Πανεπιστήμιο Δυτικής Μακεδονίας
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών

Διπλωματική Εργασία

Κατηγοριοποίηση κειμένου με χρήση εργαλείων μηχανικής μάθησης.

Text categorization using machine learning tools.

Ράπτης Χρήστος

Επιβλέπουσα Καθηγήτρια:

Μαλαματή Λούτα

Κοζάνη, Ιούνιος 2024

Περιεχόμενα

Περιεχόμενα

Πίνακας Εικόνων

Περίληψη

Η ανάπτυξη των τεχνολογιών του διαδικτύου οδήγησε στην μεγάλη διαθεσιμότητα διαδικτυακού περιεχομένου. Η πρόκληση που προέκυψε ήταν η διαχείριση μεγάλων όγκων δεδομένων, καθώς οι συμβατικές μέθοδοι δεν ήταν κατάλληλες πλέον να χρησιμοποιηθούν για το σκοπό αυτό. Οι αλγόριθμοι μηχανικής μάθησης είναι σε θέση να παράγουν πληροφορίες και συμπεράσματα από μεγάλα σύνολα δεδομένων, παρουσιάζοντας μικρή σχετικά πολυπλοκότητα. Μια δημοφιλής χρήση των τεχνικών της μηχανικής μάθησης είναι στην επεξεργασία φυσικής γλώσσας. Στην παρούσα εργασία πραγματοποιείται μία επισκόπηση των τρόπων εφαρμογής της μηχανικής μάθησης στην κατηγοριοποίηση κειμένων. Επιπλέον, εξετάζεται η αποτελεσματικότητα των μηχανισμών αυτών μέσα από την υλοποίηση τους σε γλώσσα προγραμματισμού υψηλού επιπέδου. Τα αποτελέσματα της μελέτης φανερώνουν τις υψηλές δυνατότητες της μηχανικής μάθησης για την ολοκλήρωση διαδικασιών επεξεργασίας φυσικής γλώσσας.

Λέξεις κλειδιά: Μηχανική Μάθηση, Κατηγοριοποίηση κειμένου, Επεξεργασία φυσικής γλώσσας

Abstract

The development of internet technologies has led to the great availability of online content. The challenge that arose was the management of large volumes of data, as conventional methods were no longer suitable to be used for this purpose. Machine learning algorithms are able to generate insights and inferences from large data sets, presenting relatively little complexity. A popular use of machine learning techniques is in natural language processing. In this paper, an overview of the ways of applying machine learning to text categorization is carried out. In addition, the effectiveness of these mechanisms is examined through their implementation in a high-level programming language. The results of the study reveal the high potential of machine learning to complete natural language processing processes.

Key words: Machine learning, Text categorization, Natural language processing

Ευχαριστίες

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας ολοκληρώνονται και οι σπουδές μου στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Δυτικής Μακεδονίας.

Στις σπουδές μου ήταν καθοριστική η συμβολή των καθηγητών μου στα γνωστικά αντικείμενα που παρακολούθησα, στους οποίους οφείλω να εκφράσω τις ειλικρινείς μου ευχαριστίες για τη συμβολή τους στην ολοκλήρωση των σπουδών μου.

Ιδιαίτερα επιθυμώ να ευχαριστήσω την καθηγήτριά μου, κα Μαλαματή Λούτα. Από το πρώτο κιόλας έτος ήταν ιδιαίτερη η συμπάθειά μου στο πρόσωπό της και κάπως έτσι τα έφερε ο χρόνος να εκπονήσω δικό της θέμα στη διπλωματική. Επίσης, οφείλω ένα τεράστιο ευχαριστώ στον κο Ευάγγελο Τσίπη που από την αρχή μέχρι το τέλος ήταν δίπλα μου κατά τη διάρκεια υλοποίησης της διπλωματικής. Ο συνδυασμός αυτών των δύο ανθρώπων ήταν τρομερός και τους υπερευχαριστώ για την επιστημονική και συμβουλευτική καθοδήγηση που μου προσέφεραν σε όλα τα στάδια εκπόνησης της εργασίας με τις εύστοχες και πολύ εποικοδομητικές παρατηρήσεις τους.

Οφείλω να εκφράσω τις ευχαριστίες μου προς τους συμφοιτητές μου, οι οποίοι στα εύκολα και στα δύσκολα ήταν εκεί να μου θυμίζουν το στόχο μου.

Πάρα πολύ σημαντική η συνεισφορά της ομάδας μου, “VAGIAS RACE” που επέμεναν στην κατάκτηση αυτού του στόχου ακόμη και τις στιγμές που είχα κουραστεί. Τους ευχαριστώ για όλα.

Τέλος, οφείλω να ευχαριστήσω τους πιο σημαντικούς κατ’ εμού. Την οικογένειά μου. Την μητέρα μου την Ευαγγελία, τον πατέρα μου τον Θέμη, τα μικρά μου αδέρφια, Δήμο και Άγγελο και τη σύντροφό μου την Βάσω για τη συμπαράσταση και την υπομονή τους. Χωρίς αυτούς δεν θα είχα καταφέρει το παραμικρό.

Δήλωση Πνευματικών Δικαιωμάτων

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν. 1599/1986 και τα άρθρα 2,4,6 παρ. 3 του Ν. 1256/1982, η παρούσα Διπλωματική Εργασία με τίτλο “Κατηγοριοποίηση κειμένου με χρήση εργαλείων μηχανικής μάθησης ” καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας και αναφέρονται ρητώς μέσα στο κείμενο που συνοδεύουν, και η οποία έχει εκπονηθεί στο τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών της Πολυτεχνικής Σχολής του Πανεπιστημίου Δυτικής Μακεδονίας, υπό την επίβλεψη του μέλους του Τμήματος κας Μαλαματής Λούτα αποτελεί αποκλειστικά προϊόν προσωπικής εργασίας και δεν προσβάλλει κάθε μορφής πνευματικά δικαιώματα τρίτων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή / και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα. Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και μόνο.

Εισαγωγή

Η μετάβαση στο Web2.0 δημιούργησε τις προϋποθέσεις για να αναπτυχθούν διαδικτυακές εφαρμογές με ποικίλους προσανατολισμούς. Αυτό είχε ως επακόλουθο περισσότεροι άνθρωποι να υιοθετήσουν τις εφαρμογές του και να αναζητήσουν τρόπους να βελτιώσουν την αποδοτικότητα των δραστηριοτήτων τους μέσα από τη χρήση τους. Ο πολλαπλασιασμός του πλήθους των διαθέσιμων διαδικτυακών εφαρμογών συνέβαλε αποφασιστικά στη ραγδαία αύξηση του όγκου των δεδομένων που είναι διαθέσιμα στο διαδίκτυο. Μία ακόμα αιτία που οδήγησε στην αύξηση της διαθεσιμότητας διαδικτυακών δεδομένων, ήταν και η συμμετοχή των απλών χρηστών στην παραγωγή τους. Οι άνθρωποι μέσα από τη χρήση εξειδικευμένων εφαρμογών (πχ κοινωνική δικτύωση, ιστολόγια) ή συσκευών του Διαδικτύου των Πραγμάτων (Internet of Things) τροφοδοτούν καθημερινά τις υποδομές του με μεγάλους όγκους δεδομένων. Η κατάσταση αυτή που διαμορφώθηκε περιγράφεται με τον όρο Μεγάλα Δεδομένα (Big Data).

Τα Μεγάλα Δεδομένα είναι ένας συνδυασμός δομημένων, ημιδομημένων και μη δομημένων δεδομένων που συλλέγονται κυρίως από πηγές με διαδικτυακές απολήξεις. Χρησιμοποιούνται από απλούς χρήστες του διαδικτύου και οργανισμούς προκειμένου να παράγουν χρήσιμες πληροφορίες για την λειτουργία και την εξέλιξή τους. Τα συστήματα που επεξεργάζονται και αποθηκεύουν μεγάλα δεδομένα αντιμετωπίζουν σημαντικές προκλήσεις. Αυτές οφείλονται σε μεγάλο βαθμό στα εγγενή χαρακτηριστικά τους, τα κυριότερα εκ των οποίων περιλαμβάνουν:

- Τον μεγάλο όγκο τους
- Την ποικιλία τύπων και μορφών που εμφανίζονται
- Την ταχύτητα που παράγονται και μεταβάλλονται

Η ανάλυση μεγάλων δεδομένων, η συλλογή τεράστιων ποσοτήτων δεδομένων και η μετατροπή τους σε πληροφορίες, οδηγεί συχνά στη λήψη σημαντικών αποφάσεων.

Στην κατηγορία των Μεγάλων Δεδομένων ανήκουν και αυτά που παράγονται από τις εφαρμογές κοινωνικής δικτύωσης. Αυτές αποτελούν ένα μέσο έκφρασης και επικοινωνίας για τους περισσότερους σύγχρονους ανθρώπους. Κάτι ανάλογο είναι και ένα μέσο παροχής διαδικτυακού περιεχομένου το οποίο σχετίζεται απευθείας με το κοινό. Υπό αυτή την έννοια αποτελεί μία σημαντική πηγή δεδομένων για χρήση σε διαδικασίες λήψης αποφάσεων σε οργανισμούς που η λειτουργία τους εξαρτάται σε σημαντικό βαθμό από το κοινό.

Όπως έχει ήδη αναφερθεί, τα μεγάλα δεδομένα – και αυτά που προέρχονται από τα μέσα κοινωνικής δικτύωσης, παράγονται με γρήγορους ρυθμούς. Το γεγονός αυτό σε συνδυασμό με τον όγκο τους, καθιστά σχεδόν αδύνατη την αποδοτική τους επεξεργασία με συμβατικές μεθόδους που βασίζονται σε ηλεκτρονικούς υπολογιστές, πολύ δε περισσότερο με άμεση συμμετοχή του ανθρώπινου παράγοντα. Το γεγονός αυτό εντείνει την ανάγκη για έρευνα προς την κατεύθυνση ανακάλυψης αποδοτικών τρόπων επεξεργασίας τους. Η ανάλυση δεδομένων, μέσα από την επεξεργασία φυσικής γλώσσας, αφορά μεθοδολογίες και τεχνικές που επιτρέπουν σε ηλεκτρονικές υπολογιστές να ερμηνεύουν και να αναλύουν την καθομιλουμένη γλώσσα με αυτοματοποιημένο τρόπο. Πρόκειται για έναν απαιτητικό κλάδο της τεχνητής νοημοσύνης, που έχει ήδη προσφέρει αποδοτικές διαδικασίες επεξεργασίας, με ποικίλες εφαρμογές στον πραγματικό κόσμο.

Οι μεθοδολογίες και οι τεχνικές της επεξεργασίας φυσικής γλώσσας, εφαρμόζονται ήδη στην ανάλυση του περιεχομένου των κοινωνικών δικτύων. Η ανάλυση των δεδομένων που προέρχονται από τους χρήστες τους είναι πολύτιμη πηγή δεδομένων εισόδου για διαδικασίες λήψης αποφάσεων. Μέσα από την ανάλυσή τους είναι εφικτή η αποτύπωση των τάσεων της κοινής γνώμης, του καταναλωτικού κοινού, του εκλογικού σώματος και άλλων πληθυσμών – στόχων για κάθε είδους οργανισμό. Οι μεθοδολογίες και οι τεχνικές διακρίνονται σε διάφορες κατηγορίες. Κάποιες από αυτές χρησιμοποιούν αλγορίθμους μηχανικής μάθησης προκειμένου να επεξεργαστούν μεγάλους όγκους δεδομένων φυσικής γλώσσας για τη δημιουργία συμπερασμάτων.

Στην παρούσα εργασία περιγράφονται θεωρητικά και πρακτικά οι μεθοδολογίες επεξεργασίας φυσικής γλώσσας που χρησιμοποιούνται στην κατηγοριοποίηση κειμένων που προέρχονται από το περιεχόμενο εφαρμογών κοινωνικής δικτύωσης. Ο σκοπός της είναι να ανακαλυφθεί με πρακτικό τρόπο το κατά πόσο είναι εφικτή η ανάλυση των αναρτήσεων των χρηστών των κοινωνικών δικτύων για την παραγωγή συμπερασμάτων και την κατηγοριοποίησή τους. Η μελέτη στηρίζεται στο απαραίτητο θεωρητικό υπόβαθρο.

Το υπόλοιπο του κειμένου είναι διαρθρωμένο ως εξής:

- Κεφάλαιο 1: Στο πρώτο κεφάλαιο περιγράφονται τα μέσα κοινωνικής δικτύωσης. Καταγράφονται τα χαρακτηριστικά τους και καταδεικνύεται η διεισδυτικοτητα τους στους σύγχρονους πληθυσμούς. Επιπλέον τονίζεται η επιρροή τους στην συμπεριφορά των ανθρώπων (σε διάφορους τομείς της δραστηριότητάς τους) προκειμένου να αποδειχθεί η σημασία της ανάλυσης του περιεχομένου τους. Τέλος αναφέρονται οι δημοφιλέστερες εφαρμογές κοινωνικής δικτύωσης.

- Κεφάλαιο 2: Το κεφάλαιο αυτό περιλαμβάνει μία θεωρητική περιγραφή της Μηχανικής Μάθησης. Αναλύεται το ποιες απαιτήσεις οδήγησαν στις προσεγγίσεις της ως μέρος της εξόρυξης γνώσης. Καταγράφονται τα χαρακτηριστικά της και οι διαφορετικές μεθοδολογίες ανάπτυξης των αλγορίθμων της. Ιδιαίτερη αναφορά γίνεται στην γενικότερη αλληλουχία διεργασιών που εκτελούνται κατά την ολοκλήρωση ενός έργου μηχανικής μάθησης.

- Κεφάλαιο 3: Στο τρίτο κεφάλαιο περιγράφονται οι διαδικασίες επεξεργασίας φυσικής γλώσσας. Περιλαμβάνεται ο ορισμός της και τα βασικά της χαρακτηριστικά. Επιπλέον καταγράφονται οι διαφορετικές προσεγγίσεις ανάλυσης.

- Κεφάλαιο 4: Στο σημείο αυτό του κειμένου εξειδικεύεται η θεωρητική αναφορά στο πλαίσιο της υλοποίησης που ακολουθεί. Η ανάλυση συναισθήματος συνίσταται στην κατηγοριοποίηση των κειμένων ανάλογα με το συναίσθημα που αναδύεται από αυτά. Περιγράφονται οι διαφορετικές προσεγγίσεις που έχουν προταθεί και χρησιμοποιούνται, με έμφαση σε αυτές που βασίζονται στη μηχανική μάθηση και τη χρήση λεξικών. Στο δεύτερο μέρος του κεφαλαίου περιλαμβάνονται δυο υλοποιήσεις εφαρμογών ανάλυσης συναισθήματος γραμμένες σε python. Η πρώτη βασίζεται σε αλγορίθμους μηχανικής μάθησης και η δεύτερη σε χρήση λεξικών. Για την υλοποίηση αυτή εξετάστηκε η στάση της κοινής γνώμης σε ένα επίκαιρο θέμα (ο πόλεμος στην Ουκρανία). Για τον σκοπό αυτό ανακτήθηκε και αναλύθηκε ικανός αριθμός αναρτήσεων σε δημοφιλές μέσο κοινωνικής δικτύωσης.

- Συμπεράσματα: Την ανάλυση που προηγήθηκε, ακολουθεί η καταγραφή των συμπερασμάτων που προέκυψε. Αυτά αναφέρονται στην αποδοτικότητα των διαδικασιών που χρησιμοποιήθηκαν και τη δυναμική των προσεγγίσεων της κατηγοριοποίησης κειμένων με τη χρήση τεχνικών μηχανικής μάθησης, καθώς και όλων των άλλων μεθοδολογιών και τεχνικών που αναφέρθηκαν.

Κεφάλαιο 1: Μέσα Κοινωνικής Δικτύωσης

1.1. Κοινωνικά Δίκτυα και Μέσα Κοινωνικής Δικτύωσης

Τα κοινωνικά δίκτυα (social networks) είναι μία από τις εφαρμογές του διαδικτύου και του web 2.0 που συνέβαλαν στη ραγδαία αύξηση του κοινού του. Πρόκειται για δίκτυα τα οποία μοντελοποιούν οντότητες (ανθρώπους ή φορείς) και τις μεταξύ τους σχέσεις. Όπως κάθε δίκτυο, αποτελούνται από ένα σύνολο παραγόντων (κόμβους) και συνδέσεων μεταξύ αυτών (ακμές). Υπό το πρίσμα αυτό, τα δίκτυα των οποίων οι κόμβοι αντιστοιχούν σε ανθρώπους ή φορείς ανθρώπων και οι ακμές στις μεταξύ τους σχέσεις, μπορούν να οριστούν ως κοινωνικά. Η δομή τους καθορίζει για κάθε κόμβο το σύνολο των κοινωνικών σχέσεων που τον περιβάλλουν σε διαφορετικά επίπεδα της κοινωνικής του ζωής (οικογένεια, γειτονιά, χώρος εργασίας, σχολείο κ.λπ.). Η ποιότητα των σχέσεων μεταξύ των οντοτήτων είναι εφικτό να αποτυπώνονται στις ακμές των δικτύων, καθορίζοντας την ισχύ τους. Γενικά οι ισχυρότεροι κοινωνικοί δεσμοί χαρακτηρίζονται από έντονη υποστήριξη, ενώ οι ασθενέστεροι αποτελούν πηγές πληροφόρησης για θέματα που άπτονται της ποιότητας του ατόμου. Οι ισχυροί δεσμοί γύρω από ένα άτομο τείνουν να δημιουργούν κλίκες[1], δημιουργώντας πυκνές συνιστώσες των δικτύων, ενώ οι ασθενέστεροι δεσμοί σχηματίζουν αραιά υποδίκτυα, τα οποία μπορεί να δημιουργούν μακρά μονοπάτια, εμπλουτίζοντας τις διαθέσιμες πληροφορίες για τους κόμβους (Breslin, Passant, & Decker, 2009).

Τα μέσα κοινωνικής δικτύωσης (social media) είναι οι ηλεκτρονικές πλατφόρμες πάνω στις οποίες βρίσκουν εφαρμογή τα κοινωνικά δίκτυα. Αναπτύσσονται με τη μορφή διαδικτυακών εφαρμογών και για τον λόγο αυτό είναι προσβάσιμες σε παγκόσμια κλίμακα. Το χαρακτηριστικό τους αυτό τις κάνει ικανές να προσφέρουν στους χρήστες τους άμεση και δωρεάν αλληλεπίδραση με άλλους ανθρώπους, οπουδήποτε διατίθεται πρόσβαση στο διαδίκτυο. Η υιοθέτηση του υπήρξε ταχύτατη από το καταναλωτικό κοινό, με αποτέλεσμα μεγάλο πλήθος

[1] Ως κλίκα στη θεωρία των γράφων περιγράφεται ένα μέρος του γράφου το οποίο περιλαμβάνει ένα υποσύνολο των κόμβων του που συνθέτουν ένα πλήρη γράφου (δηλαδή κάθε κόμβος του υποσυνόλου συνδέεται με κάθε άλλο κόμβο του υποσυνόλου).

ανθρώπων να διασυνδεθεί με αυτά, σε όλον τον κόσμο. Σημαντικό ρόλο σε αυτό έπαιξε και η παράλληλη ανάπτυξη των έξυπνων κινητών συσκευών, καθώς αποτέλεσαν ένα εύκολα φερόμενο και χρησιμοποιούμενο μέσο πρόσβασης στα μέσα κοινωνικής δικτύωσης. Στη σύγχρονη εποχή πολλοί άνθρωποι βασίζουν τη δραστηριότητά τους, στους λογαριασμούς που διατηρούν σε διάφορα μέσα κοινωνικής δικτύωσης (Lubbers, Verdery και Molina (Breslin, Passant, & Decker, 2009) (Breslin, Passant, & Decker, 2009) 2018).

Τα μέσα κοινωνικής δικτύωσης αποτελούν πλούσιες πηγές δεδομένων χρήσιμων για κάθε σκοπιμότητα. Όπως σε κάθε εφαρμογή του web 2.0, έτσι και στα μέσα κοινωνικής δικτύωσης, οι χρήστες συμμετέχουν δημιουργία και τη δημοσίευση του περιεχομένου τους. Στην πραγματικότητα οι ίδιοι οι χρήστες είναι οι παραγωγοί του συντριπτικά μεγαλύτερου μέρους του περιεχομένου. Οι χρήστες των μέσων κοινωνικής δικτύωσης μπορεί να δημοσιεύουν δεδομένα που σχετίζονται με κάθε πτυχή της ζωής τους (προσωπική, επαγγελματική, κοινωνική, οικονομική) ή τις απόψεις τους, τις προτιμήσεις τους και τις προσδοκίες τους. Από όλα αυτά είναι δυνατό να προκύπτουν μετά από κατάλληλη επεξεργασία γενικότερες τάσεις. Αυτό συμβαίνει γιατί σήμερα τα μέσα κοινωνικής δικτύωσης έχουν αλλάξει τις κοινωνικές σχέσεις και τις έχουν τοποθετήσει στο εικονικό πλαίσιο των υλοποιήσεών τους. Μέσω αυτών επιτυγχάνεται όχι μόνο η επικοινωνία με τους αγαπημένους μας, τους φίλους μας ή τους συνεργάτες μας, ακόμα κι αν βρίσκονται χιλιόμετρα μακριά μας, αλλά αποτελούν πλέον ίσως το σημαντικότερο μέσω άμεσης ενημέρωσης. Έτσι τα παραδοσιακά μέσα μαζικής ενημέρωσης έχουν αντικατασταθεί από τις αναρτήσεις ειδήσεων και των σχολίων πάνω σ’ αυτά στις πλατφόρμες των μέσων κοινωνικής δικτύωσης (Ellwardt, και συν. 2020).

Η κυριότερη δυνατότητα των μέσων κοινωνικής δικτύωσης που εκμεταλλεύονται οι χρήστες τους, είναι η παροχή εναλλακτικών τρόπων επικοινωνίας. Μπορούν να επικοινωνούν μεταξύ τους με:

- Αναρτήσεις στις διαδικτυακές εφαρμογές που υλοποιούν τα μέσα κοινωνικής δικτύωσης.
- Σχόλια στις παραπάνω αναρτήσεις.

- Σύγχρονα μηνύματα
- Ασύγχρονα μηνύματα

Σε κάθε επικοινωνία έχουν τη δυνατότητα να συνθέτουν τα μηνύματα τους ως υπερκείμενα (συνδυασμός κειμένου, εικόνας, ήχου, video). Επιπλέον, η επικοινωνία μπορεί να είναι ιδιωτική (το περιεχόμενο των μηνυμάτων να είναι ορατό μόνο από ένα μικρό σχετικά πλήθος χρηστών που συμμετέχουν στην επικοινωνία), ή δημόσια (το περιεχόμενο της επικοινωνίας είναι ευρύτερα ορατό σε όλους τους χρήστες του δικτύου ή σε ένα ευρύ μέρος αυτού).

Συνοψίζοντας, τα μέσα κοινωνικής δικτύωσης αποτελούν για το σύγχρονο άνθρωπο έναν πλήρη χώρο έκφρασης της κοινωνικότητάς του, της άμεσης διάδοσης των απόψεών του, των ιδεών του, των δράσεων και των συναισθημάτων του. Ταυτόχρονα είναι και ισχυρό εργαλείο επικοινωνίας για τους οργανισμούς κάθε είδους για τη συλλογή δεδομένων που σχετίζονται με τις τάσεις του καταναλωτικού κοινού.

1.2. Ιστορικό Εξέλιξης των Μέσων Κοινωνικής Δικτύωσης

Η εμφάνιση του Usenet, το 1979, μπορεί να προσδιοριστεί σαν ο πρόδρομος των σύγχρονων μέσων κοινωνικής δικτύωσης. Πρόκειται για ένα παγκόσμιο κατανεμημένο σύστημα συζήτησης στο Διαδίκτυο. Οι χρήστες του μπορούσαν να διαβάσουν και να δημοσιεύσουν μηνύματα σε μία ή περισσότερες κατηγορίες. Αυτές οι ομάδες ήταν γνωστές ως ομάδες συζήτησης. Το Usenet προέκυψε από την ανάγκη να εμπλουτιστούν οι δυνατότητες των Bulletin Board Systems (BBS)[2]. Το Usenet έδωσε τη δυνατότητα στους χρήστες του να αποθηκεύουν και προωθούν μηνύματα ο ένας στον άλλο.

Το 1984 παρουσιάστηκε το LISTSERV, το οποίο κλιμάκωσε την επικοινωνία μέσω email με έμφαση στην προσέγγιση μεγάλου αριθμού ατόμων με ένα μήνυμα. Πριν από το LISTSERV, η διαχείριση των λιστών email γινόταν με μη αυτόματο τρόπο.

[2] Ένα σύστημα πίνακα ανακοινώσεων (BBS) είναι εφαρμογή που στοχεύει στην κοινή χρήση ή ανταλλαγή μηνυμάτων ή άλλων αρχείων σε ένα δίκτυο. Το BBS χρησιμοποιήθηκε για τη δημοσίευση απλών μηνυμάτων σε κοινότητες χρηστών.

Οι άνθρωποι έπρεπε να γράψουν στον διαχειριστή που διαχειριζόταν τη λίστα και να ζητήσουν να προστεθούν ή να αφαιρεθούν επαφές. Η αυτοματοποίηση της δημιουργίας των λιστών επαφών απλοποίησε τη διαχείριση μαζικής αποστολής μηνυμάτων. Το 1988 παρουσιάστηκε το Internet Relay Chat (IRC) που αποτελεί μία δυνατότητα συνομιλίας χρηστών μέσω του διαδικτύου. Αρχικά ήταν σχεδιασμένο για ομαδική συνομιλία σε φόρουμ συζητήσεων επιτρέποντας έτσι την επικοινωνία ενός προς έναν μέσω προσωπικού μηνύματος, καθώς και τη μεταφορά δεδομένων. Το 2009, τα κορυφαία 100 δίκτυα IRC εξυπηρετούσαν περισσότερους από 500000 χρήστες τη φορά, με εκατοντάδες χιλιάδες κανάλια, που λειτουργούσαν σε περίπου 1.500 διακομιστές παγκοσμίως. Το IRC χρησιμοποιήθηκε για την αναφορά σχετικά με την απόπειρα πραξικοπήματος της Σοβιετικής Ένωσης το 1991 κατά τη διάρκεια ενός μπλακ άουτ μέσων ενημέρωσης. Παλαιότερα χρησιμοποιήθηκε με παρόμοιο τρόπο κατά τη διάρκεια του Πολέμου του Κόλπου. Ακόμα και σήμερα, το λογισμικό για IRC επικοινωνία είναι διαθέσιμο σχεδόν για κάθε σύστημα που λειτουργεί με υπολογιστή ο οποίος υποστηρίζει δικτύωση TCP/IP. Στις αρχές της δεκαετίας του '90 η πρόσβαση στο διαδίκτυο δεν ήταν εύκολη για το ευρύ κοινό. Η λειτουργία των παρόχων υπηρεσιών διαδικτύου (Internet Service Providers - ISP), διευκόλυνε την πρόσβαση στο διαδίκτυο. Αυτό είχε σαν συνέπεια την σταδιακή αύξηση του κοινού του διαδικτύου. Η δωρεάν διαθεσιμότητα του περιεχομένου του ήταν ένας ακόμα λόγος για την υιοθέτηση των δικτυακών εφαρμογών προβολής περιεχομένου. Τα Forum ήταν οι διαδικτυακές εφαρμογές της εποχής εκείνης, που έδιναν στους χρήστες τους τη δυνατότητα να δημοσιεύουν το δικό τους περιεχόμενο, το οποίο ωστόσο διαχειρίζονταν μερικοί ή γενικοί διαχειριστές.

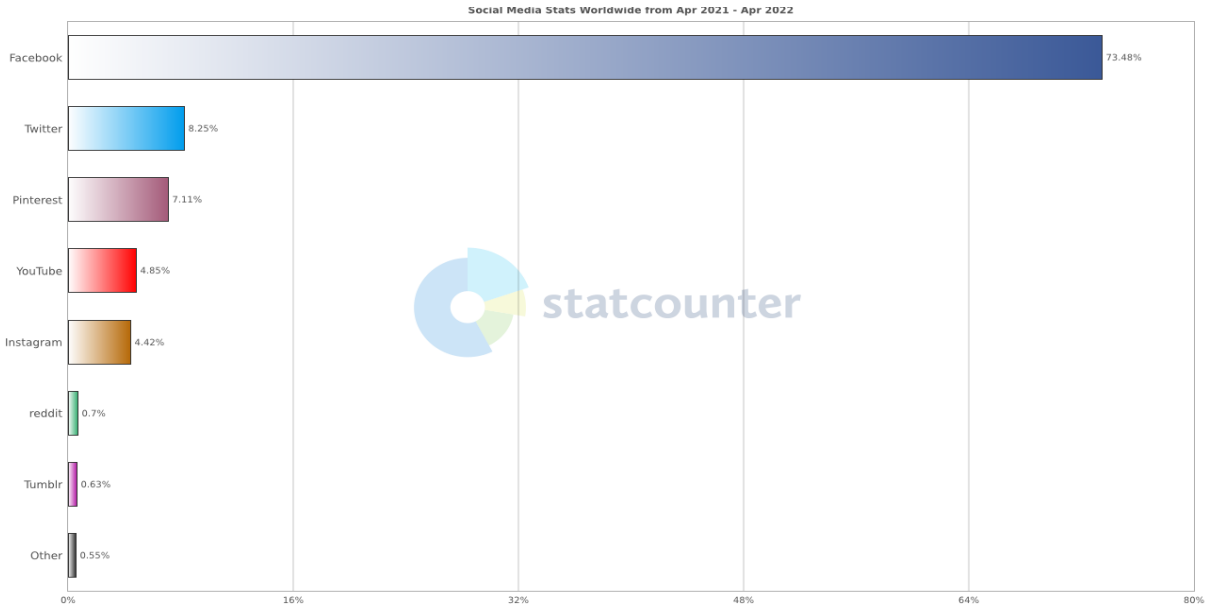
Το 1995 παρουσιάστηκε ο δικτυακός τόπος classmate.com στον οποίον οι εγγεγραμμένοι χρήστες μπορούσαν να αναζητήσουν γνωστούς από το παρελθόν τους, ανάμεσα στους υπόλοιπους χρήστες. Το classmates.com λειτουργεί ακόμα και αριθμεί σχεδόν 50 εκατομμύρια μέλη. Ένα ακόμα είδος διαδικτυακών εφαρμογών, του οποίου τα μέσα κοινωνικής δικτύωσης μπορεί να θεωρηθούν εξέλιξη τους, είναι τα blogs. Σε αυτά, ο διαχειριστής τους έδινε δικαιώματα στους χρήστες τους να αναρτούν περιεχόμενο υπερκειμένου ή/και να σχολιάζουν το υπάρχον περιεχόμενο.

Τα blogs μπορούν να φιλοξενούνται από αποκλειστικές υπηρεσίες φιλοξενίας ή σε δικτυακούς τόπους γενικής χρήσης. Η χρήση τους εξαπλώθηκε μετά το τέλος της δεκαετίας του 1990. Το 1998, το Open Dairy ήταν το πρώτο blog που επέτρεψε στους επισκέπτες του να προσθέσουν σχόλια στις καταχωρήσεις του. Το 1999, παρουσιάστηκε το blogger.com, που έδινε την δυνατότητα στους χρήστες του να δημιουργούν δικά τους blogs.

Οι εφαρμογές Web2.0 ενίσχυσαν τη διάδραση του χρήστη με τους δικτυακούς τόπους. Ο όρος Web 2.0 συνδέεται συνήθως με εφαρμογές Ιστού που δίνουν την ευκαιρία στο χρήστη να συμμετέχει στη διαμόρφωση του περιεχομένου μέσα από την επικοινωνία με άλλους χρήστες μέσα από διάδραση. Σε αυτό το πλαίσιο τοποθετούνται τα σύγχρονα μέσα κοινωνικής δικτύωσης (Sajithra και Patil 2013). Τα πρώτα μέσα κοινωνικής δικτύωσης με τη σύγχρονη μορφή τους εμφανίστηκαν στα μέσα δεκαετίας του 2000 (Facebook το 2004, YouTube το 2005 και Myspace το 2006).

1.3. Δημοφιλή Μέσα Κοινωνικής Δικτύωσης

Στην εμφάνιση των κοινωνικών δικτύων, το καταναλωτικό κοινό ανταποκρίθηκε, όπως αναφέρθηκε και πιο πάνω, θετικά με αποτέλεσμα να πολλαπλασιαστεί σε πολύ μικρό χρονικό διάστημα το διαδικτυακό κοινό. Αυτό δημιούργησε χώρο για την ανάπτυξη νέων μέσων κοινωνικής δικτύωσης . Από την εποχή της εμφάνισης των πρώτων μέσων κοινωνικής δικτύωσης μέχρι σήμερα, έχουν παρουσιαστεί πολλές πλατφόρμες γενικού ή ειδικού σκοπού. Βασικό κριτήριο για την επιτυχία τους αποτελεί το πλήθος των ενεργών χρηστών που διατηρούν λογαριασμούς σε αυτές και ο όγκος του περιεχομένου που φιλοξενούν στις ιστοσελίδες τους. Με βάση τα κριτήρια αυτά, στους πιο επιτυχημένους και σημαντικούς παρόχους υπηρεσιών κοινωνικής δικτύωσης περιλαμβάνονται το Twitter, το Facebook, το Pinterest, το YouTube και το Instagram. Στο παρακάτω διάγραμμα φαίνεται η ποσοστιαία επισκεψιμότητα των μέσων κοινωνικής δικτύωσης το τελευταίο έτος.



Εικόνα 1 Ποσοστιαία επισκεψιμότητα των κυριότερων μέσων κοινωνικής δικτύωσης (πηγή: https://japantoday.com/category/tech/japan-one-of-only-two-countries-where-twitter-beats-facebook-in-social-media-market-share-1)

1.3.1. Facebook

Το Facebook είναι μία πλατφόρμα κοινωνικής δικτύωσης που αναπτύχθηκε το 2004 από μία ομάδα φοιτητών του Χάρβαρντ για χρήση εντός των ορίων του Πανεπιστημίου. Γρήγορα η χρήση του επεκτάθηκε σε μικρότερους μαθητές. Από τον Ιούλιο του 2010 και μετά έχει περισσότερους από 500 εκατομμύρια ενεργούς χρήστες. Οι χρήστες μπορούν να δημιουργήσουν ένα προσωπικό προφίλ, να προσθέσουν άλλους χρήστες ως φίλους και να ανταλλάξουν μηνύματα, να ενημερώνονται με αυτόματες ειδοποιήσεις για δραστηριότητα χρηστών του δικτύου τους, για ανάρτηση φωτογραφιών και σχολίων. Επιπλέον, οι χρήστες του Facebook μπορούν να συμμετέχουν σε ομάδες χρηστών κοινού ενδιαφέροντος, οργανωμένες ανά χώρο εργασίας, σχολείο, κολέγιο ή άλλα χαρακτηριστικά.

1.3.2. Twitter

Το 2006 παρουσιάστηκε το Twitter που πρόσφερε ποικίλες δυνατότητες στους χρήστες του όπως το micro blogging. Η επιτυχία του βασίστηκε κυρίως στην

επιλογή πολλών διάσημων να το επιλέξουν προκειμένου να προβάλλουν τις θέσεις τους. Το Twitter χρησιμοποιεί έναν σκόπιμο περιορισμό μεγέθους μηνύματος για να διευκολύνει την αναζήτηση περιεχομένου. Οι χρήστες μπορούν να δημοσιεύουν αναρτήσεις εντάσσοντας τις σε μία ή περισσότερες θεματολογίες (και αυτός ο μηχανισμός διευκολύνει τις αναζητήσεις περιεχομένου). Η ευκολία αναζήτησης περιεχομένου και η ποικιλία της θεματολογίας των αναρτήσεων του, καθιστά το twitter μία αξιόπιστη λύση για την εκτίμηση των τάσεων της κοινής γνώμης (Gil 2021).

1.3.3. YouTube

Το YouTube, ιδρύθηκε το 2005 και αποτελεί πλέον την πιο δημοφιλή διαδικτυακή κοινότητα διαμοιρασμού βίντεο. Παρέχει ένα χώρο όπου οι χρήστες του διαδικτύου μπορούν να καταχωρούν ή να προβάλουν βίντεο που μπορεί να ανήκουν σε μία ποικιλία θεματολογίας. Τα βίντεο που φιλοξενούνται στις υποδομές του μπορεί να είναι είτε ερασιτεχνικά είτε επαγγελματικά. Αναπτύχθηκε γρήγορα και τον Ιούλιο του 2006 ανέβαιναν περισσότερα από 65.000 νέα βίντεο καθημερινά, ενώ πραγματοποιούνταν και 100 εκατομμύρια προβολές. Οι προβολές των βίντεο στο YouTube αποτελούν σημείο αναφοράς για την επιτυχία ταινιών, μουσικών κομματιών και γενικότερα κάθε δημιουργία που μπορεί να παρουσιαστεί με τη μορφή βίντεο.

1.3.4. Reddit

Το Reddit διαχωρίζει το περιεχόμενο του σε περισσότερες από ένα εκατομμύριο κοινότητες γνωστές ως "subreddits", καθεμία από τις οποίες καλύπτει ένα διαφορετικό θέμα. Η αρχική σελίδα εμφανίζει στο χρήστη διάφορες αναρτήσεις που είναι την χρονική αυτή συγκυρία δημοφιλείς. Η εφαρμογή βάζει περιορισμούς στη δυνατότητα για αναρτήσεις, ώστε να εξασφαλίζεται ότι μόνο ενεργοί χρήστες συμβάλλουν στο περιεχόμενό του. Τα Subreddits τα διαχειρίζονται εθελοντές που μπορούν να επεξεργαστούν την εμφάνιση ενός συγκεκριμένου subreddit, να υπαγορεύσουν τους τύπους περιεχομένου που επιτρέπονται στο subreddit και να

αφαιρέσουν αναρτήσεις ή περιεχόμενο ή να αποκλείσουν τους χρήστες από το subreddit. Το Reddit στο σύνολό του διοικείται από τους διαχειριστές, τους υπαλλήλους του Reddit που έχουν ισχυρά δικαιώματα στον ιστότοπο. Οι χρήστες μπορούν και να αποδοκιμάζουν αναρτήσεις (Widman 2021).

1.3.5. Instagram

Το Instagram είναι μία πλατφόρμα κοινωνικής δικτύωσης με βασικό χαρακτηριστικό των περιεχομένων του τα πολυμέσα. Παρέχει, όπως και άλλα μέσα κοινωνικής δικτύωσης, εναλλακτικές μορφές επικοινωνίας ή μηχανισμούς αξιολόγησης των αναρτήσεων άλλων μελών.

1.4. Διεισδυτικότητα και Χρήση

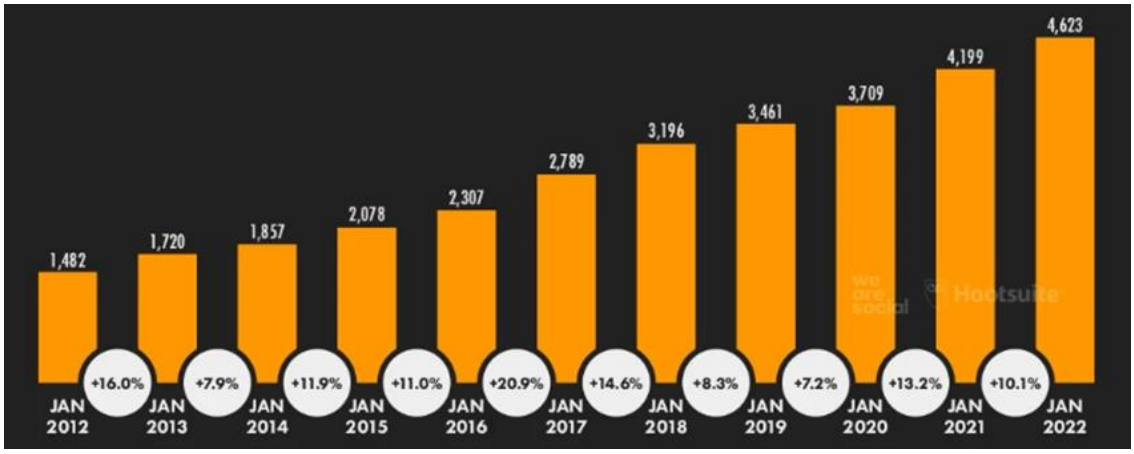
Η διεισδυτικότητα των μέσων κοινωνικής δικτύωσης καταδεικνύεται από τα σχετικά στατιστικά στοιχεία. Σχεδόν το 60% του παγκόσμιου πληθυσμού χρησιμοποιεί τα μέσα κοινωνικής δικτύωσης, ενώ εκτιμάται ότι περίπου 424 εκατομμύρια νέοι χρήστες έχουν συνδεθεί σε κάποιο κοινωνικό δίκτυο το τελευταίο έτος. Οι χρήστες των μέσων κοινωνικών δικτύων παραμένουν σε αυτά για περίπου 2,5 ώρες ημερησίως. Ο ετήσιος ρυθμός αύξησης του αριθμού των χρηστών των μέσων κοινωνικής δικτύωσης είναι 1%. Η μέση ηλικία τους είναι 31 ετών (στην Ελλάδα η μέση ηλικία είναι περίπου 46 ετών). Σχεδόν όλοι οι χρήστες εισέρχονται στις πλατφόρμες από το κινητό τους τηλέφωνο. Όπως φαίνεται στο παρακάτω διάγραμμα, οι πλατφόρμες κοινωνικής δικτύωσης είναι από τους πιο δημοφιλείς δικτυακούς τόπους.

#	WEBSITE	TOTAL VISITS	UNIQUE VISITORS	TIME PER VISIT	PAGES PER VISIT
01	GOOGLE.COM	45.41B	2.98B	21M 11S	5.74
02	YOUTUBE.COM	14.34B	1.70B	7M 43S	3.70
03	FACEBOOK.COM	11.74B	1.53B	22M 15S	5.97
04	WIKIPEDIA.ORG	5.97B	1.39B	10M 35S	2.11
05	AMAZON.COM	3.13B	0.68B	13M 11S	7.28
06	INSTAGRAM.COM	3.08B	0.74B	18M 12S	4.79
07	YAHOO.COM	2.63B	0.41B	17M 14S	3.99
08	YANDEX.RU	2.43B	0.19B	23M 32S	6.51
09	TWITTER.COM	2.43B	0.62B	14M 46S	4.45
10	PORNHUB.COM	2.29B	0.40B	14M 50S	8.32

#	WEBSITE	TOTAL VISITS	UNIQUE VISITORS	TIME PER VISIT	PAGES PER VISIT
11	REDDIT.COM	2.22B	0.39B	21M 58S	4.36
12	NAVER.COM	2.20B	0.11B	10M 44S	11.01
13	XVIDEOS.COM	2.13B	0.34B	18M 29S	8.79
14	BIT.LY	2.11B	0.82B	12M 12S	1.21
15	VK.COM	1.64B	0.18B	23M 20S	9.60
16	LIVE.COM	1.60B	0.32B	9M 15S	4.01
17	XNXX.COM	1.39B	0.24B	18M 23S	8.74
18	FANDOM.COM	1.28B	0.31B	12M 18S	3.13
19	YAHOO.CO.JP	1.23B	0.06B	13M 51S	6.22
20	TWITCH.TV	1.22B	0.14B	6M 28S	2.33

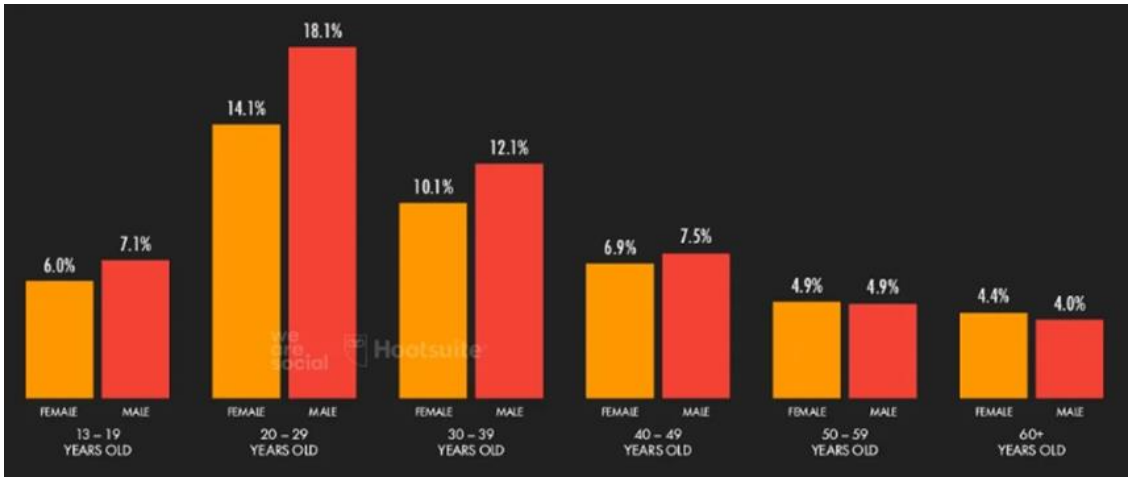
Εικόνα 2 Κατάταξη επισκεψιμότητας δικτυακών τόπων (πηγή: https://datareportal.com/reports/digital-2022-worlds-top-websites)

Η αύξηση του αριθμού των χρηστών των μέσων κοινωνικής δικτύωσης υπήρξε ραγδαία τα τελευταία δέκα χρόνια. Στην παρούσα χρονική συγκυρία φαίνεται να εμφανίζονται σταθεροποιητικές τάσεις στον αριθμό τους, με την δυναμική περαιτέρω αύξησης να διατηρείται. Στο παρακάτω γράφημα φαίνεται η εξέλιξη του πλήθους των ενεργών χρηστών από το 2012 μέχρι τις αρχές του 2022 (σε εκατομμύρια ανθρώπους). Το πλήθος των ενεργών χρηστών τους φαίνεται να ξεπερνά τα 4,6 δισεκατομμύρια ανθρώπους.



Εικόνα 3 Πλήθος χρηστών μέσω κοινωνικής δικτύωσης ανά έτος (πηγή: https://www.bebekindex.com/%CF%84%CE%B9%CE%BC%CE%B5%CF%82-%CF%83%CF%84%CF%81%CF%89%CE%BC%CE%B1%CF%84%CF%89%CE%BD-media)

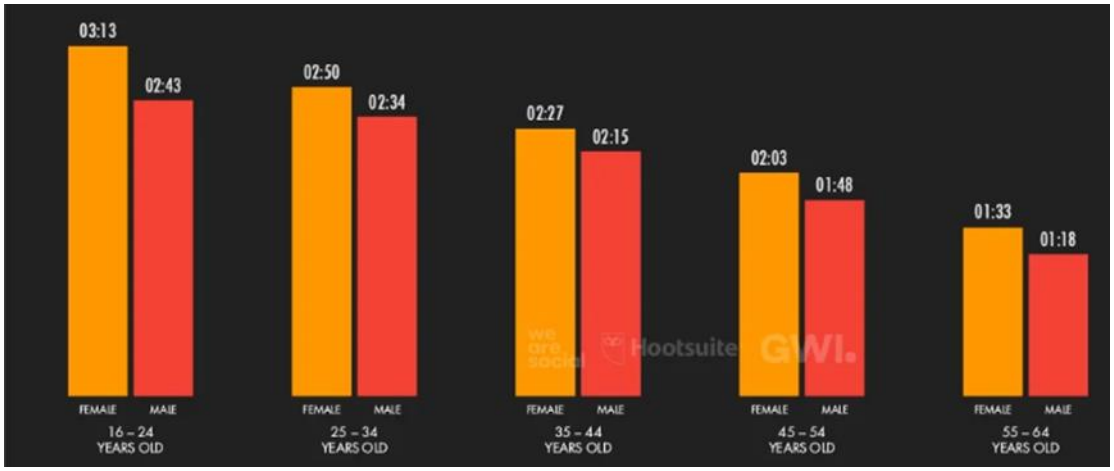
Η διεισδυτικότητα των μέσων κοινωνικής δικτύωσης είναι υψηλή σε ανθρώπους κάθε ηλικίας. Οι άνθρωποι ηλικίας 20 με 40 ετών αποτελούν πάνω από το 25% του συνόλου των χρηστών, όπως φαίνεται και από το παρακάτω γράφημα.



Εικόνα 4 Διείσδυση των μέσων κοινωνικής δικτύωσης ανά ηλικιακή ομάδα (πηγή: https://digimind.id/demografis-penting-di-instagram-yang-marketer-wajib-ketahui/)

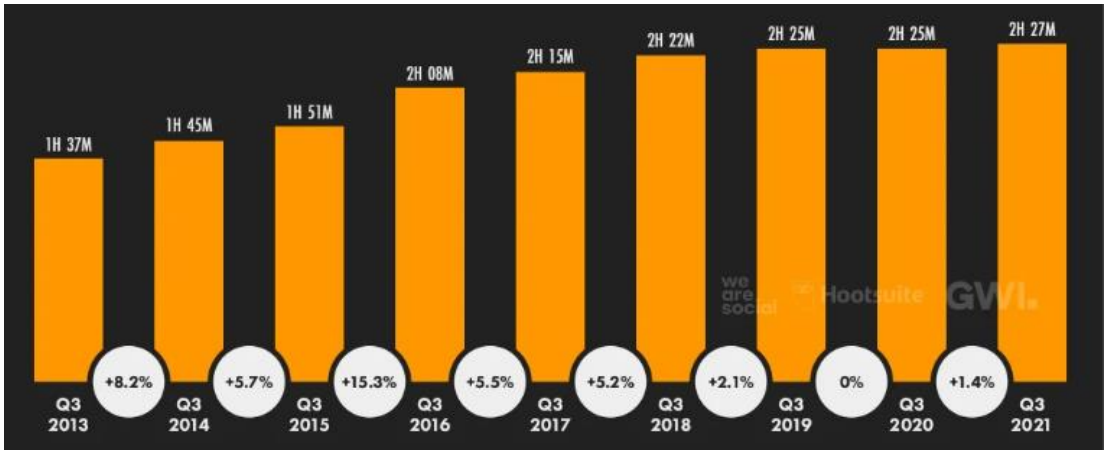
Το ποσοστό των ανηλίκων χρηστών περιορίζεται από γονικές παρεμβάσεις στη χρήση των συσκευών. Το αντίστοιχο ποσοστό των ανθρωπίνων ηλικίας από 40 ετών και πάνω, είναι πιο χαμηλό. Εκτιμάται ότι αυτό κυρίως οφείλεται στον χαμηλότερο βαθμό οικειότητας των ανθρώπων μεγαλύτερης ηλικίας με τη χρήση των διαδικτυακών εφαρμογών. Εκτιμάται ότι στο κοντινό μέλλον, οι άνθρωποι όλων των ηλικιακών ομάδων θα είναι επαρκώς εξοικειωμένοι με την χρήση των εφαρμογών κοινωνικής δικτύωσης και η κατανομή τους σε ηλικιακές ομάδες θα ομαλοποιηθεί.

Μία άλλη παράμετρος που παρουσιάζει ενδιαφέρον είναι ο χρόνος που σπαταλούν οι χρήστες κάθε ηλικιακής ομάδας στη διάδραση με τις εφαρμογές κοινωνικής δικτύωσης. Οι νεότεροι άνθρωποι, όπως φαίνεται στο παρακάτω γράφημα, έχουν την τάση να σπαταλούν περισσότερο χρόνο στη χρήση εφαρμογών κοινωνικής δικτύωσης σε σχέση με τους μεγαλύτερους.



Εικόνα 5 Χρόνος που σπαταλιέται στην χρήση μέσων κοινωνικής δικτύωσης ανά ηλικιακή ομάδα (πηγή: https://heidicohen.com/social-media-facts-2022/)

Στην προηγούμενη δεκαετία υπήρξε μία σταθερή αύξηση του χρόνου απασχόλησης των ανθρώπων με τις εφαρμογές κοινωνικής δικτύωσης. Στο παρακάτω γράφημα φαίνεται ότι η αύξηση αυτή ήταν πιο έντονη μέχρι το 2018 ενώ τα τελευταία χρόνια φαίνεται ο μέσος χρόνος ενασχόλησης να σταθεροποιείται κοντά στις 2,5 ώρες ημερησίως.



Εικόνα 6 Ημερήσια απασχόληση με τα μέσα κοινωνικής δικτύωσης ανά έτος (πηγή: https://dantri.com.vn/giao-duc-huong-nghiep/2022-tiep-tuc-la-nam-cua-nhung-nha-sang-tao-20220305101302945.htm)

Βασικός λόγος για το φαινόμενο αυτό είναι η ενσωμάτωση καινοτόμων δυνατοτήτων, η οποία τα τελευταία χρόνια φαίνεται να φτάνει σε έναν κορεσμό (Datareportal 2022).

Από τα παραπάνω προκύπτουν τα ακόλουθα συμπεράσματα:

- Η διεισδυτικότητα των μέσων κοινωνικής δικτύωσης είναι υψηλή σε ανθρώπους κάθε ηλικίας.
- Οι χρήστες των μέσων κοινωνικής δικτύωσης ξοδεύουν περίπου 2,5 ώρες καθημερινά χρησιμοποιώντας τις εφαρμογές τους.
- Ο όγκος του συνόλου των ανθρώπων που έχει πρόσβαση στο περιεχόμενο και ο χρόνος που ασχολούνται με αυτό, το καθιστά σημαντικό για τη διαμόρφωση της κοινής γνώμης.
- Η κατάθεση απόψεων και θέσεων από ένα τόσο ευρύ σύνολο ανθρώπων, καθιστά τις πλατφόρμες κοινωνικής δικτύωσης ένα ισχυρό εργαλείο ανάδειξης των τάσεων της κοινής γνώμης και του καταναλωτικού κοινού.
- Το ποσοστό των ανθρώπων που επιθυμούν να έχουν πρόσβαση στις εφαρμογές κοινωνικής δικτύωσης τα τελευταία χρόνια διατηρεί ελαφρά αυξητικές τάσεις. Αυτό σε συνδυασμό με τον μεγάλο όγκο συνόλου χρηστών σε απόλυτους αριθμούς, δείχνει ότι στο – τουλάχιστον άμεσο – μέλλον, μεγάλο ποσοστό των ανθρώπων θα ασχολείται με την διαμόρφωση και ανάγνωση του περιεχομένου των μέσων κοινωνικής δικτύωσης.

Γίνεται αντιληπτό πως η μελέτη του περιεχομένου των μέσων κοινωνικής δικτύωσης μπορεί να προσφέρει σημαντικά πλεονεκτήματα στην προσπάθεια ανάλυσης των τάσεων της κοινής γνώμης σε διαφορετικά επίπεδα της ανθρώπινης δραστηριότητας. Καθώς το περιεχόμενό τους απηχεί τις απόψεις και ανησυχίες του μεγαλύτερου μέρους του παγκόσμιου πληθυσμού και παραμένει συγκεντρωμένο σε μία σειρά από συγκεκριμένες πλατφόρμες μπορεί μέσω κατάλληλων διαδικασιών επεξεργασίας του, να αποδίδει χρήσιμα και αξιόπιστα συμπεράσματα. Επιπλέον, το γεγονός ότι κάποιες από τις πλατφόρμες κοινωνικής δικτύωσης είναι προσανατολισμένες σε συγκεκριμένες πτυχές της ανθρώπινης δραστηριότητας (π.χ. επαγγελματική δραστηριότητα, κοινωνικές σχέσεις) και ότι ακόμα και στην ίδια πλατφόρμα οι χρήστες μπορούν να οργανώνονται σε κοινότητες ή ο προσανατολισμός των περιεχομένων να είναι κατηγοριοποιημένος, διευκολύνει τη διεξαγωγή στοχευμένων ερευνών με βάση το περιεχόμενό τους.

Κεφάλαιο 2: Μηχανική Μάθηση

2.1. Οι απαιτήσεις που οδήγησαν στη Μηχανική Μάθηση

Από τη δεκαετία του 2000, η ποσοτική και ποιοτική εξέλιξη του διαδικτύου και των εφαρμογών του ήταν ραγδαία. Οι εφαρμογές του καλύπτουν πλέον το μεγαλύτερο φάσμα των ανθρώπινων δραστηριοτήτων. Αυτό έκανε περισσότερους ανθρώπους να ανακαλύψουν τις δυνατότητές του και να εμπιστευτούν τις εφαρμογές του. Παράλληλα η πρόσβασή τους διευκολύνθηκε από την ανάπτυξη των δικτυακών τεχνολογιών. Σήμερα μεγάλο μέρος του παγκόσμιου πληθυσμού έχει ευρυζωνική πρόσβαση στο διαδίκτυο. Εκτός αυτού, οι χρήστες των διαδικτυακών εφαρμογών, μπορούν να αποκτούν πρόσβαση σε αυτές, όχι μόνο από ηλεκτρονικούς υπολογιστές, αλλά και από άλλου είδους συσκευές, όπως είναι τα έξυπνα κινητά τηλέφωνα και οι συσκευές του διαδικτύου των πραγμάτων (Internet of Things). Γενικότερα, ο σύγχρονος άνθρωπος σπαταλά αρκετό χρόνο χρησιμοποιώντας διαδικτυακές εφαρμογές παρέχοντας και αναζητώντας περιεχόμενο κατά τη διάδραση με αυτές.

Η παραγωγή μεγάλου όγκου δεδομένων με ποικίλη θεματολογία και μορφή, οδηγεί σε δυσκολία εκμετάλλευσής του με τις παραδοσιακές τεχνικές, μεθοδολογίες και εργαλεία. Ενώ παλαιότερα το μείζον ζήτημα ήταν η ανεύρεση περιεχομένου, στη σύγχρονη εποχή, που η διαθεσιμότητα μεγάλου όγκου περιεχομένου είναι δεδομένη, το μείζον ζήτημα είναι να καταστεί το διαθέσιμο περιεχόμενο εκμεταλλεύσιμο. Για να είναι τα δεδομένα εκμεταλλεύσιμα, θα πρέπει να ικανοποιούνται οι παρακάτω προϋποθέσεις (Sarfin, 2021):

- Να είναι ακριβή: Τα δεδομένα που θα χρησιμοποιηθούν για κάποιο σκοπό, θα πρέπει να περιγράφουν με σαφήνεια τη σημασιολογία για την οποία θα υποστούν επεξεργασία. Σε διαφορετική περίπτωση, η όποια επεξεργασία θα οδηγήσει σε εσφαλμένα ή συγκεχυμένα αποτελέσματα.
- Να αποκτώνται έγκαιρα: Η σύγχρονη εποχή χαρακτηρίζεται από ταχύτατους ρυθμούς επικαιροποίησης των δεδομένων. Σε πολλές

περιπτώσεις δεδομένα που αποκτώνται είναι πλέον χωρίς αξία. Επομένως θα πρέπει να εξασφαλίζεται ότι τα δεδομένα αποκτώνται σε χρονική στιγμή που η επεξεργασία τους θα αποδώσει ικανοποιητικό αποτέλεσμα.

- Είναι κατάλληλα: Η σημασιολογία τους, η ποιότητα και ο όγκος τους είναι τέτοια που τα καθιστά ικανά να χρησιμοποιηθούν για την εκπλήρωση του στόχου της επεξεργασίας τους.

Η επεξεργασία των μεγάλων όγκων δεδομένων με παραδοσιακές μεθόδους δεν είναι πλέον αποδοτική. Αυτό οφείλεται στην αδυναμία των μεθόδων αυτών να παράγουν γρήγορα αποτελέσματα μέσα από την επεξεργασία μεγάλου όγκου δεδομένων. Επιπλέον παρουσιάζουν αδυναμία να εντοπίσουν τις συσχετίσεις μεταξύ τους αλλά και να οδηγήσουν στην υποστήριξη λήψης αποφάσεων, που θα συνδράμουν στην ανάπτυξη των οργανισμών (Trujillo, Kim, Jones, Garcia, & Murray, 2015).

Για την αντιμετώπιση των ζητημάτων αυτών, ήταν ανάγκη να αναπτυχθούν νέες διαδικασίες και μηχανισμοί που να είναι ικανοί να παράγουν κανόνες συσχέτισης μεταξύ δεδομένων και πρότυπα πρόβλεψης μελλοντικών καταστάσεων που βασίζονται σε τρέχουσες ή παρελθούσες παρατηρήσεις. Ο επιστημονικός τομέας της τεχνητής νοημοσύνης παρείχε αποδοτικές λύσεις στα ζητήματα αυτά. Η μηχανική μάθηση είναι ένας κλάδος της, προσανατολισμένος στην παραγωγή προτύπων μέσα από την επεξεργασία μεγάλων όγκων δεδομένων. Τα πρότυπα αυτά δύνανται να αξιολογούνται και να χρησιμοποιούνται σε ποικίλες διαδικασίες λήψης αποφάσεων. Η χρήση των μεθόδων και των τεχνικών της μηχανικής μάθησης, παρέχει ισχυρά πλεονεκτήματα σε ένα ευρύ φάσμα εφαρμογών που χρησιμοποιούνται όπως:

- Αναγνώριση Εικόνας και εντοπισμός εικόνας σε αρχείο
- Εκτίμηση τάσεων του καταναλωτικού κοινού
- Αναγνώρισης ομιλίας
- Πρόβλεψη κίνησης
- Συστήματα συστάσεων - προτάσεων
- Αυτοοδηγούμενα οχήματα

- Φιλτράρισμα ανεπιθύμητων μηνυμάτων ηλεκτρονικού ταχυδρομείου και κακόβουλου λογισμικού
- Εντοπισμός διαδικτυακής απάτης

Αυτός είναι και ο λόγος για τον οποίο έχει ενταθεί τα τελευταία χρόνια η έρευνα για τη βελτίωση των μηχανισμών της και την επέκταση της χρήσης τους σε όλο και περισσότερους τομείς της ανθρώπινης δραστηριότητας.

2.2. Περιγραφή

Η Μηχανική μάθηση είναι ο τομέας της πληροφορικής που μελετά τη δυνατότητα των ηλεκτρονικών υπολογιστών να μαθαίνουν και να προσαρμόζουν τη συμπεριφορά τους, με τη χρήση προτύπων που αναπτύσσονται από την επεξεργασία δεδομένων που έχουν συλλεχθεί κατά το παρελθόν. Τα πρότυπα αυτά καθορίζουν τη συμπεριφορά των ηλεκτρονικών υπολογιστών σε καταστάσεις που περιγράφονται από τρέχοντα δεδομένα. Η διαμόρφωση και προσαρμογή της συμπεριφοράς των ηλεκτρονικών υπολογιστών στα πρότυπα ονομάζεται εκπαίδευση. Η μηχανική μάθηση είναι κατάλληλη σε περιπτώσεις εργασιών που συνήθως εκτελούνται με την ανθρώπινη παρέμβαση ενώ η χρήση λογισμικού είναι δύσκολη ή ανέφικτη. Προκρίνεται επίσης ως καταλληλότερη όταν ο όγκος των δεδομένων που θα χρησιμοποιηθούν για επεξεργασία είναι σχετικά μεγάλος. Επιπλέον, μπορούν να συμπεριλάβουν στην εξέλιξη των μηχανισμών της και το περιβάλλον μέσα στο οποίο διεξάγεται η επεξεργασία (Carbonel, Machalski and Mitchell 1983).

Η Μηχανική Μάθηση μπορεί να παρομοιαστεί με τον τρόπο που αποκτούν γνώσεις οι ζώντες οργανισμοί και κυρίως ο άνθρωπος. Το κυριότερο στοιχείο που παράγει γνώση είναι η εμπειρία. Μέσω αυτής αναπτύσσονται τα πρότυπα των συμπεριφορών στην αντιμετώπιση των διαφορετικών καταστάσεων. Τα πρότυπα αυτά αποτελούν τον οδηγό για τη λήψη εύστοχων αποφάσεων που με τη σειρά τους οδηγούν στην ανάπτυξη των οργανισμών. Ο άνθρωπος δέχεται συνεχώς ερεθίσματα από το περιβάλλον του. Ανάλογα με το είδος των ερεθισμάτων αλλά και τις συνθήκες που επικρατούν όταν τα δέχεται, προσαρμόζει τη συμπεριφορά του, αξιολογώντας κάθε φορά τα αποτελέσματα της προσαρμογής αυτής. Η διαδικασία αυτή διαμορφώνει την εμπειρία του ανθρώπου. Η εμπειρία συνιστά τη γνώση που αποκτά

ο άνθρωπος και τον βοηθά να εκτιμά το ποιες ενέργειες θα πρέπει να κάνει ώστε να επιτυγχάνει το καλύτερο αποτέλεσμα. Η εμπειρία είναι το εργαλείο για την επιτυχή διαδικασία λήψης αποφάσεων (Alzubi, 2018).

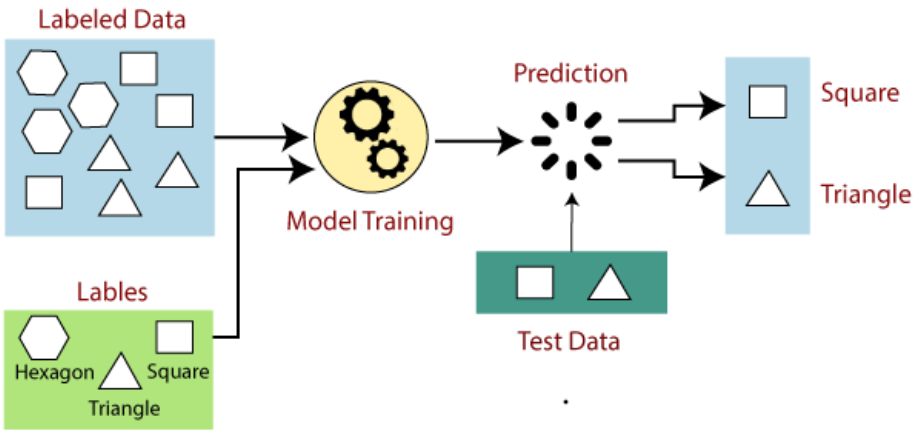
Οι σωστές αποφάσεις είναι το κρισιμότερο χαρακτηριστικό της διαχείρισης των οργανισμών στο έντονα ανταγωνιστικό παγκοσμιοποιημένο επιχειρησιακό περιβάλλον του σήμερα. Η εμπειρία αναπτύσσει πρότυπα συμπεριφοράς που βασίζονται στην ανακάλυψη της σχέσης μεταξύ αιτίου και αποτελέσματος. Η ποιότητα της εμπειρίας καθορίζει την ποιότητα των σχηματιζόμενων προτύπων. Οι μηχανές μπορούν να αποκτήσουν εμπειρία μέσα από την επεξεργασία των δεδομένων. Επομένως ο σχηματισμός των προτύπων εξαρτάται από τη διαθεσιμότητα αρκετών και ποιοτικών δεδομένων (Ge, Ding και Huang 2017). Η ανάπτυξη των σχετικών τεχνικών και των μέσων που χρησιμοποιούνται στην ανάλυση δεδομένων γίνεται τα τελευταία χρόνια με μεγάλη ταχύτητα, με αποτέλεσμα οι μηχανισμοί της μηχανικής μάθησης να γίνονται περισσότερο αποδοτικοί. Η ανάλυση των προτύπων που παράγονται, μπορούν να χρησιμοποιηθούν για την εξαγωγή σημαντικών πληροφοριών για τη συσχέτιση μεταξύ των χαρακτηριστικών που συνθέτουν τις περιγραφές των καταστάσεων. Η ανακάλυψη των συσχετίσεων μπορεί να οδηγήσει σε πληροφορίες, που χρησιμοποιούνται στην ανάπτυξη στατιστικών μοντέλων που χρησιμοποιούνται σε ποικίλες εφαρμογές στην ιατρική, στην αγορά, στην πολιτική, στη μελέτη της κοινωνίας και σε μία πληθώρα άλλων τομέων.

2.3. Κατηγορίες Μηχανικής Μάθησης

Η μεγάλη εξέλιξη των διαδικασιών και των μηχανισμών Μηχανικής Μάθησης, καθιέρωσε έναν σημαντικό αριθμό διαφορετικών προσεγγίσεων στη σχεδίαση και την υλοποίηση εφαρμογών. Οι αλγόριθμοι Μηχανικής Μάθησης μπορούν να κατηγοριοποιηθούν με διαφορετικά κριτήρια. Με βάση το είδος των προτύπων που παράγονται, διακρίνονται σε:

- Αλγόριθμοι Επιτηρούμενης μάθησης (Supervised Machine Learning): Οι αλγόριθμοι αυτής της κατηγορίας, για την εκπαίδευση των μοντέλων χρησιμοποιούνται τόσο ως δεδομένα εισόδου όσο και ως δεδομένα

εξόδου για το χαρακτηριστικό που αναζητείται κατά την διαδικασία. Αν για παράδειγμα αναζητούνται οι καιρικές συνθήκες της επόμενης ημέρας με βάση την θερμοκρασία και την υγρασία της τρέχουσας, τότε το σύνολο που θα χρησιμοποιηθεί για την παραγωγή του μοντέλου θα περιλαμβάνει δεδομένα για θερμοκρασία, υγρασία και τις αντίστοιχες καιρικές συνθήκες. Τα δεδομένα εισόδου ταξινομούνται ως προς τα δεδομένα εξόδου και η ταξινόμηση αυτή αποτελεί μία βάση για την πρόβλεψη μελλοντικών εξόδων. Με βάση τα ζεύγη αυτά εισόδου – εξόδου παράγεται το μοντέλο το οποίο θα αποτελέσει τον οδηγό για τις μελλοντικές εκτιμήσεις. Τα μοντέλα που παράγονται με τις μεθόδους της κατηγορίας αυτής χαρακτηρίζονται από σχετικά υψηλή ακρίβεια. Το βασικό μειονέκτημα των μεθόδων αυτών είναι ότι απαιτούν τον προσδιορισμό προτύπων για κάθε κατηγορία δεδομένων. Στην παρακάτω εικόνα φαίνεται σχηματικά το πως λειτουργούν αυτού του είδους οι αλγόριθμοι.



Εικόνα 7 Σχηματική Αναπαράσταση Λειτουργίας των Αλγορίθμων Εποπτευόμενης Μηχανικής Μάθησης (πηγή: https://www.dqlab.id/4-jenis-algoritma-machine-learning-yang-paling-populer-tahun-2022)

Οι αλγόριθμοι επιτηρούμενης μάθησης διακρίνονται σε δεύτερο επίπεδο σε (Singh, 2019):

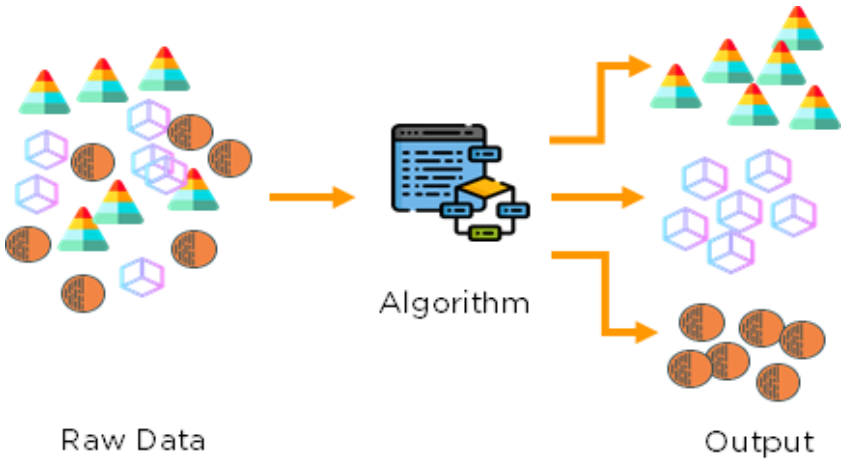
- Μεθοδολογίες ταξινόμησης (Classification): Οι μεταβλητές εξόδου αποτελούν διακριτές (κατηγορικές) τιμές που αντιστοιχούν σε κατηγορίες ταξινόμησης. Οι οντότητες που περιγράφονται από τις

παραμέτρους εισόδου, τοποθετούνται ανάλογα με τις τιμές των παραμέτρων τους στις αντίστοιχες κατηγορίες.

- Μεθοδολογίες παλινδρόμησης (Regression): Οι μεταβλητές εξόδου αντιστοιχούν σε συνεχείς τιμές. Στόχος των διαδικασιών αυτών είναι να προβλεφθεί η τιμή μιας παραμέτρου των οντοτήτων με βάση τις τιμές των υπόλοιπων παραμέτρων εισόδου.

- Αλγόριθμοι Μηχανικής Μάθησης χωρίς επιτήρηση (Gentleman & Carey, 2008): (Unsupervised Machine Learning): Η μη επιτηρούμενη μηχανική μάθηση στοχεύει στην ανίχνευση προτύπων στα ίδια τα δεδομένα εισόδου προκειμένου να μοντελοποιήσει τη δομή και την κατανομή τους. Η διαφορά τους σε σχέση με τις εποπτευόμενες τεχνικές έγκειται στο γεγονός ότι κατά την παραγωγή των μοντέλων πρόβλεψης, δεν παρέχεται η τιμή της παραμέτρου στόχος, αλλά αναζητούνται πρότυπα που να περιγράφουν με την μεγαλύτερη δυνατή ακρίβεια τις συσχετίσεις μεταξύ των οντοτήτων που απαρτίζουν τα δεδομένα εκπαίδευσης. Οι αλγόριθμοι αυτής της κατηγορίας διακρίνονται σε:

  - Αλγόριθμοι συσταδοποίησης (Clustering): Οι μεταβλητές εξόδου είναι διακριτές τιμές που αντιστοιχούν σε κατηγορίες που δεν είναι γνωστές εκ των προτέρων. Οι συστάδες των οντοτήτων διαμορφώνονται με βάση τη συνάφεια αυτών που βρίσκονται στην ίδια συστάδα και την διαφοροποίηση αυτών που βρίσκονται σε διαφορετικές. Στην παρακάτω εικόνα φαίνεται σχηματικά η λειτουργία των αλγορίθμων αυτών.



Εικόνα 8 Σχηματική Αναπαράσταση της Λειτουργίας των Αλγορίθμων Συσταδοποίησης (πηγή: https://medium.com/analytics-vidhya/beginners-guide-to-unsupervised-learning-76a575c4e942)

- Αλγόριθμοι συσχέτισης: Οι αλγόριθμοι αυτού του είδους εξελίσσονται αναζητώντας συσχετίσεις μεταξύ των χαρακτηριστικών των οντοτήτων που περιλαμβάνονται στο σύνολο εκπαίδευσης.

- Αλγόριθμοι Ενισχυτικής Μάθησης: Χαρακτηριστικό των αλγορίθμων αυτού του είδους είναι το γεγονός ότι χρησιμοποιείται πράκτορας. Ο πράκτορας είναι μία οντότητα (φυσική ή σε μορφή λογισμικού) η οποία μπορεί να ανιχνεύει την κατάσταση του περιβάλλοντος του, να αλληλοεπιδρά με τα στοιχεία του και ανάλογα να διαμορφώνει τη συμπεριφορά του. Σε κάθε βήμα της αλληλεπίδρασης ο πράκτορας λαμβάνει ως είσοδο, κάποια ένδειξη της τρέχουσας κατάστασης, του περιβάλλοντος. Τότε επιλέγει μια ενέργεια, που αποτελεί και την έξοδο του αλγόριθμου που ρυθμίζει τη συμπεριφορά του. Η ενέργεια του πράκτορα αλλάζει την κατάσταση του περιβάλλοντος και η τιμή αυτής της μετάβασης κατάστασης, του κοινοποιείται μέσω ενός σήματος βεβαρυμμένης ενίσχυσης. Η συμπεριφορά του πράκτορα, θα πρέπει να επιλέξει ενέργειες που τείνουν να αυξάνουν το μακροπρόθεσμο άθροισμα τιμών του σήματος ενίσχυσης. Μπορεί να μάθει να το κάνει αυτό με την πάροδο του χρόνου με συστηματικές δοκιμές και λάθη, καθοδηγούμενο από μια μεγάλη ποικιλία αλγορίθμων. Το πιο εύστοχο παράδειγμα των

αλγορίθμων αυτού του είδους είναι η κινήσεις ρομποτικών μηχανισμών στο χώρο (Moore και Littman 1996).

Ένα άλλο κριτήριο για τη διάκριση των κατηγοριών αλγορίθμων Μηχανικής Μάθησης είναι το κατά πόσο υπεισέρχεται η ανθρώπινη επέμβαση κατά την εξέλιξη τους (Kumar, και συν. 2018):

- Παθητικοί Αλγόριθμοι: Χαρακτηριστικό των αλγορίθμων αυτού του είδους είναι το ότι τα δεδομένα που χρησιμοποιούνται για την παραγωγή του μοντέλου προέρχονται από απλή παρατήρηση του περιβάλλοντος. Η πιο συνηθισμένη προέλευση δεδομένων για παθητικούς αλγορίθμους μηχανικής μάθησης είναι οι αισθητήρες των συσκευών του διαδικτύου των πραγμάτων. Παραδείγματα τέτοιων μηχανισμών είναι η παρατήρηση των καιρικών συνθηκών προκειμένου να εντοπιστεί η επιρροή τους σε διάφορες περιβαλλοντολογικές καταστάσεις (πχ η πυκνότητα των ρύπων στην ατμόσφαιρα).
- Ενεργητικοί Αλγόριθμοι: Στις διαδικασίες που στηρίζονται σε ενεργητικούς αλγορίθμους, τα δεδομένα διαμορφώνονται με ανθρώπινη παρέμβαση. Με τον τρόπο αυτό, ο ανθρώπινος παράγοντας επεμβαίνει στην προσαρμογή των παραγομένων μοντέλων. Η περιπτώσεις της εποπτευόμενης μηχανικής μάθησης, συνήθως κατατάσσονται στην κατηγορία αυτή, καθώς τα σύνολα δεδομένων με γνωστές ετικέτες, διαμορφώνονται από τον άνθρωπο. Για παράδειγμα δημιουργούνται σύνολα φωτογραφιών με γνωστούς χαρακτηρισμούς - ετικέτες προκειμένου να διαμορφωθούν μοντέλα εκτίμησης άλλων για τις οποίες δεν είναι γνωστή η κατηγορία στην οποία ανήκουν (Kumar, και συν. 2018).

2.4. Η Γενική Μεθοδολογία

Η διαδικασία της μηχανικής μάθησης αποτελεί μέρος της εξόρυξης γνώσης (data mining). Η διαδικασία της εξόρυξης γνώσης περιλαμβάνει μία σειρά από στάδια που εκτείνονται από τον προσδιορισμό του στόχο της μέχρι τα αποτελέσματα της. Οι

διαδικασίες είναι δυναμικές, υπό την έννοια ότι οι μεταβολές του περιβάλλοντος συμπαρασύρουν αναπροσαρμογές τους, Καθώς εξελίσσονται, δέχονται εισόδους από το περιβάλλον, οι οποίες επηρεάζουν τη ροή τους. Αυτό έχει ως αποτέλεσμα, οι έξοδοι τους να μην είναι προδιαγεγραμμένες, αλλά κάθε φορά να επικαιροποιούνται από τις καταστάσεις που βρίσκεται το περιβάλλον τους.

Τα στάδιά που περιλαμβάνει μία ολοκληρωμένη διαδικασία εξόρυξης είναι:

- Προσδιορισμός της σκοπιμότητας
- Συλλογή των δεδομένων
- Προετοιμασία των δεδομένων
- Προεπεξεργασία των δεδομένων
- Επιλογή Αλγορίθμου Εκπαίδευσης
- Εκπαίδευση και Αξιολόγηση του Μοντέλου
- Χρήση του Μοντέλου - Ανάλυση Αποτελεσμάτων

Η Μηχανική Μάθηση εξελίσσεται στα στάδια της επιλογής του αλγορίθμου και των εκπαίδευση – αξιολόγηση του μοντέλου. Στις επόμενες παραγράφους περιγράφονται συνοπτικά τα στάδια αυτά (Ge, Ding και Huang 2017).

2.4.1. Προσδιορισμός της σκοπιμότητας

Πριν ξεκινήσει οποιαδήποτε εργασία σχετίζεται με την εξόρυξη γνώσης, θα πρέπει να έχει προσδιοριστεί ο σκοπός που εξυπηρετεί. Είναι απαραίτητο να τεθούν οι στόχοι που θα πρέπει να επιτευχθούν, οι περιορισμοί που τίθενται και το κατά πόσο τα αποτελέσματα της διαδικασίας θα συμβάλλουν στην ανάπτυξη του οργανισμού. Επιπλέον, θα πρέπει να καθοριστεί η θεματολογία και η μορφή των απαιτούμενων δεδομένων εισόδου καθώς και η μορφή της εξόδου της όλης διαδικασίας (Alfred, 2005).

2.4.2. Συλλογή των Δεδομένων

Τον προσδιορισμό της σκοπιμότητας της διαδικασίας Μηχανικής Μάθησης ακολουθεί η διαδικασία συλλογής των δεδομένων. Μόνο τότε οι αναλυτές μπορούν να αναζητήσουν πηγές με κατάλληλα δεδομένα. Μία λύση που χρησιμοποιείται σε

μεγάλο βαθμό τον τελευταίο καιρό είναι η αξιοποίηση των παρόχων υπηρεσιών οπτικοποίησης δεδομένων. Οι οργανισμοί αυτοί συγκεντρώνουν δεδομένα από διαφορετικές πηγές και τα διανέμουν στους πελάτες τους ομογενοποιημένα και έτοιμα να τα χρησιμοποιήσουν στις δικές τους εφαρμογές (TIBCO 2022).

2.4.3. Προετοιμασία Δεδομένων

Μετά την συγκέντρωση των δεδομένων, αυτά θα πρέπει να εξεταστούν ως προς την καταλληλότητά τους. Στο στάδιο της προετοιμασίας των δεδομένων λαμβάνουν χώρα οι ακόλουθες διεργασίες:

- Εξέταση της δομής του συνόλου δεδομένων: Εξετάζεται αν η δομή των δεδομένων είναι κατάλληλη για την χρήση τους στην διαδικασία.
- Επιλογή των κατάλληλων δεδομένων: Εξετάζονται τα δεδομένα ως προς την σημασιολογία τους, για να διαπιστωθεί αν είναι κατάλληλα να χρησιμοποιηθούν στη διαδικασία εξόρυξης γνώσης.

Τα κριτήρια για την προετοιμασία των δεδομένων είναι ο σκοπός που θα εξυπηρετήσει η διαδικασία, το είδος των αποτελεσμάτων που αναμένονται και το περιβάλλον του εξεταζόμενου αντικειμένου. Στο τέλος αυτού του σταδίου, έχουν επιλεγεί τα δεδομένα που θα χρησιμοποιηθούν εν τέλει στη διαδικασία (Alfred, 2005).

2.4.4.Προεπεξεργασία Δεδομένων

Πριν την έναρξη της προεπεξεργασίας των δεδομένων, έχει καθοριστεί το σύνολο των δεδομένων που θα χρησιμοποιηθεί. Στη συνέχεια θα πρέπει να βελτιωθεί η ποιότητα του συνόλου δεδομένων με τρόπο τέτοιο που θα διευκολύνει την αναβάθμιση της συνολικής διαδικασίας. Αυτό επιτυγχάνεται στο στάδιο της προ επεξεργασίας των δεδομένων μέσα από κατάλληλους μετασχηματισμούς τους. Πριν τη προ επεξεργασία τους τα δεδομένα μπορεί να παρουσιάζουν τις παρακάτω αδυναμίες:

- Ημιτελή Δεδομένα: Η αδυναμία αυτή μπορεί να αφορά έλλειψη στις τιμές κάποιων γνωρισμάτων, έλλειψη ολόκληρων γνωρισμάτων ή τη διαθεσιμότητα μόνο συναθροιστικών δεδομένων.

- Ύπαρξη θορύβου: Στοιχεία των δεδομένων τα οποία δε συνάδουν με τη σημασιολογία της πλειοψηφίας τους.
- Ασυνέπειες: Αναφέρεται στις αποκλίσεις στην κωδικοποίηση που χρησιμοποιείται για την ταξινόμηση των δεδομένων ή στα ονόματα αναφοράς στα δεδομένα. Στην περίπτωση που η έξοδος των αλγορίθμων είναι αριθμητικές τιμές, εξετάζονται οι αποκλίσεις των πραγματικών τιμών από τις αναμενόμενες.

Η έλλειψη ποιότητας των δεδομένων οδηγεί σε έλλειψη ακρίβειας των αποτελεσμάτων.

Η προεπεξεργασία των δεδομένων περιλαμβάνει τις παρακάτω επιμέρους εργασίες:

- Καθαρισμός: Τα δεδομένα μπορεί να έχουν πολλά άσχετα ή/και να λείπουν κάποια στοιχεία μέρη. Για τον χειρισμό αυτού των προβλημάτων, γίνεται καθαρισμός δεδομένων. Αυτό περιλαμβάνει την κατάλληλη συμπλήρωση των στοιχείων που λείπουν, την απομάκρυνση του θορύβου και των δεδομένων που προκαλούν σημασιολογική ασυνέπεια.
- Μετασχηματισμός: Αυτό το βήμα γίνεται προκειμένου να μετατραπούν τα δεδομένα σε κατάλληλες μορφές κατάλληλες για τη διαδικασία εξόρυξης. Αυτό περιλαμβάνει τους ακόλουθους τρόπους:
    - Κανονικοποίηση: Γίνεται για να κλιμακωθούν οι τιμές δεδομένων σε ένα καθορισμένο εύρος (π.χ. -1,0 έως 1,0 ή 0,0 έως 1,0). Η κανονικοποίηση απαιτείται γενικά όταν έχουμε να κάνουμε με χαρακτηριστικά σε διαφορετική κλίμακα, διαφορετικά, μπορεί να οδηγήσει σε μείωση της αποτελεσματικότητας ενός εξίσου σημαντικού χαρακτηριστικού (σε χαμηλότερη κλίμακα) λόγω άλλων χαρακτηριστικών που έχουν τιμές σε μεγαλύτερη κλίμακα. Με απλά λόγια, όταν υπάρχουν πολλά χαρακτηριστικά αλλά τα

χαρακτηριστικά έχουν τιμές σε διαφορετικές κλίμακες, αυτό μπορεί να οδηγήσει σε κακά μοντέλα δεδομένων κατά την εκτέλεση εργασιών εξόρυξης δεδομένων. Έτσι κανονικοποιούνται ώστε να φέρουν όλα τα χαρακτηριστικά στην ίδια κλίμακα.

- Συνάθροιση (Aggregation, data cube construction): Η συλλογή ή η συνάθροιση δεδομένων είναι η μέθοδος αποθήκευσης και παρουσίασης δεδομένων σε συνοπτική μορφή. Τα δεδομένα μπορούν να ληφθούν από πολλαπλές πηγές δεδομένων για να ενσωματωθούν αυτές οι πηγές δεδομένων σε μια περιγραφή ανάλυσης δεδομένων. Αυτό είναι ένα κρίσιμο βήμα, καθώς η ακρίβεια των γνώσεων της ανάλυσης δεδομένων εξαρτάται σε μεγάλο βαθμό από την ποσότητα και την ποιότητα των δεδομένων που χρησιμοποιούνται. Η συλλογή ακριβών δεδομένων υψηλής ποιότητας και αρκετά μεγάλης ποσότητας είναι απαραίτητη για την παραγωγή σχετικών αποτελεσμάτων.
- Εξομάλυνση: Είναι μια διαδικασία που χρησιμοποιείται για την αφαίρεση του θορύβου από το σύνολο δεδομένων. Επιτρέπει την επισήμανση σημαντικών χαρακτηριστικών που υπάρχουν στο σύνολο δεδομένων. Βοηθά στην πρόβλεψη των προτύπων. Κατά τη συλλογή δεδομένων, μπορεί να γίνει χειρισμός για την εξάλειψη ή τη μείωση οποιασδήποτε διακύμανσης ή οποιασδήποτε άλλης μορφής θορύβου. Η ιδέα πίσω από την εξομάλυνση δεδομένων είναι ότι θα μπορεί να εντοπίζει απλές αλλαγές για να βοηθήσει στην πρόβλεψη διαφορετικών τάσεων και προτύπων.
- Κατασκευή χαρακτηριστικών: Δημιουργούνται και εφαρμόζονται νέα χαρακτηριστικά για να βοηθήσουν τη διαδικασία εξόρυξης από το δεδομένο σύνολο

χαρακτηριστικών. Αυτό απλοποιεί τα αρχικά δεδομένα και κάνει την εξόρυξη πιο αποτελεσματική.

- Μείωση Δεδομένων / Μείωση Διαστάσεων: Δεδομένου ότι η εξόρυξη δεδομένων είναι μια τεχνική που χρησιμοποιείται για το χειρισμό τεράστιου όγκου δεδομένων. Ενώ εργαζόταν με τεράστιο όγκο δεδομένων, η ανάλυση έγινε πιο δύσκολη σε τέτοιες περιπτώσεις. Για να απαλλαγούμε από αυτό, χρησιμοποιούμε τεχνική μείωσης δεδομένων. Στοχεύει στην αύξηση της αποδοτικότητας αποθήκευσης και στη μείωση του κόστους αποθήκευσης και ανάλυσης δεδομένων. Τα διάφορα βήματα για τη μείωση των δεδομένων είναι (Olson & Delen, 2008):
    - Συνάθροιση κύβου δεδομένων (Data Cube Aggregation): Η λειτουργία συγκέντρωσης εφαρμόζεται σε δεδομένα για την κατασκευή του κύβου δεδομένων. Αυτή η τεχνική χρησιμοποιείται για τη συγκέντρωση δεδομένων σε απλούστερη μορφή. Για παράδειγμα, έστω ότι οι πληροφορίες που συλλέχθηκαν για τα έτη 2012 έως 2014, περιλαμβάνουν τα έσοδα της εταιρείας κάθε τρεις μήνες. Αν το ζητούμενο είναι ο ετήσιος μέσος όρος, μπορούμε να συνοψίσουμε τα δεδομένα με τέτοιο τρόπο ώστε τα δεδομένα που προκύπτουν να συνοψίζουν τις συνολικές πωλήσεις ανά έτος αντί για ανά τρίμηνο.
    - Επιλογή υποσυνόλου χαρακτηριστικών: Θα πρέπει να χρησιμοποιηθούν τα εξαιρετικά σχετικά χαρακτηριστικά, ενώ το υπόλοιπο μπορεί να απορριφθεί. Για την εκτέλεση της επιλογής χαρακτηριστικών, μπορεί κανείς να χρησιμοποιήσει το επίπεδο σημασίας και την τιμή p του χαρακτηριστικού.

- Μείωση Numerosity: Αυτό επιτρέπει την αποθήκευση του μοντέλου δεδομένων αντί για ολόκληρα δεδομένα, Σε αυτήν την τεχνική μείωσης τα πραγματικά δεδομένα αντικαθίστανται με μαθηματικά μοντέλα ή μικρότερη αναπαράσταση των δεδομένων αντί για πραγματικά δεδομένα, είναι σημαντικό να αποθηκεύεται μόνο η παράμετρος μοντέλου ή μη παραμετρική μέθοδος όπως ομαδοποίηση, ιστόγραμμα, δειγματοληψία.
- Μείωση διαστάσεων: Αυτό μειώνει το μέγεθος των δεδομένων μέσω μηχανισμών κωδικοποίησης. Μπορεί να έχει απώλειες ή χωρίς απώλειες. Εάν μετά την ανακατασκευή από συμπιεσμένα δεδομένα, μπορούν να ανακτηθούν τα αρχικά δεδομένα, αυτή η μείωση ονομάζεται μείωση χωρίς απώλειες αλλιώς ονομάζεται μείωση απωλειών.
- Συμπίεση Δεδομένων: Η τεχνική συμπίεσης δεδομένων μειώνει το μέγεθος των αρχείων χρησιμοποιώντας διαφορετικούς μηχανισμούς κωδικοποίησης.
- Λειτουργία διακριτοποίησης & εννοιολογικής ιεραρχίας: Τεχνικές διακριτοποίησης δεδομένων χρησιμοποιούνται για τη διαίρεση των χαρακτηριστικών της συνεχούς φύσης σε δεδομένα με διαστήματα. Αντικαθιστούμε πολλές σταθερές τιμές των χαρακτηριστικών με ετικέτες μικρών διαστημάτων.
- Ιεραρχίες εννοιών (Generalization, concept hierarchy climbing): Μειώνει το μέγεθος των δεδομένων συλλέγοντας και στη συνέχεια αντικαθιστώντας τις έννοιες χαμηλού επιπέδου (όπως 43 για την ηλικία) σε έννοιες υψηλού επιπέδου (κατηγορικές μεταβλητές όπως η μέση ηλικία ή η ανώτερη ηλικία). Για αριθμητικά

δεδομένα μπορούν να ακολουθηθούν οι ακόλουθες τεχνικές:

- Binning: Binning είναι η διαδικασία αλλαγής αριθμητικών μεταβλητών σε κατηγορικές αντίστοιχες. Ο αριθμός των κατηγορικών αντίστοιχων εξαρτάται από τον αριθμό των δοχείων που καθορίζονται από τον χρήστη.
- Ανάλυση ιστογράμματος: Όπως και η διαδικασία δέσμευσης, το ιστόγραμμα χρησιμοποιείται για το διαμερισμό της τιμής για το χαρακτηριστικό Χ, σε διαχωρισμένες περιοχές που ονομάζονται αγκύλες. Υπάρχουν διάφοροι κανόνες κατάτμησης:
- Διαμέριση ίσης συχνότητας: Διαμερισμός των τιμών με βάση τον αριθμό των εμφανίσεων τους στο σύνολο δεδομένων.
- Κατάτμηση ίσου πλάτους: Διαμερισμός των τιμών σε ένα σταθερό κενό με βάση τον αριθμό των bins, δηλαδή ένα σύνολο τιμών που κυμαίνεται από 0-20.
- Ομαδοποίηση: Ομαδοποίηση παρόμοιων δεδομένων μαζί.

Εξισορρόπηση: Εξασφαλίζει την ισορροπία του συνόλου των δεδομένων ως προς τα βασικά χαρακτηριστικά της περιγραφής των περιεχομένων τους. Αυτό μπορεί να γίνει είτε με την προσθήκη εικονικών δεδομένων, είτε με την αφαίρεση πλεοναζόντων δεδομένων είτε με συνδυασμό αυτών (Olson & Delen, 2008).

2.4.5. Επιλογή Αλγορίθμου

Στο στάδιο αυτό επιλέγεται ο αλγόριθμος που θα χρησιμοποιηθεί για την ολοκλήρωση της διαδικασίας εξόρυξης γνώσης. Ο αλγόριθμος επιλέγεται κυρίως με βάση τα παρακάτω κριτήρια:

- Το είδος των αποτελεσμάτων που είναι επιθυμητό να επιστρέφει η διαδικασία
- Το είδος και τη μορφή των δεδομένων που θα χρησιμοποιηθούν.
- Την επιθυμητή ακρίβεια των αποτελεσμάτων
- Τον διαθέσιμο χρόνο για την παραγωγή των αποτελεσμάτων

Πολλές φορές επιλέγονται περισσότεροι του ενός αλγόριθμοι ώστε να αξιολογηθούν ως προς την απόδοση τους και για τις παραγωγικές διαδικασίες επιλέγεται αυτός με την καλύτερη απόδοση.

2.4.6. Εκπαίδευση και Αξιολόγηση του Μοντέλου

Στο στάδιο αυτό εκτελείται το κύριο μέρος της διαδικασίας της μηχανικής μάθησης. Χρησιμοποιείται ο επιλεγμένος αλγόριθμος και εκτελείται με είσοδο τα δεδομένα εκπαίδευσης. Τα μοντέλα που παράγονται αξιολογούνται και επιλέγονται τα μοντέλα που παρουσιάζουν την καλύτερη ποιότητα. Για τον καθορισμό της ποιότητας των μοντέλων που παράγονται χρησιμοποιούνται μετρικές με βάση το είδος της μηχανικής μάθησης που επιλέγεται. Για τις εποπτευόμενες μεθόδους, οι μετρικές αυτές βασίζονται στην απόκλιση που έχουν οι τιμές που επιστρέφουν τα παραγόμενα μοντέλα σε σχέση με τις πραγματικές τιμές που έχουν καταγραφεί. Οι πιο συχνά χρησιμοποιούμενες μετρικές για τα μοντέλα αυτού του είδους είναι (Minaee, 2019):

- Ακρίβεια (Precision): Είναι ο λόγος του αριθμού των σωστών προβλέψεων προς τον συνολικό αριθμό των δειγμάτων εισόδου.
- Λογαριθμική Απώλεια (Logarithmic Loss): Λειτουργεί τιμωρώντας τις ψευδείς ταξινομήσεις. Λειτουργεί καλά για ταξινόμηση σε πολλαπλές κλάσεις. Χρησιμοποιείται στις περιπτώσεις που ο ταξινομητής

επιστρέφει τη πιθανότητα το δείγμα να ανήκει σε κάθε μία από τις εξεταζόμενες κατηγορίες. Η τιμή της υπολογίζεται ως εξής:

$$Logatithmic\ Loss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} log\ (p_{ij})$$

όπου,yij, υποδεικνύει εάν το δείγμα i ανήκει στην κατηγορία j ή όχι, pij, δείχνει την πιθανότητα το δείγμα i να ανήκει στην κλάση j. Η τιμή του κινείται στο διάστημα [0, ∞). Τιμές κοντά στο 0 υποδηλώνουν μεγαλύτερη ακρίβεια.

- Πίνακας σύγχυσης (Confusion Matrix): Περιγράφει την πλήρη απόδοση του μοντέλου χρησιμοποιώντας τέσσερις έννοιες για την ταξινόμηση των δειγμάτων:
    - Αληθινά θετικά (True Positives): Οι περιπτώσεις στις οποίες προβλέφθηκε για ένα δείγμα η κατηγορία Α και η πραγματική κατηγορία ήταν η Α.
    - Αληθινά αρνητικά (True Negatives): Οι περιπτώσεις στις οποίες δεν προβλέφθηκε για ένα δείγμα η κατηγορία Α και η πραγματική κατηγορία δεν ήταν η Α.
    - Ψευδώς Θετικά (False Positives): Οι περιπτώσεις στις οποίες προβλέφθηκε για ένα δείγμα η κατηγορία Α και η πραγματική κατηγορία δεν ήταν η Α.
    - Ψευδώς Αρνητικά (False Negatives): Οι περιπτώσεις στις οποίες δεν προβλέφθηκε για ένα δείγμα η κατηγορία Α και η πραγματική κατηγορία ήταν η Α.

  Ο πίνακας έχει διαστάσεις NXN, όπου N το πλήθος των κλάσεων στις οποίες κατατάσσονται τα δείγματα. Οι γραμμές αντιστοιχούν στις πραγματικές κατηγορίες κατάταξης και οι στήλες στις προβλέψεις. Στην κύρια διαγώνιο καταγράφονται οι True Positive προβλέψεις ενώ στα επόμενα κελιά φαίνονται τα πλήθη των διαφορετικών

προβλέψεων σε σχέση με την πραγματική τιμή, για κάθε ζεύγος κλάσεων.

- Ευαισθησία (Sensitivity): Ορίζεται ο λόγος των True Positive προβλέψεων προς το άθροισμα των False Negative και True Positive. Εκφράζει το κατά πόσο κάποιο δείγμα που κατατάχθηκε σε κάποια κατηγορία, ορθώς τέθηκε εκεί.
- Ειδικότητα (Specificity): Ορίζεται ο λόγος των True Negative προβλέψεων προς το άθροισμα των False Positive and True Negative. Εκφράζει το κατά πόσο κάποιο δείγμα που δεν κατατάχθηκε σε κάποια κατηγορία, ορθώς δεν τέθηκε εκεί.
- False Positive Rate : Ορίζεται ο λόγος των False Positive προς το άθροισμα των False Positive and True Negative προβλέψεων. Αντιστοιχεί στο ποσοστό των αρνητικών πραγματικών κατατάξεων που εσφαλμένα τοποθετούνται σε δεδομένη κλάση.
- Ανάκληση (Recall): Είναι ο αριθμός των σωστών θετικών αποτελεσμάτων διαιρεμένος με τον αριθμό όλων των δειγμάτων που θα έπρεπε να έχουν αναγνωριστεί ως θετικά).
- Score F1: Η μετρική αυτή είναι ο αρμονικός μέσος όρος μεταξύ ακρίβειας και ανάκλησης. Το εύρος για το σκορ F1 είναι [0, 1] και δηλώνει πόσο ακριβής είναι ο ταξινομητής και πόσο ισχυρός είναι στο να μην κάνει εσφαλμένες προβλέψεις. Η τιμή του δίνεται από τη σχέση:

$$F1 = 2\frac{1}{\frac{1}{precision} + \frac{1}{recall}}$$

- Μέσο απόλυτο σφάλμα: Το μέσο απόλυτο σφάλμα είναι ο μέσος όρος της διαφοράς μεταξύ των αρχικών τιμών και των προβλεπόμενων τιμών. Αποδίδει το μέτρο του πόσο μακριά ήταν οι προβλέψεις από την πραγματική παραγωγή.

- Μέσο τετράγωνο σφάλματος: Το μέσο τετράγωνο σφάλμα αποδίδει τον μέσο όρο του τετραγώνου της διαφοράς μεταξύ των αρχικών τιμών και των προβλεπόμενων τιμών (Mishra, 2018).

Οι μετρικές απόδοσης των μη εποπτευόμενων τεχνικών σχετίζονται με το κατά πόσο η κατάταξη των αντικειμένων σε συστάδες είναι καταλληλότερη από κάποια άλλη. Η πιο συχνά χρησιμοποιούμενη μετρική είναι ο συντελεστής silhouette. Είναι μια μέτρηση που χρησιμοποιείται για τον υπολογισμό της καλής τεχνικής ομαδοποίησης. Η τιμή του κυμαίνεται από -1 έως 1 και υπολογίζεται ως εξής:

$$Silhouette = \frac{ba}{max\ (b, a)}$$

Όπου α= μέση απόσταση εντός του συμπλέγματος, δηλαδή η μέση απόσταση μεταξύ των ζευγών σημείων ενός συμπλέγματος και b= μέση απόσταση μεταξύ συστάδων, δηλαδή η μέση απόσταση μεταξύ όλων των συστάδων. Η ερμηνεία των ακραίων τιμών έχουν ως εξής:

- 1: Σημαίνει ότι οι συστάδες των μέσων είναι πολύ μακριά μεταξύ τους και ξεκάθαρα διακρίνονται.
- 0: Σημαίνει ότι τα συμπλέγματα είναι αδιάφορα ή μπορούμε να πούμε ότι η απόσταση μεταξύ των συστάδων δεν είναι σημαντική.
- -1: Σημαίνει ότι τα συμπλέγματα μέσων έχουν εκχωρηθεί με λάθος τρόπο (Bhardwaj, 2020).

2.4.7. Χρήση του Μοντέλου - Ανάλυση Αποτελεσμάτων

Με την ολοκλήρωση της διαδικασίας της μηχανικής μάθησης, το μοντέλο που επιλέχθηκε μετά τις αξιολογήσεις, χρησιμοποιείται για εξόρυξη δεδομένων. Οι χρήσεις των μοντέλων αυτών περιλαμβάνει ομαδοποίηση δεδομένων, μείωση διαστάσεων διανυσμάτων αναπαράστασης οντοτήτων και καταστάσεων, οπτικοποίηση δεδομένων, ανάλυση τάσεων, παρακολούθηση διαδικασίας και διάγνωση σφαλμάτων, ταξινόμηση σφαλμάτων, και πρόβλεψη ποιότητας. Τα αποτελέσματα της εφαρμογής δεδομένων παραγωγής στα μοντέλα, παράγουν

προβλέψεις που είναι κατάλληλες για την υποστήριξη διαδικασιών λήψης αποφάσεων.

Κεφάλαιο 3: Επεξεργασία Φυσικής Γλώσσας

3.1. Περιγραφή

Στον πυρήνα κάθε εργασίας επεξεργασίας φυσικής γλώσσας υπάρχει το σημαντικό ζήτημα της κατανόησης της φυσικής γλώσσας. Η διαδικασία κατασκευής εφαρμογών για χρήση σε ηλεκτρονικό υπολογιστή, που κατανοούν τη φυσική γλώσσα, περιλαμβάνει την επίλυση τριών σημαντικών προβλημάτων:

- Διαδικασία σκέψης
- Αναπαράσταση και νόημα της γλωσσικής εισόδου
- Καθολική γνώση.

Έτσι, ένα σύστημα NLP μπορεί να ξεκινήσει σε επίπεδο λέξης – για να προσδιορίσει τη μορφολογική δομή και τη φύση (όπως μέρος του λόγου, το νόημα) της λέξης. Στη συνέχεια μπορεί να προχωρήσει στο επίπεδο της πρότασης – για να καθορίσει τη σειρά των λέξεων, γραμματική, το νόημα ολόκληρης της πρότασης. Τέλος χρειάζεται να εντάξει την πρόταση στο πλαίσιο και στο συνολικό περιβάλλον. Μια δεδομένη λέξη ή μια πρόταση μπορεί να έχει μια συγκεκριμένη σημασία ή χροιά σε ένα δεδομένο πλαίσιο ή τομέα και μπορεί να σχετίζεται με πολλές άλλες λέξεις ή/και προτάσεις στο συγκεκριμένο πλαίσιο.

3.2. Ιστορικό

Η μελέτη για την επεξεργασία φυσικής γλώσσας (Natural Language Processing – NLP) ξεκίνησε τη δεκαετία του 1950. Θεωρήθηκε ένας χώρος όπου συναντιόνταν η τεχνητή νοημοσύνη και η γλωσσολογία. Η NLP και η ανάκτηση πληροφορίας (Information Retrieval - IR) από κείμενα συνέκλιναν. Η IR χρησιμοποιεί τεχνικές βασισμένες σε στατιστικά στοιχεία για την αποτελεσματική ευρετηρίαση και αναζήτηση μεγάλων όγκων κειμένου. Η εξέλιξη τόσο του NLP όσο και του IR, τα έφερε πιο κοντά με τον καιρό. Ωστόσο, η NLP είναι περισσότερο διευρυμένη καθώς συγγενεύει με αρκετά διαφορετικά πεδία (Lutkevich, 2020).

Οι πρώτες προσεγγίσεις επιχειρούσαν τη λέξη προς λέξη αυτόματη μετάφραση και νοηματική αναγνώριση. Ωστόσο οι πολλαπλές σημασιολογικές προσεγγίσεις των ίδιων κειμένων αποτέλεσε ένα σημαντικό εμπόδιο. Το 1963 παρουσιάστηκε η σημειογραφία Backus-Naur Form (BNF). Το BNF χρησιμοποιήθηκε για να καθορίσει μια γραμματική χωρίς πλαίσιο (Context Free Grammar - CFG) και βρήκε εφαρμογή στις γλώσσες προγραμματισμού. Η προδιαγραφή BNF μιας γλώσσας είναι ένα σύνολο κανόνων που επικυρώνουν ολιστικά τον κώδικα προγράμματος συντακτικά. Αργότερα προτάθηκαν οι περιοριστικές γραμματικές, που αποτέλεσαν τη βάση των regular expressions που χρησιμοποιούνται για τον εντοπισμό μοτίβων σε ένα κείμενο (Coope & Torczon, 2012).

Τη δεκαετία του 1970, οι γεννήτριες lexical-analyzer (lexer) και οι γεννήτριες parser χρησιμοποίησαν επίσης γραμματικές. Ένας lexer μετατρέπει ένα κείμενο σε στοιχεια και ο parser επικυρώνει τις ακολουθίες της. Με την χρήση τους διευκολύνονται οι υλοποιήσεις των γλωσσών προγραμματισμού. Οι regular-expressions διευκόλυναν τις αναζητήσεις προτύπων εκφράσεων σε κείμενα. Ενώ οι CFG είναι θεωρητικά ανεπαρκείς για τη φυσική γλώσσα, συχνά χρησιμοποιούνται για το NLP στην πράξη. Η γλώσσα Prolog11 επινοήθηκε αρχικά για εφαρμογές NLP. Η σύνταξή του είναι ιδιαίτερα κατάλληλη για τη σύνταξη γραμματικών (Lutkevich, 2020).

Το εξαιρετικά μεγάλο μέγεθος, η απεριόριστη φύση και η ασάφεια της φυσικής γλώσσας οδήγησαν σε δύο προβλήματα κατά τη χρήση τυπικών προσεγγίσεων ανάλυσης που βασίζονταν καθαρά σε συμβολικούς, χειροποίητους κανόνες:

Το NLP πρέπει τελικά να εξαγάγει το νόημα από το κείμενο: οι τυπικές γραμματικές που καθορίζουν τη σχέση μεταξύ των ενοτήτων κειμένου - μέρη του λόγου όπως ουσιαστικά, ρήματα και επίθετα – σχετίζονται κυρίως με το συντακτικό. Οι γραμματικές μπορούν να επεκταθούν για να αντιμετωπίσουν τη σημασιολογία της φυσικής γλώσσας. Αυτό επιτυγχάνεται με την κατηγοριοποίηση των εννοιών σε διαφορετικά επίπεδα και πρόσθετους κανόνες/περιορισμούς. Ωστόσο, οι κανόνες καθίστανται αδιαχείριστα πολυάριθμοι, συχνά αλληλεπιδρώντας απρόβλεπτα, με

πιο συχνές διφορούμενες αναλύσεις. Εκτός αυτού, στην καθομιλουμένη γλώσσα χρησιμοποιούνται άγραφοι κανόνες του λόγου που είναι δύσκολο να ενταχθούν σε αυτοματοποιημένη επεξεργασία (Lutkevich, 2020).

Αργότερα, στην NLP χρησιμοποιήθηκαν τεχνικές μηχανικής μάθησης. Αυτές χρησιμοποίησαν σε μεγάλο βαθμό τη θεωρία των πιθανοτήτων. Μεγάλα, ήδη χαρακτηρισμένα κείμενα χρησιμοποιούνται για την εκπαίδευση αλγορίθμων μηχανικής μάθησης. Οι χαρακτηρισμοί είναι αδιαμφισβήτητοι και αποτελούν τη βάση για τη δημιουργία μοντέλων εκτίμησης. Με τον τρόπο αυτό, οι πολυάριθμοι λεπτομερείς κανόνες, αντικαταστάθηκαν με πληροφορίες στατιστικής συχνότητας που αναλύονται προκειμένου να οδηγήσουν σε συμπεράσματα. Άλλες προσεγγίσεις δημιουργούν πιθανολογικούς κανόνες από χαρακτηρισμένα δεδομένα, μέσα από τη δημιουργία δέντρων απόφασης από δεδομένα οργανωμένα σε διανύσματα. Οι στατιστικές προσεγγίσεις δίνουν καλά αποτελέσματα στην πράξη απλώς και μόνο επειδή, μαθαίνοντας με άφθονα πραγματικά δεδομένα, χρησιμοποιούν τις πιο συνηθισμένες περιπτώσεις. Όσο περισσότερα και αντιπροσωπευτικά είναι τα δεδομένα, τόσο ποιοτικότερη γίνεται η επεξεργασία. (Lexalytics, 2022)

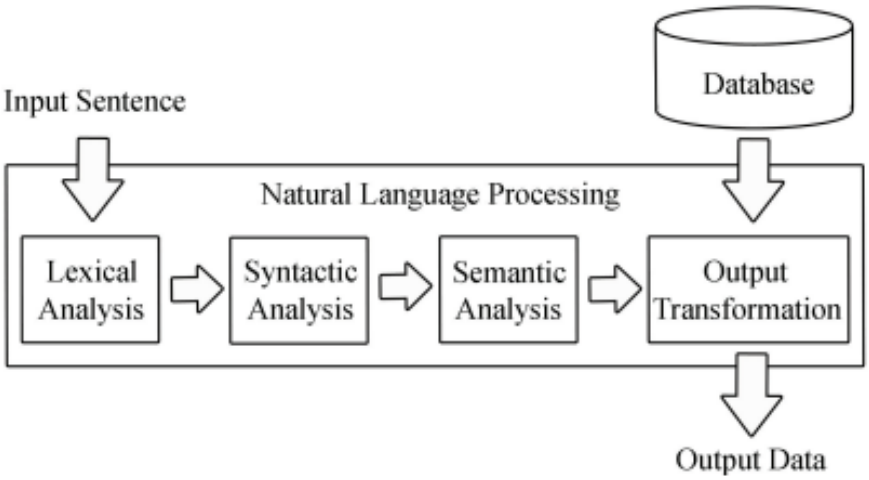
3.3. Διαδικασία NLP

Οι διαδικασίες επεξεργασίας φυσικής γλώσσας, περιλαμβάνουν τα ακόλουθα στάδια(Tapsai, Meesad & Haruechaiyasak, 2016):

- Λεξιλογική ανάλυση (lexical analysis): Αυτό το βήμα αναλύει προτάσεις φυσικής γλώσσας χωρίζοντας σε μικρά στοιχεία που το καθένα ονομάζεται Token. Επιπλέον, τα για τα Tokens προσδιορίζονται τύποι και ορισμένες βασικές πληροφορίες θα χρησιμοποιηθούν στο επόμενο βήμα.
- Συντακτική Ανάλυση (syntactic analysis): Σε αυτό το βήμα, όλα τα διακριτά στοιχεία αναλύονται με βάση προκαθορισμένη δομή προτάσεων, προκειμένου να ελέγχει ο βαθμός εγκυρότητας τους. Επιπλέον παρέχονται ορισμένες πληροφορίες που θα χρησιμοποιηθούν στη διαδικασία ανάλυσης νοήματος.

- Σημασιολογική ανάλυση (semantic analysis): Η διαδικασία σημασιολογικής ανάλυσης ερμηνεύει το νόημα μιας πρότασης αναλύοντας πληροφορίες, οι οποίες προέρχονται από το προηγούμενο βήμα, με μια σημασιολογική δομή όπως μια οντολογία ή μια δομή σημασιολογικού ιστού για να παρέχει ορισμένα δεδομένα που αντιπροσωπεύουν το νόημα μιας πρότασης.
- Διαδικασία μετασχηματισμού εξόδου (Output Transformation Process): Αυτό το βήμα μετατρέπει τα αποτελέσματα που προέρχονται από τη σημασιολογική ανάλυση σε αποτελέσματα που πληροί τους στόχους της συνολικής εργασίας των στόχων, όπως εντολές SQL για ανάκτηση πληροφοριών από βάσεις δεδομένων (Tapsai, Meesad & Haruechaiyasak, 2016).

Σχηματικά η διαδικασία φαίνεται στην παρακάτω εικόνα.



Εικόνα 9 Γενική Διαδικασία Επεξεργασίας Φυσικής Γλώσσας

Για να γίνουν κατανοητές οι φυσικές γλώσσες, είναι σημαντικό να γίνει διάκριση των ακόλουθων αλληλοεξαρτώμενων επιπέδων, που χρησιμοποιούν οι άνθρωποι για να εξάγουν νόημα από κείμενο ή ομιλούμενες γλώσσες:

- Φωνητικό ή φωνολογικό επίπεδο που ασχολείται με την προφορά
- Μορφολογικό επίπεδο που ασχολείται με τα μικρότερα μέρη λέξεων, που φέρουν νόημα, επιθήματα και προθέματα

- Λεξιλογικό επίπεδο που ασχολείται με τη λεξιλογική σημασία των λέξεων και τα μέρη του λόγου που ανήκουν
- Συντακτικό επίπεδο που ασχολείται με τη γραμματική και τη δομή των προτάσεων
- Σημασιολογικό επίπεδο που ασχολείται με τη σημασία λέξεων και προτάσεων
- Επίπεδο λόγου που ασχολείται με τη δομή διαφορετικών ειδών κειμένου χρησιμοποιώντας δομές εγγράφων
- Πραγματιστικό επίπεδο που ασχολείται με τη γνώση που προέρχεται από τον έξω κόσμο, δηλαδή από έξω από το περιεχόμενο του εγγράφου.

Οι εργασίες χαμηλού επιπέδου NLP περιλαμβάνουν:

- Ανίχνευση ορίων προτάσεων: Στο στάδιο αυτό επιχειρείται η απομόνωση διακριτών προτάσεων. Τα σημεία στίξης αποτελούν το σημαντικότερο στοιχείο οριοθέτησης προτάσεων, ωστόσο υπάρχουν λέξεις, εκφράσεις και συντομογραφίες (π.χ. οι τίτλοι - m.g. - Dr.) που περιπλέκουν αυτήν την επεξεργασία.
- Tokenization: Μέσα στην πρόταση πρέπει να προσδιοριστούν τα μεμονωμένα σημεία (λέξεις, σημεία στίξης). Ένας lexer παίζει βασικό ρόλο για αυτήν την εργασία αφού έχει τη δυνατότητα να εντοπίζει λεκτικά πρότυπα μέσα στο κείμενο.
- Ανάθεση μέρους του λόγου σε μεμονωμένες λέξεις: Στο στάδιο αυτό, οι μεμονωμένες λέξεις που εντοπίστηκαν, αντιστοιχίζονται στα μέρη του λόγου.
- Μορφολογική αποσύνθεση σύνθετων λέξεων: Στα κείμενα και τον προφορικό λόγο, συχνά χρησιμοποιούνται σύνθετες λέξεις. Για τη NLP είναι απαραίτητη η αποσύνθεση του στα μέρη που την αποτελούν.
- Λημματοποίηση: Πρόκειται για τη διεργασία κατά την οποία μία λέξη μετατρέπεται στη ρίζα της, μετά την αφαίρεση των επιθημάτων της. Με τον τρόπο αυτό ομαδοποιούνται εκφράσεις οι οποίες προέρχονται από την ίδια ρίζα – άρα έχουν το ίδιο νόημα

– στο ίδιο σύνολο, ώστε να αντιμετωπίζονται με κοινό τρόπο κατά την επεξεργασία.

- Ρηχή ανάλυση (τεμαχισμός): Στο στάδιο αυτό αναγνωρίζονται φράσεις από συστατικά διακριτικά με ετικέτα μέρους του λόγου.
- Τμηματοποίηση: Οι λέξεις που εντοπίζονται, συγκροτούν τμήματα που αποδίδουν συγκεκριμένο νόημα.

3.4. Επεξεργασία Φυσικής Γλώσσας σε Κείμενα

Ο χειρισμός κειμένων για εξαγωγή γνώσης, αυτόματη ευρετηρίαση, περίληψη, ή παραγωγή κειμένου σε επιθυμητή μορφή, έχει αναγνωριστεί ως σημαντικός τομέας έρευνας στο NLP. Αυτό ταξινομείται ευρέως ως η περιοχή επεξεργασίας κειμένου φυσικής γλώσσας που επιτρέπει τη δόμηση μεγάλων σωμάτων δεδομένων κειμένου, με σκοπό την ανάκτηση συγκεκριμένων πληροφοριών ή την εξαγωγή δομών γνώσης που μπορούν να χρησιμοποιηθούν για συγκεκριμένο σκοπό. Τα συστήματα αυτόματης επεξεργασίας κειμένου λαμβάνουν γενικά κάποια μορφή εισαγωγής κειμένου και το μετατρέπουν σε έξοδο διαφορετικής μορφής. Ο βασικός στόχος για τα συστήματα επεξεργασίας κειμένου φυσικής γλώσσας είναι η μετάφραση δυνητικά διφορούμενων ερωτημάτων και κειμένων φυσικής γλώσσας σε σαφείς εσωτερικές αναπαραστάσεις στις οποίες μπορεί να πραγματοποιηθεί αντιστοίχιση και ανάκτηση. Ένα σύστημα επεξεργασίας κειμένου φυσικής γλώσσας μπορεί να ξεκινήσει με μορφολογικές αναλύσεις. Η προέλευση των όρων, τόσο στα ερωτήματα όσο και στα έγγραφα, γίνεται για να ληφθούν οι μορφολογικές παραλλαγές των λέξεων που εμπλέκονται. Η λεξιλογική και συντακτική επεξεργασία περιλαμβάνει τη χρήση λεξικών για τον προσδιορισμό των χαρακτηριστικών των λέξεων, την αναγνώριση των τμημάτων του λόγου τους, τον προσδιορισμό των λέξεων και των φράσεων και την ανάλυση των προτάσεων (Chowdhury, 2003).

3.5. Προσεγγίσεις

Η στατιστική και η μηχανική μάθηση περιλαμβάνουν ανάπτυξη (ή χρήση) αλγορίθμων που επιτρέπουν σε ένα πρόγραμμα να συνάγει μοτίβα σχετικά με δεδομένα παραδείγματος («εκπαίδευση»), που με τη σειρά του επιτρέπουν να

«γενικεύει»— να κάνει προβλέψεις για νέα δεδομένα. Κατά τη φάση εκμάθησης , οι αριθμητικές παράμετροι που χαρακτηρίζουν το υποκείμενο μοντέλο ενός δεδομένου αλγορίθμου υπολογίζονται βελτιστοποιώντας ένα αριθμητικό μέτρο, συνήθως μέσω μιας επαναληπτικής διαδικασίας (Turing, 2022).

Γενικά, η μάθηση μπορεί να εποπτεύεται —κάθε στοιχείο στα δεδομένα εκπαίδευσης επισημαίνεται με τη σωστή απάντηση—αλλά και χωρίς επίβλεψη , όπου δεν είναι, και η διαδικασία μάθησης προσπαθεί να αναγνωρίσει μοτίβα αυτόματα (όπως στην ανάλυση συστάδων και παραγόντων) (Alzubi, 2018). Μια παγίδα σε κάθε μαθησιακή προσέγγιση είναι η πιθανότητα υπερβολικής προσαρμογής : το μοντέλο μπορεί να ταιριάζει σχεδόν τέλεια στα δεδομένα του παραδείγματος, αλλά κάνει κακές προβλέψεις για νέες, μη εμφανείς στο παρελθόν περιπτώσεις. Αυτό συμβαίνει επειδή μπορεί να μάθει τον τυχαίο θόρυβο στα δεδομένα εκπαίδευσης και όχι μόνο τα βασικά, επιθυμητά χαρακτηριστικά του. Ο κίνδυνος υπερβολικής προσαρμογής ελαχιστοποιείται με τεχνικές όπως η διασταυρούμενη επικύρωση, τα οποία διαιρούν τυχαία τα δεδομένα του παραδείγματος σε σύνολα εκπαίδευσης και δοκιμών για να επικυρώσουν εσωτερικά τις προβλέψεις του μοντέλου. Αυτή η διαδικασία κατάτμησης δεδομένων, εκπαίδευσης και επικύρωσης επαναλαμβάνεται σε πολλούς γύρους και στη συνέχεια υπολογίζεται ο μέσος όρος των αποτελεσμάτων επικύρωσης μεταξύ των γύρων (Brownlee, 2016).

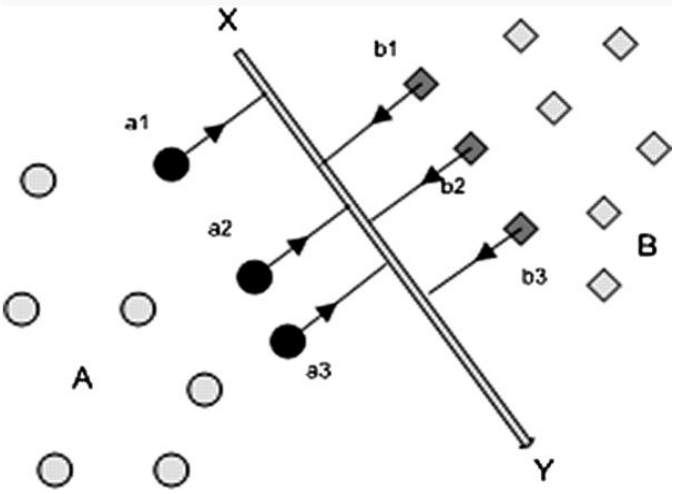
Τα μοντέλα μηχανικής μάθησης μπορούν να ταξινομηθούν ευρέως ως παραγωγικά ή διακριτικά. Οι παραγωγικές μέθοδοι επιδιώκουν να δημιουργήσουν πλούσια μοντέλα κατανομών πιθανοτήτων και ονομάζονται έτσι επειδή, με τέτοια μοντέλα, μπορεί κανείς να «δημιουργήσει» συνθετικά δεδομένα. Οι μέθοδοι διάκρισης είναι πιο χρηστικές, εκτιμώντας άμεσα τις μεταγενέστερες πιθανότητες με βάση τις παρατηρήσεις. Για παράδειγμα, για τον προσδιορισμό της γλώσσας ενός άγνωστου ομιλητή, οι παραγωγικές προσεγγίσεις θα εφαρμόσουν βαθιά γνώση πολλών γλωσσών για την εκτέλεση του αγώνα. Οι μέθοδοι διάκρισης θα βασίζονταν σε μια λιγότερο εντατική προσέγγιση της χρήσης διαφορών μεταξύ των γλωσσών για την εύρεση της πλησιέστερης αντιστοιχίας. Σε σύγκριση με τα παραγωγικά μοντέλα, τα οποία μπορούν να γίνουν δυσεπίλυτα όταν χρησιμοποιούνται πολλά χαρακτηριστικά, τα διακριτικά μοντέλα επιτρέπουν συνήθως τη χρήση περισσότερων

χαρακτηριστικών. 61 Η λογιστική παλινδρόμηση και τα τυχαία πεδία υπό όρους (CRF) είναι παραδείγματα διακριτικών μεθόδων, ενώ οι ταξινομητές Naive Bayes και τα κρυφά μοντέλα Markov (HMM) είναι παραδείγματα μεθόδων παραγωγής (Anand, 2014).

Μερικές κοινές μέθοδοι μηχανικής μάθησης που χρησιμοποιούνται σε εργασίες NLP και χρησιμοποιούνται από πολλά άρθρα σε αυτό το τεύχος, συνοψίζονται παρακάτω.

3.5.1. Support Vector Machines

Τα SVM ταξινομούν τις εισόδους (π.χ. λέξεις) σε κατηγορίες (π.χ. μέρη ομιλίας) με βάση ένα σύνολο χαρακτηριστικών. Η είσοδος μπορεί να μετασχηματιστεί μαθηματικά χρησιμοποιώντας μια συνάρτηση πυρήνα για τον γραμμικό διαχωρισμό των σημείων δεδομένων από διαφορετικές κατηγορίες. Στη γενική περίπτωση εξέτασης N χαρακτηριστικών, ο διαχωριστής είναι ένα (N−1) υπερεπίπεδο. Η πιο κοινή συνάρτηση πυρήνα που χρησιμοποιείται είναι μια Gaussian (αποτελεί τη βάση της κανονικής κατανομής στα στατιστικά στοιχεία). Η διαδικασία διαχωρισμού επιλέγει ένα υποσύνολο των δεδομένων εκπαίδευσης (τα διανύσματα υποστήριξης που αναπαριστούν σημεία δεδομένων πλησιέστερα στο υπερεπίπεδο) που διαφοροποιεί καλύτερα τις κατηγορίες. Το διαχωριστικό υπερεπίπεδο μεγιστοποιεί την απόσταση για να υποστηρίξει διανύσματα από κάθε κατηγορία.



Εικόνα 10 Σχηματική Αναπαράσταση χρήσης των Support Vector Machines

Στην παραπάνω εικόνα φαίνεται μία περίπτωση χρήσης των Support Vector Machines. Τα σημεία δεδομένων, που εμφανίζονται ως κατηγορίες A και B, μπορούν να διαχωριστούν με μια ευθεία γραμμή X–Y. Ο αλγόριθμος που προσδιορίζει το X–Y προσδιορίζει τα σημεία δεδομένων από κάθε κατηγορία που είναι πιο κοντά στην άλλη κατηγορία (a1, a2, a3 και b1, b2, b3) και υπολογίζει το X–Y έτσι, ώστε το περιθώριο που διαχωρίζει τις κατηγορίες εκατέρωθεν μεγιστοποιείται (Gandhi, 2018).

3.5.2. Hidden Markov Models

Ένα Hidden Markov Model (HMM) είναι ένα σύστημα όπου μια μεταβλητή μπορεί να εναλλάσσεται (με ποικίλες πιθανότητες) μεταξύ πολλών καταστάσεων, δημιουργώντας ένα σύνολο πιθανών εξόδων (με διαφορετικές πιθανότητες). Τα σύνολα των πιθανών καταστάσεων και των μοναδικών συμβόλων μπορεί να είναι μεγάλα, αλλά πεπερασμένα και γνωστά. Η μεθοδολογία των HMM περιλαμβάνει την επίλυση των παρακάτω προβλημάτων:

- Συμπέρασμα: Δεδομένης μιας συγκεκριμένης ακολουθίας συμβόλων εξόδου, υπολογισμός των πιθανοτήτων μιας ή περισσότερων υποψήφιων ακολουθιών μεταγωγής καταστάσεων.
- Αντιστοίχιση μοτίβων: Υπολογισμός της ακολουθίας εναλλαγής κατάστασης που είναι πιο πιθανό να έχει δημιουργήσει μια συγκεκριμένη ακολουθία συμβόλων εξόδου.
- Εκπαίδευση: Με την είσοδο παραδειγμάτων ακολουθίας συμβόλων εξόδου (εκπαίδευση), υπολογισμός των πιθανοτήτων μεταγωγής κατάστασης/εξόδου που ταιριάζουν καλύτερα σε αυτά τα δεδομένα.

Για την επίλυση αυτών των προβλημάτων, ένα HMM χρησιμοποιεί δύο απλοποιητικές υποθέσεις:

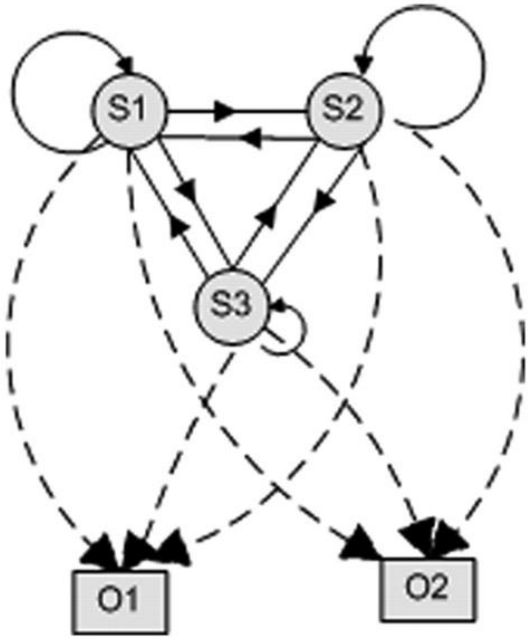
- Η πιθανότητα μετάβασης σε μια νέα κατάσταση (ή πίσω στην ίδια κατάσταση) εξαρτάται από τις προηγούμενες N καταστάσεις.

- Η πιθανότητα δημιουργίας μιας συγκεκριμένης παραγωγής σε μια συγκεκριμένη κατάσταση εξαρτάται μόνο από αυτήν την κατάσταση.

Αυτές οι παραδοχές επιτρέπουν τον υπολογισμό της πιθανότητας μιας δεδομένης ακολουθίας μεταγωγής καταστάσεων (και μιας αντίστοιχης ακολουθίας παρατήρησης εξόδου) με απλό πολλαπλασιασμό των επιμέρους πιθανοτήτων.

Τα HMM χρησιμοποιούνται ευρέως για την αναγνώριση ομιλίας, όπου η κυματομορφή μιας προφορικής λέξης (η ακολουθία εξόδου) ταιριάζει με την ακολουθία μεμονωμένων φωνημάτων (καταστάσεις) που πιθανότατα την παρήγαγαν (Jurafsky & Martin, 2023).

Στην επόμενη εικόνα φαίνεται συνοπτικά η λειτουργία ενός HMM. Οι μικροί κύκλοι S1, S2 και S3 αντιπροσωπεύουν καταστάσεις . Τα πλαίσια O1 και O2 αντιπροσωπεύουν τιμές εξόδου. Οι καταστάσεις σύνδεσης συμπαγών γραμμών/τόξων αντιπροσωπεύουν διακόπτες κατάστασης. Το βέλος αντιπροσωπεύει την κατεύθυνση του διακόπτη. Κάθε ετικέτα γραμμής/τόξου είναι η πιθανότητα αλλαγής. Μια διακεκομμένη γραμμή/τόξο που συνδέει μια κατάσταση με μια τιμή εξόδου υποδηλώνει πιθανότητα εξόδου (η πιθανότητα αυτή η τιμή εξόδου να δημιουργηθεί από τη συγκεκριμένη κατάσταση). Εάν μια συγκεκριμένη πιθανότητα διακόπτη/εξόδου είναι μηδέν, η γραμμή/το τόξο δεν σχεδιάζεται. Το άθροισμα των πιθανοτήτων του διακόπτη που αφήνει μια δεδομένη κατάσταση είναι ίσο με 1.



Εικόνα 11 Αρχιτεκτονική Hidden Markov Models

3.5.3. Conditional Random Fields

Τα CRF είναι μια οικογένεια διακριτικών μοντέλων. Τα συνηθέστερα (γραμμικής αλυσίδας) CRF μοιάζουν με HMM καθώς η επόμενη κατάσταση εξαρτάται από την τρέχουσα κατάσταση. Γενικεύουν την λογιστική παλινδρόμηση σε διαδοχικά δεδομένα. Χρησιμοποιούνται για την πρόβλεψη των μεταβλητών κατάστασης με βάση τις παρατηρούμενες μεταβλητές (Wallach, 2004).

3.5.4. N-grams

Ένα N-gram 19 είναι μια ακολουθία από N στοιχεία (γράμματα, λέξεις ή φωνήματα). Αν είναι γνωστό ότι ορισμένα ζεύγη αντικειμένων (ή τριπλέτα κοκ) είναι πιθανό να εμφανίζονται πολύ πιο συχνά από άλλα, τότε αυτό μπορεί να αποτελέσει τη βάση για την ανάπτυξη σχετικών μοτίβων. Υπάρχουν πολλοί γραμματικοί κανόνες που προσδιορίζουν ρητά ποια γράμματα μπορεί να ακολουθούν συγκεκριμένα γράμματα. Αυτό συμβαίνει και σε επίπεδο λέξεων και εκφράσεων. Αν διατίθενται επαρκή δεδομένα, είναι εφικτό να προσδιοριστούν δεδομένα κατανομής συχνότητας

για όλα τα N-grams που εμφανίζονται. Το N συνήθως περιορίζεται γιατί μεγάλες του τιμές αντιστοιχούν σε μεγάλα σύνολα εκφράσεων.

Τα N-grams είναι ένα είδος μοντέλου Markov πολλαπλών τάξεων: η πιθανότητα ενός συγκεκριμένου στοιχείου στη Nη θέση εξαρτάται από τα προηγούμενα N−1 στοιχεία και μπορεί να υπολογιστεί. Αφού υπολογιστούν, τα N-grams, είναι δυνατό να υποστηριχθούν οι παρακάτω σκοπιμότητες:

- Προτεινόμενη αυτόματη συμπλήρωση λέξεων και φράσεων στον χρήστη κατά την αναζήτηση.
- Διόρθωση ορθογραφίας: μια ανορθόγραφη λέξη σε μια φράση μπορεί να επισημανθεί και να προτείνεται μια σωστή ορθογραφία με βάση τις σωστά γραμμένες γειτονικές λέξεις.
- Αναγνώριση ομιλίας: τα ομόφωνα μπορούν να αποσαφηνιστούν πιθανολογικά με βάση τις σωστά αναγνωρισμένες γειτονικές λέξεις.
- Αποσαφήνιση λέξεων: εάν δημιουργηθούν N-grams με σημασία λέξης από ένα σχολιασμένο σώμα όπου τα ομόγραφα επισημαίνονται με τις σωστές τους έννοιες, μπορούν να χρησιμοποιηθούν οι μη διφορούμενες γειτονικές λέξεις για την πρόβλεψη της σωστής σημασίας ενός ομογράφου σε ένα έγγραφο.

Τα αρχεία που περιλαμβάνουν N-grams είναι μεγάλα σε μέγεθος αλλά καθώς οι αποθηκευτικές διατάξεις έχουν γίνει πιο προσιτές, αυτό είναι ένα ζήτημα που αντιμετωπίζεται αποδοτικά. Ειδικές δομές δεδομένων, που ονομάζονται δείκτες N-gram pointers, επιταχύνουν την αναζήτηση τέτοιων δεδομένων. Οι ταξινομητές που βασίζονται σε N-grams αξιοποιούν το ακατέργαστο κείμενο χωρίς ρητή γνώση του πλαισίου του, με ικανοποιητικά αποτελέσματα απόδοσης (Nadkarni, Ohno-Machado και Chapman 2011).

Κεφάλαιο 4: Ανάλυση Συναισθήματος

Η διαθεσιμότητα διαδικτυακού περιεχομένου τα τελευταία χρόνια έχει αυξηθεί ραγδαία. Σε αυτό συνετέλεσαν παράγοντες όπως:

- Η πρόσβαση σε υψηλής ποιότητα διαδικτυακές συνδέσεις έχει γίνει πολύ πιο προσιτή σε σχέση με το πρόσφατο παρελθόν. Αυτό οδήγησε σε αύξηση του διαδικτυακού κοινού.
- Με την αύξηση του διαδικτυακού κοινού, διευρύνθηκε η σκοπιμότητα των διαδικτυακών εφαρμογών του (για να καλυφθούν και οι απαιτήσεις των νέων χρηστών) και αυξήθηκε και όγκος του διαθέσιμου περιεχομένου στις υποδομές του διαδικτύου.

Οι χρήστες του διαδικτύου πλέον συμμετέχουν ενεργά στη διαμόρφωση του περιεχομένου του. Έχουν πλέον τη δυνατότητα να αναρτούν τις απόψεις τους σε μέσα κοινωνικής δικτύωσης, να σχολιάζουν ήδη δημοσιευμένο περιεχόμενο ή να το αξιολογούν. Η θεματολογία του περιεχομένου, στη διαμόρφωση του οποίου συμμετέχουν ποικίλει και μπορεί να περιλαμβάνει: πολιτικές εξελίξεις, γεγονότα κάθε είδους, καταναλωτικά προϊόντα κα. Μπορεί επίσης να δημοσιεύουν με τους παραπάνω τρόπους απλά και μόνο την ψυχολογική τους διάθεση. Τα δεδομένα που προκύπτουν, είναι διαθέσιμα και μπορούν να αποτελέσουν ερευνητικές διαδικασίας με απώτερο στόχο τη διάγνωση των τάσεων της κοινής γνώμης.

Η διάγνωση των τάσεων αυτών είναι πολύτιμη για κάθε είδους οργανισμό προκειμένου να χρησιμοποιηθούν στις διαδικασίες λήψης αποφάσεων. Η γνώση του τρόπου με τον οποίο σκέφτεται το κοινό, μπορεί να αποτελέσει το κριτήριο για την αλλαγή στρατηγικής ή τακτικών προσεγγίσεων ικανοποίησης των στρατηγικών στόχων. Η διαθεσιμότητα επαρκών στοιχείων στο διαδίκτυο, απαλλάσσει τους οργανισμούς από την αναγκαιότητα να προβαίνουν σε χρονοβόρες και υψηλού κόστους έρευνες σφιγμομέτρησης της κοινής γνώμης. Αντίθετα πλέον χρησιμοποιούνται τεχνικές ανάκτησης των δεδομένων που διατίθενται στο διαδίκτυο και μέσα από την κατάλληλη επεξεργασία του, δημιουργία χρήσιμων συμπερασμάτων.

Ένας από τους τρόπους εκμετάλλευσης του διαδικτυακού περιεχομένου στις διαδικασίες λήψης αποφάσεων, είναι η ανάλυση συναισθήματος. Η Ανάλυση Συναισθήματος αναφέρεται σε διαδικασίες κατά τις οποίες ο βασικός στόχος είναι να ανακαλυφθεί ο συναισθηματικός προσανατολισμό των χρηστών του διαδικτύου, σχετικά με κάποιο θέμα. Η θεματολογία μπορεί να προέρχεται από ποικίλες ανθρώπινες δραστηριότητες (πολιτικές, οικονομικές καταναλωτικές κτλ.). Για την εξυπηρέτηση των απαιτήσεων της Ανάλυσης Συναισθήματος, χρησιμοποιούνται ένα πλήθος μεθόδων, τεχνικών και αλγορίθμων. Αυτά στοχεύουν στον εντοπισμό των συναισθημάτων που αναδύονται από κείμενα που είναι αναρτημένα στο διαδίκτυο. Δεδομένου ότι πλέον ένα πολύ μεγάλο ποσοστό των ανθρώπων σε παγκόσμια κλίμακα, χρησιμοποιεί τα κοινωνικά δίκτυα και τις δυνατότητες σχολιασμού και αξιολόγησης του διαδικτυακού περιεχομένου, η Ανάλυσης Συναισθήματος από διαδικτυακό περιεχόμενο αποκτά μεγαλύτερη αξία.

Στις περισσότερες περιπτώσεις, αναζητείται το κυρίαρχο συναίσθημα σε επίπεδο πρότασης. Οι αλγόριθμοι που χρησιμοποιούνται έχουν σαν έξοδο το αν το συναίσθημα που χαρακτηρίζει την πρόταση είναι αρνητικό, θετικό ή ουδέτερο. Είναι επιπλέον πιθανό να αναζητείται και η ένταση του συναισθήματος βασισμένη σε κάποια προσδιορισμένη κλίμακα.

Δύο είναι οι βασικές προσεγγίσεις των αλγορίθμων που χρησιμοποιούνται: χρήση λεξικών συναισθημάτων και χρήση τεχνικών Μηχανικής Μάθησης. Η προσέγγιση των λεξικών συναισθημάτων στηρίζεται στη χρήση λεξικών που περιέχουν αντιστοιχίες λέξεων και εκφράσεων με συναισθήματα. Οι αλγόριθμοι της Μηχανικής Μάθησης χρησιμοποιεί τεχνικές επεξεργασίας φυσικής γλώσσας σε συνδυασμό με μηχανισμούς και τεχνικές εκπαίδευσης μοντέλων πρόβλεψης προκειμένου να αναπτυχθούν και να προετοιμαστούν αξιόπιστα μοντέλα αναγνώρισης συναισθημάτων (Mayo, 2018). Η καταλληλότητα της κάθε προσέγγισης σε διαφορετικά έργα αναγνώρισης συναισθήματος εξαρτάται:

- Από το αν για τη θεματολογία στην οποία αναφέρεται το έργο ανάλυσης συναισθήματος, υπάρχει ήδη λεξικό συναισθήματος.

- Από το αν υπάρχουν επαρκής αριθμός ήδη χαρακτηρισμένων κειμένων για να χρησιμοποιηθούν προς εκπαίδευση των μοντέλων πρόβλεψης. (Taboada, Brooke, Tofiloski, Voll, & Stede, 2014).

Πρακτικά η Ανάλυση Συναισθήματος αποτελεί μία επέκταση της επεξεργασίας φυσικής γλώσσας. Τα κείμενα που χρησιμοποιούνται αναρτώνται από απλούς χρήστες του διαδικτύου. Επομένως σε πολλές περιπτώσεις περιλαμβάνουν όλες τις σχετικές αδυναμίες όπως συντομογραφίες, χρήση εικονιδίων, «αργκό», κακή ορθογραφία, κακή σύνταξη. Είναι απαραίτητη πάντα μία προκαταρκτική διεργασία για τον καθαρισμό των κειμένων ώστε να αποκτήσουν κατάλληλη δομή και περιεχόμενο για να καταστούν επεξεργάσιμα.

4.1. Λεξικά Συναισθημάτων

Πρόκειται για απλές μεθόδους που παρέχουν ικανοποιητικά αποτελέσματα στην ανάδειξη του συναισθήματος από διαδικτυακά κείμενα. Απαραίτητη προϋπόθεση για την επιτυχία των μεθόδων αυτών είναι να παρέχονται πλήρη και αξιόπιστα λεξικά πάνω στη θεματολογία του έργου που θα υποστηρίξει η διαδικασία ανάλυσης συναισθήματος. Στα λεξικά αυτά χρειάζεται να υπάρχουν αντιστοιχία λέξεων και συναισθημάτων της μορφής:

[<ΛΕΞΗ>:<ΣΥΝΑΙΣΘΗΜΑ: [ΑΡΝΗΤΙΚΟ | ΘΕΤΙΚΟ | ΟΥΔΕΤΕΡΟ]>]

Το είδος συναισθήματος που χαρακτηρίζει ένα κείμενο, προσδιορίζεται από τη συχνότητα εμφάνισης των λέξεων ή φράσεων του λεξικού. Τα επίθετα και τα επιρρήματα είναι συνήθως τα μέρη του λόγου που καθορίζουν τον χαρακτηρισμό του κειμένου, καθώς χρησιμοποιούνται για να περιγράψουν οντότητες.

Τα λεξικά μπορεί να προσδιορίζουν και τον βαθμό αντιστοίχισης της λέξης ή της έκφρασης με το αντίστοιχο συναίσθημα, με βάση μία προκαθορισμένη κλίμακα. Επιπλέον, τα λεξικά μπορεί να επικεντρώνονται σε συγκεκριμένη θεματολογία(Musto, et al., 2015).

4.2. Λειτουργία

Η ανάλυση συναισθήματος με τη χρήση λεξικών, περιλαμβάνει αρχικά ένα στάδιο κατά το οποίο το κείμενο διαχωρίζεται στις προτάσεις που το απαρτίζουν. Στη συνέχεια κάθε πρόταση αναλύεται περαιτέρω στις λέξεις που το απαρτίζουν. Για κάθε μία λέξη ή έκφραση που εντοπίζεται, γίνεται έλεγχος για το αν περιλαμβάνεται στο λεξικό συναισθημάτων. Για κάθε μία τέτοια λέξη ή έκφραση, σημειώνεται η συχνότητα της στο κείμενο. Με τον τρόπο αυτό υπολογίζεται το πόσες και με ποια ένταση αρνητικές, θετικές ή ουδέτερες εκφράσεις περιέχονται στο κείμενο. Το είδος των εκφράσεων που παρουσιάζει μεγαλύτερη συχνότητα και ένταση, καθορίζει και τον χαρακτήρα του κειμένου.

Η διαδικασία αυτή παρουσιάζει τα ακόλουθα πλεονεκτήματα:

- Τα αρχεία των λεξικών συναισθημάτων μπορούν να χρησιμοποιηθούν σε πολλαπλές εργασίες ανάλυσης συναισθήματος. Επίσης, δύνανται να εμπλουτίζονται με νέες λέξεις και εκφράσεις.
- Τα περισσότερα αρχεία λεξικών συναισθημάτων έχουν δημιουργηθεί με σημαντική συμμετοχή του ανθρώπινου παράγοντα. Αυτό εξασφαλίζει μεγαλύτερη αξιοπιστία και ακρίβεια της σημασιολογίας των λέξεων και των εκφράσεων που περιλαμβάνουν.
- Η συνεισφορά των λεξικών στην αποδοτικότητα των εργασιών ελέγχου συναισθήματος κειμένου είναι σταθερή, ανεξάρτητα από τη θεματολογία τους.

Τα λεξικά συναισθημάτων διακρίνονται στις ακόλουθες κατηγορίες:

- Word Based: Είναι λεξικά που αναπτύσσονται με μη αυτοματοποιημένο τρόπο με ανθρώπινη παρέμβαση. Έχουν σχετικά μικρό μέγεθος και κάθε ένα αντιστοιχεί σε συγκεκριμένο θέμα. Μέσα από την συχνή ενημέρωση τους αναβαθμίζουν την αποδοτικότητα τους. Η ενημέρωση περιλαμβάνει πέρα από την προσθήκη νέων λέξεων και εκφράσεων, εμπλουτισμό με συνώνυμα και αντίθετα αλλά και τις διαφορετικές σημασίες των λέξεων συνάρτηση των λέξεων που υπάρχουν κοντά τους στο κείμενο.

- Corpus Based: Τα λεξικά αυτού τους είδους χρησιμοποιούνται με την αντιστοίχιση του συνδυασμού των λέξεων και των οντοτήτων που προσδιορίζουν, με συναίσθημα. Οι αλγόριθμοι που τα χρησιμοποιούν εξετάζουν τον τρόπο με τον οποίο συνδέονται σε ένα κείμενο εκφράσεις που έχουν μία δεδομένη συναισθηματική σημασία. Οι εκφράσεις αυτές μπορεί να ποικίλουν ως προς το μέγεθος τους. Η δημιουργία των λεξικών αυτών και η ενημέρωση τους έχουν μεγαλύτερη πολυπλοκότητα αλλά παρουσιάζουν καλύτερη ακρίβεια στη χρήση τους σε εργασίες ανάλυσης συναισθήματος.
- Dictionary Based: Παρουσιάζουν ομοιότητα με τα Corpus Based. Η διαφορά τους έγκειται στο γεγονός ότι ο χαρακτηρισμός των όρων που περιλαμβάνουν δε γίνεται σε σχέση με την οντότητα που περιγράφουν αλλά ανεξάρτητα.

Στον παρακάτω πίνακα παρουσιάζονται μερικά από τα γνωστότερα λεξικά συναισθημάτων (Loughran & McDonald, 2011).

Ονομασία	Είδος	Περιγραφή	Παρατηρήσεις
SentiWordNet	Word Based	Σε κάθε όρο αντιστοιχίζεται ένα διάνυσμα τριών αριθμών, που χαρακτηρίζουν την ένταση θετικού, αρνητικού ή ουδέτερου συναισθήματος. Προβλέπει τη συμβολή των συμφραζομένων στον χαρακτηρισμό των όρων.	Παρέχεται από διαδικτυακή υπηρεσία.
WordNet-Affect	Corpus Based	Μέσα από μία δενδρική ιεραρχική δομή των όρων που μπορεί να συνυπάρχουν σε φράσεις, αποδίδονται στα φύλλα τους το αναδυόμενο συναίσθημα.	
MPQA	Word Based	Αντιστοιχίζονται όροι με συναισθήματα με βάση μία προκαθορισμένη κλιμακα.	

Ονομασία	Είδος	Περιγραφή	Παρατηρήσεις
SenticNet	Corpus Based	Τα λεξικά αυτά είναι προσανατολισμένα σε συγκεκριμένη θεματολογία το κάθε ένα. Οι όροι αξιολογούνται ως προς το συναίσθημα με μία βαθμολογία στο διάστημα [-3,...3]	Βασιζεται στο Sentic Computing
WordStat Sentiment Dictionary	Word Based	Χρησιμοποιεί αντιστοιχία όρου και συναισθήματος και αξιολογεί και τα συμφραζόμενα.	Βασίζονται σε λεξικά του Harvard και του Pennebaker

4.3. Μηχανική Μάθηση

Η Μηχανική Μάθηση αποτελεί μέρος της Εξόρυξης Γνώσης. Στοχεύει στην ανακάλυψη προτύπων και την αξιολόγηση τους ως προς την ικανότητα τους να παράγουν σωστές προβλέψεις και εκτιμήσεις, όταν σε αυτά εφαρμόζονται τα δεδομένα που προσδιορίζουν τρέχουσες καταστάσεις. Οι αλγόριθμοι μηχανικής μάθησης βασίζονται στην αναζήτηση συσχετίσεων μεταξύ των παραμέτρων που χαρακτηρίζουν οντότητες και καταστάσεις. Η συσχετίσεις που εντοπίζονται, αποτελούν τη βάση για την ανάπτυξη των μοντέλων πρόβλεψης και εκτιμήσεων. Η ανάπτυξη των μοντέλων αυτών βασίζεται σε σύνολα δεδομένων. Η ποσότητα και η ποιότητα των δεδομένων αυτών είναι σημαντικά για την ποιότητα των μοντέλων που θα αναπτυχθούν. Με τον τρόπο αυτό οι αλγόριθμοι της μηχανικής μάθησης αναζητούν τις συσχετίσεις εκείνες των όρων που απαρτίζουν τα κείμενα που οδηγούν σε κάθε έναν από τους συναισθηματικούς χαρακτηρισμούς των κειμένων.

4.4. Μελέτη Περίπτωσης

4.4.1. Περιγραφή

Για να μελετηθεί στην πράξη η ανάλυση συναισθήματος σε αναρτήσεις κοινωνικών δικτύων, θα εξεταστούν τα συναισθήματα που διακατέχουν τους χρήστες του twitter σχετικά με τη Ρωσία και την Ουκρανία. Οι δύο χώρες τελούν σε πολεμική

σύγκρουση μεταξύ τους. Μέσα από τη μελέτη θα αναζητηθεί με ποιο μέρος συντάσσεται η πλειοψηφία του κοινού του διαδικτύου. Για τον σκοπό αυτό θα χρησιμοποιηθεί η υπηρεσία ανάκτησης αναρτήσεων του twitter. Θα γίνει η ανάκτηση δύο συνόλων δεδομένων. Το πρώτο θα περιλαμβάνει περιεχόμενο αναρτήσεων σχετικών με τον Βλαντιμίρ Πούτιν και το άλλο περιεχόμενο αναρτήσεων σχετικών με τον Βολόντιμιρ Ζελένσκι. Για τα σύνολα αυτά θα υπολογιστεί το ποσοστό των θετικών και των αρνητικών αναρτήσεων.

4.4.2. Εργαλεία

4.4.2.1 Python

Η Python είναι μια ερμηνευτική και αντικειμενοστραφής γλώσσα προγραμματισμού υψηλού επιπέδου με δυναμική σημασιολογία. Οι ενσωματωμένες δομές δεδομένων υψηλού επιπέδου, σε συνδυασμό με τη δυναμική πληκτρολόγηση και τη δυναμική σύνδεση, το καθιστούν πολύ ελκυστικό για γρήγορη ανάπτυξη εφαρμογών, καθώς και για χρήση ως γλώσσα σεναρίου ή κόλλας για τη σύνδεση υπαρχόντων στοιχείων μεταξύ τους. Η απλή, εύκολη στην εκμάθηση σύνταξη της Python δίνει έμφαση στην αναγνωσιμότητα και ως εκ τούτου μειώνει το κόστος συντήρησης του προγράμματος. Η Python υποστηρίζει λειτουργικές μονάδες και πακέτα, τα οποία ενθαρρύνουν τη σπονδυλωτότητα του προγράμματος και την επαναχρησιμοποίηση κώδικα. Ο διερμηνέας Python και η εκτεταμένη τυπική βιβλιοθήκη είναι διαθέσιμα σε πηγή ή δυαδική μορφή χωρίς χρέωση για όλες τις μεγάλες πλατφόρμες και μπορούν να διανεμηθούν ελεύθερα.

Η δυνατότητα της Python να επεκτείνεται εύκολα και να μπορεί να ικανοποιήσει ποικιλία απαιτήσεων αυξάνει την παραγωγικότητα της γλώσσας. Δεδομένου ότι δεν υπάρχει στάδιο μεταγλώττισης, ο κύκλος επεξεργασίας-δοκιμής-εντοπισμού σφαλμάτων είναι γρήγορος. Ο εντοπισμός σφαλμάτων σε προγράμματα Python είναι εύκολος. Τα σφάλματα δεν προκαλούν σφάλματα στην κύρια μνήμη, αλλά όταν ανακαλυφθεί σφάλμα δημιουργεί μια εξαίρεση. Όταν το πρόγραμμα δεν μπορεί να αντιμετωπίσει την εξαίρεση, ο διερμηνέας εκτυπώνει την κατάσταση της στοίβας. Ένα πρόγραμμα εντοπισμού σφαλμάτων σε επίπεδο κώδικα επιτρέπει την

επιθεώρηση τοπικών και καθολικών μεταβλητών, την αξιολόγηση εκφράσεων, τον ορισμό σημείων διακοπής, τον έλεγχο του κώδικα γραμμή - γραμμή. Το πρόγραμμα εντοπισμού σφαλμάτων είναι γραμμένο στην ίδια την Python.

Η Python επεξεργάζεται κατά το χρόνο εκτέλεσης από τον διερμηνέα. Δε χρειάζεται να μεταγλωττιστεί το πρόγραμμα πριν εκτελεστεί.

4.4.2.2. Jupyter

Το Jupyter Notebook είναι μια εφαρμογή web ανοιχτού κώδικα που χρησιμοποιείται για την δημιουργία και διανομή κώδικα και των αποτελεσμάτων του. Οι προγραμματιστές μπορούν να γράφουν τμήματα κώδικα σε φιλικές διεπαφές, να τρέχουν τις αντίστοιχες εντολές και να ελέγχουν τα αποτελέσματα των προγραμμάτων τους τμηματικά. Επιπλέον έχουν τη δυνατότητα να μοιραστούν τον κώδικα και το αποτέλεσμά τους σε διάφορες μορφές.

Πρόκειται για spin-off έργο του IPython. Το όνομα, Jupyter, προέρχεται από τις βασικές υποστηριζόμενες γλώσσες προγραμματισμού που υποστηρίζει: Julia, Python και R. Συνοδεύεται από τον πυρήνα IPython, ο οποίος επιτρέπει την ανάπτυξη προγραμμάτων σε Python, αλλά μπορεί να υποστηρίξει και ένα μεγάλο πλήθος άλλων πυρήνων.

Η διεπαφή του notebook υλοποιείται σε browser. Η ιστοσελίδα περιλαμβάνει κελιά στα οποία εγγράφεται ο κώδικας. Ο προγραμματιστής, αφού συμπληρώσει τα κελιά, μπορεί να τρέξει το κώδικα σε κάθε κελί και να ελέγξει το αποτέλεσμα. Με τον τρόπο αυτό η εκτέλεση του συνόλου του προγράμματος διαιρείται σε τμήματα. Η διόρθωσή του μπορεί να γίνεται στα σημεία που χρειάζεται, χωρίς να απαιτείται η επανεκτέλεση όλου το κώδικα για τον έλεγχο της. Αυτό συμβάλλει στην ταχύτερη ανάπτυξη των προγραμμάτων.

Δίνεται επίσης η δυνατότητα για προσθήκη μορφοποιημένων κειμένων σε κελιά. Τα κείμενα αυτά σχετίζονται με σχόλια, κεφαλίδες ή οδηγίες που αφορούν το πρόγραμμα και τα αποτελέσματά του. Οι δυνατότητες αυτές δίνουν την ευκαιρία

στους προγραμματιστές να ενσωματώνουν στον κώδικά τους και την αναλυτική τεκμηρίωσή του.

4.4.3. Twitter Application Programing Interface

Το Twitter API μπορεί να χρησιμοποιηθεί για την ανάκτηση και ανάλυση δεδομένων Twitter. Μπορεί να κληθεί από προγράμματα μιας ποικιλίας γλωσσών προγραμματισμού. Η αποδοχή του, οδήγησε στην εξέλιξή του με την προσθήκη πολλαπλών επιπέδων πρόσβασης, ώστε να κλιμακώνεται η πρόσβαση των διαφόρων κατηγοριών χρηστών. Η τρέχουσα έκδοσή του είναι η Twitter API v2, η οποία περιλαμβάνει προηγμένες δυνατότητες.

Στο Twitter API ο χρήστης μπορεί να κάνει εγγραφή και να αποκτήσει λογαριασμό. Η πρόσβαση σε αυτόν γίνεται με τη χρήση username και password. Για να μπορέσει να χρησιμοποιήσει ο χρήστης τις υπηρεσίες του API θα πρέπει να δημιουργήσει ένα project το οποίο και θα πρέπει να το συσχετίσει με το λογαριασμό του. Το project αυτό συσχετίζεται με μια σειρά από κλειδιά σε μορφή συμβολοσειράς. Με τη χρήση των κλειδιών αυτών, ο προγραμματιστής αποκτά τη δυνατότητα να χρησιμοποιήσει μία σειρά από υπηρεσίες. Η βασικότερη από αυτές ανακτά και επιστρέφει αναρτήσεις χρηστών του Twitter με βάση διάφορα κριτήρια.

Το σημείο που καλεί ο προγραμματιστής με την μέθοδο GET, επιστρέφει εγγραφές για τα ζητούμενα tweets. Οι εγγραφές επιστρέφονται σε μορφή JSON. Κάθε μία από αυτές περιλαμβάνει και τα ακόλουθα πεδία:

- Id: Μοναδικό αναγνωριστικό της ανάρτησης
- Text: Το περιεχόμενο της ανάρτησης
- create_at: Η ημερομηνία και η ώρα που δημιουργήθηκε του tweet
- lang: Η γλώσσα που είναι γραμμένο το tweet.
- Source: Περιγράφει το χρήστη του tweeter που έχει κάνει την ανάρτηση

Για την ανάλυση συναισθήματος, χρειάστηκε το πεδίο text.

4.4.4. Διαδικασία

Η ανάλυση συναισθήματος των περιεχομένων αναρτήσεων στο twitter, πραγματοποιήθηκε με δύο τρόπους:

- Με τη χρήση λεξικών
- Με τη χρήση τεχνικών μηχανικής μάθησης

Και για τις δύο μεθόδους ανακτήθηκαν, χρησιμοποιώντας το API του twitter, 10.000 μηνύματα που αφορούσαν τον πρόεδρο της Ρωσίας, Βλαντιμίρ Πούτιν και 10.000 μηνύματα που αφορούσαν τον πρόεδρο της Ουκρανίας, Βολόντιμιρ Ζελένσκι.

4.4.5. Μεθοδολογίες Μηχανικής Μάθησης

4.4.5.1.Λήψη Δεδομένων Εκπαίδευσης

Για την εκπαίδευση των μοντέλων εκτίμησης συναισθήματος των περιεχομένων σε αναρτήσεις κοινωνικών δικτύων, χρησιμοποιήθηκε το σύνολο δεδομένων training.1600000.processed.noemoticon.csv[1]. Το αρχείο αυτό είναι δωρεάν διαθέσιμο. Περιλαμβάνει 160.000 εγγραφές, κάθε μία από τις οποίες αντιστοιχεί σε μία ανάρτηση στο twitter. Τα πεδία που περιλαμβάνονται σε κάθε εγγραφή είναι:

1. Το είδος του συναισθήματος που χαρακτηρίζει κάθε ανάρτηση. Οι τιμές που λαμβάνει είναι είτε 0 (για αρνητικά συναισθήματα), είτε 4 (για θετικά συναισθήματα).
2. Το αναγνωριστικό του tweet .
3. Η ημερομηνία που δημοσιεύτηκε το tweet

[1] Είναι διαθέσιμο στο https://www.kaggle.com/datasets/ferno2/training1600000processednoemoticoncsv

4. Ένα πεδίο που σε κάθε εγγραφή έχει την τιμή «NO_QUERY»

5. Η ονομασία του αποστολέα του tweet

6. Το περιεχόμενο του tweet.

Τα πεδία που είναι εκμεταλλεύσιμα για την παραγωγή μοντέλων εκτίμησης του συναισθήματος είναι το 1^{ο} και το 6^{ο}. Αν και στην ιστοσελίδα παρουσίασης και λήψης του συνόλου δεδομένων δεν υπάρχει επαρκής τεκμηρίωση, τα πεδία που χρειάζονται για την παραγωγή των μοντέλων είναι προφανή. Στην ιστοσελίδα παρουσιάζονται στατιστικά στοιχεία για τα περιεχόμενα του συνόλου δεδομένων. Από αυτά προκύπτει ότι το σύνολο είναι ισορροπημένο σε σχέση με το συναίσθημα που χαρακτηρίζει κάθε tweet. Επιπλέον, προκύπτει από τα στατιστικά στοιχεία ότι οι αναρτήσεις προέρχονται από 659775 και τα περισσότερα περιεχόμενα είναι διαφορετικά μεταξύ τους (1581465 διαφορετικά tweets).

4.4.5.2. Προεπεξεργασία του Συνόλου Εκπαίδευσης

Μετά τη λήψη του, το σύνολο εκπαίδευσης έπρεπε να υποστεί μία σειρά από επεξεργασίες προκειμένου να καταστεί κατάλληλο για να αποτελέσει την είσοδο σε διαδικασίες μηχανικής μάθησης. Οι επεξεργασίες αυτές περιλαμβάνουν:

- Την απομόνωση του περιεχομένου των αναρτήσεων και των αντίστοιχων συναισθημάτων.
- Την απομάκρυνση λέξεων που δεν είναι γραμμένες στα αγγλικά.
- Το καθαρισμό των περιεχομένων από:
  - Σημεία στίξης
  - Εικονίδια
  - Ειδικούς χαρακτήρες
  - Αποστολείς και παραλήπτες των μηνυμάτων

- Stop words

- Την προσαρμογή του περιεχομένου σε N-grams

Μετά από τις ενέργειες αυτές, το σύνολο δεδομένων είναι έτοιμο να χρησιμοποιηθεί για τη παραγωγή μοντέλου πρόβλεψης.

4.4.5.3. Λήψη του Συνόλου Δεδομένων για το οποίο θα αναζητηθεί το συναίσθημα

Χρησιμοποιώντας το API του twitter, ανακτήθηκαν tweets τα οποία σχετίζονται με τους ηγέτες της Ρωσίας και της Ουκρανίας. Για τον σκοπό αυτό ανακτήθηκαν δύο σύνολα (ένα για κάθε ηγέτη). Για τον Βλαντιμίρ Πούτιν, σαν λέξεις κλειδιά δόθηκαν οι #Putin και Putin. Για τον Βολόντιμιρ Ζελένσκι δόθηκαν οι λέξεις #Zelenski και Zelenski. Ανακτήθηκαν 1094819914 tweets για τον Πούτιν και 9792 για τον Ζελένσκι.

4.4.5.4. Προεπεξεργασία του Συνόλου Δεδομένων των προς αξιολόγηση tweets

Η διαδικασία προεπεξεργασίας που ακολουθείται, είναι ακριβώς η ίδια που ακολουθείται και στην περίπτωση του συνόλου δεδομένων εκπαίδευσης.

4.4.5.5. Μετατροπή σε διανύσματα

Για την παραγωγή των μοντέλων εκτίμησης συναισθήματος, θα πρέπει τα N-grams των περιεχομένων των αναρτήσεων, να μετατραπούν σε διανύσματα. Προκειμένου τα διανύσματα να δημιουργηθούν με κοινές συνθήκες και για τα δύο σύνολα δεδομένων, αυτά συνενώνονται. Αφού ολοκληρωθεί η διαδικασία παραγωγής των αντίστοιχων διανυσμάτων, τα σύνολα δεδομένων διαχωρίζονται εκ νέου.

4.4.5.6. Παραγωγή των Μοντέλων Εκτίμησης Συναισθήματος

Για την παραγωγή των μοντέλων εκτίμησης συναισθήματος χρησιμοποιήθηκαν οι υλοποιήσεις αλγορίθμων μηχανικής μάθησης που παρέχονται από τη βιβλιοθήκη scikit-learn. Κάθε αλγόριθμος έτρεξε:

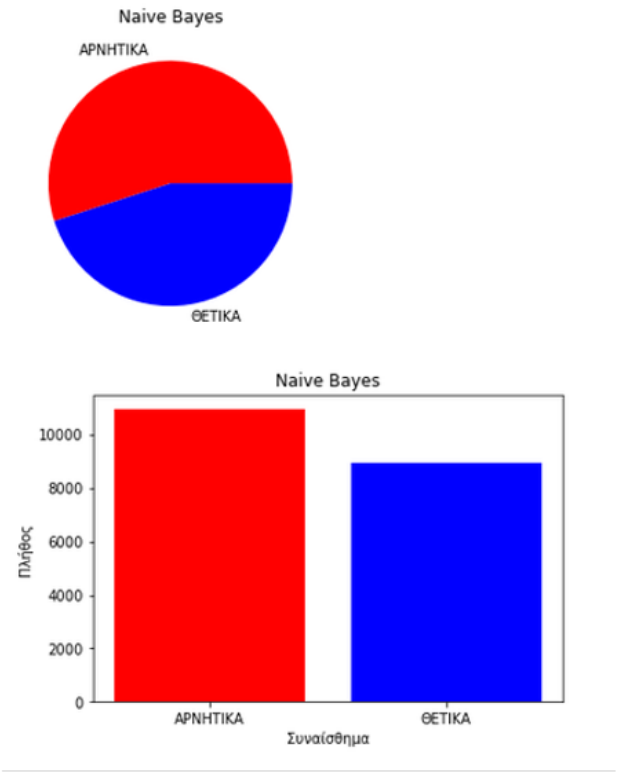
- Για το σύνολο δεδομένων που ανακτήθηκε για κάθε ηγέτη

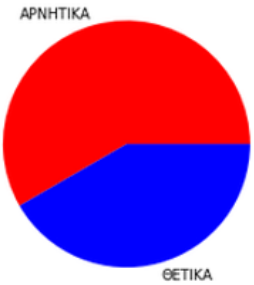
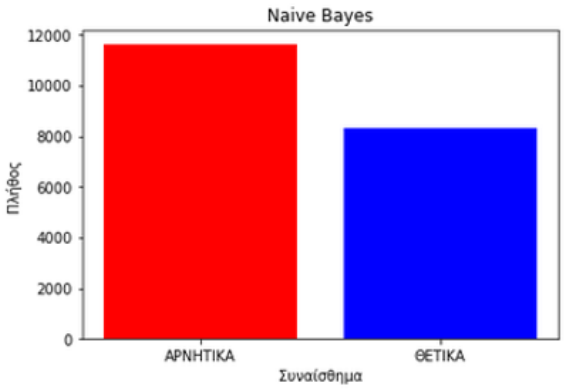
- Για 2-grams και 3-grams

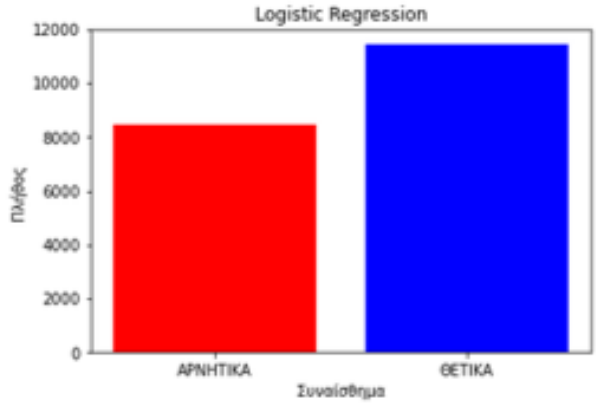
Συνολικά κάθε αλγόριθμος έτρεξε τέσσερις φορές. Παράχθηκαν συνολικά 12 μοντέλα εκτίμησης συναισθήματος. Στην επόμενη παράγραφο παρουσιάζονται τα σχετικά αποτελέσματα.

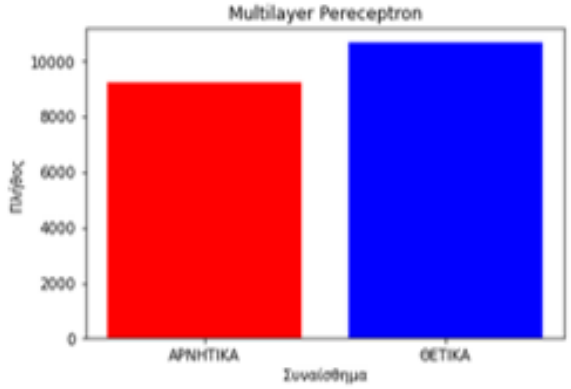
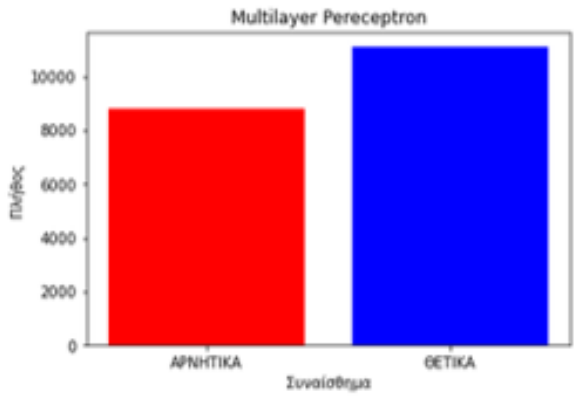
4.5. Αποτελέσματα

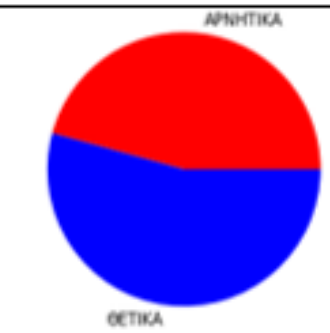
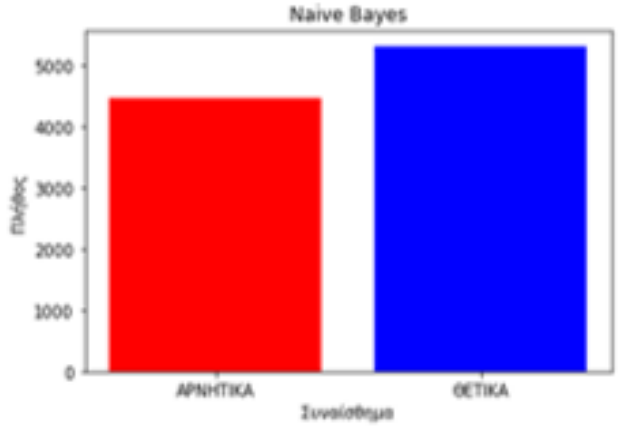
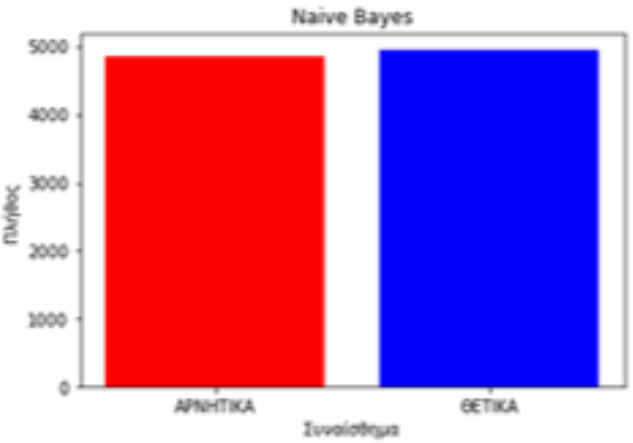
Βλαντιμίρ Πούτιν

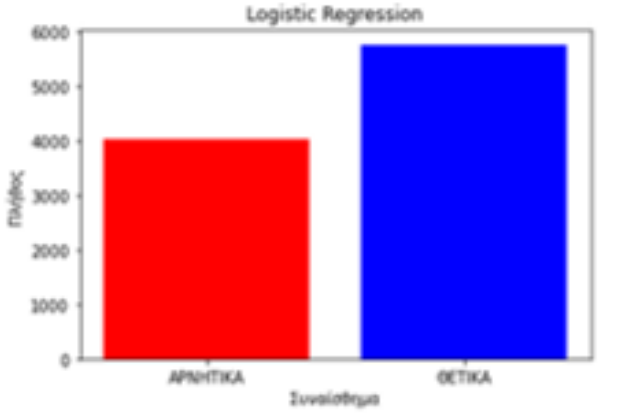
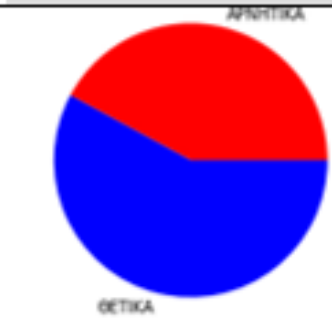
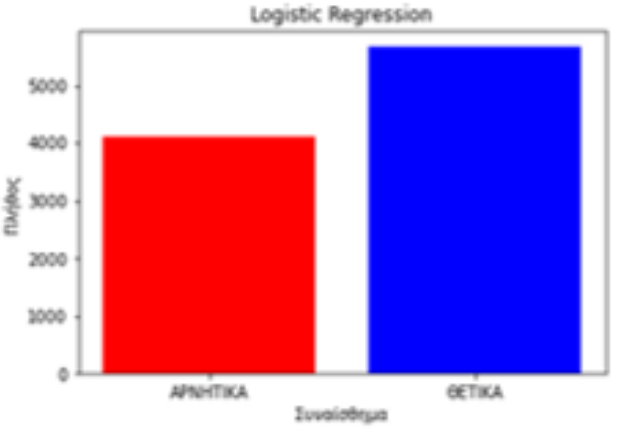
Αλγόριθμος Naïve Bayes 2-GRAMS Πλήθος Αρνητικών: 10948 Πλήθος Θετικών: 8966 Ποσοστό Αρνητικών:55% Ποσοστό Θετικών: 45%	

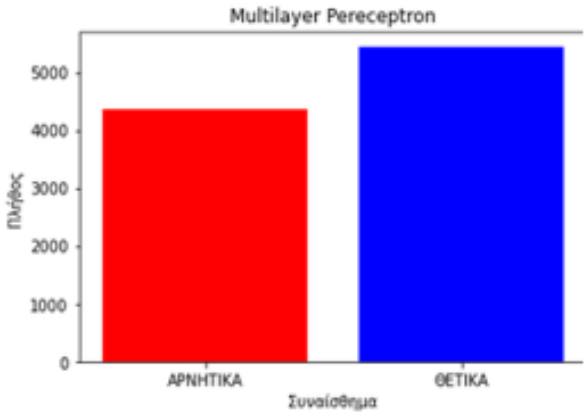
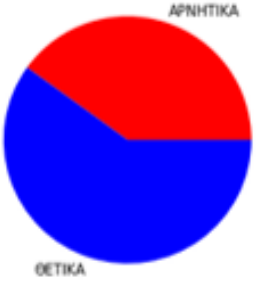
Αλγόριθμος Naïve Bayes 3-GRAMS Πλήθος Αρνητικών: 11617 Πλήθος Θετικών: 8297 Ποσοστό Αρνητικών: 58,3% Ποσοστό Θετικών: 41,7%	 

Αλγόριθμος Logistic Regression 2-GRAMS Πλήθος Αρνητικών: 8966 Πλήθος Θετικών: 10948 Ποσοστό Αρνητικών: 45% Ποσοστό Θετικών: 55%	 
Αλγόριθμος Logistic Regression 3-GRAMS Πλήθος Αρνητικών: 8483 Πλήθος Θετικών: 11431 Ποσοστό Αρνητικών: 42,6% Ποσοστό Θετικών: 57,4%	 

Αλγόριθμος Multilayer Perceptron 2-GRAMS Πλήθος Αρνητικών: 9238 Πλήθος Θετικών: 10676 Ποσοστό Αρνητικών: 46,4% Ποσοστό Θετικών: 53,6%	 
Αλγόριθμος Multilayer Perceptron 3-GRAMS Πλήθος Αρνητικών: 8811 Πλήθος Θετικών: 11103 Ποσοστό Αρνητικών: 44,2% Ποσοστό Θετικών: 55,8%	 

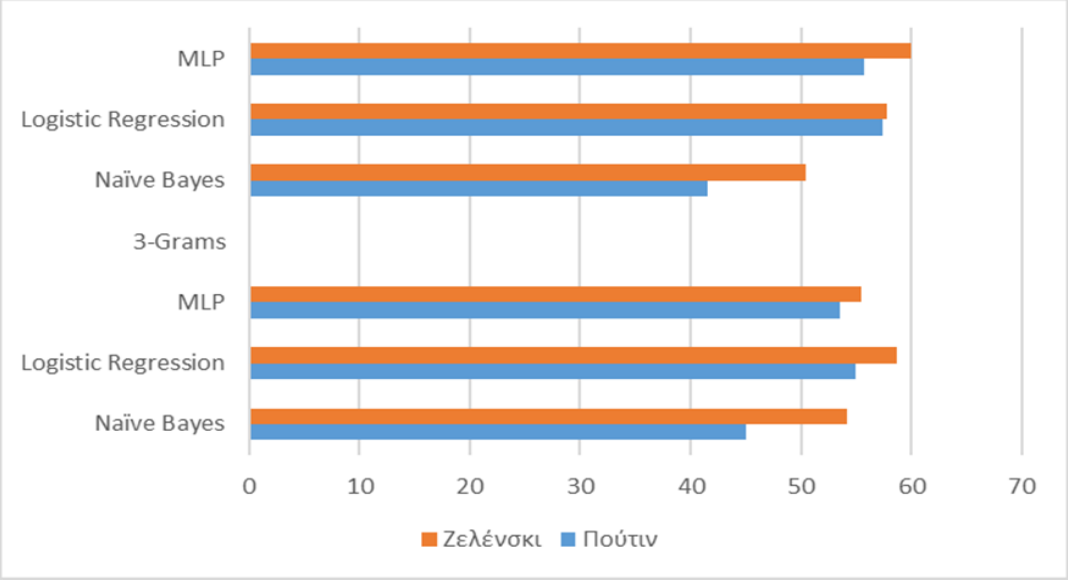
Αλγόριθμος Naïve Bayes 2-GRAMS Πλήθος Αρνητικών: 4479 Πλήθος Θετικών: 5313 Ποσοστό Αρνητικών: 45,7% Ποσοστό Θετικών: 54,3%	 
Αλγόριθμος Naïve Bayes 3-GRAMS Πλήθος Αρνητικών: 4852 Πλήθος Θετικών: 4940 Ποσοστό Αρνητικών: 49,5 Ποσοστό Θετικών: 50,5%	 

Αλγόριθμος Logistic Regression 2-GRAMS Πλήθος Αρνητικών: 4038 Πλήθος Θετικών: 5754 Ποσοστό Αρνητικών: 41,2% Ποσοστό Θετικών: 58,8%	 
Αλγόριθμος Logistic Regression 3-GRAMS Πλήθος Αρνητικών: 4125 Πλήθος Θετικών: 5667 Ποσοστό Αρνητικών: 42,1% Ποσοστό Θετικών: 57,8%	 

Αλγόριθμος Multilayer Perceptron 2-GRAMS Πλήθος Αρνητικών: b'4356' Πλήθος Θετικών: b'5436' Ποσοστό Αρνητικών: 44,5% Ποσοστό Θετικών: 55,5%	 
Αλγόριθμος Multilayer Perceptron 3-GRAMS Πλήθος Αρνητικών: 3920 Πλήθος Θετικών: 5872 Ποσοστό Αρνητικών: 40% Ποσοστό Θετικών: 60%	 

Στον παρακάτω πίνακα φαίνονται τα ποσοστά των θετικών αναρτήσεων για κάθε ηγέτη, για κάθε αλγόριθμο και κάθε προσαρμογή των όρων (N-grams).

Αλγόριθμος	**Πούτιν**	**Ζελένσκι**
2-GRAMS		
Naïve Bayes	45	**54,2**
Logistic Regression	55	**58,7**
MLP	53,6	**55,5**
3-Grams		
Naïve Bayes	41,6	**50,5**
Logistic Regression	57,4	**57,8**
MLP	55,7	**60**



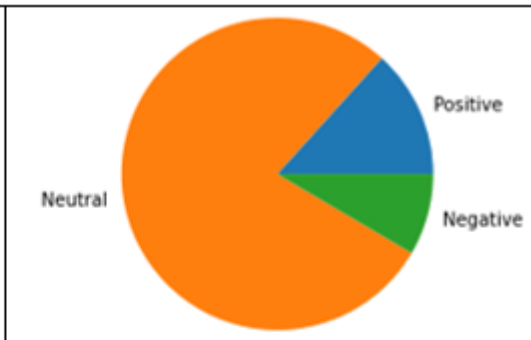
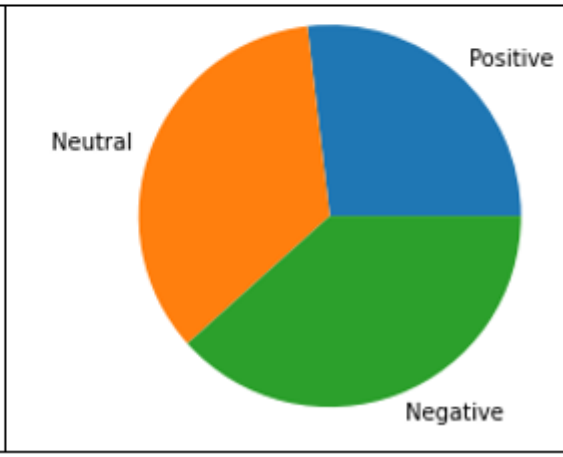
Εικόνα 12 Σύγκριση Θετικών Τάσεων για Πούτιν και Ζελένσκι

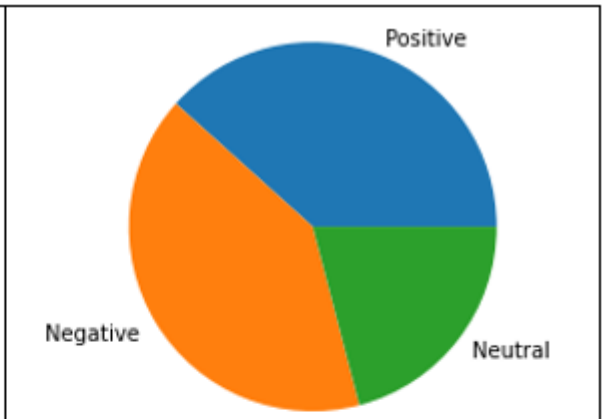
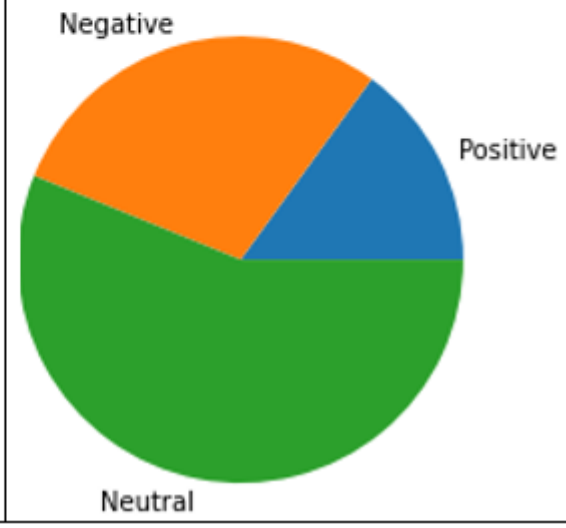
Από τα παραπάνω αποτελέσματα φαίνεται ότι η κοινή γνώμη στο twitter είναι ελαφρώς θετικά διακείμενη προς τους δύο ηγέτες. Ο Ζελένσκι συγκεντρώνει ελαφρώς καλύτερο ποσοστό θετικών συναισθημάτων σε σχέση με τον Πούτιν.

4.5.1. Ανάλυση με τη χρήση Λεξικών

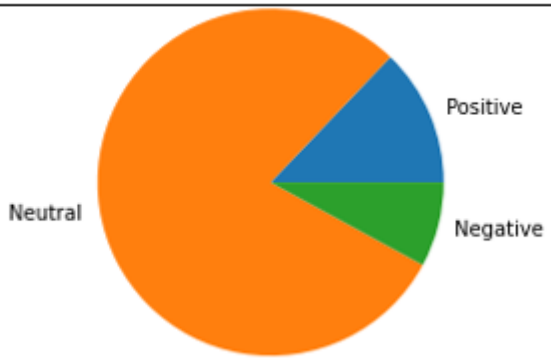
Για την αξιολόγηση της θέσης της κοινής γνώμης στο θέμα του πολέμου στην Ουκρανία, λήφθηκαν και εξετάστηκαν ένα σύνολο από αναρτήσεις στο twitter με την μέθοδο των ευρετηρίων. Εξετάστηκαν τα πιο διαδεδομένα από αυτά. Τα αποτελέσματα που προέκυψαν περιγράφονται ανά ηγέτη και χρησιμοποιούμενο λεξικό παρακάτω.

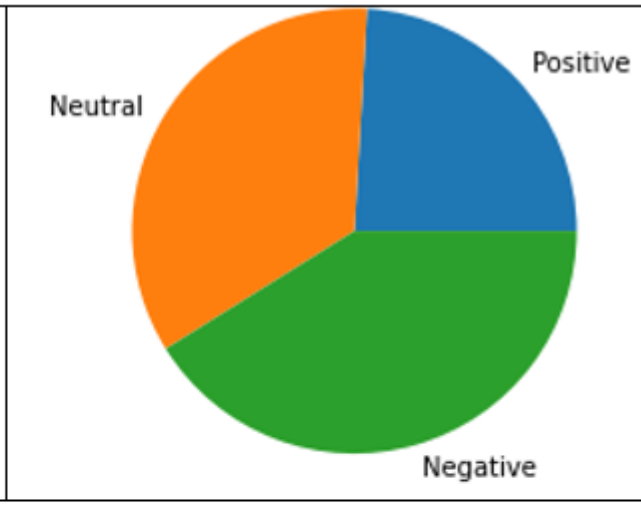
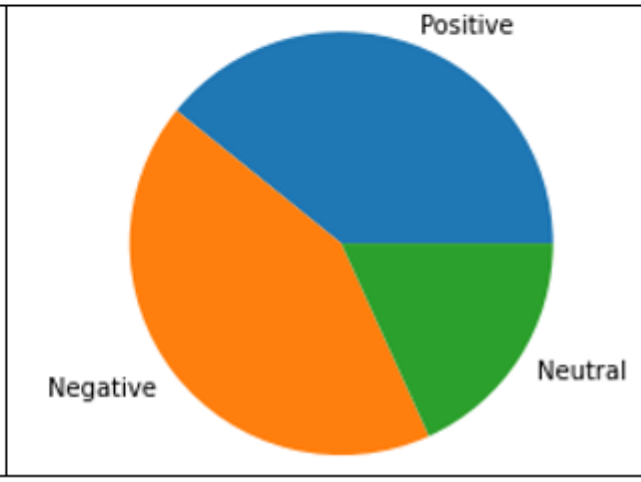
Βλαντιμίρ Πούτιν

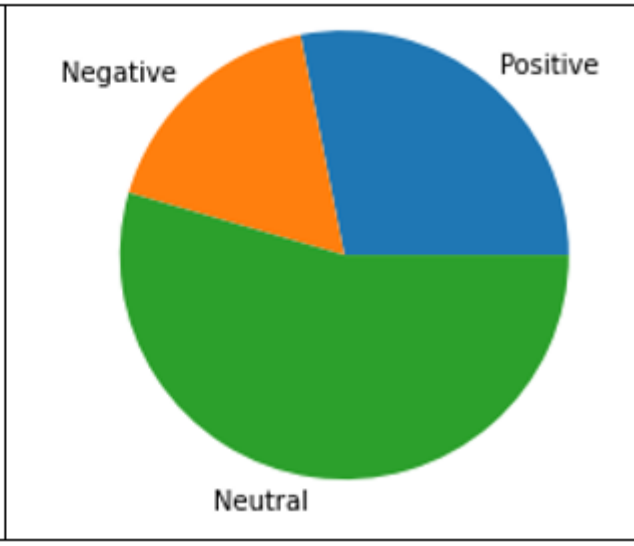
TextBlob	
Πλήθος Μοναδικών Αναρτήσεων:19914 Θετικά:2649 13,30% Αρνητικά:1683 8,51% Ουδέτερα:15582 78,24% Μέση Πολικότητα:-5.17	
Affin	
Πλήθος Μοναδικών Αναρτήσεων:19914 Θετικά:5352 26,87% Αρνητικά:7641 38.37% Ουδέτερα:6921 34,76%	
SectiWordNet	

Πλήθος Μοναδικών Αναρτήσεων:19914 Θετικά:7642 38,37% Ουδέτερα:4164 20,90% Αρνητικά:8108 40, 71%	
SenticNet	
Πλήθος Μοναδικών Αναρτήσεων:19914 Θετικά:2984 15 % Ουδέτερα:11177 56,12% Αρνητικά:57,83 28,88 %	

Βολόντιμιρ Ζελένσκι

TextBlob	
Πλήθος Μοναδικών Αναρτήσεων:9792 Θετικά:1257 12,83% Αρνητικά:774 7,9% Neutral:7761 79,25% Πολικότητα: -8,09	

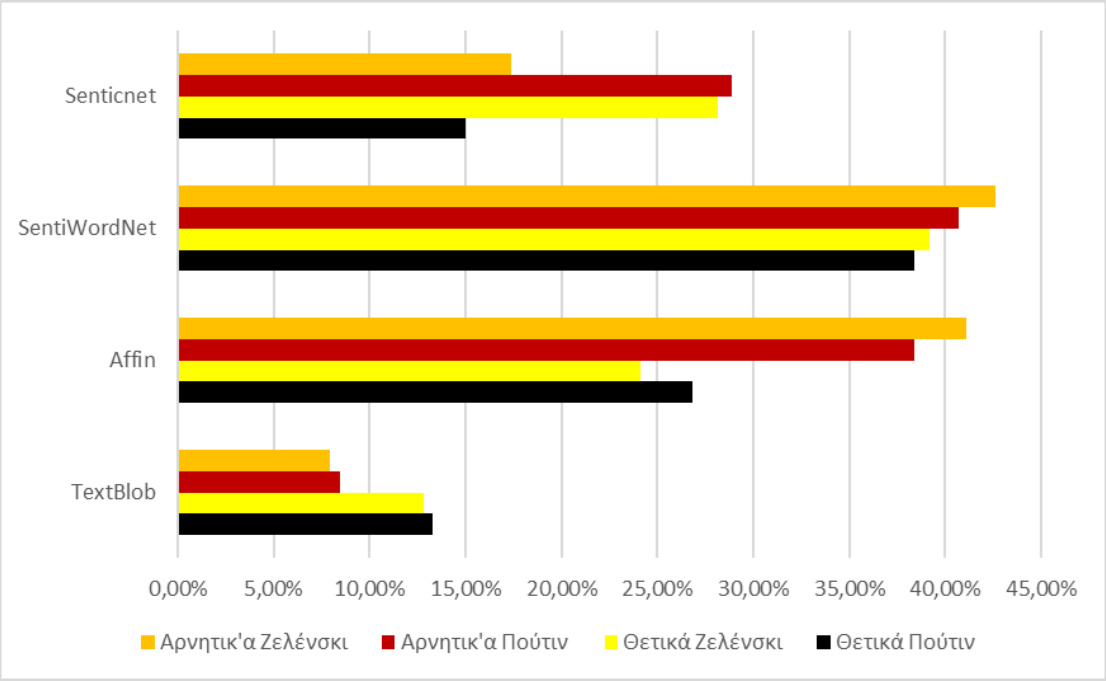
Affin	
Πλήθος Μοναδικών Αναρτήσεων:9792 Θετικά:2359 24.09% Αρνητικά:4027 41.12% Neutral:3406 34,78%	
SentiWordNet	
Πλήθος Μοναδικών Αναρτήσεων:9792 Θετικά:3838 39,19% Ουδέτερα:1782 18,19% Αρνητικά:4172 42,60%	

SenticNet	
Πλήθος Μοναδικών Αναρτήσεων:9792 Θετικά:2755 28,13% Ουδέτερα:5335 54,48% Αρνητικά:1702 17,38%	

Συγκριτικός Πίνακας

Λεξικό	Θετικά			
	Πούτιν	Ζελένσκι	Πούτιν	Ζελένσκι
TextBlob	13,30%	12,83%	8,45%	7,90%
Affin	26,85%	24,09%	38,37%	41,12%
SentiWordNet	38,37%	39,19%	40,71%	42,60%
Senticnet	14,98%	28,13%	28,88%	17,38%

Στο παρακάτω γράφημα αποτυπώνεται συνοπτικά η σύγκριση των συναισθημάτων της κοινής γνώμης για τους δύο ηγέτες όπως εκτιμήθηκαν με τις διαδικασίες χρήσης κάθε ενός από τα λεξικά.



Εικόνα 13 Σύγκριση θετικών και αρνητικων τάσεων για Πούτιν και Ζελένσκι

Από τα αποτελέσματα φαίνεται ότι αν εξαιρεθεί το λεξικό SenticNet, οι υπόλοιπες διαδικασίες φανερώνουν μία ισορροπία μεταξύ των θετικών και αρνητικών απόψεων για τους Πούτιν και Ζελένσκι. Στην περίπτωση του SenticNet φαίνεται ότι οι αρνητικές τάσεις είναι μεγαλύτερες για τον Πούτιν και οι θετικές για τον Ζελένσκι. Το γενικό συμπέρασμα δεν φαίνεται να διαφέρει πολύ από αυτό που προέκυψε με τη χρήση των μεθοδολογιών της μηχανικής μάθησης.

5. Συμπεράσματα

Η ανάπτυξη των τεχνολογιών του διαδικτύου επέφερε ραγδαία αύξηση στο περιεχόμενο που είναι διαθέσιμο στις υποδομές του. Οι παραδοσιακές μεθοδολογίες επεξεργασίας του, αποδεικνύονται αναποτελεσματικές για την επεξεργασία τους. Το πρόβλημα αυτό είναι που τροφοδοτεί την έρευνα για την ανακάλυψη νέων μεθόδων επεξεργασίας για την αρτιότερη εκμετάλλευση των άφθονων διαθέσιμων δεδομένων. Τα ζητούμενα από τις μεθοδολογίες αυτές είναι η ταχύτητα ολοκλήρωσης των διαδικασιών, η ακρίβεια των αποτελεσμάτων τους και η φιλικότητα στη χρήση τους. Η ταχύτητα και η ακρίβεια ενισχύουν την ανταγωνιστικότητα των οργανισμών επ' ωφελεία των οποίων πραγματοποιούνται οι διαδικασίες επεξεργασίας των δεδομένων. Η φιλικότητα της χρήσης τους επιτρέπει στους ανθρώπους που λαμβάνουν αποφάσεις, να συμμετέχουν πιο ενεργά και άμεσα στην επεξεργασία των δεδομένων.

Η υπερδιαθεσιμότητα σε δεδομένα δημιούργησε την ανάγκη για την εύρεση τρόπων επεξεργασίας μεγάλων όγκων δεδομένων. Η εξόρυξη γνώσης περιλαμβάνει μεθοδολογίες και τεχνικές για την ανάδειξη ωφέλιμων πληροφοριών από μεγάλους όγκους δεδομένων κάθε είδους δεδομένων (δομημένα, ημιδομημένα και αδόμητα). Οι τεχνικές της χρησιμοποιούνται και στην επεξεργασία φυσικής γλώσσας με επιτυχία. Καθίσταται δυνατή η αυτοματοποίηση των διαδικασιών επεξεργασίας κειμένων και αρχείων ήχου και η χρήση των αποτελεσμάτων τους στις διαδικασίες λήψης αποφάσεων. Η έρευνα έχει αποδώσει μεγάλο αριθμό εργαλείων που βασίζονται στη μηχανική μάθηση ή άλλες μεθόδους.

Στην παρούσα εργασία επιχειρήθηκε μία επισκόπηση των τρόπων με τους οποίους μπορεί να χρησιμοποιηθούν μεθοδολογίες της μηχανικής μάθησης αλλά και της χρήσης λεξικών, για τον χαρακτηρισμό – κατηγοριοποίηση κειμένων. Διαπιστώθηκε ότι έχουν προταθεί και χρησιμοποιούνται μία ποικιλία τεχνοτροπιών για το σκοπό αυτό. Κάθε μία από αυτές είναι κατάλληλες για συγκεκριμένου είδους εργασίες και αποδίδουν καλύτερα σε αυτές. Όσον αφορά την επεξεργασία φυσικής γλώσσας, μπορούν να χρησιμοποιηθούν πολλών ειδών τεχνικές μηχανικής μάθησης,

δίνοντας τη δυνατότητα στον προγραμματιστή να επιλέξει την καταλληλότερη για κάθε περίσταση. Επιπλέον οι ίδιες διαδικασίες μπορούν να προσαρμοστούν σε διαφορετικές περιστάσεις μέσα από μία πληθώρα επιλογών παραμετροποίησης.

Το αυξημένο ενδιαφέρον για την επεξεργασία φυσικής γλώσσας έχει συμβάλλει στο να υπάρχουν διαθέσιμα μία ποικιλία εργαλείων για τη διεκπεραίωση των διαδικασιών της. Οι προγραμματιστές έχουν τη δυνατότητα να χρησιμοποιούν υψηλού επιπέδου προγραμματιστικές διεπαφές για την ενσωμάτωση δυνατοτήτων επεξεργασίας φυσικής γλώσσας στα έργα που αναπτύσσουν. Επιπλέον διατίθενται και πακέτα λογισμικού που έχουν τη δυνατότητα να επεξεργαστούν μεγάλα δεδομένα σε μορφή φυσικής γλώσσας, με διεπαφές διαχειρίσιμες από ανθρώπους με μέση εξοικείωση με τη χρήση εφαρμογών πληροφορικής. Με τον τρόπο αυτό η επεξεργασία φυσικής γλώσσας γίνεται ολοένα και πιο προσιτή σε διάφορες δραστηριότητες και κυρίως σε εκείνες που σχετίζονται με τη λήψη αποφάσεων.

Η κατηγοριοποίηση των κειμένων εμπίπτει στην επεξεργασία φυσικής γλώσσας. Η δυνατότητα επεξεργασίας του περιεχομένου των εφαρμογών κοινωνικής δικτύωσης, δίνει τη δυνατότητα για εκτίμηση της στάσης των ανθρώπων. Μόλις πριν από λίγα χρόνια, για την εκτίμηση της στάσης της κοινής γνώμης, οι μεθοδολογίες που εφαρμόζονταν, ενέπλεξαν ένα μικρό σχετικά δείγμα ανθρώπων. Η προσπάθεια των ανθρώπων που επιχειρούσαν τέτοιες εργασίες ήταν να δημιουργήσουν δείγματα όσο το δυνατόν πιο αντιπροσωπευτικά. Σήμερα οι ερευνητές διαθέτουν εργαλεία που τους δίνουν τη δυνατότητα να χρησιμοποιήσουν πολύ μεγαλύτερους πληθυσμούς χωρίς να χρειαστεί να καταφύγουν σε παραδοσιακές μεθόδους (π.χ. διανομή ερωτηματολογίων, τηλεφωνικές έρευνες). Οι περισσότεροι άνθρωποι είναι συνδεδεμένοι σε ένα ή περισσότερα κοινωνικά δίκτυα, όπου σε ανύποπτο χρόνο καταγράφουν τις απόψεις, στάσεις και ανησυχίες τους. Τα πιο δημοφιλή κοινωνικά δίκτυα, παρέχουν προγραμματιστικές διεπαφές για τη λήψη του περιεχομένου τους (κάτω από ορισμένες προϋποθέσεις που να καλύπτουν την προστασία των προσωπικών δεδομένων των χρηστών). Οι ερευνητές των τάσεων του κοινού μπορούν να λαμβάνουν στοιχεία μετά την εφαρμογή διαφόρων φίλτρων που σχετίζονται με το αντικείμενο της έρευνας τους. Η διεύρυνση των δειγμάτων σε σχέση

με το πρόσφατο παρελθόν εξασφαλίζει τη βελτίωση της ακρίβειας των συμπερασμάτων που προκύπτουν από τις έρευνες.

Το ζήτημα που προκύπτει από τη διεύρυνση των δειγμάτων έχει να κάνει με τη δυνατότητα εκμετάλλευσης του μεγάλου όγκου δεδομένων. Με τις παραδοσιακές μορφές εκτέλεσης των ερευνών, η επεξεργασία των δεδομένων γινόταν με την καθολική σχεδόν συμμετοχή του ανθρώπινου παράγοντα. Αυτό πλέον είναι αδύνατο να συμβεί. Η λύση δίνεται από την αυτοματοποίηση που παρέχουν αλγόριθμοι μηχανικής μάθησης και οι υλοποιήσεις τους. Οι υλοποιήσεις αυτές έχουν την ικανότητα να επεξεργάζονται μεγάλους όγκους δεδομένων και να παράγουν ευανάγνωστες εξόδους που περιγράφουν τα συμπεράσματα που προκύπτουν. Με τον τρόπο αυτό, οι έρευνες των τάσεων αναβαθμίζονται και ποσοτικά (περισσότερα και πλουσιότερα δείγματα) και ποιοτικά (μεγαλύτερη ακρίβεια αποτελεσμάτων).

Στο πρακτικό μέρος της παρούσας εργασίας εξετάστηκε ένα πρόβλημα εκτίμησης των τάσεων της κοινής γνώμης για ένα επίκαιρο θέμα. Χρησιμοποιήθηκαν μηχανισμοί επεξεργασίας φυσικής γλώσσας δύο κατηγοριών. Η μία κατηγορία αφορούσε τη χρήση λεξικών και η άλλη κατηγορία αφορούσε την εφαρμογή διαδικασιών μηχανικής μάθησης. Τα δεδομένα που λήφθηκαν από το twitter πέρασαν από επεξεργασία με διαφορετικές μεθόδους και από τις δύο κατηγορίες. Τα αποτελέσματα, αν και από μέθοδο σε μέθοδο παρουσίαζαν κάποιες διαφορές, κατέδειξαν τις ίδιες τάσεις και με τις δύο κατηγορίες μεθόδων. Αυτό σημαίνει ότι οι διαδικασίες της μηχανικής μάθησης είναι σχετικά αξιόπιστες για την κατηγοριοποίηση κειμένων. Και για τους δύο τύπους μεθόδων, κρίσιμο στοιχείο αποτελεί το πλήθος των δεδομένων που χρησιμοποιούνται. Όσα περισσότερα δεδομένα χρησιμοποιούνται, τόσο πιο αξιόπιστα αναμένεται να είναι τα αποτελέσματα που παράγονται.

Χρησιμοποιήθηκαν εργαλεία ανάπτυξης βασισμένα στην γλώσσα προγραμματισμού python. Ήταν σχετικά εύκολη η εύρεση και η χρησιμοποίηση τους για την υλοποίηση της κατηγοριοποίησης κειμένων και με τις δύο κατηγορίες τεχνοτροπιών. Η python αποδείχθηκε μία πλήρης γλώσσα για την ανάπτυξη τέτοιου

είδους εργασιών. Ήταν εφικτή η υλοποίηση διεργασιών για όλο το φάσμα της διαδικασίας, από τη χρήση της προγραμματιστικής διεπαφής του twitter για την λήψη των κειμένων, την προεπεξεργασία τους και την εφαρμογή όλων των διαφορετικών μεθόδων επεξεργασίας τους. Επίσης μπόρεσε να υποστηρίξει την παρουσίαση των αποτελεσμάτων και σε μορφή κειμένων και μορφή διαγραμμάτων. Το Jupyter Notebook που χρησιμοποιήθηκε, έδωσε τη δυνατότητα για επαρκή παρουσίαση και των διαδικασιών που υλοποιήθηκαν σε γλώσσα python αλλά και των αποτελεσμάτων τους. Η υποστήριξη για τους προγραμματιστές είναι επίσης πλήρης και για την python και για το Jupyter, είτε αυτή προέρχεται από τους επίσημους φορείς διαχείρισής τους, είτε από την ευρεία διαδικτυακή κοινότητα προγραμματιστών που ασχολείται με αυτά.

6. Αναφορές

Alfred, R. (2005, 1 1). Knowledge Discovery: Enhancing Data Mining and Decision Support Integration. Ανάκτηση στις 19-9-2022 από cloudfront: https://d1wqtxts1xzle7.cloudfront.net/4249279/Knowledge_Discovery_Enhancing_Data_Mining_and_Decision_Support_Integration-with-cover-page-v2.pdf?Expires=1663190445&Signature=YRgLgmRuppNyQRQLuIjN3EZVEJaX2JzUG7Z6PIqq3yjrW75BUyZeSLiwuPNcq3YJ5ppZzxoEz7vgjGme-Q

Bhardwaj, A. (2020, 3 26). Silhouette Coefficient. Ανάκτηση στις 13-10-2022 από towardsdatascience: https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c

Carbonel, Jaime, Ryszald Machalski, και Tom Mitchell. «An Overview of Machine Learning.» *George Mason University.* 1983. http://ebot.gmu.edu/bitstream/handle/1920/1569/83-02.pdf?sequence=1&isAllowed=y (πρόσβαση 17 Ιουλίου 2022).

Chowdhury, Gobinda. «Natural Language Processing.» *syr.* 2003. https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub (πρόσβαση 7 Αυγούστου 2022).

Datareportal. *Digital 2022: Global overview report.* 31 Ιανουάριου 2022. https://datareportal.com/reports/digital-2022-global-overview-report (πρόσβαση 11 Μαϊου 2022).

Ellwardt, Lea, Rafael Wittek, Louise Hawkley, και John Cacioppo. *Social Network Characteristics and Their Associations With Stress in Older Adults: Closure and Balance in a Population-Based Sample* (πρόσβαση 7 Σεπτεμβρίου 2020). https://academic.oup.com/psychsocgerontology/article/75/7/1573/5387566 (πρόσβαση 12 Μαϊου 2022).

Ge, Zhiqiang, Steven Ding, και Biao Huang. «Data Mining and Analytics in the Process Industry: The Role of Machine Learning.» *IEEE.* (πρόσβαση 26 Σεπτεμβρίου 2022). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8051033 (πρόσβαση 15 Ιουλίου 2022).

Gil, Paul. *What Is Twitter & How Does It Work?* 12 Αυγούστου 2021. https://www.lifewire.com/what-exactly-is-twitter-2483331 (πρόσβαση 12 Μαϊου 2022).

Kumar, Arun, Ayodeji Salau, Swati Gupta, και Sandeep Arora. *A Survey of Machine Learning Methods for IoT and their Future Applications.* (πρόσβαση 13 Ιουνίου 2022). https://www.researchgate.net/publication/330702663_A_Survey_of_Machine_Learning_Methods_for_IoT_and_their_Future_Applications (πρόσβαση 1 Ιουνίου 2022).

Loughran, T, και B McDonald. «Sentiment Dictionaries.» 1 1 2011. https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/ (πρόσβαση 13 Αυγούστου, 2022).

Lubbers, Miranda Jessica, Ashton Verdery, και José Luis Molina. *Social Networks and Transnational Social Fields: A Review of Quantitative and Mixed-Methods Approaches.* 17 Δεκεμβρίου 2018. https://journals.sagepub.com/doi/full/10.1177/0197918318812343 (πρόσβαση 12 Μαϊου 2022).

Moore, Andrew, και Michael Littman. *Reinforcement Learning: A Survey.* 1996. https://www.jair.org/index.php/jair/article/download/10166/24110/ (πρόσβαση 1 Ιουνίου 2022).

Musto, Cataldo, Giovanni Semeraro, και Marco Polignano. «A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts.» 1 1 2015. http://ceur-ws.org/Vol-1314/paper-06.pdf (πρόσβαση 12 Ιουλίου 2022).

Nadkarni, Prakash, Lucila Ohno-Machado, και Wendy Chapman. «Natural language processing: an introduction.» *academic.* 2011. https://academic.oup.com/jamia/article/18/5/544/829676?login=false (πρόσβαση 1 Ιουλίου 2022).

Sajithra, K., και Rajindra Patil. *Social Media – History and Components.* 2013. https://d1wqtxts1xzle7.cloudfront.net/32137849/I0716974-with-cover-page-v2.pdf?Expires=1653757808&Signature=BUneEWAg47B3OFiZp52FEkMYcqEL5Ze5ZIg7wijCUjAwTQZphpFopq9-n6eFSTePNXT42a3ObAu49wc6RAlkT3UF66wsldIOyK2VtyWkp~OA~sy2IXW0coxHpHFhz5QrnjRKh8~tjSf7NlY95k4 (πρόσβαση 12 Ιουνίου 2022).

Tapsai, Chalermpol, Phayung Meesad, και Choochart Haruechaiyasak. «TLS-ART: Thai Language Segmentation by Automatic Ranking Trie.» *researchgate.* 2016. https://www.researchgate.net/publication/311705165_TLS-ART_Thai_Language_Segmentation_by_Automatic_Ranking_Trie?enrichId=rgreq-264e0688c324041b86336e8412cfde72-XXX&enrichSource=Y292ZXJQYWdlOzMxMTcwNTE2NTtBUzo0NDAzMjA0Njc5NjgwMDBAMTQ4MTk5MTk4ODQzMg%3D%3D&e (πρόσβαση 2 Ιουλίου 2022).

TIBCO. *What is Data Virtualization?* 2022. https://www.tibco.com/reference-center/what-is-data-virtualization (πρόσβαση 11 Ιουνίου 2022).

Widman, Jake. *What is Reddit?* Ιουλίου 2021. https://www.digitaltrends.com/web/what-is-reddit/ (πρόσβαση 21 Ιουλίου 2022).

Alzubi, J. (2018). *Machine Learning from Theory to Algorithms: An Overview.* Ανάκτηση September 2, 2022, από https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/pdf

Anand, A. (2014). *Generative vs Discriminative Models in Machine Learning.* Ανάκτηση 2022, από https://www.analyticssteps.com/blogs/generative-vs-discriminative-models-machine-learning

Breslin, J. G., Passant, A., & Decker, S. (2009, January 1). *Introduction to the Social Web (Web 2.0, social media, social software).* Ανάκτηση September 9, 2022, από https://link.springer.com/chapter/10.1007/978-3-642-01172-6_3

Brownlee, J. (2016). *Overfitting and Underfitting With Machine Learning Algorithms.* Ανάκτηση 2022, από https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Coope, K. D., & Torczon, L. (2012). *Backus-Naur Form.* Ανάκτηση 2022, από https://www.sciencedirect.com/topics/computer-science/backus-naur-form

Gandhi, R. (2018). *Support Vector Machine — Introduction to Machine Learning Algorithms.* Ανάκτηση 2022, από https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Jurafsky, D., & Martin, J. (2023). *Hidden Markov Models.* Ανάκτηση 2023, από https://web.stanford.edu/~jurafsky/slp3/A.pdf

Lexalytics. (2022). *Machine Learning (ML) for Natural Language Processing (NLP).* Ανάκτηση 2022, από https://www.lexalytics.com/blog/machine-learning-natural-language-processing/

Lutkevich, B. (2020). *Natural language processing (NLP).* Ανάκτηση 2022, από https://www.techtarget.com/searchenterpriseai/definition/natural-language-processing-NLP

Mayo, M. (2018). *The Main Approaches to Natural Language Processing Tasks.* Ανάκτηση 2022, από https://www.kdnuggets.com/2018/10/main-approaches-natural-language-processing-tasks.html

Minaee, S. (2019, Octomber 28). *20 Popular Machine Learning Metrics. Part 1: Classification & Regression Evaluation Metrics.* Ανάκτηση September 11, 2022, από https://towardsdatascience.com/20-popular-machine-learning-metrics-part-1-classification-regression-evaluation-metrics-1ca3e282a2ce

Trujillo, G., Kim, C., Jones, S., Garcia, R., & Murray, J. (2015, August 20). *Understanding the Big Data World.* Ανάκτηση September 9, 2022, από https://www.pearsonitcertification.com/articles/article.aspx?p=2427073&seqNum=2

Turing. (2022). *Introduction to Statistics for Machine Learning.* Ανάκτηση 2022, από https://www.turing.com/kb/introduction-to-statistics-for-machine-learning

Wallach, H. (2004). *Conditional Random Fields: An Introduction.* Ανάκτηση 2022, από https://repository.upenn.edu/cgi/viewcontent.cgi?article=1011&context=cis_reports