ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«*ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΟ TWITTER ΜΕ ΣΚΟΠΟ ΤΗΝ ΣΥΝΑΙΣΘΗΜΑΤΙΚΗ ΑΝΑΛΥΣΗ ΤΩΝ ΚΕΙΜΕΝΩΝ*»

Ιωάννου Γεώργιος

Επιβλέπων Καθηγητής: Κος Χαλκιόπουλος Κωνσταντίνος

Πάτρα 01/07/2022

ΠΡΟΛΟΓΟΣ

Σε αυτό το σημείο κύριως σκοπός μου είναι να ευχαριστήσω όλους αυτούς που μου στάθηκαν, όπως το οικογενειακό μου περιβάλλον, καθώς και τον επιβλέπων καθηγητή μου κο Κωνσταντίνο Χαλκιόπουλο, καθόλη τη διάρκεια εκπόνησης της πτυχιακής μου εργασίας, διότη χωρίς τη στήριξή τους τίποτα από τα παραπάνω δεν θα ήταν εφικτά.

Το θέμα μου της πτυχιακής μου « Εξόρυξη δεδομένων από το Twitter με σκοπό την συναισθηματική ανάλυση των κειμένων» προήλθε αφενώς από το ενδιαφέρων που με έχει κατακυριέυσει τελευταία για την ανάλυση δεδομένων σαν τομέα εξειδίκευσης, αφετέρου και από το πόσο σημαντική είναι η προσφορά της ανά τον κόσμο εν έτη 2022.

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αφορά την εξόρυξη δεδομένων από ένα συγκεκριμένο πεδίο εφαρμογής, το twitter. Στις πλατφόρμες των μέσων κοινωνικής δικτύωσης η εξόρυξη των γνώσεων και η ανάλυση συναισθήματος που εφαρμόζεται τα τελευταία έτη, χει οδηγήσει τους, μελετητές στην εξαγωγή κρίσιμων πληροφοριών μέσω της ανάλυσης και της ταξινόμησης του γραπτού λόγου των χρηστών. Η εργασία παρουσιάζει στα πλαίσια της βιβλιογραφικής ανασκόπησης μια γενική εισαγωγή στο Microblogging, στην Ανάλυση των Δεδομένων, των Συναισθημάτων και της φυσικής γλώσσας. Στη συνέχεια, με λεπτομέρειες παρουσιάζει τη Συναισθηματική ανάλυση, το μέσο κοινωνικής δικτύωσης twitter και τους τρόπους εξόρυξης δεδομένων. Κατόπιν, ως μελέτη περίπτωσης παρουσιάζονται συγκεκριμένα δεδομένα από το twitter και η συναισθηματική ανάλυση των κειμένων που ανεβαίνουν στην εν λόγω πλατφόρμα. Ως συγκεκριμένη μελέτη περίπτωσης παρουσιάζεται η έρευνα των Alexander Bogdanowicz και ChengHe Guan, οι οποίοι λαμβάνοντας τα δεδομένα κατά τη διάρκεια της υγειονομικής κρίσης προέβησαν στα ερευνητικά τους πορίσματα, τα οποία συμφωνούν με τη βιβλιογραφία για τη χρήση συγκεκριμένων λέξεων και όρων εκείνη τη συγκεκριμένη περίοδο χρόνου.

Λέξεις-Κλειδιά: εξόρυξη δεδομένων, Συναισθηματική ανάλυση, twitter, Microblogging, covid -19

ABSTRACT

The present work concerns the extraction of data from a specific field of application, twitter. In social media platforms, knowledge mining and emotion analysis applied in recent years, has led scholars to extract critical information through the analysis and classification of users' written speech. The paper presents in the context of the literature review a general introduction to Microblogging, Data Analysis, Emotions and natural language. It then goes into detail about Emotional analysis, twitter social media and data mining. Then, as a case study, specific data from twitter and the emotional analysis of the texts that are uploaded to this platform are presented. A specific case study is the research of Alexander Bogdanowicz and ChengHe Guan, who, taking the data during the health crisis, made their research findings, which agree with the literature on the use of specific words and terms in that particular period of time.

Keywords: data mining, Emotional analysis, twitter, Microblogging, covid -19

ΚΕΦΑΛΑΙΟ 1 Εισαγωγή

1.1 Γενική Εισαγωγή

Η τεχνολογία, σήμερα, προοδεύει αρκετά αυξητικά με αποτέλεσμα οι γνώσεις και η απόκτηση της είναι σημείο των καιρών στις σύγχρονες κοινωνίες. Διάφορες πηγές από επεξεργασμένα δεδομένα προσφέρουν την γνώση και εμπειρία της τεχνολογίας. Η δημιουργία μιας πληροφορίας που εκπορεύεται από την ορθολογική ανάλυση και την ακριβή των επεξεργασμένων δεδομένων, την καθιστούν ως γνώση.

Ο σημερινός κόσμος διακατέχεται από ένα ευρέως φάσμα γνώσεων χάριν στην αναπτυσσόμενη τεχνολογία και στην συνεχή καθημερινή χρήση κάθε πληροφοριακής συσκευής. Η γνώση, όμως αυτή, δεν παρέχεται μόνο από τα αμέτρητα βιβλία ηλεκτρονικής μορφής που διατίθενται για ανάγνωση σε όλες τις διαδικτυακές μηχανές αναζήτησης αλλά από πολλές άλλες πηγές.

Βάση μιας μελετημένης διαδικασίας προκύπτει η εξόρυξη γνώσης (data mining) που μπορεί να προσαρμοστεί σε ένα ειδικό πεδίο έρευνας, εφαρμόζοντας μια αναλυτική διαδικασία κάθε προσφερόμενου πακέτου δεδομένων, όπου είναι προερχόμενα μέσω τεχνητών και μη μέσων σε τεχνολογικά και φυσικά περιβάλλοντα όπου μπορούν να εφαρμοστούν.Το πόσο σημαντικές και δημοφιλή στη χρήση τους είναι οι microblogging πλατφόρμες κοινωνικών μέσων, έχουν οδηγήσει σε άλλα επίπεδα με μεγάλη αύξηση καθώς έχουν εξελιχθεί οι νέες τεχνολογίες σε όλες τις υπηρεσίες του διαδικτύου και στα κοινωνικά ενημέρωσης.

Η εξέλιξη αυτή είναι αποτέλεσμα της χρονοβόρας απασχόλησης των ανθρώπων να βρίσκονται online και οι εμπλεκόμενες εταιρείες να υιοθετούν τις τάσεις των μελλοντικών πελατών τους. Οι άνθρωποι μέσω του διαδικτύου και των κοινωνικών δικτύων προσφέρουν εμπειρίες, συναισθήματα, γνώσεις και στιγμές έτοιμοι να τις μοιραστούν.

1.2 Microblogging

Το microblogging είναι ένας τύπος μετάδοσης που προϋπάρχει υπό μορφής blogging. Ένα microblog έχει άλλα χαρακτηριστικά στο περιεχόμενό του από το blog παραδοσιακής μορφής, το οποίο είναι πιο μικρό στο πραγματικό και στο συνολικό μέγεθος του αρχείου. Τα microblogs προσφέρουν δυνατότητες ανταλλαγής περιεχομένου μικρότερων στοιχείων στους χρήστες όπως παραδείγματος χάριν είναι οι σύντομες προτάσεις, οι μεμονωμένες εικόνες και οι σύνδεσμοι βίντεο[9] και αυτός είναι το γεγονός που τα κάνει τόσο δημοφιλή. Τα μικρά αυτά μηνύματα συνήθως ονομάζονται και microposts.

Οι microbloggers, μπορούν να κάνουν δημοσιεύσεις από πολύ απλά θέματα όπως του τύπου "τι κάνω τώρα", ως σε πιο ουσιαστικά, όπως του τύπου τα "σπορ αυτοκίνητα"., γεγονός που υπάρχει και στο παραδοσιακό blogging. Επιπλέον, η ύπαρξη των εμπορικών microblogs συμβαίνει λόγω της προώθησης ιστοτόπων, προϊόντων και υπηρεσιών και προϊόντων πρεσβεύοντας την συνεργασία ενός οργανισμού. Υπάρχουν υπηρεσίες του microblogging με λειτουργίες ρυθμίσεων απορρήτου, όπου ο κάθε χρήστης δύναται να ελέγχει ποιος επισκέπτεται τα microblogs τους και με πολλαπλές επιλογές εναλλακτικών τρόπων δημοσίευσης των διάφορων καταχωρήσεων εκτός της

διαδικτυακής διεπαφής. Εδώ περιέχονται τα μηνύματα κειμένου, τα άμεσα μηνύματα, το ηλεκτρονικό ταχυδρομείο, και το ψηφιακό βίντεο.

Το microblogging αποτελεί μια επανάσταση στον τρόπο κατανάλωσης κάθε πληροφορίας, καθώς οι χρήστες παίζουν το ρόλο του αισθητήρα ή των πηγών πληροφορίας που θα ασκήσουν επιρροή, πρόκληση και καθοδήγηση της κάλυψης της πληροφορίας από τα μέσα ενημέρωσης. Οι άνθρωποι πλέον τείνουν να μοιράζονται όσα συμβάντα και γεγονότα συμβαίνουν στο περιβάλλον τους, καθώς και τις προσωπικές απόψεις τους για τα τρέχοντα θέματα παγκοσμίως και αυτές οι υπηρεσίες δύνανται στην αποθήκευση όλων των δεδομένων από τις δημοσιεύσεις τους, όπως είναι ο χρόνος ή η τοποθεσία.

Η ανάλυση των παραπάνω δεδομένων μπορεί να εμπεριέχει και άλλες επιπλέον διαστάσεις όπως είναι το θέμα, το συναίσθημα και η διάρθρωση του δικτύου. Αυτά αποτελούν το αντικείμενο έρευνας των ερευνητών για την κατανόηση των κοινωνικών αντιλήψεων των ανθρώπων γενικότερα στα επικείμενα ή τρέχοντα γεγονότα που προκαλούν εθνικό ή παγκόσμιο ενδιαφέρον. Ο τρόπος ανάλυσης των tweets και η αναγνώριση του φορτίου τους, κυρίως του συναισθηματικού είναι και το φλέγον ζήτημα και πεδίο παρατηρητικότητας του microblogging, όπως και αρκετές σύγχρονες μελέτες έχουν προβεί σε αναλύσεις του συναισθήματος εκ των εγγράφων ή το περιεχομένου του διαδικτύου. Υπάρχουν, όμως και αρκετά προβλήματα όταν οι εφαρμογές αυτές έχουν ως επίκεντρο το microblogging, αφού η παρατήρηση του συναισθήματος είναι αρκετά δύσκολη και δυσνόητη λόγω του μεγάλου πεδίου συζήτησης των θεμάτων συνδυασμένο με το μικρό και περιοριστικό μέγεθος των γραπτών μηνυμάτων.

1.3 Εισαγωγή στην Ανάλυση των Δεδομένων

Η Ανάλυση Δεδομένων (data analysis) είναι μια μορφή ουσιαστικής διαδικασίας για:

- συλλογή, δηλαδή παρατήρηση και απόκτηση
- επεξεργασία, δηλαδή καθαρισμό και μετατροπή και
- μοντελοποίηση των δεδομένων για την εύρεση χρήσιμων πληροφοριών προς την υποστήριξη διαφόρων ειδών για λήψεις αποφάσεων (decision-making).

Η εξόρυξη της γνώσης αποτελεί υποκατηγορία ή τεχνική ανάλυση, με στόχο την εύρεση μοντέλων (προτύπων) και την αναζήτηση γνώσεων που βάζουν στο επίκεντρο την πρόβλεψη κυρίως αλλά και την περιγραφή διάφορων φαινομένων και συμπεριφορών.

Επίσης, η διαδικασία της προγνωστικής ανάλυσης (predictive analytics) έχει ως στόχο να εφαρμόζει τα στατιστικά μοντέλα κατηγοριοποίησης και πρόβλεψης δεδομένων, όπως και την ανάλυση κειμένου (text analytics). Για την επιτυχία των παραπάνω οφείλεται να εφαρμόζονται τα στατιστικά εργαλεία συνδυασμένα με γλωσσολογικές τεχνικές έτσι ώστε να διεξάγεται και να κατηγοριοποιείται κάθε εισερχόμενη πληροφορία από χωρίς δομή δεδομένα (unstructured data).

Η επιστήμη της Ανάλυσης Δεδομένων ορίστηκε για πρώτη φορά από τον στατιστικό John Tukey το 1961 και ήταν αυτός που χρησιμοποίησε την σύντομη λέξη «bit» από το «binary digit», όταν συνεργαζόταν με τον John von Neumann στα πρώιμα στάδια της σχεδίασης υπολογιστών. Επιπλέον, ο συγκεκριμένος όρος πρωτοδημοσιεύτηκε σε ένα άρθρο του Claude Shannon το 1948 και από τότε θεωρείται η βασική μονάδα δομής των στοιχείων πληροφορίας ηλεκτρονικής μορφής. Ακόμα ο Tukey ορίζει πως η επιστήμη της Ανάλυσης Δεδομένων, είναι όλες οι διαδικασίες που αναλύουν τα δεδομένα, τις τεχνικές ερμηνείας των αποτελεσμάτων των διαδικασιών αυτών, τους

τρόπους οργάνωσης της συλλογής δεδομένων για να αναλύονται ευκολότερα και ακριβέστερα και τέλος τη μηχανική που συνδυάζει τα αποτελέσματα των μαθηματικών συναρτήσεων και των στατιστικών μεθόδων εφαρμογής της ανάλυσης δεδομένων.

1.4 Εισαγωγή στην Ανάλυση Συναισθημάτων

Η βασικότερη επιλογή για ένα μοντέλο ανάλυσης οφείλει να έχει σωστή απόδοση σε κάθε δείγμα και ποσότητα δεδομένων, χωρίς να επηρεάζεται από τον όγκο και την κατανομή και τελικά δύναται να λειτουργεί σε μεγάλη κλίμακα, εξετάζοντας τη βέλτιστη εφαρμογή του κάθε ελεγχόμενου περιβάλλοντος ενός φυσιολογικού όγκου δεδομένων. Το συναίσθημα ορίζεται διττά, από τη μια, η περιγραφή του πως αισθάνεται κάποιος και από την άλλη, πως βιώνει και σκέφτεται ένα γεγονός. Η έρευνα αυτή στοχεύει στην καθεαυτή μελέτη του συναισθήματος μιας οποιαδήποτε άποψης σε θέματα που αφορούν τη χρηματιστηριακή αγορά.

Οι απόψεις αυτές διαχωρίζονται σε θετικές και αρνητικές. Τα tweets και αυτά με τη σειρά τους ακολουθούν την συγκεκριμένη ορολογία και χωρίζονται σε positive/negative, με αποτέλεσμα να έχουν δυαδική διαδικασία κατηγοριοποίησης. Επιπλέον, η ανάλυση τους είναι συνώνυμη με αυτή της εξόρυξης γνώμης (opinion mining) και χαρακτηρίζεται από τη διαδικασία για να αναγνωριστεί το συναισθηματικό υπόβαθρο που διακατέχει ένα σώμα κειμένου. Η προβληματική ανάλυση του συναισθήματος, επομένως, προσεγγίζεται διαφορετικά για να επέλθει ανάλυση μέσω της προσέγγισης της μηχανικής μάθησης ή ευρέως κοινή ως machine learning. Ο όρος αυτός αποτελεί τον κλάδο της επιστήμης των υπολογιστών (computer science) που δίνει την ικανότητα της μάθησης στις υπολογιστικές μηχανές, προερχόμενη της εμπειρίας έτσι ώστε να υπάρχει προσομοίωση και έμπνευση από την ανθρώπινη εγκεφαλική λειτουργία.

Οι μελετητές έκαναν χρήση πολλών οικείων αλγόριθμων μηχανικής μάθησης για να έχουν μια ουσιαστική ανάλυση των συναισθημάτων. Με αυτό τον τρόπο, τα επικείμενα προβλήματα για εξαγωγή συναισθημάτων γίνονταν απλά προβλήματα ταξινόμησης και τα σύνολα δεδομένων από τα διαβαθμισμένα tweets ήταν η βάση της εκπαίδευσης των ταξινομητών, που τα έκαναν χρήση στην εξαγωγή των συναισθημάτων από τα μηνύματα.

1.5 Επεξεργασία φυσικής γλώσσας (NLP)

Οι NLP αλγόριθμοι που χρησιμοποιούνται σήμερα έχουν τα θεμέλια τους στην μηχανική μάθηση (machine learning), και ειδικότερα στην στατιστική μηχανική μάθηση. Κάποιες περασμένες εφαρμογές για την επεξεργασία γλώσσας περιείχαν μόνο την άμεση καταγραφή των μεγάλων συνόλων από κανόνες χειρόγραφα ενώ το πρότυπο της μηχανικής μάθησης, χρησιμοποιεί γενικούς αλγόριθμους μάθησης βασισμένους στη στατιστική συμπερασματολογία (statistical inference), έτσι ώστε η αυτόματη μάθηση των κανόνων να προκύπτει εκ της αναλύσεως των μεγάλων σωμάτων κειμένων (corpora) από τυπικά παραδείγματα. Επίσης, ένα σώμα (corpus, plural “corpora”) αποτελεί ένα σύνολο κειμένων ή κάποιες αυτόνομες προτάσεις με ορθές χειρόγραφες τιμές και μέσω αυτών γίνεται η εκπαίδευση του αλγορίθμου.

Στα προβλήματα NLP δοκιμάζονται διαφορετικοί τύποι αλγορίθμων μηχανικής μάθησης, οι οποίοι εισέρχονται σε μεγάλα σύνολα από χαρακτηριστικά (features), εφόσον έχουν επεξεργαστεί τα δεδομένα σε κάποια είσοδο. Χαρακτηριστικά, μέρος των πρώτων αλγορίθμων, όπως ήταν τα δέντρα-αποφάσεων, διέθεταν συστήματα από απόλυτους κανόνες 'if-then', όμοια εκείνων που προαναφέρθηκαν στην χειρόγραφη ανάθεση.

Οι πρόσφατες έρευνες επικεντρώνονται στα στατιστικά μοντέλα με απλή λήψη αποφάσεων, βάση πιθανοτήτων και υπολογισμό και ανάθεση πραγματικών τιμών-βάρη, πάνω στα χαρακτηριστικά εισόδου. Αυτά τα μοντέλα, έχουν περισσότερα πλεονεκτήματα σε ότι αφορά την έκφραση βεβαιότητας αντί για μια μοναδικής, για διάφορες πιθανές απαντήσεις, προσφέροντας πιο αξιόπιστα αποτελέσματα όταν το μοντέλο είναι μέρος ενός μεγαλύτερου συστήματος.

Κάποια από τα κάτωθι πεδία της έρευνας στην επιστήμη της NLP, χρίζουν άμεσης εφαρμογής σε πραγματικά προβλήματα, ενώ τα υπόλοιπα αξιοποίησης σε επιμέρους υπό-ενέργειες και για να επιλύονται μεγαλύτερα προβλήματα και είναι τα εξής:

- Η αυτόματη περίληψη (automatic summarization), με τη δημιουργία μιας ευανάγνωστης περίληψης από ένα μέρος κειμένου. Αυτό συμβαίνει κυρίως στη δημιουργία περιλήψεων από κείμενα γνωστού τύπου, όπως τα άρθρα ή οι οικονομικές αναλύσεις των εφημερίδων.

- Η μηχανική μετάφραση (machine translation), πάνω σε κείμενα από μια γλώσσα σε κάποια άλλη, αυτόματα. Θεωρείται από τα πιο σημαντικά προβλήματα, χαρακτηρισμένα από τον όρο 'AI-complete', καθώς αναφέρονται στα είδη γνώσης που χρήζουν επεξεργασίας από τους ανθρώπους λόγω έλλειψης γραμματικής, λεξιλογίου και σημασιολογίας για να θεωρηθούν κατάλληλα.

- Η παραγωγή φυσικής γλώσσας (natural language generation), όπου απαιτούνται βάσεις δεδομένων για να μετατραπεί η κάθε πληροφορία σε ευανάγνωστη ανθρώπινη γλώσσα.

- Η κατανόηση της φυσικής γλώσσας (natural language understanding), όπου μετατρέπονται κάποια τμήματα του κειμένου σε επίσημη αναπαράσταση. Παραδείγματος χάριν, οι πρώτης-τάξεως λογικές δομές, που μπορούν να τις διαχειριστούν πιο εύκολα τα υπολογιστικά προγράμματα.

- Η οπτική αναγνώριση χαρακτήρων (optical character recognition), όπου βάση μιας εικόνας εισόδου με τυπωμένους χαρακτήρες, αναγνωρίζει το αντίστοιχο κείμενο.

- Η απάντηση ερωτήσεων (question answering), όπου προκύπτει από την ανάλυση μιας ερώτησης σε ομιλούσα γλώσσα. Οι ερωτήσεις τυπικής φύσης έχουν μια πιο συγκεκριμένη ορθή ή καταλληλότερη απάντηση.

- Η ανάλυση συναισθημάτων (sentiment analysis), όπου κάποιες δημόσιες κριτικές (online reviews) και η εξαγωγή μιας υποκειμενικής πληροφορίας από ένα σύνολο κειμένων, καθορίζουν την «πολικότητα» των συγκεκριμένων ζητημάτων. Θεωρείται ως ένα σημαντικό χρήσιμο εργαλείο που αναγνωρίζει την κοινή γνώμη που επικρατεί στα κοινωνικά δίκτυα όπως και τις τάσεις αξιοποιώντας τα στους σκοπούς του Marketing.

- Η αναγνώριση ομιλίας (speech recognition), όπου γίνεται λεκτική αναπαράσταση της ομιλίας, μέσω ενός ηχητικού καταγεγραμμένου ντοκουμέντου ενός ομιλούντα ανθρώπου. Η μετατροπή του κειμένου σε ομιλία δεν είναι πρόβλημα τύπου «AI complete», γιατί στην φυσική ομιλία δεν

υπάρχουν παύσεις ενδιάμεσα στις διαδοχικές λέξεις. Με αυτόν τον τρόπο, η τμηματοποίηση της ομιλίας είναι μια απαραίτητη υπό-διαδικασία της αναγνώρισης. Τέλος, δύσκολη διαδικασία θεωρείται η μετατροπή του αναλογικού σήματος σε διακριτά σύμβολα, λόγω του ήχου των πολλών συνεχόμενων γραμμάτων που εμπλέκονται το ένα μέσα στο άλλο.

ΚΕΦΑΛΑΙΟ 2^{Ο} ΣΥΝΑΙΣΘΗΜΑΤΙΚΗ ΑΝΑΛΥΣΗ

2.1 Ορισμός και περιγραφή της Συναισθηματικής Ανάλυσης

Η ανάλυση συναισθήματος (sentiment analysis) και η εξόρυση γνώμης (opinion mining) αποτελούν ερευνητικά πεδία της ανάλυσης για την ταξινόμηση του κειμένου (text classification) και συχνά παρατηρούνται στις διαδικασίες εξαγωγής πληροφοριών που αφορούν τη συναισθηματική κατάσταση του γραπτού λόγου του χρήστη. Στην ουσία, είναι κάποιες αυτοματοποιημένες διαδικασίες που καταδεικνύουν την συναισθηματική πολικότητα ενός εγγράφου ή μιας φράσης. Επίσης, κάνει χρήση τεχνικών επεξεργασίας φυσικού λόγου (natural language processing - NLP), στατιστικών μεθόδων και μεθόδων μηχανικής μάθησης για να ταξινομηθεί ένα κείμενο σε κλάσεις που εκφράζουν ένα οποιοδήποτε συναίσθημα.

Γενικότερα, προκύπτει ότι η ανάλυση συναισθήματος έχει ως στόχο να εξακριβώσει το στοχεύει ύφος του ομιλητή ή του συγγραφέα σχετικά με κάποιο θέμα ή την πολικότητα του περιεχομένου κάποιου έγγραφου κειμένου ως σύνολο. Το ύφος του αφορά την εκτίμηση ή την κρίση του, τη συναισθηματική του κατάσταση ή το επικοινωνιακό συναίσθημα που θέλει να περάσει σκόπιμα στον αναγνώστη όταν θα διαβάζει το κείμενό του.

Η κατηγοριοποίηση της πολικότητας ενός κειμένου με το αν είναι θετική ή αρνητική η στάση του συγγραφέα, αποτελεί την κύρια εφαρμογή της ανάλυσης συναισθημάτων. Πέραν αυτής, υπάρχουν και κάποιες άλλες πιο σύνθετες τεχνικές, που περιγράφουν τις συναισθηματικές καταστάσεις του όπως τον θυμό, τη λύπη και τη χαρά. Η αξιολόγησή τους στο σύστημα ανάλυσης συναισθημάτων, βασίζεται στη συμφωνία τους με τη ανθρώπινη κρίση και καθορίζονται από τις κλασσικές μετρήσεις της ακρίβειας και της αξιοπιστίας ενός τέτοιου μηχανισμού. Σε κάποιες έρευνες, οι άνθρωποι που καθορίζονται σαν εκτιμητές αντίστοιχων προβλημάτων ανάλυσης, έχουν συμφωνία συνήθως στο 79% των περιπτώσεων.

Έτσι προκύπτει, ότι τα προγράμματα με 70% ακρίβεια περίπου έχουν την ίδια απόδοση με τους ανθρώπους. Επιβεβαιώνεται καθώς και ένα πρόγραμμα αν μπορούσαμε να πούμε ότι είναι 100% ακριβές, οι άνθρωποι και πάλι θα είχαν διαφωνία μαζί του γύρω στο 20% περίπου και αυτό βασίζεται στο γεγονός διαφωνίας κατά μέσο όρο για κάθε απάντηση.

Η αξιολόγηση της ανάλυσης συναισθημάτων παραμένει ένα πολύπλοκο ζήτημα, ακόμη και αν εφαρμοστούν πιο εξελιγμένες μετρήσεις. Η συσχέτιση (correlation) θεωρείται η καταλληλότερη μέθοδος μέτρησης της ακρίβειας διότι συνυπολογίζει τον βαθμό συνοχής των τιμών πρόβλεψης σχετικά με τον τελικό στόχο, για 20 διεργασίες ανάλυσης συναισθήματος μετατρέποντάς το σε μορφή κλίμακας (scale) και όχι σε δυαδικές αποφάσεις.

Γενικότερα ολόκληρος ο κλάδος της ανάλυσης συναισθημάτων διαμορφώνεται αναλόγως την εξέλιξη των κοινωνικών μέσων (social media) και ειδικότερα των κοινωνικών δικτύων (social networks) και των blogs. Η εξέλιξη του διαδικτύου ήταν ο καταλύτης στη μετάβαση της νέας γενιάς του παγκοσμίου ιστού, ευρέως γνωστή ως

web 2.0, παρέχοντας περισσότερες δυνατότητες στους χρήστες του διαδικτύου από κοινού να προβαίνουν σε ενέργειες αλληλεπίδρασης, συνεργασίας, διαμοιρασμό και ανταλλαγής πληροφοριών. Αυτή η νέα γενιά αυτή μπορεί να τα πραγματοποιήσει χωρίς να εξειδικεύεται σε γνώσεις υπολογιστών και δικτύων, αφού παρέχει τα πάντα κάθε διαδικτυακή πλατφόρμα.

Οι επιχειρήσεις υπολογίζουν αρκετά την διαδικτυακή γνώμη για την προώθηση των προϊόντων τους ή των υπηρεσιών τους, την εύρεση νέων ευκαιριών και τη βέλτιστη διαχείριση της φήμης τους στην αγορά, εξαιτίας μιας πληθώρας από κριτικές, συστάσεις αξιολογήσεις και άλλες μορφές online έκφρασης. Επίσης, η ανάλυση συναισθημάτων βοηθά τις επιχειρήσεις να βρίσκουν μορφές αυτοματοποίησης της διαδικασίας για την εξάλειψη του θορύβου, την αναγνώριση του περιεχομένου και την κατανόηση των συζητήσεων και στη λήψη ορθότερων δράσεων και αποφάσεων. Το επόμενο μελλοντικό στάδιο του διαδικτύου θα ήταν προτιμητέο να έχει τα θεμέλιά του στη δημοκρατικοποίηση της εξόρυξης γνώσης που προκύπτει από το περιεχόμενο οποιασδήποτε δημοσιευμένης πληροφορίας.

Παγκοσμίως, πανεπιστημιακές ερευνητικές ομάδες στοχεύουν να κατανοήσουν τη δύναμη και τη δυναμική του συναισθήματος στις e-κοινωνίες διαμέσου της τεχνικής της ανάλυσης συναισθημάτων, με μεγάλο τροχοπέδη την χρήση των αλγορίθμων που περιγράφουν με όρους απλούς το συναίσθημα για ένα προϊόν ή μια υπηρεσία. Το να μετατρέπεται ένα τμήμα ενός γραπτού κειμένου ή εγγράφου μόνο σε θετικό ή αρνητικό συναίσθημα, εξαιρεί πολλές διαφορετικές διακυμάνσεις που υπάρχουν ενδιάμεσα και αυτό συμβαίνει κυρίως επειδή υπάρχει αδυναμία να αποκρυπτογραφηθούν καθώς δεν μπορούν να ληφθούν οι πολιτισμικοί παράγοντες, οι γλωσσολογικές αποχρώσεις και ιδιαιτερότητες. Είναι βέβαιο, ότι ένας υπολογιστής δυσκολεύεται στην σωστή εξαγωγή

του μηνύματος, αφού και οι άνθρωποι οι ίδιοι δυσκολεύονται και όσο πιο συνοπτικό είναι το κείμενο, η δυσκολία του μεγαλώνει.

Η ανάλυση των συναισθημάτων γύρω από το microblogging, πραγματώνεται στο Twitter που θεωρείται αποδεδειγμένα σαν ένας online δείκτης του πολιτικού συναισθήματος με μεγαλύτερη αξιοπιστία, παρόλη την μικρή έκταση των κειμένων. Το πολιτικό συναίσθημα που περιέχουν τα tweets δείχνει την άμεση φανερώνει ανταπόκριση στην πολιτική καθώς και στις θέσεις και απόψεις πολιτικής φύσεως πολιτικών και κομμάτων και μελετητές έχουν αποδείξει ότι το offline πολιτικό σκηνικό αντανακλάται από και προς το περιεχόμενο των μηνυμάτων του Twitter, χωρίς αμφιβολία.

2.2 Κατηγοριοποίηση Κειμένου (Text Classification)

2.2.1 Ορισμός

Με τον όρο κατηγοριοποίηση κειμένου ονομάζεται κάθε διαδικασία ανάληψης προκαθορισμένων κατηγοριών σε έγγραφα ελεύθερου κειμένου, καθώς ο ρόλος της είναι σημαντικός στις πραγματικές εφαρμογές και η παροχή της εννοιολογικής εικόνας σπουδαία. Ένα παράδειγμα μιας τέτοιας διαδικασίας είναι η οργάνωση σε θεματικές κατηγορίες ή σε ενότητες και σε γεωγραφικούς κωδικούς. Επίσης, ακαδημαϊκές εργασίες μπορούν να ταξινομηθούν ανά τεχνικό κλάδο και υπό-κλάδους ή οι αναφορές ασθενών στους οργανισμούς υγειονομικής περίθαλψης να βασίζονται σε συγκεκριμένα χαρακτηριστικά.

Μια αρκετά γνωστή εφαρμογή της κατηγοριοποίησης ή ταξινόμησης του κειμένου είναι το 'spam filtering', καθώς εκεί τα e-mails χωρίζονται σε spam και non-spam κατηγορίες αντιστοίχως.

2.2.2 Είδος Συναισθήματος

Για να μπορέσουμε να κατηγοριοποιήσουμε την Ανάλυση Συναισθήματος αναλόγως της εξαγωγής του είδους του συναισθήματος, προκύπτουν τα εξής:

- Η Απλή Ανάλυση, όπου δείχνει αν μια δήλωση είναι θετική ή αρνητική, εξεταζόμενη για τη στάση της σε κάποιο θέμα και αναλύοντάς την απλά σε θετική ή αρνητική στάση. Σε κάποιες περιπτώσεις υπάρχουν τρεις κατηγορίες αφού προστίθεται και η ουδέτερη στάση ή πέντε βάζοντας και τον υπερθετικό βαθμό, π.χ. πολύ αρνητική- πολύ θετική αντίστοιχα.

- Η Ανάλυση Υποκειμενικότητας, όπου μια δήλωση εξετάζεται ως προς την υποκειμενικότητα ή αντικειμενικότητά της, γεγονός που την κατατάσσει δύσκολη στην ταξινόμηση. Η υποκειμενικότητα των λέξεων και φράσεων που περιέχονται σε ένα έγγραφο εξαρτάται από το πλαίσιό τους και υπάρχει πιθανότητα ένα αντικειμενικό έγγραφο να εμπεριέχει υποκειμενικές προτάσεις. Η παγκόσμια βιβλιογραφία υποστηρίζει ότι τα αποτελέσματα της βασίζονται στον ορισμό της υποκειμενικότητας που χρησιμοποιείται όταν χαρακτηρίζεται ένα κείμενο.

- Η Ανάλυση Συναισθηματικής κατάστασης, όπου το σύστημα προβαίνει στον εντοπισμό της συναισθηματικής κατάστασης του συγγραφέα για κάποιο θέμα. Αναλυτικότερα, η επιτυχία του αυτή βασίζεται στην ακριβή έννοια της συναισθηματικής κατάστασης που προσπαθεί να ορίσει. Στη wikipedia αυτό

αναλύεται ως τη γενική συναισθηματική κατάσταση του συγγραφέα την ώρα της συγγραφής του κειμένου (affective state), ή αλλιώς σαν το μεταδοτικό σκόπιμο συναίσθημα του συγγραφέα στον αναγνώστη του διαμέσου του κειμένου, ή μέσω σαν στάση/ άποψη / εκτίμηση του συγγραφέα για ένα θέμα. Εκτός από την τελευταία, οι υπόλοιπες δύο ταξινομούνται βάσει των κλάσεων έκφρασης συναισθήματος που είναι αντιληπτό από τον άνθρωπο, αναγνωρίζοντας δηλαδή τα πραγματικά συναισθήματα μες στο κείμενο.

2.3 Μέγεθος του κειμένου που εξετάζεται

Ένας άλλος διαχωρισμός που υπάρχει είναι αναλόγως του μεγέθους του κειμένου, ώστε να βρεθεί το συναίσθημα ή τη πολικότητά του (polarity) σε ολόκληρο το κείμενο (document-based sentiment analysis), σε μια πρόταση (sentence-based sentiment analysis) ή σε μεμονωμένες φράσεις (feature/aspect-based sentiment analysis) αναφερόμενες σε χαρακτηριστικά μιας οντότητας (features of an entity) όταν ψάχνουμε το συναίσθημα. Ακολουθούν αναλυτικά παραδείγματα.

1. Document-based sentiment analysis

Παρακάτω παρατίθεται μια κριτική ταινίας από την ιστοσελίδα imdb, η οποία είναι θετικής φύσεως και υπάρχει αναζήτηση του συναισθήματος. Εδώ, χρειάζεται document-based sentiment analysis. Αντίστοιχα παραδείγματα, είναι η άποψη ενός αρθρογράφου για ένα πολιτικό συμβάν μέσα από τον λόγο του στο άρθρο πολιτικής εφημερίδας ή στην άποψη ενός χρήστη για ένα προϊόν μέσα από την κριτική του στο διαδίκτυο.

«Οι πολεμικές ταινίες μεροληπτούν της μίας ή της άλλης πλευράς. Αυτή η ταινία δεν κάνει ήρωες ή εχθρούς τους Γερμανούς ναυτικούς με U-boat. Αντίθετα, συνεπαίρνει τον θεατή με τις ρεαλιστικές απεικονίσεις του πώς ήταν να είσαι ναύτης με U-boat για τους Γερμανούς στον Β' Παγκόσμιο Πόλεμο. Ξεκινά με νέους (από 17 έως 25 ετών) που έχουν γεμίσει με προπαγάνδα για την πολεμική προσπάθεια και την ένδοξη μάχη. Αφού αυτό το νεαρό πλήρωμα των ανώριμων ναυτικών αρχίσει να βιώνει τις αληθινές φρικαλεότητες του πολέμου, μπορείτε όχι μόνο να δείτε, αλλά και να βιώσετε μαζί τους την πλήξη, το γέλιο, τη συντροφικότητα, την ομαδικότητα και τον θάνατο. Σε έναν κόσμο όπου δεν υπάρχουν παράθυρα, όπου τα αυτιά σας πρέπει να είναι τα μάτια σας, όπου παίζεται το παιχνίδι της γάτας με το ποντίκι και ο χαμένος πεθαίνει, αυτοί οι νεαροί άντρες ηλικίας 10 έως 15 ετών κάνουν τον θεατή να συνειδητοποιήσει τη φρίκη του υποβρυχιακού πολέμου στο Β' Παγκόσμιος Πόλεμος. Η πιο ρεαλιστική πολεμική ταινία που έχω δει ποτέ ».

2. Sentence-based sentiment analysis

Άλλο ένα παράδειγμα είναι μία κριτική ταινίας δοσμένη σε μία πρόταση. Δύσκολα αναγνωρίζεται το συναίσθημα καθώς τα δεδομένα είναι πιο λίγα και αυτό που παίζει ρόλο είναι η σειρά και η σύνταξη των λέξεων. Η Sentence-based sentiment analysis γίνεται σε μικρό αριθμό προτάσεων, όπως εδώ ανήκουν και τα tweets.

«Είναι τόσο παλιό και νεανικό, που μόνο τα αγόρια στην εφηβεία θα μπορούσαν να το βρουν αστείο».

3. Feature/aspect-based sentiment analysis

Εδώ, υπάρχει η κριτική ενός smartphone από το διαδίκτυο και η οντότητα (entity) είναι ότι το smartphone και τα χαρακτηριστικά του είναι η οθόνη και η κάμερά του. Η

feature/aspect-based sentiment analysis αναγνωρίζει το συναίσθημα μέσα στις φράσεις της καθώς αναφέρονται στην κάμερα και στην οθόνη αντίστοιχα.

«Η οθόνη σε αυτό το smartphone είναι όμορφη με ρεαλιστικά χρώματα και μεγάλες γωνίες θέασης, ωστόσο η κάμερα θα μπορούσε να αποδώσει καλύτερα σε χαμηλό φωτισμό».

2.4 Η ανάλυση συναισθήματος στην εφαρμογή της

Η συναισθηματική ανάλυση είναι ένα νέο πεδίο έρευνας, όπου οι πρώτες προσπάθειες άρχισαν αρχές του 21ου αιώνα από τον Turney και τους Pang και Lee, ασκώντας μεγάλες προσπάθειες ταξινόμησης σε κατηγορίες μεγάλων σε μήκος κείμενα αναλόγως του συνολικού συναισθήματος που εκφράζουν. Ως τότε, υπήρχε ταξινόμηση κειμένων βάσει του θέματός τους (topic classification) και όχι του συναισθήματός τους. Έτσι, ο Turney έκανε μια προσπάθεια ταξινόμησης για διαδικτυακές κριτικές που αφορούσαν αυτοκίνητα, ταινίες, τράπεζες και ταξιδιωτικούς προορισμούς βασισμένα πάνω σε στατιστικές μεθόδους ενώ οι Pang και Lee πάνω σε κριτικές για ταινίες βασισμένες σε κλασσικούς αλγόριθμους μηχανικής και τα αποτελέσματα αυτών ήταν ικανοποιητικά για το topic classification. Η ανάλυση συναισθήματος, λοπόν, είναι ένα πεδίο που ακόμα διερευνάται και ακόμα μετά από μια δεκαπενταετία τραβάει την προσοχή χάριν στους παρακάτω παράγοντες:

1. Στην εκρηκτική εξέλιξη του διαδικτύου, ιδίως με τον Web 2.0. στην εποχή μας, το Internet, οι υπολογιστές και τα smartphones δίνουν τη δυνατότητα στους ανθρώπους παγκοσμίως έχουν πρόσβαση στο διαδίκτυο και στα πολλά δεδομένα του, όπως εικόνες, βίντεο και κείμενα, γνωστά ως big data. Το Internet πλέον είναι αναπόσπαστο κομμάτι της ζωής όλων των ανθρώπων και η σημαντικότερη πηγή ευρέσεως οποιασδήποτε πληροφορίας ή

κοινωνικοποίησης μέσω των κοινωνικών δικτύων. Τα κοινωνικά δίκτυα, Instagram, Facebook και Twitter, προσφέρουν την ανωνυμία των χρηστών και παρέχουν δυνατότητες σύνδεσης και επικοινωνίας παγκοσμίως μέσω του διαδικτύου. Από πλευράς ανάλυσης συναισθήματος σε ερευνητική φάση, βοηθούν τους ερευνητές να προβούν σε τεράστιους όγκους δεδομένων. Μεγάλη πρόοδος έχει γίνει πάνω στην υπολογιστική όραση (computer vision), τα τελευταία χρόνια, αφού πολλές εταιρείες όπως μια εξ αυτών και η Google εφαρμόζει αλγορίθμους σε μεγάλους όγκους δεδομένων.

2. Στην αύξηση των υπολογιστικών πόρων που διατίθενται στο μέσο χρήστη. Η ανάλυση συναισθήματος και η μηχανική μάθηση, γενικότερα, πέραν των πολλών δεδομένων, χρειάζονται και υπολογιστικούς πόρους. Η συνεχής και αυξανόμενη επεξεργαστική ισχύς των υπολογιστών περνά τους περιορισμούς της σημερινής τεχνολογίας. Ο κάθε χρήστης μπορεί σήμερα να χρησιμοποιεί απαιτητικούς αλγορίθμους στον προσωπικό του υπολογιστή ή και σε υπερυπολογιστές (super computers, gpu clusters) των εταιριών.

Συμπερασματικά προκύπτει ότι η πρόοδος της τεχνολογίας στους υπολογιστές και η πρόσβαση στα πολλά δεδομένα του διαδικτύου, έχουν τραβήξει την προσοχή πρόσφατα στην μηχανική μάθηση και στην τεχνητή νοημοσύνη. Η ανάλυση συναισθήματος είναι διαθέσιμη στον κάθε χρήστη και κυρίως στις εταιρείες, τους οργανισμούς και τους ερευνητές που το χρησιμοποιούν κατά κόρον.

Άλλες εφαρμογές χρήσης είναι οι κάτωθι:

- Οι Αλγόριθμοι ανάλυσης συναισθήματος από εταιρείες στα δεδομένα του διαδικτύου για να εξάγονται πληροφορίες που στοχεύουν στην ευρεία αποδοχή των προϊόντων τους από τους καταναλωτές. Έτσι, λαμβάνονται feedback για

την ποιότητα των προϊόντων και των υπηρεσιών τους (business intelligence) με ικανοποιητικά αποτελέσματα.

- Η διαδεδομένη αποδοχή των κοινωνικών δικτύων σήμερα, δείχνει ότι τα κοινωνικά δίκτυα εκφράζουν αρκετά την κοινωνία, χωρίς όμως να είναι κάτι το απόλυτο. Η ανάλυση συναισθήματος εξ αυτών βοηθά για την εξόρυξη πληροφοριών για τη στάση και σκέψη της κοινής γνώμης πάνω στα φλέγοντα ζητήματα, είτε είναι πολιτικά, κοινωνικά είτε οικονομικά.

ΚΕΦΑΛΑΙΟ 3^{Ο} ΤΟ TWITTER

3.1 Γενικές πληροφορίες για το Twitter

Μέσω των microblogging πλατφόρμων που αφορούσαν τα κοινωνικά μέσα, μια πολύ δημοφιλής πλατφόρμα είναι το Twitter, με έτος ίδρυσης το 2006. Το Twitter αποτελέι ένα δίκτυο πληροφοριών πραγματικού χρόνου που παρέχει στους χρήστες σύνδεση με τις πιο πρόσφατες πληροφορίες για ζητήματα που τους αφορούν ή ενδιαφέρουν, επιτρέποντας το διαμοιρασμό των posts (δημοσιεύσεις). Αυτό συμβαίνει εφόσον ένας χρήστης ακολουθεί κάποιον άλλον- από φυσικό πρόσωπο έως κυβερνητικές σελίδες- λαμβάνοντας τις ενημερώσεις για τα posts τους αμέσως.

Ο κεντρικός άξονας του Twitter είναι το "tweet", όπου και ονομάζεται tweet η κάθε δημοσίευση, η οποία περιέχει το μέγιστο 140 χαρακτήρες. Αυτός ο περιορισμός χαρακτήρων, προσφέρει συντομία στους χρήστες, κάνοντας ένα tweet να είναι μια συμπυκνωμένη έκρηξη από πληροφορίες, εύκολες στην ανάγνωση τους και στην

παρακολούθησή τους και συνεπώς, προσφέρουν στατιστικά στοιχεία για τις κοινωνικές τάσεις εύκολα.

Οι εκατομμύρια χρήστες του Twitter, προσφέρουν μια σημαντική πηγή υπηρεσιών για το Παγκόσμιο Ιστό μιας και μέσα στην πλατφόρμα εμπεριέχονται πολλά δεδομένα και ενδιαφέρον για τα δεδομένα αυτά για παρατήρηση και μελέτη.

3.2 Αλληλεπίδραση Twitter - Ανθρώπου

Πιο συγκεκριμένα το Twitter διαθέτει γύρω στους 200 εκατομμύρια εγγεγραμμένους χρήστες, ενώ καταγράφονται 400 εκατομμύρια επισκέπτες ανά μήνα και 50 εκατομμύρια χρήστες ανά ημέρα. Ακόμα, έχει εκτιμηθεί ότι γύρω στο 1 δισεκατομμύριο tweets δημοσιεύονται από τους χρήστες του Twitter, αριθμός που προκύπτει για κάθε πέμπτη ημέρα.

Στη δημοφιλή πλατφόρμα του Twitter, πολλές προσωπικότητες από τον καλλιτεχνικό και πολιτικό τομέα έχουν γίνει εγγεγραμμένα μέλη της, προσπαθώντας να μεγαλώσουν τη δημοτικότητα και την ευαισθητοποίησή τους και να είναι ενεργοί στα παρόντα γεγονότα. Το Twitter αποτελεί μια ατέρμονη πηγή πληροφοριών που υπάρχει σε πραγματικό χρόνο για τις παρούσες απόψεις και τάσεις της κοινωνίας και αυτό επαφίεται στο γεγονός ότι τα εκατομμύρια άτομα δημοσιεύουν τις απόψεις τους για οτιδήποτε μπορεί να φανταστεί κανείς, από την ακριβή τιμή της βενζίνης μέχρι και την εισβολή της Ρωσίας στην Ουκρανία.

Οι άνθρωποι μέσω του Twitter μοιράζονται εμπειρίες και συναισθήματα με τους λογαριασμούς που εμπλέκονται, εξάγοντας τα συναισθήματα τους γεγονός που

αντικατοπτρίζει την επικρατούσα άποψη, αφού τα Twitter posts που φέρουν το συναισθηματικό φορτίο του χρήστη, ενημερώνουν και για την συναισθηματική τους κατάσταση. Μέσω της συνεχής ροής δεδομένων γίνονται οι αντιδράσεις σε προϊόντα, υπηρεσίες, κινήματα ή πολιτικές δραστηριότητες, ονομάζοντας κάποια από αυτά. Τέλος, έχει προκύψει ότι υπάρχει μεγάλη επιρροή από αυτές τις τάσεις και τις απόψεις.

3.3 Προτίμηση Twitter από άλλα κοινωνικά δίκτυα

Τα σύρματα δικτύων (news wires) έχουν παρόμοια λειτουργία με αυτή του Twitter, διαφέροντας στο γεγονός ότι οι δημοσιεύσεις προέρχονται από οργανώσεις ειδήσεων. Ο κάθε πάροχος μεταδίδει απευθείας τα ρεύματά (streams) προς τους συνδρομητές, καθώς και άλλες υπηρεσίες τρίτων με μετάδοση πληροφοριών στους τελικούς χρήστες.[10] Οι περισσότεροι μεταδίδουν τη ροή ειδήσεων (News streams) στους ιστοτόπους του, οι οποίες βάση ακαδημαϊκής έρευνας παίζουν ρόλο στις χρηματοπιστωτικές αγορές, καθώς αποτελούν σημαντική πηγή πληροφόρησης για όσους συμμετέχουν στην αγορά και παρέχουν αξιόπιστα και έγκαιρα νέα.

Ως προκάτοχους των σημερινών κοινωνικών μέσων θεωρούνται τα φόρουμ συζητήσεων και ενώ υπάρχουν μέχρι και σήμερα, λόγω των περιορισμένων τεχνολογιών και υποδομών προόδου έχουν οδηγήσει τους χρήστες στα πιο εξελιγμένα κοινωνικά μέσα όπως το Twitter, που παρέχει συγκλονιστικό σχεδιασμό και ορθή δομή για γρήγορη και πετυχημένη ανταλλαγή πληροφοριών. Τα φόρουμ συζητήσεων ζητούν από τους χρήστες τους την ενεργή ενασχόληση με θέματα, τα υπόλοιπα σύγχρονα μέσα κοινωνικής δικτύωσης κάνουν προώθηση των πληροφοριών στους χρήστες βασισμένα σε συγκεκριμένα κριτήρια.

Σε σχέση με τα φόρουμ, τα σύγχρονα κοινωνικά μέσα όπως το Twitter, χαίρει τη εκτίμηση των χρηστών τους καθώς οι χρήστες που διαθέτουν χρήσιμες πληροφορίες είναι ορατοί και οι αλληλογραφίες που θεωρούνται ανεπιθύμητες κρύβονται. Έτσι, οι αξιόπιστες πηγές που προέρχονται από τα μεγαλύτερα πρακτορεία ειδήσεων, από τις κυβερνήσεις και από μεγάλες εταιρείες κάνουν επαλήθευση των λογαριασμών έτσι ώστε να κερδίζεται η εμπιστοσύνη των χρηστών χρήστες ότι οι πληροφορίες σε tweets υποστηριζόμενες από αυτές τις πηγές δεν είναι μόνο ένα ψευδώνυμο, αλλά κάτι παραπάνω.

Το Twitter, λοιπόν, δεν τραβά το ενδιαφέρον λόγω των μεμονωμένων χρηστών, αλλά επειδή εμπεριέχει μεγάλους οργανισμούς, γνωστές επιχειρήσεις και δημόσιες υπηρεσίες. Οι Kaplan και Haenlein κάνουν λόγο για το επιτυχημένο Twitter, χάριν στα μοναδικά επικοινωνιακά χαρακτηριστικά του, και πιο συγκεκριμένα για «τη δημιουργία ευαισθησίας ως προς το περιβάλλον, τη δημιουργία μιας μοναδικής μορφής push-push-pull επικοινωνίας και τη δυνατότητα να χρησιμεύσει ως πλατφόρμα για εικονικό εκθεσιασμό και voyeurism».

Οι Whinston και Rui τονίζουν ότι «η μοναδική καινοτομία των κοινωνικών μέσων είναι η αναγνώριση και σύνδεση της ανάγκης των ανθρώπων για ενημέρωση και προσοχή» κι έτσι, ο σχεδιασμός τους οφείλει στην διευκόλυνση μιας τέτοιας σύνδεσης. Επίσης, οι ίδιοι έχουν διαπιστώσει ότι αν υπάρχει περίπτωση ένας χρήστης να γίνει παραγωγός περιεχομένου ή καταναλωτής, αυτό θα εξαρτηθεί από τη σχέση ανάμεσα στους μισθούς κράτησης για να γίνει παραγωγός ή καταναλωτής, με τον κοινοτικό μισθό για την παραγωγή περιεχομένου. Ένας χρήστης θεωρείται παραγωγός περιεχομένου αν το κέρδος υπερβαίνει το κόστος.

Ο Bruns δήλωσε ότι ο τρόπος που λειτουργεί το Twitter έπαιξε καταλυτικό ρόλο στην επιτυχία της έως τώρα, αλλά κρούει τον κώδωνα του κινδύνου ότι η ισορροπία ανάμεσα στις ανάγκες που έχουν οι πάροχοι της πλατφόρμας, οι χρήστες και οι τρίτοι προγραμματιστές είναι αρκετά σημαντικές για να διατηρηθεί η καινοτομία και η ανάπτυξη των κοινωνικών μέσων.

Πέραν της κλασσικής του χρήσης, το Twitter έχει θεωρηθεί και ως κανάλι επικοινωνίας όταν συμβαίνουν έκτακτα περιστατικά και ειδικότερα όταν η πρόσβαση στα κοινωνικά μέσα μέσω κινητού ή άλλης συσκευής είναι η πιο κατάλληλη εναλλακτική για επικοινωνία. Οι Hughes και Palen έχουν δηλώσει ότι η χρήση των μικρο-blogs σαν δημόσιου κανάλια πληροφοριών γίνονται χρήση από τις αρχές, όταν υπάρχουν καταστάσεις έκτακτης ανάγκης.

Ακόμα, έχουν αυξηθεί οι εμπορικές επιχειρήσεις που στοχεύουν το Twitter ως μέσο για να προσεγγίσουν τους καταναλωτές και με τη σειρά του το Twitter έχει αυξήσει τα διαφημιστικά του έσοδα, πράγμα που βοηθά στην βελτιστοποίηση των λειτουργιών του. Όμως, οι ίδιοι οι χρήστες του, είναι αυτοί που αποτελούν τη βάση της ανάπτυξης του και οι φοβερές δυνατότητες που προσφέρει, όπως η ταυτοποίηση ονομάτων χρηστών προσθέτοντας @ ή προσδιορισμό θεμάτων και λέξεων-κλειδιών μέσω των hashtags (και αργότερα, cashtags).

3.4 Εισαγωγή στην πρόβλεψη των δεικτών της χρηματιστηριακής αγοράς

Πολλοί επιθυμούν να μπορούν να προβλέψουν την χρηματιστηριακή αγορά και γι' αυτό το λόγο έχουν αναπτύξει πολυάριθμα μοντέλα, τα οποία κάνουν επιστροφές σε προηγούμενα δεδομένα, αλλά δυστυχώς έχουν αποτύχει σε περαιτέρω προσπάθειες.

Αυτό εξηγείται από το γεγονός ότι η αγορά έχει δυνατότητα προσαρμογής στις νέες μεθόδους, ούτος ώστε να μην είναι πλέον κερδοφόρες πράγμα που δεν είναι ενθαρρυντικό για τους επενδυτές να κάνουν επιτυχημένο διαμοιρασμό και δυνητικά κερδοφόρα μοντέλα.

Η ανάλυση για τις χρηματιστηριακές αγορές αποτέλεσε μία από τις πρώτες εφαρμογές των ηλεκτρονικών υπολογιστών. Ο Maurice Kendall το 1953 δήλωσε ότι «δεν μπορούσε να εντοπίσει προβλέψιμα πρότυπα στις τιμές των μετοχών. Οι τιμές φαινόταν να εξελίσσονται τυχαία". Αυτά, λοιπόν, τα ευρήματα ανέπτυξαν τη θεωρία της αποδοτικής αγοράς, στην οποία τα αποθέματα καταδεικνύουν όλες τις διαθέσιμες πληροφορίες, αδυνατώντας να προβλεφθεί η κίνηση τους βάσει των προηγούμενων δεδομένων.

3.5 Η δομή ενός tweet

Το tweet είναι ο πυρήνας της πλατφόρμας, αφού αποτελεί και τον μοναδικό τρόπο δημοσίευσης και μεταφοράς της πληροφορίας σε όλους τους χρήστες, εκτός των προσωπικών μηνυμάτων. Διαθέτει και μια χαρακτηριστική δομή εκτός του περιοριστικού του μέγεθος. Αρχικά, γίνεται η χρήση των σύμβολων “#” και “@”, όπου το # λέγεται «hashtag» και η χρήση του κάνει αναφορά για κάποιο θέμα (topic), προσφέροντας τη δυνατότητα να ομαδοποιηθούν πολλά δημόσια tweet για ένα θέμα. Με το χαρακτήρα @, αναφερόμαστε σε κάποιον άλλον χρήστη (πχ. @ben_voyl).

Twitter	The brand and company
Tweet	An up to 140 character long text message
Firehose	Constant stream of all tweets in real time
Hashtag (#)	Identifies a topic (e.g. #earnings)
At (@)	Identifies a username (e.g. @CNBC)
Cashtag ($)	Identifies a stock ticker (e.g. $GS)
Followers	Users who subscribe to tweets sent by a user
Mentions	Number of times others mention a specific username
Retweet	When another user relays your tweet to their followers

Ένα άλλο σύμβολο που υπάρχει συχνά στην αρχή των tweets, είναι το “RT”, που σημαίνει «retweet» και μεταφράζεται ως η αναμετάδοση κάποιου tweet ενός χρήστη, από κάποιον άλλο χρήστη. Συχνά, παρατηρείται και το url στα tweets, όταν κάνουν επιπλέον αναφορά σε κάτι, δίνοντας μια λύση στην περιορισμένη δυνατότητα για ανάλυσης κάποιου θέματος σε τόσο μικρό χώρο.

Τα σύμβολα Hashtags, at και cashtags χρησιμοποιούνται και για να τροποποιούν το κείμενο για να δημιουργήσουν δομή. Κάθε φορά που μια λέξη έχει χαρακτηριστεί με κάποιον από τους τροποποιητές αυτούς, βάζοντας το σύμβολο ακριβώς μπροστά του (π.χ. #earnings), οι χρήστες με ένα κλικ βρίσκουν τα σχετικά tweets. Με ένα κλικ στο όνομα του χρήστη, μπορεί κανείς να δει το προφίλ του χρήστη και όσα tweets είχε δημοσιεύσει. Με ένα κλικ σε μια ταμειακή ετικέτα ή ένα hashtag, μπορεί κανείς να δει τα πιο πρόσφατα tweets που αναφέρονται στην ετικέτα. Η κάτωθι εικόνα δείχνει τους κύριους όρους και έννοιες του Twitter.

Οι Honeycutt και Herring(2009) έχουν χωρίσει τα tweets σε 10 κατηγορίες , οι οποίες είναι :

1. Το μήνυμα του αποδέκτη
2. Η ανακοίνωση/διαφήμιση
3. Η πηγή πληροφοριών στους άλλους

4. Οι πληροφορίες για τον αποστολέα
5. Ο σχολιασμός στα άλλα σχόλια
6. Τα αναρτημένα tweets για τα μέσα μαζικής ενημέρωσης
7. Οι γνώμες των χρηστών
8. Η κοινοποίηση
9. Η συλλογή πληροφοριών
10. Οι υπόλοιπες λειτουργίες

Εφόσον, κάθε tweet αποτελεί μια συνοπτική περίληψη της γνώμης ή της διάθεσης κάθε ατόμου σχετικά με κάποιο θέμα, τα συνολικά tweets για ένα θέμα εκφράζει τη συλλογική διάθεση. Συνεπώς, η δημόσια διάθεση κάνει συσχετισμό ή πρόβλεψη των οικονομικών δεικτών.

3.6 Twitter Applications

Το twitter αντλεί τις πληροφορίες, όχι τυχαία, αφού το Twitter έχει το γνώρισμα της πολυφωνίας και είναι μια τεράστια πηγή πληροφοριών. Η Πλατφόρμα έχει διαμορφωθεί έτσι ώστε να είναι φιλική στον χρήστη και στον προγραμματιστή αφού διαθέτει απλό και κατανοητό API.

Twitter ως μία πλατφόρμα για developers

Το Twitter μπορεί να δημιουργήσει third-party εφαρμογές με στόχο την συνεχή και την άνετη εκμετάλλευση του μεγάλου όγκου της διατιθέμενης πληροφορίας. Πιο συγκεκριμένα, μέσω της ασφαλής διαδικασίας μπορεί να πιστοποιήσει αυτή την εφαρμογή παρέχοντας δικαιώματα για πρόσβαση σε APIs από έναν ήδη λογαριασμό του στο Twitter, χωρίς να χρειάζεται να χρησιμοποιεί τα αντίστοιχα credentials του

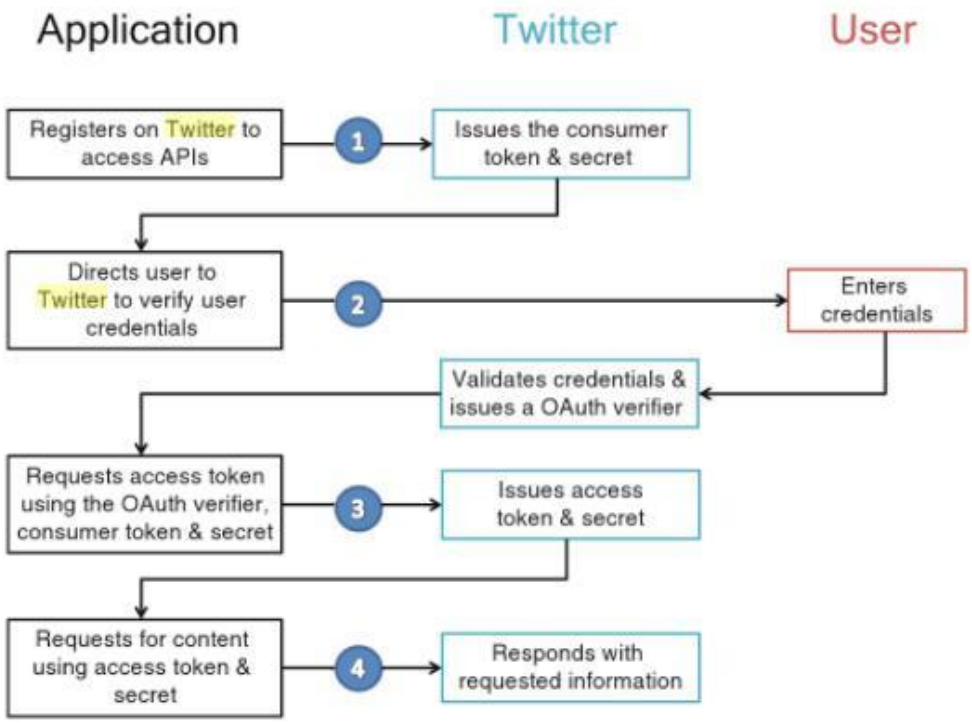
χρήστη. Μέσω developers, κάποιος μπορεί να κάνει επιλογή δημιουργίας μιας τέτοιας εφαρμογής, ονομάζοντάς το, περιγράφοντας την εφαρμογή που θέλει να φτιάξει, ένα (placeholder) website-url για την επικείμενη ιστοσελίδα της εφαρμογής αυτής και ένα callback-url για να γνωρίζει το Twitter που θα αποστέλλει τις απαντήσεις για να πιστοποιηθούν και τέλος τα ζητούμενα data από κλήσεις των παρεχόμενων υπηρεσιών. Κατόπιν, γίνεται ρύθμιση των δικαιωμάτων που διαθέτει η εφαρμογή αυτή στον λογαριασμό ενός χρήστη και η διαδικασία ολοκληρώνεται ρυθμίζοντας όλες τις παραμέτρους για την πιστοποίηση της μέσω web.

3.7 Χρήση του Twitter Oauth για πιστοποίηση

Για να μπορούν να πιστοποιηθούν οι εφαρμογές, το Twitter κάνει χρήση του OAuth, μια λειτουργία πιστοποίησης εφαρμογής (application-only), όπου πραγματοποιείται api calls αυτόνομα χωρίς την ανάμειξη του χρήστη. Μέσω της 'κατ' εξουσιοδότηση πρόσβασης' στους πόρους της εφαρμογής, προσφέρεται εξασφάλιση εκ μέρους του ιδιοκτήτη των πόρων, μέσω της λειτουργίας OAuth στάνταρντ. Πιο συγκεκριμένα, αυτό είναι που καθορίζει την διαδικασία εξουσιοδότησης που οφείλεται να ακολουθείται από τους ιδιοκτήτες των πόρων για να παραχωρείται η πρόσβαση σε τρίτους σε αυτόν τον πόρο, χωρίς διαμοιρασμό και αποκάλυψη των προσωπικών κωδικών τους που προσφέρουν πρόσβαση σε αυτόν τον πόρο.

Η χρήση του μηχανισμού αυτού, επίσης, γίνεται και από άλλες γνωστές και ευρέως διαδεδομένες εφαρμογές του Παγκοσμίου Ιστού, για να μπορούν οι χρήστες να μοιράζονται πληροφορίες από τους ίδιους τους λογαριασμούς τους, μέσω third-party εφαρμογών ή ιστοσελίδων. αναλυτικότερα, το OAuth έχει σχεδιαστεί βάσει HTTP πρωτόκολλου, επιτρέποντας στην διαδικασία να δώσει ειδικά "διακριτικά πρόσβασης"

σε τρίτους, με πιστοποίηση της εξουσιοδότησης από τον ιδιοκτήτη του πόρου. Μπορεί και να υπάρχει πρόσβαση στον ιδιωτικό πόρο από κάποιον τρίτο εάν διαθέτει τα απαραίτητα διακριτικά πρόσβασης που έχει παραχωρηθεί από τον ιδιοκτήτη του συγκεκριμένου πόρου. Το στάνταρντ OAuth δίνει στο Twitter δύο τρόπους πιστοποίησης, από τη μια την πιστοποίηση χρήστη και από την άλλη την πιστοποίηση εφαρμογής. Η πρώτη είναι η πιο συνηθισμένη μορφή πιστοποίησης στο Twitter έως σήμερα.



Το αίτημα που του έχει δοθεί δυνατότητα για να χρησιμοποιηθεί ο πόρος, δίνει πιστοποίηση της ταυτότητας της εφαρμογής όπως και την ταυτότητα του χρήστη που από τον ίδιο προσφέρεται πρόσβαση στους πόρους, όπου τοTwitter API κάνει χρήση της.

Η πιστοποίηση εφαρμογής, που θεωρείται και η ίδια μορφή πιστοποίησης, η εφαρμογή κάνει API αιτήματα εκ μέρους της και όχι από κάποιον άλλον χρήστη. Οι αιτήσεις στο API έχουν περιορισμένες δυνατότητες για κάποιες μεθόδους, όμως το φάσμα της πληροφορίας μπορεί να φτάσει στη μέγιστη δυνατή ικανότητα. Οι API αιτήσεις που κάνουν χρήση αυτής της μεθόδου πιστοποίησης, έχουν περιορισμό δύο ποσοστιαίων ορίων. Το πρώτο είναι ανα χρήστη και το δεύτερο ανά εφαρμογή, χάριν στις απαιτούμενες αιτήσεις πιστοποίησης χρήστη.

Ο ορισμός του ποσοστιαίου ορίου ανά χρήστη, πραγματοποιείται από το API στις αιτήσεις και αντιστοίχως στον όγκο των δεδομένων που διατίθενται όποτε χρειαστεί μέσω της εφαρμογής στον χρήστη. Δηλαδή, το όριο διατίθενται ανά "διακριτικό πρόσβασης". Το Twitter βάζει αυτά τα όρια για περίπου δεκαπέντε λεπτά και επιτρέπει δεκαπέντε αιτήσεις ανά δεκαπέντε λεπτά και 180 κλήσεις ανά δεκαπέντε λεπτά για τις μεθόδους αναζήτησης.

Η διαδικασία, λοιπόν, εφαρμόζεται ως κωδικοποιητής και αποστολέας του Twitter το consumer key & consumer secret προερχόμενα από της ίδια την εφαρμογή και κατόπιν η πληροφορία αποστέλεται στο OAuth ώστε να έχει μια σκυτάλη πρόσβασης (access token) και εντέλει να δημιουργεί μια pin-based χειραψία (handshake) και το Twitter να κάνει χρήση των virtual credentials. Ο χρήστης θα μεταβεί αυτομάτως στην σελίδα του λογαριασμού του όπου υπάρχει επερώτηση αποδοχής και εντολής της αντίστοιχης πιστοποίησης στην εφαρμογή που το επιζητά, και μέσω της αποδοχής, λαμβάνεται το pin που κάνει το ίδιο χρήση της εφαρμογής και ολοκληρώνει την διαδικασία. Οι κλήσεις που γίνονται στην παραγωγή των tokens, όπως και σε όλα τα api calls στο Twitter, κάνουν απαραίτητα χρήση SSL για να υπάρχει η ασφάλεια που απαιτείται και έτσι τα αιτήματα απόκτησης και χρήσης των δεδομένων ασφαλείας οφείλουν τη χρήση των https endpoints.

Twitter Api

Το Twitter έχει και τις κάτωθι κατηγορίες μεθόδων:

- Timeline Methods
- Status Methods
- OAuth Methods
- List Members Methods
- Direct Message Methods
- Notification Methods
- Block Methods
- Spam Reporting Methods
- Saved Searches Methods
- Favorite Methods
- Friendship Methods
- User Methods
- List Methods
- Account Methods
- Trends Methods
- List Subscribers Methods
- Social Graph Methods
- Geo methods
- Help Methods

3.8 Διεπαφές του Twitter – REST vs Streaming API

Χρειάζεται να υπάρχει μια διαδικασία για την αναζήτηση των επιθυμητών tweets και να δημιουργηθούν τα αναγκαία datasets για την ανάλυση μας. Έτσι, το Twitter διαθέτει 2 ξεχωριστές διεπαφές APIs, όπου μπορούν να χρησιμοποιηθούν δημιουργώντας έναν απλό web service client και θέτοντας κάποια χαρακτηριστικά στο αίτημα που θα αποσταλεί. Οι διεπαφές αυτές είναι: η REST (ή search) API, και η Streaming API.

Οι παραπάνω διεπαφές έχουν κάποιες ικανότητες και στόχος τους είναι και η εξυπηρέτηση κάποιων αναγκών ανά περίπτωση. Είναι γνωστό ότι και οι δύο ζητούν πιστοποίηση από το OAuth, αφήνοντας να γίνει εισαγωγή κάποιου searchkeyword (πχ. κάποιο #hashtag) για να περιοριστεί η αναζήτηση του, ή εισαγωγή κάποιου χρήστη ή λογαριασμού, με συγκεκριμένη χρονική περίοδο και περιορισμένη γεωγραφική αναζήτηση. Τότε, τα δεδομένα θα επανέλθουν σε json format, και θα δώσουν τη σχετική πληροφορία συγκεκριμένων tweets.

Η API REST είναι η βέλτιστη στην αρχιτεκτονική REST για την σχεδίαση API web και γίνεται με τη στρατηγική έλξης για να ανακτηθούν τα δεδομένα. Για να μπορούν να συλλεχθούν πληροφορίες, ο χρήστης οφείλει να το ζητήσει. Η REST παρέχει δυνατότητες για πιο συγκεκριμένα search-queries, προσφέροντας πολλαπλές επιλογές φιλτραρίσματος των αποτελεσμάτων και να επιστραφούν tweets από πιο πολλούς χρήστες.

Η API streaming δίνει μια συνεχόμενη ροή δημόσιων πληροφοριών από το Twitter και κάνει χρήση του push for stratedy για να ανακτηθούν τα δεδομένα. Με την αίτηση για πληροφορίες, η Streaming API δίνει τη συνεχή ροή ενημερώσεων χωρίς περισσότερες πληροφορίες από το χρήστη. Δηλαδή, η Streaming API διαθέτει μια σταθερή σύνδεση με το twitter, μεγαλώνοντας τη ροή για περισσότερα δεδομένα όσο παραμένει ανοιχτή

ή ώσπου υπάρχει ικανοποίηση του επιθυμητού μέγιστου, ορίζοντας την αναζήτηση των tweets.

Εκτός από τις παραπάνω διαφορές των 2 διεπαφών, υπάρχει και μια διαφοροποίηση στα όρια της αναζήτησης και στον αριθμό των tweets που μπορούν να αποκτηθούν και από την καθεμιά. Η streaming api φέρνει μεγαλύτερο αριθμό από tweets, ως επί των πλείστων. Επίσης, είναι γνωστά και τα εξής στοιχεία, ότι η ροή των δεδομένων προκύπτει από το 1% του συνόλου των tweets μιας συγκεκριμένης κατανομής και με μεγαλύτερο αριθμό γύρω στα 3.000 tweets/λεπτό, περιορισμός που βοηθά στην άντληση τους γύρω στα180.000 tweets την ώρα κατά προσέγγισει. Η rest api, αντιθέτως, έχει όριο 100 tweets/αναζήτηση και αφήνει γύρω στα 720 αιτήματα την ώρα, με μέγιστο αριθμό 72.000 tweets την ώρα. Φυσικά, είναι κάτι που ισχύει ανά χρήστη.

3.9 Json

Παραπάνω ειπώθηκε πως τα δεδομένα επιστρέφουν κυρίως σε json format, δίνοντας κάθε πληροφορία που έχει σχέση με συγκεκριμένα tweets. Το JavaScript Object Notation ή αλλιώς JSON, αποτελεί ένα open-standard format, δηλαδή μία λειτουργία μορφοποίησης, όπου το κείμενο μορφοποιείται κατανοητά για τον άνθρωπο, για να μεταφερθούν δεδομένα, που είναι ζευγάρια ιδιοτήτων και τιμών. Το JSON είναι η πιο συνηθισμένη μορφοποίηση για την ασύγχρονη επικοινωνία ανάμεσα σε φυλλομετρητή (browser) και σε εξυπηρετητή (server), κάνοντας αντικατάσταση της XML (Extensible Markup Language) μέσω των τεχνικών AJAX (Asynchronous JavaScript and XML).

Το JSON είναι μία μορφοποίηση, ανεξάρτητη της γλώσσας που την επεξεργάζεται. Τα αρχεία που εμπεριέχουν δεδομένα μορφής JSON έχουν κατάληξη json. Παρόλο, όμως, που πορεύεται από το JavaScript, οι πλειοψηφία των υψηλών γλωσσών

προγραμματισμού σήμερα, παρέχουν τις κατάλληλες βιβλιοθήκες για να δημιουργούνται και να αναλύονται τα δεδομένα σε μορφοποίηση JSON.

ΚΕΦΑΛΑΙΟ 4ο ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ / DATA MINING

4.1 ΟΡΙΣΜΟΣ DATA MINING

Ο αριθμός των ανθρώπων που έχουν πρόσβαση στο Διαδίκτυο αυξάνεται σε όλο τον κόσμο τα τελευταία χρόνια. Αυτή η δυνατότητα έχει οδηγήσει σε ταχεία αύξηση στη δημιουργία ιστοσελίδων και εφαρμογών. Στα παγκόσμια δίκτυα τηλεπικοινωνίων μεταδίδονται καθημερινά δεκάδες petabytes από δεδομένα. Οι μηχανικές και οι επιστημονικές πρακτικές που χρησιμοποιούνται συνεχώς παράγουν πολλά δεδομένα. Μερικές από αυτές τις δραστηριότητες περιλαμβάνουν τη τηλεπισκόπησης, τα επιστημονικά πειράματα, τις μετρήσεις διεργασιών, τη μηχανική παρατήρηση και τη περιβαλλοντική παρακολούθηση. Ο τομέας της υγείας και η ιατρική, βασίζονται στις αποθήκες δεδομένων ιατρικών αρχείων και στα δεδομένα των ασθενών. Καθημερινά μέσω μηχανών αναζήτησης γίνεται επεξεργασία δεκάδων δεδομένων petabytes. Κάθε μέρα τα μέσα κοινωνικής δικτύωσης μεταδίδουν πολύ-μεσικό υλικό (ήχος, εικόνα κλπ) και δεδομένα βάση των προτιμήσεων και των αναγκών των χρηστών που τα ακολουθούν. Οι λίστες των πηγών είναι ατελείωτες. Όλα τα παραπάνω μας οδηγούν σε μια ‘περίοδο δεδομένων’, όπου ο αριθμός τους είναι αμέτρητος και αυξάνεται καθημερινά με εκθετικό ρυθμό. Ο όγκος δεδομένων που αποθηκεύονται στις βάσεις δεδομένων και στις data warehouses δεν είναι αξιοποιήσιμος, ως έχει.

Χρειαζόμαστε νέα εργαλεία και τεχνικές για να χρησιμοποιήσουμε αποτελεσματικά τεράστιες ποσότητες δεδομένων. Αυτό θα μας επέτρεπε να τα μετατρέψουμε σε χρήσιμες πληροφορίες και σε γνώσεις. Όμως αρχικά, θα πρέπει να πραγματοποιηθούν ενέργειες για την κατάλληλη δομή των δεδομένων, έτσι ώστε να μπορούμε στην συνέχεια να εξάγουμε αξιοποιήσιμα και χρήσιμα αποτελέσματα.

Η ανάγκη αυτή ανέδειξε την επιθυμία γέννησης της Εξόρυξης Δεδομένων γνωστή και με το όνομα ανακάλυψη γνώσης των δεδομένων –KDD. Η εξόρυξη δεδομένων είναι η εξαγωγή προτύπων τα οποία αντιπροσωπεύουν τις γνώσεις που συλλαμβάνονται ή αποθηκεύονται σιωπηλά σε τεράστιες βάσεις δεδομένων, σε αποθήκες δεδομένων και πληροφοριών, στο διαδίκτυο, και σε ροές δεδομένων. Η αρχική εμφάνιση της εξόρυξης δεδομένων παρατηρήθηκε κατά το τέλος της δεκαετίας του ΄80. Κατά τη δεκαετία του ΄90 αναπτύχθηκε ραγδαία και συνεχίζει μέχρι σήμερα.

4.2 ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ DATA MINING

Το data mining χωρίζεται σε δύο κύριες κατηγορίες, το κατευθυνόμενο data mining (directed data mining) και το μη κατευθυνόμενο data mining (undirected data mining). Το directed data mining επιδιώκει να ταξινομήσει ή να εξηγήσει τα πεδία στόχους. Το undirected data mining επιδιώκει να εμφανίσει τις ομοιότητες που παρουσιάζουν οι ομάδες εγγραφών δίχως να χρησιμοποιήσει καθορισμένο πεδίο στόχου ή χωρίς τη συλλογή καθορισμένων κλάσεων. Μια διαφορετική κατηγοριοποίηση αναφέρεται στις τεχνικές που εφαρμόζει το data mining, και διαχωρίζονται στις μεθόδους «supervised» και «unsupervised».

Ο διαχωρισμός έχει ως εξής:

1. Αλγόριθμοι Εκμάθησης με Επίβλεψη ή supervised learning algorithms. Είναι αυτοί που χρησιμοποιούνται για την πρόβλεψη και τη ταξινόμηση. Στη μέθοδο αυτή τα δεδομένα πρέπει να είναι διαθέσιμα και η τιμή αποτελέσματος πρέπει να είναι γνωστή. Συγκεκριμένα, ο συγκεκριμένος όρος supervised learning δείχνει την τάση του να ανατροφοδοτήσει κάποιο αλγόριθμο με συνεχείς εγγραφές όπου μια αλλαγή στην απόκριση 18 (output variable) η οποία μελετάται, είναι αναγνωρισμένη και ο αλγόριθμος μαθαίνει να προβλέπει τιμή μετά τις νέες έγγραφες, όπου δεν έχουμε γνώση του αποτελέσματος. Δηλαδή, μοντελοποιείται κάποια μεταβλητή απόκρισης με βάση μία ή περισσότερες μεταβλητές επεξήγησης (input variable).
2. Αλγόριθμοι Εκμάθησης χωρίς Επίβλεψη ή unsupervised learning algorithms. Οι αλγόριθμοι αυτοί χρησιμοποιούνται στις περιπτώσεις όπου δεν έχουμε μεταβλητή απόκρισης η οποία πρέπει να ταξινομηθεί ή να προβλεφτεί. Συγκεκριμένα ο όρος unsupervised learning προσδιορίζει την επιχείρηση ανάλυσης κάποιου ώστε να προσδιορίσει κάτι διαφορετικό για τα υπό επεξεργασία δεδομένα, εκτός από την τιμή κάποιας μεταβλητής, π.χ. εάν ανήκει σε cluster. Οι τεχνικές unsupervised εφαρμόζονται στην περίπτωση που δεν έχουμε ένα πεδίο για πρόβλεψη, αλλά οι αλληλεπιδράσεις των δεδομένων ερευνούνται για να αποκαλυφθεί η γενική τους δομή. Αρκετά συχνά οι τεχνικές supervised και unsupervised εφαρμόζονται μαζί. Ο κατάλληλος συνδυασμός των τεχνικών αυτών εξαρτάται από το σκοπό του data mining, από το είδος των διαθέσιμων δεδομένων, όπως επίσης από τις προτιμήσεις και τις ικανότητες του data miner.

4.3 ΤΡΟΠΟΣ ΕΞΟΡΥΞΗΣ

Η μέθοδος Εξόρυξης Γνώσεων αποτελεί αμφίδρομη διαδικασία η οποία ξεκινά από την περισυλλογή των δεδομένων και φτάνει σε ένα πιο πρακτικό επίπεδο δηλαδή στην κατάλληλη αξιοποίηση των συμπερασμάτων διαμέσου επαναλαμβανόμενων σταδίων. Η διαδικασία χρίζει επανάληψης, καθώς αρκετά συχνά οι χρήστες δεν έχουν καθαρή εικόνα της πληροφορίας που τους ενδιαφέρει. Επιπλέον, σε πολλές περιπτώσεις από την εξαγωγή των πρωταρχικών συμπερασμάτων εμφανίζονται νέα ζητήματα και ερωτήσεις. Επίσης, υπάρχει το ενδεχόμενο τα συμπεράσματα της ανάλυσης δεδομένων να μην εξάγουν ασφαλή και αξιόπιστα αποτελέσματα επομένως θα πρέπει να γίνει επανασχεδιασμός της όλης διαδικασίας.

Η μέθοδος αποκάλυψης γνώσεων περιλαμβάνει συγκεκριμένα στάδια που επαναλαμβάνουν τα εξής βήματα [Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)]:

1. Καθαρισμός των δεδομένων με απομάκρυνση του θορύβου και των ασυνεπών δεδομένων.
2. Ενσωμάτωση των δεδομένων. Υπάρχει η δυνατότητα συνδυασμού πολλών πηγών από δεδομένα.
3. Επιλογή των δεδομένων. Τα δεδομένα που έχουν σχέση με την διεργασία ανάλυσης επανακτώνται από βάση δεδομένων.
4. Μετασχηματισμός των δεδομένων. Τα δεδομένα μεταβάλλονται και ενοποιούνται σε κατάλληλες μορφές για την εξόρυξη με πράξεις περιπλοκές και συγκεντρωτικές.
5. Εξόρυξη των δεδομένων. Είναι η βασική διαδικασία που εφαρμόζονται εξυπνότερες μέθοδοι για να εξάγουν πρότυπα δεδομένων.

6. Αξιολόγηση των προτύπων. Η διαδικασία γίνεται για να προσδιοριστούν τα ενδιαφέροντα πρότυπα που είναι αντιπροσωπευτικά της γνώσης με βάση τα δεδομένα ενδιαφέροντος.
7. Παρουσίαση της τελικής γνώσης. Γίνεται με χρήση τεχνικών οπτικοποίησης και αναπαράστασης της γνώσης ώστε να παρουσιαστεί η εξόρυξη.

Στα τέσσερα πρώτα στάδια εφαρμόζονται διαφορετικού είδους προεπεξεργασίες των δεδομένων. Στο τελευταίο στάδιο εφαρμόζονται αλγόριθμοι για να διεξαχθούν τα δεδομένα. Στα βήματα τέσσερα και πέντε τα μοτίβα που εμφανίζουν ενδιαφέρον δίνονται στο χρήστη για αξιολόγηση και κατανόηση, και τελικά είναι ικανά να αποθηκευτούν σαν νέες γνώσεις στην αρχική βάση γνώσεων. Η προηγούμενη εμφάνιση είναι ουσιώδης, διότι αποκαλύπτει τα κρυμμένα μοτίβα προς αξιολόγηση.

Οι τύποι των γνώσεων που αποκαλύπτονται είναι οι ακόλουθοι:

- Κανόνες Συσχέτισης. Χρησιμοποιούν ένα σύνολο αντικειμένων για να περιγράψουν το εύρος τιμών του για ένα άλλο σύνολο μεταβλητών.
- Ιεραρχίες Ταξινόμησης. Ξεκινώντας με ένα σύνολο συναλλαγών ή γεγονότων που έχουν ήδη συμβεί, επιχειρείται να φτιαχτεί ιεραρχία κλάσεων.
- Ακολουθιακά πρότυπα. Γίνεται η αναζήτηση μιας ακολουθίας γεγονότων ή ενεργειών.
- Κατηγοριοποίηση και Διαμερισμός. Ένα αρχικό σύνολο από γεγονότα ή αντικείμενα μπορεί να κατατμιστεί σε σύνολα στοιχείων, παρόμοιας μορφής.
- Πρότυπα σε Χρονοσειρές. Μια ακολουθία δεδομένων συγκρίνεται ως προς τις ομοιότητες για να διαπιστωθεί εάν υπάρχουν μοτίβα στη θέση της.

4.4 ΛΟΓΙΣΜΙΚΑ ΕΞΟΡΥΞΗΣ

RAPIDMINER:

Το RapidMiner αποτελεί εργαλείο για την ανάλυση και εξόρυξη δεδομένων και βρίσκει εφαρμογή στη διαδικασία ανάλυσης δεδομένων κα υποστήριξης διαφόρων τεχνικών σχετικών με την εξόρυξη δεδομένων (Hofmann & Klinkenberg, 2013).

Χρησιμοποιείται από πληθώρα εφαρμογών στην έρευνα, στη βιομηχανία, στην ανάπτυξη εφαρμογών και στην εκπαίδευση. Περιλαμβάνει πληθώρα λογισμικών μάθησης σχετικά με την ομαδοποίηση, την ανάλυση και την ταξινόμηση και μπορεί να παράξει reports και visualizations. Πλεονέκτημα του εργαλείου αυτού είναι το γεγονός ότι δε χρειάζεται η σύνταξη κώδικα. Επίσης υποστηρίζει κάθε τιμή δεδομένων, το οποίο σημαίνει ότι ένας χρήστης μπορεί να εισάγει δεδομένα από διάφορες πηγές για ανάλυση και εξέταση μέσα στην εφαρμογή. Εφαρμόζεται σε μεγάλο πλήθος βιομηχανιών όπως είναι η ενέργεια, οι επικοινωνίες, στους κλάδους της οικονομίας κλπ. Το RapidMiner παρέχει εκπαιδευτική άδεια ενός έτους με δυνατότητα ανανέωσης για καθηγητές και φοιτητές μέσω του προγράμματος εκπαίδευσης RapidMiner Educational License Program.

ORANGE:

Το Orange αποτελεί ένα διαδεδομένο εργαλείο για την εξόρυξη δεδομένων για προγραμματιστές που χρησιμοποιούν την Python. Ανήκει στην κατηγορία εργαλείων ανοιχτού κώδικα και είναι ιδιαίτερα ισχυρό. Η Python αποτελεί μια υψηλού επιπέδου γλώσσα γενικού προγραμματισμού και βρίσκει εφαρμογή σε κάθε είδους επιστημονικό κλάδο όπως ανάλυση δεδομένων, ανάπτυξη λογισμικού, ανάπτυξη

ιστού, μηχανική μάθηση κλπ. Η Python παρέχει μεγάλο όγκο βιβλιοθηκών επέκτασης, πολλές από τις οποίες έχουν να κάνουν με η μηχανική μάθηση. Το Orange υλοποιήθηκε προς το τέλος της δεκαετίας 1990 και αποτελεί ένα από τα παλαιότερα εργαλεία του είδους του. Δίνει έμφαση στη διαδραστικότητα και την απλότητα μέσω scripting και εστιάζει στη σχεδίαση βάσει στοιχείων (components). Προσφέρει πολλές λύσεις αναφορικά με την οπτικοποίηση δεδομένων και παρέχει εργαλεία ονόματι widgets και συνεισφέρουν στην εύρεση κρυφών δεδομένων. Όσοι χρησιμοποιούν το Orange μπορούν να το χρησιμοποιήσουν ως βιβλιοθήκη του Python αναφορικά με το χειρισμό δεδομένων και τη μεταβολή των widget. Το Orange έχει GPL άδεια, διανέμεται δωρεάν και υποστηρίζεται από πληθώρα Online σεμιναρίων. Υπεύθυνο για την ανάπτυξη και η συντήρησή του είναι το εργαστήριο Βϊοπληροφορικής στο τμήμα Η/Υ και πληροφορικής του Πανεπιστημίου στη Λουμπλιάνα της Σλοβενίας.

PHR:

είναι αφενός μια γλώσσα προγραμματισμού και αφετέρου ένα περιβάλλον λογισμικού. Η χρήση της αφορά στατιστικό υπολογισμό, δημιουργία γραφικών παραστάσεων και ανάλυση και επεξεργασία δεδομένων κατά την εξόρυξη δεδομένων. Πέραν της εξόρυξης δεδομένων παρέχονται τεχνικές στατιστικής και γραφικών μοντέλων γραμμικών όπως και μη γραμμικών, κλασσικούς ελέγχους στατιστικών, ταξινόμηση, ομαδοποίηση, ανάλυση χρονοσειρών κ.α. Η υλοποίηση της PHR βασίζεται στη γλώσσα S η οποία γράφτηκε από τον John Chambers, στο Bell Labs. Οι Robert Gentleman και Ross Ihaka του Πανεπιστήμιου του Ώκλαντ της Νέας Ζηλανδίας έγραψαν την R, η δημοφιλία της οποίας τα τελευταία έτη αυξάνεται ολοένα και περισσότερο για πολλούς λόγους, όπως π.χ. ότι είναι εύκολη στην εκμάθηση, είναι

συμβατή με τα πιο δημοφιλή λειτουργικά συστήματα (Windows, Mac Os και Linux), η ύπαρξη πολλών εύχρηστων εγχειριδίων χρήσης, τα πολλά έτοιμα διαθέσιμα πακέτα και τέλος το γεγονός ότι έχει δωρεάν άδεια χρήσης.

WEKA:

Το WEKA (Waikato Environment for Knowledge Analysis) αποτελεί σουίτα λογισμικού μηχανικής μάθησης και εξόρυξης δεδομένων. Αναπτύχθηκε Στο Πανεπιστήμιο Waikato στη Ν. Ζηλανδία. Έχει γραφτεί σε Java και έχει δωρεάν άδεια χρήσης (GNU) που επιτρέπει την ελεύθερη χρήση και τροποποίηση του λογισμικού από τους χρήστες του. Αποτελεί ένα από τα δημοφιλέστερα προγράμματα για εξόρυξη δεδομένων. Βρίσκεται σε πολλές επιστημονικές εργασίες και σε αυτό αναφέρονται πολλά βιβλία που μελετούν την εξόρυξη δεδομένων. Έχει γίνει ιδιαίτερα δημοφιλές εξαιτίας των δυνατοτήτων και των ειδικών χαρακτηριστικών που προσφέρει.

Πιο συγκεκριμένα:

- Εμπεριέχει πολλές μεθόδους για παλινδρόμηση, ταξινόμηση, ανάλυση συστάδων και επίσης πολλούς κανόνες συσχέτισης. Ακόμα, προσφέρει λειτουργία προεπεξεργασίας δεδομένων όπως και εργαλείο οπτικοποίησης.
- Αποτελεί λογισμικό ανοιχτού κώδικα, που σημαίνει ότι ο πηγαίος κώδικας του είναι διαθέσιμος στους χρήστες. Όσοι από αυτούς γνωρίζουν από προγραμματισμό μπορούν να τον τροποποιήσουν και να γράψουν αλγορίθμους.
- Έχει γραφτεί σε Java, γεγονός που του επιτρέπει να χρησιμοποιηθεί σε ποικίλες πλατφόρμες λογισμικού και υλισμικού.
- Έχει περιβάλλον που επιτρέπει ένα απλό χρήστη που δεν έχει γνώσεις προγραμματισμού να το χρησιμοποιεί.

Το WEKA προσφέρεται σε 2 εκδόσεις, μια για απλούς χρήστες και μια για προγραμματιστές.

4.5 ΕΦΑΡΜΟΓΗ

Οικονομία.

Ένας κλάδος εφαρμογής της εξόρυξης δεδομένων είναι αυτός της οικονομίας. Κύριοι αποδέκτες των οικονομικών δεδομένων είναι οι τράπεζες και άλλοι οικονομικοί οργανισμοί. Είναι ολοκληρωμένα, αξιόπιστα, υψηλής ποιότητας και για να αναλυθούν προαπαιτούν συστηματική μέθοδο. Η εξόρυξη δεδομένων προσφέρει στην οικονομία τη συλλογή και την κατανόηση των δεδομένων, στο να βελτιωθούν (data refinement) όπως και στο να δημιουργηθεί, να εκτιμηθεί και να αναπτυχτεί ένα μοντέλο.

Η ορθή ανάλυση οικονομικών δεδομένων συντελεί στη λήψη καίριων αποφάσεων βάσει της ανάλυσης της αγοράς. Βάσει των εργαλείων και των τεχνικών εξόρυξης δεδομένων που χρησιμοποιούνται, η ανάλυση των οικονομικών δεδομένων μπορεί να γίνει με τους εξής τρόπους:

- Όσα δεδομένα συλλέγονται από οικονομικά ινστιτούτα συγκεντρώνονται σε αποθήκες δεδομένων (data warehouses). Μετά εφαρμόζονται στα συλλεγμένα δεδομένα τεχνικές πολυδιάστατης ανάλυσης.
- Μέθοδοι εξόρυξης τύπου επιλογής χαρακτηριστικών (feature selection) συντελούν στην εξακρίβωση ποικίλων χαρακτηριστικών όπως το εισόδημα του πελάτη, το πιστωτικό ιστορικό, η εξόφληση οφειλών ανάλογα με τα έσοδα του κτλ. Επεξεργαζόμενη αυτά τα δεδομένα, η τράπεζα διαμορφώνει πολιτικές δανειοδότησης χαμηλού κινδύνου. Η ταξινόμηση (classification) και η

συσταδιοποίηση (clustering) αποτελούν τεχνικές που βοηθούν οικονομικά ινστιτούτα να ομαδοποιούν πελάτες με κοινά χαρακτηριστικά. Οι τράπεζες ωφελούνται από το αποτελεσματικό φιλτράρισμα και τη συσταδιοποίηση στην ταυτοποίηση μιας ομάδας πελατών, στη συσχέτιση της με ένα νέο πελάτη και την παροχή κοινών οφελών

- Ένα όφελος των εργαλείων εξόρυξης δεδομένων είναι η ανίχνευση εγκλημάτων και ενδεχόμενης απάτης από αλλοιωμένα δεδομένα βάσεων δεδομένων όπως επίσης και από το ιστορικό συναλλαγών των πελατών. Η τεχνική οπτικοποίησης βοηθά στο να παρουσιάζονται δεδομένα με ποικίλες μορφές όπως γραφήματα βασιζόμενα σε συγκεκριμένα χαρακτηριστικά. Η προβολή των δεδομένων υπό διάφορες οπτικές βοηθά την τράπεζα να διακρίνει πελάτες που έχουν παράνομες δραστηριότητες. Μια λεπτομερή έρευνα αυτών των περιπτώσεων μπορεί να συντελέσει στην εξιχνίαση απατών και εγκλημάτων.

Marketing και Πωλήσεις.

Η εξόρυξη δεδομένων δρα καταλυτικά στο marketing και στις πωλήσεις. Οι εταιρίες έχουν στη διάθεση τους τεράστιους όγκους εξαιρετικά πολύτιμων δεδομένων. Το data mining υιοθετήθηκε από νωρίς από τις τράπεζες καθότι συνετέλεσε με επιτυχία στη χρήση του machine learning για πιστοληπτική αξιολόγηση. Μέσα από αυτή τη τεχνολογία μπορεί να μειωθεί η φθορά στους πελάτες με τον εντοπισμό αλλαγών σε τραπεζικά πρότυπα τα οποία μπορεί να δηλώνουν αλλαγή τράπεζας ή αλλαγή στη ζωή του πελάτη π.χ. να μετακομίσει σε άλλη πόλη.

Το Market Basket Analysis (MBA) αποτελεί την τεχνική συσχέτισης για την αναγνώριση ομάδων βάσει στοιχείων με κοινή εμφάνιση σε συναλλαγές. Για

παράδειγμα η αναγνώριση και η ταυτοποίηση πελατών αποτελεί μια ιδιότητα που βοηθά τους εμπόρους λιανικής να καταγράφουν τις εκάστοτε αγορές που πραγματοποιούν κάθε φορά μέσω εκπτωτικών καρτών. Τα δεδομένα προσωπικού χαρακτήρα που συλλέγονται έχουν πολλαπλάσια αξία από αυτή της έκπτωσης επειδή η ταυτοποίηση επιτρέπει και την ανάλυση μοτίβων προερχόμενων από τις αγορές αλλά και τη δημιουργία και αποστολή εξειδικευμένων προσφορών στους υποψήφιους πελάτες.

Τηλεπικοινωνίες.

Ο χώρος των τηλεπικοινωνιών εξελίχθηκε μέσα από την προσφορά υπηρεσιών τηλεφωνίας τόσο σε κοντινές όσο και σε μεγαλύτερων αποστάσεων περιοχές, παρέχοντας ποικίλες υπηρεσίες επικοινωνίας, στις οποίες περιλαμβάνονται το φαξ, το κινητό τηλέφωνο, οι εικόνες, το e-mail και άλλα. Είναι επίσης σε εξέλιξη η ενσωμάτωση των τηλεπικοινωνιών, του Internet και άλλων τρόπων επικοινωνίας. Η εξόρυξη μπορεί να συντελέσει ώστε ο ανταγωνιστικός τηλεπικοινωνιακός κλάδος να ορίσει τηλεπικοινωνιακά πρότυπα, να εντοπίσει δραστηριότητες απάτης, να κάνει καλύτερη χρήση πόρων και να τελειοποιήσει την ποιότητα των παρεχόμενων υπηρεσιών. Οι τρόποι με τους οποίους μπορεί η εξόρυξη δεδομένων να συντελέσει στη βελτίωση της βιομηχανίας των τηλεπικοινωνιών είναι οι εξής:

- Αναλύοντας μοτίβα απάτης και ταυτοποιώντας ασυνήθιστα πρότυπα. Οι απάτες κοστίζουν στη βιομηχανία ετησίως εκατομμύρια δολάρια. Είναι αναγκαία η αναγνώριση πιθανώς δόλιων χρηστών, μέσω άτυπων μοτίβων που χρησιμοποιούν, η ανίχνευση μη εξουσιοδοτημένων εισόδων σε λογαριασμούς η ανακάλυψη ασυνήθιστων μοτίβων που χρίζουν ιδιαίτερης προσοχής όπως για παράδειγμα οι

επαναλαμβανόμενες κλήσεις από συσκευές π.χ. φαξ λόγω εσφαλμένου προγραμματισμού. Μεγάλος αριθμός αυτών των προτύπων μπορεί να ανακαλυφθεί διαμέσου της πολυδιάστατης ανάλυσης και της ανάλυσης συσταδοποίησης.

- Υπηρεσία κινητής τηλεφωνίας: Το διαδίκτυο, οι κινητές τηλεπικοινωνίες και οι υπηρεσίες τηλεπικοινωνιών αυξάνονται συνεχώς και αποτελούν όλο και περισσότερο αναπόσπαστο κομμάτι της δουλειάς και της ζωής του ανθρώπου. Οι χωροχρονικές πληροφορίες αποτελούν σημαντικό χαρακτηριστικό για τα δεδομένα των κινητών τηλεπικοινωνιών. Η εξόρυξη των χρονοχρονικών δεδομένων (spatiotemporal data) είναι απαραίτητη για την εξεύρεση ορισμένων προτύπων. Παραδείγματος χάριν ασυνήθιστη κίνηση σε κινητό τηλέφωνο σε κάποια περιοχή μπορεί να οφείλεται σε κάτι ασυνήθιστο στην περιοχή αυτή. Επίσης η ευκολία χρήσης αποτελεί ουσιαστικό παράγοντα στην έλευση νέων πελατών που θα υιοθετήσουν τις νέες κινητές υπηρεσίες.
- Χρήση εργαλείων για την οπτικοποίηση των τηλεπικοινωνιακών δεδομένων με σκοπό την καλύτερη ανάλυση των. Εργαλεία outlier visualization, OLAP visualization, αποτύπωση σύνδεσης, αποτύπωση συσχέτισης και συσταδοποίηση αποδείχτηκαν πολύ χρήσιμα στην ανάλυση των τηλεπικοινωνιακών δεδομένων.

Εκπαίδευση.

Τα τελευταία έτη έχει παρατηρηθεί αυξανόμενο ενδιαφέρον για η χρήση της μεθόδου εξόρυξης δεδομένων με σκοπό την κατάστρωση επιστημονικών ερωτημάτων εντός της εκπαιδευτικής έρευνας. Το πεδίο αυτό λέγεται EDM (Educational Data Mining) και με

αυτό τον όρο καλείται το κομμάτι της επιστημονικής έρευνας που σχετίζεται με ην ανάπτυξη μεθόδων που θα συντελέσουν ώστε να αντληθούν χρήσιμα συμπεράσματα από τα μοναδικά είδη δεδομένων προερχόμενα από εκπαιδευτικές τοποθεσίες. Με τη χρήση αυτών των μεθόδων κατανοούνται καλύτερα οι μαθητές και οι εγκαταστάσεις στις οποίες μορφώνονται. Παραδείγματος χάριν όσον αφορά την εξόρυξη δεδομένων αναφορικά με τον τρόπο που οι μαθητές χρησιμοποιούν το εκπαιδευτικό λογισμικό, χρίζει περισσότερης εξέτασης δεδομένα που προέρχονται από το επίπεδο της συνεδρίας, της πληκτρολόγησης, της απάντησης, των σπουδαστών, της τάξης και του σχολείου. (Baker).

Η αυξανόμενη χρήση τεχνολογίας στο εκπαιδευτικό σύστημα έχει καταλήξει σε μεγάλο αποθηκευμένο όγκο δεδομένων για τους φοιτητές, κάτι που καθιστά το EDM απαραίτητο εργαλείο για να βελτιωθούν οι διαδικασίες διδασκαλίας και μάθησης. Η χρησιμότητα της φαίνεται σε διάφορους τομείς π.χ. στον εντοπισμό σπουδαστών σε κίνδυνο, την κατηγοριοποίηση αναγκών μάθησης για ποικίλες ομάδες σπουδαστών, στην αύξηση του ποσοστού φοιτητών που αποφοιτούν, την αποτελεσματική αξιολόγηση της απόδοσης των θεσμών, στη μεγιστοποίηση των διαθέσιμων πόρων που έχει η πανεπιστημιούπολη και στη βελτιστοποίηση στην ανανέωση του περιγράμματος σπουδών.

Το EDM αναφέρεται σε έρευνες, εργαλεία και τεχνικές που έχουν σα σκοπό να εξάγεται αυτόματα γνώση από μεγάλες αποθήκες δεδομένων που προέρχονται από ή έχουν να κάνουν με μαθησιακές δραστηριότητες ανθρώπων σε περιβάλλοντα εκπαίδευσης. Αρκετά συχνά τα δεδομένα είναι εκτενή, ακριβή και συγκεκριμένα. Π.χ. συστήματα διαχείρισης μάθησης (Learning Management Systems) διαχειρίζονται πληροφορίες, κατά τη διάρκεια που ένας μαθητής χρησιμοποίησε ένα μαθησιακό αντικείμενο, τον αριθμό των φορών που είχε πρόσβαση σε αυτό και τον χρόνο που το

μαθησιακό αντικείμενο εμφανιζόταν στον υπολογιστή του μαθητή. Υπάρχουν και περιπτώσεις που τα δεδομένα είναι πιο γενικά. Π.χ. ένα πανεπιστημιακό έγγραφο ενός φοιτητή που περιέχει τη λίστα των μαθημάτων που παρακολουθεί, τι βαθμό έχει σε κάθε μάθημα και το αν ο φοιτητής διάλεξε ή άλλαξε τον κύκλο σπουδών.

Το EDM κάνει χρήση και των δυο τύπων δεδομένων ώστε να ανακαλύψει ιδιάζουσας σημασίας πληροφορίες αναφορικά με τους διάφορους τύπους μαθητών και με τον τρόπο που μαθαίνουν, πως δομείται το πεδίο γνώσης και πως επιδρούν οι εκπαιδευτικές στρατηγικές οι οποίες ενσωματώνονται σε ποικίλα περιβάλλοντα εκμάθησης. Τέτοιες αναλύσεις προσφέρουν νέες πληροφορίες που δε θα μπορούσαν να διακριθούν με την εξέταση των ακατέργαστων δεδομένων. Π.χ. αναλύοντας τα δεδομένα από ένα LMS μπορεί να αποκαλυφθεί μια ενδεχόμενη σχέση ανάμεσα στα αντικείμενα μάθησης στα οποία έχει πρόσβαση ένα φοιτητής στο μάθημα και στην τελική βαθμολογία που έχει σε αυτό. Τέτοιες πληροφορίες προσφέρουν γνώσεις για τη σχεδίαση μαθησιακών περιβαλλόντων, που επιτρέπουν σε φοιτητές, εκπαιδευτικούς, διευθυντές σχολείων και όσους διαμορφώνουν την εκπαιδευτική πολιτική να μπορούν να λάβουν τεκμηριωμένες αποφάσεις αναφορικά με το πώς αλληλεπιδρούν, παρέχονται και διαχειρίζονται οι εκπαιδευτικοί πόροι.

Οι Ryan S. Baker και Kalina Yacef όρισαν τους εξής τέσσερις στόχους για την EDM:

1. Πρόβλεψη μαθησιακής συμπεριφοράς μαθητών στο μέλλον: Με τη μοντελοποίηση μαθητών ο στόχος αυτός είναι δυνατό να επιτευχθεί δημιουργώντας μαθησιακά μοντέλα στα οποία ενσωματώνονται χαρακτηριστικά του εκπαιδευόμενου, στα οποία συμπεριλαμβάνονται λεπτομερείς πληροφορίες όπως οι γνώσεις, η συμπεριφορά και το κίνητρο για

μάθηση. Επίσης προσμετράται η εμπειρία του φοιτητή και το πόσο ικανοποιημένος είναι από τη διαδικασία τη μάθησης.

2. Βελτίωση ή ανακάλυψη μοντέλων πεδίου: Μέσα από διάφορες εφαρμογές και μεθόδους του EDM, είναι δυνατό να ανακαλυφθούν νέα μοντέλα και να βελτιωθούν τα ήδη υπάρχοντα. Παράδειγμα αυτού του στόχου είναι η παρουσίαση του εκπαιδευτικού περιεχομένου που αναδεικνύει το πόσο εμπλέκονται οι εκπαιδευόμενοι και να προσδιοριστούν οι βέλτιστες εκπαιδευτικές ακολουθίες που θα υποστηρίξουν το μαθησιακό στυλ του μαθητή.
3. Με τη μελέτη των αποτελεσμάτων από την εκπαιδευτική υποστήριξη η οποία μπορεί να πραγματοποιηθεί μέσα από συστήματα μάθησης.
4. Να προωθηθεί η επιστημονική γνώση αναφορικά με τους μαθητευόμενους και τη μάθηση οικοδομώντας και ενσωματώνοντας μαθησιακά μοντέλα, το πεδίο που ερευνά το EDM όπως επίσης το λογισμικό και την τεχνολογία που χρησιμοποιείται.

Μια λίστα με τις πρωτογενείς εφαρμογές του EDM δίδεται από τους Cristobal Romero και Sebastian Ventura και αποτελείται από τις εξής:

- Απεικόνιση και ανάλυση δεδομένων.
- Υποστήριξη των εκπαιδευομένων με παροχή σχολίων.
- Συστάσεις προς τους φοιτητές.
- Πρόβλεψη απόδοσης των μαθητών.
- Μοντελοποίηση των φοιτητών.
- Ανίχνευση μη επιθυμητών συμπεριφορών μάθησης.
- Ομαδοποίηση των μαθητών.
- Ανάλυση του κοινωνικού δικτύου.

- Προγραμματισμός και σχεδιασμός.
- Υλοποίηση μαθημάτων: Το EDM μπορεί να χρησιμοποιηθεί σε συστήματα που διαχειρίζονται μαθήματα, π.χ. το Moodle που είναι ανοιχτού κώδικα. Το Moodle περιλαμβάνει δεδομένα χρήσης από διάφορες δραστηριότητες χρηστών λόγου χάρη τα αποτελέσματα δοκιμών, τον αριθμό ολοκληρωμένων αναγνώσεων και τη παρουσία σε χώρους συζητήσεων.

Τα εργαλεία εξόρυξης δεδομένων είναι δυνατό να χρησιμοποιηθούν για την προσαρμογή των μαθησιακών δραστηριοτήτων κάθε χρήστη και την προσαρμογή του ρυθμού ολοκλήρωσης του μαθήματος από τον φοιτητή. Ιδιαίτερα ωφελημένα από αυτό είναι τα online μαθήματα όπου τα επίπεδα ικανότητας διαφέρουν. Νέα έρευνα υποδεικνύει τη χρησιμότητα της εξόρυξης δεδομένων σε περιβάλλον κινητής μάθησης. Η εξόρυξη δεδομένων είναι δυνατό να χρησιμοποιηθεί για την προσφορά εξατομικευμένου περιεχομένου σε χρήστες κινητών τηλεφώνων, αν και υπάρχουν διαφορές στο πως διαχειρίζεται το περιεχόμενο ανάμεσα στις κινητές συσκευές, στους τυπικούς υπολογιστές και στα προγράμματα περιήγησης ιστού. Οι καινούργιες εφαρμογές EDM θα έχουν ως στόχο να δίνουν τη δυνατότητα σε απλούς χρήστες να μπορούν να χρησιμοποιούν και να παραβρίσκονται σε δραστηριότητες και εργαλεία εξόρυξης δεδομένων, κάνοντας τη διαδικασία συλλογής και επεξεργασίας δεδομένων, στους χρήστες EDM, πιο προσιτή. Παραδείγματα συμπεριλαμβάνουν εργαλεία οπτικοποίησης και στατιστικής τα οποία αναλύουν την επιρροή των κοινωνικών δικτύων στην παραγωγικότητα.

Υγειονομική περίθαλψη:

Ο τομέας της υγειονομικής περίθαλψης εξάγει τεράστιες δεδομένων, στις οποίες περιλαμβάνεται το ηλεκτρονικό αρχείο υγείας (Electronic Health Record) ή τα ηλεκτρονικά ιατρικά αρχεία (Electronic Medical Records), τα δεδομένα που σχετίζονται με ανάπτυξη φαρμάκων και τα δεδομένα των ασθενών. Η εξόρυξη δεδομένων στον τομέα της υγειονομικής περίθαλψης αποσκοπεί κυρίως στη πρόβλεψη ποικίλων ασθενειών και τη δυνατότητα παροχής συμβουλών προς τους γιατρούς όταν λαμβάνουν αποφάσεις. Με τη χρήση της εξόρυξης ο χώρος της υγείας επωφελείται σε τομείς όπως τα φαρμακευτικά προϊόντα, η ιατρική έρευνα, η διαχείριση των νοσοκομείων, η ασφάλιση υγείας και δυνατότητα ανίχνευσης και πρόληψης ενδεχόμενης απάτης. Υπάρχουν πιέσεις σε αυτό τον τομέα για να μειωθεί το κόστος, ενώ παράλληλα αυξάνεται η ποιότητα των παρεχόμενων υπηρεσιών. Οι μέθοδοι εξόρυξης δεδομένων βρίσκουν εφαρμογή σε τομείς όπως ο διαβήτης, οι καρδιακές παθήσεις και οι καρδιακές προσβολές.

Μερικές εφαρμογές για την εξόρυξη δεδομένων στον τομέα της υγειονομικής περίθαλψης περιλαμβάνονται στα παρακάτω:

- Αποτελεσματικότερη διαχείριση νοσοκομειακών πόρων
- Βελτιωμένη σχέση με τον ασθενή- πελάτη
- Ελαχιστοποίηση της απάτης στον τομέα της ασφάλισης
- Βελτιωμένες τεχνικές θεραπείας.
- Καλύτερη φροντίδα για τους ασθενείς.

Η εξόρυξη δεδομένων ωφελεί τον τομέα της υγειονομικής περίθαλψης με τους παρακάτω τρόπους:

- Η υγειονομική περίθαλψη που παρέχεται στους ασθενείς είναι καλύτερη και πιο προσιτή.

- Η εξόρυξη και η ανάλυση δεδομένων χρησιμοποιείται από τους παρόχους υγειονομικής περίθαλψης για να βρεθούν οι καλύτερες πρακτικές.
- Υπάρχει η δυνατότητα από τους ασφαλιστικούς οργανισμούς να ανιχνεύουν απάτες και καταχρήσεις στην ιατρική περίθαλψη.
- Δυνατότητα λήψης καλύτερων αποφάσεων από τους παρόχους υγειονομικής περίθαλψης για τους ασθενείς.

Η ασφάλεια και η ιδιωτικότητα στα δεδομένα των ασθενών είναι ιδιάζουσας σημασίας λόγω του ευαίσθητου των δεδομένων που αφορούν την υγειονομική περίθαλψη. Με δεδομένο ότι οι πληροφορίες για την υγεία των ασθενών εμπεριέχουν ευαίσθητες, προσωπικές πληροφορίες, ελλοχεύει πάντα ο κίνδυνος παραβίασης των προσωπικών δεδομένων.

ΚΕΦΑΛΑΙΟ 5ο ΔΕΔΟΜΕΝΑ ΑΠΟ ΤΟ TWITTER ΚΑΙ ΣΥΝΑΙΣΘΗΜΑΤΙΚΗ ΑΝΑΛΥΣΗ ΚΕΙΜΕΝΟΥ

5.1 ΛΕΞΙΚΑ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Μια γνωστή μέθοδος για την ανάλυση του συναισθήματος στηρίζεται στα Λεξικά Συναισθημάτων (Lexicon - based Sentiment Analysis). Η σχετικά απλή αυτή μέθοδος ανήκει στις τεχνικές Σημασιολογικού Προσανατολισμού είναι πολύ αποτελεσματική, δεν απαιτεί επιτήρηση ούτε κάποια ενδελεχή προετοιμασία και προεπεξεργασία των στοιχείων. Χρησιμοποιεί ένα Λεξικό, στο οποίο είναι καταχωρημένες λέξεις και

εκφράσεις με συναισθηματική φόρτιση. Η συχνότητα με την οποία εμφανίζονται τέτοιας απόχρωσης λέξεις και εκφράσεις σε ένα κείμενο είναι ο οδηγός για την ανίχνευση του συναισθήματος που εκπέμπει το κείμενο. Οι λέξεις και εκφράσεις – κλειδιά, στις οποίες στηρίζεται η συγκεκριμένη μέθοδος, έχουν ήδη αντιστοιχιστεί με κάποιο, θετικό ή αρνητικό, συναίσθημα από πριν και η αντιστοιχία αυτή έχει ήδη καταγραφεί στο Λεξικό Συναισθημάτων.

Τα επίθετα και επιρρήματα είναι τα πιο ενδιαφέροντα μέρη του λόγου και αυτά που κυρίως καταχωρούνται στα λεξικά, αφού είναι τα μέρη που χρησιμοποιούνται κυρίως για να χαρακτηρίσουν πρόσωπα ή καταστάσεις σε μία πρόταση. Φυσικά, στα λεξικά περιλαμβάνονται και ρήματα, ουσιαστικά και ολόκληρες φράσεις, ενώ σε ορισμένες περιπτώσεις, ανάλογα με το θέμα της ανάλυσης συναισθήματος, επιστρατεύονται εξειδικευμένα λεξικά. Υπάρχουν επίσης και λεξικά στα οποία η αντιστοίχιση δεν έχει τη δυαδική μορφή του θετικού ή αρνητικού, αλλά η συναισθηματική δύναμη κάθε λέξης καθορίζεται από μια βαθμολογία σε κλίμακα συναισθηματικής αξιολόγησης που έχει οριστεί από πριν.

Συμπερασματικά, η διαδικασία της μεθόδου Ανάλυσης Συναισθήματος με τη βοήθεια λεξικού, στην πιο απλή μορφή της, είναι η εξής: Καταρχάς το εξεταζόμενο κείμενο εισάγεται στο μοντέλο και έπειτα χωρίζεται στις προτάσεις που το αποτελούν, οι οποίες με τη σειρά τους αποδομούνται σε λέξεις ή εκφράσεις. Η κάθε μία από αυτές τις λέξεις ή εκφράσεις αντιστοιχίζεται στο Λεξικό Συναισθημάτων με κάποιο συναίσθημα το οποίο έχει αποδοθεί από πριν και καταγράφεται η συχνότητα εμφάνισής του. Με αυτό τον τρόπο, κι αφού ενημερωθούν οι μετρητές του κάθε συναισθήματος, αποτυπώνεται το κυρίαρχο συναίσθημα του κειμένου. Αυτή σε γενικές γραμμές είναι η απλή μορφή προσέγγισης για τον εντοπισμό της συναισθηματικής χροιάς ενός κειμένου με τη χρήση

Λεξικού. Την ίδια κεντρική ιδέα χρησιμοποιούν και μια σειρά άλλων τεχνικών, οι οποίες διαφέρουν κυρίως ως προς τις λεπτομέρειες της υλοποίησής τους.

Τα Λεξικά Συναισθημάτων, κατά τη χρήση τους για την Ανάλυση Συναισθήματος, προσφέρουν τα εξής πλεονεκτήματα: δεν είναι κατάλογοι δημιουργημένοι με αυτόματο τρόπο από αλγορίθμους, αλλά έχουν φτιαχτεί από ανθρώπους, έτσι είναι ακριβέστερη η συναισθηματική αποτύπωση των όρων. Επίσης, τα λεξικά μπορούν να χρησιμοποιηθούν πολλές φορές και ανανεώνονται και εμπλουτίζονται διαρκώς με νέες λέξεις και εκφράσεις, με συνώνυμα, νεολογισμούς ή αντίθετα, γεγονός που δίνει καλύτερα αποτελέσματα στις αναλύσεις, ανεξάρτητα από το εκάστοτε θέμα τους.

Ως αδυναμία της μεθόδου αυτής μπορεί να θεωρηθεί το γεγονός ότι η Ανάλυση Συναισθήματος με τη βοήθεια του λεξικού στηρίζεται στην άποψη ότι η χροιά ενός κειμένου μπορεί να καθοριστεί με απλούς υπολογισμούς της χροιάς των λέξεων που το συνθέτουν. Όμως, πολλές φορές η πολύπλοκη φύση της γλώσσας οδηγεί σε λανθασμένα συμπεράσματα, καθώς πτυχές της, όπως για παράδειγμα η παρουσία της άρνησης, δεν λαμβάνονται υπόψη (Musto, Semeraro, & Polignano, 2014).

Ακόμη, η γλώσσα μπορεί να χρησιμοποιήσει στοιχεία που είναι δύσκολο να εντοπιστούν και να αντιστοιχιστούν, όπως τη μεταφορική ή υποδηλούμενη σημασία λέξεων και εκφράσεων, την ειρωνεία, τον σαρκασμό κ.τ.λ.

Η μέθοδος της Ανάλυσης Συναισθήματος χρησιμοποιεί κάποια γνωστά Λεξικά, όπως τα εξής:

SentiWordNet: Το λεξικό αυτό δημιουργήθηκε και παρουσιάστηκε για πρώτη φορά το 2006. Σκοπός του είναι να αποτυπώσει την πολικότητα των συναισθημάτων και να υποστηρίξει τις εφαρμογές εξόρυξης γνώσης. Οι όροι που χρησιμοποιεί είναι παρμένοι

από το WordNet, που είναι ερμηνευτικό λεξικό της αγγλικής γλώσσας με συμβατή δομή με τις τεχνικές απαιτήσεις της Επεξεργασίας Φυσικής Γλώσσας. Σε αυτό τα ρήματα, τα ουσιαστικά, τα επίθετα και τα επιρρήματα κατατάσσονται σε ομάδες γνωστικών συνωνύμων (synsets), που καθένα από αυτά δηλώνει μια ξεχωριστή έννοια κι όλα αλληλοσυνδέονται με σημασιολογικές – εννοιολογικές και λεκτικές σχέσεις (https://wordnet.princeton.edu/).

Ο συσχετισμός μεταξύ των όρων που προσφέρει το WordNet με τις έννοιές τους, δίνει τη δυνατότητα στο SentiWordNet να αποτυπώνει τη διαφοροποίηση των συναισθημάτων που απορρέουν από το κείμενο σε σχέση με τα συμφραζόμενα.

Το SentiWordNet, έχει δώσει, σε κάθε όρο s του WordNet, μέσα από μια σειρά αυτοματοποιημένων τεχνικών Μηχανικής Μάθησης (weak-supervision, semi-supervised learning step, και random-walk step), ένα διάνυσμα τριών παραμέτρων Pos(s), Neg(s), και Obj(s), οι τιμές των οποίων καταγράφουν την ένταση του θετικού, αρνητικού ή ουδέτερου συναισθήματος που προκύπτει από τον όρο s του λεξικού. Κάθε μία από τις τρεις παραμέτρους, παίρνει τιμή στο διάστημα από 0,0 έως 1,0, με τέτοιο τρόπο, ώστε το άθροισμα και των τριών παραμέτρων να έχει άθροισμα 1 (Baccianella, Esuli, Sebastiani, 2010).

WordNet-Affect: Το Λεξικό αυτό περιλαμβάνει 2.874 σύνολα (synsets) και 4.787 λέξεις. Αναπτύχθηκε χωρίς αυτοματοποιημένες διαδικασίες και στηρίζεται σε ένα υποσύνολο του WordNet Domains, επιλέγοντας όμως από αυτό μόνο όσους όρους κρίθηκαν ως οι κατάλληλοι να αναπαραστήσουν συναισθηματικές έννοιες και έχουν σχέση με λέξεις που αποπνέουν συναίσθημα. Το WordNet Domains που αναφέρθηκε πριν είναι επέκταση του λεξικού WordNet, σε πολλές γλώσσες. Κάθε όρος του λεξικού έχει μια ετικέτα που προσδιορίζει τον τομέα στον οποίο ανήκει ο όρος (ιατρική,

αθλητισμός κ.α.), που έχει επιλεγεί ανάμεσα σε περίπου διακόσιες ετικέτες που έχουν οργανωθεί ιεραρχικά.

Το WordNet-Affect, με τη χρήση της ίδιας μεθόδου, προσθέτει στα χαρακτηριστικά του WordNet Domain μία ή και περισσότερες ετικέτες συναισθήματος (a-labels), που βοηθούν στη συναισθηματική αποτύπωση του όρου με ακριβή τρόπο, ενώ χρησιμοποιούνται και επιπλέον ετικέτες για την απεικόνιση καταστάσεων που μπορούν να προκαλέσουν συναισθήματα ή συναισθηματικές αντιδράσεις. Στο WordNet-Affect περιλαμβάνονται ρήματα, ουσιαστικά, επίθετα και επιρρήματα που έχουν σχέση με τα συναισθήματα και έχουν ταξινομηθεί ως θετικά, αρνητικά, ασαφή ή ουδέτερα, ενώ υπάρχουν και άλλες επιμέρους υποκατηγορίες συναισθημάτων, όπως η έκπληξη, ο φόβος, η χαρά κ.α., που είναι ιεραρχικά τοποθετημένες γύρω από μια δενδρική δομή **1** που οι βασικοί κόμβοι είναι τα συναισθήματα και οι συσχετισμοί γονέων – παιδιών υπονοούν το σημασιολογικό συσχετισμό μεταξύ των όρων (Strapparava, Valitutti, 2004).

Afinn: Πρόκειται για ένα λεξικό που χρησιμοποιείται ευρέως στις διαδικασίες Επεξεργασίας Φυσικής Γλώσσας και κυρίως στην Ανάλυση Συναισθήματος που γίνεται με βάση τα δεδομένα του Twitter. Έχει δημιουργηθεί με αποκλειστικά χειροκίνητο τρόπο και έχει τη μορφή αρχαίου κειμένου, ενώ δομικά αποτελεί μια απλή αντιστοίχιση λέξεων, με μία ακέραια τιμή στο διάστημα (-5,5), που αφορά στη βαθμολογία πολικότητας του συναισθήματος. Χρησιμοποιεί 2.477 αγγλικές λέξεις και κάποιες φράσεις που συναντώνται κυρίως στο microblogging.

SenticNet: Αυτό το λεξικό βασίζεται στην ανάλυση του Sentic Computing και θεωρείται ως εναλλακτική προσέγγιση. Η χρήση του στην Ανάλυση Συναισθήματος είναι κυρίως εννοιολογική και δεν στηρίζεται αποκλειστικά στη συχνότητα εμφάνισης

των λέξεων και εκφράσεων, αλλά προσπαθεί να αξιολογήσει τη διευρυμένη τους έννοια, που περιλαμβάνει τον ορισμό, με κυριολεκτική σημασία, τις μεταφορές και έννοιες που βρίσκονται στο κείμενο ως υπαινιγμοί ή υποδηλωτικά.

Το SenticNet δεν αποτελεί εξέλιξη ή παραλλαγή κάποιου άλλου λεξικού, αλλά φτιάχτηκε με αυτοματοποιημένο τρόπο, με την εφαρμογή τεχνικών Εξόρυξης Γραφημάτων και άλλες πολυδιάστατες τεχνικές κλιμάκωσης σε δεδομένα λογικής, **2** τα οποία συγκεντρώθηκαν από τρία αποθετήρια, το WordNet-Affect, το OMCS (Open Mind Common Sense) και το GECKA (Game Engine for Common sense Knowledge Acquisition). Αυτά παρείχαν ένα σύνολο σημασιών, συναισθημάτων και πολικότητας που μπορούν να συσχετιστούν με 200.000 διαφορετικές έννοιες της φυσικής γλώσσας. Για να προσδιοριστεί η πολικότητα, χρησιμοποιείται μια κλίμακα συναισθήματος στο διάστημα (-1,1).

Textblob: Το Textblob είναι μια πολύ γνωστή βιβλιοθήκη ανοιχτού κώδικα για τις διαδικασίες Επεξεργασίας Φυσικής Γλώσσας και Ανάλυσης Συναισθήματος. Περιλαμβάνει ένα έγκυρο λεξικό συναισθημάτων που έχει εμπλουτιστεί με ετικέτες σημασιών που βοηθούν στις αναλύσεις και χρησιμοποιείται για να αξιοποιήσει τις διεργασίες της και είναι διαθέσιμο και σε XML μορφή αρχείου. Η πολικότητα των συναισθημάτων καταγράφεται με τη χρήση κλίμακας συναισθήματος στο διάστημα (-1,1).

Η εγκυρότητα και αποτελεσματικότητα της Α.Σ. με τη χρήση λεξικού συνδέεται άμεσα με την εγκυρότητα και την ποιότητα του λεξικού που χρησιμοποιεί και το βαθμό συμβατότητας με το κείμενο που αναλύει. Άρα, στη διαδικασία Ανάλυσης Συναισθήματος, είναι πολύ σημαντική παράμετρος η επιλογή του λεξικού βάσει του οποία θα γίνει η ανάλυση. Γι' αυτό το λόγο, η επιλογή πρέπει να γίνεται με γνώμονα

το θέμα του κειμένου και την συντακτική και λεξιλογική πολυπλοκότητα που παρουσιάζει.

5.2 ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Η Μηχανική Μάθηση είναι ένας ταχέως εξελισσόμενος επιστημονικός τομέας με διευρυμένες εφαρμογές που περιλαμβάνουν τεχνικές Εξόρυξης Γνώσης και δρα με βάση την Τεχνητή Νοημοσύνη - Artificial Intelligence. Η Μηχανική Μάθηση (ΜΜ) συμπεριλαμβάνει την έννοια ΄μάθηση΄ και αποτελεί ένα βασικό ζήτημα των ανθρωπιστικών επιστημών (ψυχολογία, παιδαγωγική, κοινωνιολογία). Παράλληλα απασχολεί τους επιστήμονες της βιολογίας και της ιατρικής. Είναι γεγονός ότι παρά τις συνεχής προσπάθειες για την διατύπωση ορισμού από διακεκριμένους επιστήμονες, δεν αναφέρεται ένας ολοκληρωμένος και κοινά αποδεκτός ορισμός.

Από την πορεία εξέλιξης του ανθρώπινου είδους αναγνωρίζουμε ότι η μάθηση αρχικά στηρίχθηκε στην παρατήρηση του φυσικού περιβάλλοντος το οποίο αποτελούσε την πηγή πληροφοριών και έτσι η γνώση αποκτήθηκε σταδιακά με την αφομοίωση των ερεθισμάτων της φύσης, αναπτύσσοντας κινητικές και νοητικές δεξιότητες, και επιλύοντας τα προβλήματα επιβίωση εμπειρικά.

Ως εκ τούτου, η μάθηση προέκυψε μέσω μιας ατελείωτης διαδικασίας ανατροφοδότησης στην οποία οι οποιαδήποτε πληροφορία φιλτράρονταν, αντιλαμβάνονταν και εφαρμόζονταν για την επίλυση κάποιου προβλήματος ή την εκτέλεση καινούργιας εργασίας, αφομοιώνοντας εμπειρικά το σύνολο των γνώσεων που με τη σειρά του διαμόρφωνε πηγές πληροφοριών που θα μπορούσαν να

αναζωογονήσουν νέες διαδικασίες για μάθηση. Με τον ίδιο ακριβώς τρόπο η Πληροφορική, με τις διαδικασίες της ΜΜ, δημιουργεί προϋποθέσεις έτσι ώστε οι ηλεκτρονικοί υπολογιστές να είναι ικανοί να ΄μάθουν΄ δια μέσω αντίστοιχης πορείας όπου τα δεδομένα που έχει καταγράφονται, αντιλαμβάνονται και εφαρμόζονται για να εκτελεστεί μια εργασία, αποκτώντας επίπεδα γνώσεων που θα ανατροφοδοτήσουν τη μάθηση.

Η ΜΜ είναι η μελέτη και δημιουργία αλγορίθμων που χρησιμοποιούν πειραματικά δεδομένα για τη δημιουργία ενός μοντέλου με ικανότητες μάθησης και αναπροσαρμογής, που μπορεί να προβλέπει ή να αποφασίζει δίχως να είναι ρητά προγραμματισμένος για να το κάνει, αλλά με την αυτοματοποιημένη αναγνώριση σημαντικών προτύπων των δεδομένων και με τη σταδιακή χρήση των εμπειριών που αποκτήθηκαν.

Εν ολίγοις, η ΜΜ χρησιμοποιεί την προηγούμενη εμπειρία για να δημιουργήσει μοντέλα που μπορούν να προβλέψουν μελλοντικά αποτελέσματα. Αναζητά κυρίως τη γνώση από την αναγνώριση βασικών συσχετίσεων και κανόνων από ένα σύνολο δεδομένων που ανήκουν στο παρελθόν, ικανά να φτιάξουν ένα μοντέλο Πρόβλεψης - Predictive Model για τις μελλοντικές καταστάσεις. Η τροφοδότηση του μοντέλου πρόβλεψης γίνεται με αναγνωρισμένες τιμής ή παρατηρήσεις του παρελθόντος και φέρει ως αποτέλεσμα συγκεκριμένη εκτίμηση για την πορεία της τιμής της παραμέτρου που ζητείται ή τελικά παίρνει μια απόφαση. Η τελική απόδοση του παραπάνω μοντέλου βασίζεται σε αρκετά μεγάλο ποσοστό από την ποιότητα και τον όγκο των πληροφοριών ελέγχου και εκπαίδευσης τα οποία χρησιμοποιούνται για να αναπτυχθεί το μοντέλο.

Οι αλγόριθμοι της ΜΜ διαχωρίζονται βάση του τρόπου εξέλιξης τους σε 3 κατηγορίες:

1) Μάθηση υπό επιτήρηση ή με επίβλεψη (Supervised Learning). Ο αλγόριθμος απαιτεί να υπάρχουν μεγάλος αριθμός δεδομένων εκπαίδευσης του παρελθόντος όπου όλες οι παράμετροι είναι γνωστές. Επίσης γνωστή θα πρέπει να είναι και η παράμετρος στόχος η οποία είναι ήδη ονοματισμένη (labeled) με το σωστό αποτέλεσμα. Στην περίπτωση αυτή έχουμε αλγορίθμους ανάπτυξης μοντέλου πρόβλεψης από την κατανόηση δεδομένων του παρελθόντος για την αποκάλυψη των σχέσεων παραμέτρων εισόδου στο μοντέλο με τα δεδομένα εξόδου τα οποία είναι γνωστά εξ αρχής. Αυτό που θα προκύψει αξιολογείται και βεβαιώνεται με την εισαγωγή στο μοντέλο ενός ικανοποιητικού ποσού ανάλογων δεδομένων ελέγχου του παρελθόντος καθώς και με τη σύγκριση των εξερχόμενων δεδομένων με τις αληθινές τιμές των αποτελεσμάτων ελέγχου. Το μοντέλο που θα εξαχθεί θα χρησιμοποιηθεί για την πρόβλεψη μελλοντικών καταστάσεων με βάση τα δεδομένα του σήμερα. Κατά την αξιολόγηση ενός ΜΜ μοντέλου, η αποτελεσματικότητά του δίνεται από το κλάσμα με αριθμητή τον αριθμό των ορθών προβλέψεων και παρονομαστή το συνολικό πλήθος των προβλέψεων. Τα δεδομένα που ελέγχονται αποτελούν μέρος των δεδομένων εκπαίδευσης που διατίθενται. Στην περίπτωση που θα υπάρξει χαμηλή αξιοπιστία κατά την αξιολόγηση, επιβάλλεται η αναπροσαρμογή του μοντέλου. Οι αλγόριθμοι της Επιτηρούμενης Μάθησης χωρίζονται με βάση τα προβλήματα που καλούνται να λύσουν σε: α) Αλγόριθμους Ταξινόμησης-(Classification). Χρησιμοποιούνται για να ταξινομήσουν δεδομένα σε καθορισμένες κλάσεις όπως την αναγνώριση του συναισθηματικού προσανατολισμού μιας παραγράφου με τη διάκριση της ΑΣ σε αρνητικό, θετικό ή ουδέτερο. Στη φάση αυτή το μοντέλο είναι εκπαιδευμένο με δεδομένα εμπειρικά τα οποία είναι γνωστά και ονοματισμένα (labeled) για να

αναγνωρίζουν τον συναισθηματικό προσανατολισμό. β) Αλγόριθμους Παλινδρόμησης – Regression. Αναφέρονται σε προβλήματα όπου τα ζητούμενα δεν είναι διακριτά αλλά συνεχή.

2) Αλγόριθμοι μη Επιτηρούμενης Μάθησης - Unsupervised Learning ή Μάθηση χωρίς επίβλεψη. Δεν γίνεται χρήση ονοματισμένων δεδομένων. Πρόκειται για χρήση εκπαιδευτικών δεδομένων στα οποία δεν ξέρουμε τις τιμές στόχου και καλείται το μοντέλο από το μηδέν να ανακαλύψει τις αλληλοσυσχετίσεις των δεδομένων και τους κανόνες που ισχύουν χωρίς να έχει την εμπειρία των δεδομένων εκπαίδευσης του παρελθόντος. Η Μη Επιτηρούμενη Μάθηση βρίσκει χρήση στην επίλυση ζητημάτων Συσταδοποίησης -Clustering, που το μοντέλο χρειάζεται αρχικά να προσδιορίσει τις ομάδες ένταξης των δεδομένων που αναλύονται.
3) Αλγόριθμοι Ενισχυτικής Μάθησης - Reinforcement Learning. Στην περίπτωση αυτή δεν χρησιμοποιούνται labeled δεδομένα. Το μοντέλο αυτό αλληλοεπιδρά με ένα περιβάλλον δυναμικό ώστε να εκπαιδευτεί από την επανάληψη και να αναπτυχθεί ένας μηχανισμός επιβράβευσης. Όταν το μοντέλο δώσει μια αληθή πρόβλεψη δέχεται σκορ επιβράβευσης, αρνητικό εάν η πρόβλεψη είναι λάθος και θετικό όταν είναι σωστή. Η Ενισχυτική Μάθηση κυρίως χρησιμοποιείται στην επιστήμη της ρομποτικής.

Η χρήση τεχνικών ανάλυσης συναισθημάτων με ΜΜ για τον εντοπισμό των τάσεων και τη σφυγμομέτρηση της κοινής γνώμης σε αναρτήσεις κειμένων σε πλατφόρμες κοινωνικής δικτύωσης είναι αρκετά αξιόπιστη και ενδιαφέρουσα. Είναι σημαντικό να υπάρχει ένα σύστημα κωδικοποίησης των κειμένων για την εξαγωγή ασφαλών συμπερασμάτων. Την διαδικασία αυτή την αντιμετωπίζει η ΕΦΓ, καθώς είναι αρκετά

απαιτητική λόγω της μη δομημένης φύσης των εγγράφων που αναρτώνται στα μέσα κοινωνικής δικτύωσης. Τα συγκεκριμένα κείμενα εμπεριέχουν λεξιλόγιο που διαφέρει από την εκπαίδευση και την ιδιοσυγκρασία του κάθε χρήστη, από την ψυχολογική του κατάσταση, τη δυνατότητα του να πληκτρολόγηση μια δεδομένη χρονική στιγμή και από μια σειρά άλλων παραγόντων. Όπως προαναφέρθηκε, τα προβλήματα της ΑΣ αφορούν τα πλαίσια Ταξινόμησης, επομένως επιβάλλεται η χρήση ΜΜ τεχνικών της κατηγορίας Επιτηρούμενης Μάθησης.

Σκοπός της ΜΜ είναι η εκπαίδευση της τεχνικής της ΑΣ. Βασική παράμετρος για να επιτύχει το μοντέλο αποστείλει η σωστή επιλογή του αλγορίθμου εκπαίδευσης. Παράλληλα, δεν πρέπει να υπάρχει αμφισβήτηση της συναισθηματικής κατάταξης του κειμένου που διδάσκει το μοντέλο της ΑΣ και ταυτόχρονα όσο περισσότερος είναι ο αριθμός των δεδομένων που εκπαιδεύονται, τόσο μεγαλύτερη θα είναι και η επιτυχία.

5.3 ΤΑ ΔΕΔΟΜΕΝΑ ΚΑΤΑ ΤΗ ΔΙΑΡΚΕΙΑ ΤΗΣ ΠΑΝΔΗΜΙΑΣ ΚΑΙ Η ΑΝΑΛΥΣΗ ΤΟΥΣ

Η ξαφνική κρίση που ξέσπασε κατά την περίοδο του Covid-19 μας οδήγησε σε μια νέα, άνευ προηγουμένου κατάσταση όπου δοκιμάζονται οι βασικές αξίες, αρχές και συνήθειες της κοινωνίας. Συνεχίζει να επηρεάζει τις ζωές των ανθρώπων σε όλο τον κόσμο. Τα χρόνια της πανδημίας από τον τεράστιο όγκο των αναρτήσεων στο δίκτυο του Twitter, τεχνικές ανάλυσης μπορούν να εξάγουν μεγάλο αριθμό συμπερασμάτων για τους φόβους, τις ανησυχίες, τις ανασφάλειες, τις ουσιαστικές δυσκολίες και γενικά την άποψη της κοινωνίας με βάση την κρίση τόσο σε επίπεδο παγκόσμιο όσο και πανελλαδικό.

Θα είναι επίσης ενδιαφέρον να δούμε ποιος είναι ο ρόλος της κόπωσης και των οικονομικών απωλειών που υπέστησαν μεγάλα τμήματα του πληθυσμού μεταξύ των παρατεταμένων περιορισμών που ξεκίνησαν το Νοέμβριο του 2020 και της σταδιακής μείωσης των μέτρων το Μάρτιο του 2021. Το Twitter αποτελεί μία δημοφιλή πλατφόρμα δικτύωσης, που τείνει να χαρακτηριστεί ώς ένα μέσο κριτικής και έκφρασης της κοινωνίας και παράλληλα χαρακτηρίζεται ως η καταλληλότερη πηγή πληροφοριών για την αξιολόγηση των τάσεων της κοινωνίας κατά τον τρόπο που διαχειρίστηκε η συγκεκριμένη υγειονομική κρίση.

Η καταλληλόλητα του συγκεκριμένου δικτύου έγκειται και στο γεγονός ότι μας δίνει δωρεάν προγραμματιστικές διεπαφές - Application Programming Interfaces ή APIs οι οποίες επιτρέπουν στον ερευνητή την ελεύθερη πρόσβαση σε πληθώρα κειμένων και αναρτήσεων με κριτήρια βασισμένα στην ορολογία των κειμένων ή στη θεματολογία που αναλύουν.

Σε αρχικό στάδιο αναζητήθηκαν πολλές δοκιμαστικές ανακτήσεις ή αλλιώς tweets, με διαφορετικά σετ όρων προς αναζήτηση ή keywords κυρίως στην ελληνική γλώσσα όπως για παράδειγμα: "#πανδημία, "πανδημία", ", "καραντίνα", "#καραντίνα", "εστίαση", "μάσκα", "Τσιόδρας", "#covid19", "#covid-19", "covid-19", "#covidgreece", "κρούσματα", "Χαρδαλιάς", "διασωληνωμένοι", είτε έγιναν ανακτήσεις των tweets στα ελληνικά, είτε γεω-εντοπισμένα στον ελλαδικό χώρο. Δυστυχώς βρέθηκε ότι ένα μεγάλο μέρος από tweets δεν είναι γεωγραφικά εντοπισμένο, επίσης η τοποθεσία στο προφίλ ενός χρήστη δεν δίνεται με ασφάλεια με συνέπεια να μην μπορεί να αξιολογηθεί για το διαχωρισμό αυτό.

Ως εκ τούτου, δεν είναι δυνατό να συγκεντρωθούν αρκετά tweets για να εξειδικεύσουμε την έρευνα στην ελληνική περιφέρεια. Τελικά, έγινε χρήση των παρακάτω όρων ανάκτησης τα οποία εξασφαλίζουν σε αρκετά μεγάλο εύρος τη

συνάφεια της θεματολογίας των tweets κατά την πορεία της πανδημίας καθώς και για τις μεταλλάξεις Epsilon και Delta:

"#deltavariant", "delta variant", "epsilon variant", "#epsilonvariant", "#deltacorona", "#coviddelta", "#covidisnotover","coronavirus", "corona", "#corona", "#coronavirus", "covid19", "covid", "#covid", "#covid-19", "#covid19", "covid-19", "sarscov2", "#sarscov2", "sars cov 2", "sars cov2", "covid_19", "#covid_19","#pandemicisnotover", "pandemic", "#vaccine", "vaccine", "vaccines", "#vaccines", "#coronavaccine", "#covidvaccine", "#coronavaccines", "#covid19vaccine", "#getvaccinated", "#vaccination", "vaccinated".

Από προεπιλογή, τα tweets εξάγονται στη γνωστή μορφή JSON που μοιάζει με λεξικό. Το κύριο στοιχείο ενός αντικειμένου tweet είναι το ίδιο το tweet και το αντικείμενο χρήστη, με ορισμένα tweets να περιέχουν επιπλέον αντικείμενα retweet, προσφοράς ή απάντησης ανάλογα με τον τύπο του tweet.

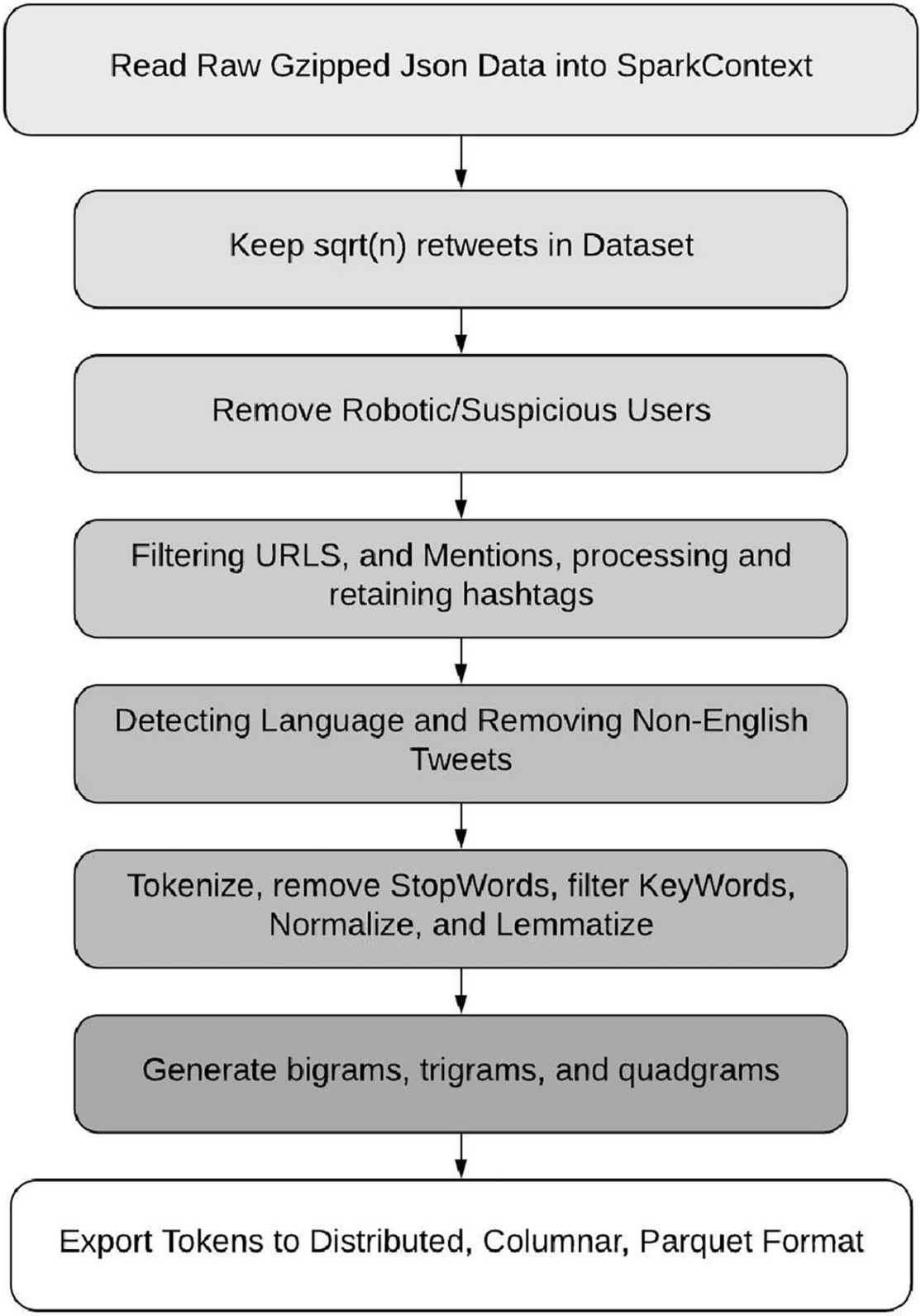
Στο ερευνητικό άρθρο "Dynamic topic modeling of twitter data during the COVID-19 pandemic" των Alexander Bogdanowicz και ChengHe Guan, οι ερευνητές μελέτησαν τα ενω λόγω δεδομένα κατά τη διάρκεια της πανδημίας. Ειδικότερα η έρευνά τους είχε ως εξής:

Το σύνολο δεδομένων εκμαιευτηκε σε μια περίοδο 14 ημερών από την 31η Μαρτίου και αποτελείται από περισσότερα από 46 εκατομμύρια tweets, κατά μέσο όρο πάνω από 3 εκατομμύρια tweets ανά ημέρα, πριν από την προεπεξεργασία.

Αρχικά ξεκίνησαν να προσλαμβάνουν και να προεπεξεργάζονται το σύνολο των δεδομένων, μεταφέροντάς το στη μνήμη, επεξεργάζοντάς το διαδικαστικά και εξάγοντάς το στον χώρο αποθήκευσης δίσκου. Αξιοποίησαν τμήματα της φυσικής

γλώσσας python του Natural Language Toolkit (NLTK) (https://github.com/nltk/nltk) καθώς και αυτών που βρίσκονται στη βιβλιοθήκη του Gensim (https://github.com/RaRe-Technologies/gensim) για την παραγωγή προκαταρκτικών αποτελεσμάτων. Ωστόσο, η χρονική πολυπλοκότητα αυτών των βιβλιοθηκών που βασίζονται σε python ήταν ένα σοβαρό εμπόδιο δεδομένου του μεγέθους του συνόλου δεδομένων, του διερμηνέα και της μνήμης του Python και των περιορισμών. Για το σκοπό αυτό, εκμεταλλεύτηκαν το δημοφιλές open source κατανεμημένο πλαίσιο μεγάλων δεδομένων, Apache Spark (https://github.com/apache/spark), το οποίο παρείχε την απαραίτητη ταχύτητα και κλίμακα και μια ισχυρή βελτιστοποιημένη βιβλιοθήκη επεξεργασίας φυσικής γλώσσας. Αυτό τους επέτρεψε να δημιουργήσουν μια γραμμή δεδομένων.

Στην κάτωθι εικόνα παρουσιάζεται η γραμμή δεδομένων


Read Raw Gzipped Json Data into SparkContext
Keep sqrt(n) retweets in Dataset
Remove Robotic/Suspicious Users
Filtering URLS, and Mentions, processing and retaining hashtags
Detecting Language and Removing Non-English Tweets
Tokenize, remove StopWords, filter KeyWords, Normalize, and Lemmatize
Generate bigrams, trigrams, and quadgrams
Export Tokens to Distributed, Columnar, Parquet Format

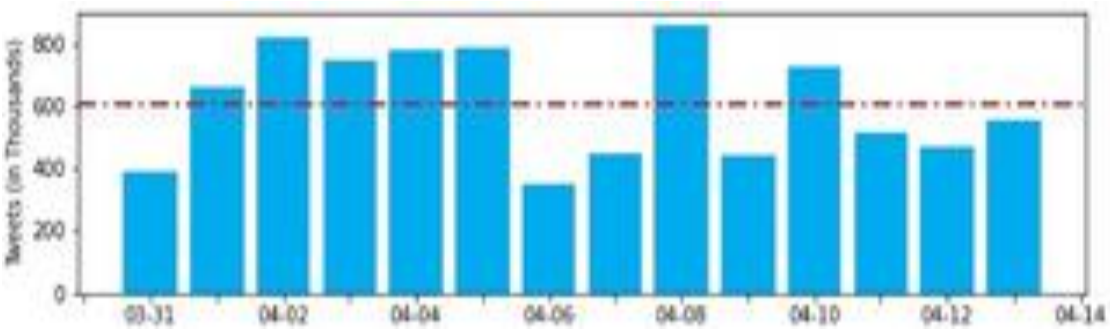
Μετέτρεψαν το ημι-δομημένο σύνολο δεδομένων με τυχαίες λέξεις (δηλαδή Tweets) σε μια αναγνώσιμη μορφή μηχανικής μάθησης. Ξεκίνησαν τη διαδικασία με δεδομένο ότι, ως έχει, πάνω από το 50% του συνόλου των tweet έχει τη μορφή δομημένων retweet.

Στο Twitter, τα retweets αντιπροσωπεύουν μια ευκαιρία για τους χρήστες να μοιράζονται ο ένας τα tweet του άλλου. Ως αποτέλεσμα, απέφυγαν να διαγράψουν τα retweet από το σύνολο δεδομένων και αντ' αυτού εστίασαν στην εκ νέου κλιμάκωση των retweet (το *N* παρουσιάζει το σύνολο δεδομένων).

Επειδή πολλοί λογαριασμοί είναι bots εξετάσαν την πιθανότητα κάποιοι χρήστες να εμφανίζουν χαρακτηριστικά εκτός του πεδίου της κανονικής χρήσης του twitter, όπως αναφέρθηκε. Ως εκ τούτου, για την κατάργηση των χρηστών και των tweets εφάρμοσαν τον κάτωθι τύπο. Αν η βαθμολογία Z είναι εκτός τριών τυπικών αποκλίσεων από τον κανόνα, καταργούνταν:

$$Z_{(\text{tweets,followers,friends})} = \frac{\log 10\text{V}(\text{tweets, followers, friends}) - \mu(\text{tweets, followers, friends})}{\sigma(\text{tweets, followers, friends})}$$

Παρακάτω παρουσιάζεται η διανομή των tweets, κατόπιν κατάργησης retweets, bots



Επίσης, είναι σημαντικό να αφαιρεθούν ασήμαντα και μη σημασιολογικά κείμενα. Καθώς τα hashtags συχνά περιέχουν δημοφιλείς ή δημοφιλείς λέξεις ή φράσεις, χωρίζουμε τα hashtags, έτσι ώστε τα hashtags της μορφής "#StayHomeCovid19" να μετατραπούν σε "Stay Home Covid 19". Η γραμματική δομή των κειμένων διατηρήθηκε για να εγγυηθεί την καλύτερη ανίχνευση γλώσσας στα τελευταία στάδια. Χρησιμοποίησαν τη δημοφιλή βιβλιοθήκη langdetect (https://github.com/Mimino666/langdetect), μια θύρα του αλγόριθμου ανίχνευσης γλώσσας java που αναπτύχθηκε από τον Nakatani Shuyo (2010) που υποστηρίζει γλώσσες, για φιλτράρισμα όλων των μη αγγλικών κειμένων.

Για να ενσωματώσουν το langdetect στη ροή εργασίας μας στο PySpark, εφάρμοσαν συναρτήσεις και πέρασαν τον αλγόριθμο langdetect στο κατανεμημένο σύνολο δεδομένων μας.

Αφού ολοκλήρωσαν όλα τα στάδια της προεπεξεργασίας, εξήγαγαν μια έκδοση με διακριτικό του συνόλου δεδομένων, σε ένα εγγενές σύστημα στηλών βασισμένο σε

σχήμα Apache Spark, γνωστό ως Apache Parquet (https://parquet.apache.org/documentation/latest/), το οποίο προσφέρει σημαντική επεκτασιμότητα, καθώς και αποτελεσματικούς χρόνους ανάγνωσης και εγγραφής στο τελικό σύνολο δεδομένων με διακριτικό.

Το τελικό μοντέλο στο οποίο κατέληξαν μετά τη συλλογή των δεδομένων χρειάστηκε 34 ώρες για να ολοκληρωθεί, περνώντας από το σύνολο δεδομένων 5 φορές κάθε επανάληψη εκπαίδευσης (δηλαδή 5 περάσματα/χρονικό κομμάτι), ενημερώνοντας τις υποθέσεις κάθε 1.000 tweets. Οι χρόνοι εκπαίδευσης των μοντέλων θα μπορούσαν να βελτιωθούν είτε μέσω της μεταγλώττισης σε cpython είτε μέσω της χρήσης ενός Apache Spark μοντέλου LDA.

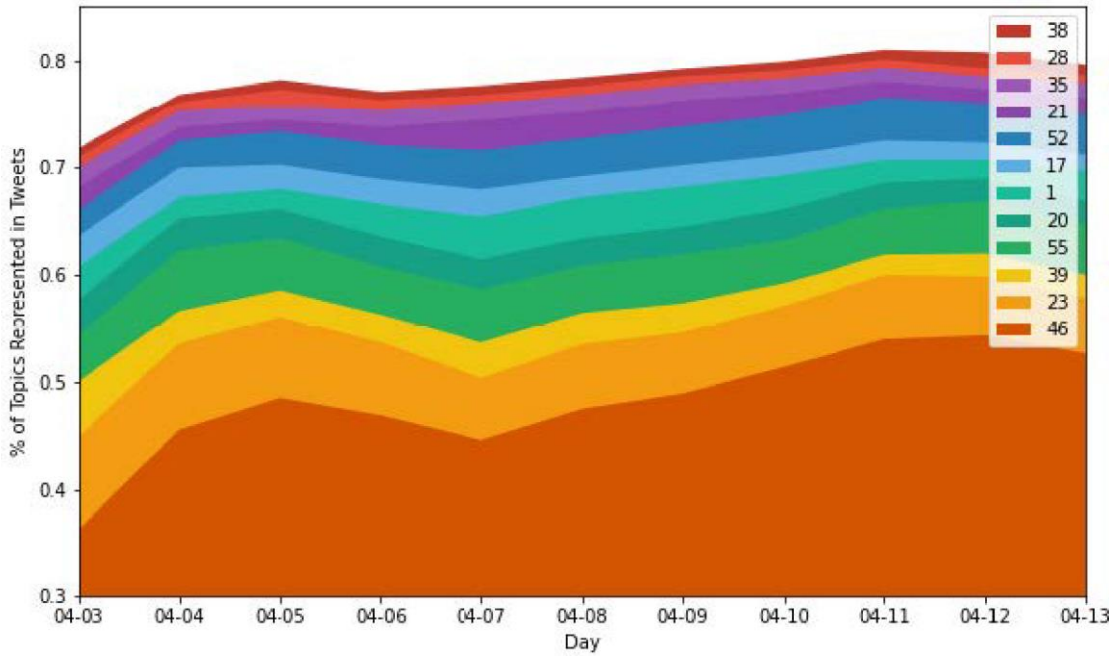
Στην παρακάτω εικόνα παρουσιάζεται η διαμόρφωση του μοντέλου των ερευνητών.

Cluster Configuration	
Nodes	2
Cores/Node	16
Memory/Node	32GB
Partition	Parallel
Model Configuration	
Dataset Passes	5
Update Model	Every 1k tweets
Scoring	Umass
Train Time	34 hours

https://doi.org/10.1371/journal.pone.0268669.t003

Ολοκληρώνοντας τη συλλογή και την καταγραφή των δεδομένων, η έρευνα κατέδειξε ένα αρκετά μεγάλο σύνολο tweets (με μέτρηση στα 100 εκατομμύρια), δημιουργώντας ένα σώμα που θα χρησιμοποιηθεί τόσο σε τρέχουσες όσο και σε μελλοντικές εργασίες. Επίσης, προώθησαν την κατανόηση τόσο των θεμάτων που αφορούν την πανδημία

COVID-19 όσο και της εξέλιξής τους με την πάροδο του χρόνου. Συγκεκριμένα, εντόπισαν 12 από τα πιο δημοφιλή θέματα που υπάρχουν στο σύνολο δεδομένων κατά την περίοδο από τις 3 Απριλίου έως τις 13 Απριλίου 2020 και συζήτησαν την ανάπτυξη και τις αλλαγές τους με την πάροδο του χρόνου, όπως παρουσιάζεται παρακάτω:



Στην εικόνα κάτωθι παρουσιάζονται τα 10 κορυφαία, που μαζί αντιπροσωπεύουν πάνω από το 70% του συνόλου των δεδομένων ανά πάσα στιγμή. Σχεδόν κάθε θέμα στη λίστα των κορυφαίων δέκα είναι μοναδικό και όλα είναι εύκολα ερμηνεύσιμα, ένα ισχυρό σημάδι ενός επιτυχημένου μοντέλου θεμάτων.

Topic Interpretations			
Topic #	Words	Label	Topic Size
46	*time, like, need, know, world, day, life, think, going, good*	Status Updates	52%
55	*trump, president, american, america, democrat, vote, response, china, republican, obama*	US Politics	5%
23	*death, test, number, testing, case, vaccine, infection, rate, data, patient*	Infection & Testing	5%
52	*case, death, new, state, new_york, total, update, county, city, reported*	Reports	4%
1	*online, business, help, student, support, resource, free, pro- gram, new, school*	Personal Finances	2%
20	*stay_home, stay_safe, social_distancing, safe, stay, home, lockdown, save_lives, healthy, easter*	Social Distancing	1.5%
17	*mask, worker, nurse, ppe, hospital, patient, medical, front- line, face_mask, healthcare_worker*	Healthcare	1.5%
39	*trump, state, january, response, election, economic, warned, american, government, warning*	American Response	1.5%
21	*support, community, thank, help, crisis, health, response, team, excellent, time*	Positive Response	1%
35	*supply, staff, ppe, company, equipment, worker, player, medical, employee, testing*	Medical Resources	1%

https://doi.org/10.1371/journal.pone.0268669.t004

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Barodiya, P., Kushwah, S. A., Kaurav, L. S. (2016). A brief study of E-Learning: Special reference in education and corporate sector. International Journal of Advanced Scientific Research, 1(2).
- Blei D., Lafferty J. (2006) Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06.
- Bogdanowicz Alexander, Guan ChengHe, Dynamic topic modeling of twitter data during the COVID-19 pandemic, 2022, https://doi.org/10.1371/journal.pone.0268669.
- Conway M., Hu M., Chapman W. (2019) Recent advances in using natural language processing to address public health research questions using social

media and consumer generated data. *Yearbook of medical informatics*, 28(1):208–217. pmid:31419834.

- Huang B., Yang Y. Mahmood A., Wang H. (2012) Microblog topic detection based on LDA model and single-pass clustering. In Yao et al. (eds) *Rough Sets and Current Trends in Computing*, 7413, 166–171, Springer: Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32115-3_19.
- Jennex, M.E. (2005). Case Studies in Knowledge Management. Idea Group Publishing: Hersley.
- Kostkova P., Szomszor M., Louis C. (2014) #swineflu: The use of Twitter as an early warning and risk communication tool in the 2009 Swine Flu pandemic. *ACM Transactions on Management Information Systems*, 5(2).
- Liu, Y., & Wang, H. (2009). A comparative study on e-learning technologies and products: from East to the West. Systems Research &Behavioral Science, Tao, Y. H., Yeh, C. R., & Sun, S. I. (2006). Improving training needs assessment processes via the Internet: system design and qualitative study. Internet Research, 16(4).
- Liu C., Liu Z., Guan C. (2021) The impacts of the built environment on the incidence rate of COVID-19: A case study of King County, Washington. *Sustainable Cities & Society* 74
- Twitter Corporation. (2021). Q1 2021 Letter to Shareholders. (2–4) https://s22.q4cdn.com/826641620/files/doc_financials/2021/q1/Q1'21-Shareholder-Letter.pdf.
- Yilmaz, Y., & Ülker, D. (2016). Learning Management Systems and Comparison of Open Source Learning Management Systems and Proprietary Learning Management Systems. Journal of Systems Integration, 7(2).

- Zhang C., Sun J. (2012) Large scale microblog mining using distributed mb-lda. In *Proceedings of the 21st International Conference on World Wide Web*, WWW ‘12 Companion, 1035–1042, Association for Computing Machinery.

- https://parquet.apache.org/documentation/latest/.

- https://github.com/Mimino666/langdetect.

- https://github.com/nltk/nltk.

- https://github.com/RaRe-Technologies/gensim

- https://github.com/apache/spark