ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΕΞΟΡΥΞΗ ΓΝΩΜΗΣ ΑΠΟ ΣΧΟΛΙΑ ΧΡΗΣΤΩΝ ΣΤΟ TWITTER ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ

ΚΟΥΓΚΟΥΛΑ ΜΑΓΔΑΛΗΝΗ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

Πλαγιανάκος Βασίλειος

Καθηγητής

Λαμία 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΙΑΤΡΙΚΗ

ΕΞΟΡΥΞΗ ΓΝΩΜΗΣ ΑΠΟ ΣΧΟΛΙΑ ΧΡΗΣΤΩΝ ΣΤΟ TWITTER ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ

ΚΟΥΓΚΟΥΛΑ ΜΑΓΔΑΛΗΝΗ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Επιβλέπων
Πλαγιανάκος Βασίλειος
Καθηγητής

ΛΑΜΙΑ 2020

Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις [1], που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. *Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί* ***χωρίς να τα περικλείω σε εισαγωγικά*** *και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάσθηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.*

2. *Δέχομαι ότι η αυτολεξεί* ***παράθεση χωρίς εισαγωγικά****, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.*

3. *Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια*

4. *Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.*

Ημερομηνία: 9/07/2020

Ο – Η Δηλ.

(Υπογραφή)

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων σκόπευε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκόπευε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ετών.

ΕΞΟΡΥΞΗ ΓΝΩΜΗΣ ΑΠΟ ΣΧΟΛΙΑ ΧΡΗΣΤΩΝ ΣΤΟ TWITTER ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ

ΚΟΥΓΚΟΥΛΑ ΜΑΓΔΑΛΗΝΗ

Τριμελής Επιτροπή:

Πλαγιανάκος Βασίλειος, Καθηγητής (επιβλέπων)

Τασουλής Σωτήριος, Επίκουρος Καθηγητής

Καρανίκας Χαράλαμπος, Λέκτορας

ΠΕΡΙΛΗΨΗ

Η ανάλυση αισθήσεων είναι ένα προοδευτικό πεδίο επεξεργασίας φυσικής γλώσσας. Είναι ένας τρόπος να ανιχνεύσουμε τη στάση, την κατάσταση του νου ή τα συναισθήματα του ατόμου προς ένα προϊόν, μια υπηρεσία, μια ταινία κλπ., αναλύοντας τις απόψεις και τις κριτικές που μοιράζονται στα κοινωνικά μέσα, τα blogs και ούτω καθεξής. Διάφορες πλατφόρμες κοινωνικών μέσων όπως Facebook, Twitter και άλλες, επιτρέπουν στους ανθρώπους να μοιραστούν τις απόψεις τους με άλλους ανθρώπους.[2] Το Twitter έχει γίνει πλέον μία από τις πιο δημοφιλής πλατφόρμες κοινωνικών μέσων που επιτρέπει στους χρήστες να μοιράζονται πληροφορίες, μέσω των σύντομων μηνυμάτων που ονομάζονται tweets σε πραγματικό χρόνο. Χιλιάδες άνθρωποι αλληλεπιδρούν μεταξύ τους ταυτόχρονα και παράγεται ένας τεράστιος όγκος δεδομένων σε δευτερόλεπτα. Για να χρησιμοποιήσουμε καλά αυτά τα δεδομένα, στο πλαίσιο της παρούσας διπλωματικής εργασίας, αναπτύσσεται ένας αλγόριθμος απεικόνισης σε πραγματικό χρόνο στο Twitter. Πρόκειται για έναν αλγόριθμο που σκοπός του είναι να χρησιμοποιήσει μια προσέγγιση κώδικα R για την ανάλυση συναισθημάτων και την απεικόνισή της, χρησιμοποιώντας ένα σύνολο πακέτων που υποστηρίζονται από τη γλώσσα R, για να απομακρύνουν τα δεδομένα σε πραγματικό χρόνο από το Twitter μέσω APIs με χρήση hashtags και λέξεων-κλειδιών . Αυτός ο αλγόριθμος μπορεί να αναλύσει τα αισθήματα ως θετικά και αρνητικά για ένα συγκεκριμένο προϊόν ή μία υπηρεσία που βοηθά οργανισμούς, πολιτικά κόμματα και κοινούς ανθρώπους, να κατανοήσουν την αποτελεσματικότητα των προσπαθειών τους και τη λήψη αποφάσεων. Τα αποτελέσματα δείχνουν ότι ο αλγόριθμος μπορεί να επεξεργάζεται δεδομένα σε πραγματικό χρόνο και να αποκτά συνεχώς οπτικοποίηση των πληροφοριών.

Λέξεις-κλειδιά: Ανάλυση συναισθημάτων, οπτικοποίηση, προσέγγιση σε πραγματικό χρόνο, Twitter

ABSTRACT

Sensory analysis is a progressive field of natural language processing. It's a way to trace a person's attitude, state of mind, or feelings toward a product, service, movie, etc., by analyzing the opinions and criticisms they share on social media, blogs, and so on. Various social media platforms such as Facebook, Twitter and others allow people to share their views with other people.[2]Twitter has now become one of the most popular social media platforms that allows users to share information through short messages called real-time tweets. Thousands of people interact with each other at the same time, producing a huge amount of data in seconds. To make good use of this data, as part of this thesis, a real-time imaging algorithm is developed on Twitter. It is an algorithm designed to use an R-code approach to emotion analysis and visualization, using a set of R-supported packages to remove real-time data from Twitter via hashtags APIs and keywords. This algorithm can analyze emotions as positive and negative for a particular product or service that helps organizations, political parties and common people understand the effectiveness of their efforts and decision making. The results show that the algorithm can process data in real time and obtain continuous visualization of information.

Keywords: Emotion Analysis, Visualization, Real-Time Approach, Twitter

Περιεχόμενα

ΕΙΣΑΓΩΓΗ 10
1.1 ΤΑ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΣΗΜΕΡΑ 12
1.2 ΕΠΙΠΤΩΣΕΙΣ ΤΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΣΤΗ ΖΩΗ ΜΑΣ 13
1.3 ΔΙΑΔΙΚΤΥΟ ΚΑΙ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΣΤΗΝ ΕΛΛΑΔΑ 15
1.4 ΧΡΗΣΤΕΣ ΤΟΥ TWITTER ΣΤΗΝ ΕΛΛΑΔΑ 16
2.1ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ Ή ΕΞΟΡΥΞΗ ΓΝΩΜΗΣ 17
2.2 ΙΣΤΟΡΙΚΟ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ 18
2.3 ΤΙ ΕΙΝΑΙ Η ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ 20
2.4 ΠΩΣ ΔΟΥΛΕΥΕΙ Η ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ 20
2.5 ΣΗΜΑΣΙΑ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΤΟΝ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟ ΤΟΜΕΑ 22
2.6 ΤΥΠΟΙ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ 23
2.7 ΑΛΓΟΡΙΘΜΟΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΒΑΣΙΣΜΕΝΟΣ ΣΕ ΚΑΝΟΝΕΣ 24
2.8 ΧΡΗΣΗ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΗΜΕΡΑ 24
3.1 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ 26
3.2 ΠΩΣ ΔΟΥΛΕΥΕΙ Η ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ 26
3.3 ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΠΩΣ ΧΡΗΣΙΜΟΠΟΙΟΥΝΤΑΙ 28
3.4 ΠΩΣ ΕΠΙΛΕΓΟΥΜΕ ΕΝΑΝ ΑΛΓΟΡΙΘΜΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ 31
3.5 ΠΡΟΚΛΗΣΕΙΣ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ 32
3.6 ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ [9] 32
3.7 ΓΙΑΤΙ ΕΙΝΑΙ Η ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΗΜΑΝΤΙΚΗ 34
4.1 ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ 35
4.2 ΤΙ ΕΙΝΑΙ Η ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ 35
4.3 ΔΙΑΦΟΡΑ ΑΝΑΜΕΣΑ ΣΤΗΝ ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ, ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΟΥ ΚΑΙ ΤΑ ΑΝΑΛΥΤΙΚΑ ΚΕΙΜΕΝΑ(TEXT ANALYTICS) 36
4.4 ΓΙΑΤΙ ΕΙΝΑΙ ΣΗΜΑΝΤΙΚΗ Η ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ 39
4.5 ΠΩΣ ΛΕΙΤΟΥΡΓΕΙ Η ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ 40
5.1 ΜΕΘΟΔΟΛΟΓΙΑ ΑΛΓΟΡΙΘΜΟΥ-ΕΙΣΑΓΩΓΗ 43
5.2 ΕΞΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ TWITTER 45
5.3 ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ ΚΑΙ ΚΑΤΑΣΚΕΥΗ WORDCLOUD 48
5.3.1 ΚΑΘΑΡΙΣΜΟΣ ΚΕΙΜΕΝΟΥ ΑΠΟ ΠΕΡΙΤΤΑ ΣΤΟΙΧΕΙΑ 49
5.3.2 ΔΗΜΙΟΥΡΓΙΑ ΠΙΝΑΚΑ ΟΡΩΝ-ΕΓΓΡΑΦΟΥ(TERM-DOCUMENT MATRIX) 51
5.3.3 ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ WORDCLOUD 52
5.4 ΕΞΟΡΥΞΗ ΚΑΙ ΟΠΤΙΚΟΠΟΙΗΣΗ ΣΥΝΑΙΣΘΗΜΑΤΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ ΛΕΞΙΚΟΥ NRC 54

6.1 ΕΞΟΡΥΞΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΑΠΟ ΣΧΟΛΙΑ ΧΡΗΣΤΩΝ ΣΤΟ TWITTER ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ 58
ΒΙΒΛΙΟΓΡΑΦΙΑ-ΑΝΑΦΟΡΕΣ.....73
ΠΗΓΕΣ ΕΙΚΟΝΩΝ.....74

ΕΙΣΑΓΩΓΗ

Τα συναισθήματα περιγράφονται ως έντονα αισθήματα που στρέφονται σε κάτι ή κάποιον ως απάντηση σε κάποιο εσωτερικό ή εξωτερικό γεγονός που έχει ιδιαίτερη σημασία για το άτομο. Και το διαδίκτυο, σήμερα, έχει γίνει ένα βασικό μέσο, μέσω του οποίου οι άνθρωποι εκφράζουν τα συναισθήματα και τις απόψεις τους. Κάθε γεγονός, νέο ή δραστηριότητα σε όλο τον κόσμο, μοιράζεται, συζητείται, δημοσιεύεται και σχολιάζεται στα κοινωνικά μέσα ενημέρωσης, από εκατομμύρια ανθρώπους. Π.χ. οι χημικές επιθέσεις στη Συρία ή εκλογική καμπάνια κάποιου υποψήφιου. Καταγραφή αυτών των συναισθημάτων στο κείμενο, ειδικά αυτά που δημοσιεύονται ή κυκλοφορούν στα κοινωνικά μέσα ενημέρωσης, μπορούν να αποτελέσουν πηγή πολύτιμων πληροφοριών, οι οποίες μπορούν να χρησιμοποιηθούν για να μελετήσουμε πόσο διαφορετικά αντιδρούν άνθρωποι σε διαφορετικές καταστάσεις και γεγονότα. Οι επιχειρηματικοί αναλυτές μπορούν να χρησιμοποιήσουν αυτές τις πληροφορίες για να εντοπίσουν τα συναισθήματα και τις απόψεις των ανθρώπων σχετικά με τα προϊόντα τους. Το πρόβλημα με το μεγαλύτερο μέρος της ανάλυσης συναισθήματος που γίνεται σήμερα είναι ότι η ανάλυση ενημερώνει ,μόνο εάν το κοινό έχει αντίδραση θετική ή αρνητική , αλλά δεν περιγράφει τα ακριβές συναισθήματα των πελατών και την ένταση της αντίδρασής τους. Με τη συναισθηματική ανάλυση για παράδειγμα, οι αναλυτές των επιχειρήσεων μπορούν να αναλύσουν την ολιστική άποψη των ανθρώπων ως απάντηση στις ενέργειές τους ή τα γεγονότα και να εργαστούν αναλόγως. Επίσης, οι αναλυτές της υγείας μπορούν να μελετήσουν τις μεταβολές της διάθεσης ατόμων ή ατόμων σε μάζες σε διαφορετικές ώρες της ημέρας. Μπορεί επίσης να χρησιμοποιηθεί για να διατυπωθεί η ψυχική ή συναισθηματική κατάσταση ενός ατόμου, μελέτη της δραστηριότητάς του για μια περίοδο του χρόνου, και ενδεχομένως να ανιχνευθεί ο κίνδυνος κατάθλιψης.

Η Ανάλυση Συναισθήματος είναι ένας τομέας έρευνας που αναπτύσσεται στο πλαίσιο της Επεξεργασίας Φυσικής Γλώσσας (NLP), που αναπτύσσει συστήματα που προσπαθούν να εντοπίσουν και να εξαγάγουν τα συναισθήματα μέσα από το κείμενο. Η ανάλυση του αισθήματος επιτρέπει στους οργανισμούς, τα πολιτικά κόμματα και τους κοινούς ανθρώπους να εντοπίζουν τα αισθήματα για οποιοδήποτε συγκεκριμένο προϊόν ή υπηρεσίες για καλύτερη λήψη αποφάσεων για τη βελτίωση του προϊόντος ή της υπηρεσίας τους. Ο στόχος της ανάλυσης συναισθήματος είναι να εντοπίσει τα συναισθήματα, τη στάση και την κατάσταση του πνεύματος των ανθρώπων σε ένα προϊόν ή μια υπηρεσία και να τα ταξινομήσει ως θετικά, αρνητικά και ουδέτερα από τον τεράστιο όγκο δεδομένων με τη μορφή αναθεωρήσεων, tweets και σχολίων.[3] Με την επέκταση του διαδικτύου παράγεται ένας τεράστιος όγκος οργανωμένων και ανοργάνωτων δεδομένων μέσω διαφόρων πλατφόρμων κοινωνικών μέσων που επιτρέπουν στους ανθρώπους να μοιράζονται τη γνώμη, το σχόλιο, την ανατροφοδότηση, την κριτική, τις σκέψεις και τις εμπειρίες τους με τον κόσμο. Μεταξύ όλων των πλατφορμών κοινωνικών μέσων όπως το Twitter, το Facebook, το YouTube, το Instagram , το Twitter γίνεται πλέον δημοφιλή κοινωνικά μέσο ενημέρωσης αυτές τις μέρες, ώστε να μοιράζονται δημόσια οι σκέψεις, τα συναισθήματα, οι απόψεις με τον κόσμο. Παρέχει τη δυνατότητα που

επιτρέπει στους χρήστες να μοιράζονται πληροφορίες μέσω των σύντομων μηνυμάτων που ονομάζονται tweets (γενικά λιγότερο από 140 χαρακτήρες, περίπου 11 λέξεις κατά μέσο όρο) σε βάση πραγματικού χρόνου. Η ανάλυση του συναισθήματος μας βοηθά να μετασχηματίζουμε αυτόματα σε μεγάλο χρονικό διάστημα ένα οργανωμένο μεγάλο όγκο δεδομένων σε οργανωμένη μορφή. Αυτά τα δεδομένα μπορούν να βοηθήσουν την κυβέρνηση, τους πολιτικούς, τους οργανισμούς, τους ερευνητές και διάφορες εμπορικές εφαρμογές.[10]

Στην παρούσα διπλωματική η οποία βασίστηκε στην εργασία των Σωτήριο Τασουλή, Αριστείδη Βραχάτη, Σπύρο Γεωργακόπουλο και Βασίλειο Πλαγιανάκο σχετικά με την ανάλυση συναισθήματος σε πραγματικό χρόνο από δεδομένα του twitter [7], προτείνεται ένας διαφορετικός αλγόριθμος απεικόνισης επίσης σε πραγματικό χρόνο για την ανάλυση του ίδιου όγκου δεδομένων. Χρησιμοποιείται μια προσέγγιση κώδικα για ανάλυση συναισθημάτων χρησιμοποιώντας ένα σύνολο πακέτων που υποστηρίζονται από τη γλώσσα R για να εξορύξουμε τα δεδομένα Twitter και να διεξάγουμε την ανάλυση των συναισθημάτων για τα tweets σε οποιοδήποτε δημοφιλές θέμα χρησιμοποιώντας tweets και λέξεις-κλειδιά και αναλύοντας τα αισθήματα των ανθρώπων σε αυτή τη λέξη-κλειδί ή tweet. Το σύστημα θα πραγματοποιήσει ανάλυση όπως η συλλογή δεδομένων, η προεπεξεργασία δεδομένων, η εξαγωγή χαρακτηριστικών, η αναγνώριση του συναισθήματος, η ταξινόμηση των αισθήσεων και η παρουσίαση των αποτελεσμάτων.

Το Twitter χρησιμοποιείται ως πηγή δεδομένων για τη συλλογή δεδομένων σε πραγματικό χρόνο πάνω σε δημοφιλή θέματα σε όλο τον κόσμο, μέσω των διεπαφών προγραμματισμού εφαρμογών twitter (API) με χρήση tweets ή λέξεων-κλειδιών. Τα δεδομένα που εξάγονται από το twitter αποθηκεύονται σε μορφή αρχείου CSV. Στη συνέχεια διεξάγουμε την ανάλυση του συναισθήματος για να πάρουμε το συναίσθημα των ανθρώπων , εκτελείται ο καθαρισμός των δεδομένων, ακολουθούμενος από θετικές, αρνητικές ή ουδέτερες κατανομές συναισθημάτων των tweets με το αντίστοιχο score. Ο αλγόριθμος χρησιμοποιεί μια προσέγγιση βασισμένη στην ταξινόμηση των συναισθημάτων και το score. Ο αλγόριθμος μας παρουσιάζει τα εξαγόμενα δεδομένα με τη μορφή διαγράμματος με την αντίστοιχη πολικότητα συναισθημάτων και βαθμολογία αισθήσεων . Τα tweets εξάγονται συνεχώς σε πραγματικό χρόνο και η παραγωγή τους ενημερώνεται συνεχώς.

1.1 ΤΑ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΣΗΜΕΡΑ

Τα μέσα κοινωνικής δικτύωσης έχουν εισβάλει για τα καλά στις ζωές όλων μας καθώς μέσα από αυτά οι άνθρωποι επικοινωνούν, ανταλλάσουν απόψεις, διακεδάζουν, διαφημίζουν εταιρίες, προιόντα και πολλές φορές λειτουργούν και σαν πλατφόρμες ηλεκτρονικών καταστημάτων. Πρόκειται για έναν τομέα που έχει συμβάλει σημαντικά στη καθοριστική αλλαγή των σημερινών κοινωνικών αλλαγών που έχουν σημειωθεί τα τελευταία χρόνια.Αυτό οφείλεται στην ραγδαία εξέλιξη της τεχνολογίας σε συνάρτηση με τον ολοένα και αυξανόμενο αριθμό των οραματιστών της νέας γενιάς μας, οι οποίοι με την εφευρετικότητα που τους διακατέχει έχουν δημιουργήσει τεράστιες παγκόσμιες και πετυχημένες πλατφόρμες κοινωνικής δικτύωσης.Μεγάλες εταιρίες κοινωνικής δικτύωσης όπως το Facebook, το Twitter κτλ έχουν συμβάλει στη δημιουργία ενός ολόκληρου νέου κόσμου όπου είμαστε ελεύθεροι να εκφράσουμε την άποψή μας και να την μοιραστούμε με τους φίλους και τους συνανθρώπους μας. Αυτός ο κόσμος των κοινωνικών μέσων ενημέρωσης, δίνει την δυνατότητα στον κάθε ένα να εκφράσει και να μοιραστεί τις ιδέες, τις σκέψεις και τα συναισθήματα, που θέλει εύκολα και γρήγορα.

Εικόνα 1. «Ενδεικτική απεικόνιση λογότυπων από διάφορα μέσα κοινωνικής δικύωσης»

1.2 ΕΠΙΠΤΩΣΕΙΣ ΤΩΝ ΜΕΣΩΝ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΣΤΗ ΖΩΗ ΜΑΣ

Αρχικά θα μπορούσε κανείς να υποστηρίξει οτι τα μέσα κοινωνικής δικτύωσης έχουν γίνει πλέον μια σημαντική πηγή ειδήσεων. Παρ'όλα αυτά η αξιοπιστία των περισσότερων πηγών μπορεί σαφώς να αμφισβητηθεί.Αξίζει να σημειώθει, ότι λόγω της χαμηλής αξιοπιστίας των ειδήσεων σύχνα το κοινό οδηγείται σε παρανοήσεις. Σε έρευνα που εγινε πρόσφατα για τα media, διαπιστώθηκε ότι σε μια εποχή αυξανόμενης αβεβαιότητας, πολλοί άνθρωποι σε όλο τον κόσμο είναι περισσότερο επιφυλακτικοί απέναντι σε αυτά που βλέπουν και διαβάζουν online, ιδιαίτερα στα κοινωνικά μέσα.Περισσότεροι από τους μισούς Έλληνες χρήστες του Διαδικτύου (56%) πιστεύουν ότι οι περισσότερες από τις ειδήσεις που βλέπουν στο διαδίκτυο είναι ψεύτικες. Παρομοίως και το 46% όλων των χρηστών παγκοσμίως συμφωνούν.Το 32% των χρηστών στην Ελλάδα (47% παγκοσμίως) λένε ότι επηρεάζονται από τις απόψεις που διαβάζουν on-line, σε σύγκριση με 38% (54% παγκοσμίως) το 2017.Το 33% των χρηστών στην Ελλάδα (47% παγκοσμίως), λένε ότι έχουν λιγότερη εμπιστοσύνη στους ειδικούς από ό,τι παλαιότερα.Επίσης το 23% των χρηστών στην Ελλάδα εμπιστεύεται τις απόψεις των bloggers/vloggers σχετικά με προϊόντα και υπηρεσίες (42% παγκοσμίως).Στον αντίποδα βέβαια η διαθεσιμότητά τους στα κοινωνικά δίκτυα καθιστά τα νέα και τις ειδήσεις πιο προσβάσιμα στο ευρύ κοινό οποιαδήποτε ώρα της ήμέρας και απο οποιδήποτε μέρος. Επιπρόσθετα τα νέα και οι ειδήσεις ταξιδεύουν με μεγάλη ταχύτητα και κάνουν το γύρο του κόσμου μέσα σε λίγα δευτερόλεπτα.[11]

Τα κοινωνικά μέσα μαζικής ενημέρωσης προωθούνται μέσω της αλληλεπίδρασης των χρηστών τους σε μια τέτοια μαζική κλίμακα που είναι δύσκολο να μην κατακλύσουν τον κόσμο της ενημέρωσης και κυρίως της πληροφορίας. Δίνουν την δυνατότητα στους ανθρώπους να έρχονται σε επαφή πιο συχνά, και πολλές φορές, με πιο στενό τρόπο από ότι συνήθιζαν παλιότερα λόγω του χρόνου και του χώρου. Οι άνθρωποι που κατοικοεδρεύουν σε διαφορετικές και μακρινές μεταξύ τους πόλεις ή και χωριά μπορούν να έρθουν σε επαφή πολύ εύκολα, και συγχρόνως τους δίνεται η δυνατότητα να γνωρίσουν ανθρώπους που ανήκουν σε διαφορετικές χώρες, ηπείρους και κουλτούρες.

Ακόμη στα κοινωνικά μέσα μαζικής ενημέρωσης παρέχεται ή δυνατότητα της δημιουργίας μιάς μεγαλύτερης πολιτικής ευαισθητοποίησης και οργάνωσης, η οποία έχει σε πολλές περιπτώσεις επαναπροσδιορίσει ολόκληρο το πολιτικό σκηνικό από την αρχή. Μερικά παραδείγματα, με τα οποία μπορούμε να επικαιροποιήσουμε την παραπάνω άποψη είναι οι εκλογές του Ιράν, η επανεκλογή του Ομπάμα για δεύτερη θητεία ως Πρόεδρος των ΗΠΑ καθώς επίσης και οι πολιτικές αναταραχές στην Αίγυπτο.

Σημαντικό ρόλο έχουν ακόμη και στην καλλιέργεια της παιδείας. Τα παιδιά της νέας γενίας που άρχισαν να χρησιμοποιούν τα μέσα κοινωνικής δικτύωσης, απέκτησαν από νωρίς δεξιότητες επικοινωνίας, και γενικά κατάφεραν να γίνουν πιο επικοινωνιακοί και ανοιχτόμαυλοι άνθρωποι. Αυτή είναι μια ενθαρρυντική εξέλιξη, και χάρη στην πολύ μεγάλη διαθεσιμότητα των πληροφοριών, που είναι τόσο απλοϊκή και εύκολα προσβάσιμη ,μέσα στο διαδίκτυο, ο καθένας μπορεί να γίνει αρκετά έξυπνος ή ευφυής ανακαλύπτοντας και αναπτύσσοντας νέες δεξιότητες. Είναι σημαντικό να αναφερθεί πως έχουν αλλάξει τα πράγματα στην επιστήμη του μάρκετινγκ. Οι εταιρείες πλέον ξοδεύουν υπέρογκα ποσά σε κανάλια για τις διαφημίσεις τους, αλλά έρχονται όλο και πιο κοντά στον καταναλωτή μέσω αλληλεπιδράσεων που πραγματοποιούνται μέσω των social media και των ιστοσελίδων τους. Οι επιχειρήσεις σήμερα είναι σε θέση να καταλάβουν καλύτερα τις ανάγκες της αγοράς , σε μεγαλύτερο βαθμό από ότι στο παρελθόν.

Συμπέρασματικά, θα λέγαμε οτι υπάρχουν τόσοι πολλοί τρόποι που τα κοινωνικά μέσα μαζικής ενημέρωσης και δικτύωσης έχουν αλλάξει τον κόσμο. Οι παραπάνω απόψεις είναι απλά από τις πιο σημαντικές. Μέχρι σήμερα τα μέσα κοινωνικής δικτύωσης έχουν αποκτήσει μια σταθερή θέση στο μέλλον μας, και ελπίζουμε ότι οι δυνατότητες αυτών των πλατφόρμων θα επεκταθούν για να μπορέσει να γίνει η καθημερινότητας μας πιο έυκολη.

1.3 ΔΙΑΔΙΚΤΥΟ ΚΑΙ ΜΕΣΑ ΚΟΙΝΩΝΙΚΗΣ ΔΙΚΤΥΩΣΗΣ ΣΤΗΝ ΕΛΛΑΔΑ

Σύμφωνα με πρόσφατη έρευνα που έγινε και τα αποτελέσματα της δημοσιεύτηκαν στην ιστοσελίδα του συνδέσμου επιχειρήσεων πληροφορικής και επικοινωνιών Ελλάδας, τουλάχιστον 3,5 ώρες την ημέρα βρίσκεται στο διαδίκτυο ο μέσος Έλληνας [12]. Στις νεαρότερες ηλικίες, από 15 έως 24 ετών, αφιερώνουν στο Internet πάνω από 5 ώρες την ημέρα.

Με την online δραστηριότητα να κερδίζει συνεχώς έδαφος, προκύπτει ότι σχεδόν οι 8 στους 10 χρήστες (78%), ηλικίας 18 έως 34 ετών, έχουν κάνει τουλάχιστον μία online αγορά το τελευταίο 6μηνο. Βέβαια, το ποσοστό αυτό μειώνεται, όπως είναι αναμενόμενο ,στις μεγαλύτερες ηλικίες και διαμορφώνεται στο 29% στην ηλικιακή ομάδα άνω των 65 ετών.

Στο μεταξύ, τα μέσα κοινωνικής δικτύωσης, αποτελούν μια καθημερινή ασχολία για περίπου 6 στους 10 τους Έλληνες χρήστες. Τα υψηλότερα ποσοστά καθημερινής χρήσης των social media (83%) καταγράφονται στις ηλικίες 13 έως 17 ετών και στις ηλικίες 18 έως 24 ετών (91%). Τα άτομα ηλικίας 25 έως 34 ετών συνδέονται καθημερινά σε ποσοστό 78%, ενώ η χρήση βαίνει μειούμενη στις μεγαλύτερες ηλικίες: οι χρήστες 35 έως 44 ετών συνδέονται καθημερινά στα μέσα κοινωνικής δικτύωσης σε ποσοστό 66%, στις ηλικίες 45 έως 54 ετών το ποσοστό καθημερινής χρήσης πέφτει στο 55%.

Συνολικά, πάντως, σύμφωνα με τα στοιχεία της Focus Bari για το τρίμηνο Οκτωβρίου-Δεκεμβρίου 2019, 1 στους 4 Έλληνες χρήστες αφιερώνει στα social media πάνω από μία ώρα ημερησίως.

Mobile internet

Συνολικά, το 90% των Ελλήνων χρηστών συνδέεται στο Διαδίκτυο με οποιαδήποτε συχνότητα, ενώ το 84% βρίσκεται online καθημερινά. Επιπλέον, 9 στα 10 νοικοκυριά (87%) έχει μια σύνδεση στο Διαδίκτυο.

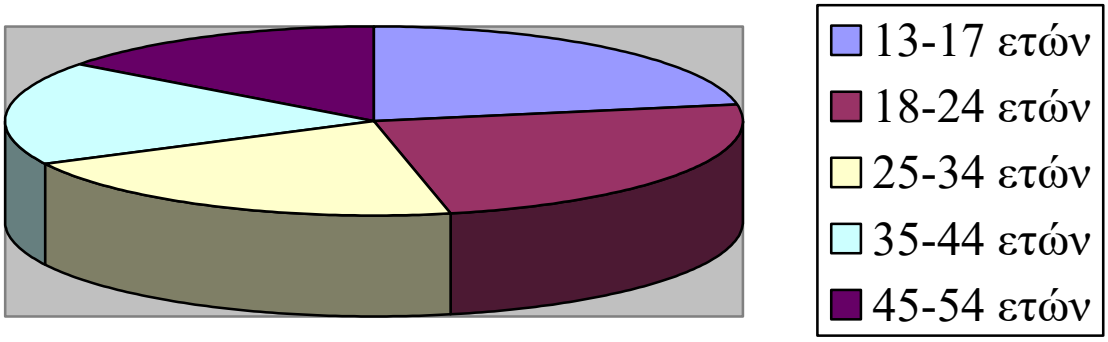
Ακολουθώντας τις διεθνείς τάσεις, οι χρήστες στην Ελλάδα προτιμούν να συνδέονται στο Διαδίκτυο από τις φορητές τους συσκευές, προκειμένου να είναι online απ' όπου κι εάν βρίσκονται. Με την κατοχή smartphone να φθάνει στη χώρα μας το 85%, είναι αναμενόμενο η χρήση Mobile Internet να κυμαίνεται στο 81%.

Μία άλλη τάση, που γίνεται καθημερινή και πιο οικεία για τον Έλληνα καταναλωτή, είναι η κατοχή συνδρομητικής τηλεόρασης, καθώς πλέον σχεδόν ένα στα δύο νοικοκυριά (45%) έχει συνδρομή σε μία σχετική υπηρεσία.

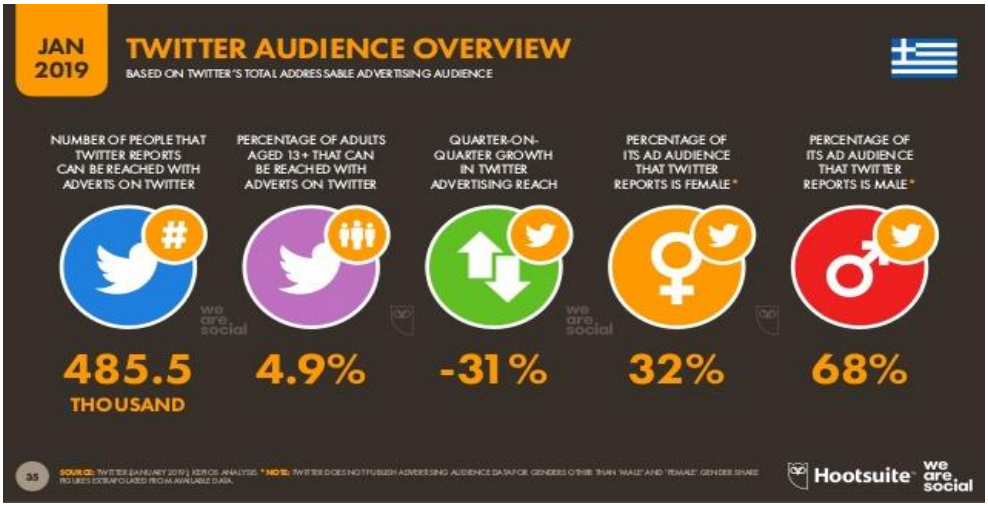
Από τα θρανία

Οι Έλληνες - όπως άλλωστε και οι πολίτες παγκοσμίως - ξεκινούν την ενασχόλησή τους με το Διαδίκτυο από πολύ τρυφερές ηλικίες. Για την ακρίβεια, 4 στα 5 παιδιά, ηλικίας 5 έως 12 ετών, ασχολείται με το Internet. Κατά μέσο όρο, η ενασχόληση με τον παγκόσμιο ιστό ξεκινά από την ηλικία των 5-6 ετών, ενώ η χρήση είναι σχεδόν καθολική στις ηλικίες 10-12 ετών.

Να σημειωθεί ότι τα smartphones είναι η πρώτη σε προτίμηση συσκευή μέσω της οποίας συνδέονται τα παιδιά με το Διαδίκτυο. Ακολουθεί ενδεικτικό διάγραμμα σε μορφή «πίτας» για την οπτικοποίηση των στατιστικών δεδομένων που προαναφέρθηκαν.



1.4 ΧΡΗΣΤΕΣ ΤΟΥ TWITTER ΣΤΗΝ ΕΛΛΑΔΑ



Εικόνα 2. «Στατιστική απεικόνιση καταμερισμού των χρηστών Twitter »

Οι χρήστες του Twitter στην Ελλάδα για το 2019, σύμφωνα με έρευνα που έγινε υπολογίζεται ότι είναι 485 χιλιάδες εκ των οποίων το 4,9% είναι ανήλικοι ηλικίας 13 και άνω, το 32% είναι γυναίκες και το 68% είναι είναι άντρες [11].

2.1ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ Ή ΕΞΟΡΥΞΗ ΓΝΩΜΗΣ

Η αυξανόμενη διαθεσιμότητα του διαδικτύου σηματοδότησε τη γέννηση ενός νέου τύπου κοινωνίας, όπου οι άνθρωποι μπορούν ελεύθερα να επικοινωνούν και να ανταλλάσσουν ιδέες και απόψεις. Οι απόψεις μπορεί να είναι πολύτιμες πληροφορίες γιατί έτσι μπορούμε να γνωρίζουμε ποιοι άνθρωποι μιλούν για ένα θέμα ή ακόμα και το πως αισθάνονται για αυτό, και έτσι διευκολύνεται η διαδικασία λήψης αποφάσεων. Χρησιμοποιώντας την εξόρυξη κειμένου, η οποία αποτελεί επέκταση της εξόρυξης δεδομένων, συγκεκριμένα δεδομένα από το περιεχόμενο της φυσικής γλώσσας (μη δομημένο κείμενο) που δεν έχει προκαθορισμένη μορφή, μπορούν να εξαχθούν, αλλά αυτές οι πληροφορίες δεν θα απαντήσουν στο ερώτημα σχετικά με το ποιές είναι οι απόψεις των ανθρώπων. Οι γνώμες είναι πολύτιμες πληροφορίες, επειδή είναι θεμελιώδες να γνωρίζουμε για το τι μιλάει ο κόσμος ή τι συναισθήματα έχει, που βοηθάει πολύ στη διαδικασία λήψης αποφάσεων. Με άλλα λόγια, η εξόρυξη κειμένου, η οποία εξάγει χρήσιμες πληροφορίες από το κείμενο δεν επαρκούν, αλλά απαιτείται μια πιο προηγμένη μέθοδος για την εξαγωγή των απόψεων που περιέχονται στα κείμενα. Αυτό δημιουργεί ένα νέο πεδίο έρευνας που ονομάζεται εξόρυξη γνώμης.

Η εξόρυξη γνώμης ή ανάλυση συναισθήματος προέκυψε ως μια αναπτυγμένη τεχνολογία για την παροχή καλύτερης μεθόδου,αναλύοντας τις απόψεις μέσα στο αδόμητο κείμενο. Δεν επικεντρώνεται μόνο στην πράξη ανάκτησης πληροφοριών από σχετικά έγγραφα και πηγές αλλά από την εξαγωγή πληροφοριών σχετικά με το γενικό συναίσθημα και απόψεις που περιέχονται στο κείμενο.

Σκοπός αυτής της ενότητας είναι να παρουσιάσει την εξέλιξη της εξόρυξης γνώμης μέχρι σήμερα, δηλαδή ένα «πέρασμα» μέσα από το χρονοδιάγραμμα της εξέλιξης της εξόρυξης γνώμης ξεκινώντας από τα συστήματα πληροφορικής που διαμορφώνουν τον άνθρωπο μέχρι την κατανόηση της φυσικής γλώσσας με τις τελευταίες προσεγγίσεις που χρησιμοποιούνται για την κατασκευή εξειδικευμένων συστημάτων εξόρυξης γνώμης. Θα επισημάνουμε επίσης τα δυνατά σημεία και τους περιορισμούς των τεχνικών εξόρυξης γνώμης σε κάθε φάση.

2.2 ΙΣΤΟΡΙΚΟ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Τεράστιος αριθμός σχολίων και ανατροφοδοτήσεων συλλέγονται καθημερινά από τους χρήστες.Είναι ένας τεράστιος όγκος δεδομένων για να αναλύθεί χειροκίνητα κάθε σχόλιο. Και εδώ μπαίνει η εξόρυξη γνώμης.Μια πρώτη ανασκόπηση της υπάρχουσας ερευνητικής βιβλιογραφίας σχετικά με τις μελέτες εξόρυξης γνώμης αποκαλύπτει ότι δεν υπάρχει μια ολοκληρωμένη ανάλυση εξέλιξης που επικεντρώνεται στο χρονοδιάγραμμα και την πρόοδο στον τομέα εξόρυξης γνώμης με εξαίρεση των Pang και Lee (2008),[1]. Επιπλέον, η μελέτη αυτή δεν έχει ταξινομημένο οδικό χάρτη και σαφή τρόπο παροχής ιστορίας εξόρυξης γνώμης και πιθανών πηγαίνει με τη ροή των χρησιμοποιούμενων αλγορίθμων και τεχνικών. Έτσι, ο σκοπός αυτής της ενότητας είναι μία διαφορετική προσέγγιση σε αυτό το θέμα. Έτσι η πρόοδος της εξόρυξης γνώμης σε ένα νοητό χρονοδιάγραμμα χωρίζεται σε 5 στάδια:

- Ερμηνεία κειμένου
- Σχολιασμός χαμηλής στάθμης
- Διαφορά μεταξύ υποκειμενικότητας και αντικειμενικότητας
- Εφαρμογές εξόρυξης δεδομένων στο Web
- Χρήση λεξικού

Πρώτη φάση από την ανασκόπηση της βιβλιογραφίας, ξεκίνησε όταν οι άνθρωποι προσπάθησαν να βρουν ένα αυτοματοποιημένο τρόπο ερμηνείας κειμένου αντί να στηρίζεται στον εγκέφαλό μας για να δουλεύει στη φυσική γλώσσα και να εκτελεί γρήγορα ανάλυση. Στην πρώτη φάση παρατηρήθηκε η γέννηση της θεωρίας του Banfield (1982), η οποία χρησιμοποιείται εκτενώς καθ 'όλη τη διάρκεια των πρώτων αναφορών στην ιστορία της εξόρυξης γνώμης.Πρότεινε τη χρήση υποκειμενικών και αντικειμενικών προτάσεων ως δεικτών, την αναζήτηση στο κείμενο με την παροχή απλών ερωτημάτων και τη χρήση του ψυχολογικού στοιχείου ως σημαντικού παράγοντα για τη φυσική επεξεργασία της γλώσσας.

Η δεύτερη φάση του χρονοδιαγράμματος απεικονίζει την ανάπτυξη της περιοχής σχολιασμού κειμένου, η οποία ορίζεται ως επισήμανση λέξεων κατα την ανάγνωση οποιασδήποτε μορφής κειμένου. Ο σχολιασμός συνδέεται στενά με τις πληροφορίες εξόρυξης λόγω της χρήσης των ταξινομητών και των ετικετών για την επισήμανση του κειμένου. Οι νέες εφαρμογές του ,οι σχολιασμοί κειμένου επέτρεψαν στην τεχνολογία εξόρυξης δεδομένων να αντικαταστήσει τις παραδοσιακές τεχνικές ερμηνείας κειμένου που οφείλονται στη χρήση των κορμών στο στάδιο της ανάλυσης.

Στις αρχές της δεκαετίας του 2000, η οποία σηματοδοτεί την τρίτη φάση του χρονοδιαγράμματος, η διαφορά μεταξύ υποκειμενικότητας και αντικειμενικότητας ήταν ο κύριος προβληματισμός που όμως παρήαγαγε πιο ακριβή ταξινόμηση πολικότητας. Διαφορετικό επίπεδο στη ταξινόμηση σε επίπεδο εγγράφου ή φράσης έδωσε ένα συνοπτικό και λεπτομερές αποτέλεσμα αντίστοιχα. Οι εξελίξεις των προσεγγίσεων ταξινόμησης στην τρίτη φάση ενισχύουν την εξόρυξη γνώμης σε νέες εφαρμογές και το έβαλε στον πυρήνα του διαδικτύου για να αναλύσει το αδόμητο κείμενο σε αρκετούς τομείς και περιβάλλοντα (δεδομένα ιστού εξόρυξη).

Η σταδιακή φάση εξόρυξης γνώμης προωθήθηκε στην τέταρτη φάση του χρονοδιαγράμματος, όπου είχαμε εφαρμογές σε πολλούς τομείς όπως η πολιτική, η υγεία, η ψυχαγωγία, το μάρκετινγκ και η κοινωνική δικτύωση.Τα Web δεδομένα,η εξόρυξη γνώμης στον ιστό,τα σχόλια χρηστών σε διάφορους ιστότοπους, βοηθούν στο να

αναπτυχθούν εφαρμογές ιστού για να βοηθήθεί περαιτέρω η ανάλυση και η αξιολόγηση προϊόντων και υπηρεσιών.

Η τελευταία φάση του χρονοδιαγράμματος απεικονίζει μια πολύ κοινή προσέγγιση για την εξόρυξη γνώμης που είναι γνωστή ως λεξική προσέγγιση.Οι λεξιλογικοί πόροι, όπως το WordNet και το SentiWordNet, παρέχουν τμήματα ομιλίας ομαδοποιημένα σε ειδική μορφή γνωστή ως γνωστική ομαδοποίηση.

Εικόνα 3. «Wordcloud ανάλυσης συναισθήματος»

2.3 ΤΙ ΕΙΝΑΙ Η ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Η ανάλυση συναισθήματος είναι ένας τομέας φυσικής επεξεργασίας γλώσσας, αφιερωμένος στη διερεύνηση υποκειμενικών απόψεων ή συναισθημάτων που συλλέγονται από διάφορες πηγές σχετικά με ένα συγκεκριμένο θέμα.Στην ανάλυση συναισθήματος συμπεριλαμβάνονται αλγόριθμοι οι οποίοι εκτελούν αναζήτηση βαθιά μέσα στο κείμενο και βρίσκουν στοιχειά που δείχνουν τη στάση απέναντι σε ενα προϊόν γενικά ή σε ένα συγκεκριμένο στοιχείο του.Με άλλα λόγια, η ανάλυση της γνώσης και της συναισθηματικής σκέψης σημαίνει μια ευκαιρία για να μπορέσουμε να εξερευνήσουμε, τη νοοτροπία των μελών του ακροατηρίου και να μελετήσουμε την κατάσταση του προϊόντος από την αντίθετη οπτική γωνία.[2],[13]

Αυτό κάνει την ανάλυση του συναισθήματος εξαιρετικό εργαλείο για:

- Διευρυμένη ανάλυση προϊόντων
- Έρευνα αγοράς
- Διαχείριση φήμης
- Στόχευση ακρίβειας
- Ανάλυση μάρκετινγκ
- Δημόσιες σχέσεις (PR)
- Κριτικές προϊόντων
- Καθαρή βαθμολογία υποκινητή
- Ανατροφοδότηση προϊόντος
- Εξυπηρέτηση πελατών

2.4 ΠΩΣ ΔΟΥΛΕΥΕΙ Η ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Η ανάλυση συναισθήματος είναι ένας αλγόριθμος ,κατά κύριο λόγο ταξινόμησης, που στοχεύει στην εξεύρεση μιας άποψης ή διάθεσης, καθώς και στην ανάδειξη των πληροφοριών που παρουσιάζουν ιδιαίτερο ενδιαφέρον για τη διαδικασία.

Τι είναι η "άποψη" στην ανάλυση των συναισθημάτων; Πριν απαντήσουμε σε αυτό το ερώτημα αξίζει να αναφερθεί ο γενικός ορισμός της άποψης: [2],[13]

"Μία άποψη ή κρίση σχηματίστηκε για κάτι, όχι κατ 'ανάγκη βασισμένο σε γεγονός ή γνώση"

Όμως από την άποψη της επιστήμης των δεδομένων, μια άποψη είναι πολύ περισσότερο από αυτό.Από τη μία πλευρά, είναι μια υποκειμενική εκτίμηση για κάτι που βασίζεται στην προσωπική εμπειρική εμπειρία. Είναι μερικώς ριζωμένη σε αντικειμενικά γεγονότα και εν μέρει κυριαρχείται από συναισθήματα.Από την άλλη πλευρά, μια άποψη μπορεί να ερμηνευτεί ως ένα είδος διάστασης στα δεδομένα που αφορούν ένα συγκεκριμένο θέμα. Πρόκειται για ένα σύνολο σημάτων που σε συνδυασμό παρουσιάζουν μια άποψη, δηλαδή μια πτυχή για το συγκεκριμένο ζήτημα. Σύμφωνα λοιπόν με αυτή την προσέγγιση, η ανάλυση συναισθήματος εφαρμόζεται για τις ακόλουθες λειτουργίες:

- Βρίσκει και εξαγάγει τα εκτιμημένα δεδομένα (γνωστά και ως δεδομένα αισθήσεων) σε μια συγκεκριμένη πλατφόρμα (υποστήριξη πελατών, κριτικές κ.λπ.)
- Προσδιορίζει την πολικότητα (θετική ή αρνητική)

- Καθορίζει το θέμα (αυτό που μιλάμε γενικά και συγκεκριμένα)
- Προσδιορίζει τον κάτοχο της γνώμης (μόνος του και σε συσχετισμό με τα υπάρχοντα τμήματα κοινού)

Ανάλογα με το σκοπό, ο αλγόριθμος ανάλυσης συναίσθημα μπορεί να χρησιμοποιηθεί στα ακόλουθα πεδία:

- Επίπεδο εγγράφου - για ολόκληρο το κείμενο.
- Επίπεδο καταδίκης - παίρνει το συναίσθημα μιας μόνο φράσης.
- Επίπεδο υποτιτλισμού - αποκτά το συναίσθημα των υπο-εκφράσεων μέσα σε μια πρόταση.

Δεδομένης της υποκειμενικής του υπόθεσης, η εξόρυξη μιας γνώμης είναι μια δύσκολη διαδικασία. Οι απόψεις διαφέρουν και ορισμένες είναι πιο πολύτιμες από τις άλλες. Τέσσερις υποκατηγορίες χαρακτηρίζουν περαιτέρω μια γνώμη:

- Άμεση γνώμη είναι αυτή που δηλώνει άμεσα κάτι. Για παράδειγμα, "η ανταπόκριση των κουμπιών στην εφαρμογή Χ είναι κακή."

- Συγκριτική γνώμη είναι εκείνη όπου το Χ συγκρίνεται με το Υ με βάση συγκεκριμένα κριτήρια. Για παράδειγμα, "η ανταπόκριση του κουμπιού στην εφαρμογή Χ είναι χειρότερη από την εφαρμογή Υ."

- Η ρητή γνώμη είναι όπου τα πάντα είναι σαφώς καθορισμένα. Για παράδειγμα, "αυτή η καρέκλα λικνίζει".

- Οι σιωπηρές απόψεις υπονοούνται αλλά δεν δηλώνονται με σαφήνεια. Για παράδειγμα, "η εφαρμογή ξεκίνησε να υστερεί σε δύο ημέρες". Είναι σημαντικό να σημειωθεί ότι οι σιωπηρές απόψεις μπορεί επίσης να έχουν ιδιώματα και μεταφορές, γεγονός που περιπλέκει τη διαδικασία ανάλυσης των αισθήσεων.

2.5 ΣΗΜΑΣΙΑ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΤΟΝ ΕΠΙΧΕΙΡΗΜΑΤΙΚΟ ΤΟΜΕΑ

Η ανάλυση του συναισθήματος ασχολείται με την αντίληψη του προϊόντος και την κατανόηση της αγοράς μέσω του φακού των δεδομένων συναισθημάτων. Υπάρχουν πολλές πηγές δημόσιων και ιδιωτικών πληροφοριών, από τις οποίες μπορείτε να αξιοποιήσετε μια εικόνα της αντίληψης του πελάτη σχετικά με το προϊόν και τη γενική κατάσταση της αγοράς.

Για να αναφέρουμε μερικά:

- Αλληλογραφία υποστήριξης πελατών (σχετικά με το προϊόν σας)
- Παρατηρήσεις για προϊόντα που δημιουργούν οι χρήστες
- Κριτικές επαγγελματικών προϊόντων (όπως στο The Verge or Wired)
- Διαδρομές κοινωνικών μέσων
- Φόρουμ γενικών και ειδικών σκοπών

Η Ανάλυση Συναλλαγών Πελατών μπορεί να σας βοηθήσει να καταστήσετε νόημα αυτά τα θραύσματα δεδομένων και να τα μετατρέψετε σε:

- μια σαφώς καθορισμένη άποψη για το τι σκέφτονται ορισμένα τμήματα των πελατών για το προϊόν ή
- γενικότερα Μια βαθιά κατάδυση στην κατάσταση της αγοράς από την άποψη του καταναλωτή.

Και στις δύο περιπτώσεις, είναι ένας σημαντικός παράγοντας για τη διατύπωση και την επεξεργασία της προτεινόμενης αξίας για ένα συγκεκριμένο τμήμα ακροατηρίου. Στην περίπτωση της έρευνας αγοράς, ο ρόλος της ανάλυσης του συναισθήματος είναι λιγότερο ολοκληρωμένος αλλά επιρροής παρ 'όλα αυτά. Δίνει μια άλλη προοπτική, προσθέτει επιπλέον χρώματα στην εικόνα της αγοράς και σας επιτρέπει να δείτε την κατάσταση από το επίπεδο του εδάφουςΕνώ στα αρχικά στάδια αυτές οι δραστηριότητες είναι σχετικά εύκολο να χειριστούν με βασικές λύσεις - σε κάποιο σημείο, αρχίζει να έχει νόημα να χρησιμοποιούν πιο περίτεχνα εργαλεία και να εξάγουν πιο περίπλοκες γνώσεις [13].

2.6 ΤΥΠΟΙ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Για να γίνει καλύτερα κατανοητό πώς εφαρμόζεται η ανάλυση συναισθημάτων θα πρέπει πρώτα να κατανοήσουμε τους διαφορετικούς τύπους.

Σε αυτή την ενότητα θα εξετάσουμε τους κύριους τύπους ανάλυσης αισθήσεων.[13]

- 1ου τύπου. Η λεπτομερής ανάλυση του συναισθήματος περιλαμβάνει τον καθορισμό της πολικότητας της γνώμης. Μπορεί να είναι μια απλή δυαδική θετική / αρνητική διαφοροποίηση συναισθημάτων. Αυτός ο τύπος μπορεί επίσης να ακολουθήσει τις πιο υψηλές προδιαγραφές (για παράδειγμα, πολύ θετικές, θετικές, ουδέτερες, αρνητικές, πολύ αρνητικές), ανάλογα με την περίπτωση χρήσης.

- 2ου τύπου. Η ανίχνευση των συναισθημάτων χρησιμοποιείται για την αναγνώριση σημείων συγκεκριμένων συναισθηματικών καταστάσεων που παρουσιάζονται στο κείμενο. Συνήθως, υπάρχει ένας συνδυασμός λεξικών και αλγορίθμων μηχανικής μάθησης που καθορίζουν τι είναι αυτό και γιατί.

- 3ου τύπου. Η ανάλυση των συναισθημάτων με βάση την πτυχή γίνεται βαθύτερη. Σκοπός του είναι να προσδιορίσει μια άποψη σχετικά με ένα συγκεκριμένο στοιχείο. Για παράδειγμα, η φωτεινότητα του φακού στο smartphone. Η βασισμένη σε θέματα ανάλυση χρησιμοποιείται συνήθως στην ανάλυση προϊόντων για να παρακολουθεί το πώς αντιλαμβάνεται το προϊόν και ποια είναι τα ισχυρά και αδύνατα σημεία από την πλευρά του πελάτη.

- 4ου τύπου. Η Ανάλυση Προσπάθειας αφορά στη δράση. Σκοπός του είναι να προσδιορίσει τι είδους πρόθεση εκφράζεται στο μήνυμα. Χρησιμοποιείται συνήθως στα συστήματα υποστήριξης πελατών για τον εξορθολογισμό της ροής εργασίας.

2.7 ΑΛΓΟΡΙΘΜΟΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΒΑΣΙΣΜΕΝΟΣ ΣΕ ΚΑΝΟΝΕΣ

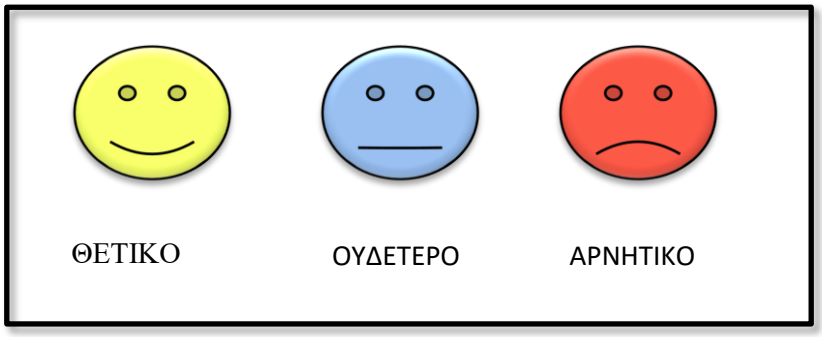
Η ανάλυση σε κανόνες βασίζεται σε έναν αλγόριθμο με σαφώς καθορισμένη περιγραφή μιας γνώμης που πρέπει να προσδιοριστεί. Περιλαμβάνει τον προσδιορισμό της υποκειμενικότητας, της πολικότητας ή του θέματος της γνώμης. Η προσέγγιση βασισμένη σε κανόνες περιλαμβάνει τη βασική ρουτίνα επεξεργασίας φυσικής γλώσσας. Περιλαμβάνει τις ακόλουθες λειτουργίες με το σώμα κειμένου[13]:

- Βλάστηση
- Τεκμηρίωση
- Μέρος της ετικέτας
- ομιλίας
- Τεχνολογία
- Ανάλυση λεξικών

Υπάρχουν δύο λίστες λέξεων. Μία από αυτές περιλαμβάνει μόνο τα θετικά, η άλλη περιλαμβάνει τα αρνητικά. Ο αλγόριθμος περνά μέσα από το κείμενο, βρίσκει τις λέξεις που ταιριάζουν με τα κριτήρια. Μετά από αυτό, ο αλγόριθμος υπολογίζει ποιος τύπος λέξεων είναι πιο διαδεδομένος στο κείμενο. Εάν υπάρχουν περισσότερες θετικές λέξεις, τότε το κείμενο θεωρείται ότι έχει θετική πολικότητα. Το πράγμα με αλγόριθμους βασισμένους σε κανόνες είναι ότι, ενώ προσφέρει κάποια αποτελέσματα - δεν διαθέτει ευελιξία και ακρίβεια που θα τα καθιστούσε πραγματικά χρησιμοποιήσιμα. Για παράδειγμα, η προσέγγιση που βασίζεται σε κανόνες δεν λαμβάνει υπόψη το πλαίσιο. Ωστόσο, μπορεί να χρησιμοποιηθεί για γενικούς σκοπούς για τον προσδιορισμό του τόνου των μηνυμάτων, τα οποία μπορεί να είναι χρήσιμα για την υποστήριξη πελατών. Αυτές τις μέρες, η βασισμένη σε κανόνες ανάλυση συναισθημάτων χρησιμοποιείται συνήθως για να τεθούν οι βάσεις για τη μετέπειτα υλοποίηση και εκπαίδευση της λύσης μηχανικής μάθησης.

2.8 ΧΡΗΣΗ ΤΗΣ ΑΝΑΛΥΣΗΣ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΣΗΜΕΡΑ

Η ανάλυση συναισθήματος σε δεδομένα μεγάλου όγκου χρηστών του διαδικτύου κερδίζει ολοένα και μεγαλύτερο έδαφος στο ακαδημαϊκό χώρο καθώς και στον επιχειρηματικό χώρο. Οι ακαδημαϊκοί ερευνητές εντοπίζουν μεγάλο ενδιαφέρον στις τεχνικές προκλήσεις που παρουσιάζει η ανάλυση συναισθήματος ενώ, οι επιχειρηματίες εντοπίζουν το ενδιαφέρον τους στις πολλά υποσχόμενες προοπτικές της. Καινοτόμες επιχειρήσεις ασχολούνται με την εξόρυξη γνώσης μέσα από τις αξιολογήσεις χρηστών σε ηλεκτρονικά καταστήματα και κοινωνικά δίκτυα με βάση την ανάλυση συναισθήματος. Τέλος, υπάρχει μεγάλη επιρροή καταναλωτών και χρηστών του διαδικτύου από τις κριτικές που υπάρχουν στον ηλεκτρονικό κόσμο πριν λάβουν κάποια απόφαση για αγορά μιας υπηρεσίας ή ενός προϊόντος.



Εικόνα 4. «Ενδεικτική απεικόνιση τριών βασικών συναισθημάτων στην συναισθηματική ανάλυση»

3.1 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Η Μηχανική Μάθηση είναι ένα σύστημα που μπορεί να μάθει από ένα παράδειγμα μέσω της αυτο-βελτίωσης και χωρίς να κωδικοποιείται ρητά από τον προγραμματιστή. Η ανακάλυψη έρχεται με την ιδέα ότι μια μηχανή μπορεί να μάθει ξεχωριστά από τα δεδομένα, για να παράγει ακριβή αποτελέσματα. Η μηχανική μάθηση συνδυάζει δεδομένα με στατιστικά εργαλεία για να προβλέψει μια παραγωγή. Αυτή η έξοδος χρησιμοποιείται στη συνέχεια από την εταιρία για να παράγει πραγματικές πληροφορίες. Η εκμάθηση μηχανών σχετίζεται στενά με την εξόρυξη δεδομένων και με τη Bayesian predictive modeling. Το μηχάνημα λαμβάνει δεδομένα ως είσοδο και χρησιμοποιεί έναν αλγόριθμο για τη διατύπωση απαντήσεων. Ένα τυπικό έργο εκμάθησης μηχανών είναι η παροχή μιας σύστασης. Για όσους έχουν λογαριασμό Netflix για παράδειγμα, όλες οι συστάσεις ταινιών ή σειρών βασίζονται στα ιστορικά δεδομένα του χρήστη. Οι εταιρείες τεχνολογίας χρησιμοποιούν μη επιτηρούμενη μάθηση για να βελτιώσουν την εμπειρία των χρηστών με την εξατομίκευση των προτάσεων. Η εκμάθηση μηχανών χρησιμοποιείται επίσης για μια ποικιλία εργασιών, όπως ανίχνευση απάτης, προβλέψιμη συντήρηση, βελτιστοποίηση χαρτοφυλακίου, αυτοματοποίηση εργασιών και ούτω καθεξής.

3.2 ΠΩΣ ΔΟΥΛΕΥΕΙ Η ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Η μηχανική μάθηση είναι ο εγκέφαλος όπου λαμβάνει χώρα όλη η εκμάθηση. Ο τρόπος με τον οποίο μαθαίνει η μηχανή είναι παρόμοιος με τον άνθρωπο. Οι άνθρωποι μαθαίνουν από την εμπειρία. Όσο περισσότερο γνωρίζουμε, τόσο πιο εύκολα μπορούμε να προβλέψουμε. Κατ 'αναλογία, όταν αντιμετωπίζουμε μια άγνωστη κατάσταση, η πιθανότητα επιτυχίας είναι χαμηλότερη από τη γνωστή κατάσταση. Τα μηχανήματα εκπαιδεύονται πάνω σε αυτό. Για να γίνει μια ακριβή πρόβλεψη, το μηχάνημα βλέπει ένα παράδειγμα. Όταν δίνουμε στο μηχάνημα ένα παρόμοιο παράδειγμα, μπορεί να καταλάβει το αποτέλεσμα. Ωστόσο, όπως ένας άνθρωπος, εάν τροφοδοτήσει ένα προηγουμένως αόρατο παράδειγμα, η μηχανή έχει δυσκολίες να προβλέψει.

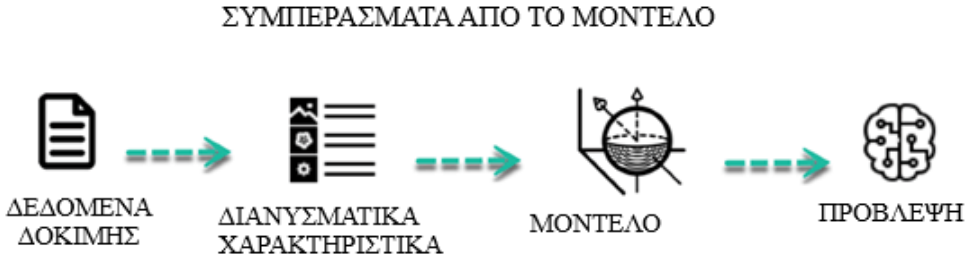
Ο βασικός στόχος της μηχανικής μάθησης είναι η εκμάθηση και ο συμπερασμός. Πρώτα απ 'όλα, η μηχανή μαθαίνει μέσω της ανακάλυψης μοτίβων. Αυτή η ανακάλυψη γίνεται χάρη στα δεδομένα.

Ένα σημαντικό μέρος του επιστήμονα δεδομένων είναι να επιλέξει προσεκτικά τα δεδομένα που πρέπει να παρέχει στο μηχάνημα. Ο κατάλογος των χαρακτηριστικών που χρησιμοποιούνται για την επίλυση ενός προβλήματος ονομάζεται διάνυσμα χαρακτηριστικών. Μπορούμε να σκεφτούμε ένα διάνυσμα χαρακτηριστικών ως υποσύνολο δεδομένων που χρησιμοποιείται για την αντιμετώπιση ενός προβλήματος. Το μηχάνημα χρησιμοποιεί κάποιους αλάνθαστους φανταχτερούς τρόπους για να απλοποιήσει την πραγματικότητα και να μετατρέψει αυτή την ανακάλυψη σε μοντέλο. Επομένως, το στάδιο μάθησης χρησιμοποιείται για να περιγράψει τα δεδομένα και να τα συνοψίσει σε ένα μοντέλο.

Για παράδειγμα,το μοντέλο μπορεί να είναι το μηχάνημα που προσπαθεί να κατανοήσει τη σχέση μεταξύ του μισθού ενός ατόμου και της πιθανότητας να πάει σε ένα φανταχτερό εστιατόριο. Αποδεικνύεται ότι το μηχάνημα βρίσκει μια θετική σχέση μεταξύ μισθού και πηγαίνοντας σε εστιατόριο υψηλής ποιότητας [14].

Εικόνα 5. «Διαδικασία κατασκευής μοντέλου στη μηχανική μάθηση»

Συμπερασματικά,όταν το μοντέλο είναι κατασκευασμένο, είναι δυνατό να δοκιμαστεί κατα πόσο ισχυρό είναι στα δεδομένα που δεν έχει δει ποτέ πριν. Τα νέα δεδομένα μετασχηματίζονται σε ένα διάνυσμα χαρακτηριστικών, περνούν από το μοντέλο και δίνουν μια πρόβλεψη. Αυτό είναι το πανέμορφο μέρος της μηχανικής μάθησης. Δεν χρειάζεται να ενημερώσουμε τους κανόνες ή να εκπαιδεύσουμε ξανά το μοντέλο.



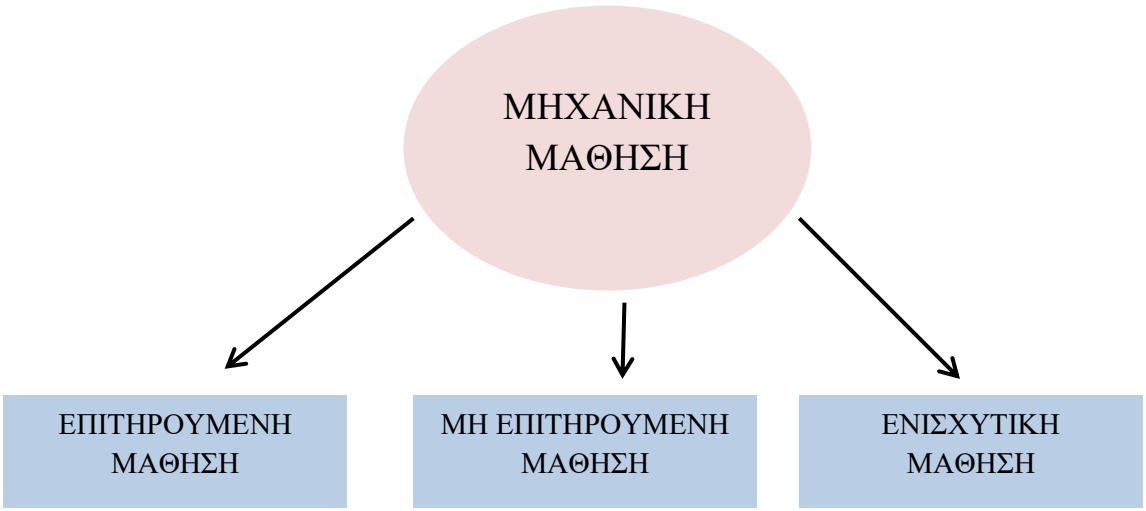
Εικόνα 6. «Στάδια επεξεργασίας για την εξόρυξη γνώσης»

Η λειτουργία των προγραμμάτων Machine Learning είναι απλή και μπορεί να συνοψιστεί στα ακόλουθα σημεία [14]:

1. Ορίζετε μια ερώτηση
2. Συλλογή δεδομένων
3. Οπτικοποίηση δεδομένων
4. Αλγόριθμος επεξεργασίας
5. Δοκιμασία αλγορίθμου
6. Συλλογή σχολίων
7. Βελτιώση αλγόριθμου
8. Επαναλαμβάνουμε τα βήματα 4 και 7 έως ότου τα αποτελέσματα ικανοποιηθούν
9. Χρησιμοποίηση του μοντέλου για να γίνει μια πρόβλεψη

Μόλις ο αλγόριθμος καταφέρει να βγάλει τα σωστά συμπεράσματα, εφαρμόζει αυτή τη γνώση σε νέα σύνολα δεδομένων.

3.3 ΑΛΓΟΡΙΘΜΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΚΑΙ ΠΩΣ ΧΡΗΣΙΜΟΠΟΙΟΥΝΤΑΙ



Εικόνα 7. «Τομείς μηχανικής μάθησης»

Η μηχανική μάθηση μπορεί να ομαδοποιηθεί σε δύο εκτεταμένα μαθησιακά καθήκοντα, με επίβλεψη και χωρίς επιτήρηση. Υπάρχουν πολλοί άλλοι αλγόριθμοι αλλά στα πλαίσια της παρούσας διπλωματικής επιλέχτηκαν να αναφερθούν τα ακόλουθα:[14],[4]

Εποπτευόμενη μάθηση

Ένας αλγόριθμος χρησιμοποιεί δεδομένα εκπαίδευσης και ανατροφοδότηση από τον άνθρωπο για να μάθει τη σχέση των δεδομένων εισόδων σε μια δεδομένη έξοδο. Για παράδειγμα, ένας επαγγελματίας μπορεί να χρησιμοποιήσει τα έξοδα μάρκετινγκ και την πρόγνωση του καιρού ως δεδομένα εισόδου για να προβλέψει τις πωλήσεις κονσερβών.

Η εποπτευόμενη μάθηση μπορεί να χρησιμοποιηθεί όταν είναι γνωστά τα δεδομένα εξόδου. Ο αλγόριθμος θα προβλέψει νέα δεδομένα.

Υπάρχουν δύο κατηγορίες εποπτευόμενης μάθησης:

- Ταξινόμηση
- Λειτουργία παλινδρόμησης

Ταξινόμηση

Έστω ότι θέλουμε να προβλέψουμε το φύλο ενός πελάτη για ένα εμπορικό. Θα αρχίσουμε να συλλέγουμε δεδομένα σχετικά με το ύψος, το βάρος, την εργασία, το μισθό, το καλάθι αγορών κ.λπ. από τη βάση δεδομένων πελατών μας. Γνωρίζουμε το φύλο του κάθε πελάτη μας, μπορεί να είναι μόνο άνδρας ή γυναίκα. Ο στόχος του ταξινομητή θα είναι να εκχωρήσει μια πιθανότητα να είναι αρσενικό ή θηλυκό (δηλαδή η ετικέτα) με βάση τις

πληροφορίες (δηλ. Τα χαρακτηριστικά που έχετε συλλέξει). Όταν το μοντέλο μάθει πώς να αναγνωρίζει αρσενικό ή θηλυκό, μπορούμε να χρησιμοποιήσουμε νέα δεδομένα για να κάνετε μια πρόβλεψη. Για παράδειγμα, μόλις πήραμε νέες πληροφορίες από έναν άγνωστο πελάτη και θέλετε να μάθετε αν είναι άντρας ή γυναίκα. Εάν ο ταξινομητής προβλέπει ανδρικό = 70%, σημαίνει ότι ο αλγόριθμος είναι σίγουρος στο 70% ότι αυτός ο πελάτης είναι άνδρας και το 30% είναι θηλυκό [4], [14].

Η ετικέτα μπορεί να αποτελείται από δύο ή περισσότερες κατηγορίες. Το παραπάνω παράδειγμα έχει μόνο δύο κατηγορίες, αλλά αν ένας ταξινομητής χρειάζεται να προβλέψει αντικείμενο, έχει δεκάδες κλάσεις (π.χ. γυαλί, τραπέζι, παπούτσια κλπ., κάθε αντικείμενο αντιπροσωπεύει μια κλάση)

Οπισθοδρόμηση

Όταν η έξοδος είναι μια συνεχής τιμή, η εργασία είναι μια παλινδρόμηση. Για παράδειγμα, ένας οικονομικός αναλυτής μπορεί να χρειαστεί να προβλέψει την αξία ενός αποθέματος με βάση ένα φάσμα χαρακτηριστικών όπως μετοχές, προηγούμενες αποδόσεις μετοχών, δείκτη μακροοικονομίας. Το σύστημα θα εκπαιδευτεί για την εκτίμηση της τιμής των αποθεμάτων με το μικρότερο δυνατό λάθος. Παρακάτω αναφέρονται κάποιοι ενδεικτικοί αλγόριθμοι καθώς και μία σύντομη περιγραφή τους [4],[14].

- Γραμμική παλινδρόμηση: Βρίσκει έναν τρόπο συσχέτισης κάθε χαρακτηριστικού με την έξοδο για να βοηθήσει στην πρόβλεψη μελλοντικών τιμών. Ανήκει στους αλγόριθμους παλινδρόμησης.

- Λογιστική παλινδρόμηση : Επέκταση γραμμικής παλινδρόμησης που χρησιμοποιείται για εργασίες ταξινόμησης. Η μεταβλητή εξόδου είναι δυαδική (π.χ. μόνο μαύρη ή λευκή) και όχι συνεχής (π.χ. μια άπειρη λίστα πιθανών χρωμάτων). Ανήκει στους αλγόριθμους ταξινόμησης.

- Δέντρο απόφασης :Υψηλά ερμηνεύσιμο μοντέλο ταξινόμησης ή παλινδρόμησης που διαιρεί τις τιμές των χαρακτηριστικών δεδομένων σε κλάδους στους κόμβους απόφασης (π.χ. εάν ένα χαρακτηριστικό είναι ένα χρώμα, κάθε πιθανό χρώμα γίνεται ένας νέος κλάδος) έως ότου φτάσει σε ένα τελικό αποτέλεσμα. Ανήκει στους αλγόριθμους παλινδρόμησης και ταξινόμησης.

- Naive Bayes : Η Bayesian μέθοδος είναι μια μέθοδος ταξινόμησης που χρησιμοποιεί το Bayesian θεώρημα. Το θεώρημα ενημερώνει την προηγούμενη γνώση ενός γεγονότος με την ανεξάρτητη πιθανότητα κάθε χαρακτηριστικού που μπορεί να επηρεάσει το συμβάν. Επίσης ανήκει στους αλγορίθμους παλινδρόμησης και ταξινόμησης.

- Support Vector Machine: Ο Support Vector Machine, ή SVM, χρησιμοποιείται συνήθως για την εργασία ταξινόμησης. Ο αλγόριθμος SVM βρίσκει ένα τρόπο που διαιρεί κατά βέλτιστο τρόπο τις κλάσεις. Χρησιμοποιείται καλύτερα με έναν μη

γραμμικό περιβάλλον. Ανήκει και αυτός στους αλγορίθμους παλινδρόμησης και ταξινόμησης.

- Random forest: Ο αλγόριθμος βασίζεται σε ένα δέντρο απόφασης για να βελτιώσει δραστικά την ακρίβεια. Το τυχαίο δάσος παράγει πολλές φορές απλά δέντρα απόφασης και χρησιμοποιεί τη μέθοδο της «πλειοψηφίας» για να αποφασίσει ποια επισήμανση θα επιστρέψει. Για την ταξινόμηση, η τελική πρόβλεψη θα είναι αυτή με την μεγαλύτερη ψήφο. ενώ για την εργασία παλινδρόμησης, η μέση πρόβλεψη όλων των δένδρων είναι η τελική πρόβλεψη. Ανήκει στους αλγορίθμους παλινδρόμησης και ταξινομησης.
- AdaBoost: Η ταξινόμηση ή η τεχνική παλινδρόμησης που χρησιμοποιεί μια πληθώρα μοντέλων για να καταλήξει σε μια απόφαση, αλλά τα ταξινομεί με βάση την ακρίβειά τους στην πρόβλεψη του αποτελέσματος. Ανήκει στους αλγορίθμους παλινδρόμησης.
- Gradient Boosting Trees: Τα δέντρα που ενισχύουν τη διαβάθμιση είναι μια σύγχρονη τεχνική ταξινόμησης / παλινδρόμησης. Επικεντρώνεται στο σφάλμα που διαπράττουν τα προηγούμενα δέντρα και προσπαθεί να το διορθώσει. Ανήκει στους αλγορίθμους παλινδρόμησης και ταξινόμησης.

Μη εποπτευόμενη μάθηση

Στην μη επιτηρούμενη μάθηση, ένας αλγόριθμος διερευνά τα δεδομένα εισόδου χωρίς να δοθεί μια ρητή μεταβλητή εξόδου (π.χ., διερευνά τα δημογραφικά δεδομένα του πελάτη για να ταυτοποιήσει μοτίβα) Μπορεί να χρησιμοποιηθεί όταν δεν ξέρουμε πώς να ταξινομήσουμε τα δεδομένα και θέλουμε ο αλγόριθμος να βρει μοτίβα και να ταξινομήσει τα δεδομένα για εμάς [4], [14].

- K-means συσσωμάτωση: Βάζει δεδομένα σε ορισμένες ομάδες (k) που περιέχουν δεδομένα με παρόμοια χαρακτηριστικά (όμως καθορίζονται από το μοντέλο, όχι εκ των προτέρων από τον άνθρωπο).Ανήκει στους αλγορίθμους συσσωμάτωσης (clustering).

- Gaussian mixture model : Μια γενίκευση της ομαδοποίησης k-means που παρέχει μεγαλύτερη ευελιξία στο μέγεθος και το σχήμα των ομάδων (συστάδες). Ανήκει στους αλγορίθμους συσσωμάτωσης.

- Ιεραρχική ομαδοποίηση :Διαχωρίζει τα clusters κατά μήκος ενός ιεραρχικού δέντρου για να σχηματίσει ένα σύστημα ταξινόμησης. Ανήκει στους αλγορίθμους συσσωμάτωσης.

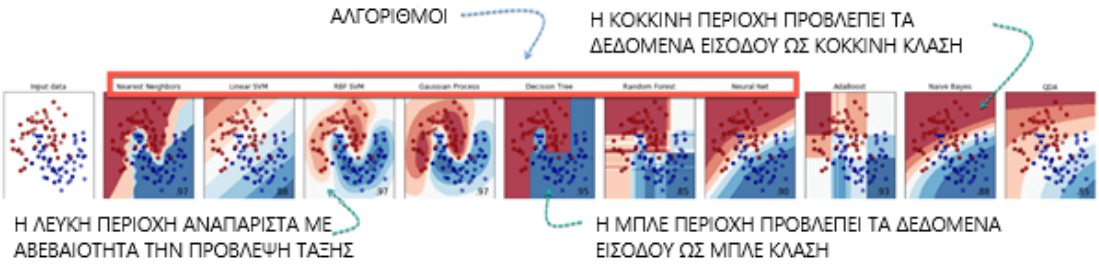
- Σύστημα συστάσεων(recommender system): Βοηθάει ώστε να καθορίστουν τα σχετικά δεδομένα για την υποβολή μιας σύστασης.Ανήκει στους αλγορίθμους συσσωμέτωσης.

- PCA / T-SNE: Χρησιμοποιείται κυρίως για τη μείωση της διαστάσεων των δεδομένων. Οι αλγόριθμοι μειώνουν τον αριθμό των χαρακτηριστικών σε 3 ή 4 διανύσματα με τις υψηλότερες διακυμάνσεις. Ανήκει στους αλγορίθμους μείωσης διάστασης.

3.4 ΠΩΣ ΕΠΙΛΕΓΟΥΜΕ ΕΝΑΝ ΑΛΓΟΡΙΘΜΟ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Υπάρχουν πολλοί αλγόριθμοι μηχανικής μάθησης. Η επιλογή του αλγορίθμου βασίζεται στον στόχο.

Στο παρακάτω παράδειγμα, η εργασία είναι να προβλέψουμε τον τύπο του λουλουδιού ανάμεσα στις τρεις ποικιλίες. Οι προβλέψεις βασίζονται στο μήκος και το πλάτος του πετάλου. Η εικόνα απεικονίζει τα αποτελέσματα δέκα διαφορετικών αλγορίθμων. Η εικόνα στο επάνω αριστερό μέρος είναι το σύνολο δεδομένων. Τα δεδομένα ταξινομούνται σε τρεις κατηγορίες: κόκκινο, γαλάζιο και σκούρο μπλε. Υπάρχουν μερικές ομάδες. Για παράδειγμα, από τη δεύτερη εικόνα, όλα στην επάνω αριστερή άκρη ανήκουν στην κόκκινη κατηγορία, στο μεσαίο τμήμα υπάρχει ένα μείγμα αβεβαιότητας και γαλάζιο, ενώ το κάτω μέρος αντιστοιχεί στη σκοτεινή κατηγορία. Οι άλλες εικόνες δείχνουν διαφορετικούς αλγόριθμους και πώς προσπαθούν να ταξινομήσουν τα δεδομένα [4], [14].



Εικόνα 8. «Πρόβλεψη τύπου λουλουδιού ανάμεσα σε τρείς ποικιλίες»

3.5 ΠΡΟΚΛΗΣΕΙΣ ΚΑΙ ΠΕΡΙΟΡΙΣΜΟΙ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Η κύρια πρόκληση της μηχανικής μάθησης είναι η έλλειψη δεδομένων ή η ποικιλομορφία στο σύνολο δεδομένων. Ένα μηχάνημα δεν μπορεί να μάθει αν δεν υπάρχουν διαθέσιμα δεδομένα. Εκτός αυτού, ένα σύνολο δεδομένων με έλλειψη διαφορετικότητας δίνει στο μηχάνημα χρόνο. Μια μηχανή πρέπει να έχει ετερογένεια για να μάθει μια ουσιαστική γνώση. Είναι γνωστό ότι ένας αλγόριθμος δεν μπορεί να εξάγει πληροφορίες όταν δεν υπάρχουν παραλλαγές. Συνιστάται να υπάρχουν τουλάχιστον 20 παρατηρήσεις ανά ομάδα για να βοηθήθεί η μηχανή και να μάθει. Αυτός ο περιορισμός οδηγεί σε κακή αξιολόγηση και πρόβλεψη.

3.6 ΕΦΑΡΜΟΓΕΣ ΤΗΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ [9]

- Αύξηση:

Μηχανική μάθηση, η οποία βοηθά τους ανθρώπους με τις καθημερινές τους εργασίες, προσωπικά ή εμπορικά χωρίς πλήρη έλεγχο της παραγωγής. Αυτή η εκμάθηση μηχανών χρησιμοποιείται με διάφορους τρόπους, όπως Virtual Assistant, Ανάλυση δεδομένων, λύσεις λογισμικού. Ο κύριος χρήστης είναι να μειώσει τα σφάλματα λόγω ανθρώπινης προκατάληψης.

- Αυτοματοποίηση:

Μηχανική μάθηση, η οποία λειτουργεί εντελώς αυτόνομα σε οποιοδήποτε τομέα χωρίς την ανάγκη για οποιαδήποτε ανθρώπινη παρέμβαση. Για παράδειγμα, τα ρομπότ που εκτελούν τα βασικά βήματα επεξεργασίας στις εγκαταστάσεις παραγωγής.

- Χρηματοοικονομική Βιομηχανία

Η μηχανική μάθηση αυξάνεται σε δημοτικότητα στη βιομηχανία χρηματοδότησης. Οι τράπεζες χρησιμοποιούν κυρίως ML για να βρουν μοτίβα μέσα στα δεδομένα, αλλά και για την πρόληψη της απάτης.

- Κυβερνητική οργάνωση

Η κυβέρνηση χρησιμοποιεί το ML για τη διαχείριση της δημόσιας ασφάλειας και των υπηρεσιών κοινής ωφέλειας. Πάρτε το παράδειγμα της Κίνας με την μαζική αναγνώριση προσώπου. Η κυβέρνηση χρησιμοποιεί Τεχνητή νοημοσύνη για να αποτρέψει το jaywalker.

- Βιομηχανία υγείας

Η υγειονομική περίθαλψη ήταν ένας από τους πρώτους κλάδους που χρησιμοποιούσαν μηχανική μάθηση με ανίχνευση εικόνων.

- Εμπορία

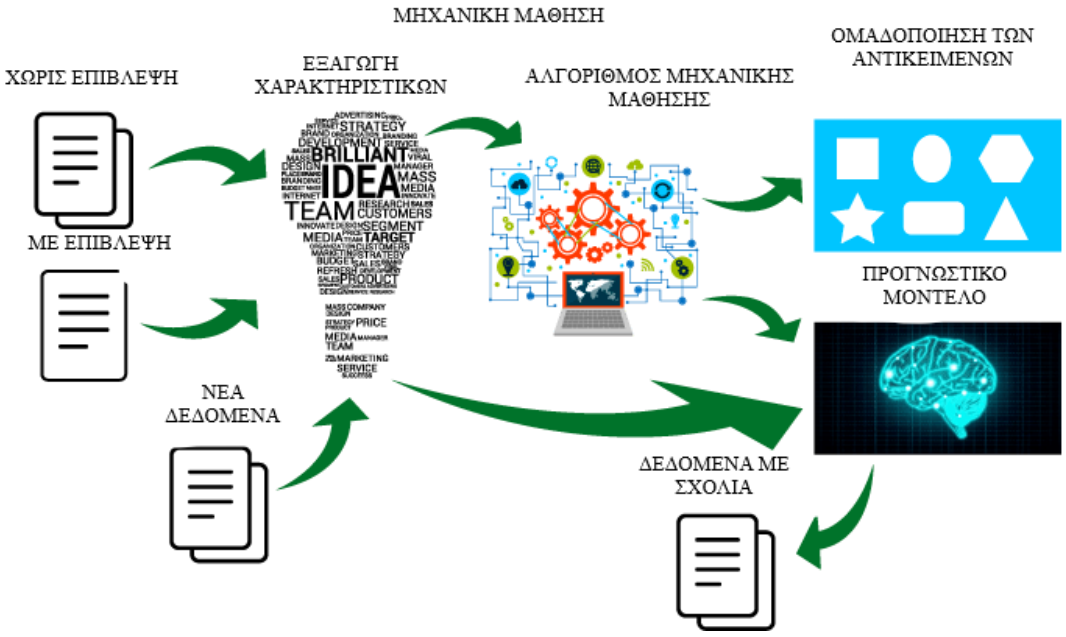
Η ευρεία χρήση της μηχανικής μάθησης γίνεται στο χώρο του μάρκετινγκ χάρη στην άφθονη πρόσβαση στα δεδομένα. Πριν από την ηλικία των μαζικών δεδομένων, οι ερευνητές αναπτύσσουν προηγμένα μαθηματικά εργαλεία όπως η Bayesian ανάλυση για την εκτίμηση της αξίας ενός πελάτη. Με την άνθηση των δεδομένων, το τμήμα μάρκετινγκ βασίζεται στη μηχανική μάθηση για τη βελτιστοποίηση της σχέσης πελατών και της εκστρατείας μάρκετινγκ [14].

Παράδειγμα εφαρμογής της μηχανικής μάθησης στην αλυσίδα εφοδιασμού ενός προιόντος

- Η μηχανική μάθηση παρέχει τεράστια αποτελέσματα για την αναγνώριση του οπτικού προτύπου, ανοίγοντας πολλές πιθανές εφαρμογές στη φυσική επιθεώρηση και συντήρηση σε ολόκληρο το δίκτυο της αλυσίδας εφοδιασμού.

- Η μη εποπτευόμενη μάθηση μπορεί γρήγορα να αναζητήσει συγκρίσιμα μοτίβα στο διαφορετικό σύνολο δεδομένων. Με τη σειρά του, το μηχάνημα μπορεί να πραγματοποιήσει έλεγχο ποιότητας σε όλο τον κόμβο εφοδιαστικής, αποστολή με ζημιά και φθορά.

- Ο διαχειριστής αποθεμάτων βασίζεται εκτενώς στην κύρια μέθοδο αξιολόγησης και πρόβλεψης του αποθέματος. Όταν συνδυάζονται μεγάλα δεδομένα και μηχανική μάθηση, έχουν εφαρμοστεί καλύτερες τεχνικές πρόβλεψης (βελτίωση κατά 20-30% σε σύγκριση με τα παραδοσιακά εργαλεία πρόβλεψης). Σε όρους πωλήσεων, αυτό σημαίνει αύξηση κατά 2 έως 3% λόγω της ενδεχόμενης μείωσης του κόστους αποθεμάτων.



Εικόνα 9 «. Machine Learning»

3.7 ΓΙΑΤΙ ΕΙΝΑΙ Η ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΣΗΜΑΝΤΙΚΗ

Η μηχανική μάθηση είναι το καλύτερο εργαλείο μέχρι στιγμής για την ανάλυση, κατανόηση και αναγνώριση ενός προτύπου στα δεδομένα. Μία από τις βασικές ιδέες πίσω από τη μηχανική μάθηση είναι ότι ο υπολογιστής μπορεί να εκπαιδευτεί για να αυτοματοποιήσει καθήκοντα που θα είναι εξαντλητικά ή αδύνατα για έναν άνθρωπο. Η ξεκάθαρη έλλειψη από την παραδοσιακή ανάλυση είναι ότι η μηχανική μάθηση μπορεί να λάβει αποφάσεις με ελάχιστη ανθρώπινη παρέμβαση.

Ας ακολουθήσουμε το ακόλουθο παράδειγμα. Ένας μεσήτης μπορεί να υπολογίσει την τιμή ενός σπιτιού με βάση τη δική του εμπειρία και τις γνώσεις του για την αγορά.Μια μηχανή μπορεί να εκπαιδευτεί για να μεταφράσει τη γνώση ενός εμπειρογνώμονα σε χαρακτηριστικά. Τα χαρακτηριστικά γνωρίσματα είναι όλα τα χαρακτηριστικά ενός σπιτιού, γειτονιάς, οικονομικού περιβάλλοντος κλπ. που κάνουν τη διαφορά τιμής. Για τον εμπειρογνώμονα, θα του πάρει αρκετό χρόνο για να εκτιμήσει την τιμής ενός σπιτιού. Η τεχνογνωσία του βελτιώνεται και βελτιώνεται μετά από κάθε πώληση.

Για το μηχάνημα, χρειάζονται εκατομμύρια δεδομένα, (δηλαδή, παράδειγμα) για να κυριαρχήσει αυτή η τέχνη. Στην αρχή της μάθησης, το μηχάνημα κάνει ένα λάθος, κάπως σαν τον κατώτερο πωλητή. Μόλις το μηχάνημα δει όλο το παράδειγμα, έχει αρκετές γνώσεις για να κάνει την εκτίμησή του. Ταυτόχρονα, με απίστευτη ακρίβεια. Το μηχάνημα μπορεί επίσης να προσαρμόσει το λάθος του ανάλογα [4].

4.1 ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ

Η εξόρυξη κειμένου, γνωστή και ως ανάλυση κειμένου, είναι η διαδικασία μετατροπής των μη δομημένων δεδομένων κειμένου σε χρήσιμες πληροφορίες και πληροφορίες. Η εξόρυξη κειμένου χρησιμοποιεί διαφορετικές τεχνολογίες τεχνητής νουμοσύνης για να επεξεργάζεται αυτόματα τα δεδομένα και να παράγει πολύτιμες πληροφορίες, επιτρέποντας στις εταιρείες να λαμβάνουν αποφάσεις βάσει δεδομένων. Για τις επιχειρήσεις, ο μεγάλος όγκος των δεδομένων που παράγονται καθημερινά αποτελεί ευκαιρία και πρόκληση. Από τη μια πλευρά, τα δεδομένα βοηθούν τις εταιρείες να αποκτήσουν έξυπνες γνώσεις σχετικά με τις απόψεις των ανθρώπων σχετικά με ένα προϊόν ή μια υπηρεσία. Χαρακτηριστικό παράδειγμα αποτελούν οι ιδέες που θα μπορούσαν να προκύψουν από την ανάλυση μηνυμάτων ηλεκτρονικού ταχυδρομείου, ανασκοπήσεις προϊόντων, αναρτήσεις κοινωνικών μέσων, σχόλια πελατών, εισιτήρια υποστήριξης κλπ. Από την άλλη πλευρά, υπάρχει το δίλημμα για τον τρόπο επεξεργασίας όλων αυτών των δεδομένων. Και αυτό είναι όπου η εξόρυξη κειμένου παίζει σημαντικό ρόλο. Όπως και τα περισσότερα πράγματα που σχετίζονται με την Επεξεργασία Φυσικής Γλώσσας (NLP), η εξόρυξη κειμένου μπορεί να ακούγεται σαν μια δύσκολη επεξεργασία. Αλλά η αλήθεια είναι ότι δεν χρειάζεται να είναι.

4.2 ΤΙ ΕΙΝΑΙ Η ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ

Η εξόρυξη κειμένου είναι μια αυτόματη διαδικασία που χρησιμοποιεί τη φυσική επεξεργασία γλώσσας για να εξαγάγει πολύτιμες πληροφορίες από αδόμητο κείμενο. Μετατρέποντας τα δεδομένα σε πληροφορίες που μπορούν να κατανοήσουν τα μηχανήματα, η εξόρυξη κειμένου αυτοματοποιεί τη διαδικασία ταξινόμησης κειμένων με βάση το συναίσθημα, το θέμα και την πρόθεση. Χάρη στην εξόρυξη κειμένου, για παράδειγμα, οι επιχειρήσεις είναι σε θέση να αναλύουν σύνθετα και μεγάλα σύνολα δεδομένων με έναν απλό, γρήγορο και αποτελεσματικό τρόπο. Ταυτόχρονα, οι εταιρείες εκμεταλλεύονται αυτό το ισχυρό εργαλείο για να μειώσουν ορισμένα από τα χειροκίνητα και επαναλαμβανόμενα καθήκοντά τους, εξοικονομώντας πολύτιμο χρόνο στις ομάδες τους και επιτρέποντας στους πράκτορες υποστήριξης πελατών να επικεντρωθούν σε αυτό που κάνουν καλύτερα. Ας υποθέσουμε ότι πρέπει να εξετάστεί η γνώμη για ένα συγκεκριμένο προιόν ή ένα γεγονός για να γίνει κατανοητό αν το κοινό στο οποίο απευθύνεται επικροτεί ή επικρίνει το συγκεκριμένο προιόν ή γεγονός. Ένας αλγόριθμος εξόρυξης κειμένου θα μπορούσε να αποβεί ιδιαίτερα χρήσιμος ώστε να εντοπίστούν τα πιο δημοφιλή θέματα που προκύπτουν στα σχόλια του κοινού και τον τρόπο με τον οποίο οι άνθρωποι αισθάνονται γι 'αυτά δηλαδή αν είναι τα σχόλια θετικά, αρνητικά ή ουδέτερα. Θα μπορούσαμε επίσης να μάθουμε τις κύριες λέξεις-κλειδιά που αναφέρονται από το κοινό σχετικά με ένα δεδομένο θέμα. Με λίγα λόγια, η εξόρυξη κειμένου βοηθά ώστε να αξιοποιήθούν στο έπακρο τα δεδομένα , γεγονός που οδηγεί σε καλύτερες επιχειρηματικές αποφάσεις που βασίζονται σε δεδομένα.

Σε αυτό το σημείο χρήσιμο θα ήταν να διευκρινιστεί πώς η εξόρυξη κειμένου πραγματοποιεί όλα τα παραπάνω. Η απάντηση βρίσκεται στην έννοια της μηχανικής μάθησης. Η μηχανική μάθηση είναι ένας τομέας που προέρχεται από την τεχνητή νουμοσύνη, η οποία επικεντρώνεται στη δημιουργία αλγορίθμων που επιτρέπουν στους υπολογιστές να μαθαίνουν εργασίες με βάση παραδείγματα. Τα μοντέλα μηχανικής μάθησης πρέπει να εκπαιδεύονται με δεδομένα, μέσα από τα οποία είναι σε θέση να

προβλέψουν με ένα συγκεκριμένο επίπεδο ακρίβειας αυτόματα. Όταν συνδυάζεται η εξόρυξη κειμένου και η μηχανική μάθηση, είναι δυνατή η αυτοματοποιημένη ανάλυση κειμένου. Για να υπαρξουν καλά επίπεδα ακρίβειας, θα πρέπει να τροφοδοτήθούν τα μοντέλα με ένα μεγάλο αριθμό παραδειγμάτων που είναι αντιπροσωπευτικά του προβλήματος που προσπαθούν να λυθούν. Εφόσον διευκρινήστηκε τι είναι η εξόρυξη κειμένου, απαραίτητο είναι να κατανοήσουμε πώς διαφοροποιείται από άλλους συνήθεις όρους, όπως η ανάλυση κειμένου και τα αναλυτικά κείμενα(text analytics).

4.3 ΔΙΑΦΟΡΑ ΑΝΑΜΕΣΑ ΣΤΗΝ ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ, ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΟΥ ΚΑΙ ΤΑ ΑΝΑΛΥΤΙΚΑ ΚΕΙΜΕΝΑ(TEXT ANALYTICS)

Η εξόρυξη κειμένου και η ανάλυση κειμένου χρησιμοποιούνται συχνά ως συνώνυμα. Ωστόσο, οι αναλύσεις κειμένων είναι μια ελαφρώς διαφορετική έννοια. Συνεπώς γεννάται το ερώτημα σχετικά με το ποια είναι η διαφορά μεταξύ εξόρυξης κειμένου και αναλυτικών στοιχείων κειμένου. Εν ολίγοις, και οι δύο σκοπεύουν να επιλύσουν το ίδιο πρόβλημα (αυτόματη ανάλυση ακατέργαστων δεδομένων κειμένου) χρησιμοποιώντας διαφορετικές τεχνικές. Η εξόρυξη κειμένου εντοπίζει σχετικές πληροφορίες μέσα σε ένα κείμενο και συνεπώς παρέχει ποιοτικά αποτελέσματα. Ωστόσο, οι αναλύσεις κειμένου επικεντρώνονται στην εύρεση προτύπων και τάσεων σε μεγάλα σύνολα δεδομένων, με αποτέλεσμα περισσότερα ποσοτικά αποτελέσματα. Τα αναλυτικά στοιχεία κειμένου χρησιμοποιούνται συνήθως για τη δημιουργία γραφημάτων, πινάκων και άλλων ειδών οπτικών αναφορών. Η εξόρυξη κειμένου συνδυάζει τις έννοιες της στατιστικής, της γλωσσολογίας και της μηχανικής μάθησης για τη δημιουργία μοντέλων που μαθαίνουν από τα δεδομένα κατάρτισης και μπορούν να προβλέψουν αποτελέσματα σε νέες πληροφορίες με βάση την προηγούμενη εμπειρία τους. Από την άλλη πλευρά, τα αναλυτικά στοιχεία κειμένου χρησιμοποιούν αποτελέσματα από αναλύσεις που εκτελούνται από μοντέλα εξόρυξης κειμένου, για τη δημιουργία γραφημάτων και όλων των ειδών οπτικοποιήσεις δεδομένων. Η επιλογή της σωστής προσέγγισης εξαρτάται από το είδος των πληροφοριών που είναι διαθέσιμες. Στις περισσότερες περιπτώσεις, οι δύο προσεγγίσεις συνδυάζονται για κάθε ανάλυση, οδηγώντας σε πιο συναρπαστικά αποτελέσματα.

Υπάρχουν διάφορες μέθοδοι και τεχνικές για την εξόρυξη κειμένου. Σε αυτή την ενότητα, θα καλύψουμε μερικές από τις πιο συχνές [5].

1. Βασικές μέθοδοι

- Συχνότητα λέξεων
  Η συχνότητα των λέξεων μπορεί να χρησιμοποιηθεί για να προσδιορίσει τους πιο επαναλαμβανόμενους όρους ή έννοιες σε ένα σύνολο δεδομένων. Η εύρεση των πιο αναφερθέντων λέξεων σε αδόμητο κείμενο μπορεί να είναι ιδιαίτερα χρήσιμη όταν αναλύονται για παράδειγμα οι αναθεωρήσεις πελατών, οι συνομιλίες των κοινωνικών μέσων ενημέρωσης ή τα σχόλια των πελατών. Για παράδειγμα, αν οι λέξεις ακριβές, υπερτιμημένες εμφανίζονται συχνά στις αναθεωρήσεις πελατών , ενδέχεται να υποδεικνύουν ότι πρέπει να προσαρμοστούν οι τιμές των προιόντων.
- Συνδυασμός λέξεων

Ο συνδυασμός λέξεων αναφέρεται σε μια ακολουθία λέξεων που εμφανίζονται συνήθως κοντά σε κάποια άλλη. Οι πιο συνηθισμένοι τύποι συνδυασμών είναι τα bigrams (ένα ζευγάρι λέξεων που είναι πιθανό να πάνε μαζί) και trigrams (ένας συνδυασμός τριών λέξεων). Ο προσδιορισμός των συνδυασμών - και η μέτρησή τους ως μία μόνο λέξη - βελτιώνει την λεπτομερειακότητα του κειμένου, επιτρέπει καλύτερη κατανόηση της σημασιολογικής του δομής και, τελικά, οδηγεί σε ακριβέστερα αποτελέσματα εξόρυξης κειμένου.

- Συμφωνία(concordance)
  Συμφωνία χρησιμοποιείται για να αναγνωρίσει το συγκεκριμένο πλαίσιο ή την περίπτωση στην οποία εμφανίζεται μια λέξη ή σύνολο λέξεων. Όλοι γνωρίζουμε ότι η ανθρώπινη γλώσσα μπορεί να είναι διφορούμενη καθώς μία ίδια λέξη μπορεί να χρησιμοποιηθεί σε πολλά διαφορετικά πλαίσια. Η ανάλυση της συναίνεσης μιας λέξης μπορεί να βοηθήσει στην κατανόηση του ακριβούς της νοήματος βάσει του πλαισίου.

2. Προηγμένες μέθοδοι

Ταξινόμηση κειμένου

Η ταξινόμηση κειμένου είναι η διαδικασία ανάθεσης κατηγοριών (ετικετών) σε μη δομημένα δεδομένα κειμένου. Αυτό το βασικό καθήκον της Επεξεργασίας Φυσικής Γλώσσας (NLP) διευκολύνει την οργάνωση και τη δομή σύνθετου κειμένου, μετατρέποντάς το σε σημαντικά δεδομένα. Χάρη στην ταξινόμηση κειμένου, για παράδειγμα οι επιχειρήσεις μπορούν να αναλύσουν κάθε είδους πληροφορίες, από τα μηνύματα ηλεκτρονικού ταχυδρομείου έως τα εισιτήρια υποστήριξης, και να αποκτήσουν πολύτιμες γνώσεις με γρήγορο και οικονομικό τρόπο. Παρακάτω, θα αναφερθούμε σε μερικά από τα πιο δημοφιλή καθήκοντα ταξινόμησης κειμένου - ανάλυση θέματος, ανάλυση συναισθημάτων, ανίχνευση γλώσσας και ανίχνευση προθέσεων [5].

- Ανάλυση θεμάτων:

  βοηθά στην καλύτερη κατανόηση των κύριων θεμάτων ή των θεμάτων ενός κειμένου και είναι ένας από τους κύριους τρόπους οργάνωσης των δεδομένων κειμένου. Για παράδειγμα, ένα κλεισμένο εισιτήριο που έχει τα σχόλια ότι η ηλεκτρονική κράτηση δεν έχει γίνει, μπορεί να χαρακτηριστεί και να καταχωρηθεί στα προβλήματα κρατήσεων.

- Ανάλυση συναισθημάτων:

  αποτελείται από την ανάλυση των συναισθημάτων που υποκρύπτουν κάθε δεδομένο κείμενο. Ας υποθέσουμε ότι αναλύουμε μια σειρά από κριτικές σχετικά με μία εφαρμογή για κινητά. Μπορεί να διαπιστωθεί ότι τα πιο συχνά αναφερόμενα θέματα σε αυτές τις αναθεωρήσεις είναι UI-UX ή Ευκολία Χρήσης, αλλά αυτό δεν παρέχει αρκετές πληροφορίες για να καταλήξουμε σε συμπεράσματα. Η ανάλυση συναισθημάτων μας βοηθά να κατανοήσουμε τη γνώμη και τα συναισθήματα σε ένα κείμενο και να τα κατατάξουμε ως θετικά,

αρνητικά ή ουδέτερα. Η ανάλυση των συναισθημάτων έχει πολλές χρήσιμες εφαρμογές στην επιχείρηση, από την ανάλυση των δημοσιεύσεων των κοινωνικών μέσων σε αναθεωρήσεις ή εισιτήρια. Από την άποψη της υποστήριξης πελατών, για παράδειγμα, ίσως μπορεί να εντοπιστούν γρήγορα οι θυμωμένοι πελάτες και να δώθεί προτεραιότητα στα προβλήματά τους πρώτα.

- Γλωσσική ανίχνευση:

  Μας επιτρέπει να ταξινομήσουμε ένα κείμενο με βάση τη γλώσσα του. Μία από τις πιο χρήσιμες εφαρμογές της είναι η αυτόματη δρομολόγηση των εισιτηρίων υποστήριξης στη σωστή γεωγραφικά τοποθετημένη ομάδα. Η αυτοματοποίηση αυτού του έργου είναι πολύ απλή και βοηθά τις ομάδες να εξοικονομούν πολύτιμο χρόνο.

- Ανίχνευση κινήσεων:

  θα μπορούσσαμε να χρησιμοποιήσουμε έναν ταξινομητή κειμένου για να αναγνωρίσουμε αυτόματα τις προθέσεις ή το σκοπό πίσω από ένα κείμενο. Αυτό μπορεί να είναι ιδιαίτερα χρήσιμο όταν για παράδειγμα αναλύονται συνομιλίες με πελάτες. Επίσης, θα μπορούσαμε να εξετάσουμε διαφορετικές απαντήσεις ηλεκτρονικού ταχυδρομείου πωλήσεων και να προσδιορίσουμε τις εταιρίες που ενδιαφέρονται για το προϊόν από εκείνες που δεν ενδιαφέρονται ή εκείνες που επιθυμούν να διαγραφούν.

Εξαγωγή κειμένου

Η εξαγωγή κειμένου είναι μια τεχνική ανάλυσης κειμένου που εξάγει συγκεκριμένα κομμάτια δεδομένων από ένα κείμενο, όπως λέξεις-κλειδιά, ονόματα οντοτήτων, διευθύνσεις, ηλεκτρονικά ταχυδρομεία κ.λπ. Χρησιμοποιώντας εξαγωγή κειμένου, μπορούν να αποφευχθούν τυχόν δυσκολίες στη διαλογή δεδομένων και πληροφορίων. Στις περισσότερες περιπτώσεις, μπορεί να είναι χρήσιμο να συνδυάσουμε την εξαγωγή κειμένου με την ταξινόμηση κειμένου στην ίδια ανάλυση. Παρακάτω θα αναφερθούμε σε μερικά από τα κύρια καθήκοντα της εξαγωγής κειμένου - την εξαγωγή λέξεων-κλειδιών, την αναγνώριση της οντότητας και την εξαγωγή χαρακτηριστικών [5].

- Εξαγωγή λέξεων-κλειδιών:

οι λέξεις-κλειδιά είναι οι πιο συναφείς όροι μέσα σε ένα κείμενο και μπορούν να χρησιμοποιηθούν για να συνοψίσουν το περιεχόμενό τους. Χρησιμοποιώντας ένα εργαλείο εξαγωγής λέξεων-κλειδιών, μπορούμε να ταξινομήσουμε τα δεδομένα προς αναζήτηση, να συνοψίσουμε το περιεχόμενο ενός κειμένου ή να δημιουργήσουμε σύννεφα ετικετών, μεταξύ άλλων.

- Αναγνώριση ονομαστικής οντότητας:

μας επιτρέπει να αναγνωρίζουμε και να εξάγουμε για παράδειγμα ονόματα εταιρειών, οργανισμών ή προσώπων από ένα κείμενο.

- Εξαγωγή χαρακτηριστικών:

βοηθά στην αναγνώριση συγκεκριμένων χαρακτηριστικών ενός προϊόντος ή μιας υπηρεσίας σε ένα σύνολο δεδομένων. Για παράδειγμα, εάν αναλύσουμε περιγραφές προϊόντων, θα μπορούσαμε εύκολα να εξαγάσουμε χαρακτηριστικά όπως χρώμα, μάρκα, μοντέλο κλπ.

4.4 ΓΙΑΤΙ ΕΙΝΑΙ ΣΗΜΑΝΤΙΚΗ Η ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ

Τα άτομα και οι οργανισμοί παράγουν τόνους δεδομένων κάθε μέρα. Τα στατιστικά στοιχεία υποστηρίζουν ότι σχεδόν το 80% των υπαρχόντων δεδομένων κειμένου δεν είναι δομημένα, δηλαδή δεν είναι οργανωμένα με προκαθορισμένο τρόπο, δεν είναι αναζητήσιμα και είναι σχεδόν αδύνατο να τα διαχειριστούμε. Με άλλα λόγια,η εξόρυξη κειμένου δεν είναι απλά χρήσιμη. Η ικανότητά της να οργανώνει, να ταξινομεί και να συλλαμβάνει σχετικές πληροφορίες από ακατέργαστα δεδομένα αποτελεί μείζονα ανησυχία και πρόκληση. Η εξόρυξη κειμένου είναι ζωτικής σημασίας για την αποστολή αυτή. Τα μη δομημένα δεδομένα κειμένου μπορούν να περιλαμβάνουν μηνύματα ηλεκτρονικού ταχυδρομείου, αναρτήσεις κοινωνικής δικτύωσης, συζητήσεις, εισιτήρια υποστήριξης, έρευνες κλπ. Η ταξινόμηση μέσω όλων αυτών των τύπων πληροφοριών με μη αυτόματο τρόπο οδηγεί συχνά σε αποτυχία. Όχι μόνο επειδή είναι χρονοβόρο και ακριβό, αλλά και επειδή είναι ανακριβές και αδύνατο να κλιμακωθεί. Η εξόρυξη κειμένου, ωστόσο, αποδείχθηκε αξιόπιστος και αποδοτικός ως προς το κόστος τρόπος επίτευξης της ακρίβειας, της επεκτασιμότητας και των χρόνων απόκρισης. Ακολουθούν μερικά από τα κύρια πλεονεκτήματά της:

Ευελιξία:

με εξόρυξη κειμένου είναι δυνατό να αναλυθούν μεγάλοι όγκοι δεδομένων σε λίγα μόνο δευτερόλεπτα. Με την αυτοματοποίηση συγκεκριμένων εργασιών, μπορούν να εξοικονομήθεί αρκετός χρόνος που μπορεί να χρησιμοποιηθε'ι για να επικεντρωθεί σε άλλα καθήκοντα.

Ανάλυση σε πραγματικό χρόνο:

χάρη στην εξόρυξη κειμένου, οι εταιρείες μπορούν να δώσουν προτεραιότητα σε επείγοντα θέματα ανάλογα, μεταξύ άλλων, εντοπίζοντας μια πιθανή κρίση και ανακαλύπτοντας ελαττώματα προϊόντων ή αρνητικές κριτικές σε πραγματικό χρόνο. Γιατί είναι τόσο σημαντικό αυτό; Επειδή για παράδειγμα μπορεί να επιτρέψει σε επιχειρήσεις να λάβουν γρήγορα μέτρα.

Συνεπές κριτήριο:

όταν εργάζονται σε επαναλαμβανόμενες, χειρωνακτικές εργασίες, οι άνθρωποι είναι πιο πιθανό να κάνουν λάθη. Επίσης, είναι δύσκολο να διατηρηθεί η συνοχή και να αναλυθούν δεδομένα υποκειμενικά. Ας πάρουμε, για παράδειγμα το tagging. Για τις περισσότερες

ομάδες, η προσθήκη κατηγοριών σε μηνύματα ηλεκτρονικού ταχυδρομείου ή εισιτήρια υποστήριξης είναι μια χρονοβόρα εργασία που συχνά οδηγεί σε σφάλματα και ασυνέπειες. Η αυτοματοποίηση αυτής της εργασίας όχι μόνο εξοικονομεί πολύτιμο χρόνο, αλλά και επιτρέπει ακριβέστερα αποτελέσματα και διασφαλίζει ότι εφαρμόζονται ενιαία κριτήρια σε κάθε εισιτήριο.

4.5 ΠΩΣ ΛΕΙΤΟΥΡΓΕΙ Η ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ

Η εξόρυξη κειμένου βοηθά στην ανάλυση μεγάλων ποσοτήτων πρωτογενών δεδομένων και στην εύρεση συναφών στοιχείων. Σε συνδυασμό με την εκμάθηση μηχανών, μπορεί να δημιουργήσει μοντέλα ανάλυσης κειμένου που μαθαίνουν να ταξινομούν ή να εξαγάγουν συγκεκριμένες πληροφορίες βάσει προηγούμενης εκπαίδευσης. Παρόλο που η εξόρυξη κειμένων μπορεί να μοιάζει με περίπλοκο ζήτημα, μπορεί να είναι αρκετά απλό να ξεκινήσουμε. Το πρώτο βήμα για να ξεκινήσουμε και να εργαστούμε με την εξόρυξη κειμένου είναι η συλλογή των δεδομένων μας. Ας υποθέσουμε ότι θέλουμε να αναλύσουμε τις συνομιλίες με χρήστες μέσω της ζωντανής συνομιλίας μιας εταιρείας.

Το πρώτο που θα πρέπει να κάνουμε είναι να δημιουργήσουμε ένα έγγραφο που περιέχει αυτά τα δεδομένα. Τα δεδομένα μπορούν να είναι εσωτερικά (αλληλεπιδράσεις μέσω συζητήσεων, μηνύματα ηλεκτρονικού ταχυδρομείου, έρευνες, υπολογιστικά φύλλα, βάσεις δεδομένων κ.λπ.) ή εξωτερικά (πληροφορίες από κοινωνικά μέσα, ιστοτόπους ανασκόπησης, ειδησεογραφικά πρακτορεία και οποιεσδήποτε άλλες ιστοσελίδες).

Το δεύτερο βήμα είναι η προετοιμασία των δεδομένων μας. Τα συστήματα εξόρυξης κειμένου χρησιμοποιούν διάφορες τεχνικές NLP - όπως τοξοκίνηση, ανάλυση, απομάκρυνση και τερματισμός της απομάκρυνσης - για την κατασκευή των εισροών του μοντέλου εκμάθησης μηχανών. Έπειτα θέση λαμβάνει η ίδια η ανάλυση κειμένου. Σε αυτή την ενότητα θα εξηγήσουμε πώς λειτουργούν οι δύο πιο κοινές μέθοδοι εξόρυξης κειμένου: ταξινόμηση κειμένου και εξαγωγή κειμένου [15].

Ταξινόμηση κειμένου

Η ταξινόμηση κειμένου είναι η διαδικασία εκχώρησης ετικετών ή κατηγοριών σε κείμενα, με βάση το περιεχόμενό τους. Χάρη στην αυτοματοποιημένη ταξινόμηση κειμένου, είναι δυνατή η ετικέτα ενός μεγάλου συνόλου δεδομένων κειμένου και η επίτευξη καλών αποτελεσμάτων σε πολύ σύντομο χρονικό διάστημα, χωρίς να απαιτούνται επίπονες και χρονοβόρες διαδικασίες. Αυτό έχει συναρπαστικές εφαρμογές σε διάφορους τομείς.

- Συστήματα βασισμένα σε κανόνες

Αυτοί οι τύποι συστημάτων ταξινόμησης κειμένου βασίζονται σε γλωσσικούς κανόνες. Με κανόνες, εννοούμε τους ανθρώπινους συνδυασμούς μεταξύ ενός συγκεκριμένου γλωσσικού προτύπου και μιας ετικέτας. Μόλις ο αλγόριθμος κωδικοποιηθεί με αυτούς τους κανόνες, μπορεί να ανιχνεύσει αυτόματα τις διαφορετικές γλωσσικές δομές και να αντιστοιχίσει τις αντίστοιχες ετικέτες. Οι κανόνες γενικά συνίστανται σε αναφορές σε συντακτικά, μορφολογικά και λεξικά πρότυπα. Μπορούν επίσης να σχετίζονται με σημασιολογικές ή φωνολογικές πτυχές.

Για παράδειγμα, αυτό θα μπορούσε να είναι ένας κανόνας για την ταξινόμηση των περιγραφών προϊόντων με βάση το χρώμα ενός προϊόντος:

(Μαύρο | Γκρι | Λευκό | Μπλε) → Χρώμα

Σε αυτή την περίπτωση, το σύστημα θα αντιστοιχίσει την ετικέτα COLOR κάθε φορά που ανιχνεύει κάποια από τις παραπάνω λέξεις. Τα συστήματα βασισμένα σε κανόνες είναι εύκολα κατανοητά, καθώς αναπτύσσονται και βελτιώνονται από τον άνθρωπο. Ωστόσο, η προσθήκη νέων κανόνων σε έναν αλγόριθμο συχνά απαιτεί πολλές δοκιμές για να διαπιστωθεί εάν θα επηρεάσουν τις προβλέψεις άλλων κανόνων, κάνοντας το σύστημα δύσκολο να κλιμακωθεί. Επιπλέον, η δημιουργία σύνθετων συστημάτων απαιτεί ειδικές γνώσεις σχετικά με τη γλωσσολογία και τα δεδομένα που θέλουμε να αναλύσουμε.

- Συστήματα που βασίζονται στη μηχανική μάθηση

Τα συστήματα ταξινόμησης κειμένου που βασίζονται στη μηχανική μάθηση μπορούν να μάθουν από προηγούμενα δεδομένα (παραδείγματα). Για να γίνει αυτό, πρέπει να εκπαιδεύονται με σχετικά παραδείγματα κειμένου - γνωστά ως δεδομένα κατάρτισης - τα οποία έχουν επισημανθεί σωστά. Τα δείγματα εκπαίδευσης πρέπει να είναι συνεπή και αντιπροσωπευτικά, έτσι ώστε το μοντέλο να μπορεί να κάνει ακριβείς προβλέψεις. Αλλά σε αυτό το σημείο προκύπτει το ερώτημα σχετικά με το πώς λειτουργεί ένας ταξινομητής κειμένου. Οι μηχανές πρέπει να μετατρέψουν τα δεδομένα εκπαίδευσης σε κάτι που μπορούν να καταλάβουν σε αυτή την περίπτωση, φορείς (μια συλλογή αριθμών με κωδικοποιημένα δεδομένα). Οι φορείς αντιπροσωπεύουν διαφορετικά χαρακτηριστικά των υφιστάμενων δεδομένων. Μια από τις πιο κοινές προσεγγίσεις για διάνυσμα ονομάζεται σακούλα λέξεων και συνίσταται στο να μετράτε πόσες φορές μια λέξη - από ένα προκαθορισμένο σύνολο λέξεων - εμφανίζεται στο κείμενο που θέλουμε να αναλύσουμε, Τα δεδομένα κειμένου μετασχηματισμένα σε φορείς, μαζί με τις αναμενόμενες προβλέψεις (ετικέτες), τροφοδοτούνται σε έναν αλγόριθμο μηχανικής μάθησης, δημιουργώντας ένα μοντέλο ταξινόμησης που αναλύθηκε διεξοδικά στην προηγούμενη ενότητα.

- Υβριδικά συστήματα

Τα υβριδικά συστήματα συνδυάζουν συστήματα βασισμένα σε κανόνες με μηχανικά συστήματα μάθησης. Συμπληρώνονται μεταξύ τους για να αυξήσουν την ακρίβεια των αποτελεσμάτων.

Εκτίμηση

Η απόδοση ενός ταξινομητή κειμένου μετράται με διαφορετικές παραμέτρους: ακρίβεια, ανάκληση και βαθμολογία(F1). Η κατανόηση αυτών των μετρήσεων θα μας επιτρέψει να δούμε πόσο καλό είναι το μοντέλο ταξινομητή μας στην ανάλυση κειμένων. Μπορούμε να αξιολογήσουμε τον ταξινομητή μας μέσω ενός σταθερού συνόλου δοκιμών - δηλαδή ενός συνόλου δεδομένων για το οποίο γνωρίζουμε ήδη τις αναμενόμενες ετικέτες - ή χρησιμοποιώντας τη διασταυρούμενη επικύρωση. Αυτή είναι μια διαδικασία που χωρίζει τα δεδομένα εκπαίδευσης μας σε δύο υποσύνολα: ένα μέρος των δεδομένων χρησιμοποιείται για την εκπαίδευση και το άλλο μέρος, για σκοπούς δοκιμής. Αυτή η ενότητα θα περάσει από τις διαφορετικές μετρήσεις για να αναλύσει την απόδοση του ταξινομητή κειμένου μας και θα εξηγήσει πώς λειτουργεί η διασταυρούμενη επικύρωση:

Η ακρίβεια υποδεικνύει τον αριθμό των σωστών προβλέψεων που έχει καταρτίσει ο ταξινομητής διαιρούμενος με τον συνολικό αριθμό των προβλέψεων. Ωστόσο, η ακρίβεια δεν είναι πάντα η καλύτερη μέτρηση για την αξιολόγηση της απόδοσης ενός ταξινομητή. Μερικές φορές, όταν οι κατηγορίες είναι ανισορροπημένες (αυτό σημαίνει ότι όταν υπάρχουν πολλά περισσότερα παραδείγματα για μια κατηγορία από ότι για άλλα), μπορεί να αντιμετωπίσετε ένα παράδοξο ακρίβειας: το μοντέλο είναι πιο πιθανό να κάνει μια καλή πρόβλεψη, καθώς τα περισσότερα δεδομένα ανήκουν σε ένα μόνο των κατηγοριών. Όταν συμβεί αυτό, είναι καλύτερα να εξετάσετε άλλες μετρήσεις όπως η ακρίβεια και η ανάκληση.

- Η ακρίβεια αξιολογεί τον αριθμό των σωστών προβλέψεων του ταξινομητή σχετικά με τον συνολικό αριθμό προβλέψεων για μια δεδομένη ετικέτα (συμπεριλαμβανομένων των σωστών ή των εσφαλμένων προβλέψεων). Μια μέτρηση υψηλής ακρίβειας δείχνει ότι υπήρχαν λιγότερα ψευδώς θετικά. Είναι σημαντικό να εξετάσουμε, ωστόσο, ότι η ακρίβεια μετρά μόνο τις περιπτώσεις όπου ο ταξινομητής προβλέπει ότι ένα κείμενο ανήκει σε μια συγκεκριμένη ετικέτα. Ορισμένες εργασίες, όπως οι αυτοματοποιημένες απαντήσεις μέσω ηλεκτρονικού ταχυδρομείου, απαιτούν μοντέλα με υψηλό επίπεδο ακρίβειας, ώστε να παρέχουν απάντηση σε ένα χρήστη μόνο όταν είναι πολύ πιθανό η πρόβλεψη να είναι σωστή.
- Η ανάκληση υποδεικνύει τον αριθμό των κειμένων που προβλέφθηκαν σωστά, πάνω από τον συνολικό αριθμό που θα έπρεπε να κατηγοριοποιηθεί με μια δεδομένη ετικέτα. Μια υψηλή μέτρηση ανάκλησης σημαίνει ότι υπήρχαν λιγότερα ψευδώς αρνητικά. Αυτή η μέτρηση είναι ιδιαίτερα χρήσιμη όταν χρειάζεται να δρομολογήσουμε εισιτήρια υποστήριξης στις σωστές ομάδες. Θέλουμε για παραδειγμα να δρομολογήσουμε αυτόματα όσο το δυνατόν περισσότερα εισιτήρια για μια συγκεκριμένη ετικέτα σε βάρος της απόκτησης εσφαλμένης πρόβλεψης στην πορεία.
- Η βαθμολογία F1 συνδυάζει τις παραμέτρους της ακρίβειας και της ανάκλησης για να μας δώσει μια ιδέα για το πόσο καλά λειτουργεί ο ταξινομητής μας. Αυτή η μέτρηση είναι ένας καλύτερος δείκτης από την ακρίβεια για να κατανοήσουμε πόσο καλές προβλέψεις είναι για όλες τις κατηγορίες του μοντέλου μας.

Διασταυρωμένη επικύρωση

Η διασταυρούμενη επικύρωση χρησιμοποιείται συχνά για τη μέτρηση της απόδοσης ενός ταξινομητή κειμένου. Αποτελείται από τη διαίρεση των δεδομένων κατάρτισης σε διαφορετικά υποσύνολα, με τυχαίο τρόπο. Για παράδειγμα, θα μπορούσαμε να έχουμε 4 υποσύνολα δεδομένων εκπαίδευσης, καθένα από τα οποία περιέχει το 25% των αρχικών δεδομένων. Στη συνέχεια, όλα τα υποσύνολα, εκτός από ένα, χρησιμοποιούνται για την κατάρτιση ενός ταξινομητή κειμένου. Αυτός ο ταξινομητής κειμένου χρησιμοποιείται για την πραγματοποίηση προβλέψεων πάνω από το υπόλοιπο υποσύνολο δεδομένων (δοκιμή). Μετά από αυτό, υπολογίζονται όλες οι μετρήσεις απόδοσης - συγκρίνοντας την πρόβλεψη με την πραγματική προκαθορισμένη ετικέτα - και η διαδικασία ξεκινά ξανά, έως ότου όλα τα υποσύνολα δεδομένων έχουν χρησιμοποιηθεί για έλεγχο. Το τελευταίο βήμα είναι η συγκέντρωση των αποτελεσμάτων όλων των υποσυνόλων δεδομένων για να επιτευχθεί μέση απόδοση κάθε μετρικής.

5.1 ΜΕΘΟΔΟΛΟΓΙΑ ΑΛΓΟΡΙΘΜΟΥ-ΕΙΣΑΓΩΓΗ

Η μεθοδολογία που εφαρμόστηκε στην παρούσα διπλωματική, εφαρμόστηκε σε ένα σύνολο δεδομένων twitter, που μεταδόθηκε από τις 15-03-2018 έως τις 24-03-2018, χρησιμοποιώντας το hashtag(#) "theresamay". Προφανώς, αυτός ο όρος αναφέρεται στην Theresa Mary May, πρωθυπουργό του Ηνωμένου Βασιλείου και ηγέτιδα του Συντηρητικού Κόμματος από το 2016. Αυτό το θέμα επιλέχτηκε για να εξυπηρετήσει τους σκοπούς της παρούσας διπλωματικής, διότι τα hashtags του twitter που σχετίζονται με πολιτική προσφέρουν μία αρκετά ικανοποιητική βάση για δοκιμές αλλαγών στο συναίσθημα. Η άποψη αυτή προκύπτει από το γεγονός ότι οι διάφορες απόψεις δεν είναι μονόπλευρες, διότι οι υποστηρικτές μπορεί να είναι και στις δύο πλευρές, ενώ οι κριτικοί μπορεί να εμπλέκονται επίσης.Συνεπώς είναι αρκετά δύσκολο να εξαχθεί ένα γενικό συναίσθημα γύρω από ένα πολιτικό πρόσωπο ή ζήτημα [7].

Ένας επιπλέον λόγος που επιλέχτηκε αυτό το tweet για ανάλυση, είναι ότι κατα τη διάρκεια αυτής της περιόδου το θέμα για την έξοδο της Βρετανίας από την ευρωπαική ένωση επέστησε την προσοχή ,λόγω περαιτέρω συζητήσεων μεταξύ πολιτικών υψηλού επιπέδου, σχετικά με τις σχέσεις μεταξύ του Ηνωμένου Βασίλειου και της Ευρωπαϊκής Ένωσης. Στο σύνολο δεδομένων ροής περιλαμβάνονται 15491 θέσεις(hashtags), με τον περιορισμό της απόρριψης μη αγγλικής γλώσσας στις δημοσιεύσεις.

Η γλώσσα επεξεργασίας του αλγορίθμου εξόρυξης γνώμης που επειλέχτηκε για την συγκεκριμένη διπλωματική είναι η R. Πρόκειται για μία γλώσσα που σχεδιάστηκε ειδικά για στατιστική ανάλυση, γεγονός που τη καθιστά ιδιαίτερα κατάλληλη για εφαρμογές επιστήμης δεδομένων. Αν και η διαδικασία μάθησης για προγραμματισμό με το R μπορεί να είναι σχετικά δύσκολη, ειδικά για άτομα χωρίς προηγούμενη εμπειρία προγραμματισμού, τα εργαλεία που είναι τώρα διαθέσιμα για την ανάλυση κειμένου στο R καθιστούν εύκολη την εκτέλεση ισχυρών αναλύσεων κειμένου αιχμής χρησιμοποιώντας μόνο μερικές απλές εντολές.

Ένα από τα κλειδιά για την εκρηκτική ανάπτυξη του R ήταν η πυκνοκατοικημένη συλλογή βιβλιοθηκών λογισμικού επέκτασης, γνωστή στην ορολογία R ως πακέτα, που παρέχονται και συντηρούνται από την εκτεταμένη κοινότητα χρηστών της R. Κάθε πακέτο επεκτείνει τη λειτουργικότητα της βασικής γλώσσας R και των βασικών πακέτων και εκτός από τις λειτουργίες και τα δεδομένα πρέπει να περιλαμβάνει τεκμηρίωση και παραδείγματα, συχνά με τη μορφή σημάτων που αποδεικνύουν τη χρήση του πακέτου. Η πιο γνωστή αποθήκη πακέτων, το Comprehensive R Archive Network (CRAN), έχει σήμερα πάνω από 10.000 πακέτα που δημοσιεύονται.

Ειδικότερα, η ανάλυση κειμένων έχει εδραιωθεί στο R. Υπάρχει μια τεράστια συλλογή ειδικών κειμένων για την επεξεργασία κειμένων και την ανάλυση κειμένων, από λειτουργίες χαμηλών επιπέδων έως προηγμένες τεχνικές μοντελοποίησης κειμένου, όπως η τοποθέτηση μοντέλων Latent Dirichlet Allocation, η R τα παρέχει όλα. Ένα από τα κύρια πλεονεκτήματα της εκτέλεσης της ανάλυσης κειμένου στο R είναι ότι είναι συχνά πιθανό και σχετικά εύκολο να γίνει εναλλαγή μεταξύ διαφορετικών πακέτων ή να τα συνδυαστούν. Οι πρόσφατες προσπάθειες μεταξύ της κοινότητας προγραμματιστών αναλύσεων κειμένων R έχουν σχεδιαστεί για να προωθήσουν αυτήν τη διαλειτουργικότητα για να μεγιστοποιήσουν την ευελιξία και την επιλογή μεταξύ των χρηστών.Ως αποτέλεσμα, η εκμάθηση των βασικών στοιχείων για την ανάλυση κειμένου στο R παρέχει πρόσβαση σε ένα ευρύ φάσμα προηγμένων χαρακτηριστικών ανάλυσης κειμένου.

Η πηγή των δεδομένων που χρησιμοποιήθηκε στην παρούσα πτυχιακή είναι το Twitter. Πρόκειται για ένα online εργαλείο microblogging που διαδίδει περισσότερα από 400 εκατομμύρια μηνύματα την ημέρα, συμπεριλαμβανομένων τεράστιων πληροφοριών σχετικά με όλες σχεδόν τις βιομηχανίες, από ψυχαγωγία έως αθλητισμό, υγεία σε επιχειρήσεις κλπ. Ένα από τα καλύτερα πράγματα όσο αναφορά το twitter είναι η προσβασιμότητά του. Είναι εύκολο στη χρήση τόσο για την ανταλλαγή πληροφοριών όσο και για τη συλλογή τους. Το Twitter παρέχει άνευ προηγουμένου πρόσβαση στην καθημερινότητα μας και στις προσωπικότητες μας, καθώς και στις ειδήσεις καθώς συμβαίνουν. Το Twitter επίσης αποτελεί σημαντική πηγή δεδομένων για τα επιχειρηματικά μοντέλα των τεράστιων εταιρειών.

Όλα τα παραπάνω χαρακτηριστικά κάνουν το twitter το καλύτερο μέρος για να συλλέξει κανείς σε πραγματικό χρόνο έως και τα τελευταία δεδομένα για να αναλύσει και να κάνει οποιαδήποτε έρευνα που αναζητά για πραγματικές καταστάσεις της ζωής.

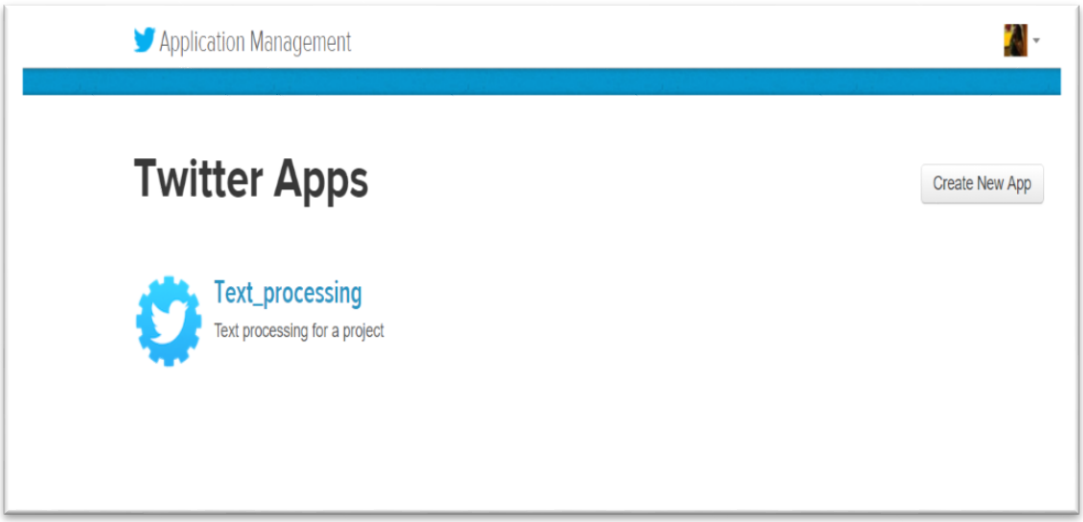
5.2 ΕΞΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ TWITTER

Αρχικά, απαραίτητο θα ήταν να αναφερθεί το πρόβλημα σχετικά με τα ζητήματα ανάκτησης δεδομένων από το twitter. Μετά την θέσπιση νομοθεσίας περι προσωπικών δεδομένων τον Μάιο του 2018 , η δυσκολία άντλησης δεδομένων από το twitter αυξήθηκε καθώς για να αποκτηθεί πρόσβαση στα δεδομένα του κάθε χρήστη ή σελίδας χρειάζεται άδεια από αυτόν. Για να υπερπηδηθεί η δυσκολία αυτή και στα πλαίσια της διπλωματικής έγινε χρήση προσωπικών σελίδων του twitter ώστε να γίνει η ανάλυση συναισθήματος που χρειάστηκε.

Η υλοποίηση των αλγορίθμων στην παρούσα διπλωματική έγινε χρησιμοποιώντας τη γλώσσα προγραμματισμού R. Πρόκειται για γλώσσα προγραμματισμού που χρησιμοποιείται διαρκώς για στατιστικούς υπολογισμούς, ανάλυση και εξόρυξη δεδομένων. Μερικές από τις εργασίες που εκτελεί άπτονται πολλαπλών προβλημάτων γραμμικής αλλα και μη γραμμικής μοντελοποίησης (li- 3.3 Προεπεξεργασία Δεδομένων 41 near & non linear modeling), συσταδοποίησης (clustering), περιγραφικής στατιστικής, κατηγοριοποίησης (classification), ανάλυσης χρονοσειρών κλπ.. Οι δυνατότητες που προσφέρει επεκτείνονται εκτεταμένα μέσω της χρήσης βιβλιοθηκών (libraries ή packages), που δημιουργούνται από προγραμματιστές- χρήστες και παρέχουν επιπλέον εργαλεία ανάλυσης. Η γλώσσα R είναι ανοιχτό λογισμικό αλλα και διαθέσιμο δωρεάν σε όλους τους χρήστες. Στην παρούσα διπλωματική εργασία, χρησιμοποιήσαμε την πλατφόρμα του R Studio για την πλήρη υλοποίηση του απαραίτητου κώδικα.

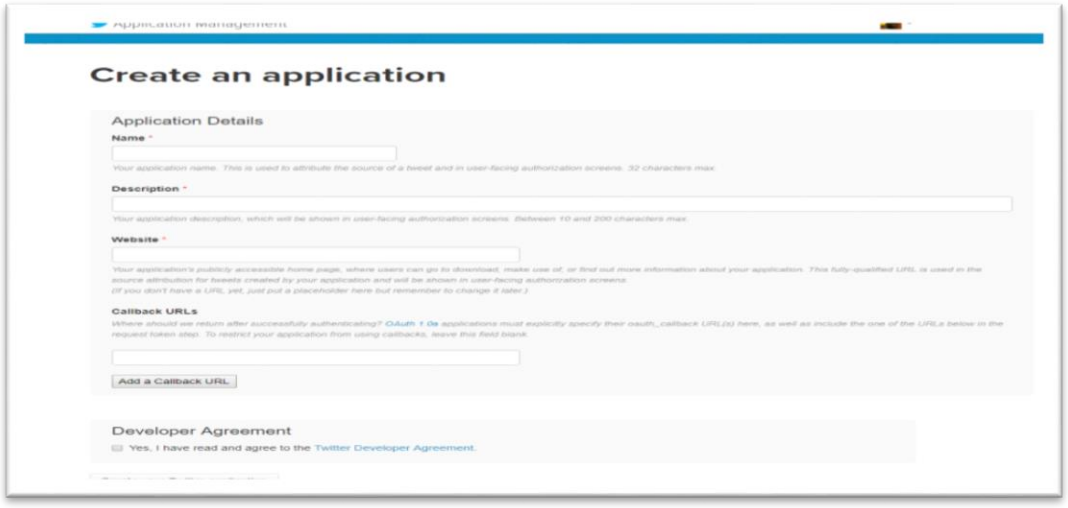
Για να εξαγάγουμε τα δεδομένα του twitter πρέπει να δημιουργήσουμε μια εφαρμογή twitter.Αυτό μπορεί να γίνει με τρία απλά βήματα που παρουσιάζονται στις εικόνες 11,12 και 13.

1.Δημιουργία Twitter App



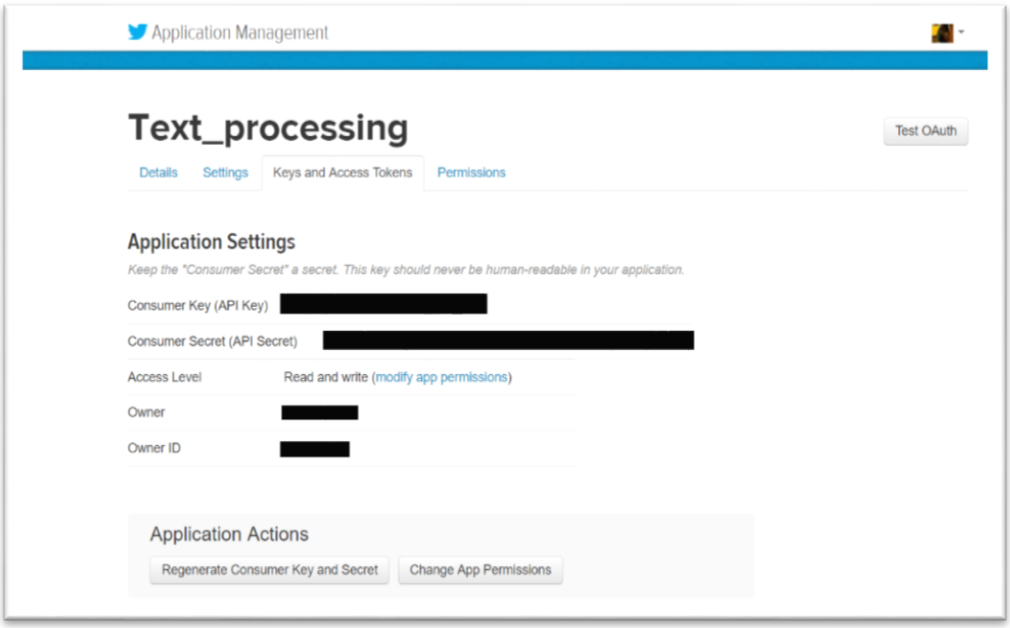
Εικόνα 10. «Δημιουργία Twitter App»

2. Κάνοντας κλίκ στο κουμπί “create new app” και συμπληρώνοντας τη φόρμα.



Εικόνα 11. «Δημιουργία της εφαρμογής»

3 Μόλις συμπληρωθούν όλες οι λεπτομέρειες και επαληθευτούν, θα έχουμε κλειδιά πελάτη και τα κλειδιά πρόσβασης.



Εικόνα 12. « Ρυθμίσεις και στοιχεία της εφαρμογής»

Το R έρχεται με ένα τυποποιημένο σύνολο πακέτων. Ένας αριθμός άλλων πακέτων είναι διαθέσιμοι για λήψη και εγκατάσταση. Για τους σκοπούς αυτής της διπλωματικής, θα χρειαστούμε τα ακόλουθα πακέτα:

- ROAuth: Παρέχει μια διεπαφή στην προδιαγραφή OAuth 1.0, επιτρέποντας στους χρήστες να επαληθεύουν μέσω OAuth τον εξυπηρετητή της επιλογής τους.
- rtweet: Παρέχει μια διεπαφή στο API του ιστού Twitter.

Με τη βοήθεια αυτών των πακέτων,συνδέεται ο λογαριασμός στο twitter στο R, για να εξαχθούν τα απαιτούμενα tweets εφόσον έχουμε συμπληρώσει τα απαραίτητα στοιχεία σχετικά με το twitter app που δημιουργήσαμε παραπάνω.Η μορφή του αρχείου που δημιουργείται από τα tweets που συλλέξαμε θα έχει την μορφή που παρουσιάζεται στην εικόνα 14:

	status_id	created_at	user_id	screen_name	text	source
17946	974168869718937601	2018-03-15 06:22:12	884358893782142976	sinamiladi	RT @EnglishAlwaght: Russia Vows Reaction to Britai...	Twitter Web Clier
17945	974169093296283648	2018-03-15 06:23:05	36063034	gibbonape	In fundraising speech, #Trump says he made up tra...	Twitter for iPad
17944	974169119040987141	2018-03-15 06:23:11	304441070	TuffeTu	RT @MoscowTimes: #TheresaMay says Britain will e...	Twitter for iPhone
17943	974169541612679168	2018-03-15 06:24:52	849811792645599233	searchworld10	RT @MoscowTimes: #TheresaMay says Britain will e...	Twitter Web Clier
17942	974169733548466176	2018-03-15 06:25:38	750357775423836161	tisy47	RT @JerryHicksUnite: Whatever #TheresaMay &...	Twitter Web Clier
17941	974169816801136640	2018-03-15 06:25:58	905252540	XboxWrld	#TheresaMay has announced 23 Russian diplomats ...	Tween
17940	974170185526636544	2018-03-15 06:27:26	330576967	ELMaracuXo	RT @RealNewsLine: The UK is expelling 23 Russian ...	Twitter for Andro
17936	974170346680184834	2018-03-15 06:28:04	2635376577	JaredThreepoint	RT @beingrichard: #Irish MSM doesn't buy it: "#The...	Twitter for iPhone
17935	974170394579144705	2018-03-15 06:28:15	752761530	VindobonNorikum	And #TheresaMay was in office as the Home Secret...	Twitter for Andro
17934	974170615161720832	2018-03-15 06:29:08	142842961	tonguelash	In support of #UK the #Irish #Republic will #Boyco...	Twitter for Andro
17933	974170962471079936	2018-03-15 06:30:31	845541935683158016	CoraCoralez	RT @cortezshine1: The Best Book Of The Year! https...	Twitter for Andro
17932	974171222299893760	2018-03-15 06:31:33	4509990922	CorkNCentral	RT @tonguelash: In support of #UK the #Irish #Rep...	Twitter for Andro
17931	974171457403179009	2018-03-15 06:32:29	3629331433	EnglishAlwaght	RT @EnglishAlwaght: Russia Vows Reaction to Britai...	Twitter Web Clier
17930	974171492513730560	2018-03-15 06:32:37	4511680467	CorkSouthC	RT @tonguelash: In support of #UK the #Irish #Rep...	Twitter for Andro
17929	974172437863481344	2018-03-15 06:36:23	821063666422267908	wasregan	RT @EnglishAlwaght: Russia Vows Reaction to Britai...	Twitter for iPad
17928	974172751408893953	2018-03-15 06:37:37	3997870336	PrincessBibiRF_	RT @EnglishAlwaght: Russia Vows Reaction to Britai...	Twitter for iPhone
17927	974172767154331648	2018-03-15 06:37:41	3358213042	ChrisKeelty	RT @ChrisKeelty: #Tory fundraiser see sporting his ...	Twitter Web Clier
17926	974172775815331840	2018-03-15 06:37:43	20063640	MysteryHunters	RT @MoscowTimes: #TheresaMay says Britain will e...	Twitter Web Clier

Εικόνα 13. «Παράδειγμα RDS αρχείου που δημιουργείται»

Στη συνέχεια αφού έχει πραγματοποιηθεί η εξαγωγή των tweets που θέλουμε, πρέπει να διαμορφωθούν έτσι ώστε να είναι σε μία πιο εύκολα διαχειρίσιμη μορφή από τον υπόλοιπο κώδικα.

5.3 ΕΞΟΡΥΞΗ ΚΕΙΜΕΝΟΥ ΚΑΙ ΚΑΤΑΣΚΕΥΗ WORDCLOUD

Οι μέθοδοι εξόρυξης κειμένου επιτρέπουν να εντοπίσουμε τις πιο συχνά χρησιμοποιούμενες λέξεις-κλειδιά σε ένα σύνολο δεδομένων. Ένας τρόπος που θα μπορούσε να γίνει αυτό είναι η δημιουργία ένος σύννεφου λέξεων(wordcloud), που αναφέρεται επίσης ως σύννεφο κειμένου ή σύννεφο ετικετών, το οποίο είναι μια οπτική αναπαράσταση δεδομένων κειμένου.

Η διαδικασία δημιουργίας σύννεφου λέξεων είναι πολύ απλή στο R αν κανείς γνωρίζει τα διάφορα βήματα που πρέπει να εκτελέσει. Το πακέτο εξόρυξης κειμένου (tm) και το πακέτο γεννήτρια σύννεφων λέξης (wordcloud) είναι διαθέσιμα στο R για να μας βοηθήσουν να αναλύσουμε κείμενα και να απεικονίσουμε γρήγορα τις λέξεις-κλειδιά ως σύννεφο λέξεων.

Οι λόγοι για τους οποίους επιλέχτηκε το σύννεφο λέξεων ως μέσο παρουσίασης των δεδομένων της παρούσας διπλωματικής είναι οι εξής [16] :

- Τα σύννεφα λέξης προσθέτουν απλότητα και σαφήνεια διότι οι πιο χρησιμοποιούμενες λέξεις-κλειδιά ξεχωρίζουν καλύτερα σε ένα σύννεφο λέξεων.
- Τα σύννεφα λέξης είναι ένα ισχυρό εργαλείο επικοινωνίας διότι είναι εύκολο να κατανοηθούν, να μοιραστούν και να επεξεργαστούν.
- Τα σύννεφα λέξεων είναι οπτικά πιο ελκυστικά από τα δεδομένα ενός απλού πίνακα

Σε αυτό το σημείο, θα ήταν χρήσιμο να τονιστεί η ευρεία χρήση των συννέφων λέξεων από ερευνητές για την αναφορά ποιοτικών δεδομένων, από εμπειρογνώμονες για την προβολή των αναγκών των πελατών, από εκπαιδευτικούς για την υποστήριξη και παρουσίαση βασικών ζητημάτων, απο πολιτικούς και δημοσιογράφους καθώς και από ιστότοπους κοινωνικών μέσων για τη συλλογή και ανάλυση των συναισθημάτων των χρηστών.

Αρχικά, για την κατασκευή του συννέφου λέξεων και ακολουθώντας τον κώδικα της παρούσας διπλωματικής, φορτώνουμε τις απαραίτητες βιβλιοθήκες tm και wordcloud. Η βιβλιοθήκη tm είναι ένα πλαίσιο για εφαρμογές εξόρυξης κειμένου μέσα στην R και αντίστοιχα η βιβλιοθήκη wordcloud αποτελεί ένα τρόπο για να δημιουργήσουμε σύννεφα λέξεων, να απεικονίσουμε τις διαφορές και την ομοιότητα μεταξύ των εγγράφων και να αποφύγουμε την υπερπαραγωγή διαγραμμάτων συχνοτήτων με κείμενο.

Ακολούθως, ως στόχο έχουμε την μετατροπή του συνόλου δεδομένων που συλλέξαμε, σε μορφή RDS, σε διάνυσμα χαρακτήρων που θα μπορέι να γίνει αποδεκτό από τις ακόλουθες συναρτήσεις του κώδικα που χρησιμοποιήθηκε για την κατασκευή του σύννεφου λέξεων. Σε αυτό το σημείο η αξιοποίηση της συνάρτησης VectorSource() θα μπορούσε να αποβεί ιδιαίτερη χρήσιμη.

5.3.1 ΚΑΘΑΡΙΣΜΟΣ ΚΕΙΜΕΝΟΥ ΑΠΟ ΠΕΡΙΤΤΑ ΣΤΟΙΧΕΙΑ

΄Οπως έχει ήδη προαναφερθεί, για την υλοποίηση της διπλωματικής εργασίας δημιουργήσαμε ένα δικό μας αρχείο από tweets που συλλέχθηκαν από την εφαρμογή που δημιουργήθηκε μέσω του Twitter. Η εφαρμογή δίνει τα tweets άθικτα, όπως έχουν δημοσιευτεί από το χρήστη, δίχως να έχουν υποστεί καποια επεξεργασία. Πριν την εφαρμογή οποιασδήποτε από τις τεχνικές εξόρυξης γνώμης απαραίτητο είναι να ¨καθαριστούν¨ τα μηνύματα, από κάθε περιττό ¨θόρυβο¨ ο οποίος δυσχαιρένει την κατηγοριοποίηση τους. ΄Οι συντομογραφίες, τα ορθογραφικά λάθη και οι ειδικοί χαρακτήρες απροπροσανατολίζουν τους αλγορίθμους και οδηγούν σε λανθασμένα συμπεράσματα. Η κανονικοποίηση αυτών των μηνυμάτων βελτιώνει τα αποτελέσματα της απόδοσης του αλγορίθμου, επιταχύνει ιεραρχική κατηγοριοποίηση των μηνυμάτων και δίνει καλύτερα αποτελέσματα. Συνεπώς η προεπεξεργασία αποτελεί θεμελιώδες βήμα μη παραλείψιμο. Επίσης απαραίτητο βήμα πριν τον καθαρισμό των tweets είναι η μετατροπή των δεδομένων σε lower case για την ομαλή διεξαγωγή του αλγορίθμου.

Συγκεκριμένα, τα βήματα καθαρισμού των tweets είναι [8] :

- Αφαίρεση υπερσυνδέσμων

Αφού τα μηνύματα έρθουν σε παρόμοια μορφή μετά αφαιρούνται οι υπερσύνδεσμοι. Οι υπερ-σύνδεσμοι ανακατευθύνουν σε περιεχόμενο άλλων ιστοσελίδων. Αν και πολλές φορές έχουν υπάρξει απόπειρες εξόρυξης γνώμης με χρήση τους, στην παρούσα διπλωματική εργασία δεν εξετάζεται η συμβολή τους στην πολικότητα των μηνυμάτων γιαυτό και αφαιρούνται.

- Αφαίρεση ειδικών χαρακτήρων

΄Εχει αναφερθεί ήδη η χρήση ειδικών χαρακτήρων από τους χρήστες των μέσων κοινωνικής δικτύωσης, οι οποίοι κατά πλειοψηφία αποτελούν θόρυβο για τα σύνολο δεδομένων. Οι αναφορές σε άλλους χρήστες (=Tag)(πχ @όνομαχρήστη), οι χαρακτήρες RT που συμβολίζουν αναδημοσίευση κειμένου από άλλον χρήστη αλλα και οι χαρακτήρες της δίεσης (=#,hashtag), οι οποίοι συνοδεύουν λέξεις κλειδιά στις οποίες αναφέρονται οι χρήστες προφανώς δεν προσφέρουν τίποτα στην κατηγοριοποίηση των tweets με βάση την πολικότητα τους αρα πρέπει πάντα να αφαιρούνται πριν την εφαρμογή κάθε αλγορίθμου. Αξίζει βέβαια να αναφερθεί πως στα hashtags, αφαιρείται μόνο ο χαρακτήρας δίεση (#) και όχι η λέξη που συνοδεύει, καθώς αποτελεί αναπόσπαστο στοιχείο στη θεματικής ταυτότητας του tweet.

- Αφαίρεση σημείων στίξης

΄Οπως συμβαίνει και στην χρήση των κεφαλαίων χαρακτήρων οι χρήστες χρησιμοποιούν επαναλαμβανόμενα σημεία στίξης ωστε να τονίσουν το περιεχόμενο του μηνύματος τους . Ακόμα και σε αυτή την περίπτωση, τα σημεία στίξης χρησιμεύουν μόνο στον εντοπισμό της έντασης του συναισθήματος αλλα όχι στον εντοπισμό του tweet αυτού . Συνεπώς, τα σημεία στίξης είναι περιττά και πρέπει να αφαιρούνται.

- Αφαίρεση αριθμών

Οι αριθμοί δεν παρέχουν επιπλέον πληροφορία και δεν βοηθούν καθόλου τη διαδικασία της εξόρυξης συναισθήματος, άρα αφαιρούνται και αυτοί από κάθε σύνολο δεδομένων.

- Αφαίρεση stop words

Τα stop words είναι οι συχνά χρησιμοποιούμενες λέξεις μίας γλώσσας, οπου το περιεχόμενο των οποίων δεν είναι φορτισμένο συναισθηματικά. Κάποια πραδείγματα τέτοιων λέξεων είναι οι σύνδεσμοι, οι προθέσεις, τα άρθρα, οι αντωνυμίες ή και ορισμένα ρήματα (π.χ. έχω, είμαι) και ουσιαστικά (π.χ. πληκτρολόγιο, καρέκλα, μολύβι κ.α.). Για τις πλέον διαδεδομένες λέξεις, όπως τα αγγλικά, οι γλώσσες προγραμματισμού, όπως η R που χρησιμοποιούμε, έχουν ενσωματωμένες έτοιμες λίστες stop words, στις οποίες ο κάθε χρήστης μπορεί να προσθέσει κατά βούληση επιπλέον και άλλες λέξεις. Όσον αφορά τα ελληνικά, κυκλοφορούν και στο διαδίκτυο διάφορες λίστες stop words οι οποίες είναι βασισμένες σε λεξικά της νέας ελληνικής. Όταν εργαζόμαστε με εφαρμογές εξόρυξης κειμένου, ακούμε συχνά τον όρο "stop words" ή "stop word list" ή ακόμα και "λίστα διακοπών". Οι λέξεις σταματήματος είναι βασικά ένα σύνολο λέξεων που χρησιμοποιούνται συνήθως σε οποιαδήποτε γλώσσα, όχι μόνο στα αγγλικά. Ο λόγος για τον οποίο οι λέξεις διακοπής είναι κρίσιμες για πολλές εφαρμογές είναι ότι, αν αφαιρέσουμε τις λέξεις που χρησιμοποιούνται πολύ συχνά σε μια δεδομένη γλώσσα, μπορούμε να επικεντρωθούμε στις σημαντικές λέξεις. Οι λέξεις σταματήματος θεωρούνται γενικά ότι είναι ένα "ενιαίο σύνολο λέξεων". Μπορεί πρακτικά να σημαίνει διαφορετικά πράγματα σε διαφορετικές εφαρμογές. Για παράδειγμα, σε ορισμένες εφαρμογές, η κατάργηση όλων των stop words όπως προσδιοριστές, προθέσεις (π.χ. παραπάνω, απέναντι, πριν) και ορισμένα επίθετα, σε ορισμένες εφαρμογές, μπορεί να είναι επιζήμιο. Για παράδειγμα, στην ανάλυση συναισθημάτων, η κατάργηση των επιθέτων όπως «καλό» και «ωραίο» καθώς και οι αρνητικές απαντήσεις, όπως «όχι», μπορούν να πετάξουν αλγορίθμους από τις διαδρομές τους. Σε τέτοιες περιπτώσεις, μπορεί κανείς να επιλέξει να χρησιμοποιήσει έναν ελάχιστο κατάλογο διακοπών που αποτελείται απλώς από προσδιοριστές ή προσδιοριστές με προθέσεις ή απλώς συντονιστικούς συνδέσμους ανάλογα με τις ανάγκες της εφαρμογής. Αξίζει να σημειωθεί ότι η πληροφοριακή αξία των "διακοπών" είναι σχεδόν μηδενική λόγω του γεγονότος ότι είναι τόσο συνηθισμένες σε μια γλώσσα. Η κατάργηση αυτού του είδους των λέξεων είναι χρήσιμη πριν από περαιτέρω αναλύσεις. Για τα "stopwords", οι υποστηριζόμενες γλώσσες είναι δανικά, ολλανδικά, αγγλικά, φινλανδικά, γαλλικά, γερμανικά, ουγγρικά, ιταλικά, νορβηγικά, πορτογαλικά, ρωσικά, ισπανικά και σουηδικά. Στα ονόματα γλωσσών γίνεται διάκριση πεζών-κεφαλαίων.

- Αφαίρεση κενού λευκού χώρου

Ο κενός λευκός χώρος ανάμεσα στις λέξεις που μπορεί να υπάρχει δεν δίνει καμία πληροφορία για την ανάλυση συναισθήματος, συνεπώς είναι σημαντικό να αφαιρεθεί.

- Αφαίρεση επιλεγμένων «ουδέτερων» λέξεων

Στον κώδικα που αναπτύχθηκε και χρησιμοποιήθηκε στα πλαίσια της παρούσας διπλωματικής,προστέθηκε και η δυνατότητα αφαίρεσης ορισμένων λέξεων που χαρακτηρίζονται ως ουδέτερες για την αγγλική γλώσσα όπως για παράδειγμα οι λέξεις "get","told","gave","took","get","can","said","asked","will","even","spoke","got","given","really". Η δυνατότητα αυτή, όπως και η αφαίρεση όλων των παραπάνω για τον καθαρισμό των tweets, έγινε μέσω της συνάρτησης tm_map().

5.3.2 ΔΗΜΙΟΥΡΓΙΑ ΠΙΝΑΚΑ ΟΡΩΝ-ΕΓΓΡΑΦΟΥ(TERM-DOCUMENT MATRIX)

Ο πίνακας εγγράφων είναι ένας πίνακας που περιέχει τη συχνότητα των λέξεων. Η δημιουργία του αποτελεί απαραίτητο βήμα για την κατασκευή του σύννεφου λέξεων καθώς το wordcloud αποτελεί έναν τρόπο εντοπισμού και εμφάνισης των πιο συχνών ή σημαντικών λέξεων ανάμεσα σε ένα σύνολο δεδομένων. Τα ονόματα των στηλών είναι λέξεις και τα ονόματα γραμμών είναι έγγραφα. Η συνάρτηση TermDocumentMatrix () από το πακέτο εξόρυξης κειμένου μπορεί να χρησιμοποιηθεί για την κατασκευή του συγκεκριμένου πίνακα. Για την παρούσα διπλωματική και το σύνολο δεδομένων που χρησιμοποιήθηκε ο πίνακας έχει την μορφή της εικόνας 15:

	word	freq
cna	cna	46344
theresamay	theresamay	32355
false	false	18578
twitter	twitter	14542
true	true	12745
amp	amp	8192
russia	russia	7619
brexit	brexit	5609
conservatives	conservatives	5067
euvoteleaverd	euvoteleaverd	4680
tory	tory	4544
android	android	4502
web	web	4492

Showing 1 to 14 of 32,685 entries, 2 total columns

Εικόνα 14. « Πίνακας ορών-εγγράφων για το σύνολο δεδομένων με tweet “Therresa May”»

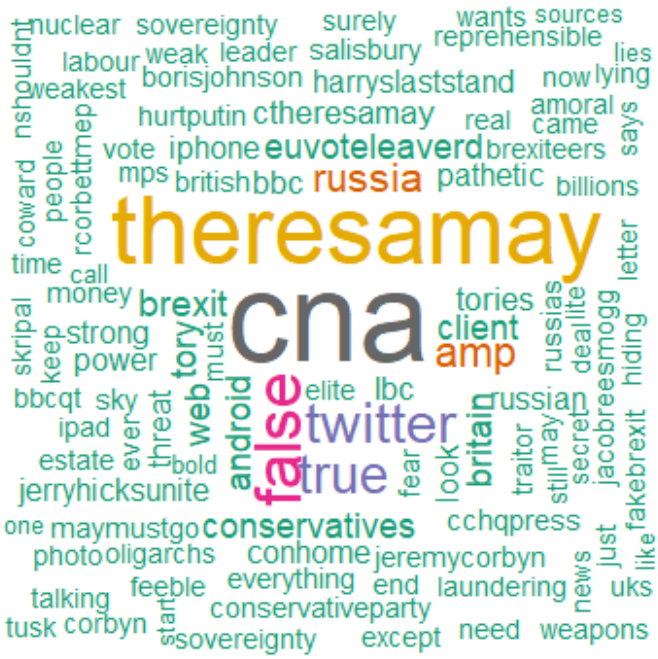
5.3.3 ΔΗΜΙΟΥΡΓΙΑ ΤΟΥ WORDCLOUD

Η συχνότητα των λέξεων μπορεί να απεικονιστεί ως σύννεφο λέξης με χρήση της συνάρτησης wordcloud() ,όπου με τον κατάλληλο ορισμό των παραμέτρων της μπορούμε να οπτικοποιήσουμε ποικίλα σύννεφα λέξεων. Ενδεικτικά οι παράμετροι της συνάρτησης μπορούν να αφορούν το πληθός των λέξεων που θα συμπεριληφθούν στο σύννεφο, τις αντίστοιχες συχνότητες εμφάνισής τους καθώς και την μικρότερη συχνότητα κάτω από την οποία δεν θα συμπεριλαμβάνονται οι αντίστοιχε λέξεις. Επίσης παρέχεται και η δυνατότητα μορφοποίησης του σύννεφου λέξεων όπως ή σειρά κατάταξης των λέξεων, ο χρωματισμός καθώς και η αναλογία λέξεων με περιστροφή τους ανα μοίρες. Ο κώδικας που αναπτύχθηκε για την δημιουργία του wordcloud στην παρούσα διπλωματική παρουσιάζεται στην εικόνα 16:

```
library(tm)
library(wordcloud)
testdat <- readRDS(file.choose())
corpus<- Corpus(VectorSource(testdat))
corpus[1][1]
#text cleaning
#convert the text to lower case
corpus<- tm_map(corpus,content_transformer(tolower))
#remove numbers
corpus<- tm_map(corpus,removeNumbers)
corpus<- tm_map(corpus,removeWords,stopwords("english"))
corpus<- tm_map(corpus,removePunctuation)
corpus<- tm_map(corpus,stripWhitespace)
corpus<-
tm_map(corpus,removeWords,c("get","told","gave","took","get","can","s
aid","asked",

"will","even","spoke","got","given","really"))

corpus[1][1]
tdm<- TermDocumentMatrix(corpus)
corpus
m<-as.matrix(tdm)
v<-sort(rowSums(m),decreasing=TRUE)
d<- data.frame(word=names(v),freq=v)
wordcloud(d$word,d$freq,
          random.order=FALSE,
          rot.per=0.3,scale=c(4,.8),
          max.words=500,
          colors=brewer.pal(8,"Dark2"))
```

Εικόνα 15: « Κώδικας σε γλώσσα R για την κατασκευή wordcloud»

Το αποτέλεσμα για το σύνολο δεδομένων με hashtag «Therresa May» απεικονίζεται στην εικόνα 17:



Εικόνα 16. «Κατασκευή συννέφου λέξεων με rot.per=0.3,scale=c(4,.8), max.words=200, colors=brewer.pal(8,"Dark2")»

Στην εικόνα 16 στην οποία παρουσιάζεται το σύννεφο λέξεων για την παρούσα διπλωματική, δείχνει ότι οι λέξεις «Theresa May», «cna» , «false», «twitter», και «true» είναι οι πέντε πιο συχνές λέξεις που αναφέρθηκαν.

5.4 ΕΞΟΡΥΞΗ ΚΑΙ ΟΠΤΙΚΟΠΟΙΗΣΗ ΣΥΝΑΙΣΘΗΜΑΤΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΤΟΥ ΛΕΞΙΚΟΥ NRC

Στην ταξινόμηση με χρήση λεξικού, προτεύοντα ρόλο διαδραματίζει το λεξικό συναισθήματος, καθώς χρησιμοποιείται για να ορίσει την πολικότητα των μηνυμάτων. Το γεγονός αυτό επιτυγχάνεται,με την μελέτη της έντασης συναισθήματος των λέξεων, που αποτελούν τα διάφορα παραδείγματα. Επιπλέον στη ταξινόμηση λεξικού και στη ταξινόμηση με αλγορίθμους μηχανικής μάθησης υπάρχει η δυνατότητα να γίνουν και με βάσει συντακτικών και στατιστικών τεχνικών. Για τις πρώτες τεχνικές μπορούμε να υπογραμμίσουμε πώς στηρίζονται αρκετά σε κανόνες συντακτικούς και είναι ικανές να εντοπίσουν μέρη του λόγου για εκάστοτε λέξεις. Οι τεχνικές που βασίζονται σε πιθανοθεωρίες,βρίσκουν με βάση αυτές τα δεδομένα που μας αφορούν.Τέλος,είναι αξιοσημείωτο το γεγονός ότι με χρήση λεξικού, η κατηγοριοποίηση αποτελεί μία εύκολη διαδικασία αφού δεν χρειάζεται εκπαίδευση των δεδομένων.

Υπάρχουν διάφορες μέθοδοι και λεξικά που χρησιμοποιούνται για την αξιολόγηση της γνώμης ή του συναισθήματος στο κείμενο. Το πακέτο tidytext παρέχει πρόσβαση σε αρκετά λεξικά συναισθημάτων. Τρία λεξικά γενικού σκοπού είναι:

- AFINN από την Finn Årup Nielsen,
- bing από τον Bing Liu και τους συνεργάτες,
- και nrc από τον Saif Mohammad και τον Peter Turney.

Και τα τρία αυτά λεξικά βασίζονται σε μονόγραμμα, δηλαδή σε μεμονωμένες λέξεις. Αυτά τα λεξικά περιέχουν πολλές αγγλικές λέξεις και οι λέξεις αποδίδουν βαθμολογίες για θετικό / αρνητικό συναίσθημα, και πιθανώς συναισθήματα όπως χαρά, οργή, θλίψη κ.ο.κ. Το λεξικό nrc κατηγοριοποιεί τις λέξεις με δυαδικό τρόπο ("ναι" / "όχι") σε κατηγορίες θετικών, αρνητικών, θυμού, αναμονής, αηδιασμού, φόβου, χαράς, θλίψης, έκπληξης και εμπιστοσύνης. Το bing lexicon κατηγοριοποιεί τις λέξεις με δυαδικό τρόπο σε θετικές και αρνητικές κατηγορίες. Το λεξικό AFINN εκχωρεί λέξεις με βαθμολογία που κυμαίνεται μεταξύ -5 και 5, με αρνητικές βαθμολογίες που δείχνουν αρνητικό συναίσθημα και θετικές βαθμολογίες που δείχνουν θετικό συναίσθημα. Στην παρούσα διπλωματική εργασία επιλέχτηκε η ανάλυση συναισθήματος των δεδομένων με τη βοήθεια του NRC λεξικού.

Οι βιβλιοθήκες που είναι απαραίτητες για αυτήν την διαδικασία είναι οι εξής:

- Το πακέτο plotrix έχει σκοπό να παρέχει μια μέθοδο για να πάρει πολλά είδη εξειδικευμένων σχεδιαγραμμάτων γρήγορα, αλλά επιτρέπει ταυτόχρονα και την εύκολη προσαρμογή αυτών των σχεδιαγραμμάτων χωρίς ιδιαίτερη εξειδικευμένη σύνταξη.
- Το ggplot2 είναι ένα σύστημα δηλωτικής δημιουργίας γραφικών, βασισμένο στη γραμματική γραφικών. Παρέχουμε τα απαιτούμενα δεδομένα σε αυτό, υποδεικνύουμε στο ggplot2 πώς να χαρτογραφήσει τις μεταβλητές, και οσο αναφορά το αισθητικό κομμάτι,καθορίζουμε ποια γραφικά πρωτότυπα πρέπει να χρησιμοποιήσει και αυτό φροντίζει τις λεπτομέρειες.

Σε αυτό το σημείο θα χρειαστεί να γίνει χρήση της ίδιας μεθόδου καθαρισμού των δεδομένων, όπως έγινε και για την κατασκευή του wordcloud, ώστε τα tweets που συλλέξαμε να έχουν την μορφή της εικόνας 17:

1		x
2		Russia Vows Reaction to Britain's Unprecedented Provocation Russia Britain Skripal
	1	TheresaM...
3	2	In fundraising speech Trump says he made up trade claim in meeting with Justin Trudeau revealing what everyone suspectedwhat KimJongUnour own TheresaMay are up against
4		TheresaMay says Britain will expelRussian diplomats and freeze Russian state assets in response to Skripal attack
	3	ht...
5		TheresaMay says Britain will expelRussian diplomats and freeze Russian state assets in response to Skripal attack
	4	ht...
6		Whatever TheresaMayBlairite Labour MPs say this is the JeremyCorbyn response to the SalisburyChemicalAttack
	5	•...
7	6	TheresaMay has announcedRussian diplomats will be expelled from the UK
8	7	The UK is expellingRussian diplomats as punishment over an ex spy's poisoningTheresaMay Skripal...
9	8	Irish MSM doesnt buy it TheresaMay's scenario that the Kremlin was directly involved seems unlikely For Vladimir...
10	9	And TheresaMay was in office as the Home SecretaryNo wonder she tries to switch a public attention to vicious Russians

Εικόνα 17. «Ενδεικτικό παράδειγμα του αρχείου που περιλαμβάνει τα tweets χωρίς περριτά στοιχεία»

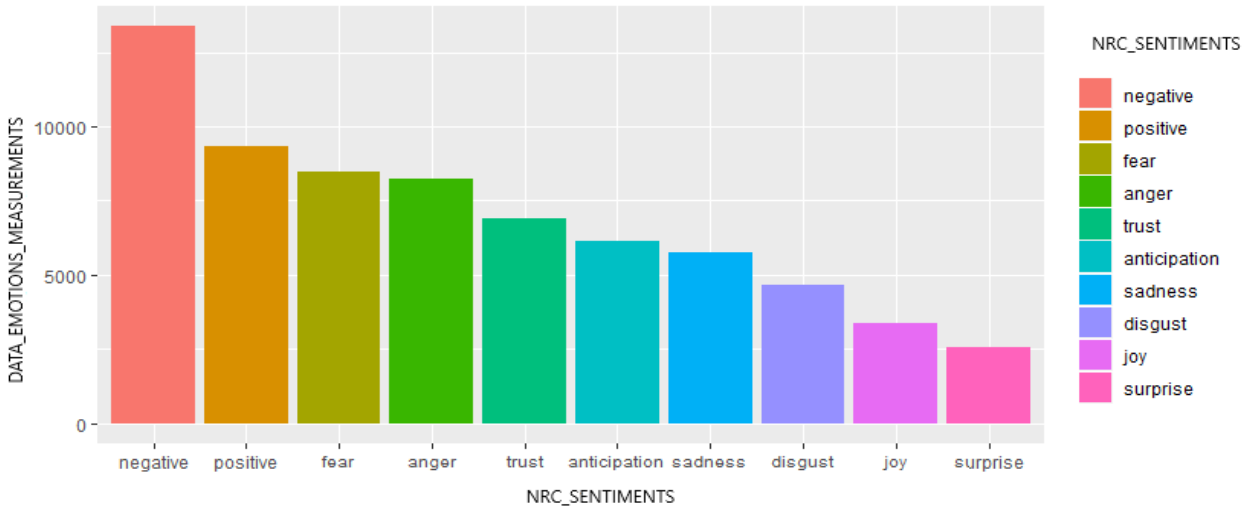
Ακολούθως, ,με τη βοήθεια της συνάρτησης get_nrc_sentiment() καλούμε το λεξικό συναισθημάτων NRC να υπολογίσει την παρουσία οκτώ διαφορετικών συναισθημάτων και το αντίστοιχο score τους για το αρχείο κειμένου με τα καθαρά από θόρυβο tweets, που προαναφέρθηκε.Έπειτα υπολογίζουμε το αντίστοιχο συνολικό score για το κάθε συναίσθημα ξεχωριστά.Τέλος, με τη βοήθεια κατάλληλων συναρτήσεων οπτικοποιούμε τα αποτελέσματά μας.

Προτεινόμενη επιλογή είναι η συνάρτηση ggplot που μπορεί να χρησιμοποιηθεί για να δηλώσει το πλαίσιο δεδομένων εισόδου για ένα γραφικό και να καθορίσει το σύνολο της αισθητικής γραφής που προορίζεται να είναι κοινό σε όλες τις επόμενες στρώσεις, εκτός εάν έχει ξεπεραστεί συγκεκριμένα. Ο κώδικας που εκπονήθηκε,στα πλαίσια της παρούσας διπλωματικής, για την εξόρυξη συναισθημάτων με τη χρήση λεξικού NRC παρουσιάζεται στην εικόνα 19:

```
library("plotrix")
library("plotly")
library("SnowballC")
library("ggplot2") # for graph
library("syuzhet")

tweets.df=readLines(file.choose())
head(tweets.df)
#getting emotions using in-built function
emotions <- get_nrc_sentiment(tweets.df)
emo_bar = colSums(emotions)
#calculating total score for each sentiment
emo_sum = data.frame(count_emo=emo_bar, emotion=names(emo_bar))
emo_sum$emotion = factor(emo_sum$emotion,
levels=emo_sum$emotion[order(emo_sum$count_emo,  decreasing = TRUE)])
#emotion.df2 <- cbind(tweets.df, emotion)
# Visualize the emotions from NRC sentiments

ggplot(emo_sum, aes(x=emo_sum$emotion,y=emo_sum$count_emo,
                    fill=emo_sum$emotion))+
  geom_col(position="stack" )+
  theme()
```

Εικόνα 18: «Κώδικας σε γλώσσα R για εξόρυξη συναισθημάτων με χρήση λεξικού NRC»

Το οπτικοποιημένο αποτέλεσμα για το σύνολο δεδομένων της παρούσας διπλωματικής παρουσιάζεται στην εικόνα 19:



Εικόνα 19. «Διάγραμμα ανάλυσης συναισθημάτων με χρήση NRC λεξικού»

Από το γράφημα στην εικόνα 19, φτάνουμε στο συμπέρασμα ότι οι απόψεις σχετικά με το tweet «Therresa May» είναι με αρκετά μεγάλη διαφορά αρνητικές, καθώς επίσης ότι και τα υπόλοιπα αρνητικά συναισθήματα είναι σε εξαιρετικά μεγάλα επίπεδα.

6.1 ΕΞΟΡΥΞΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΑΠΟ ΣΧΟΛΙΑ ΧΡΗΣΤΩΝ ΣΤΟ TWITTER ΣΕ ΠΡΑΓΜΑΤΙΚΟ ΧΡΟΝΟ

Σε αυτή την ενότητα θα γίνει μία διαφορετική προσέγγιση σχετικά με τη συναισθηματική ανάλυση των δεδομένων σε συνάρτηση με το χρόνο. Αφού γίνουν οι απαραίτητες ενέργειες μέσω του twitter API και πάρουμε τα δεδομένα που χρειαζόμαστε σε μορφη αρχείου RDS, επόμενο βήμα θα είναι να φορτώσουμε τις απαραίτητες βιβλιοθήκες και πακέτα που θα χρειαστούμε για αυτή την ανάλυση.

Σε αυτό το σημείο,απαραίτητο θα είναι να γίνει μία εκτενή αναφορά στους αλγόριθμους εντοπισμού αλλαγών σε χρονοσειρές, καθώς και να διευκρινιστεί ποιος αλγόριθμος χρησιμοποιήθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας.

Η ανίχνευση αλλαγών ή ανίχνευση σημείου αλλαγής προσπαθεί να εντοπίσει τα χρονικά σημεία όπου η κατανομή πιθανότητας μιας στοχαστικής διαδικασίας ή χρονοσειράς αλλάζει. Σε γενικές γραμμές το πρόβλημα αφορά τόσο την ανίχνευση του αν έχει συμβεί ή όχι, μια ή περισσότερες τέτοιες αλλαγές, και τον εντοπισμό των χρόνου εμφάνισης αυτών των αλλαγών. Συγκεκριμένες εφαρμογές, όπως step detection, ασχολούνται με τις αλλαγές στην μέση τιμή, την διακύμανση, την συσχέτιση, ή την φασματική πυκνότητα της διαδικασίας. Γενικότερα στην τεχνική αυτή ανοίκει και η ανίχνευση ανώμαλων συμπεριφορών.

1. <u>Step Detection αλγόριθμος</u>

Ο αλγόριθμος Step Detection (ανίχνευση βήματος) [γνωστός και ως step smoothing (λείανση βήματος), step filtering(φιλτράρισμα βήματος), shift detection (ανίχνευση μετατόπισης), jump detection (ανίχνευση πηδήματος) ή edge detection (ανίχνευση άκρης)] αποτελεί την διαδικασία εντοπισμού ξαφνικών αλλαγών (βημάτων, μετατοπίσεων, πηδημάτων) σε μια χρονοσειρά ή κάποιο σήμα. Συχνά, το βήμα αυτό είναι μικρό και οι χρονοσειρές έχουν αλλοιωθεί εξαιτίας κάποιου είδους θορύβου, και αυτό καθιστά το πρόβλημα δύσκολο, διότι το βήμα μπορεί να κρυφτεί λόγω του θορύβου. Το πρόβλημα ανίχνευσης βήματος εμφανίζεται σε μεγάλο αριθμό επιστημονικών και μηχανικών πλαισίων, μερικά από τα οποία αποτελούν την γενετική, την βιοφυσική και την επεξεργασία εικόνων. Οι περισσότεροι offline αλγόριθμοι για την ανίχνευση βήματος ψηφιακών δεδομένων μπορούν να κατηγοριοποιηθούν ως top-down, bottom-up, sliding window, ή global μέθοδοι.[9],[6].

- Top-down μέθοδος

Αυτοί οι αλγόριθμοι ξεκινούν με την παραδοχή ότι δεν υπάρχουν βήματα και εισάγουν πιθανά βήματα ένα την φορά, τα οποία ελέγχουν για να βρουν αυτό το οποίο ελαχιστοποιεί κάποια κριτήρια.

- Bottom-up μέθοδος

Εδώ οι αλγόριθμοι λειτουργούν με την ακριβώς αντίθετη λογική από τους topdown. Αρχικά θεωρούν ότι υπάρχει βήμα μεταξύ όλων των δειγμάτων ενός ψηφιακού σήματος, και στην συνέχεια συγχωνεύουν βήματα σύμφωνα με κάποια κριτήρια με τα οποία ελέγχονται όλα τα βήματα.

- Sliding window μέθοδος

Σε αυτή την μέθοδο εξετάζεται το σήμα μέσω ενός παραθύρου. Οι αλγόριθμοι ψάχνουν για στοιχεία βήματος που εμφανίζονται εντός του παραθύρου, το οποίο μετά από κάθε ολοκληρωμένο έλεγχο μετατοπίζεται(σείρεται) κατα μία θέση κάθε φορά, όπου και

ξαναγίνεται έλεγχος μέσω αυτού. Την μέθοδο αυτή ακολουθούν αρκετά φίλτρα που έχουν στόχο την μείωση του θορύβου διατηρώντας όμως τα απότομα βήματα του σήματος.

- Global μέθοδος

Οι αλγόριθμοι αυτοί εξετάζουν το σήμα με μια κίνηση και προσπαθούν να εντοπίσουν βήματα στο σήμα με κάποιο είδος διαδικασίας βελτιστοποίησης.

2. Online αλγόριθμοι ανίχνευσης αλλαγών

Όταν η ανίχνευση βήματος πρέπει να εκτελείται με το που φτάνουν τα δεδομένα, τότε βρισκόμαστε στην περίπτωση της σειριακής ανάλυσης. Στην σειριακή ανάλυση, κάθε αλγόριθμος παρουσιάζει 3 αρνητικά χαρακτηριστηκά αντιστρόφως ανάλογα μεταξύ τους [9],[6] :

• Ποσοστό αρνητικών συναγερμών(false alarm)

• Ποσοστό αποτυχίας εντοπισμού αλλαγής(misdetection

• Καθυστέρηση ανίχνευσης

Ο κάθε αλγόριθμος στοχεύει στο να ελαχιστοποιήσει αυτά τα χαρακτηρηστικα χωρίς όμως να μπορεί να εξαλείψει και τα 3. Οι περισσότεροι online αλγόριθμοι ανίχνευσης αλλαγών είναι και ταυτόχρονα αλγόριθμοι συσταδοποίησης, καθώς κάθε αλγόριθμος συσταδοποίησης μπορεί να τροποποιηθεί για ανίχνευση αλλαγών.

Ωστόσω υπάρχουν και αυθεντικοί αλγόριθμοι online ανίχνευσης αλλαγών, με πιο δημοφιλή τον CuSum. Επίσης πολύ συχνα χρησιμοποιούνται παραλλαγές του CuSum, ο οποίος τροποποιείται για να καλύψει τις ανάγκες του κάθε χρήστη. Μια τέτοια παραλλαγή υλοποιήθηκε στα πλαίσια αυτής της εργασίας και αναλύεται παρακάτω.

3. Cumulative Sum

Ο αλγόριθμος Cumulative Sum ή αλλιώς Cusum είναι ένας σειριακός αλγόριθμος ελέγχου αλλαγής κατάστασης, και συγκεκριμένα ο ποιο γνωστός του είδους. Όπως λέει και το όνομά του, ο Cusum υπολογίζει το συσωρευτικό άθροισμα των στοιχείων μιας διαδικασίας και αυτό είναι που τον κάνει σειριακό.

Τα στοιχεία x_n αναθέτονται βάρη ω_n και αθροίζονται ως εξής [9],[6] :

- $S_0=0$
- $S_{n+1}=\max(0, Sn + x_n - \omega_n)$

Όταν η τιμή του S ξεπεράσει ένα κατώφλι h το οποίο έχει τεθεί από τον χρήστη, εντοπίζεται αλλαγή. Σε περίπτωση που θέλουμε να εξετάσουμε και για αρνητικές τιμές, θα πρέπει να υπολογιστεί και ένα δεύτερο άθροισμα:

$$S^2_{n+1}=\min(0, Sn - x_n + \omega_n)$$

Όπου το κατώφλι εδώ θα έχει αρνητική τιμή. Τα βάρη ω_n αναθέτονται από τον χρήστη και αποτελούν την κανονική κατάσταση την οποία εξετάζουμε.

4. Προσαρμοστικός Cumulative Sum

Ο κλασικός αλγόριθμος Cusum έχει το μειονέκτημα ότι δεν μπορεί να προσαρμοστεί σε αλλαγές. Ένα δείγμα μπορεί πολυ εύκολα να έχει καταστάσεις(συστάδες) με διαφορετική διακύμανση. Αυτό σημαίνει ότι χρειάζεται διαφορετικό βάρος ω για τον έλεγχο της κάθε κατάστασης, καθώς εαν το βάρος είναι κατάλληλο για την συστάδα με την μικρότερη διακύμανση, τότε θα παρουσιάζεται αρνητικός συναγερμός, ενώ αν είναι κατάλληλο για την άλλη συστάδα τότε παρουσιάζει αποτυχία στον εντοπισμό κάποιων αλλαγών.

Για αυτό τον λόγο υλοποιήθηκε αλγόριθμος με μεταβλητό ω_n. Για την υλοποίησή του χρησιμοποιήθηκε η μέθοδος του σειρόμενου παραθύρου (Sliding window). Δεν δίνεται από τον χρήστη η αρχική κατάσταση, αλλά, μέσω μιας αρχικής φάσης εκπαίδευσης, ο αλγόριθμος υπολογίζει το ω_n ως την μέση τιμή των τελευτέων ν στοιχείων, με ν σταθερός αριθμός που δίνεται στην αρχή. Στην συνέχεια το συσωρευτικό άθροισμα υπολογίζεται με τον ίδιο ακριβώς τρόπο και με κάθε νέο στοιχείο, το ω υπολογίζεται εκ νέου. Με αυτόν τον τρόπο, εάν παρατήρηθει συνεχής απόκλιση από μια «σωστή» κατάσταση, ο αλγόριθμος θα αρχίσει να προσαρμόζεται σε αυτή τη νέα κατασταση, θεωρώντας πλέον αυτή ως σωστή. Ωστόσω για τις ανάγκες μερικών συνόλων δεδομένων μπορεί να χρειαστεί να επηρεάσουμε την αυστηρότητα του αλγορίθμου. Αυτό μπορούμε να το κάνουμε, εκτός με το να μεταβάλουμε το κατώφλι h, αυξάνοντας ή μειώνοντας το βάρος ω κάθε φορά που υπολογίζεται εκ νέου [9],[6].

5. Shewhart Controller

Ένας ακόμη αρκετά γνωστός αλγόριθμος ανίχνευσης αλλαγών είναι ο Shewhart Controller, του οποίου όμως η λειτουργία είναι εντελώς διαφορετική από τον CuSum. Ο CuSum ανιχνεύει απότομες αλλαγές(βήματα) λαμβάνοντας υπόψη μόνο τα στοιχεία που εμφανίστηκαν μετά το τελευταίο βήμα, δηλαδή ελέγχει αν υπάρχει απόκλιση από την τωρινή κατάσταση (συσταδα). Αντίθετα ο Shewhart Controller λαμβάνει υπόψη ολόκληρο το σύνολο δεδομένων, από το πρώτο στοιχείο του μέχρι το τελευταίο. Στον Shewhart Controller η κανονικότητα ενός στοιχείου X_n καθορίζεται από δύο όρια: το Άνω Όριο Ελέγχου (Upper Control Limit/UCL) και το Κάτω Όριο Ελέγχου (Lower Control Limit/LCL). Αυτά τα όρια ελέγχου υπολογίζονται ως εξής [9],[6] :

- $UCL = \mathbf{x}_n + a*\sigma_n$
- $LCL = \mathbf{x}_n - a*\sigma_n$

Όπου $\mathbf{x}_n$ η μέση τιμή του συνόλου δεδομένων, σn η τυπική απόκλιση και a μία σταθερά που καθορίζεται από τον χρήστη (συνήθως 2 ή 3).

Κάθε φορά που εμφανίζεται καινούργιο στοιχείο xn, ο αλγόριθμος ελέγχει αν ξεπερνάει ένα από τα δύο όρια. Αν $x_n > UCL_n$ τότε ο αλγόριθμος επιστρέφει 1, αν $x_n < UCL_n$ επιστρέφει -1, αλλιώς επιστρέφει 0, το οποίο σημαίνει ότι δεν υπάρχει αλλαγή. Τέλος υπολογίζεται η καινούργια μέση τιμή και διακύμανση.

Σύμφωνα με τον κώδικα της παρούσας διπλωματικής, θα χρειαστούν οι βιβλιοθήκες lubridate() και nabor() για να μπορέσουμε να ταξινομήσουμε τα tweets με βάση το χρόνο, διότι όταν εξάγουμε τα tweets από το twitter δεν θα είναι ταξινομημένα χρονικά.

Η βιβλιοθήκη lubridate() χρησιμοποιείται γενικά για την εργασία με ημερομηνίες και χρονικές περιόδους. Μπορεί να χαρακτηριστεί ως ένα αρκετά γρήγορο και χρήσιμο εργαλείο για ανάλυση δεδομένων ημερομηνίας και ώρας όπως εξαγωγή και ενημέρωση στοιχείων ημερομηνίας και ώρας (έτη, μήνες, ημέρες, ώρες, λεπτά και δευτερόλεπτα),και αλγεβρικό χειρισμό σε αντικείμενα ημερομηνίας και χρόνου. Το πακέτο «lubridate» έχει

μια συνεπή και εύκολη στην απομνημόνευση σύνταξη που κάνει την εργασία με τις ημερομηνίες εύκολη και ευέλικτη [17].

Όσο αναφορά το πακέτο nabor() περιβάλλει τη βιβλιοθήκη libnabo, μια γρήγορη βιβλιοθήκη K Nearest Neighbor για χώρους χαμηλού επιπέδου που έχουν γραφτεί σε C ++. Το πακέτο παρέχει και μια αυτόνομη λειτουργία για βασικά ερωτήματα κατά μήκος μιας επιλογής για την παραγωγή ενός αντικειμένου που περιέχει τη δομή αναζήτησης k-d όταν κάνετε πολλαπλά ερωτήματα στα ίδια σημεία στόχου [17].

Στη συνέχεια,πραγματοποιούμε ταξινόμηση των tweets με το χρόνο με κατάλληλες ενέργειες όπως:

- Με τη χρήση της συνάρτησης ymd_hms(), η οποία αναλύει ημερομηνίες που περιέχουν στοιχεία ωρών, λεπτών ή δευτερολέπτων
- Με την συνάρτηση as.numeric() η οποία μετατρέπει το όρισμά της σε αριθμητικό τύπο (είτε ακέραιος είτε πραγματικός) και επιστρέφει TRUE αν το όρισμα του είναι τύπου real ή type integer και FALSE διαφορετικά [17].
- και με την συνάρτηση order() η οποία επιστρέφει μια μεταβλητή που αναδιατάσσει το πρώτο στοιχείο της σε αύξουσα ή φθίνουσα σειρά, σπάζοντας τους δεσμούς με περαιτέρω στοιχεία.

Το τμήμα του κώδικα που εκτελεί τις ενέργειες που περιγράφησαν παρουσιάζεται στην εικόνα 20:

```
testdat <- readRDS(file.choose())


# sort tweets with time
library(lubridate)
library(nabor)
testdat$created_at<-ymd_hms(testdat$created_at)
testdat$timesort<-as.numeric(testdat$created_at)
testdat <- testdat[order(testdat$timesort),]
```

Εικόνα 20: «Τμήμα του κώδικα σε γλώσσα R για την εξόρυξη συναισθήματος»

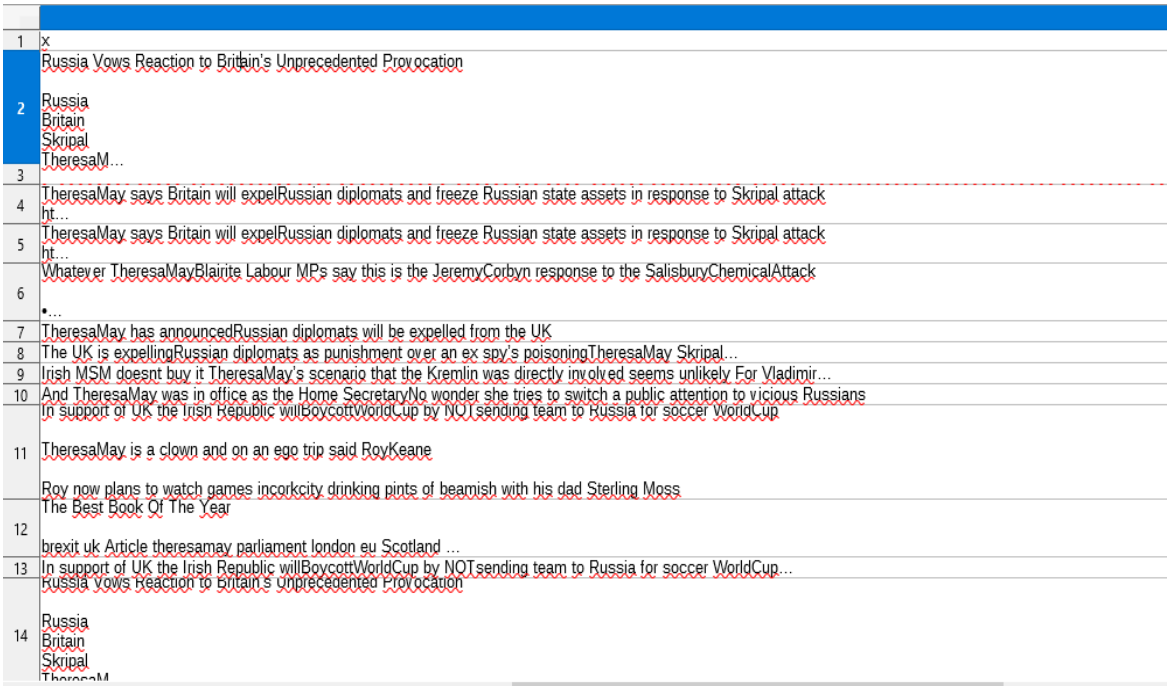
Μετά το πέρας της παραπάνω ταξινόμησης,στόχος είναι όπως και στις προηγούμενες ενότητες να καθαρίσουμε τα tweets από περριτά στοιχεία όπως τα μη αγγλικά tweets,διότι όπως αναφέρθηκε και στην προηγούμενη ενότητα η εξορυξη γνώμης που θα πραγματοποιηθεί στη συγκεκριμένη εργασία δεν θα περιλαμβλάνει tweets σε άλλες γλώσσες πέραν της αγγλικής. Σημαντικό βήμα έπειτα θα είναι να βρεθεί ο αριθμός των retweet, των tweet δηλαδή που έχουν αναδημοσιευτεί έτσι ώστε να μην συμπεριληφθούν στην συναισθηματική ανάλυση που θα πραγματοποιηθέι απο τον αλγόριθμο. Σύμβολα, σημεία στίξης και αναφορές σε συνδέσμους επίσης θα πρέπει με τις κατάλληλες εντολές να αφαιρεθούν.

Για να γίνει ο καθαρισμός των tweets θα πρέπει φυσικά να φορτωθούν και οι κατάλληλες βιβλιοθήκες:

- Η βιβλιοθήκη tidyverse() είναι ένα συνεκτικό σύστημα πακέτων για χειρισμό, εξερεύνηση και οπτικοποίηση δεδομένων που μοιράζονται μια κοινή φιλοσοφία σχεδίασης. Τα πακέτα Tidyverse αποσκοπούν στην αύξηση της παραγωγικότητας των στατιστικών και των επιστημόνων δεδομένων, καθοδηγώντας τους μέσω ροών εργασίας που διευκολύνουν την επικοινωνία και οδηγούν σε αναπαραγωγικά προϊόντα εργασίας [17].
- Η βιβλιοθήκη tidytext() εφαρμόζει αρχές τακτοποιημένων δεδομένων για να διευκολύνει την εξόρυξη κειμένου και σύμφωνα με τα εργαλεία που ήδη χρησιμοποιούνται ευρέως [17].
- Η βιβλιοθήκη glue() δέχεται ως όρισμα τις εκφράσεις που περικλείονται από τις παρενθέσεις οι οποίες θα εκτιμηθούν από τον κώδικα R. Οι μακροσκελείς συμβολοσειρές σπάνε κατά γραμμή και συγκολλούνται μαζί. Τα κενά και οι κενές γραμμές από την πρώτη και την τελευταία γραμμή κόβονται αυτόματα [17].
- Η βιβλιοθήκη stringr() που περιλαμβάνει απλές λειτουργίες κοινών συμβολοσειρών

Ο καθαρισμός του κειμένου από περιττά στοιχεία θα γίνει με τη βοήθεια της συνάρτησης gsub(), η οποία έχει τη δυνατότητα να αντικαθιστά όλες τις αντιστοιχίσεις μιας συμβολοσειράς. Αν η παράμετρος είναι ένα διάνυσμα συμβολοσειρών, επιστρέφει ένα διάνυσμα συμβολοσειρών του ίδιου μήκους και με τα ίδια χαρακτηριστικά (μετά από πιθανό εξαναγκασμό στον χαρακτήρα). Τα στοιχεία των διανυσμάτων string που δεν αντικαθίστανται θα επιστραφούν αμετάβλητα (συμπεριλαμβανομένης οποιασδήποτε δήλωσης κωδικοποίησης).

Έπειτα τα δεδομένα θα αποθηκευτούν σε ένα csv αρχείο που θα έχει την μορφή της εικόνας 21:



Εικόνα 21. «Ενδεικτικό παράδειγμα των tweets χώρις περριτά στοιχεία»

Το αρχείο που προκύπτει, clean_tweet2.csv,αποτελείται από 15492 γραμμές και 2 στήλες και φυσικά η απεικόνιση της εικόνας 21 μπορεί να διαφέρει ανάλογα με το πρόγραμμα που θα επιλέξουμε να ανοίξουμε το αρχείο.

Ο κώδικας που δημιουργήθηκε στα πλαίσια της παρούσας διπλωματικής εργασίας, για την επεξεργασία των δεδομένων με βάση το χρόνο, παρουσιάζεται στην εικόνα 22

```
testdat <- readRDS(file.choose())


# sort tweets with time
library(lubridate)
library(nabor)
testdat$created_at<-ymd_hms(testdat$created_at)
testdat$timesort<-as.numeric(testdat$created_at)
testdat <- testdat[order(testdat$timesort),]

# ------------------------------------------------------

# remove non english tweets

testdat <- testdat[which(testdat$lang=="en"),]


# find the number of retweets
length(which(testdat$is_retweet))

library(tidyverse)
library(tidytext)
library(glue)
library(stringr)
#

# cleaning the tweets
# -----------
clean_tweet = gsub("&amp", "", testdat$text)
clean_tweet = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", clean_tweet)
clean_tweet = gsub("@\\w+", "", clean_tweet)
clean_tweet = gsub("[[:punct:]]", "", clean_tweet)
clean_tweet = gsub("[[:digit:]]", "", clean_tweet)
clean_tweet = gsub("http\\w+", "", clean_tweet)
clean_tweet = gsub("[ \t]{2,}", "", clean_tweet)
clean_tweet = gsub("^\\s+|\\s+$", "", clean_tweet)
# ref: ( Hicks , 2014)
# ------
getwd()

# save the tweets after cleaning
write.csv(clean_tweet, file = "clean_tweet2.csv")


# example prossesing
# ==============================
# tokenize
# ---------------
source("twitter_functions.R")
sentiments2 <- sapply(clean_tweet[1:15491],FUN = GetSentiment)
plot(sentiments2)
```

Εικόνα 22 : «Κώδικας εξόρυξης συναισθήματος»

Στη συνέχεια με χρήση συνάρτησης που δημιουργήθηκε για την εξυπηρέτηση των σκοπών αυτής της διπλωματικής, πραγματοποιείται η εξαγωγή συναισθήματος για κάθε tweet που υπάρχει στο σύνολο δεδομένων. Απαραίτητο σε αυτό το σημείο και πριν προχωρήσουμε στην περαιτέρω ανάλυση του αλγορίθμου, είναι να γίνει επεξήγηση της χρήσης και λειτουργικότητας της συνάρτησης που προαναφέρθηκε.

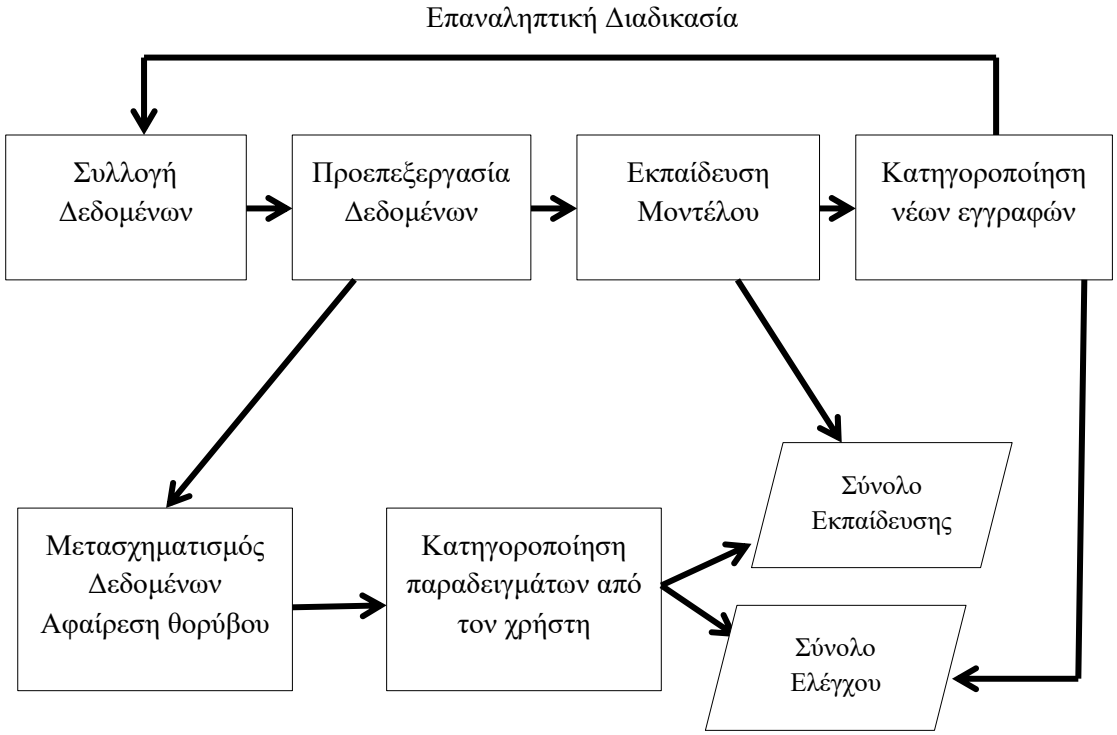
Στην εικόνα 23 παρουσιάζεται ο κωδικας της συνάρτησης που εκτελεί την εξόρυξη συναισθμάτων:

```
# write a function that takes the name of a file and returns the # of
postive
# sentiment words, negative sentiment words, the difference & the
normalized difference
GetSentiment <- function(file){
# get the file

  tokens <-tibble( text = file) %>% unnest_tokens(word, text)
  # tokenize

  testt<- tokens %>%
  inner_join(get_sentiments("bing")) %>% # pull out only sentimen
words
  count(sentiment) %>% # count the # of positive & negative words
  spread(sentiment, n, fill = 0)

  if (length(names(testt)) > 0){

    if (names(testt) == "negative"){
      testt$positive=0
    }else if (names(testt) == "positive"){
      testt$negative=0
    }

    result <- testt$positive - testt$negative

  } else {
    result <- 0
  }

  # result <- testt$positive - testt$negative

  # return(sentiment)
  return(result)
}
```

Εικόνα 23: «Κώδικας συνάρτησης για εξόρυξη συναισθήματος»

Η συνάρτηση δέχεται σαν όρισμα ένα όνομα αρχείου, και επιστρέφει το πλήθος από τις θετικά και αρνητικά συναισθηματικά λέξεις, τη διαφορά θετικών και αρνητικών καθώς και την κανονικοποιημένη διαφορά αυτών. Για την εξόρυξη συναισθήματος μέσα από την συνάρτηση χρησιμοποιείται το λεξικό BING γεγονός που μας οδηγεί στο συμπέρασμα ότι έχουμε να κάνουμε με έναν αλγόριθμό επιβλεπόμενης μάθησης. Η επιβλεπόμενη μάθηση είναι μία κατηγορία μηχανικής μάθησης, στόχος της οποίας είναι ο χαρακτηρισμος δεδομένων με βάση κάποια δεδομένα εκπαίδευσης. Τα δεδομένα εκπαίδευσης

απότελούνται από ένα σύνολο παραδειγμάτων τα οποία χρησιμοποιούνται για εκπαίδευση μοντέλων. Στην επιβλεπόμενη μάθηση κάθε παράδειγμα αποτελείται από ένα σύνολο εισόδου (συνήθως ένα διάνυσμα από χαρακτηριστικά) και μία επιθυμητή τιμή εισόδου. Οι αλγόριθμοι επιβλεπόμενης μάθησης αναλύουν τα δεδομένα εκπαίδευσης και παράγουν ένα μοντέλο το οποίο μπορεί να χρησιμοποιηθεί για να χαρακτηρίσει νέα παραδείγματα. Το βέλτιστό σενάριο επιτρέπει στον αλγόριθμο να καθορίσει σωστά την ετικέτα της κατηγορίας για άγνωστα μέχρι τώρα παραδείγματα. Για να επιτευχθεί αυτό απαιτείται ο αλγόριθμος μάθησης να γενικέυει από τα δεδομένα εκπαίδευσης σε «αθέατες» καταστάσεις με ένα «λογικό» τρόπο.



Εικόνα 24. «Διάγραμμα ροής αλγορίθμου επιβλεπόμενης μάθησης»

Οι ενέργειες της συνάρτησης διεξάγωνται με χρήση ποικίλων συναρτήσεων όπως :

- Η συνάρτηση tibble().Η συνάρτηση αυτή είναι ένας καλός τρόπος για να δημιουργηθεί ένα πλαίσιο δεδομένων. Ενσωματώνει τις βέλτιστες πρακτικές για τα πλαίσια δεδομένων και ποτέ δεν αλλάζει τον τύπο μιας εισόδου. Εκεί εκχωρείται το όνομα του φακέλου που είχαμε ΄δωσει ως είσοδο στην συνάρτηση [17].
- Σημαντικό ρόλο διεξάγει και ο τελεστής %>% . Πρόκειται για ένα εμπρόσθιο τελεστή σωλήνωσης. Μπορούμε να τον χρησιμοποιήσουμε για να περάστεί μία είσοδος στην αριστερή πλευρά μέσω του τελεστή της δεξιάς πλευράς. Σε μαθηματικούς όρους, είναι η ακόλουθη ενέργεια [18]:

x%>% f που μεταφράζεται σε f (x)

Για την καλύτερη κατανόηση του όρου, ακολουθεί παράθεση παραδείγματος όπου δημιουργείται ένα διάνυσμα τιμών, λαμβάνεται το τετράγωνο ρίζας κάθε αριθμού και στη συνέχεια υπολογίζεται το άθροισμα όπως φαίνεται στην εικόνα 25:

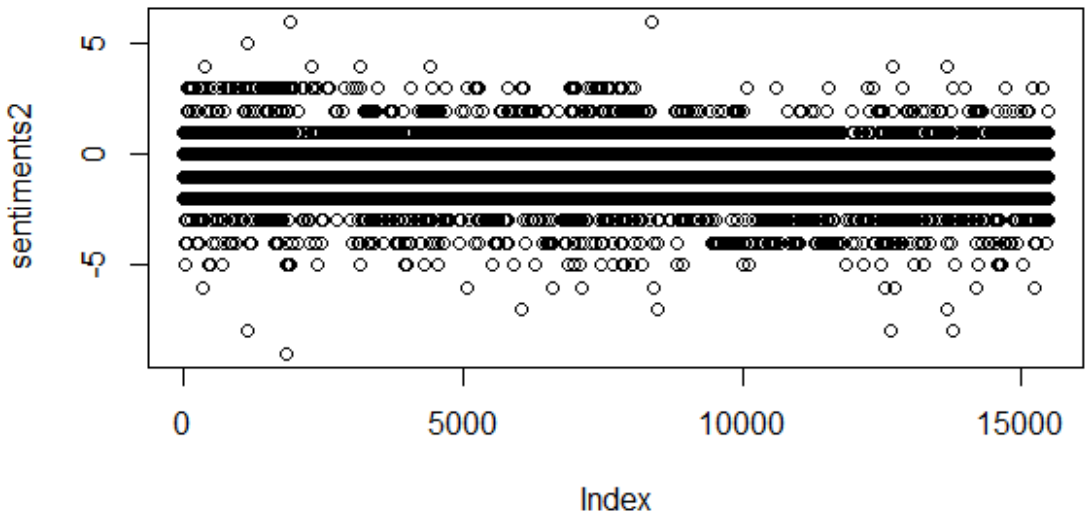
```
c(1,2,3,4) %>% Map(sqrt,.) %>% Reduce(sum,.)
#The output: [1] 6.14264
```

Εικόνα 25. « Παράδειγμα χρήσης του τελεστή»

Επίσης ο τελεστής αυτός είναι πολύ χρήσιμος όταν πρέπει να εφαρμόσουμε πολλούς διαφορετικούς μετασχηματισμούς στα δεδομένα μας και δεν θέλουμε να αποθηκεύσουμε τα ενδιάμεσα αποτελέσματα ή να έχουμε πολλές ανοιχτές και κλειστές παρενθέσεις.

- Η συνάρτηση unnest_tokens() που διαχωρίζει μια στήλη σε τμήματα χρησιμοποιώντας το πακέτο tokenizers, διαιρώντας τον πίνακα με βάση ένα διακριτικό ανά σειρά.

Αφού αναλύθηκε η λειτουργία της συνάρτησης, στην συνέχεια του κώδικα της εικόνας 22, με χρήση κατάλληλων συναρτήσεων όπως η συνάρτηση sapply() εξάγουμε αποτέλεσμα για όλα τα δεδομένα. Με τη χρήση μίας απλής συνάρτησης plot έχουμε το αποτέλεσμα της εικόνας 26:



Εικόνα 26. « Διάγραμμα score συναισθήματος»

Όπως είναι εμφανές και από την εικόνα 26, τα περρισότερα score συναισθήματος συγκεντρώνονται σε αρνητικούς αριθμούς, κάποια στο 0 και αρκετά στην τιμή 1. Η διασπορά για αριθμούς μεγαλύτερους του 1 και μικρότερους απο το -3 είναι αρκετά μικρότερη. Καταλήγουμε λοιπόν στο συμπέρασμα ότι τα tweets που αναφέρονται στην Theressa May είναι αρνητικού συναισθήματος ως προς την πλειοψηφία τους.

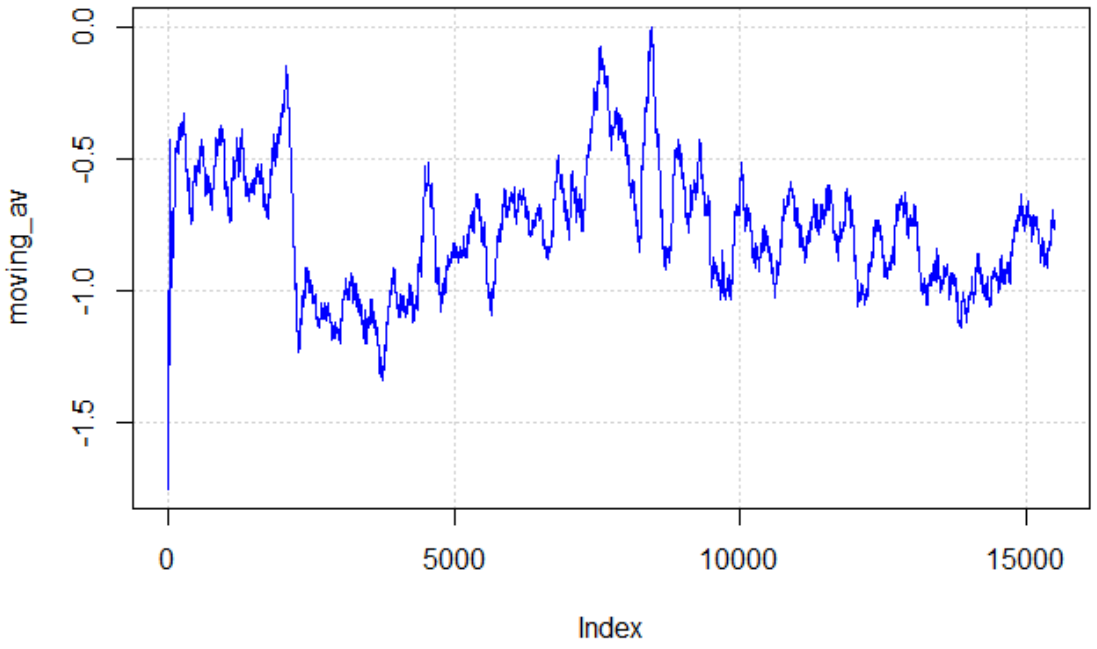
Στη συνέχεια θα υπολογίσουμε μέσω του αλγορίθμου που αναπτύχθηκε για την παρούσα διπλωματική έναν μέσο όρο ανα 200 tweets και θα εξάγουμε ένα score το οποίο στη συνέχεια θα το οπτικοποιήσουμε. Αυτά τα score θα μας φανούν στη συνέχεια του αλγορίθμου αρκετά χρησιμα για την εξαγωγή συναισθήματος με βάση το χρόνο.

```
library(pracma)
#calculate the moving average with window size 200
moving_av <- movavg(sentiments2,200,type=c("s"))
plot(moving_av,type="l",col="blue",panel.first=grid())
```

Εικόνα 27. «Κώδικας σε γλώσσα R για υπολογισμό ‘κινούμενου’μέσου όρου»

Για να υπολογιστεί ο μέσος όρος, θα χρειαστεί η βιβλιοθήκη pracma η οποία παρέχει μεγάλο αριθμό λειτουργιών από την αριθμητική ανάλυση και γραμμική άλγεβρα, αριθμητική βελτιστοποίηση, διαφορικές εξισώσεις, χρονοσειρές, συν μερικές γνωστές ειδικές μαθηματικές λειτουργίες. Έπειτα χρησιμοποιώντας την συνάρτηση movavg() υπολογίζεται ο μέσος όρος ανά 200 tweets(δηλαδή 200 είναι το μέγεθος του

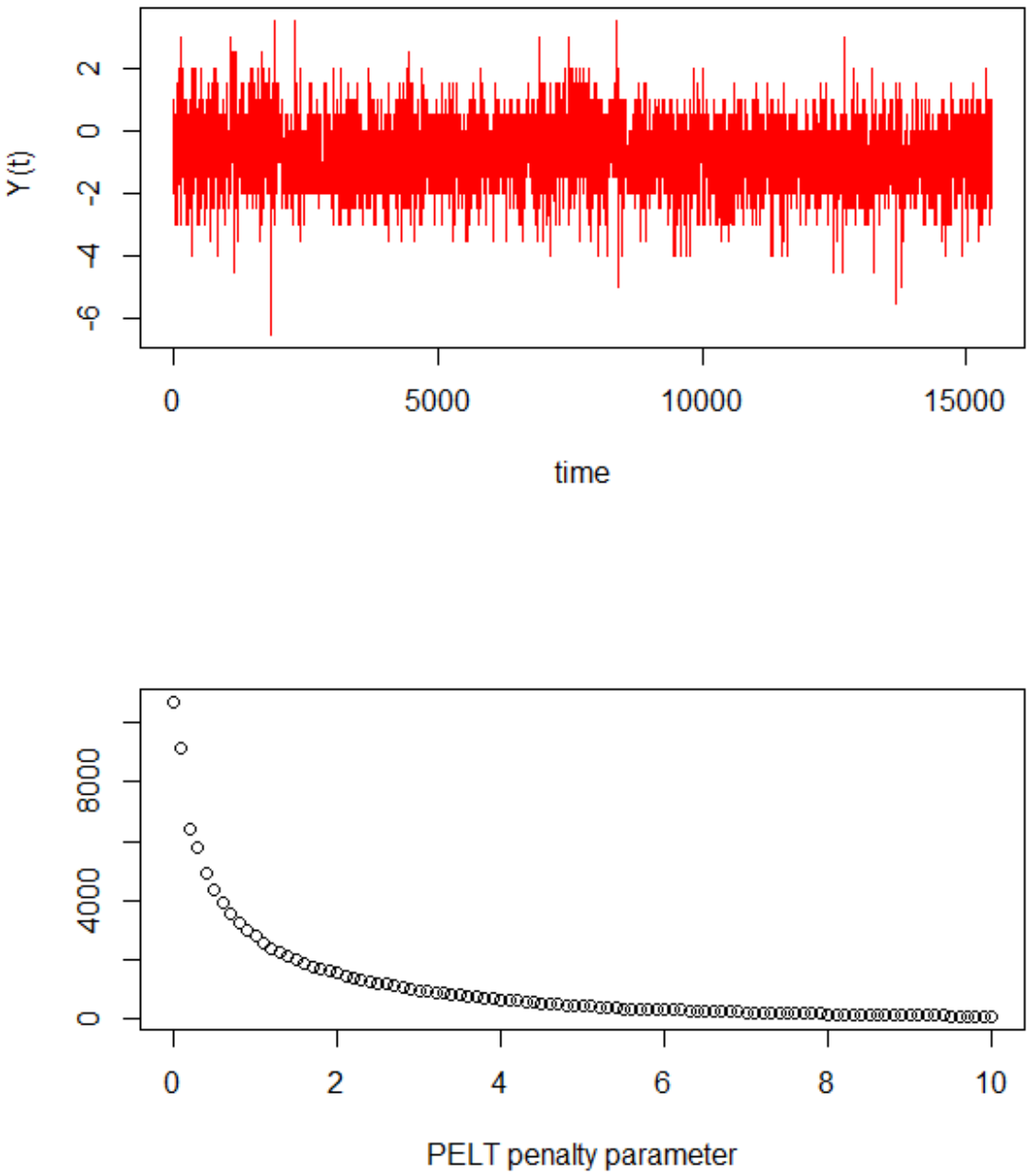
«κινούμενου» παραθύρου που χρησιμοποιούμε) και στη συνέχεια με plot() οπτικοποιούμε το αποτέλεσμα που παρουσιάζεται στην εικόνα 28:



Εικόνα 28. « Διάγραμμα μέσου όρου με μέγεθος παραθύρου 200»

Για να μπορέσουμε να έχουμε και μία προσέγγιση εξόρυξης γνώμης σε πραγματικό χρόνο,πρέπει να προσδιορίσουμε τις παραμέτρους ποινής για για να ελέγξουμε την ανίχνευση σημείου αλλαγής συναισθήματος. Στα πλαίσια της παρούσας διπλωματικής θα χρησιμοποιήσουμε τον αλγόριθμο PELT. Μπορούμε να το κάνουμε αυτό κάνοντας σχεδιαγράμματα και χρησιμοποιώντας την τιμή παραμέτρου ποινής στα σχεδιάγραμματα [19]. Θα ορίσουμε μια συνάρτηση cptfn για την εκτέλεση μιας ακολουθίας διαφορετικών παραμέτρων ποινής και στη συνέχεια σχεδιάζουμε για κάθε χρονική σειρά ένα αντίστοιχο σχεδιάγραμμα.

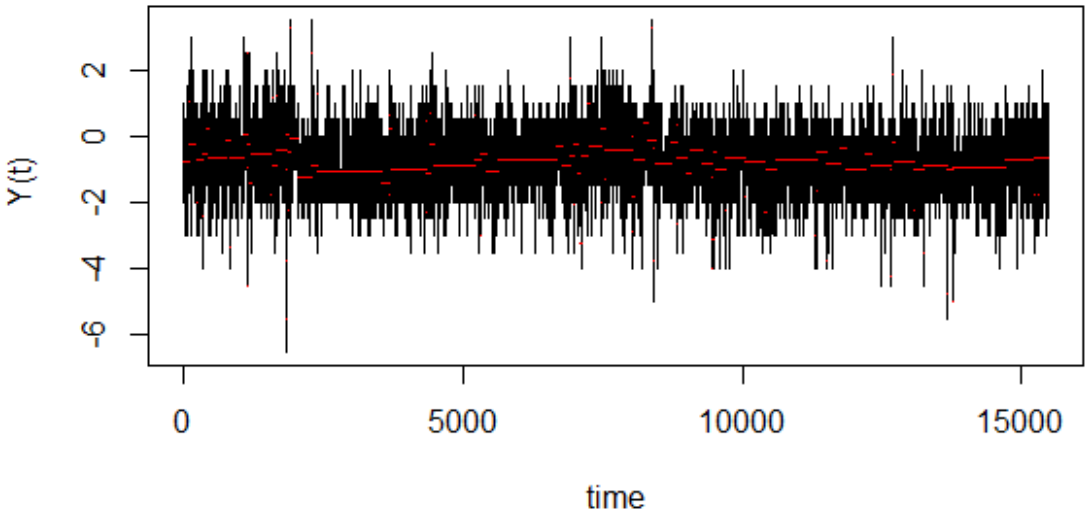
Σημαντικό σημείο του αλγορίθμου αποτελεί η συνάρτηση seq() που προσδιορίζει τις τιμές από τις οποίες ξεκινάει και τελειώνει η ακολουθία που δημιουργείται καθώς και την σταδιακή αύξηση αυτής. Τα αποτελέσματα παρουσιάζονται στην εικονα 30:



Εικόνα 29. «Διάγραμμα κυματομορφής και διάγραμμα παραμέτρου ποινής»

Από τις γραφικές παραστάσεις της εικόνας 29 και συγκεκριμένα από την δευτερη οριζόντια γραφική παράσταση , μπορούμε να δούμε ότι μια τιμή παραμέτρου ποινής (penalty.val) περίπου 0 ή 1 θα πρέπει να είναι επαρκής για να αποφευχθεί ανίχνευση σημείου αλλαγής πορείας. Μπορούμε λοιπόν να εφαρμόσουμε τη συνάρτηση μέσης μεταβολής (cpt.mean) στα σήματα της σειράς μας χρησιμοποιώντας μια τιμή παραμέτρου ποινής 8 ή 9 και να δούμε αν μπορούμε να προσδιορίσουμε σωστά πού συμβαίνουν τα σημεία αλλαγής.

Το γραφικό αποτέλεσμα:

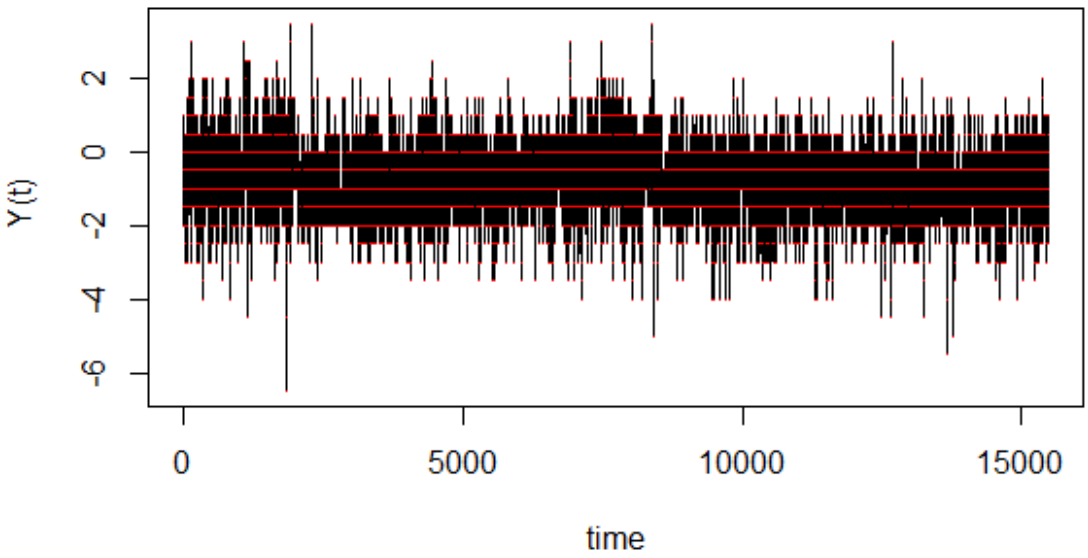


Εικόνα 30. «Διάγραμμα παρουσίασης change point detection»

Ο προσδιορισμός της τιμής παραμέτρου ποινής έχει ιδιαίτερη σημασία καθώς δεν θέλουμε έναν αριθμό ο οποίος θα κάνει τον αλγόριθμο πολύ ευαίσθητο στις αλλαγές διοτι τότε η διεξαγωγή συμπεράσματος δεν θα ήταν εφικτή. Η γραφική παράσταση για το pelt parameter που έγινε παραπάνω εξυπηρετεί αυτό τον σκοπό, να γινέι δηλαδή αντιληπτό σε ποιά τιμή της παραμέτρου ο αλγόριθμος παρουσιάζει πιο ομαλή καμπύλη και συνεπώς δεν θα έχει πολλές διακυμάνσεις.

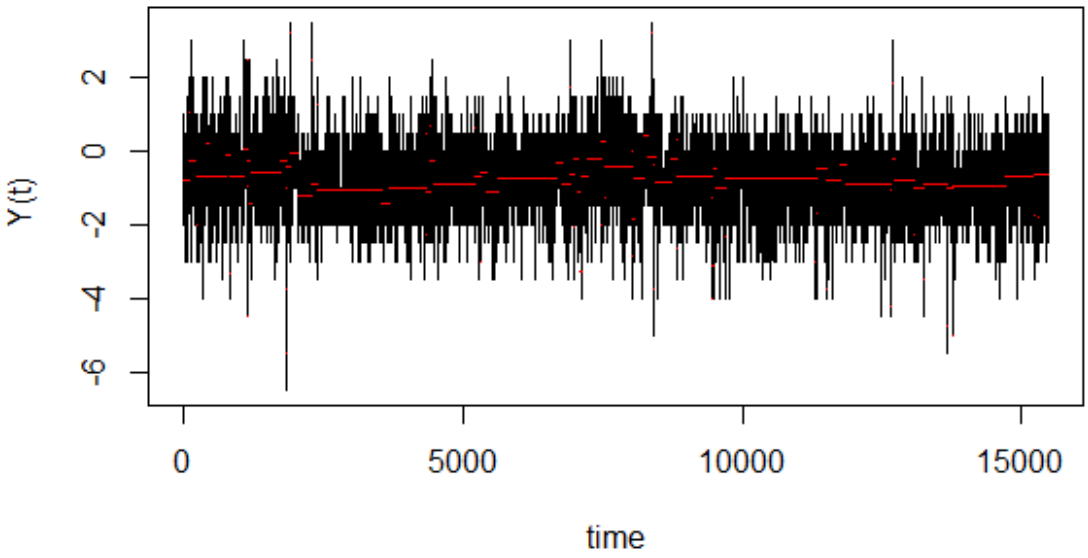
Εάν το ζητούμενο μας για κάποιο άλλο παράδειγμα απαιτεί την ευαισθησία του αλγορίθμου τότε εννοείται πως επιλέγουμε μικρότερες τιμές για παράμετρο ποινής.

Στη συνέχεια, αναφέρονται μερικά παραδείγματα διεξαγωγής του αλγορίθμου για μικρότερες ή πολυ μεγαλύτερες τιμές της παραμέτρου ποινής για να γίνουν καλύτερα αντιληπτά τα όσα αναφέρθηκαν.

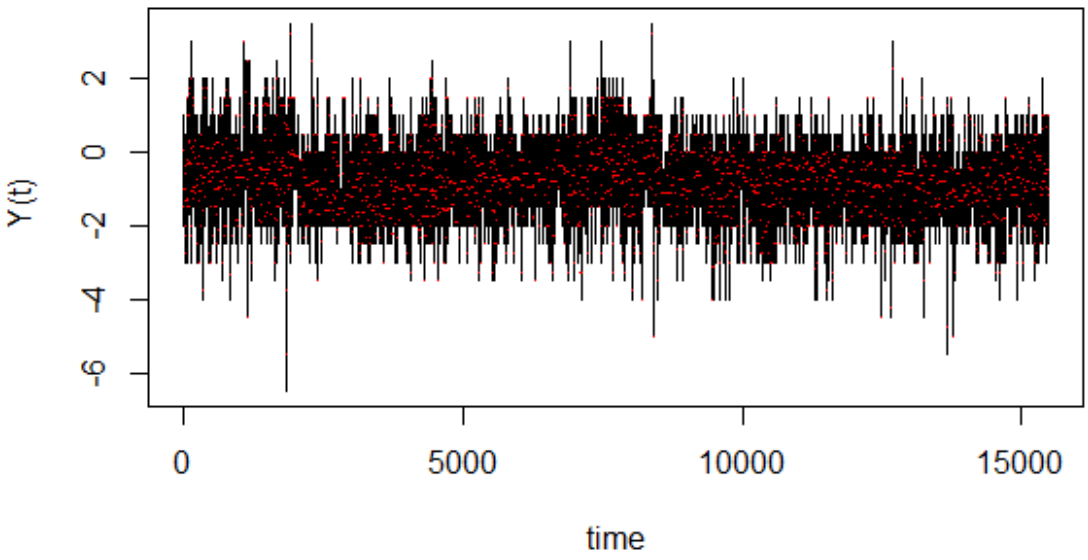


Εικόνα 31. «Διάγραμμα παρουσίασης change point detection με παράμετρο ποινής ίση με μηδέν»

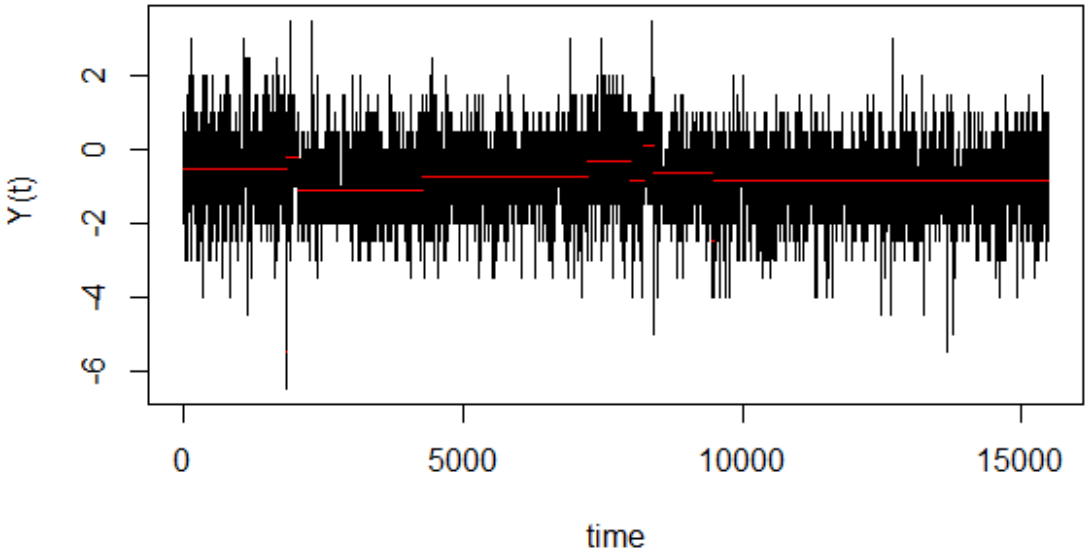
PENALTY VALUE = 0 ,παρατηρούνται πολύ συχνές αλλαγές στο σχεδιάγραμμα.



Εικόνα 32. «Διάγραμμα παρουσίασης change point detection με παράμετρο ποινής ίση με 10»



Εικόνα 33. «Διάγραμμα παρουσίασης change point detection με παράμετρο ποινής ίση με δύο»



Εικόνα 34. «Διάγραμμα παρουσίασης change point detection με παράμετρο ποινής ίση με 30»

PENALTY VALUE=30 για τιμές εκτός της κλίμακας (1-10) που επιλέξαμε δεν παρατηρούνται πολλές αλλαγές σημείων.

ΒΙΒΛΙΟΓΡΑΦΙΑ-ΑΝΑΦΟΡΕΣ

[1] Bo Pang and Lillian Lee. “Opinion Mining and Sentiment Analysis”. Foundations and Trends in Information Retrieval:Vol 2: No. 1-2, pages 1-135. July 07, 2008

[2] Bing Liu. “Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies”. Morgan and Claypool publishers. May 23, 2012

[3] Bing Liu. “Sentiment Analysis: Mining Opinions, Sentiments and Emotions”. Cambridge University Press. June 04, 2015

[4] Management Association, Information Resources. “Machine Learning: Concepts, Methodologies, Tools and Applications.” IGI Global, July 31, 2011

[5] Michael W.BerryJacob Kogan. “Text Mining: Applications and Theory”. John Wiley and Sons, February 25, 2010

[6] O A Grigg, V T FareWell, D J Spiegelhalter. “Use of risk-adjusted CUSUM and PSPRT charts for monitoring in medical contexts”. April 1, 2003. Volume 12, issue 2, pages: 147-170, PubMed

[7] Aristidis G.Vrahatis, Sotiris K Tasoulis, Spiros V. Georgakopoulos, Vasilis P. Plagianakos. “Real Time Sentiment Change Detection of Twitter Data Streams”. Lamia, Greece. Department of Computer Science and Biomedical Informatics, University of Thessaly. April 02, 2018. Pages 6. Διαθέσιμο: https://arxiv.org/pdf/1804.00482.pdf

[8] Ραυτόπουλος Ιωάννης. “ Ανάλυση Συναισθήματος σε Κοινωνικά Δίκτυα”. Πάτρα. Σχολή Διοίκησης Επιχειρήσεων Μεταπτυχιακό Μ.Β.Α, Διπλωματική Εργασία, Πανεπιστήμιο Πατρών. 2019. Σελίδες 61. Διαθέσιμο: https://nemertes.lis.upatras.gr/jspui/bitstream/10889/12734/1/ΡΑΥΤΟΠΟΥΛΟΣ%20ΙΩΑΝΝΗΣ%20MBA%202019.pdf

[9] Γεράσιμος Ε. Σκαράκης. “ Ανίχνευση Αλλαγών και Συσταδοποίηση σε πραγματικό χρόνο”. Αθήνα. Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, σχολή θετικών επιστημών, τμήμα πληροφορικής και τηλεπικοινωνιών, πτυχιακή εργασία. Μάϊος 2015. Σελίδες 53. Διαθέσιμο: https://pergamos.lib.uoa.gr/uoa/dl/frontend/file/lib/default/data/1324086/theFile/1324087

[10] Ranjan Baitha. “Twitter Text Mining Analytics using R and Hadoop”. DEPT OF MCA DSCE 2014-2015,pages 7.Διαθέσιμο:https://www.academia.edu/15083048/Twitter_Analytics_Using_R

[11] https://wearesocial.com/global-digital-report-2019 . Ανώνυμος.“We are social”. 2019

[12] http://www.sepe.gr/gr/research-studies/article/15221847/toulahiston-35-ores-tin-imera-vrisketai-online-o-mesos-ellinas/ . Ανώνυμος. “Σύνδεσμος Επιχειρήσεων Πληροφορικής και Επικοινωνιών Ελλάδας” .2020

[13] https://theappsolutions.com/blog/development/sentiment-analysis/#contents_0 . Bily V. “The App solutions”.2020

[14] https://www.guru99.com/machine-learning-tutorial.html . Ανώνυμος. “Machine Learning Tutorial for Beginners”. 2020

[15] https://monkeylearn.com/text-mining/ . Garveta R. “Text Mining: The Beginner Guide”. 2019

[16] http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know . Kassambara A. “STHDA, Statistical tools for high-throughput data analysis”. 2020

[17] https://cran.r-project.org/web/packages/index.html . CRAN. 2020

[18] https://www.datacamp.com/community/tutorials/pipe-r-tutorial . Willems K. “Pipes in R Tutorial for Beginners”. November 16, 2017

[19] https://rpubs.com/richkt/269908 . T.Rich K. 2020. “RPubs: Change Point Detection”. 2020

ΠΗΓΕΣ ΕΙΚΟΝΩΝ

Εικόνα 1 : https://colemanlegalpartners.ie/banners/social-media-moderator-trauma-compensation/top-10_best-social-media-sites-apps/

Εικόνα 2: Hootsuite, https://hootsuite.com/resources/digital-in-2019

Εικόνα 3: The Information Lab, Alteryx Global Partner of the year 2019
https://www.theinformationlab.co.uk/2018/12/21/sentiment-analysis-3-ways-in-alteryx/

Εικόνα 4: Δημιουργία εικόνας μέσω Word

Εικόνες 5,6,9: [14]

Εικόνα 7: Δημιουργία εικόνας μέσω Word

Εικόνα 8: Introduction to machine learning and deep learning,
https://medium.com/@sanchittanwar75/introduction-to-machine-learning-and-deep-learning-bd25b792e488

Εικόνα 10: Machine Learning - Machine Learning Algorithms Deep Learning Artificial Intelligence Computer Science, https://favpng.com/png_view/machine-learning-machine-learning-algorithms-deep-learning-artificial-intelligence-computer-science-png/cBe2H9vF

Εικόνα 11,12,13: Text Processing and Sentiment Analysis of Twitter Data, https://hackernoon.com/text-processing-and-sentiment-analysis-of-twitter-data-22ff5e51e14c

Εικόνα 14: Αρχείο που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 15: Αρχείο που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 16: Κώδικας που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 17: Αρχείο που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 18: Αρχείο που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 19: Κώδικας που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 20: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 21: Κώδικας που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 22: Αρχείο που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 23: Κώδικας που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 24: Κώδικας που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 25: [8]

Εικόνα 26: [18]

Εικόνα 27: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 28: Κώδικας που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 29: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 30: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 31: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 32: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 33: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 34: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio

Εικόνα 35: Αποτέλεσμα που δημιουργήθηκε από την επεξεργασία στο R studio