



ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΕΞΑΓΩΓΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΑΠΟ ΣΩΜΑΤΑ ΚΕΙΜΕΝΩΝ ΤΟΥ TWITTER ΜΕ ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Παπαφιλίππου Γιώργος (Π2015153)

Επιβλέπων καθηγητής: Κερμανίδου Κάτια

ΣΕΠΤΕΜΒΡΙΟΣ 2020

**ΕΞΑΓΩΓΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΑΠΟ ΣΩΜΑΤΑ ΚΕΙΜΕΝΩΝ ΤΟΥ TWITTER ΜΕ
ΧΡΗΣΗ ΤΕΧΝΙΚΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ**

Ονοματεπώνυμο Φοιτητή: Παπαφιλίππου Γιώργος

Αριθμός Μητρώου: Π2015153

Επόπτης της Π.Ε. (επιβλέπουσα καθηγήτρια): Κερμανίδου Κάτια

Μέλος: Μυλωνάς Φοίβος

Μέλος: Έξαρχος Θεμιστοκλής

ΔΗΛΩΝΩ υπεύθυνα ότι εκπονώ πτυχιακή εργασία, αφού μελέτησα τον Κανονισμό Πτυχιακής Εργασίας και με την προϋπόθεση ότι δεν έχει εκπονηθεί ή εκπονείται η ίδια πτυχιακή εργασία από άλλο φοιτητή του Τμήματος. Αποτελεί προϊόν προσωπικής μελέτης και εργασίας και όλες οι πηγές που χρησιμοποιήθηκαν για τη συγγραφή της δηλώνονται σαφώς και στο βιβλιογραφικό κατάλογο.

ΠΕΡΙΛΗΨΗ

Η παρούσα πτυχιακή εργασία παρουσιάζει την διαδικασία εξόρυξης γνώσης από δεδομένα και την εξαγωγή συμπερασμάτων από αυτά μέσω των συναισθημάτων που εκφράζουνε. Αύτη η διαδικασία ονομάζεται sentiment analysis (ανάλυση συναισθημάτων). Μέσω αυτής της διαδικασίας έχουμε την ικανότητα να κατανοούμε τα συναισθήματα μεγάλων πληθυσμών σχετικά με ένα θέμα.

Η διαδικασία του sentiment analysis χρησιμοποιείται για την κατανόηση συναισθημάτων πάνω σε πολλά παγκόσμια ζητήματα. Χρησιμοποιείται για την μελέτη συναισθημάτων σε καταστάσεις όπως φυσικές καταστροφές [5], για την μελέτη εκλογικών αποτελεσμάτων στον χώρο των πολιτικών [8], για την μελέτη ασθενειών στον χώρο της ιατρικής [7], για την μελέτη δυσκολιών κοινωνικοποίησης στον χώρο της ψυχανάλυσης [6], και για την μελέτη απόδοσης ενός προϊόντος μίας επιχείρησης.

Σε αυτή την εργασία θα ασχοληθούμε με το τελευταίο παράδειγμα καθώς παρουσιάζεται όλη η διαδικασία μίας ανάλυσης συναισθημάτων από την αρχή της με την περισυλλογή των δεδομένων, μέχρι την ανάλυση των αποτελεσμάτων και την εξόρυξη συμπερασμάτων από αυτά. Στην έρευνα μας μελετάμε ένα από τα τελευταία οχήματα της εταιρίας Tesla το cybertruck. Το cybertruck αποτελεί ένα πολύ καινοτόμο όχημα με πολλές νέες και διαφορετικές λειτουργίες. Αυτό το χρίζει πολύ διαφορετικό για την αγορά του και αρκετά ασυνήθιστο για τους καταναλωτές. Κάθε καινούργια διαφορά για την αγορά προκαλεί εντελώς διαφορετικές αντιδράσεις στην αγορά. Αυτό λοιπόν αποτελεί τον πυρήνα της εργασίας μας. Με την μελέτη αυτών των αντιδράσεων μπορούμε να οδηγηθούμε σε συμπεράσματα τα οποία θα αποτελέσουν

σημαντικά σημάδια για την κατεύθυνση των πωλήσεων και την επιτυχία του νέου αυτού οχήματος.

Το πρώτο βήμα της παρούσας έρευνας είναι η μελέτη παρόμοιων εργασιών, με σκοπό την παρακολούθηση της διαδικασίας που ακολούθησαν. Κάποιες από τις εν λόγω εργασίες ασχολούνται με διαφορετικό αντικείμενο μελέτης αλλά η διαδικασία παραμένει ίδια στην βάση της. Με την μελέτη τέτοιων εργασιών μπορούμε να παρατηρήσουμε τον τρόπο προσέγγισης στο αντικείμενο καθώς η κάθε έρευνα ακολουθεί διαφορετική προσέγγιση σε όλα τα στάδια μέχρι να φτάσει στην ανάλυση των συναισθημάτων. Μέσω αυτών των ερευνών μελετάμε καλύτερα την διαδικασία και έτσι επιλέγουμε έτσι τον βέλτιστο τρόπο για να πραγματοποιήσουμε την έρευνα.

Στην συνέχεια ακολουθεί η παρουσίαση της ανάλυσης συναισθημάτων για το cybertruck όπου παρουσιάζεται η συλλογή των δεδομένων. Η συλλογή των δεδομένων έγινε μέσω του εργαλείου twitter developer. Τα δεδομένα αποτελούν tweets και με την χρήση κατάλληλων keywords έχουν να κάνουν σχέση με το cybertruck και την παρουσίαση του. Με την χρήση κώδικα απλοποιούμε τα δεδομένα για την ευκολότερη αναγνώριση των συναισθημάτων τους από τον υπολογιστή. Το πρόγραμμα αναγνωρίζει τα μέρη του λόγου και τις λέξεις που έχουν μεγαλύτερη συναισθηματική βαρύτητα και πραγματοποιεί την ανάλυση του συναισθήματος. Ο κώδικας μας παρουσιάζει στο τέλος το sentiment για το κάθε tweet και τον συνολικό μέσο όρο του sentiment των δεδομένων.

Στην συνέχεια προχωράμε στην αξιολόγηση και στην ανάλυση των αποτελεσμάτων, όπου έχουμε την δυνατότητα να εξάγουμε συμπεράσματα σχετικά με το συναίσθημα των δεδομένων

που αναλύσαμε. Γίνεται μελέτη του συνολικού μέσου όρου καθώς αποτελεί την γενική εικόνα για την αντιμετώπιση του οχήματος. Σημαντική επίσης είναι η μελέτη των δεδομένων ατομικά, καθώς μπορεί να υπήρχαν αποκλίσεις μεταξύ της ανάλυσης του συναισθήματος και του πραγματικού συναισθήματος των δεδομένων. Για αυτό παρουσιάζουμε μέσω confusion matrix μεταξύ των προβλεπόμενων και των πραγματικών αποτελεσμάτων της ανάλυσης.

Τέλος οδηγούμαστε στα συμπεράσματα της γενικής έρευνας όπου παρουσιάζεται η κρίση της γενικής εικόνας της εργασίας. Μελετάμε τροποποιήσεις και βελτιώσεις που θα μπορούσαν να προσφέρουν περισσότερα στην έρευνα. Η έρευνα αυτή αποτελεί ένα καλό παράδειγμα μιας βασικής μελέτης ενός προϊόντος με την χρήση sentiment analysis καθώς θα μπορούσε σε δεύτερο χρόνο να γίνει ακόμη καλύτερη.

Στο παράρτημα της διπλωματικής εργασίας παρουσιάζεται ο κώδικας. Έχουν προστεθεί σημειώσεις για την πιο εύκολη κατανόηση του και παρουσιάζεται όλη η διαδικασία του προγράμματος.

Σε πρώτη φάση πραγματοποιείται το κομμάτι της άντλησης των δεδομένων και η διαχείριση τους. Για την άντληση των δεδομένων χρησιμοποιούμε keywords για να καταλήξουμε σε κατάλληλο σετ tweets και στην συνέχεια τα εισάγουμε σε ένα dataframe.

Αργότερα ακολουθεί η διαδικασία του preprocessing όπου τα δεδομένα θα επιδεχθούν αλλαγές για την απλοποίησή τους και την ευκολότερη κατανόηση από τον υπολογιστή. Έτσι γίνεται ευκολότερη η ανάλυση του συναισθήματος από το κάθε tweet ξεχωριστά.

Εν τέλει αφού όλα τα δεδομένα βρίσκονται στην πιο απλή τους μορφή ακολουθεί η διαδικασία της ανάλυσης του συναισθήματος κατά τη διάρκεια της οποίας το πρόγραμμα μελετάει τα δεδομένα και στο τέλος μας εξάγει τα αποτελέσματα τους.

Σκοπός της παρούσας εργασίας είναι η παρουσίαση της διαδικασίας μελέτης δεδομένων με sentiment analysis, η εξακρίβωση της πιστότητας των αποτελεσμάτων και η εύρεση μελλοντικών εξελίξεων στο παρόν πρόγραμμα και στον χώρο της έρευνας.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

ΚΕΦΑΛΑΙΟ 1 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ (SENTIMENT ANALYSIS)

ΚΕΦΑΛΑΙΟ 2 ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΙΣΗ

2.1. Sentiment Analysis of Twitter Data

2.2. Building Large-Scale Twitter-Specific Sentiment Lexicon : A Representation Learning Approach

2.3. Exploring Feature Definition and Selection for Sentiment Classifiers

2.4. Expanding Domain Sentiment Lexicon through Double Propagation

2.5. Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods

2.6. Data-Mining Twitter And The Autism Spectrum Disorder: A Pilot Study

2.7. Mining Of Health And Disease Events On Twitter: Validating Search Protocol Within The Setting Of Indonesia

2.8. A Method For Predicting The Winner Of The USA Presidential Elections Using Data extracted From Twitter

ΚΕΦΑΛΑΙΟ 3 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΓΙΑ ΤΟ CYBERTRUCK ΤΗΣ TESLA

3.1 Συλλογή δεδομένων

3.1.1 Παρουσίαση Twitter Developer Και Kaggle

3.1.2. Import Βασικών Βιβλιοθηκών

3.1.3 Tracking Keywords

3.1.4 Συγκέντρωση Δεδομένων Και Προετοιμασία Για Προεπεξεργασία

3.2 Προεπεξεργασία δεδομένων

3.2.1 Παρουσίαση Προεπεξεργασίας δεδομένων

3.2.2 Διαδικασία προεπεξεργασίας δεδομένων

3.3 Ανάπτυξη εργαλείου ανάλυσης συναισθήματος

ΚΕΦΑΛΑΙΟ 4 ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΚΕΦΑΛΑΙΟ 5 ΣΥΜΠΕΡΑΣΜΑΤΑ

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΠΑΡΑΡΤΗΜΑ

ΕΙΣΑΓΩΓΗ

Στην καθημερινότητα της ανθρωπότητας εμφανίζονται διάφορες νέες ενέργειες και προκλήσεις οι οποίες χρειάζεται να αντιμετωπισθούν. Μπορεί αυτές οι ενέργειες να αφορούν την παραγωγή ενός νέου προϊόντος μιας επιχείρησης ή ακόμη και την απόφαση που πρέπει να πάρει ο πρωθυπουργός μιας χώρας για ένα παγκόσμιο ζήτημα. Τα περασμένα χρόνια παρακολουθήσαμε τις επιπτώσεις και τα αποτελέσματα αυτών των αποφάσεων πολύ “βίαια”. Μπορεί να παρατηρούσαμε μεγάλη πτώση των μετοχών σε μία περίπτωση όπως το πρώτο παράδειγμα ή ακόμη και την παρουσίαση επιθετικών διαμαρτυριών απέναντι σε μια κυβέρνηση όπως το δεύτερο παράδειγμα. Ακόμη επιπλέον θέματα, τις επιπτώσεις των οποίων παρακολουθούμε, ήταν θέματα υγείας (την εξάπλωση ενός ιού), φυσικές καταστροφές, αλλαγές σε νόμους και ψυχικά προβλήματα. Πλέον έχουμε αναπτύξει εργαλεία και μεθόδους τα οποία μας βοηθάνε στην πρόληψη και σωστή αντιμετώπιση τέτοιων ζητημάτων. Μία από αυτές τις μεθόδους είναι η σωστή χρήση της πληροφορίας.

Η πληροφορία βρίσκεται παντού στο διαδίκτυο, ψάχνοντας στα σωστά μέρη μπορούμε να αντλήσουμε πολλές πληροφορίες και σε πιο εξειδικευμένες θεματικές. Ερευνώντας τους σχολιασμούς των καταναλωτών σχετικά με ένα προϊόν μπορούμε να εντοπίσουμε την έκφραση τους και το πώς νιώθουν σχετικά με αυτό το προϊόν. Αντίστοιχα με την έρευνα εκθέσεων και άρθρων διαφορετικών πολιτικά ενεργών πολιτών σχετικά με τις αποφάσεις μιας κυβέρνησης μπορούμε να αναγνωρίσουμε το πώς αισθάνονται για αυτές. Αυτό μπορεί να γίνει και σε άλλους τομείς οι οποίοι αποτελούν πιο στιγμιαίες και προσωρινές καταστάσεις όπως την εμφάνιση μιας νέας ίωσης ή το ξέσπασμα μιας φυσικής καταστροφής. Με την συλλογή δεδομένων από άτομα που βιώνουν αυτή την κατάσταση μπορεί να οργανωθεί καλύτερα το σχέδιο για την δράση και

τα μέτρα που πρέπει να παρθούν ώστε να αντιμετωπιστεί αποτελεσματικά μία τέτοια κατάσταση. Τέτοιου είδους παραδείγματα μας οδηγούν στο συμπέρασμα πως η πληροφορία με την σωστή χρήση της μπορεί να μας βοηθήσει στην πρόβλεψη, την διαχείριση και την αποτελεσματικότερη οργάνωση πολλών προβλημάτων και καταστάσεων.

Μια από τις μεθόδους που χρησιμοποιούμε για την αναγνώριση και μελέτη τέτοιων καταστάσεων είναι το sentiment analysis. Στην εργασία πραγματοποιούμε μια μελέτη περίπτωσης καθώς περνάμε από όλα τα στάδια μιας διαδικασίας sentiment analysis. Από την δημιουργία του κώδικα, στην εκτέλεση του, την ανάλυση του, την παραγωγή αποτελεσμάτων και τέλος και την παρουσίαση των συμπερασμάτων.

Το sentiment analysis αποτελεί μια αρκετά καινούργια μέθοδο στον χώρο της πληροφορικής. Αναπτύσσεται όλο και περισσότερο με τα χρόνια καθώς οι πληροφορίες που έχουμε πλέον στην διάθεσή μας είναι απεριόριστες. Οι τεχνικές που χρησιμοποιούμε στον κώδικα εξελίσσονται καθημερινά . Προσπαθούμε να αποδώσουμε καλύτερα στον υπολογιστή το συναίσθημα που εμπεριέχει ένα κείμενο καθημερινού λόγου.

Ο υπολογιστής με την χρήση λεξικών ακολουθεί μια διαδικασία αναγνώρισης συναισθημάτων και παράγει αποτελέσματα που μας παρουσιάζουν το πόσο θετικό ή αρνητικό συναίσθημα εμφανίζει μια πρόταση. Αυτή η ανάλυση μπορεί να πραγματοποιηθεί σε μικρά κείμενα ή σύντομες προτάσεις πχ tweets ή comments, αλλά μπορούμε να το χρησιμοποιήσουμε και σε μεγαλύτερα κείμενα όπως δοκίμια ή άρθρα σχετικά με κάποιο θέμα. Μέσω αυτού καταλήγουμε να έχουμε μεγαλύτερη κατανόηση για το πως αισθάνεται ο συγγραφέας για κάποιο θέμα και έτσι

οδηγούμαστε πιο εύκολα στην λύση του.

Με την εξέλιξη αυτή λαμβάνουμε καλύτερα και ευστοχότερα αποτελέσματα. Αυτό συνεπάγεται στην αποτελεσματικότερη και πιο αποδοτική λύση προβλημάτων. Επίσης έχουμε την δυνατότητα έρευνας παλαιότερων δεδομένων που θα μπορούσαν να μας προσφέρουν καινούριες πληροφορίες χρησιμοποιώντας νεότερα εργαλεία και λειτουργίες. Ανατρέχοντας σε παλαιότερα σετ δεδομένων μπορούμε να οδηγηθούμε σε ανακαλύψεις οι οποίες μας παρέχουν εφόδια για νέες έρευνες.

Πλέον μέθοδοι όπως το sentiment analysis εφαρμόζονται καθημερινά για την πρόβλεψη και την σωστή αντιμετώπιση βασικών προβλημάτων. Τέτοιου είδους τεχνολογίες έχουν αποδειχθεί χρήσιμες στους περισσότερους τομείς φέρνοντας θετικά αποτελέσματα σε πολλά ζητήματα. Με την συνεχή χρήση τέτοιων τεχνολογιών παρακολουθούμε την καθημερινή εξέλιξή της και την επέκτασή της σε άλλους τομείς, προσφέροντας λύσεις και διευκολύνσεις σε συχνά προβλήματα καθημερινών εργασιών. Ταυτόχρονα καθώς τέτοιου είδους διαδικασίες γίνονται όλο και πιο δημοφιλείς, μπορούμε να δούμε την ραγδαία ανάπτυξη τους και εξέλιξη τους στο πέρασμα του χρόνου. Καθώς όλο και περισσότερος κόσμος μελετά το αντικείμενο αυτό, μέσω της συνεχούς έρευνας βλέπουμε πολλές καινοτόμες αλλαγές στον χώρο.

Εμείς μπορούμε να γίνουμε κομμάτι αυτής της εξέλιξης παρακολουθώντας και λαμβάνοντας συμμετοχή σε μέρη αυτής της τεχνολογίας αποσκοπώντας στην εξέλιξή της και παρατηρώντας κάθε φορά τα αποτελέσματα που φέρει στον πραγματικό κόσμο. Από αυτά τα αποτελέσματα στην συνέχεια μαθαίνουμε, προσπαθούμε να διορθώσουμε πιθανά λάθη και στο τέλος

καταβάλουμε προσπάθεια για την βελτίωση του αντικειμένου στοχεύοντας σε ακόμη πιο βοηθητικά αποτελέσματα.

Στην εργασία αυτή παρουσιάζεται μια παρόμοια μελέτη. Αρχικά πραγματοποιείται η παρακολούθηση παρόμοιων εργασιών για την παρατήρηση τεχνικών που έχουν χρησιμοποιηθεί σε παρόμοιες έρευνες. Παρατηρούμε επίσης τρόπους προσέγγισης και μεθοδολογίες που χρησιμοποιήθηκαν από άλλους ερευνητές για την αποτελεσματικότερη μελέτη του αντικειμένου. Αφού γίνει η μελέτη παρόμοιων εργασιών επιλέγουμε τον τρόπο προσέγγισης που θα ακολουθήσουμε και προσαρμόζουμε αντίστοιχα τις τεχνικές με βάση του σετ δεδομένων που επιλέξαμε.

Το σετ δεδομένων που θα μελετήσουμε αφορά ένα νέο προϊόν της εταιρίας Tesla. Το προϊόν αυτό ονομάζεται cybertruck και έχει κερδίσει τις εντυπώσεις με την καινοτομία του και τις καινούριες λειτουργίες του. Ο λόγος που εστιάζουμε σε ένα τέτοιο προϊόν είναι η ιδιαιτερότητα του. Το γεγονός ότι αυτό το προϊόν είναι αρκετά ιδιαίτερο προκαλεί πολλές αντιδράσεις οι οποίες κρίνουν ενδιαφέρον παράδειγμα προς μελέτη. Με την χρήση sentiment analysis (ανάλυση συναισθημάτων) μπορούμε να οδηγηθούμε σε συμπεράσματα για την γενική εικόνα των συναισθημάτων των καταναλωτών. Έτσι μελετώντας τα αποτελέσματα μπορούμε να γνωρίζουμε άμα μια τέτοια καινοτομία προκαλεί θετικά συναισθήματα ή αρνητικά σε πιθανούς αγοραστές.

Μέσω της παρακολούθησης παρόμοιων εργασιών κατασκευάζουμε έναν κώδικα ο οποίος μπορεί να εξυπηρετήσει τις δικές μας ανάγκες και να παράγει πιο ακριβή αποτελέσματα. Σκοπός της εργασίας είναι η μελέτη μίας ιδιαίτερης περίπτωσης ενός προϊόντος και η συναισθηματική

ανάλυση των δεδομένων που αντλούμε από πιθανούς καταναλωτές. Τέλος αναλύουμε τα αποτελέσματα της έρευνας και μελετάμε τα συμπεράσματα που μπορούν να προκύψουν από αυτά.

ΚΕΦΑΛΑΙΟ 1 – ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ **(SENTIMENT ANALYSIS)**

Το sentiment analysis αποτελεί έναν όρο αρκετά συζητήσιμο στις μέρες μας. Αρκετοί ισχυρίζονται πως ο όρος δεν είναι τίποτα παραπάνω από μια πιο ευπαρουσίαστη μορφή για να εκφράσουμε τα καθήκοντα ταξινόμησης (classification tasks). Για να αποφύγουμε μία τέτοιου είδους διαμάχη μπορούμε να χρησιμοποιήσουμε και τον όρο sentiment classification.

Το sentiment analysis (ανάλυση συναισθημάτων) αφορά την μελέτη του συνόλου των συναισθημάτων που εκφράζει μια ομάδα ανθρώπων σχετικά με κάποιο ζήτημα. Συλλέγοντας δεδομένα «έκφρασης» των ανθρώπων που βιώνουν αυτή την κατάσταση έχουμε την δυνατότητα να εξάγουμε χρήσιμες πληροφορίες οι οποίες θα μας βοηθήσουν στην διαδικασία εύρεσης λύσης για την σωστή οργάνωση και αντιμετώπιση του θέματος.

Πέρα από την αντιμετώπιση προβλημάτων η ανάλυση του συναισθήματος ενός μεγάλου συνόλου μπορεί να μας προϋδεάσει για τα αποτελέσματα μιας κατάστασης. Αυτό μπορεί να οδηγήσει μία ομάδα ανθρώπων στις σωστές κινήσεις για την ανατροπή των αποτελεσμάτων. Για παράδειγμα με την σωστή χρήση των δεδομένων μπορούμε να προβλέψουμε με μεγάλα ποσοστά ακριβείας την εξέλιξη μιας πολιτικής εκλογής και μαθαίνοντας από αυτά μπορούμε να προχωρήσουμε σε κινήσεις για την βελτίωση των αποτελεσμάτων υπέρ μας [8]. Όπως μπορούμε και να μελετήσουμε τις αντιδράσεις πιθανών καταναλωτών σχετικά με την εξέλιξη των πωλήσεων ενός νέου προϊόντος και τις πιθανές αλλαγές ή μετατροπές που μπορούμε να πραγματοποιήσουμε για την αύξηση της πιθανότητας επιτυχίας του προϊόντος αυτού όπως μελετάμε και σε αυτή την εργασία.

Καθώς το sentiment analysis αποτελεί μια αρκετά καινούρια μέθοδο παρατηρούμε και την συνεχή ανάπτυξη της στον χρόνο. Έχοντας ευκολότερη πρόσβαση σε μεγάλα σύνολα πληροφοριών, μας προσφέρεται η δυνατότητα για πολυάριθμες μελέτες σε διαφορετικού χώρους.

Με τέτοιου είδους ευκαιρίες, έχουμε την δυνατότητα εξερεύνησης αποτελεσμάτων που μπορούν να μας προσφέρουν νέα και διαφορετικά δεδομένα.

Λόγο αυτής της καθημερινής χρήσης τέτοιων τεχνικών ανάλυσης και πρόβλεψης παρατηρούμε όλο και μεγαλύτερη εξέλιξη στον χώρο. Βλέπουμε πως με την αύξηση της χρήσης του sentiment analysis είναι πιο προσβάσιμη η μελέτη άλλων εργασιών και η γέννηση νέων τεχνικών ανάλυσης δεδομένων. Κύριος στόχος του sentiment analysis αποτελεί η σωστή απόδοση των δεδομένων στον υπολογιστή για την ευκολότερη και πιο ακριβή κατανόηση του συναισθήματος από τα δεδομένα. Έχοντας ολοκληρώσει μια ανάλυση χρησιμοποιώντας μια σωστή διαδικασία, έχουμε μεγαλύτερα ποσοστά επιτυχίας στα αποτελέσματά μας και καταλήγουμε σε πιο ακριβή έρευνα.

Κάθε sentiment analysis (συναισθηματική ανάλυση) δεν έχει το ίδιο είδος δεδομένων. Για αυτό το λόγο πρέπει να γίνει η σωστή επιλογή για την αποτελεσματικότερη προσέγγιση των δεδομένων ώστε να έχουμε πιο ακριβή αποτελέσματα. Οι δύο βασικές προσεγγίσεις που ακολουθεί το sentiment analysis είναι η προσέγγιση με λεξικό (lexicon based) και η προσέγγιση μέσω machine learning.

Η προσέγγιση με λεξικό (lexicon based) κάνει τη χρήση λεξικών τα οποία εμπεριέχουν λέξεις

που άμα αναγνωριστούν στο κείμενο μπορούν να μας φανερώσουν συναίσθημα. Αναγνωρίζοντας το συναίσθημα των λέξεων μπορούμε να καταλάβουμε το πόσο θετική είναι μια πρόταση αναγνωρίζοντας έτσι τον χαρακτήρα ενός κειμένου. Για παράδειγμα λέξεις όπως το “good” μπορούν να αναγνωριστούν θετικά και λέξεις όπως “not good” αρνητικά. Με την χρήση αυτού του λεξικού μελετάμε τα δεδομένα και μπορούμε να διεξάγουμε αποτελέσματα σχετικά με το συναίσθημα που εκφράζει το κείμενο.

Η προσέγγιση με λεξικό (lexicon based) συνεχώς εξελίσσεται. Συνεχείς ανανεώσεις και διαφορετικά λεξικά εμφανίζονται για την προσθήκη ποιο ουσιαστικών εννοιών για την καλύτερη κατανόηση του συναισθήματος. Πέρα από μεμονωμένες λέξεις των δεδομένων πολύ σημαντικός είναι επίσης και ο χαρακτήρας του συντάκτη την ώρα που πληκτρολογεί την λέξη. Οι λέξεις μπορούν να αποκτήσουν διαφορετικό νόημα ανάλογα με το ύφος του συντάκτη. Λέξεις που εκφράζουν κάτι θετικό, με την προσθήκη ειρωνείας ή υπεροψίας μπορούν να αλλάξουν αμέσως νόημα. Το ίδιο μπορεί να γίνει με άσχημους χαρακτηρισμούς όπως μια βρισιά που μπορεί να δηλώνει ενθουσιασμό. Πολλές φορές λοιπόν σημαντικό είναι να δίνεται μεγαλύτερη βαρύτητα στα μέρη του λόγου. Μέρη του λόγου όπως το ρήμα έχουν μεγαλύτερη βαρύτητα στο νόημα αλλά χρησιμοποιώντας το λάθος υποκείμενο ή αντικείμενο για την αποσαφήνιση του μπορεί να αλλάξει εντελώς το συναίσθημα της πρότασης. Τέτοιου είδους λεπτομέρειες κάνουν την προσέγγιση με λεξικό πολύ χρήσιμη και κρίνουμε πως μπορεί να υπάρξει μεγάλη εξέλιξη στο μέλλον μέσω αυτής.

Αντιθέτως η προσέγγιση μέσω machine learning με την χρήση αλγορίθμων σύγκρισης αναλύει τα δεδομένα και μαθαίνει από αυτά ώστε να παρέχει πιο ακριβή αποτελέσματα. Με την

εκτέλεση συγκρίσεων έχουμε την δυνατότητα να καταλήγουμε σε μεγαλύτερα ποσοστά επιτυχίας, γνωρίζοντας με μεγαλύτερη ακρίβεια τα δεδομένα με πιο έντονα συναισθήματα.

Αλγόριθμοι κατηγοριοποίησης όπως νευρωνικά δίκτυα, δέντρα αποφάσεων και Naive Bayes αναλύουν τα δεδομένα παράγοντας αποτελέσματα με πολύ υψηλά επίπεδα επιτυχίας στην αποσαφήνισή του sentiment των δεδομένων. Καθώς υπάρχει επίσης πολύ συχνή χρήση της προσέγγισης μέσω machine learning συνεχών παρουσιάζονται καινούριοι και πιο εξειδικευμένοι για κάθε κατάσταση αλγόριθμοι οι οποίοι μέσω συγκρίσεων έχουν την δυνατότητα να παράγουν πιο ακριβή αποτελέσματα.

Το sentiment analysis χρησιμοποιείται σήμερα στους περισσότερους τομείς, έχοντας σκοπό την βελτιστοποίηση διαδικασιών, την λύση προβλημάτων και την πρόβλεψη αποτελεσμάτων. Με την χρήση του sentiment analysis μπορούμε να πραγματοποιήσουμε μελέτες στην ιατρική, στην πολιτική, στην παρακολούθηση συναισθημάτων μετά από μία φυσική καταστροφή, σε μια κατάσταση πανδημίας και στην μελέτη απόδοσης πωλήσεων ενός νέου προϊόντος. Είναι μια τεχνική η οποία καθημερινά εξελίσσεται καθώς θεωρείται πλέον απαραίτητη στους περισσότερους τομείς.

Καθώς αυξάνονται συνεχώς οι πηγές δεδομένων και τα δεδομένα είναι άπειρα, οι δυνατότητες για έρευνα και μελέτη είναι απεριόριστες. Αυτό αυτομάτως συνεπάγεται και με την διαρκή ανάπτυξη νέων τεχνικών για την πραγματοποίηση τέτοιων αναλύσεων. Σκοπός είναι η βελτιωμένη απόδοση των δεδομένων στον υπολογιστή για την καλύτερη κατανόηση του συναισθήματος, έχοντας ως αποτέλεσμα την επίλυση περισσότερων προβλημάτων και την

βελτίωση ήδη υπαρχόντων καταστάσεων.

Στην εργασία μελετάμε ένα απλό case study με την χρήση Sentiment analysis για την διεξαγωγή αποτελεσμάτων πάνω σε ένα θέμα χρησιμοποιώντας ένα απλό set δεδομένων. Θα αναλύσουμε μετά τις μεθόδους που χρησιμοποιούμε, γιατί επιλέξαμε αυτές και την εγκυρότητα των αποτελεσμάτων.

Στην εργασία πραγματοποιούμε μια μελέτη περίπτωσης με την χρήση sentiment analysis (ανάλυση συναισθημάτων) όπου μελετάμε τα συναισθήματα σχετικά με το νέο όχημα (cybertruck) της εταιρίας Tesla. Παρουσιάζονται όλα τα στάδια της διαδικασίας.

Αρχικά παρουσιάζεται ο τρόπος συλλογής των δεδομένων, η οργάνωση τους και η προετοιμασία για την ανάλυση τους. Στην συνέχεια παρουσιάζεται ο τρόπος με τον οποίο τα απλοποιούμε για την ευκολότερη κατανόηση τους από τον υπολογιστή για την παραγωγή ευστοχότερων αποτελεσμάτων. Μετά προχωράμε στην ανάλυση του συναισθήματος που πραγματοποιείται με προσέγγιση με λεξικό και παρουσιάζεται η ανάλυση των δεδομένων, η ακρίβεια του προγράμματος και τα συμπεράσματα που προκύπτουν από αυτά τα δεδομένα. Τέλος βλέπουμε τα συμπεράσματα, όπου παρουσιάζονται οι δυνατότητες που υπάρχουν με την χρήση του κώδικα σε άλλες περιπτώσεις μελέτης. Επίσης παρουσιάζονται αλλαγές και τροποποιήσεις που θα μπορούσε να επιδεχθεί ο κώδικας για την βελτίωση του αλλά και μελλοντικούς στόχους για μια αποδοτικότερη και πιο ακριβή μελέτη με την χρήση νεότερων τεχνικών.

ΚΕΦΑΛΑΙΟ 2 – ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

2.1. Sentiment Analysis of Twitter Data [1]

Αντικείμενο

Το αντικείμενο της παραπάνω έρευνας είναι αρκετά σχετικό με το αντικείμενο μελέτης μας. Στόχος της έρευνας είναι η μελέτη ενός σετ δεδομένων. Το σετ εμπεριέχει 11,875 δεδομένα τα οποία έχουν αντληθεί από το twitter.

Στην συγκεκριμένη έρευνα υπάρχουν δυο διαφορετικές προσεγγίσεις για την μελέτη των δεδομένων. Ο διαχωρισμός του sentiment των δεδομένων σε θετικά και αρνητικά και ο διαχωρισμός τους σε θετικά αρνητικά και ουδέτερα για την καλύτερη κατανόηση των αποτελεσμάτων και τον μεθόδων που χρησιμοποιήθηκαν για την μελέτη τους.

Μεθοδολογία

Όπως προαναφέρθηκε χρησιμοποιήθηκαν 11,875 δεδομένα τα οποία αντλήθηκαν από το twitter από μια διαφημιστική πηγή. Τα μοντέλα που χρησιμοποιήθηκαν ήταν το unigram model, το feature based model και το tree kernel based model. Χρησιμοποιείται αρχικά το unigram model γιατί γνωρίζουμε ότι λειτουργεί καλά με δεδομένα του twitter. Από τα πειράματα φαίνεται ότι το unigram model είναι μια πολύ καλή βάση για την σύγκριση των άλλων δυο μοντέλων. Παρατηρείται επίσης πως το feature based model χρησιμοποιεί μόνο 100 χαρακτηριστικά και πετυχαίνει την ίδια ακρίβεια με το unigram model το οποίο χρησιμοποιεί 10,000 χαρακτηριστικά. Ακόμη το tree kernel based model υπερτερεί και των δύο μοντέλων με σημαντική διαφορά.

Στην διαδικασία του preprocessing επίσης εμπεριέχεται ένα λεξικό emoticon και ένα λεξικό για ακρωνύμια. Το λεξικό με τα emoticon εμπεριέχει 170 emoticons και την συναισθηματική τους κατάσταση κατηγοριοποιημένη ως Extremely-positive, Extremely-negative, Positive, Negative, και Neutral.

Το preprocessing ακολουθεί τα εξής βήματα: α) αντικατάσταση όλων των emoticons με την συναισθηματική τους κατάσταση χρησιμοποιώντας το λεξικό των emoticons, β) αντικατάσταση όλων των URLs με την ταμπέλα ||U||, γ) αντικατάσταση των twitter handles με την ταμπέλα ||T||, δ) αντικατάσταση όλων των αρνήσεων με την ταμπέλα “NOT” και ε) γίνεται αντικατάσταση επαναλαμβανόμενων χαρακτήρων με τριπλούς χαρακτήρες (πχ γίνεται η μετατροπή του cooooooooo σε cool). Ο λόγος που δεν γίνεται η μετατροπή σε δύο χαρακτήρες είναι επειδή θέλουμε να δώσουμε έμφαση στην χρήση της λέξης.

Στην συνέχεια ακολουθούμε μια διαδικασία για την μετατροπή του tweet σε μία αναπαράστασή δέντρου. Καθορίζουμε το βασικό δέντρο σε “ROOT” στην συνέχεια το tree kernel δημιουργεί όλα τα πιθανά υπό-δέντρα και τα συγκρίνει μεταξύ τους. Και έτσι καταλήγουμε στο καλύτερο αποτέλεσμα.

Συμπεράσματα

Όπως αναφέρεται στην αρχή, χρησιμοποιούμε το unigram model ως βάση και αναφέρεται ότι παρατηρούμε ένα συνολικό κέρδος άνω του 4% για δύο καθήκοντα ταξινόμησης. Το δυαδικό που μας δίνει ή θετικά ή αρνητικά αποτελέσματα και το τριαδικό που μας προσφέρει θετικά,

αρνητικά και ουδέτερα.

Στην συνέχεια παρατηρούμε ότι το tree kernel ξεπερνά το unigram και το features model για 4.02% και 4.29%. Επίσης επισημαίνεται πως αυτή η διαφορά είναι περισσότερο αισθητή και με την σύγκριση της δυαδικής ταξινόμησης καθηκόντων. Στην έρευνα παρατηρήθηκε ότι και το μοντέλο tree kernel και το feature based model ξεπερνά σε απόδοση το Unigram baseline.

Το τελικό συμπέρασμα της έρευνας είναι ότι το sentiment analysis σε ένα data set με δεδομένα που έχουν αντληθεί από το twitter δεν έχει μεγάλη διαφορά από ένα sentiment analysis σε οποιοδήποτε άλλο είδος.

2.2. Building Large-Scale Twitter-Specific Sentiment Lexicon : A Representation Learning Approach [2]

Αντικείμενο

Στην παραπάνω έρευνα παρουσιάζεται η πρόταση για την κατασκευή ενός μεγάλου σε μέγεθος συναισθηματικού λεξικού με δεδομένα που έχουν αντληθεί από το twitter όπου θα χρησιμοποιείται προσέγγιση εκμάθησης εκπροσώπησης.

Πιο συγκεκριμένα παρουσιάζεται η ανάπτυξη μιας καθαρά αφιερωμένης νευρωτικής αρχιτεκτονικής όπου ενσωματώνει τις συναισθηματικές πληροφορίες του λόγου (πχ. προτάσεις ή tweets) σε μία λειτουργία για τη μάθησης συναισθηματικά ειδικών ενσωματωμένων φράσεων (sentiment - specific phrase embedding SSPE).

Στην συνέχεια το νευρωτικό δίκτυο εκπαιδεύεται από ένα ογκώδες σύνολο tweets τα οποία έχουν συλλεχθεί με θετικά ή αρνητικά emoticons χωρίς κάποιον ήδη υπάρχων σχολιασμό στον χαρακτήρα τους. Ακόμη παρουσιάζεται Urban Dictionary για την διεύρυνση σπόρων συναισθημάτων στον λόγο για την ανάκτηση training data για την δημιουργία sentiment classifiers με βάση τις φράσεις.

Μεθοδολογία

Όπως προαναφέρθηκε μαθαίνουμε συναισθηματικά ειδικές ενσωματωμένες φράσεις (sentiment-specific phrase embedding SSPE) από tweets που εμπεριέχουν θετικά ή αρνητικά emoticons ως training set χωρίς αυτά τα δεδομένα να εμπεριέχουν επιπλέον σημειώσεις. Για την ανάκτηση ακόμη περισσότερων training data για την ανάπτυξη του φραστικού sentiment classifier εκμεταλλευόμαστε παρόμοιες λέξεις από το Urban Dictionary όπου αποτελεί πηγή πόρων που προέρχονται από κάποιο κοινό για την διεύρυνση των συναισθηματικών seeds.

Στην συνέχεια χρησιμοποιούμε τον classifier για να προβλέψουμε το συναισθηματικό score της κάθε φράσης στο λεξιλόγιο του SSPE με συμπέρασμα το συναισθηματικό λεξικό.

Τέλος συγκρίνουμε την απόδοση του λεξικού με τα υπόλοιπα ήδη υπάρχοντα παρατηρώντας τις διαφορές στην απόδοση και στη μεθοδολογία που ακολουθεί το καθένα για την πραγματοποίηση της έρευνας του.

Συμπεράσματα

Μετά την παρουσίαση της δημιουργίας του συναισθηματικού λεξικού με την χρήση του SSPE παρουσιάζονται στο τέλος τα αποτελέσματα σχετικά με την απόδοση του λεξικού και η

σύγκριση του με τα ήδη υπάρχοντα λεξικά.

Η αποτελεσματικότητα του λεξικού έχει επιβεβαιωθεί μέσα από την εφαρμογή του σε εποπτευόμενη μαθησιακή δομή για την δημιουργία συναισθηματική κατηγοριοποίηση για δεδομένα του twitter. Τα αποτελέσματα της έρευνας μα δείχνουν ότι το λεξικό ξεπερνά προηγούμενα λεξικά καθώς θεωρείται ένα περισσότερο βελτιωμένο λεξικό σε σχέση με τα υπόλοιπα.

Τέλος παρουσιάζεται πως μελλοντικός στόχος είναι η εφαρμογή του λεξικού σε μαθησιακές δομές χωρίς επίβλεψη σε δεδομένα του twitter για συναισθηματική κατηγοριοποίηση.

2.3. Exploring Feature Definition and Selection for Sentiment Classifiers [3]

Αντικείμενο

Σε αυτή την έρευνα παρουσιάζεται η εξερεύνηση για τον ορισμό και στρατηγικές επιλογής σχετικά με την πολικότητα συναισθηματικής κατηγοριοποίησης. Η κείμενη πολική κατηγοριοποίηση θεωρείται ένα από τα βασικά καθήκοντα της συναισθηματικής ανάλυσης (Sentiment Analysis). Και όπως έχουμε προαναφέρει συνεχίζει και αναπτύσσεται όλο και περισσότερο.

Υπάρχουν πολλές έρευνες που έχουν αναφερθεί σε διάφορες τεχνικές ορισμού χαρακτηριστικών για την συναισθηματική ανάλυση αλλά δεν είναι ξεκάθαρο ποιο από αυτά τα χαρακτηριστικά λειτουργεί καλύτερα σε μία έρευνα. Σκοπός της έρευνας είναι η μελέτη αυτών των

χαρακτηριστικών και η σύγκριση τους.

Μεθοδολογία

Για την καλύτερη κατανόηση της αξίας των ήδη υπαρχόντων τεχνικών, στην έρευνα μελετάμε τα χαρακτηριστικά για συναισθηματική ανάλυση σε δύο διαστάσεις. Αρχικά μελετάμε τις βασικές μονάδες που έχουν εξαχθεί από το κείμενο όπως λέξεις, n-grams και φράσεις. Στην συνέχεια διερευνούμε την επιλογή χαρακτηριστικών θεωρώντας τις στρατηγικές και με βάση την συχνότητα τους και πιθανολογικά.

Στην συνέχεια γίνεται μελέτη τριών διαφορετικών λεξικών (το εξήχθη από το Affect Control Theoretical sociological studies of emotion (Mejova 2010) και δύο extensions του WordNet: SentiWordNet (Esuli and Sebastiani 2006) και WordNet-Affect (Strapparava and Vlitutti 2004)). Τέλος γίνεται ο έλεγχος αυτών των τεχνικών σε τρία datasets διαφορετικών μεγεθών αποδεικνύοντας έτσι ότι το μέγεθος του κάθε dataset επηρεάζει την απόδοση μερικών τεχνικών.

Συμπεράσματα

Στην έρευνα αυτή χρησιμοποιούνται μερικές από τις πιο σύγχρονες τεχνικές ορισμού και επιλογής χαρακτηριστικών και γίνεται ο έλεγχος πάνω σε τρία διαφορετικά dataset για τον έλεγχο της απόδοσης των τεχνικών σχετικά με την συναισθηματική ανάλυση (Sentiment Analysis).

Μέσω της έρευνας γίνεται να επιβεβαιώσουμε κάποιες υποθέσεις συμπεριλαμβανομένου ότι τα επιρρήματα είναι σημαντικά όσο αναφορά την πολική κατηγοριοποίηση. Επίσης η χρήση

stemming δεν επηρεάζει την απόδοση των τεχνικών. Ακόμη παρουσιάζεται πως η βοήθεια συγκεκριμένων τεχνικών εξαρτάται από την φύση των dataset συμπεριλαμβανομένου του μεγέθους τους και της ισορροπίας των κλάσεων του.

Τέλος παρουσιάζεται το κόστος της ανάλυσης, δηλαδή ο χώρος που χρησιμοποιήθηκε για την αποθήκευση των datasets και ο χρόνος που χρειάστηκε για την επεξεργασία του ξεχωριστά με την κάθε τεχνική.

2.4. Expanding Domain Sentiment Lexicon through Double Propagation [4]

Αντικείμενο

Οι λέξεις που εμπεριέχουν συναίσθημα (sentiment), είναι λέξεις που εκφράζουν μια θετική ή αρνητική πολικότητα. Για αυτό χρειάζεται ένα αναλυτικό συναισθηματικό λεξικό για την καλύτερη ανάλυση του συναισθήματος.

Στις περισσότερες εφαρμογές ανάλυσης συναισθήματος το συναισθηματικό λεξικό έχει βασικό ρόλο. Είναι σχεδόν αδύνατο να συλλέξουμε και να διατηρήσουμε ένα γενικό λεξικό για όλες τις εφαρμογές σε όλες τις βάσεις όπου μπορούν να διαφέρουν λεκτικά.

Σε αυτή την έρευνα παρουσιάζεται μια νέα propagation προσέγγιση όπου εκμεταλλεύεται τις σχέσεις μεταξύ συναισθηματικών λέξεων και θεμάτων όπου οι συναισθηματικές λέξεις τροποποιούνται και συναισθηματικών λέξεων και θεμάτων όπου παράγουν από μόνες τους νέες συναισθηματικές λέξεις.

Μεθοδολογία

Η προτεινόμενη προσέγγιση ακολουθεί αρχικά την εξαγωγή μερικών συναισθηματικών λέξεων και χαρακτηριστικών χρησιμοποιώντας ένα συναισθηματικό λεξικό. Στην συνέχεια χρησιμοποιεί αυτές τις λέξεις και χαρακτηριστικά για να βρει εκ νέου άλλες συναισθηματικές λέξεις και χαρακτηριστικά.

Αυτή η διαδικασία επαναλαμβάνεται μέχρι να μην μπορούν να βρεθούν άλλες λέξεις και χαρακτηριστικά καθώς ταυτόχρονα προβλέπεται η ταυτότητα των νέων συναισθηματικών λέξεων (άμα είναι θετικές ή αρνητικές ως προς το sentiment).

Συμπεράσματα

Με αυτό τον τρόπο προσέγγισης ο αλγόριθμος εξάγει εξαρτώμενες σχέσεις για την μελέτη των σχέσεων μεταξύ χαρακτηριστικών και συναισθηματικών λέξεων. Παρουσιάζεται επίσης πως πολλοί εμπειρικοί κανόνες σχεδιαστήκαν για την εξαγωγή συναισθηματικών λέξεων και χαρακτηριστικών με βάση ήδη υπάρχοντων συναισθηματικών λέξεων.

Τέλος παρουσιάζονται τα αποτελέσματα που αποδεικνύουν ότι αυτή η μέθοδος είναι αποτελεσματική και προτείνονται νέοι μέθοδοι σχετικά με τον τρόπο ανάθεσης συναισθημάτων στις λέξεις που προκύπτουν.

2.5. Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods [5]

Αντικείμενο

Χρησιμοποιώντας tweets που έχουν εξαχθεί από το tweeter κατά την περίοδο των πλημμυρών της Αυστραλίας 2010-2011 χρησιμοποιήθηκαν τεχνικές για την παραγωγή και την ανάλυση που είχαν εμφανιστεί εκείνη την περίοδο.

Στόχος ήταν η ανάπτυξη κατανόησης σχετικά με τις ενεργές κοινότητες των πλημμυρών του Queensland, του New South Wales και τις πλημύρες της Victorian ώστε να αναγνωρισθούν ενεργοί πολίτες και η επιρροή τους στη διάδοση κρίσιμων πληροφοριών. Το μεγαλύτερο ποσοστό των ενεργών πολιτών είχε αποδειχθεί πως ήταν κυρίως τοπικές αρχές.

Επιπλέον σε περιοχές όπως το New South Wales παρατηρήθηκαν σχεδόν μηδενικά επίπεδα ενέργειας μέσα από το twitter. Τα μεγαλύτερα ποσοστά που παρατηρήθηκαν από το New South Wales και τις πλημύρες της Victorian ήταν εθελοντές οι οποίοι ήταν ενεργοί κατά την διάρκεια των πλημμυρών του Queensland.

Μεθοδολογία

Η ανάλυση κοινωνικού δικτύου (Social network analysis (SNA)) αποτελεί μια διαδικασία για την διερεύνηση κοινωνικών δομών σε όρους κόμβων. Η ανάλυση κοινωνικού δικτύου έχει πλέον εμφανιστεί ως μια βασική τεχνική στην μελέτη των κοινωνικών και συμπεριφορικών επιστημών.

Ο κύριος στόχος της ανάλυση κοινωνικού δικτύου είναι οι σχέσεις μεταξύ οντοτήτων (όπως

επικοινωνίες μεταξύ των μελών μιας ομάδας) και κάνει την χρήση μιας ποικιλίας στατιστικών και οπτικών αναλύσεων για την επίτευξη αυτού.

Στα πλαίσια της μελέτης αυτής κάνουμε χρήση της ανάλυσης κοινωνικού δικτύου για να αποκτήσουμε καλύτερη κατανόηση σχετικά με τύπους δικτύων. Αρχικά όταν κάποιος χρήστης το tweeter απαντά σε ένα tweet, ένα δίκτυο με twitterers δημιουργείται από κόμβους που παρουσιάζει τους twitterers και τα link τα οποία παρουσιάζουν τις απαντήσεις σε συγκεκριμένα tweets από συγκεκριμένους twitterers. Από την στιγμή που μία απάντηση μεταφέρεται μεταξύ ενός ανταποκριτή σε κάποιον παραλήπτη, τα links σε αυτό το δίκτυο είναι κατευθυνόμενα.

Ένα δεύτερο είδος δικτύου που μπορεί να κατασκευαστεί από tweets είναι ένα δίκτυο “ενεργών πόρων”, καθώς πολλά tweets εμπεριέχουν links σε ιστοσελίδες καθώς το όριο των 140 λέξεων εμποδίζει μια πιο αναλυτική περιγραφή από το “τι συμβαίνει”. Τέτοιου είδους δίκτυα εμπεριέχουν δύο ειδών κόμβους, κυρίως twitterers και πόρους. Αυτά τα δίκτυα είναι γνωστά και ως bimodal ή bipartite δίκτυα στην βιβλιογραφία της ανάλυσης κοινωνικού δικτύου.

Τα δεδομένα που μελετήθηκαν επιλέχθηκαν με μια σειρά από hashtag για να υπάρχει σχετικότητα με το συμβάν (πχ #qldfloods, #nswfloods, #vicfloods). Τα δεδομένα αυτά κατέβηκαν κατά την περίοδο 3 με 20 Φεβρουαρίου το 2011. Στην συνέχεια αυτά τα δεδομένα αποθηκεύτηκαν σε τρεις ξεχωριστούς πίνακες μιας βάσης δεδομένων. Μία για το Queensland, μια για το New South Wales και μια για τη Victoria.

Συμπεράσματα

Όπως πρότειναν οι ενεργοί πόροι μας κανένα λογικό συμπέρασμα δεν μπορεί να παρθεί. Οι περισσότερες πληροφορίες αφορούσαν αντικείμενα γενικής φύσης παρά κρίσιμες πληροφορίες. Αυτό μπορεί να είναι κατανοητό καθώς τα δεδομένα αντλήθηκαν μετά το στάδιο της σύγκρουσης στις πλημύρες του Queensland και ταυτόχρονα η συμμετοχή στο New South Wales και στις πλημύρες της Victorian ήταν πολύ χαμηλότερες.

Ένας από τους αναγνωρισμένους περιορισμούς της έρευνας είναι ο ανεπαρκής αριθμός των tweets που συλλέχθηκαν για να υπήρχε η ικανότητα να γενικεύσουμε τα αποτελέσματα. Έτσι ένα λογικό πρώτο βήμα σε μελλοντικό χρόνο είναι η βελτίωση της συλλογής των δεδομένων και του αριθμού τους για την απόκτηση μεγαλύτερου όγκου δεδομένων. Αυτό την ίδια στιγμή αποτελεί κάτι πολύ δύσκολο καθώς το twitter εμποδίζει την άντληση μεγάλων αριθμών δεδομένων.

2.6. Data-Mining Twitter And The Autism Spectrum Disorder: A Pilot Study [6]

Αντικείμενο

Οι διαταραχές στο φάσμα του αυτισμού (autism spectrum disorder (ASD)) αναγνωρίζεται όλο και περισσότερο ως ένα μεγάλο δημόσιο θέμα το οποίο επηρεάζει ένα ποσοστό 0.5-0.6% του πληθυσμού.

Στόχος της παρούσας εργασίας είναι η διεξαγωγή μιας έρευνας για το αν το twitter ως μια

δημοφιλής πλατφόρμα για ανταλλαγή πληροφοριών, μπορεί να χρησιμοποιηθεί ως πηγή εξόρυξης δεδομένων γ οποία θα μπορούσε να βοηθήσει στις προαναφερθέντες προκλήσεις. Πιο συγκεκριμένα, με την χρήση μεγάλων set δεδομένων που έχουν συλλεχθεί από tweets παρουσιάζεται μία σειρά πειραμάτων τα οποία πραγματοποιούν μελέτες σχετικά με τις λεκτικές και σημασιολογικές πτυχές των μηνυμάτων που έχουν δημοσιευτεί από άτομα που εμφανίζουν ενδιαφέρον σχετικά με τις διαταραχές στο φάσμα του αυτισμού.

Παρά την πρόκληση της μελέτης δεδομένων ενός στενευμένου πληθυσμού, είναι ξεκάθαρο ότι τα ευρήματα αυτής της έρευνας ανοίγουν πολλούς δρόμους για μελλοντική δουλειά με μεγάλες προοπτικές βελτίωσης της δημόσιας υγείας. Αυτά τα οφέλη εμπεριέχουν τον εμπλουτισμό των εγχειριδίων γνώσεων σχετικά με την πάθηση και την βελτίωση της κατανόησης των πρακτικών και των ανησυχιών ατόμων που επηρεάζονται από διαταραχές στο φάσμα του αυτισμού.

Μεθοδολογία

Η συλλογή των δεδομένων έγινε με tweets τα οποία ήταν γραμμένα στα αγγλικά. Για την πραγματοποίηση μιας συγκριτικής ανάλυσης συλλέχθηκαν δυο set δεδομένων. Το πρώτο set δεδομένων εμπεριέχει tweets τα οποία αφορούν τις διαταραχές στο φάσμα του αυτισμού και παρουσιάζεται ως το ASD subset. Πιο συγκεκριμένα αποτελείται από tweets τα οποία εμπεριέχουν ένα από τα παρακάτω keywords “autism”, “adhd”, “autismawareness”, “asperger”, and “aspie” και το control subset εμπεριείχε όλα τα υπόλοιπα tweets.

Η κατηγοριοποίηση πραγματοποιήθηκε με τον Naïve Bayes και με logistic regression-based

classifiers. Πιο συγκεκριμένα μπορούμε να δούμε ότι το 79% των tweets είχαν κατηγοριοποιηθεί σωστά. Αυτό αποτελεί αξιοσημείωτη απόδοση λαμβάνοντας υπόψη τη συντομία των tweets και το γεγονός ότι μερικές από τις πιο πληροφοριακές λέξεις δεν χρησιμοποιήθηκαν στην κατηγοριοποίηση. Σημαντική επίσης παρατήρηση αποτελεί ότι υπήρξε μεγαλύτερη ακρίβεια στο control subset (84-85%) παρά στο ASD subset (71-74%).

Συμπεράσματα

Σκοπός της έρευνας αυτής ήταν η έρευνα των δυνατοτήτων της εξόρυξης δεδομένων σχετικά με μηνύματα του twitter ώστε να μάθουμε για τις ανησυχίες και τις πρακτικές για τα άτομα που ενδιαφέρονται για τις διαταραχές στο φάσμα του αυτισμού

Η προσέγγιση του θέματος αυτού πραγματοποιήθηκε αρχικά με την άντληση μεγάλου αριθμού δεδομένων από το twitter σχετικά με τις διαταραχές στο φάσμα του αυτισμού. Με αυτά τα δεδομένα πραγματοποιήσαμε μια σειρά πειραμάτων τα οποία ανέλυσαν συχνά και διαφορετικά χαρακτηριστικά των μηνυμάτων. Το σημαντικότερο εύρημα είναι η αξία αυτών των αποτελεσμάτων καθώς παρουσιάζεται ο όγκος των πληροφοριών που μπορούμε να εξάγουμε από τέτοιου είδους δεδομένα και το όφελος που μπορούν να προσφέρουν στην δημόσια υγεία. Μελλοντικοί στόχοι της έρευνας αυτής είναι η αυτοματοποίηση των διαδικασιών για την ευκολότερη αναγνώριση των μηνυμάτων από άτομα που ενδιαφέρονται σχετικά με το ASD προς όφελος των ιατρικών επιστημών.

2.7. Mining Of Health And Disease Events On Twitter: Validating Search Protocol Within The Setting Of Indonesia [7]

Αντικείμενο

Σε αυτή την έρευνα παρουσιάζεται η προσπάθεια κατανόησης της τορινής κατάστασης υγείας χρησιμοποιώντας δεδομένα μέσα από το twitter με την χρήση όρων υγείας και πληροφορίες που μπορούμε να κατανοήσουμε μέσα από αυτά. Γίνεται η απορρόφηση δεδομένων από συζητήσεις με θέματα υγείας και σχετικά post με παρουσίαση ασθενειών. Αυτή η διαδικασία γίνεται μέσω προ ορισμένων λέξεων, την συχνότητα και την χρονικότητα τέτοιου είδους περιεχομένου και συζητήσεων ώστε να ορισθεί πόσο ικανό ήταν το πρωτόκολλο έρευνας για να παρουσιάσει ικανοποιητικά δεδομένα.

Για την εκπαίδευση και για το δοκιμαστικό πρωτόκολλο έρευνας χρησιμοποιήθηκε ο αλγόριθμος Classification and Regression Trees (CART). Η ακρίβεια των προβλέψεων φάνηκε να είναι καλή και η έρευνα μας παρουσιάζει ότι η παρακολούθηση του δημοσίου συναισθήματος μπορεί να χρησιμοποιηθεί για πραγματικά συμβάντα υγείας.

Μεθοδολογία

Αρχικά έγινε η ανάπτυξη ομάδων λέξεων και φράσεων σχετικά με τα συμπτώματα και τα αποτελέσματα της υγείας για να πραγματοποιηθεί σωστή απορρόφηση υλικού από το twitter search. Από την στιγμή που αυτή η έρευνα ήθελε να φανερώσει τις σχέσεις της υγείας εντός της Ινδονησίας όλα τα keywords είναι γραμμένα στα Bahasa:

"rumah OR sakit OR rawat OR inap OR demam OR panas -cuaca OR berdarah OR pendarahan OR tombosit OR badan OR muntah OR badan OR tua"

Έγινε η συλλογή 390 tweets γραμμένων στα Bahasa. Στην συνέχεια πραγματοποιήθηκε η απλοποίηση των δεδομένων αφαιρώντας ειδικούς χαρακτήρες, κεφαλαία, αριθμοί και τα stop-words της Bahasa.

Στην συνέχεια έγινε η χρήση του αλγορίθμου CART. Οι προβλέψεις που συλλέχθηκαν ήταν 22 από πολλές λέξεις που είχαν μεγάλη συχνότητα καθώς έγινε η καταμέτρηση των συχνών λέξεων κάθε tweet. Τέλος τα δεδομένα χωρίστηκαν σε δυο κατηγορίες. Τα 273 tweets (70%) χωρίστηκαν για training και στο τέλος εφαρμόστηκε το μοντέλο στα 117 tweet (30%).

Συμπεράσματα

Παρουσιάστηκε πως λέξεις όπως hati (καρδιά), rasa (αφή) και perut (στομάχι) ήταν οι πιο συχνά συνδεδεμένες λέξεις με την λέξη sakit (ναυτία, αρρώστια, πόνος). Άλλες λέξεις εμφάνισαν σύνδεση κατώτερη του 0.1. Ταυτόχρονα οι προβλέψεις που βασίστηκαν σε tweets που περιείχαν πληροφορία σχετικά με αρρώστια παρουσιάστηκαν 6,109 True αποτελέσματα από 1.145.649 tweets που έγιναν real time από το twitter. Αυτό σημαίνει πως τα αποτελέσματα σχετικά με θέματα υγείας ήταν πολύ μικρά (0.5%).

Ως συμπέρασμα παρουσιάζεται πως ογκώδη και ανισόρροπα set δεδομένων μειώνουν απότομα την ικανότητα του μοντέλου να πετύχει. Συνεπώς σε επόμενη φάση της έρευνας είναι απαραίτητη η σύγκριση τωρινών αποτελεσμάτων με κάποιον πιο προχωρημένο αλγόριθμο που

μπορεί να αντιμετωπίσει τέτοιου είδους set δεδομένων.

2.8. A Method For Predicting The Winner Of The USA Presidential Elections Using Data extracted From Twitter [8]

Αντικείμενο

Αυτή η εργασία παρουσιάζει δεδομένα που έχουν απορροφηθεί από το twitter με σκοπό την πρόβλεψη των αποτελεσμάτων των τελευταίων εκλογών των ΗΠΑ. Η μελέτη αφορά τρεις βασικές πολιτείες την Florida, το Ohio και την N. Carolina και επικεντρώνεται στους δυο βασικούς υποψήφιους Donald J. Trump και Hilary Clinton.

Η εργασία ακολουθεί τα εξής βήματα. Τη διαδικασία του preprocessing και την ανάλυση θετικού ή αρνητικού συναισθήματος προς τους δύο υποψήφιους και την πρόβλεψή τους για τον νικητή των εκλογών.

Μεθοδολογία

Η κατηγοριοποίηση των δεδομένων θα καταταχθεί σε θετικό, ουδέτερο ή αρνητικό. Ο τρόπος που παρουσιάζεται αυτή η κατηγοριοποίηση είναι με την χρήση του μοντέλου Naïve Bayes καθώς αποτελεί ένα απλό αλλά αποτελεσματικό εργαλείο. Το μοντέλο χρειάζεται να εκπαιδευθεί για τα γίνει η χρήση του. Το training dataset αποτελούνταν από 1.578.627 tweets τα οποία είχαν ήδη κατηγοριοποιηθεί ως θετικά, ουδέτερα ή αρνητικά και η ακρίβεια του μοντέλου ήταν μεγαλύτερη από το 75%.

Συμπεράσματα

Αφού παρατηρείται πως τα αποτελέσματα τη έρευνας ήταν κοινά με τα αποτελέσματα των εκλογών τον Νοέμβριο του 2016, μπορούμε να φτάσουμε στο συμπέρασμα πως το πείραμα κρίνεται επιτυχές. Σημειώνεται πως η ανάλυση των δεδομένων δεν αποτελεί ιδανικό τρόπο για την πρόβλεψη των δεδομένων καθώς μεγάλο ποσοστό των ψηφοφόρων δεν έχει λογαριασμό στο twitter, παρόλα αυτά μπορούμε να μελετήσουμε την άποψη ενός μεγάλου μέρους των πολιτών και να παρατηρήσουμε την άποψη τους σχετικά με άλλα θέματα.

Σε αυτή την δουλειά παρουσιάζεται το πώς οι ψηφοφόροι αισθάνονται σχετικά με τους υποψήφιους και δεν εστιάζεται η έρευνα στον αριθμό των ψήφων. Αυτό αυτομάτως σημαίνει ότι δεν μπορεί να γίνει η πρόβλεψη του νικητή. Παρόλα αυτά με τέτοιου είδους έρευνα είναι φανερό πως η πρόβλεψη του νικητή των εκλογών δεν είναι απλό ζήτημα. Τέλος παρουσιάζονται πιθανές αλλαγές στο μοντέλο καθώς θα μπορούσαμε να έχουμε παρόμοια αποτελέσματα με δέντρα αποφάσεων.

ΚΕΦΑΛΑΙΟ 3 – ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ ΓΙΑ ΤΟ CYBERTRUCK ΤΗΣ TESLA

3.1 Συλλογή Δεδομένων

3.1.1 Παρουσίαση Twitter Developer Και Kaggle

Στην εργασία πραγματοποιούμε μια μελέτη περίπτωσης με την χρήση sentiment analysis σε ένα set δεδομένων. Τα δεδομένα αυτά τα αντλούμε μέσα από το twitter με την χρήση του εργαλείου twitter developer [15].

Το twitter developer είναι μια κοινότητα προγραμματιστών οι οποίοι δουλεύουν με την πλατφόρμα του twitter. Δημιουργώντας μια νέα αίτηση, με τον σκοπό της έρευνας και τον λόγο που επιθυμούμε να χρησιμοποιήσουμε tweets, έχουμε πρόσβαση σε κλειδιά για να αντλήσουμε τα tweets που επιθυμούμε. Το twitter developer μας επιτρέπει να αντλούμε 2000-3000 ~ tweets για κάθε έρευνα, για να διατηρείται η ασφάλεια των δεδομένων. Σε περίπτωση που κάποιος χρήστης κάνει πολλά αιτήματα για να αντλήσει δεδομένα τότε διακόπτεται το δικαίωμα του για την χρήση αυτών. Αυτό σημαίνει πως έχουμε πρόσβαση σε ένα περιορισμένο set δεδομένων το οποίο όμως είναι επαρκές για μία μελέτη περίπτωσης σε ένα τέτοιο θέμα.

Οι περισσότερες έρευνες που πραγματοποιούνται με βάση το sentiment analysis βασίζονται πάνω σε δεδομένα που έχουν αντληθεί από το twitter. Αποτελεί μια πολύ ισχυρή πηγή για την άντληση των δεδομένων καθώς μας παρουσιάζεται απλά και κατανοητά μέσα από 280 χαρακτήρες το συναίσθημα και η άποψη ενός χρήστη για κάποιο θέμα. Επίσης μέσω των hashtag μπορούμε να απορροφάμε στενευμένα τα tweets καθώς αποτελούν κλειδί για την σχετικότητα των δεδομένων μας. Πολύ σημαντικοί παράγοντες θεωρούνται επίσης οι φορές που κάποιος

μπορεί να αναπαράγει το tweet-άποψη κάποιου χρήστη, κάνοντάς το retweet. Αυτό μας βοηθά καθώς η αναπαραγωγή κάποιας άποψης σημαίνει θετικό συναίσθημα και συμφωνία απόψεων. Ακόμη ένα σημαντικό προσόν που μας παρέχεται από τα tweets είναι τα emoticons, τα οποία έχουν αρχίσει να εμφανίζονται σε πιο πρόσφατες έρευνες, καθώς μπορούμε να αξιοποιήσουμε το συναίσθημα που εκφράζουν.

Η συγγραφή του κώδικα έγινε μέσα από την πλατφόρμα του kaggle [14]. Το kaggle είναι μια πλατφόρμα/κοινότητα που αποτελείται από αναλυτές δεδομένων. Μέσα από το kaggle έχουμε την δυνατότητα σύνθεσης κώδικα και της δημοσίευσής του. Έτσι δημιουργείται μια κοινότητα στην οποία μπορούν άλλοι χρήστες να επισκεφτούν την έρευνας μας να μελετήσουν τον κώδικα μας και να προτείνουν βελτιώσεις.

Επίσης έχουμε την δυνατότητα να επικοινωνήσουμε με άλλους χρήστες και να επιλύσουμε προβλήματα που αντιμετωπίσαμε σε δική μας έρευνα. Από την στιγμή που ανήκουμε και εμείς σε μια τόσο μεγάλη κοινότητα, έχουμε την δυνατότητα να μελετήσουμε ή να συμβουλευτούμε άλλες εργασίες παρόμοιες με την δική μας για να παρατηρήσουμε τεχνικές ή προσεγγίσεις που έχουν γίνει σε παρόμοια θέματα.

Το θέμα για το οποίο αντλήσαμε δεδομένα είναι το cyber truck της εταιρίας tesla. Το cyber truck είναι το νέο όχημα της εταιρείας tesla το οποίο έφερε μεγάλες αλλαγές στον χώρο λόγω πολλών καινοτομιών σχετικά με την αυτονομία του, το είδος των καυσίμων του, την τεχνητή νοημοσύνη του οχήματος και τον σχεδιασμό του. Μεγάλες αλλαγές όμως συχνά φέρνουν και μεγάλες αντιδράσεις το οποίο έκρινε ενδιαφέρουσα την μελέτη της ανταπόκρισης του κοινού. Σκοπός της

έρευνας είναι η μελέτη των απόψεων σχετικά με την παρουσίαση στην αγορά ενός τόσο διαφορετικού οχήματος.

Όλες οι κεφαλίδες που παρουσιάζονται με hashtag(#) αποτελούν σημειώσεις το κώδικα και μπορούμε να τις μελετήσουμε αναλυτικά στο παράρτημα.

3.1.2. Import Βασικών Βιβλιοθηκών

Ξεκινάμε κάτω από την σημείωση #INSTALLING LIBRARIES και εγκαθιστούμε - προσθέτουμε βιβλιοθήκες στον κώδικα μας. Το πρώτο library είναι το tweepy το οποίο συνδέεται απευθείας στο tweeter και είναι υπεύθυνο ώστε να μας παρέχονται τα δεδομένα που θέλουμε να τα χρησιμοποιήσουμε. Στην συνέχεια κάνουμε install το library nltk [11][17] (natural language tool kit). Είναι ένα library που μας παρέχει εργαλεία για την επεξεργασία της φυσική γλώσσας, τα οποία χρησιμοποιούμε παρακάτω για να κάνουμε επεξεργασία τον λόγο του κειμένου μας. Μετά, εγκαθιστούμε το library textblob [9] όπου θα το χρησιμοποιήσουμε στο sentiment analysis και θα μας παρέχει μια κατάταξη του tweet μας για το πόσο θετικό (με μέγιστη τιμή 1), αρνητικό (με μέγιστη τιμή -1) ή ουδέτερο είναι.

Στην συνέχεια κάνουμε import το numpy [10] ένα από τα πιο βασικά libraries της python το οποίο περιέχει πολυδιάστατους πίνακες και μας βοηθάει στην εκτέλεση μαθηματικών πράξεων. Κάνουμε import τη re [16] (regex), το όνομα του προέρχεται από τα regular expressions τα οποία τα συναντάμε σε κάθε γλώσσα προγραμματισμού και αφορά την επεξεργασία κειμένου. Μετά το import της re, κάνουμε import και τις βιβλιοθήκες nltk [11][17] και textblob [9] που κάναμε install πριν. Το install πραγματοποιείται μόνο την πρώτη φορά, ενώ τις επόμενες απλά ελέγχει άμα υπάρχει η εγκατάσταση τους.

Προχωράμε κάνοντας import συγκεκριμένα εργαλεία (stopwords, WordNetLemmatizer) που θα χρησιμοποιήσουμε αργότερα στην επεξεργασία κειμένου. Κάνουμε import αυτά τα εργαλεία στην αρχή για να έχουμε πιο απλή την σύνταξη του κώδικα μας. Μετά κάνουμε import το tweepy [13] που κάναμε install στην αρχή του κώδικα και στην συνέχεια το pandas [18]. Η βιβλιοθήκη pandas μας βοηθάει αργότερα στην δημιουργία πινάκων για την οργάνωση των δεδομένων που θα αντλήσουμε. Μετά κάνουμε import την βιβλιοθήκη os σε περίπτωση που θέλουμε να έχουμε πρόσβαση σε αρχεία του υπολογιστή μας, για κάποια μελλοντική επεξεργασία δεδομένων και αυτό επιτρέπει στο πρόγραμμά μας να είναι extrovert. Δηλαδή να έχει την δυνατότητα να επικοινωνεί με τον υπολογιστή που χρησιμοποιούμε. Τέλος η πλατφόρμα του kaggle μας προσθέτει ένα function

```
[ for dirname, _ , filenames in os.walk('/kaggle/input'):  
    for filename in filenames: print(os.path.join(dirname, filename)) ]
```

κάτω από το κομμάτι που πραγματοποιήσαμε τα imports των libraries όπου μας προσφέρει να βλέπουμε αρχεία που έχουμε κάνει upload. Είναι ένα function το οποίο προσθέτει από μόνο του το kaggle by default σε όλα τα project.

Στην συνέχεια στο κομμάτι #KEYS περνάμε σε τέσσερις διαφορετικές μεταβλητές (ckey, csecret, atoken, asecret) τα αντίστοιχα keys. Οι μεταβλητές για κάθε key είναι απλό string κείμενο τα οποία τα παρέχει το twitter μετά από αίτημα μέσα από το tweeter developer. Παρακάτω επικαλούμαστε την κλάση και τα functions του tweepy τα οποία κάνουν αίτημα στο

twitter για να μας δώσει δεδομένα. Ταυτόχρονα όμως μέσω των keys που δηλώσαμε προηγουμένως κάνουν authenticate εμάς και άμα επιτρέπεται η δυνατότητα να αντλήσουμε δεδομένα από το tweeter.

3.1.3 Tracking Keywords

Στο κομμάτι #SETUP μέσα στο TRACKING_KEYWORDS δημιουργούμε μια λίστα με tracking keywords τα οποία αποτελούν τις λέξεις που στοχεύουμε σε κάθε tweet. Τα δικά μας tracking keywords είναι Cybertruck, cubertruck, CYBERTRUCK καθώς ενδιαφερόμαστε για tweets που εμπεριέχουν αυτόν τον όρο. Μία εναλλακτική επιλογή που έχουμε σε περίπτωση που θα θέλαμε να στοχεύσουμε αντί για μεμονωμένα tweet, έναν συγκεκριμένο λογαριασμό έχουμε το TARGET_ACCOUNT. Αυτό θα μας επέτρεπε να αντλήσουμε tweets σχετικά με τα tracking words που αναφέραμε αλλά μόνο μέσα από τα tweets ενός συγκεκριμένου λογαριασμού. Στην δική μας περίπτωση δεν το επιθυμούμε αυτό καθώς εμείς ενδιαφερόμαστε για την άποψη των καταναλωτών του προϊόντος και όχι αποκλειστικά για τα tweet του παραγωγού. Τέλος στο TWEETS_TO_CAPTURE βάζουμε τον αριθμό των tweets που θέλουμε να αντλήσουμε. Το tweeter developer μας επιτρέπει να αντλήσουμε συγκεκριμένο αριθμό tweet (περίπου 3000). Αυτό γίνεται γιατί το tweeter αποτελεί μια από τις μεγαλύτερες δεξαμενές πληροφοριών σχετικά με την επικαιρότητα, για αυτόν τον λόγο επιτρέπουν την ελεγχόμενη άντληση δεδομένων.

3.1.4 Συγκέντρωση Δεδομένων Και Προετοιμασία Για Προεπεξεργασία

Στην συνέχεια στο #D1 δημιουργούμε δυο λίστες (την data_raw, την location) η λίστα data_raw περιέχει ένα στοιχείο, ενώ η location είναι άδεια. Την data_raw θα την χρησιμοποιήσουμε στην συνέχεια για να περάσουμε τα tweets μέσα σε αυτήν και ώστε αργότερα να τα περάσουμε μέσα σε ένα dataframe. Η λίστα location θα μας παρέχει απλά την τοποθεσία του κάθε tweet που

αντλούμε. Μετά χρησιμοποιούμε την μέθοδο `tweepy.Cursor` σε μία επανάληψη και μέσω των `tracking keywords` που προσθέσαμε προηγουμένως αρχίζει να συλλέγει τα `items`. Καθώς συλλέγει τα `items` προσθέτει το `text` από το κάθε `tweet` στην λίστα `data_raw` και την τοποθεσία του στην λίστα `location`. Αφού ολοκληρωθεί η επανάληψη και έχουμε συλλέξει τον αριθμό των `tweets` μας αφαιρούμε από τον πίνακα `data_raw` το στοιχείο 0, δηλαδή το πρώτο στοιχείο την λέξη (“Initializer”). Και μετά εμφανίζουμε την λίστα `data_raw` και μήνυμα `tweets downloaded`.

Στο κομμάτι `#Creates Dataframe and passes text inside` προχωράμε στην δημιουργία του πίνακα. Δημιουργούμε τον πίνακα `df` και χρησιμοποιούμε τη library `pandas (pd)` και την μέθοδο του `DataFrame`. Τα δεδομένα του `dataframe` είναι τα στοιχεία από την λίστα `data_raw` κάτω από την στήλη `text` και αριθμούνται από το `index=None` το οποίο από `default` είναι αύξον αριθμός. Στο επόμενο κομμάτι `#Passes location to dataframe` προσθέτουμε και την τοποθεσία από την λίστα `location` μέσα στο `dataframe` μέσω μιας επανάληψης κάτω από την στήλη `location`.

Παρακάτω βρισκόμαστε στο κομμάτι `#DOWNLOADS ENGLISH STOPWORDS` σε αυτό το σημείο δημιουργούμε την μεταβλητή `stop_words`. Η μεταβλητή αυτή είναι τύπου `set` και μέσα περνάμε όλα τα `English stopwords` για να τα χρησιμοποιήσουμε στην επεξεργασία κειμένου.

3.2 Προεπεξεργασία δεδομένων

3.2.1 Παρουσίαση Προεπεξεργασίας δεδομένων

Αφού αντλήσουμε τα σχετικά δεδομένα περνάμε στο στάδιο της προεπεξεργασίας δεδομένων. Στο στάδιο προεπεξεργασίας σκοπός μας είναι να φέρουμε τα δεδομένα μας στην πιο απλή τους μορφή. Έχοντας τα δεδομένα μας απλοποιημένα παρέχουμε στον υπολογιστή μεγαλύτερη κατανόηση για αυτά. Έτσι παρουσιάζεται μεγαλύτερος αριθμός ακριβείας στα αποτελέσματα

μας. Σκοπός της διαδικασίας είναι να φτάσουμε σε σημείο όπου θα έχουμε απλές λέξεις οι οποίες κρύβουν sentiment, ώστε ο υπολογιστής να κατανοεί το συναίσθημα από κάθε tweet.

3.2.2 Διαδικασία προεπεξεργασίας δεδομένων

Αφού έχουμε συγκεντρώσει τα δεδομένα από τα tweets σε ένα dataframe προχωράμε στο κομμάτι της επεξεργασίας κειμένου **#PREPROCESSING**. Σε αυτό το κομμάτι θα απλοποιήσουμε την μορφή των tweet αφαιρώντας χαρακτήρες και στοιχεία που δεν θα χρησιμοποιήσουμε για την απορρόφηση sentiment από το tweet. Κάθε tweet αναλύεται ξεχωριστά για την αφαίρεση των στοιχείων που δεν μας ενδιαφέρουν. Στο τέλος της επανάληψης το tweet θα επανέρχεται στην αρχική του μορφή χωρίς αυτά τα στοιχεία.

Στο κομμάτι **#LOWERING** μετατρέπουμε όλα τα γράμματα σε μικρά, ώστε να μην υπάρχουν κεφαλαία σε όλα τα tweet. Στην συνέχεια στο **#URL REMOVAL** μέσω της re [16] (regular expression) αφαιρούμε τα url που βρίσκουμε μέσα σε κάθε tweet και τα αντικαθιστούμε με κενά στην θέση τους. Στην συνέχεια κάτω από την σημείωση **#USER REMOVAL** πάλι με την χρήση re (regular expression) χρησιμοποιούμε την ίδια μέθοδο με την αφαίρεση των url απλά έχοντας διαφορετικό κανόνα για την αφαίρεση των users.

Στο κομμάτι **#NUMBER & HTML REMOVAL** στοχεύουμε στην αφαίρεση των αριθμών και όλων των στοιχείων html που μπορούν να βρεθούν στα tweets. Επίσης προσθέτουμε και την αφαίρεση του στοιχείου '&' γιατί παρατηρούμε ότι είναι κομμάτι της html που δεν αναγνωρίστηκε.

Στην συνέχεια περνάμε στο #EMOJI REMOVAL όπου αφαιρούμε τα emoji από κάθε tweet. Πλέον τα emoji εμπεριέχουν sentiment και μέσω νέων ανεπτυγμένων μεθόδων μπορούμε να τα λάβουμε υπόψιν μας στα αποτελέσματα. Τέτοιου είδους τεχνικές παρατηρούμε σε νεότερες εργασίες και ακόμη δεν είναι τόσο εύκολο να εφαρμοστούν. Τα emoticons/emoji μπορούν να προσφέρουν σε μεγάλο βαθμό καθώς τα ίδια αποτελούν ουσιαστικά απεικόνιση του συναισθήματος ενός μηνύματος. Αν και θα μπορούσε να μας προσφέρει μεγάλη αύξηση της ακριβείας των αποτελεσμάτων δεν θα χρησιμοποιηθούν σε αυτή την μελέτη.

Παρακάτω στο κομμάτι #PUNCTUATION REMOVAL αφαιρούμε τα σημεία στίξης ('[.,:;\#]'). Τα σημεία στίξης προσφέρουν στον λόγο και μπορούν να μας δώσουν στοιχεία sentiment. Μια τελεία ή ένα κόμμα, έχει την δυνατότητα να αλλάξει μια φράση ή μια ολόκληρη πρόταση, καθώς προσφέρει έμφαση ή ολοκλήρωση. Όπως και τα emoji βέβαια αποτελεί έναν πολύ εξελιγμένο τρόπο αναγνώρισης sentiment το οποίο δεν θα μελετήσουμε σε αυτή την εργασία.

Στο κομμάτι #TOKENIZATION χρησιμοποιούμε το εργαλείο word_tokenize από το library nltk [17] το οποίο σπάει το string του tweet μας ξεχωριστά σε λέξεις οι οποίες χωρίζονται με κόμμα (,). Αυτό γίνεται γιατί θέλουμε στο sentiment analysis να στοχεύσουμε ξεχωριστά σε κάθε λέξη. Αυτό το επιτυγχάνουμε με τον διαχωρισμό του string σε μία λίστα λέξεων.

Στην συνέχεια στο #REMOVE STOPWORDS & OTHER WORDS μέσω μίας επανάληψης ελέγχουμε το tweet μας για stopwords. Κάθε γλώσσα έχει την ξεχωριστή της λίστα από stopwords. Στην συγκεκριμένη εργασία έχουμε την λίστα με τα Αγγλικά stopwords που αποτελείται από συνθετικές λέξεις και λέξεις καθημερινού λόγου οι οποίες αφαιρούνται από τα

tweets γιατί δεν προφέρουν στο sentiment ενός μηνύματος. Λέξεις που δεν ανήκουν στα stopwords τις περνάμε στην μεταβλητή assistant. Στο τέλος αφού ελέγξουμε ολόκληρο το κείμενο τότε το βάζουμε το tweet μας να ισούται με την μεταβλητή assistant όπου εμπεριέχονται όλες οι “καθαρές” πλέον λέξεις.

Πλέον αφού έχουμε καταλήξει με μία λίστα με λέξεις, στο κομμάτι #LEMMATIZATION επιστρέφουμε την κάθε λέξη της λίστας στην ρίζα της. Αυτό το κάνουμε για να είναι πιο κατανοητές από τον υπολογιστή. Ο υπολογιστής αντιλαμβάνεται τις λέξεις διαφορετικά όταν είναι σε διαφορετικούς χρόνους οπότε με αυτή την απλοποίηση πετυχαίνουμε μεγαλύτερη ακρίβεια στο sentiment.

3.3 Ανάπτυξη εργαλείου ανάλυσης συναισθήματος

Το εργαλείο που χρησιμοποιήθηκε για την ανάλυση του συναισθήματος είναι το Textblob. Η Textblob είναι μια ισχυρή βιβλιοθήκη με προσέγγιση λεξικού (lexicon-based) της python η οποία χρησιμοποιείται σε πολλές εργασίες πλέον σχετικές με την εξόρυξη δεδομένων. Η Textblob έχει την δυνατότητα να μας παρέχει έναν API ο οποίος μας βοηθάει στην εκτέλεση διαφορετικών επεξεργασιών της φυσικής γλώσσας. Ένα από τα αντικείμενα που επικεντρώνεται είναι η συναισθηματική και η υποκειμενική ανάλυση (Sentiment and Subjectivity Analysis). Η υποκειμενική ανάλυση επικεντρώνεται στο μέρος του λόγου κάθε λέξης σε μία πρόταση και εστιάζει στις λέξεις που εκφράζουν το μεγαλύτερο συναίσθημα. Μέρη του λόγου όπως τα ρήματα, τα αντικείμενα και τα επιρρήματα αποτελούν λέξεις που εμπεριέχουν μεγαλύτερο συναίσθημα από άλλες. Αυτό μας βοηθάει στην καλύτερη κατανόηση του κάθε tweet ξεχωριστά και έτσι επιτυγχάνουμε μεγαλύτερα ποσοστά επιτυχίας και κατανόησης των συναισθημάτων.

Αφού έχουμε ολοκληρώσει την επεξεργασία του κειμένου, περνάμε στο κομμάτι #REMAKING TWEET INTO STRING όπου με το εργαλείο join επαναφέρουμε το tweet μας σε μορφή string. Αυτό γίνεται γιατί στην συνέχεια στο κομμάτι #ITERATE TON PINAKA -blob- θέλουμε να μετατρέψουμε τα tweet μας σε μορφή blob με το εργαλείο Text.Blob. Για να μετατραπεί μια μεταβλητή σε μορφή blob χρειάζεται μεταβλητές τύπου string. Το text blob μετατρέπει το string ξανά σε λίστα και βρίσκει τι μέρος του λόγου είναι κάθε λέξη.

Με βάση του τι μέρος του λόγου είναι η κάθε λέξη δίνει διαφορετική βαρύτητα σε αυτές έχοντας έτσι μεγαλύτερη ακρίβεια στο sentiment. Μέσα στην ίδια επανάληψη δημιουργούμε την στήλη sentiment όπου περνάμε με το εργαλείο sentiment.polarity την τιμή του sentiment για το κάθε tweet. Όπως προαναφέραμε οι τιμές για το sentiment αρχίζουν από -1 ως αρνητικό sentiment και φτάνουν μέχρι το 1 ως θετικό sentiment. Αφού έχουμε στο data frame πλέον και το sentiment, στο κομμάτι #SORTING DATAFRAME BY SENTIMENT κάνουμε sort το dataframe με βάση το sentiment ώστε να μπορούμε να κάνουμε πιο εύκολη μελέτη των αποτελεσμάτων.

Τέλος στο κομμάτι #OVERALL SENTIMENT βγάζουμε τον μέσο όρο όλων των sentiment από τα tweets που μελετήσαμε και το εκτυπώνουμε για να δούμε ένα γενικό δείγμα για το sentiment του θέματος.

ΚΕΦΑΛΑΙΟ 4 – ΑΞΙΟΛΟΓΗΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Αφού ολοκληρωθεί το πρόγραμμα ξεκινάμε την μελέτη των αποτελεσμάτων. Τα αποτελέσματα τα χωρίζουμε σε τρεις κατηγορίες. Στα θετικά tweet (κείμενα με βαθμολογία μεγαλύτερη του 0), στα αρνητικά tweet (κείμενα με βαθμολογία μικρότερη του 0) και στα ουδέτερα (κείμενα με βαθμολογία ίση με 0).

Όπως προαναφέραμε τα αποτελέσματα περιείχαν τα tweet στην απλή τους μορφή, την τοποθεσία που δημοσιεύτηκε το κάθε tweet (θα μπορούσε να χρησιμοποιηθεί για μελλοντική έρευνα για τον τρόπο αντιμετώπισης ενός προϊόντος με βάση την γεωγραφική θέση του συντάκτη του tweet) και τέλος το sentiment για το κάθε tweet. Ένα από τα προβλήματα άντλησης αποτελεσμάτων ακριβείας ήταν το όριο δύναμης του tweeter developer. Ένα συχνό πρόβλημα για την άντληση δεδομένων από το tweeter developer είναι η επαναλαμβανόμενη άντληση όμοιων tweets. Το tweeter developer επιτρέπει πολύ περιορισμένη άντληση δεδομένων καθώς και είναι πολιτική του tweeter να μην δίνει ελεύθερα τα δεδομένα σε μεγάλο όγκο. Βέβαια ο αριθμός που χρησιμοποιήθηκε για την έρευνα ήταν πολύ ικανοποιητικός.

Παρατηρείται επίσης στο σετ δεδομένων, πώς πολλά tweets εμφανίζονται παραπάνω από μία φορά. Αυτό σημαίνει πως άλλοι χρήστες έκαναν αναπαραγωγή αυτών των δεδομένων (μέσω retweet). Από αυτό μπορούμε να συμπεράνουμε θετικά αποτελέσματα, καθώς η αναπαραγωγή ενός tweet συμβάλει στην υπεράσπιση της άποψης κάποιου άλλου χρήστη.

Η ακρίβεια των αποτελεσμάτων βρίσκεται στο 86.8% το οποίο είναι ένα πολύ θετικό και πετυχημένο αποτέλεσμα αρκετά ανώτερο από τον μέσο όρο που παρουσιάζεται σε άλλες

εργασίες. Παρατηρώντας τα δεδομένα μπορούμε να δούμε πως υπάρχει μεγάλος αριθμός θετικών και ουδέτερων tweet ενώ υπάρχουν αρκετά λιγότερα αρνητικά. Βλέποντας τον μέσο όρο της βαθμολογίας του sentiment των δεδομένων μας (overall sentiment average = 0.102423) μπορούμε να αντλήσουμε την πληροφορία πως το αντικείμενο που ερευνούμε (το cybertruck) έχει κινήσει θετικά τους καταναλωτές, καθώς η βαθμολογία μας κυμαίνεται κυρίως σε βαθμολογίες sentiment ανώτερες του μηδέν. Ο μέσος όρος μας δίνει μια πολύ καλή εικόνα για την γενική αντιμετώπιση του θέματος από τους χρήστες, αλλά για βαθύτερη κατανόηση χρειάζεται να μελετήσουμε μεμονωμένα σχόλια και την ακρίβεια του Sentiment που παρήγαγε ο αλγόριθμος.

Στην συνέχεια έγινε ή κάθε παρακολούθηση του κάθε tweet ξεχωριστά ώστε να οριστεί το πραγματικό sentiment σε ακραίες αποκλείσεις (δηλαδή για περιπτώσεις tweets που εξέφραζαν ειρωνεία και το πρόγραμμα δεν ήταν ικανό να τα αναγνωρίσει). Αρχίζοντας την μελέτη του κάθε tweet, παρατηρούμε πως υπήρχαν σε κάποια tweets, αποκλείσεις από τις βαθμολογημένες προβλέψεις του sentiment.

Τα περισσότερα ουδέτερα tweets ήταν απλές παρατηρήσεις σχετικά με το προϊόν. Πολλά από αυτά ήταν ακόμη παρατηρήσεις από την παρουσίαση του προϊόντος ή σχετικά με την αναφορά των χαρακτηριστικών του. Κάποια από αυτά βέβαια μπορεί να παρουσίαζαν κάποιον βαθμό ειρωνείας στον λόγο τους μειώνοντας έτσι το προϊόν. Πολλά tweet επίσης εμπεριείχαν κάποιον βαθμό θαυμασμού. Τέτοια σημεία μας αποδεικνύουν πως το sentiment κρίνεται ξεχωριστά από τις λέξεις και όχι από το ολικό νόημα μίας φράσης.

Παρατηρούμε επίσης, πως η βαθμολογία των περισσότερων δεδομένων ήταν αρκετά απόλυτη. Δηλαδή παρατηρούμε πως τα περισσότερα δεδομένα είχαν τιμές πιο κοντινές στο 1, στο 0 και στο -1, παρά σε ενδιάμεσες τιμές και δεκαδικά. Για αυτό το κομμάτι πρέπει να ευθύνεται το textblob που όπως προαναφέραμε, δίνει μεγαλύτερη βαρύτητα σε διαφορετικά μέρη του λόγου (ρήματα, επιρρήματα, αντικείμενα). Οπότε τέτοια μέρη του λόγου είχαν την ικανότητα να αυξήσουν ή να μειώσουν σε μεγάλο βαθμό την βαθμολογία του sentiment από κάθε tweet.

Τέλος τα αποτελέσματα που πήραμε είχαν ικανοποιητικό ποσοστό επιτυχίας το οποίο θα μπορούσε να αυξηθεί με κάποιες μελλοντικές αλλαγές στον τρόπο επεξεργασίας των δεδομένων. Η εικόνα που παίρνουμε για το προϊόν είναι κατά μέσο όρο θετική μέσα από το sentiment των καταναλωτών οπότε παίρνουμε άμεσα τα δεδομένα που αναζητούσαμε μέσα από σχόλια του κοινού.

Predicted: Negative	275	Actually: Negative	377
Predicted: Neutral	1132	Actually: Neutral	834
Predicted: Positive	1093	Actually: Positive	1289

1 Πίνακας Prediction - Actual

M.O. (Overall Sentiment Average)	0.102423
----------------------------------	----------

2 Πίνακας M.O. αποτελεσμάτων

	Predicted: Negative	Predicted: Neutral	Predicted: Positive
Actually: Negative	268	86	23
Actually: Neutral	0	834	0
Actually: Positive	7	212	1070

3 Πίνακας Confusion Matrix

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην έρευνα αυτή χρησιμοποιήσαμε πολλά εργαλεία για να φτάσουμε στην απόδοση συναισθήματος. Με την χρήση του kaggle συνθέσαμε τον κώδικα για την ανάλυση των δεδομένων, τα δεδομένα αντλήθηκαν με την χρήση keywords μέσω του twitter developer και ο κώδικας μας πρόσφερε δεδομένα τα οποία μελετήσαμε και στο τέλος πραγματοποιήσαμε μια σύγκριση για την απόδοση του αλγορίθμου κατηγοριοποίησης συναισθημάτων.

Με αυτό τον τρόπο προσέγγισης που επιλέξαμε μπορούμε να παρατηρήσουμε ότι υπάρχουν περιθώρια βελτίωσης, αναπτύσσοντας έναν κώδικα όπου παράγει ακόμη πιο ακριβή αποτελέσματα χρησιμοποιώντας την ίδια τεχνική κατηγοριοποίησης συναισθημάτων.

Λαμβάνοντας υπόψη επιπρόσθετες σύγχρονες παραμέτρους θα μπορούσαμε να έχουμε αρκετά διαφορετικά αποτελέσματα. Όπως είδαμε και σε παρόμοιες έρευνες στην βιβλιογραφία, πολύ συχνή είναι και η μελέτη χαρακτηριστικών όπως emoticons που συμπεριλαμβάνονται μέσα στα δεδομένα, τα οποία μπορούν να αποδώσουν πιο ξεκάθαρα το συναίσθημα των δεδομένων ακόμα και όταν υπάρχει σύγχυση σχετικά με το λεκτικό νόημα.

Σημεία τα οποία θα μπορούσαν να βελτιωθούν επίσης, είναι η επέκταση του λεξικού με την δυνατότητα αναγνώρισης συναισθήματος λέξεων που δεν ανήκουν στο λεξικό. Όπως συναντήσαμε στην βιβλιογραφία υπάρχει η δυνατότητα επέκτασης του λεξικού, αναγνωρίζοντας το συναίσθημα νέων λέξεων με την χρήση των ήδη υπαρχόντων λέξεων ή φράσεων που βρίσκονται ήδη στο λεξικό.

Τέλος μπορούμε να δούμε πως τα αποτελέσματα είναι πολύ ικανοποιητικά κρίνοντας από τον αριθμό των δεδομένων που διαθέτουμε και από την σύγκριση της απόδοσης του αλγορίθμου κατηγοριοποίησης μέσα από το confusion matrix. Μελλοντικός στόχος για την επέκταση της έρευνας θα μπορούσε να είναι η αναβάθμιση του κώδικα λαμβάνοντας υπόψη μας περισσότερες παραμέτρους αλλά και η μελέτη ενός ακόμη μεγαλύτερου dataset.

Ευχαριστίες

Η εργασία αυτή αποτελεί το τελικό στάδιο των σπουδών μου καθώς με την αποπεράτωσή της ολοκληρώνεται η φοίτησή μου στο Τμήμα Πληροφορικής και συνάμα στην Κέρκυρα, ενός δημιουργικού και ιδιαίτερα ευχάριστου κεφαλαίου της ζωής μου.

BIBΛΙΟΓΡΑΦΙΑ

- [1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau. Columbia University. Sentiment Analysis of Twitter Data
- [2] Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, Ting Liu. Harbin Institute of Technology, China Microsoft Research, Beijing, China, Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach
- [3] Yelena Mejova, Padmini Srinivasan. University of Iowa, Exploring Feature Definition and Selection for Sentiment Classifiers
- [4] Guang Qiu, Bing Liu, Jiajun Bu, Chun Chen. Zhejiang University, University of Illinois at Chicago, Expanding Domain Sentiment Lexicon through Double Propagation
- [5] France Cheong, Christopher Cheong. RMIT University. Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods
- [6] Adham Beykikhoshk, Ognjen Arandjelovic, Dinh Phung, Svetha Venkatesh, Terry Caelli. Centre for Pattern Recognition and Data Analytics Deakin Univeristy, Australia, National ICT Australia (NICTA) Melbourne, Australia. Data-Mining Twitter and the Autism Spectrum Disorder: A Pilot Study
- [7] Aditya L. Ramadonaa, Rendra Agusta, Sulistyawati, Lutfan Lazuardi, Anwar D. Cahyono, Åsa Holmner, Fatwa S.T. Dewi, Hari Kusnanto, Joacim Rocklöv. Department of Public Health and Clinical Medicine, Epidemiology and Global Health, Umeå University. Mining of health and disease events on Twitter: validating search protocols within the setting of Indonesia
- [8] Lazaros Oikonomou, Christos Tjortjis. The Data Mining and Analytics Research Group, School of Science & Technology, International Hellenic University. A Method for Predicting the Winner of the USA Presidential Elections using Data extracted from Twitter
- [9] Steven Loria, textblob Documentation release 0.16.0, April 26 2020
- [10] NumPy community, NymPy Reference release 1.19.0, June 29 2020
- [11] Steven Bird, Ewan Klein, and Edward Loper, Natural Language Processing Toolkit with python Analysing Text With Natural Language Toolkit. Link: <https://www.nltk.org/book/>
- [12] Wes McKinney and the Pandas Development Team, pandas: powerful Python data analysis toolkit Release 1.1.4, October 30 2020. Link: <https://pandas.pydata.org/docs/index.html>
- [13] Tweepy Documentation. Link: <http://docs.tweepy.org/en/latest/#tweepy-documentation>

Διαδικτυακοί τόποι

[14] Kaggle Documentation. Link: <https://www.kaggle.com/docs>

[15] Twitter Developer. Link: <https://developer.twitter.com/en/docs>

[16] Regular Expressions. Link: <https://docs.python.org/3/library/re.html>

[17] Natural Language Processing Toolkit NLTK. Link: <https://www.nltk.org/book/>

[18] Pandas. Link: <https://pandas.pydata.org/docs/index.html>

IIAPAPTHMA

INSTALLING LIBRARIES

```
!pip install tweepy # eggathistoy n ta packages
!pip install nltk
!pip install textblob
import numpy as np # linear algebra den xreisimopoieitai
import re
import nltk # stopwords
import textblob
from nltk.corpus import stopwords
from nltk.stem.wordnet import WordNetLemmatizer
import tweepy # twitter related actions
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

#KEYS

```
ckey="XEcKX41E6s9t2ruPogK7MerMn"
csecret="W9exWka6Z4a5AVGse2x71Mz9RDMv62l3Sf9cYFkmuj1rEvaeNH"
atoken="1201827837965148161-DQxxLKwx2GdgROX7nptii9EyiYniQw"
asecret="qsHgWpkXZZGMHSBSjfTdr41pkjwTTHBypWYCPGi9dJw6i"
```

```
auth = tweepy.OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
api = tweepy.API(auth)
```

```
#SETUP 2tracking h target den ginontai kai ta dyo mx tweets 3200
TRACKING_KEYWORDS = ["Cybertruck", "cybertruck", "CYBERTRUCK"]
TARGET_ACCOUNT = ""
TWEETS_TO_CAPTURE = 2500
```

#D1

```
data_raw = ["Initializer"]
location = []
for status in tweepy.Cursor(api.search,q=TRACKING_KEYWORDS, tweet_mode="extended",
lang="en").items(TWEETS_TO_CAPTURE):
```

```
data_raw.append(status.full_text) # .full_text affects data type pulled, status alone grabs everything
location.append(status.user.location)
del data_raw[0]
print ("Tweets downloaded", len(data_raw))
```

```
#Creates Dataframe and passes text inside
df = pd.DataFrame(data=data_raw, index=None, columns=["text"], dtype=str)
print(df)
```

```
#Passes location to dataframe
for i in range(len(location)):
    df.loc[i, "location"] = location[i]
```

```
#DOWNLOADS ENGLISH STOPWORDS
stop_words=set(stopwords.words("english"))
```

```
#PREPROCESSING
```

```
assistant = ""
```

```
for i in df.index.values:
```

```
    # LOWERING
```

```
    assistant = df.loc[i, "text"]
```

```
    df.loc[i, "text"] = assistant.lower()
```

```
    # URL REMOVAL
```

```
    assistant = df.loc[i, "text"]
```

```
    assistant = re.sub(r"http\S+", "", assistant)
```

```
    df.loc[i, "text"] = assistant
```

```
    # USER REMOVAL
```

```
    assistant = df.loc[i, "text"]
```

```
    assistant = re.sub(r"(?:\@|https?\:\/\/)\S+", "", assistant)
```

```
    df.loc[i, "text"] = assistant
```

```
    # NUMBER & HTML REMOVAL
```

```
    assistant = df.loc[i, "text"]
```

```
    assistant = re.sub('[0-9]+', "", assistant)
```

```
    assistant = re.sub('&amp;', "", assistant)
```

```
    df.loc[i, "text"] = assistant
```

```
# EMOJI REMOVAL
assistant = df.loc[i, "text"]
assistant = re.sub(r'^\x00-\x7F+', '', assistant)
df.loc[i, "text"] = assistant
```

```
# PUNCTUATION REMOVAL
assistant = df.loc[i, "text"]
assistant = re.sub(r'[.,:;\#\']', '', assistant)
df.loc[i, "text"] = assistant
```

```
# TOKENIZATION
assistant = df.loc[i, "text"]
assistant = nltk.word_tokenize(assistant)
df.loc[i, "text"] = assistant
```

```
# REMOVE STOPWORDS & OTHER WORDS
other = ["'s"]
stop_words = list(stop_words)
stop_words.extend(other)
assistant = []
for word in df.loc[i, "text"]:
    if word not in stop_words and word != "rt":
        assistant.append(word)
df.loc[i, "text"] = assistant
```

```
# LEMMATIZATION
lemmatizer = WordNetLemmatizer()
assistant = []
for word in df.loc[i, "text"]:
    word = lemmatizer.lemmatize(word)
    assistant.append(word)
df.loc[i, "text"] = assistant
```

```
# REMAKING TWEET INTO STRING
assistant = df.loc[i, "text"]
assistant = " ".join(assistant)
df.loc[i, "text"] = assistant
```

```
# ITERATE TON PINAKA -blob-
assistant = ""
```

```
for i in df.index.values:
    blob = textblob.TextBlob(df.loc[i, "text"])
    df.loc[i, "sentiment"] = blob.sentiment.polarity
```

```
# SORTING DATAFRAME BY SENTIMENT
df = df.sort_values(by='sentiment', ascending=True)
```

```
# OVERALL SENTIMENT
assistant = 0
amount_of_tweets = len(df.index.values)
for i in df.index.values:
    assistant = df.loc[i, "sentiment"] + assistant

overall_sentiment = assistant / amount_of_tweets
print("THE OVERALL SENTIMENT IS",overall_sentiment )
```