ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΧΡΗΜΑΤΟΟΙΚΟΝΝΟΜΙΚΗΣ ΚΑΙ ΣΤΑΤΙΣΤΙΚΗΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΣΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΗΣ ΕΠΙΣΤΙΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ

ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΑΓΝΩΡΙΣΗ ΤΑΣΗΣ ΣΕ ΡΟΕΣ ΔΕΔΟΜΕΝΩΝ ΚΕΙΜΕΝΟΥ

ΠΕΤΡΟΠΑΝΑΓΙΩΤΑΚΗ ΚΥΡΙΑΚΗ

Διπλωματική Εργασία που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς , Ιούνιος 2023

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη Συνέλευση του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμό 20/13 – 7-2022 συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

-Τασουλής Σωτήριος, Επίκουρος Καθηγητής (Επιβλέπων)

-Μπερσίμης Σωτήριος, Αναπληρωτής Καθηγητής

- Πελέκης Νικόλαος, Αναπληρωτής Καθηγητής

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

SCHOOL OF FINANCE AND STATISTICS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN

APPLIED STATISTICS

MSC DISSERTATION

SENTIMENT ANALYSIS

IN TEXT DATA STREAMS

PETROPANAGIOTAKI KYRIAKI

MSc Dissertation submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus Greece, June 2023

Στην οικογένεια και

στους φίλους μου.

Περίληψη

Η παρούσα διατριβή διερευνά την ακρίβεια της μηχανικής μάθησης και των μοντέλων νευρωνικών δικτύων στην ανάλυση συναισθήματος σε δεδομένα του Twitter κατά τη διάρκεια της πανδημίας COVID-19. Με την αυξανόμενη σημασία των δεδομένων των μέσων κοινωνικής δικτύωσης για τις επιχειρήσεις και την κοινωνία, η ανάλυση συναισθήματος έχει γίνει ένα κρίσιμο εργαλείο για την κατανόηση της κοινής γνώμης και τη λήψη τεκμηριωμένων αποφάσεων. Η μελέτη επικεντρώνεται στην ανάλυση συναισθήματος αγγλικών tweets και συγκρίνει τις επιδόσεις πέντε μοντέλων μηχανικής μάθησης - Naive Bayes, Decision Tree, Random Forest, SVM, Gradiaent Boost- και δύο μοντέλων νευρωνικών δικτύων - RNN ,CNN - με βάση την συνολική ακρίβεια, την ακρίβεια, την ανάκληση και το F1 score. Η μεθοδολογία περιλαμβάνει τη συλλογή και προ επεξεργασία δεδομένων, την εξαγωγή χαρακτηριστικών και την αξιολόγηση των μοντέλων με τη χρήση διαφόρων μετρικών αξιολόγησης. Στόχος της μελέτης είναι να συμβάλει στη βιβλιογραφία σχετικά με την ανάλυση συναισθήματος στο Twitter με χρήση μηχανικής μάθησης και νευρωνικών δικτύων και να παράσχει πληροφορίες σχετικά με την απόδοση διαφόρων μοντέλων στην ανάλυση συναισθήματος δεδομένων Twitter. Τα αποτελέσματα της μελέτης θα μπορούσαν να είναι χρήσιμα για επιχειρήσεις, κυβερνητικές υπηρεσίες και άλλους οργανισμούς στην κατανόηση της κοινής γνώμης που σχετίζεται με το COVID-19 και στη λήψη τεκμηριωμένων αποφάσεων με βάση την ανάλυση συναισθήματος.

Abstract

This thesis aims to evaluate the accuracy of machine learning and neural network models in sentiment analysis of Twitter data. With the increasing importance of social media data for business and society, sentiment analysis has become a crucial tool to understand public opinion and make informed decisions. The study will focus on sentiment analysis of English tweets and compare the performance of four models Naive Bayes, Decision Tree, Random Forest, SVM, Gradiaent Boost- and two neural networks - RNN, CNN - based on evaluation metrics such as accuracy, precision, recall, and F1 score. The method involves data collection and pre-processing, feature extraction, and the implementation of machine learning and neural network models. The evaluation process will compare the performance of the models and visualize the results. The study aims to contribute to the literature on sentiment analysis in Twitter using machine learning and neural networks and supply insights into the performance of various models in sentiment analysis of Twitter data. The results of the study could be useful for businesses, government agencies, and other organizations in understanding public opinion related to COVID-19 and making informed decisions based on the sentiment analysis.

Περιεχόμενα

Λίστα Γραφημάτων

Λίστα Πινάκων

1 Εισαγωγή

1.1 Σκοπός της εργασίας

Ο σκοπός της παρούσας μελέτης είναι η ανάλυση των συναισθημάτων που εκφράζονται σε tweets σχετικά με τον COVID-19. Μέσω της εφαρμογής τεχνικών επεξεργασίας φυσικής γλώσσας , όπως είναι οι αλγόριθμοι μηχανικής μάθησης και τα νευρωνικά δίκτυα, σκοπός είναι η πρόβλεψη των συναισθημάτων των πολιτών σχετικά με την πανδημία του COVID-19. Στόχος είναι η κατανόηση του τρόπου που αντιλαμβάνονται οι χρήστες των κοινωνικών δικτύων την πανδημία και η ανίχνευση των κυρίαρχων συναισθημάτων που εκφράζονται στο διαδίκτυο, με σκοπό τη βελτίωση της κατανόησης της αντίληψης του κοινού για τον COVID-19 και των αντιδράσεων που προκάλεσε στην κοινωνία.

1.2 Η πανδημία του Covid-19

Η έξαρση του Covid-19 είναι μια από τις σημαντικότερες υγειονομικές κρίσεις της σύγχρονης εποχής. Ο Covid19 ανήκει στην οικογένεια των κοροναϊών και προκαλείται από τον SARS-CoV-2, οποίος μεταδίδεται κυρίως μέσω των αναπνευστικών σταγονιδίων. Η εξάπλωση του ξεκίνησε από την πόλη Wuhan τον Δεκέμβριο του 2019 και λόγω του τρόπου μετάδοσής του, ήταν ταχιά και σε παγκόσμιο επίπεδο. Αποτέλεσμα αυτού ήταν ο Παγκόσμιο Οργανισμό Υγείας να κηρύξει παγκόσμια πανδημία στις 11 Μαρτίου του 2020 το οποίο προκάλεσε δραματικές αλλαγές σε όλες τις πτυχές της ζωής. Αρχικά, οι κυβερνήσεις σε όλον τον κόσμο έλαβαν μέτρα, όπως ήταν τα lockdowns και η υποχρεωτική χρήση μάσκας σε δημόσιους χώρους, προκυμμένου να ελέγξουν την εξάπλωσης του ιού και σε μια προσπάθεια τους να μειώσουν τα απρόβλεπτα ποσοστά θνησιμότητας Όμως, η εφαρμογή αυτών των μέτρων είχε αντίκτυπος στην καθημερινή ζωή πολλών ανθρώπων με αποτέλεσμα να επηρεαστούν οι συμπεριφορές τους, οι συνήθειες τους αλλά και η ψυχική τους υγεία. Οι άνθρωποι απομακρύνθηκαν μεταξύ τους και περιόρισαν τις κοινωνικές δραστηριότητες στο ελάχιστο δυνατό. Σε αυτό το κεφάλαιο, παρουσιάζεται μια επισκόπηση των παγκόσμιων επιπτώσεών της πανδημίας και σημειώνεται ο ρόλο της ανάλυσης συναισθήματος για την κατανόηση της κοινής γνώμης την συγκεκριμένη περιόδου.

1.2.1 Οι επιπτώσεις της πανδημίας

Η πανδημία είχε σημαντικό αντίκτυπο στη σωματική και ψυχική υγεία του πληθυσμού. Από τα πρώτα προβλήματα που προέκυψαν στην αρχή της πανδημίας ήταν οι πολλές ελλείψεις σε ιατρικές προμήθειες και σε προσωπικό, που προκλήθηκαν από την αύξηση της ζήτηση για εξοπλισμό ατομικής προστασίας και άλλες βασικές ιατρικές προμήθειες για την αντιμετώπιση της. Η αυξημένη ζήτηση οδήγησε στην δημιουργία ελλείψεων σε παγκόσμια κλίμακα και στην σταδιακή κατάρρευση του συστήματος υγείας. Η σωστή λειτουργία των περισσότερων υπηρεσιών υγεία περιορίστηκε, ενώ αυξήθηκαν οι καθυστερήσεις στην περίθαλψη ασθενών που δεν νοσούσαν από COVID-19. Επιπλέον, η ψυχική υγεία πολλών ανθρώπων κλονίστηκε και αυτό αποτυπώνεται εύστοχα στην μελέτη «Mental health in the COVID-19 pandemic» [1]. Σύμφωνα με αυτή, οι πολίτες δεν φοβόντουσαν μόνο για την μόλυνσή τους αλλά ήταν επίσης ανήσυχοι για το μέλλον τους και την οικονομική τους ασφάλεια. Όλα τα προηγούμενα σε συνδυασμό με την κοινωνική απομόνωση αύξησε σε πολλούς πολίτες τα επίπεδα άγχους, στρες και κατάθλιψης. Η δημιουργία του εμβολίου θεωρήθηκε ως η λύση για την αντιμετώπιση της υγειονομικής και ανθρωπιστικής κρίσης. Ωστόσο, συνέβη κάτι αναπάντεχο όταν ήρθε η ώρα του εμβολιασμού, ένα μεγάλο μέρος της κοινότητας όχι μόνο δεν επιθυμούσε να εμβολιαστεί, αλλά ήταν και εναντίον του, εκφράζοντας την αρνητική του άποψη ελεύθερα στα μέσα κοινωνικής δικτύωσης. Αυτό με την σειρά του δημιούργησε ψευδής ειδήσεις, η διάδοση των οποίων αύξησε την δυσπιστία των πολιτών ως προς της ασφάλεια και την εγκυρότητα του εμβόλιού και που εν γένει καθυστέρησε της επιστροφή της κανονικότητας.

Ένας τομέας που επηρεάστηκε σημαντικά από την πανδημία COVID-19 ήταν αυτός της παγκόσμιας οικονομίας, που προκλήθηκε λόγω του πλήγματος των επιχειρήσεων και των βιομηχανιών. Η βασικότερη επίπτωσή αφορούσε την απώλεια των θέσεων εργασίας και το κλείσιμο των επιχειρήσεων. Αυτό ήταν αποτέλεσμα κυρίως των πολλών lockdowns και των περιοριστικών μέτρων που οδήγησαν στην μείωση των καταναλωτικών δαπανών. Γενικότερα, μειώθηκε σε μεγάλο βαθμό η οικονομική δραστηριότητα, προκαλώντας οικονομική ύφεση και το κλείσιμο πολλών επιχειρήσεων που δεν μπορούσαν να ανταπεξέλθουν σε αυτές τις συνθήκες. Επίσης, η παγκόσμια αλυσίδα εφοδιασμού υπέστη σημαντικές διαταραχές τόσο στην παραγωγή όσο και τη διανομή αγαθών. Ως εκ τούτου, δημιουργήθηκαν ελλείψεις σε βασικά αγαθά ενώ σε άλλα υπήρχαν αυξήσεις των τιμών τους. Οι κυβερνήσεις προκυμμένου να ανταπεξέλθουν και να αντιμετωπίσουν την οικονομική κρίση, εφάρμοσαν μια σειρά από

μέτρα για την τόνωση της οικονομίας και για τη οικονομική στήριξη των ατόμων και των επιχειρήσεων που επλήγησαν. Οι επιχειρήσεις, που είχαν την δυνατότητα, υιοθέτησαν την χρήση ψηφιακών τεχνολογιών, όπως το ηλεκτρονικό εμπόριο και η εργασία εξ αποστάσεως, ως μέσο άμβλυνσης των οικονομικών επιπτώσεων της πανδημίας. Με αυτό τον τρόπο επιταχύνθηκε ο ψηφιακός μετασχηματισμός στους περισσότερους τομείς και αυξήθηκαν οι επενδύσεις σε τεχνολογικές υποδομές και υπηρεσίες [2].

Η πανδημίας προκάλεσαν σημαντικές αλλαγές στον κοινωνικό ιστό. Συγκεκριμένα, οι κοινωνικές αλληλεπιδράσεις ήταν οι πρώτες που επηρεάστηκαν από αυτή, αφού έπρεπε να περιοριστούν για την πρόληψη της εξάπλωσης του ιού και για την ευρύτερη ευημερία της κοινωνίας. Η υιοθέτηση μέτρων φυσικής απομάκρυνσης, όπως ήταν η καραντίνα, περιόρισαν τις προσωπικές συναναστροφές μεταξύ των πολιτών, τόνισαν σοβαρά προβλήματα της κοινωνίας , όπως ήταν η ενδοοικογενειακής βίας, και αλλοίωσαν τη συνοχή της. Όμως, υπό αυτές τις συνθήκες οι άνθρωποι αναγκάστηκαν να προσαρμοστούν και να βρουν νέους τρόπους επικοινωνίας και αντιμετώπισης της νέας πραγματικότητας. Για αυτό οι περισσότεροι στράφηκαν στα μέσα κοινωνικής δικτύωσης που τους πρόσφεραν άμεση επικοινωνία και αλληλεπίδραση με φίλους και οικογένεια με την χρήση των video calls. Με αυτό τον τρόπο αναδείχτηκε ο ρόλο των ψηφιακών τεχνολογιών στη διατήρηση των κοινωνικών σχέσεων και στην ανακούφιση συναισθημάτων όπως η απομόνωση και η μοναξιά. Επιπλέον η περίοδος των lockdowns και η εξ αποστάσεως μάθηση επηρέασαν την κοινωνική, συναισθηματική και ακαδημαϊκή ανάπτυξή των μαθητών. Σύμφωνα με την UNESCO [3], η εξ αποστάσεως μάθηση οδήγησε σε μείωση των ακαδημαϊκών επιδόσεων των μαθητών, αύξησε την εκπαιδευτική ανισότητα και επηρέασε την κοινωνική ζωή των μαθητών, με την πλειονότητα τους να δηλώνει ότι τους λείπουν οι φίλοι και οι καθηγητές τους. Τέλος έφερε στην επιφάνεια την έλλειψη πρόσβασης στην τεχνολογία και στο διαδίκτυο και τη άνιση κατανομή των εκπαιδευτικών πόρων στους μαθητές.

1.2.2 Ο ρόλος στην ανάλυσης συναισθήματος στην πανδημία

Η ανάλυση συναισθήματος αποτέλεσε ένα πολύτιμο εργαλείο για την κατανόηση των συναισθημάτων και των στάσεων των πολιτών απέναντι στην πανδημία COVID-19. Μέσω της ανάλυσης του περιεχομένου των μέσων κοινωνικής δικτύωσης, των διαδικτυακών συζητήσεων, των ειδησεογραφικών άρθρων και άλλων ψηφιακών πηγών, η ανάλυση

συναισθήματος παρείχε πληροφορίες για τις απόψεις, τις ανησυχίες και τα συναισθήματα των πολιτών σχετικά με τον αντίκτυπο της πανδημίας στην υγεία, την οικονομία και την κοινωνία.

Μια γνωστή ιδιότητα της ανάλυσης συναισθημάτων είναι η καταγραφή των συναισθημάτων και των απόψεων σε πραγματικό χρόνο και μάλιστα για μεγάλο όγκο δεδομένων. Αυτό σημαίνει ότι μπορεί γρήγορα και με την χρήση λιγοστών πόρων να αναλύει μεγάλα σύνολα δεδομένων και να παρέχει επίκαιρες πληροφορίες για διάφορα ζητήματα που αφορούν το κοινό. Κατά την διάρκεια της πανδημίας, η ιδιότητα αυτή αξιοποιήθηκε από τους ερευνητές για να αντλήσουν πληροφορίες αναφορικά με τις απόψεις και τα συναισθήματα των ανθρώπων εκείνη την περίοδο. Πιο συγκεκριμένα, χρησιμοποίησαν τις πλατφόρμες κοινωνικής δικτύωσης για να συλλέξουν δεδομένα και αναλύοντας το περιεχόμενό τους εντόπιζαν τα συχνότερα θέματα συζήτησης σχετικά με την πανδημία καθώς και το συνολικό συναίσθημα ως προς αυτά. Μια τέτοια εφαρμογή ήταν η μελέτη των Hu κ.ά.[4] που ανέλυσε δεδομένα του Twitter .Τα αποτελέσματα της μελέτης ανέδειξαν ότι τα πιο συχνά αναφερόμενα θέματα κατά τη διάρκεια της πανδημίας ήταν το άγχος και η κατάθλιψη. Πραγματοποιώντας τέτοιου είδους αναλύσεις σε τακτά χρονικά διαστήματα, επέτρεψε στις υγειονομικές αρχές και στους υπεύθυνους λήψης μέτρων να κατανοήσουν τον τρόπο με τον οποίο το κοινό αντιδρούσε στην πανδημία καθώς και τα συναισθήματα του απέναντι σε θέματα που αφορούσαν την ψυχική υγεία. Έτσι μπορούσαν σε σύντομο χρονικό διάστημα να αναπτύξουν αποτελεσματικότερες στρατηγικές για την αντιμετώπισή της πανδημίας και να βελτιώσουν τις ήδη υπάρχοντες. Τέλος, τους επέτρεψε να εξετάσουν την αποτελεσματικότητα των εκστρατειών και των παρεμβάσεων που κάναν στην υγείας αφού είχαν την δυνατότητα της ανάλυσης του κοινού αισθήματος πριν και μετά την εφαρμογή μιας εκστρατείας .

Κατά τη διάρκεια της πανδημίας του COVID-19, η διάδοση ψευδών ειδήσεων και ανακριβών πληροφοριών ήταν ανεξέλεγκτη στα μέσα κοινωνικής δικτύωσης και αποτέλεσε σοβαρή απειλή για τη δημόσια υγεία. Σε αυτό το πλαίσιο, η ανάλυση συναισθήματος αναδείχθηκε ως ένα ισχυρό εργαλείο για την παρακολούθηση, τον εντοπισμό της παραπληροφόρησης και των ανακριβειών σχετικά με τον COVID-19, διασφαλίζοντας έτσι την ενημέρωση του κοινού με αξιόπιστες και επιστημονικά τεκμηριωμένες πληροφορίες. Αυτό επιτεύχθηκε λόγω της ικανότητας των αλγόριθμών της ανάλυσης συναισθήματος να αναλύουν τον τόνο ενός κειμένου και να εντοπίζουν μοτίβα που πιθανόν να υποδηλώνουν παραπλανητική

πληροφορία. Για παράδειγμα. ένας αλγόριθμος εξετάζοντας ένα κείμενο μπορεί να αποκαλύψει μια έξαρση του αρνητικού συναισθήματος προς ένα συγκεκριμένο εμβόλιο COVID-19, που αυτό μπορεί να δηλώνει την παρουσία παραπληροφόρησης ή φημών σχετικά με την ασφάλεια ή την αποτελεσματικότητα του εμβολίου. Επομένως, χρησιμοποιώντας αυτούς τους αλγόριθμους οι υγειονομικές αρχές και οι υπεύθυνοι χάραξης πολιτικής, αρχικά, ήταν σε θέση να εντοπίσουν και να επισημάνουν γρήγορα ψευδείς ή παραπλανητικές πληροφορίες που θα μπορούσαν να βλάψουν τη δημόσια υγεία. Στη συνέχεια, αξιοποιούσαν αυτές τις πληροφορίες μπορούσαν να αναπτύξουν στοχευμένες εκστρατείες ευαισθητοποίησης του κοινού, σε συνεργασία με διαδικτυακές κοινότητες, για αντιμετωπίσουν τις ανησυχίες των πολιτών και να προωθήσουν ακριβείς πληροφορίες.

1.3 Σχετικές εργασίες

Η ανάλυση συναισθήματος έχει προσελκύσει μεγάλο ενδιαφέρον τα τελευταία χρόνια λόγω των δυντικών εφαρμογών της σε διάφορους τομείς, όπως το μάρκετινγκ, η ανάλυση των σχολίων πελατών και η ανάλυση των μέσων κοινωνικής δικτύωσης. Ειδικά, τα μέσα όπως κοινωνικής δικτύωσης είναι ένας τομέας όπου μπορεί να αξιοποιηθεί εύκολα, καθώς εκεί οι χρήστες εκφράζουν χωρίς περιορισμούς τις σκέψεις τους, τις ανησυχίες τους και τα συναισθήματά τους. Ανά χρονικές περιόδου έχουν υπάρξει πολλές ερευνητικές μελέτες που έχουν εφαρμόσει είτε τεχνικές μηχανικής μάθησης είτε τεχνικές βαθιάς μάθησης είτε και συνδυασμό τους προκυμμένου να εκτελέσουν ανάλυσης συναισθημάτων σε δεδομένα Twitter, με σκοπό την κατανόηση συγκριμένων θεμάτων. Οι μελέτες αυτές έχουν προσφέρει πολύτιμες γνώσεις στον τομέα της επεξεργασίας της φυσικής γλώσσα. Ωστόσο, εξακολουθεί να υπάρχει ανάγκη για περισσότερη έρευνα και για εύρεση πιο ακριβών και αξιόπιστων μοντέλων που να είναι σε θέση να συλλάβουν αποτελεσματικά τις διαφοροποίησης της ανθρώπινης γλώσσας. Στην συγκεκριμένη παράγραφο θα αναφερθούν σημαντικές έρευνες που αποτέλεσαν πηγή για την μετέπειτα υλοποίηση της πειραματικής εργασίας.

Η πρώτη μελέτη που αναφέρεται εδώ είναι αυτή που διεξήχθη από τους Pak και Paroubek [5] και αποσκοπούσε στην ταξινόμηση των tweets σε τρεις κατηγορίες : θετικό, αρνητικό και ουδέτερο συναίσθημα. Η μελέτη αυτή ήταν μια από τις πρώτες που εξέτασαν τη χρήση του Twitter ως σώμα δεδομένων, δηλαδή εισήγαγε ένα μεγάλο σύνολο δεδομένων, και ως εκ τούτου αποτέλεσε σημείο αναφοράς και για μελλοντικές έρευνες. Οι ερευνητές δημιούργησαν

μια συλλογή 200.000 tweets, τα οποία προ επεξεργάστηκαν και στη συνέχεια χρησιμοποίησαν τρεις διαφορετικές μεθόδους : την ανάλυση με χρήση λεξικού, την μηχανική μάθησή και μια υβριδική προσέγγιση που συνδύαζέ τις άλλες δύο μεθόδους. Για τη πρώτη προσέγγιση, χρησιμοποίησαν το λεξικό SentiWordNet για να αποδώσουν μια βαθμολογία της πολικότητας σε κάθε λέξη που υπήρχε στο tweet και έπειτα υπολόγισαν το συνολικό συναίσθημα του tweet με βάση το άθροισμα των βαθμολογιών της πολικότητας. Για την δεύτερη προσέγγιση, ο αλγόριθμος μηχανικής μάθησης που εφάρμοσαν ήταν ο SVM για να ταξινομήσουν τα tweets στις αντίστοιχες κατηγορίες. Η εκπαίδευση του έγινε με ένα σύνολο δεδομένων από χειροκίνητα σχολιασμένα tweets. Για την υβριδική προσέγγιση, οι συγγραφείς συνδύασαν τις δύο προηγούμενες, με στόχο να βελτιώσουν την ακρίβεια της ανάλυσης συναισθήματος. Χρησιμοποίησαν το λεξικό SentiWordNet για την εξαγωγή χαρακτηριστικών από τα tweets και στη συνέχεια χρησιμοποίησαν τον ταξινομητή SVM για την ταξινόμηση των tweets με βάση αυτά τα χαρακτηριστικά. Οι συγγραφείς σημείωσαν ότι η ενσωμάτωση των emoticons ως χαρακτηριστικό βελτίωσε την ακρίβεια της ταξινόμησης. Συνολικά, η μέθοδος SVM σε συνδυασμό με λεξικά και συντακτικά χαρακτηριστικά, συμπεριλαμβανομένων των emoticons, βρέθηκε να είναι η καλύτερη προσέγγιση για την ανάλυση συναισθήματος στο Twitter σε αυτή τη μελέτη.

Ακολουθεί η αναφορά των Hussai κ.ά. [6] όπου οι συγγραφείς πραγματοποίησαν ανάλυση συναισθήματος δεδομένων κοινωνικών μέσων ενημέρωσης χρησιμοποιώντας τους αλγορίθμους Naive Bayes, Decision Trees και Random Forest. Συγκεκριμένα, και εδώ τα συναισθήματα κατηγοριοποιήθηκαν ως θετικά, αρνητικά και ουδέτερα με βάση δεδομένα χρηστών του Twitter της Ινδονησίας. Τα αποτελέσματα έδειξαν ότι οι χρήστες έτειναν να εκφράσουν περισσότερο ουδέτερες απόψεις. Οι συγγραφείς συνέκριναν τις τρεις μεθόδους χρησιμοποιώντας εργαλεία γρήγορης εξόρυξης και διαπίστωσαν ότι ο Naive Bayes πέτυχε το υψηλότερο ποσοστό ακρίβειας (86.43%), ξεπερνώντας τις άλλες δύο μεθόδους. Οι αλγόριθμοι Decision Tree και Random Forest πέτυχαν ακρίβεια 82,91%. Συνοπτικά, η μελέτη αναδεικνύει την αποτελεσματικότητα του Naive Bayes στην ανάλυση συναισθήματος, ιδίως για τα δεδομένα των μέσων κοινωνικής δικτύωσης.

Μια μελέτη που εφάρμοσε τεχνικές μηχανικής μάθησης για την ταξινόμηση των tweets που σχετίζονται με το δημόσιο αίσθημα για τον COVID-19 ήταν των Samuel κ.ά. [7]. Οι

συγγραφείς συνέλεξαν ένα μεγάλο σύνολο δεδομένων από tweets που σχετίζονται με το COVID-19 και εφάρμοσαν διάφορες τεχνικές προ επεξεργασίας κειμένου για να καθαρίσουν και να προετοιμάσουν τα δεδομένα για μηχανική μάθηση. Για την εξαγωγής χαρακτηριστικών, εφαρμόστηκαν μέθοδοι όπως bag-of-words, n-grams και TF-IDF, για να αναπαραστήσουν τα tweets ως αριθμητικά διανύσματα. Στην συνέχεια, εφάρμοσαν διάφορους αλγορίθμους μηχανικής μάθησης, συμπεριλαμβανομένων των Naive Bayes, Decision Tree, Random Forest και Support Vector Machines, για να ταξινομήσουν τα tweets σε διαφορετικές κατηγορίες συναισθήματος. Τα πειραματικά αποτελέσματα έδειξαν ότι ο αλγόριθμος Support Vector Machines υπερείχε των άλλων αλγορίθμων όσον αφορά την ακρίβεια ταξινόμησης, με ακρίβεια 88.3%. Διαπιστώθηκε επίσης ότι οι πιο συχνά χρησιμοποιούμενες λέξεις στα θετικά tweets σχετίζονταν με την ανάρρωση, την ευγνωμοσύνη και την ελπίδα, ενώ οι πιο συχνά χρησιμοποιούμενες λέξεις στα αρνητικά tweets σχετίζονταν με τον θάνατο, τον φόβο και την απομόνωση. Οι συγγραφείς προτείνουν ότι οι μέθοδοι που παρουσιάζονται στη μελέτη μπορούν να χρησιμοποιηθούν για την ανάπτυξη κινητήριων λύσεων και στρατηγικών για την αντιμετώπιση της εξάπλωσης του φόβου, του πανικού και της απόγνωσης που σχετίζονται με την πανδημία και την καλύτερη κατανόηση του συναισθήματος και των προσδοκιών των καταναλωτών για τις εταιρείες και τις μικρές επιχειρήσεις.

Αξίζει να αναφερθούν δύο μελέτες που χρησιμοποιούν νευρωνικά δίκτυα για να υλοποιήσουν ανάλυση συναισθήματος. Η πρώτη αξιοσημείωτη μελέτη είναι των Smith κ.ά. [8] η οποία προτείνει τη χρήση βαθιάς μάθησης για την ανάλυση συναισθήματος σε σύντομα κείμενα, όπως είναι οι αναρτήσεις στα μέσα κοινωνικής δικτύωσης και οι κριτικές προϊόντων. Η μελέτη βασίζεται σε αρχιτεκτονική CNN και στην ενσωματώνει τεχνικών όπως είναι word embeddings και το max-pooling για την εξαγωγή χαρακτηριστικών από τα κείμενα εισόδου. Οι συγγραφείς αξιολόγησαν το προτεινόμενο μοντέλο εφαρμόζοντάς το σε διάφορα σύνολα δεδομένων και συγκρίνοντάς το με άλλα σύγχρονα μοντέλα. Καταλήξαν στο συμπεράσματα ότι τα CNN είναι ικανά να αναγνωρίζουν δυαδικό συναίσθημα σε δεδομένα Twitter, αλλά οι επιδόσεις τους δεν είναι τόσο καλές όσο άλλες εργασίες ταξινόμησης συναισθήματος. Τόνισαν επίσης ότι τα μοντέλα βαθιάς μάθησης πρέπει να χρησιμοποιούν ένα ευρύτερο και πιο ολοκληρωμένο σύνολο δεδομένων προκειμένου να έχουν καλύτερες επιδόσεις στην ανάλυση συναισθημάτων. Η δεύτερη μελέτη [9] χρησιμοποιεί ένα διαφορετικό τύπο αρχιτεκτονικής ο οποίος είναι ο RNN. Το προτεινόμενο μοντέλο σε αυτή τη μελέτη αποτελείται από ένα κοινό

επίπεδο RNN που εκπαιδεύεται από κοινού σε πολλαπλές συναφείς εργασίες, ακολουθούμενο από επίπεδα εξόδου για συγκεκριμένες εργασίες. Σκοπός αυτής της αρχιτεκτονικής ήταν η βελτίωση της αναπαράστασης του κειμένου και η ενίσχυση της απόδοσης ταξινόμησης. Οι συγγραφείς αξιολόγησαν το προτεινόμενο μοντέλο με διάφορα σύνολα δεδομένων αναφοράς για διαφορετικές εργασίες ταξινόμησης κειμένου, συμπεριλαμβανομένης της ανάλυσης συναισθήματος, της ταξινόμησης θεμάτων και της ταξινόμησης ερωτήσεων. Τα πειραματικά αποτελέσματα έδειξαν ότι το προτεινόμενο μοντέλο υπερτερεί έναντι άλλων σύγχρονων μοντέλων στα περισσότερα σύνολα δεδομένων. Επιπλέον, οι συγγραφείς έδειξαν ότι το προτεινόμενο μοντέλο μπορεί να αξιοποιήσει αποτελεσματικά πρόσθετες συναφείς εργασίες για τη βελτίωση της απόδοσης ακόμη και όταν υπάρχουν περιορισμένα δεδομένα με ετικέτες για την κύρια εργασία. Συνοπτικά, η εργασία παρουσιάζει μια πολλά υποσχόμενη προσέγγιση για την ταξινόμηση κειμένου με χρήση RNN και μάθηση πολλαπλών εργασιών.

2 Θεωρητικό υπόβαθρο

2.1 Η Συναισθηματική Ανάλυση στα μέσα κοινωνικής δικτύωσης.

2.1.1 Εισαγωγή

Τα μέσα κοινωνικής δικτύωσης διαδραματίζουν σημαντικό ρόλο στη σύγχρονη επικοινωνία, αφού οι περισσότεροι άνθρωποι ξοδεύον μεγάλο μέρος της ημέρας τους σε αυτά εκφράζοντας τις απόψεις και τα συναισθήματά τους άλλα και για να ενημερώνονται για τις τελευταίες εξελίξεις. Με την πανδημία, έχει αυξηθεί ραγδαία η δημοτικότητά τους και ο όγκος των παραγόμενων δεδομένων. Για την συγκεκριμένη περίοδο αλλά και εν γένει, έχουν αποτελέσει σημαντική πηγή δεδομένων για επιχειρήσεις και κυβερνήσεις καθώς και ένα ενδιαφέρον πεδίο έρευνας για τους ακαδημαϊκούς ερευνητές. Ο πιο ενδιαφέρον ερευνητικός τομέας τους είναι η ανάλυση συναισθημάτων, η οποία εντοπίζει και κατηγοριοποιεί αυτόματα τις απόψεις και τα συναισθήματα που εκφράζονται στα περιεχόμενα των μέσων κοινωνικής δικτύωσης. Το παρόν κεφάλαιο παρέχει μια επισκόπηση των μεθόδων που χρησιμοποιούνται για την ανάλυση συναισθήματος, περιγράφει τις προκλήσεις που σχετίζονται με την επεξεργασία της φυσικής γλώσσας και τις διαφορετικές εφαρμογές αυτών των μεθόδων.

2.1.2 Ορισμός και μέθοδοι συναισθηματικής ανάλυσης.

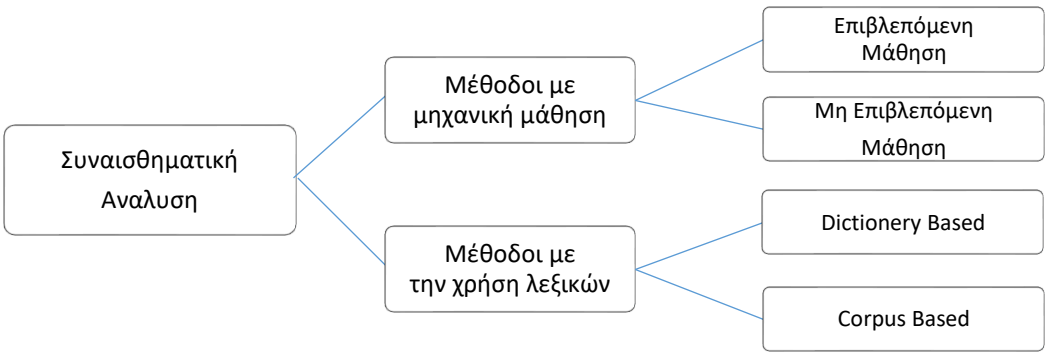
Η ανάλυση συναισθήματος, που συνήθως αναφέρεται και ως εξόρυξη γνώμης (opinion mining), είναι ένας κλάδος της επεξεργασίας της φυσικής γλώσσας (natural language processing). Πρόκειται για μια υπολογιστική τεχνική που στοχεύει στην κατηγοριοποίηση των απόψεων, των στάσεων και των συναισθημάτων που εκφράζονται σε ένα δοσμένο κείμενο. Ειδικότερα, η ανάλυση συναισθήματος αποσκοπεί στην εξαγωγή πληροφοριών από ένα κείμενο και στην αυτόματη αναγνώριση της πολικότητα του συναισθήματος που εκφράζεται σε αυτό. Η πολικότητα του συναισθήματος μπορεί να είναι θετική, αρνητική ή ουδέτερη και βάση αυτής γίνεται η ταξινόμηση του κειμένου στην αντίστοιχη κατηγορία. Για παράδειγμα, εάν ένας πελάτης εκφράζει την ικανοποίησή του για την ποιότητα ενός προϊόντος σε μια κριτική, το κείμενο κατατάσσεται στη θετική κατηγορία, καθώς το συναίσθημα που εκφράζεται είναι θετικό. Αντίθετα, αν ο πελάτης εκφράσει απογοήτευση για το προϊόν, το κείμενο θα ταξινομηθεί στην αρνητική κατηγορία, καθώς το συναίσθημα που εκφράζεται είναι αρνητικό.

Έχουν προταθεί διάφορες μέθοδοι για την ανάλυση συναισθήματος. Οι μέθοδοι αυτές χωρίζονται σε δύο κύριες κατηγορίες. Ακολουθεί ένας περιγραφικός κατάλογος αυτών των κατηγοριών.

- **Τεχνικές με χρήση λεξικών**: Οι μέθοδοι με την χρήση λεξικών βασίζονται σε προϋπάρχοντας λεξικά λέξεων και στις ανάλογες βαθμολογίες συναισθήματος των λέξεων τους για την ανάλυση του συναισθήματος σε ένα κείμενο [10]. Η εν λόγω μέθοδοι περιλαμβάνουν την απόδοση μιας βαθμολογίας συναισθήματος σε κάθε λέξη ενός κειμένου και έπειτα τον συνδυασμό αυτών των βαθμολογιών για τον υπολογισμό μιας συνολικής βαθμολογίας συναισθήματος για το κείμενο. Η αποδοτικότητα τους εξαρτάται σε μεγάλο βαθμό από την πληρότητα των λεξικών που χρησιμοποιούνται.

- **Τεχνικές με χρήση μηχανικής μάθησης**: Οι μέθοδοι αυτοί χρησιμοποιούν αλγορίθμους μηχανικής μάθησης για την εκπαίδευση μοντέλων σε δεδομένα με ετικέτες και την μετέπειτα πρόβλεψη του συναισθήματος σε νέα δεδομένα κειμένου. Τα μοντέλα βασίζονται σε διάφορες τεχνικές ανάλογα το είδος του κειμένου, όπως η τεχνική Naive Bayes και η Μηχανών Διανυσμάτων Υποστήριξης (SVM).

Το Γράφημα 2.1. απεικονίζει τις δύο βασικές κατηγορίες και τις διακρίσεις των τεχνικών τους.



Γράφημα 2.1: Κατηγορίες μεθόδων ανάλυσης συναισθήματος

2.1.3 Επίπεδα ανάλυσης συναισθήματος.

Υπάρχουν τρία επίπεδα συναισθηματικής ανάλυσης και η διάκρισή τους βασίζεται στην έκταση του κειμένου που αναλύεται καθώς και στον τρόπο με τον οποίο γίνεται αντιληπτό το συναίσθημα που εκφράζεται σε αυτό. Πιο συγκεκριμένα, η ανάλυση σε επίπεδο κειμένου (document-level) αφορά την ανίχνευση του συναισθήματος σε ένα ολόκληρο κείμενο, όπως είναι ένα άρθρο ειδήσεων ή οι κριτικές των πελατών. Η ανάλυση θεωρεί ότι κάθε κείμενο εκφράζει απόψεις για μια μόνο οντότητα [11]. Επίσης, στοχεύει στον εντοπισμό της συνολική διάθεση που εκφράζεται στο κείμενό, χωρίς να επικεντρώνεται σε μεμονωμένες φράσεις ή λέξεις. Η ανάλυση συναισθήματος σε επίπεδο πρότασης (sentence-level) αναλύει μεμονωμένες προτάσεις και με αυτό τον τρόπο παρέχει μια πιο λεπτομερή ανάλυση του συναισθήματος. Η διαδικασία περιλαμβάνει τον εντοπισμό των προτάσεων που εκφράζουν μια άποψη ή ένα συναίσθημα, τον διαχωρισμό τους από τις αντικειμενικές προτάσεις και τέλος τον προσδιορισμό της πολυπλοκότητα τους. Η ανάλυση σε επίπεδο λεκτικής μονάδας (aspect-level), γίνεται βάση οντοτήτων ή πτυχών και εστιάζει στο συναίσθημα συγκεκριμένων πτυχών ή χαρακτηριστικών ενός προϊόντος ή μιας υπηρεσίας αντί για το συνολικό συναίσθημα ενός κειμένου ή μεμονωμένων προτάσεων. Χρησιμοποιείται κυρίως σε επιχειρήσεις αφού τους προσφέρει μια βαθύτερη κατανόηση του πώς αισθάνονται οι πελάτες τους για συγκεκριμένες πτυχές των προσφορών ή των υπηρεσιών τους.

2.1.4 Προκλήσεις της συναισθηματικής ανάλυσης.

Η παρούσα εργασία ασχολείται με την ανάλυση συναισθήματος σε tweets, τα οποία διακρίνονται για τη συνοπτικότητα και το περιορισμένο μέγεθος τους, καθώς συνήθως αποτελούνται από 10 έως 15 λέξεις. Επομένως, λόγω του μικρούς τους μεγέθους, η ανάλυση τους πραγματοποιείται σε επίπεδο πρότασης. Η ανάλυσή αυτή αποτελεί πρόκληση, ιδιαίτερα για τα tweets τα οποία εκ φύσεως είναι μη δομημένα και χρησιμοποιούν ασυνήθιστο λεξιλόγιο, το οποίο περιλαμβάνει επαναλαμβανόμενες λέξεις, αργκό και σύμβολα. Οι συγκεκριμένες ιδιαιτερότητες σε συνδυασμό με την ανάλυση σε επίπεδο πρότασης, καθιστούν δύσκολη την ανίχνευση και την εξαγωγή συναισθήματος. Η συγκεκριμένη ενότητα επικεντρώνεται στις δυσκολίες που προκύπτουν κατά την διάρκεια της ανάλυσης συναισθήματος και τονίζει τα σημεία που συμβάλουν στην ανάπτυξη αξιόπιστων μεθοδολογιών.

Η κύρια πρόκληση που συναντάται κατά την πραγματοποίηση της ανάλυση συναισθήματος στα μέσα κοινωνικής δικτύωσης είναι η γλωσσική ασάφεια. Συγκεκριμένα, οι αλγόριθμοι ανάλυσης συναισθήματος λαμβάνουν υπόψη το πλαίσιο στο οποίο χρησιμοποιούνται οι λέξεις προκειμένου να προσδιορίσουν με ακρίβεια το συναίσθημα που εκφράζεται στο κειμένου. Ωστόσο, η γλώσσα που χρησιμοποιείται συνήθως από τους χρήστες είναι άτυπη, περιλαμβάνοντας αργκό και συντομογραφίες, ,γεγονός που καθιστά δύσκολο τον ακριβή προσδιορισμό αυτού του πλαισίου. Ένα τέτοιο παράδειγμα είναι η λέξη "sick" στα αγγλικά η οποία μπορεί να έχει θετική ή αρνητική σημασία ανάλογα την φράση. Οι χρήστες χρησιμοποιούν επίσης συχνά σαρκασμό, ειρωνεία ή χιούμορ για να εκφράσουν τα συναισθήματά και τις απόψεις τους σχετικά με ένα συγκεκριμένο θέμα ή προϊόν. Όμως, οι έννοιές αυτές είναι δύσκολο να εντοπιστούν και να ερμηνευτούν σωστά από τις τεχνικές ανάλυσης συναισθήματος, διότι το πλαίσιο στο οποίο χρησιμοποιούνται οι λέξεις δεν είναι σαφώς προκαθορισμένο. Επομένως, για να προσδιοριστούν με ακρίβεια τα συναισθήματα που εκφράζονται στις αναρτήσεις στα μέσα κοινωνικής δικτύωσης, απαιτούνται αλγόριθμοι που μπορούν να κατανοήσουν τις λεπτές και ποικίλες αποχρώσεις της γλώσσας.

Μια άλλη πρόκληση είναι η ακρίβεια των αποτελεσμάτων του αλγορίθμου. Η ακρίβεια μπορεί να επηρεαστεί από διάφορους παράγοντες, συμπεριλαμβανομένης της μεροληψίας των δεδομένων, του θόρυβος και της υποκειμενικότητα. Ειδικότερα, η ακρίβεια των προβλέψεων των μοντέλων της ανάλυσης συναισθήματος βασίζεται στην ακρίβεια των δεδομένων εκπαίδευσης. Ωστόσο, εάν τα δεδομένα εκπαίδευσης στρέφονται προς ένα συγκεκριμένο συναίσθημα, τότε τα μοντέλα είναι πιθανό να δώσουν μεροληπτικά αποτελέσματα. Επιπλέον, ο θόρυβο των μέσων κοινωνικής δικτύωσης, που συχνά συναντάται με την μορφή ορθογραφικών λαθών, emoticon και hashtags, μπορεί να προκαλεί σύγχυσή στους αλγορίθμους και να επηρεάσει την ακρίβεια των αποτελεσμάτων τους. Συνεπώς, η απομάκρυνσή του είναι σημαντική για την επίτευξη καλύτερης εκπαίδευσης και άρα καλύτερων προβλέψεων. Η ανάλυση συναισθήματος είναι εν γένει υποκειμενική, πράγμα που σημαίνει ότι διαφορετικοί ερευνητές μπορεί να έχουν διαφορετικές ερμηνείες του ίδιου κειμένου. Το γεγονός αυτό πιθανό να οδηγήσει σε διαφορετικά αποτελέσματα της ανάλυσης συναισθήματος για το ίδιο κείμενο, το οποίο με την σειρά του μπορεί να επηρεάσει τη διαδικασία λήψης αποφάσεων. Ως εκ τούτου, η εκπαίδευση και η αξιολόγηση των αλγορίθμων θα πρέπει να πραγματοποιείται με τη

συμμετοχή εμπειρογνωμόνων στον τομέα της ανάλυσης συναισθήματος, και τα όρια και οι παραδοχές της διαδικασίας θα πρέπει να ορίζονται σαφώς.

Η τελευταία πρόκληση που συναντούν οι αλγόριθμοι ανάλυσης συναισθήματος είναι η έλλειψη δεδομένων. Είναι γεγονότος ότι οι αλγόριθμοι ανάλυσης συναισθήματος απαιτούν αρκετά δεδομένα για την εκπαίδευση τους και τον έλεγχο της ακρίβειάς τους. Παρόλο που οι πλατφόρμες των μέσων κοινωνικής δικτύωσης παράγουν έναν τεράστιο όγκο δεδομένων, δεν σημαίνει πως όλα αυτά τα δεδομένα είναι σχετικά ή χρήσιμα για την ανάλυση συναισθημάτων. Για παράδειγμα, υπάρχουν tweets που περιέχουν μόνο μια διεύθυνση URL ή ένα hashtag και επομένως δεν παρέχουν χρήσιμες πληροφορίες σχετικά με το συναίσθημα. Επιπροσθέτως , σε κάποιες πλατφόρμες μέσων κοινωνικής δικτύωσης υπάρχουν περιορισμοί ως προς την πρόσβαση ή τη χρήση δεδομένων, λόγω ηθικών ζητημάτων ή πολιτικής της πλατφόρμας. Τέλος, όταν πραγματοποιείται έρευνα για ένα συγκεκριμένο θέμα, όπως είναι η υγειονομική περίθαλψη, οι αλγόριθμοι αυτοί απαιτούν δεδομένα που προέρχονται από τον συγκεκριμένο τομέα προκυμμένου να δώσουν ακριβή ανάλυση των συναισθημάτων των χρηστών. Όλα τα παραπάνω αποτελούν παράγοντες που εμποδίζουν τόσο την ανάπτυξη όσο και την βελτίωση αυτών των αλγορίθμων.

2.1.5 Εφαρμογές της ανάλυσης συναισθήματος.

Αρκετοί κλάδοι, συμπεριλαμβανομένων των επιχειρήσεων, της πολιτικής και της εξυπηρέτησης πελατών, χρησιμοποιούν ανάλυση συναισθήματος. Στην συνέχεια παρουσιάζονται οι πιο δημοφιλείς χρήσεις της ανάλυσης συναισθημάτων σε διάφορους τομείς.

1. Επιχειρήσεις

Σε ένα έντονα ανταγωνιστικό επιχειρηματικό περιβάλλον, η ικανότητα απόκτησης ανταγωνιστικού πλεονεκτήματος είναι ζωτικής σημασίας. Η ανάλυση συναισθήματος παρέχει μια τέτοια ευκαιρία και για αυτό αποτελεί πλέον αναπόσπαστο μέρος των επιχειρήσεων. Ο κύριος τρόπος με τον οποίο χρησιμοποιείται είναι για την ανάλυση των αξιολογήσεων που αφήνουν οι πελάτες στα μέσα κοινωνικής δικτύωσης. Μέσω των σχολίων, οι επιχειρήσεις μπορούν να παρακολουθούν την ικανοποίηση των καταναλωτών, να αποκτούν πληροφορίες σχετικά με τις προτιμήσεις τους και να εντοπίσουν προβλήματα που σχετίζονται με τα προϊόντα τους και τις υπηρεσίες τους. Οι πληροφορίες αυτές μπορούν στη συνέχεια να χρησιμοποιηθούν

για τη λήψη αποφάσεων για τη βελτίωσή των προϊόντων και των υπηρεσιών. Ένας άλλος τρόπος που χρησιμοποιείται είναι για έρευνα αγοράς και ανάλυση των ανταγωνιστικών προϊόντων. Ειδικότερα, η διάδοση των μέσων κοινωνικής δικτύωση έδωσε την δυνατότητα στις επιχειρήσεις να λαμβάνουν σε πραγματικό χρόνο σχόλια κα ανατροφοδοτήσεις πελατών, τόσο για τα δικά τους προϊόντα όσο και για των ανταγωνιστών τους. Ως εκ τούτο, αναλύοντας το δικό τους περιεχόμενο, μπορούν να αποκτούν μια εικόνα για τα δυνατά και αδύνατα τους σημεία. Εντοπίζοντας έγκαιρα τόσο τα δικά τους αδύναμα σημεία όσο και τα ζητήματα των πελατών τους και αντιμετωπίζοντας τα εγκαίρως, οι επιχειρήσεις μπορούν να βελτιώνουν την εξυπηρέτηση των πελατών τους και να οικοδομήσουν ισχυρότερες σχέσεις με τους πελάτες τους.

2. Πολιτική

Η ανάλυση συναισθήματος αποτελεί βασικό εργαλείο στον τομέα της πολιτικής ανάλυσης, καθώς συμβάλλει στην κατανόηση της κοινής γνώμης και των πολιτικών τάσεων. Στα μέσα κοινωνικής δικτύωσης, η ανάλυση συναισθήματος χρησιμοποιείται συχνά για την πρόβλεψη των εκλογών. Ο λόγος είναι ότι οι πλατφόρμες των μέσων κοινωνικής δικτύωσης αποτελούν μόνιμή πηγή πολιτικού λόγου και παρέχουν μεγάλο όγκο πληροφοριών που είναι απαραίτητος για την αποτελεσματικότητα των μεθόδων της συναισθηματικής ανάλυσης. Τα πολιτικά κόμματα εφαρμόζουν αυτή τη μέθοδο για να εκτιμήσουν το κοινό αίσθημα απέναντι τόσο στις πολιτικές τους όσο και στους πολιτικούς τους υποψηφίους, προκυμμένου λάβουν τις απαραίτητες ενέργειες για να αυξηθούν οι πιθανότητες εκλογή τους. Μια μελέτη που κατέδειξε την δυνατότητα της ανάλυσης συναισθήματος στην πρόβλεψη εκλογών, ήταν αυτή του Tumasjan [12]. Η μελέτη αυτή εξέτασε περισσότερα από 100.000 tweets που ήταν σχετικά με τις γερμανικές ομοσπονδιακές εκλογές του 2009 και χρησιμοποίησε έναν αλγόριθμο ανάλυσης συναισθήματος για να προσδιορίσει το συναίσθημα απέναντι στα πολιτικά κόμματα και τους υποψηφίους τους. Η μελέτη κατέληξε στο συμπέρασμα ότι ο αλγόριθμος ανάλυσης συναισθήματος ήταν σε θέση να προβλέψει τα εκλογικά αποτελέσματα με ακρίβειας άνω του 80%, ξεπερνώντας τις παραδοσιακές μεθόδους δημοσκοπήσεων. Η ανάλυση συναισθήματος στα δεδομένα των μέσων κοινωνικής δικτύωσης χρησιμοποιείται ακόμα για να βοηθήσει τους πολιτικούς αναλυτές να κατανοήσουν καλύτερα τις απόψεις και τις στάσεις του κοινού απέναντι σε πολιτικά ζητήματα. Αυτό συμβαίνει γιατί είναι ικανή να αναγνωρίσει τα κυρίαρχα

αισθήματα και τις απόψεις που επικρατούν στα κοινωνικά μέσα και να προβλέψει πιθανές αλλαγές στην κοινή γνώμη στο μέλλον. Ένα τέτοιο παράδειγμα είναι η μελέτη του διεξήγαγε ο Thelwall [13] η οποία ανέλυσε αναρτήσεις στο Twitter που σχετίζονται με τις πολιτικές μεταρρύθμισης της πρόνοιας του Ηνωμένου Βασιλείου. Η μελέτη είχε ως στόχο να διερευνήσει εάν το Twitter είναι εν δύναμη μια χρήσιμη πηγή δεδομένων για την ανάλυση της στάσης του κοινού απέναντι σε πολιτικά ζητήματα. Η μελέτη διαπίστωσε ότι η κοινό γνώμη προς τις μεταρρύθμισης ήταν κατά κύριο λόγο αρνητική, με περίπου το 63% των tweets να εκφράζει αρνητικό συναίσθημα. Το συμπέρασμα του Thelwall ήταν ότι η ανάλυση συναισθήματος σε δεδομένα των μέσων κοινωνικής δικτύωσης είναι ένα χρήσιμο εργαλείο για ανάλυση των απόψεων των πολιτών απέναντι στα πολιτικά ζητήματα.

3.Τουρισμός

Η επιτυχία και η δημοτικότητα των τουριστικών επιχειρήσεων εξαρτώνται σε μεγάλο βαθμό από την ικανοποίηση των πελατών τους σχετικά με τις υπηρεσίες που προσφέρουν. Για το λόγο αυτό, είναι σημαντικό οι επιχειρήσεις αυτές να αφουγκράζονται τις απόψεις και τα συναισθήματα των πελατών τους ώστε να βελτιώνουν και να καθιστούν πιο ελκυστικές τις υπηρεσίες τους. Μέσω της ανάλυσης των διαδικτυακών κριτικών για ξενοδοχεία, αξιοθέατα και προορισμούς, οι επιχειρήσεις μπορούν να κατανοήσουν τα πλεονεκτήματα και τα μειονεκτήματα των υπηρεσιών τους και να εντοπίσουν τις επικρατούσες τάσεις στα ταξίδια. Αυτό τους επιτρέπει, σύμφωνα πάντα με το επικρατέστερο κλίμα, να αναλαμβάνουν ανάλογες διορθωτικές δράσεις. Ακόμα, η ανάλυση συναισθήματος μπορεί να χρησιμοποιηθεί για τον εντοπισμό δημοφιλών προορισμών ή αξιοθέατων μεταξύ των τουριστών, διότι επιτρέπει στις επιχειρήσεις να αναπτύξουν στοχευμένες στρατηγικές μάρκετινγκ για την προσέλκυση των τουριστών. Μια μελέτη που δείχνει τις τάσεις ανάμεσα στους τουρίστες ήταν αυτή του Gretzel [14] που ανέλυσε το περιεχόμενο των μέσων κοινωνικής δικτύωσης που αφορούσε την πόλη του Λονδίνου. Τα δεδομένα προέρχονταν από διαφορετικές πλατφόρμας και περιείχαν κριτικές στο TripAdvisor και φωτογραφίες στο Instagram. Η μελέτη διαπίστωσε ότι οι τουρίστες ήταν θετικοί ως προς τα πολιτιστικά αξιοθέατα και τα ιστορικά μνημεία του Λονδίνου, πληροφορία που αξιοποιήθηκε υπερ των τουριστικών επιχειρήσεων.

2.2 Τεχνικές με χρήση λεξικών

2.2.1 Εισαγωγή

Η ανάλυση συναισθήματος βάσει λεξικού είναι μια ευρέως διαδεδομένη προσέγγιση στον τομέα της επεξεργασίας φυσικής γλώσσας, η οποία βασίζεται στη χρήση προκαθορισμένων λεξικών συναισθήματος. Η προσέγγιση αυτή είναι ευρύτερα αποδεκτή και διακρίνεται για την απλότητα και την αποτελεσματικότητά της. Η παρούσα ενότητα παρέχει μια επισκόπηση αυτής της προσέγγισης, συμπεριλαμβανομένης της διαδικασίας δημιουργίας των λεξικών συναισθήματος, του υπολογισμού των βαθμολογιών συναισθήματος καθώς και των δυσκολιών που συνδέονται με αυτή την προσέγγιση.

2.2.2 Περιγραφή τεχνικής και σημασία επιλογής λεξικού.

Η τεχνική που βασίζεται στην χρήση λεξικών (lexicon - base) χρησιμοποιεί λεξικά που αποτελούνται από προκαθορισμένες λέξεις και εκφράσεις με τις αντίστοιχες βαθμολογίες συναισθήματος ή αλλιώς κατηγορίες συναισθήματος. Τα λεξικά αυτά ονομάζονται λεξικά συναισθήματος και περιέχουν λίστες με φράσεις και λέξεις που συνήθως χρησιμοποιούνται για να εκφράσουν κάποιο συναισθήματα. Τέτοια παράδειγμα αποτελούν οι λέξεις "καταπληκτικό" και "εξαιρετικό", που εκφράζουν θετικό συναίσθημα, η λέξη "απαίσιος" και η φράση "Μου κόστισε μια περιουσία " που εκφράζουν αρνητικό συναίσθημά. Η τεχνική αυτή περιλαμβάνει αρχικά, την ανάθεση μιας κατηγορίας συναισθήματος σε κάθε λέξη που περιέχει το κείμενο. Στην συνέχεια, οι μεμονωμένες βαθμολογίες αθροίζονται με σκοπό τον υπολογισμό μιας συνολικής βαθμολογίας συναισθήματος για το κείμενο. Η συνολική βαθμολογία είναι αυτή που καθορίζει και ταξινομεί το κείμενο στην αντιστοιχεί συναισθηματική κατηγορία.

Η επιλογή του λεξικού συναισθήματος αποτελεί σημαντική παράμετρο σε αυτή την μέθοδο διότι μπορεί να έχει αντίκτυπο στην ακρίβεια και την αξιοπιστία των αποτελεσμάτων της ανάλυσης συναισθήματος. Η ανάθεση της βαθμολογίας συναισθήματος σε κάθε λέξη εξαρτάται κατά κύριο λόγο από το λεξικό που χρησιμοποιείται. Διαφορετικά λεξικά μπορεί να δώσουν διαφορετικές κατηγορίες συναισθήματος για το ίδιο κείμενο. Το γεγονός αυτό οφείλεται στο υποκειμενικό χαρακτήρα που έχουν οι βαθμολογίες συναισθήματος που αποδίδονται στις λέξεις των λεξικών συναισθήματος και που πιθανό να διαφέρουν ανάλογα

την κρίση των δημιουργών τους. Για παράδειγμα, μία λέξη όπως το "εξαιρετικό" μπορεί να χαρακτηρίζεται από ένα λεξικό ως έντονα θετική, ενώ από ένα άλλο λεξικό ως μέτρια θετική.

Οι ερευνητές σε μια προσπάθεια να ξεπεράσουν τους περιορισμούς που προκύπτουν με την χρήσης ενός μόνο λεξικού συναισθήματος και για να βελτιώσουν ταυτόχρονα και την ακρίβεια, πρότειναν τη χρήση πολλαπλών λεξικών συναισθημάτων. Συγκεκριμένα, μια έρευνα που πραγματοποιήθηκε από τους Mohammadi και Davarpanah Jazi [15] έδειξε ότι ο συνδυασμός πολλαπλών λεξικών συναισθημάτων, όπως των SentiWordNet, AFINN και Opinion Lexicon, βελτίωσε σημαντικά την ακρίβεια της ανάλυσης συναισθήματος για τα δεδομένα του Twitter. Επιπλέον, μια άλλη έρευνα των Baccianella και Esul [16] που πραγματοποιήθηκε σε ιταλικό κείμενο διαπίστωσε ότι με την χρήση πολλών λεξικών, η ακρίβεια της ταξινόμησης συναισθημάτων βελτιώθηκε. Τα παραπάνω αποτελέσματα υποδηλώνουν ότι ο συνδυασμός πολλαπλών λεξικών συναισθημάτων έχει την δυνατότητα να βελτιώσει την ακρίβεια και την αξιοπιστία της ανάλυσης συναισθήματος. Συνεπώς, είναι σημαντικό να επιλεχτεί ο κατάλληλος συνδυασμός από λεξικά που να είναι κατάλληλα για τον αντίστοιχο τομέα και την γλώσσα που αναλύεται, προκειμένου να διασφαλιστεί το βέλτιστο αποτέλεσμα.

2.2.3 Τύποι λεξικών.

Όπως προαναφέρθηκε, τα λεξικά συναισθήματος αποτελούν κρίσιμο παράγοντα για την αξιολόγηση των αποτελεσμάτων της συναισθηματικής ανάλυσης. Υπάρχει ποικιλία ως προς τον τύπο λεξικών που χρησιμοποιούνται και η διαφοροποίηση τους έγκειται στο περιεχόμενο των λέξεων και τον τομέα στον οποίο αναφέρονται. Οι τύποι λεξικών περιλαμβάνουν:

- **Λεξικά γενικών συναισθημάτων**. Τα συγκείμενα λεξικά περιέχουν μια ποικιλία από λέξεις που ου μπορούν να συνδεθούν με διάφορα συναισθήματα, όπως η χαρά, η λύπη, η έκπληξη κ.λ.π. Αυτά τα λεξικά βασίζονται σε μια λίστα λέξεων που ήδη έχουν θετικές, αρνητικές ή ουδέτερες βαθμολογίες συναισθήματος και έχουν σχεδιαστεί με στόχο να καταγράφουν τη γενική συναισθηματική πολικότητα σε ένα κείμενο. Λόγω αυτού, είναι ικανά για να λειτουργούν σε διαφορετικούς τομείς και περιβάλλοντα. Παρόλο που έχουν ένα ευρύ πεδίο εφαρμογής και καλύπτουν μεγάλο αριθμό λέξεων, δεν αντικατοπτρίζουν πλήρως τη πολυπλοκότητα την ανθρώπινης γλώσσας αφού μπορεί να παραλείπουν

συγκεκριμένες εκφράσεις συναισθημάτων και αποχρώσεις που είναι απαραίτητες σε συγκεκριμένους τομείς. Παραδείγματα τέτοιων λεξικών είναι το λεξικό AFINN, το λεξικό SentiWordNet και το λεξικό VADER.

- **Λεξικά για συγκεκριμένους τομείς**. Αυτά τα λεξικά είναι προσαρμοσμένα για να χρησιμοποιούνται σε συγκεκριμένους τομείς, όπως είναι η υγεία η οικονομία, η πολιτική. Σε αντίθεση με τα λεξικά γενικών συναισθημάτων, αυτά έχουν σχεδιαστεί για να περιλαμβάνουν εκφράσεις και ορολογία για συγκεκριμένους τομείς. Συνήθως έχουν μικρό εύρος και περιορισμένο αριθμό λέξεων, αλλά πολλές φορές είναι ικανά παρέχουν πιο ακριβή αποτελέσματα στην ανάλυση συναισθήματος για τον τομέα που αναφέρονται. Παραδείγματα τέτοιων λεξικών είναι το λεξικό Financial Sentiment, το Political Sentiment Lexicon και το Health Opinion Lexicon.

- **Λεξικά για αργκό (Slang lexicons)**. Αυτά τα λεξικά περιέχουν λέξεις και εκφράσεις που χρησιμοποιούνται κυρίως στα μέσα κοινωνικής δικτύωσης ή σε διαδικτυακές πλατφόρμες. Τα συγκεκριμένα λεξικά είναι χρήσιμα για την ανάλυση συναισθημάτων άτυπων δεδομένων κειμένου, όπως είναι τα tweets, αλλά είναι δύσκολο να διατηρηθούν και να ενημερωθούν καθώς οι όροι της αργκό εξελίσσονται και αλλάζουν συνεχώς. Παραδείγματα τέτοιων είναι Emoticon Sentiment και το Sentiment140 Lexicon.

- **Λεξικά συναισθημάτων**. Αυτά τα λεξικά περιλαμβάνουν μια λίστα λέξεων που σχετίζονται με συγκεκριμένα συναισθήματα και έχουν σχεδιαστεί ειδικά για να αποτυπώνουν το συναισθηματικό περιεχόμενο των κειμένου. Έχουν την δυνατότητα να διακρίνουν τον τρόπο με τον οποίο ένα άτομο εκφράζει τα συναισθήματά του καθώς και την υποκειμενικότητα του. Συχνά χρησιμοποιούνται για να συμπληρώσουν τα γενικά λεξικά συναισθημάτων και να παρέχουν μια καλύτερη εικόνα του συναισθήματος που εκφράζεται σε ένα κείμενο. Παραδείγματα τέτοιων λεξικών είναι το NRC Emotion Lexicon και το WordEmotion lexicon.

- **Πολυγλωσσία λεξικά**. Αυτά τα λεξικά έχουν δημιουργηθεί για να επιτρέπουν την ανάλυση συναισθημάτων σε διαφορετικές γλώσσες και να παρέχουν βαθμολογίες συναισθημάτων για λέξεις και εκφράσεις σε διαφορετικές γλώσσες με διαφορετικά πολιτιστικά πλαίσια. Σε

περιπτώσεις πολυγλωσσιών κειμένων, όπως για παράδειγμα τα δεδομένα των μέσων κοινωνικής δικτύωσης, είναι χρήσιμα όταν υπάρχει περιορισμένη κάλυψη και ακρίβεια για συγκεκριμένες γλώσσες. Παραδείγματα τέτοιων λεξικών είναι το λεξικό SentiWordNet Multilingual και το λεξικό MultiLexNorm.

Συνοψίζοντας τα προηγούμενα, προκύπτει ότι η επιλογή του κατάλληλου λεξικού για την ανάλυση συναισθήματος εξαρτάται τόσο από τον τομέα της εφαρμογής της όσο και από το πλαίσιο στο οποίο χρησιμοποιείται. Συχνά, χρησιμοποιείται ένας συνδυασμός διαφορετικών λεξικών για την επίτευξη καλύτερων αποτελεσμάτων στην ανάλυση συναισθήματος.

2.2.4 Δημιουργία λεξικών συναισθημάτων.

Η δημιουργία λεξικών συναισθημάτων αποτελεί ένα κρίσιμο βήμα στη χρήση της τεχνικής lexicon-based. Υπάρχουν διάφορες προσεγγίσεις για τη δημιουργία των απαραίτητων λεξικών ,οι κυριότερες από τις οποίες είναι οι εξής:

- **Μέθοδος βασισμένη σε λεξικό (dictionary-based)**: Η μέθοδος αυτή περιλαμβάνει τη δημιουργία ενός λεξικού συναισθήματος κάνοντας χρήση ενός λεξικού. Η διαδικασία ξεκινάει χρησιμοποιώντας ένα μικρό σύνολο λέξεων συναισθήματος που έχει συλλεχτεί χειροκίνητα και έχει εκχωρημένο την συναισθηματική πολικότητα. Στη συνέχεια, το σύνολο αυτό επεκτείνεται με την αναζήτηση συνωνύμων και αντωνυμιών σε ένα διαδικτυακό λεξικό όπως το WordNet. Η διαδικασία αυτή επαναλαμβάνεται έως ότου δεν είναι δυνατόν να βρεθούν άλλες νέες λέξεις. Η αντιστοίχιση μεταξύ των λέξεων και των συναισθηματικών τους πολικοτήτων μπορεί να γίνει είτε χειροκίνητα είτε χρησιμοποιώντας αυτοματοποιημένες μεθόδους όπως είναι η μηχανική μάθηση. Τα λεξικά που βασίζονται σε αυτή την μέθοδο προσαρμόζεται ευκολότερα σε συγκεκριμένους τομείς ή γλώσσες και ερμηνεύονται χωρίς ιδιαίτερη δυσκολία. Ωστόσο, υστερούν στον εντοπισμό των σημασιολογικών αποχρώσεων και δεν αντικατοπτρίζουν τις αλλαγές στη χρήση της γλώσσας με την πάροδο του χρόνου.

- **Μέθοδος των σωμάτων κείμενων (corpus-based)**: Η μέθοδος αυτή χρησιμοποιεί μεγάλες συλλογές κειμένων ή σωμάτων για την δημιουργία των συναισθηματικών λεξικών. Όπως

διαπιστώνεται από την βιβλιογραφία, υπάρχουν δύο διαφορετικοί τρόποι που εφαρμόζεται αυτή η τεχνική. Ο πρώτος τρόπος χρησιμοποιεί μια γνωστή λίστα συναισθηματικών λέξεων και η μέθοδος προσπαθεί να ανακαλύψει άλλες συναισθηματικές λέξεις βάση της λίστας, λαμβάνοντας υπόψιν της το συντακτικό και τα συμφραζόμενα. Ο δεύτερος τρόπος αφορά την προσαρμογή ενός λεξικού γενικής χρήσης σε ένα νέο λεξικό συναισθημάτων χρησιμοποιώντας ένα κείμενο του αντίστοιχου τομέα [17]. Το επόμενο βήμα, ανεξαρτήτως τρόπου, είναι η εκχώρηση της πολικότητας συναισθήματος με την χρήση αλγορίθμων μηχανικής μάθησης, ενώ το τελευταίο βήμα είναι η αξιολόγηση του λεξικού χρησιμοποιώντας στατιστικές μεθόδους και μέτρα αξιολόγησης. Τα λεξικά τα οποία δημιουργούνται με αυτή την μέθοδο προσαρμόζονται τόσο στον τομέα όσο και στις αλλαγές στη χρήση της γλώσσας με την πάροδο του χρόνου. Ωστόσο, απαιτούν μεγάλο όγκο επισημασμένων δεδομένων για την δημιουργία τους και την εκπαίδευση των αντίστοιχων αλγορίθμων που προσδιορίζουν την πολικότητα.

- **Μέθοδος βασισμένη σε γλωσσικούς κανόνες (linguistic rules-based)**: Η μέθοδος αυτή κάνει χρήση γλωσσικών κανόνων που έχουν αναπτυχθεί με σκοπό να αποτυπώνονται τα γλωσσικά πρότυπα της φυσικής γλώσσας και να αναγνωρίζουν τα συναισθήματα που συνδέονται με συγκεκριμένες λέξεις ή φράσεις στο κείμενο. Αυτοί οι κανόνες βασίζονται στη δομή και τη σύνταξη της γλώσσας και συνήθως περιέχουν και κανόνες που επικεντρώνονται σε συγκριμένα γλωσσικά χαρακτηριστικά που πιθανόν να επηρεάσουν την πολικότητα μιας λέξης ή μιας φράσης. Τα βασικά πλεονεκτήματα της μεθόδου είναι ότι μπορεί να εφαρμοστεί σε γλώσσες και τομείς για τους οποίους δεν υπάρχουν λεξικά συναισθημάτων και επιπλέον παρέχει αξιόπιστα αποτελέσματα σε κείμενα που ακολουθούν συγκεκριμένες δομές. Όμως στα αρνητικά της είναι η δυσκολία στον ορισμό ενός ολοκληρωμένου συνόλου κανόνων, η αδυναμία της στον εντοπισμό συναισθημάτων σε άτυπη ή μη τυποποιημένη γλώσσα και η συνεχή επιμέλεια των κανόνων.

2.2.5 Προκλήσεις της τεχνικής με χρήση λεξικών.

Η τεχνική lexicon-base εφαρμόζεται σε διάφορους τομείς και δεν απαιτεί ιδιαίτερη γνώση ή εξειδίκευση πάνω σε έναν συγκεκριμένο αντικείμενο. Παρόλο που η μέθοδος αυτή είναι σχετικά εύκολη στην εφαρμογή της και δεν απαιτεί μεγάλο αριθμό δεδομένων για την

εκπαίδευσή της, παρουσιάζει ορισμένες αδυναμίες , όπως είναι αδυναμία της κάλυψης της φυσικής γλώσσας και της κατανόησης των εκφράσεων που σχετίζονται με αρνητικά συναισθήματα. Στην παρούσα ενότητα αναλύονται λεπτομερώς οι προαναφερθείσες αδυναμίες.

Η αντιμετώπιση της άρνησης είναι η κύρια αδυναμία αυτής της μεθόδου. Συγκεκριμένα, η άρνηση αναφέρεται στο φαινόμενο κατά το οποίο μια λέξη αναιρεί τη σημασία μιας άλλης λέξης, αντιστρέφοντας έτσι τη πολικότητα της έκφρασης ή της λέξης. Για παράδειγμα, η φράση "δεν είμαι ευτυχισμένος" εκφράζει αρνητικό συναίσθημα, αλλά λόγω της λέξης "ευτυχισμένος" μπορεί λανθασμένα να παρερμηνευτεί ως θετική. Επομένως, η σωστή διαχείριση της άρνηση είναι σημαντική και για αυτό έχει προταθεί η χρήση μεθόδου που βάσει κανόνων ανιχνεύει λέξεις που υποδηλώνουν άρνηση ,όπως είναι το "μη" και το "ποτέ". Η μέθοδο αυτή είναι αποτελεσματική για απλές περιπτώσεις, αλλά αποτυγχάνει να συλλάβει πιο σύνθετες περιπτώσεις άρνησης. Ένα παράδειγμα πιο σύνθετής άρνησης είναι η φράση "Δεν μπορώ να πιστέψω πόσο καλό είναι αυτό το προϊόν". Παρόλο που η λέξη "δεν μπορώ" δηλώνει άρνηση, το συναίσθημα της έκφρασης "πόσο καλό είναι αυτό το προϊόν" παραμένει θετικό. Προκειμένου να αντιμετωπιστούν οι περιορισμοί της προηγούμενης μεθόδου, έχουν αναπτυχθεί πιο προηγμένες τεχνικές επεξεργασίας φυσικής γλώσσας όπως είναι η χρήση αλγορίθμων μηχανικής μάθησης. Με την χρήση τους τα μοντέλα εκπαιδεύονται σε μοτίβα άρνησης από δεδομένα με ετικέτες και είναι σε θέση να τα αναγνωρίζουν και να διαχειρίζονται την πολυπλοκότητα σε νέα δεδομένα. Η προσέγγιση αυτή έχει αποδειχθεί αποτελεσματική σε πιο σύνθετες περιπτώσεις άρνησης, όπου η απλή αναζήτηση λέξεων που υποδηλώνουν άρνηση δεν είναι αρκεί.

Η κάλυψη της φυσικής γλώσσας είναι εξίσου σημαντική αδυναμία της μεθόδου lexicon-base. Αναλυτικότερα, τα λεξικά συναισθήματος κατασκευάζονται από ανθρώπους οι οποίοι αποδίδουν βαθμολογίες συναισθήματος σε λέξεις και φράσεις με βάση τις υποκειμενικές τους κρίσεις. Ως αποτέλεσμα, τα λεξικά συναισθήματος δεν περιλαμβάνουν όλες τις λέξεις και φράσεις που εκφράζουν το συναίσθημα, ιδίως σε σπάνιες γλώσσες ή γλώσσες που το συντακτικό και η γραμματική τους δομή είναι πολύπλοκη. Για παράδειγμα, αν κάποια λεξικά συναισθήματος αναπτυχθούν χρησιμοποιώντας την μέθοδος των σωμάτων κείμενων σκοπού και τα κείμενα που χρησιμοποιούνται είναι άρθρα ειδήσεων ή αναρτήσεις στα μέσα κοινωνικής

δικτύωσης, τότε είναι πιθανό να μην περιλαμβάνουν τύπους εκφράσεων συναισθήματος, όπως είναι οι ιδιωματισμοί ή η αργκό. Επίσης κάποια λεξικά συναισθήματος έχουν αναπτυχθεί για συγκεκριμένους τομείς, όπως κριτικές προϊόντων ή πολιτικές ομιλίες, με αποτέλεσμα να είναι ακατάλληλά για την ανάλυση συναισθήματος σε άλλους τομείς. Αυτός ο περιορισμός οδηγήσει σε ανακριβή ή ελλιπή ανάλυση συναισθήματος, λόγω του ότι ορισμένες εκφράσεις συναισθήματος δεν καταγράφονται από το λεξικό. Για να ξεπεραστεί το πρόβλημα της περιορισμένης κάλυψης, οι ερευνητές έχουν αναπτύξει διάφορες μεθόδους η μια εκ των οποίων είναι η ανάπτυξή λεξικών που είναι προσαρμοσμένα για συγκεκριμένους τομείς. Ωστόσο, η πιο διαδιδομένη τεχνική είναι αυτή της επέκτασης των λέξεων συναισθήματος προσθέτοντάς συνώνυμα, αντωνυμία και άλλες συναφείς λέξεις για την αύξηση της κάλυψής τους.

2.3 Ανάλυση συναισθήματος με μηχανική μάθηση

2.3.1 Εισαγωγή

Η μηχανική μάθηση είναι ένα πεδίο έρευνας που συνδυάζει τρεις διαφορετικές επιστημονικές περιοχές: τη στατιστική, την τεχνητή νοημοσύνη και την επιστήμη των υπολογιστών [18]. Η μηχανική μάθηση χρησιμοποιεί μεθόδους που στοχεύουν στην εκμάθηση των μηχανών με χρήση γνωστών δεδομένων και στην βελτίωση της απόδοσης τους με την πάροδο του χρόνου. Στόχος της είναι η αυτόματη αναγνώριση μοτίβων στα δεδομένα, η πρόβλεψη και η ταξινόμηση νέων δεδομένων καθώς και η λήψη αποφάσεων κάτω από συγκεκριμένες συνθήκες. Η ανάλυση συναισθήματος αποτελεί ένα σημαντικός τομέας εφαρμογής για τη μηχανική μάθησης. Αυτό οφείλεται στο γεγονός ότι οι αλγόριθμοι που χρησιμοποιούνται διευκολύνουν τη εκμάθηση των μηχανών και μέσω της οποίας μπορούν να επεξεργαστούν μεγάλα σύνολα δεδομένων και να εντοπίσουν σύνθετα μοτίβα στα κείμενα. Στην παρούσα ενότητα περιγράφονται οι διάφοροι τύποι μηχανικής μάθησης και οι βασικοί αλγόριθμοι που χρησιμοποιούνται στο υπόλοιπο της παρούσας εργασίας.

Στο πλαίσιο της συναισθηματικής ανάλυσης, οι τεχνικές μηχανικής μάθησης που εφαρμόζονται είναι η επιβλεπόμενη μάθηση (supervised learning),η μη επιβλεπόμενη μάθηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning).

2.3.1.1 Επιβλεπόμενη μάθηση

Η επιβλεπόμενη μηχανική μάθηση αποτελεί μία από τις πλέον διαδεδομένες τεχνικές της μηχανικής μάθησης, η οποία χρησιμοποιείται για την πρόβλεψη ενός αποτελέσματος δεδομένου μιας συγκεκριμένης εισόδου. Ειδικότερα, ο αλγόριθμος λαμβάνει ως είσοδο ένα σύνολο δεδομένων, το οποίο αποτελείται από δεδομένα με ετικέτες ή κατηγορίες μαζί με τις επιθυμητές εξόδους (σύνολο εκπαίδευσης). Με βάση αυτό το σύνολο, ο αλγόριθμος επιδιώκει να ανακαλύψει έναν γενικό κανόνας, ο οποίος θα αντιστοιχεί τα χαρακτηριστικά της εισόδου με τις ετικέτες εξόδου. Στην ουσία αυτός ο κανόνας είναι μια συνάρτηση που απεικονίζει τα δεδομένα του συνόλου εκπαίδευσης στις ήδη γνωστές τους εξόδου. Κατά την διάρκεια της εκπαίδευσης, η συνάρτηση γενικεύεται έτσι ώστε να είναι σε θέση να πραγματοποιεί προβλέψεις για νέα δεδομένα που δεν περιέχουν ετικέτες. Η αξιολόγηση για τη ικανότητα του αλγορίθμου να γενικεύει την πρόβλεψή του σε νέα δεδομένα γίνεται μέσω του συνόλου

επαλήθευσης στο οποίο η επιθυμητή έξοδος δεν είναι γνωστή εξαρχής (σύνολο ελέγχου). Η αξιολόγηση αυτή είναι σημαντική για την επιλογή των κατάλληλων παραμέτρων του αλγορίθμου και τη βελτίωση της απόδοσής του στο μέλλον.

Η επιβλεπόμενη μάθηση χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης (classification) και σε προβλήματα παλινδρόμησης (regression). Ακολουθεί μια συνοπτική περιγραφή αυτών των προβλημάτων.

1. Κατηγοριοποίηση

Η κατηγοριοποίηση αποτελεί μια σημαντική εργασία στον χώρο της μηχανικής μάθησης. Στο γενικό πλαίσιο, το πρόβλημα περιλαμβάνει την αντιστοίχιση ενός αντικειμένου σε μία ή περισσότερες προκαθορισμένες κατηγορίες. Οι αλγόριθμοι κατηγοριοποίησης λειτουργούν μαθαίνοντας μια συνάρτηση στόχου (target function) f, χρησιμοποιώντας ένα σύνολο εκπαίδευσης. Η f είναι το μοντέλο που δημιουργεί ο αλγόριθμος προκυμμένου να απεικονίζει κάθε σύνολο γνωρισμάτων x σε μια από τις προαποφασισμένες ετικέτες-κλάσεις y. Παραδείγματα τέτοιων αλγορίθμων κατηγοριοποίησης είναι η λογιστική παλινδρόμηση, ο απλοϊκός Naive Bayes, ο αλγόριθμο SVM, τα δέντρα απόφασης και ο αλγόριθμο k-nearest neighbor.

Η κατηγοριοποίηση μπορεί να είναι δυαδική, όπου τα δεδομένα χωρίζονται σε δύο κλάσεις και ο αλγόριθμος εκπαιδεύεται προκειμένου υλοποιεί την διάκριση των ετικετών βάση αυτών των δύο κλάσεων. Επίσης μπορεί να είναι πολλαπλή, όπου τα δεδομένα χωρίζονται σε περισσότερες από δύο κλάσεις και ο αλγόριθμο ομοίως εκπαιδεύεται στο αντίστοιχο πλήθος κλάσεων. Στα περισσότερα προβλήματα επιλέγεται το πλήθος των κλάσεων εξόδου να είναι μικρό και διακριτό και σπανίως ξεπερνά τις δυο. Ένα κλασικό παράδειγμα ταξινόμησης είναι το spam filtering, στο οποίο το μοντέλο δέχεται σαν είσοδο email, μαθαίνει μέσω των χαρακτηριστικών να αναγνωρίζει αν είναι spam ή non-spam, και τα ταξινομεί στις αντίστοιχες κλάσεις ανάλογα το περιεχόμενο τους.

2. Παλινδρόμηση

Η παλινδρόμηση αποτελεί μία μέθοδο επιβλεπόμενης μάθησης, η οποία εφαρμόζεται κυρίως σε προβλήματα όπου η μεταβλητή εξόδου λαμβάνει συνεχείς τιμές. Στη διαδικασία

αυτή, ο αλγόριθμος αναζητείται μία συνάρτηση, η οποία αντιστοιχίζει τα δεδομένα εισόδου σε μία συνεχόμενη μεταβλητή εξόδου, με στόχο την παραγωγή πιο ακριβών προβλέψεων για τις μελλοντικές εξόδους. Παραδείγματα τέτοιων αλγορίθμων είναι η γραμμική παλινδρόμηση και τα δέντρα απόφασης. Ένα παράδειγμα παλινδρόμησης είναι η ανάλυση δεδομένων ασθενών για να γίνει πρόβλεψη της πιθανότητας ο ασθενής να επιβιώσει ή να νοσήσει .

2.3.1.2 Μη επιβλεπόμενη μάθηση

Εκτός από την επιβλεπόμενη μηχανική μάθηση που περιεγράφηκε στην 2.3.1..1, αναφέρθηκε και η ύπαρξη ενός άλλου τύπου μηχανικής μάθησης: η μη επιβλεπόμενη μηχανική μάθηση. Στη μάθηση χωρίς επίβλεψη, ο αλγόριθμος γνωρίζει μόνο τα δεδομένα εισόδου και ανακαλύπτει κρυμμένα μοτίβα ή ομαδοποιήσεις που υπάρχουν μέσα σε αυτά. Δηλαδή, ο στόχος είναι η εξαγωγή γνώσης από τα δεδομένα εισόδου χωρίς να υπάρχει εκ των προτέρων γνώση των ετικετών και των κατηγοριών των δεδομένων. Επομένως, με την έλλειψη των επισημασμένων δεδομένων, είναι σαφές ότι δεν είναι δυνατή η εκτίμηση πιθανών σφαλμάτων. Αυτή είναι η σημαντική διαφορά μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης.

Όπως και στην επιβλεπόμενη μάθηση, έτσι και στη μη επιβλεπόμενη μάθηση, υπάρχουν δύο υποκατηγορίες προβλημάτων, που αναφέρονται ως μείωση διαστάσεων (dimensionality reduction) και συσταδοποίηση (clustering).

1. Μείωση διαστάσεων

Η μείωση των διαστάσεων αναφέρεται σε μια τεχνική που αποσκοπεί στην αναπαράσταση ενός συνόλου διανυσμάτων υψηλής διάστασης σε έναν χώρο χαμηλής διάστασής. Με άλλα λόγια, στοχεύει στη μείωση του αριθμού των χαρακτηριστικών εισόδου, διατηρώντας παράλληλα τις σημαντικές πληροφορίες του συνόλου δεδομένων. Η τεχνική αυτή εφαρμόζεται συνήθως στην προ επεξεργασία των δεδομένων για την απομάκρυνση ανεπιθύμητου θορύβου. Σε αυτό το στάδιο, συμβάλλει στη μείωση των αποστάσεων μεταξύ των δεδομένων στον χώρο, στην ελάττωση του όγκου τους και στη βελτίωση της υπολογιστικής αποδοτικότητας των αλγορίθμων. Παραδείγματα τέτοιων αλγορίθμων είναι ανάλυση κυρίων συνιστωσών και η γραμμική διαχωριστική ανάλυση.

2. Συσταδοποίηση

Η μέθοδος συσταδοποίησης είναι η τεχνική εξόρυξης δεδομένων κατά την οποία διαχωρίζεται ένα σύνολο μη επισημασμένων δεδομένων σε υποομάδες ή αλλιώς συστάδες μέσω ενός αλγορίθμου. Ο σκοπός της ομαδοποίησης είναι τα αντικείμενα εντός της ίδια ομάδας να έχουν κοινά χαρακτηριστικά και να παρουσιάζουν μεγαλύτερη ομοιότητα μεταξύ τους, συγκριτικά με τα αντικείμενα που ανήκουν σε μια άλλη ομάδα. Επομένως, τα αντικείμενα εντός μιας συστάδας παρουσιάζουν μία σχετική ομοιογένειά ενώ τα αντικείμενα που ανήκουν σε διαφορετικές συστάδες διαφέρουν όσο το δυνατό περισσότερο. Ο αριθμός των συστάδων στις οποίες χωρίζονται τα δεδομένα καθώς και το περιεχόμενό τους δεν είναι γνωστά εκ των προτέρων. Για το λόγο αυτό η τεχνική της συσταδοποίησης κατατάσσεται στην κατηγορία της μη επιβλεπόμενης μάθησης. Παραδείγματα τέτοιων αλγορίθμων είναι ο SVD, ο BIRCH και ο DBSCAN.

2.3.1.3 Ενισχυτική μάθηση

Η ενισχυτική μάθηση είναι η τελευταία υποκατηγορία της μηχανικής μάθησης που θα αναφερθεί και ασχολείται με προβλήματα λήψης αποφάσεων σε δυναμικά περιβάλλοντα. Ειδικότερα ασχολείται με την εκπαίδευση ενός πράκτορα (agent) μέσω της αλληλεπίδρασής ενός με το περιβάλλον για την βελτιστοποίηση της συμπεριφοράς του. Η αλληλεπίδραση ενός πράκτορα με το περιβάλλον είναι συνεχής και λαμβάνει χώρα σε διακριτά χρονικά διαστήματα. Ο πράκτορας εκτελεί διάφορες ενέργειες και συμπεριφορές στις οποίες το περιβάλλον ανταποκρίνεται με ανατροφοδότηση στη μορφή ανταμοιβών ή τιμωριών. Στόχος του πράκτορα είναι να μάθει μια βέλτιστη στρατηγική που να μεγιστοποιεί τη συνολική του ανταμοιβή με την πάροδο του χρόνου. Ένα παράδειγμα ενισχυτικής μάθησης είναι το σκάκι ,όπου ο πράκτορας αποφασίζει για διάφορες κινήσεις ώστε να κερδίσει ή να χάσει [19].

2.3.2 Μοντέλα επιβλεπόμενης μάθησης για προβλήματα ταξινόμηση

Η αναγνώριση του συναισθήματος των tweets είναι μια εργασία της επιβλεπόμενης μηχανικής μάθησης, καθώς απαιτούνται επισημασμένα δεδομένα για την εκπαίδευση του μοντέλου. Η ανάλυση του συναισθήματος αντιμετωπίζεται ως ένα πρόβλημα ταξινόμησης κειμένου, όπου ο στόχος είναι να αναγνωριστεί το συναίσθημα ενός tweet και να ταξινομηθεί σε μία από τις προκαθορισμένες κατηγορίες. Η αποτελεσματικότητα αυτής της διαδικασίας

εξαρτάται από την ποιότητα των επισημασμένων δεδομένων, την τεχνική προ επεξεργασίας που χρησιμοποιείται και τον αλγόριθμο που επιλέγεται. Υπάρχουν διάφοροι αλγόριθμοι που μπορούν να χρησιμοποιηθούν για την ανάλυση συναισθήματος των tweets, ενδεικτικά αναφέρονται ο Naive Bayes και ο Support Vector Machine. Στη παρούσα ενότητα περιγράφονται οι αλγόριθμοι που θα χρησιμοποιηθούν στην συνέχεια της εργασίας.

2.3.2.1 Naive Bayes classifier

Ο αλγόριθμος Naive Bayes ανήκει στην κατηγορία των πιθανολογικών μοντέλων μηχανικής μάθησης και είναι δημοφιλής για την επίλυση προβλημάτων ταξινόμησης κειμένου. Βασίζεται στο θεώρημα του Bayes βάσει του οποίου, ο αλγόριθμος είναι ικανός να προβλέψει την κατηγορία-ετικέτα στην οποία ανήκει ένα δοσμένο κείμενο δεδομένων των χαρακτηριστικών του (π.χ. λέξεις). Ο όρος "πιθανολογικός" υπογραμμίζει τη δυνατότητα του να υπολογίζει την πιθανότητα της κάθε κατηγορίας για δοσμένο κείμενο και να επιλέγει την κατηγορία που έχει την μεγαλύτερη πιθανότητα.

Η βασική ιδέα του αλγορίθμου είναι ότι δοθέντος ενός έγγραφου που αντιπροσωπεύεται από ένα σύνολο χαρακτηριστικών (features), $X = \{x_1, x_2, ... ,x_n\}$, και ενός σύνολού κατηγοριών, $C = \{c_1, c_2, ...,c_k\}$, μπορεί να υπολογιστεί η την εκ των υστέρων πιθανότητα κάθε κατηγορίας δεδομένων των χαρακτηριστικών του εγγράφου. Αυτό γίνεται με βάση το θεώρημα του Bayes, σύμφωνα με το οποίο η εκ των υστέρων πιθανότητα μπορεί να υπολογιστεί ως εξής:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (1)$$

όπου P(X|C) είναι η πιθανότητα των χαρακτηριστικών δεδομένης της κατηγορίας, P(C) είναι η εκ των προτέρων πιθανότητα της κατηγορίας και P(X) είναι η πιθανότητα των χαρακτηριστικών του εγγράφου.

Ο αλγόριθμος υποθέτει υπό όρους την ανεξαρτησία μεταξύ των χαρακτηριστικών και της κλάσης κατηγοριοποίησης, δηλαδή ότι τα χαρακτηριστικά-λέξεις σε ένα έγγραφο είναι ανεξάρτητα μεταξύ τους. Αυτό σημαίνει ότι η P(X) είναι σταθερή σε όλες τις κατηγορίες και δεν επηρεάζει την απόφαση για την ταξινόμηση. Άρα η σχέση της πιθανότητα των χαρακτηριστικών δεδομένης της κατηγορίας παραγοντοποιείται ως :

$$P(X|C) = P(x_1|C) * P(x_2|C) * \ldots * P(x_n|C) \quad (2)$$

όπου $P(x_i|C)$ είναι η πιθανότητα παρατήρησης του i-οστού χαρακτηριστικού δεδομένης της κατηγορίας. Για την εκτίμηση της $P(x_i|C)$, δηλαδή της πιθανότητας εμφάνισης κάθε χαρακτηριστικού σε κάθε κατηγορία, χρησιμοποιούνται τα δεδομένα εκπαίδευσης μετρώντας τον αριθμό των φορών που εμφανίζεται κάθε χαρακτηριστικό σε κάθε κατηγορία. Η τελική απόφαση ταξινόμησης λαμβάνεται με την επιλογή της κατηγορίας με την υψηλότερη εκ των υστέρων πιθανότητα.

Παρόλο που η υπόθεση της ανεξαρτησίας μεταξύ των χαρακτηριστικών σπάνια ισχύει στην πραγματική ζωή, ο Naive Bayes έχει αποδειχθεί αποτελεσματικός σε πολλά προβλήματα ταξινόμησης κειμένου, ειδικά όταν ο αριθμός των χαρακτηριστικών είναι μεγάλος και τα δεδομένα εκπαίδευσης περιορισμένα. Αυτό οφείλεται στην υπολογιστική του απλότητας η οποία τον καθιστά εύκολο στη χρήση και γρήγορο στην υλοποίησή του. Επιπλέον, ο Naive Bayes δεν απαιτεί υψηλή μνήμη που σημαίνει ότι μπορεί να εκπαιδευτεί σε σύντομο χρονικό διάστημα.

Ανάλογα με την υπόθεση που γίνεται για την κατανομή που ακολουθούν τα χαρακτηριστικά, δημιουργούνται και διαφορετικοί Naive Bayes ταξινομητές, για την επίλυση των διαφόρων προβλημάτων κατηγοριοποίησης. Μερικοί από αυτούς είναι: ο Multinomial , Complement και ο Bernoulli Naive Baye. Στην συγκεκριμένη διπλωματική εφαρμόζονται και εξετάζονται τόσο ο multinomial όσο και ο complement ταξινομητή.

- **Multinomial Naive Bayes**

Ο Multinomial Naive Bayes (MNB) είναι μια παραλλαγή του αλγορίθμου Naive Bayes που έχει σχεδιαστεί ειδικά για εργασίες ταξινόμησης κειμένου. Ο MNB διαφέρει από τον απλό ταξινομητή Naive Bayes γιατί χρησιμοποιεί διακριτά χαρακτηριστικά, τα οποία αντιπροσωπεύουν τις συχνότητα εμφάνισης των λέξεων στο έγγραφο και θεωρεί ότι η κατανομή της πιθανότητας των χαρακτηριστικών ενός εγγράφου δεδομένης μιας συγκεκριμένης κατηγορίας είναι πολυωνυμική.

Το πρώτο βήμα στον αλγόριθμο MNB είναι η αναπαράσταση κάθε εγγράφου ως διάνυσμα συχνοτήτων των λέξεων. Έστω, λοιπόν, D το σύνολο των εγγράφων, V το λεξιλόγιο,

δηλαδή το σύνολο των διακριτών λέξεων, και di το i-οστό έγγραφο. Τότε, το d_i αναπαρίσταται ως ένα διάνυσμα x_i , όπου x_{ij} είναι ο αριθμός των φορών που η j-οστή λέξη του λεξιλογίου εμφανίζεται στο d_i. Επόμενο βήμα είναι η εκτίμηση της πιθανότητας κάθε κατηγορίας-ετικέτας, δεδομένου του εγγράφου, κάνοντας χρήση του θεωρήματος Bayes. Έστω C το σύνολο των πιθανών κατηγοριών με c_j να είναι η j-οστή κατηγορία. Τότε, η πιθανότητα του c_j δεδομένου του εγγράφου d_i γράφεται ως εξής:

$$P(c_j|d_i) = \frac{P(d_i|c_j)P(c_j)}{P(d_i)} \quad (3)$$

όπου $P(d_i|c_j)$ είναι η πιθανότητα παρατήρησης του εγγράφου d_i δεδομένης της ετικέτας c_j, $P(c_j)$ είναι η εκ των προτέρων πιθανότητα της ετικέτας c_j και $P(d_i)$ είναι η πιθανότητα παρατήρησης του εγγράφου d_i. Η εκ των προτέρων πιθανότητα της ετικέτας c_j εκτιμάται ως το ποσοστό των εγγράφων στο σύνολο εκπαίδευσης που έχουν την ετικέτα c_j. Η κατανομή των κειμένων δεν είναι δυνατό να εκτιμηθεί απευθείας και για αυτό γίνεται η υπόθεση ότι τα κείμενα αποτελούνται από τις λέξεις οι οποίες κατανέμονται ανεξάρτητα. Η πιθανότητα παρατήρησης του εγγράφου d_i δεδομένης της ετικέτας c_j εκτιμάται ως :

$$P(d_i|c_j) = \prod_j P(x_{ij}|c_j) \quad (4)$$

όπου x_{ij} είναι η συχνότητα της j-οστής λέξης στο λεξιλόγιο στο d_i και $P(x_{ij}|c_j)$ είναι η πιθανότητα παρατήρησης x_{ij} εμφανίσεων της j-οστής λέξης στο έγγραφό d_i δεδομένου ότι ανήκει στην κατηγορία c_j. Για την εκτίμηση της πιθανότητα $P(x_{ij}|c_j)$ υπάρχει ο τύπος:

$$P(x_{ij}|c_j) = \frac{N_{xc} + a}{N_c + a * |V|} \quad (5)$$

όπου N_{xc} είναι ο αριθμός των φορών που η j-οστή λέξη εμφανίζεται σε έγγραφα με ετικέτα c_j, N_c είναι ο συνολικός αριθμός όλων των λέξεων στα έγγραφα της κλάσης c_j και το |V| είναι το μέγεθος του λεξιλογίου. Υπάρχουν περιπτώσεις στο σύνολο εκπαίδευσης όπου μια λέξη μπορεί να μην εμφανιστεί σε κανένα από τα έγγραφα μιας συγκεκριμένης κλάσης. Το γεγονός αυτό οδηγήσει σε μηδενική πιθανότητα και επομένως στην μηδενίσει ολόκληρης της πιθανότητας του εγγράφου σε μια κλάση. Για να αποφευχθεί αυτό το πρόβλημα, προστέθηκε στον τύπο (5) το a το οποίο είναι μια παράμετρος που αποτρέπει την ύπαρξη μηδενικών πιθανοτήτων.

Συνήθως παίρνει μικρή θετική τιμή, όπως το 1, η οποία στην προστίθεται σε κάθε λέξη για κάθε κλάση.

Τελειώνοντας ο αλγόριθμος υπολογίζει τη πιθανότητα παρατήρησης του εγγράφου d_i κάνοντας χρήση το νόμο της ολικής πιθανότητας και έτσι προκύπτει:

$$P(d_i|x) = P(d_i) \prod_t P(x_{ij}|c_j)^{x_{ij}} \quad (6)$$

Το κριτήριο για την επιλογή της κατηγορίας είναι η υψηλότερη εκ των υστέρων πιθανότητα. Στην περίπτωση που οι όροι ενός εγγράφου δεν δείχνουν σαφώς σε ποια κατηγορία ανήκει σε σχέση με μια άλλη, τότε επιλέγεται η κατηγορία που έχει υψηλότερη πιθανότητα να ανήκει σε αυτήν, βάσει προηγούμενων εμπειριών ή στατιστικών δεδομένων.

- **Complement Naive Bayes**

Ο αλγόριθμος Complement Naive Bayes (CNB) είναι μια παραλλαγή του αλγορίθμου Naive Bayes που προτάθηκε για την αντιμετώπιση του προβλήματος των μη ισορροπημένων συνόλων δεδομένων. Μη ισορροπημένο σύνολο δεδομένων είναι αυτό στο οποίο μία ή περισσότερες κλάσεις έχουν πολύ μικρότερο αριθμό δεδομένων από τις άλλες. Σε τέτοιες περιπτώσεις, ο απλός αλγόριθμος Naive Bayes είναι μεροληπτικός ως προς τις κλάσεις με τα περισσότερα δεδομένα. Αυτό έχει ως αποτέλεσμα οι πιθανότητες για τα παρατηρούμενα χαρακτηριστικά σε κλάσεις με λίγα παραδείγματα να είναι χαμηλές και άρα μειώνεται σημαντικά η απόδοση της ταξινόμησης αυτών των κλάσεων.

Ο CNB, σε αντίθεση με τον απλό Naive Bayes, υπολογίζει την πιθανότητα ένα στοιχείο να ανήκει σε όλες τις κλάσεις και όχι μόνο σε μια συγκεκριμένη κλάση. Έστω ένα σύνολο εκπαίδευσης το οποίο αποτελείται από έγγραφα με τις ετικέτες τους και τα αντίστοιχα διανύσματα των χαρακτηριστικών τους. Ο CNB υπολογίζει την συμπληρωματική πιθανότητα για κάθε χαρακτηριστικό t και κάθε κλάση c. Η πιθανότητα αυτή στην ουσία είναι η μη εμφάνισης ενός χαρακτηριστικού σε μία δεδομένη κλάση και ορίζεται ως :

$$P'_c(t) = \frac{N - N_{ct} + a}{N - N_c + a * V} \quad (7)$$

όπου N_{ct} είναι ο αριθμός εμφανίσεων του χαρακτηριστικού t σε έγγραφα που ανήκουν στη κατηγορία c, N_c είναι ο αριθμός των χαρακτηριστικών σε έγγραφα της κλάση C, V είναι το μέγεθος του λεξιλόγιο και το N είναι ο συνολικός αριθμός χαρακτηριστικών στο σύνολο εκπαίδευσης. Το α είναι η παράμετρο που βοηθά στην αποφυγή του προβλήματος των μηδενικών πιθανοτήτων και βελτιώνει την ακρίβεια του ταξινομητή. Η συμπληρωματική πιθανότητάς έχει σχεδιαστεί έτσι ώστε να δίνει μεγαλύτερη βαρύτητα στα σπάνια χαρακτηριστικά σε κάθε κλάση, γεγονός που βελτιώνει την απόδοση ταξινόμησης για τις μικρότερες κλάσεις. Αυτό οφείλεται στον όρο $(N-N_{ct}+\alpha)$ στον αριθμητή ο οποίος «τιμωρεί» τα χαρακτηριστικά που εμφανίζονται συχνά στην κλάση C, ενώ ο όρος $(N_{ct}+\alpha V)$ στον παρονομαστή που κανονικοποιεί τις πιθανότητες.

Στη συνέχεια, ο αλγόριθμος εκτιμά στο σύνολο εκπαίδευσης, την εκ των προτέρων πιθανότητα κάθε κατηγορίας . Συγκεκριμένα η P'(C) υπολογίζεται ως:

$$P'(c) = \frac{N - N_c}{N} \quad (8)$$

Τέλος, ο αλγόριθμος πραγματοποιεί μια πρόβλεψη, δηλαδή δεδομένου ενός νέου εγγράφου με διάνυσμα χαρακτηριστικών υπολογίζει τη συμπληρωματική πιθανότητα της κλάσης C δεδομένου του διανύσματος χαρακτηριστικών x ως εξής:

$$P(c|x) = P'(c) \prod P'_c(t)^{x_t} \quad (9)$$

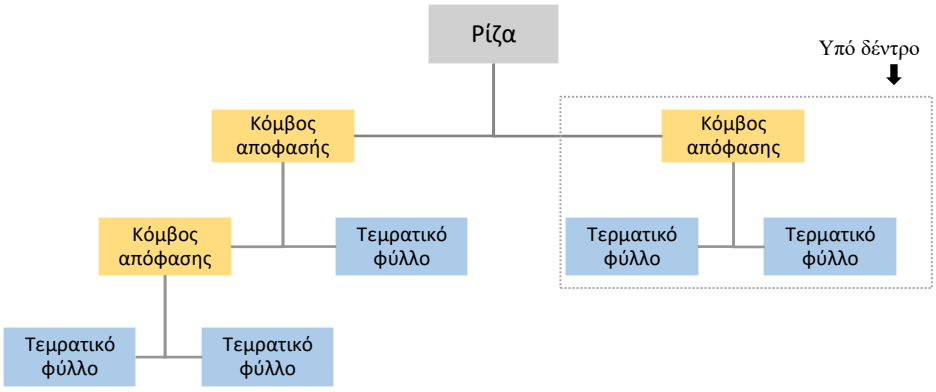
όπου x_t είναι η συχνότητα του χαρακτηριστικού t στο έγγραφο. Από διαφορές μελέτες έχει αποδειχτεί ότι το CNB (Complement Naive Bayes) είναι πιο αποτελεσματικό από τον κανονικό Naive Bayes για μη ισορροπημένα σύνολα δεδομένων, ωστόσο σε ισορροπημένα σύνολα δεδομένων ή όταν οι κλάσεις μειοψηφίας έχουν πολλά παραδείγματα, το CNB μπορεί να είναι λιγότερο αποδοτικός.

2.3.2.2 Decision tree classifier

Τα δέντρα αποφάσεων είναι ένας θεμελιώδης αλγόριθμος στον τομέα της μηχανικής μάθησης και μελετώνται ευρέως λόγω της απλότητας και της εύκολής ερμηνεία τους. Πρόκειται για έναν αλγόριθμο που διαμερίζει αναδρομικά τα δεδομένα εισόδου σε μικρότερα

υποσύνολα προκειμένου να εντοπίσει τα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν για την διαδικασία της κατηγοριοποίηση. Ωστόσο, ένα μεμονωμένο δέντρο απόφασης δεν χειρίζεται καλά το θόρυβο και ενέχει τον κίνδυνο για υπερβολική προσαρμογή των δεδομένων εκπαίδευσης, με αποτέλεσμα την μειωμένη απόδοση του όταν γενικεύεται. Για να αντιμετωπιστεί αυτός ο περιορισμός, έχουν αναπτυχθεί μέθοδοι που χρησιμοποιούν δέντρα απόφασης ως δομικά στοιχεία για την δημιουργία ισχυρότερων μοντέλων. Ο αλγόριθμος Random Forest και ο Gradient boosting είναι δύο αποτελεσματικές μέθοδοι που χρησιμοποιούν δέντρα αποφάσεων ως βάση και που εφαρμόστηκαν στην παρούσα εργασία. Για την κατανόηση αυτών των αλγορίθμων, είναι απαραίτητη η αναφορά της βασική θεωρία των δέντρων απόφασης η οποία αποτελεί το αντικείμενο της συγκεκριμένης υποενότητας.

Η δομή ενός δέντρου απόφασης κατασκευάζεται με χρήση ευριστικής διαίρεσης και ορίζεται από την ρίζα (root), τους κόμβους απόφασής (decision nodes) και τα τερματικά φύλλα (terminal nodes). Η ρίζα αντιπροσωπεύει ολόκληρο το σύνολο των δεδομένων και το δέντρο, έχοντας ως αυτό το σημείο έναρξής, αναπτύσσεται από εκεί και κάτω. Οι κόμβοι απόφασης αντιπροσωπεύουν τα σημεία απόφασης ή τις ερωτήσεις σχετικά με τα χαρακτηριστικά εισόδου και οι εξερχόμενοι κλάδοι αντιπροσωπεύουν τις πιθανές απαντήσεις σε αυτή την ερώτηση. Τα τερματικά φύλλα είναι τα σημεία ενός κλάδου που δεν χωρίζονται περαιτέρω και αντιπροσωπεύουν την τελική απόφαση ταξινόμησης για μια δεδομένη είσοδο. Συχνά σε ένα δέντρο απόφασης συναντιούνται υπό δέντρα (subtree) που προέρχεται από έναν συγκεκριμένο κόμβο απόφασης και αντιπροσωπεύουν ένα σύνολο αποφάσεων που οδηγούν σε ένα συγκεκριμένο αποτέλεσμα. Για την καλύτερη κατανόηση της δομής των δέντρων αποφάσεων παρουσιάζεται το Γράφημα 2.2.



Γράφημα 2.2 : Δομή των δέντρων αποφάσεων

Η διαδικασία δημιουργίας ενός δέντρου αποφάσεων γίνεται με την βοήθεια του συνόλου εκπαίδευσης και περιλαμβάνει τον επαναληπτικό διαχωρισμό των δεδομένων σε υποσύνολα με βάση τις τιμές των χαρακτηριστικών εισόδου. Έχοντας ως γνώμονα ένα συγκεκριμένο κριτήριο διαχωρισμού γίνεται η επιλογή του καλύτερου χαρακτηριστικού για διαχωρισμό σε κάθε κόμβο απόφασης. Μόλις επιλεγεί το χαρακτηριστικό και το σημείο διαίρεσης, ο χώρος εισόδου χωρίζεται σε δύο ή περισσότερα υποσύνολα, τα οποία στη συνέχεια χωρίζονται αναδρομικά χρησιμοποιώντας την ίδια διαδικασία. Για την απόφαση του διαχωρισμού ελέγχεται εάν το χαρακτηριστικό είναι μεγαλύτερο ή μικρότερο από την προκαθορισμένη σταθερά. Αυτή η διαδικασία συνεχίζεται μέχρι να επιτευχθεί ένα κριτήριο διακοπής και να φτάσει στον τερματικό κόμβο.

2.3.2.3 Random forest classifier

Ο Random Forest είναι μια μέθοδος που συνδυάζει πολλαπλά δέντρα αποφάσεων τα οποία συγχωνεύονται σε ένα "δάσος" προκυμμένου να πραγματοποιήσει μια πρόβλεψη. Κάθε δέντρο απόφασης δημιουργείται χρησιμοποιώντας ένα τυχαίο υποσύνολο των δεδομένων εκπαίδευσης με επανατοποθέτηση έτσι ώστε να δημιουργείται κάθε φορά ένα διαφορετικό σύνολο εκπαίδευσης. Αυτή η μέθοδος αναφέρεται ως "bagging" και ο λόγος που εφαρμόζεται είναι γιατί μειώνεται η διακύμανση των μεμονωμένων δέντρων και βελτιώνεται η απόδοσης τους όταν αυτά γενικεύονται. Στην συνέχεια, για κάθε υποσύνολο αναπτύσσεται ένα δέντρο όπου σε κάθε κόμβο επιλέγεται ένα τυχαίο υποσύνολο των χαρακτηριστικών εισόδου. Αυτό το τυχαίο υποσύνολο χρησιμεύει για τον διαχωρισμό του δέντρου των χαρακτηριστικών και συμβάλλει στη μείωση της συσχέτισης μεταξύ των μεμονωμένων δέντρων. Η μείωση της συσχέτισης οφείλεται στο γεγονός ότι τα δέντρα εστιάσουν σε διαφορετικές πτυχές των δεδομένων και έτσι αποφεύγεται η υπερβολική προσαρμογή σε συγκεκριμένα χαρακτηριστικά ή συνδυασμούς χαρακτηριστικών. Η παραπάνω διαδικασία επαναλαμβάνεται έως ότου δημιουργηθεί ένας συγκριμένος αριθμός δέντρων, ο οποίος μπορεί να ρυθμιστεί για να βελτιστοποιήσει την απόδοση ενός μοντέλου. Συνήθως, η αύξησή του αριθμού αυτού βελτιώνει την απόδοση, αλλά ταυτόχρονα αυξάνει τη υπολογιστική πολυπλοκότητα και το χρόνο εκπαίδευσης του αλγορίθμου.

Στο στάδιο της πρόβλεψης, για την ταξινόμηση ενός νέου κειμένου, το κείμενο εισόδου περνά από κάθε δέντρο απόφασης του δάσους. Κάθε δέντρο παράγει μια πρόβλεψή για την

κλάση και η κλάση που θα λαμβάνει τις περισσότερες ψήφους από τα μεμονωμένα δέντρα είναι αυτή που επιλέγεται ως η τελική πρόβλεψη για το κείμενο. Αυτή η προσέγγιση βελτιώνει την ακρίβειας αφού επηρεάζεται λιγότερο από το θόρυβο και τις ακραίες τιμές στα δεδομένα.

2.3.2.4 Gradient Boosting classifier

Ο ταξινομητής Gradient Boosting (GBC) είναι ένας αλγόριθμος εκμάθησης συνόλου (ensemble learning) που σημαίνει ότι συνδυάζονται πολλαπλά μοντέλα με σκοπό την δημιουργία ενός νέου πιο ακριβούς μοντέλου. Η βασική ιδέα πίσω από τον GBC είναι ο συνδυάζοντάς πολλαπλών αδύναμών ταξινομητών, για την δημιουργία ενός ισχυρότερου ταξινομητή. Αδύναμος ταξινομητής καλείται ένα απλό μοντέλο που λαμβάνει απόφαση με βάση ένα μόνο χαρακτηριστικό ή μια ιδιότητα των δεδομένων. Συγκεκριμένα, ο GBC λειτουργεί προσθέτοντας επαναληπτικά δέντρα απόφασης στο μοντέλο και κάθε νέο δέντρο που προστίθεται έχει εκπαιδευτεί ώστε να διορθώνει τα σφάλματα που έχουν γίνει στα προηγούμενα. Ο αλγόριθμος εκπαιδεύει πρώτα ένα απλό δέντρο αποφάσεων με βάση τα δεδομένα εισόδου. Έπειτα υπολογίζει τα υπολειπόμενα σφάλματα του τρέχοντος μοντέλου και εκπαιδεύει το νέο δέντρο απόφασης ώστε να συμπεριλαμβάνει τα υπολειπόμενα σφάλματα. Στη συνέχεια, το νέο δέντρο προστίθεται στο μοντέλο και η διαδικασία επαναλαμβάνεται έως ότου επιτευχθεί το κριτήριο διακοπής.

Η μαθηματική θεωρία πίσω από το Gradient Boosting είναι αρκετά περίπλοκη ωστόσο αναφέρεται συνοπτικά στα παρακάτω βήματα:

1. Έστω X ένας πίνακας μεταβλητών εισόδου και y ένα διάνυσμα μεταβλητών στόχου. Ο στόχος είναι η εκμάθηση μιας συνάρτησης πρόβλεψης F(x) που ελαχιστοποιεί τη συνάρτηση απώλειας L(y, F(x)).
2. Πραγματοποιείται η αρχικοποίηση του μοντέλου με μια σταθερά και ο λόγος είναι η ύπαρξη ενός λογικού σημείο εκκίνησης για τον αλγόριθμο. Επομένως :

$$F_0 = \arg min \sum_i L(y_i, c) \quad (10)$$

όπου c είναι μια σταθερά.

3. Για κάθε επανάληψη t = 1, 2, ..., N, προσαρμόζεται ένα δέντρο απόφασης στην αρνητική κλίση της συνάρτησης απώλειας σε σχέση με την τρέχουσα συνάρτηση πρόβλεψης, δηλαδή:

$$r_i = -\left[\frac{\partial L\left(y_i, F(x_i)\right)}{\partial F(x_i)}\right]_{F(x)=F_{t-1}(x)} \quad (11)$$

όπου r_i είναι η αρνητική κλίση για τη i-οστη παρατήρηση από το σύνολο εκπαίδευσης και $F_{t-1}(x)$ είναι η τρέχουσα συνάρτηση πρόβλεψης στην επανάληψη t-1. Στην ουσία το r_i είναι ένα μέτρο του πόσο απέχει η τρέχουσα συνάρτηση πρόβλεψης F(X) από τις πραγματικές τιμές y. Με την προσαρμογή ενός δέντρου αποφάσεων με βάση την αρνητική κλίση, ο αλγόριθμος βρίσκει τα χαρακτηριστικά εισόδου και τα αντίστοιχα σημεία διαχωρισμού που συμβάλουν στην μείωση της διαφοράς μεταξύ της τρέχουσας πρόβλεψης και των πραγματικών τιμών.

4. Στην συνέχεια, προσαρμόζεται ένα δέντρο απόφασης h_t (X) στην αρνητική κλίση και με αυτό τον τρόπο ο αλγόριθμος βρίσκει τα χαρακτηριστικά εισόδου και τα αντίστοιχα σημεία διαχωρισμού που συμβάλουν στην μείωση της διαφοράς μεταξύ της τρέχουσας πρόβλεψης και των πραγματικών τιμών. Ο τύπος για την προβλεπόμενη τιμή του δέντρο απόφασης για την είσοδο X είναι :

$$h_t(X) = \arg\min_h \sum_i L\left(y_i, F_{(t-1)}(x_i) + \lambda h_t(x_i)\right) \quad (12)$$

όπου h_t είναι ένα δέντρο απόφασης με παραμέτρους θ, λ. Το θ ελέγχει τη δομή και την πολυπλοκότητα του δέντρου, συμπεριλαμβανομένου του αριθμού των διαχωρισμών και του βάθους του δέντρου. Η λ είναι μια παράμετρος συρρίκνωσης που συνήθως είναι ένας μικρός θετικός αριθμός που χρησιμοποιείται για τον έλεγχο του ρυθμού μάθησης του αλγορίθμου, αφού ελέγχει τη συμβολή του νέου δέντρου στη συνολική πρόβλεψη.

5. Τέλος, ενημερώνεται η συνάρτηση πρόβλεψης με ένα σταθμισμένο άθροισμα της προηγούμενης πρόβλεψης και του νέου δέντρου:

$$F_t(X) = F_{(t-1)}(X) + \lambda h_t(X) \quad (13)$$

όπου $F_t(X)$ είναι η νέα συνάρτηση πρόβλεψης στην επανάληψη t.

6. Η παραπάνω διαδικασία επαναλαμβάνεται έως ότου ικανοποιηθεί ένα κριτήριο διακοπής και η τελική συνάρτηση που προκύπτει είναι ένα σταθμισμένο άθροισμα όλων των προηγούμενων δέντρων απόφασης.

2.3.2.5 Support vector machine classifier

Οι Μηχανές Υποστήριξης Διανυσμάτων (SVM) ανήκουν σε μια κατηγορία κατηγοριοποιητών που προσεγγίζουν τα δεδομένα γραμμικά και αναφέρθηκαν πρώτη φορά από τους Vapnik και Chermovenkis το 1963. Είναι κατάλληλοι για προβλήματα όπου υπάρχει μεγάλο πλήθος χαρακτηριστικών και το πλήθος των παραδειγμάτων είναι μικρός. Η βασική ιδέα των SVM είναι η εύρεση ενός ορίου απόφασης που να διαχωρίζει τις παρατηρήσεις στις διάφορες κατηγορίες, π.χ. σε αρνητικές και θετικές. Συγκεκριμένα, οι SVM λαμβάνοντας ως δεδομένο το σύνολο εκπαίδευσης, στοχεύουν στην δημιουργία ενός υπερεπιπέδου, δηλαδή ενός ορίου απόφασης, που να διαχωρίζει τις κλάσεις με τρόπο που να μπορεί να κατηγοριοποιούνται εξίσου καλά και νέα δεδομένα. Στην περίπτωση της ταξινόμησης κειμένου τα σημεία δεδομένων είναι τα έγγραφα και οι κλάσεις είναι οι κατηγορίες που ανήκουν τα κείμενα.

Έστω, λοιπόν, ένα σύνολο από N ζευγάρια δειγμάτων $\{x_i , y_i\}$ με $i = 1, \dots . N$, το x_i είναι ένα διάνυσμα χαρακτηριστικών για το i-οστό σημείο δεδομένων και το $y_i \epsilon \{ -1,+1\}$ είναι η ετικέτα της κλάσης του. Το υπερεπιπέδου ορίζεται ως :

$$w^T x + b = 0 \quad (14)$$

όπου w είναι το διάνυσμα βάρους που είναι κάθετο στο υπερεπίπεδο και b είναι η μεροληψία. Το διάνυσμα βάρους w αναπαριστά τον προσανατολισμό του υπερεπιπέδου στο χώρο και αντιπροσωπεύει την σημασία κάθε χαρακτηριστικού στη διαδικασία της ταξινόμησης. Ο στόχος του αλγόριθμου είναι η εύρεση του βέλτιστου διανύσματος βάρους που μεγιστοποιεί το περιθώριο μεταξύ των δύο κλάσεων. Το b καθορίζει τη θέση του υπερεπιπέδου και αντιπροσωπεύει την ελάχιστη απόσταση που χρειάζεται έτσι ώστε ένα σημείο να ταξινομηθεί σε μια συγκεκριμένη κλάση.

Τα σημεία που βρίσκονται πλησιέστερα στο υπερεπίπεδο ονομάζονται διανύσματα υποστήριξης (support vectors) και αποτελούν κρίσιμα σημεία για το διαχωρισμό των δύο

κλάσεων. Για τον διαχωρισμό πρέπει να βρεθεί το βέλτιστο υπερεπιπέδο και ως κριτήριο για την εύρεση του χρησιμοποιείται η μεγιστοποίηση του περιθώριο (margin). Το margin ορίζεται ως η απόσταση μεταξύ του υπερεπιπέδου και της πλησιέστερης θετικής ή αρνητική παρατήρησης. Στα μαθηματικά αυτό εκφράζεται ως :

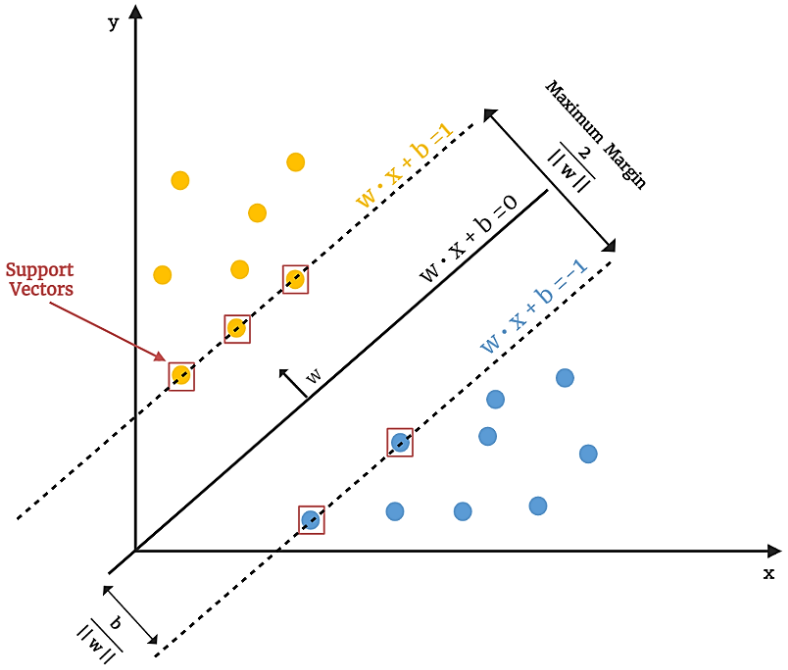
$$\max \frac{2}{||w||} \quad (15)$$

με τον περιορισμό :

$$y_i \, (\, b + \, w^T \, x_i - 1) \geq 0 \quad (16)$$

δηλαδή ότι όλα τα θετικά παραδείγματα βρίσκονται στη μία πλευρά του υπερεπιπέδου και όλα τα αρνητικά παραδείγματα βρίσκονται στην άλλη πλευρά (όπως φαίνεται στο Γράφημα 2.3). Όλο αυτό μπορεί να διατυπωθεί ως πρόβλημα βελτιστοποίησης ως εξής :

$$\min \frac{||w||}{2} \; subject \; to : y_i \, (\, b + \, w^T \, x_i - 1) \geq 0 \quad (17)$$



Γράφημα 2.3 : Υπερεπίπεδο μεταξύ δυο κλάσεων

Ο πρώτος όρος της σχέσης (17) "τιμωρεί" την πολυπλοκότητα του μοντέλου αφού του προσθέτει ένα κόστος όταν το μοντέλο αποδίδει υπερβολική σημασία σε οποιοδήποτε μεμονωμένο χαρακτηριστικό. Ο δεύτερος όρος είναι η συνάρτηση απώλειας που μετρά το σφάλμα ταξινόμησης. Ελαχιστοποιώντας τον πρώτο όρο και το σφάλμα ταξινόμησης, ο αλγόριθμος βρίσκει το υπερεπίπεδο με το μέγιστο περιθώριο που διαχωρίζει τα σημεία στις διαφορετικές κλάσεις. Η λύση του προβλήματος δίνεται μέσω της ελαχιστοποίησης την συνάρτηση Lagrange:

$$\mathrm{L} = \frac{||w||^2}{2} - \sum_{i=1}^{N} a[y_i\,(\,b + \,w\,x_i - 1)] \quad a \geq 0 \;\; (18)$$

θέτοντας τις μερικές παραγώγους της ως προς w, b και x_i ίσες με το μηδέν και λύνοντας το αντίστοιχο σύστημα εξισώσεων.

2.4 Νευρωνικά δίκτυα

2.4.1 Εισαγωγή

Οι παραδοσιακές μέθοδοι ταξινόμησης κειμένου και ειδικότερα της συναισθηματικής ανάλυσης, βασίζονταν σε μεθόδους που απαιτούσαν σε μεγάλο βαθμό την παρέμβαση του ανθρώπου όπως είναι οι bag-of words ή η n-grams. Στην συνέχεια, χρησιμοποιήθηκαν μέθοδοι μηχανικής μάθησης, όπως οι Naive Bayes, Support Vector Machines ή Decision Trees. Ενώ αυτές οι μέθοδοι είναι αποτελεσματικές σε κάποιο βαθμό, έχουν περιορισμούς ως προς την ικανότητά τους να καλύπτουν τις πολύπλοκες σχέσεις και τα μοτίβα που υπάρχουν στη φυσική γλώσσα. Επιπλέον, απαιτούν σχολαστική εξαγωγή των χαρακτηριστικών, δηλαδή την επιλογή, την εξαγωγή και τη μετατροπή των ακατέργαστων δεδομένων κειμένου σε ένα σύνολο αριθμητικών χαρακτηριστικών που μπορούν να χρησιμοποιηθούν ως είσοδος σε έναν αλγόριθμο μηχανικής μάθησης. Είναι φανερό ότι η διαδικασία αυτή είναι χρονοβόρα και επιρρεπής σε σφάλματα αφού απαιτεί εξειδίκευση στον τομέα και καλή κατανόηση των δεδομένων.

Τα νευρωνικά δίκτυα, από την άλλη πλευρά, μπορούν να εξάγουν αυτόματα τα κατάλληλα χαρακτηριστικά από τα ακατέργαστα δεδομένα κειμένου, μειώνοντας έτσι την ανάγκη για την εξαγωγή χαρακτηριστικών. Αυτό οφείλεται στην ικανότητά τους να μαθαίνουν πολύπλοκες μη γραμμικές σχέσεις μεταξύ εισόδων και εξόδων, γεγονός που τους επιτρέπει να συλλαμβάνουν αποχρώσεις και μοτίβα που υπάρχουν στη φυσική γλώσσα. Ειδικότερα, τα μοντέλα που βασίζονται σε μετασχηματιστές, όπως το ChatGPT, μπορούν να συλλάβουν το πλαίσιο και τις αποχρώσεις της γλώσσας και να ανιχνεύσουν τις λεπτές αλλαγές στο συναίσθημα που συχνά διαφεύγουν από τις παραδοσιακές προσεγγίσεις. Στο παρόν κεφάλαιο περιγράφονται οι βασικές έννοιες των νευρωνικών δικτύων και τα νευρωνικά δίκτυα που χρησιμοποιούνται ευρέως σήμερα στην συναισθηματική ανάλυση.

2.4.2 Βασικές έννοιες των νευρωνικών δικτύων

Η ιδέα και ο σχεδιασμός των νευρωνικών δικτύων εμπνεύστηκε από τη δομή και τη λειτουργία των βιολογικών νευρώνων του εγκεφάλου. Τα νευρωνικά δίκτυα προτάθηκαν για πρώτη φορά το 1944 από τους Warren McCullough και Walter Pitts, και έπειτα ακολούθησαν και άλλες μελέτες. Ωστόσο, οι σύγχρονες μορφές των νευρωνικών δικτύων, που

περιλαμβάνουν πολλαπλά επίπεδα και μη γραμμικές συναρτήσεις ενεργοποίησης, αναπτύχθηκαν στη δεκαετία του 1980. Έκτοτε, τα νευρωνικά δίκτυα έχουν γίνει όλο και πιο δημοφιλή και εφαρμόζονται σε πολλούς διαφορετικούς τομείς.

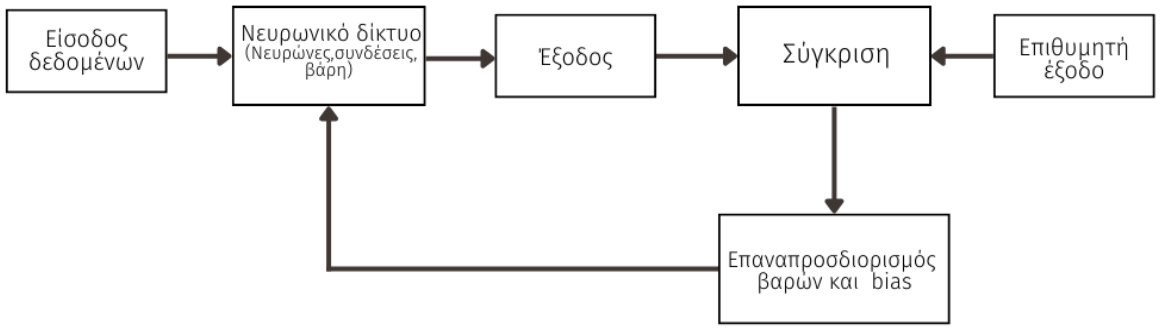
Η δομή ενός νευρωνικού δικτύου αποτελείται από διασυνδεδεμένους κόμβους, που ονομάζονται νευρώνες (neurons), οι οποίοι είναι οργανωμένοι σε επίπεδα (layers). Το πρώτο επίπεδο του νευρωσικού δικτύου ονομάζεται επίπεδο εισόδου (input layer) και δέχεται τα ακατέργαστα δεδομένα που εισάγονται στο μοντέλο. Το τελευταίο επίπεδο είναι το επίπεδο εξόδου (output layer), το οποίο παράγει το τελικό αποτέλεσμα, είτε της πρόβλεψης είτε της ταξινόμησης. Μεταξύ των επιπέδων εισόδου και εξόδου, υπάρχουν ένα ή περισσότερα κρυφά επίπεδα (hidden layers) τα οποία επεξεργάζονται τα δεδομένα και εξάγουν τα σχετικά χαρακτηριστικά. Αυτή η δομή επιτρέπει στα νευρωνικά δίκτυα να μοντελοποιούν τις πολύπλοκες σχέσεις μεταξύ εισόδων και εξόδων και να μαθαίνουν πληροφορίες σχετικά με τα δεδομένα εισόδου που είναι χρήσιμες για το εκάστοτε πρόβλημα.

Για την εργασία της ταξινόμηση κειμένου, ένα απλό νευρωνικό δίκτυο δέχεται στο επίπεδο εισόδου τα δεδομένο κείμενου, τα οποία έχουν υποβληθεί σε προ επεξεργασία και συχνά αναπαρίστανται ως διάνυσμα αριθμητικών τιμών, που συμβολίζονται ως x. Κάθε τιμή εισόδου, δηλαδή κάθε λέξη, συνδέεται με ένα βάρος που συμβολίζεται με w και αντιπροσωπεύει την ισχύ της σύνδεσης μεταξύ της εισόδου και του νευρώνα που είναι στο επόμενο επίπεδο. Όσο μεγαλύτερα η τιμή του βάρους, τόσο ισχυρότερη η σύνδεση που υποδηλώνει. Οι τιμές εισόδου πολλαπλασιάζονται με τα αντίστοιχα βάρη τους και τα αποτελέσματα αθροίζονται και έτσι παράγεται μια ενδιάμεση τιμή που ονομάζεται σταθμισμένο άθροισμα. Σε ορισμένες περιπτώσεις, εκτός από τα βάρη, ο νευρώνας έχει και ένα εσωτερικό συνοπτικό βάρος w_o, το οποίο ονομάζεται bias, και προστίθεται στο σταθμισμένο άθροισμα. Το bias μπορεί να θεωρηθεί ως μια μοναδιαία είσοδο, δηλαδή $x_o=1$, πολλαπλασιασμένη με το αντίστοιχο βάρος w_o. Αυτό δίνει την ακόλουθη σχέση:

$$S = w_0 + \sum_{i=1}^{n} w_i * x_i \quad (19)$$

οπού το w_i αντιπροσωπεύει το βάρος που σχετίζεται με την i-οστή είσοδο xi.

Κατά την διάρκεια της εκπαίδευσης, το νευρωνικό δίκτυο μαθαίνει να προσαρμόζει τα βάρη ώστε να ελαχιστοποιεί τη διαφορά μεταξύ της προβλεπόμενης εξόδου και της επιθυμητής εξόδου. Πιο συγκεκριμένα, το σταθμισμένο άθροισμα περνάει από μια συνάρτηση ενεργοποίησης (activation function) που συμβολίζεται ως f, η οποία εισάγει μη γραμμικότητα στο νευρωνικό δίκτυο και του επιτρέπει να μαθαίνει πολύπλοκα μοτίβα από δεδομένα εισόδου. Η συνάρτηση ενεργοποίησης αποφασίζει εάν ο νευρώνας πρέπει να "ενεργοποιηθεί" ή όχι με βάση την τιμή του σταθμισμένου αθροίσματος. Συγκρίνει το σταθμισμένο άθροισμα με μια τιμή κατωφλιού θ. Εάν S > θ, ο νευρώνας ενεργοποιείται διαφορετικά δεν ενεργοποιείται. Η έξοδος της συνάρτησης ενεργοποίησης χρησιμεύει ως είσοδος για το επόμενο επίπεδο του νευρωνικού δικτύου. Η ίδια διαδικασία επαναλαμβάνεται για κάθε επόμενο κρυφό επίπεδο στο νευρωνικό δίκτυο, δηλαδή τα επεξεργασμένα δεδομένα περνούν από τα κρυφά επίπεδα μέχρι να φτάσουν στο επίπεδο εξόδου. Το επίπεδο εξόδου είναι υπεύθυνο για την παραγωγή της τελικού αποτελέσματος που πολλές περνάει από μια τελική συνάρτηση ενεργοποίησης η οποία είναι προσαρμοσμένη στο συγκεκριμένο πρόβλημα. Στο Γράφημα 2.4 απεικονίζεται γραφικά η διαδικασία της εκπαίδευσης του νευρωνικού δικτύου.



Γράφημα 2.4 : Διαδικασία εκπαίδευσης νευρωνικού δικτύου.

Θα πρέπει σημειωθεί ότι η ενεργοποίηση ή όχι ενός νευρώνα εξαρτάται από τα βάρη των συνδέσεων και την τιμή κατωφλίου του εκάστοτε νευρώνα. Επιπλέον, δεν είναι αναγκαίο όλοι οι νευρώνες να υλοποιούν την ίδια συνάρτηση ενεργοποίησης. Τέλος, ο αριθμός των κρυφών επιπέδων και ο αριθμός των νευρώνων σε κάθε κρυφό επίπεδο μπορεί να ποικίλλει ανάλογα με την αρχιτεκτονική του νευρωνικού δικτύου.

2.4.2.1 Συνάρτηση απώλειας

Στόχος του νευρωνικού δικτύου κατά τη διάρκεια της εκπαίδευσης είναι να ελαχιστοποιήσει την τιμή της συνάρτησης απώλειας (loss function), η οποία υποδεικνύει την απόδοση ή όχι του μοντέλου στην πραγματοποίηση ακριβών προβλέψεων. Στο πλαίσιο της ταξινόμησης κειμένου, η συνάρτηση απώλειας είναι μια μαθηματική συνάρτηση που ποσοτικοποιεί τη διαφορά μεταξύ της προβλεπόμενης τιμής και της πραγματικής εξόδου (στόχου) για ένα δοσμένο κείμενο. Η επιλογή της συνάρτησης απώλειας εξαρτάται από το εκάστοτε πρόβλημα που επιλύεται και ποικίλλει ανάλογα με την αρχιτεκτονική του μοντέλου. Ορισμένες από τις συναρτήσεις απωλειών που χρησιμοποιούνται συχνά σε προβλήματα ταξινόμησης κειμένου είναι οι εξής:

1. **Δυαδική Διασταυρούμενη Εντροπία (Binary Cross-Entropy Loss)**

   Αυτή η συνάρτηση απώλειας χρησιμοποιείται για εργασίες ταξινόμησης κειμένου όπου υπάρχουν μόνο δύο κατηγορίες, για παράδειγμα στη ανάλυση συναισθήματος (θετικό ή αρνητικό συναίσθημα). Η συνάρτηση απώλειας μετρά την ανομοιότητα μεταξύ της προβλεπόμενης πιθανότητας της θετικής κλάσης και της πραγματικής ετικέτας για ένα δεδομένο κείμενο εισόδου. Πιο συγκεκριμένα, οι προβλεπόμενες πιθανότητες και οι πραγματικές ετικέτες αναπαρίστανται ως διανύσματα, μεγέθους 2, που αντιστοιχεί στις δύο κατηγορίες. Η συνάρτηση απώλειας στοχεύει στην βελτίωση των προβλέψεων "τιμωρώντας" το μοντέλο για λανθασμένες προβλέψεις, π.χ. όταν η προβλεπόμενη πιθανότητα αποκλίνει από την πραγματική ετικέτα, και εκπαιδεύοντας το μοντέλο να παράγει πιθανότητες που είναι πιο κοντά στην πραγματική ετικέτα. Ο τύπος της Binary Cross-Entropy Loss είναι ο ακόλουθος:

$$\text{Log loss} = \frac{1}{N}\sum_{i=1}^{N} -[y_i * log\,(p_i) + (1 - y_{i)} * log\,(1 - p_i)] \quad (20)$$

   Το y είναι η πραγματική ετικέτα, η οποία παίρνει την τιμή 0 ή 1, όπου το 0 αντιπροσωπεύει την αρνητική κλάση και το 1 τη θετική κλάση. Το p_i είναι η προβλεπόμενη πιθανότητα της θετικής κλάσης, που υπολογίζεται από το μοντέλο νευρωνικών δικτύων. Αντιπροσωπεύει την εκτιμώμενη πιθανότητα ότι το κείμενο εισόδου ανήκει στη θετική κατηγορία. Ενώ το $1\text{-}p_i$ είναι η πιθανότητα της αρνητικής κλάσης.

Η Binary Cross-Entropy Loss αυξάνεται όταν η αναμενόμενη τιμή της πιθανότητας αποκλίνει από την πραγματική ετικέτα. Επομένως, όταν η προβλεπόμενη πιθανότητα p είναι κοντά στην πραγματική ετικέτα y, δηλαδή η πιθανότητα είναι κοντά στο 1, η απώλεια είναι μικρή και υποδηλώνει μια καλή πρόβλεψη. Αντίστοιχα, όταν η προβλεπόμενη πιθανότητα απέχει πολύ από την πραγματική ετικέτα, δηλαδή η πιθανότητα είναι κοντά στο 0, η απώλεια γίνεται μεγαλύτερη, υποδηλώνοντας κακή πρόβλεψη.

2. **Κατηγορηματική Διασταυρούμενη Εντροπία (Categorical Cross-Entropy Loss).**

Αυτή η συνάρτηση απώλειας χρησιμοποιείται σε εργασίες ταξινόμησης κειμένου πολλών κλάσεων, όπου υπάρχουν περισσότερες από δύο κατηγορίες, όπως η ταξινόμηση θεμάτων ή η κατηγοριοποίηση ειδήσεων. Μετρά την ανομοιότητα μεταξύ των προβλεπόμενων πιθανοτήτων όλων των κλάσεων και της πραγματικής ετικέτας πολλαπλών κλάσεων για ένα δεδομένο κείμενο εισόδου. Ο τύπος για της Categorical Cross-Entropy Loss έχει ως εξής:

$$\text{Log } loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j}^{M} y_{ij}\ log\,(p_{ij}) \quad (21)$$

όπου το M είναι ο αριθμός των κλάσεων, N είναι ο αριθμός των γραμμών. Το y είναι ένα ενιαίο κωδικοποιημένο διάνυσμα που αντιπροσωπεύει την πραγματική ετικέτα του κείμενο εισόδου, δηλαδή περιέχει μηδενικές τιμές σε όλες τις θέσεις, εκτός από τη θέση που αντιστοιχεί στη σωστή κλάση, όπου υπάρχει ο αριθμός 1. Το p είναι η προβλεπόμενη πιθανότητα κάθε κλάσης και αναπαρίστανται ως ένα διάνυσμα μήκους M, όπου κάθε στοιχείο αντιπροσωπεύει την προβλεπόμενη πιθανότητα για την κατηγορία i, με i = 1…M. Το άθροισμα των προβλεπόμενων πιθανοτήτων είναι 1, το οποίο βρίσκονται στο εύρος [0, 1] και υποδηλώνει το επίπεδο εμπιστοσύνης του μοντέλου στην εκχώρηση του δείγματος δεδομένων εισόδου σε κάθε κλάση. Όσο υψηλότερη είναι η προβλεπόμενη πιθανότητα κλάσης, τόσο πιο πιθανό είναι το μοντέλο να αντιστοιχίσει το δείγμα δεδομένων εισόδου σε εν λόγω κλάση. Για παράδειγμα, έστω ένα πρόβλημα ταξινόμησης με 3 κλάσεις (M=3) με το διάνυσμα εξόδου να είναι το εξής [0.5, 0.2, 0.3]. Το p_1 αντιπροσωπεύει την προβλεπόμενη πιθανότητα για την κλάση 1που είναι 0.5, p_2 για την κλάση 2 που είναι 0.2 και το p_3 για την κλάση 3 που είναι 0.3.Σε αυτήν την περίπτωση, το μοντέλο προβλέπει ότι το δείγμα δεδομένων εισόδου έχει μεγαλύτερη πιθανότητα να ανήκει στην κλάση 1.

Όπως και η Binary Cross-Entropy έτσι και η Loss Categorical Cross-Entropy Loss υπολογίζει την διαφορά μεταξύ των προβλεπόμενων πιθανοτήτων και των πραγματικών ετικετών πολλαπλών κατηγοριών. Στην συνέχεια, τιμωρεί το μοντέλο για λανθασμένες προβλέψεις. Οι προβλεπόμενες πιθανότητες συγκρίνονται με τις πραγματικές πιθανότητες κατηγορίας και η απώλεια ελαχιστοποιείται κατά τη διάρκεια της εκπαιδευτικής διαδικασίας για τη βελτιστοποίηση της απόδοσης του μοντέλου.

2.4.2.2 Συνάρτηση ενεργοποίησης

Μια άλλη σημαντική παράμετρος στα νευρωνικά δίκτυα είναι η επιλογή της συνάρτησης ενεργοποίησης, η οποία εξαρτάται από τις ειδικές απαιτήσεις και τα χαρακτηριστικά του προβλήματος ταξινόμησης κειμένου. Ορισμένες από τις πιο συνηθισμένες συναρτήσεις ενεργοποίησης που χρησιμοποιούνται σε προβλήματα ταξινόμησης κειμένου είναι οι εξής:

1. **Σιγμοειδή συνάρτηση ενεργοποίησης** (Sigmoid Activation Function)**.**
   Αυτή η συνάρτηση ενεργοποίησης εφαρμόζεται συχνά σε διαδικασίες ταξινόμησης κειμένου όπου υπάρχουν μόνο δύο κλάσεις. Είναι ιδανική για την αναπαράσταση πιθανοτήτων, καθώς μετατρέπει την έξοδο ενός κόμβου σε μια τιμή μεταξύ 0 και 1, που μπορεί να ερμηνευθεί ως η πιθανότητα της θετικής κλάσης. Ο τύπος της για μια είσοδο x είναι :

$$f(x) = \frac{1}{1 + \ exp(-x)} \quad (22)$$

   Η έξοδος της συνάρτησης αποτελεί πιθανότητα και μπορεί να χρησιμοποιηθεί ως κατώφλι για την κατηγοριοποίηση των δεδομένων σε δύο κλάσεις, ανάλογα με το αν είναι μεγαλύτερη ή μικρότερη από την τιμή του κατωφλίου.

   Η σιγμοειδής συνάρτηση ενεργοποίησης προσφέρει πολλά πλεονεκτήματα στο πλαίσιο της εκπαίδευσης των νευρωνικών δικτύων. Ένα από αυτά τα πλεονεκτήματα της είναι η ομαλή της κλίση, η οποία είναι σχήματος S, και επιτρέπει ομαλές μεταβάσεις στις τιμές εξόδου όταν υπάρχουν μικρές αλλαγές στις τιμές εισόδου. Έτσι εξασφαλίζεται ο εύκολος υπολογισμός της κλίση της συνάρτησης και η χρήση της σε αλγορίθμους βελτιστοποίησης. Ακόμη, η σιγμοειδής συνάρτηση είναι διαφοροποιήσιμη, πράγμα που την καθιστά συμβατή

με διάφορες τεχνικές βελτιστοποίησης οι οποίες βασίζονται σε παραγώγους για την ενημέρωση των βαρών του μοντέλου. Είναι σημαντικό να σημειωθεί ότι η σιγμοειδής συνάρτηση είναι επιρρεπής όταν οι κλίσεις της συνάρτησης γίνονται πολύ μικρές για ακραίες τιμές. Αυτό προκαλεί αργή σύγκλιση και μειωμένη απόδοση του μοντέλου. Σε αυτή την περιπτώσεις είναι προτιμότερες άλλες συναρτήσεις ενεργοποίησης, όπως η Rectified Linear Unit (ReLU) ή Leaky ReLU.

2. **Συνάρτηση ενεργοποίησης υπερβολικής εφαπτομένης** (Tanh Activation Function).
   Η συνάρτηση αύτη είναι μια άλλη ευρέως διαδεδομένη συνάρτηση ενεργοποίησης στην ταξινόμηση κειμένου και αποτελεί μια βελτιωμένη έκδοση της συνάρτησης σιγμοειδούς ενεργοποίησης. Παρέχει μεγαλύτερη περιοχή απόκρισης αφού χαρτογραφεί την έξοδο ενός κόμβου σε μια τιμή μεταξύ -1 και 1, επιτρέποντας την αποτελεσματική καταγραφή τόσο θετικών όσο και αρνητικών τιμών. Η συνάρτηση υπερβολικής εφαπτομένης για μια είσοδο x ορίζεται μαθηματικά ως:

$$f(x) = \frac{exp(x) \ - \ exp(-x)}{exp(x) \ + \ exp(-x)} \quad (23)$$

   Από την σχέση (23) προκύπτει ότι η συνάρτηση αυτή είναι συμμετρική ως προς το μηδέν, αφού η τιμή της είναι μηδέν όταν η είσοδος της είναι μηδέν, δηλαδή όταν x=0. Αυτή η ιδιότητα της επιτρέπει την εύκολη αντιστοίχιση των τιμών εξόδου σε έντονα αρνητικά, ουδέτερα ή έντονα θετικά συναισθήματα. Επίσης, η συγκεκριμένη συνάρτηση χρησιμοποιείται στα κρυφά επίπεδα των νευρωνικών δικτύων επειδή οι τιμές εξόδου της βρίσκονται μεταξύ -1 και 1. Αυτό σημαίνει ότι βοηθά στην κεντροποίηση των δεδομένων και εξασφαλίζει ότι ο μέσος όρος για το κρυφό επίπεδο είναι κοντά στο 0 ή ακριβώς 0 στην ιδανική περίπτωση. Επομένως η διαδικασία μάθησης για τα επόμενα επίπεοδα του νευρωνικού δικτύου γίνεται πιο εύκολη αφού το πρόβλημα των μικρών κλίσεων αποφεύγεται και η διαδικασία βελτιστοποίησης καθιστάτε πιο σταθερή.

3. **Συνάρτηση ενεργοποίησης ReLU** (Rectified Linear Unit Activation Function).
   Η συγκεκριμένη συνάρτηση αποτελεί μια συνάρτηση ενεργοποίησης που εφαρμόζεται στις εξόδους των νευρώνων ενός νευρωνικού δικτύου και έχει τον ακόλουθο τύπο για μια δεδομένη είσοδο x :

$$f(x) = max(0, x) \quad (24)$$

Από τον παραπάνω τύπο φαίνεται ότι η συνάρτηση ReLU θέτει όλες τις αρνητικές τιμές ίσες με το μηδέν, ενώ αφήνει αμετάβλητες τις θετικές τιμές. Αυτό το χαρακτηριστικό της ReLU καθιστά αραιό το δίκτυο ως προς τις ενεργοποιήσεις, καθώς πολλοί από τους νευρώνες θα είναι ανενεργοί ,δηλαδή θα έχουν μηδενικές εξόδους. Ως αποτέλεσμα, η ReLU μπορεί να επιταχύνει σημαντικά τη διαδικασία εκπαίδευσης και να μειώσει το υπολογιστικό κόστος σε εργασίες ταξινόμησης κειμένου μεγάλης κλίμακας.

4. **Συνάρτηση ενεργοποίησης Softmax** (Softmax Activation Function).
   Η συνάρτηση softmax χρησιμοποιείται συνήθως για εργασίες ταξινόμησης πολλαπλών κλάσεων στην ταξινόμηση κειμένου, όπου υπάρχουν περισσότερες από δύο κλάσεις. Μετατρέπει την έξοδο των κόμβων ενός νευρωνικού δικτύου σε μια κατανομή πιθανοτήτων σε όλες τις πιθανές κλάσεις, εξασφαλίζοντας ότι οι προβλεπόμενες πιθανότητες αθροίζουν στη μονάδα. Η συνάρτηση softmax ορίζεται μαθηματικά ως εξής:

$$f(x_i) = \frac{exp(x_i)}{\sum_j exp(x_j)} \quad (25)$$

όπου x_i είσοδο στον i-οστό κόμβο ,δηλαδή x_i και στον παρανομαστή είναι το άθροισμά των εκθετικών τιμών όλων των κόμβων εξόδου. Το βασικό της πλεονέκτημα είναι ότι αντιμετωπίζει την ανισορροπία που μπορεί να υπάρξει ανάμεσα στις κλάσεις. Μέσω της ανάθεσης πιθανοτήτων σε κάθε κλάση, βασιζόμενες στα χαρακτηριστικά εισόδου, η συνάρτηση softmax εξασφαλίσει ότι το μοντέλο εκπαιδεύεται για να κάνει ακριβείς προβλέψεις για όλες τις κλάσεις, ανεξαρτήτως του αριθμού των δειγμάτων ανά κλάση στα δεδομένα εκπαίδευσης.

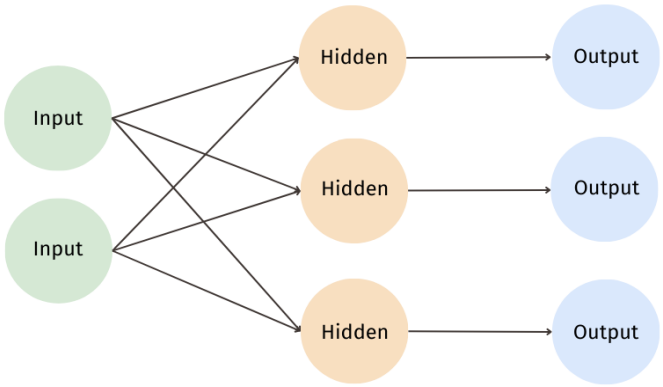
Εκτός από την συνάρτηση απώλειας και της ενεργοποίησης, το πιο σημαντικό μέρος των νευρωνικών δικτύων είναι η αρχιτεκτονική τους. Η επιλογή της αρχιτεκτονικής ενός νευρωνικού δικτύου εξαρτάται από παράγοντες όπως οι ειδικές απαιτήσεις της εργασίας, το μέγεθος του συνόλου δεδομένων, η διαθεσιμότητα των υπολογιστικών πόρων και το επιθυμητό επίπεδο ακρίβειας. Οι πιο συνηθισμένες αρχιτεκτονικές που συναντώνται σε προβλήματα

ανάλυσης συναισθήματος είναι η Feedforward Neural Networks (FNNs), η Convolutional Neural Networks (CNNs), η Recurrent Neural Networks (RNNs) και οι Transformers, οι οποίες θα συζητηθούν στις επόμενες υποπαραγράφους.

2.4.3 Feedforward νευρωνικά δίκτυα

Τα Feedforward νευρωνικά δίκτυα (FNN) ή αλλιώς multilayer perceptrons (MLP), αποτελούν τον πιο διαδεδομένο τύπο νευρωνικού δικτύου που έχει πολλές εφαρμογές, όπως είναι η αναγνώρισης εικόνας και ομιλίας, η επεξεργασίας φυσικής γλώσσας και η πρόβλεψης χρονοσειρών. Ο όρος "feedforward" δηλώνει ότι η πληροφορία ρέει μόνο προς μία κατεύθυνση, δηλαδή από την είσοδο προς την έξοδο, χωρίς συνδέσεις ανατροφοδότησης.

Ένα τυπικό FNN αποτελείται από νευρώνες οργανωμένους ανά επίπεδα, χωρίς συνδέσεις μεταξύ νευρώνων του ίδιου επιπέδου. Σημειώνεται ότι δεν υπάρχει ανατροφοδότηση μεταξύ των επιπέδων. Μια τυπική μορφή του FNN περιλαμβάνει ένα επίπεδο εισόδου, ένα ή περισσότερα κρυφά επίπεδα και ένα επίπεδο εξόδου. Το επίπεδο εισόδου δέχεται τα δεδομένα εισόδου και με βάση τη διάσταση τους, ο αριθμός των κόμβων που περιέχονται σε αυτό, καθαρίζεται ανάλογα με το μέγεθός του, με κάθε κόμβο να αντιπροσωπεύει ένα χαρακτηριστικό. Η έξοδος του επιπέδου εισόδου είναι η είσοδο του πρώτου κρυφού επιπέδου. Ομοίως, η έξοδός του πρώτου κρυφού επιπέδου γίνεται η είσοδο του δεύτερου κρυφού επιπέδου κ.ο.κ. Ο αριθμός των κόμβων σε κάθε κρυφό επίπεδο και ο αριθμός των κρυφών επιπέδων καθορίζονται κατά τον σχεδιασμό της αρχιτεκτονικής του νευρωνικού δικτύου. Αυτή η διαδικασία επαναλαμβάνεται μέχρι το επίπεδο εξόδου όπου παράγεται η προγνωστική ταξινόμηση για τα δεδομένα εισόδου. Ο αριθμός των κόμβων στο επίπεδο εξόδου καθορίζεται από τον αριθμό των κλάσεων που είναι διαθέσιμες στο πρόβλημα της ταξινόμησης. Στο Γράφημα 2.5. παρουσιάζεται η δομή ενός τυπικού FNN.



Γράφημα 2.5 : Δομή ενός τυπικού FNN

Σε προβλήματα ταξινόμησης, η πραγματική συνάρτηση f *(x) = y εκχωρεί τα δεδομένα εισόδου x στις αντίστοιχες κατηγορίες τους y. Ο στόχος του FNN είναι η χαρτογράφηση μιας προβλεπόμενης συνάρτησης f(x;θ)=y ελαχιστοποιώντας την απόκλιση της από την πραγματική. Τα θ είναι τα βάρη, δηλαδή είναι αριθμητικοί παράμετροι που δηλώνουν την σημασία των συνδέσμων μεταξύ των νευρώνων στο δίκτυο. Για την προσαρμογή των βαρών χρησιμοποιούνται αλγόριθμοι εκπαίδευσης, έτσι ώστε το νευρωνικό δίκτυο να μπορεί να παράγει τις επιθυμητές εξόδους για ένα σύνολο εισόδων. Οι βασικοί αλγόριθμοι εκπαίδευσης που εφαρμόζονται κατά την εκπαίδευση των FNN είναι ο αλγόριθμος forward propagation και ο back propagation. Ειδικότερα, κατά την forward propagation τα δεδομένα εισόδου τροφοδοτούνται μέσω του δικτύου με κατεύθυνση από το επίπεδο εισόδου στο επίπεδο εξόδου, περνώντας μέσα από τα κρυφά επίπεδα. Κάθε νευρώνας υπολογίζει μια έξοδο βάση τις σταθμισμένες εισόδους και τη συνάρτηση ενεργοποίησης. Μετά τη forward propagation, οι υπολογισμένες έξοδοι συγκρίνονται με τις εξόδους στόχους για τον υπολογισμό του σφάλματος ή της απώλειας. Αυτό το σφάλμα αντιπροσωπεύει τη διαφορά μεταξύ των προβλεπόμενων εξόδων και των πραγματικών εξόδων και χρησιμοποιείται ως μέτρο για την καλή απόδοση του νευρωσικού δικτύου. Το σφάλμα χρησιμοποιείται για την ενημέρωση των βαρών σε μια διαδικασία που ονομάζεται back propagation, κατά την οποία υπολογίζεται η διαβάθμιση του σφάλματος σε σχέση με τα βάρη ,τα οποία ενημερώνονται ανάλογα χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης, όπως το gradient descent . Η διαδικασία αυτή επαναλαμβάνεται πολλές φορές με σκοπό το FNN να μάθει τις βέλτιστες τιμές των παραμέτρων θ που ελαχιστοποιούν την προκαθορισμένη συνάρτηση απώλειας. Με αυτό τον τρόπο η προβλεπόμενη έξοδο είναι όσο το δυνατόν πιο κοντά στη αληθινή ετικέτα y για δεδομένο

x.Μόλις ολοκληρωθεί η διαδικασία της εκπαίδευσης το νευρωνικό δίκτυο αποθηκεύει τις τιμές του θ που βοηθούν στην καλύτερη προσέγγιση της άγνωστης συνάρτησης f*. Αυτό του δίνει την δυνατότητα να προβλέψει με ακρίβεια την κατηγορία y για νέα, αόρατα δεδομένα εισόδου x.

2.4.4 Συνελικτικά νευρωνικά δίκτυα

Τα συνελικτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNN) είναι ένας τύπος αρχιτεκτονικής βαθιάς μάθησης που εφαρμόζεται σε προβλήματα όπως η ταξινόμηση εικόνων και η ανίχνευση αντικειμένων. Ωστόσο, μπορεί να χρησιμοποιηθεί για προβλήματα ταξινόμησης κειμένου, όπου ο στόχος είναι η κατηγοριοποίηση του κειμένου σε προκαθορισμένες κατηγορίες ή κλάσεις. Πιο συγκεκριμένα, ένα μοντέλο CNN εφαρμόζει μια σειρά από φίλτρα που του επιτρέπουν την εκμάθηση των χαρακτηριστικών των δεδομένων και την ανάδυση σημαντικών παραμέτρων. Στην ουσία το CNN μαθαίνει αυτόματα τα φίλτρα, κάτι που στα κλασσικά νευρωνικά δίκτυα γίνεται χειροκίνητα.

Η βασική δομή ενός CNN αποτελείται από διάφορα επίπεδα που συνεργάζονται μεταξύ τους για να επεξεργαστούν τα δεδομένα εισόδου και να εξάγουν χρήσιμα χαρακτηριστικά. Κάθε επίπεδο εκτελεί διαφορετική επεξεργασία με τις πιο βασικές, που είναι η συνέλιξη (convolution) και η συγκέντρωση (pooling), να λαμβάνουν χώρα στο επίπεδο συνέλιξης και συγκέντρωσής αντίστοιχα. Τα βασικά επίπεδα ενός CNN είναι τα εξής:

1. **Επίπεδο εισόδου**: Το επίπεδο εισόδου είναι υπεύθυνο για τη λήψη και την επεξεργασία των ακατέργαστων δεδομένων πριν αυτά εισαχθούν στο CNN. Αυτό περιλαμβάνει βήματα προ-επεξεργασίας, όπως είναι η κανονικοποίηση και η κωδικοποίηση, με σκοπό τη μετατροπή των δεδομένων σε μορφή που μπορεί να χρησιμοποιηθεί από το CNN. Συνήθως, το τελικό αποτέλεσμα των δεδομένων κειμένου αναπαρίσταται ως πίνακας, όπου κάθε γραμμή αντιπροσωπεύει μια λέξη ή μια ακολουθία λέξεων και κάθε στήλη αντιπροσωπεύει ένα διάνυσμα χαρακτηριστικών που αποτυπώνει τη σημασία της λέξης.
2. **Συνελικτικά επίπεδα:** Το επίπεδο συνέλιξης εφαρμόζει ένα φίλτρο συνέλιξης στα δεδομένα εισόδου για την εξαγωγή τοπικών χαρακτηριστικών και μοτίβων.
3. **Συνάρτησης ενεργοποίησης:** Αυτό το επίπεδο εφαρμόζεται μια μη γραμμική συνάρτηση ενεργοποίησης, όπως η ReLU ή η Σιγμοειδή, στην έξοδο του επιπέδου συνέλιξης,

εισάγοντας μη γραμμικότητα στο μοντέλο και να επιτρέψει στο CNN να μάθει σύνθετα μοτίβα στα δεδομένα κειμένου.

4. **Συγκεντρωτικά επίπεδα:** Το επίπεδο αυτό εκτελεί υποδειγματοληψία στην έξοδο του επιπέδου συνέλιξης για να μειώσει τη χωρική διάσταση των χαρακτηριστικών και να συλλάβει τις πιο σημαντικές πληροφορίες.
5. **Πλήρως συνδεδεμένο επίπεδο**: Μετά τον επαναλαμβανόμενο συνδυασμό των βημάτων της συνέλιξης και της συγκέντρωσης, τα αποτελέσματα εισάγονται σε ένα πλήρως συνδεδεμένο επίπεδο (Fully Connected Layer), όπου κάθε νευρώνας συνδέεται με κάθε νευρώνα του προηγούμενου επιπέδου σχηματίζοντας ένα πυκνό δίκτυο συνδέσεων. Οι αρχικές πληροφορίες από το κείμενο συνοψίζονται και μετασχηματίζονται σε έναν ή περισσότερους νευρώνες τελικής πρόβλεψης, οι οποίοι είναι υπεύθυνοι για την τελική απόφαση της ταξινόμησης του κειμένου.
6. **Επίπεδο εξόδου**: Το τελευταίο επίπεδο του μοντέλου είναι το επίπεδο εξόδου, όπου οι νευρώνες τελικής πρόβλεψης παράγουν την έξοδο του μοντέλου, δηλαδή παράγουν τις προβλεπόμενες πιθανότητες κλάσης για κάθε κλάση του προβλήματος. Για κάθε κλάση εφαρμόζεται και μια κατάλληλη συνάρτηση ενεργοποίησης, όπως η softmax για ταξινόμηση πολλαπλών κλάσεων ή η σιγμοειδής για δυαδική ταξινόμηση.

2.4.4.1 Συνέλιξη

Η συνέλιξη είναι η μαθηματική πράξη που πραγματοποιείται στο συνελικτικό επίπεδο για την εξαγωγή χαρακτηριστικών ή μοτίβων από δεδομένα εισόδου. Στο πλαίσιο της ταξινόμησης κειμένου, η συνέλιξη λειτουργεί σε δεδομένα κειμένου που αναπαρίστανται ως πίνακες με κάθε γραμμή να αναπαριστά μια λέξη ή μια ακολουθία λέξεων. Ειδικότερα, η διαδικασία περιλαμβάνει την εφαρμογή ενός φίλτρου W που στην ουσία είναι ένας μικρός πίνακας με βάρη που έχουν προκύψει από την εκπαίδευση του μοντέλου. Αυτό το φίλτρο κινείται πάνω από τα δεδομένα εισόδου με συγκεκριμένο βήμα το οποίο ονομάζεται "βήμα συνέλιξης" και καθορίζει την απόσταση που διανύει το φίλτρο στον οριζόντιο και στον κάθετο άξονα κάθε φορά. Σε κάθε βήμα της συνέλιξης, το φίλτρο έχει τοποθετηθεί σε μια συγκεκριμένη περιοχή των δεδομένων εισόδου και υπολογίζει το εσωτερικό γινόμενο μεταξύ των βαρών του φίλτρου και των αντίστοιχων τιμών των χαρακτηριστικών του κειμένου στην εκάστοτε περιοχή. Αυτό το γινόμενο αθροίζεται σε μια ενιαία τιμή η οποία αποθηκεύεται σε

ένα νέο πίνακα μικρότερων διαστάσεων που ονομάζεται χάρτης χαρακτηριστικών (feature map). Ο χάρτης χαρακτηριστικών διαμορφώνεται πλήρως με την ολίσθηση του φίλτρου σε όλα τα δεδομένα εισόδου, την εφαρμογή της πράξης συνέλιξης σε κάθε βήμα και την αποθήκευση των τιμών που προκύπτουν σε αυτόν. Οι τιμές του αντιπροσωπεύουν το βαθμό ταύτισης ανάμεσα στο φίλτρο συνέλιξής και στις τοπικές περιοχές των δεδομένων εισόδου που έχει εξετάσει. Υψηλότερες τιμές υποδηλώνουν ισχυρότερη απόκριση του φίλτρου σε ένα συγκεκριμένο μοτίβο ή χαρακτηριστικό των δεδομένων εισόδου, ενώ χαμηλότερες τιμές υποδηλώνουν ασθενέστερη απόκριση ή μηδενική απόκριση. Αυτές οι τιμές χρησιμοποιούνται στη συνέχεια για περαιτέρω επεξεργασία σε επόμενα στρώματα του CNN, για να γίνουν προβλέψεις ή ταξινομήσεις με βάση τα εξαγόμενα χαρακτηριστικά.

2.4.4.2 Pooling

Το pooling είναι μια τεχνική που χρησιμοποιείται στα CNN για τη διατήρηση των πιο σημαντικών χαρακτηριστικών που έχουν ληφθεί από τη συνέλιξη και την παράλληλη μείωση των διαστάσεων τους. Το pooling πραγματοποιείται σε κάθε χάρτη χαρακτηριστικών και χρησιμοποιεί ένα φίλτρο που προσπαθεί να αναγνωρίζει αν ένα χαρακτηριστικό υπάρχει στο κείμενο ανεξαρτήτως της μεταβολής της μορφή του. Αυτό το φίλτρο μετακινείται κατά μήκος του χάρτη χαρακτηριστικών με ένα ορισμένο βήμα, το οποίο καθορίζει το μέγεθος της μετατόπισης ή της επικάλυψης μεταξύ γειτονικών περιοχών. Το αποτέλεσμα της διαδικασίας είναι η αναπαράσταση των χαρτών χαρακτηριστικών εισόδου με μειωμένη χωρική διάσταση.

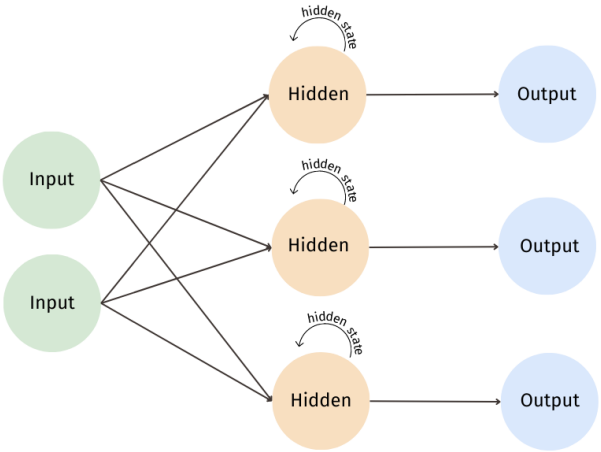
Υπάρχουν διάφορές μορφές pooling με τις δύο κύριες να είναι η μέγιστη συγκέντρωση (max pooling) και η μέση συγκέντρωση (average pooling). Στη max pooling, μια τοπική περιοχή χωρίζεται σε μη επικαλυπτόμενες περιοχές και η μέγιστη τιμή σε κάθε περιοχή επιλέγεται ως η αντιπροσωπευτική τιμή για την εν λόγω περιοχή. Αυτό συμβάλλει στη διατήρηση του πιο σημαντικού χαρακτηριστικού σε αυτή την περιοχή και στην απόρριψη λιγότερο σημαντικών χαρακτηριστικών. Από την άλλη πλευρά, κατά τη average pooling, η μέση τιμή εντός κάθε τοπικής περιοχής λαμβάνεται ως αντιπροσωπευτική τιμή. Αυτό συμβάλλει στη διατήρηση μιας πιο γενικευμένης αναπαράστασης των χαρακτηριστικών στην εν λόγω περιοχή. Σε γενικές γραμμές, η max pooling χρησιμοποιείται συχνότερα σε μοντέλα βαθιάς μάθησης λόγω της ικανότητάς της να καταγράφει τα κυρίαρχα χαρακτηριστικά, ενώ η

average pooling είναι χρήσιμη σε ορισμένες περιπτώσεις όπου είναι επιθυμητή μια πιο γενικευμένη αναπαράσταση των χαρακτηριστικών.

2.4.5 Αναδρομικά νευρωνικά δίκτυα

Τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks-RNN) είναι ένας τύπος νευρωνικού δικτύου που εφαρμόζεται για την επεξεργασία διαδοχικών δεδομένων, όπως είναι το κείμενο και ο ήχος. Το κύριο χαρακτηριστικό τους είναι η ικανότητά τους να διατηρούν τη μνήμη από προηγούμενες εισόδους. Αυτό τους επιτρέπει να κατανοούν τις ακολουθίες των δεδομένων και μέσω από αυτές, να μαθαίνουν σύνθετα μοτίβα και συσχετίσεις μεταξύ των στοιχείων της ακολουθίας. Για αυτούς τους λόγους είναι ιδιαίτερα αποτελεσματικά για την ταξινόμηση κειμένου.

Το RNN αποτελείται από επαναλαμβανόμενους κόμβους που συνδέονται μεταξύ τους με αναδρομικούς συνδέσμους (recurrent connection) οι οποίοι παρέχουν τη δυνατότητα να ανταλλαγής πληροφοριών και "θυμούνται" προηγούμενες καταστάσεις. Οι αναδρομικοί σύνδεσμοι αναφέρονται στους συνδέσμους που επιτρέπουν τη ροή των πληροφοριών από τον ένα κόμβο στον άλλο μέσω χρονικών βημάτων. Με άλλα λόγια, η έξοδος ενός κόμβου στο χρονικό βήμα t-1 χρησιμοποιείται ως είσοδος του ίδιου κόμβου στο επόμενο χρονικό βήμα t, δημιουργώντας μια αναδρομική σχέση. Σε κάθε χρονικό βήμα η "μνήμη" ή αλλιώς το hidden state του RNN ενημερώνεται. Καθώς στο hidden state επεξεργάζονται νέες εισροές, επιτρέπει στο RNN να συλλαμβάνει πληροφορίες από ολόκληρη την ακολουθία εισόδου και να μάθει πολύπλοκα μοτίβα με την πάροδο του χρόνου. Στο Γράφημα 2.6 παρουσιάζεται η δομή ενός τυπικού RNN.



Γράφημα 2.6 : Δομή ενός τυπικού RNN

Ένας τρόπος διάκρισης των RNN είναι ο χρόνος λειτουργίας τους , τα οποία μπορούν να ταξινομηθούν σε δύο κατηγορίες: τα αναδρομικά δίκτυα με χρονική καθυστέρηση και σύγχρονα αναδρομικά δίκτυα. Τα αναδρομικά δίκτυα με χρονική καθυστέρηση, γνωστά και ως RNN διακριτού χρόνου, είναι αρχιτεκτονικές RNN όπου η έξοδος κάθε νευρώνα σε ένα δεδομένο χρονικό βήμα χρησιμοποιείται ως είσοδος για το επόμενο χρονικό βήμα. Σε ένα RNN διακριτού χρόνου, κάθε νευρώνας έχει ένα συσχετισμένο hidden state που ενημερώνεται σε κάθε χρονικό βήμα με βάση την είσοδο, τις επαναλαμβανόμενες συνδέσεις και μια συνάρτηση ενεργοποίησης. Το ενημερωμένο hidden state σε κάθε χρονικό βήμα χρησιμοποιείται στη ως είσοδος για το επόμενο χρονικό βήμα, επιτρέποντας στο δίκτυο να συλλάβει πληροφορίες από προηγούμενα χρονικά βήματα και να τις ενσωματώσει στον τρέχοντα υπολογισμό. Αυτή η αναδρομική σύνδεση επιτρέπει στο δίκτυο να έχει μνήμη και να καταγράφει χρονικές εξαρτήσεις στα δεδομένα. Από την άλλη πλευρά, τα σύγχρονα αναδρομικά δίκτυα, επίσης γνωστά ως RNN συνεχούς χρόνου, είναι ένας τύπος αρχιτεκτονικής που εξελίχθηκαν από τα παραδοσιακά feedforward νευρωνικά δίκτυα, προσθέτοντας τη δυνατότητα της μνήμης και του χρόνου στην ανάλυση των δεδομένων. Στα RNN συνεχούς χρόνου οι νευρώνες εξελίσσουν συνεχώς το hidden state με την πάροδο του χρόνου, ενημερώνοντας το συνέχεια και ταυτόχρονα, χωρίς να βασίζονται σε διακριτά χρονικά βήματα. Χρησιμοποιώντας σύγχρονα αναδρομικά δίκτυα, η συνεχής εξέλιξη των δεδομένων μπορεί να μοντελοποιηθεί και να προσομοιωθεί με πιο ακριβή και αποτελεσματικό τρόπο σε σύγκριση με τις προσεγγίσεις διακριτού χρόνου.

Ένας τύπος RNN που ανήκει στην κατηγορία των RNN διακριτού χρόνου , είναι ο Long Short-Term Memory (LSTM) ο οποίο αναπτύχθηκε από τους Sepp Hochreiter και Jürgen Schmidhuber το 1997. Αποτελεί μία εξέλιξη των παραδοσιακών RNN, προσθέτοντας μηχανισμούς μνήμης που επιτρέπουν την αποθήκευση και την εκμετάλλευση πληροφοριών με μακροπρόθεσμη εξάρτηση στο χρόνο. Η βασική καινοτομία του LSTM είναι η χρήση εξειδικευμένων μονάδων μνήμης (cells) που μπορούν να αποθηκεύουν και να ενημερώνουν πληροφορίες για μεγάλες ακολουθίες, επιτρέποντας στο δίκτυο να μαθαίνει και να διατηρεί σχετικές πληροφορίες για μεγάλα χρονικά διαστήματα. Οι μονάδες μνήμης του LSTM αποτελούνται από τέσσερις βασικές θύρες (gates) που είναι οι εξής: input gate, output gate, state gate και forget gate. Οι θύρες προσθέτουν ευελιξία στο νευρωνικό δίκτυο αφού κάθε θύρα είναι υπεύθυνη για τη ρύθμιση της ροής της πληροφορίας. Κατά την επεξεργασία των

δεδομένων στην LSTM μονάδα, η input gate αποφασίζει ποιες νέες πληροφορίες θα εισέλθουν στην μονάδα και μέχρι πιο σημείο, η state gate ρυθμίζει την ενημέρωση του hidden state της μονάδας, η forget gate αποφασίζει ποιες πληροφορίες θα αφαιρεθούν από το του hidden state, και η output gate ελέγχει ποιες πληροφορίες θα εξέλθουν από την μονάδα. Η εισαγωγή της forget gate επιτρέπει στην LSTM να διατηρεί πληροφορίες από προηγούμενες εισόδους για μεγαλύτερο χρονικό διάστημα στη μνήμη της. Αυτό επιτρέπει στην LSTM να επεξεργάζεται ακολουθίες με μεγάλο χρονικό ορίζοντα, κρατώντας σημαντικές πληροφορίες από προηγούμενες εισόδους, ακόμα και αν έχουν περάσει αρκετά χρονικά βήματα.

2.4.6 Transformers

Οι Transformers εισάχθηκαν πρώτη φορά στο τομέα της επεξεργασίας φυσικής γλώσσας από την εταιρεία Google DeepMind μέσω της εργασία "Attention is All You Need" το 2017. Η εργασία πρότεινε μια νέα προσέγγιση που αντικατέστησε τα RNN με μηχανισμούς αυτό-προσοχής (self-attention), επιτρέποντας πιο αποτελεσματική επεξεργασία των ακολουθιών και εξαιρετική απόδοσης. Έκτοτέ, η αρχιτεκτονική του Transformer έχει γίνει το θεμέλιο για πολλά μοντέλα της σύγχρονής τεχνολογίας που έχει υιοθετηθεί και αναπτυχθεί ευρέως από την ερευνητική κοινότητα, οδηγώντας σε σημαντικές προόδους στην κατανόηση της φυσικής γλώσσας.

Η βασική καινοτομία της αρχιτεκτονικής του Transformer είναι ο μηχανισμός αυτό-προσοχής ο οποίος επιτρέπει στο μοντέλο να παρακολουθεί διαφορετικά μέρη της ακολουθίας εισόδου με διαφορετικά βάρη. Αυτός ο μηχανισμός επιτρέπει στο μοντέλο να καταγράφει μακροπρόθεσμες εξαρτήσεις μεταξύ των λέξεων ή των στοιχείων της ακολουθίας, ταυτόχρονα χωρίς τη χρήση αναδρομικών συνδέσμων όπως στα παραδοσιακά αναδρομικά νευρωνικά δίκτυα ,κάτι που είναι σημαντικό για πολλές εργασίες NLP.

Ένας μετασχηματιστής αποτελείται από δύο κύρια στοιχεία: τον κωδικοποιητή (encoder) και τον αποκωδικοποιητή (decoder) ,και σχεδιάστηκε αρχικά για εργασίες που περιλαμβάνουν συνεχόμενες λειτουργίες όπως η μηχανική μετάφραση, η απάντηση ερωτήσεων κ.λ.π. Ο κωδικοποιητής λαμβάνει ως είσοδο μια ακολουθία λέξεων ή tokens και παράγει μια ακολουθία συνεχών αναπαραστάσεων. Στην συνέχεια, ο αποκωδικοποιητής λαμβάνει τις συνεχείς αναπαραστάσεις που παράγονται από τον κωδικοποιητή και δημιουργεί την ακολουθία εξόδου,

η οποία μπορεί να είναι μια ακολουθία λέξεων, μεταφράσεων ή οποιαδήποτε άλλη επιθυμητή έξοδο. Σημειώνεται ότι τόσο ο κωδικοποιητής όσο και ο αποκωδικοποιητής περιλαμβάνουν πολλαπλά επίπεδα και κάθε ένα από αυτά αποτελείται από μικρότερα δομικά στοιχεία που ονομάζονται υπό-επίπεδα και θα αναλυθούν στην συνέχεια.

Συγκεκριμένα, ο κωδικοποιητής αποτελείται από δύο κύρια υπό-επίπεδα: το Multi-head self-attention επίπεδο και το Position-wise feed-forward επίπεδο. Το πρώτο υπολογίζει την αυτό-προσοχή για κάθε λέξη στην ακολουθία εισόδου και τα βάρη προσοχής (self-attention weights) για κάθε λέξη λαμβάνοντας υπόψη τις σχέσεις μεταξύ όλων των άλλων λέξεων της ακολουθίας. Ο μηχανισμός της αυτό-προσοχής επιτρέπει στο μοντέλο να παρακολουθεί διαφορετικά μέρη της ακολουθίας εισόδου με διαφορετικά βάρη, με αποτέλεσμα να αντιλαμβάνεται τόσο τοπικές όσο και γενικές πληροφορίες που προκύπτουν από τα συμφραζόμενα. Στο δεύτερο επίπεδο ο κωδικοποιητής χρησιμοποιεί ένα FNN σε κάθε αναπαράσταση λέξης με σκοπό εφαρμόσει έναν μη γραμμικό μετασχηματισμό σε αυτή και έτσι επιτρέπει στο μοντέλο να καταγράψει περίπλοκες αλληλεπιδράσεις μεταξύ λέξεων. Παρόμοια με τον κωδικοποιητή, ο αποκωδικοποιητής αποτελείται επίσης από τρία κύρια υπό-επίπεδα: το Multi-head self-attention επίπεδο,το Encoder-decoder attention επίπεδο και Position-wise feed-forward επίπεδο. Το πρώτο επίπεδο όπως και στον κωδικοποιητή, χρησιμοποιείται για να καταγράψει τις σχέσεις μεταξύ των λέξεων στην ακολουθία εξόδου. Αυτό επιτρέπει στον αποκωδικοποιητή να παρακολουθεί διαφορετικά μέρη της ακολουθίας εξόδου με διαφορετικά βάρη, συλλαμβάνοντας το πλαίσιο που απαιτείται για τη δημιουργία της εξόδου. Το επίπεδο επιτρέπει στον αποκωδικοποιητή να παρακολουθεί τις αναπαραστάσεις εξόδου του κωδικοποιητή. Υπολογίζει τα βάρη προσοχής μεταξύ της εισόδου του αποκωδικοποιητή και της εξόδου του κωδικοποιητή, επιτρέποντας στον αποκωδικοποιητή να ευθυγραμμίσει τις ακολουθίες εισόδου και εξόδου και να συλλάβει τις σχετικές πληροφορίες από τις αναπαραστάσεις του κωδικοποιητή. Τέλος, στο τρίτο επίπεδο ο αποκωδικοποιητής εφαρμόζει ένα FNN σε κάθε αναπαράσταση λέξης ανεξάρτητα, παρόμοιο με τον κωδικοποιητή.

2.4.7 Εφαρμογή του ChatGTP στην ανάλυση συναισθήματος

Το ChatGPT (Generative Pre-trained Transformer) είναι ένα καινοτόμο γλωσσικό μοντέλο που αναπτύχθηκε από την OpenAI και έχει προσφέρει πολλές δυνατότητες στο πεδίο της NLP. Το ChatGPT χρησιμοποιεί μια αρχιτεκτονική νευρωνικού δικτύου που βασίζεται

στους Transformers, η οποία του επιτρέπει να μαθαίνει και να παράγει απαντήσεις χωρίς να χρειάζεται να του ειπωθεί ρητά ποια είναι η σωστή απάντηση. Αυτό το καθιστά ένα ισχυρό εργαλείο για ένα ευρύ φάσμα εργασιών NLP, συμπεριλαμβανομένης της ανάλυσης συναισθήματος, αφού μπορεί να ρυθμιστεί λεπτομερώς για την συγκεκριμένη εργασία, εφαρμόζοντας μια κατάλληλη προσέγγιση της μάθησης με επίβλεψη. Σε σύγκριση με τις παραδοσιακές μεθόδους ανάλυσης συναισθήματος, οι οποίες βασίζονται σε χαρακτηριστικά και σε συστήματα βασισμένα σε κανόνες που δημιουργήθηκαν από ανθρώπινο παράγοντα, το ChatGPT προσφέρει μια πιο ευέλικτη και προσαρμόσιμη προσέγγιση στην ανάλυση συναισθήματος, καθιστώντας το μια δημοφιλή επιλογή. Σε αυτό το κεφάλαιο, παρέχεται μια επισκόπηση του ChatGPT και των βασικών χαρακτηριστικών του, συμπεριλαμβανομένης της αρχιτεκτονικής του καθώς και τις ρυθμίσεις που μπορούν υλοποιηθούν στο ChatGPT για τη εφαρμογή της ανάλυσης συναισθήματος.

2.4.7.1 Αρχιτεκτονική ChatGPT

Όπως αναφέρθηκε η αρχιτεκτονική του ChatGPT βασίζεται στο μοντέλο Transformer, το οποίο επιτρέπει την παράλληλη επεξεργασία δεδομένων εισόδου με την κωδικοποίηση κάθε λέξης σε μια ακολουθία σε σχέση με το περιεχόμενό της. Η διαδικασία της εκπαίδευσης του μοντέλου αποτελείται από δύο βασικά μέρη: την προ-εκπαίδευση (pre-training) και την τελειοποίηση (fine-tuning). Αναλυτικότερα, στη προ-εκπαίδευση, το μοντέλο εκπαιδεύεται σε ένα τεράστιο όγκο δεδομένων κειμένου, που προέρχονται από βιβλία, websites και άλλες πηγές, χρησιμοποιώντας μια προσέγγιση μη επιβλεπόμενης μάθησης. Κατά τη διάρκεια της προ-εκπαίδευσης, το μοντέλο εκπαιδεύεται για να προβλέπει την επόμενη λέξη σε μια πρόταση με βάση τα συμφραζόμενα των προηγούμενων λέξεων. Όμως η προ-εκπαίδευση δεν είναι επαρκής και εξακολουθεί να χρειάζεται περαιτέρω βελτίωση ώστε το μοντέλο να παρέχει εξατομικευμένες και πιο ακριβείς εξόδους. Σε αυτό το σημείο υπεισέρχεται το δεύτερο μέρος που είναι η τελειοποίησης. Η τελειοποίηση αποσκοπεί στη βελτίωση της απόδοσης ενός προ-εκπαιδευμένου μοντέλου σε μια συγκεκριμένη εργασία και αυτό το επιτυγχάνει με την εκπαίδευσή του σε ένα μικρότερο σύνολο δεδομένων με ετικέτες. Στην περίπτωση του ChatGPT, η τελειοποίηση περιλαμβάνει την προσαρμογή των βαρών του προ-εκπαιδευμένου μοντέλου με στόχο τη βελτίωση της απόδοσής του σε εργασίες όπως η ταξινόμηση κειμένου και η ανάλυση συναισθήματος. Με την τελειοποίηση του μοντέλου για τις συγκεκριμένες

εργασίες, το ChatGPT μπορεί να προσαρμοστεί ώστε να παρέχει πιο ακριβείς και εξατομικευμένες απαντήσεις στους χρήστες. Μετά τη φάση της τελειοποίησης, γίνεται εφαρμογή της Reinforcement Learning with Human Feedback (RLHF) η οποία χρησιμοποιείται για την περαιτέρω βελτίωση της ποιότητας των αποκρίσεων του μοντέλου. Το RLHF περιλαμβάνει τη χρήση ανθρώπινης ανατροφοδότησης για την ενίσχυση και τη βελτίωση της ποιότητας των αποκρίσεων του μοντέλου με την πάροδο του χρόνου. Τα τρία βασικά βήματα του RLHF είναι:

1.**Supervised Fine-tuning:** Σε αυτό το βήμα το μοντέλο εκπαιδεύεται μέσω επιβλεπόμενης μάθησης, να αναγνωρίζει μοτίβα στα δεδομένα με τη χρήση χαρακτηρισμένων παραδειγμάτων. Συγκεκριμένα, στο μοντέλο παρέχονται παραδείγματα εισόδου και εξόδου που πρέπει να μάθει. Στην περίπτωση του ChatGPT, τα παραδείγματα αυτά έχουν δημιουργηθεί από ανθρώπινους σχολιαστές που παρήγαγαν κατάλληλες απαντήσεις σε ένα σύνολο δεδομένων με βάση τις προτροπές των χρηστών.

2.**Reward Model:** Το εκπαιδευμένο μοντέλο από το προηγούμενο βήμα έχει πράξει πολλαπλές προβλέψεις για διάφορες προτροπές του χρήστη. Σε αυτό το βήμα, οι ανθρώπινοι σχολιαστές καλούνται να βαθμολογήσουν τις προβλέψεις με βάση τη χρησιμότητά τους, από τη λιγότερο χρήσιμη έως την πιο χρήσιμη. Έτσι το μοντέλο ανταμοιβής (reward model) χρησιμοποιεί αυτά τα δεδομένα και εκπαιδεύεται πάνω σε αυτά για να προβλέπει πόσο χρήσιμη ή όχι είναι μια απάντηση σε μια δεδομένη προτροπή.

3.**Reinforcement Learning:** Στο τελευταίο βήμα εφαρμόζεται η διαδικασία Reinforcement Learning χρησιμοποιείται για να βελτιώσει την απόδοση του Supervised Fine-tuning μοντέλου, το οποίο χρησιμοποιείται ως πράκτορας που μεγιστοποιεί την ανταμοιβή από το μοντέλο ανταμοιβής. Το μοντέλο παράγει μια απάντηση σε μια προτροπή του χρήστη, η οποία αξιολογείται από το μοντέλο ανταμοιβής. Στη συνέχεια, το μοντέλο Supervised Fine-tuning ενημερώσει τις προβλέψεις του για να επιτύχει μεγαλύτερες ανταμοιβές για μελλοντικές προβλέψεις.

Η προσέγγιση (RLHF) που υιοθετείται από το ChatGPT επιτρέπει στο μοντέλο να παράγει απαντήσεις υψηλής ποιότητας σε ένα πλαίσιο συνομιλίας χωρίς να βασίζεται χωρίς την ανάγκη ρητών κανόνων ή επισημασμένων δεδομένων. Αξιοποιώντας την ανθρώπινη

ανατροφοδότηση για τη βελτιστοποίηση της απόδοσής του, το μοντέλο μαθαίνει από τα ίδια του τα λάθη και βελτιώνει συνεχώς τις ικανότητές του στη δημιουργία απαντήσεων. Η προσέγγιση αυτή είναι ιδιαίτερα χρήσιμη σε σενάρια όπου τα επισημασμένα δεδομένα είναι ελάχιστα ή η πολυπλοκότητα της εργασίας καθιστά δύσκολο τον σχεδιασμό ρητών κανόνων. Επιπλέον, η προσέγγιση RLHF μπορεί να προσαρμοστεί σε διαφορετικούς τομείς και πλαίσια, καθιστώντας την ένα ευέλικτο και ισχυρό εργαλείο για ένα ευρύ φάσμα εργασιών.

2.4.7.2 ChatGPT και συναισθηματική ανάλυση

Η αρχιτεκτονική του ChatGPT αποτελείται από διάφορα μοναδικά χαρακτηριστικά που μπορούν να αξιοποιηθούν αποτελεσματικά για την ανάλυσης συναισθήματος. Ένα από τα βασικά χαρακτηριστικά του ChatGPT που μπορεί να αξιοποιηθεί για την βελτίωση της απόδοσης ενός μοντέλου συναισθηματικής ανάλυσης είναι η προ-εκπαίδευσή του σε τεράστιες ποσότητες δεδομένων κειμένου. Πιο συγκεκριμένα, η διαδικασία της προ-εκπαίδευσης επιτρέπει στο μοντέλο να μάθει μια πλούσια αναπαράσταση της γλώσσας, η οποία αποτυπώνει τις σημασιολογικές σχέσεις μεταξύ των λέξεων και των προτάσεων. Ως αποτέλεσμα αυτού είναι η βελτίωση τόσο της ικανότητας του να κατανοεί το νόημα ενός δοσμένου κειμένου όσο και της γενίκευσή του σε νέα παραδείγματα. Η εκπαίδευση του μοντέλου σε τεράστιές ποσότητες δεδομένων συνεπάγεται την έκθεσή του ποικίλες πηγές δεδομένων. Επομένως, η αντίληψη του ως προς τις αποχρώσεις και την πολυπλοκότητα της φυσικής γλώσσας αυξάνεται, κάτι που είναι ιδιαίτερα χρήσιμο στη ανάλυσης συναισθήματος, όπου μικρές διαφορές στη γλώσσα μπορούν να αλλάξουν δραστικά το συναίσθημα που εκφράζεται σε μια δήλωση. Ακόμη, ένα πρόβλημά που συναντάται στην ανάλυση συναισθήματος είναι η έλλειψη δεδομένων που συμβαίνει όταν τα επισημασμένα δεδομένα είναι περιορισμένα ή να είναι δύσκολο να αποκτηθούν. Η προ-εκπαίδευση βοηθάει το μοντέλο να ξεπεράσει αυτό το πρόβλημα, αφού εκπαιδεύεται πάνω μεγάλες ποσότητες δεδομένων χωρίς ετικέτες και εξάγει χρήσιμα χαρακτηριστικά από το κείμενο εισόδου, τα οποία μπορούν να ρυθμιστούν σε ένα μικρότερο σύνολο δεδομένων με ετικέτες για να δοθούν στο μοντέλο για τη συγκεκριμένη εργασία. Ο συνδυασμός όλων των παραπάνω επιτρέπει στο μοντέλο να κατανοεί ένα ευρύ φάσμα γλωσσικών γνώσεων και μοτίβων και ταυτόχρονα βελτιώνει την ικανότητά του να ταξινομεί με ακρίβεια το συναίσθημα.

Ένα ακόμα σημαντικό χαρακτηριστικό του ChatGPT είναι η τελειοποίηση, η οποία περιέχει τη λήψη του προ-εκπαιδευμένου μοντέλου και την περαιτέρω εκπαίδευσή του σε μια συγκεκριμένη εργασία, που στην συγκεκριμένη περίπτωση είναι η ανάλυσης συναισθήματος. Η περαιτέρω εκπαίδευση περιλαμβάνει την προσαρμογή των βαρών του νευρωνικού δικτύου, ώστε να είναι πιο κατάλληλα για τη ανάλυση συναισθήματος. Αρκετές μελέτες έχουν αποδείξει την αποτελεσματικότητα της προσαρμογής της τελειοποίησης του ChatGPT για εργασίες ανάλυσης συναισθήματος. Ένα τέτοιο παράδειγμα είναι η έρευνα των Wang κ.ά. [20] που προσάρμοσαν κατάλληλά την τελειοποίηση του ChatGPT σε ένα σύνολο δεδομένων με κριτικές ταινιών για ένα πρόβλημα δυαδικής ταξινόμησης του συναισθήματος. Χρησιμοποίησαν μια προσέγγιση με επιβλεπόμενη μάθηση όπου το μοντέλο εκπαιδεύτηκε σε ένα επισημασμένο σύνολο δεδομένων για να προβλέψει το συναίσθημα νέων, μη επισημασμένων δεδομένων κειμένου. Το σύνολο δεδομένων που χρησιμοποιήθηκε στη μελέτη αποτελούνταν από 50.000 κριτικές ταινιών, οι οποίες χωρίζονταν ομοιόμορφα σε θετικές και αρνητικές κριτικές. Στη συνέχεια, εφάρμοσαν την τεχνική της τελειοποίησης στο προ-εκπαιδευμένο μοντέλο του ChatGPT με σύνολο δεδομένων με τις ετικέτες των κριτικών ταινιών, προσαρμόζοντας τα βάρη και τα bias του δικτύου ώστε να ταιριάζει καλύτερα στην ταξινόμησης δυαδικού συναισθήματος. Οι Wang κ.ά. συνέκριναν την απόδοση του νέου μοντέλου ChatGPT με άλλα σύγχρονα μοντέλα ανάλυσης συναισθήματος, όπως το BERT και το XLNet και διαπίστωσαν ότι πέτυχε την υψηλότερη ακρίβεια ξεπερνώντας τα άλλα μοντέλα με σημαντική διαφορά. Ομοίως, οι Keskar κ.ά. [21] εφάρμοσαν την τεχνική της τελειοποίησης στο γλωσσικό μοντέλο GPT-2, μια παραλλαγή του ChatGPT, σε ένα σύνολο δεδομένων με κριτικές προϊόντων για την ταξινόμηση συναισθήματος πολλαπλών κλάσεων (θετικό, αρνητικό ή ουδέτερο). Οι ερευνητές χρησιμοποίησαν διάφορα σύνολα δεδομένων αναφοράς και πειραματίστηκαν με διαφορετικές παραμέτρους για να βελτιστοποιήσουν την απόδοση του μοντέλου. Συνέκριναν τις επιδόσεις του με παραδοσιακά μοντέλα μηχανικής μάθησης και διαπίστωσαν ότι το προ-εκπαιδευμένο γλωσσικό μοντέλο τα ξεπέρασε στα περισσότερα σύνολα δεδομένων. Διερεύνησαν επίσης τις επιδράσεις του μεγέθους των δεδομένων εκπαίδευσης και του μεγέθους του προ-εκπαιδευμένου μοντέλου στην απόδοση του λεπτομερώς ρυθμισμένου μοντέλου. Η μελέτη κατέδειξε την αποτελεσματικότητα της τελειοποίησης του προ-εκπαιδευμένου γλωσσικού μοντέλου για εργασίες ανάλυσης συναισθήματος.

Αν και το ChatGPT έχει αποδειχθεί ένα ισχυρό εργαλείο για τη ανάλυσης συναισθήματος, υπάρχουν ορισμένοι περιορισμοί κατά τη χρήση του. Ένας περιορισμός του ChatGPT στην ανάλυση συναισθήματος είναι η αδυναμία του να κατανοήσει με ακρίβεια την πολυπλοκότητα της γλώσσα, ιδίως έννοιες όπως ο σαρκασμός και η ειρωνεία. Αυτό συμβαίνει επειδή το μοντέλο βασίζεται σε στατιστικά μοτίβα που τα εντοπίζει από τα δεδομένα εκπαίδευσης και έννοιες όπως ο σαρκασμός να μην ταιριάζουν πάντα σε αυτά τα μοτίβα. Επιπλέον, η ακρίβεια του μοντέλου εξαρτάται σε μεγάλο βαθμό από την ποιότητα και την ποικιλομορφία των δεδομένων εκπαίδευσης. Εάν τα δεδομένα εκπαίδευσης είναι μεροληπτικά ή περιορισμένα, το μοντέλο είναι πιθανό να εξακολουθεί να είναι επιρρεπές κατά την διάρκεια της εκπαίδευσης. Ως εκ τούτου, το μοντέλο μπορεί να παρερμηνεύσει το συναίσθημα που εκφράζεται στο κείμενο, οδηγώντας σε ανακριβή ή προβληματική ταξινόμηση συναισθήματος, ειδικά σε ευαίσθητα ή αμφιλεγόμενα θέματα, με αποτέλεσμα το μοντέλο να μην γενικεύεται καλά σε νέα δεδομένα, οδηγώντας σε κακές επιδόσεις. Ένας ακόμα περιορισμός του ChatGPT έγκειται από την διαδικασία της λεπτομερούς ρύθμισης του προ-εκπαιδευμένου μοντέλου ,η οποία είναι αναγκαία για να αποτυπώσει το μοντέλο με ακρίβεια την ανάλυση συναισθήματος. Ωστόσο, η διαδικασία αυτή απαιτεί μεγάλος όγκος επισημασμένων δεδομένων, η απόκτηση των οποίων είναι δύσκολη και δαπανηρή, ιδίως για εξειδικευμένες θέματα. Τέλος, για την εκτέλεση της χρειάζεται τόσο η υπολογιστική ισχύ όσο και χρόνος , κάτι που καθιστά δύσκολο την αποτελεσματική χρήση ενός τέτοιου μοντέλου από όσους δεν έχουν πρόσβαση σε ισχυρούς υπολογιστικούς πόρους. Από τα παραπάνω, συμπεραίνεται ότι το ChatGPT είναι ένα μοντέλο γενικής γλώσσας που μπορεί να μην είναι πάντα κατάλληλο για συγκεκριμένες εργασίες και χρειάζεται συγκεκραμένες προϋπόθεσης τόσο κατά την εκπαίδευση όσο και κατά την τελειοποίηση του προκυμμένου να είναι αποτελεσματικό.

3 Μεθοδολογία

3.1 Εισαγωγή

Στο συγκεκριμένο κεφάλαιο γίνεται η περιγράφει τη διαδικασία διεξαγωγής ανάλυσης συναισθήματος σε δεδομένα του Twitter κατά τη διάρκεια της πανδημίας COVID-19 με τη χρήση αλγορίθμων μηχανικής μάθησης με επίβλεψη και νευρωνικών δικτύων. Το κεφάλαιο αυτό παρέχει λεπτομερή περιγραφή της συλλογής και της προ επεξεργασίας των δεδομένων, της εξαγωγής χαρακτηριστικών, της επιλογής και της αξιολόγησης μοντέλων. Η παρούσα μελέτη αποσκοπεί στην σύγκρισή διαφοετετικών μοντέλων για την ανάλυση συναισθήματος των συγκεκριμένων δεδομένων λαμβάνοντας υπόψιν της την υπάρχουσα βιβλιογραφία. Στόχος των μοντέλων είναι να προβλέψουν την κατηγορία συναισθήματος στην οποία ανήκει το κάθε δοσμένο tweet. Στις επόμενες ενότητες αναφέρεται βήμα προς βήμα η διαδικασία διεξαγωγής των δεδομένων και δίνεται έμφαση στη μεθοδολογία που χρησιμοποιήθηκε στην παρούσα μελέτη.

3.2 Συλλογή δεδομένων

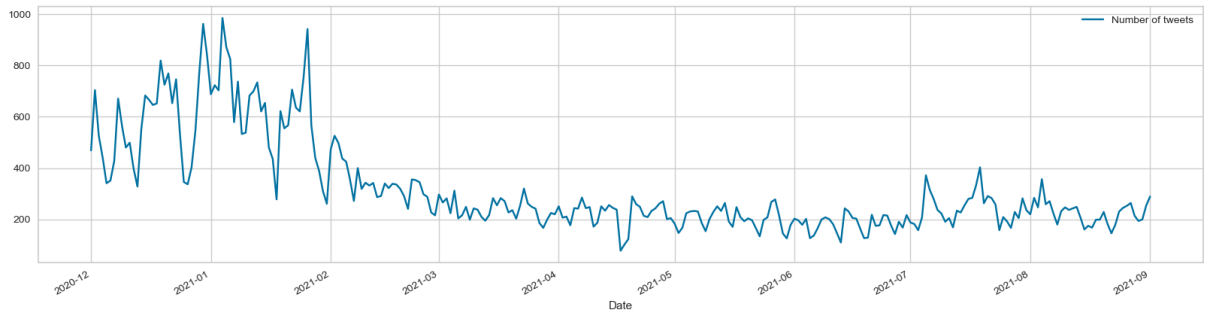
Για την εκπόνηση της εργασίας και του ελέγχου των μοντέλων χρησιμοποιήθηκε τροποποιημένο το παρακάτω σύνολο δεδομένων :

> OpenICPSR. (2021). Twitter COVID dataset - Sep2021: [tweetid_ userid_keyword_ sentiments_ emotions_United Kingdom.csv (Version V11)]

το οποίο συλλέχθηκε βάση των λέξεων κλειδιών “corona”, “wuhan”, “nCov” και “covid” και περιλαμβάνει tweets, που έχουν ως χώρα προέλευσης το Ηνωμένο Βασίλειο για την χρονική περίοδο από τον Ιανουαρίου το 2020 μέχρι Σεπτεμβρίου του 2021. Ειδικότερα, το αρχείο περιλαμβάνει διάφορες στήλες με πληροφορίες για κάθε tweet όπως tweet ID, user ID, και επιπλέον δεκαεπτά διαφορετικές ετικέτες που αντιστοιχούν σε διάφορες σημασιολογικές ιδιότητες. Το Tweet ID είναι ένας μοναδικός αναγνωριστικός αριθμό που αντιστοιχεί σε κάθε tweet και χρησιμοποιείται για την ταυτοποίηση και αναφορά ενός συγκεκριμένου tweet στο Twitter. Στην συγκεκριμένη εργασία χρησιμοποιήθηκε για την ανάκτηση των αρχικών tweets

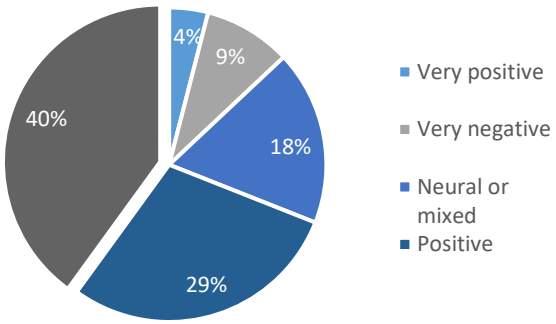
η οποία πραγματοποιήθηκε μέσο της εφαρμογής Hydrator App, η οποία έλαβε ως είσοδο το tweet ID και επίστρεψε το αντίστοιχο tweet.

Μετά την ανάκτηση των αρχικών tweets, ο όγκος των δεδομένων αυξήθηκε ραγδαία. Λόγω περιορισμένων υπολογιστικών πόρων το αρχικό σύνολο δεδομένων τροποποιήθηκε μειώνοντας την χρονική περίοδο από τον Δεκέμβριο του 2020 μέχρι τον Σεπτέμβριο του 2021 και επιλέγοντας μόνο μια σημασιολογική κατηγορία, δηλαδή της συναισθηματικής ανάλυσης. Στο τροποποιημένο σύνολο δεδομένων πραγματοποιήθηκε τυχαία δειγματοληψία και το τελικό σύνολο περιείχε 106.335 tweets με τις αντίστοιχες ετικέτες συναισθήματος. Στο Γράφημα 3.1 φαίνεται το πλήθος των tweets ανά ημέρα για την χρονική περίοδο Δεκέμβριο του 2020 με Σεπτέμβριος 2021.



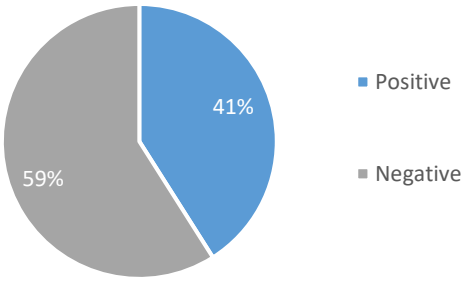
Γράφημα 3.1 : Πλήθος tweets ανά ημέρα

Το τελικό σύνολο δεδομένων περιλαμβάνει τρεις στήλες εκ των οποίων η πρώτη είναι η "created_at" που δηλώνει την ημερομηνία δημιουργίας του tweet, η δεύτερη είναι η "full_text" που είναι το πλήρες tweet και η τρίτη είναι η "sentiment" που είναι η ετικέτα συναισθήματος που αντιπροσωπεύει το συναίσθημα που εκφράζεται σε κάθε tweet. Οι πέντε κατηγορίες συναισθήματος που περιέχονται στο σύνολο είναι πολύ αρνητικό (Very negative), αρνητικό (Negative), ουδέτερο (Neutral or mixed), θετικό (Positive) και πολύ θετικό (Very positive). Όπως παρουσιάζεται και στο Γράφημα 3.2 το 40.1 % των tweets ανήκουν στην κατηγορία "Negative", το 29.4% στην κατηγορία "Positive", το 18.10% στην κατηγορία "Neutral or mixed" , το 8.52% στην κατηγορία "Very negative" ενώ 3.89% στην κατηγορία "Very positive"



Γράφημα 3.2 : Πλήθος tweets ανά κατηγορία συναισθήματος στο αρχικό σύνολο

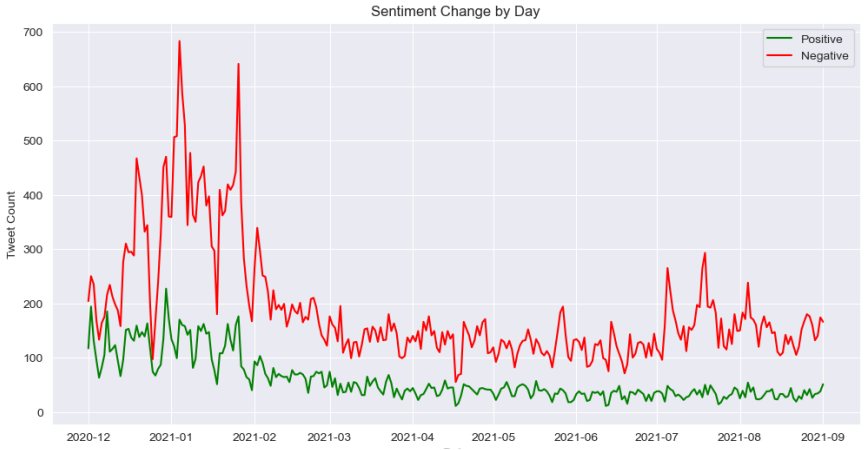
Όμως για να εφαρμοστούν αλγόριθμοι ταξινόμησης σε ένα σύνολο δεδομένων με συναισθήματα, χρειάζεται να μειωθεί ο αριθμός των κατηγοριών σε δύο. Συγκεκριμένα, αφαιρείται η κατηγορία "Neutral or mixed", και οι κατηγορίες "Very negative" και "Negative" ενώνονται σε μια κατηγορία, την "Negative", ενώ οι κατηγορίες "Very positive" και "Positive" ενώνονται στην κατηγορία "Positive". Έτσι, το σύνολο δεδομένων μειώνεται σε 87.101 εκ των οποίων το 59.3 % είναι στην κατηγορία ”Negative ”και 40.7% στην κατηγορία Positive” (Γράφημα 3.3).



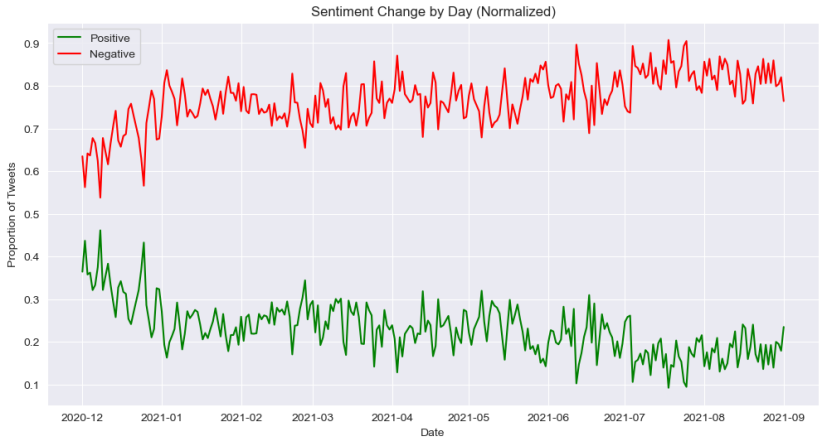
Γράφημα 3.3 : Πλήθος tweets ανά κατηγορία συναισθήματος

Στην συνέχεια παρουσιάζεται το Γράφημα 3.4 που απεικονίζει πώς το συναίσθημα των tweets μεταβάλλεται ανά ημέρα κατά τη διάρκεια της περιόδου μελέτης. Το Γράφημα 3.5 απεικονίζει ακριβώς το ίδιο με μόνη διαφορά ότι οι τιμές του συναισθήματος κανονικοποιήθηκαν για να ληφθούν υπόψη οι διαφορές στον αριθμό των tweets κάθε ημέρας.

Τόσο το διάγραμμα του Γράφημα 3.4 όσο και το Γράφημα 3.5 δείχνουν ότι το ποσοστό των θετικών tweets ήταν υψηλότερο στην αρχή της περιόδου μελέτης και μειώθηκε σταδιακά με την πάροδο του χρόνου, ενώ το ποσοστό των αρνητικών tweets αυξήθηκε με την πάροδο του χρόνου. Γενικά είναι φανερό ότι το ποσοστό των αρνητικών tweets υπερτερεί το ποσοστό των θετικών tweets καθ' όλη τη διάρκεια της περιόδου μελέτης. Τα ευρήματα αυτά υποδηλώνουν ότι το συναίσθημα των tweets που μπορεί να έχει μετατοπιστεί με την πάροδο του χρόνου και ότι το συναίσθημα ήταν γενικά πιο αρνητικό όσο περνούσε ο καιρός.



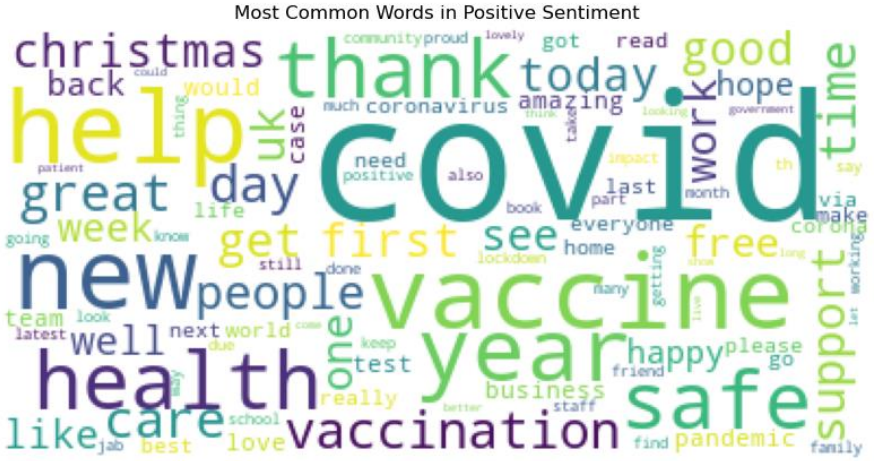
Γράφημα 3.4 : Συναίσθημα ανά ημέρα



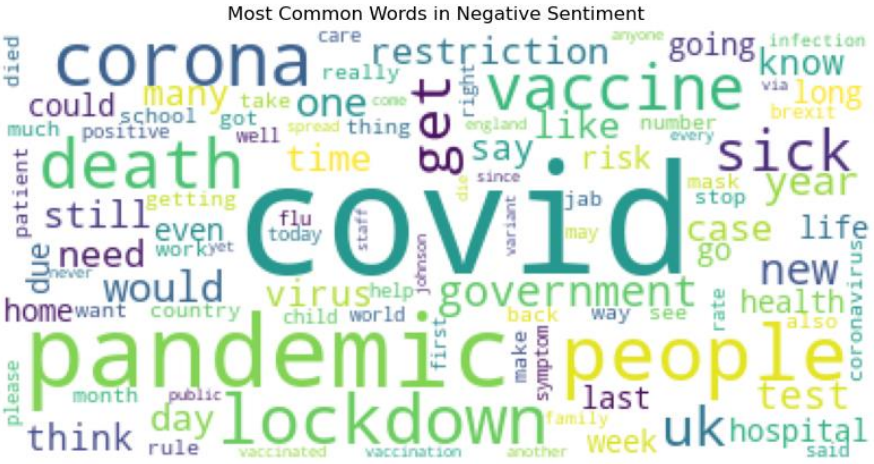
Γράφημα 3.5 : Συναίσθημα ανά ημέρα (κανονικοποιημένο)

Για την καλύτερη κατανόηση των βασικών θεμάτων που υπάρχουν στα δεδομένα κειμένου, δημιουργήθηκαν τα word clouds τα οποία είναι οπτικές που απεικονίζουν τις πιο

συχνά εμφανιζόμενές λέξεις στο σύνολο δεδομένων. Το μέγεθος μιας λέξης σε ένα word cloud είναι ανάλογο με τη συχνότητα εμφάνισής της στα δεδομένα κειμένου. Αυτό σημαίνει ότι οι λέξεις που εμφανίζονται συχνά στο κείμενο αναπαρίστανται με μεγαλύτερο μέγεθος γραμματοσειράς, ενώ οι λέξεις που εμφανίζονται λιγότερο συχνά αναπαρίστανται με μικρότερο μέγεθος γραμματοσειράς. Στην Εικόνα 3.1 είναι το word cloud για την θετική κατηγορία ενώ στη Εικόνα 3.2 για την αρνητική.



Εικόνα 3.1 : Word Cloud για την θετική κατηγορία



Εικόνα 3.2 : Word Cloud για την αρνητική κατηγορία

Στην Εικόνα 3.1 φαίνονται λέξεις όπως είναι "help" , "health", "vaccine" , "vaccination" , "safe" και "care" , που εστιάζουν στην υγειονομική περίθαλψη με ιδιαίτερη έμφαση στο εμβολιασμό του Covid-19. Στην Εικόνα 3.2 φαίνονται λέξεις όπως "death", "restriction", "pandemic", "lockdown" και "sick" που επικεντρώνονται περισσότερο στα μέτρα ασφαλείας και στις κατά την περίοδο την πανδημίας καθώς και των επιπτώσεων της.

3.3 Προ επεξεργασία δεδομένων

Η προ επεξεργασία αποτελεί ένα απαραίτητο βήμα στην ανάλυση συναισθήματος, καθώς συμβάλει στη μετατροπή των ακατέργαστων δεδομένων κειμένου σε μορφή που μπορεί να χρησιμοποιηθεί από μοντέλα μηχανικής μάθησης και νευρωνικών δικτύων. Τα δεδομένα του Twitter παρουσιάζουν αρκετές προκλήσεις λόγω της μη δομημένης φύσης τους, του μικρού μήκους κειμένου και του θορυβώδους του περιεχομένου τους, τα οποία μπορούν να επηρεάσουν την ακρίβεια των μοντέλων. Τα βήματα που υλοποιήθηκαν κατά την προ επεξεργασία των δεδομένων είναι τα εξής:

1. **Καθαρισμός δεδομένων:** Το βήμα αυτό περιλαμβάνει την αφαίρεση άσχετων δεδομένων, όπως είναι οι ειδικοί χαρακτήρες, URL, hashtags, τα σημεία στίξης και τα σύμβολα. Αυτό το βήμα είναι σημαντικό επειδή τα συγκεκριμένα στοιχεία δεν συμβάλλουν στην εύρεση του συναισθήματος του tweet και η ύπαρξή τους επηρεάζει την ακρίβεια της ανάλυσης συναισθήματος.
2. **Tokenization:** Το Tokenization είναι η διαδικασία διαχωρισμού του κειμένου σε μικρότερες μονάδες που ονομάζονται tokens, οι οποίες είναι συνήθως μεμονωμένες λέξεις ή φράσεις. Υπάρχουν διαφορετικές προσεγγίσεις όσον αφορά τα κριτήρια διαχωρισμού των δεδομένων, όπως η χρήση του κενό μεταξύ των λέξεων ή η χρήση των σημείων στίξης. Για παράδειγμα η φράση «I love reading books» μπορεί να διαχωριστεί στις ακόλουθες λέξεις: "I", "love", "reading", "books". Με αυτό το βήμα τα δεδομένα κειμένου διασπώνται σε αριθμητική μορφή, η οποία μπορεί στην συνέχεια να χρησιμοποιηθεί από τα μοντέλα μηχανικής μάθησης και νευρωνικών δικτύων για την ανάλυση. Το tokenization επιτρέπει επίσης τον εντοπισμό των πιο σημαντικών λέξεων σε ένα tweet, που μπορούν να χρησιμοποιηθούν ως χαρακτηριστικά για τα μοντέλα.

3. **Αφαίρεση stopwords** : Τα stopwords είναι κοινές λέξεις σε μια γλώσσα που δεν έχουν σημαντικό νόημα και δεν συμβάλλουν στο συνολικό συναίσθημα μιας πρότασης ή ενός εγγράφου. Παραδείγματα τέτοιων λέξεων είναι οι τις λέξεις "the", "a", "an", "in", "of" και "to". Η αφαίρεση αυτών τον λέξεων μειώνει τον αριθμό των λέξεων στο σύνολο δεδομένων και το ποσό των υπολογισμών που απαιτούνται να γίνουν από το μοντέλο. Αυτό βελτιώνει την απόδοση του μοντέλου από άποψη ταχύτητας και αποδοτικότητας.
4. **Lemmatization** : Το Lemmatization είναι μια τεχνική για τον εντοπισμό της βασικής μορφής μιας λέξης, που ονομάζεται λήμμα, λαμβάνοντας υπόψη τα συμφραζόμενα και το μέρος του λόγου της λέξης. Ο lemmatizer λειτουργεί αναλύοντας τη μορφολογία μιας λέξης και αφαιρώντας άλλες τροποποιήσεις , όπως είναι η κατάληξη, το πρόθεμα, οι κλίσεις, οι πτώσεις και οι χρόνο , που δεν είναι απαραίτητες για την αναγνώριση του λήμματος. Για παράδειγμα, το lemmatizer αναγνώριζε ότι οι λέξεις "τρώω", "τρώει", "τρώνε" αντιστοιχούν στο ίδιο λήμμα που είναι το "τρώω". Αυτή η τεχνική είναι χρήσιμη για την ανάλυση συναισθήματος, καθώς βελτιώνει την ακρίβεια του μοντέλου μειώνοντας τον αριθμό των μοναδικών λέξεων στο σύνολο δεδομένων, διατηρώντας παράλληλα τη σημασία των λέξεων.
5. **Stemming** : Το Stemming είναι μια τεχνική που περιλαμβάνει την μείωση των καταλήξεων των λέξεων ώστε όλες οι λέξεις να αποκτήσουν το ίδιο στέλεχος (stem). Η διαδικασία αυτή επιτυγχάνεται με τη χρήση κανόνων, όπως η αφαίρεση των καταλήξεων "- ng", "-ed" και "-s", είτε με τη χρήση πιο σύνθετων αλγορίθμων. Για παράδειγμα, οι λέξεις "running", "runner" και "run" έχουν όλες το ίδιο στέλεχος το "run" και μπορούν να δηλωθούν στη βασική μορφή "run". Αυτή η τεχνική είναι σημαντική καθώς μειώνει τον αριθμό των μοναδικών λέξεων στο σύνολο δεδομένων, βελτιώνοντας έτσι την αποτελεσματικότητα του μοντέλου και να μειώνοντας τον κίνδυνο υπερπροσαρμογής.

Μετά την προ επεξεργασία , παρατηρείται μείωση του όγκου των λέξεων σε κάθε tweet και ειδικότερα φαίνεται ότι κατά μέσο όρο ο αριθμός των λέξεων για κάθε tweet πριν την διαδικασία του καθαρισμού ήταν 29 λέξεις ενώ μετά ανήλθε στις 15 λέξεις. Με αυτό τον τρόπο, η διαδικασία επιλογής λέξεων, που θα αποτελέσουν μετέπειτα τα χαρακτηριστικά για τους αλγορίθμους, καθιστάτε πιο εύκολη.

3.4 Επιλογή χαρακτηριστικών

Η επιλογή χαρακτηριστικών είναι επίσης πολύ σημαντική για την δημιουργία μοντέλων στην συναισθηματική ανάλυση. Υπάρχουν διάφοροι μέθοδοι χαρακτηριστικών και για την συγκεκριμένη εργασία χρησιμοποιήθηκαν οι εξής :

1. **Bag-of-words (BoW)**

Η BoW είναι μια τεχνική η οποία αναπαριστά δεδομένα κειμένου ως αριθμητικά χαρακτηριστικά, τα οποία χρησιμοποιούνται ως είσοδοι σε μοντέλα μηχανικής μάθησης ή νευρωνικών δικτύων. Το πρώτο βήμα είναι η δημιουργίας ενός λεξιλογίου που περιέχει όλες τις μοναδικές λέξεις που υπάρχουν σε ολόκληρο το σώμα των tweets. Στη συνέχεια κάθε tweet αναπαρίσταται ως διάνυσμα μήκους ίσου με το μέγεθος του λεξιλογίου, όπου κάθε στοιχείο του διανύσματος αντιστοιχεί στον αριθμό ή τη συχνότητα μιας συγκεκριμένης λέξης στο tweet. Το διάνυσμα που προκύπτει χρησιμοποιείται στη συνέχεια ως αριθμητική αναπαράσταση του κειμένου. Επομένως, στόχος της BoW είναι η αναπαράσταση ενός έγγραφου κειμένου ως "συλλογή" λέξεων, αγνοώντας τη σειρά και το πλαίσιο των λέξεων στο έγγραφο.

Για την υλοποίηση της τεχνικής BoW με γλώσσα προγραμματισμού Python , έγινε χρήση της συνάρτησης CountVectorizer που ανήκει στην βιβλιοθήκη scikit-learn. Ειδικότερα, η CountVectorizer δέχεται μια λίστα tweets ως είσοδο και εκτελεί διάφορα βήματα προ επεξεργασίας, όπως tokenization, αφαίρεση stopword και stemming. Στη συνέχεια δημιουργεί έναν πίνακα όπου κάθε γραμμή αντιπροσωπεύει ένα tweet και κάθε στήλη αντιπροσωπεύει μια μοναδική λέξη σε ολόκληρο το σώμα. Οι τιμές στον πίνακα αντιπροσωπεύουν τη συχνότητα εμφάνισης κάθε λέξης σε κάθε tweet.

Ένα πλεονέκτημα του BoW είναι η απλότητα και η ευκολία στην ερμηνεία της. Ωστόσο, ένας περιορισμός της είναι ότι αγνοεί τη σειρά και το πλαίσιο των λέξεων στα tweets και δεν αποτυπώνει τις σημασιολογικές σχέσεις μεταξύ των λέξεων. Επιπλέον, εάν το μέγεθος του λεξιλογίου στην BoW είναι μεγάλο τότε παράγεται ένας χώρο χαρακτηριστικών υψηλών διαστάσεων, γεγονός που οδηγήσει σε υπερπροσαρμογή και μειωμένη απόδοση του μοντέλου.

2. **Term Frequency-Inverse Document Frequency (TF-IDF)**

Η TF-IDF είναι μια τεχνική που αναπαριστά δεδομένα κειμένου ως σταθμισμένο άθροισμα συχνοτήτων λέξεων, όπου τα βάρη εξαρτώνται από τη συχνότητα της λέξης στο

tweet και σε ολόκληρο το σώμα κειμένων. Σε αυτή την τεχνική, όσο υψηλότερη είναι η συχνότητα εμφάνισης μιας λέξεις, τόσο χαμηλότερη είναι αξία που έχει. Με άλλα λόγια, εκτός από τη συχνότητα εμφάνισης μιας λέξης, παρέχεται και η σημασία της λέξης. Η μέθοδος αυτή ξεπερνά ορισμένους από τους περιορισμούς της BoW, καθώς δίνει μεγαλύτερη βαρύτητα στις σπάνιες σε ολόκληρο το σώμα κειμένων και μικρότερη βαρύτητα στις κοινές λέξεις που δεν είναι κατατοπιστικές για την ανάλυση συναισθήματος.

Στην Python, η τεχνική υλοποιήθηκε χρησιμοποιώντας τη συνάρτηση TfidfVectorizer της βιβλιοθήκης scikit-learn. Η μέθοδος μετατρέπει μια συλλογή εγγράφων κειμένου, όπως είναι τα tweets, σε έναν αριθμητικό πίνακα χαρακτηριστικών, όπου κάθε έγγραφο αναπαρίσταται ως ένα διάνυσμα αριθμητικών τιμών. Συγκεκριμένα, η TfidfVectorizer στο scikit-learn υπολογίζει τα tf-idf score για κάθε λέξη σε μια συλλογή εγγράφων και επιστρέφει έναν πίνακα όπου κάθε γραμμή αντιπροσωπεύει ένα έγγραφο και κάθε στήλη αντιπροσωπεύει μια μοναδική λέξη. Κάθε στοιχείο του πίνακα αντιπροσωπεύει το tf-idf score για μια συγκεκριμένη λέξη σε ένα συγκεκριμένο έγγραφο. Το tf-idf score είναι ένα στατιστικό μέτρο που αξιολογεί πόσο σχετική είναι μια λέξη με ένα έγγραφο σε μια συλλογή εγγράφων. Υπολογίζεται πολλαπλασιάζοντας το πόσες φορές εμφανίζεται μια λέξη σε ένα έγγραφο και την αντίστροφη συχνότητα εγγράφων της λέξης σε ένα σύνολο εγγράφων. Η αντίστροφη συχνότητα δείχνει την συχνότητα με την οποία εμφανίζεται η λέξη σε όλα τα έγγραφα του σώματος.

Ένα από τα κύρια πλεονεκτήματα αυτής της μεθόδου είναι ότι δεν επηρεάζεται από το μήκος των εγγράφων. Αναλυτικότερα, κατά τη χρήση απλών μετρήσεων συχνότητας λέξεων, τα έγγραφα μεγαλύτερου μήκους έχουν υψηλότερες συχνότητες λέξεων λόγω του μεγάλου αριθμού λέξεων που περιέχουν. Αυτό οδηγεί σε μεροληψία της ανάλυσης προς τα έγγραφα μεγαλύτερο μήκος και καθιστά πιο δύσκολη τη σύγκριση των συχνοτήτων λέξεων μεταξύ εγγράφων με διαφορετικό μήκος. Η μέθοδος TF-IDF αντιμετωπίζει αυτό το πρόβλημα με την κανονικοποίηση των μετρήσεων διαιρώντας τη συχνότητα λέξεων με το συνολικό αριθμό των όρων στο έγγραφο, επιτρέποντας τη σύγκριση μεταξύ εγγράφων διαφορετικού μήκους. Ένα άλλο πλεονέκτημα είναι ότι μειώνει την επίδρασης των stopwords. Κατά τον υπολογισμό της TF-IDF για ένα όρο σε ένα έγγραφο, η συχνότητα του όρου πολλαπλασιάζεται με την αντίστροφη συχνότητα του εγγράφου, η οποία δίνει λιγότερη αξία στους όρους που είναι κοινοί σε όλα τα έγγραφα, συμπεριλαμβανομένων των stopwords, ενώ δίνει περισσότερη αξία στους όρους που είναι σπάνιοι ή μοναδικοί σε ένα συγκεκριμένο έγγραφο.

3. Word embeddings

Σε αντίθεση με τις προσεγγίσεις bag-of-words και TF-IDF, οι οποίες αναπαριστούν κάθε λέξη ως ανεξάρτητο χαρακτηριστικό, η τεχνική Word embeddings λαμβάνει υπόψιν της την σημασιολογική συγγένεια των λέξεων, δηλαδή αποτυπώνει τις σημασιολογικές και συντακτικές σχέσεις μεταξύ των λέξεων σε έναν διανυσματικό χώρο. Η εκμάθηση αυτής της τεχνικής γίνεται μέσω της εκπαίδευσης μοντέλων νευρωνικών δικτύων σε μεγάλο όγκο δεδομένων κειμένου ,όπως είναι τα ειδησεογραφικά άρθρα ή οι ιστοσελίδες. Βασικό της πλεονέκτημα είναι η καταγραφή σημασιολογικών και συντακτικών σχέσεων μεταξύ των λέξεων χωρίς την αύξηση των διαστάσεων, γεγονός που μειώνει την υπολογιστική πολυπλοκότητα.

Δύο δημοφιλείς αλγόριθμοι για την εκμάθηση της είναι οι Word2Vec και GloVe, και είναι αυτοί που εφαρμόστηκαν στα μοντέλα των νευρωνικών δικτύων αυτής της εργασίας. Ακολουθεί συνοπτική περιγραφή αυτών των αλγορίθμων.

3.1 Αλγόριθμος Word2Vec

Ο Word2Vec είναι ένα γλωσσικό μοντέλο νευρωνικού δικτύου που μαθαίνει διανυσματικές αναπαραστάσεις (word embeddings) των λέξεων σε έναν χώρο υψηλών διαστάσεων. Η βασική του ιδέα είναι ότι οι λέξεις που εμφανίζονται σε παρόμοια συμφραζόμενα είναι πιθανό να έχουν παρόμοιες σημασίες. Η εκπαίδευσης του περιλαμβάνει την τροφοδοσία του μοντέλου με μια συλλογή κειμένων (corpus) και την προσαρμογή των βαρών του ώστε να προβλέψει την πιθανότητας των λέξεων δεδομένου ενός συγκεκριμένου πλαισίου. Το πλαίσιο στην ουσία είναι οι λέξεις που βρίσκονται πριν και μετά τη λέξη-στόχο. Το πρώτο βήμα για την εκπαίδευσή του είναι η μετατροπή των δεδομένων κειμένου σε μεμονωμένες λέξεις και η δημιουργία ενός λεξιλογίου που αποτελείται από τις μοναδικές λέξεις στη συλλογή κειμένων. Έπειτα ακολουθεί η κωδικοποίηση του συνόλου των λέξεων χρησιμοποιώντας αναπαράσταση one-hot μεγέθους V, όπου V είναι το μέγεθος του λεξιλογίου, δηλαδή κάθε λέξη αναπαριστάτε ως διάνυσμα που έχει τιμή 1 στη θέση που αντιστοιχεί στο δείκτη της λέξης στο λεξιλόγιο και 0 αλλού. Τελευταίο βήμα, είναι η εκπαίδευση νευρωνικού δικτύου για την πρόβλεψη γειτονικών λέξεων δεδομένης μιας λέξης-στόχου ή για την πρόβλεψη της λέξης-στόχου δεδομένης ενός παραθύρου γειτονικών λέξεων. Ο πρώτος αλγόριθμος είναι ο Skip-Gram ενώ ο δεύτερος είναι ο Continuous Bag of words (CBOW).

Ειδικότερα, ο CBOW δέχεται σαν είσοδο ένα σύνολο λέξεων (context) που αντιστοιχούν σε μια λέξη στόχος, και προσπαθεί να προβλέψει την λέξη στόχο βάση της εισόδου του context. Δηλαδή στόχος της είναι να προβλέψει την τρέχουσα λέξη με βάση τις λέξεις του περιβάλλοντος. Από την άλλη πλευρά, ο Skip-gram δέχεται σαν είσοδο μια λέξη-στόχο που βρίσκεται μέσα σε μια πρόταση. Βάση αυτής, οι λέξεις που την περιβάλλουν, ελέγχονται και επιλέγεται κάποια από αυτές με τυχαίο τρόπο. Επομένως προβλέπει τις λέξεις του περιβάλλοντος δεδομένης της τρέχουσας λέξης. Το αποτέλεσμα είναι η σωστή πιθανότητα για κάθε λέξη του λεξιλογίου δεδομένης της λέξης-στόχου. Αφού εκπαιδευτεί το μοντέλο Word2Vec με έναν από τους δύο αλγορίθμους, δημιουργούνται τα word embeddings που χρησιμοποιούνται ως χαρακτηριστικά εισόδου για μοντέλα ανάλυσης συναισθήματος. Στην Python, το Word2Vec μπορεί να υλοποιηθεί χρησιμοποιώντας τη βιβλιοθήκη gensim.

3.2 Αλγόριθμος Glove

Παρόμοια με το Word2Vec, το GloVe παράγει διανυσματικές αναπαραστάσεις λέξεων που αποτυπώνουν τη σημασιολογική τους σημασία και το περιεχόμενό τους σε ένα δεδομένο σώμα κειμένου. Ο Glove είναι ένας αλγόριθμος μη επιβλεπόμενης μάθησης που δημιουργεί διανυσματικές αναπαραστάσεις για λέξεις και αποτελεί προέκταση του Word2Vec. Ο Glove παράγει διανυσματικές αναπαραστάσεις ενσωματώνοντας την συνολική (global) πληροφορία από σώμα κειμένων για τη δημιουργία word embeddings και όχι τοπικά παράθυρα από συμφραζομένα όπως στο Word2Vec. Συγκεκριμένα, ο GloVe κατασκευάζει έναν πίνακα συν-εμφάνισης (co-occurrence matrix) μεταξύ των λέξεων, όπου κάθε στοιχείο του πίνακα αντιπροσωπεύει τον αριθμό των φορών που η λέξη i εμφανίζεται στο πλαίσιο της λέξης j στο σώμα κειμένων. Στη συνέχεια, ο αλγόριθμος GloVe παραγοντοποιεί αυτόν τον πίνακα χρησιμοποιώντας τεχνικές παραγοντοποίησης πινάκων για τη δημιουργία των word embeddings που αποτυπώνουν το συνολικό πλαίσιο και τη σημασία των λέξεων. Στην Python για να εφαρμοστεί η μέθοδος αρχικά κατέβηκαν διάφορα προ-εκπαιδευμένα GloVe embeddings μοντέλα από την ιστοσελίδα της ομάδας NLP του Stanford [22]. Μετά την φόρτωσή τους, χρησιμοποιήθηκαν για να γίνουν οι αντίστοιχες αναπαραστάσεις των λέξεων στο σύνολό δεδομένων των tweets.

Η χρήση και η σύγκριση διαφορετικών μεθόδων εξαγωγής χαρακτηριστικών αποτελεί κοινή πρακτική στην έρευνα της ανάλυση συναισθήματος, προκυμμένου να γίνει η βελτίωση της απόδοσης των μοντέλων μηχανικής μάθησης και των νευρωνικών δικτύων. Κάθε μέθοδος εξαγωγής χαρακτηριστικών έχει τα δικά της πλεονεκτήματα και μειονέκτημα στην αναπαράσταση δεδομένων κειμένου και η επιλογή της μεθόδου εξαρτάται από τα συγκεκριμένα χαρακτηριστικά του συνόλου δεδομένων,

Αφού έχει πραγματοποιηθεί η προ επεξεργασία των δεδομένων και η επιλογή χαρακτηριστικών ακολουθεί ο διαχωρισμός τους. Στόχος των μοντέλων ταξινόμησης είναι η πρόβλεψη της κατηγορίας συναισθήματος που ανήκει το κάθε tweet. Επομένως, στο συγκεκριμένο πρόβλημα το X ορίζεται το προ επεξεργασμένο tweet και ως y η αντίστοιχη κατηγορία συναισθήματος. Προκειμένου να εφαρμοστούν και να αξιολογηθούν οι αλγόριθμοι απαιτείται ο διαχωρισμός του X και του y σε δυο επιμέρους υποσύνολα. Το πρώτο σύνολο είναι το σύνολο δεδομένων εκπαίδευσης (train set) και το δεύτερό σύνολο είναι το σύνολο ελέγχου (test set).Το πρώτο αποτελείται από το 70% των δεδομένων ενώ το δεύτερο από το υπόλοιπο 30 %. Η επιλογή συγκεκριμένου διαχωρισμού έγινε γιατί είναι ο πιο συνήθης συμφώνα με την βιβλιογραφία.

3.5 Μέτρα αξιολόγησης μοντέλων

Η αξιολόγηση των μοντέλων γίνεται με τον έλεγχο της απόδοσης του εκάστοτε μοντέλου μέσο των δεδομένων εκπαίδευσης. Αποτελεί κρίσιμο βήμα , αφού δείχνει κατά πόσο τα δεδομένα εκπαίδευσης έχουν καταχωρηθεί σωστά και την δυνατότητα του κάθε μοντέλου να λαμβάνει υπόψιν του τα χαρακτηριστικά των κλάσεων και να πραγματοποιεί σωστή πρόβλεψη.

Συνήθως, για την επίδοση των μοντέλων σχετικά με την σωστή πρόβλεψη, χρησιμοποιείται το μετρικό της συνολική ακρίβεια (Accuracy). Η συνολική ακρίβεια υπολογίζεται ως το πηλίκο των σωστά ταξινομημένων κειμένων που ανήκουν στο σύνολο ελέγχου, προς το πληθικό αριθμό του συνόλου ελέγχου. Επομένως, μετρά το ποσοστό των σωστών προβλέψεων που κάνει το μοντέλο επί του συνολικού αριθμού των προβλέψεων. Ωστόσο, η συνολική ακρίβεια μπορεί να είναι παραπλανητική σε περιπτώσεις όπου το σύνολο δεδομένων είναι ανισόρροπο, δηλαδή η μία κλάση έχει σημαντικά περισσότερα παραδείγματα από την άλλη, καθώς το μοντέλο επιτυγχάνει υψηλή ακρίβεια προβλέποντας απλώς την

πλειοψηφούσα κλάση για όλα τα παραδείγματα. Σχετικά με την δυνατότητα του μοντέλου να "μαθαίνει" την κάθε κατηγορία χρησιμοποιούνται διαφορετικά μετρικά τα οποία ονομάζονται ακρίβεια (Precision) και ανάκληση (Recall). Το Precision δείχνει το ποσοστό των σωστών ταξινομήσεων σε μια συγκεκριμένη κατηγορία, δηλαδή αντιπροσωπεύει το ποσοστό των σωστά προβλεπόμενων θετικών (ή αρνητικών) tweets συναισθήματος επί όλων των tweets που προβλέπονται ως θετικά (ή αρνητικά). Το Recall δείχνει το ποσοστό των δειγμάτων που ταξινομήθηκαν σωστά ανάμεσα σε όλα τα δείγμα που ανήκουν σε αυτή την κατηγορία. Στο πλαίσιο της ανάλυσης συναισθήματος, το Recall αντιπροσωπεύει το ποσοστό των σωστά προβλεπόμενων θετικών (ή αρνητικών) συναισθηματικών tweets από όλα τα πραγματικά θετικά (ή αρνητικά) συναισθηματικά tweets στο σύνολο δεδομένων. Ο αρμονικός μέσος των δυο προηγούμενων είναι το F1 score και αποτελεί έναν καλό δείκτη της συνολικής απόδοσης του μοντέλου. Μια υψηλή τιμή υποδεικνύει ότι το μοντέλο έχει καλή ισορροπία μεταξύ ακρίβειας και ανάκλησης, πράγμα που σημαίνει ότι κάνει ακριβείς προβλέψεις, ενώ ταυτόχρονα εντοπίζει ένα μεγάλο ποσοστό θετικών (ή αρνητικών) συναισθηματικών tweets στο σύνολο δεδομένων. Η καλύτερη τιμή που λαμβάνει είναι το 1, ενώ η χειρότερη το 0.

Εκτός από τα παραπάνω μέτρα αξιολόγησης, υπάρχει και ο confusion matrix και η μετρική AUC-ROC που μπορούν επίσης να χρησιμοποιηθούν για την αξιολόγηση των μοντέλων ανάλυσης συναισθήματος. Ο confusion matrix είναι ένας πίνακας που παρέχει μια λεπτομερή ανάλυση της απόδοσης του μοντέλου συγκρίνοντας τις προβλεπόμενες και τις πραγματικές ετικέτες για κάθε κλάση. Ο πίνακας αποτελείται από τέσσερις τιμές:

- True Positive (TP), που είναι ο αριθμός των πραγματικών θετικών δειγμάτων που ταξινομήθηκαν σωστά. Δηλαδή, ο αριθμό των θετικών συναισθηματικών tweets που προβλέφθηκαν σωστά από το μοντέλο.
- False Positive (FP), που είναι ο αριθμός των αρνητικών δειγμάτων που ταξινομήθηκαν λανθασμένα ως θετικά. Δηλαδή, ο αριθμό των αρνητικών συναισθηματικών tweets που προβλέφθηκαν εσφαλμένα ως θετικά.
- False Negative (FN), που είναι ο αριθμός των θετικών δειγμάτων που ταξινομήθηκαν λανθασμένα ως αρνητικά. Δηλαδή, ο αριθμό των θετικών συναισθηματικών tweets που προβλέφθηκαν εσφαλμένα ως αρνητικά.

- True Negative (TN), που είναι ο αριθμός των πραγματικών αρνητικών δειγμάτων που ταξινομήθηκαν σωστά. Δηλαδή, αριθμό των αρνητικών συναισθηματικών tweets που προβλέφθηκαν σωστά από το μοντέλο.

Με βάση αυτές τις τέσσερις τιμές μπορούν να υπολογιστούν:

- Accuracy = (TP + TN) / (TP + FP + TN + FN)
- Recall = TP / (TP + FN)
- Precision = TP / (TP + FP)

Επομένως αυτός ο πίνακας είναι ένα χρήσιμο εργαλείο καθώς παρέχει μια λεπτομερή και ποσοτική ανάλυση των προβλέψεων του μοντέλου. Επιπλέον εντοπίζει συγκεκριμένα μοτίβα στις προβλέψεις του μοντέλου, όπως είναι οι περιπτώσεις όπου ορισμένες κλάσεις ταξινομούνται σταθερά λανθασμένα ή περιπτώσεις όπου υπάρχουν υψηλά ποσοστά ψευδώς θετικών ή ψευδώς αρνητικών αποτελεσμάτων. Οι πληροφορίες αυτές μπορούν να χρησιμοποιηθούν για την καθοδήγηση περαιτέρω ανάλυσης και τη βελτίωση της απόδοσης του μοντέλου.

Το AUC-ROC είναι ένα μέτρο που δείχνει κατά πόσο το μοντέλο είναι σε θέση να διακρίνει σωστά μεταξύ θετικών και αρνητικών συναισθηματικών tweets. Η καμπύλη ROC απεικονίζει το αληθώς θετικό ποσοστό (TPR) έναντι του ψευδώς θετικού ποσοστού (FPR) για διαφορετικά κατώφλια ταξινόμησης. Το Area under Curve έχει οριστεί ως το εμβαδόν κάτω από την καμπύλη ROC, και παρέχει μια συνολική πληροφορία για τον ταξινομητή. Πιο συγκεκριμένα, αν AUC είναι ίσο με 1, σημαίνει ότι η ταξινόμηση είναι τέλεια, αν είναι ίσο με 0.5, τότε γίνεται αντιληπτό ότι ο ταξινομητής δεν είναι ικανός να καταλάβει τις διαφορετικές κλάσεις, δηλαδή η ταξινόμηση έγινε τυχαία.

Στην πειραματική διαδικασία χρησιμοποιήθηκαν από την βιβλιοθήκη sklearn.metrics οι συναρτήσεις accuracy_score, f1_score, precision_score και recall_score. Ως επιμέρους παράμετροι σε αυτές τις συναρτήσεις υπήρχε το average όπου επιλέχθηκε το “macro” , δηλαδή το macro averaging score. Η επιλογή αυτή οφείλεται στο γεγονός ότι τα δεδομένα δεν παρουσιάζουν μεγάλη ανισορροπία. Τέλος, για την μετρική AUC-ROC και τον confusion matrix εφαρμόστηκαν αντίστοιχα η roc_auc_score και confusion_matrix από το sklearn.metrics.

4 Αποτελέσματα

Στην παρούσα ενότητα παρουσιάζεται η σύγκριση των επιδόσεων των διαφόρετικών μοντέλων μηχανικής μάθησης και νευρωνικών δικτύων για την ανάλυση συναισθήματος σε δεδομένα Twitter. Οι μετρικές αξιολόγησης που χρησιμοποιήθηκαν είναι αυτές που αναφέρονται στην ενότητα 3.1.4. Τα αποτελέσματα αναλύονται και συγκρίνονται για τον εντοπισμό του πιο αποτελεσματικού μοντέλου για την ανάλυση συναισθήματος σε δεδομένα Twitter.

4.1 Εφαρμογή αλγορίθμων ταξινόμησης μηχανικής μάθησης

Σε αυτή την ενότητα παρουσιάζονται τα αποτελέσματα και οι ερμηνείες από την εφαρμογή των εξής αλγορίθμων :

- Naive Bayes Multinomial Classifier
- Naïve Bayes Complement Classifier
- Random Forest Classifier
- Linear Support Vector Machine
- Gradient Boosting Classifier

Σε κάθε αλγόριθμο τα δεδομένα εξήχθησαν πρώτα με βάση την μέθοδο CountVectorizer και έπειτα με την μέθοδο TF-IDF.

4.1.1 Naïve Bayes Multinomial Classifier

Η πρώτη μέθοδος που εφαρμόστηκε είναι η Naive Bayes Multinomial Classifier (NBM). Τα αποτελέσματα με την μέθοδο CountVectorizer μελετήθηκαν και αξιολογήθηκαν βάση των τιμών μέτρων που αναφέρθηκαν προηγουμένως.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	91%	92 %	92 %	87.21%
Positive	74%	72%	73%	

Πίνακας 4.1: Μέτρα αξιολόγησης του NBM (CountVectorizer)

	Predicted label	
True label	Negative	Positive
Negative	33.351	2.883
Positive	3.194	8.079

Πίνακας 4.2 : Confusion πίνακάς του NBM (CountVectorizer)

Βλέποντας τα παραπάνω είναι φανερό ότι έχει επιτευχθεί ένα ικανοποιητικό ποσοστό ακρίβειας του μοντέλου στις προβλέψεις. Με βάση αυτές τις τιμές, φαίνεται ότι το μοντέλο ταξινόμησε σωστά 8.079 tweets ως θετικά (TP) και 33.351 tweets ως αρνητικά (TN). Ωστόσο, ταξινόμησε εσφαλμένα 3.194 θετικά tweets ως αρνητικά (FN) και 2.883 αρνητικά tweets ως θετικά (FP). Για να αξιολογήσετε την απόδοση του μοντέλου , ελέγχεται το Accuracy που είναι 87,17 % που σημαίνει ότι και για τις δυο κλάσεις (Negative, Positive) το μοντέλο πρόβλεψε σωστά τα 41.430 από τα συνολικά 47.507 tweets. Όμως, το Accuracy από μόνο του μπορεί να μην είναι επαρκές για τον προσδιορισμό της αποτελεσματικότητας του μοντέλου. Έτσι υπολογίστηκε το Precision ως εξής:

$$\text{Precision} = TP/(TP+FP) = 8.079/ (8.079+2.883) = 0.737 \approx 0.74$$

που σημαίνει ότι από τα 10.962 tweets που προβλέφθηκαν σαν θετικά, το 74% αφορούσαν θετικά tweets, περίπου τα 8.079 tweets, και το 26% αφορούσαν αρνητικά tweets, δηλαδή τα 2.883 . Το Recall αντίστοιχά υπολογίστηκε ως εξής:

$$\text{Recall} = TP / (TP + FN) = 8{,}079 / (8{,}079 + 3{,}194) = 0.717 \approx 0.72$$

που σημαίνει ότι από 11.273 tweets που ήταν πραγματικά θετικά τα 8.079 προβλέφθηκαν ως πραγματικά θετικά. Το F1 score για την κατηγορία των αρνητικών tweets είναι 92%, υποδεικνύοντας ότι το μοντέλο πέτυχε μια καλή ισορροπία μεταξύ Precision και Recall για τα αρνητικά tweets. Το F1 score για την κλάση των θετικών tweets είναι 73%, υποδεικνύοντας ότι το μοντέλο έχει χαμηλότερη ισορροπία μεταξύ Precision και Recall για τα θετικά tweets. Με βάση τις παρεχόμενες πληροφορίες, το μοντέλο φαίνεται να είναι αρκετά καλό στην πρόβλεψη του συναισθήματος από ένα tweet, με υψηλό Accuracy, Precision, Recall και F1 score για τα αρνητικά tweets, αλλά χαμηλότερη Precision, Recall και F1 score για τα θετικά tweets.

Ο ίδιος αλγόριθμος εφαρμόστηκε ξανά με μόνη διαφορά ότι τα δεδομένα εξήχθησαν με βάση τη μέθοδο TF-IDF. Τα αποτελέσματα παρουσιάζονται στην συνέχεια:

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	80%	100 %	89 %	80.55%
Positive	98%	18%	31%	

Πίνακας 4.3: Μέτρα αξιολόγησης του NBM (TD-IDF)

	Predicted label	
True label	Negative	Positive
Negative	36.197	37
Positive	9.202	2.071

Πίνακας 4.4: Confusion πίνακάς του NBM (TD-IDF)

Παρατηρείται πτώση στο Accuracy του μοντέλου που πλέον είναι 80.55 %.Το Precision είναι στα 98% που δηλώνει ότι από 2.108 tweets που προβλέφθηκαν σαν θετικά, τα 2.071 αφορούσαν θετικά tweets. Επιπλέον το Recall 18% που σημαίνει ότι από τα συνολικά 11.273 θετικά tweets , ως θετικά προβλέφθηκαν μόνο τα 2071, πράγμα που σημαίνει ότι δεν είναι καλό στο να αναγνωρίζει σωστά όλα τα θετικά tweets. Από την άλλη πλευρά, έχει τέλειο Recall για την κατηγορία των αρνητικών tweets υποδεικνύοντας ότι είναι καλό στο να αναγνωρίζει σωστά τα αρνητικά tweets. Παρατηρείται μεγάλη αστοχία στην ικανότητα πρόβλεψης των θετικών tweets και για αυτό μπορεί να μην είναι ένα καλό μοντέλο για την πρόβλεψη του συναισθήματος από ένα tweet.

4.1.2 Naïve Bayes Complement Classifier

Ο επόμενος αλγόριθμος που εφαρμόστηκε ο Complement Naive Bayes (NBC) ταξινομητή. Τα αποτελέσματα που ακολουθούν αναφέρονται στην μέθοδο CountVectorizer.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	93%	87%	90 %	85.28%
Positive	66%	80%	72%	

Πίνακας 4.5: Μέτρα αξιολόγησης του NBC (CountVectorizer)

	Predicted label	
True label	Negative	Positive
Negative	31.515	4.719
Positive	2.272	9.001

Πίνακας 4.6: Confusion πίνακας του NBC (CountVectorizer)

Το Accuracy του αλγορίθμου είναι 85.28 % δηλαδή από συνολικά 47.507 tweets τα 40.516 προβλέφθηκαν σωστά. Το Precision είναι στα 66% που σημαίνει ότι από 13.720 tweets που προβλέφθηκαν ως θετικά, τα 9.001 αφορούσαν θετικά tweets. Το Recall είναι 80 % που σημαίνει ότι από τα συνολικά 11.273 θετικά tweets, ως θετικά προβλέφθηκαν μόνο τα 9.001. Φαίνεται το Precision και το Recall για τις δύο κλάσεις είναι επίσης σχετικά υψηλές, με F1 score 90% για την κλάση των αρνητικών tweets και 72% για την κλάση των θετικών tweets. Επιπλέον, ο confusion πίνακας παρουσιάζει σχετικά χαμηλά FP και FN, υποδεικνύοντας ότι το μοντέλο έχει καλή απόδοση στον ορθό εντοπισμό τόσο του θετικού όσο και του αρνητικού συναισθήματος στα tweets.

Εφαρμόστηκε ο ίδιος αλγόριθμος με μόνη διαφορά ότι τα δεδομένα εξήχθησαν με βάση τη μέθοδο TF-IDF. Τα αποτελέσματα παρουσιάζονται στην συνέχεια:

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	87%	97 %	92 %	86.74%
Positive	86%	53%	65%	

Πίνακας 4.7: Μέτρα αξιολόγησης του NBC (TF-IDF)

	Predicted label	
True label	Negative	Positive
Negative	35.238	996
Positive	5.302	5.971

Πίνακας 4.8: Confusion πίνακας του NBC (TF-IDF)

Το Accuracy είναι 86.74% από συνολικά 47.507 tweets τα 41.209 προβλέφθηκαν σωστά. Το Precision είναι στα 86% που σημαίνει ότι από 6.967 tweets που προβλέφθηκαν ως θετικά, τα 5.971 αφορούσαν θετικά tweets. Το Recall είναι 53 % που σημαίνει ότι από τα συνολικά 11.273

θετικά tweets ,ως θετικά προβλέφθηκαν μόνο τα 5.971. Ωστόσο, η απόδοση για την κατηγορία με θετικό συναίσθημα είναι χαμηλότερη από την κατηγορία με το αρνητικό συναίσθημα, με χαμηλότερο Recall και F1 score.

4.1.3 Random Forest Classifier

Ο επόμενος αλγόριθμος για ταξινόμησης που εξετάστηκε είναι Random Forest (RFC). Αναφέρεται ότι κατά την εφαρμογή του χρησιμοποιήθηκαν οι εξής παράμετροι :

- n_estimators=50
- min_samples_split = 5
- min_samples_leaf = 8
- oob_score= True
- n_jobs=4

Ειδικότερα, η παράμετρος n_estimators καθορίζει τον αριθμό των δειγμάτων με τα οποία θα λειτουργήσει ο αλγόριθμος και στην συνέχεια θα συγκεντρώσει για να δώσει το τελικό αποτέλεσμα. Η παράμετρος min_samples_split ορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτούνται για την διαίρεση ενός εσωτερικού κόμβου. Αντίστοιχα, το min_samples_leaf ορίζει τον ελάχιστο αριθμό δειγμάτων που απαιτείται να βρίσκονται σε ένα κόμβο φύλλου. Το n_jobs είναι ο αριθμός των εργασιών που θα εκτελεστούν παράλληλα για προσαρμογή και πρόβλεψη. Τέλος oob_score που είναι true σημαίνει ότι χρησιμοποιήθηκαν δείγματα out-of-bag για την εκτίμηση της απόδοσης. Τα αποτελέσματα που ακολουθούν αναφέρονται στην μέθοδο CountVectorizer.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	83%	99%	90 %	83.07%
Positive	88%	33%	48%	

Πίνακας 4.9: Μέτρα αξιολόγησης του RFC (CountVectorizer)

	Predicted label	
True label	Negative	Positive
Negative	35.734	500
Positive	7.545	3.728

Πίνακας 4.10: Confusion πίνακας του RFC (CountVectorizer)

Το Accuracy είναι 83.07% από συνολικά 47.507 tweets τα 39.462 προβλέφθηκαν σωστά. Το Precision είναι στα 88% που σημαίνει ότι από 4.228 tweets που προβλέφθηκαν ως θετικά, τα 3.728 αφορούσαν θετικά tweets. Το Recall είναι 33 % που σημαίνει ότι από τα συνολικά 11.273 θετικά tweets, ως θετικά προβλέφθηκαν μόνο τα 3.728. Το F1 score για την κατηγορία των θετικών tweets είναι αρκετά χαμηλό (48%) και το Recall είναι επίσης αρκετά χαμηλό (33%), υποδεικνύοντας ότι το μοντέλο δεν αναγνωρίζει πολύ καλά τα θετικά tweets. Ωστόσο, το Precision για την κατηγορία των αρνητικών tweets είναι σχετικά υψηλή (88%), πράγμα που σημαίνει ότι όταν το μοντέλο προβλέπει ότι ένα tweet είναι θετικό, είναι συνήθως σωστό.

Τα αποτελέσματα που ακολουθούν αναφέρονται στην μέθοδο TF-IDF.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	82%	99 %	90 %	82.53%
Positive	93%	29%	44%	

Πίνακας 4.11: Μέτρα αξιολόγησης του RFC (TF-IDF)

	Predicted label	
True label	Negative	Positive
Negative	35.981	253
Positive	8.046	3.227

Πίνακας 4.12: Confusion πίνακας του RFC (TF-IDF)

Το Precision αυξήθηκε σχετικά με πριν και από 88% είναι 93%. Αυτό σημαίνει ότι από 3.480 tweets που προβλέφθηκαν ως θετικά, τα 3.227 αφορούσαν θετικά tweets. Το Recall όμως μειώθηκε και από 33 % που ήταν πλέον είναι 29% και σημαίνει ότι από τα συνολικά 11.273 θετικά tweets ,ως θετικά προβλέφθηκαν μόνο τα 3.227. Με βάση τις μετρικές που παρέχονται, φαίνεται ότι το μοντέλο έχει υψηλό Precision και για τις δύο κλάσεις, αλλά πολύ χαμηλό Recall

για την κλάση με τα θετικά tweets. Αυτό σημαίνει ότι το μοντέλο είναι πολύ καλό στον σωστό εντοπισμό αρνητικών tweets, αλλά του διαφεύγουν πολλά θετικά tweets.

4.1.4 Support Vector Machine Linear Classifier

Ένας άλλος γνωστός αλγόριθμος που εξετάστηκε είναι ο γραμμικός SVM. Όπως αναφέρθηκε είναι ο γραμμικός SVM επομένως ως kernal μπήκε ο linear. Επιπλέον, η παράμετρο C αποτελεί την ποινή του όρου σφάλματος. Πραγματοποιώντας διάφορες δοκιμές , τελικά επιλέχτηκε ως C η τιμή 0.1.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	88%	96%	92 %	87.45%
Positive	83%	59%	69%	

Πίνακας 4.13: Μέτρα αξιολόγησης του SVM (CountVectorizer)

	Predicted label	
True label	Negative	Positive
Negative	34.850	1.384
Positive	4.579	6.694

Πίνακας 4.14: Confusion πίνακας του SVM (CountVectorizer)

Το Precision είναι 83% και δηλώνει ότι από τα 8.078 tweets που προβλέφθηκαν ως θετικά, τα 6.694 αφορούσαν θετικά tweets, γεγονός που υποδηλώνει ότι το μοντέλο έχει χαμηλό ποσοστό FP αποτελεσμάτων. Το Recall είναι 59% και σημαίνει ότι από τα συνολικά 11.273 θετικά tweets, ως θετικά προβλέφθηκαν τα 6.694, υποδεικνύοντας ότι το μοντέλο δυσκολεύεται να εντοπίσει ορισμένα θετικά tweets.

Τα αποτελέσματα που ακολουθούν αναφέρονται στην μέθοδο TF-IDF.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	80%	100%	89 %	80.87%
Positive	98%	20%	33%	

Πίνακας 4.15: Μέτρα αξιολόγησης του SVM (TF-IDF)

	Predicted label	
True label	Negative	Positive
Negative	36.194	40
Positive	9.049	2.224

Πίνακας 4.16: Confusion πίνακας του SVM (TF-IDF)

Συγκριτικά με πριν παρατηρείται μεγάλη πτώση στο Recall, που είναι 20% . Αυτό δηλώνει ότι από τα συνολικά 11.273 θετικά tweets, ως θετικά προβλέφθηκαν μόνο τα 2.224. Επιπλέον, το F1 score για την κλάση των θετικών tweets είναι μόνο 33%, που είναι αρκετά χαμηλό και υποδηλώνει κακή απόδοση για την ταξινόμηση θετικού συναισθήματος. Φαίνεται ότι το μοντέλο δεν αποδίδει καλά στην πρόβλεψη του συναισθήματος των tweets, ιδίως για την κατηγορία των θετικών tweets.

4.1.5 Grandian Boosting Classifier

Ο τελευταίος αλγόριθμος ταξινόμησης που χρησιμοποιήθηκε είναι ο Gradient Boosting (GBC). Κατά την εφαρμογή του έχει χρησιμοποιηθεί η παράμετρος n_estimators =100, είναι ο αριθμός των δέντρων που υπάρχουν στο μοντέλο. Όσο μεγαλύτερος είναι ο αριθμός τόσο καλύτερα μαθαίνει το μοντέλο τα δεδομένα. Ωστόσο, αυξάνει πολύ τον χρόνο εκπαίδευσης του μοντέλου. Τα αποτελέσματα που ακολουθούν αναφέρονται στην μέθοδο CountVectorizer.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	86%	95%	90%	83.97%
Positive	74%	50%	60%	

Πίνακας 4.17: Μέτρα αξιολόγησης του GBC (CountVectorizer)

	Predicted label	
True label	Negative	Positive
Negative	34.292	1.942
Positive	5.675	5.598

Πίνακας 4.18: Confusion πίνακας του GBC (CountVectorizer)

Το Precision είναι 74% και δηλώνει ότι από τα 7.538 tweets που προβλέφθηκαν ως θετικά, τα 5.598 αφορούσαν θετικά tweets. Το Recall είναι 50% και σημαίνει ότι από τα συνολικά 11.273

θετικά tweets, ως θετικά προβλέφθηκαν τα 5.598. Το F1 score για την κλάση των αρνητικών tweets είναι 90%, που είναι σχετικά υψηλή, ενώ το F1 score για την κλάση των θετικών tweets είναι 60%, που είναι χαμηλότερο από τα F1 score ορισμένων άλλων μοντέλων.

Τα αποτελέσματα που ακολουθούν αναφέρονται στην μέθοδο TF-IDF.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	85%	93%	89 %	82.94%
Positive	70%	49%	82%	

Πίνακας 4.19: Μέτρα αξιολόγησης του GBC (TF-IDF)

	Predicted label	
True label	Negative	Positive
Negative	33.836	2.390
Positive	5.707	5.566

Πίνακας 4.20: Confusion πίνακας του GBC (TF-IDF)

Το Precision είναι 70% και δηλώνει ότι από τα 7.956 tweets που προβλέφθηκαν ως θετικά, τα 5.566 αφορούσαν θετικά tweets. Το Recall είναι 49% και σημαίνει ότι από τα συνολικά 11.273 θετικά tweets, ως θετικά προβλέφθηκαν τα 5.566. Αυτό δηλώνει ότι το μοντέλο δεν αναγνωρίζει όσα θετικά tweets θα έπρεπε.

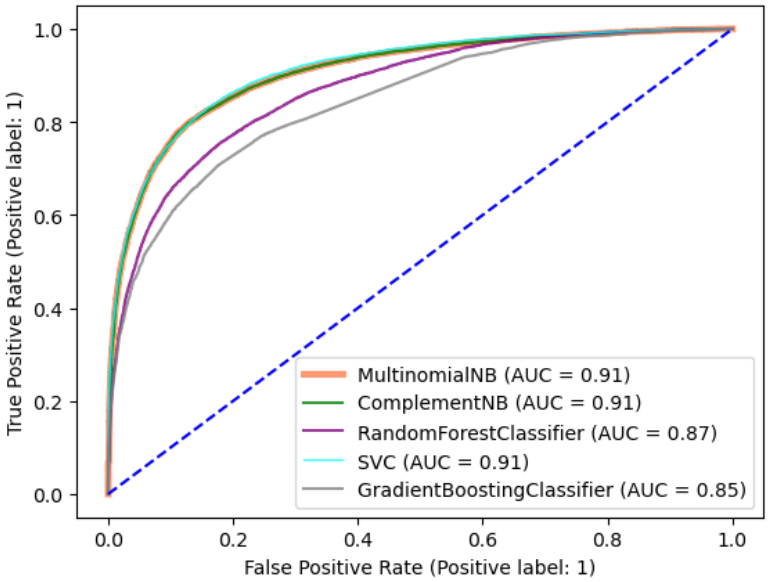
4.1.6 Συγκέντρωση αποτελεσμάτων

Ακολουθεί ο συγκεντρωτικός πίνακας των αποτελεσμάτων των αλγορίθμων με την μέθοδο CountVectorizer. Στο Γράφημα 4.1 φαίνεται οι καμπύλες ROC και οι τιμές του AUC για κάθε ένα από τα μοντέλα που εφαρμόστηκαν προηγουμένως.

	Precision score		Recall score		F1- score		Test Accuracy	Macro Average Precision score	Macro Average Recall score	Macro Average F1 – score
Class	Negative	Positive	Negative	Positive	Negative	Positive				
NBM	91%	74%	92 %	72%	92 %	73%	87.21%	82.48%	81.86%	82.16%
NBC	93%	66%	87%	80%	90 %	72%	85.28%	79.44%	83.41%	81.02%
RFC	83%	88%	99%	33%	90 %	48%	83.07%	85.60%	65.59%	68.70%

SVM	88%	83%	96%	59%	92 %	69%	87.45%	85.63%	77.78%	80.65%
GBC	86%	74%	95%	50%	90%	60%	83.97%	80.02%	72.15%	74.76%

Πίνακας 4.21: Συγκεντρωτικός πίνακας των μέτρων αξιολόγησης των αλγορίθμων (CountVectorizer)



Γράφημα 4.1: Καμπύλη Roc και τιμές AUC(CountVectorixer)

Κοιτάζοντας το Γράφημα 4.1 φαίνεται ότι η μετρική περιοχή κάτω από την καμπύλη (AUC) είναι σχετικά υψηλή σε όλα τα μοντέλα και κυμαίνεται από 85% έως 91%, που σημαίνει ότι μπορούν να αντιληφθούν τις διαφορές μεταξύ των δυο κλάσεων. Ωστόσο, οι επιδόσεις των μοντέλων ποικίλλουν σε διάφορες μετρικές αξιολόγησης, γεγονός που υποδηλώνει ότι ορισμένα μοντέλα μπορεί να είναι καταλληλότερα για αυτή την εργασία από άλλα.

Με βάση τον Πίνακα 4.21 είναι φανερό ότι η απόδοση κάθε μοντέλου ποικίλλει στις διάφορες μετρήσεις και κανένα μοντέλο δεν φαίνεται να υπερέχει έναντι των άλλων σε όλες τις μετρικές. Αναλυτικότερα, φαίνεται ότι όλα τα μοντέλα πέτυχαν υψηλότερα Precision για την κατηγορία 1 (αρνητικά tweets) από ό,τι για την κατηγορία 2 (θετικά tweets). Το NBC πέτυχε το υψηλότερο Precision για την κατηγορία 1 (93%), ακολουθούμενο από το NBM (91%). Για την κατηγορία 2, το RFC το υψηλότερο Precision (88%) ακολουθούμενο από το SVM (83%). Οι υψηλές βαθμολογίες για το Precision υποδηλώνουν ότι τα μοντέλα έχουν

χαμηλά ποσοστά ψευδώς θετικών αποτελεσμάτων, πράγμα που σημαίνει ότι αναγνωρίζουν με ακρίβεια τα αρνητικά tweets ως αρνητικά και τα θετικά tweets ως θετικά.

Με βάση τον Πίνακα 4.21 φαίνεται ότι όλα τα μοντέλα πέτυχαν υψηλότερη βαθμολογία Recall για την κατηγορία 1 από ό,τι για την κατηγορία 2. Το RFC πέτυχε την υψηλότερη βαθμολογία Recall για την κλάση 1 (99%), ακολουθούμενο από το SVM (96%). Για την κλάση 2, το NBC πέτυχε το υψηλότερο Recall (80%) ακολουθούμενο από το NBC (72%). Ωστόσο, στην ίδια κλάση εντοπίζεται και το χαμηλότερο Recall για το RFC που δηλώνει την αδυναμία του στον εντοπισμό των θετικών tweets. Οι υψηλές βαθμολογίες του Recall υποδηλώνουν ότι τα μοντέλα έχουν χαμηλό ποσοστό ψευδώς αρνητικών, πράγμα που σημαίνει ότι αναγνωρίζουν με ακρίβεια τα αρνητικά tweets ως αρνητικά και τα θετικά tweets ως θετικά.

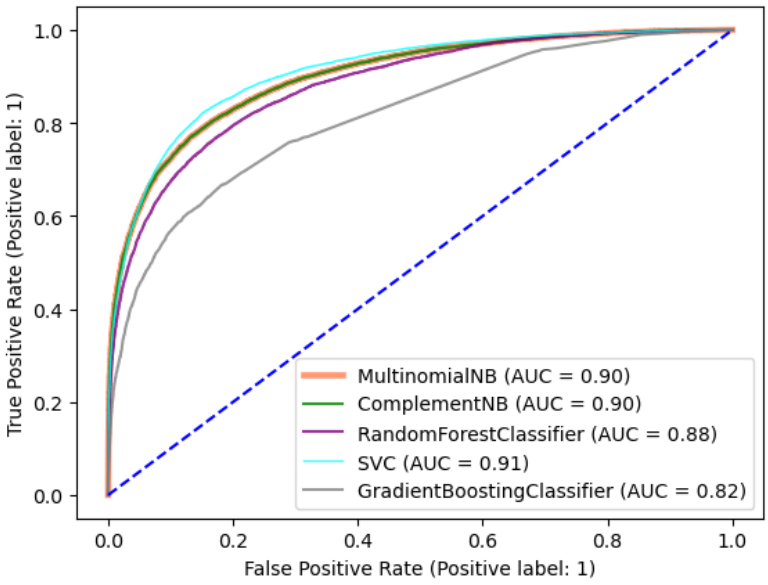
Το F1-score παρέχει ένα συνολικό μέτρο της απόδοσης του μοντέλου ενώ το Accuracy το ποσοστό των σωστά ταξινομημένων περιπτώσεων στο σύνολο δοκιμής. Από τον Πίνακα 4.1.12 φαίνεται ότι το NBM πέτυχε το υψηλότερο F1-score (82.16%) που επαληθεύεται και από τα επιμέρους F1-score για την κατηγορία 1 (92%) και για την κατηγορία 2 (73%).Οι υψηλές βαθμολογίες F1 για την κλάση 1 δείχνουν ότι τα μοντέλα ταξινομούν με ακρίβεια τα αρνητικά tweets, ενώ οι χαμηλότερες βαθμολογίες F1 για την κλάση 2 υποδηλώνουν ότι υπάρχει περιθώριο βελτίωσης στον εντοπισμό θετικών tweets. Τέλος, το SVM πέτυχε το υψηλότερο Accuracy (87,45%) και ακολούθησε το NBM (87,21%).

Λαμβάνοντας υπόψιν όλα τα παραπάνω φαίνεται πως συνολικά το μοντέλο NBM είναι μια καλή επιλογή για την ανάλυση συναισθήματος. Αυτό οφείλεται στο γεγονός ότι έχει υψηλότερό Precision την κατηγορία 1 (91%) σε σύγκριση με όλα τα άλλα μοντέλα, πράγμα που σημαίνει ότι όταν το μοντέλο προβλέπει σωστά ένα tweet ως αρνητικό. Μπορεί το Precision για την κλάση 2 (74%) να είναι ελαφρώς χαμηλότερό από ορισμένα άλλα μοντέλα, ωστόσο για την ίδια κλάση έχει το υψηλότερο Recall. Το Recall για την κλάση 1 είναι εξίσου υψηλό (92%) και το F1 score του μοντέλου είναι το υψηλότερο. Τέλος, τα μοντέλα NBM είναι γνωστό ότι λειτουργούν καλά σε εργασίες ταξινόμησης κειμένου με διακριτά χαρακτηριστικά, όπως ο αριθμός των λέξεων, κάτι που συμβαδίζει με τα αποτελέσματα της ανάλυση συναισθήματος.

Ακολουθεί ο συγκεντρωτικός πίνακας των αποτελεσμάτων των αλγορίθμων με την μέθοδο TF-IDF και στο Γράφημα 4.2 φαίνεται η καμπύλη ROC και η τιμή του AUC.

	Precision score		Recall score		F1- score		Test Accuracy	Macro Average Precision score	Macro Average Recall score	Macro Average F1 – score
Class	Negative	Positive	Negative	Positive	Negative	Positive				
NBM	80%	98%	100 %	18%	89 %	31%	80.55%	88.99%	59.13%	59.82%
NBC	87%	86%	97 %	53%	92 %	65%	86.74%	86.31%	75.11%	78.63%
RFC	82%	93%	99 %	29%	90 %	44%	82.53%	87.58%	64.78%	67.77%
SVM	80%	98%	100%	20%	89 %	33%	80.87%	89.12%	59.81%	60.85%
GBC	85%	70%	93%	49%	89 %	82%	82.94%	77.73%	71.38%	73.59%

Πίνακας 4.22: Συγκεντρωτικός πίνακας των μέτρων αξιολόγησης των αλγορίθμων (TF-IDF)



Γράφημα 4.2: Καμπύλη Roc και τιμές AUC(TF-IDF)

Συγκριτικά με την προηγούμενη μέθοδο φαίνεται ότι η απόδοση των μοντέλων ήταν πιο μικτή με την χρήση της μεθόδου TF-IDF. Ορισμένα μοντέλα έχουν υψηλότερό Precision και Recall και για τις δύο κλάσεις (π.χ. RFC), ενώ άλλα έχουν υψηλότερό Precision για τη μία κλάση και υψηλότερο Recall για την άλλη (π.χ. SVM και NBM). Συνολικά για όλα τα μοντέλα, το Macro Average Precision είναι υψηλότερο ενώ το Accuracy, Macro Average Recall, Macro Average F1 score είναι χαμηλότερα σε σύγκριση με την προηγούμενη μέθοδο.

Εξετάζοντας τα αποτελέσματα από τον Πίνακα 4.22 φαίνεται ότι το Precision του μοντέλου NBC έχει την υψηλότερη τιμή για την κατηγορία 1 με 87%. Από την άλλη πλευρά, τα μοντέλα NBM και SVM έχουν την υψηλότερη τιμή για την κατηγορία 2 με 98%. Το Recall των μοντέλων NBM και SVM έχουν την υψηλότερη τιμή για την κλάση 1 στο 100%, ενώ το μοντέλο NBC έχει την υψηλότερη τιμή για την κλάση 2 στο 53%. Το μοντέλο NBC έχει το υψηλότερο F1 score για την κατηγορία 1 σε 92%, ενώ το μοντέλο GBC έχει το υψηλότερο F1 score για την κατηγορία 2 σε 65%.Το μοντέλο NBC έχει το υψηλότερο Accuracy ενώ το μοντέλα SVM έχει την υψηλότερη AUC (91%) και με μικρή διαφορά ακολουθούν τα NBM και NBC με 90%.

Λαμβάνοντας όλα τα παραπάνω υπόψιν, μοντέλο NBC φαίνεται να είναι η καλύτερη επιλογή μεταξύ των συγκεκριμένων μοντέλων για τη συγκεκριμένη εργασία. Ένας λόγος γι' αυτό είναι ότι το μοντέλο NBC έχει το υψηλότερο F1 score τόσο για την κλάση όσο και για την κλάση 2 , υποδεικνύοντας ότι έχει καλές επιδόσεις όσον αφορά το Precision και το Recall και για τις δύο κλάσεις. Επιπλέον, το μοντέλο NBC έχει σχετικά υψηλό Accuracy (86.74%) και AUC (90%), υποδεικνύοντας ότι μπορεί να διακρίνει αποτελεσματικά μεταξύ αρνητικών και θετικών tweets. Τέλος, το NBC είναι ότι είναι σχετικά απλό και υπολογιστικά αποδοτικό μοντέλο σε σύγκριση με άλλα μοντέλα όπως το GBC.

Συνολικά παρατηρείται ότι η μέθοδος επιλογής χαρακτηριστικών TF-IDF έχει χαμηλότερη απόδοση συγκριτικά με την Bow. Αυτό οφείλεται στο γεγονός ότι τα tweets είναι μικρά σε έκταση και με την μέθοδο TF-IDF προστίθεται επιπλέον θόρυβος ο οποίος δεν συμβάλει στον εντοπισμό των κατάλληλων χαρακτηριστικών.

4.2 Εφαρμογή νευρωνικών δικτύων

Σε αυτή την ενότητα δημιουργήθηκαν δύο μοντέλα με διαφορετική αρχιτεκτονική. Τα μοντέλα αυτά εκπαιδεύτηκαν με δυο διαφορετικούς αλγορίθμους ο ένας εκ των οποίον είναι ο Word2Vec και ο άλλος ο Glove. Για την εκπαίδευση των μοντέλων χρησιμοποιείται η μέθοδος fit. Για όλα τα μοντέλα κατά την εκπαίδευση επιλέχτηκε batch_size ίσο με 1024. Αυτό σημαίνει ότι 1204 δείγματα από το σύνολο των δεδομένων εκπαίδευσης χρησιμοποιήθηκαν σε μία επανάληψη. Επιπλέον ορίστηκε το epoch ίσο με 5, όπου το ένα epoch αναφέρεται στην εκτέλεση μιας πλήρους διαδικασίας εκπαίδευσης στο σύνολο των δεδομένων εκπαίδευσης.

Κατά τη διάρκεια ενός epoch, το μοντέλο επεξεργάζεται και ενημερώνεται με κάθε ζεύγος δεδομένων εισόδου-εξόδου στο σύνολο εκπαίδευσης. Ως loss συνάρτηση δηλώθηκε binary_ crossentropy και ως optimizer , που είναι ο αλγόριθμός που ενημερώνει τα βάρη και τα bias του μοντέλου προκειμένου να ελαχιστοποιηθεί η συνάρτηση απώλειας, ο adam.

Για την απόδοση των μοντέλων σχετικά με την ταξινόμηση χρησιμοποιούνται τα μέτρα που αναφέρονται στη παράγραφο 3.1.4 .Επιπλέον προθέτονται οι καμπύλες Accuracy και Loss κατά την διάρκεια της εκπαίδευσης. Η πρώτη δείχνει πόσο καλά το μοντέλο είναι σε θέση να ταξινομήσει σωστά τα δεδομένα σε κάθε epoch. Είναι ένα μέτρο του ποσοστού των σωστά ταξινομημένων δεδομένων. Καθώς η ακρίβεια αυξάνεται, υποδεικνύει ότι το μοντέλο μαθαίνει τα μοτίβα στα δεδομένα και βελτιώνει την απόδοσή του. Η δεύτερη δείχνει πόσο καλά το μοντέλο ελαχιστοποιεί το σφάλμα του κατά τη διάρκεια της εκπαίδευσης. Αντιπροσωπεύει τη διαφορά μεταξύ της προβλεπόμενης εξόδου και της πραγματικής εξόδου του μοντέλου. Όσο η απώλεια μειώνεται, σημαίνει ότι το μοντέλο βελτιώνει τις προβλέψεις του και μαθαίνει τα μοτίβα στα δεδομένα. Ο συνδυασμός αυτών των καμπυλών συμβάλει στην κατανόηση της κατεύθυνσης με την οποία μαθαίνει το νευρωνικό μοντέλο, δηλαδή αν υπάρχει overfitting, underfitting ή επιτυγχάνει καλά αποτελέσματα.

4.2.1 Δημιουργία και αποτελέσματα πρώτου μοντέλου

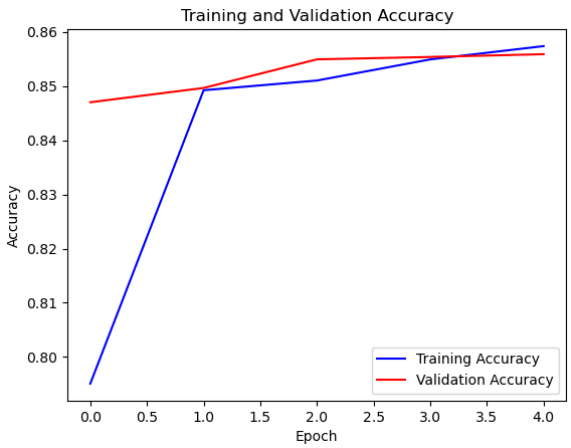
Το πρώτο μοντέλο νευρωνικού δικτύου είναι ένας τύπος RNN και η δομή του φαίνεται στην Εικόνα 4.1. Συγκεκριμένα το πρώτο επίπεδο του μοντέλου είναι το embedding επίπεδο, το οποίο αντιστοιχεί κάθε λέξη στο κείμενο εισόδου σε μια διανυσματική αναπαράσταση σε έναν χώρο χαμηλότερης διάστασης. Αυτό επιτυγχάνεται με τη χρήση προ-εκπαιδευμένων word embeddings,, οι οποίες λαμβάνονται από ένα μεγάλο σώμα δεδομένων κειμένου και αποτυπώνουν τη σημασιολογική σημασία των λέξεων. Το δεύτερο επίπεδο του μοντέλου είναι ένα bidirectional LSTM επίπεδο, το οποίο είναι ένας τύπος αναδρομικού νευρωνικού δικτύου που μπορεί να επεξεργάζεται διαδοχικά δεδομένα. Το επίπεδο LSTM δέχεται τα word embeddings ως είσοδο και μαθαίνει να μοντελοποιεί τις χρονικές εξαρτήσεις μεταξύ των λέξεων στο κείμενο. Η αρχιτεκτονική του bidirectional επιτρέπει στο LSTM να επεξεργάζεται το κείμενο τόσο προς τα εμπρός όσο και προς τα πίσω, γεγονός που βελτιώνει την ικανότητά του να καταγράφει εξαρτήσεις μεγάλης εμβέλειας. Το επόμενο επίπεδο είναι ένα πλήρως συνδεδεμένο dense επίπεδο με 128 κόμβους και μια συνάρτηση ενεργοποίησης ReLU, η οποία

συμβάλλει στην εισαγωγή μη γραμμικότητας στο μοντέλο και στη βελτίωση της δυναμικής του. Μετά από αυτό προστίθεται ένα dropout επίπεδο για να αποφευχθεί η υπερπροσαρμογή με την τυχαία αφαίρεση ορισμένων κόμβων κατά τη διάρκεια της εκπαίδευσης. Στη συνέχεια, το μοντέλο περνάει από ένα άλλο dense επίπεδο με 64 κόμβους και συνάρτηση ενεργοποίησης ReLU πριν φτάσει στο επίπεδο εξόδου. Το επίπεδο εξόδου έχει δύο κόμβους και μια συνάρτηση ενεργοποίησης softmax, η οποία εξάγει μια κατανομή πιθανότητας πάνω στις δύο κλάσεις (θετικό και αρνητικό συναίσθημα).

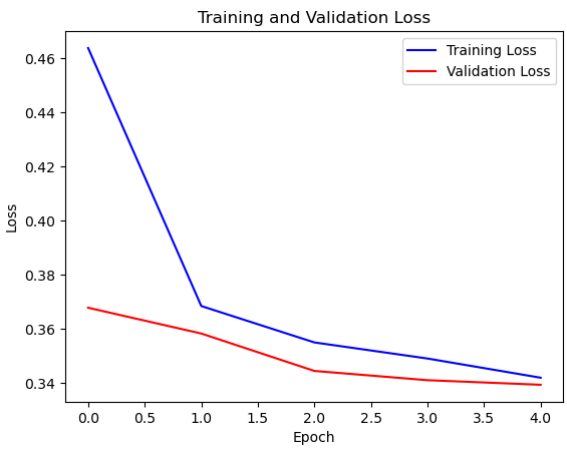
```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 300, 300)          11854500

 bidirectional (Bidirectiona  (None, 256)              439296
 l)

 dense (Dense)               (None, 128)               32896

 dropout (Dropout)           (None, 128)               0

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 2)                 130

=================================================================
Total params: 12,335,078
Trainable params: 480,578
Non-trainable params: 11,854,500
_________________________________________________________________
```

Εικόνα 4.1: Αρχιτεκτονική 1ου μοντέλου

Ακολουθούν τα αποτελέσματα του πρώτου μοντέλου με την χρήση της μεθόδου Word2Vec για την εκπαίδευση του .



Γράφημα 4.3: Accuracy plot για το 1ο μοντέλο (Word2Vec)



Γράφημα 4.4: Loss plot για το 1ο μοντέλο(Word2Vec)

Από το Γράφημα 4.3 φαίνεται ότι η γραμμή για του Training Accuracy είναι κάτω από την γραμμή του Validation Accuracy που υποδηλώνει ότι το μοντέλο έχει καλή απόδοση στα δεδομένα επικύρωσης, υποδεικνύοντας ότι μπορεί να γενικευτεί καλά σε νέα δεδομένα. Προς το τέλος φαίνεται η είναι Training Accuracy να είναι ελαφρώς υψηλότερη από την Validation Accuracy που υποδηλώνει ότι το μοντέλο έχει αρχίσει να προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης. Από το Γράφημα 4.4 φαίνεται ότι η γραμμή του Validation Loss είναι χαμηλότερη συγκριτικά με την Training Loss. Αυτό υποδεικνύει ότι το μοντέλο γενικεύει καλά σε νέα δεδομένα και ότι δεν προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης. Συνοπτικά τα γραφήματα δείχνουν ότι το μοντέλο αποδίδει αρκετά καλά.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	90%	95%	92 %	88.10%
Positive	81%	65%	72%	

Πίνακας 4.23 Μέτρα αξιολόγησης του μοντέλου 1 (Word2Vec)

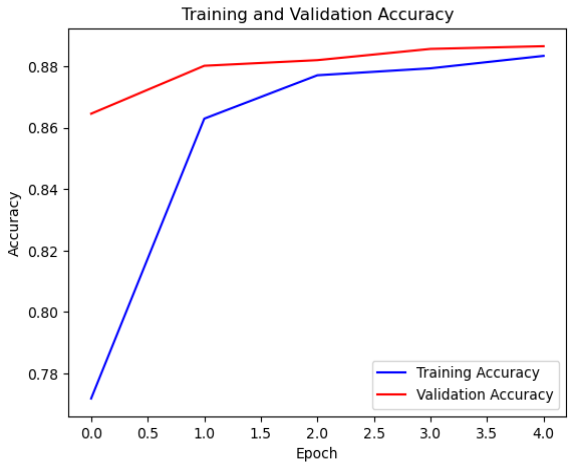
	Predicted label	
True label	Negative	Positive
Negative	9.933	477
Positive	1.138	2.093

Πίνακας 4.24 : Confusion του μοντέλου 1 (Word2Vec)

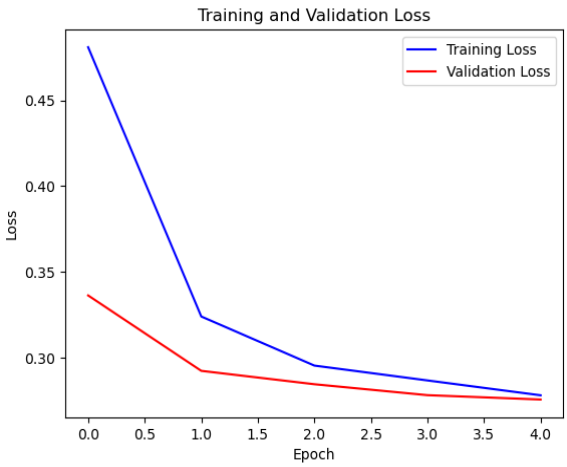
Από τον Πίνακα 4.1.24 φαίνεται ότι το μοντέλο προέβλεψε σωστά μεγάλο αριθμό αρνητικών tweets (TN=9.933), αλλά χαρακτήρισε εσφαλμένα αρνητικό ένα σημαντικό αριθμό θετικών tweets (FP=477) και έχασε ένα σημαντικό αριθμό θετικών tweets (FN=1.138). Ο αριθμός των αληθώς θετικών (TP=2.093) είναι σχετικά χαμηλός σε σύγκριση με τον αριθμό των αρνητικών tweets, υποδεικνύοντας ότι το μοντέλο είναι προβλέπει με ευκολότερα τα αρνητικά tweets. Επιπλέον έχει Accuracy 88,10 % που σημαίνει ότι και για τις δυο κλάσεις το μοντέλο πρόβλεψε σωστά τα 12.093 από τα συνολικά 13.641 tweets. Το μέτρο Precision είναι 81% που σημαίνει ότι από τα 2.570 tweets που προβλέφθηκαν σαν θετικά, τα 2.093 αφορούσαν θετικά tweets και 477 αφορούσαν αρνητικά tweets. Το μέτρο Recall υπολογίστηκε ως 65 %, που σημαίνει ότι από 3.231 tweets που ήταν θετικά τα 2.093 προβλέφθηκαν ως θετικά. Το Recall για την κατηγορία 1 (θετικά tweets) είναι υψηλό (95%), γεγονός που υποδηλώνει ότι το

μοντέλο αναγνώρισε σωστά ένα μεγάλο ποσοστό αρνητικών tweets. Ωστόσο, για την κατηγορία 2 (θετικά tweets) είναι μόνο 65%, γεγονός που υποδηλώνει ότι το μοντέλο έχασε σημαντικό αριθμό θετικών tweets κάτι το οποίο επιβεβαιώνεται και από τον confusion πίνακα.

Στην συνέχεια εφαρμόζεται το 1ο μοντέλο με μόνη αλλαγή ότι χρησιμοποιείται ο Glove για την εκπαίδευση και ακολουθούν τα αποτελέσματα.



Γράφημα 4.5: Accuracy plot για το 1ο μοντέλο (Glove)



Γράφημα 4.6: Loss plot για το 1ο μοντέλο (Glove)

Από το Γράφημα 4.5 φαίνεται ότι το Training Accuracy είναι κάτω από την γραμμή του Validation Accuracy για όλα τα epoch που σημαίνει υποδηλώνει ότι το μοντέλο προσαρμόζεται καλά στα δεδομένα εκπαίδευσης. Από το Γράφημα 4.6 φαίνεται ότι η γραμμή του Validation Loss είναι χαμηλότερη συγκριτικά με την Training Loss, οπότε το μοντέλο ο μοντέλο γενικεύει καλά.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	89%	97%	93 %	88.65%
Positive	86%	62%	72%	

Πίνακας 4.25: Μέτρα αξιολόγησης του μοντέλου 1 (Glove)

	Predicted label	
True label	Negative	Positive
Negative	10.076	334
Positive	1.213	2.018

Πίνακας 4.26: Confusion πίνακας του μοντέλου 1 (Glove)

Από τον Πίνακα 4. 25 φαίνεται ότι το μοντέλο ταξινόμησε σωστά 10.076 αρνητικά tweets (TN) και 2.018 θετικά tweets (TP). Ωστόσο, ταξινόμησε επίσης εσφαλμένα 1.213 αρνητικά tweets (FN) ως θετικά και 334 θετικά tweets (FP) ως αρνητικά. Το Accuracy παραμένει υψηλό και είναι 88,65 % που σημαίνει ότι και για τις δυο κλάσεις το μοντέλο πρόβλεψε σωστά τα 12.094 από τα συνολικά 13.641 tweets. Το μέτρο Precision είναι 86% που σημαίνει ότι από τα 2.352 tweets που προβλέφθηκαν σαν θετικά, τα 2.018 αφορούσαν θετικά tweets. Το μέτρο Recall υπολογίστηκε ως 62 % , ελαφρώς μικρότερο με πριν, που σημαίνει ότι από 3.231 tweets που ήταν θετικά τα 2.018 προβλέφθηκαν ως θετικά. Το F1 score για την κλάση 1 (93%) είναι ένας σταθμισμένος μέσος όρος του Recall και του Precision και υποδεικνύει μια καλή ισορροπία μεταξύ των δύο. Το F1 score για την κλάση 2 είναι 72% είναι επίσης λογικό, αλλά χαμηλότερη από την κλάση 1. Συνοπτικά, το μοντέλο αυτό έχει υψηλό Accuracy και λογικό F1 score και για τις δύο κατηγορίες. Ωστόσο, το υψηλό ποσοστό ψευδώς αρνητικών αποτελεσμάτων για την κλάση 1 και η σχετικά χαμηλή ανάκληση για την κλάση 2 υποδηλώνουν ότι υπάρχουν περιθώρια βελτίωσης.

4.2.2 Δημιουργία και αποτελέσματα δεύτερου μοντέλου

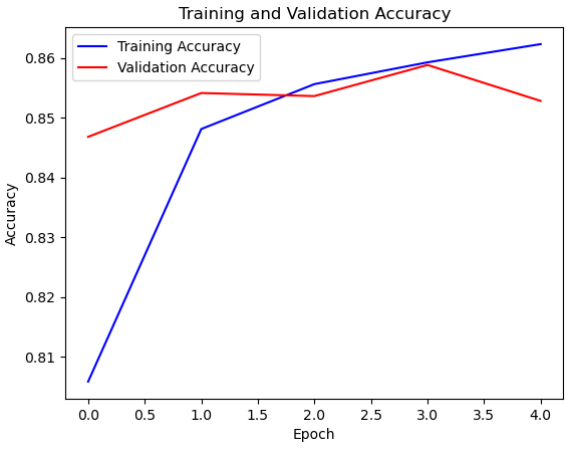
Το δεύτερο μοντέλο νευρωνικού δικτύου είναι ένας τύπος CNN και η δομή του φαίνεται στην Εικόνα 4.2. Το μοντέλο ξεκινά με ένα embedding επίπεδο, το οποίο μετατρέπει κάθε λέξη στην ακολουθία εισόδου σε μια πυκνή διανυσματική αναπαράσταση. Ακολουθεί ένα convolutional επίπεδο 1D με 64 φίλτρα, μεγέθους 2 και με συνάρτηση ενεργοποίησης ReLU. Το padding ορίζεται σε "same", το οποίο εξασφαλίζει ότι η έξοδος έχει το ίδιο σχήμα με την είσοδο. Στη συνέχεια εφαρμόζεται ένα max pooling επίπεδο για να μειωθεί η διάστασή της εξόδου από το convolutional επίπεδο, λαμβάνοντας τη μέγιστη τιμή εντός ενός συρόμενου παραθύρου μεγέθους 2. Η έξοδος από το pooling επίπεδο ισοπεδώνεται και τροφοδοτείται σε ένα πλήρως συνδεδεμένο dense επίπεδο με 1024 κόμβους και μια συνάρτηση ενεργοποίησης ReLU. Ένα άλλο dense επίπεδο με 512 κόμβους και μια συνάρτηση ενεργοποίησης ReLU προστίθεται πριν από το τελικό dense επίπεδο με τον ίδιο αριθμό μονάδων με τον αριθμό των κλάσεων στη μεταβλητή-στόχο, στην προκειμένη περίπτωση 2, και μια συνάρτηση ενεργοποίησης softmax. Η συνάρτηση softmax παράγει μια κατανομή πιθανοτήτων πάνω στις κλάσεις, με κάθε τιμή να αντιπροσωπεύει την πιθανότητα η είσοδος να ανήκει στη

συγκεκριμένη κλάση. Το επίπεδο εξόδου αυτού του μοντέλου είναι ένα πλήρως συνδεδεμένο επίπεδο με συνάρτηση ενεργοποίησης softmax και 2 κόμβους στην έξοδο, μία για κάθε κλάση.

```
Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_2 (Embedding)     (None, 300, 300)          11854500

 conv1d_1 (Conv1D)           (None, 300, 64)           38464

 max_pooling1d_1 (MaxPooling  (None, 150, 64)          0
 1D)

 flatten_1 (Flatten)         (None, 9600)              0

 dense_5 (Dense)             (None, 1024)              9831424

 dense_6 (Dense)             (None, 512)               524800

 dense_7 (Dense)             (None, 2)                 1026

=================================================================
Total params: 22,250,214
Trainable params: 10,395,714
Non-trainable params: 11,854,500
_________________________________________________________________
```

Εικόνα 4.2: Αρχιτεκτονική 2ου μοντέλου

Ακολουθούν τα αποτελέσματα του δεύτερου μοντέλου με την χρήση της μεθόδου Word2Vec για την εκπαίδευση του .



Γράφημα 4.7: Accuracy plot για το 2ο μοντέλο (Word2Vec)

Γράφημα 4.8: Loss plot για το 2ο μοντέλο (Word2Vec)

Από το Γράφημα 4.7 φαίνεται ότι το Training Accuracy είναι κάτω από την γραμμή του Validation Accuracy ωστόσο περίπου στην μέση το Training Accuracy ξεπερνά το Validation

Accuracy. Από το Γράφημα 4.8 φαίνεται ότι η γραμμή του Validation Loss είναι χαμηλότερη συγκριτικά με την Training Loss , οπότε το μοντέλο ο μοντέλο γενικεύει καλά.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	86%	97%	91 %	85.27%
Positive	83%	48%	60%	

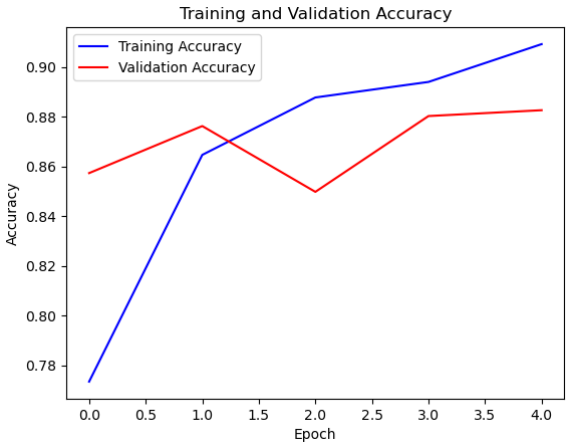
Πίνακας 4.27: Μέτρα αξιολόγησης του μοντέλου 2(Word2Vec)

	Predicted label	
True label	Negative	Positive
Negative	10.098	312
Positive	1.696	1.535

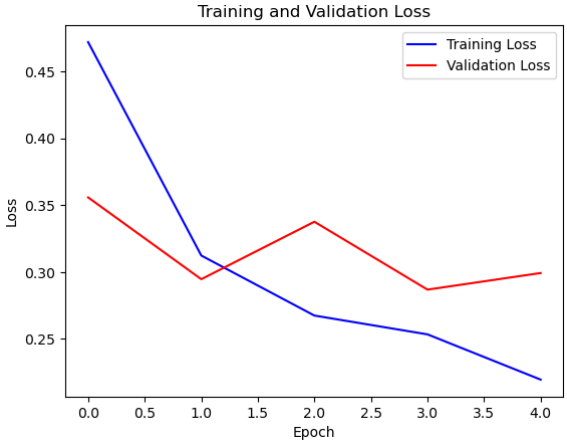
Πίνακας 4.28: Confusion πίνακας του μοντέλου 2 (Word2Vec)

Από τον Πίνακα 4.1.27 φαίνεται ότι το μοντέλο ταξινόμησε σωστά 10.098 αρνητικά tweets (TN) και 1.535 θετικά tweets (TP). Ωστόσο, ταξινόμησε επίσης εσφαλμένα 1.696 αρνητικά tweets (FN) ως θετικά και 312 θετικά tweets (FP) ως αρνητικά. 85.27%, που σημαίνει ότι από τα συνολικά 10.410 tweets και για τις δύο κλάσεις προβλέφθηκαν σωστά τα 11.633. Το μέτρο Precision είναι 83% που σημαίνει ότι από τα 1847 tweets που προβλέφθηκαν σαν θετικά, τα 1535 αφορούσαν θετικά tweets. Το μέτρο Recall υπολογίστηκε ως 48%, που σημαίνει ότι από 3.231 tweets που ήταν θετικά τα 1.535 προβλέφθηκαν ως θετικά. Το F1 score για την κλάση 1 (91%), είναι ένας σταθμισμένος μέσος όρος της ακρίβειας και της ανάκλησης και υποδεικνύει μια καλή ισορροπία μεταξύ των δύο. Το F1 score για την κλάση 2 του 60% είναι επίσης λογική, αλλά χαμηλότερη από την κλάση 1.

Στην συνέχεια εφαρμόζεται το ίδιο μοντέλο (2ο μοντέλο) με μόνη αλλαγή στον αλγόριθμο όπου χρησιμοποιείται είναι ο Glove.



Γράφημα 4.9: Accuracy plot για το 2ο μοντέλο (Glove)



Γράφημα 4.10: Loss plot για το 2ο μοντέλο (Glove)

Από το Γράφημα 4.9 φαίνεται ότι το Training Accuracy είναι κάτω από την γραμμή του Validation Accuracy ωστόσο μετά από ένα το Training Accuracy ξεπερνά το Validation Accuracy. Από το Γράφημα 4.10 το Training Loss είναι κάτω από τοValidation Loss.

Class	Precision score	Recall score	F1- score	Test Accuracy
Negative	90%	96%	93 %	88.25%
Positive	82%	64%	72%	

Πίνακας 4.29: Μέτρα αξιολόγησης του μοντέλου 2 (Glove)

	Predicted label	
True label	Negative	Positive
Negative	9.971	439
Positive	1.163	2.068

Πίνακας 4.30 : Confusion πίνακας του μοντέλου 2 (Glove)

Από τον Πίνακα 4.1.28 φαίνεται ότι το μοντέλο ταξινόμησε σωστά 9.971αρνητικά tweets (TN) και 2068 θετικά tweets (TP). Ωστόσο, ταξινόμησε επίσης εσφαλμένα 1.163 αρνητικά tweets (FN) ως θετικά και 439 θετικά tweets (FP) ως αρνητικά. Το μέτρο Precision είναι 90% που σημαίνει ότι από τα 2.507 tweets που προβλέφθηκαν σαν θετικά, τα 2.068 αφορούσαν θετικά tweets. Το μέτρο Recall υπολογίστηκε ως 64 %, που σημαίνει ότι από 3.232 tweets που ήταν θετικά τα 2.068προβλέφθηκαν ως θετικά. Τέλος, Το F1 score για την κλάση 1(93%) είναι

ένας σταθμισμένος μέσος όρος της ακρίβειας και της ανάκλησης και υποδεικνύει μια καλή ισορροπία μεταξύ των δύο. Το F1 score για την κατηγορία 2, 72%, είναι επίσης λογική, αλλά χαμηλότερη από την κατηγορία 1.

4.2.3 Συγκέντρωση αποτελεσμάτων

Για την ευκολότερη σύγκριση των αποτελεσμάτων ακολουθεί συγκεντρωτικός πίνακας των αποτελεσμάτων των μοντέλων με την χρήση του αλγορίθμου word2Vec. Επίσης έχουν υπολογιστεί και προστεθεί τα AUC για τα δύο μοντέλα.

	Precision score		Recall score		F1- score		Test Accuracy	Test Loss	Macro Average Precision score	Macro Average Recall score	Macro Average F1 – score	AUC
Class	Negative	Positive	Negative	Positive	Negative	Positive						
1ο μοντέλο	90%	81%	95%	65%	92 %	72%	88.10%	40%	81%	78%	79%	91.9%
2ο μοντέλο	86%	83%	97%	48%	91 %	60%	85.27%	53%	84%	72%	76%	87.9%

Πίνακας 4.31 : Συγκεντρωτικός πίνακας μέτρων αξιολόγησης των μοντέλων (Word2Vec.)

Από τον Πίνακας 4.31 φαίνεται ότι το μοντέλο 1 έχει υψηλότερο Precision και Recall για την κατηγορία 1 (αρνητικά tweets) και ελαφρώς χαμηλότερο αλλά αξιοπρεπές Precision για την κατηγορία 2 (θετικά tweets). Ωστόσο, έχει σχετικά χαμηλό Recall για την κλάση 2, γεγονός που υποδηλώνει ότι μπορεί να δυσκολεύεται να εντοπίσει σωστά τα θετικά tweets. Το F1 score για την κλάση 1 είναι επίσης υψηλότερο από αυτό για την κλάση 2, υποδεικνύοντας ότι το μοντέλο αποδίδει καλύτερα στον εντοπισμό αρνητικών tweets. Από την άλλη πλευρά, το μοντέλο 2 έχει υψηλότερο Precision για την κλάση 2 και υψηλότερο και Recall για την κλάση 1, υποδεικνύοντας ότι είναι καλύτερο στον σωστό εντοπισμό θετικών tweets, ενώ εξακολουθεί να έχει καλή απόδοση για τα αρνητικά tweets. Ωστόσο, έχει χαμηλότερο F1 score για την κλάση 1 από το Μοντέλο 1, υποδεικνύοντας ότι η απόδοσή του για τα αρνητικά tweets δεν είναι τόσο ισχυρή. Το μοντέλο 1 είναι καλύτερο στον εντοπισμό αρνητικών δειγμάτων, ενώ το μοντέλο 2 είναι καλύτερο στον εντοπισμό θετικών δειγμάτων. Ωστόσο, όταν εξετάζουμε και τις δύο κλάσεις μαζί και τη συνολική απόδοση του μοντέλου, το μοντέλο 1 φαίνεται να είναι καλύτερη επιλογή, καθώς έχει υψηλότερο, Precision και F1 score, καθώς και υψηλο AUC και Accuracy.

Στην συνέχεια παρουσιάζεται ο αντίστοιχος συγκεντρωτικός πίνακας των αποτελεσμάτων των μοντέλων με την χρήση του αλγορίθμου Glove.

	Precision score		Recall score		F1- score		Test Accuracy	Test Loss	Macro Average Precision score	Macro Average Recall score	Macro Average F1 – score	AUC
Class	Negative	Positive	Negative	Positive	Negative	Positive						
1ο μοντέλο	89%	86%	97%	62%	93 %	72%	88.65%	36.7%	88%	80%	72%	92.6%
2ο μοντέλο	90%	82%	96%	64%	93%	72%	88.25%	29.9%	86%	80%	82%	91.9%

Πίνακας 4.32 : Συγκεντρωτικός πίνακας μέτρων αξιολόγησης των μοντέλων (Glove)

Συγκρίνοντας τα δύο μοντέλα, βλέπουμε ότι το μοντέλο 1 έχει μεγαλύτερο Precision για την κλάση 1 και την κλάση 2, Recall για την κλάση 1, F1 score για την κλάση 1 και την κλάση 2 και AUC. Ωστόσο, το Μοντέλο 2 έχει υψηλότερο Precision για την Κλάση 2 και Macro Average Precision score. Το Macro Average Recall score και το Macro Average F1-score είναι σχεδόν τα ίδια και για τα δύο μοντέλα. Συνολικά, με βάση αυτές τις μετρήσεις, μπορούμε να πούμε ότι το μοντέλο 1 είναι ελαφρώς καλύτερο από το μοντέλο 2 στο σενάριο 2.

Συνολικά, τα περισσότερα μοντέλα που δημιουργήθηκαν δυσκολεύονταν στην ακριβή αναγνώριση των θετικών tweets. Ωστόσο το μοντέλο NBM με χρήση του BoW επιτυγχάνει το μεγαλύτερο Recall για την κατηγορία των θετικών tweets, ενώ δεύτερο ακολουθεί το RNN. Λαμβάνοντας υπόψιν την απλότητα ως προς την κατανόηση και την εφαρμογή φαίνεται πως το NBM είναι μια καλή επιλογή για το συγκεκριμένο πρόβλημα και τα συγκεκριμένα δεδομένα.

5 Προτάσεις μελλοντικής βελτίωση

Σ’ αυτήν την εργασία, επιχειρήθηκε μια εφαρμογή συναισθηματικής ανάλυσης σε δεδομένα που ερχόντουσαν από Twitter με θέμα την πανδημία του Covid-19. Εκτελέστηκαν πειράματα χρησιμοποιώντας διαφορετικά μοντέλα μηχανικής μάθησης και δυο διαφορετικοί τύποι νευρωνικών δικτύων. Για την επιλογή χαρακτηριστικών χρησιμοποιήθηκαν τα στην μηχανική μάθηση ο CountVectorizer και ο TD-IDF, ενώ για τα νευρωνικά δίκτυα ο Word2vect και ο Glove.

Τα πειράματα έδειξαν ότι η μέθοδος που εφαρμόζεται για την επιλογή χαρακτηριστικών επηρεάζει σε μεγάλο βαθμό τα αποτελέσματα των μοντέλων της μηχανικής μάθησης. Αυτό οδηγεί στο συμπέρασμα ότι είναι αναγκαία η προσεκτική επιλογή της μεθόδου για την εξαγωγή χαρακτηριστικών και ότι υπάρχει ανάγκη πειραματισμού με διάφορες μεθόδους προκειμένου να επιλεγεί η καταλληλότερη για τα δεδομένα που επεξεργάζονται. Επίσης, για τα μοντέλα των νευρωνικών δικτύων ήταν σαφές πως η αρχιτεκτονική παίζει σημαντικό ρόλο στην διαμόρφωση του αποτελέσματος, με άλλες αρχιτεκτονικές να αποδίδουν καλύτερα για το συγκεκριμένο σύνολο δεδομένων και να μην ενέχουν μεγάλο κίνδυνο για overfitting. Όπως και στα μοντέλα μηχανικής μάθησης, έτσι και στα μοντέλα των νευρωνικών δικτύων τα προ-εκπαιδευμένα γλωσσικά μοντέλα επηρεάζουν τις αποδώσεις τους. Για αυτό προτείνεται ο πειραματισμός με διαφορά γλωσσικά μοντέλα έτσι ώστε να γίνει η επιλογή του καταλληλότερου που συμβάλει στην καλύτερη κατανόηση των αποχρώσεων στην γλώσσα.

Η πειραματική μελέτη που πραγματοποιήθηκε είχε κάποιους περιορισμούς που θα μπορούσαν να ληφθούν υπόψιν για να βελτιωθούν μελλοντικές μελέτες. Αρχικά, εξαιτίας της περιορισμένης υπολογιστικής δυνατότητας που υπήρχε, το σύνολο των δεδομένων που χρησιμοποιήθηκε περιορίστηκε μέσω τυχαίας δειγματοληψίας. Αυτό μείωσε τα δεδομένα που ήταν διαθέσιμα για εκπαίδευση, οπότε περιορίστηκε η ακρίβεια και η απόδοση των μοντέλων. Ακόμη στην μηχανική μάθηση εξετάστηκαν μεμονωμένοι αλγόριθμοι και δεν υλοποιήθηκε κάποιο πείραμα με συνδυασμό αλγορίθμων (ensemble methods), π.χ. χρήση SVM για την πρόβλεψη της μια κατηγορίας και NBM για την πρόβλεψη της δεύτερης κατηγορίας. Ο συνδυασμός πολλών μεθόδων θα μπορούσε να λύσει το πρόβλημα της ελλιπούς ταξινόμησης των θετικών tweets, να μειώσει την πιθανότητα overfitting και γενικά να βελτίωση την σταθερότητα και την γενίκευση του μοντέλου.

Αναφορικά με τα νευρωνικά δίκτυα, υπάρχουν μερικές αλλαγές που θα μπορούσαν να βελτιώσουν την απόδοσή τους. Αρχικά θα μπορούσαν να είχαν δημιουργηθεί και άλλα μοντέλα νευρωνικών δικτύων, είτε με περισσότερα layers είτε μια διαφορετικές αρχιτεκτονικές. Ένα παράδειγμα αρχιτεκτονικής που θα μπορούσε να εφαρμοστεί είναι οι Transformers και ειδικότερα ο αλγόριθμος του ChatGTP. Αυτός ο αλγόριθμος, με την υλοποίηση κατάλληλων αλλαγών για την προσαρμογή του στο συγκεκριμένο πρόβλημα, φαίνεται πολλά υποσχόμενος για την βαθύτερη κατανόηση των αποχρώσεων της γλώσσα και εννοιών όπως η ειρωνεία και ο σαρκασμός. Μια άλλη εξίσου αποδοτική αρχιτεκτονική σύμφωνα με την βιβλιογραφία, είναι αυτή των μοντέλων BERT, η οποία συμβάλλει στην λήψη πιο σύνθετων σχέσεων και εξαρτήσεων. Επιπλέον, η ενσωμάτωση μηχανισμών προσοχής στα νευρωνικά δίκτυα θα βοηθούσε το μοντέλο να εστιάσει στα πιο σημαντικά τμήματα των δεδομένων εισόδου, κάτι που είναι ιδιαίτερα χρήσιμο για κείμενα όπου τα διάφορα μέρη μιας πρότασης μπορεί να έχουν διαφορετικό βαθμό σπουδαιότητας. Τέλος, οι τεχνικές επαύξησης δεδομένων μπορούν να χρησιμοποιηθούν για την αύξηση του μεγέθους του συνόλου δεδομένων εκπαίδευσης, αφού δημιουργούν νέα δείγματά δεδομένων από το υπάρχον σύνολο δεδομένων, και έτσι συμβάλλουν στη μείωση της υπερπροσαρμογής.

.

Βιβλιογραφία

[1] Cullen, W., Gulati, G., & Kelly, B. D. (2020). Mental health in the COVID-19 pandemic. QJM: An International Journal of Medicine, 113(5), 311-312.

[2] Coe, N. M., Hess, M., Yeung, H. W. C., Dicken, P., & Henderson, J. (2021). Global production networks in the pandemic and beyond: Strategies of resilience. Environment and Planning A: Economy and Space, 53 (7), 1377-1385.

[3] UNESCO. (2020). Adverse consequences of school closures. Retrieved from https://en.unesco.org/covid19/educationresponse/consequences

[4] Hu, X., Zhang, Y., Zhang, X., & Yu, B. (2021). Public sentiment analysis on social media in the early phase of COVID-19 outbreak.

[5] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10) (pp. 1320-1326). European Language Resources Association.

[6] Hussain, A., Tahir, A., Hussain, Z., et al. (2021). Artificial intelligence-enabled analysis of public attitudes on Facebook and Twitter toward COVID-19 vaccines in the United Kingdom and the United States: Observational Study. Journal of Medical Internet Research, Published 2021 Apr 5. doi:10.2196/26627.

[7] Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. arXiv preprint arXiv:2005.10898.

[8] Smith, J., Brown, A., & Lee, K. (2021). A deep learning framework for sentiment analysis of short texts. arXiv preprint arXiv:2211.09733.

[9] Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

[10] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5 (1), 1-167.

[11] Ghosh, D., & Veale, T. (2018). Document level sentiment analysis: A survey. In Journal of the Association for Information Science and Technology ,69(6), 855-869.

[12] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with Twitter: What 140 characters reveal about political sentiment. ICWSM, 10(1), 178-185.

[13] Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment in Twitter events. Journal of the American Society for Information Science and Technology, 63(1), 143-152.

[14] Gretzel, U., Sigala, M., Xiang, Z., & Koo, C. (2015). Recommending and marketing destinations with user-generated content: Use of sentiment analysis and geo-tagging information. Journal of Travel Research, 54(4), 437-451.

[15] Mohammadi, E., & Davarpanah Jazi, S. (2018). Combining multiple sentiment dictionaries for sentiment analysis of Twitter data. International Journal of Web Information Systems, 14(4), 363-380.

[16] Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10) ,2200-2204. Retrieved from: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf

[17] Smith, J., Brown, A., & Lee, K. (2021). A Deep Learning Framework for Sentiment Analysis of Short Texts. arXiv preprint arXiv:2211.09733.

[18] Muller, A. C., & Guido, S. (2017). Introduction to machine learning with Python: A guide for data scientists. Sebastopol, CA: O'Reilly Media, Inc.

[19] Raschka, S. (2015). Unlock deeper insights into machine learning with vital guide to cutting-edge predictive analytics. Packt Publishing.

[20] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In J. May and K. Knight (Eds.), Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and

Interpreting Neural Networks for NLP (pp. 353-355). Association for Computational Linguistics.

[21] Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2019). On large-batch training for deep learning: Generalization gap and sharp minima. Proceedings of the International Conference on Learning Representations (ICLR).

[22] Stanford NLP Group. (n.d.). GloVe: Global Vectors for Word Representation. Retrieved from: https://nlp.stanford.edu/projects/glove/

[23] Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of Naive Bayes text classifiers. In M. G. Wallace & A. K. D. Wong (Eds.), Proceedings of the Twentieth International Conference on Machine Learning (ICML-03) (pp. 616-623). AAAI Press.

[24] Manning, C. D., Raghavan, P., Schütze, H., & Baeza-Yates, R. (2008). Introduction to information retrieval (Vol. 1). Cambridge University Press.

[25] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

[26] Huang, X., & Ling, C. X. (2005). Using A Complement Naive Bayes Classifier for Text Classification. Journal of Machine Learning Research, 5, 1223-1255.

[28] Syachrani, S., Syadaruddin, J., Jeong, H., & Chung, C. (2013). Decision tree–based deterioration model for buried wastewater pipelines. Journal of Performance of Constructed Facilities, 27(6), 793-803.

[29] Goh, C. H., & Liang, X. (2017). Text classification using machine learning: A review. Journal of Computer Science and Technology, 32(3), 434-448.

[30] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

[31] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[32] Jaeger, H. (2002). Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the "echo state network" approach (2nd revision) (GMD Report 159). German National Research Center for Information Technology.

[33] Singh, P., & Kumar, A. (2018). Sentiment analysis of twitter data: A review of techniques. Journal of Information Processing Systems, 14(3), 675-689. DOI: 10.3745/JIPS.04.0077

[34] TensorFlow. (n.d.). Keras: Deep Learning API for TensorFlow. TensorFlow. Retrieved from: https://www.tensorflow.org/api_docs/python/tf/keras

[35] Scikit-learn. (n.d.). scikit-learn: Machine learning in Python. DevDocs. Retrieved May 12, 2023, from https://devdocs.io/scikit_learn/

[36] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[37] Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–1751). Association for Computational Linguistics.