Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής

Πρόγραμμα Μεταπτυχιακών Σπουδών

«Ψηφιακός Πολιτισμός, Έξυπνες Πόλεις, IoT και Προηγμένες Ψηφιακές Τεχνολογίες»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Εξόρυξη Δεδομένων στα μέσα κοινωνικής δικτύωσης** **Data Mining in Social Media**
Ονοματεπώνυμο Φοιτητή	**Σταυρούλα Διακουμάκου**
Πατρώνυμο	**Γεώργιος**
Αριθμός Μητρώου	**ΨΠΟΛ20018**
Επιβλέπων	**Διονύσης Σωτηρόπουλος, Επίκουρος Καθηγητής**

Ημερομηνία παράδοσης Δεκεμβρίου 2023

Τριμελής Εξεταστική Επιτροπή

Δ. Σωτηρόπουλος	Δ. Βέργαδος	Ε. Σκόνδρας
Επίκουρος Καθηγητής	Καθηγητής	Διδάσκων ΠΜΣ

Αθήνα, Σεπτέμβριος 2023

Πίνακας Περιεχομένων

Περίληψη

Η διατριβή αυτή αναλύει εκτενώς την εφαρμογή των τεχνικών μηχανικής μάθησης στον σημαντικό τομέα της ανάλυσης συναισθημάτων. Πραγματευόμαστε τον τρόπο με τον οποίο η μηχανική μάθηση μπορεί να χρησιμοποιηθεί για την εξαγωγή σημαντικών πληροφοριών σχετικά με το συναισθηματικό περιεχόμενο σε κείμενα, εστιάζοντας ιδιαίτερα στην ανάλυση κριτικών ταινιών. Μέσα από την παρουσίαση της διαδικασίας που ακολούθησε η έρευνα, προσφέρουμε μια ενδιαφέρουσα εισαγωγή στον κόσμο της Sentiment Analysis.

Η αφετηρία της έρευνας αποτέλεσε μια συλλογή από 50000 κριτικές ταινιών που προέρχονταν από τη δημοφιλή ιστοσελίδα IMDB. Ωστόσο, δεν αρκεί απλώς να συλλέξουμε κείμενα. Ήταν απαραίτητη η εφαρμογή διάφορων τεχνικών προεπεξεργασίας για τον καθαρισμό και την προετοιμασία αυτών των κειμένων. Αυτό διευκολύνει τη διαδικασία ανάλυσης και εξάλειψης των παραμορφώσεων που μπορούν να επηρεάσουν τα αποτελέσματα.

Το κύριο ερώτημα που αντιμετωπίσαμε ήταν πώς μπορούμε να κατηγοριοποιήσουμε τις κριτικές αυτές σε κατηγορίες συναισθημάτων. Χρησιμοποιήσαμε διάφορους αλγορίθμους μηχανικής μάθησης για την επίτευξη αυτού του στόχου. Οι αλγόριθμοι αυτοί παρέχουν τη δυνατότητα ταξινόμησης των κριτικών σε διάφορες κατηγορίες συναισθημάτων, όπως "θετικές," "αρνητικές," ή "ουδέτερες."

Σημαντικό μέρος της έρευνάς μας αποτελεί η αξιολόγηση της απόδοσης των αλγορίθμων. Χρησιμοποιήσαμε μετρικές όπως η ακρίβεια, η ανάκληση και το F1-score για να αξιολογήσουμε την ακρίβεια και την αποτελεσματικότητα των μοντέλων μας.

Τα αποτελέσματα της έρευνάς μας αποδεικνύουν ότι οι τεχνικές Μηχανικής Μάθησης μπορούν να χρησιμοποιηθούν αποτελεσματικά για την ανάλυση συναισθημάτων σε κριτικές ταινιών. Αυτό το εύρημα ανοίγει νέους δρόμους για την εφαρμογή της μηχανικής μάθησης σε διάφορους τομείς της βιομηχανίας κινηματογράφου και της ανάλυσης κειμένου.

Επιπλέον, αυτή η διατριβή δεν αφορά απλώς τη θεωρία και την εφαρμογή τεχνικών, αλλά επιδιώκει την πρακτική εφαρμογή των ευρημάτων της. Προσφέρει χρήσιμες πληροφορίες και εργαλεία για τους επαγγελματίες του κινηματογράφου, επιτρέποντάς τους να κατανοήσουν καλύτερα την αντίδραση του κοινού στις ταινίες τους.

Τα συμπεράσματα και οι μελλοντικές επεκτάσεις που παρουσιάζονται στη διατριβή ενισχύουν τη σημασία της έρευνας σε αυτόν τον τομέα. Επισημαίνουν τη σημασία της συνεχούς ανάπτυξης και βελτίωσης των τεχνικών ανάλυσης συναισθημάτων, ενισχύοντας την αντιμετώπιση πιο σύνθετων και προκλητικών προβλημάτων ανάλυσης συναισθημάτων σε κείμενα.

Συνολικά, αυτή η διατριβή αποτελεί μια πολύτιμη πηγή γνώσης για εκείνους που ασχολούνται με την ανάλυση συναισθημάτων, τη μηχανική μάθηση και τη βιομηχανία του κινηματογράφου. Επιπλέον, προτρέπει για περαιτέρω εξερεύνηση και ανάπτυξη του θέματος, ενθαρρύνοντας περαιτέρω επιστημονική έρευνα σε αυτόν τον τομέα.

Λέξεις-κλειδιά: Ανάλυση Συναισθημάτων, Μηχανική Μάθηση, Κριτικές Ταινιών, Χαρακτηριστικά Κειμένου, Ταξινόμηση, Προεπεξεργασία, Δημόσια Αντίδραση, Βιομηχανία Κινηματογράφου.

Εισαγωγή

Τι είναι το data mining;

Το Data Mining αποτελεί κεντρική τεχνολογία στον χώρο της πληροφορικής και της ανάλυσης δεδομένων. Καθώς η ποσότητα των δεδομένων αυξάνεται εκθετικά, το Data Mining αναδεικνύεται ως κρίσιμο εργαλείο για την εξαγωγή πληροφοριών. Στοχεύει στην ανακάλυψη κρυμμένων προτύπων, σχέσεων και τάσεων σε μεγάλα σύνολα δεδομένων.

Η εφαρμογή αλγορίθμων και τεχνικών αναλυτικής επεξεργασίας δεδομένων γίνεται γνωστή ως Knowledge Discovery in Databases (KDD). Ο στόχος είναι η ανακάλυψη κρυμμένων, πολύτιμων πληροφοριών που διαφεύγουν της αντίληψης μας στα δεδομένα.

Η σημασία του Data Mining αντικατοπτρίζεται στη δυνατότητά του να μετατρέπει ακατέργαστα δεδομένα σε πληροφορίες με εφαρμογή και νόημα. Επιτρέπει την ανακάλυψη κρυμμένων μοτίβων, την πρόβλεψη μελλοντικών γεγονότων και τη λήψη αποφάσεων βασισμένων σε δεδομένα.

Βασικές έννοιες περιλαμβάνουν την εξόρυξη προτύπων, που αναγνωρίζει και εξάγει κοινά σύνολα δεδομένων με συγκεκριμένα πρότυπα. Επίσης, η κατηγοριοποίηση ταξινομεί τα δεδομένα σε κατηγορίες, ενώ η πρόβλεψη εκτιμά τιμές βασισμένες στα δεδομένα.

Οι τεχνικές περιλαμβάνουν τη συσταδοποίηση, που χωρίζει τα δεδομένα σε ομάδες βάσει ομοιότητας, και τα δέντρα αποφάσεων που αναπαριστούν διακλαδώσεις αποφάσεων, χρησιμοποιούμενα για κατηγοριοποίηση.

Η εξόρυξη προτύπων αποτελεί κεντρική διαδικασία στο Data Mining. Πρόκειται για την αναγνώριση τυπικών συμπεριφορών ή σχέσεων στα δεδομένα. Οι αλγόριθμοι εξόρυξης προτύπων αναζητούν συχνά εμφανιζόμενα σύνολα αντικειμένων ή γεγονότων που εμφανίζονται μαζί. Για παράδειγμα, σε δεδομένα αγοραπωλησίας, η εξόρυξη προτύπων μπορεί να ανακαλύψει ότι οι πελάτες που αγοράζουν προϊόντα Α και Β τείνουν επίσης να αγοράζουν και το προϊόν C.

Η κατηγοριοποίηση αποτελεί μια μορφή επιβλεπόμενης μάθησης, όπου τα δεδομένα εκπαιδεύονται σε ένα μοντέλο για να ταξινομηθούν σε προκαθορισμένες κατηγορίες. Παράδειγμα είναι η κατηγοριοποίηση email ως "spam" ή "μη spam". Αντίθετα, η πρόβλεψη εστιάζει στο να προβλέπει μελλοντικές τιμές, όπως η τιμή μιας μετοχής ή η πρόβλεψη πωλήσεων.

Η συσταδοποίηση είναι μια τεχνική που αποσκοπεί στο να ομαδοποιήσει τα δεδομένα σε σύνολα, ή συστάδες, βάσει της ομοιότητας τους. Αυτή η τεχνική μπορεί να αποκαλύψει μοτίβα και σχέσεις που μπορεί να μην είναι φανερά ορατά. Για παράδειγμα, σε δεδομένα καταναλωτικής συμπεριφοράς, η συσταδοποίηση μπορεί να ομαδοποιήσει πελάτες με παρόμοια αγοραπωλησιακή συμπεριφορά.

Τέλος, τα Δέντρα Αποφάσεων είναι μια μορφή αλγορίθμου που χρησιμοποιείται για την κατηγοριοποίηση. Το δέντρο αποφάσεων αναπαριστά μια σειρά από αποφάσεις βασισμένες σε χαρακτηριστικά των δεδομένων. Κάθε κόμβος αντιπροσωπεύει μια ερώτηση ή μια απόφαση, ενώ οι κλαδιές αντιπροσωπεύουν τις πιθανές απαντήσεις ή πορείες που μπορεί να ακολουθηθούν.

Αυτά τα στοιχεία αποτελούν μόνο ένα μικρό δείγμα του τι μπορεί να επιτευχθεί με το Data Mining. Οι προκλήσεις του πεδίου περιλαμβάνουν τη διαχείριση μεγάλων όγκων δεδομένων, την αντιμετώπιση της ανασφάλειας και τον σεβασμό στην ιδιωτικότητα των δεδομένων. Παράλληλα, οι μελλοντικές εξελίξεις εστιάζονται στην ενσωμάτωση νέων τεχνολογιών όπως τα νευρωνικά δίκτυα και η αυτόματη μάθηση.

Στον πραγματικό κόσμο, το Data Mining εφαρμόζεται σε ποικίλους τομείς, από την υγεία έως την επιχειρηματικότητα, παρέχοντας πρακτικά παραδείγματα αξιοποίησης της τεχνολογίας.

Παρά τις ευκαιρίες, το Data Mining αντιμετωπίζει προκλήσεις. Ωστόσο, ενδεχόμενες εξελίξεις και κατευθύνσεις αναλύονται, αναδεικνύοντας τη συνεχή εξέλιξη του πεδίου.

Συνολικά, το Data Mining αποτελεί κρίσιμο εργαλείο για την ανακάλυψη γνώσης από τα δεδομένα, με δυνατότητα καθοδήγησης αποφάσεων και δημιουργίας καινοτόμων λύσεων. [1],[2].

Η ιστορία της

Η ιστορία του data mining ξεκινά περίπου τη δεκαετία του 1960, αν και οι ρίζες του σχετίζονται με προηγούμενες εξελίξεις στον τομέα της στατιστικής και της θεωρίας της πληροφορίας. Κατά

τη διάρκεια αυτής της περιόδου, οι ερευνητές και οι επιστήμονες άρχισαν να αντιλαμβάνονται την αξία των δεδομένων και τη δυνατότητα ανάκτησης κρυμμένων πληροφοριών από αυτά.

Το πρώτο σημαντικό βήμα προς το data mining ήταν η ανάπτυξη της στατιστικής ανάλυσης δεδομένων και των μεθόδων ανακάλυψης γνώσης από τις βάσεις δεδομένων. Η εφαρμογή στατιστικών μεθόδων, όπως οι γραμμικές παλινδρομήσεις και οι αναλύσεις διακύμανσης, επέτρεψε στους ερευνητές να ανακαλύπτουν πρότυπα και συσχετίσεις μεταξύ δεδομένων.

Οι πρώτες δεκαετίες του 20ού αιώνα έφεραν την ανάπτυξη των πρώτων υπολογιστών και τη χρήση τους για την αποθήκευση μεγάλων ποσοτήτων δεδομένων. Ο όρος "data mining" αναφέρεται για πρώτη φορά στη δεκαετία του 1960 από τον χημικό και στατιστικό προγραμματιστή John W. Tukey, ο οποίος χρησιμοποίησε τον όρο "data analysis" για να περιγράψει τη διαδικασία εξαγωγής χρήσιμης πληροφορίας από μεγάλα σύνολα δεδομένων.

Στη δεκαετία του 1980, η ανάπτυξη των συστημάτων διαχείρισης βάσεων δεδομένων (DBMS) παρείχε τη δυνατότητα αποθήκευσης και επεξεργασίας μεγάλων ποσοτήτων δεδομένων. Αυτό άνοιξε τον δρόμο για περαιτέρω έρευνα στον τομέα του data mining. Οι ερευνητές άρχισαν να αναπτύσσουν μεθόδους για τον αναγνωρισμό μοντέλων και προτύπων από δεδομένα.

Στα τέλη της δεκαετίας του 1980 και τις αρχές της δεκαετίας του 1990, η ανάπτυξη των αλγορίθμων μηχανικής μάθησης, όπως οι δέντρα αποφάσεων και οι ταξινομητές k-πλησιέστερων γειτόνων, άρχισε να επιτρέπει την αυτόματη ανακάλυψη προτύπων από τα δεδομένα.

Η δεκαετία του 1990 έφερε την εμφάνιση περισσότερων συστημάτων data mining και εργαλείων λογισμικού που επέτρεπαν σε επιχειρήσεις και ερευνητές να εφαρμόζουν αποτελεσματικά τις τεχνικές data mining στα πραγματικά δεδομένα τους.

Συνολικά, η ιστορία του data mining αντιπροσωπεύει μια συνεχή εξέλιξη από τη χρήση βασικών στατιστικών μεθόδων στη δεκαετία του 1960 έως την υιοθέτηση προηγμένων αλγορίθμων μηχανικής μάθησης στις μέρες μας. Η ανάπτυξη τεχνολογιών όπως τα σημασιολογικά μοντέλα και τα νευρωνικά δίκτυα συνεχίζει να διαμορφώνει τον τομέα και να επεκτείνει τα όρια του.[3].

Χρησιμότητα της εξόρυξης δεδομένων

Η εξόρυξη δεδομένων μπορεί να βοηθήσει σε πολλούς τρόπους μερικοί από τους οποίους αναφέρονται παρακάτω [1], [2]:

- ***Ανακάλυψη Κρυμμένων Πληροφοριών:*** Οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να αποκαλύψουν κρυμμένες συσχετίσεις, τάσεις ή μοτίβα μεταξύ των δεδομένων. Αυτή η ανακάλυψη πληροφοριών μπορεί να οδηγήσει στην αναγνώριση νέων ευκαιριών ή προκλήσεων.
- ***Πρόβλεψη Μελλοντικών Συμβάντων:*** Οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να αναπτύξουν μοντέλα πρόβλεψης που βασίζονται στα ιστορικά δεδομένα. Αυτά τα μοντέλα μπορούν να προβλέψουν μελλοντικά συμβάντα ή τάσεις, βοηθώντας έτσι στη λήψη αποφάσεων και στρατηγικών προβλέψεων.
- ***Κατανόηση των Πελατών:*** Με την εξόρυξη δεδομένων, είναι δυνατή η ανάλυση των προτιμήσεων, συμπεριφοράς και αναγκών των πελατών. Αυτή η κατανόηση μπορεί να βοηθήσει στην προσαρμογή των προϊόντων και των υπηρεσιών προς τις ατομικές ανάγκες και στην αύξηση της ικανοποίησης των πελατών.

- ***Ανίχνευση Απάτης και Ασφάλειας:*** Οι αλγόριθμοι εξόρυξης δεδομένων μπορούν να ανακαλύψουν ατυχήματα, απάτες ή παραβάσεις ασφάλειας από τα δεδομένα. Αυτό μπορεί να βοηθήσει στην πρόληψη και ανίχνευση κακόβουλων ενεργειών.

Πως χρησιμοποιείται η εξόρυξη δεδομένων;

Πριν τη χρήση ενός Μοντέλου Γλωσσικής Επεξεργασίας (ΜΓΕ), χρησιμοποιούμε την τεχνική της εξόρυξης δεδομένων για να παρέχουμε απαντήσεις και να παράγουμε γνώση από τον όγκο των δεδομένων που έχουμε αποθηκευμένα. Αυτή η διαδικασία επιτρέπει έπειτα στο ΜΓΕ να εξάγει σημαντικές πληροφορίες και τάσεις από τα δεδομένα, προσφέροντας ουσιαστικές απαντήσεις στα ερωτήματα των χρηστών.

Επιχειρηματική Ανάλυση:

Πρόβλεψη Συμπεριφοράς Πελατών: Στον χώρο της επιχειρηματικής, το data mining μπορεί να χρησιμοποιηθεί για να προβλέψει τη συμπεριφορά των πελατών, προτείνοντας προϊόντα και υπηρεσίες που θα τους ενδιαφέρουν.

Ανάλυση Κόστους-Οφέλους: Στη λήψη αποφάσεων, το data mining βοηθά στην ανάλυση του κόστους και του οφέλους, επιτρέποντας στις επιχειρήσεις να προβλέπουν τα οικονομικά αποτελέσματα διάφορων σεναρίων.

Υγεία και Ιατρική:

Διάγνωση Ασθενειών: Στον τομέα της υγείας, το data mining μπορεί να χρησιμοποιηθεί για την ανάλυση δεδομένων ασθενών για να κατανοήσει πιθανούς παράγοντες κινδύνου και να βοηθήσει στη διάγνωση ασθενειών.

Βελτιστοποίηση Θεραπείας: Μέσω του data mining, μπορεί να αναγνωριστούν πρότυπα στην απόκριση σε διάφορες θεραπείες, βοηθώντας στην εξατομίκευση της ιατρικής περίθαλψης.

Εκπαίδευση:

Προσαρμοστική Διδασκαλία: Στον τομέα της εκπαίδευσης, το data mining μπορεί να βοηθήσει στην προσαρμογή των μεθόδων διδασκαλίας στις ανάγκες του κάθε μαθητή, βασιζόμενο στα προηγούμενα τους επιτεύγματα.

Κοινωνικά Δίκτυα:

Ανάλυση Συναισθήματος: Στα κοινωνικά δίκτυα, το data mining χρησιμοποιείται για την ανάλυση συναισθημάτων και απόψεων χρηστών, προκειμένου να κατανοήσει τις τάσεις και τις προτιμήσεις τους.

Επιστημονική Έρευνα:

Ανακάλυψη Προτύπων: Στην επιστημονική έρευνα, το data mining εφαρμόζεται για την ανακάλυψη προτύπων σε μεγάλα σύνολα δεδομένων, επιτρέποντας στους ερευνητές να βγάλουν συμπεράσματα και να διατυπώσουν υποθέσεις.

Η χρήση του data mining σε αυτό το πλαίσιο μας επιτρέπει να προσφέρουμε εξατομικευμένες και ενημερωμένες απαντήσεις, καθώς εξάγουμε γνώση από την ανάλυση μεγάλων ποσοτήτων δεδομένων [1], [2].

Εξόρυξη δεδομένων και Κοινωνικά Μέσα

Η εξόρυξη δεδομένων (Data Mining) και τα κοινωνικά δίκτυα αποτελούν σημαντικούς πυλώνες στον κόσμο της πληροφορικής, παρέχοντας πλούσιες πηγές δεδομένων και ευκαιρίες ανάλυσης. Η εξόρυξη δεδομένων επιτρέπει τον εντοπισμό προτύπων, τάσεων και συσχετίσεων σε μεγάλα σύνολα δεδομένων, ενώ τα κοινωνικά δίκτυα δημιουργούν περιβάλλοντα όπου οι άνθρωποι αλληλεπιδρούν, εκφράζουν απόψεις και μοιράζονται πληροφορίες.

Η συνένωση αυτών των δύο πεδίων ανοίγει νέες προοπτικές για την κατανόηση των κοινωνικών δικτύων και των διαδικασιών που συμβαίνουν σε αυτά. Αναπτύσσονται προηγμένες τεχνικές εξόρυξης δεδομένων που εφαρμόζονται στα κοινωνικά μέσα για την ανακάλυψη προτύπων συμπεριφοράς και την εξαγωγή συμπερασμάτων.

Η ανάλυση συναισθημάτων αποτελεί ένα σημαντικό κομμάτι αυτής της συνέργιας. Η εξόρυξη δεδομένων μπορεί να αναλύσει μαζικά όγκους κειμένων από κοινωνικά μέσα για να κατανοήσει το συναισθηματικό φόρτο που συνοδεύει τις δημοσιεύσεις. Αυτό μπορεί να χρησιμοποιηθεί για να μετρήσει τη στάση του κοινού προς συγκεκριμένα θέματα, προϊόντα ή γεγονότα.

Οι τεχνικές πρόβλεψης είναι επίσης ευρέως εφαρμόσιμες σε αυτό το πλαίσιο. Χρησιμοποιώντας δεδομένα από τα κοινωνικά μέσα, μπορεί να γίνει πρόβλεψη συγκεκριμένων γεγονότων ή τάσεων. Αυτό μπορεί να είναι χρήσιμο σε πολλούς τομείς, όπως η διαχείριση κρίσεων, η πρόβλεψη των πωλήσεων, και η αντιληπτική των κοινωνικών εξελίξεων.

Η κατηγοριοποίηση και η συσταδοποίηση αποτελούν δυνατά εργαλεία για την οργάνωση του χάους των δεδομένων από τα κοινωνικά μέσα. Μπορούν να βοηθήσουν στην αυτόματη ταξινόμηση των δημοσιεύσεων ή στην αναγνώριση των θεμάτων που ενώνουν τους χρήστες. Αυτό καταστέλλει την υπερβολική πολυπλοκότητα και επιτρέπει την αποτελεσματική παρουσίαση των πληροφοριών.

Το ζήτημα της ασφάλειας και της προστασίας των δεδομένων είναι, επίσης, ιδιαιτέρως σημαντικό σε αυτό το πλαίσιο. Ενώ η εξόρυξη δεδομένων στα κοινωνικά μέσα παρέχει χρήσιμες πληροφορίες, πρέπει να λαμβάνονται υπόψη η ιδιωτικότητα και η ευαισθησία των δεδομένων.

Συνοψίζοντας, η εξόρυξη δεδομένων στα κοινωνικά δίκτυα αντιπροσωπεύει έναν σημαντικό πυλώνα στην κατανόηση των συμπεριφορών και των διαδικασιών σε αυτά, παρέχοντας τη δυνατότητα για βαθύτερη ανάλυση και καλύτερη κατανόηση της κοινωνίας και των ατόμων που την απαρτίζουν.[4], [5].

Εισαγωγή στην Ανάλυση Συναισθηματικού Τόνου (Sentiment Analysis)

Η δυναμική της ψηφιακής εποχής και η πρωτοφανής εξάπλωση των κοινωνικών μέσων, των ιστοσελίδων, και της ψηφιακής επικοινωνίας έχουν δημιουργήσει έναν ανεξάντλητο πηγαίο όγκο κειμένων που περιέχουν πληροφορίες για την ανθρώπινη συμπεριφορά, τις αντιδράσεις των ανθρώπων σε επιστημονικά, κοινωνικά, πολιτικά και καταναλωτικά θέματα. Η Ανάλυση Συναισθηματικού Τόνου, γνωστή και ως Sentiment Analysis, έχει ως στόχο να εξάγει και να κατανοήσει το συναισθηματικό περιεχόμενο αυτών των κειμένων.

Αναπτύσσεται ως ένα σημαντικό πεδίο εντός της φυσικής γλώσσας και της τεχνητής νοημοσύνης. Ουσιαστικά αποσκοπεί στην αυτοματοποιημένη ανίχνευση και ανάλυση του συναισθηματικού περιεχομένου ενός κειμένου, με σκοπό την κατηγοριοποίηση του ως θετικού, αρνητικού ή ουδέτερου.

Η Sentiment Analysis είναι ένας τομέας της φυσικής γλώσσας και της τεχνητής νοημοσύνης μπορεί να “αναγνωρίσει” συναισθήματα όπως η χαρά, η θλίψη, ο θυμός, η έκπληξη, και η απογοήτευση, τα οποία στη συνέχεια θα αξιολογηθούν και θα ταξινομηθούν ως θετικά, αρνητικά ή ουδέτερα ανάλογα με τον τρόπο που εκφράζονται στο κείμενο.

Ο σκοπός της Sentiment Analysis είναι να παρέχει μια κατανόηση του τρόπου με τον οποίο τα συναισθήματα εκφράζονται και ερμηνεύονται στα κείμενα. Αυτό μπορεί να γίνει μέσω ανθρώπων που αξιολογούν τα κείμενα, αλλά και μέσω αυτοματοποιημένων συστημάτων υπολογιστών που χρησιμοποιούν αλγορίθμους για να αναλύσουν το συναισθηματικό περιεχόμενο.

Η Sentiment Analysis παρέχει πολλές ευκαιρίες και πλεονεκτήματα σε διάφορους τομείς:

α) Εμπόριο και Μάρκετινγκ: Η ανάλυση του συναισθηματικού τόνου των καταναλωτών μπορεί να βοηθήσει τις επιχειρήσεις να κατανοήσουν τις αντιδράσεις των πελατών στα προϊόντα και τις υπηρεσίες τους.

β) Κοινωνικά Δίκτυα: Η Sentiment Analysis χρησιμοποιείται για την παρακολούθηση των συναισθημάτων που εκφράζονται σε κοινωνικά δίκτυα και για την αξιολόγηση των κοινωνικών τάσεων.

γ) Ειδησεογραφία: Η ανάλυση του συναισθηματικού τόνου των ειδήσεων μπορεί να βοηθήσει στον εντοπισμό και τον αντικειμενικό χειρισμό των συναισθηματικά φορτισμένων ειδήσεων.

δ) Πολιτική και κοινωνική ανάλυση: Στον τομέα της πολιτικής και κοινωνικής ανάλυσης, η Sentiment Analysis μπορεί να χρησιμοποιηθεί για την παρακολούθηση των πολιτικών απόψεων και την αξιολόγηση των κοινωνικών αντιδράσεων σε σημαντικά γεγονότα.

Πολύ σημαντικό ρόλο διαδραματίζει στην αξιολόγηση προϊόντων και υπηρεσιών. Οι καταναλωτές χρησιμοποιούν όλο και περισσότερο τις διαδικτυακές αξιολογήσεις για να καθορίσουν την ποιότητα ενός προϊόντος πριν την αγορά του. Με τη Sentiment Analysis, είναι δυνατή η αυτόματη επεξεργασία των αξιολογήσεων αυτών προκειμένου να αποσπαστεί η γενική τάση των συναισθηματικών αντιδράσεων. Η διαδικασία αυτή επικεντρώνεται στη χρήση ποικίλων μεθόδων, συμπεριλαμβανομένης της μηχανικής μάθησης (machine learning) και της βαθιάς μάθησης (deep learning), για την αποτελεσματική αναγνώριση και κατανόηση των συναισθηματικών πτυχών των κειμένων. Αυτή η διαδικασία μπορεί να επιτευχθεί σε διάφορα επίπεδα, από την ανάλυση μεμονωμένων λέξεων και φράσεων μέχρι τον πλήρη συναισθηματικό χαρακτήρα του κειμένου.

Παρά την ανάπτυξή της, υπάρχουν αναπάντητα ερωτήματα σχετικά με την ακρίβεια και την αξιοπιστία των αποτελεσμάτων της Sentiment Analysis, λόγω της πολυπλοκότητας της ανθρώπινης γλώσσας. Οι ερευνητές συνεχίζουν να αναζητούν καινοτόμες προσεγγίσεις και νέες τεχνικές για την βελτίωση της απόδοσης των αλγορίθμων. Μία ακόμη πρόκληση που δημιουργείται είναι οι πολιτικές διαφορές καθώς, οι πολιτικές, πολιτισμικές και κοινωνικές διαφορές μπορούν να επηρεάσουν τον τρόπο με τον οποίο αξιολογούνται τα συναισθήματα σε διάφορες περιοχές του κόσμου. Τέλος, τα συναίσθημα και πολυπλοκότητα τους ανάγονται στις προκλήσεις της Sentiment Analysis. Τα συναισθήματα συχνά εκφράζονται σε πολύπλοκες και αναφορικές δομές, κάτι που απαιτεί προχωρημένες τεχνικές για την ανάλυσή τους.

Συνολικά, η Sentiment Analysis αντιπροσωπεύει μια σημαντική τεχνολογία που διευκολύνει την κατανόηση της συναισθηματικής φόρτισης των κειμένων στον ψηφιακό κόσμο.

Θεωρητικό Υπόβαθρο

Μεθοδολογίες Βασισμένες σε Μηχανική Μάθηση για την Ανάλυση Συναισθηματικού Τόνου

Η επίκαιρη ανάπτυξη των μεθοδολογιών βασισμένων σε μηχανική μάθηση για την Sentiment Analysis έχει καταλάβει το ενδιαφέρον της επιστημονικής κοινότητας. Αυτές οι μεθοδολογίες βασίζονται στη χρήση αλγορίθμων και μοντέλων μηχανικής μάθησης για την ανίχνευση και ανάλυση των συναισθηματικών χαρακτηριστικών του κειμένου.

Στον τομέα της Sentiment Analysis, οι μεθοδολογίες που βασίζονται στη Μηχανική Μάθηση έχουν αποδειχθεί ως αποτελεσματικές για την αυτοματοποιημένη αξιολόγηση του συναισθηματικού τόνου σε κείμενα. Αυτές οι μεθοδολογίες επιτρέπουν στους υπολογιστές να μάθουν από τα δεδομένα και να εκπαιδεύσουν μοντέλα που μπορούν να αναγνωρίσουν και να κατηγοριοποιήσουν τα συναισθήματα σε κείμενα.

Μία από τις βασικές προκλήσεις στην Ανάλυση Συναισθηματικού Τόνου είναι η αποτελεσματική αναγνώριση και κατανόηση των συναισθηματικών εκφράσεων και συμφράσεων σε διάφορα είδη κειμένων. Οι μεθοδολογίες βασισμένες σε μηχανική μάθηση προσπαθούν να αντιμετωπίσουν αυτήν την πρόκληση εκπαιδεύοντας μοντέλα μεγάλης κλίμακας πάνω σε δεδομένα που περιλαμβάνουν προηγούμενες αξιολογήσεις και συναισθηματικές εκφράσεις.

Οι κύριες μεθοδολογίες που χρησιμοποιούνται στη Sentiment Analysis, είναι η “ταξινόμηση κειμένου”, “Προεπεξεργασία Δεδομένων” και “Αξιολόγηση”.

α) Ταξινόμηση Κειμένου

Η ταξινόμηση κειμένου είναι η κύρια μεθοδολογία που χρησιμοποιείται στη Sentiment Analysis. Σε αυτήν την προσέγγιση, κάθε κείμενο ταξινομείται σε μία από τις τρεις κατηγορίες: θετικό, αρνητικό ή ουδέτερο. Αυτό γίνεται μέσω μοντέλων μηχανικής μάθησης, όπως οι Νευρωνικοί Δίκτυα, οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) και οι K-B Κοντινότεροι Γείτονες (K-NN). Αυτά τα μοντέλα εκπαιδεύονται σε σύνολα δεδομένων με ετικέτες που περιλαμβάνουν το συναισθηματικό τόνο των κειμένων και στη συνέχεια χρησιμοποιούνται για την ταξινόμηση νέων κειμένων.

β) Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων είναι κρίσιμη για την αποδοτική Sentiment Analysis. Αυτή περιλαμβάνει τα εξής στάδια:

Καθαρισμός Κειμένου: Αφαιρείται η στίξη, η πεζογραφία, οι ειδικοί χαρακτήρες και άλλοι μη απαραίτητοι χαρακτήρες που δεν προσφέρουν πληροφορίες για το συναισθηματικό περιεχόμενο.

Διαχείριση Στεμμάτων: Τα στέμματα των λέξεων μειώνουν τις λέξεις στη βασική τους μορφή, π.χ. "τρώνε" αντί για "τρώει".

Διαχείριση Στοπ-Λέξεων: Οι στοπ-λέξεις είναι λέξεις που δεν προσφέρουν πολλή πληροφορία για το συναισθηματικό περιεχόμενο και μπορούν να αφαιρεθούν, π.χ. "και," "το," "αλλά."

Διανυσματοποίηση: Οι λέξεις μετατρέπονται σε αριθμητικές αναπαραστάσεις, όπως τα διανύσματα λέξεων, για να μπορούν να χρησιμοποιηθούν από τα μοντέλα μηχανικής μάθησης.

γ) Αξιολόγηση

Η αξιολόγηση των αποτελεσμάτων είναι απαραίτητη για την εκτίμηση της απόδοσης των μοντέλων Sentiment Analysis. Συνήθως, χρησιμοποιούνται μετρικές αξιολόγησης, όπως η ακρίβεια, η ανάκλιση και η F1-μέτρηση. Επιπλέον, μπορεί να χρησιμοποιηθεί δια συνδυασμένη αξιολόγηση, που λαμβάνει υπόψη την προβλεπόμενη κατηγορία και την πραγματική αξιολόγηση του συναισθηματικού τόνου.

Επιπλέον των μοντέλων μηχανικής μάθησης, υπάρχουν και άλλες τεχνικές που μπορούν να χρησιμοποιηθούν στη Sentiment Analysis και περιλαμβάνουν:

α) Μοντέλα Συνεχούς Διανυσματοποίησης: Χρησιμοποιούνται για την αναπαράσταση λέξεων σε πολυδιάστατους χώρους και την ανίχνευση συναισθηματικών προτύπων.

β) Βαθιά Μάθηση: Οι βαθείς νευρωνικοί δίκτυοι (Deep Learning) χρησιμοποιούνται για την ανάλυση συναισθημάτων σε πιο πολύπλοκα κείμενα, όπως αναλύσεις κοινωνικών μέσων.

Η Sentiment Analysis είναι ένας συναρπαστικός τομέας έρευνας και ανάπτυξης, αλλά αντιμετωπίζει προκλήσεις, όπως η πολυπλοκότητα της φυσικής γλώσσας, η πολυπλοκότητα της ανθρώπινης συμπεριφοράς και οι πολιτισμικές διαφορές. Με την συνεχή εξέλιξη των τεχνικών και των μοντέλων, όμως, αναμένεται να βελτιωθεί σημαντικά η ακρίβεια και η αποτελεσματικότητα της Sentiment Analysis στο μέλλον.

Τέλος, μεταξύ των προσεγγίσεων που χρησιμοποιούνται είναι οι αλγόριθμοι μηχανικής μάθησης όπως τα Νευρωνικά Δίκτυα (Neural Networks), τα Διανύσματα Μηχανής (Support Vector Machines), Logistic Regression, Supervised Classification, Sentiment Lexicon Analysis, Pattern Recognition και Semantic Models. Αυτές οι μεθοδολογίες έχουν επιτύχει εντυπωσιακά αποτελέσματα, επιτρέποντας την ακριβή και αυτόματη κατηγοριοποίηση του συναισθηματικού τόνου των κειμένων.

Κατηγοριοποίηση Με Επίβλεψη (Supervised Classification): Στην κατηγοριοποίηση με επίβλεψη, χρησιμοποιούνται ετικετοποιημένα δεδομένα εκπαίδευσης που περιέχουν παραδείγματα με συναισθηματικό τόνο. Τα μοντέλα μηχανικής μάθησης, όπως οι Support Vector Machines, οι Random Forests και οι Νευρωνικοί Δίκτυα, εκπαιδεύονται σε αυτά τα δεδομένα για να μπορούν να κατηγοριοποιούν νέα δεδομένα σε συγκεκριμένες κατηγορίες συναισθηματικού τόνου.

Ανάλυση Συναισθηματικού Λεξιλογίου (Sentiment Lexicon Analysis): Αυτή η μέθοδος χρησιμοποιεί ένα λεξιλόγιο με λέξεις που συσχετίζονται με συγκεκριμένα συναισθήματα. Κάθε

λέξη ανατίθεται μια βαθμολογία συναισθηματικού τόνου, και ο συνολικός τόνος ενός κειμένου υπολογίζεται με βάση τις βαθμολογίες αυτές.

Αναγνώριση Συναισθηματικών Προτύπων (Pattern Recognition): Σε αυτήν τη μέθοδο, τα μοντέλα μηχανικής μάθησης εκπαιδεύονται να αναγνωρίζουν πρότυπα συναισθηματικού τόνου μέσα από τα δεδομένα. Αυτό μπορεί να περιλαμβάνει την ανίχνευση συγκεκριμένων λέξεων, φράσεων, ή ακόμη και προτύπων στη σύνταξη που σχετίζονται με συγκεκριμένα συναισθήματα.

Χρήση Νευρωνικών Δικτύων (Neural Networks): Τα βαθιά νευρωνικά δίκτυα, όπως τα αναδραστικά νευρωνικά δίκτυα (Recurrent Neural Networks - RNNs) και τα αναφορικά νευρωνικά δίκτυα (Long Short-Term Memory - LSTM), έχουν χρησιμοποιηθεί με επιτυχία στην ανάλυση συναισθηματικού τόνου. Τα νευρωνικά δίκτυα μπορούν να αποκτήσουν κατανόηση του συμφραζόμενου και να εξάγουν πιο πολύπλοκα χαρακτηριστικά από τα δεδομένα.

Ενσωμάτωση Σημασιολογικών Μοντέλων (Semantic Models): Αυτή η μέθοδος χρησιμοποιεί σημασιολογικά μοντέλα που κατανοούν τη σημασία των λέξεων και των προτάσεων σε ένα κείμενο. Η ενσωμάτωση σημασιολογικών μοντέλων μπορεί να βελτιώσει την ακρίβεια της ανίχνευσης συναισθηματικού τόνου.

Καθεμία από αυτές τις μεθοδολογίες έχει τα πλεονεκτήματά της και εφαρμόζεται ανάλογα με τα δεδομένα και τις απαιτήσεις της συγκεκριμένης εφαρμογής.

Κατηγοριοποίηση με Επίβλεψη (Supervised Classification)

Βασικές Αρχές:

Συλλογή Δεδομένων: Η διαδικασία ξεκινά με τη συλλογή ενός σημαντικού όγκου δεδομένων. Αυτά τα δεδομένα πρέπει να είναι ετικετοποιημένα με τον συναισθηματικό τους τόνο, όπως θετικό, αρνητικό, ή ουδέτερο.

Διαχωρισμός Δεδομένων: Τα δεδομένα διαιρούνται σε σύνολα εκπαίδευσης και δοκιμής. Το σύνολο εκπαίδευσης χρησιμοποιείται για την εκπαίδευση του μοντέλου, ενώ το σύνολο δοκιμής χρησιμοποιείται για τον έλεγχο της απόδοσης του μοντέλου σε νέα δεδομένα.

Αλγόριθμοι Μηχανικής Μάθησης:

Υποψήφιοι Αλγόριθμοι: Χρησιμοποιούνται διάφοροι αλγόριθμοι μηχανικής μάθησης, όπως οι Support Vector Machines (SVM), οι K-Κοντινότεροι Γείτονες (K-Nearest Neighbors), ή τα Νευρωνικά Δίκτυα.

Επιλογή Χαρακτηριστικών: Προτού την εκπαίδευση, μπορεί να πραγματοποιηθεί ανάλυση χαρακτηριστικών για την επιλογή των σημαντικότερων χαρακτηριστικών του κειμένου.

Διαδικασία Εκπαίδευσης:

Αναπαράσταση Κειμένου: Το κείμενο πρέπει να αναπαρασταθεί κατάλληλα για το μοντέλο. Αυτό μπορεί να γίνει με χρήση μεθόδων όπως οι Bag-of-Words (BoW), Word Embeddings, ή τα Transformer-based μοντέλα.

Εκπαίδευση Μοντέλου: Το μοντέλο εκπαιδεύεται χρησιμοποιώντας το σύνολο εκπαίδευσης. Κατά τη διάρκεια αυτής της διαδικασίας, το μοντέλο προσαρμόζει τα βάρη του για να αναγνωρίζει συναισθηματικά πρότυπα στα δεδομένα.

Αξιολόγηση και Πρόβλεψη:

Αξιολόγηση Επίδοσης: Το μοντέλο αξιολογείται στο σύνολο δοκιμής χρησιμοποιώντας μετρικές όπως η ακρίβεια, η ανάκληση και η F1-score, για να μετρήσει πόσο καλά προβλέπει τον συναισθηματικό τόνο.

Πρόβλεψη Νέων Δεδομένων: Αφού το μοντέλο εκπαιδευτεί, μπορεί να χρησιμοποιηθεί για την πρόβλεψη του συναισθηματικού τόνου σε νέα κείμενα που δεν είχαν χρησιμοποιηθεί κατά την εκπαίδευση.

Προκλήσεις:

Ανεπαρκής Δεδομένα: Η επίτευξη υψηλής ακρίβειας απαιτεί συχνά μεγάλα και ισορροπημένα σε συναισθηματικούς τόνους σύνολα δεδομένων.

Πολυπλοκότητα Φυσικής Γλώσσας: Η αντιμετώπιση της πολυπλοκότητας της φυσικής γλώσσας, όπως η σημασιολογία, η συντακτική δομή και η ποικιλία του λεξιλογίου, απαιτεί προσεκτική επεξεργασία και επιλογή χαρακτηριστικών.

Υπερεκπαίδευση: Μεγάλα μοντέλα, όπως τα νευρωνικά δίκτυα, μπορεί να υπερεκπαιδεύονται εάν δεν διαχειριστούν κατάλληλα, ενδεχομένως οδηγώντας σε κακή απόδοση σε νέα δεδομένα.

Στην πράξη, η Κατηγοριοποίηση με Επίβλεψη χρησιμοποιείται σε πολλές εφαρμογές, όπως:

Αναγνώριση Εικόνων:
Στον τομέα της όρασης υπολογιστών, η κατηγοριοποίηση χρησιμοποιείται για τον αυτόματο εντοπισμό και ταξινόμηση αντικειμένων σε εικόνες.

Φυσική Γλώσσα και Επεξεργασία Κειμένου:
Στον τομέα της επεξεργασίας φυσικής γλώσσας, η κατηγοριοποίηση είναι κρίσιμη για την αυτόματη αναγνώριση θεμάτων, την ανάλυση συναισθημάτων, και την ταξινόμηση κειμένων.

Ιατρική Διάγνωση:
Στην ιατρική, η κατηγοριοποίηση χρησιμοποιείται για τον αυτόματο εντοπισμό και διάγνωση ασθενειών μέσω της ανάλυσης ιατρικών δεδομένων.

Χρηματοοικονομική Πρόβλεψη:
Στη χρηματοοικονομία, η κατηγοριοποίηση χρησιμοποιείται για τον προσδιορισμό της απόδοσης επενδύσεων και τον προσδιορισμό χρηματοοικονομικών τάσεων.

Κατηγοριοποίηση Κλιμάκων:
Στον τομέα των ενεργειακών συστημάτων, η κατηγοριοποίηση χρησιμοποιείται για τον αυτόματο έλεγχο και την προσαρμογή των ενεργειακών κλιμάκων.

Η Κατηγοριοποίηση Με Επίβλεψη είναι ισχυρή μέθοδος για την ανάλυση συναισθηματικού τόνου, αλλά η απόδοσή της εξαρτάται σημαντικά από την ποιότητα των δεδομένων και την κατάλληλη επιλογή και παραμετροποίηση του μοντέλου. [23]

Ανάλυση Συναισθηματικού Λεξιλογίου (Sentiment Lexicon Analysis)

Η Ανάλυση Συναισθηματικού Λεξιλογίου είναι μια μέθοδος για την αξιολόγηση του συναισθηματικού περιεχομένου ενός κειμένου με βάση τις λέξεις που περιέχει. Αυτή η προσέγγιση βασίζεται στον υπολογισμό της συνολικής συναισθηματικής φόρτισης ενός κειμένου, χρησιμοποιώντας ένα λεξιλόγιο στο οποίο κάθε λέξη έχει αντιστοιχηθεί με ένα συναισθηματικό βάρος.

Βασικά Χαρακτηριστικά:

Συλλογή Συναισθηματικού Λεξιλογίου: Η πρώτη διαδικασία είναι η δημιουργία ενός συναισθηματικού λεξιλογίου. Σε αυτό το λεξιλόγιο, κάθε λέξη έχει αντιστοιχηθεί με ένα συναισθηματικό βάρος, που μπορεί να είναι θετικό, αρνητικό ή ουδέτερο. Υπάρχουν πολλά διαθέσιμα συναισθηματικά λεξιλόγια, και μερικά από αυτά είναι το SentiWordNet, το AFINN, και το NRC Emotion Lexicon.

Υπολογισμός Συνολικής Συναισθηματικής Φόρτισης: Κατά την ανάλυση ενός κειμένου, υπολογίζεται η συνολική συναισθηματική φόρτιση χρησιμοποιώντας τα συναισθηματικά βάρη των λέξεων που περιλαμβάνει. Συνήθως, οι αθροίσεις ή οι μέσοι όροι των συναισθηματικών βαρών καθορίζουν εάν το κείμενο είναι θετικό, αρνητικό ή ουδέτερο.

Πλεονεκτήματα:

Απλότητα και Ευελιξία: Η Ανάλυση Συναισθηματικού Λεξιλογίου είναι απλή στην υλοποίηση και ευέλικτη, καθώς μπορεί να προσαρμοστεί σε διάφορα πεδία και γλώσσες.

Κατανοητότητα: Είναι εύκολο να κατανοήσει κανείς τον τρόπο λειτουργίας της, καθώς βασίζεται στη συναισθηματική φόρτιση των λέξεων.

Περιορισμοί:

Έλλειψη Συμφραζομένου: Δεν λαμβάνει υπόψη τη σημασιολογική σύνδεση μεταξύ των λέξεων ή το συμφραζόμενο, πράγμα που μπορεί να οδηγήσει σε περιορισμένη ακρίβεια.

Ευαισθησία στην Πολυσημία: Κάποιες λέξεις μπορεί να έχουν διάφορες σημασίες ανάλογα με το περιεχόμενο, και η μέθοδος δεν τις διακρίνει αποτελεσματικά.

Ανεπάρκεια για Πολυπλοκά Κείμενα: Σε πολύπλοκα κείμενα, η μέθοδος μπορεί να αδυνατεί να αντιμετωπίσει τις πολλαπλές διαστάσεις του συναισθηματικού περιεχομένου.

Η Ανάλυση Συναισθηματικού Λεξιλογίου είναι μια αποτελεσματική, αν και απλή, προσέγγιση για την ανάλυση συναισθηματικού τόνου, αλλά πρέπει να χρησιμοποιείται με επιφύλαξη σε συγκεκριμένα περιβάλλοντα και χρήσεις. [23], [24]

Αναγνώριση Συναισθηματικών Προτύπων (Pattern Recognition)

Η Αναγνώριση Συναισθηματικών Προτύπων είναι μια κατηγορία της μηχανικής μάθησης που επικεντρώνεται στο να εκπαιδεύει μοντέλα να αναγνωρίζουν και να εκτιμούν τα συναισθηματικά πρότυπα και τις εκφράσεις σε δεδομένα, όπως κείμενα, ήχοι, εικόνες και βίντεο. Αυτή η διαδικασία έχει εφαρμογές σε πολλούς τομείς, συμπεριλαμβανομένων των κοινωνικών μέσων, της αναγνώρισης ομιλίας, της ρομποτικής και άλλων.

Κάποια βασικά στοιχεία της Αναγνώρισης Συναισθηματικών Προτύπων περιλαμβάνουν:

Είδη Συναισθημάτων: Τα μοντέλα αναγνώρισης προσπαθούν να διακρίνουν διάφορα είδη συναισθημάτων, όπως χαρά, θλίψη, οργή, φόβο, κ.ά. Αυτό μπορεί να γίνει μέσω της ανάλυσης των λειτουργιών, της συντακτικής δομής, του λεξιλογίου, και άλλων στοιχείων στα δεδομένα.

Επεξεργασία Συναισθηματικού Περιεχομένου: Η αναγνώριση συναισθηματικών προτύπων εφαρμόζεται σε πολλά μέσα, όπως κείμενα, ήχοι, εικόνες και βίντεο. Για παράδειγμα, μπορεί να αξιολογεί το συναισθηματικό περιεχόμενο ενός κειμένου ή το συναίσθημα που εκφράζεται σε μια φωνητική εγγραφή.

Εκπαίδευση με Επίβλεψη: Οι αλγόριθμοι αναγνώρισης προτύπων συνήθως χρειάζονται ετικετοδοτημένα δεδομένα για εκπαίδευση, δηλαδή δεδομένα όπου το συναίσθημα έχει ήδη ετικετοδοτηθεί. Οι αλγόριθμοι μάθησης με επίβλεψη, όπως οι μέθοδοι μηχανικής μάθησης, εκπαιδεύονται σε αυτά τα δεδομένα για να αναγνωρίσουν παρόμοια συναισθηματικά πρότυπα σε νέα δεδομένα.

Χρήση Σημαντικών Χαρακτηριστικών: Οι μέθοδοι αναγνώρισης προτύπων χρησιμοποιούν σημαντικά χαρακτηριστικά για να εκπαιδεύσουν τα μοντέλα τους. Για παράδειγμα, στο κείμενο, μπορεί να εξετάζει τις συναισθηματικές λέξεις, τη συντακτική δομή, και τον τόνο για να αναγνωρίσει τα συναισθηματικά πρότυπα.

Η Αναγνώριση Συναισθηματικών Προτύπων έχει εφαρμογές σε πολλούς τομείς, συμβάλλοντας στην κατανόηση των συναισθημάτων και των αντιδράσεων του ανθρώπινου ψυχισμού και βοηθώντας στη βελτίωση των συστημάτων αλληλεπίδρασης με τους ανθρώπους. [25], [26]

Νευρωνικά Δίκτυα

Τα Νευρωνικά Δίκτυα είναι μια κατηγορία αλγορίθμων μηχανικής μάθησης που εμπνέονται από τη λειτουργία του ανθρώπινου εγκεφάλου. Τα δίκτυα αυτά χρησιμοποιούνται για κατηγοριοποίηση, πρόβλεψη, εξαγωγή χαρακτηριστικών και πολλά άλλα.

Ένα νευρωνικό δίκτυο αποτελείται από εκατοντάδες ή ακόμα και εκατοντάδες χιλιάδες μικρούς "νευρώνες" που λειτουργούν παρόμοια με τους νευρώνες του ανθρώπινου εγκεφάλου. Οι νευρώνες αυτοί συνεργάζονται για την επεξεργασία πληροφορίας. Κάθε νευρώνας λαμβάνει εισόδους, επεξεργάζεται την πληροφορία, και εκδίδει μια έξοδο. Η συλλογή των νευρώνων και ο τρόπος με τον οποίο συνδέονται αποτελεί την αρχιτεκτονική του δικτύου.

Οι νευρωνικοί νευρώνες χρησιμοποιούν συνήθως συναρτήσεις ενεργοποίησης για να ρυθμίσουν την έξοδο τους, όπως η σιγμοειδής συνάρτηση. Οι συνδέσεις μεταξύ των νευρώνων

έχουν βάρη που προσαρμόζονται κατά τη διάρκεια της εκπαίδευσης με σκοπό να επιτευχθεί η βέλτιστη απόδοση του δικτύου.

Ένα σημαντικό χαρακτηριστικό των νευρωνικών δικτύων είναι η ικανότητά τους να αναγνωρίζουν πολύπλοκα μη γραμμικά μοτίβα στα δεδομένα. Αυτό επιτυγχάνεται με την χρήση βαθιών νευρωνικών αρχιτεκτονικών, όπως τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs) και τα Αναδραστικά Νευρωνικά Δίκτυα (RNNs).

Τα νευρωνικά δίκτυα έχουν επιτύχει εξαιρετική απόδοση σε πολλά προβλήματα, όπως η αναγνώριση εικόνων, η φωνητική αναγνώριση, η μετάφραση γλωσσών, και η ανάλυση συναισθημάτων. Παρόλα αυτά, το εκπαιδευτικό σύνολο δεδομένων και οι παράμετροι πρέπει να ρυθμιστούν προσεκτικά για την επίτευξη των βέλτιστων αποτελεσμάτων [21].

Βασική Δομή των Νευρωνικών Δικτύων:

Νευρώνες (Neurons): Οι βασικές μονάδες των νευρωνικών δικτύων, που λειτουργούν ως "τεχνητά νευρώνες." Κάθε νευρώνας λαμβάνει είσοδο από διάφορες πηγές, εφαρμόζει μια συνάρτηση ενεργοποίησης, και παράγει έξοδο.

Στρώματα (Layers): Οι νευρώνες οργανώνονται σε στρώματα. Ένα νευρωνικό δίκτυο μπορεί να έχει πολλά στρώματα, συμπεριλαμβανομένων του εισόδου (input layer), των κρυφών (hidden layers), και της εξόδου (output layer).

Συνδέσεις (Connections): Οι συνδέσεις μεταξύ νευρώνων καθορίζουν τον τρόπο με τον οποίο περνάει η πληροφορία μεταξύ τους. Κάθε σύνδεση έχει ένα βάρος που προσδιορίζει το πόσο επηρεάζει μια είσοδος τον συγκεκριμένο νευρώνα.

Εκπαίδευση των Νευρωνικών Δικτύων:

Υποβοηθούμενη Εκπαίδευση (Supervised Learning): Σε αυτήν τη μέθοδο, το δίκτυο εκπαιδεύεται χρησιμοποιώντας ετικετοδοτημένα δεδομένα, όπου γνωρίζουμε την επιθυμητή έξοδο για κάθε είσοδο.

Μη Υποβοηθούμενη Εκπαίδευση (Unsupervised Learning): Σε αυτήν τη μέθοδο, το δίκτυο εκπαιδεύεται χωρίς ετικετοδοτημένα δεδομένα. Το δίκτυο πρέπει να ανακαλύψει πρότυπα και δομές από τα δεδομένα μόνο του.

Συναρτήσεις Ενεργοποίησης (Activation Functions):

Οι συναρτήσεις που χρησιμοποιούνται για να ρυθμίσουν την έξοδο ενός νευρώνα. Συμβολίζονται συνήθως ως συναρτήσεις ενεργοποίησης και μπορεί να είναι σιγμοειδείς, υπερβολικοί τόνοι, ή άλλες.

Συνάρτηση Κόστους (Cost Function):

Χρησιμοποιείται για να μετρήσει πόσο καλά το δίκτυο αποδίδει σε σύγκριση με τα πραγματικά αποτελέσματα. Σκοπός είναι να ελαχιστοποιήσει τη διαφορά μεταξύ των προβλέψεων και των πραγματικών ετικετών.

Εφαρμογές Νευρωνικών Δικτύων:

Οι νευρωνικοί αλγόριθμοι χρησιμοποιούνται σε πολλές εφαρμογές, συμπεριλαμβανομένων της εικονικής αναγνώρισης, της φωνητικής αναγνώρισης, της αναγνώρισης συναισθημάτων, και της αυτόματης οδήγησης.

Η χρήση νευρωνικών δικτύων έχει επιτρέψει την αντιμετώπιση προβλημάτων που παρουσιάζουν πολύπλοκες δομές και αλληλεπιδράσεις στα δεδομένα, κάνοντας τα ιδανικά για εφαρμογές όπως η αναγνώριση προτύπων και η πρόβλεψη. [27],[28]

Ενσωμάτωση Σημασιολογικών Μοντέλων (Semantic Models)

Η Ενσωμάτωση Σημασιολογικών Μοντέλων (Semantic Models) αναφέρεται σε μια κατηγορία τεχνικών που χρησιμοποιούνται για να αντιληφθούν και να αναπαραστήσουν τη σημασία των λέξεων και των φράσεων σε ένα κείμενο. Ο σκοπός της είναι να δώσει στις μηχανές τη δυνατότητα να κατανοούν το περιεχόμενο με τρόπο παρόμοιο με αυτόν που το κατανοούν οι άνθρωποι.

Η Ενσωμάτωση Σημασιολογικών Μοντέλων (Semantic Models) αποτελεί, επίσης, έναν σημαντικό τομέα έρευνας στον χώρο της επεξεργασίας φυσικής γλώσσας και της ανάλυσης κειμένου. Σε αυτόν τον τομέα, τα σημασιολογικά μοντέλα αναζητούν τρόπους να εκφράσουν τη σημασία των λέξεων και των προτάσεων με τρόπο που να είναι κατανοητός από τις υπολογιστικές συστοιχίες.

Ένα από τα πιο διαδεδομένα σημασιολογικά μοντέλα είναι τα Word Embeddings. Αυτά τα μοντέλα αντιστοιχίζουν λέξεις σε πολυδιάστατους χώρους, όπου η απόσταση μεταξύ των λέξεων αντικατοπτρίζει τη σημασιολογική τους ομοιότητα. Τα Word Embeddings εκπαιδεύονται συνήθως με μεγάλα σώματα κειμένων και μπορούν να αναπαριστούν πλούσια σημασιολογική πληροφορία, καθιστώντας τα ιδανικά για πολλές εφαρμογές, όπως η ανάλυση συναισθημάτων, η μηχανική μετάφραση και η κατηγοριοποίηση κειμένου.

Πέρα από τα Word Embeddings, υπάρχουν και πιο προηγμένα σημασιολογικά μοντέλα, όπως τα BERT (Bidirectional Encoder Representations from Transformers) και τα GPT (Generative Pre-trained Transformers). Τα BERT μπορούν να αντιληφθούν το πλαίσιο και τον συμφραζόμενο της κάθε λέξης σε μια πρόταση, ενώ τα GPT μπορούν να δημιουργήσουν φυσικό κείμενο με βάση ένα δεδομένο κείμενο εκπαίδευσης.

Τα σημασιολογικά μοντέλα έχουν επίσης εφαρμογές στην ανάλυση συναισθημάτων, καθώς μπορούν να αντιληφθούν τον συναισθηματικό φόρτο των λέξεων και των προτάσεων. Αυτή η ικανότητα είναι χρήσιμη σε πολλές εφαρμογές, όπως η παρακολούθηση κοινωνικών μέσων και η ανάλυση συναισθηματικών αντιδράσεων σε προϊόντα και υπηρεσίες.

Αναπαράσταση Λέξεων:

Word Embeddings: Οι αναπαραστάσεις λέξεων αναπαριστούν τις λέξεις ως διανύσματα σε χώρους χαρακτηριστικών, όπου η σημασία μιας λέξης αντικατοπτρίζεται από τη θέση της στο χώρο.

Word2Vec, GloVe, FastText: Αλγόριθμοι και μοντέλα που δημιουργούν αυτές τις αναπαραστάσεις, λαμβάνοντας υπόψη το περιβάλλον των λέξεων σε ένα κείμενο.

Αναπαράσταση Φράσεων και Προτάσεων:

Sentence Embeddings: Παρόμοιες τεχνικές εφαρμόζονται για να αναπαριστάνουν φράσεις και προτάσεις σε ένα χώρο χαρακτηριστικών. Η σημασία του περιεχομένου ενσωματώνεται στο διάνυσμα αναπαράστασης.

InferSent, Universal Sentence Encoder: Παραδείγματα μοντέλων που παρέχουν αναπαραστάσεις φράσεων και προτάσεων με βάση τη σημασία τους.

Σημασιολογική Συνένωση (Semantic Fusion):

Η ικανότητα σύνθεσης αναπαραστάσεων για πιο πολύπλοκες μονάδες, όπως προτάσεις, έγγραφα ή κείμενα.

Skip-Thought Vectors, BERT (Bidirectional Encoder Representations from Transformers): Μοντέλα που ενσωματώνουν την σημασία του περιεχομένου και τη συνάρτηση που έχει μεταξύ των λέξεων σε πληθώρα επιπέδων.

Εφαρμογές:

Η Ενσωμάτωση Σημασιολογικών Μοντέλων χρησιμοποιείται σε εφαρμογές όπως η κατηγοριοποίηση κειμένου, η αναζήτηση πληροφοριών, η σύνοψη κειμένου και η αναγνώριση οντοτήτων.

Επίσης, χρησιμοποιείται σε προηγμένες εφαρμογές όπως η διακριτική σημασιολογική ανάλυση (fine-grained semantic analysis) και η συναρτησιακή ανάλυση (functional analysis).

Η Ενσωμάτωση Σημασιολογικών Μοντέλων αποτελεί κρίσιμη πτυχή της εξέλιξης της φυσικής γλώσσας στη μηχανική μάθηση, προσφέροντας βελτιωμένες δυνατότητες κατανόησης του περιεχομένου κειμένων από τους υπολογιστές.

Συνολικά, η Ενσωμάτωση Σημασιολογικών Μοντέλων αντιπροσωπεύει ένα σημαντικό βήμα προς την κατανόηση και την αναπαράσταση της σημασίας του φυσικού γλωσσικού περιεχομένου, επιτρέποντας στις υπολογιστικές συστοιχίες να αντιλαμβάνονται και να αλληλεπιδρούν με τη γλώσσα με πιο φυσικό και προηγμένο τρόπο. [27|,[29]

Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines)

Οι Μηχανές Διανυσματικής Υποστήριξης (Support Vector Machines - SVMs) ανήκουν στην κατηγορία των αλγορίθμων μηχανικής μάθησης και χρησιμοποιούνται για την κατηγοριοποίηση και την πρόβλεψη τιμών. Είναι ιδιαίτερα αποτελεσματικοί σε περιπτώσεις όπου τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα στον χώρο των χαρακτηριστικών. Οι SVM λειτουργούν επιλέγοντας το βέλτιστο υπερεπίπεδο που διαχωρίζει τα δεδομένα διαφορετικών κλάσεων, ενώ ταυτόχρονα μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων.

Οι Μηχανές Διανυσματικής Υποστήριξης (SVM) αποτελούν ένα ισχυρό εργαλείο στον χώρο της μηχανικής μάθησης και ειδικότερα της κατηγοριοποίησης. Η βασική τους αρχή είναι η εύρεση ενός υπερεπιπέδου που διαχωρίζει δύο κλάσεις στον χώρο των χαρακτηριστικών. Αυτό το υπερεπίπεδο επιλέγεται έτσι ώστε οι αποστάσεις από τα πλησιέστερα σημεία δεδομένων των δύο κλάσεων, γνωστά και ως σημεία διανυσματικής υποστήριξης, να είναι μέγιστες.

Μια ενδιαφέρουσα ιδιότητα των SVM είναι η δυνατότητα χρήσης συναρτήσεων πυρήνα, οι οποίες επιτρέπουν την αντιμετώπιση μη γραμμικών προβλημάτων. Με τη διάχυση δεδομένων στο χώρο, οι SVM μετατρέπουν τα χαρακτηριστικά σε έναν χώρο υψηλής διαστατικότητας, επιτρέποντας τη γραμμική διαχωριστική επιφάνεια.

Σημαντικό στοιχείο των SVM είναι τα σημεία διανυσματικής υποστήριξης, τα οποία είναι τα σημεία που βρίσκονται πιο κοντά στο υπερεπίπεδο και επηρεάζουν τη θέση και την κλίση του.

Η συνάρτηση υποστήριξης μετρά την απόσταση από κάθε σημείο προς το υπερεπίπεδο. Αν η απόσταση υπερβαίνει ένα καθορισμένο κατώφλι, τότε το σημείο ταξινομείται σε μία από τις δύο κλάσεις.

Επιπλέον, οι παράμετροι των SVM, όπως το κόστος C και ο παράμετρος του πυρήνα, πρέπει να επιλεγούν προσεκτικά, καθώς επηρεάζουν την απόδοση και την ικανότητα γενίκευσης του μοντέλου.

Οι SVM παρέχουν ένα αξιόπιστο μέσο κατηγοριοποίησης και έχουν εφαρμογές σε πολλούς τομείς της μηχανικής μάθησης, από την αναγνώριση προτύπων έως την ανάλυση εικόνας και τον εντοπισμό ανωμαλιών.

Παρακάτω αναλύω το πως λειτουργούν τα SVMs:

Υπερεπίπεδο: Σε ένα δισδιάστατο χώρο, το υπερεπίπεδο είναι μια γραμμή που διαχωρίζει τα δεδομένα διαφορετικών κλάσεων. Σε χώρους με περισσότερες διαστάσεις, είναι ένα υπερεπίπεδο, το οποίο μπορεί να εννοηθεί ως μια πολυδιάστατη επιφάνεια.

Περιθώριο: Το περιθώριο είναι η απόσταση μεταξύ του υπερεπιπέδου και των πλησιέστερων δεδομένων κάθε κλάσης. Οι SVM στοχεύουν στο να μεγιστοποιήσουν αυτό το περιθώριο, καθώς ένα μεγαλύτερο περιθώριο συνήθως οδηγεί σε καλύτερη γενίκευση σε μη ορατά δεδομένα.

Διανύσματα Υποστήριξης: Αυτά είναι τα δεδομένα που βρίσκονται κοντά στο όριο απόφασης (υπερεπίπεδο). Αποτελούν τα κρίσιμα στοιχεία στους SVM, καθώς καθορίζουν τη θέση και την κατεύθυνση του υπερεπιπέδου.

Πυρήνας: Οι SVM μπορούν να χειριστούν μη γραμμικές διαχωρίσεις μέσω της μετασχηματισμένης αναπαράστασης των δεδομένων σε έναν χώρο υψηλότερων διαστάσεων, με τη χρήση μιας συνάρτησης πυρήνα. Αυτή η μετασχηματισμένη αναπαράσταση μπορεί να κάνει δυνατή τη γραμμική διαχωρισιμότητα των δεδομένων, ακόμα κι αν δεν ήταν στον αρχικό χώρο χαρακτηριστικών.

Παράμετρος C: Οι SVM έχουν μια παράμετρο που ονομάζεται "C," και ελέγχει την αντίθεση μεταξύ του να μεγιστοποιεί το περιθώριο και την ελαχιστοποίηση του σφάλματος κατηγοριοποίησης στα δεδομένα εκπαίδευσης. Μια μικρότερη τιμή C οδηγεί σε μεγαλύτερο περιθώριο, αλλά μπορεί να ανεχθεί περισσότερα λανθασμένα κατηγοριοποιημένα σημεία, ενώ μια μεγαλύτερη τιμή C μπορεί να οδηγήσει σε μικρότερο περιθώριο, αλλά λιγότερα λανθασμένα κατηγοριοποιημένα σημεία.

Οι SVM έχουν χρησιμοποιηθεί ευρέως σε διάφορους τομείς όπως η κατηγοριοποίηση κειμένων, η αναγνώριση εικόνων, η βιοπληροφορική και άλλοι. Είναι ισχυρά εργαλεία τόσο για δυαδικές όσο και για πολυκλασικές κατηγοριοποιήσεις, και η ικανότητά τους να χειριστούν πολύπλοκα όρια απόφασης μέσω του πυρήνα τους τους καθιστά ευέλικτους και αποτελεσματικούς σε πολλά σενάρια [19,20].

Logistic Regression

Η λογιστική παλινδρόμηση (Logistic Regression) είναι μια στατιστική μέθοδος που χρησιμοποιείται για την πρόβλεψη της πιθανότητας ενός γεγονότος, έχοντας ως αποτέλεσμα μια δυαδική έξοδο (0 ή 1). Παρόλο που το όνομά της περιλαμβάνει τη λέξη "παλινδρόμηση," η λογιστική παλινδρόμηση χρησιμοποιείται κυρίως για προβλήματα ταξινόμησης, όχι για προβλήματα παλινδρόμησης.

Ο βασικός σκοπός της λογιστικής παλινδρόμησης είναι να εκτιμήσει την πιθανότητα ενός περιστατικού να συμβεί βάσει ενός ή περισσοτέρων εξηγητικών μεταβλητών. Το μοντέλο της λογιστικής παλινδρόμησης χρησιμοποιεί τη σιγμοειδή συνάρτηση (σιγμοειδή καμπύλη) για να προβλέψει την πιθανότητα ανήκει κάποιο παρατηρούμενο γεγονός στην κατηγορία 1.

Βασικά στοιχεία και τα βήματα της λογιστικής παλινδρόμησης:

Σιγμοειδή Συνάρτηση (Sigmoid Function): Η σιγμοειδής συνάρτηση είναι κεντρικό στοιχείο της λογιστικής παλινδρόμησης. Αυτή η συνάρτηση μετατρέπει κάθε αριθμητική είσοδο σε ένα εύρος μεταξύ 0 και 1, αντιπροσωπεύοντας την πιθανότητα.

Συνάρτηση Κόστους (Cost Function): Η λογιστική παλινδρόμηση χρησιμοποιεί μια συνάρτηση κόστους για την εκτίμηση του πόσο καλά το μοντέλο προβλέπει τις πιθανότητες. Η συνάρτηση κόστους στοχεύει στο να ελαχιστοποιήσει το σφάλμα μεταξύ των προβλέψεων του μοντέλου και των πραγματικών τιμών.

Εκπαίδευση (Training): Κατά τη διάρκεια της φάσης εκπαίδευσης, το μοντέλο προσαρμόζει τα βάρη του με στόχο την ελαχιστοποίηση της συνάρτησης κόστους. Αυτό επιτυγχάνεται μέσω μεθόδων όπως ο αλγόριθμος κατάβασης κλίσης.

Απόφαση (Decision): Τελικά, το μοντέλο λογιστικής παλινδρόμησης λαμβάνει αποφάσεις βάσει των προβλέψεών του. Συνήθως, εάν η προβλεπόμενη πιθανότητα είναι μεγαλύτερη από ένα κατώφλι (συνήθως 0.5), το μοντέλο κατατάσσει την παρατήρηση στην κατηγορία 1, αλλιώς στην κατηγορία 0.

Η λογιστική παλινδρόμηση είναι δημοφιλής λόγω της απλότητάς της και της ικανότητάς της να παράγει αξιόπιστες προβλέψεις σε προβλήματα δυαδικής ταξινόμησης. Επιπλέον, είναι ευέλικτη και μπορεί να χρησιμοποιηθεί σε ποικίλες εφαρμογές, όπως η κατηγοριοποίηση email ως spam ή μη spam, η διάκριση μεταξύ διαφημίσεων και μη διαφημίσεων, και πολλά άλλα.[17,18].

Εμπειρική ανάλυση

Περιγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν σε αυτήν τη μελέτη προέρχονται από το IMDB Dataset, ένα αναγνωρισμένο και δημοφιλές σύνολο δεδομένων που έχει καταστεί βασικό εργαλείο στον χώρο της επεξεργασίας φυσικής γλώσσας και της ανάλυσης κειμένου. Το IMDB Dataset

περιλαμβάνει συνολικά 50.000 κριτικές ταινιών, που διαμοιράζονται σε δύο σύνολα: ένα σύνολο εκπαίδευσης (training) με 25.000 κριτικές και ένα σύνολο δοκιμής (testing) με τις υπόλοιπες 25.000 κριτικές.

Κάθε κριτική αποτελείται από δύο βασικά μέρη:

Κείμενο Κριτικής: Αυτό το μέρος περιλαμβάνει το περιεχόμενο της κριτικής της ταινίας, προσφέροντας τις απόψεις και τα σχόλια των χρηστών σχετικά με την ταινία που έχουν παρακολουθήσει.

Ετικέτα Συναισθημάτων: Κάθε κριτική συνοδεύεται από μια ετικέτα που υποδεικνύει το συναισθηματικό χαρακτήρα της κριτικής. Αυτές οι ετικέτες καθορίζουν αν η κριτική χαρακτηρίζεται ως "θετική" ή "αρνητική," ανάλογα με την συναισθηματική αξιολόγηση που έχει προσδοθεί από τον συγγραφέα της κριτικής.

Αυτά τα δεδομένα αποτελούν τη βάση εκπαίδευσης για το μοντέλο μηχανικής μάθησης και χρησιμοποιούνται για την εκμάθηση των προτύπων που αφορούν την αναγνώριση συναισθημάτων σε κείμενα κριτικών ταινιών. Τα συγκεκριμένα δεδομένα αποτελούν αξιόπιστη πηγή πληροφοριών και χρησιμοποιούνται ευρέως στην επιστημονική κοινότητα για την ανάπτυξη και τον έλεγχο μοντέλων ανάλυσης συναισθημάτων σε κείμενα.

Review	**Sentiment**
One of the other reviewers has mentioned that after watching just 1 Oz episode you'll be hooked...	Positive
Basically, there's a family where a little boy (Jake) thinks there's a zombie in his closet & his par...	Negative
An awful film! It must have been up against some real stinkers to be nominated for the Golden Globe....	Negative

Προεπεξεργασία Δεδομένων

Για την προεπεξεργασία των δεδομένων έγιναν τα εξής βήματα:

Μέθοδος Προεπεξεργασίας	**Πριν την προεπεξεργασία**	**Μετά την προεπεξεργασία**
Lower Casing	“ The costumes weren't great. ”	“ the costumes weren't great. ”
Αφαίρεση HTML tags	“ the costumes weren't great. ”	“the costumes weren't great.”
Αφαίρεση συντομογραφιών	“the costumes weren't great.”	“the costumes were not great.”

Αφαίρεση stopwords	“the costumes were not great.”	“costumes not great.”
Lemmatization	“costumes not great.”	“costume not great.”
Αφαίρεση σημείων στίξης	“costume not great.”	“costume not great”
Tokenization	“costume not great”	[“costume“, “not”, “great”]

Οι παραπάνω μέθοδοι προεπεξεργασίας βοηθούν στη ελάττωση του θορύβου, βοηθώντας μας να συγκεντρωθούμε στο ουσιώδες περιεχόμενο. Επίσης, μειώνουν το μέγεθος του τελικού λεξιλογίου που θα δημιουργηθεί κάνοντας το ξεκάθαρο, καθώς και επιτρέπουν ταχύτερο model training.

Τεχνικές γνώσεις

Τι είναι η Python;

Η Python αποτελεί μια υψηλού επιπέδου, γενικού σκοπού γλώσσα προγραμματισμού που διακρίνεται για τη συντηρητική της σύνταξη, την ευανάγνωστη και καθαρή γραφή κώδικα, καθώς και την ευελιξία και ισχυρή της κοινότητα. Σχεδιάστηκε από τον Guido van Rossum και πρωτοεμφανίστηκε το 1991, καθιερώνοντας τον εαυτό της ως μία από τις πιο δημοφιλείς γλώσσες προγραμματισμού σε παγκόσμιο επίπεδο.

Η Python ξεχωρίζει για την απλότητα της και την ικανότητά της να καλύπτει μια ευρεία γκάμα εφαρμογών, από απλά σενάρια συντομεύσεων μέχρι και πολύπλοκα συστήματα λογισμικού. Μια από τις κύριες αρχές της Python είναι η "αναγνωσιμότητα του κώδικα", που υπογραμμίζει τη σημασία του να είναι ο κώδικας ευανάγνωστος και κατανοητός, κάτι που συμβάλλει στη συντηρησιμότητα και τη συνεργασία σε μεγάλα έργα.

Μια από τις ισχυρές πλευρές της Python είναι η πολυπαραδειγματική φύση της, υποστηρίζοντας πολλαπλασιαστικό προγραμματισμό και αντικειμενοστραφή ανάπτυξη. Αυτό την καθιστά κατάλληλη για διάφορες εφαρμογές, όπως ανάπτυξη λογισμικού, ανάλυση δεδομένων, επιστημονική έρευνα, αυτοματισμό συστήματος, και διαχείριση ιστού.

Ένα ιδιαίτερο χαρακτηριστικό της Python είναι η εκτεταμένη κοινότητά της και τα χιλιάδες πακέτα (libraries) που υποστηρίζονται. Η Python Package Index (PyPI) παρέχει πρόσβαση σε έναν τεράστιο αριθμό πακέτων που επεκτείνουν τη λειτουργικότητα της Python σε ποικίλους τομείς, όπως επεξεργασία εικόνων, επιστημονικοί υπολογισμοί, παιχνίδια, και άλλα.

Η Python έχει γίνει ακόμη πιο ευρέως αποδεκτή στην επιστημονική κοινότητα λόγω των βιβλιοθηκών όπως η NumPy, η Pandas και η TensorFlow που υποστηρίζουν ανάλυση δεδομένων και μηχανική μάθηση.

Συνοπτικά, η Python είναι μια πολυδιάστατη γλώσσα προγραμματισμού που συνδυάζει την απλότητα με τη δυνατότητα επεκτασιμότητας, κάνοντάς την ιδανική για αρχάριους προγραμματιστές και επαγγελματίες του χώρου. [6], [7], [8].

Τι είναι το Jupyter;

Το Jupyter είναι ένα περιβάλλον ανάπτυξης (IDE) που χρησιμοποιείται κυρίως για την εκτέλεση προγραμμάτων Python και άλλων γλωσσών προγραμματισμού. Ονομάζεται έτσι εξαιτίας της σύνθετης λέξης "Julia", "Python" και "R", οι οποίες είναι γλώσσες προγραμματισμού που μπορούν να χρησιμοποιηθούν μέσα στο περιβάλλον Jupyter.

Το Jupyter παρέχει ένα διαδραστικό περιβάλλον όπου μπορείτε να γράψετε, να εκτελέσετε και να αποτυπώσετε κώδικα σε μια οργανωμένη μορφή. Ένα από τα βασικά χαρακτηριστικά του Jupyter είναι οι "notebooks", που επιτρέπουν τη συνδυασμένη χρήση κώδικα, εξηγήσεων, γραφικών και εξόδων. Αυτό το καθιστά ένα ισχυρό εργαλείο για την ανάπτυξη και την αναπαράσταση αναλυτικών διαδικασιών, αποσφαλμάτωσης και οπτικοποίησης δεδομένων.

Εκτός από την Python, το Jupyter υποστηρίζει και άλλες γλώσσες προγραμματισμού, όπως η Julia, η R, η Scala και πολλές άλλες. Μπορεί επίσης, να χρησιμοποιηθεί για τη δημιουργία αναφορών, παρουσιάσεων και ανάλυσης δεδομένων.

Συνοπτικά, το Jupyter είναι ένα ισχυρό περιβάλλον προγραμματισμού και ανάλυσης δεδομένων που επιτρέπει την αλληλεπίδραση, την οπτικοποίηση και την αναπαράσταση προγραμμάτων και δεδομένων.

Εξόρυξη Δεδομένων με χρήση της Python

Υπάρχουν διάφορες βιβλιοθήκες και εργαλεία που μπορούν να χρησιμοποιηθούν για την εξόρυξη δεδομένων με χρήση της Python. Παρακάτω αναφέρονται τα βασικά βήματα προκειμένου κάποιος να μπορέσει να ξεκινήσει data mining [7], [9], [10], [11]:

Εισαγωγή και προετοιμασία των δεδομένων

Χρησιμοποιώντας βιβλιοθήκες όπως η Pandas, μπορεί να γίνει η ανάγνωση αρχείων από διάφορες πηγές (αρχεία CSV, βάσεις δεδομένων κλπ.) και να εκτελεστούν διάφορες διαδικασίες όπως η επεξεργασία και ο καθαρισμός τους που θα βοηθήσουν έτσι ώστε τα δεδομένα να είναι έτοιμα για την επόμενη φάση.

Εξερεύνηση και ανάλυση των δεδομένων

Χρησιμοποιώντας την Pandas και τη NumPy, μπορεί να γίνει εξερεύνηση δεδομένων, υπολογιστικές και στατιστικές αναλύσεις, καθώς και οπτικοποίηση των δεδομένων για την αντίληψη των προτύπων και των τάσεων.

Εφαρμογή αλγορίθμων data mining

Η Python διαθέτει πληθώρα βιβλιοθηκών που παρέχουν υλοποιημένους αλγορίθμους data mining. Για παράδειγμα, η scikit-learn παρέχει μια ευρεία γκάμα αλγορίθμων μηχανικής μάθησης, ενώ η NLTK παρέχει εργαλεία για επεξεργασία φυσικής γλώσσας.

Αξιολόγηση και ερμηνεία των αποτελεσμάτων

Μετά την εφαρμογή των αλγορίθμων, μπορεί να γίνει η αξιολόγηση των αποτελεσμάτων και να ερμηνευτούν τα ανακτηθέντα πρότυπα και οι πληροφορίες. Η Python παρέχει εργαλεία για την αξιολόγηση της απόδοσης των μοντέλων, όπως οι μετρικές ακρίβειας, ανάκλησης και το F1-score.

Βασικές βιβλιοθήκες που χρησιμοποιήθηκαν για την συγγραφή του κώδικα

Pandas

Το Python Pandas είναι μια ισχυρή βιβλιοθήκη ανοικτού κώδικα για την επεξεργασία και ανάλυση δεδομένων, χτισμένη πάνω στη γλώσσα προγραμματισμού Python. Παρέχει εύκολες στη χρήση δομές δεδομένων και εργαλεία ανάλυσης δεδομένων, καθιστώντας το μια δημοφιλή επιλογή για την εργασία με δομημένα δεδομένα. Το Pandas προσφέρει αποδοτικές, ευέλικτες και εκφραστικές δομές δεδομένων, όπως ο DataFrame και ο Series, που επιτρέπουν την αποτελεσματική επεξεργασία, μετασχηματισμό και ανάλυση δεδομένων.

Η κύρια συνιστώσα του Pandas είναι ο DataFrame, ένας δισδιάστατος πίνακας δεδομένων που αποτελείται από γραμμές και στήλες. Επιτρέπει εύκολο δείκτη, αποκοπή, αναδιαμόρφωση και φιλτράρισμα των δεδομένων, παρόμοια με την εργασία με ένα φύλλο εργασίας ή μια πίνακα SQL. Ο DataFrame είναι ικανός να χειριστεί ανισόμορφους τύπους δεδομένων, καθιστώντας το ευέλικτο για μια ευρεία γκάμα εργασιών ανάλυσης δεδομένων.

Το Pandas παρέχει μια πλούσια συλλογή από συναρτήσεις και μεθόδους για τον καθαρισμό, τη μετασχηματισμό και τη διαμόρφωση των δεδομένων. Επιτρέπει εργασίες όπως η αντιμετώπιση απουσιάζουσων τιμών, η συγχώνευση και η ένωση συνόλων δεδομένων, η ομαδοποίηση και η συγκέντρωση δεδομένων, η εφαρμογή συναρτήσεων στα δεδομένα και η δημιουργία πινάκων περιστροφής. Επιπλέον, το Pandas ενσωματώνεται καλά με άλλες βιβλιοθήκες της Python, όπως ο NumPy, ο Matplotlib και ο Scikit-learn, επεκτείνοντας περαιτέρω τις δυνατότητες του για ανάλυση και οπτικοποίηση δεδομένων.

Η λειτουργική και εκφραστική σύνταξή του επιτρέπει αποτελεσματικές και συνοπτικές λειτουργίες επεξεργασίας δεδομένων. Υποστηρίζει μια ευρεία γκάμα μορφών εισόδου και εξόδου δεδομένων, συμπεριλαμβανομένων των CSV, Excel, SQL βάσεων δεδομένων και περισσότερων, επιτρέποντας την άψογη ενσωμάτωση με διάφορες πηγές δεδομένων. Το Pandas παρέχει επίσης ισχυρή λειτουργικότητα χρονοσειρών, επιτρέποντας τον χειρισμό και ανάλυση των δεδομένων που βασίζονται στο χρόνο.

Το Pandas έχει κερδίσει σημαντική δημοτικότητα στις κοινότητες της επιστήμης των δεδομένων και της ανάλυσης δεδομένων χάρη στην ευελιξία, την απόδοση και την εκτεταμένη λειτουργικότητα του. Χρησιμοποιείται ευρέως σε διάφορους τομείς, όπως οι οικονομικές επιστήμες, η οικονομία, οι κοινωνικές επιστήμες και το μηχανικό μάθημα [14].

Natural Language Toolkit (NLTK)

Το Natural Language Toolkit (NLTK) είναι μια βιβλιοθήκη για τη γλώσσα προγραμματισμού Python που παρέχει εργαλεία και κυρίως αλγορίθμους για την επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP). Η NLTK παρέχει μια συλλογή από διάφορες λειτουργίες, αλγορίθμους και πόρους που είναι χρήσιμοι για την ανάλυση και την επεξεργασία κειμένου [13].

Η βιβλιοθήκη NLTK περιλαμβάνει μια ευρεία γκάμα εργαλείων και λειτουργιών, όπως:

- Τοκενιστές (Tokenizers): Χρησιμοποιούνται για τον διαχωρισμό του κειμένου σε μικρότερες μονάδες, όπως λέξεις ή φράσεις.
- Μορφολογική ανάλυση (Morphological Analysis): Επιτρέπει την αναγνώριση της μορφολογίας των λέξεων, όπως την αναγνώριση της κλίσης και της κατάστασης των λημμάτων.
- Ανάλυση συντακτικής δομής (Parsing): Χρησιμοποιείται για την αναγνώριση της συντακτικής δομής των προτάσεων και τη δημιουργία δέντρων συντακτικής ανάλυσης.
- Εξόρυξη πληροφοριών από κείμενο (Information Extraction): Χρησιμοποιείται για τον εντοπισμό και την ανάκτηση σημαντικών πληροφοριών από το κείμενο, όπως οντότητες, συμβάντα, σχέσεις κλπ.
- Κατηγοριοποίηση κειμένου (Text Classification): Επιτρέπει την κατηγοριοποίηση του κειμένου σε διάφορες κατηγορίες, όπως θετικό ή αρνητικό συναίσθημα, θέμα κλπ.
- Επεξεργασία κειμένου (Text Preprocessing): Παρέχει εργαλεία για τον καθαρισμό, την αφαίρεση στοπ λέξεων, τον χωρισμό σε προτάσεις κλπ.

Python Matplotlib

Το Python Matplotlib είναι μια ισχυρή βιβλιοθήκη ανοικτού κώδικα για την οπτικοποίηση δεδομένων και τη δημιουργία γραφημάτων στη γλώσσα προγραμματισμού Python. Αποτελεί ένα από τα πιο δημοφιλή εργαλεία για την απεικόνιση δεδομένων λόγω της ευελιξίας, της ποικιλίας των διαθέσιμων γραφημάτων και της ευκολίας χρήσης της.

Το Matplotlib παρέχει ένα πλούσιο σύνολο λειτουργιών για τη δημιουργία γραφημάτων. Μπορεί να δημιουργήσει διαγράμματα γραμμής, διαγράμματα προβολής, διαγράμματα χρωμάτων, διαγράμματα στήλης, πίτσας, scatter plots, καθώς και πολλά άλλα. Οι χρήστες έχουν τη δυνατότητα να προσαρμόσουν και να ελέγξουν κάθε πτυχή του γραφήματος, όπως τον τύπο των γραφικών στοιχείων, τα χρώματα, την κλίμακα, τις ετικέτες και τους άξονες.

Ένα από τα ισχυρά χαρακτηριστικά του Matplotlib είναι η δυνατότητά του να παρέχει ευέλικτες επιλογές για την ενσωμάτωση των γραφημάτων σε διάφορες εφαρμογές και περιβάλλοντα. Μπορεί να εξαχθούν γραφήματα σε διάφορες μορφές αρχείων, όπως εικόνες PNG, JPEG, PDF και SVG, και να ενσωματωθούν σε εφαρμογές διαδικτύου ή αναφορές επιστημονικών εργασιών.

Η ευκολία χρήσης του Matplotlib το καθιστά ιδανικό για αρχάριους χρήστες, αλλά παράλληλα προσφέρει και προηγμένες δυνατότητες για τους προγραμματιστές που αναζητούν ακρίβεια και εξειδίκευση. Με το Matplotlib, οι χρήστες μπορούν να δημιουργήσουν εκπληκτικά γραφήματα και οπτικοποιήσεις για να αναδείξουν και να εξερευνήσουν τα δεδομένα τους [15].

Python WordCloud

Το Python WordCloud είναι μια βιβλιοθήκη ανοικτού κώδικα για τη δημιουργία και οπτικοποίηση επαναλαμβανόμενων λέξεων σε μορφή επικάλυψης σύννεφου λέξεων. Αποτελεί ένα από τα πιο δημοφιλή εργαλεία για την ανάλυση κειμένου και την οπτικοποίηση των συχνότερων λέξεων που εμφανίζονται σε ένα κείμενο.

Η βιβλιοθήκη WordCloud χρησιμοποιεί αλγορίθμους για να υπολογίσει τη σημαντικότητα των λέξεων μέσω της συχνότητας εμφάνισής τους. Στη συνέχεια, δημιουργεί ένα γραφικό στοιχείο που παρουσιάζει τις λέξεις μεγέθυνσης και τοποθετεί τις πιο σημαντικές λέξεις σε περιοχές μεγαλύτερης έντασης.

Το Python WordCloud παρέχει πολλές δυνατότητες προσαρμογής και ελέγχου για τη δημιουργία εντυπωσιακών σύννεφων λέξεων. Οι χρήστες μπορούν να προσαρμόσουν την εμφάνιση, το χρώμα, το μέγεθος και τον τύπο γραμματοσειράς των λέξεων, καθώς και να προσθέσουν προσαρμοσμένα σχήματα και χρώματα φόντου.

Η οπτικοποίηση των λέξεων με το Python WordCloud μπορεί να χρησιμοποιηθεί για να απεικονίσει μοτίβα και τάσεις σε ένα κείμενο, να αναδείξει τις κυρίαρχες έννοιες και τις συχνότερες λέξεις κλειδιά, και να παρουσιάσει επικοινωνιακά μηνύματα σε μια ευανάγνωστη και εντυπωσιακή μορφή [16].

Αποτελέσματα

Πίνακας Σύγχυσης (Confusion Matrix)

Ο πίνακας σύγχυσης είναι ένας πίνακας που χρησιμοποιείται στη μηχανική μάθηση για να αξιολογήσει την απόδοση ενός μοντέλου. Καταγράφει την ταξινόμηση των πραγματικών δεδομένων σε κατηγορίες σε σχέση με τις προβλέψεις του μοντέλου, παρέχοντας μια επισκόπηση των αληθών θετικών, αληθών αρνητικών, ψευδών θετικών και ψευδών αρνητικών προβλέψεων.

Αξιολόγηση ταξινόμησης (Classification Report)

Το "classification report" της sklearn είναι ένα όργανο αξιολόγησης της απόδοσης μοντέλων ταξινόμησης. Παρέχει μετρικές όπως η ακρίβεια (precision), ανάκληση (recall) και το f1-score για κάθε κατηγορία της ταξινόμησης, καθώς και τον συνολικό μέσο όρο αυτών των μετρικών.

Multinomial Naive Bayes:

Metric name	**Precision**	**Recall**	**F1-score**
Negative	0.84	0.89	0.87
Positive	0.88	0.84	0.86
Average	0.86	0.86	0.86

Linear SVM:

Metric name	**Precision**	**Recall**	**F1-score**
Negative	0.91	0.89	0.90
Positive	0.90	0.91	0.90
Average	0.90	0.90	0.90

Logistic Regression:

Metric name	**Precision**	**Recall**	**F1-score**
Negative	0.91	0.89	0.90
Positive	0.89	0.91	0.90
Average	0.90	0.90	0.90

Εμπειρικά Αποτελέσματα

Συγκρίνοντας τους τρεις αλγορίθμους Μηχανικής Μάθησης (Multinomial Naive Bayes, Linear SVM και Logistic Regression) για τη Sentiment Analysis, μπορούμε να επεκτείνουμε τη σύγκριση με περισσότερες λεπτομέρειες:

Απόδοση Multinomial Naive Bayes:

Ο αλγόριθμος Multinomial Naive Bayes εμφάνισε μέση τιμή F1-score 0.86, που είναι αξιοπρεπής, αλλά είναι χαμηλότερος από τους άλλους δύο αλγορίθμους.

Είναι γνωστό ότι ο Naive Bayes λαμβάνει υπόψη του τη στατιστική εξάρτηση μεταξύ των λέξεων, αλλά παραδοσιακά υποθέτει ανεξαρτησία μεταξύ των χαρακτηριστικών. Αυτή η απλότητα μπορεί να τον καθιστά λιγότερο αποδοτικό στην αντιμετώπιση πολύπλοκων γλωσσικών προβλημάτων.

Απόδοση Linear SVM:

Ο αλγόριθμος Linear SVM εμφάνισε υψηλή απόδοση, με μέση τιμή F1-score 0.90 για και τις δύο κατηγορίες.
Οι SVM είναι ιδιαίτερα καλοί στην αντιμετώπιση προβλημάτων ταξινόμησης, καθώς επιδιώκουν να βρουν ένα υπερεπίπεδο που διαχωρίζει τα δεδομένα. Η γραμμική παραλλαγή (Linear SVM) είναι απλή, αλλά αποτελεσματική στην αναγνώριση συναισθημάτων σε κείμενα.

Απόδοση Logistic Regression:

Ο αλγόριθμος Logistic Regression εμφάνισε επίσης υψηλή απόδοση, με μέση τιμή F1-score 0.90 για και τις δύο κατηγορίες.
Το Logistic Regression είναι ένα από τα πιο δημοφιλή μοντέλα για ταξινόμηση και παρουσιάζει αξιοπιστία στην αναγνώριση συναισθημάτων σε κείμενα.

Τα συμπεράσματα από τη σύγκριση των αλγορίθμων είναι τα εξής:

Οι αλγόριθμοι Linear SVM και Logistic Regression είναι πιο αποδοτικοί και αξιόπιστοι στη Sentiment Analysis σε σχέση με τον αλγόριθμο Multinomial Naive Bayes.

Επιλέγοντας μεταξύ των δύο πρώτων, η επιλογή μεταξύ Linear SVM και Logistic Regression εξαρτάται από τη συγκεκριμένη εφαρμογή και τα απαιτούμενα χαρακτηριστικά.

Οι υψηλές τιμές F1-score για όλους τους αλγορίθμους δείχνουν ότι η Sentiment Analysis είναι αποδοτική και αξιόπιστη μέθοδος για την αναγνώριση συναισθημάτων σε κείμενα.

Τελικά, η επιλογή του αλγορίθμου εξαρτάται από τις συγκεκριμένες απαιτήσεις της εφαρμογής και τον τύπο των δεδομένων που διαχειρίζεται. Παρ' όλα αυτά, η Sentiment Analysis αποτελεί έναν ισχυρό εργαλείο για την αντιληπτική κατανόηση και αξιολόγηση του συναισθηματικού τόνου σε κείμενα, και η εξέλιξη της τεχνολογίας συνεχίζει να την καθιστά όλο και πιο ισχυρή και αποτελεσματική.

Συμπεράσματα και Μελλοντικές Επεκτάσεις

Στη διπλωματική εργασία αυτή, εξερευνήσαμε τον αναπτυσσόμενο τομέα της Sentiment Analysis. Ξεκινήσαμε με μια λεπτομερή εισαγωγή στην έννοια της Sentiment Analysis, τη σημασία της και τις εφαρμογές της σε διάφορους τομείς. Στη συνέχεια, αναλύσαμε τις βασικές μεθοδολογίες και τεχνικές που χρησιμοποιούνται στη Sentiment Analysis, συμπεριλαμβανομένης της Μηχανικής Μάθησης.

Βάσει των αποτελεσμάτων και της ανάλυσης που πραγματοποιήθηκε στη διπλωματική εργασία, προκύπτουν τα εξής συμπεράσματα:

Η Sentiment Analysis είναι ένας σημαντικός τομέας έρευνας που βρίσκει εφαρμογές σε πολλούς τομείς, όπως τον εμπόριο, τα κοινωνικά μέσα, την καταναλωτική έρευνα και την πολιτική ανάλυση.

Οι αλγόριθμοι Μηχανικής Μάθησης, όπως οι Linear SVM και Logistic Regression, εμφανίζουν υψηλή απόδοση στη Sentiment Analysis, κάνοντας τους κατάλληλους για την αυτοματοποιημένη ανάλυση συναισθηματικού τόνου σε κείμενα.

Υπάρχουν πολλές μελλοντικές επεκτάσεις και δυνητικές κατευθύνσεις για περαιτέρω έρευνα στον τομέα της Sentiment Analysis:

Βελτίωση των αλγορίθμων: Οι αλγόριθμοι Μηχανικής Μάθησης μπορούν να βελτιωθούν με περαιτέρω βελτιστοποίηση των υπερ παραμέτρων και την χρήση πιο προηγμένων μοντέλων.

Χρήση βαθιών νευρωνικών δικτύων: Τα βαθιά νευρωνικά δίκτυα, όπως τα Συνελευστικά Νευρωνικά Δίκτυα (CNN) και τα Ανατροφοδοτούμενα Δίκτυα (RNN), μπορούν να βελτιώσουν την απόδοση στη Sentiment Analysis.

Συνεχής παρακολούθηση: Η Sentiment Analysis μπορεί να ωφεληθεί από συστήματα παρακολούθησης συναισθημάτων σε πραγματικό χρόνο για την αντιληπτική ανάλυση της κοινής γνώμης.

Πολυγλωσσική ανάλυση: Η επέκταση της Sentiment Analysis σε πολλές γλώσσες και η αντιμετώπιση της πολυπλοκότητας των διαφόρων γλωσσικών δομών αποτελούν προκλήσεις για μελλοντική έρευνα.

Εφαρμογές στον κοινωνικό τομέα: Η Sentiment Analysis μπορεί να χρησιμοποιηθεί για την παρακολούθηση της κοινής γνώμης σε θέματα κοινωνικής σημασίας, όπως η δημόσια υγεία και η κοινωνική αλληλεπίδραση.

Συνολικά, η Sentiment Analysis αποτελεί έναν συναρπαστικό τομέα έρευνας με πλούσιες προοπτικές και ευρεία εφαρμογή σε διάφορους τομείς. Η συνεχής εξέλιξη των τεχνικών και των μοντέλων ανοίγει τον δρόμο για περαιτέρω ανάπτυξη και βελτίωση στον τομέα αυτόν, με την δυνατότητα αντιμετώπιση πιο σύνθετων και προκλητικών προβλημάτων ανάλυσης συναισθημάτων σε κείμενα.

Βιβλιογραφικές Αναφορές

1. Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.
2. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
3. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-54.
4. Zeng, D., Chen, H., & Lusch, R. (2010). Social media analytics and intelligence. IEEE Intelligent Systems, 25(6), 13-16.
5. Aggarwal, C. C. (2018). Data mining: the textbook. Springer.
6. Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. CreateSpace.
7. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

8. Grus, J. (2015). Data Science from Scratch: First Principles with Python. O'Reilly Media.
9. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media.
10. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
11. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
12. Jupyter documentation (https://jupyter.org/documentation)
13. Natural Language Toolkit (NLTK) (https://www.nltk.org/)
14. Pandas Development Team. (2021). pandas-dev/pandas: Pandas - Python Data Analysis Library. Ανακτήθηκε από https://github.com/pandas-dev/pandas
15. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95. DOI: 10.1109/MCSE.2007.55
16. Ameya, S. (2019).amueller/word_cloud: Word Clouds in Python. Ανακτήθηκε από https://github.com/amueller/word_cloud
17. *Logistic regression* (2023) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Logistic_regression (Accessed: 14 August 2023).
18. *12.1 - logistic regression* (no date) *12.1 - Logistic Regression | STAT 462*. Available at: https://online.stat.psu.edu/stat462/node/207/ (Accessed: 14 August 2023).
19. *Support Vector Machine* (2023) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Support_vector_machine (Accessed: 14 August 2023).
20. *1.4. Support Vector Machines* (no date) *scikit*. Available at: https://scikit-learn.org/stable/modules/svm.html (Accessed: 14 August 2023).
21. *Neural network* (2023) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Neural_network (Accessed: 14 August 2023).
22. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews
23. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.
24. Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Pearson.
25. Ekman, P. (1992). An argument for basic emotions. Cognition & Emotion, 6(3-4), 169-200.
26. Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning, 412-420.
27. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.
28. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
29. Kiros, R., Zhu, Y., Salakhutdinov, R., Zemel, R., Torralba, A., Urtasun, R., & Fidler, S. (2015). Skip-thought vectors. In Advances in neural information processing systems (pp. 3294-3302).

Παράρτημα

Preprocessing

```
import pandas as pd
import nltk
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.sentiment.util import mark_negation
import csv
import re
import contractions
from tqdm import tqdm


def preprocess_text(text):
    # Convert text to lowercase
    text = text.lower()

    # Remove HTML tags
    text = re.sub(r'<[^>]+>', '', text)

    # Replace contractions with their expanded form
    text = contractions.fix(text)

    # Tokenize the text into words    "I am Stavroula"
    words = nltk.word_tokenize(text) # ["I", "am", "Stavroula"]

    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    negative_words = {'no', 'not', 'don', "don't", 'ain', 'aren',
"aren't", 'couldn', "couldn't", 'didn',
                      "didn't", 'hadn', "hadn't", 'hasn', "hasn't",
'haven', "haven't", 'isn', "isn't",
                      'mightn', "mightn't", 'mustn', "mustn't",
'needn', "needn't", 'shan', "shan't",
                      'shouldn', "shouldn't", 'wasn', "wasn't",
'weren', "weren't", 'won', "won't",
```

```
                    'wouldn', "wouldn't"}
    words = [word for word in words if word.lower() not in stop_words
- negative_words]

    # Lemmatize the words
    lemmatizer = WordNetLemmatizer()
    words = [lemmatizer.lemmatize(word) for word in words]

    # Remove punctuation
    words = [word for word in words if word.isalnum()]

    return words



# Define a function to convert the sentiment labels to 'positive' or
'negative'
def convert_sentiment_label(sentiment):
    if sentiment == 'positive':
        return 'positive'
    else:
        return 'negative'

# Load CSV file into a pandas DataFrame
df = pd.read_csv('IMDB Dataset.csv')

print('Preprocessing text...')
# Preprocess the movie reviews
preprocessed_reviews = []
list_of_reviews = df['review']
for review in tqdm(list_of_reviews):
    preprocessed_review = preprocess_text(review)
    preprocessed_reviews.append(preprocessed_review)
preprocessed_reviews

# Convert the sentiment labels to 'positive' or 'negative'
sentiments = df['sentiment'].apply(convert_sentiment_label)
print('Text preprocessed.\n')

print('Creating preprocessed_movie_reviews.csv...')
# Write the preprocessed data to a new CSV file
with open('preprocessed_movie_reviews.csv', mode='w', newline='',
encoding="utf-8") as file:
    writer = csv.writer(file)
```

```
    writer.writerow(['review', 'preprocessed_review', 'sentiment']) #
Write the header row
    for i in tqdm(range(len(preprocessed_reviews))):
        preprocessed_review = ' '.join(preprocessed_reviews[i])
        sentiment = sentiments[i]
        writer.writerow([df.loc[i, 'review'], preprocessed_review,
sentiment])
print('Created preprocessed_movie_reviews.csv\n')
```

Training, Testing and Evaluating

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.svm import LinearSVC
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report

# Load the preprocessed data
data = pd.read_csv("preprocessed_movie_reviews.csv")

# Split the data into training, validation, and test sets
train_data, test_data, train_labels, test_labels =
train_test_split(data["review"], data["sentiment"], test_size=0.2,
random_state=42)

# Convert the raw text into TF-IDF vectors
vectorizer = TfidfVectorizer()
train_vectors = vectorizer.fit_transform(train_data)
test_vectors = vectorizer.transform(test_data)

def run_model(classifier, classifier_name):
    classifier.fit(train_vectors, train_labels)
    predictions = classifier.predict(test_vectors)
    accuracy = accuracy_score(test_labels, predictions)
    print(classifier_name, "validation accuracy:
{:.2f}%".format(accuracy * 100))
```

```
    # Compute and display confusion matrix
    cm = confusion_matrix(test_labels, predictions)
    print("Confusion matrix:")
    print(cm)

    # Compute and display classification report
    cr = classification_report(test_labels, predictions)
    print("Classification report:")
    print(cr)
    print('\n')

run_model(MultinomialNB(),"Multinomial Naive Bayes")
run_model(LinearSVC(),"Linear SVM")
run_model(LogisticRegression(),"Logistic Regression")
```