αναρτήσεις στα μέσα κοινωνικής δικτύωσης, να μεταφράζουν κείμενο ή ομιλία από τη μια γλώσσα στην άλλη, να συνοψίζουν μεγάλα έγγραφα και πολλά άλλα. Το NLP έχει πολυάριθμες εφαρμογές σε διάφορους τομείς, όπως η υγειονομική περίθαλψη, τα οικονομικά, η εξυπηρέτηση πελατών και το μάρκετινγκ.

2.2.2 Βασικές αρχές και εφαρμογές

Το NLP περιλαμβάνει ένα ευρύ φάσμα θεμελιωδών εργασιών και εφαρμογών. Μερικές από τις θεμελιώδεις εργασίες στο NLP περιλαμβάνουν:

- Ανάλυση συναισθήματος (Opinion Mining): Προσδιορίζει και ταξινομεί το συναίσθημα που εκφράζεται σε ένα κείμενο, θετικό, αρνητικό ή ουδέτερο. Εφαρμόζεται σε τομείς όπως η παρακολούθηση μέσων κοινωνικής δικτύωσης, η διαχείριση εταιρικής φήμης και η ανάλυση σχολίων πελατών.
- Σύνοψη κειμένου: Χρησιμοποιεί τεχνικές NLP για την επεξεργασία μεγάλων όγκων ψηφιακού κειμένου και τη δημιουργία συνοπτικών περιλήψεων. Αυτές οι περιλήψεις είναι χρήσιμες για τη δημιουργία δεικτών, ερευνητικών βάσεων δεδομένων ή για την εξυπηρέτηση αναγνωστών με περιορισμένο χρόνο που προτιμούν να μην διαβάσουν ολόκληρο το κείμενο.
- Μηχανική μετάφραση: Περιλαμβάνει την αυτόματη μετάφραση κειμένου από τη μια γλώσσα στην άλλη. Είναι κρίσιμο για πολλές εφαρμογές, ειδικά στον ολοένα και πιο παγκοσμιοποιημένο κόσμο μας, συμπεριλαμβανομένης της μετάφρασης ιστοτόπων, της μετάφρασης εγγράφων και της μετάφρασης σε πραγματικό χρόνο σε εφαρμογές επικοινωνίας.
- Αναγνώριση ομιλίας: Μετατρέπει την προφορική γλώσσα σε γραπτό κείμενο. Εφαρμόζεται σε διάφορες εφαρμογές, συμπεριλαμβανομένων των βοηθών που ενεργοποιούνται με φωνή (π.χ. Siri, Alexa ή Google Assistant) και συστημάτων με φωνητικό έλεγχο.
- Αναγνώριση επώνυμης οντότητας (NER): Προσδιορίζει και κατηγοριοποιεί επώνυμες οντότητες σε ένα κείμενο σε προκαθορισμένες κατηγορίες, όπως ονόματα ατόμων, οργανισμών, τοποθεσίες, χρονικές εκφράσεις, ποσότητες, χρηματικές αξίες κ.λπ. Θεμελιώδης για πολλές εφαρμογές NLP, όπως η απάντηση σε ερωτήσεις και η αυτόματη μετάφραση .

2.2.3 Προκλήσεις στην Επεξεργασία Φυσικής Γλώσσας

Παρά τις εξελίξεις στην Επεξεργασία Φυσικής Γλώσσας (NLP), πολλές προκλήσεις εξακολουθούν να υφίστανται, συμβάλλοντας στην πολυπλοκότητα της αποτελεσματικής επεξεργασίας και κατανόησης της ανθρώπινης γλώσσας.

- Αμφισημία και κατανόηση των συμφραζομένων

  Η ανθρώπινη γλώσσα είναι εγγενώς διφορούμενη, στηρίζεται στο πλαίσιο για την ερμηνεία. Οι λέξεις έχουν συχνά πολλαπλές σημασίες και η ίδια λέξη μπορεί να μεταφέρει διαφορετικές αποχρώσεις με βάση το περιβάλλον. Η επίλυση αυτής της ασάφειας είναι μια σημαντική πρόκληση στο NLP, καθώς οι μηχανές πρέπει να κατανοούν με ακρίβεια το επιδιωκόμενο νόημα μέσα σε διαφορετικά πλαίσια.

- Χειρισμός άτυπης γλώσσας

  Η φυσική γλώσσα είναι πλούσια με ανεπίσημες εκφράσεις, καθομιλουμένες, αργκό και πολιτισμικές αναφορές. Η κατανόηση και η επεξεργασία της άτυπης γλώσσας αποτελεί πρόκληση για τα συστήματα NLP, ειδικά όταν έχουμε να κάνουμε με περιεχόμενο που δημιουργείται από χρήστες σε πλατφόρμες κοινωνικών μέσων. Η προσαρμογή στη δυναμική φύση της γλωσσικής εξέλιξης είναι ζωτικής σημασίας για την αποτελεσματική επικοινωνία.

- Έλλειψη Κοινής Γνώσης

  Η ανθρώπινη γλώσσα συχνά βασίζεται σε κοινές γνώσεις και εμπειρίες. Τα συστήματα NLP μπορεί να δυσκολεύονται όταν έρχονται αντιμέτωπα με πληροφορίες ή αναφορές έξω από την προϋπάρχουσα βάση γνώσεων τους. Η γεφύρωση αυτού του χάσματος απαιτεί μηχανισμούς για τη συνεχή απόκτηση, ενημέρωση και ενσωμάτωση της γνώσης.

- Βασική ασάφεια και ανάλυση συνάφειας

  Η κατανόηση σε αναφορές αντωνυμιών και ο προσδιορισμός των οντοτήτων στις οποίες αναφέρονται οι αντωνυμίες είναι μια πολύπλοκη εργασία. Η επίλυση της ασάφειας στη γλώσσα, ιδιαίτερα σε περιβάλλοντα με πολλαπλές οντότητες, είναι κρίσιμη για την ακριβή ερμηνεία και τον ουσιαστικό διάλογο.

2.2.4 Πεδία εφαρμογής

Το Natural Language Processing βρίσκει διαφορετικές εφαρμογές σε διάφορους τομείς, αποδεικνύοντας την ευελιξία και τον αντίκτυπό του στη διαμόρφωση ευφυών συστημάτων.

- Υγειονομική περίθαλψη

  Στον τομέα της υγειονομικής περίθαλψης, το NLP βοηθά στην εξαγωγή πολύτιμων πληροφοριών από ιατρικά αρχεία, κλινικές σημειώσεις και ερευνητικά άρθρα. Διευκολύνει τη διάγνωση, τον σχεδιασμό της θεραπείας και την ιατρική έρευνα με την επεξεργασία και την ανάλυση τεράστιων ποσοτήτων δεδομένων κειμένου.

- Χρηματοοικονομική Ανάλυση

  Το NLP χρησιμοποιείται σε χρηματοπιστωτικά ιδρύματα για την ανάλυση συναισθήματος των ειδήσεων της αγοράς, την αξιολόγηση κινδύνου και τον εντοπισμό απάτης. Αναλύοντας οικονομικές αναφορές και άρθρα ειδήσεων, τα συστήματα NLP βοηθούν στη λήψη τεκμηριωμένων επενδυτικών αποφάσεων.

- Εξυπηρέτηση πελατών

  Στην εξυπηρέτηση πελατών, το NLP ενισχύει τις αλληλεπιδράσεις μέσω chatbot και εικονικών βοηθών. Επιτρέπει αυτοματοποιημένες απαντήσεις, ανάκτηση πληροφοριών και ανάλυση συναισθήματος, οδηγώντας σε βελτιωμένη ικανοποίηση των πελατών και αποτελεσματική επίλυση ερωτημάτων.

- Μάρκετινγκ και μέσα κοινωνικής δικτύωσης

  Το NLP διαδραματίζει κρίσιμο ρόλο στην ανάλυση των συναισθημάτων, των τάσεων και των προτιμήσεων των καταναλωτών στις πλατφόρμες μέσων κοινωνικής δικτύωσης. Οι έμποροι χρησιμοποιούν την ανάλυση συναισθήματος για να κατανοήσουν τις αντιδράσεις του κοινού, να προσαρμόσουν τις στρατηγικές διαφήμισης και να ενισχύσουν τη φήμη της επωνυμίας.

- Ανάλυση νομικών εγγράφων

  Οι επαγγελματίες νομικοί επωφελούνται από το NLP για την ανάλυση και τη σύνοψη μεγάλων όγκων νομικών εγγράφων. Το NLP βοηθά στην εξαγωγή πληροφοριών, στην ανάλυση συμβάσεων και στη νομική έρευνα, εξοικονομώντας χρόνο και βελτιώνοντας την ακρίβεια.

2.2.5 Μελλοντικές Τάσεις στην Επεξεργασία Φυσικής Γλώσσας

- Συνεχής ενσωμάτωση της βαθιάς μάθησης

  Η ενσωμάτωση των τεχνικών βαθιάς μάθησης αναμένεται να συνεχίσει να διαμορφώνει το μέλλον του NLP. Οι προηγμένες αρχιτεκτονικές νευρωνικών δικτύων και τα προεκπαιδευμένα μοντέλα έχουν επιδείξει αξιοσημείωτες βελτιώσεις στην απόδοση σε διάφορες εργασίες NLP.

- Επεξηγησιμότητα και ηθικά ζητήματα

  Η αντιμετώπιση της φύσης του «μαύρου κουτιού» των μοντέλων βαθιάς μάθησης και η διασφάλιση της διαφάνειας στη λήψη αποφάσεων αποτελούν κρίσιμες προκλήσεις. Οι μελλοντικές εξελίξεις στο NLP πιθανότατα θα τονίσουν την επεξήγηση, την ερμηνευτικότητα και τα ηθικά ζητήματα για την οικοδόμηση εμπιστοσύνης στα συστήματα AI.

- Πολυτροπικό NLP

  Η σύγκλιση του NLP με άλλες μεθόδους, όπως εικόνες και βίντεο, είναι μια αναδυόμενη τάση. Το Multimodal NLP στοχεύει στη δημιουργία ολοκληρωμένων μοντέλων που μπορούν να κατανοήσουν και να δημιουργήσουν περιεχόμενο σε διάφορους τύπους δεδομένων.

- Προσαρμογή και προσαρμοστικότητα

  Η προσαρμογή των συστημάτων NLP σε συγκεκριμένους τομείς ή ανάγκες των χρηστών είναι ένας τομέας συνεχούς έρευνας. Οι μελλοντικές τάσεις μπορεί να επικεντρωθούν στην ανάπτυξη προσαρμόσιμων και προσαρμόσιμων μοντέλων NLP που μπορούν να βελτιστοποιηθούν για συγκεκριμένους κλάδους ή εφαρμογές.

Συμπερασματικά, η Επεξεργασία Φυσικής Γλώσσας έχει εξελιχθεί σημαντικά με τα χρόνια, ξεπερνώντας διάφορες προκλήσεις και βρίσκοντας εφαρμογές σε διάφορους τομείς. Καθώς η τεχνολογία συνεχίζει να προοδεύει, η αντιμετώπιση των προκλήσεων που απομένουν και η εξερεύνηση των αναδυόμενων τάσεων θα ενισχύσει περαιτέρω τις δυνατότητες του NLP, καθιστώντας το αναπόσπαστο μέρος των ευφυών συστημάτων στο μέλλον.

2.3 Ανάλυση Συναισθήματος

2.3.1 Η σημασία της και τα εμπόδια που συναντά

Η ανάλυση συναισθήματος επικεντρώνεται στον προσδιορισμό του συναισθήματος που εκφράζεται σε ένα κομμάτι κειμένου, ως θετικό, αρνητικό ή ουδέτερο. Η τεχνική αυτή περιλαμβάνει την αξιοποίηση της επεξεργασίας φυσικής γλώσσας, της υπολογιστικής γλωσσολογίας και της ανάλυσης κειμένου για την εξαγωγή υποκειμενικών πληροφοριών και την απόκτηση πληροφοριών σχετικά με τον συναισθηματικό τόνο του περιεχομένου.

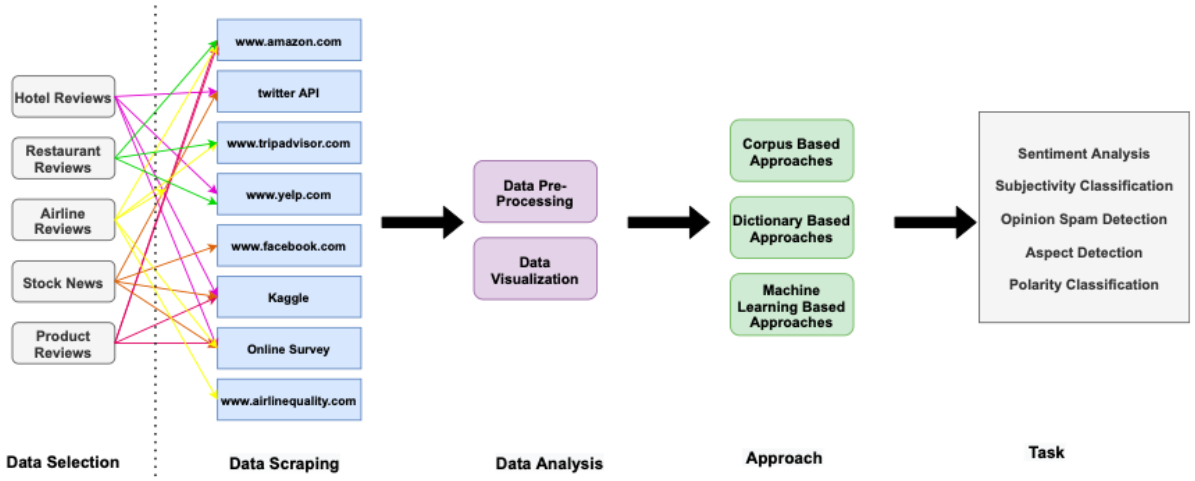
Ο πρωταρχικός στόχος της ανάλυσης συναισθήματος είναι η αξιολόγηση και η κατανόηση της συναισθηματικής στάσης ή γνώμης ενός ατόμου ή μιας ομάδας σχετικά με ένα συγκεκριμένο θέμα. Η ανάλυση συναισθήματος αποσκοπεί στην αποκάλυψη των συναισθημάτων που εκφράζουν οι πελάτες σε κριτικές, απαντήσεις σε έρευνες ή αλληλεπιδράσεις στα μέσα κοινωνικής δικτύωσης. Εντοπίζοντας τα συναισθήματα, οι επιχειρήσεις μπορούν να μετρήσουν την ικανοποίηση των πελατών, να συλλάβουν την κοινή γνώμη και να λάβουν αποφάσεις βάσει δεδομένων για τη βελτίωση των προϊόντων ή των υπηρεσιών τους.

Για τις επιχειρήσεις, η ενσωμάτωση των ευρημάτων της ανάλυσης συναισθήματος στη συνολική στρατηγική τους είναι ζωτικής σημασίας. Δεν πρόκειται απλώς για τον εντοπισμό συναισθημάτων, αλλά για την κατανόηση του τρόπου με τον οποίο αυτά τα συναισθήματα επηρεάζουν την αντίληψη της εταιρείας, την αφοσίωση των πελατών και την τοποθέτησή τους στην αγορά. Οι εταιρείες πρέπει να χρησιμοποιούν την ανάλυση συναισθήματος ως εργαλείο για τη λήψη τεκμηριωμένων αποφάσεων, προσαρμόζοντας τις προσεγγίσεις τους με βάση την ανατροφοδότηση των πελατών για να παραμείνουν ανταγωνιστικές και να ανταποκρίνονται στις εξελισσόμενες προτιμήσεις των καταναλωτών.

Παρά την ευρεία χρήση της, η ανάλυση συναισθήματος αντιμετωπίζει προκλήσεις. Μια σημαντική πρόκληση είναι η ασάφεια και η πολυπλοκότητα της ανθρώπινης γλώσσας. Ο σαρκασμός, η ειρωνεία ή οι πολιτισμικές αποχρώσεις μπορεί να οδηγήσουν σε παρερμηνεία των συναισθημάτων. Επιπλέον, τα συναισθήματα δεν δηλώνονται πάντα ρητά και μπορεί να απαιτούν κατανόηση του πλαισίου στο οποίο ανήκει το κείμενο. Η δυναμική φύση της γλώσσας και η συνεχής εξέλιξη των εκφράσεων καθιστούν την ανάλυση συναισθήματος ένα πολύπλοκο έργο, που απαιτεί συνεχή βελτίωση των αλγορίθμων και των μοντέλων.

Η ποικιλομορφία των γλωσσών και των πολιτισμών προσθέτει άλλο ένα επίπεδο πολυπλοκότητας στην ανάλυση συναισθήματος. Η μετάφραση των σχολίων από τη μία γλώσσα στην άλλη ενέχει τον κίνδυνο να χαθούν οι πολιτισμικές και γλωσσικές αποχρώσεις, επηρεάζοντας ενδεχομένως την ακρίβεια της

ανάλυσης συναισθήματος. Οι ερευνητές και οι αναλυτές πρέπει να αντιμετωπίσουν αυτές τις προκλήσεις για να διασφαλίσουν ότι τα αποτελέσματα της ανάλυσης συναισθήματος είναι αξιόπιστα και αντικατοπτρίζουν τις διαφορετικές οπτικές που εκφράζονται σε διαφορετικές γλώσσες και πολιτισμικά πλαίσια.



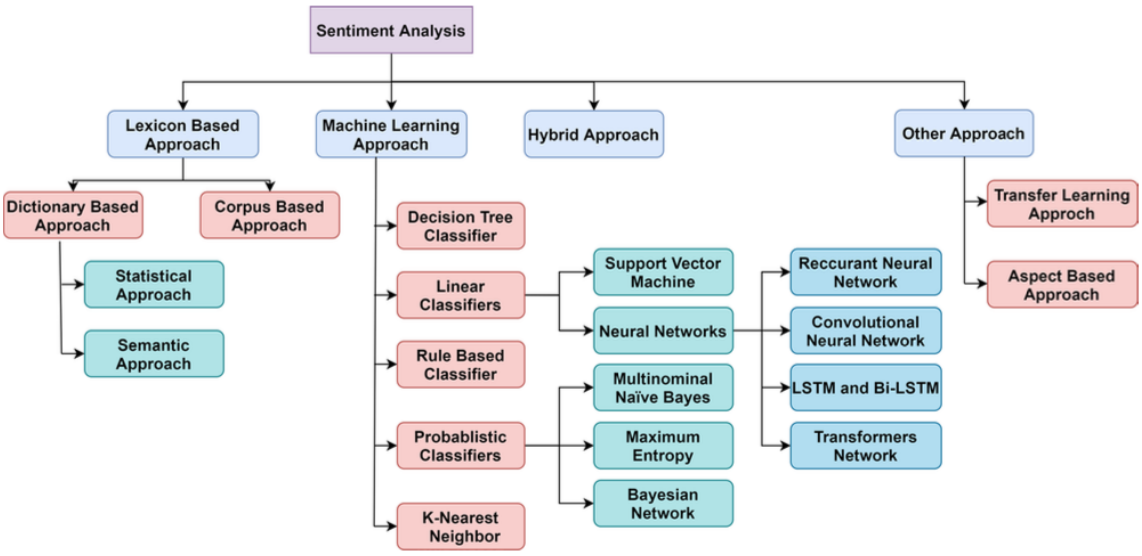
Σχήμα 2.1: Γενική διαδικασία συναισθηματικής ανάλυσης

2.3.2 Μέθοδοι Συναισθηματικής Ανάλυσης

Η ανάλυση συναισθήματος χρησιμοποιεί διάφορες μεθόδους για την αποκρυπτογράφηση και την κατηγοριοποίηση των συναισθημάτων μέσα σε δεδομένα κειμένου. Οι μέθοδοι αυτές λειτουργούν σε διαφορετικά επίπεδα, παρέχοντας πληροφορίες σχετικά με την πολικότητα και τη φύση των συναισθημάτων. Οι κυρίως χρησιμοποιούμενες προσεγγίσεις για την ανάλυση συναισθημάτων είναι τρείς :

1. Προσέγγιση Βασισμένη σε Λεξικό / Lexicon Based Approach
2. Προσέγγιση Μηχανικής Μάθησης / Machine Learning Approach
3. Υβριδική Προσέγγιση / Hybrid Approach.

Οι ερευνητές προσπαθούν συνεχώς να βρουν καλύτερους τρόπους για να πετύχουν τον στόχο τους με καλύτερη ακρίβεια και χαμηλότερο υπολογιστικό κόστος. Η επισκόπηση διαφόρων μεθόδων που χρησιμοποιούνται στην ανάλυση συναισθημάτων φαίνεται στο Σχ. 2. Επιπλέον η γενική ροή εργασίας από την συλλογή δεδομένων, την επιλογή μεθόδου κ.ο.κ φαίνονται στο Σχ.1.



Σχήμα 2.2: Προσεγγίσεις συναισθηματικής ανάλυσης

1. Προσέγγιση Βασισμένη σε Λεξικό / Lexicon Based Approach

Η ανάλυση συναισθήματος που χρησιμοποιεί αυτή τη προσέγγιση βασίζεται στη χρήση λεξικών, τα οποία είναι συλλογές διακριτικών (tokens) με προκαθορισμένες βαθμολογίες που υποδεικνύουν την ουδέτερη, θετική ή αρνητική φύση τους. Σε κάθε token εκχωρείται μια βαθμολογία, συνήθως στην περιοχή από +1 έως -1, όπου το +1 υποδηλώνει εξαιρετικά θετικά και το -1 υποδηλώνει εξαιρετικά αρνητικά συναισθήματα.

Σε αυτή τη μέθοδο, το συναίσθημα ενός κειμένου καθορίζεται με τη συγκέντρωση των βαθμολογιών των μεμονωμένων token. Οι βαθμολογίες, που αντιπροσωπεύουν θετικά, αρνητικά και ουδέτερα συναισθήματα, αθροίζονται χωριστά και η συνολική πολικότητα του κειμένου εκχωρείται με βάση την υψηλότερη τιμή μεταξύ αυτών των μεμονωμένων βαθμών. Αυτή η διαδικασία περιλαμβάνει τη διαίρεση του εγγράφου σε μονολεκτικά token, τον υπολογισμό της πολικότητας καθενός από αυτά και τη συγκέντρωση των βαθμολογιών για τον προσδιορισμό του συνολικού συναισθήματος.

Η μέθοδος αυτή είναι ιδιαίτερα κατάλληλη για ανάλυση συναισθήματος σε επίπεδο προτάσεων. Είναι μια μέθοδος χωρίς επίβλεψη που δεν απαιτεί δεδομένα εκπαίδευσης. Ωστόσο, ένα αξιοσημείωτο μειονέκτημα είναι η ευαισθησία του στις διαφορές των σημασιών ανά τομέα. Οι λέξεις μπορεί να έχουν διαφορετικές έννοιες και πολικότητες σε διαφορετικούς τομείς, οδηγώντας σε πιθανές παρερμηνείες. Για παράδειγμα, η λέξη "μικρό" μπορεί να γίνει αντιληπτή ως αρνητική στο πλαίσιο μιας οθόνης τηλεόρασης αλλά θετική όταν περιγράφεται μια φορητή κάμερα.

Για την αντιμετώπιση της εξάρτησης από τον τομέα, η δημιουργία λεξικών συναισθημάτων για συγκεκριμένο τομέα ή η προσαρμογή των υπαρχόντων καθίσταται απαραίτητη. Παρά αυτόν τον περιορισμό, η προσέγγιση αυτή επαινείται για την απλότητά της, την έλλειψη εξάρτησης από δεδομένα εκπαίδευσης και την καταλληλότητα για ορισμένες εφαρμογές.

1.1 Corpus Based

Η μέθοδος αυτή χρησιμοποιεί σημασιολογικά και συντακτικά μοτίβα για να διακρίνει το συναίσθημα των προτάσεων. Ξεκινώντας με ένα προκαθορισμένο σύνολο όρων και προσανατολισμών συναισθήματος, αυτή η μέθοδος διερευνά συντακτικά μοτίβα σε ένα μεγάλο σώμα για να προσδιορίσει τα διακριτικά συναισθήματος και τους προσανατολισμούς τους. Ένα παράδειγμα είναι η χρήση συσχετιστικών συνδέσμων για να συναχθεί η συνέπεια του συναισθήματος. Παρά τη δημοτικότητά του, η επίτευξη συνεπών αποτελεσμάτων με αυτήν την προσέγγιση μπορεί να είναι δύσκολη.

Στο πλαίσιο αυτής της μεθόδου υπάρχουν δύο υποτύποι:

- Στατιστική Προσέγγιση

Η στατιστική προσέγγιση προσδιορίζει λέξεις ή μοτίβα συνύπαρξης μέσω στατιστικής ανάλυσης. Με την εξέταση της συχνότητας των συνεμφανίσεων των tokens σε παρόμοια πλαίσια, προσδιορίζονται τα συναισθήματα. Αυτή η προσέγγιση χρησιμοποιείται συνήθως σε εφαρμογές όπως η ανίχνευση χειραγωγημένων κριτικών, όπου χρησιμοποιούνται δοκιμές (test) τυχαιότητας.

- Σημασιολογική Προσέγγιση

Στη σημασιολογική προσέγγιση, οι βαθμολογίες ομοιότητας μεταξύ των tokens υπολογίζονται χρησιμοποιώντας εργαλεία όπως το WordNet. Αυτή η μέθοδος διευκολύνει τον εντοπισμό συνωνύμων και αντωνύμων, βοηθώντας στην ανάλυση συναισθήματος. Βρίσκει εφαρμογές στην κατασκευή μοντέλων λεξικών για την περιγραφή επιθέτων, ρημάτων και ουσιαστικών στην ανάλυση συναισθημάτων.

1.2 Dictionary Based

Η δεύτερη αυτή κατεύθυνση περιλαμβάνει μια προκαθορισμένη λίστα λέξεων γνώμης που συλλέγονται χειροκίνητα. Τα συνώνυμα και τα αντώνυμα προέρχονται από εργαλεία όπως ο θησαυρός ή το WordNet. Ο θησαυρός είναι ένα εργαλείο που παραθέτει λέξεις με παρόμοιες ή αντίθετες έννοιες. Ενώ αυτή η προσέγγιση βασίζεται στην υπόθεση ότι τα συνώνυμα μοιράζονται την ίδια πολικότητα και τα αντώνυμα έχουν αντίθετη πολικότητα, προκύπτουν προκλήσεις στην εύρεση λέξεων γνώμης για συγκεκριμένο τομέα.

Παρά τους περιορισμούς που σχετίζονται με την εξάρτηση από τον τομέα που αναφέρεται το κείμενο, οι προσεγγίσεις που βασίζονται σε λεξικό προσφέρουν απλότητα και αποτελεσματικότητα σε συγκεκριμένα σενάρια συναισθηματικής ανάλυσης. Η χειροκίνητη επιμέλεια και προσαρμογή εκτελούνται συχνά για τη βελτίωση της ποιότητας των λεξικών και τον μετριασμό των προκλήσεων που σχετίζονται με τον τομέα εφαρμογής.

2. Προσέγγιση Μηχανικής Μάθησης / Machine Learning Approach

Η ανάλυση συναισθήματος, η διαδικασία αυτή της αναγνώρισης και κατηγοριοποίησης του συναισθήματος από κείμενο ή ήχο, έχει αποκτήσει εξέχουσα θέση λόγω της εφαρμογής αλγορίθμων μηχανικής μάθησης. Η μηχανική μάθηση προσφέρει δύο κύριες προσεγγίσεις για την ανάλυση συναισθήματος:

2.1 Εποπτευόμενη Μηχανική Μάθηση / Supervised Learning

Η εποπτευόμενη μηχανική μάθηση περιλαμβάνει αλγόριθμους εκπαίδευσης σε επισημασμένα σύνολα δεδομένων για την ακριβή πρόβλεψη συναισθημάτων. Αυτή η μέθοδος αξιοποιεί συντακτικούς και γλωσσικούς παράγοντες, αντιμετωπίζοντας την ταξινόμηση συναισθημάτων ως τυπικό πρόβλημα ταξινόμησης κειμένου. Το μοντέλο κατηγοριοποίησης συσχετίζει χαρακτηριστικά του κειμένου με προκαθορισμένες ετικέτες κλάσεων, επιτρέποντας την πρόβλεψη συναισθημάτων για καινούρια κείμενα.

Οι συνήθως χρησιμοποιούμενοι αλγόριθμοι μηχανικής μάθησης για ανάλυση συναισθημάτων περιλαμβάνουν:

- Naive Bayes (NB)

  Η τεχνική Naive Bayes, που βασίζεται στην ταξινόμηση Bayes, είναι ένας πιθανολογικός ταξινομητής. Προβλέπει την πιθανότητα ενός δεδομένου συνόλου χαρακτηριστικών που ανήκουν σε μια συγκεκριμένη κατηγορία συναισθημάτων. Η NB είναι ιδιαίτερα χρήσιμη όταν το μέγεθος των δεδομένων εκπαίδευσης είναι μικρό. Ωστόσο, μπορεί να αντιμετωπίσει προκλήσεις στην ακριβή ταξινόμηση των αρνητικών συναισθημάτων. Έχουν προταθεί βελτιωμένες εκδόσεις για την αντιμετώπιση αυτού του ζητήματος.

- Supervised Vector Machine (SVM)

Η τεχνική SVM χρησιμοποιεί υπερεπίπεδα για να αναλύσουν και να καθορίσουν τα όρια απόφασης. Αυτές οι μη πιθανολογικές εποπτευόμενες τεχνικές μάθησης χρησιμοποιούνται ευρέως για την ανάλυση συναισθήματος. Το SVM επιδιώκει να προσδιορίσει το υπερεπίπεδο που διαχωρίζει καλύτερα τα δεδομένα σε διακριτές κατηγορίες συναισθημάτων. Τα γραμμικά μοντέλα SVM έχουν επιδείξει υψηλή ακρίβεια στις εργασίες ταξινόμησης συναισθημάτων.

- Logistic Regression (LR)

  Η λογιστική παλινδρόμηση είναι μια πιθανολογική ανάλυση παλινδρόμησης που χρησιμοποιείται για εργασίες δυαδικής ταξινόμησης. Υπολογίζει την αναλογία πιθανοτήτων για εφαρμογές δυαδικής ταξινόμησης, καθιστώντας τον έναν κοινά αναπτυσσόμενο αλγόριθμο. Το LR προσδιορίζει τις ιδιότητες εισόδου που είναι πιο χρήσιμες για τον εντοπισμό θετικών και αρνητικών κλάσεων.

- Δέντρο αποφάσεων (DT)

  Οι ταξινομητές δένδρων αποφάσεων δημιουργούν ένα δέντρο χρησιμοποιώντας παραδείγματα εκπαίδευσης για την ταξινόμηση της πολικότητας του κειμένου. Χρησιμοποιείται αναδρομική διαίρεση δεδομένων με βάση τις συνθήκες. Το Random Forest (RF), το οποίο συνδυάζει πολλαπλά δέντρα απόφασης, προτιμάται συχνά για να αποφευχθεί η υπερβολική προσαρμογή και να ενισχυθεί η ακρίβεια.

- Μέγιστη Εντροπία (ΜΕ)

  Οι ταξινομητές μέγιστης εντροπίας κωδικοποιούν σύνολα χαρακτηριστικών με ετικέτα και υπολογίζουν τα βάρη χαρακτηριστικών για να προβλέψουν την πιο πιθανή ετικέτα για ένα δεδομένο σύνολο χαρακτηριστικών. Η εντροπία, ένα μέτρο της μη προβλεψιμότητας, χρησιμοποιείται για τον προσδιορισμό της αβεβαιότητας στην ταξινόμηση.

- Κ-Πλησιέστεροι γείτονες (KNN)

  Το K-Nearest Neighbors λειτουργεί με την προϋπόθεση ότι η ταξινόμηση ενός δείγματος δοκιμής θα είναι παρόμοια με τους κοντινούς γείτονες. Αν και δεν χρησιμοποιείται εκτενώς, το KNN έχει δείξει καλά αποτελέσματα όταν εκπαιδεύεται προσεκτικά.

2.2 Ημί-εποπτευόμενη Μηχανική Μάθηση / Semi-Supervised Learning

Η ημι-εποπτευόμενη μάθηση χρησιμοποιείται όταν το σύνολο δεδομένων εκπαίδευσης περιέχει δεδομένα τόσο με ετικέτα όσο και χωρίς ετικέτα. Αυτή η προσέγγιση αποδεικνύεται πολύτιμη σε πραγματικές καταστάσεις όπου η συλλογή δεδομένων χωρίς ετικέτα είναι ευκολότερη από τη λήψη δεδομένων με ετικέτα. Συνδυάζει αλγόριθμους προεπεξεργασίας και ταξινόμησης για σύνολα δεδομένων χωρίς ετικέτα.

Η επιλογή της προσέγγισης μηχανικής μάθησης εξαρτάται από παράγοντες όπως το μέγεθος του συνόλου δεδομένων εκπαίδευσης, η ανάγκη για ερμηνευσιμότητα και τα ειδικά χαρακτηριστικά της εργασίας ανάλυσης συναισθήματος.

3. Υβριδική Προσέγγιση / Hybrid Approach.

Η έρευνα για τις υβριδικές αρχιτεκτονικές βρίσκεται σε εξέλιξη, με μελέτες που διερευνούν τον συνδυασμό τεχνικών μάθησης που βασίζονται στο λεξικό και αυτοματοποιημένης εκμάθησης για τη βελτίωση των αποτελεσμάτων. Αυτά τα υβριδικά μοντέλα έχουν δείξει πολλά υποσχόμενα αποτελέσματα, μειώνοντας τον συνολικό αριθμό χαρακτηριστικών, ενώ επιτυγχάνουν καλύτερες μετρήσεις απόδοσης σε σύγκριση με μεμονωμένα μοντέλα. Υπάρχουν περαιτέρω ευκαιρίες έρευνας για τη βελτίωση και τη βελτιστοποίηση υβριδικών μοντέλων για ανάλυση συναισθήματος.

Συμπερασματικά, οι τεχνικές ανάλυσης συναισθήματος περιλαμβάνουν μια ποικιλία προσεγγίσεων, καθεμία με τα πλεονεκτήματα και τα μειονεκτήματά της. Η επιλογή της προσέγγισης εξαρτάται από τις ειδικές απαιτήσεις της εργασίας και τη φύση των δεδομένων που αναλύονται. Η συνεχιζόμενη έρευνα στην ανάλυση συναισθήματος συνεχίζει να διερευνά νέες μεθόδους και συνδυασμούς υφιστάμενων προσεγγίσεων για τη βελτίωση της ακρίβειας και της εφαρμογής σε διαφορετικά περιβάλλοντα.

ΚΕΦΑΛΑΙΟ 3

ΜΕΘΟΔΟΛΟΓΙΑ & ΕΡΓΑΛΕΙΑ

3.1 Εισαγωγή

Αυτό το κεφάλαιο περιγράφει τη μεθοδολογία που χρησιμοποιήθηκε για την ανάπτυξη του συστήματος ανάλυσης συναισθήματος, περιγράφοντας λεπτομερώς τα βήματα που έγιναν για την επεξεργασία και την ανάλυση δεδομένων κειμένου. Επιπλέον, συζητούνται τα εργαλεία και οι τεχνολογίες που χρησιμοποιούνται στην υλοποίηση.

3.2 Συλλογή δεδομένων

Το πρώτο βήμα για τη δημιουργία του συστήματος ανάλυσης συναισθημάτων ήταν η συλλογή δεδομένων. Ένα ποικίλο σύνολο δεδομένων που περιέχει δεδομένα κειμένου και αντίστοιχες ετικέτες συναισθήματος ήταν απαραίτητο για την εκπαίδευση και την αξιολόγηση του μοντέλου. Το κύριο σύνολο δεδομένων που χρησιμοποιήθηκε για αυτό το έργο ανάλυσης συναισθήματος είναι το "Amazon Review Data (2018)" που παρέχεται από τον Jianmo Ni. Αυτό το σύνολο δεδομένων χρησιμεύει ως πολύτιμος πόρος λόγω της περιεκτικής φύσης του, που περιλαμβάνει διάφορες πτυχές, όπως κριτικές, μεταδεδομένα προϊόντων και συνδέσμους. Συγκεκριμένα, το σύνολο δεδομένων περιλαμβάνει αξιολογήσεις, κείμενο κριτικής, ψήφους χρήσεως, περιγραφές προϊόντων, πληροφορίες κατηγορίας, τιμή, λεπτομέρειες επωνυμίας και χαρακτηριστικά εικόνας. Το σύνολο δεδομένων είναι μια ενημερωμένη έκδοση του συνόλου δεδομένων κριτικών Amazon 2014 και περιλαμβάνει ένα υποσύνολο όπου όλοι οι χρήστες και τα στοιχεία έχουν τουλάχιστον 5 κριτικές. Επιλέχθηκε ένα υποσύνολο αυτών των δεδομένων, στοχεύοντας συγκεκριμένα σε κριτικές λογισμικού. Αυτό το υποσύνολο ενισχύει τη συνάφεια της ανάλυσης συναισθημάτων μας, παρέχοντας ένα πιο συγκεντρωμένο και συγκεκριμένο σύνολο δεδομένων για εκπαίδευση και αξιολόγηση.

3.3 Προεπεξεργασία Δεδομένων

Η προεπεξεργασία δεδομένων είναι μια κομβική φάση στη μεθοδολογία που ακολουθείται για την συναισθηματική ανάλυση. Αυτή η πολύπλευρη διαδικασία περιλαμβάνει πολλές βασικές τεχνικές που στοχεύουν στη βελτίωση της ποιότητας και της ομοιομορφίας του συνόλου δεδομένων.

1. Καθαρισμός κειμένου:

Ανεξάρτητα από την πηγή, τα ακατέργαστα κειμενικά δεδομένα συχνά συνοδεύονται από ξένα στοιχεία, συμπεριλαμβανομένων άσχετων χαρακτήρων, σημείων στίξης και ειδικών συμβόλων. Για την ενίσχυση της συνέπειας και τη βελτίωση της αναγνωσιμότητας του συνόλου δεδομένων, εφαρμόστηκε μια αυστηρή διαδικασία καθαρισμού κειμένου. Αυτό περιλάμβανε την αφαίρεση περιττών χαρακτήρων, διασφαλίζοντας ότι το σύνολο δεδομένων διατηρεί μια συνεκτική δομή χωρίς τεχνουργήματα που θα μπορέσουν να εισάγουν θόρυβο στην ανάλυση.

2. Tokenization:

Για να ξεκλειδωθεί η εγγενής σημασιολογική δομή των κειμενικών δεδομένων, χρησιμοποιήθηκε το tokenization ως βασική τεχνική προεπεξεργασίας. Αυτό περιλάμβανε την κατανομή των προτάσεων σε μεμονωμένες λέξεις ή διακριτικά (tokens). Αποσυναρμολογώντας το κείμενο στα βασικά του στοιχεία, στοχεύεται η διευκόλυνση των πιο λεπτομερών και λεπτών αναλύσεων του συναισθήματος, επιτρέποντας στο μοντέλο να διακρίνει συναισθήματα που σχετίζονται με συγκεκριμένες λέξεις και φράσεις.

3. Κανονικοποίηση:

Η ομοιομορφία στην αναπαράσταση λέξεων είναι ζωτικής σημασίας για την αποτελεσματικότητα των μοντέλων επεξεργασίας φυσικής γλώσσας. Καθώς η γλώσσα παρουσιάζει φυσικές παραλλαγές σε πεζά (κεφαλαία και πεζά), η κανονικοποίηση του κειμένου περιλάμβανε τη μετατροπή όλων των γραμμάτων σε πεζά. Αυτό το βήμα κανονικοποίησης μετριάστηκε τον κίνδυνο παρερμηνείας λέξεων από το μοντέλο λόγω διαφοροποιήσεων σε πεζά γράμματα, διασφαλίζοντας έτσι μια συνεπή και τυποποιημένη αναπαράσταση σε όλο το σύνολο δεδομένων.

Αναλαμβάνοντας αυτά τα βήματα προεπεξεργασίας, η προσέγγιση που ακολουθήθηκε δίνει προτεραιότητα στη βελτίωση και την τυποποίηση του συνόλου δεδομένων. Αυτό όχι μόνο συνέβαλε στην ευρωστία των μεταγενέστερων αναλύσεων, αλλά έθεσε τις βάσεις για την αποτελεσματική εφαρμογή της μηχανικής εκμάθησης και των αλγορίθμων επεξεργασίας φυσικής γλώσσας για την εξαγωγή ουσιαστικών πληροφοριών από τα δεδομένα κειμένου.

3.4 Εξαγωγή χαρακτηριστικών

Για να αξιοποιηθούν αποτελεσματικά τα μοντέλα μηχανικής μάθησης, τα δεδομένα κειμένου υπέστησαν έναν κρίσιμο μετασχηματισμό μέσω τεχνικών εξαγωγής χαρακτηριστικών. Στη μεθοδολογία μας, χρησιμοποιήσαμε δύο εξέχουσες μεθόδους για το σκοπό αυτό: το μοντέλο Bag-of-Words (BoW) και Term Frequency-Inverse Document Frequency (TF-IDF).

1. Μοντέλο Bag-of-Words (BoW):

Το μοντέλο Bag-of-Words είναι μια θεμελιώδης τεχνική που αναπαριστά το κείμενο ως ένα μη ταξινομημένο σύνολο λέξεων, αγνοώντας τη γραμματική και τη σειρά των λέξεων, αλλά διατηρώντας πληροφορίες σχετικά με τη συχνότητα των λέξεων. Σε αυτό το βήμα, χρησιμοποιήθηκε το CountVectorizer για να εφαρμοστεί το μοντέλο BoW. Το CountVectorizer μετέτρεψε το κείμενο σε μια μήτρα μετρήσεων διακριτικών, δημιουργώντας μια αριθμητική αναπαράσταση της συχνότητας των λέξεων μέσα στο σύνολο δεδομένων.

2. Συχνότητα όρου-Αντίστροφη συχνότητα εγγράφου (TF-IDF):

Αναγνωρίζοντας τη σημασία της καταγραφής όχι μόνο της συχνότητας των λέξεων αλλά και της σημασίας των λέξεων σε ολόκληρο το σύνολο δεδομένων, χρησιμοποιήθηκε το TfidfVectorizer. Αυτή η τεχνική αξιολογεί τη σημασία μιας λέξης σε ένα κείμενο σε σχέση με τη συχνότητά της σε ολόκληρο το σώμα. Οι προκύπτουσες τιμές TF-IDF παρείχαν αριθμητικά διανύσματα που μετέφεραν τόσο την τοπική σημασία των λέξεων σε συγκεκριμένες ανασκοπήσεις όσο και τη συνολική σημασία τους σε ολόκληρο το σύνολο δεδομένων.

Με την ενσωμάτωση αυτών των τεχνικών εξαγωγής χαρακτηριστικών, εξασφαλίστηκε ότι τα δεδομένα κειμένου μετατράπηκαν σε μορφή συμβατή με μοντέλα μηχανικής εκμάθησης. Η χρήση του CountVectorizer για BoW και του TfidfVectorizer για το TF-IDF εμπλούτισε το σύνολο δεδομένων με αριθμητικές αναπαραστάσεις που περιείχαν τη συχνότητα και τη σημασία των λέξεων, ανοίγοντας το δρόμο για μεταγενέστερη μοντελοποίηση και ανάλυση.

3.5 Μοντέλα Ανάλυσης Συναισθημάτων

Στην προσπάθειά μας να αποκαλυφθούν τα συναισθήματα που περικλείονται στις κριτικές του Amazon, χρησιμοποιήθηκε μια ποικιλία μοντέλων ανάλυσης συναισθημάτων. Αναγνωρίζοντας τις περιπλοκές της ερμηνείας των συναισθημάτων, επιλέξαμε ένα μείγμα μεθόδων για να συλλάβουμε διεξοδικά τις πτυχές που υπάρχουν στα δεδομένα κειμένου.

1. VADER (Valence Aware Dictionary and Sentiment Reasoner): Ακολουθώντας μια προσέγγιση βασισμένη στο λεξικό, χρησιμοποιήθηκε το εργαλείο ανάλυσης συναισθήματος VADER. Αυτή η μέθοδος βασίζεται σε ένα προκατασκευασμένο λεξικό συναισθημάτων για τη μέτρηση της πολικότητας των λέξεων, παρέχοντας μια συνολική βαθμολογία συναισθήματος για ένα δεδομένο κείμενο.
2. Naive Bayes (NB): Η τεχνική NB χρησιμοποιείται τόσο για κατηγοριοποίηση όσο και για εκπαίδευση. Η NB είναι μια προσέγγιση ταξινόμησης που βασίζεται στο θεώρημα των Bayes. Η NB είναι ένας πιθανοτικός ταξινομητής που χρησιμοποιεί το θεώρημα Bayes για να

προβλέψει την πιθανότητα ενός δεδομένου συνόλου χαρακτηριστικών ως μέρος οποιασδήποτε συγκεκριμένης ετικέτας. Η υπό όρους πιθανότητα να συμβεί το συμβάν Α λαμβάνοντας υπόψη τις επιμέρους πιθανότητες των Α και Β και την υπό όρους πιθανότητα εμφάνισης του γεγονότος Β. Εδώ θεωρείται ότι τα χαρακτηριστικά δεν εξαρτώνται. Το μοντέλο BoW μπορεί να χρησιμοποιηθεί για εξαγωγή χαρακτηριστικών. Γενικά, η NB εφαρμόζεται όταν το μέγεθος των δεδομένων εκπαίδευσης είναι μικρό. Η NB ταξινομήθηκε ως θετική κατά 10% μεγαλύτερη ακρίβεια από την αρνητική ταξινόμηση. Αυτό οδήγησε σε μείωση της μέσης ακρίβειας κατά τη λήψη.

3. Support vector machine (SVM): Η προσέγγιση SVM, η οποία χρησιμοποιεί υπερ-επίπεδα, χρησιμοποιείται για την ανάλυση δεδομένων και τον καθορισμό των ορίων απόφασης σε αυτήν την τεχνική. Το SVM είναι ένας τύπος μη πιθανολογικής εποπτευόμενης τεχνικής μάθησης που χρησιμοποιείται συχνά για εργασίες ταξινόμησης. Ο πρωταρχικός στόχος του SVM είναι να προσδιορίσει το υπερεπίπεδο που διαχωρίζει καλύτερα τα δεδομένα σε διακριτές κλάσεις. Ως αποτέλεσμα, η SVM αναζητά το υπερπλάνο με το υψηλότερο δυνατό περιθώριο.
4. Decision Tree (DT): Το DT Classifier είναι μια εποπτευόμενη τεχνική εκμάθησης όπου ένα δέντρο κατασκευάζεται χρησιμοποιώντας το παράδειγμα εκπαίδευσης για την ταξινόμηση της πολικότητας του κειμένου. Το DT χρησιμοποιεί μια συνθήκη για να διαιρεί τα δεδομένα σε μέρη αναδρομικά. Οι ραδιοσυχνότητες χρησιμοποιούνται συχνά από το DT που συνδυάζει πολλαπλά DT για την αποφυγή υπερβολικής προσαρμογής και τη βελτίωση της ακρίβειας.

3.6 Εργαλεία και Τεχνολογίες

Για την δημιουργία του συστήματος χρησιμοποιήθηκε ένα σύνολο εργαλείων και τεχνολογιών ειδικά επιλεγμένο για συναισθηματική ανάλυση. Τα βασικότερα, αναφέρονται:

- Γλώσσα προγραμματισμού:

Η γλώσσα προγραμματισμού που χρησιμοποιήθηκε είναι η Python. Με την αφθονία των βιβλιοθηκών και των πλαισίων προσαρμοσμένων για την επεξεργασία φυσικής γλώσσας (NLP) και τη μηχανική μάθηση, η Python παρείχε μια σταθερή βάση για τις προσπάθειες ανάπτυξης του συστήματος.

- Βιβλιοθήκες NLP:

Η αξιοποίηση της ισχύος του NLP ήταν πρωταρχικής σημασίας για την εξάπλωση της πολυπλοκότητας των δεδομένων κειμένου. Το NLTK χρησιμοποιήθηκε για να εκτελεστούς βασικές εργασίες NLP. Διαδραμάτισε κεντρικό ρόλους σε καθήκοντα προεπεξεργασίας κειμένου.

- Βιβλιοθήκες μηχανικής μάθησης:

Η ανάλυση του συναισθήματος επωφελήθηκε σημαντικά από την ανδρεία του Scikit-Learn. Αυτή η βιβλιοθήκη διευκόλυνε την εφαρμογή των βασικών αλγορίθμων μηχανικής μάθησης, συμπεριλαμβανομένων των SVM, της Naive Bayes (NB) και του Decision Tree.

- Εργαλεία για την Lexicon Based Προσέγγιση:

Το Vader μπήκε στο επίκεντρο για την ανάλυση που βασίζεται στο λεξικό, και βοήθησε να δούμε την προσέγγιση αυτή σε βάθος και να συγκριθεί με τις υπόλοιπες.

ΚΕΦΑΛΑΙΟ 4

ΤΕΧΝΙΚΗ ΥΛΟΠΟΙΗΣΗ

4.1 Εισαγωγή

Στο παρόν κεφάλαιο θα παρουσιαστούν κομμάτια κώδικα, με λεπτομερή επεξήγηση για την λειτουργικότητα που προσφέρουν στο σύστημα που υλοποιήθηκε.

4.2 Διερεύνηση Δεδομένων

Το πρώτο κομμάτι κώδικα για την υλοποίηση του συστήματος είναι η διερεύνηση των δεδομένων. Ο κώδικας βρίσκεται στο αρχείο ‘data_exploration.py’ και υλοποιεί βασικές λειτουργίες όπως η φόρτωση δεδομένων, η εμφάνιση πληροφοριών για το σύνολο δεδομένων, η δειγματοληψία και η οπτικοποίηση τους. Παρακάτω λοιπόν θα αναλυθεί η λειτουργία του κομμάτι προς κομμάτι.

Αρχικά στον κώδικα πραγματοποιούνται τα εξής:

1. Εισαγωγή βιβλιοθηκών (libraries)
   - Η βιβλιοθήκη pandas εισάγεται ως pd για την επεξεργασία δεδομένων σε μορφή DataFrame.
   - Η βιβλιοθήκη json εισάγεται για την εργασία με δεδομένα JSON.

2. Φόρτωση και ανάλυση των δεδομένων :
   - Η μεταβλητή file_path καθορίζει τη διαδρομή προς το JSON αρχείο (Software_5.json).
   - Το αρχείο ανοίγεται και διαβάζεται γραμμή προς γραμμή (file.readlines()).
   - Κάθε γραμμή, υποθέτοντας ότι είναι ένα αντικείμενο JSON, αναλύεται με τη χρήση της json.loads() και αποθηκεύεται στη μεταβλητή json_objects, δημιουργώντας μια λίστα από λεξικά (αντικείμενα JSON).

3. Δημιουργία DataFrame:

- Η εντολή pd.DataFrame(json_objects) μετατρέπει τη λίστα των αντικειμένων JSON (json_objects) σε ένα DataFrame της pandas (df), όπου κάθε λεξικό αντιπροσωπεύει μια σειρά στο DataFrame.

Τα τμήματα που έχουν μετατραπεί σε σχόλια, μιας και δεν αποτελούν απαραίτητη διαδικασία, δείχνουν επιπλέον βήματα που μπορεί να είναι μέρος της διαδικασίας εξερεύνησης:

- Βασικές Πληροφορίες για το Dataset: Εκτύπωση πληροφοριών σχετικά με το DataFrame, όπως οι τύποι των στηλών και η χρήση μνήμης (df.info()).
- Δειγματοληψία Δεδομένων: Περιορισμός του DataFrame στις πρώτες 500 γραμμές (df.head(500)).
- Εμφάνιση Δεδομένων: Εκτύπωση των πρώτων λίγων γραμμών του DataFrame (df.head()).
- Πρόσβαση σε Δεδομένα: Πρόσβαση και εκτύπωση ενός παραδείγματος κειμένου από τη στήλη reviewText της πρώτης γραμμής (df['reviewText'][0]).

Τέλος, γίνεται οπτικοποίηση της κατανομής των αξιολογήσεων:

- Οπτικοποίηση: Χρησιμοποιεί την sns.countplot() για να σχεδιάσει το πλήθος των αξιολογήσεων (overall στήλη).
- Προσαρμογή του Γραφήματος: Ορίζονται ετικέτες και τίτλος με χρήση των plt.xlabel(),

4.3 Προεπεξεργασία Δεδομένων

Στη συνέχεια ακολουθεί η διαδικασία της προεπεξεργασίας των δεδομένων, που περιλαμβάνει τον καθαρισμό κειμένων, τη μετατροπή τους σε πεζά γράμματα, τον διαχωρισμό τους σε λέξεις, την αφαίρεση κοινών λέξεων και την κατηγοριοποίηση των αξιολογήσεων σε συναισθηματικές κατηγορίες (θετικές, ουδέτερες, αρνητικές). Ο κώδικας βρίσκεται στο αρχείο ‘data_preprocessing.py’.

Στον κώδικα πραγματοποιούνται τα εξής:

1. Εισαγωγές βιβλιοθηκών:
   - Εισάγονται βιβλιοθήκες για συνήθεις εκφράσεις (re), επεξεργασία φυσικής γλώσσας (από το NLTK: stopwords και word_tokenize), και το DataFrame από την προηγούμενη ενότητα (df).

2. Συνάρτηση Προεπεξεργασίας Κειμένου (preprocess_text):
   - Αφαίρεση των HTML tags από το κείμενο.
   - Μετατροπή του κειμένου σε πεζά γράμματα.
   - Διαχωρισμός του κειμένου σε λέξεις (tokenization).
   - Αφαίρεση των κοινών λέξεων (stopwords) και διατήρηση μόνο των αλφαριθμητικών λέξεων.
   - Επανασύνδεση των φιλτραρισμένων λέξεων σε μια ενιαία συμβολοσειρά.
3. Επεξεργασία του DataFrame:
   - Αφαίρεση των γραμμών που έχουν κενές τιμές στη στήλη reviewText.
   - Εφαρμογή της συνάρτησης προεπεξεργασίας στη στήλη reviewText και δημιουργία μιας νέας στήλης preprocessed_review.

Σε αυτό το σημείο υλοποιείται η συνάρτηση κατηγοριοποίησης συναισθήματος (categorize_sentiment):

- Αν η αξιολόγηση (rating) είναι μεγαλύτερη ή ίση με 4.0, η συναισθηματική κατηγορία είναι "positive".
- Αν η αξιολόγηση είναι ίση με 3.0, η συναισθηματική κατηγορία είναι "neutral".
- Αν η αξιολόγηση είναι μικρότερη από 3.0, η συναισθηματική κατηγορία είναι "negative".

Έπειτα γίνεται εφαρμογή της παραπάνω συνάρτησης:

- Εφαρμογή της συνάρτησης κατηγοριοποίησης στις αξιολογήσεις (overall) και δημιουργία μιας νέας στήλης sentiment.

4.4 Lexicon Based

Στη συνέχεια υλοποιείται η ανάλυση συναισθήματος με προσέγγιση λεξικού χρησιμοποιώντας το VADER που έχει αναφερθεί προηγουμένως, στο αρχείο ‘lexicon_based_sa_vader.py’.

Εδώ γίνονται τα εξής:

1. Εισαγωγές βιβλιοθηκών:
   - Η βιβλιοθήκη SentimentIntensityAnalyzer από το NLTK για την ανάλυση συναισθήματος.

- Η βιβλιοθήκη tqdm για την εμφάνιση progress bars κατά την εκτέλεση βρόχων.
- Η βιβλιοθήκη pandas για την επεξεργασία δεδομένων σε μορφή DataFrame.
- Η βιβλιοθήκη seaborn για τη δημιουργία γραφημάτων.
- Η βιβλιοθήκη matplotlib.pyplot για την απεικόνιση γραφημάτων.
- Το DataFrame από την προηγούμενη ενότητα προεπεξεργασίας δεδομένων (df).

2. Ανάλυση Συναισθήματος με τον VADER:
   - Αρχικοποίηση του αναλυτή συναισθήματος VADER.
   - Εκτέλεση της ανάλυσης συναισθήματος για το αρχικό κείμενο στο dataset και αποθήκευση των αποτελεσμάτων σε ένα λεξικό res_original.
   - Δημιουργία ενός DataFrame vaders_original από τα αποτελέσματα της ανάλυσης για το αρχικό κείμενο.
   - Εκτέλεση της ανάλυσης συναισθήματος για το προεπεξεργασμένο κείμενο στο dataset και αποθήκευση των αποτελεσμάτων σε ένα λεξικό res_preprocessed.
   - Δημιουργία ενός DataFrame vaders_preprocessed από τα αποτελέσματα της ανάλυσης για το προεπεξεργασμένο κείμενο.

3. Δημιουργία Γραφημάτων:
   - Σχήμα 1: Σύγκριση του compound score ανάμεσα στις αρχικές και προεπεξεργασμένες κριτικές.
   - Σχήμα 2: Παρουσίαση των θετικών, ουδέτερων και αρνητικών σκορ για τις αρχικές κριτικές.
   - Σχήμα 3: Παρουσίαση των θετικών, ουδέτερων και αρνητικών σκορ για τις προεπεξεργασμένες κριτικές.

4.5 Machine Learning – Naive Bayes

Ακολουθεί η προσέγγιση με αλγόριθμο μηχανικής μάθησης, και συγκεκριμένα του Naive Bayes. Ο κώδικας εκπαιδεύει και αξιολογεί ένα μοντέλο Naive Bayes με χρήση των TF-IDF και CountVectorizer. Αναφερόμαστε στο αρχείο ‘ml_sa_naive_bayes.py’

Με τον κώδικα καταφέρνουμε τα εξής:

1. Εισαγωγές:
   - Η βιβλιοθήκη train_test_split από το sklearn.model_selection για τον διαχωρισμό του dataset σε εκπαιδευτικά και δοκιμαστικά σύνολα.
   - Οι διανυσματοποιητές κειμένου TfidfVectorizer και CountVectorizer από το sklearn.feature_extraction.text για τη μετατροπή του κειμένου σε διανύσματα χαρακτηριστικών.
   - Ο αλγόριθμος MultinomialNB από το sklearn.naive_bayes για την εκπαίδευση του μοντέλου Naive Bayes.
   - Οι συναρτήσεις accuracy_score και classification_report από το sklearn.metrics για την αξιολόγηση του μοντέλου.
   - Η συνάρτηση compute_class_weight από το sklearn.utils για τον υπολογισμό των βαρών κλάσεων.
   - Το DataFrame df από την προηγούμενη ενότητα προεπεξεργασίας δεδομένων.

2. Συνάρτηση Εκπαίδευσης και Αξιολόγησης (run_naive_bayes_classification):
   - Βήμα 1: Διάσπαση των δεδομένων σε 80% εκπαιδευτικά και 20% δοκιμαστικά σύνολα.
   - Βήμα 2: Διανυσματοποίηση των δεδομένων κειμένου χρησιμοποιώντας τον συγκεκριμένο διανυσματοποιητή.
   - Βήμα 3: Υπολογισμός των βαρών των κλάσεων για την αντιμετώπιση τυχόν ανισορροπίας στις κλάσεις.
   - Βήμα 4: Εκπαίδευση του μοντέλου Naive Bayes με τα εκπαιδευτικά δεδομένα.
   - Βήμα 5: Προβλέψεις στο δοκιμαστικό σύνολο.
   - Βήμα 6: Αξιολόγηση του μοντέλου και εκτύπωση της ακρίβειας και της αναφοράς ταξινόμησης.

3. Εκτέλεση και Αξιολόγηση:
   - Εκτέλεση της συνάρτησης με χρήση του TfidfVectorizer για τη διανυσματοποίηση του κειμένου.
   - Εκτέλεση της συνάρτησης με χρήση του CountVectorizer για τη διανυσματοποίηση του κειμένου.

4.6 Machine Learning – SVM

Συνεχίζοντας με την προσέγγιση αλγορίθμων μηχανικής μάθησης, ακολουθεί ο SVM. Παρόμοια με προηγουμένως, εκπαιδεύεται και αξιολογείται το μοντέλο SVM με τη χρήση των TF-IDF και CountVectorizer. Ο κώδικας βρίσκεται στο αρχείο 'ml_sa_svm.py'.

Σε αυτό τον κώδικα υλοποιούνται τα εξής:

1. Εισαγωγές:
    - Η βιβλιοθήκη train_test_split από το sklearn.model_selection για τον διαχωρισμό του dataset σε εκπαιδευτικά και δοκιμαστικά σύνολα.
    - Οι διανυσματοποιητές κειμένου TfidfVectorizer και CountVectorizer από το sklearn.feature_extraction.text για τη μετατροπή του κειμένου σε διανύσματα χαρακτηριστικών.
    - Ο αλγόριθμος SVC από το sklearn.svm για την εκπαίδευση του μοντέλου SVM.
    - Οι συναρτήσεις accuracy_score και classification_report από το sklearn.metrics για την αξιολόγηση του μοντέλου.
    - Η συνάρτηση compute_class_weight από το sklearn.utils για τον υπολογισμό των βαρών κλάσεων.
    - Το DataFrame df από την προηγούμενη ενότητα προεπεξεργασίας δεδομένων.

2. Συνάρτηση Εκπαίδευσης και Αξιολόγησης (run_svm_classification):
    - Βήμα 1: Διάσπαση των δεδομένων σε 80% εκπαιδευτικά και 20% δοκιμαστικά σύνολα.
    - Βήμα 2: Διανυσματοποίηση των δεδομένων κειμένου χρησιμοποιώντας τον συγκεκριμένο διανυσματοποιητή.
    - Βήμα 3: Υπολογισμός των βαρών των κλάσεων για την αντιμετώπιση τυχόν ανισορροπίας στις κλάσεις.
    - Βήμα 4: Εκπαίδευση του μοντέλου SVM με τα εκπαιδευτικά δεδομένα.
    - Βήμα 5: Προβλέψεις στο δοκιμαστικό σύνολο.
    - Βήμα 6: Αξιολόγηση του μοντέλου και εκτύπωση της ακρίβειας και της αναφοράς ταξινόμησης.

3. Εκτέλεση και Αξιολόγηση:

- Εκτέλεση της συνάρτησης με χρήση του TfidfVectorizer για τη διανυσματοποίηση του κειμένου.
- Εκτέλεση της συνάρτησης με χρήση του CountVectorizer για τη διανυσματοποίηση του κειμένου.

4.7 Machine Learning – DT

Τέλος υλοποιείται η προσέγγιση με Decision Tree. Ο κώδικας περιλαμβάνει τη διανυσματοποίηση του κειμένου, την εκπαίδευση του μοντέλου, τη βελτιστοποίηση των υπερπαραμέτρων με Grid Search, και την αξιολόγηση της απόδοσης του μοντέλου. Αναφερόμαστε στο αρχείο 'ml_sa_decision_tree.py'.

Συγκεκριμένα:

1. Εισαγωγές:
   - Η βιβλιοθήκη train_test_split από το sklearn.model_selection για τον διαχωρισμό του dataset σε εκπαιδευτικά και δοκιμαστικά σύνολα.
   - Η GridSearchCV για την εύρεση των καλύτερων υπερπαραμέτρων μέσω Grid Search.
   - Το DecisionTreeClassifier από το sklearn.tree για τη δημιουργία του μοντέλου Decision Tree.
   - Οι διανυσματοποιητές κειμένου TfidfVectorizer και CountVectorizer από το sklearn.feature_extraction.text για τη μετατροπή του κειμένου σε διανύσματα χαρακτηριστικών.
   - Οι συναρτήσεις accuracy_score και classification_report από το sklearn.metrics για την αξιολόγηση του μοντέλου.

2. Βήμα 1: Διάσπαση Δεδομένων:
   - Χωρίζει τα δεδομένα σε εκπαιδευτικό και δοκιμαστικό σύνολο, με 80% των δεδομένων να χρησιμοποιούνται για εκπαίδευση και 20% για δοκιμή.

3. Βήμα 2: Διανυσματοποίηση Δεδομένων:
   - Χρησιμοποιείται ο TfidfVectorizer (ή ο CountVectorizer) για τη μετατροπή των προεπεξεργασμένων κειμένων σε αριθμητικά διανύσματα χαρακτηριστικών.

4. Βήμα 3: Ορισμός Πλέγματος Παραμέτρων:

- Ορίζεται ένα πλέγμα υπερπαραμέτρων για το βάθος του δέντρου (max_depth), το ελάχιστο πλήθος δειγμάτων για τον διαχωρισμό κόμβου (min_samples_split), και το ελάχιστο πλήθος δειγμάτων σε φύλλο κόμβου (min_samples_leaf).

5. Βήμα 4: Δημιουργία Μοντέλου:
   - Δημιουργείται το μοντέλο DecisionTreeClassifier.

6. Βήμα 5: Εκτέλεση Grid Search:
   - Εκτελείται το Grid Search με διασταυρούμενη επικύρωση για να βρεθούν οι καλύτερες υπερπαράμετροι για το μοντέλο Decision Tree.

7. Βήμα 6: Λήψη Καλύτερων Παραμέτρων:
   - Εκτυπώνονται οι καλύτερες υπερπαράμετροι που βρέθηκαν από το Grid Search.

8. Βήμα 7: Προβλέψεις με το Καλύτερο Μοντέλο:
   - Χρησιμοποιείται το βελτιστοποιημένο μοντέλο για να γίνουν προβλέψεις στο δοκιμαστικό σύνολο δεδομένων.

9. Βήμα 8: Αξιολόγηση Μοντέλου:
   - Υπολογίζεται και εκτυπώνεται η ακρίβεια του μοντέλου.
   - Εκτυπώνεται η αναφορά ταξινόμησης για περισσότερες λεπτομέρειες.

ΚΕΦΑΛΑΙΟ 5

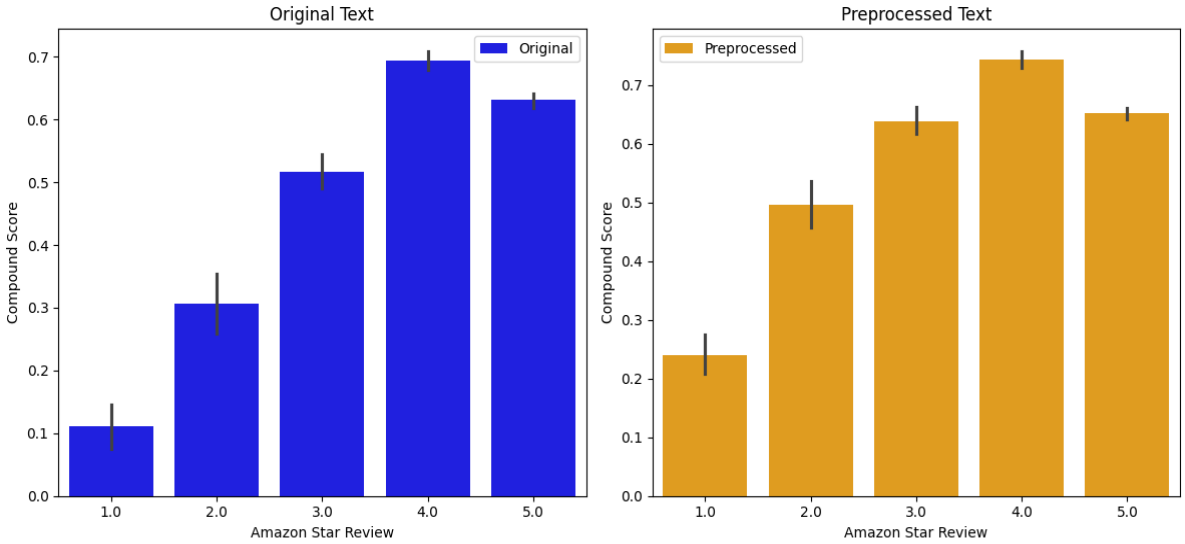
ΑΠΟΤΕΛΕΣΜΑΤΑ ΚΑΙ ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1 Εισαγωγή

Σε αυτό το τελευταίο κεφάλαιο θα παρουσιαστούν τα αποτελέσματα των μεθόδων που παρουσιάστηκαν στα προηγούμενα κεφάλαια. Επιπλέον, μετά την αξιολόγηση τους θα καταλήξουμε φυσικά σε συμπεράσματα που αφορούν ολόκληρη την διπλωματική εργασία.

5.2 Αποτελέσματα από την Ανάλυση Συναισθήματος με VADER

Όσον αφορά τις συνολικές βαθμολογίες ανά αξιολόγηση ο κώδικας μας παρήγαγε το εξής διάγραμμα:



Σχήμα 5.1: Διάγραμμα συνολικών βαθμολογιών αρχικού και προεπεξεργασμένου κειμένου

Για το αρχικό κείμενο:

Παρατηρούμε πως οι βαθμολογίες γενικά αυξάνονται με την αύξηση των αξιολογήσεων αστεριών. Συγκεκριμένα οι αξιολογήσεις με 1 αστέρι έχουν τις χαμηλότερες βαθμολογίες, υποδεικνύοντας πιο αρνητικό συναίσθημα. Οι αξιολογήσεις με 4 αστέρια έχουν τις υψηλότερες σύνθετες βαθμολογίες,

υποδεικνύοντας πιο θετικό συναίσθημα. Υπάρχει μια ελαφρά πτώση στη σύνθετη βαθμολογία για τις αξιολογήσεις με 5 αστέρια σε σύγκριση με αυτές με 4 αστέρια.

Για το προεπεξεργασμένο κείμενο:

Παρόμοια με το αρχικό κείμενο, οι σύνθετες βαθμολογίες αυξάνονται με την αύξηση των αξιολογήσεων αστεριών. Οι αξιολογήσεις με 1 αστέρι έχουν τις χαμηλότερες σύνθετες βαθμολογίες και οι αξιολογήσεις με 4 αστέρια έχουν τις υψηλότερες. Τα προεπεξεργασμένο δεδομένα δείχνουν μια πιο σταθερή αύξηση στις βαθμολογίες σε όλες τις αξιολογήσεις αστεριών σε σύγκριση με τα αρχικά. Η πτώση στη σύνθετη βαθμολογία για τις αξιολογήσεις με 5 αστέρια σε σύγκριση με αυτές με 4 αστέρια είναι λιγότερο έντονη στο προεπεξεργασμένο κείμενο.

Συνεχίζοντας ας παρατηρήσουμε τις βαθμολογίες μεμονωμένες στα αρχικά δεδομένα:

Σχήμα 5.2: Διάγραμμα βαθμολογιών αρχικού κειμένου

Οι θετικές βαθμολογίες αυξάνονται με την αύξηση των αξιολογήσεων αστεριών. Οι αξιολογήσεις με 5 αστέρια έχουν τις υψηλότερες θετικές βαθμολογίες, ενώ οι αξιολογήσεις με 1 αστέρι έχουν τις χαμηλότερες.

Οι ουδέτερες βαθμολογίες παραμένουν σχετικά σταθερές σε όλες τις αξιολογήσεις αστεριών, με ελαφρές διακυμάνσεις. Επομένως βλέπουμε πως οι αξιολογήσεις είναι κυρίως ουδέτερες σε όλες τις κατηγορίες αστεριών.

Οι αρνητικές βαθμολογίες μειώνονται με την αύξηση των αξιολογήσεων αστεριών. Οι αξιολογήσεις με 1 αστέρι έχουν τις υψηλότερες αρνητικές βαθμολογίες, ενώ οι αξιολογήσεις με 5 αστέρια έχουν τις χαμηλότερες, όπως και αναμένεται λογικά.

Τέλος ας παρατηρήσουμε το διάγραμμα για τα προεπεξεργασμένα δεδομένα

Σχήμα 5.3: Διάγραμμα βαθμολογιών προεπεξεργασμένου κειμένου

Οι θετικές βαθμολογίες αυξάνονται με την αύξηση των αξιολογήσεων αστεριών, παρόμοια με το αρχικό κείμενο.Οι αξιολογήσεις με 5 αστέρια έχουν τις υψηλότερες θετικές βαθμολογίες, και οι αξιολογήσεις με 1 αστέρι έχουν τις χαμηλότερες. Όμως η αύξηση στις θετικές βαθμολογίες είναι πιο έντονη στο προεπεξεργασμένο κείμενο.

Οι ουδέτερες βαθμολογίες είναι ελαφρώς χαμηλότερες σε σύγκριση με το αρχικό κείμενο αλλά παραμένουν σχετικά σταθερές σε όλες τις αξιολογήσεις αστεριών. Επιπλέον υπάρχει μια ελαφρά μείωση στις ουδέτερες βαθμολογίες για τις υψηλότερες αξιολογήσεις αστεριών.

Οι αρνητικές βαθμολογίες μειώνονται με την αύξηση των αξιολογήσεων αστεριών, παρόμοια με το αρχικό κείμενο. Οι αξιολογήσεις με 1 αστέρι έχουν τις υψηλότερες αρνητικές βαθμολογίες, και οι αξιολογήσεις με 5 αστέρια έχουν τις χαμηλότερες. Άρα η μείωση στις αρνητικές βαθμολογίες είναι ελαφρώς πιο έντονη στο προεπεξεργασμένο κείμενο.

Συνοψίζοντας τα παραπάνω, η προεπεξεργασία των δεδομένων τείνει να κάνει τις βαθμολογίες συναισθημάτων πιο έντονες. Οι θετικές βαθμολογίες αυξάνονται εμφανώς και οι αρνητικές βαθμολογίες μειώνονται, με τις υψηλότερες αξιολογήσεις αστεριών να είναι στο προεπεξεργασμένο κείμενο σε σύγκριση με το αρχικό κείμενο.

Και τα δύο, το αρχικό και το προεπεξεργασμένο κείμενο, δείχνουν μια συνεπή τάση όπου οι υψηλότερες αξιολογήσεις αστεριών συνδέονται με πιο θετικό συναίσθημα και λιγότερο αρνητικό συναίσθημα.

Όσον αφορά τις ουδέτερες βαθμολογίες, παραμένουν σχετικά σταθερές σε διαφορετικές αξιολογήσεις αστεριών, υποδεικνύοντας ότι πολλές αξιολογήσεις περιέχουν σημαντική ποσότητα ουδέτερου συναισθήματος.

Συμπερασματικά, καταλήγουμε πως η προεπεξεργασία ενισχύει την καθαρότητα των αποτελεσμάτων της ανάλυσης συναισθημάτων, καθιστώντας τα θετικά και αρνητικά συναισθήματα πιο διακριτά σε διάφορες αξιολογήσεις αστεριών.

5.3 Αποτελέσματα από την Ανάλυση Συναισθήματος με Naive Bayes

Στη συνέχεια θα παρατηρήσουμε τα αποτελέσματα της προσέγγισης με Naive Bayes και θα εξάγουμε συμπεράσματα.

NB με TfidfVectorizer:

Ακρίβεια: 46.19%

Precision: Οι αξιολογήσεις με 1, 2 και 3 αστέρια έχουν πολύ υψηλή precision, αλλά αυτό δεν αντικατοπτρίζει πραγματική ακρίβεια λόγω του χαμηλού recall.

Recall: Οι αξιολογήσεις με 1, 2 και 3 αστέρια έχουν πολύ χαμηλό recall, υποδηλώνοντας ότι πολλά δείγματα αυτών των κατηγοριών δεν αναγνωρίζονται σωστά.

F1-Score: Οι αξιολογήσεις με 5 αστέρια έχουν το υψηλότερο f1-score (0.63), υποδεικνύοντας καλύτερη απόδοση σε αυτήν την κατηγορία, ενώ οι άλλες κατηγορίες έχουν πολύ χαμηλά f1-scores.

Μέσοι Όροι:

Macro Average: 0.85 precision, 0.21 recall, 0.15 f1-score.

Weighted Average: 0.70 precision, 0.46 recall, 0.30 f1-score.

NB με CountVectorizer:

Ακρίβεια: 59.59%

Precision: Η ακρίβεια βελτιώνεται για τις περισσότερες κατηγορίες, ιδιαίτερα για τις αξιολογήσεις με 2 και 5 αστέρια.

Recall: Οι αξιολογήσεις με 5 αστέρια έχουν το υψηλότερο recall (0.80), υποδεικνύοντας ότι αυτή η κατηγορία αναγνωρίζεται περισσότερο σωστά σε σύγκριση με τις άλλες.

F1-Score: Οι αξιολογήσεις με 5 αστέρια έχουν το υψηλότερο f1-score (0.74), δείχνοντας καλύτερη συνολική απόδοση σε αυτήν την κατηγορία.

Μέσοι Όροι:

Macro Average: 0.60 precision, 0.46 recall, 0.46 f1-score.

Weighted Average: 0.59 precision, 0.60 recall, 0.57 f1-score.

Η ακρίβεια αυξάνεται από 46.19% με το TfidfVectorizer σε 59.59% με το CountVectorizer, δείχνοντας ότι η απλή μέθοδος μέτρησης των λέξεων μπορεί να είναι πιο αποτελεσματική για το Naive Bayes σε αυτή την περίπτωση.

Επίσης είναι υψηλότερη για τις αξιολογήσεις με 5 αστέρια σε σύγκριση με τις άλλες κατηγορίες, και στα δύο μοντέλα. Οι αξιολογήσεις με 2 αστέρια έχουν πολύ χαμηλό recall και στα δύο μοντέλα, υποδεικνύοντας δυσκολία στην αναγνώρισή τους.

Οι μέσοι όροι macro είναι χαμηλότεροι από τους weighted averages, δείχνοντας ότι η απόδοση είναι καλύτερη στις κατηγορίες με περισσότερα δείγματα (όπως η κατηγορία με 5 αστέρια).

Τα αποτελέσματα υποδεικνύουν ότι η επιλογή του vectorizer μπορεί να επηρεάσει σημαντικά την απόδοση του Naive Bayes. Το CountVectorizer φαίνεται να αποδίδει καλύτερα σε αυτή την περίπτωση. Ωστόσο, υπάρχουν δυσκολίες στην κατηγοριοποίηση των ενδιάμεσων αξιολογήσεων (2 και 3 αστέρια), και η απόδοση μπορεί να βελτιωθεί με την περαιτέρω βελτιστοποίηση του μοντέλου ή την χρήση πιο εξελιγμένων τεχνικών προεπεξεργασίας κειμένου.

5.4 Αποτελέσματα από την Ανάλυση Συναισθήματος με SVM

Ακολουθούν τα αποτελέσματα της προσέγγισης με SVM και η εξαγωγή συμπερασμάτων.

SVM με TfidfVectorizer:

Ακρίβεια: 58.45%

Precision: Οι αξιολογήσεις με 1 και 5 αστέρια έχουν τις υψηλότερες τιμές precision, ιδιαίτερα η κατηγορία με 5 αστέρια (0.79).

Recall: Οι κατηγορίες με 1 και 5 αστέρια έχουν υψηλό recall, ενώ οι ενδιάμεσες κατηγορίες (2, 3 και 4 αστέρια) έχουν χαμηλότερο recall.

F1-Score: Οι κατηγορίες με 1 και 5 αστέρια έχουν υψηλότερα f1-scores (0.66 και 0.71 αντίστοιχα), υποδεικνύοντας καλύτερη συνολική απόδοση σε αυτές τις κατηγορίες.

Μέσοι Όροι:

Macro Average: 0.52 precision, 0.55 recall, 0.53 f1-score.

Weighted Average: 0.61 precision, 0.58 recall, 0.59 f1-score.

SVM με CountVectorizer:

Ακρίβεια: 58.69%

Precision: Οι αξιολογήσεις με 1 και 5 αστέρια έχουν και πάλι τις υψηλότερες τιμές precision, με την κατηγορία 5 αστέρων να έχει precision 0.73.

Recall: Η κατηγορία με 5 αστέρια διατηρεί υψηλό recall (0.73), ενώ οι ενδιάμεσες κατηγορίες (2, 3 και 4 αστέρια) έχουν χαμηλότερο recall.

F1-Score: Οι κατηγορίες με 1 και 5 αστέρια έχουν υψηλότερα f1-scores (0.62 και 0.73 αντίστοιχα), δείχνοντας καλύτερη απόδοση σε αυτές τις κατηγορίες.

Μέσοι Όροι:

Macro Average: 0.51 precision, 0.53 recall, 0.51 f1-score.

Weighted Average: 0.59 precision, 0.59 recall, 0.59 f1-score.

Συγκεκριμένες Παρατηρήσεις:

Η ακρίβεια είναι σχεδόν η ίδια και με τα δύο vectorizers (58.45% με TfidfVectorizer και 58.69% με CountVectorizer), δείχνοντας ότι και οι δύο μέθοδοι μέτρησης των λέξεων έχουν παρόμοια απόδοση για το SVM.

Οι αξιολογήσεις με 5 αστέρια έχουν σταθερά υψηλότερο precision και recall και στα δύο μοντέλα, υποδεικνύοντας ότι το SVM μπορεί να αναγνωρίσει καλύτερα αυτήν την κατηγορία. Οι ενδιάμεσες κατηγορίες (2, 3 και 4 αστέρια) έχουν χαμηλότερο precision και recall, δείχνοντας δυσκολία στην ακριβή αναγνώρισή τους.

Οι μέσοι όροι macro είναι χαμηλότεροι από τους weighted averages, υποδηλώνοντας ότι η απόδοση είναι καλύτερη στις κατηγορίες με περισσότερα δείγματα (όπως η κατηγορία με 5 αστέρια).

Αυτά τα αποτελέσματα δείχνουν ότι το SVM παρουσιάζει σταθερή απόδοση, ανεξάρτητα από τον επιλεγμένο vectorizer. Οι κατηγορίες με 5 αστέρια αναγνωρίζονται καλύτερα, αλλά η απόδοση στις ενδιάμεσες κατηγορίες παραμένει χαμηλότερη. Η βελτιστοποίηση των παραμέτρων θα μπορούσε να βελτιώσει περαιτέρω την απόδοση.

5.5 Αποτελέσματα από την Ανάλυση Συναισθήματος με Decision Tree

Τέλος θα αξιολογήσουμε τα αποτελέσματα της μεθόδου με Decision Tree.

Ακρίβεια: 52.56%

Precision: Οι αξιολογήσεις με 1 αστέρι έχουν precision 0.49, δείχνοντας μέτρια ικανότητα αναγνώρισης. Οι αξιολογήσεις με 5 αστέρια έχουν το υψηλότερο precision (0.61), δείχνοντας την καλύτερη απόδοση στην αναγνώριση αυτής της κατηγορίας.

Recall: Η κατηγορία με 5 αστέρια έχει το υψηλότερο recall (0.78), υποδεικνύοντας ότι το μοντέλο μπορεί να αναγνωρίσει την πλειοψηφία αυτών των δειγμάτων. Οι ενδιάμεσες κατηγορίες (2, 3 και 4 αστέρια) έχουν χαμηλότερο recall, με την κατηγορία 2 αστέρων να έχει το χαμηλότερο recall (0.22).

F1-Score: Η κατηγορία με 5 αστέρια έχει το υψηλότερο f1-score (0.68), ενώ οι ενδιάμεσες κατηγορίες (2, 3 και 4 αστέρια) έχουν χαμηλότερα f1-scores (0.27, 0.28, και 0.39 αντίστοιχα).

Μέσοι Όροι:

Macro Average: 0.44 precision, 0.40 recall, 0.41 f1-score.

Weighted Average: 0.50 precision, 0.53 recall, 0.51 f1-score.

Η συνολική ακρίβεια του Δέντρου Αποφάσεων είναι 52.56%, η οποία είναι μέτρια και χαμηλότερη σε σύγκριση με άλλα μοντέλα όπως το SVM.

Η κατηγορία με 5 αστέρια έχει το υψηλότερο precision και recall, δείχνοντας ότι το μοντέλο αναγνωρίζει καλύτερα τις θετικές αξιολογήσεις. Ωστόσο, οι ενδιάμεσες κατηγορίες (2, 3 και 4

αστέρια) παρουσιάζουν χαμηλότερη απόδοση, με το precision και recall να είναι σημαντικά χαμηλότερα.

Οι μέσοι όροι macro είναι χαμηλότεροι από τους weighted averages, υποδηλώνοντας ότι η απόδοση είναι καλύτερη στις κατηγορίες με περισσότερα δείγματα, ιδιαίτερα στην κατηγορία με 5 αστέρια.

Το Δέντρο Αποφάσεων παρουσιάζει μέτρια απόδοση στην αναγνώριση των αξιολογήσεων. Η βελτίωση του μοντέλου μπορεί να επιτευχθεί με την εφαρμογή τεχνικών όπως η επιλογή χαρακτηριστικών, η αύξηση του βάθους του δέντρου, ή η χρήση άλλων προχωρημένων τεχνικών προεπεξεργασίας δεδομένων.

5.6 Τελικά Συμπεράσματα

Μπορούμε να διαπιστώσουμε πως η συναισθηματική ανάλυση είναι μια πολύπλοκη διαδικασία και εξαρτάται από πάρα πολλούς παραμέτρους. Η κύρια παράμετρος είναι η ποιότητα και η ποσότητα των δεδομένων. Για αυτό το λόγο έχει τεράστια σημασία η επιλογή των δεδομένων αλλά και η επεξεργασία τους, πριν τα τροφοδοτήσουμε σε κάποια μοντέλο. Τα αποτελέσματα που πήραμε από τις διάφορες μεθόδους που χρησιμοποιήσαμε δεν ήταν η ιδανικά και ο λόγος είναι βρίσκεται τόσο στην έλλειψη συνόλου δεδομένων αλλά και στο γεγονός ότι η τελειοποίηση ενός συστήματος χρειάζεται πολλές δοκιμές και αλλαγές στην παραμετροποίηση του κάθε μοντέλου.

Το σύνολο δεδομένων αποτελείται από κριτικές του Amazon με βαθμολογίες που κυμαίνονται από 1 έως 5 αστέρια. Συνήθως, τέτοια σύνολα δεδομένων είναι ανισόρροπα, με μεγαλύτερο αριθμό θετικών κριτικών (4-5 αστέρια) σε σύγκριση με τις αρνητικές κριτικές (1-2 αστέρια). Αυτό φαίνεται και στα διαγράμματα που εξετάσαμε παραπάνω. Επίσης οι κριτικές είναι δεδομένα κειμένου, τα οποία πρέπει να μετατραπούν σε αριθμητικές αναπαραστάσεις για αλγορίθμους μηχανικής μάθησης. Για αυτό και είναι τόσο σημαντική η διαδικασία που εφαρμόζεται για τη μετατροπή του κειμένου σε αριθμητικά χαρακτηριστικά.

Η ακρίβεια δείχνει τη συνολική ορθότητα. Η υψηλότερη ακρίβεια υποδηλώνει καλύτερη απόδοση, αλλά μπορεί να είναι παραπλανητική σε μη ισορροπημένα σύνολα δεδομένων. Η υψηλότερη ακρίβεια παρατηρήθηκε με το Naive Bayes με CountVectorizer (59,59%).

Η ακρίβεια και ανάκληση είναι ιδιαίτερα σημαντικές σε περιπτώσεις όπου το κόστος των ψευδώς θετικών και των ψευδώς αρνητικών είναι μεγάλο. Η ισορροπημένη ακρίβεια και ανάκληση είναι ιδανικές συνθήκες. Το Naive Bayes με CountVectorizer παρουσίασε ισορροπημένη ακρίβεια και ανάκληση.

Το F1-Score εξισορροπεί την ακρίβεια και την ανάκληση, και είναι ιδιαίτερα χρήσιμο για μη ισορροπημένα σύνολα δεδομένων. Το SVM με TfidfVectorizer και το Naive Bayes με CountVectorizer παρουσίασαν τον υψηλότερο σταθμισμένο μέσο όρο F1-Score (0,59 και 0,57 αντίστοιχα).

Με βάση τις παραπάνω μετρήσεις και απεικονίσεις η μέθοδος Naive Bayes με CountVectorizer φαίνεται να αποδίδει συνολικά καλύτερα από άποψη ακρίβειας και ισορροπημένης ακρίβειας/ανάκλησης.

Σε δεύτερη θέση η μέθοδος SVM με το TfidfVectorizer έχει επίσης καλές επιδόσεις, ιδίως όσον αφορά το F1-Score.

Επιλέγουμε το Naive Bayes με CountVectorizer αν η προτεραιότητα βρίσκεται στην ακρίβεια και την ισορροπημένη απόδοση σε όλες τις μετρικές. Εν΄β εάν η διερμηνεία και η ευστάθεια είναι πιο κρίσιμες, θα εξετάσουμε με το SVM με TfidfVectorizer.

Τα αποτελέσματα της παρούσας μελέτης υπογραμμίζουν τη σημασία της διαδικασίας της τελειοποίησης και προσαρμογής των μοντέλων συναισθηματικής ανάλυσης. Η ανάγκη για επαναλαμβανόμενες δοκιμές και αλλαγές στην παραμετροποίηση των μοντέλων αποτελεί ουσιαστικό μέρος της διαδικασίας βελτίωσης της απόδοσής τους. Με τη συνεχή ανάπτυξη των ερευνητικών προσεγγίσεων και τη βελτίωση της διαθεσιμότητας κατάλληλων δεδομένων, είναι πιθανό να επιτευχθεί η ανάπτυξη πιο ακριβών και αποτελεσματικών μοντέλων για τη συναισθηματική ανάλυση σε μελλοντικές έρευνες και εφαρμογές.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Ni, Jianmo, Jiacheng Li, and Julian Mcauley. n.d. Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects.

[2] Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies, #37. Morgan & Claypool Publishers.

[3] What is Natural Language Processing? | IBM. (n.d.). What Is Natural Language Processing? | IBM. Ανακτήθηκε 25 Ιουνίου 2023, από https://www.ibm.com/topics/natural-language- processing