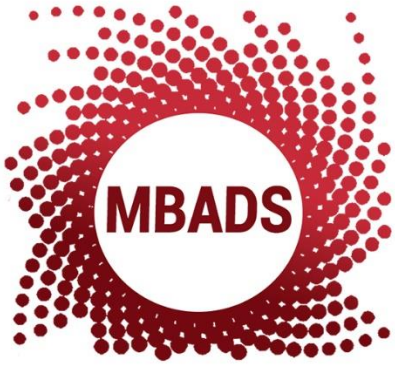


ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ
ΑΝΑΛΥΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ ΚΑΙ ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

Πρόγραμμα Μεταπτυχιακών Σπουδών

στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική Εργασία

Τεχνικές NLP και Twitter Sentiment Analysis

Ζωή Κεσίδου του Αλέξανδρου

Φεβρουάριος 2024

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω από τα βάθη της καρδιάς μου την οικογένεια μου, για την ανεκτίμητη στήριξή τους, τόσο στην παρούσα διπλωματική, όσο και σε ολόκληρη την διάρκεια των σπουδών μου.

Θα ήθελα επίσης να ευχαριστήσω θερμά τους φίλους μου για τη συμπαράσταση και την υπομονή που μου προσέφεραν καθ' όλη τη διάρκεια του μεταπτυχιακού μου.

Τέλος, οφείλω να ευχαριστήσω τον καθηγητή μου και επιβλέποντα της παρούσα διπλωματική εργασία, κ. Κωνσταντάρα Ιωάννη, για την επιστημονική και συμβουλευτική καθοδήγηση που μου προσέφερε κατά την διάρκεια της εκπόνησης της εργασίας με τις εύστοχες παρατηρήσεις του.

Περίληψη

Η ταχεία εξέλιξη του Παγκόσμιου Ιστού (WWW) τα τελευταία χρόνια, σε συνδυασμό με τον αυξανόμενο ενθουσιασμό για τα κοινωνικά δίκτυα, έχει αναμφισβήτητα προσφέρει ένα ευνοϊκό περιβάλλον για την ανάπτυξη της Ανάλυσης Δεδομένων. Συγκεκριμένα, αναφερόμαστε στην ανάλυση δεδομένων που προέρχονται από τα κοινωνικά δίκτυα, δεδομένου του τεράστιου όγκου δεδομένων που διαθέτουν.

Αυτή η διπλωματική διερευνά την πρακτική εφαρμογή τεχνικών επεξεργασίας φυσικής γλώσσας (NLP) για την ανάλυση του συναισθήματος στο Twitter και την ανίχνευση ομιλίας μίσους. Το Κεφάλαιο 1 εισάγει τη σημασία του NLP στην κατανόηση και την ανάλυση των μοτίβων της ανθρώπινης γλώσσας.

Στο Κεφάλαιο 2 διεξάγεται μια ολοκληρωμένη βιβλιογραφική ανασκόπηση, που καλύπτει τις θεμελιώδεις έννοιες του NLP, τα βήματα που εμπλέκονται στην επεξεργασία NLP και τις βασικές εφαρμογές του NLP, όπως η ανίχνευση σαρκασμού, η ανάκτηση πληροφοριών, η μεταφραστική μηχανική και η ανάλυση των συναισθημάτων. Στη συνέχεια, η εστίαση μετατοπίζεται στην ανάλυση των συναισθημάτων, τα στάδια, τα οφέλη και την εφαρμογή του στο περιβάλλον Twitter.

Το Κεφάλαιο 3 ασκεί την εφαρμογή της έρευνας, περιγράφοντας λεπτομερώς τις μεθοδολογίες συλλογής δεδομένων και της επακόλουθης διαδικασίας ανάλυσης δεδομένων. Αυτό περιλαμβάνει τη φόρτωση και την εξερεύνηση του συνόλου δεδομένων, την προεπεξεργασία δεδομένων, την ανάλυση διερευνητικών δεδομένων (EDA), την εξαγωγή χαρακτηριστικών, την εκπαίδευση μοντέλων και την τελειοποίηση τους.

Τέλος, το Κεφάλαιο 4 παρουσιάζει συμπεράσματα που προέρχονται από την ανάλυση και παρέχουν προτάσεις για μελλοντικές οδηγίες έρευνας. Μέσω εμπειρικού πειραματισμού και κριτικής αξιολόγησης, αυτή η εργασία συμβάλλει στην προώθηση της κατανόησης των τεχνικών NLP, στην αντιμετώπιση των προκλήσεων της ανάλυσης των συναισθημάτων και της ανίχνευσης ομιλίας μίσους στις πλατφόρμες των κοινωνικών μέσων.

Abstract

The rapid development of the World Wide Web (WWW) in recent years, combined with the growing enthusiasm for social networks, has undoubtedly provided a favourable environment for the development of Data Analysis. Specifically, we refer to data analysis derived from social networks, given the huge amount of data they possess.

This thesis explores the practical application of Natural Language Processing (NLP) techniques for sentiment analysis on Twitter and hate speech detection. Chapter 1 introduces the importance of NLP in understanding and analyzing human language patterns.

Chapter 2 conducts a comprehensive literature review, covering the fundamental concepts of NLP, the steps involved in NLP processing, and the key applications of NLP, such as sarcasm detection, information retrieval, translation engineering, and sentiment analysis. The focus then shifts to sentiment analysis, its stages, benefits and application in the Twitter environment.

Chapter 3 exercises the application of the research, detailing the data collection methodologies and the subsequent data analysis process. This includes data set loading and exploration, data pre-processing, exploratory data analysis (EDA), feature extraction, model training and refinement.

Finally, Chapter 4 presents conclusions drawn from the analysis and provides suggestions for future research directions. Through empirical experimentation and critical evaluation, this thesis contributes to advancing the understanding of NLP techniques in addressing the challenges of sentiment analysis and hate speech detection on social media platforms.

Πίνακας Περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Διαγραμμάτων

Κεφάλαιο 1. Εισαγωγή

Ένα από τα βασικά συστατικά της Τεχνητής Νοημοσύνης (Artificial Intelligence – AI) και της Υπολογιστικής Γλωσσολογίας (Computational Linguistics), είναι η Επεξεργασίας της Φυσικής Γλώσσας (Natural Language Processing – NLP) (Abram, Mancini & Parker, 2020). Η Επεξεργασία της Φυσικής Γλώσσας έχει σαν αντικείμενο τη διερεύνηση και εξέταση της αλληλεπίδρασης ανάμεσα στη φυσική γλώσσα και στον υπολογιστή (Abram, Mancini & Parker, 2020). Προσφέρει σημαντικά οφέλη και ειδικότερα βοηθάει στη διασφάλιση ομαλής αλληλεπίδρασης μεταξύ ανθρώπου και υπολογιστή, δίνοντας τη δυνατότητα στους υπολογιστές να κατανοήσουν την ανθρώπινη ομιλία μέσα από τη βοήθεια της Μηχανικής Μάθησης (Machine Learning) (Abram, Mancini & Parker, 2020). Περίπου στα τέλη της δεκαετίας του 1940 έκανε την εμφάνιση της Επεξεργασίας Φυσικής Γλώσσας, ως απόρροια της ανάπτυξης της επιστήμης των υπολογιστών. Με την πάροδο των χρόνων και τις διάφορες εξελίξεις που σημειώθηκαν σε τεχνολογικό επίπεδο, δημιουργήθηκε και η επίσημη θεωρία της Επεξεργασίας Φυσικής Γλώσσας.

Η Επεξεργασίας Φυσικής Γλώσσας χρησιμοποιεί ένα σύνολο τεχνικών και μεθόδων προκειμένου να επιτευχθεί η αποτελεσματική αναγνώριση των κειμένων, μέσα από την μετατροπή διακριτών ή συνεχών συνδυαστικών δομών (όπως πίνακες, δέντρα, διανύσματα κ.α.) (Alaei, Becken & Stantic, 2019). Σε πρακτικό επίπεδο η NLP είναι αντίστοιχη της διδασκαλίας της γλώσσας σε μικρά παιδιά καθώς περιέχει εργασίες όπως κατανόηση λέξεων και προτάσεων, διαμόρφωση ορθών δομικών προτάσεων σε γραμματικό και συντακτικό επίπεδο, στοιχεία δηλαδή που γίνονται και στον φυσικό-πραγματικό κόσμο του ανθρώπου (Alaei, Becken & Stantic, 2019). Εδώ θα πρέπει να σημειωθεί ότι ως φυσική γλώσσα αντιλαμβάνεται οποιαδήποτε γλώσσα η οποία αξιοποιείται για τη διασφάλιση της επικοινωνίας ανάμεσα στους ανθρώπους. Είναι επίσης η γλώσσα την οποία μαθαίνει ο άνθρωπος ήδη από τη νεαρή του ηλικία από το περιβάλλον του προκειμένου να καταφέρει να εκφράζει τα συναισθήματα, τις σκέψεις και τις γνώσεις του και να μπορεί να μεταφέρει τις απαντήσεις του σε άλλους ανθρώπους.

Τα τελευταία χρόνια αναπτύχθηκε μια άλλη μορφή γλώσσας, οι γλώσσες προγραμματισμού, οι οποίες έκαναν την εμφάνιση τους περίπου στα μέσα του 20^{ου} αιώνα με γνώμονα την επικοινωνία του ανθρώπου με τα μηχανήματα. Με την ολοένα και

αυξανόμενη ανάπτυξη των υπολογιστών, οι γλώσσες προγραμματισμού έλαβαν πολύ μεγάλη σημασία. Τόσο οι γλώσσες προγραμματισμού όσο και η φυσική γλώσσα εμφανίζουν κάποια κοινά στοιχεία και ομοιότητες μεταξύ τους, όπως ύπαρξη κανόνων σύνταξης και δόμησης, η κύρια διαφορά τους στηρίζεται όμως στο γεγονός ότι οι φυσικές γλώσσες είναι διφορούμενες και ως εκ τούτου πολύ δύσκολο να κατανοηθούν από τις μηχανές (Agarwal, 2022). Παράλληλα θα πρέπει να αναφερθεί ότι οι φυσικές γλώσσες έχουν και άλλα στοιχεία όπως σαρκασμό, ρητορικές εκφράσεις, διπλή άρνηση, τα οποία ενισχύουν την πολυπλοκότητα της. Ως εκ τούτου οι μηχανές προκειμένου να αναγνωρίσουν τα διάφορα αυτά στοιχεία και το νόημα τους, θα πρέπει να έχουν τις κατάλληλες κωδικοποιήσεις, προκειμένου να μπορούν να αποδώσουν/επικοινωνήσουν το αντίστοιχο νόημα (Agarwal, 2022).

Το γεγονός αυτό βοήθησε σημαντικά στην εξέλιξη και ανάπτυξη της NLP με πολλούς μελετητές να αξιοποιούν τη συγκεκριμένη γλώσσα είτε υιοθετώντας μια στοχαστική προσέγγιση, είτε υιοθετώντας μια συμβολική προσέγγιση σε επίπεδο μοντελοποίησης της γλώσσας. Με την εμφάνιση των μέσων κοινωνικής δικτύωσης, αναπτύχθηκε ραγδαία η βιομηχανία της ανάλυσης συναισθημάτων (sentiment analysis), με την NLP να βοηθάει σημαντικά στον τομέα αυτό.

Αντικείμενο της παρούσας εργασίας είναι είναι να εμβαθύνει στις τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) και στην πρακτική εφαρμογή τους στον τομέα της ανάλυσης συναισθημάτων Twitter και της ανίχνευσης της ρητορικής μίσους. Μέσα από εξερεύνηση και ανάλυση, η εργασία στοχεύει να παεουσίασει τις λειτουργίες των αλγορίθμων και μεθοδολογιών NLP, ρίχνοντας φως στην αποτελεσματικότητα και τους περιορισμούς τους στην αποκρυπτογράφηση του συναισθήματος και των γλωσσικών προτύπων που επικρατούν στα κείμενα των μέσων κοινωνικής δικτύωσης. Αξιοποιώντας σύγχρονα μοντέλα μηχανικής μάθησης και τεχνικές προεπεξεργασίας δεδομένων, η εργασία προσπαθεί να εκπαιδεύσει μοντέλα για τον εντοπισμό και τον μετριασμό περιπτώσεων ρητορικής μίσους και τοξικής συμπεριφοράς στο Twitter, συμβάλλοντας έτσι στον ευρύτερο στόχο της προώθησης μιας πιο ασφαλής διαδικτυακής κοινότητας. Μέσω εμπειρικού πειραματισμού και κριτικής αξιολόγησης, η παρούσα διπλωματική επιδιώκει την κατανόησή μας για την περίπλοκη αλληλεπίδραση μεταξύ γλώσσας, τεχνολογίας και κοινωνίας, προσπαθώντας τελικά να ενδυναμώσει τους ενδιαφερόμενους με ερεθίσματα και εργαλεία για την καταπολέμηση της διαδικτυακής κατάχρησης και την προώθηση θετικών ψηφιακών αλληλεπιδράσεων.

Κεφάλαιο 2. Βιβλιογραφική Επισκόπηση

2.1 Εννοιολογική προσέγγιση Επεξεργασίας Φυσικής Γλώσσας

Η Επεξεργασία Φυσικής Γλώσσας, αποτελεί μια από τις δημοφιλέστερες εφαρμογές της Τεχνικής Νοημοσύνης και περιέχει τη συλλογή γνώσεων σχετικά με τους τρόπους που τα ανθρώπινα όντα αντιλαμβάνονται και κάνουν χρήση της γλώσσας (Arnarsson et.al., 2021). Απώτερος στόχος είναι η δημιουργία κατάλληλων εργαλείων και τεχνικών οι οποίες θα βοηθήσουν τα συστήματα των υπολογιστών να κατανοήσουν και να είναι σε θέση να χειριστούν τις φυσικές γλώσσες προκειμένου να εκτελέσουν τις διάφορες αναγκαίες εργασίες που πρέπει να γίνουν (Alharbi et.al., 2021).

Θα πρέπει να αναφερθεί ότι μέχρι σήμερα δεν υπάρχει αποκλειστικά και μόνο ένας κοινά αποδεκτός ορισμός σχετικά με την Επεξεργασία Φυσικής Γλώσσας και ως εκ τούτου η έννοια περιέχει ένα μεγάλο εύρος αντιληπτών ενεργειών, τεχνικών και αποτελεσμάτων. Σε γενικές γραμμές η NLP μπορεί να προσδιοριστεί σε εννοιολογικό επίπεδο σαν ένα άθροισμα υπολογιστικών τεχνικών που χρησιμοποιούνται για την ανάλυση και την αναπαράσταση κειμένων σε ένα ή ακόμη και περισσότερα επίπεδα γλωσσολογικής ανάλυσης, με απώτερο στόχο την επίτευξη της αναγκαίας γλωσσικής επεξεργασίας η οποία μοιάζει με την αντίστοιχη του ανθρώπου (Alharbi et.al., 2021). Πιο απλά η NLP βοηθάει στην εξαγωγή νοήματος από τις μηχανές έτσι ώστε να είναι κατανοητό και αντιληπτό από την ανθρώπινη γλώσσα (Arnarsson et.al., 2021). Ο αλγόριθμός για να το πετύχει αυτό, εκτελεί διάφορες διεργασίες εξαγωγής νοημάτων από τα κείμενα που έχουν εισαχθεί, συλλέγοντας βασικά δεδομένα από κάθε πρόταση του κειμένου (Arnarsson et.al., 2021). Ως πρακτική έχει εφαρμογή σε διάφορους κλάδους λαμβάνοντας διαφορετικές μορφές και διαφορετικούς τίτλους όπως Εξόρυξη Κειμένου (Text Mining), Κειμενική Ανάλυση (Textual Analysis), Ανάλυση Περιεχομένου (Content Analysis) και Υπολογιστική Γλωσσολογία (Computational Linguistics).

Ιδιαίτερο σημαντικό είναι το γεγονός ότι τα διάφορα εργαλεία που υπάρχουν και οι εφαρμογές που αξιοποιούνται στην NLP βασίζονται σε τεχνολογίες ανοιχτού κώδικα, κάτι το οποίο δηλώνει ότι οι διάφορες λύσεις είναι ελεύθερα διαθέσιμες σε οποιονδήποτε το επιθυμεί, θέλει να τις αξιοποιήσει ή να συνεισφέρει σε αυτές (Baldwin et.al., 2022). Τα δεδομένα του κειμένου μπορεί να εμφανίζονται δομημένα ή ακόμη και μη δομημένα, ενώ τις περισσότερες φορές αναγκαία είναι η εφαρμογή διαφόρων τεχνικών και απαιτούνται πολλά βήματα προ-επεξεργασίας των δεδομένων, προκειμένου να

ετοιμαστούν για ανάλυση (Baldwin et.al., 2022). Τέλος, θα πρέπει να αναφερθεί ότι οι επιχειρήσεις εφαρμόζουν διάφορες εφαρμογές NLP όπως chatbots, σύστημα ανίχνευσης μηνυμάτων ανεπιθύμητης αλληλογραφίας (spam), σύστημα αναγνώρισης ομιλίας, σύστημα αυτόματης μετάφρασης γλώσσας, σύστημα ανάλυσης συναισθημάτων και πολλές άλλες εφαρμογές (Baldwin et.al., 2022).

2.2 Βήματα που περιέχονται στην Επεξεργασία Φυσικής Γλώσσας

Πληθώρα παραγόντων επηρεάζουν το βαθμό δυσκολίας της NLP, όπως για παράδειγμα το γεγονός ότι υπάρχουν πάρα πολλές φυσικές γλώσσες με καθεμία από αυτές να διαθέτει διαφορετικούς δομικούς και κανόνες σύνταξης (Bilbao-Jayo & Almeida, 2018). Παράλληλα η σημασία και το νόημα των λέξεων είναι άρρηκτα συνδεδεμένο με το περιβάλλον, δημιουργώντας σημαντικά εμπόδια στην NLP. Παράγοντες που σχετίζονται με το επίπεδο των χαρακτήρων του κειμένου επηρεάζουν και αυτοί την NLP (Bilbao-Jayo & Almeida, 2018). Για παράδειγμα θα πρέπει να ληφθεί υπόψη το σχήμα κωδικοποίησης που περιέχεται στο κείμενο, ο βαθμός ευαισθησίας του κειμένου στα πεζά ή στα κεφαλαία γράμματα, τα διάφορα σημεία στίξης καθώς και οι αριθμοί, με όλα τα άνωθι στοιχεία να απαιτούν ειδικές επεξεργασίες (Bohlouli, et.al., 2015).

Σε μια φυσική συνομιλία, δηλαδή μια συνομιλία ανάμεσα σε ανθρώπους, κάποια στοιχεία μπορεί να μην ειπωθούν ή μπορεί να εννοηθούν μέσα από κάποια έκφραση, τη στάση του σώματος, ή άλλες ενδείξεις. Ενώ όμως οι άνθρωποι μπορούν να κατανοήσουν την υποκείμενη πρόθεση μιας συνομιλίας, οι υπολογιστές δεν έχουν αυτήν την ικανότητα. Μια ακόμη δυσκολία υφίσταται αναφορικά με την ασάφεια των προτάσεων, η οποία μπορεί να αφορά το επίπεδο της λέξης, το επίπεδο της ίδιας της πρότασης ή ακόμη και της σημασίας (έννοιας) που έχει η πρόταση (Bohlouli, et.al., 2015). Η NLP περιέχει μια πληθώρα διαφορετικών τεχνικών για την αποτελεσματική ερμηνεία της ανθρώπινης γλώσσας, με τις τεχνικές να κυμαίνονται από τη χρήση στατιστικών μεθόδων (statistical methods) και μεθόδων Μηχανικής Μάθησης, μέχρι τεχνικές οι οποίες είναι βασισμένες σε συγκεκριμένους κανόνες (rule-based methods) και αλγοριθμικές προσεγγίσεις (algorithmic approaches) (Chang et.al., 2021).

Καθώς τα δεδομένα τα οποία υπάρχουν σε ένα κείμενο έχουν μια μεγάλη ποικιλία, αναγκαία είναι η ύπαρξη διαφορετικών προσεγγίσεων και πρακτικών εφαρμογών (Chang et.al., 2021). Κάποιες από τις βασικές εργασίας της NLP περιέχουν την διακριτοποίηση (tokenization), τη λημματοποίηση (lemmatization), τη συντακτική ανάλυση (parsing), τη

στελέχωση (stemming), την επισήμανση των μερών του λόγου (part-of-speech tagging), τον εντοπισμό της γλώσσας (language detection) και την αναγνώριση των σημασιολογικών σχέσεων που υπάρχουν (identification of semantic relationships) (Chang et.al., 2021). Εδώ θα πρέπει να αναφερθεί ότι δεν υπάρχει αποκλειστικά και μόνο ένας ορθός τρόπος για να γίνει ανάλυση NLP, καθώς ο χρήστης μπορεί να εφαρμόσει πολλαπλούς τρόπους και να υιοθετήσει διάφορες προσεγγίσεις αναφορικά με τη διαχείριση των δεδομένων που περιέχονται σε ένα κείμενο (Chau et.al., 2023).

Από άποψη Μηχανικής Μάθησης όμως υπάρχουν πέντε βασικά βήματα τα οποία θα πρέπει να ακολουθήσει ο χρήστης προκειμένου να μπορέσει να προετοιμάσει τα δεδομένα του κειμένου για ανάλυση. Τα βήματα αυτά αφοράνε το διάβασμα του σώματος κειμένου (corpus), τη διακριτοποίηση (tokenization), τον καθαρισμό του κειμένου και την παράλληλη αφαίρεση των τερματικών όρων (stopword removal), την στελέχωση (stemming) και τέλος την μετατροπή του κειμένου σε διάνυσμα (Chau et.al., 2023).

2.2.1 Προεπεξεργασία Δεδομένων

Κάθε εργασία NLP απαιτεί την προεπεξεργασία του κειμένου και των αντίστοιχων δεδομένων ευρύτερα, πριν να γίνει η εκπαίδευση του μοντέλου (Deng et.al., 2020). Τα μοντέλα δεν έχουν τη δυνατότητα να χρησιμοποιήσουν ένα ακατέργαστο κείμενο άμεσα, επομένως ο χρήστης θα πρέπει να «καθαρίσει» το κείμενο από μόνος του πριν το εισάγει ως δεδομένα (Deng et.al., 2020). Ανάλογα τη φύση της εργασίας που θα πρέπει να γίνει καθώς και το αντίστοιχο πρόβλημα που έχει οριστεί, υπάρχουν διαφορετικές μέθοδοι προεπεξεργασίας που μπορούν να αξιοποιηθούν (Deng et.al., 2020). Για την γλωσσική ανάλυση ενός ψηφιακού κειμένου φυσικής γλώσσας, είναι αναγκαίος ο καθορισμός των χαρακτήρων, των λέξεων και των προτάσεων οποιουδήποτε εγγράφου με σαφήνεια (Dimisianos, 2019). Ο καθορισμός όμως αυτός παρουσιάζει προκλήσεις με βάση τη γλώσσα που αξιοποιείται για την επεξεργασία, την πηγή προέλευσης των εγγράφων καθώς και την ποικιλία των διαφόρων συστημάτων γραφής που υπάρχουν (Dimisianos, 2019).

Θα πρέπει να αναφερθεί πως οι διάφορες φυσικές γλώσσες καθώς και τα συστήματα γραφής περιέχουν εγγενείς αμφισημίες, δηλαδή στοιχεία τα οποία έχουν διττή ερμηνεία και ως εκ τούτου η γλώσσα καθίσταται διφορούμενη (Ding et.al., 2023) (για παράδειγμα το κάθομαι, έχει τη βασική σημασία του ‘κάθομαι (στην καρέκλα)’, αλλά μπορεί να σημαίνει και ‘μένω, διαμένω’. Έτσι η ερώτηση Πού κάθεσαι; είναι δυνάμει αμφίσημη,

και η αμφισημία οφείλεται στην πολυσημία της λέξης κάθομαι). Οι εφαρμογές NLP καλούνται να επιλύσουν τη σημαντική αυτή πρόκληση των αμφισημιών. Θεμελιώδη στοιχεία για την επίλυση των προκλήσεων αυτών είναι οι χαρακτήρες, οι λέξεις και οι προτάσεις που περνάνε στα περαιτέρω στάδια της επεξεργασίας (Ding et.al., 2023). Αυτά τα επιπρόσθετα στάδια επεξεργασίας μπορεί να απορρέουν από στοιχεία ανάλυσης και επισήμανσης, όπως μορφολογικούς αναλυτές (morphological analyzers) και εύρεση των μερών του λόγου (part-of-speech taggers), μέχρι και διάφορες εφαρμογές όπως ανάκτηση πληροφοριών (information retrieval) και τα συστήματα αυτόματης μετάφρασης (machine translation systems) (Ding et.al., 2023). Ως βασικά βήματα της προεπεξεργασίας, μετά την απόκτηση ενός κειμένου και με βάση το είδος της εκάστοτε εφαρμογής που αξιοποιείται, μπορεί να είναι εργασίες όπως μετατροπή γραμμάτων σε πεζά, μετατροπή αριθμών σε λέξεις ή ακόμη και κατάργηση αριθμών, κατάργηση διάφορων σημείων στίξης, σημείων έμφασης και άλλων διακριτικών στοιχείων, κατάργηση λευκών διαστημάτων και επέκταση των συντομογραφιών (Ding et.al., 2023), όπως φαίνεται και στην παρακάτω Εικόνα 1. Στην εικόνα η εντολή (input) είναι: *The 5 biggest countries by population in 2017 are China, India, United States, Indonesia and Brazil.* Μετά από την επεξεργασία φαίνεται ξεκάθαρα ότι υπάρχει αφαίρεση/κατάρφηση αριθμών, μετατροπή κεφαλαίων σε πεζά γράμματα καθώς και διαφόρων σημείως στίξης με αποτέλεσμα το τελικό κείμενο να είναι: *the biggest countries by population in are china indua united states Indonesia and brazil.*

```
In [1]: import re
        import string
        input_str = "\t The 5 biggest countries by population in 2017 are China, India, \
        United States, Indonesia, and Brazil.!?"
        input_str = input_str.strip()
        input_str = input_str.lower()
        input_str = re.sub(r'\d+', '', input_str)
        text_p = "".join([char for char in input_str if char not in string.punctuation])
        print(text_p)

        the  biggest countries by population in  are china india united states indonesia and brazil
```

Εικόνα 1. Παράδειγμα βασικής προεπεξεργασίας κειμένου (Ιδία επεξεργασία)

2.2.2 Αναγνώριση μέρους του λόγου

Η διαδικασία της Αναγνώρισης Μέρους του Λόγου (Parts-of-Speech Tagging (POS)) αφορά την επισήμανση κάθε λέξης σχετικά με το μέρος του λόγου όπου ανήκει (Doğan,

Balcioglu & Elçi, 2024). Κάθε λέξη που υπάρχει σε μια πρόταση επισημαίνεται σε μια ετικέτα η οποία με τη σειρά της υποδηλώνει τη χρήσης της λέξης στην πρόταση (Doğan, Balcioglu & Elçi, 2024). Οι ετικέτες αυτές υποδηλώνουν μια συντακτική ταξινόμηση, όπως ρήμα, ουσιαστικό, επίρρημα κτλ, ενώ άλλες φορές μπορεί να περιέχουν και πρόσθετες πληροφορίες με δείκτες (Doğan, Balcioglu & Elçi, 2024). Αποτελεί ένα σημαντικό βήμα για πολλές εφαρμογής NLP καθώς μέσα από τον προσδιορισμό του μέρους του λόγου μιας λέξης, μπορεί να γίνει καλύτερη εκτίμηση της σημασίας της με βάση τα συμφραζόμενα (Djohari, 2016) (πχ η έννοια και η σημασία μιας λέξης μπορεί να είναι διαφορετικά στοιχεία όταν γίνεται χρήση της λέξης ως ουσιαστικό, σε σύγκριση με το εάν αξιοποιείται σαν επίθετο).

Υπάρχουν δυο βασικές προσεγγίσεις οι οποίες ακολουθούνται στη διαδικασία της Αναγνώρισης Μέρους του Λόγου, η Προσέγγιση Βάση Κανόνα (Rule Based Approach) και η Στοχαστική Προσέγγιση (Stochastic Approach) (Djohari, 2016). Η προσέγγιση βάση κανόνα χρησιμοποιεί μια μεγάλη βάση δεδομένων από λέξεις μαζί με προκαθορισμένους κανόνες αποσαφήνισης (Djohari, 2016). Η Στοχαστική Προσέγγιση από την άλλη πλευρά αξιοποιεί διάφορες στατιστικές πληροφορίες για την πιθανότητα της εκχώρησης ετικέτας σε λέξεις (Djohari, 2016). Ως προσέγγιση στηρίζεται σε στατιστικούς αλγορίθμους και όχι σε γραμματικούς κανόνες, με τις στοχαστικές ετικέτες να έχουν ευρεία εφαρμογή σε σύγκριση με τις ετικέτες βάσει κανόνων λόγω της δυνατότητας παροχής υψηλότερου βαθμού ακρίβειας. Ωστόσο, θα πρέπει να αναφερθεί ότι η συγκεκριμένη μεθοδολογία είναι αρκετά περίπλοκη και απαιτεί τεράστιες ποσότητες αποθηκευμένων πληροφοριών (Duan et.al., 2016).

2.2.3 Διαδικασία Διακριτοποίησης

Ένα από τα αρχικά βήματα της προεπεξεργασίας του κειμένου, το οποίο συντελεί σημαντικά στη δημιουργία ενός λεξιλογίου, είναι ο διαχωρισμός των εγγράφων ή των προτάσεων σε πιο μικρά κομμάτια τα οποία καλούνται λεξικογραφικές μονάδες (tokens) (Duan et.al., 2016). Κάθε λεξικογραφική μονάδα φέρνει μια σημασιολογική έννοια η οποία είναι σχετική με αυτή. Η συγκεκριμένη διαδικασία διαχωρισμού ονομάζεται διακριτοποίηση (tokenization) (Fang & Wang, 2022). Η λεξικογραφική μονάδα αποτελεί ένα κομμάτι ενός συνόλου, έτσι μια λέξη αποτελεί λεξικογραφική μονάδα σε μια πρόταση και μια πρόταση είναι μια λεξικογραφική μονάδα σε μια παράγραφο (Fang & Wang, 2022).

Ως διαδικασία είναι θεμελιώδης και πρέπει να γίνει σε οποιαδήποτε δραστηριότητα της επεξεργασίας κειμένου (Fang & Wang, 2022). Είναι η τεχνική της τμηματοποίησης στην οποία γίνεται προσπάθεια διάσπασης μεγαλύτερων κομματιών κειμένων σε μικρότερα, για τα οποία το νόημα τους θα είναι ευκολότερο να κατανοηθεί (Βλέπε Εικόνα 2). Σε γενικές γραμμές περιέχουν αριθμούς και λέξεις, αλλά μπορούν και να επεκταθούν ώστε να περιέχουν σημεία στίξης, σύμβολα ή και συνδυασμούς αυτών των δύο (Gárdos, et.al., 2023).

```
In [1]: #!pip install nltk
        import nltk
        #nltk.download()
        text = "It is a truth universally acknowledged, that a single man in\
        possession of a good fortune, must be in want of a wife."
        from nltk.tokenize import word_tokenize
        tokens = word_tokenize(text)
        print(tokens)

        ['It', 'is', 'a', 'truth', 'universally', 'acknowledged', ',', 'that', 'a', 'single', 'man', 'inposse
        ssion', 'of', 'a', 'good', 'fortune', ',', 'must', 'be', 'in', 'want', 'of', 'a', 'wife', '.']
```

Εικόνα 2. Παράδειγμα Διακριτοποίησης (Ιδία επεξεργασία)

Τέλος, θα πρέπει να αναφερθεί ότι υπάρχουν διαφορετικοί τύποι διακριτοποίησης οι οποίοι μπορούν να αξιοποιηθούν για τη σωστή διαμόρφωση του κειμένου με βάση πάντα το αντίστοιχο πρόβλημα ή την εφαρμογή που χρησιμοποιείται (Hamaz & Benchikha, 2017). Οι κυρίαρχοι τύποι διακριτοποίησης είναι ο Regular expressions-based tokenizers, ο Treebank tokenizer και ο TweetTokenizer, με κάθε έναν από τους τύπους να έχουν διαφορετικό σκοπό και τρόπο χρήσης.

2.2.4 Αφαίρεση των Τερματικών Όρων

Στη διαδικασία της προεπεξεργασίας, μια από τις σημαντικότερες μορφές είναι το φιλτράρισμα των άχρηστων δεδομένων (Hamaz & Benchikha, 2017). Στην NLP, οι άχρηστες λέξεις (τα δεδομένα) αναφέρονται ως τερματικοί όροι ή αλλιώς ως λέξεις διακοπής (stop words). Θα πρέπει να γίνει η αφαίρεση των λέξεων οι οποίες δεν έχουν κάποιο σημασιολογικό περιεχόμενο (Hartmann & Netzer, 2023). Οι τερματικοί όροι αποτελούν συνήθως λέξεις σύνδεσης οι οποίες δεν συμβάλλουν στην έννοια της πρότασης, αλλά απαιτούνται να συμπληρωθούν για την γραμματική ορθότητα της πρότασης (Hartmann & Netzer, 2023). Οι λέξεις αυτές είναι ως επί το πλείστον οι πιο κοινές λέξεις σε μια φυσική γλώσσα και μπορούν εύκολα να επεξεργαστούν στις περισσότερες εργασίες της NLP βοηθώντας παράλληλα στη μείωση του λεξιλογίου ή του χώρου αναζήτησης (Hartmann & Netzer, 2023).

Ειδικότερα θα πρέπει να αναφερθεί ότι οι μηχανές αναζήτησης είναι έτσι προγραμματισμένες στο να αγνοούν τους τερματικούς όρους, τόσο κατά την ευρετηρίαση καταχωρίσεων για αναζήτηση, όσο και στη διαδικασία ανάκτησης τους ως απόρροια ενός ερωτήματος αναζήτησης (He et.al., 2019). Οι τερματικοί αυτοί όροι θα πρέπει να αφαιρεθούν για να μην καταλαμβάνουν αναγκαίο χώρο στη βάση δεδομένων ή πολύτιμο χρόνο στην επεξεργασία (He et.al., 2019).

2.2.5 Διαδικασία στελέχωσης

Η ύπαρξη μεγάλου βαθμού ποικιλίας στις λέξεις και ειδικότερα η ύπαρξη μεγάλου αριθμού μορφολογικών παραλλαγών είναι μια σημαντική πρόκληση που καλείται να αντιμετωπίσει η NLP (Heath et.al., 2023). Η διαδικασία της στελέχωσης (stemming) αποτελεί μια τεχνική προεπεξεργασίας η οποία έχει τη δυνατότητα να διαχειριστεί τις παραλλαγές αυτές (Heath et.al., 2023). Είναι μια διαδικασία συγχώνευσης διαφόρων παραλλαγών μιας λέξης σε κοινό στέλεχος (stem), δηλαδή σε λέξεις οι οποίες προέρχονται από την ίδια ρίζα, στις οποίες στη συνέχεια αποκόπτεται η κατάληξη τους και γίνεται υποβίβαση στη ρίζα τους προκειμένου να ανεξαρτητοποιηθούν από τις μορφολογικές παραλλαγές (Heath et.al., 2023) (*πχ οι λέξεις παίζαμε, παίξαμε, παίζουν, μπορούν να μειωθούν σε ένα κοινό στέλεχος το οποίο είναι η λέξη παίζω*).

Η στελέχωση ως διαδικασία είναι πολύ σημαντική καθώς μειώνει το μέγεθος του ευρετηρίου καθώς και του χώρου που απαιτείται για την αποθήκευση διάφορων δομών (Hossain & Rahman, 2023). Επίσης βοηθάει σημαντικά στην επίλυση του προβλήματος της αναντιστοιχίας λεξιλογίου (vocabulary mismatch), συμβάλλοντας σημαντικά στη βελτίωση των συστημάτων Ανάκτησης Πληροφοριών σε ζητήματα ακρίβειας (precision) και ανάκλησης (recall) (Hossain & Rahman, 2023). Βέβαια θα πρέπει να αναφερθεί ότι μπορεί να προκύψουν πιθανά σφάλματα (προβλήματα) στη διαδικασία της στελέχωσης τα οποία χαρακτηρίζονται ως σφάλματα υπερστελέχωσης (over stemming errors) ή σφάλματα υποστελέχωσης (under stemming errors) (Hossain & Rahman, 2023). Στην περίπτωση της υποστελέχωσης δυο διαφορετικές λέξεις πρέπει να ομαδοποιηθούν στην ίδια ρίζα κάτι το οποίο δεν γίνεται και καλείται false negative, ενώ αντιθέτως στην υπερστελέχωση δυο διαφορετικές λέξεις με διαφορετικά στελέχη έχουν ομαδοποιηθεί στην ίδια ρίζα, κάτι το οποίο καλείται false positive (Hossain & Rahman, 2023).

2.2.6 Διαδικασία Λημματοποίησης

Τέλος, ένα ακόμη σημαντικό και θεμελιώδες στοιχείο στην NLP είναι η διαδικασία της λημματοποίησης (lemmatization) (Hutchinson, 2020). Ως διαδικασία αποτελεί τον μορφολογικό μετασχηματισμό ο οποίος μετατρέπει μια λέξη που εμφανίζεται σε κάποιο κείμενο στη βασική της μορφή ή ακόμη καλύτερη στη μορφή του λεξικού της, με την μορφή αυτή να καλείται λήμμα (Hutchinson, 2020). Το λήμμα αποτελεί τον βασικότερο και πιο χαρακτηριστικό τύπο με τον οποίο γράφεται και εξετάζεται μια λέξη. Έχει το ρόλο της διαδικασίας ομαλοποίησης στην οποία διάφορες μορφολογικές παραλλαγές των λέξεων χαρτογραφούνται στο ίδιο υποκείμενο λήμμα, προκειμένου στη συνέχεια να μπορούν να αναλυθούν ως ένα μόνο στοιχείο (όρος ή έννοια) (Hutchinson, 2020). Μέσα από τη μείωση του συνολικού αριθμού διακριτών όρων, η διαδικασία της λημματοποίησης μειώνει την πολυπλοκότητα του κειμένου που πρέπει να αναλυθεί και ως εκ τούτου αποφέρει σημαντικά οφέλη στα μετέπειτα στοιχεία επεξεργασίας του κειμένου (Kehl, Jackson & Fergnani, 2020).

Εν αντιθέσει με τη διαδικασία της στελέχωσης, στην οποία όπως ειπώθηκε και προηγουμένως κάποιοι χαρακτήρες αφαιρούνται από τις λέξεις μέσα από τη χρήση ακατέργαστων μεθόδων, η λημματοποίηση είναι μια διαδικασία στην οποία το περιεχόμενο του κειμένου σε συνολικό επίπεδο, αξιοποιείται για την μετατροπή μιας λέξης σε μια ουσιαστική μορφή βάσης (Kehl, Jackson & Fergnani, 2020). Έτσι ενώ στη στελέχωση οι λέξεις μειώνονται σε στελέχη, στη λημματοποίηση οι λέξεις μειώνονται σε γλωσσολογικά λήμματα (Kehl, Jackson & Fergnani, 2020). Η διαφορά αυτή είναι περισσότερο αντιληπτή σε γλώσσες οι οποίες έχουν περίπλοκες μορφολογικές ιδιαιτερότητες (Kehl, Jackson & Fergnani, 2020). Καθώς όμως η μέθοδος της στελέχωσης έχει επίδραση πάνω σε μια μόνο λέξη την φορά, χωρίς να υπάρχει γνώση για το περιεχόμενο του κειμένου, το γεγονός αυτό την καθιστά δυσκολότερα ικανή για την διάκριση παρόμοιων λέξεων οι οποίες όμως μπορεί να έχουν διαφορετικό νόημα (Keramatfar & Amirkhani, 2019).

Οι διάφοροι αλγόριθμοι λημματοποίησης (lemmatizers), προσπαθούν να αναγνωρίσουν το λήμμα μιας λέξης λαμβάνοντας όμως υπόψη τους το περιβάλλον/τη γειτονία της λέξης, τις ετικέτες του μέρους του λόγου, την ευρύτερη σημασίας της λέξης και άλλα στοιχεία (Keramatfar & Amirkhani, 2019). Ως γειτονιά της λέξης καλείται το λήμμα το οποίο είναι σε κοντινές λέξεις, προτάσεις ή ακόμη και σε ολόκληρο το έγγραφο στο οποίο

θα περιέχεται μέσα η λέξη (Keramatfar & Amirkhani, 2019). Όπως γίνεται αντιληπτό οι ίδιες λέξεις μπορούν να έχουν διαφορετικά λήμματα με βάση το περιβάλλον/τη γειτονιά τους. Ο πιο συχνά χρησιμοποιούμενος αλγόριθμος λημματοποίησης είναι ο WordNet lemmatizer, ενώ άλλοι ευρέως γνωστοί αλγόριθμοι είναι ο Spacy lemmatizer, ο TextBlob lemmatizer και ο Gensim lemmatizer.

2.3 Σώμα κειμένου (Corpus) και σημαντικότητα ύπαρξης του

Οι διάφορες εφαρμογές οι οποίες σχετίζονται και χρησιμοποιούνται για την Επεξεργασία Φυσικής Γλώσσας διαχειρίζονται και επεξεργάζονται τεράστιες ποσότητες δεδομένων. Μια συλλογή δεδομένων κειμένου καλείται σώμα κειμένου ή με πιο απλά λόγια καλείται απλώς ως σώμα (corpus) (Khan et.al., 2022). Ένα σώμα κειμένου αποτελεί λοιπόν τη συλλογή των εγγράφων κειμένου (π.χ. τα διάφορα ηλεκτρονικά μηνύματα που υπάρχουν σε έναν λογαριασμό, τα οποία θα αξιοποιηθούν για επεξεργασία και ανάλυση, αποτελούν μια ομάδα κειμένου που καλείται σώμα κειμένου) (Khan et.al., 2022). Σε επίπεδο ορισμού, το σώμα κειμένου (corpus) αναφέρεται σε μια συλλογή από υλικό φυσικής γλώσσας η οποία είναι σε γραπτή ή προφορική μορφή, αποθηκευμένη σε κάποιον υπολογιστή και αξιοποιείται προκειμένου το άτομο να μάθει πώς χρησιμοποιείται η γλώσσα (Khan et.al., 2022). Μπορεί να χρησιμοποιηθεί είτε ως αφετηρία της γλωσσικής περιγραφής είτε σαν ένα μέσο για την επαλήθευση των υποθέσεων σε μια γλώσσα (Kwartler, 2021).

Γίνεται επομένως αντιληπτό ότι το σώμα κειμένου αποτελεί μια συλλογή δεδομένων που έχουν επιλεχθεί με περιγραφικό ή εφαρμοστό τρόπο για την επίτευξη ενός σκοπού (Kwartler, 2021). Το σώμα αυτό ωστόσο θα πρέπει να διαθέτει ένα σύνολο κοινών αποδεκτών θεμελιωδών ιδιοτήτων. Θα πρέπει να είναι αντιπροσωπευτικό, θα πρέπει να διαθέτει πεπερασμένο μέγεθος και υποχρεωτικά να είναι σε ηλεκτρονική μορφή (Kwartler, 2021). Το μέγεθος ενός σώματος κειμένου θα πρέπει να περιορίζεται σε ένα συγκεκριμένο αριθμό λέξεων (π.χ. 1.000.000 λέξεις) και θα πρέπει να έχει καθοριστεί εκ των προτέρων κατά τη φάση του σχεδιασμού (Lin & Yu, 2023). Στις περιπτώσεις που γίνεται συνεχής ενημέρωση του σώματος κειμένου, αυτό ονομάζεται ως συλλογή κειμένου (text collection), θα πρέπει πάλι να διασφαλίζεται η αντιπροσωπευτικότητα του σώματος με την πάροδο του χρόνου (Lin & Yu, 2023). Αναγκαία είναι λοιπόν η ύπαρξη ενός σώματος κειμένου, το οποίο όπως παρουσιάστηκε και προηγουμένως, θα πρέπει να είναι είτε γραπτό είτε προφορικό υλικό φυσικής γλώσσας σε ηλεκτρονικό υπολογιστή

(Lin & Yu, 2023). Το συγκεκριμένο υλικό ή τα δεδομένα, αξιοποιούνται ως δεδομένα εισόδου και γίνονται οι κατάλληλες προσπάθειες για τον εντοπισμό των γεγονότων που θα βοηθήσουν στην ανάπτυξη των κατάλληλων εφαρμογών NLP (Lin & Yu, 2023).

Σε κάποιες περιπτώσεις οι εφαρμογές NLP κάνουν χρήση ενός σώματος κειμένου (corpus) ως είσοδο, ενώ σε άλλες περιπτώσεις χρησιμοποιούν πολλά σώματα κειμένου (corpora) ως είσοδο (Liu et.al., 2019). Η χρήση των σωμάτων κειμένου είναι καθοριστικής σημασίας και πραγματοποιείται για πληθώρα λόγων. Μέσα από τη βοήθεια του σώματος, ο χρήστης μπορεί να πραγματοποιήσει στατιστικές αναλύσεις, αλληλεξαρτήσεις λέξεων και να καθορίσει ή ακόμη και να επικυρώσει τους κανόνες γλωσσολογίας για τις διάφορες εφαρμογές NLP (Liu et.al., 2019). Για παράδειγμα, εάν ο χρήστης ήθελε να δημιουργήσει ένα σύστημα διόρθωσης γραμματικής, μπορεί να αξιοποιήσει το σώμα κειμένου και να προσπαθήσει να εντοπίσει τις γραμματικές εσφαλμένες περιπτώσεις και στη συνέχεια να ορίσει κανόνες γραμματικής οι οποίοι θα τον βοηθήσουν να διορθώσει τις περιπτώσεις αυτές. Σημαντικό στοιχείο είναι το γεγονός ότι μέσα από το σώμα κειμένου ο χρήσης μπορεί να ορίσει ορισμένους γλωσσικούς κανόνες οι οποίοι είναι συνδεδεμένοι και εξαρτώνται από τη χρήση της γλώσσας (Liu et.al., 2019). Με την αξιοποίηση ενός συστήματος το οποίο έχει βασιστεί σε κανόνες (rulebased system), γίνεται καλύτερη επικύρωση των γλωσσικών κανόνων (Liu et.al., 2019).

Σε ένα σώμα (corpus), η συλλογή των δεδομένων μπορεί να λαμβάνει τη μορφή των δεδομένων κειμένου (δηλαδή υλικό το οποίο είναι γραπτό) ή να λαμβάνει τη μορφή των δεδομένων ομιλίας (δηλαδή υλικό το οποίο είναι προφορικό) (Liu et.al., 2019). Τα δεδομένα κειμένου αποτελούν μια συλλογή γραπτών πληροφοριών όπως ηλεκτρονικά βιβλία, ψηφιακές βιβλιοθήκες, ιστοσελίδες, ιστολόγια, άρθρα ειδήσεων κα. Από την άλλη πλευρά, ένα σώμα δεδομένων ομιλίας αποτελείται από ένα αρχείο ήχου και από ένα απομαγνητοφωνημένο κείμενο. Σε γενικές γραμμές υπάρχουν τρεις διακριτοί τύποι σωμάτων, το μονογλωσσικό σώμα (monolingual corpus) το οποίο ως τύπος σώματος έχει μια γλώσσα, το δίγλωσσο σώμα (bilingual corpus) το οποίο ως τύπος σώματος έχει δυο γλώσσες και τέλος το πολύγλωσσο σώμα (multilingual corpus) το οποίο ως τύπος σώματος έχει περισσότερες από μια γλώσσες (Liu et.al., 2019).

Σε οποιαδήποτε εφαρμογή NLP καθοριστική και αναγκαία είναι η ύπαρξη σώματος, καθώς αποτελεί το βασικό δομικό στοιχείο, προσφέροντας ποσοτικά δεδομένα τα οποία

στη συνέχεια αξιοποιούνται για τη δημιουργία των εφαρμογών (Liu et.al., 2019). Υπάρχουν όμως και σημαντικές προκλήσεις τις οποίες καλείτε να ξεπεράσει ο χρήστης και οι οποίες μπορούν να μειώσουν σημαντικά την αποτελεσματικότητα, τη δυναμική και την ποιότητα της εργασίας των εφαρμογών NPL. Ο χρήστης θα πρέπει να έχει διασφαλίσει τη διαθεσιμότητα των δεδομένων, ενώ παράλληλα θα πρέπει να υπάρχει συνεχής έλεγχος σχετικά με την ποιότητα τους (Luri, Schau & Ghosh, 2023). Τέλος, αναγκαία είναι η διασφάλιση της επάρκειας των δεδομένων αναφορικά με την ποσότητα τους και η επιλογή των κατάλληλων τύπων δεδομένων που θα αξιοποιηθούν για την επίλυση της δήλωσης του προβλήματος (Luri, Schau & Ghosh, 2023).

2.4 Βασικές Εφαρμογές Επεξεργασίας Φυσικής Γλώσσας

Οι εφαρμογές οι οποίες στηρίζονται στην NLP έχουν σημειώσει ραγδαία αύξηση τα τελευταία χρόνια. Προκειμένου να μπορέσει ο τομέας της NLP να συμβαδίσει με την συνεχώς εξελισσόμενη ζήτηση, επιτακτική είναι η ανάγκη για εύρεση αποτελεσματικών λύσεων και εφαρμογών προκειμένου να επιλυθούν διάφορα πολυσύνθετα προβλήματα. Οι βασικότερες εφαρμογές Επεξεργασίας Φυσικής Γλώσσας είναι η ανάλυση συναισθημάτων, η ανίχνευση σαρκασμού σε κείμενο, η ανάκτηση πληροφορίας, η μηχανική μετάφραση και τα chatbots.

2.4.1 Ανίχνευση σαρκασμού σε κάποιο κείμενο

Ο σαρκασμός αποτελεί μια μορφή επικοινωνίας η οποία από τη φύση της είναι αμφίσημη και έχει σαν αποτέλεσμα το άτομο να δηλώνει κάτι διαφορετικό από αυτό που πραγματικά εννοεί (Majhi & Mukherjee, 2023). Η ανίχνευση του σαρκασμού αποτελεί ένα αρκετά περιορισμένο ερευνητικό πεδίο της NLP, αποτελώντας μια συγκεκριμένη κατηγορία ανάλυσης συναισθημάτων στην οποία αντί να πραγματοποιείται η ανίχνευση του συναισθήματος σε όλο το φάσμα, γίνεται εστίαση στον σαρκασμό (Majhi & Mukherjee, 2023). Απώτερος στόχος του συγκεκριμένου πεδίου είναι η ανίχνευση εάν το κείμενο το οποίο εξετάζεται είναι σαρκαστικό ή δεν είναι (Mus et.al., 2023). Υπάρχει όμως ένα πολύ σημαντικό πρόβλημα καθώς σε αντίθεση με τη συναισθηματική ανάλυση στην οποία οι κατηγορίες των συναισθημάτων είναι αρκετά σαφείς (π.χ. στην περίπτωση της αγάπης αυτή συνδέεται με ένα αντικειμενικά θετικό συναίσθημα, ενώ στην περίπτωση του μίσους αυτό συνδέεται με ένα αρνητικό συναίσθημα, ανεξαρτήτως από τη γλώσσα ομιλίας ή του ατόμου το οποίο ρωτάται), ενώ στην περίπτωση του σαρκασμού τα όρια δεν είναι καθορισμένα (Mus et.al., 2023).

Καθοριστικής σημασίας σε αρχικό επίπεδο είναι η κατανόηση του όρου σαρκασμού και τι ακριβώς ο σαρκασμός περιέχει (Musa et.al., 2023). Ο σαρκασμός αποτελεί τη χρήση της φυσικής γλώσσας η οποία όμως δηλώνει το αντίθετο από αυτό το οποίο ειπώθηκε, με επιδίωξη την περιφρόνηση του άλλου ατόμου ή την κοροϊδία του (Nee et.al., 2022). Με πιο απλά λόγια, ο σαρκασμός αποτελεί την έντονη κοροϊδευτική ειρωνεία και χρησιμοποιείται προκειμένου να κοροϊδευτεί ένα τρίτο πρόσωπο με πλάγιο τρόπο. Αποτελεί μια πρακτική στην οποία οι χρήστες της φυσικής γλώσσας αξιοποιούν θετικές λέξεις προκειμένου να μεταδώσουν ένα αρνητικό μήνυμα (Nee et.al., 2022). Φυσικά θα πρέπει να αναφερθεί ότι ο σαρκασμός ποικίλλει από άτομο σε άτομο και είναι πλήρως συνδεδεμένος με τα διάφορα δημογραφικά, κοινωνικά και πολιτιστικά στοιχεία του ατόμου (Nee et.al., 2022). Αυτό σημαίνει ότι τα άτομα ανάλογα το εθνικό, φυλετικό και πολιτιστικό τους υπόβαθρο μπορούν να αντιληφθούν το σαρκασμό διαφορετικά, ενώ παράλληλα εάν κάποιο άτομο είναι σαρκαστικό δεν σημαίνει αυτό ότι θα γίνει πάντα αντιληπτό από τους υπολοίπους (Panda & Kaur, 2023).

Το γεγονός αυτό δημιουργεί υψηλό βαθμό υποκειμενικότητας στον σαρκασμό, αποτελώντας σημαντικό εμπόδιο για τα διάφορα μοντέλα NLP τα οποία δημιουργούνται με στόχο την ανίχνευση του (Panda & Kaur, 2023). Τα άτομα μέσα από τη χρήση των μέσων κοινωνικής δικτύωσης μπορούν να εκφράσουν τις απόψεις, τις σκέψεις, τις ιδέες τους γενικότερα για διάφορα ζητήματα και θέματα (Panda & Kaur, 2023). Ο σαρκασμός μπορεί να βρεθεί σε αρκετά κοινωνικά δίκτυα αλλά και σε ιστότοπους μικρο-ιστολογίων, με τις πλατφόρμες αυτές να ενθαρρύνουν τους χρήστες να σχολιάζουν τις αναρτήσεις και να εκφράζουν τις απόψεις τους (Paschen, Kietzmann & Kietzmann, 2019). Η ανάγνωση της σαρκαστικής σημασίας των απόψεων αλλά και της κυριολεκτικής της σημασίας είναι καθοριστική και αναγκαία προκειμένου να γίνει ορθή κατανόηση των απόψεων των χρηστών σχετικά με διάφορα θέματα στα κοινωνικά μέσα (Paschen, Kietzmann & Kietzmann, 2019). Ειδικά τα τελευταία χρόνια, ο εντοπισμός σαρκαστικών δημοσιεύσεων στα μέσα κοινωνικής δικτύωσης έχει λάβει από την επιστημονική κοινότητα ιδιαίτερη προσοχή, κυρίως επειδή σαρκαστικά σχόλια που εμφανίζονται με τη μορφή tweets περιέχουν συχνά θετικές λέξεις, οι οποίες όμως αντιπροσωπεύουν αρνητικά ή ανεπιθύμητα χαρακτηριστικά στοιχεία (Paul et.al., 2023). Ειδικότερα, η χρήση Μηχανών Διανυσμάτων Υποστήριξης μπορούν να βοηθήσουν σημαντικά στην αποτελεσματική ανίχνευση του σαρκασμού στο Twitter, ενώ παράλληλα ο συνδυασμός

εφαρμογών όπως τα Συνελικτικά Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης παρέχουν υψηλό ποσοστό ακρίβειας πρόβλεψης (Paul et.al., 2023).

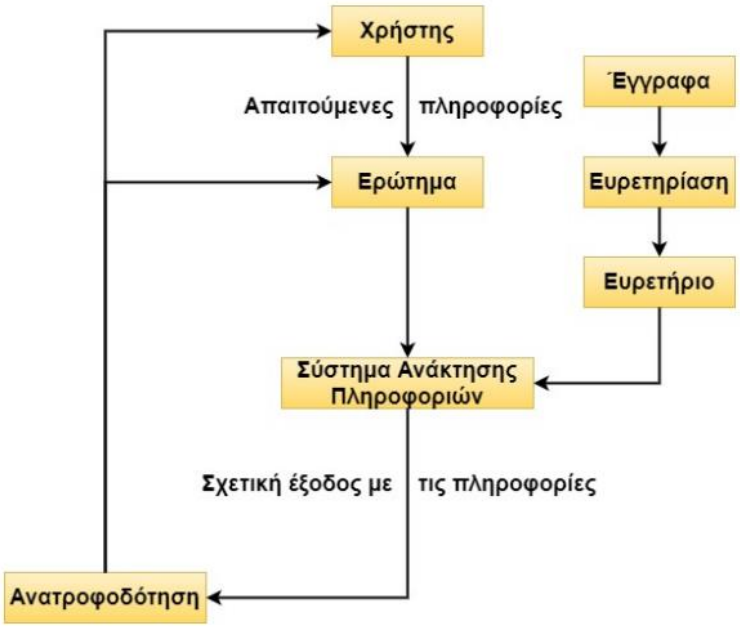
2.4.2 Ανάκτηση Πληροφορίας

Η εφαρμογή της Ανάκτησης Πληροφοριών (Information Retrieval – IR) αποτελεί μια από τις πιο δημοφιλείς και ευρέως εφαρμοσμένες εφαρμογές της NLP. Το πιο απλό παράδειγμα όπου γίνεται χρήση της Ανάκτησης Πληροφοριών είναι η αναζήτηση της Google, όπου ο χρήστης εισάγει ένα ερώτημα εισόδου και ο αλγόριθμος Ανάκτησης Πληροφοριών προσπαθεί να ανακτήσει τις πληροφορίες οι οποίες είναι σχετικές με το ερώτημα αυτό. Η Ανάκτηση Πληροφοριών αποτελεί τη διαδικασία της απόκτησης των πιο σχετικών πληροφοριών τις οποίες χρειάζεται άμεσα και τη δεδομένη χρονική στιγμή ο χρήστης, ενώ θα πρέπει να αναφερθεί ότι υπάρχουν διάφοροι τρόποι μέσα από τους οποίους ο χρήστης μπορεί να απευθυνθεί στο σύστημα προκειμένου να αναζητήσει πληροφορίες και εν τέλει το σύστημα να φέρει τις πληροφορίες οι οποίες είναι πιο σχετικές (Šandor & Bagić Babac, 2023).

Τα δεδομένα τα οποία διαχειρίζεται ένα σύστημα Ανάκτησης Πληροφοριών είναι ημιδομημένα ή μη δομημένα καθώς αυτά αφοράνε κείμενο ή έγγραφα (Šandor & Bagić Babac, 2023). Η Ανάκτηση Πληροφοριών θα μπορούσε επίσης να οριοθετηθεί σαν το πρόγραμμα λογισμικού το οποίο ασχολείται με την οργάνωση, την αποθήκευση, την ανάκτηση και την αξιολόγηση των πληροφοριών (κυρίως όμως πληροφορίες κειμένου) από διάφορα αποθετήρια εγγράφων (Šandor & Bagić Babac, 2023). Το σύστημα παρέχει την κατάλληλη βοήθεια στους χρήστες σχετικά με την εύρεση πληροφοριών τις οποίες χρειάζονται, ωστόσο δεν επιστρέφει ρητά τις απαντήσεις των ερωτήσεων. Ενημερώνει για την ύπαρξη αλλά και τη θέση των εγγράφων που πιθανότητα να περιέχουν τις απαιτούμενες πληροφορίες (Sarmet et.al., 2023). Τα έγγραφα τα οποία ικανοποιούν τις απαιτήσεις του χρήστη καλούνται σχετικά έγγραφα (relevant documents), ενώ ένα ιδανικό σύστημα Ανάκτησης Πληροφοριών θα πρέπει ιδεατά να ανακτά μόνο σχετικά έγγραφα (Sarmet et.al., 2023). Όπως φαίνεται και στην Εικόνα 3, ο χρήστης χρειάζεται τις πληροφορίες και επομένως διατυπώνει ένα αίτημα με τη μορφή ερωτήματος σε φυσική γλώσσα.

Το σύστημα Ανάκτησης Πληροφοριών στη συνέχεια ανταποκρίνεται ανακτώντας τη σχετική έξοδο με τη μορφή εγγράφων, αναφορικά με τις απαιτούμενες πληροφορίες (Sarmet et.al., 2023). Τα ερωτήματα όμως που θέτει ο χρήστης έχουν αρκετά μεγάλη

ασάφεια και ως εκ τούτου τα αποτελέσματα έρχονται με μια σειρά κατάταξης, βαθμολογημένα αναλόγως της συνάφειας τους ως προς το ερώτημα το οποίο έχει τεθεί στη φυσική γλώσσα (Saurwein, Brantner & Möck, 2023). Τα συστήματα Ανάκτησης Πληροφοριών προσπαθούν να βρουν και να ανακτήσουν τα σχετικά έγγραφα τα οποία φαίνεται να έχουν συνάφεια με το ερώτημα του χρήστη, ενώ συγχρόνως επιδιώκουν να μην αποκτήσουν κάποιο έγγραφο το οποίο δεν έχει βαθμό συσχέτισης με την ερώτηση του χρήση (Saurwein, Brantner & Möck, 2023). Ως εκ τούτου, τα αποτελέσματα ταξινομούνται με βάση το ποσοστό συσχέτισης τους (relevance), κάτι το οποίο δείχνει το βαθμό ομοιότητας τους με το ερώτημα καθώς και τη σημαντικότητα τους προς αυτό (Saurwein, Brantner & Möck, 2023).



Εικόνα 3. Διαδικασία ανάκτησης πληροφοριών (Ιδία επεξεργασία) (ShabbirHusain et.al., 2023)

2.4.3 Μηχανική Μετάφραση

Ένα από τα πρώτα προβλήματα που επιδίωξε να αντιμετωπίσει η NLP ήταν αυτό της μετάφρασης της γλώσσας. Στο αποκορύφωμα του Ψυχρού Πολέμου, οι Αμερικανοί ερευνητές είχανε επιτακτική ανάγκη να μεταφράσουν διάφορα Ρωσικά έγγραφα στα Αγγλικά, χρησιμοποιώντας τις διάφορες τεχνικές Τεχνητής Νοημοσύνης. Ειδικότερα το έτος 1964 η κυβέρνηση των Ηνωμένων Πολιτειών Αμερικής Η.Π.Α. ανέπτυξε μια διεπιστημονική επιτροπή κορυφαίων επιστημόνων, γλωσσολόγων και ερευνητών προκειμένου να διερευνήσει τη σκοπιμότητα της Μηχανικής Μετάφρασης, η οποία έφερε

τον τίτλο της Automatic Language Processing Advisory Committee (ALPAC) (ShabbirHusain et.al., 2023). Ωστόσο η συγκεκριμένη ALPAC δεν κατάφερε να φέρει εις πέρας κάποια σημαντική ανακάλυψη, κάτι το οποίο προκάλεσε σημαντικό προβληματισμό σχετικά με τη σκοπιμότητα της τεχνολογίας της Τεχνητής Νοημοσύνης, με αποτέλεσμα να πραγματοποιηθούν τεράστιες περικοπές χρηματοδότησης και μειωμένο ενδιαφέρον για την έρευνα στο συγκεκριμένο τομέα στη δεκαετία του 1970 (Singh, 2019).

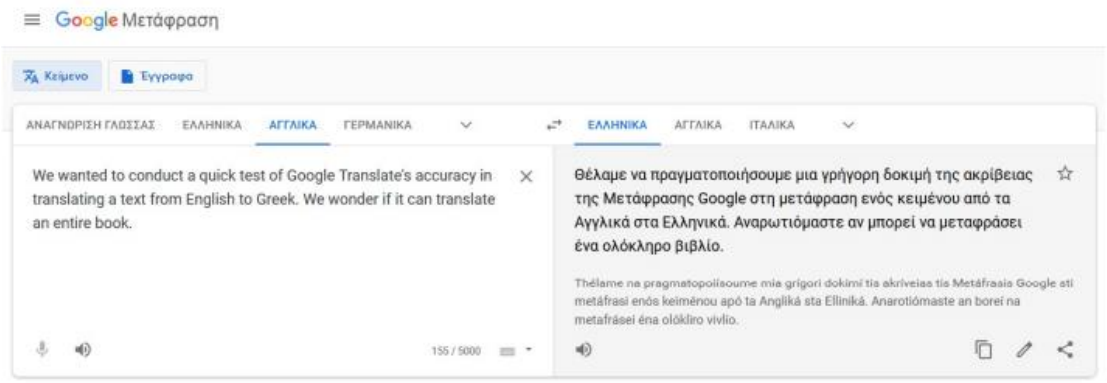
Η κατάσταση αυτή έχει αλλάξει σημαντικά τα τελευταία χρόνια καθώς υπάρχουν μεταφραστές με υψηλό επίπεδο ακρίβειας, με την αξία της Μηχανικής Μετάφρασης (Machine translation) να είναι αρκετά διαδεδομένη και σημαντική. Παρά το γεγονός ότι πολλές επιχειρήσεις αξιοποιούν και στηρίζονται σε ανθρώπινους μεταφραστές προκειμένου να μεταφραστούν σημαντικά έγγραφα όπως νομικά συμβόλαια, η χρήση εφαρμογών NLP στον τομέα αυτό έχει αυξηθεί ραγδαία (Singh, 2019). Οι σύγχρονες προσεγγίσεις NLP για τη μετάφραση των εγγράφων στηρίζονται στη Βαθειά Μάθηση και στην αναγνώριση προτύπων, κάτι το οποίο έχει συμβάλλει σημαντικά στη βελτίωση της ακρίβειας των μεταφράσεων (Singh, 2019).

Η μετάφραση της Google (google translate) χρησιμοποιεί για παράδειγμα ένα σύστημα το οποίο στηρίζεται σε ένα Τεχνητό Νευρωνικό Δίκτυο το οποίο προβλέπει την πιθανή ακολουθία των μεταφρασμένων λέξεων. Ο πιο εύκολος τρόπος για να κατανοηθεί η Μηχανική Μετάφραση είναι η παρατήρηση του τρόπου με τον οποίο οι άνθρωποι κάνουν μετάφραση από μια γλώσσα σε μια άλλη γλώσσα. Το μυαλό του ανθρώπου σε αρχικό επίπεδο αναλύει τη δομή των προτάσεων και επιδιώκει να κατανοήσει την πρόταση. Μόλις η πρόταση γίνει κατανοητή, τότε το μυαλό προσπαθεί να αντικαταστήσει τις λέξεις από την αρχική γλώσσα με τις λέξεις από τη γλώσσα-στόχο που επιθυμεί το άτομο να μεταφράσει. Κατά τη διαδικασία της αντικατάστασης, γίνεται χρήση κανόνων γραμματικής της πρότασης στόχου και εν τέλει επιτυγχάνεται η σωστή μετάφραση. Γενικότερα υπάρχει μια ποικιλία προσεγγίσεων μέσα από τις οποίες μπορεί ένα κείμενο να μεταφραστεί από μια γλώσσα σε μια άλλη. Οι βασικότερες είναι τρεις, με την άμεση μετάφραση να είναι μια από αυτές (Ta-Johnson, Suss & Lande, 2023).

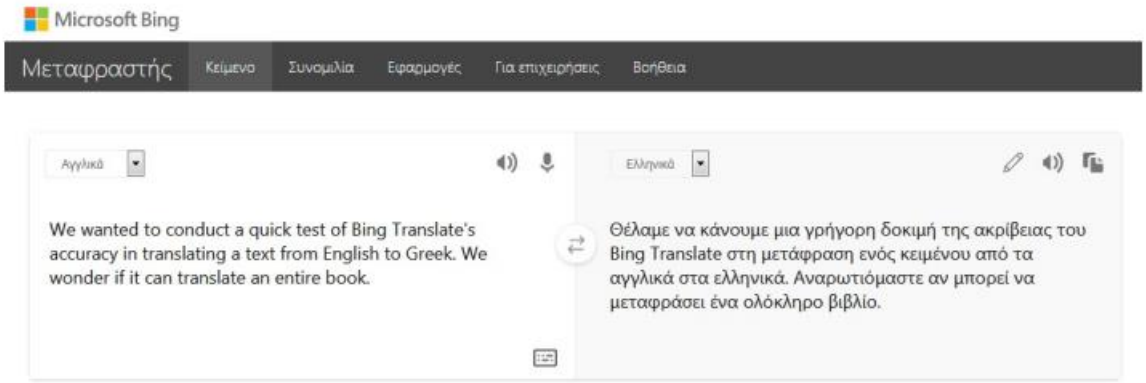
Η άμεση μετάφραση σαν μέθοδος βασίζεται στο λεξικό και απαιτεί την ύπαρξη ενός τεράστιου σώματος κειμένου, τόσο της αρχικής γλώσσας (της γλώσσας πηγής), όσο και της γλώσσας στην οποία ο χρήστης θέλει να μεταφράσει το κείμενο (γλώσσα στόχος)

(Ta-Johnson, Suss & Lande, 2023). Σαν εφαρμογή είναι ιδιαίτερη απλή και δημοφιλής. Η δεύτερη προσέγγιση είναι αυτή της συντακτικής μετάφρασης στην οποία γίνεται προσπάθεια δημιουργίας ενός αναλυτή της γλώσσας πηγής (Taskin & Al, 2019). Γίνεται επιλογή του κατάλληλου αναλυτή οι οποίοι με τη σειρά τους στη συνέχεια πραγματοποιούν την αντικατάσταση της κάθε λέξης στόχου με την τελική γλώσσα-στόχο (Taskin & Al, 2019). Τέλος, η τρίτη προσέγγιση είναι αυτή της στατιστικής μετάφρασης (statistical machine translation SMT) στην οποία υπάρχει ένας συγκεκριμένος αριθμός αλγορίθμου ο οποίος βοηθάει την πρόβλεψη της μετάφρασης της γλώσσας πηγής στη γλώσσα στόχο (Taskin & Al, 2019).

Για την καλύτερη κατανόηση της συγκεκριμένης διεργασίας έγινε μια γρήγορη δοκιμή της ακρίβειας της Μετάφρασης της Google καθώς και της ακρίβειας της Μετάφρασης της Bing, στη μετάφραση ενός κειμένου από τα Αγγλικά στα Ελληνικά (Βλέπε Εικόνα 4 & 5), προκειμένου να αποτυπωθούν τα αποτελέσματα των μεταφράσεων.



Εικόνα 4. Μετάφραση Google

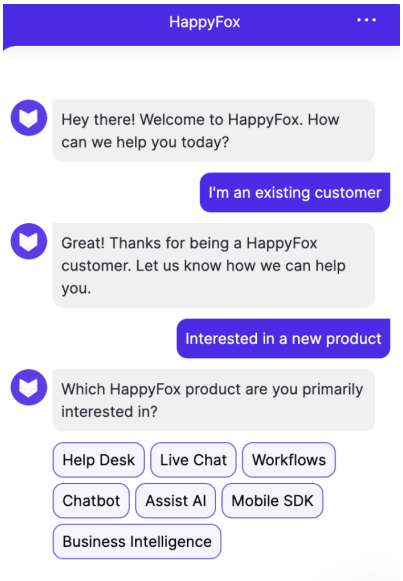


Εικόνα 5. Μετάφραση Bing

Η μετάφραση και στις δυο περιπτώσεις ήταν άμεση, κάτι το οποίο σημαίνει ότι ο χρόνος εκτέλεσης της αυτόματης μετάφρασης έχει μειωθεί σε σημαντικό επίπεδο. Με βάση τις δοκιμές που έγιναν φαίνεται ότι υπάρχουν ακόμη αρκετά περιθώρια βελτίωσης καθώς και αύξησης της συνάφειας και βαθμού ποιότητας του μεταφρασμένου κειμένου.

2.4.4 Chatbot

Τα διάφορα συστήματα Τεχνητής Νοημοσύνης που λειτουργούν ως διεπαφές για την πραγματοποίηση αλληλεπιδράσεων ανάμεσα σε ανθρώπους και μηχανές, είτε μέσω κειμένου, είτε μέσω της φωνής, ονομάζονται Chatbots (Yuan et.al., 2021). Τα chatbots αποτελούν ένα λογισμικό το οποίο στηρίζεται στην τεχνολογία της Τεχνητής Νοημοσύνης και δίνει τη δυνατότητα να πραγματοποιούνται συνομιλίες με ανθρώπους σε φυσικές γλώσσες (Yuan et.al., 2021). Απώτερος στόχος είναι η μίμηση του ανθρώπινου λόγου σε γραπτό και προφορικό επίπεδο όσο το δυνατόν καλύτερα γίνεται (Yuan et.al., 2021). Τα chatbots αποτελούν μια εφαρμογή η οποία έχει την ικανότητα να συνομιλεί με τους ανθρώπους προκειμένου να λύσει ένα πρόβλημα ή να προσφέρει απαντήσεις σε ένα συγκεκριμένο ερώτημα (Zhang & Mu, 2022) (Βλέπε Εικόνα 6).



Εικόνα 6. Παράδειγμα συνομιλίας με chatbot (https://www.brevo.com/blog/chatbot-examples/)

Το chatbot ως εφαρμογή έχει ευνοήσει σε μεγάλο επίπεδο διάφορες επιχειρήσεις, μειώνοντας τον χρόνο απόκρισης τους, ενισχύοντας την αποδοτικότητα τους και προσφέροντας σημαντικά ανταγωνιστικά πλεονεκτήματα (Zhang & Mu, 2022). Τα chatbots μπορεί να έχουν αναπτυχθεί βασισμένα σε συγκεκριμένους κανόνες, αλλά

μπορεί να είναι και πιο εξελιγμένα με βάση τις απαιτήσεις του εκάστοτε οργανισμού (Zhou et.al., 2023). Τα περισσότερα chatbots τα οποία έχουν αναπτυχθεί εκπαιδεύονται προκειμένου να κατευθύνουν τους χρήστης στην κατάλληλη πηγή πληροφοριών ή για να απαντήσουν σε ερωτήματα τα οποία σχετίζονται με ένα καθορισμένο ερώτημα (Zhou et.al., 2023). Η δημιουργία των chatbot θα πρέπει να λαμβάνει υπόψη το κοινό το οποίο πρόκειται να τα χρησιμοποιήσει και επομένως είναι απαραίτητο να ξέρουμε τα δημογραφικά τους στοιχεία (Zhou et.al., 2023).

2.4.5 Ανάλυση Συναισθημάτων

Το chatbot ως εφαρμογή έχει ευνοήσει σε μεγάλο επίπεδο διάφορες επιχειρήσεις, μειώνοντας τον χρόνο απόκρισης τους, ενισχύοντας την αποδοτικότητας τους και προσφέροντας σημαντικά ανταγωνιστικά πλεονεκτήματα (Zhang & Mu, 2022). Τα chatbots μπορεί να έχουν αναπτυχθεί βασισμένα σε συγκεκριμένους κανόνες, αλλά μπορεί να είναι και πιο εξελιγμένα με βάση τις απαιτήσεις του εκάστοτε οργανισμού (Zhou et.al., 2023). Τα περισσότερα chatbots τα οποία έχουν αναπτυχθεί εκπαιδεύονται προκειμένου να κατευθύνουν τους χρήστης στην κατάλληλη πηγή πληροφοριών ή για να απαντήσουν σε ερωτήματα τα οποία σχετίζονται με ένα καθορισμένο ερώτημα (Zhou et.al., 2023). Η δημιουργία των chatbot θα πρέπει να λαμβάνει υπόψη το κοινό το οποίο πρόκειται να τα χρησιμοποιήσει και επομένως είναι απαραίτητο να ξέρουμε τα δημογραφικά τους στοιχεία (Zhou et.al., 2023).

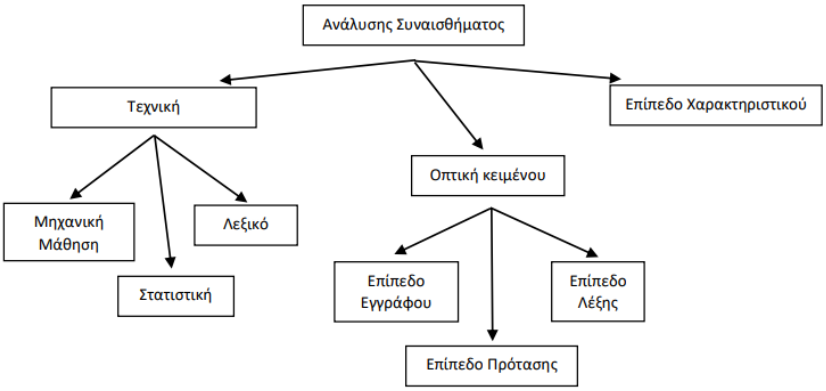
2.4.5 Ανάλυση Συναισθημάτων

Η Ανάλυση Συναισθημάτων (Sentiment Analysis) αποτελεί ένα σύνολο αλγορίθμων και διαφόρων τεχνικών που αξιοποιούνται για την ανίχνευση του συναισθήματος, είτε αυτό είναι θετικό, αρνητικό ή ουδέτερο, ενός δεδομένου κειμένου με τρόπο αυτοματοποιημένο (Abram, Mancini & Parker, 2020). Είναι μια πολύ δυνατή εφαρμογή της NLP η οποία έχει μεγάλο εύρος εφαρμογής σε διάφορες βιομηχανίες (Abram, Mancini & Parker, 2020). Δίνει τη δυνατότητα στους οργανισμούς να εξορίξουν απόψεις από ένα αρκετά ευρύ κοινό με χαμηλό κόστος (Abram, Mancini & Parker, 2020). Παραδοσιακά, οι εταιρείες συγκέντρωναν σχόλια μέσα από έρευνες, από κλειστές ομάδες χρηστών, συνεντεύξεις και άλλα ποιοτικά ή ποσοτικά εργαλεία τα οποία ήταν ακριβά και απαιτούσαν σημαντικό χρόνο για την εξαγωγή αποτελεσμάτων (Abram, Mancini &

Parker, 2020). Καθώς όμως στη σύγχρονη εποχή υπάρχει ένας μεγάλος όγκος πληροφοριών (αξιολογήσεις πχ προϊόντων ή υπηρεσιών) ηλεκτρονικά, οι οποίες σχετίζονται με τα συναισθήματα των ανθρώπων, η Ανάλυση Συναισθημάτων έχει αποκτήσει σημαντικό ενδιαφέρον, τόσο σε ακαδημαϊκό όσο και σε επιχειρηματικό επίπεδο (Alaei, Becken & Stantic, 2019). Τα συστήματα αυτά μπορούν να στηριχθούν είτε σε απλά λεξικά ή σε μεθόδους Μηχανικής Μάθησης και σε μεθόδους Βαθιάς Μάθησης (Alaei, Becken & Stantic, 2019). Η επιλογή της κατάλληλης μεθόδου εξαρτάται από μια πληθώρα παραγόντων καθώς και από τα αντίστοιχα οφέλη και μειονεκτήματα που επιφέρει η κάθε προσέγγιση ανάλογα τους αντίστοιχους στόχους (Alaei, Becken & Stantic, 2019). Στο επόμενο κεφάλαιο γίνεται λεπτομερέστερη παρουσίαση της Ανάλυσης Συναισθήματος.

2.5 Η Ανάλυση Συναισθήματος (Sentiment Analysis) και τα επίπεδα της

Οι διάφορες διαδικασίες που αξιοποιούνται για την Ανάλυση Συναισθημάτων (Sentiment Analysis) ταξινομούνται από διάφορες οπτικές γωνίες (Agarwal, 2022). Η ταξινόμηση αυτή επηρεάζεται από την τεχνική που έχει αξιοποιηθεί, από την οπτική του κειμένου, από το επίπεδο της λεπτομέρειας σχετικά με την ανάλυση του κειμένου, από τη βαθμολόγηση και άλλα στοιχεία (Agarwal, 2022). Αναφορικά με την οπτική γωνία της τεχνικής, υπάρχει η Μηχανική Μάθηση, οι Τεχνικές Βασισμένες σε λεξικό και οι Στατιστικές Τεχνικές (Agarwal, 2022) (Βλέπε Εικόνα 7).



Εικόνα 7. Ταξινόμηση Ανάλυσης Συναισθήματος (Ιδία επεξεργασία) (Agarwal, 2022)

Η αξιοποίηση της προσέγγισης της Μηχανικής Μάθησης σχετίζεται με τη χρήση διάφορων αλγορίθμων μάθησης προκειμένου να γίνει προσδιορισμός του συναισθήματος από τον εκπαιδευόμενο από ένα γνωστό σύνολο δεδομένων (Bilbao-Jayo & Almeida,

2018). Από την άλλη πλευρά, η προσέγγιση η οποία βασίζεται στο λεξικό περιέχει τον υπολογισμό της πολικότητας των συναισθημάτων που βρίσκονται σε ένα κείμενο, χρησιμοποιώντας το σημασιολογικό προσανατολισμό των λέξεων ή των προτάσεων που περιέχονται σε αυτό (*Η πολικότητα των συναισθημάτων αναφέρεται στην ύπαρξη θετικών και αρνητικών συναισθημάτων. Τα θετικά συναισθήματα, όπως η χαρά, η ικανοποίηση και ο ενθουσιασμός, συνδέονται συχνά με θετική διάθεση. Από την άλλη πλευρά, τα αρνητικά συναισθήματα, όπως ο θυμός, ο φόβος και η λύπη, συνδέονται συχνά με αρνητική διάθεση*) (Bilbao-Jayo & Almeida, 2018). Ο σημασιολογικός προσανατολισμός αποτελεί ένα σημαντικό μέτρο υποκειμενικότητας και γνώμης στο κείμενο (Doğan, Balcioglu & Elçi, 2024). Βέβαια στο σημείο αυτό θα πρέπει να γίνει σύντομη αναφορά σχετικά με τους βασικούς περιορισμούς τους οποίους καλούνται να αντιμετωπίσουν οι τεχνικές που είναι βασισμένες σε λεξικό. Ειδικότερα μια βασική αδυναμία είναι η άρνηση (negation), καθώς οι συγκεκριμένες τεχνικές αγνοούν την άρνηση στις λέξεις (Doğan, Balcioglu & Elçi, 2024). Η ύπαρξη όμως μιας άρνησης λέξης επηρεάζει σημαντικά το νόημα της λέξης ή των υπολοίπων λέξεων που ακολουθούν αντιστρέφοντας με τον τρόπο αυτό το polarity (Doğan, Balcioglu & Elçi, 2024).

Για παράδειγμα η πρόταση *The movie was good* σχετίζεται με μια θετική κριτική ενώ η πρόταση *The movie was not good* σχετίζεται με μια αρνητική κριτική. Η προφανής επιλογή αντιμετώπισης του φαινομένου αυτού και αυτή που χρησιμοποιείται συνήθως είναι η αντιστροφή του polarity των λέξεων που ακολουθούν την λέξη άρνησης μέχρι το επόμενο σημείο στίξης ή κάποιον αντιθετικό σύνδεσμο, πχ. but, however κ.ά (Fang & Wang, 2022). Όμως η επιλογή αυτή έχει αδυναμίες. Μια ακόμη σημαντική αδυναμία των τεχνικών βασισμένες σε λεξικό είναι η μετατόπιση της έντασης/του σθένους (valence shifters), καθώς πέραν της άρνησης η οποία επηρεάζει το νόημα των λέξεων που ακολουθούνε, υπάρχουν παράλληλα και λέξεις μεταβολής της έντασης οι οποίες αυξάνουν (intensifiers) ή μειώνουν (downtoners) την ένταση της επόμενης λέξης (Fang & Wang, 2022). Η εξέταση των συγκεκριμένων λέξεων είναι ιδιαίτερα σημαντικά ειδικά όταν γίνεται προσπάθεια για πολυεπίπεδη συναισθηματική κατάταξη (Fang & Wang, 2022).

Τρίτη αδυναμία αποτελεί το γεγονός ότι οι μέθοδοι αυτοί αγνοούν τη σειρά των λέξεων που εμφανίζονται στο κείμενο, με αποτέλεσμα να πραγματοποιείται μια Bow (Bag of Words) μοντελοποίηση του κειμένου (Gárdos et,a.l., 2023). Παράλληλα οι διάφοροι αντιθετικοί σύνδεσμοι (adversative conjuctions) που υπάρχουν σε μια πρόταση δεν

λαμβάνονται υπόψη (Gárdos et,a.l., 2023). Οι σύνδεσμοι αυτοί συνδέουν δυο φράσεις αντίθετης πολικότητας (λέξεις όπως το but, however, altought), ενώ συνήθως το polarity της συνολικής πρότασης καθορίζεται από το δεύτερο συστατικό της (Gárdos et,a.l., 2023). Για παράδειγμα, στην πρόταση *The car is nice but expensive* η προδιάθεση του συγγραφέα ή ομιλητή είναι εναντίον της αγοράς του αυτοκινήτου, ενώ στην πρόταση *The car is expensive but nice*, η προδιάθεση του συγγραφέα ή ομιλητή ως προς την αγορά του αυτοκινήτου είναι θετική.

Μια ακόμη σημαντική αστοχία της συγκεκριμένης μεθόδου είναι ότι μέσα από τη μελέτη της κάθε λέξης σε ξεχωριστό επίπεδο, γίνεται αγνόηση της ύπαρξης φράσεων των οποίων οι επιμέρους λέξεις προσδίδουν ένα ιδιαίτερο συνολικό νόημα (Hamaz & Benchikha, 2017). Προκειμένου να αντιμετωπισθεί η κατάσταση αυτή αναγκαία είναι η χρήση ειδικού λεξικού το οποίο περιέχει συγκεκριμένους ιδιωματισμούς ώστε η ανάλυση του κειμένου να μην γίνει σε επίπεδο λέξεων, αλλά να γίνει σε επίπεδο φράσεων (Hamaz & Benchikha, 2017). Ένα από τα δυσκολότερα εμπόδια που καλείται επίσης να αντιμετωπίσει η συγκεκριμένη τεχνική, είναι αυτό της αμφισημίας. Η αμφισημία είναι το φαινόμενο όπου μια λέξη ή φράση έχει διαφορετικό νόημα με βάση το ευρύτερο νοηματικό πλαίσιο στο οποίο αξιοποιείται (Hamaz & Benchikha, 2017).

Δυσκολία υπάρχει αναφορικά και με την κατανόηση της ειρωνείας (Hartmann & Netzer, (2023). Η ειρωνεία είναι πολλές φορές δύσκολη να κατανοηθεί ακόμη και από τον άνθρωπο, πόσο μάλλον από μια μηχανή. Το πρόβλημα αυτό μπορεί να αντιμετωπισθεί μέσα από την υιοθέτηση και εφαρμογή διάφορων μεθόδων ταξινόμησης των προτάσεων ως σαρκαστική ή όχι, διαδικασία η οποία αυξάνει ακόμη περισσότερο το βαθμό πολυπλοκότητας των τεχνικών βασισμένων σε λεξικό (Hartmann & Netzer, (2023). Τέλος, ένα ακόμη πρόβλημα είναι αυτό των πολλαπλών στόχων (Hartmann & Netzer, (2023). Πολλές φορές σε μια κριτική μπορεί να γίνει αναφορά σε παραπάνω από μια οντότητες, π.χ. σε πρόσωπα, σε γεγονότα και σε προϊόντα ή ακόμη και σε διαφορετικά χαρακτηριστικά (aspects) της ίδιας της οντότητας (Hartmann & Netzer, (2023). Στις περιπτώσεις αυτές όμως το ενδιαφέρον δεν στρέφεται στην ταξινόμηση του κειμένου συνολικά ως μια θετική αρνητική άποψη, αλλά εξετάζεται το κείμενο σε επίπεδο χαρακτηριστικών (aspect level) (Hossain & Rahman, 2023). Η ανάλυση συναισθήματος σε επίπεδο χαρακτηριστικών (Aspect-level SA) στοχεύει στον προσδιορισμό του συναισθήματος ως προς συγκεκριμένα aspects των οντοτήτων και διαφέρει από την ανάλυση συναισθήματος σε επίπεδο κειμένου (Document-level SA) (Hossain & Rahman,

2023). Η εξαγωγή όμως συμπερασμάτων σχετικά με την πολικότητα της άποψης είναι μια εξαιρετικά δύσκολη αν όχι αδύνατη διαδικασία (Hossain & Rahman, 2023). Πολλές φορές τα μηνύματα μπορεί να περιέχουν συντομογραφίες ή ακόμη και ορθογραφικά λάθη λόγω επιπολαιότητας, να περιέχουν λέξεις που είναι σε περισσότερες από μια γλώσσα (πχ χρήση greeklish), καθιστώντας τη διαδικασίας της Ανάλυσης Συναισθήματος πολύ δύσκολη (Hossain & Rahman, 2023).

Για να υπάρχει το αντίστοιχο ποιοτικό αποτέλεσμα αναγκαία είναι η χρήση ενός λεξικού στο οποίο θα αφομοιώνονται οι τάσεις των χρηστών του διαδικτύου, ενσωματώνοντας διάφορες συντομογραφίες λέξεων, misspellings, αργκό και άλλα στοιχεία τα οποία όμως είναι αδύνατο να κατασκευαστούν (Hutchinson, 2020). Σε αντίθεση με τη μηχανική μάθηση, η βασισμένη σε λεξικό μέθοδος δεν απαιτεί την εκπαίδευση ενός ταξινομητή πάνω σε επισημειωμένα δεδομένα εξοικονομώντας έτσι σημαντικό χρόνο (Hutchinson, 2020). Ωστόσο, απαιτεί ένα συναισθηματικό λεξικό, περιορίζοντας την εφαρμογή της μεθόδου στην ανάλυση κειμένων γραμμένων στην γλώσσα του λεξικού και την απόδοσή της στην ποιότητα ή πληρότητα του λεξικού (Hutchinson, 2020). Από την άλλη πλευρά, οι μέθοδοι μηχανικής μάθησης πετυχαίνουν συνήθως καλά αποτελέσματα με αντίτιμο την ανάγκη εκπαίδευσης που μπορεί να είναι χρονοβόρα (Kehl, Jackson & Fergnani, 2020). Γι' αυτό, η επιστημονική έρευνα τα τελευταία χρόνια προσανατολίζεται στη χρήση υβριδικών μεθόδων που συνδυάζουν λεξικό με μηχανική μάθηση ώστε να επωφεληθούν από τα πλεονεκτήματα των επιμέρους μεθόδων, δηλαδή της ταχύτητας της lexicon-based προσέγγισης και της ακρίβειας της machine learning προσέγγισης (Kehl, Jackson & Fergnani, 2020).

Τέλος, η προσέγγιση η οποία στηρίζεται στη στατιστική ανάλυση, έχει σαν απώτερο στόχο την αντιμετώπιση των δυσκολιών που προκύπτουν από την αδυναμία της διαχείρισης κάποιων λέξεων ή κάποιων κειμένων από τις άλλες δυο μεθόδους (Keramatfar & Amirkhani, 2019). Η συγκεκριμένη προσέγγιση περιέχει τη δημιουργία ενός συναισθηματικού λεξικού από ένα μεγάλο σύνολο εγγράφων προκειμένου να χρησιμοποιηθεί για να προσφέρει ένα σημασιολογικό προσανατολισμό σε κάθε λέξη αναλόγως της συχνότητας που εμφανίζεται στα έγγραφα, τα οποία με τη σειρά τους έχουν χαρακτηριστεί σε θετικό ή αρνητικό προσανατολισμό (Keramatfar & Amirkhani, 2019).

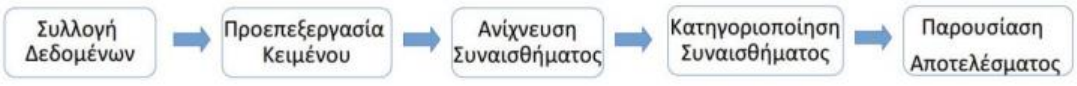
Μια άλλη μορφή ταξινόμησης σχετίζεται περισσότερο με τη δομή του κειμένου και αφορά είτε το επίπεδο του εγγράφου, είτε το επίπεδο των προτάσεων είτε την ταξινόμηση

του επιπέδου των λέξεων ή των χαρακτηριστικών (Keramatfar & Amirkhani, 2019). Ειδικότερα, η ταξινόμηση σε επίπεδο εγγράφου επιδιώκει να βρει μια πολικότητα συναισθημάτων σε ολόκληρο το κείμενο, ενώ αντιθέτως η ταξινόμηση σε επίπεδο προτάσεων ή σε επίπεδο λέξεων έχει τη δυνατότητα να εκφράσει μια πολικότητα συναισθημάτων για κάθε πρόταση ενός κειμένου ή ακόμη και για κάθε λέξη (Kwartler, 2021). Τέλος, υπάρχει και η ταξινόμηση με βάση το επίπεδο του χαρακτηριστικού η οποία στοχεύει στην κατάταξη του συναισθήματος ενός κειμένου με βάση το συναίσθημα κάθε χαρακτηριστικού που έχει εντοπιστεί εντός του κειμένου (Kwartler, 2021).

Θα πρέπει να αναφερθεί ότι κάθε προσέγγιση έχει τα δικά της διακριτά χαρακτηριστικά στοιχεία, με τις μεθόδους που στηρίζονται στα λεξικά να έχουν υψηλό βαθμό ευχρηστίας, ειδικά σε εφαρμογές που αξιοποιούν μεγάλα σύνολα δεδομένων (Kwartler, 2021). Η αυτόματη προσθήκη της ετικέτας σε κάθε λέξη που περιέχεται στο κείμενο, ενισχύει σημαντικά την ανάλυση του σημασιολογικού τους προσανατολισμού (Lin & Yu, 2023). Από την άλλη πλευρά όμως υπάρχει και ένας κίνδυνος ο οποίος σχετίζεται με την εξάρτηση από συναισθηματικά λεξικά και μπορεί να επηρεάσει σημαντικά την ικανότητα του συστήματος συναισθηματικής ανάλυσης με βάση την εξειδίκευση του λεξικού (Lin & Yu, 2023). Η προσέγγιση της τεχνικής της χρήσης λεξικών, η οποία στηρίζεται περισσότερο στη Μηχανική Μάθηση, έχει τη δυνατότητα να λειτουργήσει με μεγάλη επιτυχία κάνοντας χρήση μικρών συνόλων δεδομένων εκπαίδευσης, πετυχαίνοντας προβλέψεις σε πρωτοεμφανιζόμενα δεδομένα, ακόμη και όταν κάποιες λέξεις δεν υπάρχουν στη βάση τους (Lin & Yu, 2023). Τέλος, οι μέθοδοι οι οποίες έχουν βασιστεί στις παραδοσιακές στατιστικές μεθόδους, είναι ικανές να πετύχουν την συναισθηματική κατηγοριοποίηση ενός κειμένου με αποδεκτό βαθμό ακριβείας, μόνο εάν τους δοθεί ένας μεγάλος όγκος συλλογής κειμένων ως είσοδο, με το σύστημα να αποδίδει καλά σε επίπεδο παραγράφου και όχι σε επίπεδο πρότασης (Luri, Schau & Ghosh, 2023).

2.5.1 Στάδια Ανάλυσης Συναισθήματος

Σε γενικές γραμμές, υπάρχουν πέντε βασικά στάδια την Ανάλυση Συναισθήματος όπως φαίνεται και στην Εικόνα 8 παρακάτω.



Εικόνα 8. Στάδια Ανάλυσης Συναισθήματος (Ιδία Επεξεργασία) (Luri, Schau & Ghosh, 2023)

Στο πρώτο στάδιο πραγματοποιείται η διαδικασία της συλλογής των δεδομένων. Παραδείγματος χάριν, οι καταναλωτές εκφράζουν τα συναισθήματα τους σε διάφορα δημόσια φόρουμ όπως τα blogs, σε διάφορα σχόλια προϊόντων καθώς και σε ιδιωτικά αρχεία (Luri, Schau & Ghosh, 2023), όπως και στα διάφορα Μέσα Κοινωνικής Δικτύωσης όπως το Facebook, το Twitter και το TikTok. Οι διάφορες απόψεις και τα διάφορα συναισθήματα αποτυπώνονται με διαφορετικό τρόπο, με διαφορετικό λεξιλόγιο, πλαίσιο γραφής, συντομογραφίες και αργκό με αποτέλεσμα να δημιουργείται ένας μεγάλος όγκος δεδομένων τα οποία εμφανίζονται αρκετά αποδιοργανωμένα (Majhi & Mukherjee, 2023). Καθώς η χειρωνακτική ανάλυση των δεδομένων συναισθημάτων είναι σχεδόν αδύνατη να πραγματοποιηθεί, χρησιμοποιούνται ειδικές γλώσσες προγραμματισμού όπως η R, προκειμένου να γίνει η κατάλληλη επεξεργασία και ανάλυση των δεδομένων (Majhi & Mukherjee, 2023).

Στο δεύτερο στάδιο πραγματοποιείται η προεπεξεργασία του κειμένου η οποία σχετίζεται με το φιλτράρισμα των εξαγόμενων δεδομένων πριν την ανάλυση τους (Majhi & Mukherjee, 2023). Γίνεται εντοπισμός και εξάλειψη του περιεχομένου το οποίο δεν είναι κείμενο καθώς και του περιεχομένου το οποίο δεν έχει σχέση με τον τομέα της μελέτης των δεδομένων (Musa et.al., 2023). Στο συγκεκριμένο στάδιο προκειμένου να επιτευχθεί η Ανάλυση Συναισθημάτων, αξιοποιούνται διάφορες τεχνικές της Επεξεργασίας Φυσικής Γλώσσας και της Ανάκτησης Πληροφορίας (Musa et.al., 2023) (έχουν παρουσιαστεί σε προηγούμενο ενότητα τα βασικά στάδια της προεπεξεργασίας κειμένου).

Στη συνέχεια, στο τρίτο στάδιο, πραγματοποιείται η Ανίχνευση Συναισθήματος, όπου κάθε πρόταση εξετάζεται αναφορικά με την υποκειμενικότητα της (Musa et.al., 2023). Οι προτάσεις οι οποίες έχουν υποκειμενικές εκφράσεις διατηρούνται και αυτές οι οποίες έχουν αντικειμενικές εκφράσεις απορρίπτονται (Nee et.al., 2022). Η Ανάλυση του Συναισθήματος πραγματοποιείται σε διάφορα επίπεδα όπως έχουν παρουσιαστεί προηγουμένως (Nee et.al, 2022). Στο τέταρτο στάδιο πραγματοποιείται η κατηγοριοποίηση του συναισθήματος, με τα συναισθήματα να κατηγοριοποιούνται σε δυο ομάδες στις θετικές και αρνητικές (Nee et.al, 2022). Κάθε υποκειμενική δράση η οποία έχει ανιχνευθεί, κατηγοριοποιείται είτε στις θετικές ομάδες είτε στις αρνητικές (καλές ή κακές) (Panda & Kaur, 2023). Τέλος, το πέμπτο στάδιο, αφορά την παρουσίαση του αποτελέσματος, με την βασική ιδέα της Ανάλυσης Συναισθήματος να αφορά την

μετατροπή ενός αδόμητου κειμένου σε χρήσιμες πληροφορίες (Panda & Kaur, 2023). Μετά την ολοκλήρωση της ανάλυσης, τα αποτελέσματα του κειμένου εμφανίζονται σε γραφήματα (όπως πχ το διάγραμμα πίτας, τα γραμμικά γραφήματα κ.α.) (Panda & Kaur, 2023).

2.5.2 Κύρια οφέλη της Ανάλυσης Συναισθήματος

Τα διάφορα συστήματα Ανάλυσης Συναισθημάτων προσφέρουν μια ποικιλία από οφέλη στους οργανισμούς που τα αξιοποιούν, δημιουργώντας πολύ σημαντικές πληροφορίες μέσα από τον τεράστιο όγκο μη δομημένων κειμένων, εξοικονομώντας πολλές ώρες μη αυτόματης επεξεργασίας δεδομένων (Paschen, Kietzmann & Kietzmann, 2019). Οι επιχειρήσεις αναπτύσσουν σημαντικά τη δυνατότητα κλιμάκωσης καθώς η Ανάλυση Συναισθημάτων επιτρέπει την επεξεργασία των δεδομένων σε κλίμακα με οικονομικά αποδοτικό τρόπο και κάτω από υψηλούς δείκτες ποιότητας (Paschen, Kietzmann & Kietzmann, 2019). Σημαντικό πλεονέκτημα είναι το γεγονός ότι η ανάλυση πραγματοποιείται σε πραγματικό χρόνο και μπορεί να αξιοποιηθεί η εφαρμογή για τον εντοπισμό πολύ σημαντικών πληροφοριών οι οποίες επιτρέπουν την ευαισθητοποίηση της κατάστασης κατά τη διάρκεια καθορισμένων σεναρίων (Paschen, Kietzmann & Kietzmann, 2019). Για παράδειγμα ένας θυμωμένος πελάτης μπορεί να δημιουργήσει προβλήματα για την επιχείρηση, έτσι ένα σύστημα Ανάλυσης Συναισθήματος μπορεί να βοηθήσει την επιχείρηση να τον εντοπίσει και να αναλάβει άμεσα κάποια δράση.

Παράλληλα, μέσα από τη χρήση ενός συστήματος Ανάλυσης Συναισθημάτων, οι οργανισμοί μπορούν να έχουν σαφή κριτήρια για όλα τα δεδομένα τους, ενισχύοντας σημαντικά την αντικειμενικότητα και ρεαλιστικότητα τους (Šandor & Bagić Babac, 2023). Αυτό βοηθάει επίσης σημαντικά στη μείωση των διαφόρων σφαλμάτων καθώς και στη βελτίωση του βαθμού συνέπειας των δεδομένων (Šandor & Bagić Babac, 2023). Παράλληλα, η ανάλυση συναισθήματος είναι μια πολύ χρήσιμη διαδικασία για την παρακολούθηση των Μέσων Κοινωνικής Δικτύωσης, δίνοντας τη δυνατότητα στους οργανισμούς να έχουν προτεραιότητα δράσης (Šandor & Bagić Babac, 2023). Μέσα από την Ανάλυση Συναισθήματος γίνεται ιεράρχηση των διαφόρων αναφορών που έχουν θετικά ή αρνητικά σχόλια, προσφέροντας σημαντικές πληροφορίες για το τμήμα Μάρκετινγκ του οργανισμού σχετικά με τα προϊόντα, τις υπηρεσίες και τις καμπάνιες τους (Šandor & Bagić Babac, 2023).

Σε επίπεδο παρακολούθησης της επωνυμίας, η Ανάλυση Συναισθημάτων είναι ιδιαίτερα χρήσιμη καθώς βοηθάει στην κατανόηση της εξέλιξης της φήμης με την πάροδο του χρόνου (Saurwein, Brantner & Möck, 2023). Η επιχείρηση μπορεί επίσης να βελτιώσει σημαντικά τη διαδικασία εξυπηρέτησης πελατών καθώς μπορεί να παρακολουθήσει καλύτερα τα βασικά μηνύματα που προέρχονται από τις απόψεις και τις σκέψεις των πελατών (Saurwein, Brantner & Möck, 2023). Με τον τρόπο αυτό το τμήμα εξυπηρέτησης πελατών έχει καλύτερη επίγνωση σχετικά με διάφορα θέματα ή προβλήματα, μπορεί να κατανοήσει καλύτερα τις ανάγκες και επιθυμίες των πελατών και να έχουν μια ευρύτερη εικόνα σχετικά με τα προβλήματα που καλείται να αντιμετωπίσει ο οργανισμός (Saurwein, Brantner & Möck, 2023). Η επιχείρηση μέσα από την Ανάλυση Συναισθημάτων έχει τη δυνατότητα να ενεργήσει άμεσα και γρήγορα σε δυσμενείς παρατηρήσεις πελατών, μειώνοντας σημαντικά τον χρόνο απόκρισης για διάφορα ζητήματα/προβλήματα που καλείται να αντιμετωπίσει η εταιρεία (Saurwein, Brantner & Möck, 2023).

Σημαντικό όφελος αποτελεί επίσης το γεγονός ότι μέσα από την Ανάλυση Συναισθημάτων, η επιχείρηση έχει τη δυνατότητα να αναπτύξει ποιοτικότερα προϊόντα, καθώς έχει καλύτερη εικόνα σχετικά με τις τρέχουσες τάσεις και τις προτιμήσεις των πελατών (Ta-Johnson, Suss & Lande, 2023). Οι διάφορες απαντήσεις των πελατών μπορούν να αξιοποιηθούν σαν μια κατευθυντήρια γραμμή προκειμένου να βελτιωθεί η ποιότητα των προσφερόμενων προϊόντων και υπηρεσιών, να βελτιωθεί ο τρόπος παρουσίασης τους και γενικότερα να ενισχυθούν διάφορα στοιχεία που συντελούν στην ανάπτυξη μιας συγκεκριμένης εικόνας σχετικά με την ποιότητα (Ta-Johnson, Suss & Lande, 2023). Ένα ακόμη ιδιαίτερα σημαντικό πλεονέκτημα είναι η δυνατότητα του οργανισμού να βελτιώσει τις αντιλήψεις στα διάφορα Μέσα Κοινωνικής Δικτύωσης, μέσα από τη χρήση συστημάτων Ανάλυσης Συναισθημάτων (Ta-Johnson, Suss & Lande, 2023). Ο οργανισμός έχει τη δυνατότητα καλύτερης παρακολούθησης των Μέσων Ενημέρωσης και των ενεργειών των δημοσιογράφων, αρθρογράφων, αναλυτών αγοράς, συγγραφέων κτλ, προκειμένου να προληφθούν φαινόμενα παρερμηνείας σε επικοινωνιακό επίπεδο και να ενισχυθούν τα φαινόμενα sharing (Ta-Johnson, Suss & Lande, 2023). Η επιχείρηση δηλαδή ετοιμάζεται το κατάλληλο υλικό έτσι ακριβώς όπως τα θέλουν τα Μέσα Ενημέρωσης προκειμένου αυτό στη συνέχεια να πλασαριστεί στους καταναλωτές (Taskin & Al, 2019).

Έρευνες έχουν δείξει επίσης ότι η Ανάλυση Συναισθημάτων βοηθάει σημαντικά στην αύξηση των εσόδων από τις πωλήσεις (Taskin & Al, 2019). Ειδικότερα αποτυπώνει τις διάφορες εντυπώσεις και τις διαθέσεις των πελατών, οδηγώντας στην βελτίωση των κερδών (Taskin & Al, 2019). Καθώς ο οργανισμός έχει τη δυνατότητα να εντοπίσει αρνητικά βασικά μηνύματα, στη συνέχεια μπορεί να πραγματοποιήσει τις κατάλληλες κινήσεις προκειμένου να πετύχει την υψηλότερη δυνατή χρηματική απόδοση (Zhang & Mu, 2022). Παράλληλα, οι πελάτες νιώθουν ότι ακούγονται και ότι η επιχείρηση φροντίζει για τις ανάγκες, τις απαιτήσεις, τις επιθυμίες τους, βελτιώνοντας με τον τρόπο αυτό την πιστότητα, την αντίληψη τους για την μάρκα και εν τέλει τα έσοδα από τις πωλήσεις (Zhang & Mu, 2022). Τέλος, ως εφαρμογή βοηθάει σημαντικά στη βελτίωση της διαχείρισης των κρίσεων, δίνοντας τη δυνατότητα στον οργανισμό να αναλάβει έγκαιρες προληπτικές ενέργειες για την εξάλειψη μιας κρίσης επικοινωνίας στο διαδίκτυο, πριν αυτή εξαπλωθεί σε αυτό σε λίγα δευτερόλεπτα οδηγώντας σε σημαντικές αρνητικές επιπτώσεις για την επωνυμία (Zhang & Mu, 2022).

2.5.3 Εφαρμογή Ανάλυσης Συναισθήματος στο Twitter

Ο τρόπος που έχει δομηθεί και αναπτυχθεί το διαδίκτυο έχει αλλάξει σημαντικά λόγω της εμφάνισης των Μέσων Κοινωνικής Δικτύωσης. Τα παραδοσιακά μέσα ενημέρωσης έχουν σημειώσει σημαντικά μείωση της χρήσης τους, καθώς τα Μέσα Κοινωνικής Δικτύωσης συμβάλλουν σημαντικά στην αύξηση της διαδραστικότητας καθώς και στην άμεση αλληλεπίδραση μεταξύ των χρηστών (Abram, Mancini & Parker, 2020). Το γεγονός ότι ο χρήστης έχει σημαντική ευκολία στην απόκτηση και ανάρτηση οποιασδήποτε μορφής πληροφορίας (κείμενο, πολυμέσα, σύνδεσμοι), δημιούργησε την ανάγκη αξιολόγησης και αξιοποίησης της πληροφορίας από διάφορους κλάδους (Abram, Mancini & Parker, 2020). Οι διάφορες ιστοσελίδες Κοινωνικής Δικτύωσης αποτελούν εικονικές κοινότητες στις οποίες υπάρχουν εγγεγραμμένοι χρήστες που έχουν δημιουργήσει εικονικό προφίλ και έχουν αναπτύξει ένα δίκτυο επαφών (Abram, Mancini & Parker, 2020).

Παράλληλα οι χρήστες αυτοί έχουν στη διάθεση τους διάφορα εργαλεία προκειμένου να μπορούν να επικαιροποιούν το προφίλ τους, να κάνουν διάφορες αναρτήσεις και αν σχολιάζουν σε αναρτήσεις άλλων χρηστών (Alaei, Becken & Stantic, 2019). Φυσικά θα πρέπει να σημειωθεί ότι η δημοφιλία και η χρήση της κάθε πλατφόρμας κοινωνικής δικτύωσης εξαρτάται και επηρεάζεται από τους χρήστες και την ποσότητα χρηστών

(Alaei, Becken & Stantic, 2019). Ένα Μέσο Κοινωνικής Δικτύωσης μπορεί να γίνει αντιληπτό ως μια δομή κόμβων, με τους κόμβους να συνδέονται μεταξύ τους με διαφορετικούς τρόπους αλληλεξάρτησης (Alaei, Becken & Stantic, 2019). Τα μέσα κοινωνικής δικτύωσης ορίζονται σαν ένα σύνολο από διαδικτυακές εφαρμογές που βασίζονται στα ιδεολογικά και τεχνολογικά θεμέλια του Web 2.0 και επιτρέπουν τη δημιουργία και την ανταλλαγή περιεχομένου από τους χρήστες (Bilbao-Jayo & Almeida, 2018). Θα πρέπει να πληρούνται πέντε βασικές συνιστώσες, με τη συμμετοχή (participation) να είναι μια από αυτές. Θα πρέπει λοιπόν να δίνεται η δυνατότητα στα διάφορα εμπλεκόμενα μέρη να μπορούν να αλληλεπιδράσουν μεταξύ τους μέσω της ατομικής τους συμβολής σχετικά με την κατάρτιση του περιεχομένου των αναρτήσεων τους (Bilbao-Jayo & Almeida, 2018). Μια ακόμη σημαντική συνιστώσα είναι αυτή της διαφάνειας (openness). Θα πρέπει οι μηχανισμοί να παρέχουν τις κατάλληλες πληροφορίες για τη σύνθεση και το διαμοιρασμό του περιεχομένου, δίνοντας τη δυνατότητα στους χρήστες να παρέχουν ανατροφοδότηση, πέραν της δυνατότητας συμμετοχής τους, κάτω από ελάχιστους περιορισμούς (Bilbao-Jayo & Almeida, 2018).

Τρίτη βασική συνιστώσα είναι η συνομιλία (conversation), η οποία σε αντίθεση με τα παραδοσιακά μέσα τα οποία απευθύνονται σε παθητικούς χρήστης, στα Μέσα Κοινωνικής Δικτύωσης υπάρχει η επικοινωνία διπλής κατεύθυνσης (Doğan, Balcioglu & Elçi, 2024). Η κοινότητα (community) αποτελεί ένα ακόμη βασικό στοιχείο τους καθώς δίνεται η ευκαιρία δημιουργίας κοινοτήτων χρηστών με βάση κοινά ενδιαφέροντα, κοινές αξίες, θεωρήσεις, πεποιθήσεις, αντιλήψεις κ.α (Doğan, Balcioglu & Elçi, 2024). Τέλος το τελευταίο βασικό στοιχείο είναι αυτό της συνεκτικότητας (connectedness), η οποία θεωρείται και ως μέτρο της εγγύτητας και ταύτισης ανάμεσα στους χρήστες του Μέσου Κοινωνικής Δικτύωσης (Doğan, Balcioglu & Elçi, 2024). Η συνεκτικότητα μπορεί να ενισχυθεί σημαντικά κατά τη χρήση συνδέσεων με άλλες ιστοσελίδες, πόρους ή χρήστες βοηθώντας στην επίτευξη καλύτερης ανθρώπινης αλληλεπίδρασης (Doğan, Balcioglu & Elçi, 2024). Αυτό που τα καθιστά όμως τόσο σημαντικό ερευνητικό πεδίο είναι η διάδοσή τους σε παγκόσμια κλίμακα, το γεγονός ότι όλα συμβαίνουν σε πραγματικό χρόνο και το χαμηλό κόστος πρόσβασης στα δεδομένα τους (Fang & Wang, 2022).

Στα διάφορα Κοινωνικά Δίκτυα, η εξόρυξη δεδομένων από αυτά αποτελεί έναν πολύ σημαντικό τομέα καθώς δίνεται η δυνατότητα καλύτερης μελέτης των ανθρωπίνων σχέσεων και των συναισθημάτων σε ένα μεγάλο δείγμα που σε αντίθετες περιπτώσεις θα ήταν πολύ δύσκολα προσβάσιμο (Fang & Wang, 2022). Η εξόρυξη η οποία λαμβάνει

χώρα στα Κοινωνικά Δίκτυα θεωρείται ως μια σειρά κατάλληλων ενεργειών οι οποίες μπορούν να οδηγήσουν στην εξαγωγή προτύπων από δεδομένα αξιοποιώντας τα κατάλληλα εργαλεία ανάλυσης και μοντελοποίησης προκειμένου να διερευνηθούν δεδομένα τα οποία είναι μεγάλα σε όγκο (Fang & Wang, 2022). Καθώς τα Μέσα Κοινωνικής Δικτύωσης επηρεάζουν τις τάσεις και αντιλήψεις της αγοράς και γενικότερα της καταναλωτική συμπεριφορά, διάφορα επιχειρήσεις αναλόγως και του κλάδου δραστηριοποίησης τους, έχουν επενδύσει σημαντικούς πόρους για την εφαρμογή συστημάτων Ανάλυσης δεδομένων, Συναισθημάτων και άλλων στοιχείων που είναι χρήσιμα για το μάρκετινγκ του οργανισμού (Hamaz & Benchikha, 2017). Η ανάλυση αυτών των δεδομένων παρέχει τη δυνατότητα να βελτιώσουν τις επιδόσεις τους σε διαφορετικούς τομείς και φυσικά, να διατηρούνται ανταγωνιστικές.

Από την άλλη πλευρά όμως, η προσβασιμότητα των επιχειρήσεων στις κατάλληλες πληροφορίες είναι μια χρονοβόρα διαδικασία για την οποία πρέπει να λάβουν υπόψη τους μια σειρά από παραμέτρους (Hamaz & Benchikha, 2017). Ειδικότερα, θα πρέπει να γίνει η κατάλληλη επιλογή του μοντέλου που θα αξιοποιηθεί προκειμένου να είναι σε θέση να επεξεργάζεται έναν μεγάλο όγκο δεδομένων στο αντίστοιχο χρονικό διάστημα (Hamaz & Benchikha, 2017). Θα πρέπει επίσης να υπάρχει κάποια μορφή προεπεξεργασίας των δεδομένων ή ανοχή του μοντέλου στο θόρυβο που μπορεί να υπάρχει στα δεδομένα (Hartmann & Netzer, 2023). Στα δεδομένα, κυρίως στη μορφή κειμένου, συναντώνται συχνά σημαντικά προβλήματα που σχετίζονται με το λεκτικό περιεχόμενο καθώς και το συντακτικό μέρος του κειμένου (Hartmann & Netzer, 2023). Το γεγονός αυτό μπορεί να επηρεάσει την ανάλυση και σε κάποιες περιπτώσεις το τελικό αποτέλεσμα.

Ειδικότερα ένας πολύ σημαντικός τομέας ο οποίος έχει ραγδαία ανάπτυξη και εξέλιξη στα Μέσα Κοινωνικής Δικτύωσης, είναι αυτός της Ανάλυσης Συναισθήματος. Μέσα από τα Μέσα Κοινωνικής Δικτύωσης, ο χρήστης μπορεί να ανταλλάσει απόψεις, να δημοσιοποιεί απόψεις άλλων, να δημιουργεί περιεχόμενο, να παραθέτει τις απόψεις του μέσω αυτοματοποιημένους τρόπους για κάποιο περιεχόμενο (π.χ. like) ή να αξιοποιεί διάφορα εικονίδια τα οποία δηλώνουν την κατάσταση του ή την άποψη για κάποιο θέμα (Hartmann & Netzer, 2023). Τα Μέσα Κοινωνικής Δικτύωσης είναι ιδιαίτερα πλούσια σε επίπεδο συναισθημάτων και οι διάφορες πληροφορίες που μπορούν να εξαχθούν αποτελούν σημαντική πηγή δεδομένων και έχουν μεγάλη αξία για κυβερνήσεις, οργανισμούς, ερευνητικά κέντρα και άλλα (Hossain & Rahman, 2023). Μια από τις πιο δημοφιλείς πλατφόρμες Κοινωνικής Δικτύωσης είναι το Twitter, το οποίο

δημιουργήθηκε το 2006 και επιτρέπει την επικοινωνία ανάμεσα στους χρήστες μέσα από μηνύματα, με μέγιστο αριθμό χαρακτήρα τους 140.

Η επικοινωνία αυτή ονομάζεται "tweet", το οποίο αποτελεί κείμενο το οποίο εκτός από λέξεις μπορεί να περιέχει ειδικά σύμβολα καθώς και συνδέσμους. Τα δυο βασικά σύμβολα που μπορούν να χρησιμοποιηθούν στο twitter είναι το "@"και το "#". Το πρώτο σύμβολο συνοδεύεται στη συνέχεια με κάποιο όνομα χρήστη (username) το οποίο χρησιμοποιεί ο συγγραφέας του tweet προκειμένου να αναφερθεί σε κάποιον άλλο χρήστη του Twitter. Το δεύτερο σύμβολο καλείται ως hashtag και εφαρμόζεται ως ένα κλειδί το οποίο είναι σχετικό με το θέμα του tweet στο οποίο αναφέρεται ή έχει ως στόχο να αναφερθεί ο συντάκτης του κειμένου (Hossain & Rahman, 2023). Ειδικότερα, μέσα από τη χρήση των hashtag, γίνεται ομαδοποίηση των σχολίων σε διάφορα ειδικά θέματα. Τέλος, υπάρχει επίσης και το "RT" το οποίο προστίθεται στις περιπτώσεις αναμετάδοσης ενός μηνύματος από το χρήστη (Hossain & Rahman, 2023).

Το Twitter ως πλατφόρμα θεωρείται αρκετά εύχρηστη συγκριτικά με τα υπόλοιπα Μέσα Κοινωνικής Δικτύωσης αναφορικά με την Ανάλυση Συναισθήματος. Τα μηνύματα είναι βέβαια μικρά σε μέγεθος αφού περιέχουν 140 χαρακτήρες, όμως τα περισσότερα δεδομένα είναι ελευθέρα σε τρίτους (μέσω του Streaming API) και παραχωρούνται από το τμήμα Twitter Developer. Παράλληλα υπάρχει και η χρονοσφραγίδα προκειμένου να γίνει σωστή ταξινόμηση μιας συζήτησης η οποία εξελίσσεται στην πλατφόρμα (Keramatfar & Amirkhani, 2019). Παρόλο αυτά υπάρχουν κάποια στοιχεία τα οποία δυσκολεύουν σημαντικά στην Ανάλυση Συναισθήματος, με ένα από αυτά να είναι η έκταση του κειμένου (Keramatfar & Amirkhani, 2019). Τα σχόλια τα οποία έχουν γραφτεί είναι μικρά σε μέγεθος και πολλές φορές υπάρχει μεγάλη δυσκολία εντοπισμού του εννοιολογικά σημαντικού στοιχείου της πρότασης, ακόμη και από τον ίδιο τον άνθρωπο. Αυτό έχει σαν αποτέλεσμα να μην γίνει ξεκάθαρη η πολικότητα του κειμένου (Kwartler, 2021). Παράλληλα η ύπαρξη ορθογραφικών και συντακτικών λαθών (πχ λάθη σε τονισμούς, λάθη στη σειρά σύνταξης των προτάσεων κα) προκαλούν θόρυβο στα κείμενα παρασύροντας τον υπολογιστή σε λανθασμένες επιλογές ταξινόμησης των λέξεων (Kwartler, 2021). Τέλος, η ύπαρξη πολυγλωσσίας, δηλαδή της χρήσης άλλων λέξεων και εκφράσεων παράλληλα με την κυρίαρχη γλώσσα, επηρεάζει σημαντικά το βαθμό αποτελεσματικότητας των εφαρμογών και συστημάτων Μηχανικής Μάθησης και Τεχνητής Νοημοσύνης (Lin & Yu, 2023).

Κεφάλαιο 3. Εφαρμογή Έρευνας (Συλλογή, Μεθοδολογία και Ανάλυση Δεδομένων)

Στον τομέα της επεξεργασίας φυσικής γλώσσας (NLP), η κατανόηση του συναισθήματος μέσω της ανάλυσης δεδομένων μέσων κοινωνικής δικτύωσης έχει αναδειχθεί ως κεντρικός τομέας έρευνας και εφαρμογής. Αυτό το κεφάλαιο εμβαθύνει στη μεθοδολογία, τη συλλογή δεδομένων και τις τεχνικές ανάλυσης που χρησιμοποιούνται για την επιδίωξη αποκάλυψης γνώσεων σχετικά με την ανάλυση συναισθημάτων (sentiment analysis) Twitter, εστιάζοντας ιδιαίτερα στον εντοπισμό της ρητορικής μίσους (hate speech).

Οι πλατφόρμες μέσων κοινωνικής δικτύωσης, ιδιαίτερα το Twitter, έχουν γίνει αναπόσπαστα στοιχεία της σύγχρονης επικοινωνίας, προσφέροντας μια πλούσια πηγή κειμενικών δεδομένων που αντικατοπτρίζουν διαφορετικές απόψεις, συναισθήματα και αισθήματα. Η κατανόηση του συναισθήματος που ενσωματώνεται σε αυτό το τεράστιο σώμα περιεχομένου (corpus) που δημιουργείται από χρήστες είναι ζωτικής σημασίας για διάφορους σκοπούς, που κυμαίνονται από την έρευνα αγοράς και την ανάλυση συναισθήματος επωνυμίας έως τον κοινωνικοπολιτικό λόγο και την παρακολούθηση της κοινής γνώμης.

Μεταξύ των μυριάδων συναισθημάτων που εκφράζονται στο Twitter, η ρητορική μίσους ξεχωρίζει ως σημαντικό και διάχυτο ζήτημα, που συχνά συμβάλλει στην τοξικότητα στο διαδίκτυο, στον κυβερνοεκφοβισμό και στον κοινωνικό διχασμό. Ο εντοπισμός και ο μετριασμός της ρητορικής μίσους όχι μόνο προάγει ένα ασφαλέστερο διαδικτυακό περιβάλλον, αλλά υποστηρίζει επίσης τις αρχές της συμμετοχής, της διαφορετικότητας και του σεβασμού.

Αυτό το κεφάλαιο διερευνά τη μεθοδολογία που χρησιμοποιείται για την ανάλυση του συναισθήματος στο Twitter, με ιδιαίτερη έμφαση στον εντοπισμό της ρητορικής μίσους. Για λόγους απλότητας, λέμε ότι ένα tweet περιέχει ρητορική μίσους εάν έχει ρατσιστικό ή σεξιστικό αίσθημα που σχετίζεται με αυτό. Έτσι, το ζητούμενο είναι να ταξινομηθούν τα ρατσιστικά ή σεξιστικά tweets από άλλα tweets. Περιγράφει τα βήματα που έγιναν για τη συλλογή σχετικών δεδομένων, την προεπεξεργασία του περιεχομένου του κειμένου και την εξαγωγή σημαντικών χαρακτηριστικών για επακόλουθη ανάλυση. Επιπλέον, διευκρινίζει τις διάφορες αναλυτικές τεχνικές που χρησιμοποιούνται για τη διάκριση

μοτίβων, τάσεων και γνώσεων από τα δεδομένα που συλλέγονται, ρίχνοντας φως στην επικράτηση, τη φύση και τη δυναμική της ρητορικής μίσους στο Twitter.

Η μεθοδολογία που υιοθετήθηκε σε αυτή τη μελέτη περιλαμβάνει μια πολύπλευρη προσέγγιση, ενσωματώνοντας τεχνικές από το NLP, τη μηχανική μάθηση και την ανάλυση συναισθημάτων. Αξιοποιώντας αλγόριθμους και μεθοδολογίες αιχμής, αυτό το κεφάλαιο διευκρινίζει τη διαδικασία βήμα προς βήμα που εμπλέκεται στην απόκτηση δεδομένων, την προεπεξεργασία, την εξαγωγή χαρακτηριστικών και την ανάπτυξη μοντέλων.

Συνοψίζοντας, το παρόν κεφάλαιο διευκρινίζει το μεθοδολογικό πλαίσιο που διέπει την επεξεργασία φυσικής γλώσσας και την ανάλυση του συναισθήματος στο Twitter, με ιδιαίτερη έμφαση στην ανίχνευση της ρητορικής μίσους. Περιγράφοντας τη μεθοδολογία, τις στρατηγικές συλλογής δεδομένων και τις αναλυτικές τεχνικές που χρησιμοποιήθηκαν, θέτει τα θεμέλια για μια ολοκληρωμένη κατανόηση σχετικά με τη μελέτη της δυναμικής του διαδικτυακού συναισθήματος και της ανίχνευσης επιβλαβούς λόγου σε ψηφιακούς χώρους.

3.1 Συλλογή Δεδομένων

Αρχικό στοιχείο αυτής της προσπάθειας είναι η απόκτηση ενός ολοκληρωμένου συνόλου δεδομένων που περιλαμβάνει αναρτήσεις στο Twitter (tweets) με σχολιασμούς με ετικέτες (labels) που υποδεικνύουν την παρουσία ή την απουσία ρητορικής μίσους. Η διαδικασία συλλογής δεδομένων περιλαμβάνει την αναζήτηση και εύρεση τεχνικών εξόρυξης επιμελημένων συνόλων δεδομένων για τη συλλογή ενός διαφορετικού και αντιπροσωπευτικού δείγματος συνομιλιών στο Twitter που καλύπτουν διάφορα θέματα, τομείς και δημογραφικά στοιχεία.

3.1.1 Επικοινωνία με το Twitter - Twitter API

Κάθε μέρα, εκατομμύρια μηνύματα γνωστά ως tweets δημιουργούνται στο Twitter, συνολικά πάνω από 500 εκατομμύρια. Αυτά τα tweets χρησιμεύουν ως πολύτιμος πόρος για ερευνητές και άλλα άτομα που ενδιαφέρονται να έχουν πρόσβαση σε αυτά μέσω του Twitter API, ενός εξειδικευμένου εργαλείου που παρέχεται από την πλατφόρμα. Το Twitter είναι γνωστό για την αφθονία των πληροφοριών του, καθιστώντας το ιδανική πλατφόρμα για τη συλλογή πληροφοριών και την ανάλυση των συναισθημάτων που εκφράζονται στα tweets.

Το Twitter API (συντομογραφία του Application Programming Interface), δηλαδή η Διεπαφή Προγραμματισμού Εφαρμογών, διευκολύνει την επικοινωνία μεταξύ διαφορετικών εφαρμογών. Λειτουργώντας ως ενδιάμεσος, επεξεργάζεται αιτήματα και επιστρέφει απαντήσεις, επιτρέποντας στους χρήστες να αλληλεπιδρούν με τα δεδομένα του Twitter. Αυτή η λειτουργία περιλαμβάνει διάφορες ενέργειες, όπως δημιουργία tweet, πρόσβαση σε προφίλ χρηστών και ανάκτηση μεγάλου όγκου tweet με βάση συγκεκριμένες λέξεις-κλειδιά ή θέματα ενδιαφέροντος, με τα δεδομένα που συλλέγονται να αποθηκεύονται συνήθως σε μορφή JSON.

Για να αξιοποιήσετε τις δυνατότητες του Twitter API, είναι απαραίτητο να έχουμε έναν προσωπικό λογαριασμό Twitter, ο οποίος όχι μόνο παρέχει πρόσβαση στις λειτουργίες API που δεν είναι προσβάσιμες σε μη πιστοποιημένους χρήστες, αλλά καθορίζει επίσης την ταυτότητα μας στην κοινότητα και ως προγραμματιστές αλλά και ως χρήστες. Έτσι, η δημιουργία ενός λογαριασμού Twitter είναι απαραίτητη για την εξερεύνηση του πλήρους φάσματος των λειτουργιών του API.

Το API του Twitter χρησιμεύει ως πολύτιμος πόρος για ερευνητές, προγραμματιστές και επιχειρήσεις, προσφέροντας πρόσβαση σε δεδομένα που παρέχουν πληροφορίες για τη στάση των ατόμων απέναντι σε διάφορα θέματα. Αυτή η προσβασιμότητα σε δεδομένα tweet διευκολύνει την ανάλυση και την κατανόηση του δημόσιου λόγου στην πλατφόρμα. Σε αυτήν την ενότητα, θα μάθουμε πώς να χρησιμοποιούμε αυτό το εργαλείο για να λαμβάνουμε πληροφορίες από το Twitter και να αναλύουμε τα συναισθήματα των ανθρώπων στα tweet τους.

Για να δημιουργηθεί ενός λογαριασμού (προφίλ) χρειάζονται:

1. Όνομα χρήστη (username)
2. Κωδικός πρόσβασης (password)
3. Έγκυρη διεύθυνση ηλεκτρονικού ταχυδρομείου (email)

Καθώς υπάρχουν και ορισμένοι περιορισμοί:

- Τα usernames πρέπει να είναι λιγότερα από 15 χαρακτήρες, οι οποίοι μπορούν να είναι γράμματα και αριθμοί αλλά δεν μπορούν να περιέχουν το "admin" ή "X", προκειμένου να αποφευχθεί η σύγχυση της επωνυμίας.
- Μία διεύθυνση email δεν μπορεί να χρησιμοποιηθεί για την δημιουργία περισσότερων από έναν λογαριασμών.
- Η αλλαγή ονόματος χρήστη μπορεί να γίνει εύκολα από τις ρυθμίσεις του λογαριασμού οποιαδήποτε στιγμή, καθώς το νέο όνομα χρήστη δεν χρησιμοποιείται ήδη από κάποιον άλλον λογαριασμό / χρήστη.

3.1.1.1 Τύποι του Twitter API

Το Twitter API v2 αποτελεί το κύριο API του Twitter, όμως η πλατφόρμα υποστηρίζει προς το παρόν και παλαιότερες εκδόσεις (v1.1, Gnip 2.0). Το Twitter API v2 έχει διάφορα επίπεδα πρόσβασης για να επιτρέπει μεγαλύτερη χρήση της πλατφόρμας. Υπάρχουν τέσσερις κατηγορίες εξουσιοδότησης για πρόσβαση στο Twitter API, με τα ακόλουθα χαρακτηριστικά:

1. **Free**: Για περιπτώσεις χρήσης μόνο για εγγραφή και δοκιμές του API του Twitter.
    - Περιορισμένη πρόσβαση σε v2 tweet ανάρτησης (posting) και μεταφόρτωσης πολυμέσων (media upload) endpoints.
      - 1.500 tweets ανά μήνα - όριο δημοσίευσης σε επίπεδο εφαρμογής (app level)
    - Δημιουργία ενός Project
    - Δημιουργία μιας εφαρμογή ανά έργο (App / Project)
    - 1 περιβάλλον (Environment) (Ανάπτυξη/ Παραγωγή/ Σταδιοποίηση) (Development/ Production/ Staging)
    - Σύνδεση με το Twitter
    - Πρόσβαση στο API διαφημίσεων (Twitter Ads API)
    - Κόστος: Δωρεάν
2. **Basic**: Για ασχολίες τύπου “χόμπι” ή σπουδαστές.

- Περιορισμένη πρόσβαση στη "σουίτα υπηρεσιών" v2 endpoints, αλλά με επιπλέον πρόσβαση σε endpoints και δεδομένα καθώς επίσης και σε διάφορα περιβάλλοντα App
- 3.000 tweets το μήνα - όριο ανάρτησης σε επίπεδο χρήστη (user level)
- 50.000 tweets το μήνα - όριο ανάρτησης σε επίπεδο εφαρμογής (app level)
- 10.000/μήνα ανώτατο όριο ανάγνωσης Tweets
- Δημιουργία ενός Project
- Δημιουργία δύο εφαρμογών ανά έργο (App / Project) με μοναδικό περιβάλλον (Environment) (Ανάπτυξη/ Παραγωγή/ Σταδιοποίηση) (Development/ Production/ Staging)
- Σύνδεση με το Twitter
- Πρόσβαση στο API διαφημίσεων (Twitter Ads API)
- Κόστος: $100 ανά μήνα

3. **Pro**: Για νεοσύστατες επιχειρήσεις (startups) που κλιμακώνουν την επιχείρησή τους.
    - Περιορισμένη πρόσβαση στη "σουίτα υπηρεσιών" v2 endpoints, συμπεριλαμβανομένης της αναζήτησης και της φιλτραρισμένης ροής (search and filtered stream)
    - 1.000.000 Tweets ανά μήνα - GET σε επίπεδο εφαρμογής (app level)
    - 300.000 Tweets ανά μήνα - όριο αποστολής σε επίπεδο εφαρμογής (posting limit)
    - Δημιουργία ενός Project
    - Δημιουργία τριών εφαρμογών ανά έργο (App / Project) με μοναδικό περιβάλλον (Environment) (Ανάπτυξη/ Παραγωγή/ Σταδιοποίηση) (Development/ Production/ Staging)
    - Σύνδεση με το Twitter
    - Πρόσβαση στο API διαφημίσεων (Twitter Ads API)
    - Κόστος: $5,000 ανά μήνα
4. **Enterprise**: Για επιχειρήσεις και εμπορικά έργα.
    - Πρόσβαση εμπορικού επιπέδου που ανταποκρίνεται στις δικές μας ανάγκες και τις ανάγκες των πελατών μας
    - Υπηρεσίες διαχείρισης από ειδική ομάδα λογαριασμού
    - Πλήρεις ροές: επανάληψη, μετρήσεις αφοσίωσης και άλλα χαρακτηριστικά

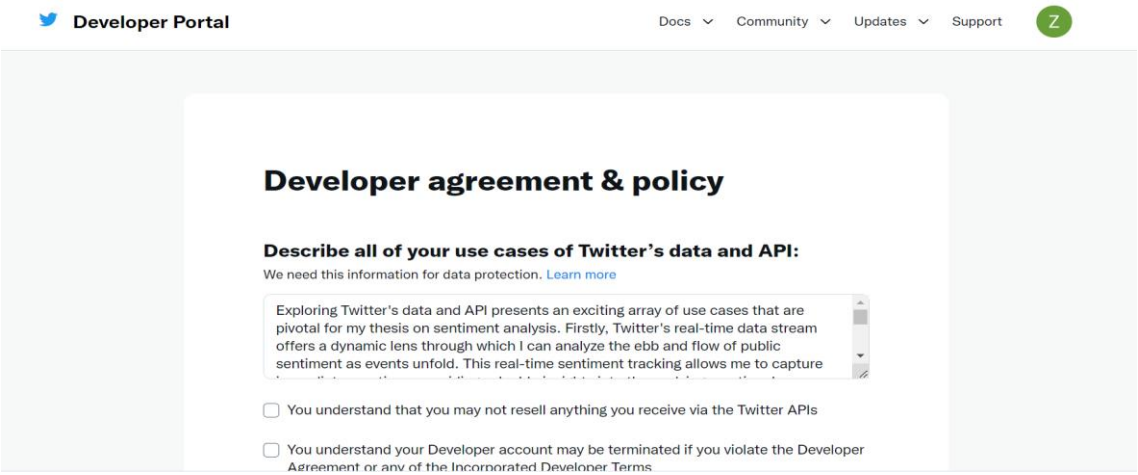
- Κόστος: Ξεκινάει στα $42,000 ανά μήνα

	Free	Basic	Pro	Enterprise
Getting access	Get Started	Get Started	Get Started	Get Started
Price	Free	$100/month	$5000/month	
Access to Twitter API v2	✔ (Only Tweet creation)	✔	✔	
Access to standard v1.1	✔ (Only Media Upload, Help, Rate Limit, and Login with Twitter)	✔ (Only Media Upload, Help, Rate Limit, and Login with Twitter)	✔ (Only Media Upload, Help, Rate Limit, and Login with Twitter)	
Project limits	1 Project	1 Project	1 Project	
App limits	1 App per Project	2 Apps per Project	3 Apps per Project	
Tweet caps - Post	1,500	3,000	300,000	
Tweet caps - Pull	✘	10,000	1,000,000	
Filteres stream API	✘	✘	✔	
Access to full-archive search	✘	✘	✔	
Access to Ads API	✔	✔	✔	

Εικόνα 9. Είδη Twitter API (Twitter Developer Platform)

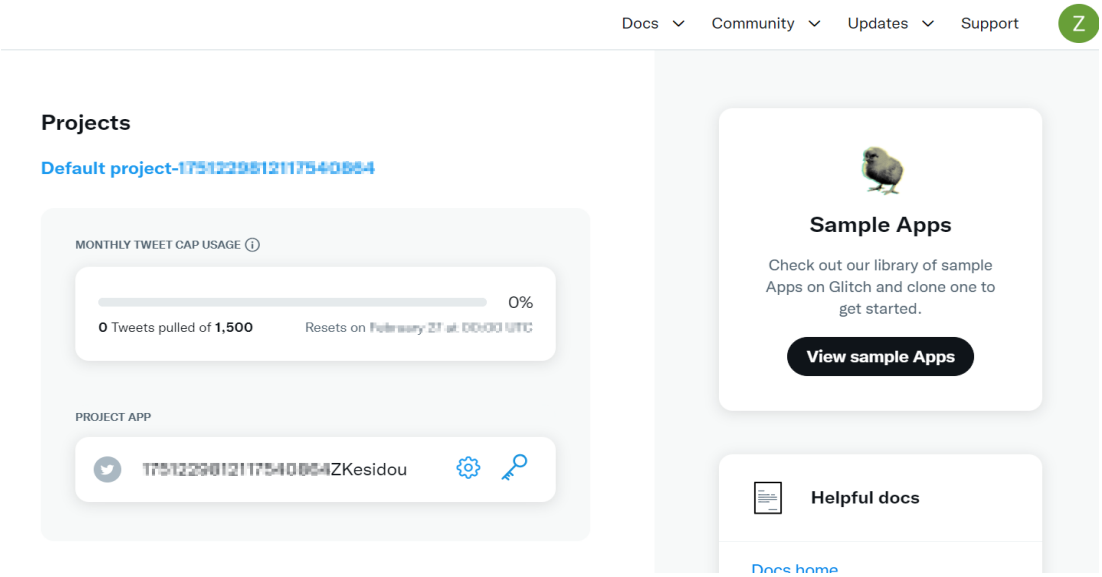
3.1.1.2 Πρόσβαση στα δεδομένα του Twitter

Για να αποκτήσουμε πρόσβαση στα API του Twitter και να καταφέρουμε να εξάγουμε δεδομένα από το Twitter μέσω προγραμματισμού, πρέπει να υποβάλετε αίτηση για πρόσβαση προγραμματιστή. Αυτό περιλαμβάνει την επίσκεψη στον ιστότοπο του Twitter Developer (developer.twitter.com) και τη σύνδεση με τον λογαριασμό Twitter που ήδη έχουμε δημιουργήσει. Οι αιτούντες πρέπει να παρέχουν απαντήσεις σε βασικά ερωτήματα σχετικά με το σκοπό πίσω από την πρόσβαση στο API του Twitter και την προβλεπόμενη χρήση των δεδομένων. Μετά την ολοκλήρωση της διαδικασίας υποβολής αίτησης, αποστέλλεται ειδοποίηση μέσω email που υποδεικνύει εάν η αίτηση έχει εγκριθεί ή απορριφθεί από την πλατφόρμα.



Εικόνα 10. Αίτημα για πρόσβαση στο Twitter API

Εάν η αίτηση γίνει δεκτή, ο χρήστης μπορεί να χρησιμοποιήσει το Twitter API για να δημιουργήσει το δικό του έργο (project) και εφαρμογή (app). Κάθε project έχει τα δικά του μοναδικά "κλειδιά" (keys) που επιτρέπουν στον χρήστη να έχει πρόσβαση στα δεδομένα.

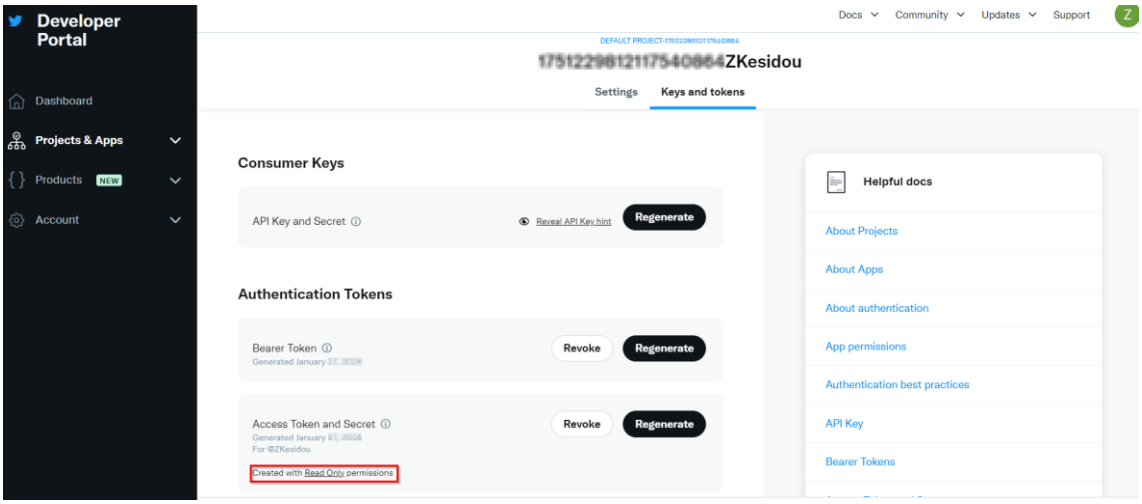


Εικόνα 11. Δημιουργία Project στο Twitter Developer Platform

Πιο συγκεκριμένα τα κλειδιά που απαιτούνται για την σύνδεση μας με την διεπαφή του Twitter είναι τα εξής:

- **API KEY** και **API SECRET**: αποτελούν το όνομα χρήστη (username) και τον κωδικό (password) του APP, ώστε το πρόγραμμα να μπορεί να έχει πρόσβαση στην ανάκτηση δεδομένων.

- **ACCESS TOKEN** και **ACCESS TOKEN SECRET**: θεωρούνται οι ειδικοί κωδικοί πρόσβασης του χρήστη για την επικύρωση και την επαλήθευση των δεδομένων πριν από τη χορήγηση πρόσβασης στα δεδομένα του Twitter για το APP.



Εικόνα 12. Twitter Tokens

Μετά την απόκτηση των απαραίτητων κωδικών, ακολούθησε μια δοκιμή του Twitter API. Κατα την διάρκεια αυτής της δοκιμής διαπιστώθηκε ότι είναι αδύνατη η απόκτηση δεδομένων (tweets) με την "Free" κατηγορία του Twitter API και είναι απαραίτητη τουλάχιστον η "Basic" κατηγορία (που είναι επί πληρωμή, όπως αναφέρθηκε στην ενότητα 3.3.2.1). Εξαιτίας αυτού του περιστατικού η δημιουργία dataset μέσω του Twitter API είναι αδύνατη στη συγκεκριμένη ερευνητική εργασία.

Εικόνα 13. Τμήμα κώδικα που απεικονίζει την αδυναμία εξαγωγής tweets

3.1.2 Kaggle

Μετά την διαπίστωση ότι δεν μπορεί να γίνει χρήση του Twitter API, έγινε έρευνα για την απόκτηση dataset (tweets) με κάποιον άλλον τρόπο. Γνωστό στον κόσμο της επιστήμης δεδομένων, το Kaggle έχει γίνει ένας τόπος συγκέντρωσης ερευνητών, επαγγελματιών και λάτρεις των δεδομένων. Από την ίδρυσή του το 2010, το Kaggle έχει προσφέρει ένα ευρύ φάσμα εργαλείων και ευκαιριών σε άτομα που ενδιαφέρονται να αναπτύξουν τις ικανότητές τους στην επιστήμη δεδομένων, να συνεργαστούν σε έργα και να ανταγωνίζονται.

Για τους data scientists, το Kaggle είναι μια εικονική "παιδική χαρά" που παρέχει μια πληθώρα συνόλων δεδομένων (datasets), διαγωνισμών, μαθημάτων (tutorials) και φόρουμ συζήτησης για την προώθηση της μάθησης και της συνεργασίας. Το Kaggle χρησιμεύει ως πλατφόρμα για τους λάτρεις της επιστήμης δεδομένων για να συμμετέχουν σε πρακτικές προκλήσεις και να κάνουν πολύτιμες συνεισφορές σε αξιόλογες πρωτοβουλίες . Όλα αυτά τα χαρακτηριστικά προσθέτουν στο Kaggle μια ολοκληρωμένη πλατφόρμα για μάθηση και συνεργασία.

Διαγωνισμοί: Ένα από τα πιο αξιοσημείωτα χαρακτηριστικά του Kaggle είναι οι διαγωνισμοί του, στους οποίους data scientists από όλο τον κόσμο ανταγωνίζονται για κορυφαίες κατατάξεις εφαρμόζοντας τεχνικές στατιστικής μοντελοποίησης και μηχανικής μάθησης (machine learning) για την επίλυση δύσκολων προκλήσεων. Μια ποικιλία βιομηχανιών καλύπτονται από διαγωνισμούς Kaggle, όπως η μηχανική όραση (computer vision / AI), τα οικονομικά και η υγειονομική περίθαλψη. Εκτός από τη μάθηση από συνομηλίκους, οι συμμετέχοντες έχουν την ευκαιρία να αναδείξουν τα ταλέντα τους και ίσως να κερδίσουν χρηματικά έπαθλα και ευκαιρίες καριέρας που παρέχονται από χορηγούς εταιρείες.

Datasets: Συλλογές datasets από ένα ευρύ φάσμα πεδίων, συμπεριλαμβανομένης της επεξεργασίας φυσικής γλώσσας (NLP), είναι διαθέσιμες στο Kaggle. Αυτά τα σύνολα δεδομένων είναι χρήσιμα εργαλεία για μάθηση, πειραματισμό και έρευνα. Τα σύνολα δεδομένων του Kaggle επιλέγονται προσεκτικά και επαληθεύονται, διασφαλίζοντας ότι οι χρήστες θα έχουν πρόσβαση σε μια ποικιλία αξιόπιστων και αξιόπιστων πηγών δεδομένων για τα έργα τους.

Kernel: Ένα άλλο σημαντικό στοιχείο του Kaggle είναι τα kernels, τα οποίοι επιτρέπουν στους χρήστες να δημιουργούν και να εκτελούν προγράμματα σε cloud περιβάλλον χωρίς

να απαιτείται τοπική εγκατάσταση ή προετοιμασία. Πολλοί συγγραφείς τονίζουν ποσο βολικά και προσαρμόσιμα είναι τα kernels καθώς διευκολύνουν την εξερεύνηση, την ανάλυση και την εμφάνιση δεδομένων. Επιπλέον, τα kernels είναι εργαλεία διδασκαλίας που επιτρέπουν στους χρήστες να αποκτήσουν αξιολόγηση και να μοιραστούν τον κώδικα και τις σκέψεις τους με την κοινότητα.

Φόρουμ συζητήσεων: Τα φόρουμ συζήτησης του Kaggle δίνουν στους χρήστες τη δυνατότητα να κάνουν ερωτήσεις, να μοιράζονται ιδέες και να έχουν συζητήσεις σχετικά με τη μηχανική μάθηση (machine learning) και την επιστήμη δεδομένων (data science). Αξιοσημείωτη είναι η συνεργατική πτυχή της κοινότητας του Kaggle, όπου τα μέλη μοιράζονται ανοιχτά πληροφορίες, συζητούν ιδέες και βοηθούν το ένα το άλλο για να ξεπεράσουν τα εμπόδια.

Συνοψίζοντας, το Kaggle είναι μια βασική πλατφόρμα στο οικοσύστημα της επιστήμης δεδομένων που παρέχει σε επαγγελματίες και ερευνητές άφθονες πληροφορίες και ευκαιρίες, που παραδόξως είναι εντελώς δωρεάν. Με τους διαγωνισμούς, τα σύνολα δεδομένων, τους πυρήνες και τους πίνακες συζητήσεων, το Kaggle δίνει τη δυνατότητα στους χρήστες να προωθήσουν τις γνώσεις τους, να συνεργαστούν σε έργα και να συνεισφέρουν σημαντικά στην κοινότητα της επιστήμης δεδομένων. Πρόσφατα άρθρα και οδηγοί έχουν δείξει πόσο σημαντικό είναι το Kaggle για να επηρεάσει το μέλλον της επιστήμης δεδομένων και να προωθήσει τη δημιουργικότητα στην ψηφιακή εποχή.

Η έναρξη ενός λογαριασμού στο Kaggle δεν είναι δύσκολη. Στον ιστότοπο του Kaggle, οι χρήστες μπορούν να εγγραφούν και να εξερευνήσουν τις δυνατότητες και τις επιλογές της πλατφόρμας. Στα πλαίσια της παρούσας εργασίας, έγινε αναζήτηση στην πλατφόρμα για την εύρεση dataset αποτελούμενο από tweets και εντοπίστηκαν αρκετά, επιλέχθηκε μια συλλογή δεδομένων αποτελούμενη από tweets και ετικέτες (labels), όπου η ετικέτα '1' υποδηλώνει ότι το tweet είναι ρατσιστικό/σεξιστικό και η ετικέτα '0' υποδηλώνει ότι το tweet δεν είναι ρατσιστικό/σεξιστικό. Στην εργασία αυτή αναλύθηκε ένα αρχείο δεδομένων CSV από το Kaggle που περιείχε 31.935 tweets.

Dataset	
id	Αναγνωριστικό/σειριακός αριθμός
label	Κλάση του tweet: 0 μη ρατσιστικό/σεξιστικό 1 ρατσιστικό/σεξιστικό

tweet	Το περιεχόμενο το tweet

3.1.3 Python

Η Python ξεχωρίζει ως μια γλώσσα προγραμματισμού που γνωρίζει ταχεία ανάπτυξη. Η δημοτικότητά της μεταξύ των προγραμματιστών πηγάζει από τη φήμη της ως γλώσσας που είναι τόσο εύκολη στην κατανόηση όσο και αποτελεσματική για σκοπούς κωδικοποίησης. Η Python μπορεί να υπερηφανεύεται για μια εκτεταμένη συλλογή βιβλιοθηκών, ιδιαίτερα στην επιστημονική πληροφορική και την επιστήμη δεδομένων, καθιστώντας την μια προτιμώμενη επιλογή για πολλές μεγάλες εταιρείες, συμπεριλαμβανομένων των Google, Yahoo, YouTube, Dropbox και NASA. Επιπλέον, η ευελιξία της Python επεκτείνεται στην υποστήριξη διαφόρων τομέων όπως η μηχανική εκμάθηση, η ανάπτυξη λογισμικού και η ανάπτυξη ιστοσελίδων. Ως γλώσσα γενικής χρήσης, αντικειμενοστραφής και υψηλού επιπέδου, η Python κατέχει σημαντική θέση στον τομέα της επιστήμης δεδομένων και χρησιμεύει ως ακρογωνιαίος λίθος για τη δημιουργία εξελιγμένων αλγορίθμων βαθιάς μάθησης (Rajaraman, 2020). Πολλές βιβλιοθήκες για επεξεργασία φυσικής γλώσσας (NLP), είναι διαθέσιμες στους χρήστες της γλώσσας προγραμματισμού Python, για αυτό και αποτελεί την προτιμόμενη γλώσσα για την υλοποίηση του πρακτικού τμήματος αυτής της εργασίας.

Για να γίνει Ανάλυση Συναισθήματος των tweets,είναι βασική η χρήση των παρακάτω Library/ βιβλιοθηκών οι οποίες είναι χρήσιμες για να βγάλουμε αξιόπιστα αποτελέσματα όσον αφορά την Sentiment Analysis:

1. Pandas
2. Numpy
3. Wordcloud
4. Re (Regular Expressions)
5. Matplotlib
6. Seaborn
7. NLTK (Natural Language Toolkit)
8. Sklearn (Scikit-Learn)

Συγκεκριμένα αυτές οι βιβλιοθήκες είναι οι κυριότερες που χρησιμοποιήθηκαν και λειτουργούν ως έχει:

1. **Pandas**:

Η βιβλιοθήκη Pandas στην Python είναι ένα ευέλικτο εργαλείο σχεδιασμένο για χειρισμό και ανάλυση δεδομένων. Προσφέρει δομές δεδομένων όπως DataFrames, επιτρέποντας στους χρήστες να οργανώνουν και να επεξεργάζονται δομημένα δεδομένα αποτελεσματικά. Η Pandas χρησιμοποιείται ευρέως για εργασίες όπως ο καθαρισμός δεδομένων, η εξερεύνηση, ο μετασχηματισμός και η στατιστική ανάλυση, καθιστώντας την απαραίτητη στις εργασίες της επιστήμης δεδομένων για την απλότητα και την αποτελεσματικότητά τους στο χειρισμό δεδομένων σε πίνακα. (Python Libraries)

2. **Numpy**:

Η βιβλιοθήκη NumPy στην Python παρέχει ισχυρά εργαλεία για αριθμητικούς υπολογισμούς, προσφέροντας αποτελεσματικές δομές δεδομένων και λειτουργίες για το χειρισμό μεγάλων πινάκων (arrays) και matrixs. Χρησιμοποιείται εκτενώς σε επιστημονικούς και μαθηματικούς υπολογισμούς λόγω των δυνατοτήτων υψηλής απόδοσης και του πλούσιου συνόλου συναρτήσεων για χειρισμό πινάκων, γραμμικής άλγεβρας και δημιουργία τυχαίων αριθμών. Η ευελιξία της NumPy την καθιστά ακρογωνιαίο λίθο σε διάφορους τομείς, συμπεριλαμβανομένης της ανάλυσης δεδομένων, της μηχανικής μάθησης και της επιστημονικής έρευνας, επιτρέποντας ταχύτερους και αποτελεσματικούς αριθμητικούς υπολογισμούς. (Python Libraries)

3. **Wordcloud**:

Η βιβλιοθήκη Wordcloud στην Python είναι ένα εργαλείο για τη δημιουργία "σύννεφων λέξεων" (word clouds) από δεδομένα κειμένου, οπτικοποιώντας τις συχνότητες λέξεων με οπτικά ελκυστικό τρόπο. Επιτρέπει στους χρήστες να δημιουργούν προσαρμόσιμα σύννεφα λέξεων, όπου το μέγεθος κάθε λέξης αντιστοιχεί στη συχνότητά της στο κείμενο. Τα σύννεφα λέξεων χρησιμοποιούνται συνήθως για διερευνητική ανάλυση δεδομένων, για τη σύνοψη πληροφοριών κειμένου και για την απόκτηση γνώσεων σχετικά με τις πιο συχνά εμφανιζόμενες λέξεις σε ένα σώμα. Είναι ιδιαίτερα χρήσιμα για τον εντοπισμό βασικών θεμάτων, τάσεων και μοτίβων σε μεγάλους όγκους δεδομένων κειμένου. (Python Libraries)

4. **Re**:

Η βιβλιοθήκη "Re" στην Python, συντομογραφία για "κανονικές εκφράσεις" (Regular Expressions), παρέχει υποστήριξη για εργασίες αντιστοίχισης προτύπων και χειρισμού κειμένου. Προσφέρει ένα ισχυρό σύνολο εργαλείων για αναζήτηση, εξαγωγή και αντικατάσταση μοτίβων εντός συμβολοσειρών, καθιστώντας το ανεκτίμητο για εργασίες όπως ο καθαρισμός δεδομένων, η ανάλυση κειμένου και η αναγνώριση μοτίβων. Με την ευέλικτη και εκφραστική σύνταξη της, η βιβλιοθήκη "re" επιτρέπει στους χρήστες να εκτελούν σύνθετες λειτουργίες συμβολοσειρών αποτελεσματικά, συμβάλλοντας σε διάφορες εφαρμογές στην επεξεργασία δεδομένων, την ανάλυση κειμένου και πολλά άλλα. (Python Libraries)

5. **Matplotlib**:

Η βιβλιοθήκη "Matplotlib" στην Python είναι μια ολοκληρωμένη βιβλιοθήκη σχεδίασης που χρησιμοποιείται για τη δημιουργία στατικών, κινούμενων και διαδραστικών απεικονίσεων. Προσφέρει ένα ευρύ φάσμα συναρτήσεων γραφικής παράστασης για τη δημιουργία διαφόρων τύπων γραφικών παραστάσεων, συμπεριλαμβανομένων γραμμικών γραφημάτων, διαγραμμάτων ράβδων, διαγραμμάτων διασποράς, ιστογραμμάτων και άλλων. Η Matplotlib παρέχει μια ευέλικτη και προσαρμόσιμη διεπαφή για τη δημιουργία ποιοτικών γραφημάτων, καθιστώντας την κατάλληλη για εξερεύνηση, ανάλυση και παρουσίαση δεδομένων σε τομείς όπως η επιστήμη δεδομένων και η επιστημονική έρευνα. Η ενσωμάτωσή του με άλλες βιβλιοθήκες Python, όπως οι NumPy και Pandas, ενισχύει περαιτέρω τις δυνατότητές του για εργασίες οπτικοποίησης και ανάλυσης δεδομένων. (Python Libraries)

6. **Seaborn**:

Η βιβλιοθήκη "Seaborn" στην Python είναι χτισμένη πάνω στο Matplotlib και παρέχει μια διεπαφή υψηλού επιπέδου για τη δημιουργία ελκυστικών και ενημερωτικών στατιστικών γραφικών. Απλοποιεί τη διαδικασία δημιουργίας πολύπλοκων απεικονίσεων όπως scatter plots, box plots και violin plots με ελάχιστο κώδικα. Το Seaborn προσφέρει ενσωματωμένα θέματα και χρωματικές παλέτες, ενισχύοντας την αισθητική των γραφημάτων και παρέχει υποστήριξη για την αποτελεσματική οπτικοποίηση κατηγορικών και σχεσιακών δεδομένων. Χρησιμοποιείται ευρέως σε

εργασίες ανάλυσης δεδομένων και εξερεύνησης, ιδιαίτερα σε τομείς όπως η επιστήμη των δεδομένων, η μηχανική μάθηση και η στατιστική μοντελοποίηση. (Python Libraries)

7. **NLTK**:

Η βιβλιοθήκη "NLTK" (Natural Language Toolkit) στην Python είναι μια ολοκληρωμένη πλατφόρμα για εργασία με δεδομένα ανθρώπινης γλώσσας. Προσφέρει διάφορα εργαλεία και πόρους για εργασίες όπως το tokenization, το stemming, το lemmatization, το part of-speech tagging και το parsing. Η NLTK χρησιμοποιείται ευρέως σε εφαρμογές επεξεργασίας φυσικής γλώσσας (NLP), συμπεριλαμβανομένης της ανάλυσης συναισθήματος, της ταξινόμησης κειμένων, της μηχανικής μετάφρασης και της εξαγωγής πληροφοριών, λόγω της εκτεταμένης συλλογής σωμάτων, λεξιλογικών πόρων και αλγορίθμων προσαρμοσμένων για εργασίες επεξεργασίας κειμένου. (Python Libraries)

8. **Sklearn**:

Η βιβλιοθήκη "Sklearn" (Scikit-learn) στην Python είναι μια ισχυρή βιβλιοθήκη μηχανικής εκμάθησης που παρέχει απλά και αποτελεσματικά εργαλεία για εξόρυξη δεδομένων και ανάλυση δεδομένων. Προσφέρει ένα ευρύ φάσμα εποπτευόμενων και μη εποπτευόμενων αλγορίθμων μάθησης, όπως ταξινόμηση, παλινδρόμηση, ομαδοποίηση, μείωση διαστάσεων και αξιολόγηση μοντέλων. Το Sklearn χρησιμοποιείται ευρέως για τη δημιουργία μοντέλων μηχανικής μάθησης, την εκτέλεση εξαγωγής χαρακτηριστικών και την αξιολόγηση της απόδοσης του μοντέλου λόγω της φιλικής προς το χρήστη διεπαφής, της εκτεταμένης τεκμηρίωσης και της αποτελεσματικής εφαρμογής αλγορίθμων. (Python Libraries)

3.1.3.1 Google Colaboratory – Colab

Στην παρούσα εργασία, οι εντολές θα γραφτούν στο Google Collaboratory, συχνά γνωστό ως Colab. Το Colab είναι ένα δωρεάν φορητό περιβάλλον Jupyter Notebook που φιλοξενείται στο cloud και συγκεκριμένα στο Google Drive του χρήστη, ενώ επιτρέπει τη σύνταξη, την εκτέλεση και την κοινή χρήση κώδικα Python μέσω ενός προγράμματος περιήγησης ιστού (web browser). Χρησιμοποιείται συχνά σε έργα που επικεντρώνονται στην ανάλυση δεδομένων, τη μηχανική εκμάθηση αλλά και άλλα θέματα.

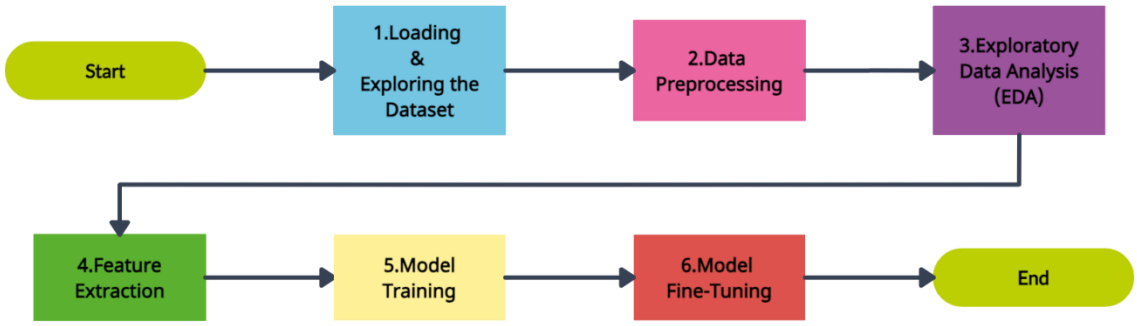
Εικόνα 14. Logo της Python και του Colab

3.2 Μεθοδολογία και Ανάλυση των Δεδομένων

Η μεθοδολογία που χρησιμοποιήθηκε σε αυτή τη μελέτη αποτελεί τη ραχοκοκαλιά της προσπάθειάς μας να αξιοποιήσουμε τις τεχνικές Επεξεργασίας Φυσικής Γλώσσας (NLP) για την ανάλυση συναισθήματος Twitter, εστιάζοντας ειδικά στον εντοπισμό της ρητορικής μίσους. Σε αυτήν την ενότητα, περιγράφουμε τη διαδικασία βήμα προς βήμα μέσω της οποίας προσεγγίσαμε την ανάλυση, από τη ανάγνωση δεδομένων έως την εκπαίδευση μοντέλων με λεπτομέρεια.

Αρχικά, είναι επιτακτική ανάγκη να υπογραμμιστεί η σημασία της μεθοδολογικής αυστηρότητας στον χειρισμό ευαίσθητων θεμάτων όπως ο εντοπισμός της ρητορικής μίσους. Δεδομένης της διαφοροποιημένης φύσης της γλώσσας και των πιθανών συνεπειών της εσφαλμένης ταξινόμησης, η μεθοδολογία μας ακολουθεί από τις βέλτιστες πρακτικές και χρησιμοποιεί αρκετά ισχυρές τεχνικές για να διασφαλίσει την ακεραιότητα και την αξιοπιστία των ευρημάτων μας.

Η μεθοδολογία μας είναι δομημένη σε έξι βασικές φάσεις, καθεμία από τις οποίες εξυπηρετεί έναν ξεχωριστό σκοπό.



Εικόνα 15. Τα βήματα της μεθοδολογίας που εφαρμόστηκε σε μορφή flow chart (Δημιουργία μέσω του app.creately.com)

3.2.1 Φόρτωση και εξερεύνηση του dataset

Εισαγωγή Βιβλιοθηκών – Import Libraries

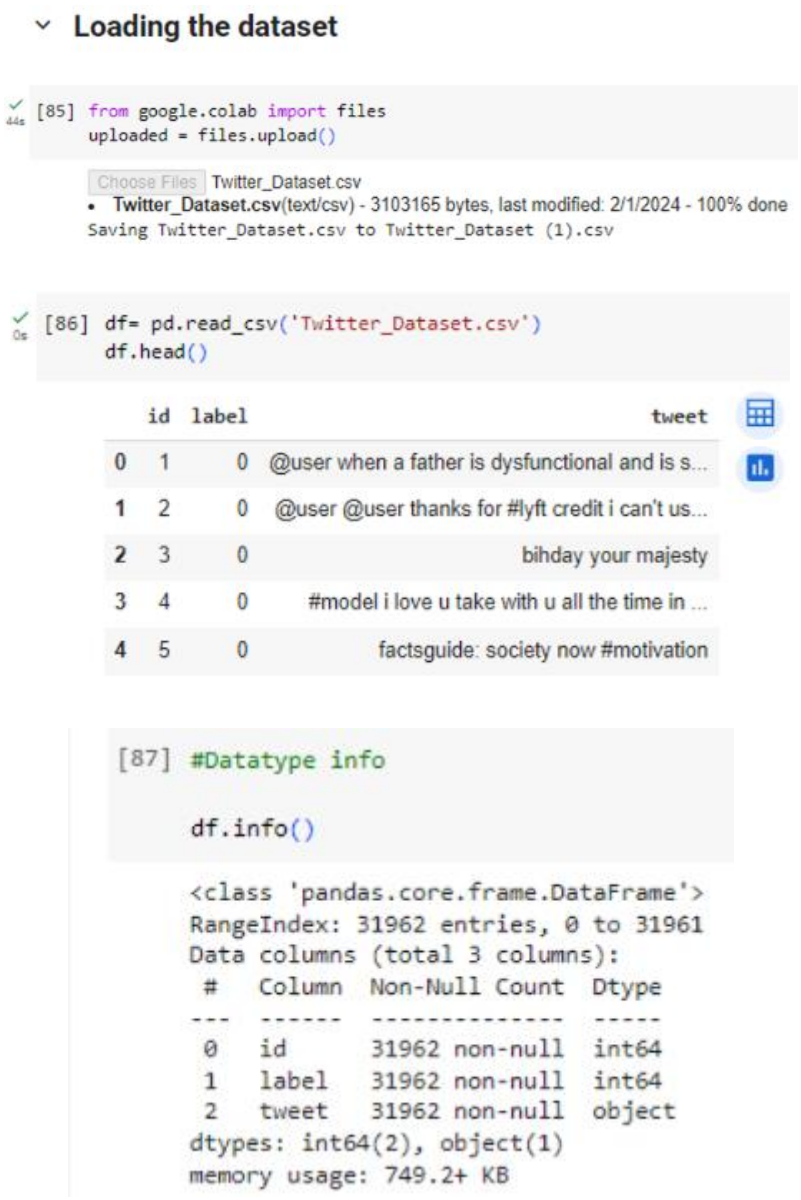
Το πράγμα με το οποίο ξεκινήσαμε τον κώδικα μας στο νέο μας notebook του colab είναι να εισάγουμε κάποιες βασικές βιβλιοθήκες (κατά την διάρκεια συγγραφής του κώδικα γεννήθηκε η ανάγκη για την εισαγωγή κι άλλων βιβλιοθηκών, που τις εισάγαμε την στιγμή που τις χρειαστήκαμε).

```
[84] import pandas as pd
     import numpy as np
     import matplotlib.pyplot as plt
     import seaborn as sns
     import re
     import string
     import nltk
     import warnings
     %matplotlib inline

     warnings.filterwarnings('ignore')
```

Εικόνα 16. Τμήμα κώδικα για την εισαγωγή βασικών βιβλιοθηκών

Φόρτωση του Dataset

Αρχικά φορτώνουμε το σύνολο δεδομένων στο περιβάλλον ανάλυσης. Το απόσπασμα κώδικα δείχνει τη διαδικασία φόρτωσης ενός συνόλου δεδομένων που ονομάζεται "Twitter_Dataset.csv" χρησιμοποιώντας τη δυνατότητα μεταφόρτωσης αρχείων του Google Colab. Το σύνολο δεδομένων περιέχει 31.962 εγγραφές με τρεις στήλες: "id", "label" και "tweet". Η στήλη "label" υποδηλώνει εάν ένα tweet ταξινομείται ως ρατσιστικό/σεξιστικό (1) ή όχι (0). Στη συνέχεια, το σύνολο δεδομένων εμφανίζεται χρησιμοποιώντας τη συνάρτηση head() για να δώσει μια ματιά στη δομή των δεδομένων, ακολουθούμενη από τη συνάρτηση info() για να ληφθούν πληροφορίες σχετικά με τις στήλες του συνόλου δεδομένων, τις μηδενικές μετρήσεις και τους τύπους δεδομένων.



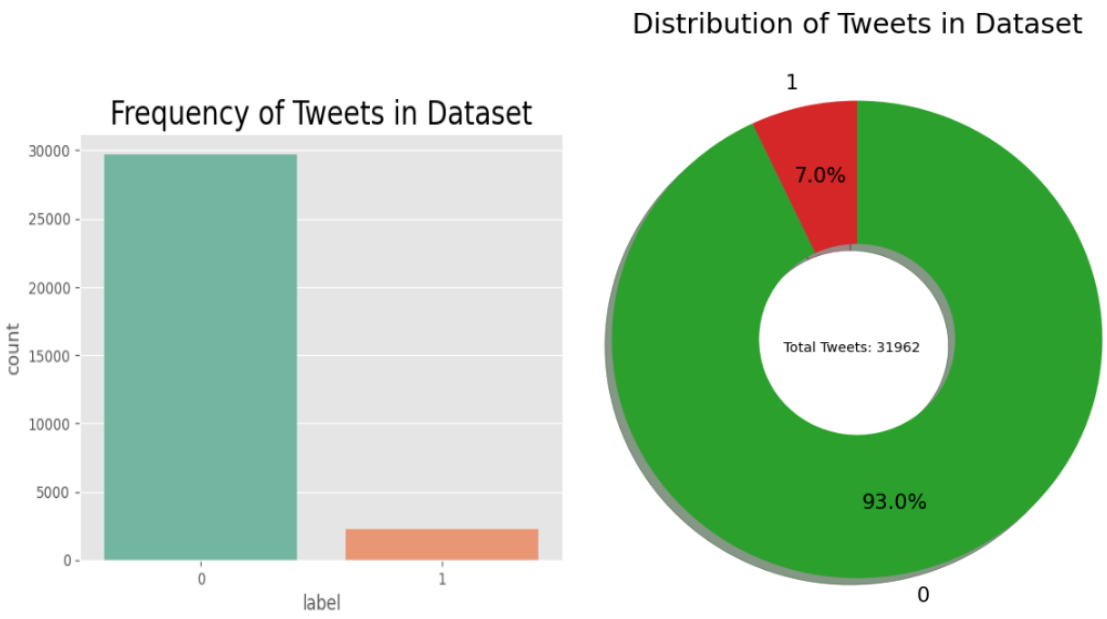
Εικόνα 17. Τμήμα κώδικα για την φόρτωση του Dataset

Μετά την εισαγωγή του συνόλου δεδομένων και πριν ξεκινήσουμε την εξερεύνηση του, το σύνολο δεδομένων υποβλήθηκε σε μερικές μικρές ενέργειες προεπεξεργασίας για να διασφαλιστεί η καθαρότητα και η ακεραιότητα των δεδομένων πριν από την εξερεύνηση. Οι ενέργειες αυτές αφορούν την αφαίρεση των διπλότυπων εγγραφών χρησιμοποιώντας τη συνάρτηση drop_duplicates() και τον εντοπισμό τυχόν τιμών που λείπουν στις στήλες "label" και "tweet" το isna().sum(), αποκαλύπτοντας ότι δεν υπάρχουν περιπτώσεις διπλότυπων ή χαμένων τιμών. Αυτά τα βήματα εξασφάλισαν ένα ισχυρό σύνολο δεδομένων για μετέπειτα ανάλυση.

Εικόνα 18. Τμήμα κώδικα για την αφαίρεση των διπλότυπων εγγραφών και το εντοπισμό τιμών που λείπουν

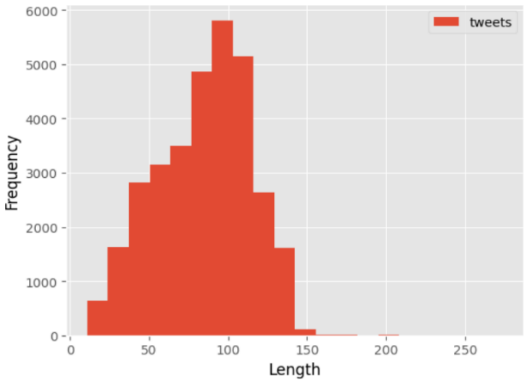
Εξερεύνηση του Dataset

Στη φάση της εξερεύνησης, η κατανομή της τάξης του συνόλου δεδομένων εξετάστηκε για να κατανοηθεί η ισορροπία μεταξύ των tweets που χαρακτηρίζονται ως μη ρατσιστικά/σεξιστικά (0) και εκείνων που χαρακτηρίζονται ως ρατσιστικά/σεξιστικά (1). Η μέτρηση των τιμών με βάση τον "label" αποκάλυψαν μια ταξική ανισορροπία, με 29.720 tweets (~93%) να χαρακτηρίζονται ως μη ρατσιστικά/σεξιστικά και μόνο 2.242 tweets (~7%) χαρακτηρισμένα ως ρατσιστικά/σεξιστικά. Η απεικόνιση αυτής της κατανομής μέσω μιας γραφικής παράστασης μέτρησης και ενός γραφήματος πίτας κάνει πιο κατανοητή τη δυσαναλογία της τάξης, υπογραμμίζοντας την ανάγκη για προσεκτική εξέταση κατά τη διάρκεια της εκπαίδευσης μοντέλων για την αποφυγή μεροληπτικών προβλέψεων προς την πλειοψηφική τάξη. Παρόλα αυτά εφόσον το ποσοστό των ρατσιστικών/ σεξιστικών tweets είναι κοντά στο 10% δεν χρειάζεται να καταφύγουμε σε μεθόδους εξομάλυνσης του συνόλου δεδομένων.



Διάγραμμα συχνότητας και διάγραμμα κατανομής των tweets στο dataset

Συνεχίσαμε με την προσθήκη μιας νέας στήλης "length" στο πλαίσιο δεδομένων που υποδηλώσει τον αριθμό χαρακτήρων κάθε tweet και σχεδιάσαμε ένα ιστόγραμμα για να απεικονίσει την κατανομή των μηκών των tweet. Το ιστόγραμμα εμφανίζει τη συχνότητα των μηκών tweet σε όλο το σύνολο δεδομένων. Η ανάλυση του ιστογράμματος αποκαλύπτει ότι η πλειονότητα των tweets εμπίπτει στην περιοχή από 90 έως 103 χαρακτήρες, με 5809 tweets να εμπίπτουν σε αυτό το εύρος. Καθώς το μήκος του tweet αυξάνεται ή μειώνεται από αυτό το εύρος, η συχνότητα των tweet μειώνεται γενικά, υποδεικνύοντας μια κάπως κανονική κατανομή με μια κορυφή γύρω από το μέσο μήκος του tweet.



Διάγραμμα συχνότητας των μηκών tweet

Ο κώδικας μας συνέχισε με την εντολή “df.describe()”, όπου η συνάρτηση "describe()" παρέχει συνοπτικά στατιστικά στοιχεία για τις αριθμητικές στήλες στο σύνολο

δεδομένων. Αυτά τα στατιστικά στοιχεία παρέχουν πληροφορίες σχετικά με τη κατανομή και τα χαρακτηριστικά σύνοψης, βοηθούν στην κατανόηση της κεντρικής τάσης, της μεταβλητότητας και της εξάπλωσης των δεδομένων.

	id	label	length
count	31962.000000	31962.000000	31962.000000
mean	15981.500000	0.070146	84.739628
std	9226.778988	0.255397	29.455749
min	1.000000	0.000000	11.000000
25%	7991.250000	0.000000	63.000000
50%	15981.500000	0.000000	88.000000
75%	23971.750000	0.000000	108.000000
max	31962.000000	1.000000	274.000000

Εικόνα 19. Πίνακας στατιστικών στοιχείων

Από την εικόνα 18, αξίζει να σχολιάσουμε:

- Η μέση τιμή της στήλης "label" που είναι 0,07 υποδηλώνει ότι ένα μικρό ποσοστό των tweets (7,01%) χαρακτηρίζονται ως ρατσιστικά/σεξιστικά (που συμπίπτει με τα διαγράματα συχνοτήτων και κατανομής που είχαν προηγηθεί).
- Η στήλη "length" που αντιπροσωπεύει τον αριθμό χαρακτήρων κάθε tweet, έχει μέσο μήκος περίπου 84,74 χαρακτήρες και τυπική απόκλιση περίπου 29,46. Το ελάχιστο και το μέγιστο μήκη υποδεικνύουν το εύρος των μηκών των tweets στο σύνολο δεδομένων, από 11 έως 274 χαρακτήρες. Οι τιμές τεταρτημορίου παρέχουν πληροφορίες σχετικά με την κατανομή των μηκών των tweets, με το 25% των tweets να έχουν μήκος 63 χαρακτήρες ή λιγότερο, το 50% να έχουν μήκος 88 χαρακτήρες ή λιγότερους (διάμεσος) και το 75% να έχουν μήκος 108 χαρακτήρες ή λιγότεροι.

‘Εχοντας αυτά τα στοιχεία θέλαμε για λόγους περιέργειας να εμφανίσουμε μερικά tweets. Επομένως έχουμε:

- Tweet με το ελάχιστο μήκος χαρακτήρων,
- Tweet με το μέγιστο μήκος χαρακτήρων και
- Tweet με το μέσο μήκος χαρακτήρων:

```
df[df['length'] == min(df['length'])]['tweet'].iloc[0]
```

```
'i love you '
```

```
df[df['length'] == max(df['length'])]['tweet'].iloc[0]
```

```
' @user lmfao pathetic #soit   #growup #funny #noonethere #iknowwhoitis ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f
\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x
9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f\x98±ð\x9f\x98±ð\x9f¤\x97ð\x9f¤\x97ð\x9f\x98±ð
\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97ð\x9f¤\x97
ð\x9f¤\x97ð\x9f¤\x97ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ð\x9f\x98\x82ã\x80¦'
```

```
df[df['length'] == 84]['tweet'].iloc[0]
```

```
'my mom shares the same bihday as @user   bihday snake! see you this weekend ð\x9f\x99\x8cð\x9f\x8f¾'
```

Εικόνα 20. Tweets με διάφορα μήκος χαρακτήρων

Προχωράμε με την δημιουργία δύο νέων DataFrames, τα "nohate_t" και "hate_t", τα οποία διαχωρίζουν τα tweets με βάση την ετικέτα τους, υποδεικνύοντας εάν περιέχουν ρατσιστικό/σεξιστικό περιεχόμενο ή όχι.

Για τα tweets χωρίς ρατσιστικό/σεξιστικό περιεχόμενο (nohate_t):

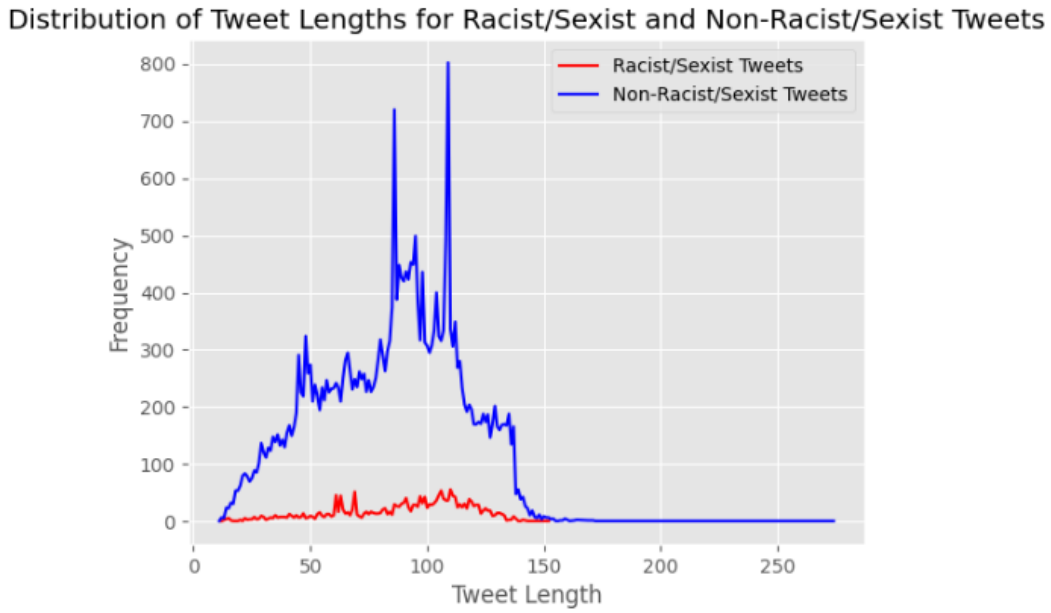
	id	label	length
count	29720.000000	29720.0	29720.000000
mean	15974.454441	0.0	84.328634
std	9223.783469	0.0	29.566484
min	1.000000	0.0	11.000000
25%	7981.750000	0.0	62.000000
50%	15971.500000	0.0	88.000000
75%	23965.250000	0.0	107.000000
max	31962.000000	0.0	274.000000

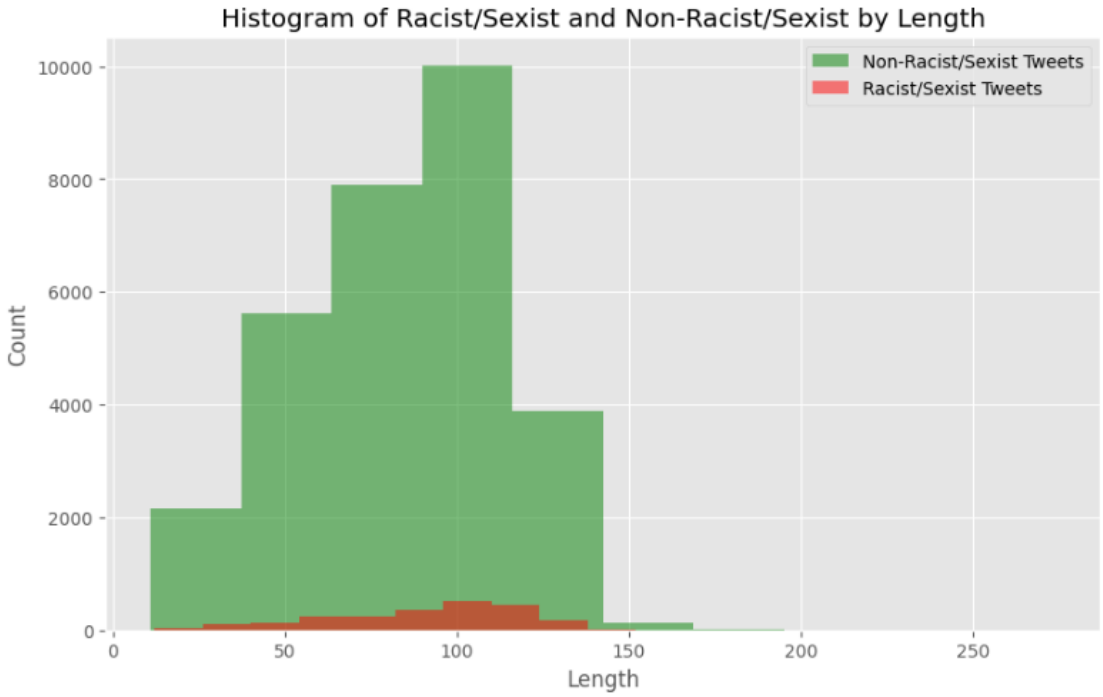
- Υπάρχουν 29.720 tweets που ταξινομούνται ως μη ρατσιστικού/σεξιστικού περιεχομένου (label=0).
- Το μέσο μήκος tweet είναι περίπου 84,33 χαρακτήρες, με τυπική απόκλιση περίπου 29,57 χαρακτήρες.
- Τα μήκη των tweet κυμαίνονται από 11 έως 274 χαρακτήρες, με τις τιμές τεταρτημορίου να υποδεικνύουν την κατανομή των μηκών των tweet.

Για τα tweets που περιέχουν ρατσιστικό/σεξιστικό περιεχόμενο (hate_t):

	id	label	length
count	2242.000000	2242.0	2242.000000
mean	16074.896075	1.0	90.187779
std	9267.955758	0.0	27.375502
min	14.000000	1.0	12.000000
25%	8075.250000	1.0	69.000000
50%	16095.000000	1.0	96.000000
75%	24022.000000	1.0	111.000000
max	31961.000000	1.0	152.000000

- Υπάρχουν 2.242 tweets που έχουν ταξινομηθεί ως που περιέχουν ρατσιστικό/σεξιστικό περιεχόμενο (ετικέτα=1).
- Το μέσο μήκος του tweet είναι ελαφρώς μεγαλύτερο σε περίπου 90,19 χαρακτήρες, με τυπική απόκλιση περίπου 27,38 χαρακτήρες.
- Τα μήκη tweet για αυτά τα tweet κυμαίνονται από 12 έως 152 χαρακτήρες, με τις τιμές τεταρτημορίου να παρέχουν πληροφορίες σχετικά με την κατανομή των μηκών tweet.





Διαγράμματα συχνότητας και κατανομής

Αυτά τα αποτελέσματα υποδηλώνουν ότι υπάρχει διαφορά στη διάρκεια των tweet μεταξύ των tweet με και χωρίς ρατσιστικό/σεξιστικό περιεχόμενο, με τα tweets που περιέχουν τέτοιο περιεχόμενο γενικά να είναι ελαφρώς μεγαλύτερα κατά μέσο όρο.

3.2.2 Προεπεξεργασία των δεδομένων

Η διαδικασία προεπεξεργασίας δεδομένων είναι ζωτικής σημασίας για την προετοιμασία των δεδομένων κειμένου για ανάλυση και μοντελοποίηση. Σε αυτή τη φάση, εφαρμόζονται διάφορες τεχνικές για τον καθαρισμό και την τυποποίηση του κειμένου, καθιστώντας το κατάλληλο για περαιτέρω ανάλυση. Αυτό διασφαλίζει ότι τα δεδομένα είναι συνεπή, ακριβή και απαλλαγμένα από θόρυβο που θα μπορούσε να επηρεάσει δυνητικά την απόδοση μεταγενέστερων εργασιών, όπως η ανάλυση συναισθήματος. Τα ακόλουθα βήματα χρησιμοποιήθηκαν για την προεπεξεργασία δεδομένων:

1. Κατάργηση των χρηστών Twitter (@user)
2. Αφαίρεση σημείων στίξης, αριθμών και ειδικών χαρακτήρων
3. Αντικατάσταση των «αργκό» λέξεων με τις αρχικές λέξεις
4. Αφαίρεση σύντομων λέξεων
5. Αφαίρεση της λέξης "hmm" και των παραλλαγών της
6. Κανονοποίηση κειμένου (Text Normalization)
    1. Διακριτοποίηση (Tokenization)

2. Λημματοποίηση (Lemmitization)
3. Στελέχωση (Stemming)
4. Επανένωση του κειμένου

Με τη συστηματική εφαρμογή αυτών των τεχνικών προεπεξεργασίας, τα δεδομένα κειμένου γίνονται τυποποιημένα, καθαρότερα και καταλληλότερα για επακόλουθες εργασίες ανάλυσης και μοντελοποίησης. Ακολουθεί ανάλυση των παραπάνω βημάτων.

Κατάργηση των χρηστών Twitter (@user)

Αφαιρούμε το όνομα του χρήστη, που υποδηλώνονται με το "@user", καθώς δεν συμβάλλει στην ανάλυση του συναίσθηματος ή στο περιεχόμενο του tweet. Για να βλέπουμε τις αλλαγές μας στο περιεχόμενο των tweets μετά την εφαρμογή του κάθε βήματος προεπεξεργασίας, δημιουργήσαμε μία νέα στήλη με την ονομασία "clean_tweet" για να βλέπουμε το αποτέλεσμα αλλά ταυτόχρονα να μπορούμε να συγκρίνουμε το επεξεργασμένο tweet με το αρχικό tweet που βρίσκεται στην στήλη "tweet".

	id	label	tweet	length	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	102	when a father is dysfunctional and is so sel...
1	2	0	@user @user thanks for #lyft credit i can't us...	122	thanks for #lyft credit i can't use cause th...
2	3	0	bihday your majesty	21	bihday your majesty
3	4	0	#model i love u take with u all the time in ...	86	#model i love u take with u all the time in ...
4	5	0	factsguide: society now #motivation	39	factsguide: society now #motivation

Εικόνα 21. Tweets μετά των κατάργηση του "@user"

Αφαίρεση σημείων στίξης, αριθμών και ειδικών χαρακτήρων

Αυτά τα στοιχεία είναι συχνά άσχετα με την εργασία ανάλυσης συναισθήματος και μπορούν να αφαιρεθούν με ασφάλεια για να μειωθεί ο θόρυβος και να βελτιωθεί η υπολογιστική απόδοση. Η "εκφραση" (expression) [^a-zA-Z#] που είναι ένα τυπικό μοτίβο έκφρασης, χρησιμοποιείται για την αφαίρεση χαρακτήρων που δεν είναι γράμματα (και κεφαλαία και πεζά) ή το σύμβολο "#".

	id	label	tweet	length	clean_tweet
14	15	1	no comment! in #australia #opkillingbay #se...	101	no comment in #australia #opkillingbay #se...
15	16	0	ouch...junior is angryð□□□#got7 #junior #yugyo...	56	ouch junior is angry #got #junior #yugyo...
16	17	0	i am thankful for having a paner. #thankful #p...	58	i am thankful for having a paner #thankful #p...
17	18	1	retweet if you agree!	22	retweet if you agree
18	19	0	its #friday! ð□□□ smiles all around via ig use...	78	its #friday smiles all around via ig use...
19	20	0	as we all know, essential oils are not made of...	58	as we all know essential oils are not made of...
20	21	0	#euro2016 people blaming ha for conceded goal ...	127	#euro people blaming ha for conceded goal ...

Εικόνα 22. Tweets μετά την αφαίρεση σημείων στίξης, αριθμών και ειδικών χαρακτήρων

Αντικατάσταση των "αργκό" λέξεων με τις αρχικές λέξεις

Οι όροι και οι συντομογραφίες της αργκό αντικαθίστανται με τις αντίστοιχες πρωτότυπες λέξεις για να διασφαλιστεί η συνέπεια και να βελτιωθεί η ερμηνευτικότητα του κειμένου. Για τον λόγο αυτό, ήταν ανάγκη να δημιουργήσουμε ένα λεξικό (dictionary) που περιέχει αντιστοιχίσεις λέξεων της αργκό στην αρχική τους μορφή. Κάθε ζεύγος κλειδιού-τιμής αντιπροσωπεύει μια λέξη αργκό και την αντίστοιχη αρχική της μορφή.

```
slang_dictionary = {
    "$" : " dollar ",
    "€" : " euro ",
    "4ao" : "for adults only",
    "a.m" : "before midday",
    "a3" : "anytime anywhere anyplace",
    "aamof" : "as a matter of fact",
    "acct" : "account",
    "adih" : "another day in hell",
    "af" : "as fuck",
    "afaic" : "as far as i am concerned",
    "afaict" : "as far as i can tell",
    "afaik" : "as far as i know",
    "afair" : "as far as i remember",
```

Εικόνα 23. Απόσπασμα του λεξικού της αργκό

id	label	tweet	length	clean_tweet
4	0	#model i love u take with u all the time in ...	86	#model i love you take with you all the time i...

Εικόνα 24. Tweet μετά την αντικατάσταση των "αργκό" λέξεων

Αφαίρεση σύντομων λέξεων

Οι σύντομες λέξεις αφαιρούνται καθώς συχνά έχουν μικρή σημασιολογική σημασία και μπορούν να θεωρηθούν θόρυβος στο σύνολο δεδομένων. Στην παρούσα εργασία επιλέξαμε να αφαιρέσουμε τις λέξεις που έχουν τρεις ή λιγότερους χαρακτήρες.

id	label	tweet	length	clean_tweet
1	0	@user when a father is dysfunctional and is s...	102	when father dysfunctional selfish drags kids i...

Εικόνα 25. Tweet μετά την αφαίρεση σύντομων λέξεων

Αφαίρεση της λέξης "hmm" και των παραλλαγών της

Λέξεις όπως "hmm" που μεταφέρουν ελάχιστο συναίσθημα ή πληροφορίες αφαιρούνται για να επικεντρωθούν σε πιο ουσιαστικό περιεχόμενο.

	id	label	tweet	length	clean_tweet
22512		1	hmmm...sounds familiar. @user @user @user @use...	80	sounds familiar hasn held news

Εικόνα 26. Tweet μετά την αφαίρεση της λέξης "hmm"

Κανονικοποίηση κειμένου (Text Normalization)

Η κανονικοποίηση κειμένου είναι η διαδικασία μετατροπής δεδομένων κειμένου σε τυποποιημένη μορφή, που διευκολύνει την επεξεργασία και την ανάλυση. Αυτό συνήθως περιλαμβάνει τεχνικές όπως η διακριτοποίηση (tokenization), η λημματοποίηση (lemmitization) και η στελέχωση (stemming) για τη μείωση των παραλλαγών στο κείμενο και τη βελτίωση της συνέπειας για εργασίες. Για παράδειγμα, η αναγωγή όρων όπως loves, loving και lovable στη βασική τους λέξη, π.χ., "love". χρησιμοποιούνται συχνά στο ίδιο πλαίσιο. Αν μπορούμε να τα αναγάγουμε στη ρίζα τους, που είναι "love", θα βοηθήσει στη μείωση του συνολικού αριθμού μοναδικών λέξεων στα δεδομένα μας χωρίς να χάσουμε σημαντικό όγκο πληροφοριών.

Διακριτοποίηση (Tokenization)

Αναλύουμε ή αλλίως "σπάμε" την πρόταση/ το κειμένου του tweet σε μεμονωμένες λέξεις ή tokens.

```
tokenized_tweet = df['clean_tweet'].apply(lambda x: x.split())
tokenized_tweet.head()
```

```
0    [when, father, dysfunctional, selfish, drags, ...
1    [thanks, #lyft, credit, cause, they, offer, wh...
2                               [bihday, your, majesty]
3                    [#model, love, take, with, time]
4                  [factsguide, society, #motivation]
Name: clean_tweet, dtype: object
```

Εικόνα 27. Tweets μετά την εφαρμογή του tokenization

Λημματοποίηση (Lemmitization)

Μετατρέπουμε τις λέξεις στη βασική τους μορφή για μείωση μορφών και παραλλαγών κλίσης.

Στελέχωση (Stemming)

Ανάγουμε τις λέξεις στη ρίζα τους αφαιρώντας προθέματα και επιθήματα.

```
0    [when, father, dysfunct, selfish, drag, kid, i...
1    [thank, #lyft, credit, caus, they, offer, whee...
2                                [bihday, your, majesti]
3                      [#model, love, take, with, time]
4                          [factsguid, societi, #motiv]
Name: clean_tweet, dtype: object
```

Εικόνα 28. Tweets/tokens μετά την εφαρμογή του lemmatization και stemming

Επανένωση του κειμένου

Ανακατασκευάζουμε το κείμενο (πρόταση) του tweet μετά την κανονικοποίηση για να διατηρήσει την αρχική του δομή και αναγνωσιμότητα.

	id	label	tweet	length	clean_tweet
0	1	0	@user when a father is dysfunctional and is s...	102	when father dysfunct selfish drag kid into dys...
1	2	0	@user @user thanks for #lyft credit i can't us...	122	thank #lyft credit caus they offer wheelchair ...
2	3	0	bihday your majesty	21	bihday your majesti
3	4	0	#model i love u take with u all the time in ...	86	#model love take with time
4	5	0	factsguide: society now #motivation	39	factsguid societi #motiv

Εικόνα 29. Tweets/tokens μετά την επανένωση των tokens

Μετά την εφαρμογή των βημάτων προεπεξεργασίας των κειμένων/ tweets, προσθέσαμε ακόμη μια στήλη "clean_length", η οποία αντιπροσοπεύση το μήκος του κειμένου μετά τον "καθαρισμό" του. Παρατηρούμε ότι το μήκος των tweets έχει μειωθεί αισθητά.

	id	label	tweet	length	clean_tweet	clean_length
0	1	0	@user when a father is dysfunctional and is s...	102	when father dysfunct selfish drag kid into dys...	56
1	2	0	@user @user thanks for #lyft credit i can't us...	122	thank #lyft credit caus they offer wheelchair ...	70
2	3	0	bihday your majesty	21	bihday your majesti	19
3	4	0	#model i love u take with u all the time in ...	86	#model love take with time	26
4	5	0	factsguide: society now #motivation	39	factsguid societi #motiv	24

Εικόνα 30. Tweets με την προσθήκη της στήλης clean_length

3.2.3 Διερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis - EDA)

Στη φάση της Διερευνητικής Ανάλυσης Δεδομένων (EDA), εμβαθύνουμε στο σύνολο δεδομένων για να αποκαλύψουμε μοτίβα, τάσεις και σχέσεις που μπορούν να παρέχουν πολύτιμες πληροφορίες. Μέσω της οπτικοποίησης, στοχεύουμε να αποκτήσουμε μια βαθύτερη κατανόηση των χαρακτηριστικών. Αυτή η διαδικασία είναι ζωτικής σημασίας για τα επόμενα βήματα της ανάλυσης.

Οπτικοποίηση όλων των λέξεων των tweets

Αρχικά στην εξερεύνηση του συνόλου δεδομένων, αξιοποιήσαμε τη δύναμη της βιβλιοθήκης wordcloud για να απεικονίσουμε τη συχνότητα των λέξεων που χρησιμοποιούνται σε όλα τα tweets. Αυτή η προσέγγιση παρέχει μια πανοραμική άποψη των πιο διαδεδομένων όρων στο σώμα (corpus), προσφέροντας πληροφορίες για επικρατέστερα θέματα ή συναισθήματα που εκφράζονται στο σύνολο δεδομένων. Αντιπροσωπεύοντας οπτικά τη συχνότητα λέξεων, η "λέξη σύννεφο" (word cloud) βοηθά στον εντοπισμό βασικών όρων που μπορεί να δικαιολογούν περαιτέρω έρευνα και χρησιμεύει ως πολύτιμο σημείο εκκίνησης για την κατανόηση των γενικών θεμάτων στα δεδομένα του tweet.

Εικόνα 31. Οπτικοποίηση όλων των λέξεων των tweets

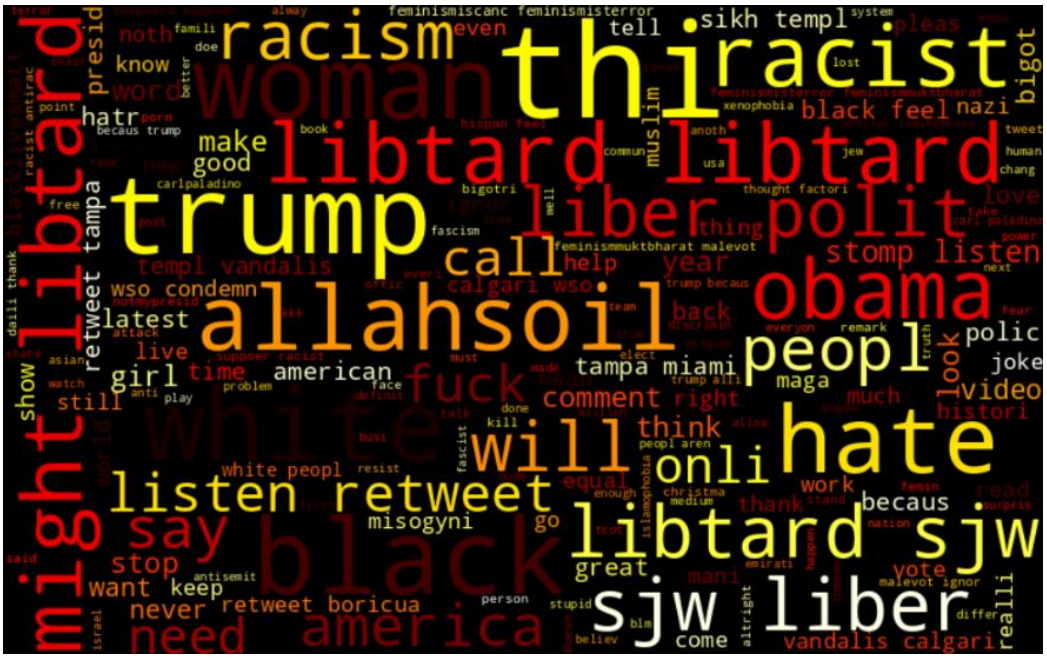
Μια σαφής εικόνα όλων των συχνών λέξεων, οι πιο συχνές λέξεις εμφανίζονται μεγαλύτερες και οι λιγότερο συχνές λέξεις εμφανίζονται μικρότερες. Η οπτικοποίηση εμφανίζει πολλές θετικές λέξεις και μερικές αρνητικές λέξεις. Επίσης χρησιμοποιήθηκε το λογότυπο του Twitter για μια πιο προτότυπη απεικόνιση.

Οπτικοποίηση λέξεων από μη ρατσιστικά/σεξιστικά tweets και λέξεων από ρατσιστικά/σεξιστικά tweets

Εμβαθύναμε στο περιεχόμενο των tweets που κατηγοριοποιήθηκαν ως μη ρατσιστικά/σεξιστικά και εκείνων που κατηγοριοποιήθηκαν ως ρατσιστικά/σεξιστικά. Οπτικοποιώντας τις λέξεις που εμφανίζονται πιο συχνά σε κάθε κατηγορία, αποκτήσαμε πληροφορίες για το ξεχωριστό λεξιλόγιο και τα θέματα που επικρατούν σε αυτά τα tweets. Αυτές οι οπτικοποιήσεις τόνισαν μερικές από τις λέξεις-κλειδιά και φράσεις που χρησιμοποιούνται σε κάθε κατηγορία, ρίχνοντας φως στη γλώσσα και τα θέματα που συνήθως συνδέονται με τον μη ρατσιστικό/σεξιστικό λόγο, καθώς και εκείνα που ενδεικτικά ρατσιστικού/σεξιστικού περιεχομένου.

Εικόνα 32. Οπτικοποίηση λέξεων των tweets από μη ρατσιστικά/σεξιστικά tweets

Από την εικόνα καταλαβαίνουμε ότι πολλα μη ρατσιστικά/σεξιστικά tweets μιλούν για την αγάπη (love), την ζωή(life), την φιλία (friend) και πολλά άλλα. Για την δημιουργία αυτής της εικόνας χρησιμοποιήθηκαν "χαρούμενα" και έντονα χρώματα για να ταιριάζουν με το περιεχόμενο.

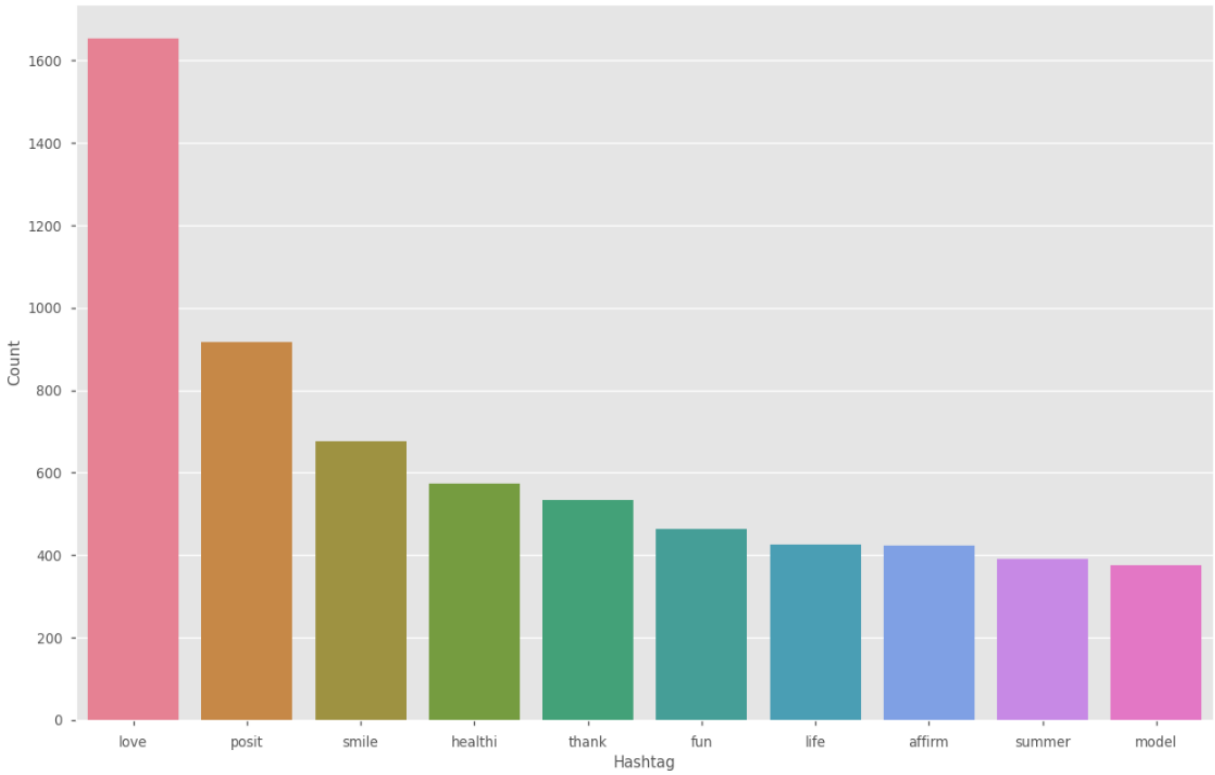


Εικόνα 33. Οπτικοποίηση λέξεων των tweets από ρατσιστικά/σεξιστικά tweets

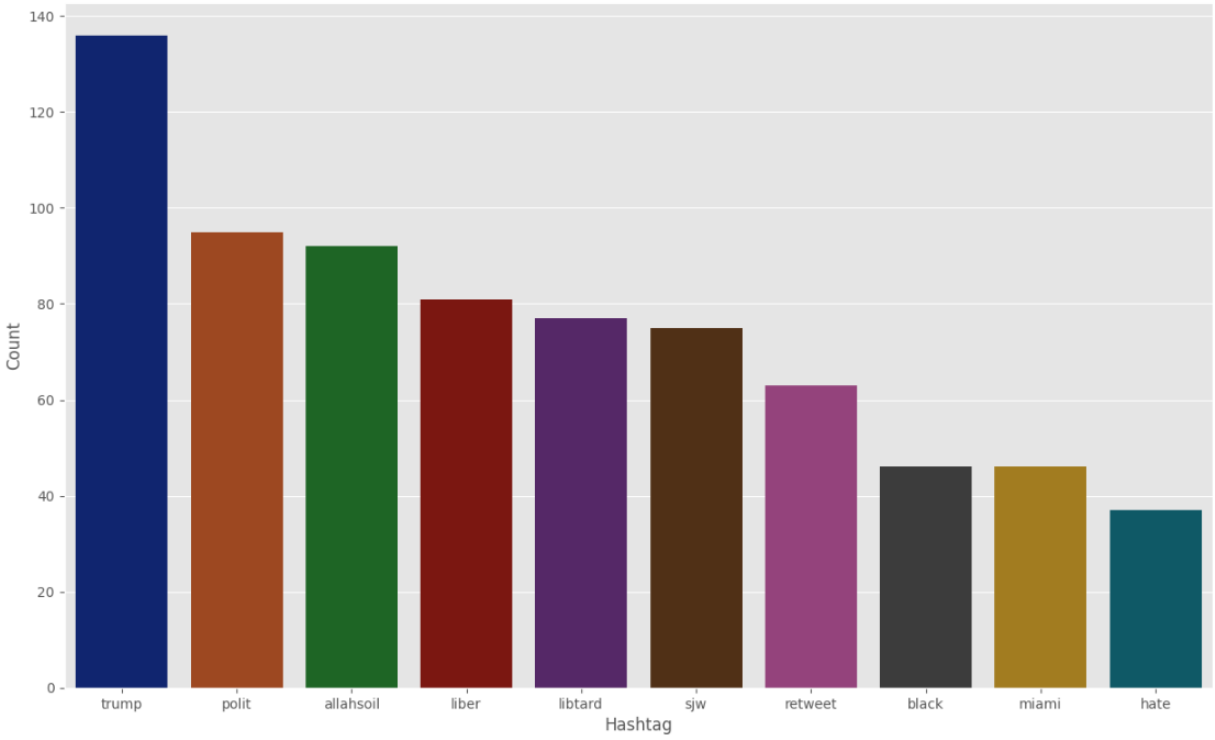
Παρατηρούμε από τον αποτέλεσμα της εικόνας πως πολλα ρατσιστικά/σεξιστικά tweets αφορούν τις πολιτικές πεποιθήσεις (Trump, Obama, litard), το γυναικείο φύλο (woman), εθνικά θέματα (America), φυλετικά θέματα (black) και άλλα. Για την δημιουργία αυτής της εικόνας χρησιμοποιήθηκαν πιο σκούρα και έντονα χρώματα για να ταιριάζουν με το περιεχόμενο.

Συλλογή και οπτικοποίηση hashtags (#)

Στην ανάλυσή μας, επεκτείναμε την εξερεύνηση μας για να συμπεριλάβουμε τη συλλογή των hashtag, αναγνωρίζοντας τη σημασία τους για την κατανόηση των θεμάτων των tweets. Εξάγοντας hashtags τόσο από μη ρατσιστικά/σεξιστικά όσο και από ρατσιστικά/σεξιστικά tweets, στοχεύουμε να διακρίνουμε τα διαδεδομένα θέματα σε κάθε κατηγορία. Χρησιμοποιώντας αυτές τις πληροφορίες, κατασκευάσαμε δύο ραβδογράμματα που παρουσιάζουν τα κορυφαία 10 hashtag που εμφανίζονται πιο συχνά σε κάθε κατηγορία. Αυτές οι απεικονίσεις παρέχουν πολύτιμες πληροφορίες για τα μοντέρνα θέματα και τις διαδεδομένες συζητήσεις στο πλαίσιο του μη ρατσιστικού/σεξιστικού και ρατσιστικού/σεξιστικού λόγου στο Twitter.



Ραβδόγραμμα 10 κορυφέων hashtag σε μη ρατσιστικά/ σεξιστικά tweets



Ραβδόγραμμα 10 κορυφέων hashtag σε ρατσιστικά/ σεξιστικά tweets

Παρατηρούμε ότι η θεματολογία που πραγματεύονται τα παραπάνω hashtags και στις δύο κατηγορίες tweets, συμπίπτει με την θεματολογία που σχολιάσαμε στις wordcloud εικόνες.

Ανάλυση συναισθημάτων (Sentiment Analysis)

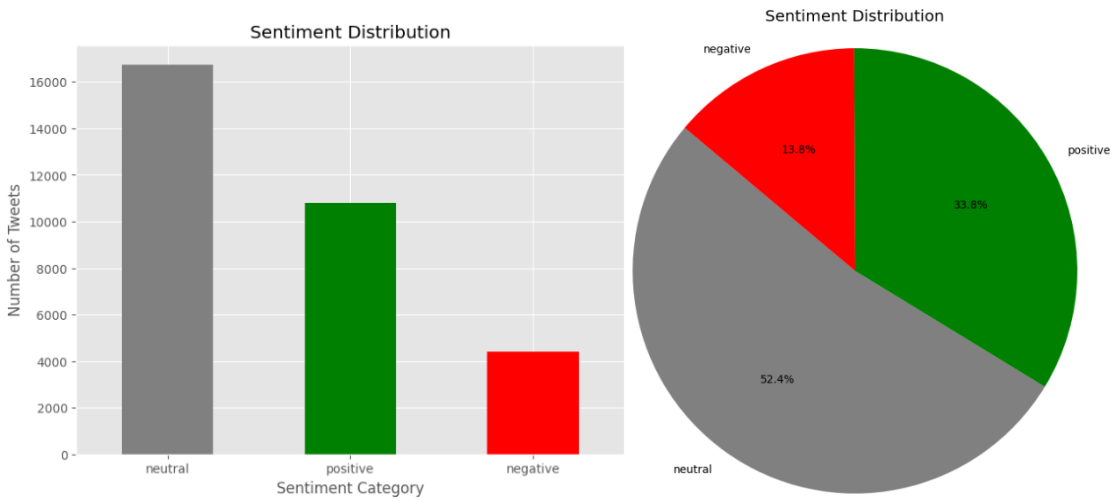
Στην επόμενη φάση της ανάλυσής μας, εμβαθύναμε στην ανάλυση συναισθημάτων για να διακρίνουμε τον συναισθηματικό τόνο κάθε tweet. Αξιοποιώντας τον αναλυτή συναισθήματος VADER από την βιβλιοθήκη Natural Language Toolkit (NLTK), υπολογίσαμε τις βαθμολογίες συναισθήματος (sentiment score) για κάθε tweet, ενσωματώνοντας τα συναισθήματα ως θετικά, ουδέτερα ή αρνητικά. Αυτές οι βαθμολογίες προήλθαν από τη βαθμολογία σύνθετου συναισθήματος, παρέχοντας μια γενική αξιολόγηση της πολικότητας του συναισθήματος. Ενσωματώνοντας σαν στήλες τις βαθμολογίες συναισθήματος (sentiment score) και κατηγορίες (sentiment) στο πλαίσιο δεδομένων μας, προσπαθούμε να αποκτήσαμε βαθύτερες γνώσεις σχετικά με τα κυρίαρχα συναισθήματα που εκφράζονται στο σύνολο δεδομένων Twitter. Αυτό το βήμα χρησιμεύει ως βασικό συστατικό για την κατανόηση του συναισθήματος και τον εντοπισμό προτύπων στο συναισθηματικό περιεχόμενο των tweets.

```
nltk.download('vader_lexicon')
from nltk.sentiment.vader import SentimentIntensityAnalyzer

# Initialize the VADER sentiment analyzer
sid = SentimentIntensityAnalyzer()

# Perform sentiment analysis on each tweet and extract sentiment scores
sentiment_scores = df['clean_tweet'].apply(lambda x: sid.polarity_scores(x))

# Extract compound sentiment scores (overall sentiment)
compound_scores = sentiment_scores.apply(lambda x: x['compound'])

# Add sentiment score as a new column to the dataframe
df['sentiment_score'] = compound_scores

# Categorize tweets as positive, neutral, or negative based on compound scores
sentiment_categories = compound_scores.apply(lambda x: 'positive' if x > 0 else ('neutral' if x == 0 else 'negative'))

# Add sentiment categories as a new column to the dataframe
df['sentiment'] = sentiment_categories
df.head(10)
```

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
```

	id	label	tweet	length	clean_tweet	clean_length	sentiment_score	sentiment
0	1	0	@user when a father is dysfunctional and is s...	102	when father dysfunct selfish drag kid into dys...	56	-0.6124	negative
1	2	0	@user @user thanks for #lyft credit i can't us...	122	thank #lyft credit caus they offer wheelchair ...	70	0.6249	positive
2	3	0	bihday your majesty	21	bihday your majesti	19	0.0000	neutral
3	4	0	#model i love u take with u all the time in ...	86	#model love take with time	26	0.6369	positive
4	5	0	factsguide: society now #motivation	39	factsguid societi #motiv	24	0.0000	neutral
5	6	0	[2/2] huge fan fare and big talking before the...	116	huge fare talk befor they leav chao disput whe...	74	0.3182	positive
6	7	0	@user camping tomorrow @user @user @user @use...	74	camp tomorrow danni	19	0.0000	neutral
7	8	0	the next school year is the year for exams.ð□□...	143	next school year year exam think about that #s...	106	0.0000	neutral
8	9	0	we won!!! love the land!!! #allin #cavs #champ...	87	love land #allin #cav #champion #cleveland #cl...	59	0.6369	positive
9	10	0	@user @user welcome here ! i'm it's so #gr...	50	welcom here	11	0.0000	neutral

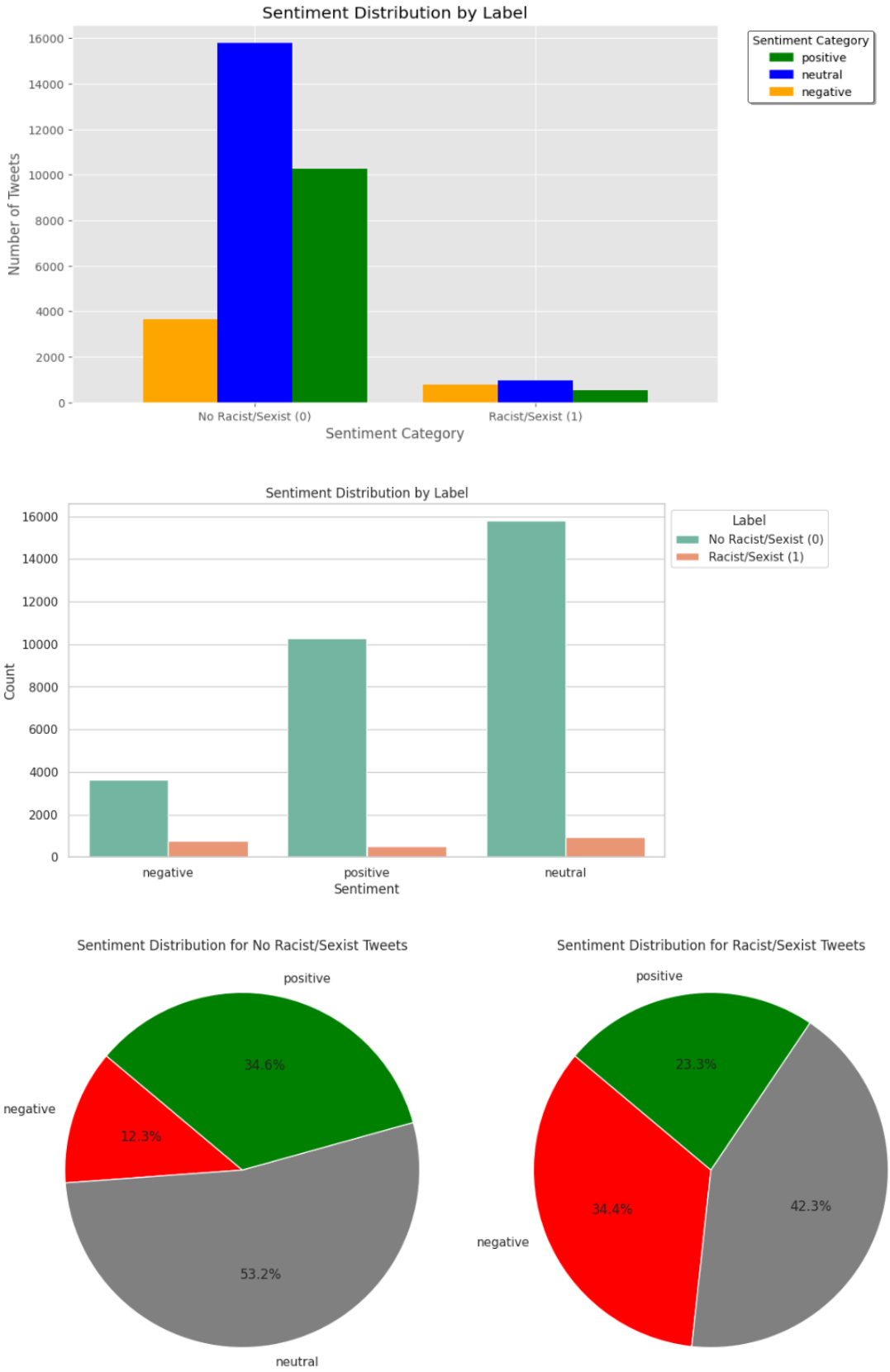
Εικόνα 34. Τμήμα κώδικα για το Sentiment analysis και δείγμα του dataset με τα σχετικά στοιχεία



Διαγράμματα κανατομής συναισθημάτων στο σύνολο του dataset

Η πλειοψηφία των tweet ταξινομούνται ως ουδέτερα, αποτελώντας το 52,4% του συνόλου δεδομένων. Τα θετικά συναισθήματα αποτελούν το 33,8% των tweets, ενώ τα αρνητικά αντιπροσωπεύουν το 13,8%. Αυτή η οπτικοποίηση παρέχει μια επισκόπηση της κατανομής συναισθημάτων, υπογραμμίζοντας την επικράτηση των ουδέτερων συναισθημάτων και τις σχετικές αναλογίες θετικών και αρνητικών συναισθημάτων στο σύνολο δεδομένων.

Εκτός από την ίδια την ανάλυση συναισθήματος, επιδιώξαμε να διερευνήσουμε τη σχέση μεταξύ των βαθμολογιών/κατηγοριών συναισθήματος και των ετικετών που αποδίδονται στα tweets, κάνοντας διάκριση μεταξύ ρατσιστικών/σεξιστικών και μη ρατσιστικών/σεξιστικών tweets. Αυτή η προσπάθεια είχε ως στόχο να διακρίνει τυχόν ευδιάκριτα μοτίβα ή συσχετισμούς μεταξύ του συναισθήματος που εκφράζεται στα tweets και της κατηγοριοποίησής τους ως είτε ρατσιστές/σεξιστές είτε μη ρατσιστές/σεξιστές. Αντιπαραθέτοντας βαθμολογίες/κατηγορίες συναισθημάτων με ετικέτες tweet, στοχεύαμε να αποκαλύψουμε πληροφορίες σχετικά με το πώς ο συναισθηματικός τόνος των tweet μπορεί να ποικίλλει σε διαφορετικές κατηγορίες, ρίχνοντας έτσι φως στην αλληλεπίδραση μεταξύ συναισθήματος και περιεχομένου tweet.



Διαγράμματα κανατομής συναισθημάτων ανά κατηγορία tweet

Για τα tweets που κατηγοριοποιούνται ως μη ρατσιστικά/σεξιστικά, η πλειονότητα των συναισθημάτων εμπίπτει στις κατηγορίες "ουδέτερο" και "θετικό", με 53,2% και 34,6%,

αντίστοιχα. Αυτό δείχνει ότι ένα σημαντικό μέρος αυτών των tweets εμφανίζει ουδέτερο ή θετικό συναίσθημα. Ωστόσο, ένα μικρότερο ποσοστό των tweets (12,3%) εμφανίζει αρνητικό συναίσθημα.

Από την άλλη πλευρά, για τα tweets που κατηγοριοποιούνται ως ρατσιστικά/σεξιστικά, η κατανομή των συναισθημάτων διαφέρει αισθητά. Ενώ η πλειονότητα αυτών των tweets είναι ακόμα ουδέτερα (42,3%), υπάρχει μια σημαντική αύξηση στο ποσοστό των tweets με αρνητικό συναίσθημα (34,4%). Το ποσοστό των tweets με θετικό συναίσθημα είναι σημαντικά χαμηλότερο σε αυτήν την κατηγορία (23,3%) σε σύγκριση με την κατηγορία " μη ρατσιστικά/σεξιστικά".

Συνολικά, αυτά τα αποτελέσματα υποδηλώνουν ότι υπάρχει μια ευδιάκριτη σχέση μεταξύ του συναισθήματος των tweets και της κατηγοριοποίησής τους είτε ως "μη ρατσιστικά/σεξιστικά" είτε ως "ρατσιστικά/σεξιστικά". Τα tweets που χαρακτηρίζονται ρατσιστικά/σεξιστικά τείνουν να έχουν υψηλότερο ποσοστό αρνητικού συναισθήματος σε σύγκριση με τα tweets που είναι μη ρατσιστικά/σεξιστικά, όπου τα ουδέτερα και θετικά συναισθήματα είναι πιο διαδεδομένα.

Το να έχουμε ένα "ρατσιστικό/σεξιστικό" tweet με θετικό συναίσθημα μπορεί να φαίνεται αντιφατικό με την πρώτη ματιά, αλλά είναι σημαντικό να λάβουμε υπόψη ότι οι αλγόριθμοι ανάλυσης συναισθήματος αξιολογούν τον συνολικό συναισθηματικό τόνο του κειμένου χωρίς απαραίτητα να κατανοούν το πλαίσιο ή τη σημασία συγκεκριμένων λέξεων ή φράσεων. Μερικά σενάρια όπου μπορεί να συμβεί αυτή η φαινομενική αντίφαση:

- Σαρκασμός ή ειρωνεία: Το tweet θα μπορούσε να χρησιμοποιήσει σαρκασμό ή ειρωνεία για να μεταφέρει ένα θετικό συναίσθημα σε ένα πλαίσιο που συζητά ρατσιστικά ή σεξιστικά θέματα. Για παράδειγμα, ένα tweet μπορεί να επαινεί σαρκαστικά τη μεροληπτική συμπεριφορά για να τονίσει τον παραλογισμό του.
- Παράθεση ή αναφορά: Το tweet μπορεί να παραθέτει ή να αναφέρει ρατσιστικό ή σεξιστικό περιεχόμενο χωρίς να το εγκρίνει. Σε τέτοιες περιπτώσεις, το συναίσθημα μπορεί να αντικατοπτρίζει τη συνολική θετικότητα του μηνύματος ή της πρόθεσης του tweet, παρά το συναίσθημα του αναφερόμενου περιεχομένου.
- Ενδυνάμωση ή Ανθεκτικότητα: Το tweet θα μπορούσε να μεταφέρει ένα θετικό συναίσθημα παρά τη συζήτηση για εμπειρίες ρατσισμού ή σεξισμού ως μορφή

ενδυνάμωσης, ανθεκτικότητας ή αλληλεγγύης. Μπορεί να επικεντρωθεί στην υπέρβαση των προκλήσεων ή στην υποστήριξη θετικής αλλαγής.

- Παράθεση ή αντίθεση: Το tweet μπορεί να αντιπαραθέτει ρατσιστικό ή σεξιστικό περιεχόμενο με θετικά συναισθήματα για να τονίσει την ανισότητα ή την αντίθεση μεταξύ των δύο. Αυτή η αντιπαράθεση μπορεί να εξυπηρετήσει ρητορικούς ή εκπαιδευτικούς σκοπούς.
- Ανακρίβεια της Ανάλυσης Συναισθήματος: Τέλος, αξίζει να σημειωθεί ότι οι αλγόριθμοι ανάλυσης συναισθήματος, αν και χρήσιμοι, δεν είναι τέλειοι και μπορεί μερικές φορές να ταξινομούν εσφαλμένα συναισθήματα, ειδικά σε πολύπλοκα ή διαφοροποιημένα πλαίσια. Σε τέτοιες περιπτώσεις, το αποτέλεσμα της ανάλυσης συναισθήματος μπορεί να μην αντικατοπτρίζει με ακρίβεια το αληθινό συναίσθημα του κειμένου.

Συνολικά, αν και μπορεί αρχικά να φαίνεται αντιφατικό, η παρουσία ενός "ρατσιστικού/σεξιστικού"tweet με θετικό συναίσθημα υπογραμμίζει την πολυπλοκότητα της γλώσσας και τις προκλήσεις της ερμηνείας του συναισθήματος σε δεδομένα κειμένου πραγματικού κόσμου. Η κατανόηση των συμφραζομένων και η ανθρώπινη κρίση είναι συχνά απαραίτητα για την πλήρη κατανόηση του νοήματος και των συνεπειών τέτοιων tweets.

Ανάλυση συσχέτισης (Correlation Analysis)

Σε συνέχεια της σχέσης του συναισθήματος που έχει ένα tweet και την κατηγοριοποίηση του ως ρατσιστικό/σεξιστικό ή μη, υπολογίσαμε την συσχέτιση μεταξύ του "label" και του "sentiment_score", καθώς αυτό αποτελεί αριθμητική μεταβλητή σε αντίθεση με το sentiment.

```
correlation = df['label'].corr(df['sentiment_score'])

print("Correlation between label and sentiment_score:", correlation)
```

```
Correlation between label and sentiment_score: -0.14339429492462175
```

Εικόνα 35. Τμήμα κώδικα σχετικά με το correlation analysis

Ένας συντελεστής συσχέτισης -0,14 υποδεικνύει μια ασθενή αρνητική συσχέτιση μεταξύ της μεταβλητής "label" (που υποδεικνύει εάν ένα tweet έχει ταξινομηθεί ως

ρατσιστικό/σεξιστικό ή όχι) και της στήλης sentiment_score (υπολογισμένη με την ανάλυση συναισθήματος VADER). Μπορείτε να ερμηνεύσουμε αυτήν τη συσχέτιση ως εξής:

- Δεδομένου ότι ο συντελεστής συσχέτισης είναι αρνητικός, υποδηλώνει ότι όσο αυξάνεται η βαθμολογία συναισθήματος (δηλαδή πιο θετικό συναίσθημα), η πιθανότητα το tweet να ταξινομηθεί ως ρατσιστικό/σεξιστικό μειώνεται ελαφρώς.
- Ωστόσο, το μέγεθος του συντελεστή συσχέτισης είναι κοντά στο 0, υποδηλώνοντας ασθενή συσχέτιση. Αυτό σημαίνει ότι η σχέση μεταξύ των μεταβλητών δεν είναι πολύ ισχυρή και η συσχέτιση μπορεί να μην είναι στατιστικά σημαντική.
- Σε πρακτικούς όρους, αυτός ο συσχετισμός υποδηλώνει ότι υπάρχει ελάχιστη έως καθόλου γραμμική σχέση μεταξύ της βαθμολογίας συναισθήματος και της πιθανότητας ένα tweet να ταξινομηθεί ως ρατσιστικό/σεξιστικό. Άλλοι παράγοντες μπορεί να διαδραματίσουν πιο σημαντικό ρόλο στον καθορισμό του εάν ένα tweet θεωρείται ρατσιστικό/σεξιστικό.

Συνολικά, ενώ υπάρχει μια μικρή τάση για tweets με πιο θετικό συναίσθημα να είναι λιγότερο πιθανό να ταξινομηθούν ως ρατσιστικά/σεξιστικά, αυτή η σχέση είναι αδύναμη και θα πρέπει να ερμηνεύεται με προσοχή.

3.2.4 Εξαγωγή χαρακτηριστικών (Feature Extraction)

Η εξαγωγή χαρακτηριστικών είναι ένα κρίσιμο βήμα στη διαδικασία της επεξεργασίας φυσικής γλώσσας (NLP) και της μηχανικής μάθησης, όπου τα ακατέργαστα δεδομένα κειμένου μετατρέπονται σε μορφή κατάλληλη για ανάλυση και μοντελοποίηση. Σε αυτή τη φάση, οι πληροφορίες κειμένου μετατρέπονται σε ένα σύνολο χαρακτηριστικών ή αριθμητικών αναπαραστάσεων που αποτυπώνουν βασικά χαρακτηριστικά του κειμένου. Αυτά τα χαρακτηριστικά χρησιμεύουν ως μεταβλητές εισόδου για αλγόριθμους μηχανικής μάθησης, δίνοντάς τους τη δυνατότητα να μαθαίνουν μοτίβα, να κάνουν προβλέψεις και να εξάγουν πληροφορίες από τα δεδομένα κειμένου. Οι τεχνικές εξαγωγής χαρακτηριστικών περιλαμβάνουν μια σειρά μεθόδων, συμπεριλαμβανομένης της διανυσματοποίησης (vectorization), του μετασχηματισμού (transformation) και της μείωσης διαστάσεων (dimensionality reduction), προσαρμοσμένες στις ειδικές απαιτήσεις της εργασίας NLP. Εξάγοντας σημαντικά χαρακτηριστικά από δεδομένα

κειμένου, μπορούμε να αναπαραστήσουμε και να αναλύσουμε αποτελεσματικά τις πληροφορίες κειμένου, διευκολύνοντας εργασίες.

Μετατρέψαμε τα καθαρισμένα δεδομένα κειμένου σε αριθμητικά χαρακτηριστικά χρησιμοποιώντας τις τεχνικές:

- Bag-of-Words (BoW)
- TF-IDF (Term Frequency-Inverse Document Frequency)

Bag-of-Words (BoW): Αυτή η τεχνική αντιπροσωπεύει δεδομένα κειμένου μετρώντας τη συχνότητα εμφάνισης κάθε λέξης σε ένα έγγραφο χωρίς να λαμβάνεται υπόψη η σειρά με την οποία εμφανίζονται. Δημιουργεί ένα λεξιλόγιο μοναδικών λέξεων που υπάρχουν στο σώμα και κατασκευάζει ένα διάνυσμα χαρακτηριστικών για κάθε έγγραφο, όπου κάθε διάσταση αντιστοιχεί σε μια λέξη στο λεξιλόγιο και η τιμή αντιπροσωπεύει τη συχνότητα αυτής της λέξης στο έγγραφο. Το BoW είναι απλό και αποτελεσματικό, αλλά αγνοεί τη σημασιολογία και το πλαίσιο των λέξεων. (McTear, Callejas & Griol, 2016)

```
[70] from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer
```

1. Bag-of-Words (BoW) Features

```
[71] # Initialize CountVectorizer
     vectorizer = CountVectorizer(max_features=1000, ngram_range=(1, 2), stop_words='english')

     # Fit and transform the cleaned text data
     bow = vectorizer.fit_transform(df['clean_tweet'])

     bow.shape
```

```
(31962, 1000)
```

```
# Print feature names
print("Feature names:", vectorizer.get_feature_names_out())
```

```
Feature names: ['abl' 'absolut' 'accept' 'account' 'act' 'action' 'actor' 'actual'
 'adapt' 'adapt environ' 'adventur' 'affirm' 'afternoon' 'agre' 'ahead'
 'aist' 'album' 'aliv' 'allahsoil' 'allow' 'alon' 'alreadi' 'altwaystoh'
 'altwaystoh healthi' 'alway' 'amaz' 'america' 'american' 'angel' 'angri'
 'angri polar' 'anim' 'anniversari' 'announc' 'anoth' 'answer' 'anti'
 'anxieti' 'anymor' 'anyon' 'anyth' 'appl' 'applic' 'appreci' 'aren'
 'arriv' 'asian' 'ask' 'attack' 'attack bull' 'august' 'avail' 'award'
 'away' 'awesom' 'babi' 'balanc' 'ball' 'bday' 'beach' 'bear' 'bear climb'
 'bear live' 'beat' 'beauti' 'beauti followm' 'becaus' 'becom' 'beer'
 'befor' 'begin' 'believ' 'best' 'best friend' 'bestfriend' 'better'
 'bihday' 'bihday bihday' 'bike' 'bing' 'bing bong' 'bird' 'bitch' 'black'
 'blame' 'bless' 'block' 'blog' 'blog silver' 'blogger' 'blond' 'blue'
 'blur' 'blur sun' 'bodi' 'bong' 'bong bing' 'book' 'bore' 'bought' 'boy'
 'boyfriend' 'brand' 'break' 'breakfast' 'brexit' 'bride' 'bring' 'broken'
 'brother' 'buffalo' 'buffalo simul' 'buffalo vicin' 'build' 'bull'
 'bull chase' 'bull direct' 'bull domin' 'bull game' 'bull hill' 'busi'
 'cake' 'calm' 'came' 'camp' 'campaign' 'cantwait' 'card' 'care' 'case'
 'cat' 'caus' 'celebr' 'challeng' 'chanc' 'chang' 'chase' 'chase leav'
 'check' 'cheer' 'child' 'chill' 'choic' 'choos' 'christian' 'christma'
```

Εικόνα 36. Τμήμα κώδικα σχετικά με το BoW και κομμάτι του δημιουργημένου του λεξιλογίου

TF-IDF (Term Frequency-Inverse Document Frequency): Το TF-IDF είναι ένα στατιστικό μέτρο που χρησιμοποιείται για την αξιολόγηση της σημασίας μιας λέξης σε

ένα έγγραφο σε σχέση με ένα σώμα. Λαμβάνει υπόψη τόσο τη συχνότητα ενός όρου σε ένα έγγραφο (Term Frequency, TF) όσο και τη σπανιότητα του όρου στο σώμα (Inverse Document Frequency, IDF). Το TF-IDF εκχωρεί υψηλότερα βάρη σε όρους που είναι συχνοί στο έγγραφο αλλά σπάνιοι στο σώμα, βοηθώντας στον εντοπισμό βασικών όρων που είναι διακριτοί για ένα συγκεκριμένο έγγραφο. Αντιμετωπίζει τους περιορισμούς του BoW δίνοντας μεγαλύτερη σημασία σε λέξεις που είναι πιο ενημερωτικές και λιγότερο κοινές στα έγγραφα. (Rajaraman & Ullman, 2011)

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

```
[73] # Initialize TfidfVectorizer
     tfidf_vectorizer = TfidfVectorizer(max_features=1000, ngram_range=(1, 2), stop_words='english')

     # Fit and transform the cleaned text data
     tfidf = tfidf_vectorizer.fit_transform(df['clean_tweet'])

     tfidf.shape

     (31962, 1000)
```

```
# Print feature names
print("TF-IDF Feature names:", tfidf_vectorizer.get_feature_names_out())

TF-IDF Feature names: ['abl' 'absolut' 'accept' 'account' 'act' 'action' 'actor' 'actual'
 'adapt' 'adapt environ' 'adventur' 'affirm' 'afternoon' 'agre' 'ahead'
 'aist' 'album' 'aliv' 'allahsoil' 'allow' 'alon' 'alreadi' 'altwaystoh'
 'altwaystoh healthi' 'alway' 'amaz' 'america' 'american' 'angel' 'angri'
 'angri polar' 'anim' 'anniversari' 'announc' 'anoth' 'answer' 'anti'
 'anxieti' 'anymor' 'anyon' 'anyth' 'appl' 'applic' 'appreci' 'aren'
 'arriv' 'asian' 'ask' 'attack' 'attack bull' 'august' 'avail' 'award'
 'away' 'awesom' 'babi' 'balanc' 'ball' 'bday' 'beach' 'bear' 'bear climb'
 'bear live' 'beat' 'beauti' 'beauti followm' 'becaus' 'becom' 'beer'
 'befor' 'begin' 'believ' 'best' 'best friend' 'bestfriend' 'better'
 'bihday' 'bihday bihday' 'bike' 'bing' 'bing bong' 'bird' 'bitch' 'black'
 'blame' 'bless' 'block' 'blog' 'blog silver' 'blogger' 'blond' 'blue'
```

Εικόνα 37. Τμήμα κώδικα σχετικά με το TF-IDF και κομμάτι του δημιουργημένου του λεξιλογίου

Και στις δυο εφαρμογές έχουμε λάβει υπόψη μόνο τις κορυφαίες 1000 λέξεις που εμφανίζονται πιο συχνά στο σώμα (corpus). Αυτό βοηθά στη μείωση της διάστασης του χώρου των χαρακτηριστικών, ο οποίος μπορεί να είναι χρήσιμος για τη διαχείριση της μνήμης και των υπολογιστικών πόρων, ειδικά όταν πρόκειται για μεγάλα σύνολα δεδομένων κειμένου.

3.2.5 Εκπαίδευση Μοντέλου (Model Training)

Σε αυτό το στάδιο εμβαθύνουμε στη διαδικασία δημιουργίας μοντέλων μηχανικής μάθησης για να προβλέψουμε εάν ένα tweet περιέχει ρατσιστικό/σεξιστικό περιεχόμενο με βάση τα εξαγόμενα χαρακτηριστικά από τα δεδομένα κειμένου. Αυτή η φάση περιλαμβάνει την επιλογή κατάλληλων αλγορίθμων, την εκπαίδευση των μοντέλων σε

δεδομένα και την αξιολόγηση της απόδοσής τους. Μέσω αυτής της επαναληπτικής διαδικασίας, στοχεύουμε να αναπτύξουμε μοντέλα που μπορούν να ταξινομήσουν αποτελεσματικά τα tweets σε σχετικές κατηγορίες, συμβάλλοντας στον ευρύτερο στόχο του εντοπισμού της ρητορικής μίσους και της ανάλυσης συναισθημάτων στα δεδομένα των μέσων κοινωνικής δικτύωσης.

Τόσο για το Bag of words όσο και για το TFIDF, εκτελούμε 4 αλγόριθμους ταξινόμησης κατάλληλους για δεδομένα κειμένου, δηλαδή:

1. Logistic Regression
2. Naive Bayes
3. Support Vector Machines (SVM)
4. Random Forest

Αρχικά διαχωρίζουμε τα δεδομένα σε σετ εκπαίδευσης (training set) και δοκιμών (testing set) ξεχωριστά για αναπαραστάσεις Bag-of-Words (BoW) και TF-IDF χρησιμοποιώντας τη συνάρτηση train_test_split. Αυτό το βήμα διασφαλίζει ότι έχουμε ξεχωριστά σύνολα δεδομένων για εκπαίδευση και αξιολόγηση των μοντέλων. Από το σύνολο δεδομένων μας το 80% το ορίζουμε σαν training set και το υπολοιπο 20% ως το testing set.

1. Split the Data

Split the dataset into training and testing sets.

```
[76] from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.naive_bayes import MultinomialNB
     from sklearn.svm import SVC
     from sklearn.ensemble import RandomForestClassifier
     from sklearn.metrics import accuracy_score

     # Split the data into training and testing sets for BoW
     X_train_bow, X_test_bow, y_train, y_test = train_test_split(bow, df['label'], test_size=0.2, random_state=42)

     # Split the data into training and testing sets for TF-IDF
     X_train_tfidf, X_test_tfidf, y_train, y_test = train_test_split(tfidf, df['label'], test_size=0.2, random_state=42)
```

Εικόνα 38. Τμήμα κώδικα σχετικά με τον διαχωρισμό του dataset

Συνεχίζουμε με την εκπαίδευση και αξιολόγηση μοντέλου. Για κάθε μοντέλο, εκπαιδεύει το μοντέλο στα δεδομένα εκπαίδευσης (τόσο με BoW όσο TF-IDF) και αξιολογεί την απόδοσή του χρησιμοποιώντας τη συνάρτηση train_and_evaluate_model και την ακρίβεια (accuracy) ως μέτρηση αξιολόγησης.

```
[81] # Train and evaluate Model
     def train_and_evaluate_model(model, X_train, X_test, y_train, y_test):
         model.fit(X_train, y_train)
         predictions = model.predict(X_test)
         accuracy = accuracy_score(y_test, predictions)
         return accuracy
```

```
[79] # Initialize models
     logistic_regression = LogisticRegression()
     naive_bayes = MultinomialNB()
     svm = SVC()
     random_forest = RandomForestClassifier()
```

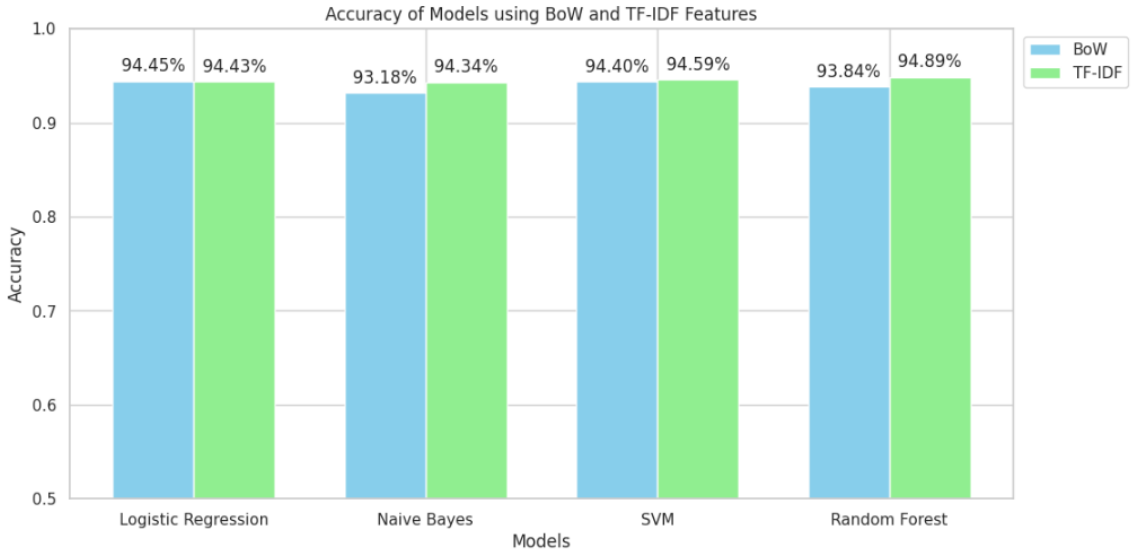
Train and evaluate models using BoW features

```
[80] accuracy_lr_bow = train_and_evaluate_model(logistic_regression, X_train_bow, X_test_bow, y_train, y_test)
     accuracy_nb_bow = train_and_evaluate_model(naive_bayes, X_train_bow, X_test_bow, y_train, y_test)
     accuracy_svm_bow = train_and_evaluate_model(svm, X_train_bow, X_test_bow, y_train, y_test)
     accuracy_rf_bow = train_and_evaluate_model(random_forest, X_train_bow, X_test_bow, y_train, y_test)

     accuracies_bow=[accuracy_lr_bow, accuracy_nb_bow, accuracy_svm_bow, accuracy_rf_bow]
```

Train and evaluate models using TF-IDF features

```
[84] accuracy_lr_tfidf = train_and_evaluate_model(logistic_regression, X_train_tfidf, X_test_tfidf, y_train, y_test)
     accuracy_nb_tfidf = train_and_evaluate_model(naive_bayes, X_train_tfidf, X_test_tfidf, y_train, y_test)
     accuracy_svm_tfidf = train_and_evaluate_model(svm, X_train_tfidf, X_test_tfidf, y_train, y_test)
     accuracy_rf_tfidf = train_and_evaluate_model(random_forest, X_train_tfidf, X_test_tfidf, y_train, y_test)

     accuracies_tfidf=[accuracy_lr_tfidf, accuracy_nb_tfidf, accuracy_svm_tfidf, accuracy_rf_tfidf]
```

Εικόνα 39. Τμήμα κώδικα σχετικά με την εκπαίδευση και αξιολόγηση των μοντέλων



Ραβδόγραμμα σχετικά με το accuracy σε όλα τα μοντέλα

Συνολικά, αυτό το τμήμα κώδικα προετοιμάζει τα δεδομένα, αρχικοποιεί μοντέλα ταξινόμησης, τα εκπαιδεύει και αξιολογεί την απόδοσή τους χρησιμοποιώντας την ακρίβεια ως μέτρηση. Μέχρι τώρα την μεγαλύτερη ακρίβεια φαίνεται να την έχει το μοντέλο Random Forest TF-IDF.

‘Επειτα αναλύουμε το confusion matrix για το κάθε μοντέλο. Ένας πίνακας confusion matrix είναι ένα χρήσιμο εργαλείο σε εργασίες ταξινόμησης, για την αξιολόγηση της απόδοσης ενός μοντέλου μηχανικής μάθησης. Στην παρούσα εργασία, το confusion matrix μας βοηθά να κατανοήσουμε πόσο καλά αποδίδει το μοντέλο στην ταξινόμηση των tweets είτε ως ρατσιστικά/σεξιστικά είτε ως μη ρατσιστικά/σεξιστικά. Οργανώνει τις προβλέψεις σε τέσσερις κατηγορίες:

1. **True Positive (TP):** Περιπτώσεις όπου το μοντέλο προβλέπει σωστά ένα tweet ως ρατσιστικό/σεξιστικό.
2. **True Negative (TN):** Περιπτώσεις όπου το μοντέλο προβλέπει σωστά ένα tweet ως μη ρατσιστικό/σεξιστικό.
3. **False Positive (FP):** Περιπτώσεις όπου το μοντέλο προβλέπει εσφαλμένα ένα μη ρατσιστικό/σεξιστικό tweet ως ρατσιστικό/σεξιστικό (σφάλμα τύπου I).
4. **False Negative (FN):** Περιπτώσεις όπου το μοντέλο προβλέπει εσφαλμένα ένα ρατσιστικό/σεξιστικό tweet ως μη ρατσιστικό/σεξιστικό (σφάλμα τύπου II).

Αναλύοντας τον confusion matrix πίνακα, μπορούμε να υπολογίσουμε διάφορες μετρήσεις αξιολόγησης, όπως accuracy, precision, recall και F1 score. Αυτές οι μετρήσεις παρέχουν πληροφορίες για την απόδοση του μοντέλου και βοηθούν στον εντοπισμό περιοχών προς βελτίωση. Επομένως, η χρήση ενός πίνακα σύγχυσης είναι απαραίτητη για την αξιολόγηση της αποτελεσματικότητας και της αξιοπιστίας του μοντέλου.

Confusion Matrix για μοντέλα εκπαιδευμένα σε χαρακτηριστικά Bag-of-Words (BoW)

Confusion Matrix - Logistic Regression (Bow)

Ο πίνακας σύγχυσης για το μοντέλο Logistic Regression που έχει εκπαιδευτεί σε τεχνική Bag-of-Words (BoW) παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 225
- True Negatives (TN): 5761
- False Positives (FP): 176
- False Negatives (FN): 231

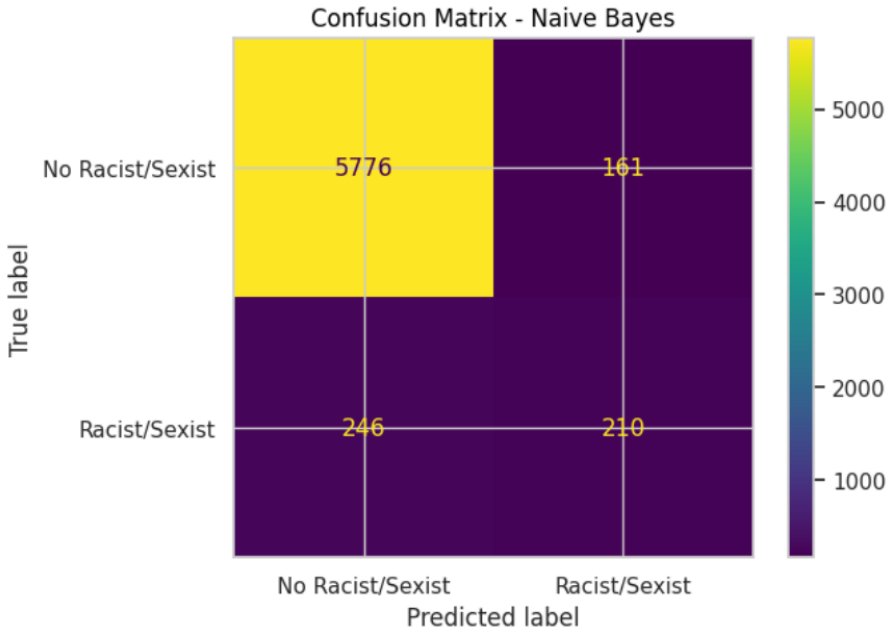
Εικόνα 40. Confusion Matrix - Logistic Regression (Bow)

Αυτές οι μετρήσεις δείχνουν ότι το μοντέλο εντόπισε σωστά 225 περιπτώσεις ρατσιστικών/σεξιστικών tweet και 5761 περιπτώσεων μη ρατσιστικών/σεξιστικών tweets. Ωστόσο, χαρακτήρισε εσφαλμένα 176 μη ρατσιστικά/σεξιστικά tweets ως ρατσιστικά/σεξιστικά (σφάλμα τύπου I) και 231 ρατσιστικά/σεξιστικά tweets ως μη ρατσιστικά/σεξιστικά (σφάλμα τύπου II). Αυτές οι πληροφορίες είναι ζωτικής σημασίας για την κατανόηση των δυνατών και των αδυναμιών του μοντέλου Logistic Regression στην ταξινόμηση των tweets.

Confusion Matrix – Naive Bayes (Bow)

Ο πίνακας σύγχυσης για το μοντέλο Naïve Bayes που έχει εκπαιδευτεί σε χαρακτηριστικά Bag-of-Words (BoW) παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 210
- True Negatives (TN): 5776
- False Positives (FP): 161
- False Negatives (FN): 246

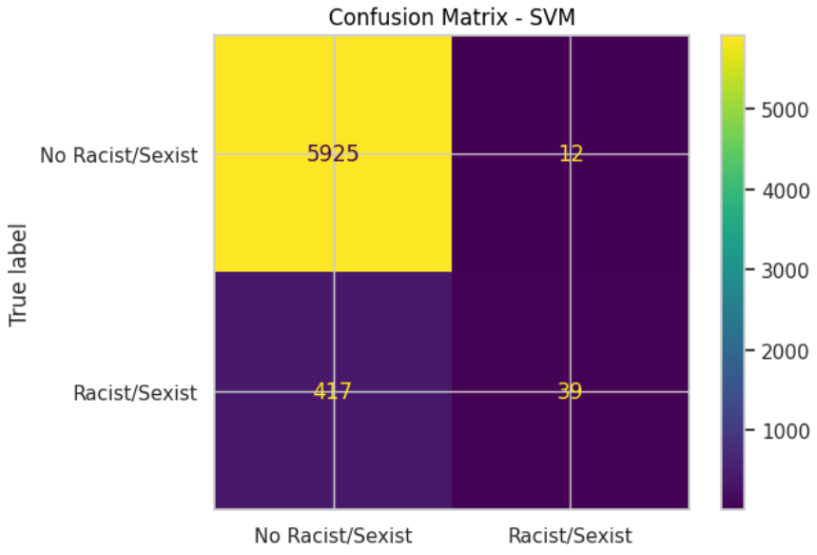


Εικόνα 41. Confusion Matrix - Naive Bayes (Bow)

Confusion Matrix – SVM (Bow)

Ο πίνακας σύγχυσης για το μοντέλο SVM που έχει εκπαιδευτεί σε χαρακτηριστικά Bag-of-Words (BoW) παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 39
- True Negatives (TN): 5925
- False Positives (FP): 12
- False Negatives (FN): 417

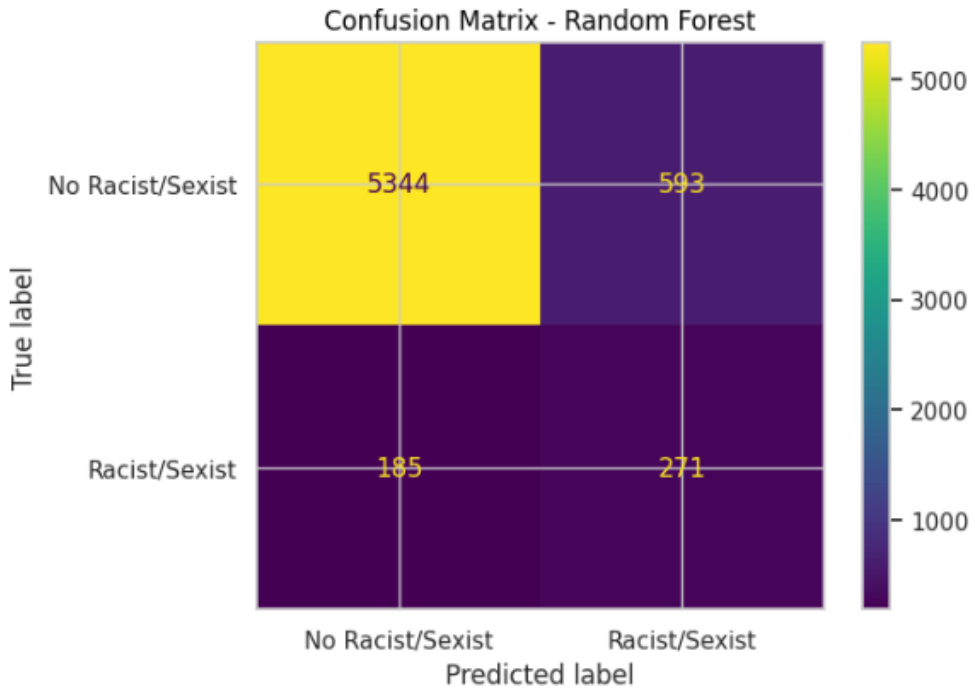


Εικόνα 42. Confusion Matrix – SVM (Bow)

Confusion Matrix – Random Forest (Bow)

Ο πίνακας σύγχυσης για το μοντέλο Random Forest που έχει εκπαιδευτεί σε χαρακτηριστικά Bag-of-Words (BoW) παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 271
- True Negatives (TN): 5344
- False Positives (FP): 593
- False Negatives (FN): 185



Εικόνα 43. Confusion Matrix – Random Forest (Bow)

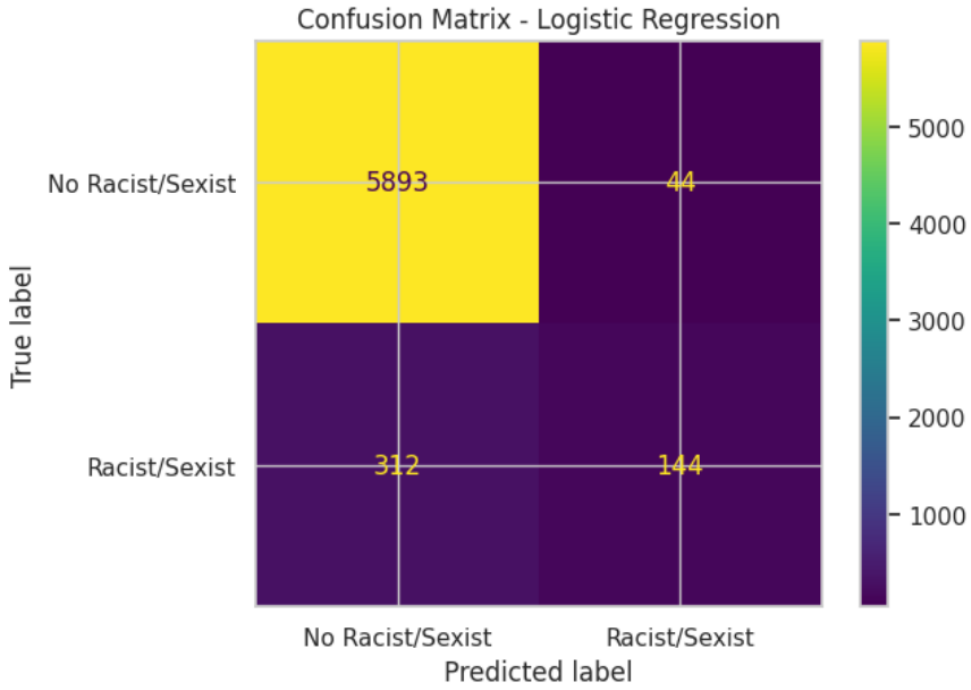
Η Logistic Regression έχει τον υψηλότερο αριθμό True Positives (TP: 225) και το Random Forest έχει τον χαμηλότερο αριθμό False Negatives (FN: 185) σε σύγκριση με την Logistic Regression (FN: 231). Επομένως, με βάση τον συνδυασμό αληθινών θετικών και ψευδών αρνητικών, το Random Forest αναδεικνύεται ως το προτιμώμενο μοντέλο μεταξύ των αξιολογηθέντων με βάση το σύνολο χαρακτηριστικών Bow. Ωστόσο, μια ολοκληρωμένη αξιολόγηση με χρήση πρόσθετων μετρήσεων είναι απαραίτητη για ένα οριστικό συμπέρασμα.

Confusion Matrix για μοντέλα εκπαιδευμένα σε χαρακτηριστικά TF-IDF

Confusion Matrix - Logistic Regression (TF-IDF)

Ο πίνακας σύγχυσης για το μοντέλο Logistic Regression που έχει εκπαιδευτεί σε χαρακτηριστικά TF-IDF παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 144
- True Negatives (TN): 5893
- False Positives (FP): 44
- False Negatives (FN): 312

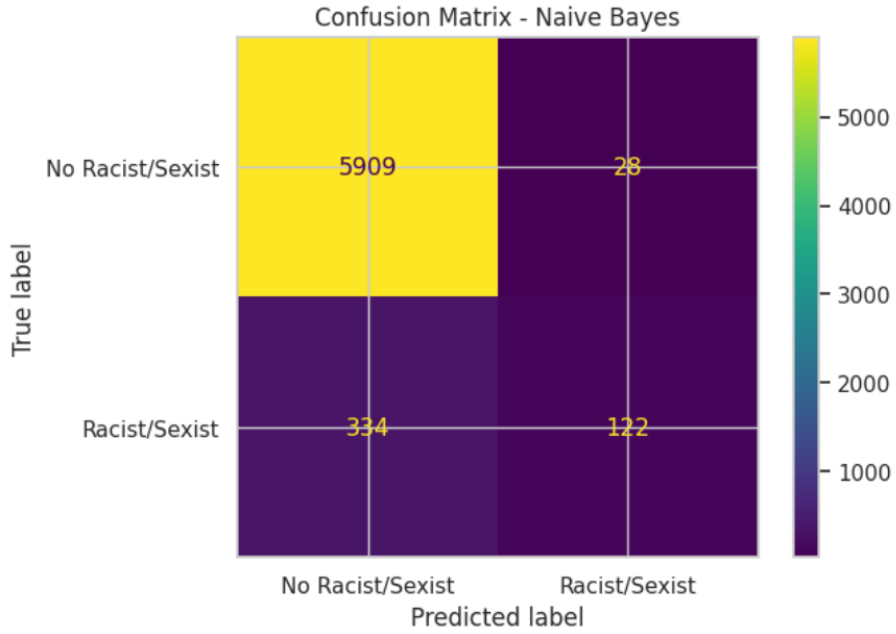


Εικόνα 44. Confusion Matrix - Logistic Regression (TF-IDF)

Confusion Matrix – Naive Bayes (TF-IDF)

Ο πίνακας σύγχυσης για το μοντέλο Naïve Bayes που έχει εκπαιδευτεί σε χαρακτηριστικά TF-IDF παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 122
- True Negatives (TN): 5909
- False Positives (FP): 28
- False Negatives (FN): 334

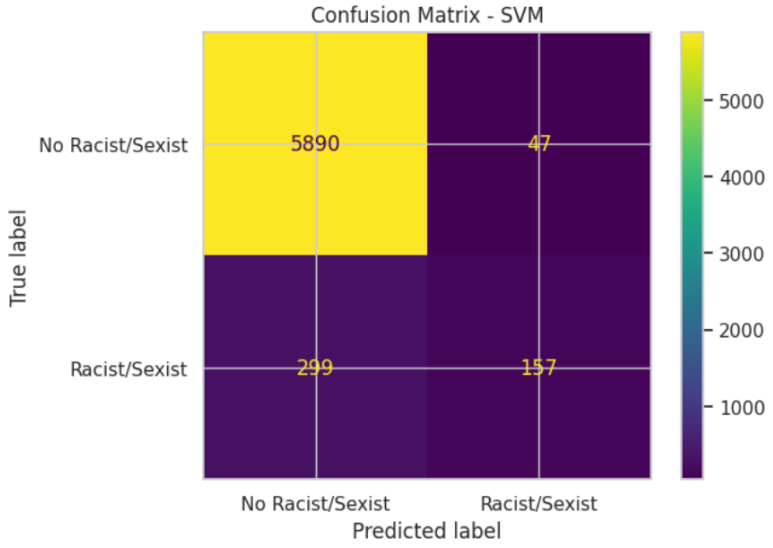


Εικόνα 45. Confusion Matrix - Naive Bayes (TF-IDF)

Confusion Matrix – SVM (TF-IDF)

Ο πίνακας σύγχυσης για το μοντέλο SVM που έχει εκπαιδευτεί σε χαρακτηριστικά TF-IDF παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 157
- True Negatives (TN): 5890
- False Positives (FP): 47
- False Negatives (FN): 299

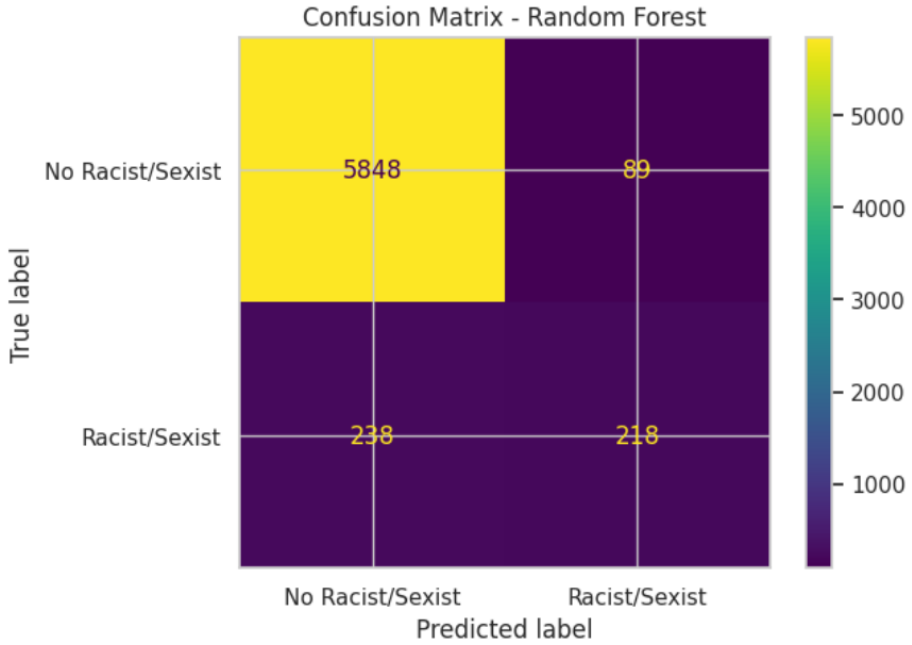


Εικόνα 46. Confusion Matrix – SVM (TF-IDF)

Confusion Matrix – Random Forest (TF-IDF)

Ο πίνακας σύγχυσης για το μοντέλο Random Forest που έχει εκπαιδευτεί σε χαρακτηριστικά TF-IDF παρέχει πολύτιμες πληροφορίες για την απόδοσή του:

- True Positives (TP): 218
- True Negatives (TN): 5848
- False Positives (FP): 89
- False Negatives (FN): 238



Εικόνα 47. Confusion Matrix – Random Forest (TF-IDF)

Αναλύοντας τις μετρήσεις απόδοσης για κάθε μοντέλο, το Random Forest (TF-IDF) ξεχωρίζει με τον υψηλότερο αριθμό αληθινών θετικών (TP: 218) και σχετικά χαμηλότερα ψευδώς αρνητικά (FN: 238), υποδεικνύοντας την αποτελεσματικότητά του στον σωστό εντοπισμό ρατσιστικών/σεξιστικών tweets .

Υπολογισμός μετρικών αξιολόγησης

Για να επιλέξουμε το καλύτερο μοντέλο για αυτό το θέμα, συνήθως θέλουμε να επιλέξουμε αυτό που έχει την καλύτερη απόδοση σύμφωνα με τη μέτρηση αξιολόγησης που σχετίζεται περισσότερο με το πρόβλημά μας. Ακολουθεί μια σύντομη εξήγηση κάθε μέτρησης:

1. **Accuracy**: Η αναλογία των σωστά ταξινομημένων περιπτώσεων επί του συνόλου των περιπτώσεων. Είναι μια καλή μέτρηση όταν η κατανομή κλάσεων είναι ισορροπημένη.

$$accuracy = \frac{true\ positives + true\ negatives}{true\ positives + true\ negatives + false\ negatives + false\ positives}$$

Εικόνα 48. Τύπος υπολογισμού accuracy

2. **Precision**: Το ποσοστό των αληθινών θετικών προβλέψεων από όλες τις θετικές προβλέψεις που έγιναν από το μοντέλο. Είναι μια καλή μέτρηση όταν η τιμή των ψευδώς θετικών είναι υψηλό.

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Εικόνα 49. Τύπος υπολογισμού precision

3. **Recall**: Το ποσοστό των αληθινών θετικών προβλέψεων από όλες τις πραγματικές θετικές περιπτώσεις στα δεδομένα. Είναι μια καλή μέτρηση όταν η τιμή των ψευδώς αρνητικών είναι υψηλό.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Εικόνα 50. Τύπος υπολογισμού recall

4. **F1-score**: Ο αρμονικός μέσος της ακρίβειας και της ανάκλησης. Παρέχει μια ισορροπία μεταξύ ακρίβειας (precision) και ανάκλησης (recall).

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Εικόνα 51. Τύπος υπολογισμού F1-score

Με βάση τις συγκεκριμένες απαιτήσεις μας και τα χαρακτηριστικά των δεδομένων μας, ενδέχεται να δώσουμε προτεραιότητα σε μία μέτρηση έναντι των άλλων. Για παράδειγμα:

- Εάν θέλουμε να ελαχιστοποιήσουμε τα ψευδώς θετικά (δηλαδή, να προσδιορίσουμε εσφαλμένα ένα tweet χωρίς ρητορική μίσους ως ρητορική μίσους), μπορεί να δώσουμε προτεραιότητα στο precision.

- Εάν θέλουμε να ελαχιστοποιήσουμε τα ψευδώς αρνητικά (δηλαδή, να προσδιορίσουμε εσφαλμένα ένα tweet ρητορικής μίσους ως μη ρητορική μίσους), μπορεί να δώσουμε προτεραιότητα στο recall.
- Εάν θέλουμε μια ισορροπία μεταξύ ακρίβειας και ανάκλησης, μπορεί να δώσουμε προτεραιότητα στο F1-score.

```
Logistic Regression (BoW) Metrics:
  Accuracy: 0.9445
  Precision: 0.7463
  Recall: 0.3355
  F1-score: 0.4629

Logistic Regression (TF-IDF) Metrics:
  Accuracy: 0.9443
  Precision: 0.7660
  Recall: 0.3158
  F1-score: 0.4472

Naive Bayes (BoW) Metrics:
  Accuracy: 0.9318
  Precision: 0.5229
  Recall: 0.5000
  F1-score: 0.5112

Naive Bayes (TF-IDF) Metrics:
  Accuracy: 0.9434
  Precision: 0.8133
  Recall: 0.2675
  F1-score: 0.4026

SVM (BoW) Metrics:
  Accuracy: 0.9440
  Precision: 0.7579
  Recall: 0.3158
  F1-score: 0.4458

SVM (TF-IDF) Metrics:
  Accuracy: 0.9459
  Precision: 0.7696
  Recall: 0.3443
  F1-score: 0.4758

Random Forest (BoW) Metrics:
  Accuracy: 0.9393
  Precision: 0.5766
  Recall: 0.5614
  F1-score: 0.5689

Random Forest (TF-IDF) Metrics:
  Accuracy: 0.9506
  Precision: 0.7381
  Recall: 0.4759
  F1-score: 0.5787
```

Εικόνα 52. Μετρικές αξιολόγησεις για όλα τα μοντέλα

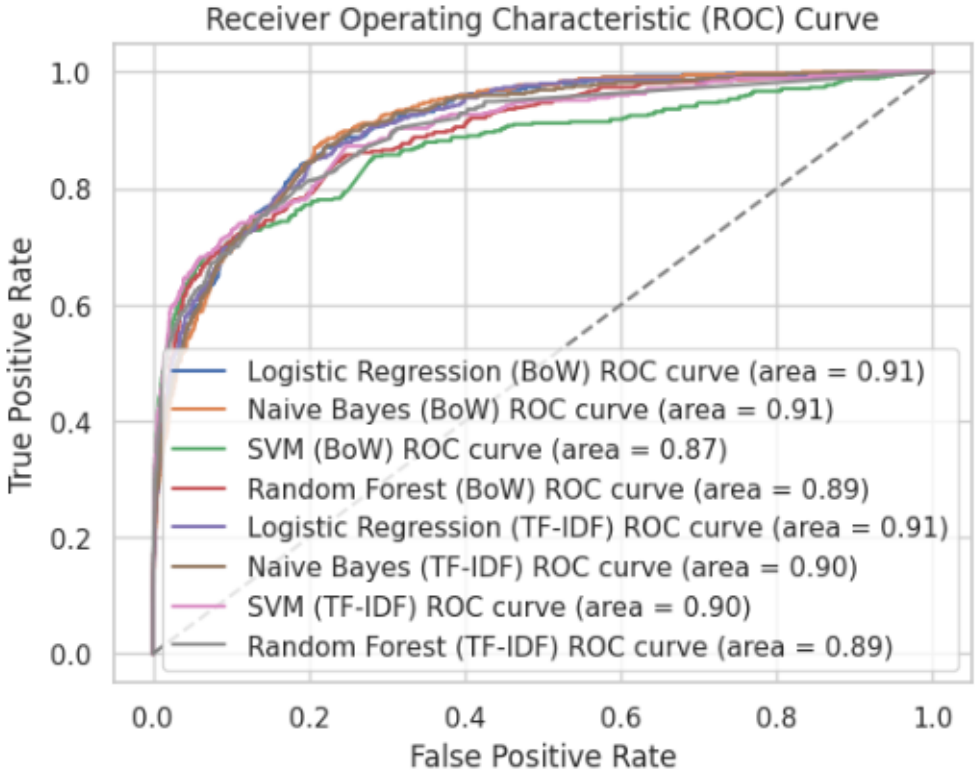
Με βάση αυτές τις μετρήσεις, η καλύτερη επιλογή μπορεί να διαφέρει ανάλογα με τα συγκεκριμένα κριτήρια που δίνουμε προτεραιότητα. Ακολουθούν ορισμένες σκέψεις:

- Accuracy: Το Random Forest (TF-IDF) έχει την υψηλότερη ακρίβεια 0,9512.
- Precision: Το Naive Bayes (TF-IDF) έχει την υψηλότερη ακρίβεια 0,8133.
- Recall: Το Random Forest (TF-IDF) έχει την υψηλότερη ανάκληση 0,4781.
- Score F1: Το Random Forest (TF-IDF) έχει την υψηλότερη βαθμολογία F1 0,5829.

Εάν δίνουμε προτεραιότητα στην ακρίβεια, το Random Forest με χαρακτηριστικά TF-IDF μπορεί να είναι η καλύτερη επιλογή. Ωστόσο, εάν δίνουμε προτεραιότητα στο precision, το Naive Bayes με χαρακτηριστικά TF-IDF μπορεί να είναι προτιμότερο. Ομοίως, εάν δίνετε προτεραιότητα στην ανάκληση ή τη βαθμολογία F1, το Random Forest με χαρακτηριστικά TF-IDF φαίνεται να είναι η καλύτερη επιλογή.

3.2.5 Fine Tuning

Η εκτέλεση Fine-tuning με τον αλγόρυθμο Grid Search είναι μια καλή επιλογή, επειδή μας επιτρέπει να εξερευνούμε συστηματικά μια σειρά υπερπαραμέτρων για κάθε μοντέλο και να επιλέγουμε τον συνδυασμό που μεγιστοποιεί τις μετρήσεις απόδοσης, όπως το accuracy, precision, recall και F1-score. Παρά την επίτευξη σχετικά υψηλών accuracies άλλες μετρικές όπως το recall δεν ήταν ιδιαίτερα υψηλά και αξιόπιστες, έτσι μπορεί να υπάρχει ακόμα περιθώριο βελτίωσης με τη λεπτομερή ρύθμιση των υπερπαραμέτρων. Το Grid Search βοηθά στη βελτιστοποίηση της απόδοσης του μοντέλου περαιτέρω πραγματοποιώντας εξαντλητική αναζήτηση σε ένα καθορισμένο πλέγμα παραμέτρων και επιλέγοντας τον συνδυασμό παραμέτρων με την καλύτερη απόδοση. Αυτή η διαδικασία μπορεί να οδηγήσει σε καλύτερη γενίκευση και πιο εύρωστα μοντέλα, ενισχύοντας τη συνολική αποτελεσματικότητα της ταξινόμησής μας.



Διάγραμμα καμπύλης ROC όλων των μοντέλων μετά την εφαρμογή Fine-tuning

Η καμπύλη Receiver Operating Characteristic (ROC) είναι μια γραφική αναπαράσταση που απεικονίζει την απόδοση ενός μοντέλου ταξινόμησης σε διάφορα thresholds. Σχεδιάζει το πραγματικό θετικό ποσοστό (TPR) έναντι του ψευδώς θετικού ποσοστού (FPR) σε διαφορετικές ρυθμίσεις threshold.

Στην καμπύλη ROC, κάθε σημείο αντιπροσωπεύει ένα διαφορετικό όριο και η διαγώνια γραμμή από την κάτω αριστερή γωνία στην επάνω δεξιά γωνία αντιπροσωπεύει την απόδοση ενός τυχαίου ταξινομητή. Ένας τέλειος ταξινομητής θα είχε την καμπύλη ROC του να διέρχεται από την επάνω αριστερή γωνία, υποδεικνύοντας TPR 1 και FPR 0, με αποτέλεσμα μια περιοχή κάτω από την καμπύλη (AUC) 1.

Η περιοχή κάτω από την καμπύλη ROC (AUC) είναι ένα συνοπτικό μέτρο της απόδοσης του ταξινομητή. Μια υψηλότερη τιμή AUC υποδηλώνει καλύτερη διάκριση μεταξύ των θετικών και αρνητικών κλάσεων. Γενικά, μια τιμή AUC πιο κοντά στο 1 υποδηλώνει ένα μοντέλο με καλύτερη απόδοση.

Στα αποτελέσματά μας, παρατηρούμε ότι όλα τα μοντέλα έχουν σχετικά υψηλές τιμές AUC, που κυμαίνονται από 0,87 έως 0,91, υποδεικνύοντας καλή ικανότητα διάκρισης. Τα μοντέλα Logistic Regression που έχουν εκπαιδευτεί τόσο σε χαρακτηριστικά Bag-of-Words (BoW) όσο και σε TF-IDF παρουσιάζουν σταθερά τις υψηλότερες τιμές AUC, υποδηλώνοντας ανώτερη απόδοση στη διάκριση μεταξύ των δύο κατηγοριών. Το Naive Bayes έχει επίσης καλή απόδοση, με τιμές AUC γύρω στο 0,90, ενώ τα μοντέλα SVM και Random Forest εμφανίζουν ελαφρώς χαμηλότερες αλλά και πάλι σεβαστές τιμές AUC που κυμαίνονται από 0,87 έως 0,89. Συνολικά, η ανάλυση της καμπύλης ROC επιβεβαιώνει την αποτελεσματικότητα των μοντέλων στη διάκριση μεταξύ ρατσιστικών/σεξιστικών και μη ρατσιστικών/σεξιστικών tweet, με την Logistic Regression να δείχνει ιδιαίτερα υποσχόμενα αποτελέσματα.

Επίσης μετά την εφαρμογή του fine tuning έχουμε τα ακόλουθα αποτελέσματα:

Model	precision	recall	f1-score	support
Logistic Regression (BoW)	0.9348425924904029	0.9433755670264352	0.9356960091179108	6393
Naive Bayes (BoW)	0.9306555084404423	0.9319568277803848	0.9312873163974903	6393
SVM (BoW)	0.9456962296098194	0.9511966213045518	0.9461789553296187	6393
Random Forest (BoW)	0.9373675276195866	0.9377444079461912	0.9375539847107465	6393
Logistic Regression (TF-IDF)	0.9358133700257165	0.9441576724542469	0.9360415461578472	6393
Naive Bayes (TF-IDF)	0.9346669046399494	0.9430627248553105	0.9314784965527807	6393
SVM (TF-IDF)	0.9447988663926806	0.9504145158767402	0.9455824749582247	6393
Random Forest (TF-IDF)	0.9431664582662278	0.9491631471922415	0.9440539627296846	6393

Εικόνα 53. Πίνακας αποτελεσμάτων Fine-Tuing

Για τα μοντέλα Logistic Regression, παρατηρούμε τιμές υψηλού precision που κυμαίνονται από περίπου 0,935 έως 0,936 υποδεικνύοντας την ικανότητα των μοντέλων να ταξινομούν σωστά τα θετικά στιγμιότυπα από όλες τις περιπτώσεις που προβλέπονται ως θετικές. Επιπλέον, τα μοντέλα εμφανίζουν ισχυρές τιμές recall που κυμαίνονται από

0,943 έως 0,944 υποδηλώνοντας το ποσοστό των πραγματικών θετικών περιπτώσεων που εντοπίστηκαν σωστά. Συνεπώς, το F1-score, το οποίο είναι το αρμονικό μέσο precision και recall, κυμαίνεται από περίπου 0,935 έως 0,936 επιδεικνύοντας μια ισορροπία μεταξύ precision και recall.

Ομοίως, τα μοντέλα Naive Bayes επιτυγχάνουν ανταγωνιστικές τιμές precision που κυμαίνονται από περίπου 0,930 έως 0,935 σε συνδυασμό με τις τιμές recall που κυμαίνονται από 0,931 έως 0,943. Αυτά τα μοντέλα αποδίδουν επίσης αξιέπαινα F1-score που κυμαίνονται από περίπου 0,931 έως 0,932, υποδεικνύοντας μια αρμονική ισορροπία μεταξύ precision και recall.

Τα μοντέλα SVM επιδεικνύουν τις υψηλότερες τιμές precision μεταξύ όλων των μοντέλων, που κυμαίνονται από περίπου 0,945 έως 0,946 υπογραμμίζοντας την ικανότητα των μοντέλων να κάνουν ακριβείς θετικές προβλέψεις. Επιπλέον, τα μοντέλα SVM επιτυγχάνουν τιμές recall που κυμαίνονται από 0,950 έως 0,951, υποδεικνύοντας υψηλή ευαισθησία στον εντοπισμό πραγματικών θετικών περιπτώσεων. Κατά συνέπεια, τα F1-score για τα μοντέλα SVM κυμαίνονται από περίπου 0,945 έως 0,946 επιδεικνύοντας μια ισχυρή ισορροπία μεταξύ precision και recall.

Τέλος, τα μοντέλα Random Forest επιτυγχάνουν τιμές precision που κυμαίνονται από περίπου 0,937 έως 0,943 συνοδευόμενες από τιμές ανάκλησης που κυμαίνονται από 0,938 έως 0,949. Αυτά τα μοντέλα παρουσιάζουν επίσης ανταγωνιστικά F1-score που κυμαίνονται από περίπου 0,938 έως 0,944 υποδεικνύοντας μια ισχυρή ισορροπία μεταξύ precision και recall.

Συνολικά, αυτά τα αποτελέσματα υπογραμμίζουν την αποτελεσματικότητα όλων των μοντέλων στην ταξινόμηση των tweets σε ρατσιστικές/σεξιστικές και μη ρατσιστικές/σεξιστικές κατηγορίες, με τα μοντέλα SVM να επιδεικνύουν ελαφρώς ανώτερη απόδοση όσον αφορά το precision, το recall και το F1-score.

Κεφάλαιο 4. Συμπεράσματα και Προτάσεις

Το αποτέλεσμα αυτής της ανάλυσης δείχνει την αποτελεσματικότητα των διαφόρων μοντέλων μηχανικής μάθησης σε tweets είτε περιέχουν ρατσιστικό/σεξιστικό περιεχόμενο είτε ως μη ρατσιστικό/σεξιστικό περιεχόμενο. Μέσω εκτεταμένης εξερεύνησης, προεπεξεργασίας, εξαγωγής χαρακτηριστικών και εκπαίδευσης μοντέλων, έχουμε εμβαθύνει στις αποχρώσεις των τεχνικών επεξεργασίας φυσικής γλώσσας και της ανάλυσης συναισθημάτων, ρίχνοντας φως στη δυνατότητα εφαρμογής τους σε σενάρια πραγματικού κόσμου.

Τα μοντέλα, εκπαιδευμένα με τεχνικές Bag-of-Words (BoW) και TF-IDF, έχουν επιδείξει αξιέπαινη απόδοση σε πολλαπλές μετρήσεις αξιολόγησης. Οι ταξινομητές Logistic Regression, Naive Bayes, Support Vector Machine (SVM) και Random Forest έχουν αποδείξει την ικανότητά τους στην ακριβή ταξινόμηση των tweets, με τα μοντέλα SVM να παρουσιάζουν ένα μικρό πλεονέκτημα όσον αφορά το precision, το recall και το F1-score. Επιπλέον, η ανάλυση συναισθήματος παρείχε πολύτιμες γνώσεις για το συναισθηματικό πλαίσιο των tweets, επιτρέποντας μια βαθύτερη κατανόηση των υποκείμενων συναισθημάτων που εκφράζονται στο σύνολο δεδομένων. Συνολικά, ακόμη κι αν αρχικά φαίνεται αντιφατική η παρουσία ενός «ρατσιστικού/σεξιστικού» tweet με θετικό συναίσθημα υπογραμμίζει την πολυπλοκότητα της γλώσσας και τις προκλήσεις της ερμηνείας του συναισθήματος σε δεδομένα κειμένου πραγματικού κόσμου. Αντίστοιχα μπορεί να φαίνεται απροσδόκητο να συναντήσουμε ένα «μη ρατσιστικό/σεξιστικό» tweet με αρνητικό συναίσθημα, διάφοροι παράγοντες, συμπεριλαμβανομένων των αποχρώσεων της γλώσσας και των περιορισμών της ανάλυσης συναισθημάτων, μπορούν να συμβάλουν σε αυτό το φαινόμενο. Η κατανόηση των συμφραζομένων και η προσεκτική ερμηνεία είναι ζωτικής σημασίας για την ακριβή ερμηνεία του συναισθήματος τέτοιων tweets.

Επιπλέον, η εξερεύνηση των confusion matrices έχει αποσαφηνίσει την ικανότητα των μοντέλων να εντοπίζουν σωστά τα αληθινά θετικά και τα αληθινά αρνητικά, ελαχιστοποιώντας τα ψευδώς θετικά και τα ψευδώς αρνητικά. Αυτή η ανάλυση έχει παράσχει πολύτιμες πληροφορίες για τα δυνατά και αδύνατα σημεία κάθε μοντέλου, ενημερώνοντας για μελλοντικές επαναλήψεις και βελτιώσεις.

Συμπερασματικά, τα ευρήματα αυτής της μελέτης υπογραμμίζουν τις δυνατότητες της μηχανικής εκμάθησης και των τεχνικών επεξεργασίας φυσικής γλώσσας για τον εντοπισμό και την κατηγοριοποίηση διαδικτυακού περιεχομένου, ιδιαίτερα στο πλαίσιο του εντοπισμού και του μετριασμού της επιβλαβούς ή καταχρηστικής γλώσσας στις πλατφόρμες κοινωνικών μέσων. Καθώς η διαδικτυακή συζήτηση συνεχίζει να εξελίσσεται, η ανάπτυξη ισχυρών και αποτελεσματικών μοντέλων για την ανάλυση συναισθημάτων και τον εντοπισμό της ρητορικής μίσους παραμένει επιτακτική για την προώθηση ενός ασφαλέστερου και περιεκτικού διαδικτυακού περιβάλλοντος.

Βιβλιογραφία

Abram, M. D., Mancini, K. T., & Parker, R. D. (2020). Methods to Integrate Natural Language Processing Into Qualitative Research. *International Journal of Qualitative Methods*, *19*. https://doi.org/10.1177/1609406920984608

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, *58*(2), 175-191. https://doi.org/10.1177/0047287517747753

Al-Taie, M. Z., Salim, N., & Obasa, A. I. (2017). Successful Data Science Projects: Lessons Learned from Kaggle Competition. *Kurdistan Journal of Applied Research*, *2*(3), 40-49. https://doi.org/10.24017/science.2017.3.18

Agarwal, S. (2022). Deep Learning-based Sentiment Analysis: Establishing Customer Dimension as the Lifeblood of Business Management. *Global Business Review*, *23*(1), 119-136. https://doi.org/10.1177/0972150919845160

Alharbi, M., Roach, M., Cheesman, T., & Laramee, R. S. (2021). VNLP: Visible natural language processing. *Information Visualization*, *20*(4), 245-262. https://doi.org/10.1177/14738716211038898

Arnarsson IÖ, Frost O, Gustavsson E, Jirstrand M, Malmqvist J. (2021). Natural language processing methods for knowledge management—Applying document clustering for fast search and grouping of engineering documents. *Concurrent Engineering*, 29(2), 142-152. doi:10.1177/1063293X20982973

Baldwin P, Mee J, Yaneva V, et al. (2022). A Natural-Language-Processing-Based Procedure for Generating Distractors for Multiple-Choice Questions. *Evaluation & the Health Professions*, 45(4), 327-340. doi:10.1177/01632787211046981

Bilbao-Jayo A, Almeida A. (2018). Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data. *International Journal of Distributed Sensor Networks*, 14(11). doi:10.1177/1550147718811827

Bohlouli, M., Dalter, J., Dornhöfer, M., Zenkert, J., & Fathi, M. (2015). Knowledge discovery from social media using big data-provided sentiment analysis

(SoMABiT). *Journal of Information Science*, *41*(6), 779-798. https://doi.org/10.1177/0165551515602846

Chang, T., DeJonckheere, M., Vydiswaran, V. G. V., Li, J., Buis, L. R., & Guetterman, T. C. (2021). Accelerating Mixed Methods Research With Natural Language Processing of Big Text Data. *Journal of Mixed Methods Research*, *15*(3), 398-412. https://doi.org/10.1177/15586898211021196

Chau, X. T. D., Nguyen, T. T., Jo, J., Quach, S., Ngo, L. V., Pham, H., & Thaichon, P. (2023). Simplifying Sentiment Analysis on Social Media: A Step-by-Step Approach. *Australasian Marketing Journal*, *0*(0). https://doi.org/10.1177/14413582231217126

Chau, X. T. D., Nguyen, T. T., Jo, J., Quach, S., Ngo, L. V., Pham, H., & Thaichon, P. (2023). Simplifying Sentiment Analysis on Social Media: A Step-by-Step Approach. *Australasian Marketing Journal*, *0*(0). https://doi.org/10.1177/14413582231217126

Deng H, Wang Q, Turner DP, et al. (2020). Sentiment analysis of real-world migraine tweets for population research. *Cephalalgia Reports*, 3. doi:10.1177/2515816319898867

Dimisianos, N. (2019). Political Campaigns, Social Media, and Analytics: The Case of the GDPR, Visvizi, A. and Lytras, M.D. (Ed.) *Politics and Technology in the Post-Truth Era* (*Emerald Studies in Politics and Technology*), Emerald Publishing Limited, Leeds, pp. 73-88. https://doi.org/10.1108/978-1-78756-983-620191006

Ding, Q., Ding, D., Wang, Y., Guan, C. and Ding, B. (2023). Unraveling the landscape of large language models: a systematic review and future perspectives, *Journal of Electronic Business & Digital Economics*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JEBDE-08-2023-0015

Doğan, B., Balcioglu, Y.S. and Elçi, M. (2024). Multidimensional sentiment analysis method on social media data: comparison of emotions during and after the COVID-19 pandemic, *Kybernetes*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/K-09-2023-1808

Djohari, N. (2016). Trading in unicorns: The role of exchange etiquette in managing the online second-hand sale of sentimental babywearing wraps. *Journal of Material Culture*, *21*(3), 297-316. https://doi.org/10.1177/1359183515619455

Duan, W., Yu, Y., Cao, Q., & Levy, S. (2016). Exploring the Impact of Social Media on Hotel Service Performance: A Sentimental Analysis Approach. *Cornell Hospitality Quarterly*, *57*(3), 282-296. https://doi.org/10.1177/1938965515620483

Fang, X., & Wang, T. (2022). Using Natural Language Processing to Identify Effective Influencers. *International Journal of Market Research*, *64*(5), 611-629. https://doi.org/10.1177/14707853221101565

Gárdos, J., Egyed-Gergely, J., Horváth, A., Pataki, B., Vajda, R. and Micsik, A. (2023). Identification of social scientifically relevant topics in an interview repository: a natural language processing experiment, *Journal of Documentation*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/JD-12-2022-0269

Hamaz, K. and Benchikha, F. (2017). A novel method for providing relational databases with rich semantics and natural language processing, *Journal of Enterprise Information Management*, Vol. 30 No. 3, pp. 503-525. https://doi.org/10.1108/JEIM-01-2015-0005

Hartmann, J. and Netzer, O. (2023). Natural Language Processing in Marketing, Sudhir, K. and Toubia, O. (Ed.) *Artificial Intelligence in Marketing* (*Review of Marketing Research, Vol. 20*), Emerald Publishing Limited, Leeds, pp. 191-215. https://doi.org/10.1108/S1548-643520230000020011

He, W., Zhang, W., Tian, X., Tao, R. and Akula, V. (2019). Identifying customer knowledge on social media through data analytics, *Journal of Enterprise Information Management*, Vol. 32 No. 1, pp. 152-169. https://doi.org/10.1108/JEIM-02-2018-0031

Heath JK, Clancy CB, Pluta W, et al. (2023). Natural Language Processing of Learners' Evaluations of Attendings to Identify Professionalism Lapses. *Evaluation & the Health Professions*, 46(3), 225-232. doi:10.1177/01632787231158128

Hossain, M. S., & Rahman, M. F. (2023). Customer Sentiment Analysis and Prediction of Insurance Products' Reviews Using Machine Learning Approaches. *FIIB Business Review*, *12*(4), 386-402. https://doi.org/10.1177/23197145221115793

Hutchinson, T. (2020). Natural language processing and machine learning as practical toolsets for archival processing, *Records Management Journal*, Vol. 30 No. 2, pp. 155-174. https://doi.org/10.1108/RMJ-09-2019-0055

Kaggle platform [Online] Available at https://www.kaggle.com/

Kehl, W., Jackson, M., & Fergnani, A. (2020). Natural Language Processing and Futures Studies. *World Futures Review*, *12*(2), 181-197. https://doi.org/10.1177/1946756719882414

Keramatfar, A., & Amirkhani, H. (2019). Bibliometrics of sentiment analysis literature. *Journal of Information Science*, *45*(1), 3-15. https://doi.org/10.1177/0165551518761013

Khan, U., Khan, H.U., Iqbal, S. and Munir, H. (2022). Four decades of image processing: a bibliometric analysis, *Library Hi Tech*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/LHT-10-2021-0351

Kwartler, T. (2021). Text Analytics and Natural Language Processing, Einhorn, M., Löffler, M., de Bellis, E., Herrmann, A. and Burghartz, P. (Ed.) *The Machine Age of Customer Insight*, Emerald Publishing Limited, Leeds, pp. 119-128. https://doi.org/10.1108/978-1-83909-694-520211012

Lin, Y. and Yu, Z. (2023). A bibliometric analysis of artificial intelligence chatbots in educational contexts, *Interactive Technology and Smart Education*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/ITSE-12-2022-0165

Liu, S., Peng, X., Cheng, H. N. H., Liu, Z., Sun, J., & Yang, C. (2019). Unfolding Sentimental and Behavioral Tendencies of Learners' Concerned Topics From Course Reviews in a MOOC. *Journal of Educational Computing Research*, *57*(3), 670-696. https://doi.org/10.1177/0735633118757181

Luri, I., Schau, H. J., & Ghosh, B. (2023). Metaphor-Enabled Marketplace Sentiment Analysis. *Journal of Marketing Research*, *0*(0). https://doi.org/10.1177/00222437231191526

Majhi, D. and Mukherjee, B. (2023). Identifying research fronts in NLP applications in library and information science using meta-analysis approaches, *Digital Library Perspectives*, Vol. 39 No. 3, pp. 393-411. https://doi.org/10.1108/DLP-12-2022-0099

Musa, I. H., Zamit, I., Xu, K., Boutouhami, K., & Qi, G. (2023). A comprehensive bibliometric analysis on opinion mining and sentiment analysis global research output. *Journal of Information Science*, *49*(6), 1506-1516. https://doi.org/10.1177/01655515211061866

Nee, J., Smith, G. M., Sheares, A., & Rustagi, I. (2022). Linguistic justice as a framework for designing, developing, and managing natural language processing tools. *Big Data & Society*, *9*(1). https://doi.org/10.1177/20539517221090930

Panda, S. and Kaur, N. (2023). Revolutionizing language processing in libraries with SheetGPT: an integration of Google Sheet and ChatGPT plugin, *Library Hi Tech News*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/LHTN-03-2023-0051

Paschen, J., Kietzmann, J. and Kietzmann, T.C. (2019). Artificial intelligence (AI) and its implications for market knowledge in B2B marketing, *Journal of Business & Industrial Marketing*, Vol. 34 No. 7, pp. 1410-1419. https://doi.org/10.1108/JBIM-10-2018-0295

Paul, S., Spain, R., Min, W., Pande, J., & Lester, J. (2023). Evaluating the Classification Performance of Natural Language Processing-Driven Team Communication Analysis Models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *67*(1), 2181-2186. https://doi.org/10.1177/21695067231192449

Python Libraries, n.d. [Online] Available at https://docs.python.org/3/library/index.html

Šandor, D. and Bagić Babac, M. (2023). Sarcasm detection in online comments using machine learning, *Information Discovery and Delivery*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/IDD-01-2023-0002

Sarmet M, Kabani A, Coelho L, dos Reis SS, Zeredo JL, Mehta AK. (2023). The use of natural language processing in palliative care research: A scoping review. *Palliative Medicine*, 37(2), 275-290. doi:10.1177/02692163221141969

Saurwein, F., Brantner, C., & Möck, L. (2023). Responsibility networks in media discourses on automation: A comparative analysis of social media algorithms and social companions. *New Media & Society*, *0*(0). https://doi.org/10.1177/14614448231203310

ShabbirHusain, R.V., Pathak, A.A., Chandrasekaran, S. and Annamalai, B. (2023). The power of words: driving online consumer engagement in Fintech, *International Journal of Bank Marketing*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/IJBM-11-2022-0519

Singh, B. (2019). Social Media, Cultural Activism and Place-making in Contemporary Lucknow. *Society and Culture in South Asia*, *5*(1), 1-18. https://doi.org/10.1177/2393861718787869

Ta-Johnson, V., Suss, J. and Lande, B. (2023). Using natural language processing to measure cognitive load during use-of-force decision-making training, *Policing: An International Journal*, Vol. 46 No. 2, pp. 227-242. https://doi.org/10.1108/PIJPSM-06-2022-0084

Taskin, Z. and Al, U. (2019). Natural language processing applications in library and information science, *Online Information Review*, Vol. 43 No. 4, pp. 676-690. https://doi.org/10.1108/OIR-07-2018-0217

Twitter Developer Platform [Online] Available at https://developer.twitter.com/en

Yuan, H., Tang, Y., Xu, W. and Lau, R.Y.K. (2021). Exploring the influence of multimodal social media data on stock performance: an empirical perspective and analysis, *Internet Research*, Vol. 31 No. 3, pp. 871-891. https://doi.org/10.1108/INTR-11-2019-0461

Zhang, L., & Mu, X. (2022). Tweets on a horror movie: An investigation into relationships between sentiment strength, cognitive language and tweet virality. *Journal of Information Science*, *0*(0). https://doi.org/10.1177/01655515221116516

Zhou, H., Gao, B., Tang, S., Li, B. and Wang, S. (2023). Intelligent detection on construction project contract missing clauses based on deep learning and NLP, *Engineering, Construction and Architectural Management*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/ECAM-02-2023-0172