Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Πανεπιστήμιο Θεσσαλίας

Ανάλυση δεδομένων κοινωνικών δικτύων με εφαρμογή στο Twitter

Data Analysis of Social Networks with Application on Twitter

Ουρανία Τσιλομήτρου

Επιβλέποντες: Μιχαήλ Βασιλακόπουλος, Αναπληρωτής Καθηγητής

Ασπασία Δασκαλοπούλου, Επίκουρος Καθηγήτρια

Στην οικογένειά μου

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή κ. Μιχαήλ Βασιλακόπουλο που με την καθοδήγηση και τις συμβουλές του συνέβαλε στην ολοκλήρωση της Διπλωματικής μου Εργασίας, καθώς και την επιβλέπουσα καθηγήτρια κα. Ασπασία Δασκαλοπούλου.

Θα ήθελα επίσης να ευχαριστήσω θερμά την οικογένεια μου για την στήριξη που μου προσέφερε καθ' όλη τη διάρκεια των σπουδών μου.

Τέλος, ευχαριστώ όλους τους φίλους και συναδέλφους με τους οποίους συμπορευτήκαμε όλα τα χρόνια των σπουδών μας. Ένα ιδιαίτερο ευχαριστώ στον Θάνο Γκαντσίδη για την καθοριστική συμβολή και την υπομονή του.

Περίληψη

Η ραγδαία ανάπτυξη του Παγκόσμιου Ιστού (WWW) τα τελευταία χρόνια, καθώς και το όλο και αυξανόμενο ενδιαφέρον που παρουσιάζεται για τα κοινωνικά δίκτυα, έχουν αδιαμφισβήτητα δημιουργήσει εύφορο έδαφος για την ανάπτυξη του τομέα της Ανάλυσης Δεδομένων. Αναφερόμαστε κυρίως στην Ανάλυση Δεδομένων που προέρχονται από τα κοινωνικά δίκτυα, μιας και ο όγκος των δεδομένων που περιέχονται σε αυτά είναι τεράστιος.

Η παρούσα Διπλωματική Εργασία πραγματεύεται την Ανάλυση Δεδομένων τα οποία προέρχονται από το κοινωνικό δίκτυο Twitter, το οποίο αποτελεί κατά γενική ομολογία το μέσο που προωθεί τον διαδικτυακό διάλογο περισσότερο από κάθε άλλο. Οι εκατομμύρια χρήστες του δημιουργούν καθημερινά τόση πληροφορία ώστε να απαιτείται η ανάλυση αυτής. Ο σκοπός αυτής της ανάλυσης είναι από τη μία να εξασφαλίζεται η εξατομικευμένη προσέγγιση του κάθε χρήστη και από την άλλη να βγαίνουν συμπεράσματα για την άποψη της κοινής γνώμης με σκοπό της περαιτέρω αξιοποίηση αυτών.

Με βάση τα παραπάνω, ασχοληθήκαμε με την υλοποίηση δύο διαφορετικών υπηρεσιών ώστε να καλύψουμε και τις δύο δυνατές προσεγγίσεις της Ανάλυσης Δεδομένων στα κοινωνικά δίκτυα. Το πρώτο κομμάτι του συστήματος μας έχει να κάνει με την Ανάλυση των Δεδομένων που συλλέξαμε από το Twitter, με σκοπό τον εντοπισμό των πιο διαδεδομένων θεμάτων συζήτησης μεταξύ των χρηστών του Twitter και την Συναισθηματική Ανάλυση αυτών. Τα συμπεράσματα που προκύπτουν από μία τέτοια ανάλυση μπορούν να αξιοποιηθούν δημιουργικά από τρίτους.

Το δεύτερο μέρος της υλοποίησης μας αποτελεί μία προσπάθεια ώστε μέσω της Ανάλυσης Δεδομένων να μπορέσουμε να προσφέρουμε στους χρήστες του Twitter μία υπηρεσία πρότασης άλλων χρηστών για σύναψη διαδικτυακής φιλίας. Ο αλγόριθμος που αναπτύξαμε βασίζεται στη σκέψη ότι οι χρήστες του Twitter επιθυμούν να παρακολουθούν τις δημοσιεύσεις ατόμων με τα οποία μοιράζονται κοινές απόψεις για κάποια θέματα. Η ολοκλήρωση της ανάπτυξης της υπηρεσίας αυτής δεν θα ήταν δυνατή χωρίς την εφαρμογή και πάλι ενός αλγορίθμου Συναισθηματικής Ανάλυσης.

Το σύστημα το οποίο αναπτύξαμε υποβλήθηκε σε πειραματική αξιολόγηση, τα αποτελέσματα της οποίας ήταν ενθαρρυντικά και παρατίθενται αναλυτικά στην εργασία αυτή.

Abstract

The rapid growth of the World Wide Web in parallel to the constantly increasing interest for social networks have undeniably created an opportunity for the development of the field of Data Analysis. Specifically, the field that we refer to is Analysis on Data collected from the social networks. This occurs because of the great volume of information that are derived from the social networks.

In this thesis, we deal with data collected from the social network Twitter. Twitter is known as the social media that promotes the dialogue among its users more than every other. Twitter's millions of users create every day huge quantities of information that constitute our dataset. We conduct analysis on this dataset in order to ensure a personalized approach for every user. For instance, by conducting sentiment analysis on a user's timeline we are capable of proposing to him/her other users to follow who have the same beliefs with him/her..

Moreover, by conducting sentiment analysis on Twitter's stream, we are capable of finding the top trends on Twitter for a specific time lapse. This provides important knowledge, especially if someone wants to know the public opinion, for example, about a product or a company.

In order to cover these two different approaches, we implement two different tasks. Firstly, we find the ten top trending topics on Twitter for a specific time lapse and users' opinion about these. Secondly, we come up with a list of users that could be suitable for following from a specific user. This list is being created based on the results of Sentiment Analysis on every user that refers to a trending topic.

We conducted a great number of experiments using our system that gave us positive results. These results are presented in detail in this thesis.

Περιεχόμενα

Ευχαριστίες 3
Abstract 6
Λίστα Εικόνων 9
Λίστα Πινάκων 10
1. Εισαγωγή 11
1.1. Περιγραφή 11
1.2. Διάρθρωση της διπλωματικής 13
2. Θεωρητικό Υπόβαθρο 14
2.1 Κοινωνικά Δίκτυα (Social Media) 14
2.2 Το κοινωνικό δίκτυο Twitter 15
2.2.1 Γενικά για το Twitter 15
2.2.2 Επικοινωνία με το Twitter 17
2.2.3 Εξελίξεις στο Twitter 18
2.3 Εξόρυξη Δεδομένων (Data Mining) 19
2.4 Ανάλυση Συναισθήματος (Sentiment Analysis) 20
2.4.1 Τεχνικές με Επιβλεπόμενη Μηχανική Μάθηση 21
2.4.2 Τεχνικές με Μη Επιβλεπόμενη Μηχανική Μάθηση 21
2.4.3 Τεχνικές Ανάλυσης Συναισθήματος με Χρήση Συναισθηματικού Λεξικού 22
2.4.4 Εργαλεία Ανάλυσης Συναισθήματος 24
2.5 Διαδικασία Αφαίρεσης Καταλήξεων (Stemming) 25
3. Υλοποίηση Συστήματος 27
3.1 Αρχιτεκτονική Συστήματος 28
3.2 Εντοπισμός Σημαντικών Θεμάτων και Εφαρμογή σε αυτά Ανάλυσης Συναισθήματος 29
3.2.1 Επικοινωνία με το Twitter και Συλλογή Δεδομένων 29
3.2.2 Φιλτράρισμα Δεδομένων 31
3.2.3 Εύρεση των Σημαντικότερων Θεμάτων 32
3.2.4 Αφαίρεση καταλήξεων (stemming) 34
3.2.5 Συναισθηματική Ανάλυση (Sentiment Analysis) 35
3.3 Στοχευμένη πρόταση φιλίας μεταξύ των χρηστών 38
3.3.1 Επικοινωνία με το Twitter και Συλλογή Δεδομένων 40

3.3.2 Εύρεση ταυτότητας του κάθε χρήστη 41
3.3.3 Συλλογή Δεδομένων της Δραστηριότητας των Χρηστών 42
3.3.4 Επεξεργασία Δεδομένων 44
3.3.5. Συναισθηματική Ανάλυση για κάθε χρήστη 46
3.3.6 Αποτελέσματα αλγορίθμου πρότασης φίλων 47
4. Πειραματική Αξιολόγηση 49
4.1 Δεδομένα 49
4.2 Παρουσίαση πρώτου πειράματος 49
4.3 Παρουσίαση δεύτερου πειράματος 59
5.Σχετική Έρευνα 62
6.Συμπεράσματα 65
6.1 Μελλοντική Έρευνα 66
Αναφορές-Βιβλιογραφία 67
Παράρτημα Α 69
Παράρτημα Β 70
Παράρτημα Γ 79
Παράρτημα Δ 81

Λίστα Εικόνων

Figure 1 Στατιστικά κοινωνικών δικτύων 2017 15
Figure 2 Στοιχεία σχετικά με το Twitter 16
Figure 3 Σχεδιασμός των αλλαγών που ετοιμάζει το Twitter. 18
Figure 4 Διαδικασία Εξόρυξης Δεδομένων. 20
Figure 5 Αναπαράσταση πολικότητας Λέξεων 23
Figure 6 Αναπαράσταση Ανάλυσης Συναισθήματος 24
Figure 7 Αναπαράσταση του αλγορίθμου αφαίρεσης καταλήξεων του Portet Stemmer 25
Figure 8 Τα πρώτα βήματα του συστήματός μας. 28
Figure 9 Διαδικασία απόκτισης αναγνωριστικών για σύνδεση με τις διεπαφές του Twitter 30
Figure 10 Παράδειγμα εντοπισμού των tweets που δεν περιέχουν hashtag και κατηγοριοποίηση. 34
Figure 11 Βήματα ολόκληρης της διαδικασίας. 35
Figure 12 Διαδικασία σύγκρισης των λέξεων ενός tweet με τις λέξεις του λεξικού μας. 37
Figure 13 Η δομή ενός tweet. 40
Figure 14 Παράδειγμα ανάδειξης των Ids των χρηστών που αναφέρθηκαν σε ένα hashtag. 41
Figure 15 Παράδειγμα συλλογής των 20 tweets όλων των χρηστών ενός hashtag. 43
Figure 16 Ολόκληρη η διαδικασία της δεύτερης υπηρεσίας μας. 46
Figure 17 Αναπαράσταση της μορφής των αποτελεσμάτων μας για όλα τα hashtags. 48
Figure 18 Hashtag σε συνδυασμό με πλήθος των tweets που τα περιέχουν. 52
Figure 19 Αποτελέσματα ανάλυσης συναισθήματος για το πρώτο πείραμά μας. 53
Figure 20 Αποτελέσματα εύρεσης συγκεκριμένου συναισθήματος για κάποια από τα hashtags. 54
Figure 21 Χρήστες που σχολιάζουν θετικά ή αρνητικά κα σε πόα tweets. 57
Figure 22 Hashtag σε συνδυασμό με πλήθος των tweets που τα περιέχουν. 59
Figure 23 Αποτελέσματα Συναισθηματικής Ανάλυσης γισ το δεύτερο πείραμα. 60

Λίστα Πινάκων

Πίνακας 1 Λέξεις που αναζητούμε για συλλογή tweets. 31
Πίνακας 2 Παράδειγμα με κάποιες stopwords. 32
Πίνακας 3 Τα δεδομένα του πρώτου πειράματος. 49
Πίνακας 4 Εντοπισμός δημοφιλών λέξεων για κάθε hashtag. 50
Πίνακας 5 Πόσα tweets κατηγοριοποιήσαμε για κάθε hashtag. 51
Πίνακας 6 Τα tweets που εντοπισαμε για κάθε Id και ο\η πολικότητά τους. 56
Πίνακας 7 Αποτελέσματα για το #cyprus. 58
Πίνακας 8 Δεδομένα δεύτερου πειράματος. 59
Πίνακας 9 Αποτελέσματα για το #paobc. 61

1. Εισαγωγή

1.1. Περιγραφή

Ο σύγχρονος άνθρωπος καταναλώνει πολύ χρόνο από την καθημερινότητά του στο Διαδίκτυο (Internet) το οποίο αποτελεί πια το κυριότερο μέσο επικοινωνίας και ενημέρωσης. Αφενός πρόκειται για τη βασική πηγή ενημέρωσης για θέματα Οικονομικού, Κοινωνικού, Πολιτικού και Επιστημονικού ενδιαφέροντος και αφετέρου συμβάλει στη δημιουργία αλλά και τη διακίνηση πληροφορίας.

Τα μέσα κοινωνικής δικτύωσης είναι ψηφιακές πλατφόρμες οι οποίες αποτελούνται από χρήστες που έχουν δημιουργήσει ο καθένας ένα δικό του προφίλ και μοιράζεται με τους υπόλοιπους χρήστες πληροφορίες με περιεχόμενο διαφόρων μορφών. Κάποια από τα πιο διαδεδομένα μέσα κοινωνικής δικτύωσης είναι το Facebook, το Twitter και το Instagram. Λόγω της ραγδαίας εξέλιξης τους τα κοινωνικά δίκτυα αποτελούν πια την κυριότερη πηγή άμεσης πληροφόρησης του ατόμου, ενώ προσφέρουν στους χρήστες τεράστιες δυνατότητες για έκφραση ιδεών και ανταλλαγή απόψεων με ένα μεγάλο πλήθος ατόμων που αποτελεί τους διαδικτυακούς φίλους τους. Αυτό έχει σαν συνέπεια να αντανακλάται στα μέσα κοινωνικής δικτύωσης η άποψη της κοινής γνώμης ενώ παράλληλα τα ίδια τα μέσα κοινωνικής δικτύωσης να συμβάλλουν σκόπιμα ή μη στη άμεση και έμμεση διαμόρφωση αυτής.

Οι παραπάνω παρατηρήσεις μας οδηγούν στο συμπέρασμα ότι οι χρήστες των κοινωνικών δικτύων είναι σε θέση να βγάζουν συμπεράσματα για προϊόντα ή υπηρεσίες μέσα από τις σχετικές απόψεις άλλων χρηστών που εκφράζονται στα κοινωνικά δίκτυα. Επιπλέον, συμπεραίνουμε ότι όλη αυτή η πληροφορία είναι ιδιαίτερα χρήσιμη, καθότι μπορεί να αξιοποιηθεί με ποικίλους τρόπους. Ένας από αυτούς είναι η αξιοποίηση της πληροφορίας που προκύπτει από τα κοινωνικά μέσα από εταιρίες ώστε να αποκτήσουν άποψη για το πως βλέπουν οι καταναλωτές τα προϊόντα τους. Ένας άλλος τρόπος αξιοποίησης των συγκεκριμένων δεδομένων είναι η προσπάθεια για στοχευμένη διαφήμιση στα κοινωνικά δίκτυα. Οι πληροφορίες που μπορεί να αντλήσει κανείς από τα

κοινωνικά δίκτυα είναι σε θέση ακόμα και να δείξουν την άποψη που διατηρούν οι χρήστες για κρίσιμα εθνικά ζητήματα.

Ταυτόχρονα, λόγω του ότι ο όγκος της πληροφορίας που καλείται να επεξεργαστεί κάθε χρήστης των κοινωνικών δικτύων είναι τεράστιος και ίσως συχνά εκτός των ενδιαφερόντων του, γίνεται έκδηλη η ανάγκη για ανάπτυξη του τομέα της ανάλυσης δεδομένων κοινωνικών δικτύων ως προς όφελος των χρηστών. Πρόκειται για τη μελέτη δεδομένων που είναι διαθέσιμα στα κοινωνικά δίκτυα και την αυτοματοποιημένη εξαγωγή συμπερασμάτων σχετικά με τις προτιμήσεις των χρηστών με σκοπό μελλοντικές προτάσεις σε αυτούς.

Η ανάλυση μέσω της οποίας βγάζουμε συμπέρασμα σχετικά με την γνώμη ενός χρήστη πάνω σε ένα θέμα λέγεται Ανάλυση Συναισθήματος (Sentiment Analysis). Έχει ως στόχο τον χαρακτηρισμό μίας γνώμης ως θετική, αρνητική ή ουδέτερη και την εύρεση του συγκεκριμένου συναισθήματος που εκφράζεται σε κάθε περίπτωση (θυμός, αηδία, φόβος, χαρά, λύπη, έκπληξη). Στην πλειοψηφία των περιπτώσεων, η Ανάλυση Συναισθήματος εφαρμόζεται στην Αγγλική γλώσσα καθώς αποτελεί μία εύκολη επιλογή. Ένας από τους λόγους είναι ο περιορισμένος αριθμός λέξεων που την χαρακτηρίζει σε σύγκριση με την Ελληνική γλώσσα.

Στην εργασία αυτή επιλέξαμε την ανάλυση δεδομένων στην Ελληνική γλώσσα που προέρχονται από το κοινωνικό μέσο Twitter. Αναλύσαμε τα tweets που δημοσιεύθηκαν σε ένα συγκεκριμένο χρονικό διάστημα ώστε να εντοπίσουμε τα δέκα πιο σημαντικά θέματα συζήτησης και εφαρμόσαμε ανάλυση συναισθήματος σε αυτά. Επιπλέον, για κάθε ένα από τα σημαντικά θέματα κατηγοριοποιήσαμε τους χρήστες ανάλογα με τη γνώμη τους για το θέμα και με το πόσο έντονη είναι η ενασχόληση τους με αυτό, με σκοπό την στοχευμένη πρόταση φιλίας μεταξύ των χρηστών.

1.2. Διάρθρωση της διπλωματικής

Το υπόλοιπο της παρούσας διπλωματικής είναι οργανωμένο όπως φαίνεται παρακάτω.

Το Κεφάλαιο 2 αποτελεί την ανάλυση του Θεωρητικού Υποβάθρου που είναι σχετικό με αυτή την Διπλωματική Εργασία. Αναλύονται έννοιες όπως η Εξόρυξη Δεδομένων, η Ανάλυση Συναισθήματος και Διαδικασία Αφαίρεσης Καταλήξεων. Γίνεται επίσης αναφορά στα κοινωνικά δίκτυα στη σημερινή εποχή και συγκεκριμένα στο Twitter πάνω στο οποίο δουλέψαμε.

Στο Κεφάλαιο 3 περιγράφουμε την ολοκληρωμένη υλοποίηση του συστήματος μας που έχει ως στόχο την παροχή συγκεκριμένων πληροφοριών για τη δραστηριότητα των χρηστών στο κοινωνικό δίκτυο Twitter.

Στο Κεφάλαιο 4 αναλύουμε τα αποτελέσματα της Πειραματικής Αξιολόγησης του συστήματος μας παραθέτοντας όλες τις λεπτομέρειες της διεξαγωγής των πειραμάτων και τα αποτελέσματα αυτών.

Στο Κεφάλαιο 5 αναφερόμαστε στη Σχετική Δουλειά που έχουμε εντοπίσει και τη θεωρούμε αξιοσημείωτη στα πλαίσια της παρούσας Διπλωματικής Εργασίας.

Στο Κεφάλαιο 6 παραθέτουμε τα Συμπεράσματα στα οποία μας οδήγησε η υλοποίηση του συστήματος μας.

2. Θεωρητικό Υπόβαθρο

Το κεφάλαιο αυτό αποτελεί την ανασκόπηση βιβλιογραφίας της παρούσας Διπλωματικής Εργασίας. Παρακάτω αναλύονται κάποιοι από τους βασικούς όρους που μας απασχόλησαν και στους οποίους βασίζεται η εν λόγω Διπλωματική Εργασία. Συγκεκριμένα, αναφερόμαστε γενικά στα Κοινωνικά Δίκτυα αλλά και ειδικότερα στο κοινωνικό δίκτυο πάνω στο οποίο εργαστήκαμε, δηλαδή το Twitter. Στη συνέχεια, αναφερόμαστε στον τομέα της Εξόρυξης Δεδομένων, στην Ανάλυση Συναισθήματος και στη διαδικασία του stemming, δηλαδή της αφαίρεσης καταλήξεων από λέξεις με σκοπό την εύρεση των ριζών τους. Πρόκειται για ενδιαφέροντες και συνεχώς εξελίξιμους τομείς που αποτελούν κομμάτι των τελευταίων τεχνολογιών στον τομέα της Ανάλυσης Δεδομένων.

2.1 Κοινωνικά Δίκτυα (Social Media)

Τα τελευταία χρόνια έκαναν την εμφάνιση τους στο Internet τα Κοινωνικά Δίκτυα (Social Media) φέρνοντας μαζί τους σημαντικές αλλαγές τόσο στον τρόπο χρήσης του Internet, όσο και στον τρόπο επικοινωνίας μεταξύ των ατόμων. Κάποια από τα πιο δημοφιλή κοινωνικά δίκτυα είναι το Facebook, το Twitter και το Instagram. Η παράλληλη ανάπτυξη του τομέα των έξυπνων τηλεφώνων (smartphones), οδήγησε στην αναπόσπαστη ένταξη των μέσων κοινωνικής δικτύωσης σε κάθε πτυχή της καθημερινότητας του ανθρώπου.

Οι δημοσιεύσεις του κάθε χρήστη στα κοινωνικά δίκτυα δεν περιέχουν μόνο αναρτήσεις απλών κειμένων που περιλαμβάνουν απόψεις και ιδέες, αλλά και πολυμέσων όπως φωτογραφίες και βίντεο. Η συνεχής αλληλεπίδραση του ατόμου με τα κοινωνικά δίκτυα δημιούργησε μια νέα δυνατότητα για ανάλυση και αξιοποίηση των δεδομένων που οι χρήστες μοιράζονται σε αυτά.

Η ανίχνευση σημαντικών θεμάτων στα κοινωνικά δίκτυα είναι ένας τομέας που γνωρίζει διαρκή ανάπτυξη τα τελευταία χρόνια. Δίνει τη δυνατότητα για παρακολούθηση διεθνών εξελίξεων, έκτακτων συμβάντων και άλλων περιστατικών, σε πραγματικό χρόνο και από πολλές και διαφορετικές πηγές.

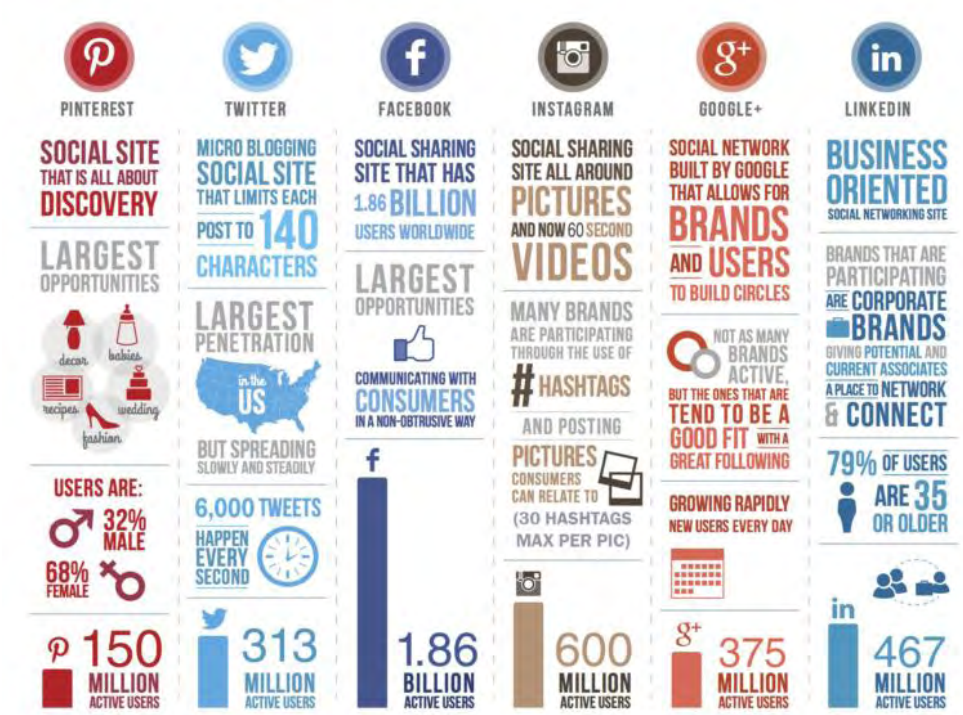

Figure 1 Στατιστικά κοινωνικών δικτύων 2017

2.2 Το κοινωνικό δίκτυο Twitter

2.2.1 Γενικά για το Twitter

Χαρακτηρίζεται ως το κοινωνικό μέσο που προωθεί περισσότερο από όλα τον δημόσιο διάλογο και την ανταλλαγή απόψεων. Πρόκειται για ένα από τα πιο δημοφιλή μέσα κοινωνικής δικτύωσης με σχεδόν τριακόσια εκατομμύρια χρήστες παγκοσμίως συνδεδεμένους μηνιαίως, μέσω του οποίου οι χρήστες επικοινωνούν με δημοσιεύσεις (tweets) με μέγιστο πλήθος λέξεων τις 140 λέξεις. Η ύπαρξη ορίου λέξεων οδηγεί τους χρήστες σε επιλογή στοχευμένων εκφράσεων και περιεχομένου των tweets τους. Ωστόσο, στη γλώσσα που χρησιμοποιούν οι χρήστες του Twitter συνήθως είναι η καθομιλουμένη και άρα ανεπίσημη γλώσσα. Το γεγονός αυτό, σε συνδυασμό μάλιστα και με το ότι στα tweets χρησιμοποιούνται συχνά συντομογραφίες και αργκό κάνει πιο σύνθετη την επεξεργασία του κειμένου που προέρχεται από tweets σε σχέση με άλλα κείμενα.

Λόγω της γρήγορης και άμεσης επικοινωνίας που χαρακτηρίζει όλα τα μέσα κοινωνικής δικτύωσης προκύπτουν κάποια θέματα ως προς την ορθότητα της γραμματικής και του συντακτικού. Ο «θόρυβος» αυτός που προκύπτει δημιουργεί πρόβλημα στο κομμάτι της αυτοματοποιημένης διαδικασίας αναγνώρισης συναισθήματος και η αφαίρεσή του αποτελεί ένα μεγάλο στοίχημα.

Στο κοινωνικό μέσο Twitter κάθε χρήστης έχει τη δυνατότητα να επιλέξει ο ίδιος τους χρήστες που θέλει να ακολουθήσει (follow). Πρόκειται για τους χρήστες των οποίων τα tweets θέλει να βλέπει απευθείας στο χρονολόγιο του (timeline). Ο κάθε χρήστης έχει τη δυνατότητα να αναδημοσιεύσει ένα tweet αυτούσιο και με αναφορά στον αρχικό του δημιουργό (retweet). Επιπλέον, μπορεί να μαρκάρει ένα tweet ως κάτι που του αρέσει (favorite). Τα παραπάνω αποτελούν χρήσιμες μετρικές αν προσπαθήσουμε να χαρακτηρίσουμε έναν χρήστη σχετικά με τη δραστηριότητα του στο κοινωνικό δίκτυο Twitter.

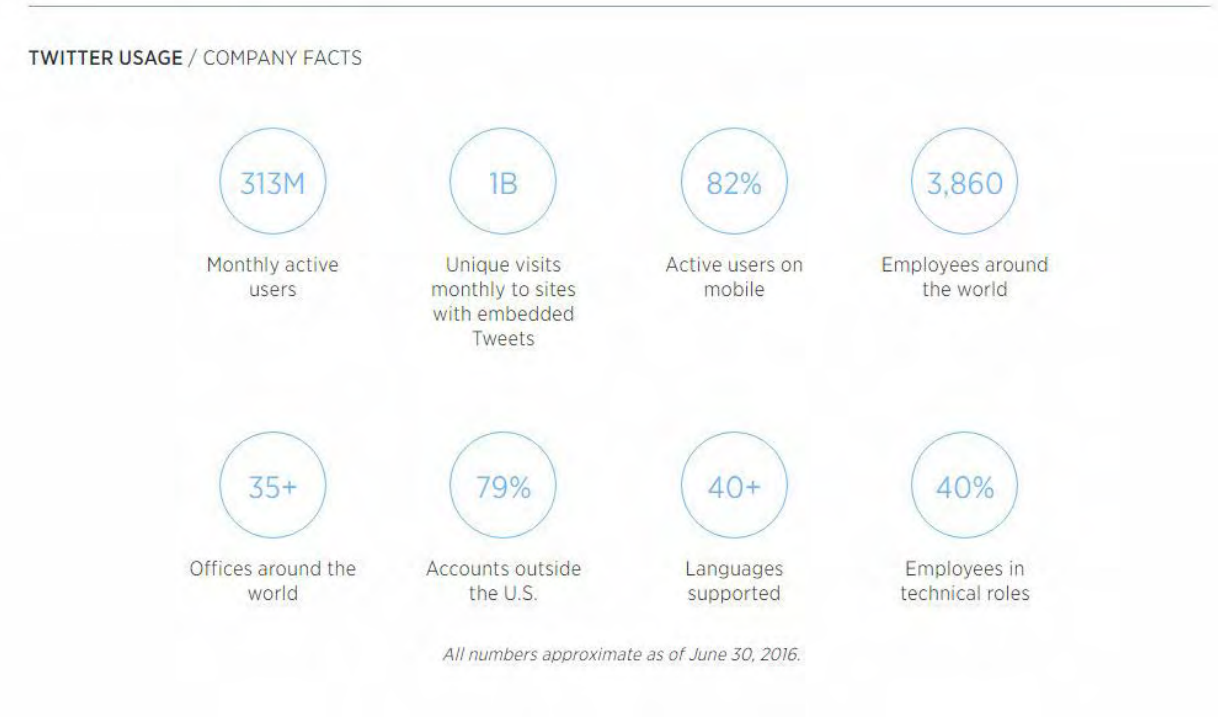

Figure 2 Στοιχεία σχετικά με το Twitter

2.2.2 Επικοινωνία με το Twitter

Σε αντίθεση με άλλα κοινωνικά δίκτυα, το Twitter παρέχει μία σειρά από δημόσιες διεπαφές (Application Programming Interface (API)) (2017). Οι διεπαφές αυτές, αποτελούν το δομικό στοιχείο πολλών άλλων εφαρμογών. Ανάλογα με τις ανάγκες που υπάρχουν σε κάθε εφαρμογή μπορεί να επιλεγεί και η αντίστοιχη διεπαφή του Twitter. Πιο συγκεκριμένα, το Twitter παρέχει 2 είδη διεπαφών το REST API και το Streaming API.

Η πρώτη κατηγορία διεπαφών, το REST API, παρέχει δυνατότητα τόσο για ανάγνωση όσο και για δημοσίευση μιας σειράς από tweets. Κάποιες από τις πιο σημαντικές μεθόδους αυτής της κατηγορίας αποτελούν η μέθοδος που σχετίζεται με το χρονολόγιο (timeline) μεμονωμένων χρηστών, η μέθοδος για την εύρεση των tweets που συνδέονται με ένα συγκεκριμένο μέρος και η μέθοδος που σχετίζεται με την αναζήτηση στο Twitter.

Η δεύτερη κατηγορία διεπαφών, το Streaming API, παρέχει πρόσβαση στην παγκόσμια ροή των δεδομένων του Twitter και ενδείκνυται αν ο στόχος μας είναι να λαμβάνουμε tweets σε πραγματικό χρόνο. Η συλλογή των tweets σε πραγματικό χρόνο μπορεί να γίνει είτε συλλέγοντας την παγκόσμια ροή δεδομένων (Public Stream), είτε συλλέγοντας τη ροή δεδομένων για έναν συγκεκριμένο χρήστη (User Stream) είτε τέλος για πολλούς αλλά προκαθορισμένους χρήστες (Site Stream).

2.2.3 Εξελίξεις στο Twitter

Το Twitter αποτελεί ένα διαρκώς εξελισσόμενο κοινωνικό μέσο. Μάλιστα σήμερα βρίσκεται σε μεταβατική περίοδο καθώς επέλεξε να κάνει κάποιες αλλαγές ώστε να εξασφαλίσει στους χρήστες του επιπλέον χαρακτήρες για τα tweets τους χωρίς όμως να χάσει τη σπουδαιότητα που προσδίδουν στο μέσο οι σύντομες δημοσιεύσεις. Άλλωστε, αυτό το χαρακτηριστικό του σε συνδυασμό βέβαια με τις σπουδαίες του διεπαφές, είναι που το καθιστούν το κοινωνικό μέσο που έχει μελετηθεί περισσότερο από οποιοδήποτε άλλο.

Συγκεκριμένα, άμεσα πρόκειται να σταματήσουν να συμπεριλαμβάνονται στους 140 χαρακτήρες οριοθέτησης των tweets, τα media που ο χρήστης θέλει να συμπεριλάβει και το πεδίο του @όνομα που εμφανίζεται αυτόματα στην αρχή κάθε tweet απάντησης. Αυτή η φαινομενικά αμελητέα τροποποίηση επιφέρει μεγάλες αλλαγές ακόμα και στις διεπαφές του Twitter.

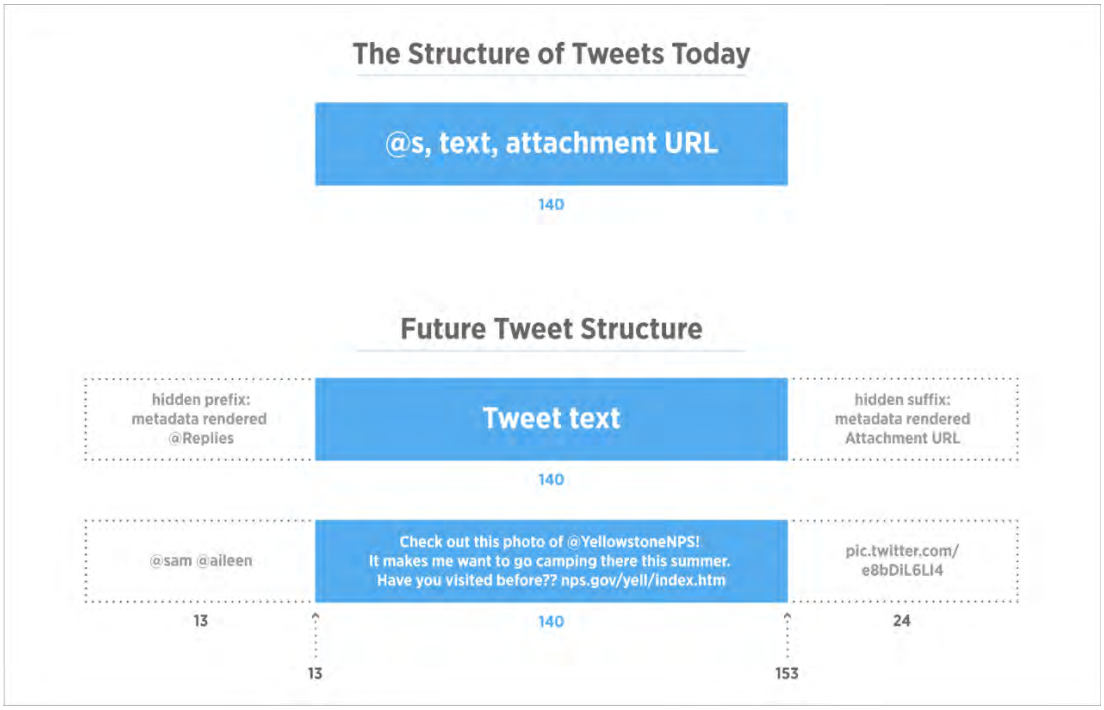

Figure 3 Σχεδιασμός των αλλαγών που ετοιμάζει το Twitter.

2.3 Εξόρυξη Δεδομένων (Data Mining)

Ο όρος αυτός αναφέρεται στη διαδικασία εξαγωγής κάποιων χρήσιμων και ενδιαφερόντων προτύπων ή πληροφοριών από μεγάλες ποσότητες δεδομένων με αυτοματοποιημένο τρόπο. Με τον όρο πρότυπο εννοούμε μία έκφραση σε κάποια γλώσσα, η οποία περιγράφει ένα υποσύνολο δεδομένων. Το παραπάνω επιτυγχάνεται χρησιμοποιώντας αλγορίθμους συσταδοποίησης (clustering), κατηγοριοποίησης (classification) αλλά και αλγορίθμους στατιστικής. Στο βιβλίο τους οι Leskovec et al. (2014) αναλύουν πολλές από τις πτυχές του θέματος της Εξόρυξης Δεδομένων. Ο τελικός στόχος της Εξόρυξης Δεδομένων έχει να κάνει με την απόκτηση πληροφορίας ή προτύπου που θα συμβάλει στην διαδικασία απόφασης για κάποιο θέμα από μεγάλες βάσεις δεδομένων.

Ο συγκεκριμένος τομέας αποτελεί στη σημερινή εποχή αναπόσπαστο κομμάτι πολλών άλλων επιστημονικών περιοχών. Συγκεκριμένα, η εξόρυξη δεδομένων εφαρμόζεται στο marketing με σκοπό την κατηγοριοποίηση των πελατών και την πρόβλεψη της συμπεριφοράς τους, στον τομέα των επενδύσεων με στόχο την επίτευξη προσοδοφόρων επενδύσεων και ευρέως στο Διαδίκτυο για βελτίωση της εμπειρίας των πολιτών. Η εξόρυξη δεδομένων στο Διαδίκτυο είναι απαραίτητη καθώς η πληροφορία που υπάρχει σε αυτό είναι αδύνατο ακόμα και να μετρηθεί.

Ένα πολύ ενδιαφέρον πεδίο αποτελεί η εξόρυξη δεδομένων από κάποιο κοινωνικό δίκτυο σε πραγματικό χρόνο. Αυτή η πλευρά της θα μας απασχολήσει στην παρούσα διπλωματική εργασία. Η σπουδαιότητα της έγκειται στο γεγονός ότι δεν υπάρχει αποδοτικός τρόπος για την απόκτηση ουσιαστικής πληροφορίας από τα κοινωνικά δίκτυα χωρίς την ύπαρξη εξόρυξης δεδομένων και αυτό λόγω των εκατομμύρια εγγεγραμμένων χρηστών. Ταυτόχρονα, λόγω της υψηλής δυναμικότητας που παρουσιάζουν τα δεδομένα των κοινωνικών δικτύων για την εξαγωγή συμπεράσματος σχετικά με οποιοδήποτε θέμα είναι απαραίτητη η εφαρμογή τεχνικών Εξόρυξης Δεδομένων. Στη σημερινή εποχή της συστηματικής καθημερινής χρήσης του Internet, η Εξόρυξη Δεδομένων από κάποιο κοινωνικό δίκτυο δεν είναι μόνο ένας αρωγός για τις θετικές επιστήμες ή ένα ενδιαφέρον αντικείμενο μελέτης σε ακαδημαϊκό επίπεδο. Τα αποτελέσματα της αποτελούν

αντικείμενο μελέτης πολλών επιστημονικών κλάδων και ιδιαίτερα των ανθρωπιστικών επιστημών.

Η κοινωνιολογία και η ψυχολογία αποτελούν παράδειγμα τέτοιων επιστημών αφού χρησιμοποιώντας τα αποτελέσματα της Εξόρυξης Δεδομένων σε κοινωνικά δίκτυα έρχονται αντιμέτωποι με μια τεράστια ευκαιρία να βγάλουν συμπεράσματα για την ανθρώπινη συμπεριφορά ανακαλύπτοντας σπουδαίες πλευρές της κοινωνικής συμπεριφοράς και αλληλεπίδρασης.

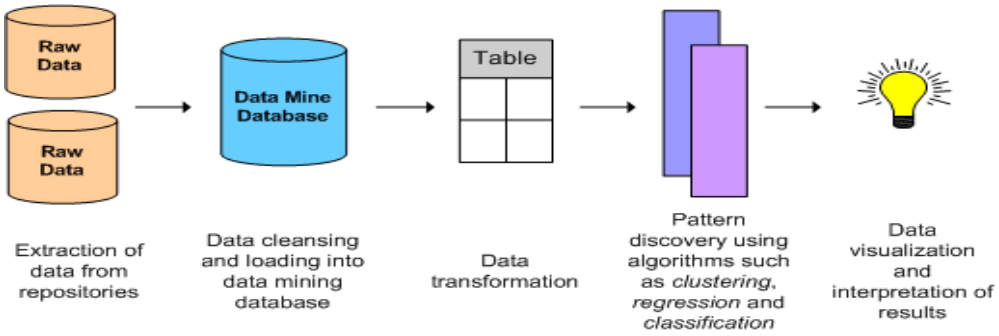

Figure 4 Διαδικασία Εξόρυξης Δεδομένων.

2.4 Ανάλυση Συναισθήματος (Sentiment Analysis)

Η τεχνική ανάλυσης του περιεχομένου ενός κειμένου ως προς το συναίσθημα είναι γνωστή ως Ανάλυση Συναισθήματος (Sentiment Analysis). Πρόκειται για μία διαδικασία άρρηκτα συνδεδεμένη με τον τομέα της Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing (NLP)). Ο εν λόγω τομέας ανήκει στα πεδία της Γλωσσολογίας και της Επιστήμης των Υπολογιστών και ασχολείται με την αλληλεπίδραση μεταξύ της φυσικής γλώσσας και της γλώσσας των υπολογιστών.

Ο Paul Ekman (2003), ψυχολόγος και πρωτοπόρος της έρευνας σχετικά με τα συναισθήματα ήταν αυτός που υποστήριξε την ύπαρξη έξι βασικών συναισθημάτων τα οποία είναι κοινά σε όλους τους ανθρώπους, όλων των πολιτισμών. Τα έξι αυτά

συναισθήματα είναι ο θυμός, η αηδία, ο φόβος, η χαρά, η λύπη και η έκπληξη. Αυτή η θεωρία σε συνδυασμό με τη χρήση πολικότητας (polarity) αποτελεί τη βάση για τη μελέτη των συναισθημάτων στην παρούσα Διπλωματική Εργασία.

Οι ερευνητές του συγκεκριμένου θέματος έχουν προτείνει δύο διαφορετικές προσεγγίσεις για το θέμα της Ανάλυση Συναισθήματος. Η πρώτη προσέγγιση είναι αυτή της Μηχανικής Μάθησης για την οποία έχουν αναπτυχθεί δύο τεχνικές. Πρόκειται για τη Μηχανική Μάθηση με Επίβλεψη (Supervised machine learning methods) και τη Μηχανική Μάθηση χωρίς Επίβλεψη (Unsupervised machine learning methods). Η δεύτερη προσέγγιση είναι αυτή της ανάλυσης συναισθήματος με χρήση λεξικού (lexicon - based).Οι αλγόριθμοι της Μηχανικής Μάθησης αναλύονται στο βιβλίο των Theodoridis και Koutroumbas (2012). Δεν είναι λίγες οι φορές που αυτές οι τεχνικές εφαρμόζονται και συνδυαστικά μεταξύ τους όπως έδειξαν και οι Giatsogloua et al. (2016).

2.4.1 Τεχνικές με Επιβλεπόμενη Μηχανική Μάθηση

Αρχικά, θα αναλύσουμε τη Μηχανική Μάθηση με Επίβλεψη. Σκοπός της είναι η κατηγοριοποίηση (classification) των δεδομένων σε κάποιες προκαθορισμένες κλάσεις με βάση κάποια δεδομένα εκπαίδευσης τα οποία είναι εκ των προτέρων γνωστά. Ένας συχνός διαχωρισμός που γίνεται είναι με βάση την πολικότητα των δεδομένων σε θετικά, αρνητικά ή ουδέτερα.

Κάποιοι από τους πιο διαδεδομένους αλγορίθμους επιβλεπόμενης μηχανικής μάθησης είναι ο Απλοϊκός Ταξινομητής κατά Bayes (Naive Bayes Classification), ο Ταξινομητής Μέγιστης Εντροπίας (maximum entropy classification) και οι Μηχανές Διανυσματικής Στήριξης (Support Vector Machines (SVM)). Οι παραπάνω αλγόριθμοι έχουν χρησιμοποιηθεί σε πληθώρα ερευνών πάνω στο θέμα της ομαδοποίησης δεδομένων με βάση την Ανάλυση Συναισθήματος. Μία πρωτοπόρα έρευνα στον τομέα αυτό διεξάχθηκε από τους Pang et al (2002).

2.4.2 Τεχνικές με Μη Επιβλεπόμενη Μηχανική Μάθηση

Όσον αφορά την Ανάλυση Συναισθήματος με Μη Επιβλεπόμενη Μηχανική Μάθηση, πρόκειται για την περίπτωση όπου τα δεδομένα δεν είναι ολοκληρωτικά γνωστά εκ των προτέρων. Έχουμε ένα σύνολο από δεδομένα για τα οποία προσπαθούμε να

αναδείξουμε τις ομοιότητες τους και να τα ομαδοποιήσουμε (clustering) χωρίς προηγούμενη εκπαίδευση. Η ανάδειξη των ομοιοτήτων των δεδομένων είναι ένα σοβαρό ζήτημα που προκύπτει στις Τεχνικές Μη Επιβλεπόμενης Μηχανικής Μάθησης, καθώς υπάρχουν πολλά διαφορετικά αλγοριθμικά σχήματα που μπορούν να ακολουθηθούν. Συνεπώς, ο κάθε ερευνητής πρέπει να είναι σε θέση να εξάγει κάθε φορά τα αντίστοιχα συμπεράσματα.

Τέτοια προβλήματα εμφανίζονται σε πολλές εφαρμογές τόσο των κοινωνικών επιστημών όσο και των τεχνολογικών επιστημών. Χαρακτηριστικά παραδείγματα αποτελούν η κατάτμηση εικόνας (image segmentation) και η κωδικοποίηση εικόνας και ήχου.

2.4.3 Τεχνικές Ανάλυσης Συναισθήματος με Χρήση Συναισθηματικού Λεξικού

Πρόκειται για μία τεχνική που προϋποθέτει την ύπαρξη λεξικού συναισθήματος που αναθέτει στις λέξεις του κάποιο συναίσθημα. Αυτό το συναίσθημα υποδηλώνεται με τη χρήση θετικών αριθμών για τις λέξεις θετικής πολικότητας, αρνητικών αριθμών για τις λέξεις αρνητικής πολικότητας και του μηδενός για τις λέξεις ουδέτερης πολικότητας. Ωστόσο, η πληροφορία σχετικά με το συναίσθημα μίας λέξης δεν σταματάει εκεί. Είναι χρήσιμο για την ολοκληρωμένη ανάλυση συναισθήματος να περιλαμβάνεται σε ένα συναισθηματικό λεξικό πληροφορία που να υποδηλώνει το είδος του συναισθήματος, λαμβάνοντας για παράδειγμα υπόψη τη θεωρία του Paul Ekman (2003) που αναφέρεται παραπάνω. Η πληροφορία για το είδος του συναισθήματος υποδηλώνεται με μία αριθμητική διαβάθμιση από το 0 ως το 5 που εκφράζει την ένταση του συναισθήματος σε μία λέξη (θυμός, αηδία, φόβος, χαρά, λύπη και έκπληξη).

Η τεχνική αυτή βασίζεται στην προσέγγιση ότι ο προσδιορισμός του συναισθήματος ενός κειμένου προκύπτει από το την πολικότητα κάθε λέξης που αυτό περιέχει. Αντίστοιχα και για τον εντοπισμό του συγκεκριμένου συναισθήματος που εκφράζεται σε ένα κείμενο. Σε πολλές μελέτες παρουσιάζεται ένας συνδυασμός της χρήσης λεξικού και της μη επιβλεπόμενης μηχανικής μάθησης.

Ειδικά για τα κοινωνικά δίκτυα, η χρήση λεξικών για την ανάλυση συναισθήματος παρουσιάζει κάποιες ιδιαιτερότητες όπως ανέδειξαν οι Ngoc και Yoo (2014). Μία από αυτές είναι το γεγονός ότι, όπως αναφέρθηκε και παραπάνω, η γλώσσα που

χρησιμοποιείται σε αυτά τα μέσα δεν είναι πάντοτε επίσημη και συνεπώς ένα συναισθηματικό λεξικό θα πρέπει να περιέχει λέξεις και φράσεις της καθομιλουμένης, αργκό ακόμα και εσκεμμένα ορθογραφικά λάθη ώστε να καλύψει όσο το δυνατό περισσότερες λέξεις. Μία ακόμη ιδιαιτερότητα που παρατηρείται είναι το γεγονός ότι στα κοινωνικά δίκτυα συχνά τα κείμενα των χρηστών (δηλαδή στην περίπτωση του Twitter τα tweets), αποτελούνται από παραπάνω από μία γλώσσες. Αυτό δυσκολεύει τη χρήση λεξικών.

Ωστόσο, αν στοχεύουμε στην συναισθηματική ανάλυση κειμένων μίας συγκεκριμένης γλώσσας το παραπάνω δεν επηρεάζει τα αποτελέσματα μας. Σε αντίθετη περίπτωση, έχει επιχειρηθεί ακόμα και μετάφραση λέξεων άλλης γλώσσας από αυτή του λεξικού πριν την συναισθηματική ανάλυση ενός κειμένου.

Ένα από τα θετικά που παρουσιάζουν οι τεχνικές με λεξικά είναι η απλότητα τους και άρα η δυνατότητα άμεσης εφαρμογής τους. Δεν απαιτούν προηγούμενη εκπαίδευση για κάθε ξεχωριστή εφαρμογή και άρα είναι ευέλικτες από την άποψη ότι μπορούν να χρησιμοποιηθούν αυτούσιες σε διαφορετικές εφαρμογές. Όσον αφορά την εφαρμογή τους σε κοινωνικά δίκτυα όπως το Twitter, έρευνες δείχνουν πως αν ένα λεξικό περιλαμβάνει αργκό και λέξεις του προφορικού λόγου, τα αποτελέσματα που θα δώσει είναι πολύ ικανοποιητικά.

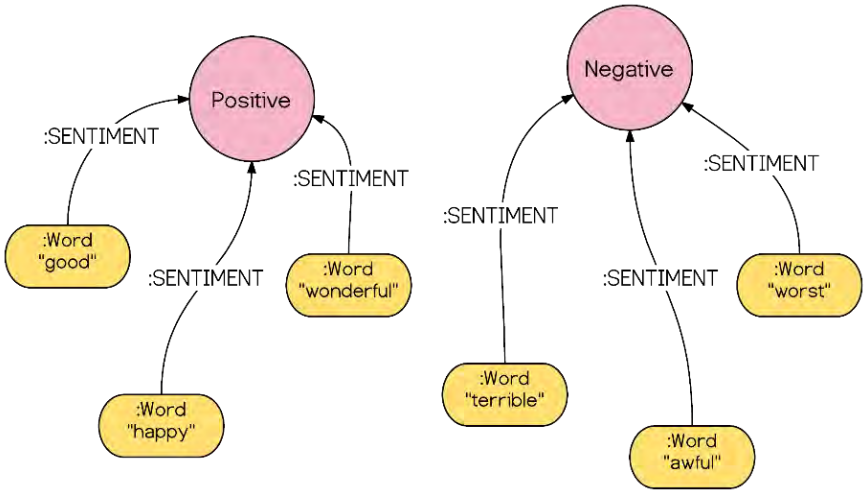

Figure 5 Αναπαράσταση πολικότητας Λέξεων

Από τα πιο γνωστά λεξικά που χρησιμοποιούνται ευρέως για Συναισθηματική Ανάλυση της Αγγλικής γλώσσας είναι το Bing Liu's Opinion Lexicon (2012). Πρόκειται για ένα λεξικό το οποίο περιέχει 2006 θετικές και 4783 αρνητικές λέξεις. Περιέχει επίσης ορθογραφικά λάθη, αργκό και είναι ιδανικό για χρήση στα social-media όπως το Twitter. Άλλο ένα δημοφιλές λεξικό είναι το SenticNet το οποίο μελετάτε εκτενώς από τους Esuli και Sebastiani (2006). Το λεξικό αυτό παρέχει ένα σύνολο από τριάντα χιλιάδες έννοιες της αγγλικής γλώσσας καθώς και πληροφορία σχετικά με τις πέντε έννοιες που είναι σημασιολογικά κοντά με την έννοια της λέξης εισόδου και την πολικότητα της λέξης εισόδου.

2.4.4 Εργαλεία Ανάλυσης Συναισθήματος

Λόγω της πολύπλευρης σπουδαιότητας του τομέα της Συναισθηματικής Ανάλυσης, αναπτύσσονται διαρκώς εργαλεία, βιβλιοθήκες αλλά και διεπαφές (Application programming interface (API)) που συμβάλουν στην αυτοματοποίηση της εν λόγω διαδικασίας.

Figure 6 Αναπαράσταση Ανάλυσης Συναισθήματος

Το Natural Language Toolkit (NLTK) (2017) είναι μια πλατφόρμα της γλώσσας Python η οποία διευκολύνει ουσιαστικά την διαχείριση της Φυσικής Γλώσσας (Natural Language Processing (NLP)). Πρόκειται για ένα πρόγραμμα ελεύθερου λογισμικού με τεράστια συμβολή στον τομέα της επεξεργασίας φυσικής γλώσσας μέσω της γλώσσας Python και με εφαρμογή τόσο στην εκπαίδευση όσο και στη βιομηχανία. Αναφέρεται σε πολλές φυσικές γλώσσες και όχι μόνο στην Αγγλική όπως τα περισσότερα εργαλεία αυτού του είδους. Περιλαμβάνει μία διεπαφή (Application programming interface (API)) μέσω της οποίας υλοποιούνται κάποιες σημαντικές εργασίες σχετικά με την επεξεργασία φυσικής γλώσσας. Μερικές από αυτές είναι ο χωρισμός προτάσεων σε λέξεις (tokenization), η κατηγοριοποίηση (classification) και η ανάλυση συναισθήματος.

Ένα άλλο εργαλείο στον τομέα της αυτοματοποίησης της ανάλυσης συναισθήματος είναι το Stanford CoreNLP (2017) το οποίο είναι υλοποιημένο σε Java και παρέχει λειτουργικότητα κυρίως για την Αγγλική γλώσσα, αλλά υποστηρίζει σε περιορισμένο επίπεδο και κάποιες ακόμη γλώσσες όπως τα Ισπανικά.

2.5 Διαδικασία Αφαίρεσης Καταλήξεων (Stemming)

Λόγω της εκτεταμένης χρήσης τόσο του διαδικτύου όσο και των κοινωνικών δικτύων προκύπτει μία ακόμη ανάγκη στον τομέα της ανάλυσης των δεδομένων των κοινωνικών δικτύων. Πρόκειται για την ανάγκη για αναζήτηση της ρίζας (Stemming) κάθε λέξης και άρα της αφαίρεσης της κατάληξης της. Εφαρμόζεται έτσι ώστε να βελτιωθεί η διαδικασία αυτοματοποιημένης συναισθηματικής ανάλυσης και να αντιμετωπίζεται μία λέξη με τον ίδιο τρόπο ακόμα και αν της έχουμε προσδώσει κάποια κατάληξη, για παράδειγμα υποκοριστικού.

Ωστόσο, η παραπάνω διαδικασία δεν είναι καθόλου απλή, ειδικά αν μιλάμε για γλώσσες με μεγάλη μορφολογική ποικιλία και πολλές κλίσεις όπως είναι τα Ελληνικά. Οι περισσότερες υλοποιήσεις που υπάρχουν σήμερα εφαρμόζονται στην Αγγλική γλώσσα. Μία από τις ευρύτερα διαδεδομένες υλοποιήσεις αφαίρεσης καταλήξεων είναι ο Porter Stemming Αλγόριθμος ή Porter Stemmer (1980). Πρόκειται για μία διαδικασία αφαίρεσης των πιο συχνών μορφολογικών καταλήξεων από λέξεις της Αγγλικής γλώσσας, η οποία εφαρμόζεται κατά κύριο λόγο σε συστήματα ανάκτησης πληροφορίας Information Retrieval systems (IRS). Ο αλγόριθμος του Porter Stemmer αφού χρησιμοποιήθηκε κατά

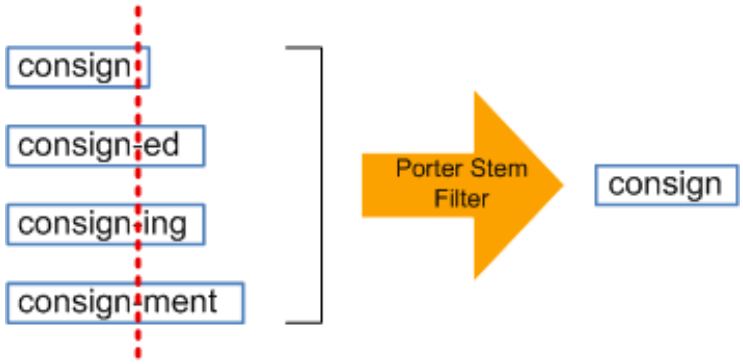

Figure 7 Αναπαράσταση του αλγορίθμου αφαίρεσης καταλήξεων του Portet Stemmer

κόρων τα τελευταία 20 χρόνια, σήμερα υπάρχει υλοποιημένος επίσημα σε τουλάχιστον δεκαπέντε γλώσσες προγραμματισμού.

Όσον αφορά την υλοποίηση αλγορίθμου αφαίρεσης καταλήξεων (stemmer) για την Ελληνική γλώσσα, όσοι ερευνητές το επιχείρησαν ήρθαν αντιμέτωποι με το πρόβλημα που δημιουργεί ο πλούτος της Ελληνικής γλώσσας, οι πολλές κλίσεις της και οι λέξεις ποικίλης σημασίας. Η πρώτη προσπάθεια που έγινε για υλοποίηση ενός τέτοιου αλγορίθμου ήταν ο Αλγόριθμος TZK του Καλαμπούκη (1995), ενώ ακολούθησε ο Αλγόριθμος AMP των Tambouratzis και Carayannis (2001) και ο αλγόριθμος του Νταή (2006). Πάνω στον αλγόριθμο του Νταή βασίστηκε, ο πιο πρόσφατος αλγόριθμος για αφαίρεση καταλήξεων στα Ελληνικά. Πρόκειται για τον αλγόριθμο του Αλέξανδρου Καλοπόδη (2013) ο οποίος χρησιμοποιείται κατά κόρων και φαίνεται να έχει ικανοποιητικά αποτελέσματα.

3. Υλοποίηση Συστήματος

Σε αυτό το κεφάλαιο γίνεται ανάλυση όλης της διαδικασίας ανάπτυξης του συστήματός μας που στόχο έχει την παροχή συγκεκριμένων υπηρεσιών και πληροφοριών βασιζόμενο στο κοινωνικό δίκτυο Twitter. Παρουσιάζεται ο γενικός σχεδιασμός του συστήματος και δίνονται πληροφορίες αναφορικά με όλη τη διαδικασία ανάπτυξης και σχεδίασης του. Στόχος αυτού του κεφαλαίου είναι τόσο η παρουσίαση της δουλειάς που έγινε στη συγκεκριμένη Διπλωματική Εργασία όσο και η ενημέρωση σχετικά με τις τεράστιες δυνατότητες που παρουσιάζονται από προγραμματιστικής σκοπιάς όταν μία εφαρμογή σαν το Twitter παρέχει σημαντικές διεπαφές.

Οι βασικές υπηρεσίες που παρέχει το σύστημα μας είναι οι παρακάτω:

- Ανάλυση των tweets που δημοσιεύθηκαν σε ένα συγκεκριμένο χρονικό διάστημα κατά το οποίο αντλούμε tweets, ώστε να εντοπιστούν τα δέκα πιο σημαντικά θέματα συζήτησης σε πραγματικό χρόνο και να εφαρμόσουμε ανάλυση συναισθήματος σε αυτά.
- Για κάθε ένα από τα σημαντικά θέματα που εντοπίστηκαν παραπάνω, κάνουμε κατηγοριοποίηση των χρηστών ανάλογα με τη γνώμη τους για το θέμα στο οποίο συμμετέχουν και με το πόσο έντονη είναι η ενασχόληση τους με αυτό, με σκοπό την στοχευμένη πρόταση φιλίας μεταξύ των χρηστών.

Επιλέχθηκε η υλοποίηση των συγκεκριμένων υπηρεσιών επειδή αποτελούν ενδιαφέροντα θέματα μελέτης ειδικά για δεδομένα στην Ελληνική γλώσσα που παρουσιάζει πολλές ιδιαιτερότητες. Επιπλέον, μας δίνουν τη δυνατότητα ενασχόλησης και τριβής με ένα μεγάλο μέρος των δυνατοτήτων που προσφέρουν οι διεπαφές του Twitter.

3.1 Αρχιτεκτονική Συστήματος

Το πρώτο στάδιο της ανάπτυξης του συστήματος μας είναι η συλλογή των δεδομένων που θέλουμε να αναλύσουμε. Πρόκειται για δεδομένα που αντλούνται απευθείας από το Twitter. Το δεύτερο στάδιο καταλαμβάνει η προεπεξεργασία στην οποία πρέπει να υποβληθούν τα δεδομένα μας ώστε να είναι σε κατάλληλη μορφή για την ανάλυση που επιθυμούμε να κάνουμε στην συνέχεια. Το στάδιο της ανάλυσης είναι και το επόμενο στάδιο στη διαδικασία ανάπτυξης του συστήματός μας. Αποτελείται από πολλές επιμέρους εργασίες στις οποίες θα αναφερθούμε με περισσότερες λεπτομέρειες στη συνέχεια.

Η γλώσσα που επιλέχθηκε για την υλοποίηση του συστήματος μας είναι η Python και ο λόγος αυτής της επιλογής είναι οι πολλές δυνατότητες που προσφέρει στον τομέα της Ανάλυσης Δεδομένων (Data Analysis).

Figure 8 Τα πρώτα βήματα του συστήματός μας.

Σε αυτό το στάδιο δίνουμε μόνο τη γενική ιδέα της αρχιτεκτονικής μας, επειδή η κάθε μία από τις δύο υπηρεσίες που υλοποιήσαμε ακολουθεί μία δική της ξεχωριστή επιμέρους πορεία. Μάλιστα οι διαφορές τους ξεκινάνε ακόμα και από το πρώτο στάδιο της υλοποίησης καθώς για τη μία υπηρεσία χρησιμοποιούμε το Streaming API του Twitter και για την άλλη χρησιμοποιούμε το REST API του Twitter.

3.2 Εντοπισμός Σημαντικών Θεμάτων και Εφαρμογή σε αυτά Ανάλυσης Συναισθήματος

Πρόκειται για τη διαδικασία ανάλυσης των tweets που δημοσιεύθηκαν σε ένα συγκεκριμένο χρονικό διάστημα κατά το οποίο συλλέγουμε tweets. Ο στόχος μας είναι να εντοπιστούν τα δέκα πιο σημαντικά θέματα συζήτησης στο κοινωνικό μέσο Twitter σε πραγματικό χρόνο και στη συνέχεια να εφαρμόσουμε ανάλυση συναισθήματος σε αυτά, ώστε να ανακαλύψουμε τη γνώμη της πλειοψηφίας των χρηστών που σχολιάζουν σχετικά με τα εν λόγω δέκα θέματα.

Για την υλοποίηση της παραπάνω υπηρεσίας απαιτείται αρχικά άντληση tweets σε πραγματικό χρόνο, και στη συνέχεια επεξεργασία των δεδομένων που συλλέξαμε. Το κομμάτι της επεξεργασίας των δεδομένων μας, χωρίζεται σε δύο επιμέρους στάδια. Πρόκειται για το στάδιο φιλτραρίσματος των δεδομένων (cleaning) και το στάδιο της αφαίρεσης καταλήξεων (stemming) των λέξεων που αποτελούν τα δεδομένα μας.

Για πρώτα αυτά βήματα αλλά και για το κομμάτι της ανάλυσης των δεδομένων μας, παρουσιάζονται παρακάτω αναλυτικά όλες οι λεπτομέρειες της υλοποίησης μας.

3.2.1 Επικοινωνία με το Twitter και Συλλογή Δεδομένων

Όσον αφορά το πρώτο κομμάτι της παρούσας Διπλωματικής Εργασίας, για την επικοινωνία μας με το Twitter χρειάστηκε να χρησιμοποιήσουμε το Streaming API που προσφέρει το Twitter. Στην παράγραφο 2.2 υπάρχει αναλυτική αναφορά στις διεπαφές του κοινωνικού μέσου Twitter καθώς και στις διαφορές τους. Για τη συγκεκριμένη υπηρεσία χρησιμοποιήσαμε το Streaming API διότι θέλαμε να αντλήσουμε δεδομένα (tweets) σε πραγματικό χρόνο.

Επιπλέον, για την απόκτηση των απαραίτητων στοιχείων διασύνδεσης με το δίκτυο δημιουργήσαμε μία εφαρμογή μέσω της πλατφόρμας εφαρμογών του Twitter που απευθύνεται σε προγραμματιστές (Twitter Developers).

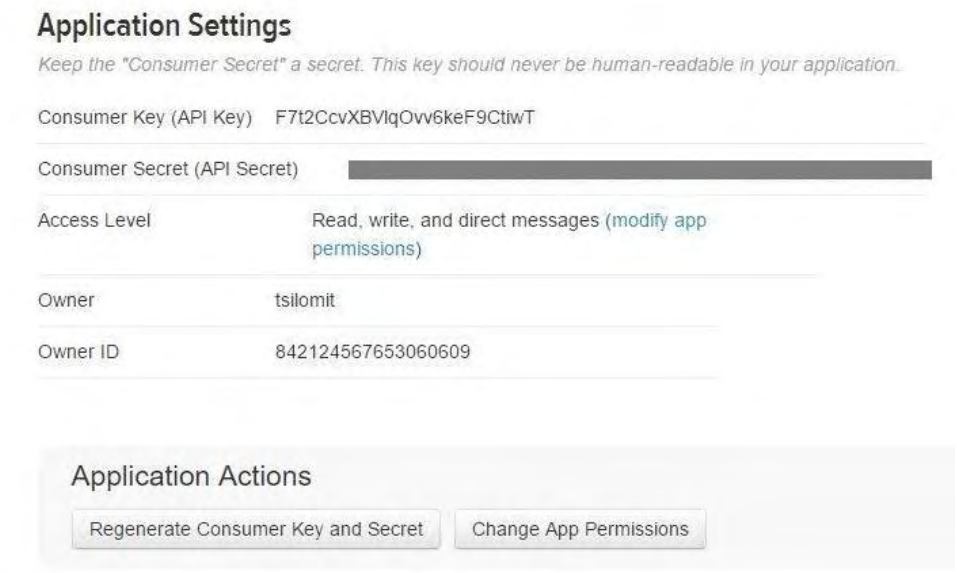

Figure 9 Διαδικασία απόκτησης αναγνωριστικών για σύνδεση με τις διεπαφές του Twitter.

Αποκτήσαμε τα απαραίτητα κλειδιά που αποτελούν τα προσωπικά μας στοιχεία πρόσβασης και χωρίς την επικύρωση των οποίων δεν είναι δυνατή η πρόσβασή μας στις διεπαφές του Twitter. Συγκεκριμένα, τα κλειδιά που χρειαζόμαστε για να συνδεθούμε με τις διεπαφές του Twitter είναι τα Consumer Key, Consumer Secret, Access Token και Access Token Secret.

Μία σημαντική βιβλιοθήκη της Python είναι η Tweepy. Πρόκειται για μία βιβλιοθήκη η οποία διευκολύνει πολύ τη χρήση του Twitter Streaming API παρέχοντας μεθόδους που χειρίζονται όλη τη διαδικασία της σύνδεσης.

Για τους σκοπούς αυτής της Διπλωματικής Εργασίας, χρησιμοποιήθηκε η μέθοδος filter της Tweepy. Η μέθοδος filter επιστρέφει tweets σε μορφή json και παρέχει δυνατότητα υψηλής παραμετροποίησης των δεδομένων που συλλέγονται. Στη συγκεκριμένη εργασία, παραμετροποιήσαμε τη συλλογή δεδομένων ως προς τη γλώσσα των tweets που επιθυμούμε να συλλεχθούν, δηλαδή τα Ελληνικά. Τέλος, η μέθοδος filter απαιτεί την ύπαρξη του πεδίου track το οποίο πρέπει να περιέχει τις λέξεις με βάση τις οποίες θα

επιλεγούν τα tweets. Επειδή ο στόχος ήταν να μαζευτούν γενικά tweets στην Ελληνική γλώσσα και όχι tweets πάνω σε κάποιο συγκεκριμένο θέμα, το πεδίο track στην παρούσα εργασία περιέχει μία σειρά από πολύ συχνές λέξεις, όπως άρθρα. Αυτές οι λέξεις φαίνονται στον παρακάτω πίνακα:

και	η	ένας	οι	τα
από	ή	μια	που	της
απ	ο	μία	πού	τι
για	το	ένα	σου	τις
δε	τον	θα	στις	τους
δεν	τη	ότι	στο	του
να	την	με	στον	των

Πίνακας 1 Λέξεις που αναζητούμε για συλλογή tweets.

Με την παραπάνω διαδικασία έχουμε εξασφαλίσει ότι θα αντλήσουμε δεδομένα από το κοινωνικό δίκτυο Twitter τα οποία θα είναι γραμμένα στα Ελληνικά. Έτσι, συλλέξαμε τα δεδομένα τα οποία θα μελετήσουμε για να εντοπίσουμε τα δέκα σημαντικότερα θέματα σε πραγματικό χρόνο και να εφαρμόσουμε σε αυτά συναισθηματική ανάλυση. Η υλοποίηση της λειτουργικότητας που περιγράψαμε μέχρι τώρα φαίνεται στο Παράρτημα Α.

3.2.2 Φιλτράρισμα Δεδομένων

Όπως περιμέναμε, το κείμενο στα διάφορα tweets που συλλέξαμε εκτός από λέξεις περιέχει σε μεγάλο βαθμό και υπερσυνδέσμους (URLs), εικονίδια συναισθήματος (emoticons) και αριθμούς. Επιπλέον, συχνά το κείμενο ενός tweet μπορεί να περιέχει το σύμβολο @ ακολουθούμενο από κάποιο όνομα ενός χρήστη. Αυτό εμφανίζεται όταν πρόκειται για ένα tweet το οποίο αποτελεί αναδημοσίευση (retweet) από κάποιον άλλον χρήστη. Για τις ανάγκες της παρούσας Διπλωματικής Εργασίας επιλέξαμε να αφαιρέσουμε τις εν λόγω ειδικές περιπτώσεις όπου αυτές εντοπίστηκαν. Αυτή η επιλογή έγινε αφενός επειδή αυτού του είδους τα δεδομένα δεν παρουσιάζουν κάποιου είδους χρησιμότητα στην συγκεκριμένη υλοποίηση και αφετέρου για να «καθαρίσουμε» τα δεδομένα μας από θόρυβο.

Στη συνέχεια, αφού χρησιμοποιήσαμε το εργαλείο Natural Language Toolkit (NLTK) το οποίο αναλύεται στο κεφάλαιο 2.4.4, χωρίσαμε το κείμενο κάθε tweet σε λέξεις (tokens) ώστε να διευκολύνουμε τη διαδικασία επεξεργασίας των δεδομένων μας από εδώ και πέρα.

Το επόμενο στάδιο του φιλτραρίσματος των δεδομένων μας αποτελεί η αφαίρεση λέξεων (stopwords), οι οποίες δεν θέλουμε να υπάρχουν στο σετ των δεδομένων μας ώστε να μην επηρεάζουν τον αλγόριθμό μας. Πρόκειται για λέξεις οι οποίες κρίναμε ότι δεν είναι κρίσιμης σημασίας στην παρούσα Διπλωματική Εργασία και που πειράματα μας έδειξαν ότι δημιουργούν περιττό «θόρυβο». Αναφερόμαστε κυρίως σε συνδέσμους παρόμοιους με αυτούς που χρησιμοποιήσαμε για να συλλέξουμε ελληνικά tweets γενικού σκοπού.

Και	Αυτή	Άλλος	Ένας	Από	Αυτοί	Εμείς
κι	Αυτός	Άλλη	Μία	Ως	Αυτές	Εσείς
θα	αυτό	άλλο	ένα	που	αυτά	αυτοί

Πίνακας 2 Παράδειγμα με κάποιες stopwords.

3.2.3 Εύρεση των Σημαντικότερων Θεμάτων

Η διαδικασία εύρεσης σημαντικών θεμάτων στα κοινωνικά δίκτυα είναι ίσως το πιο διαδεδομένο θέμα για το οποίο γίνεται ανάλυση από τους ερευνητές των κοινωνικών δικτύων. Αποτελεί έναν τομέα που συνεχώς εξελίσσεται και πάνω στον οποίο έχουν αναπτυχθεί πολλές προσεγγίσεις. Όσον αφορά το κοινωνικό μέσο Twitter το οποίο μελετάμε στην παρούσα Διπλωματική Εργασία, υπάρχει ένας πολύ απλός τρόπος να βγάλει κάποιος συμπεράσματα σχετικά με το ποια θέματα είναι αυτά που συζητούνται περισσότερο. Πρόκειται για την παρακολούθηση των λεγόμενων hashtags, δηλαδή λέξεων οι οποίες ξεκινάνε με το σύμβολο δίεση (#) και χρησιμοποιούνται από τους χρήστες για να υποδηλώσουν το θέμα της δημοσίευσης τους. Αυτή η λειτουργικότητα παρέχεται από το ίδιο το Twitter στο ευρύ κοινό. Αυτή η προσέγγιση είναι η πιο απλοϊκή που μπορεί να γίνει.

Για την ανάπτυξη της πρώτης υπηρεσίας που σχεδιάστηκε στα πλαίσια της εργασίας αυτής, σχεδιάστηκε ένας αλγόριθμος για τον εντοπισμό των σπουδαιότερων θεμάτων για τα οποία γίνεται συζήτηση στο Twitter. Στο σύστημά μας, αρχικά αναζητούμε τα δέκα (10) πιο διαδεδομένα hashtags στη χρονική διάρκεια για την οποία συλλέγουμε tweets. Ταυτόχρονα, εντοπίζουμε τα συγκεκριμένα tweets στα οποία χρησιμοποιούνται αυτά τα δέκα πιο διαδεδομένα hashtags. Σε αυτά τα tweets αναζητούμε για κάθε hashtag τις τρεις (3) πιο διαδεδομένες λέξεις που εμφανίζονται στα αντίστοιχα tweets. Στο τέλος αυτής της διαδικασίας, έχουμε τα δέκα πιο διαδεδομένα tweets και τρεις λέξεις για κάθε ένα από αυτά .Ο λόγος που μπήκαμε σε αυτή τη διαδικασία είναι για να κατηγοριοποιήσουμε τα tweets που δεν περιέχουν hashtags. Η ανάγκη αυτή προέκυψε από το γεγονός ότι παρατηρήσαμε από τα πειράματα μας πως τα περισσότερα tweets για τη συγκεκριμένη περίοδο που συλλέγαμε δεδομένα κάθε φορά, δεν περιείχαν ούτε ένα hashtag στο κείμενό τους. Συνεπώς, θα ήταν παράλειψη μας να μην ασχοληθούμε με αυτή την κατηγορία από tweets. Αφού λοιπόν εντοπίσαμε τα δέκα πιο διαδεδομένα tweets, για κάθε ένα από αυτά ψάχναμε τις τρεις πιο συχνές λέξεις που το συνοδεύουν στα tweets χωρίς κάποιο hashtag. Στο τέλος αυτής της διαδικασίας κατηγοριοποιούσαμε το κάθε tweet στη λίστα με εκείνα που αντιστοιχούν στο hashtag για το οποίο βρήκαμε τις περισσότερες κοινές λέξεις.

Στο παρακάτω σχήμα βλέπουμε ένα στιγμιότυπο του αλγορίθμου που υλοποιήσαμε. Στη συγκεκριμένη περίπτωση, βλέπουμε ότι δύο από τις τρεις πιο δημοφιλείς λέξεις που αντιστοιχίζονται στο πρώτο δημοφιλέστερο hashtag εντοπίζονται σε ένα συγκεκριμένο tweet. Στο ίδιο tweet εντοπίστηκε και μία από τις τρεις πιο δημοφιλείς λέξεις που αντιστοιχίζονται στο δέκατο δημοφιλέστερο hashtag. Άρα τελικά κατηγοριοποιούμε αυτό το tweet στη λίστα με τα tweets που περιέχουν το πρώτο πιο δημοφιλές hashtag. Αντίστοιχα, σε ένα άλλο tweet εντοπίζεται μία λέξη από τις τρεις πιο δημοφιλείς του δέκατου hashtag. Το tweet αυτή τη φορά κατηγοριοποιείται στη λίστα με τα tweets του δέκατου hashtag.

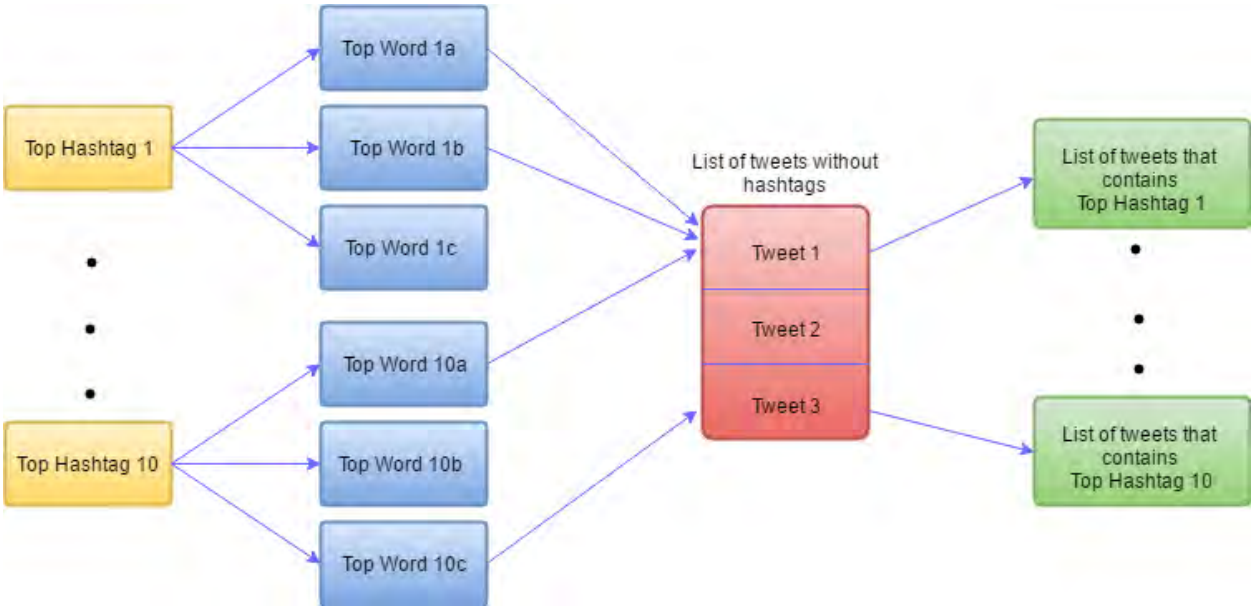

Figure 10 Παράδειγμα εντοπισμού των tweets που δεν περιέχουν hashtag και κατηγοριοποίηση.

Ανακεφαλαιώνοντας, τα βήματα που έχουμε ολοκληρώσει μέχρι τώρα μας οδήγησαν στην εύρεση των δέκα δημοφιλέστερων hashtags και στη δημιουργία λιστών η κάθε μία από τις οποίες περιέχει όλα τα tweets που αντιστοιχούν σε ένα hashtag. Όπου αναφερόμαστε σε tweets που αντιστοιχούν σε ένα hashtag, εννοούμε τόσο αυτά που περιείχαν το hashtag εξ' αρχής, όσο και αυτά που προσθέσαμε με τον αλγόριθμο που αναλύουμε παραπάνω.

3.2.4 Αφαίρεση καταλήξεων (stemming)

Επειδή ο τελικός μας στόχος είναι να εντοπίσουμε την άποψη της κοινής γνώμης όπως αυτή εκφράζεται στο κοινωνικό δίκτυο Twitter, θα πρέπει αφού εντοπίσαμε τα δέκα πιο σημαντικά θέματα να εφαρμόσουμε συναισθηματική ανάλυση στα Tweets που αναφέρονται σε καθένα από αυτά ξεχωριστά. Για τις ανάγκες αυτής της διαδικασίας, αφαιρούμε την κατάληξη κάθε λέξης του κάθε tweet που μας ενδιαφέρει.

Όπως αναφέραμε και στην παράγραφο 2.5, η διαδικασία αφαίρεσης καταλήξεων από τις λέξεις που αποτελούν τα tweets των δεδομένων που συλλέξαμε είναι μία απαραίτητη διαδικασία επεξεργασίας στην οποία πρέπει να υποβάλουμε τα δεδομένα μας. Αυτό συμβαίνει τόσο λόγω των πολλών κλίσεων που υπάρχουν στην ελληνική γλώσσα, όσο

και λόγω του ότι τα δεδομένα μας έχουν μεγάλο όγκο και άρα οποιαδήποτε ταξινόμηση θα μας βοηθήσει στην ανάλυση τους. Ωστόσο, εκτός από χρήσιμη αυτή η διαδικασία είναι και αρκετά απαιτητική για γλώσσες με πλούσιο λεξιλόγιο σαν την Ελληνική.

Στην παρούσα Διπλωματική Εργασία για την αφαίρεση των καταλήξεων των λέξεων των tweets, χρησιμοποιήθηκε ο αλγόριθμος του Αλέξανδρου Καλοπόδη. Σε πειράματα μας παρατηρήσαμε την ορθή λειτουργία του συγκεκριμένου αλγορίθμου. Μία σημαντική παρατήρηση ωστόσο, είναι ότι δουλεύει μόνο για λέξεις σε κεφαλαία Ελληνικά και όχι σε πεζά. Τέλος, για τη σωστή λειτουργία του συγκεκριμένου αλγορίθμου πρέπει να έχουμε προβλέψει να αφαιρέσουμε τους τόνους που πιθανών να υπάρχουν στα δεδομένα μας.

3.2.5 Συναισθηματική Ανάλυση (Sentiment Analysis)

Για την ολοκληρωμένη κατανόηση της δουλειάς μας πριν φτάσουμε στην περιγραφή της υλοποίησης αυτού του βήματος σε ότι αφορά το κομμάτι της προεπεξεργασίας των δεδομένων μας μπορούμε να παρατηρήσουμε τα βήματα στο παρακάτω σχήμα.

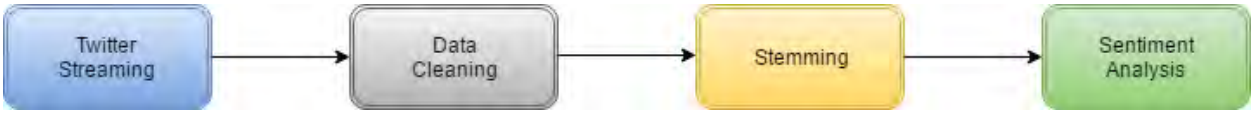

Figure 11 Βήματα ολόκληρης της διαδικασίας.

Είμαστε πλέον σε θέση να εφαρμόσουμε συναισθηματική ανάλυση στα δεδομένα μας. Ο στόχος μας είναι να εντοπίσουμε για καθένα από τα θέματα που έχουμε εντοπίσει ως πιο διαδεδομένα την άποψη της κοινής γνώμης. Η άποψη αυτή θα ανταποκρίνεται στη γνώμη των χρηστών όπως αυτή διαμορφώθηκε στο χρονικό διάστημα κατά το οποίο συλλέγαμε tweets.

Σε αυτό το σημείο θα ασχοληθούμε με το πως υλοποιήθηκε η Συναισθηματική Ανάλυση στα δεδομένα μας. Όπως αναφέραμε και στο Κεφάλαιο 2 και συγκεκριμένα στην παράγραφο 2.4, υπάρχουν δύο τρόποι για συναισθηματικοί ανάλυση. Ο πρώτος αφορά τη Συναισθηματική Ανάλυση με μηχανική μάθηση (επιβλεπόμενη και μη) και ο δεύτερος

αφορά τη Συναισθηματική Ανάλυση με χρήση Συναισθηματικού Λεξικού. Για τις ανάγκες της εργασίας αυτής χρησιμοποιήσαμε τη μέθοδο του Συναισθηματικού Λεξικού.

Το συναισθηματικό Λεξικό που χρησιμοποιήσαμε είναι αυτό που σχεδιάστηκε από το Ελληνικό Κέντρο Ερευνών (Information Technologies Institute of CERTH) (2014). Περιέχει 2315 όρους και για κάθε έναν από αυτούς παρέχει πληροφορία σχετικά με την πολικότητα του (polarity), ενώ συνοδεύεται από αντίστοιχες βαθμολογίες μεταξύ 0 και 5, οι οποίες εκφράζουν κατά πόσο μία λέξη εκφράζει κάποιο συγκεκριμένο συναίσθημα από τα έξι τα οποία υπάρχουν σύμφωνα με τη θεωρία του Paul Ekman. Τα συναισθήματα αυτά είναι ο θυμός, η αηδία, ο φόβος, η χαρά, η λύπη και η έκπληξη. Περισσότερες πληροφορίες σχετικά με αυτά παρέχονται στην παράγραφο 2.4 του Κεφαλαίου 2. Ερευνητές έχουν συγκρίνει το εν λόγω Συναισθηματικό Λεξικό με άλλα αυτής της κατηγορίας και συμπέραναν ότι παρουσιάζει μεγάλα ποσοστά κάλυψης συγκεκριμένα σε κοινωνικά δίκτυα.

Χρησιμοποιώντας λοιπόν το εν λόγω Συναισθηματικό Λεξικό, ο στόχος μας είναι να εντοπίσουμε τόσο την πολικότητα του κάθε διαδεδομένου θέματος (+, -, 0), όσο και το βαθμό στον οποίο αντιστοιχεί κάθε ένα από τα δέκα πιο διαδεδομένα θέματα που εντοπίσαμε παραπάνω, σε ένα από τα έξι συναισθήματα που μελετάμε. Το σκεπτικό σύμφωνα με το οποίο προσεγγίσαμε το συγκεκριμένο θέμα ήταν για κάθε tweet ενός από τα δέκα πιο διαδεδομένα θέματα, να εντοπίζουμε τα παραπάνω στοιχεία του και αθροιστικά να βγάζουμε συμπέρασμα για ολόκληρο το θέμα.

Επειδή ένα Λεξικό έχει περιορισμένο αριθμό λέξεων, για να αυξήσουμε την πιθανότητα ταιριάσματος μίας λέξης των δεδομένων μας με έναν από τους όρους του λεξικού μας, φροντίσαμε και στις δύο πλευρές σύγκρισης να έχουν αφαιρεθεί οι καταλήξεις (stemming). Όπως περιγράφεται και παραπάνω, οι καταλήξεις των λέξεων στα tweets μας έχουν ήδη αφαιρεθεί. Με τον ίδιο τρόπο βρίσκουμε τις ρίζες των λέξεων που περιέχει το λεξικό μας. Το μόνο που μένει από εδώ και πέρα είναι η σύγκριση των δεδομένων μας με αυτά του λεξικού.

Στο παραπάνω σχήμα βλέπουμε ένα παράδειγμα του πως δουλεύει το σύστημα μας για ένα μόνο tweet ενός δημοφιλούς θέματος. Εφαρμόζουμε το ίδιο για όλα τα tweets που αντιστοιχούν στο ίδιο δημοφιλές θέμα και παίρνουμε το αποτέλεσμα σχετικά με τα

συναισθήματα που εκφράστηκαν από τους χρήστες του Twitter για το συγκεκριμένο θέμα στο χρονικό διάστημα κατά το οποίο συλλέξαμε τα δεδομένα μας.

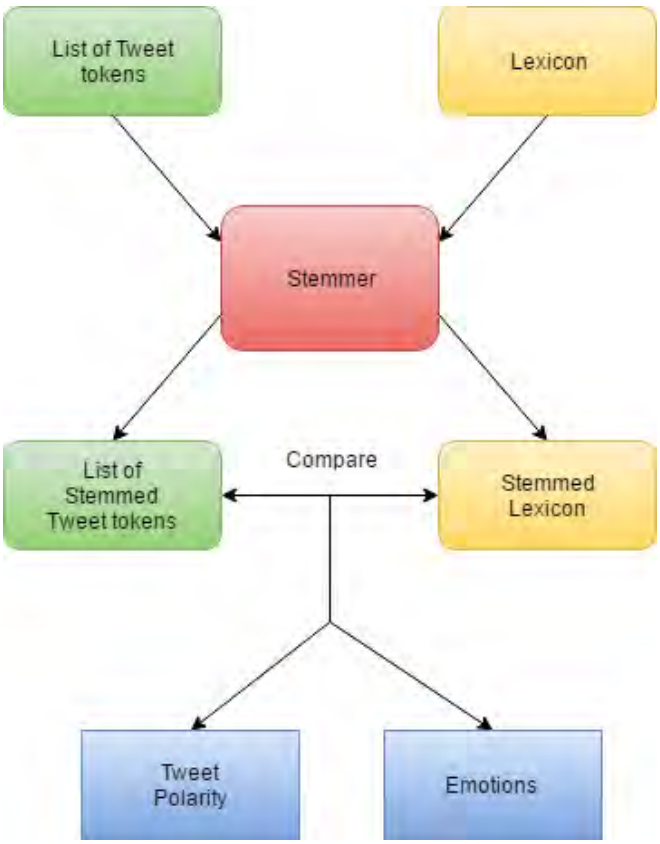

Figure 12 Διαδικασία σύγκρισης των λέξεων ενός tweet με τις λέξεις του λεξικού μας.

Με τον ίδιο τρόπο εργαζόμαστε για όλα τα δέκα διαφορετικά δημοφιλή θέματα που εντοπίσαμε. Τα αποτελέσματα αυτής της διαδικασίας θα μας δώσουν τη δυνατότητα να βγάλουμε συμπεράσματα σχετικά με όλα τα θέματα που απασχόλησαν μεγάλη μερίδα των χρηστών για το χρονικό διάστημα που επιλέγουμε να συλλέξουμε δεδομένα. Για παράδειγμα, μπορούμε να συμπεράνουμε πια θέματα αντιμετωπίζει θετικά η κοινή γνώμη και ποια αρνητικά. Ακόμη, μπορούμε να εντοπίσουμε τα θέματα συζήτησης για τα οποία οι χρήστες του Twitter εκφράζουν συναισθήματα όπως οργή ή έκπληξη. Στο Παράρτημα Β βλέπουμε την υλοποίηση των όσων περιγράψαμε παραπάνω.

Πειραματική αξιολόγηση της συγκεκριμένης υπηρεσίας που υλοποιήθηκε στα πλαίσια της παρούσας Διπλωματικής Εργασίας θα παρουσιαστεί στο Κεφάλαιο 4.

3.3 Στοχευμένη πρόταση φιλίας μεταξύ των χρηστών

Ο αριθμός των χρηστών που είναι συνδεδεμένοι σε ένα κοινωνικό δίκτυο ποικίλει ανάλογα με τη δημοτικότητα του κοινωνικού δικτύου. Το πιο δημοφιλές είναι το κοινωνικό δίκτυο Facebook στο οποίο είναι συνδεδεμένοι πάνω από ένα δισεκατομμύριο χρήστες μηνιαίως. Το λιγότερο δημοφιλές είναι το Pinterest με 150 εκατομμύρια χρήστες συνδεδεμένους μηνιαίως. Από τους παραπάνω αριθμούς βλέπουμε ότι ακόμα και το λιγότερο δημοφιλές κοινωνικό δίκτυο της σημερινής περιόδου έχει πάρα πολλούς χρήστες.

Η παρατήρηση αυτή μας οδηγεί στο συμπέρασμα ότι, ακριβώς λόγω των πολλών χρηστών, κάποιος που χρησιμοποιεί ένα κοινωνικό δίκτυο υπάρχει μεγάλη πιθανότητα να μην μπορεί επιτυχημένα να εντοπίσει χρήστες με ένα συγκεκριμένο μοτίβο. Για παράδειγμα, κάποιος χρήστης αφού εντοπίσει κάποια θέματα που τον ενδιαφέρουν από την επικαιρότητα ή ακόμα και από τηλεοπτικές εκπομπές θα ήταν χρήσιμο να μπορεί να εντοπίσει άτομα που ασχολούνται με αυτά τα θέματα.

Στο Twitter ο τρόπος που υπάρχει για να εντοπίσουμε τα άτομα που ασχολούνται με ένα συγκεκριμένο θέμα είναι αν το θέμα αυτό είναι hashtag, να το επιλέξουμε και τότε να μας εμφανιστεί μία λίστα με tweets που περιέχουν αυτό το hashtag στο κείμενό τους. Ωστόσο, όπως αναφέραμε και στην παράγραφο 3.2, δεν είναι λίγες οι φορές που το κείμενο ενός tweet δεν περιέχει hashtags. Επομένως, το Twitter δεν προσφέρει στους χρήστες του κάποιον τρόπο να εντοπίσουν άλλους χρήστες που ασχολούνται με ένα συγκεκριμένο θέμα αν αυτοί δεν έχουν βάλει το αντίστοιχο hashtag στη δημοσίευση τους.

Καταλήξαμε ότι η υλοποίηση μας θα γίνει ακόμη πιο χρήσιμη αν ασχοληθούμε με αυτό σε συνδυασμό και με το κομμάτι των συναισθημάτων κάθε χρήστη. Δηλαδή, για κάθε ένα από τα σημαντικά θέματα που εντοπίστηκαν όπως περιγράφεται στην παράγραφο 3.2, κάνουμε κατηγοριοποίηση των χρηστών ανάλογα με τη γνώμη τους για το θέμα στο οποίο συμμετέχουν και ανάλογα με το πόσο έντονη είναι η ενασχόληση τους με αυτό. Σκοπός μας είναι η στοχευμένη πρόταση φιλίας μεταξύ των χρηστών οι οποίοι όχι μόνο ασχολούνται με το ίδιο θέμα αλλά έχουν και την ίδια θετική ή αρνητική άποψη για αυτό.

3.3.1 Επικοινωνία με το Twitter και Συλλογή Δεδομένων

Το σύνολο των δεδομένων μας αποτελείται σε πρώτη φάση, από τα δεδομένα που συλλέξαμε από το Twitter όπως περιγράφεται στην παράγραφο 3.3.2. Στην παρακάτω εικόνα βλέπουμε ένα tweet όπως αυτό είναι αποθηκευμένο στη συλλογή δεδομένων μας.

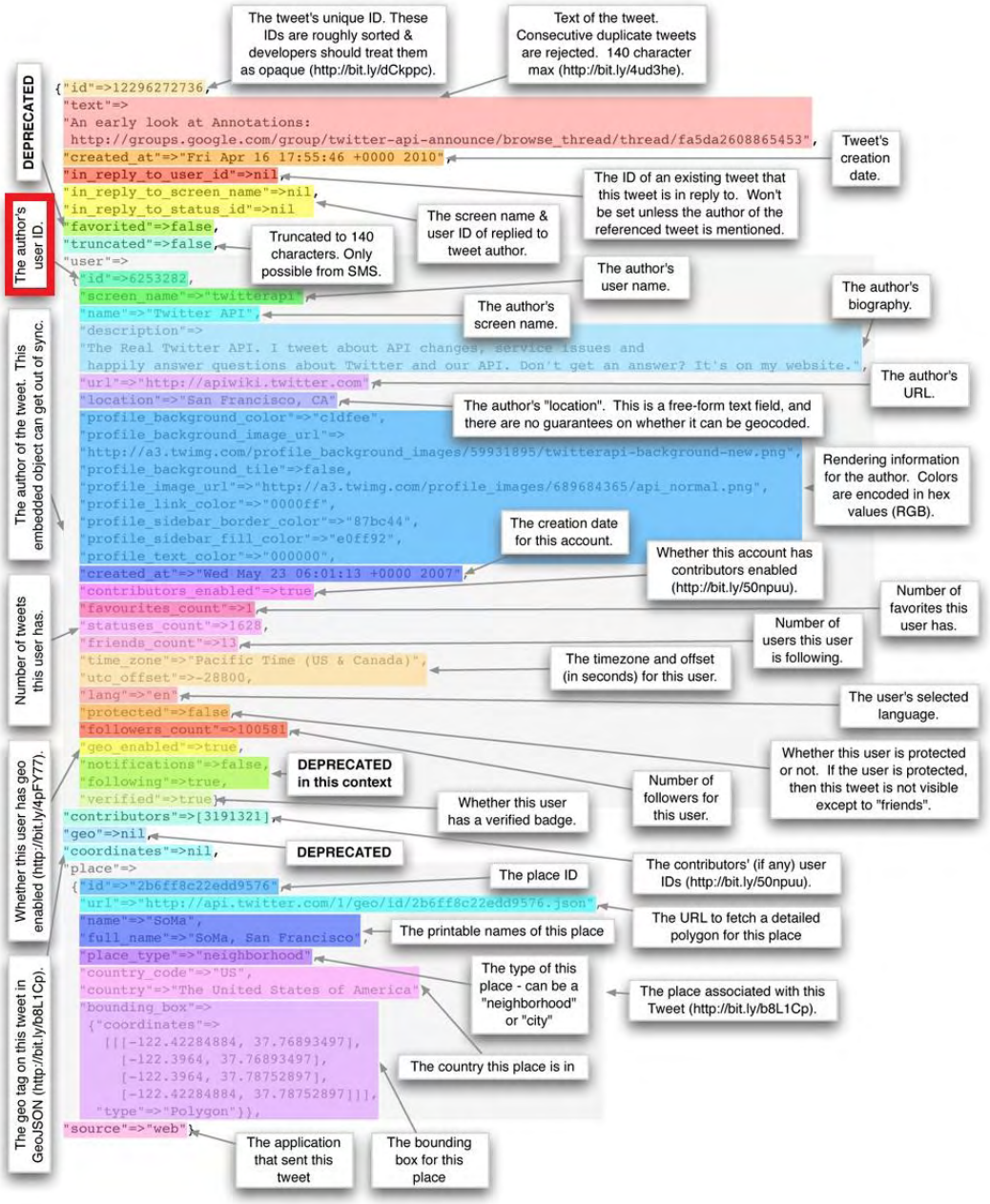

Figure 13 Η δομή ενός tweet.

Η παραπάνω εικόνα μας βοηθάει να κατανοήσουμε τη δομή ενός tweet. Από κάθε tweet πληροφορούμαστε εκτενώς για το ίδιο το tweet αλλά και για το προφίλ του χρήστη που το δημοσίευσε. Σε κόκκινο πλαίσιο έχουμε βάλει το πεδίο της ταυτότητας του χρήστη (The author's user id) στον οποίο ανήκει το συγκεκριμένο tweet. Πρόκειται για ένα πολύ χρήσιμο πεδίο που θα χρησιμοποιήσουμε για την υλοποίηση της συγκεκριμένης υπηρεσίας της παρούσας Διπλωματικής Εργασίας με τον τρόπο που θα αναφέρουμε στη συνέχεια.

3.3.2 Εύρεση ταυτότητας του κάθε χρήστη

Κατά τη διαδικασία της υλοποίησης της υπηρεσίας εντοπισμού σημαντικών θεμάτων που περιγράψαμε στην παράγραφο 3.2, κατασκευάσαμε λίστες που περιέχουν η κάθε μία όλα τα tweets που αντιστοιχίζονται σε ένα σημαντικό θέμα. Υπενθυμίζουμε ότι αυτή η αντιστοίχιση των tweets έγινε σε πρώτη φάση με βάση τα hashtags και σε δεύτερη φάση με βάση τις τρεις πιο δημοφιλείς λέξεις για κάθε hashtag.

Αφού επεξεργαστήκαμε κατάλληλα τα tweets της κάθε κατηγορίας, καταφέραμε να πάρουμε το πεδίο της ταυτότητας του χρήστη (The author's user id) που δημοσίευσε το κάθε tweet. Επομένως, μάθαμε ποιοι χρήστες ασχολήθηκαν στο Twitter με κάποιο θέμα ιδιαίτερα δημοφιλές και ποιο είναι αυτό το συγκεκριμένο θέμα.

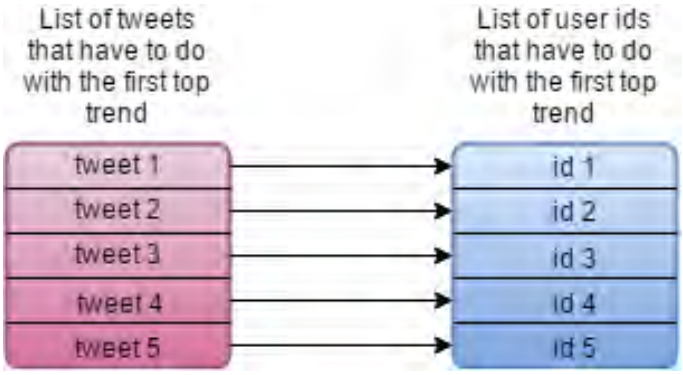

Figure 14 Παράδειγμα ανάδειξης των Ids των χρηστών που ναφέρθηκαν σε ένα hashtag.

Στην παραπάνω εικόνα βλέπουμε ένα παράδειγμα αυτής της διαδικασίας για το πρώτο δημοφιλέστερο θέμα που αναδείξαμε στην παράγραφο 3.2. Για το συγκεκριμένο παράδειγμα, υποθέσαμε ότι υπάρχουν πέντε tweets τα οποία αναφέρονται στο

δημοφιλέστερο θέμα μας. Με τον ίδιο τρόπο δουλέψαμε για όλα τις λίστες από tweets. Τελικά, δημιουργήσαμε δέκα λίστες από ταυτότητες χρηστών που σχολίασαν για κάθε ένα από τα δέκα θέματα που συμπεράναμε ότι είναι τα σημαντικότερα.

Για τις ανάγκες της εργασίας αυτής, το πεδίο της ταυτότητας κάθε χρήστη είναι το μοναδικό που μας ενδιαφέρει και για αυτό το λόγο κρατήσαμε μόνο αυτό το πεδίο στην παρούσα ανάλυση. Ωστόσο, όπως φαίνεται στην εικόνα 13, οι πληροφορίες που συνοδεύουν το κείμενο ενός tweet είναι πολλές και μας δίνουν τη δυνατότητα να σκεφτούμε και να πειραματιστούμε σχετικά με αυτές. Κάποια από τα υπόλοιπα πεδία ενδεικτικά είναι ο μετρητής των αναδημοσιεύσεων ενός tweet (retweet_count) και ο μετρητής των φορών που ένα tweet έχει μαρκαριστεί ως αγαπημένο από κάποιον χρήστη (favorite_count). Επιπλέον, για κάθε χρήστη μπορούμε να αντλήσουμε και άλλες πληροφορίες εκτός από την ταυτότητα του, όπως πληροφορίες που έχει εισάγει ο ίδιος στο προφίλ του στο στο Twitter.

3.3.3 Συλλογή Δεδομένων της Δραστηριότητας των Χρηστών

Αφού λοιπόν εντοπίσαμε τις ταυτότητες των χρηστών που μας ενδιαφέρουν, η επόμενη εργασίας μας έχει να κάνει με την ανίχνευση της δραστηριότητας του κάθε χρήστη. Όπως αναφέραμε και νωρίτερα ο στόχος μας είναι να εντοπίσουμε χρήστες που ασχολούνται κατά κόρων με ένα θέμα ώστε να τους προτείνουμε σε άλλους χρήστες που ενδιαφέρονται για το συγκεκριμένο θέμα. Σε δεύτερη φάση, μας ενδιαφέρει να κατηγοριοποιήσουμε τους χρήστες με βάση το αν μιλάνε θετικά η αρνητικά για το εν λόγω σημαντικό θέμα.

Ξεκινώντας λοιπόν τη διαδικασία ανίχνευσης των τελευταίων δραστηριοτήτων του κάθε χρήστη που μας ενδιαφέρει, το πρώτο βήμα που κάναμε έχει να κάνει με τη συλλογή δεδομένων από τα είκοσι (20) τελευταία δημοσιευμένα tweets του. Για την συλλογή αυτών των δεδομένων, χρειάστηκε να συνδεθούμε για ακόμα μία φορά με μία από τις διεπαφές που προσφέρει το Twitter. Σε αντίθεση με την προηγούμενη φορά, για αυτή τη συλλογή δεδομένων συνδεθήκαμε με το REST API του Twitter. Πληροφορίες σχετικά με το REST API του Twitter παρατίθενται στην παράγραφο 2.2.2.

Για τη σύνδεση με το REST API του Twitter χρησιμοποιήσαμε τα κλειδιά (Consumer Key, Consumer Secret, Access Token και Access Token Secret) που αποκτήσαμε με την

εφαρμογή που δημιουργήσαμε όπως περιγράφεται στην παράγραφο 3.2.1. Επιπλέον, χρησιμοποιήσαμε για ακόμη μία φορά την βιβλιοθήκη της Python, Tweepy. Για τους σκοπούς αυτής της υπηρεσίας που αναπτύχθηκε στην εργασία αυτή, χρησιμοποιήθηκαν οι μέθοδοι get_user και user_timeline της Tweepy. Η μέθοδος get_user επιστρέφει πληροφορίες σχετικά με ένα χρήστη ενώ η μέθοδος user_timeline επιστρέφει τα είκοσι (20) τελευταία tweets που δημοσίευσε ένας συγκεκριμένος χρήστης. Για να δηλώσουμε για ποιον χρήστη θέλουμε να συλλέξουμε τα είκοσι τελευταία του tweets δίνουμε σαν όρισμα στη συγκεκριμένη μέθοδο κάποιο αναγνωριστικό του όπως την ταυτότητα του (user id).

Στο τέλος αυτής της διαδικασίας έχουμε συλλέξει για κάθε χρήστη που μας ενδιαφέρει τα είκοσι τελευταία του tweets. Έχουμε δημιουργήσει έτσι ένα νέο σύνολο δεδομένων πάνω στο οποίο θα δουλέψουμε από εδώ και στο εξής.

Όσον αφορά την επιλογή μας να χρησιμοποιήσουμε μία μέθοδο της βιβλιοθήκης Tweepy η οποία να επιστρέφει είκοσι από τα προηγούμενα Tweets, δεν ήταν μία τυχαία επιλογή. Ενώ υπάρχουν και άλλες μέθοδοι οι οποίες επιστρέφουν περισσότερα tweets ενός χρήστη, για τη συγκεκριμένη υλοποίηση επιλέξαμε αυτή την προσέγγιση. Θεωρήσαμε ότι για την κλίμακα στην οποία αναπτύσσεται το σύστημα μας είναι το κατάλληλο πλήθος tweets.

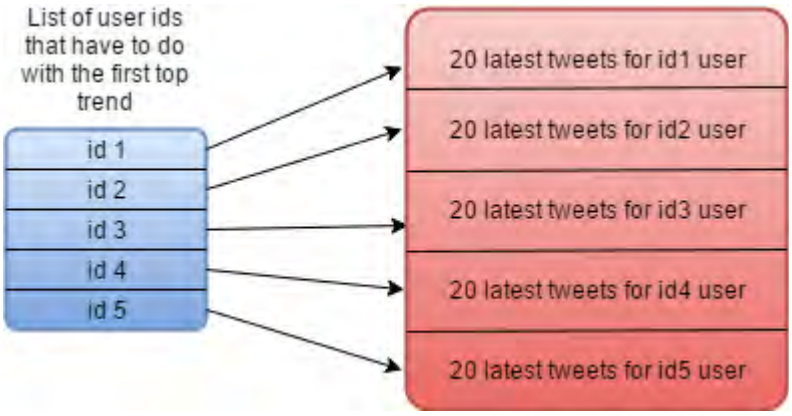

Figure 15 Παράδειγμα συλλογής των 20 tweets όλων των χρηστών ενός hashtag.

Στο παραπάνω σχήμα, βλέπουμε ένα παράδειγμα της λειτουργίας που περιγράψαμε. Έστω ότι οι χρήστες που έχουν ασχοληθεί με ένα σημαντικό θέμα είναι πέντε και οι ταυτότητες τους είναι id1, id2, id3, id4 και id5. Αφού συνδεθήκαμε με την κατάλληλη διεπαφή του Twitter συλλέξαμε μία λίστα με τα είκοσι τελευταία tweets κάθε χρήστη.

Υπάρχει ένα σημείο της παραπάνω διαδικασίας το οποίο πρέπει να το προσέξουμε ιδιαίτερα. Πρόκειται για την περίπτωση όπου κάποιος χρήστης δεν θέλει να διατηρεί δημόσιο προφίλ στο Twitter και για αυτό δεν επιτρέπει σε όλους, εκτός από τους φίλους του, να έχουν πρόσβαση σε αυτό. Επομένως, θα πρέπει να ελέγχουμε αν κάποιος από τους χρήστες που μας ενδιαφέρουν είναι σε αυτή την κατηγορία και να ενημερώνουμε κατάλληλα το σύστημα μας ώστε να πηγαίνουμε στον επόμενο χρήστη.

3.3.4 Επεξεργασία Δεδομένων

Σε αυτό το στάδιο έχουμε συλλέξει είκοσι tweets για κάθε χρήστη για κάθε ένα από τα δέκα σπουδαιότερα θέματα συζήτησης στο Twitter. Στόχος μας είναι να εντοπίσουμε μέσα σε αυτά τα είκοσι tweets αν εμφανίζεται σε αυτά τόσο το hashtag, όσο και οι τρεις πιο συχνά εμφανιζόμενες λέξεις για αυτό όπως αυτές εντοπίστηκαν κατά τη διάρκεια υλοποίησης της πρώτης υπηρεσίας της παρούσας εργασίας. Επιπλέον, θέλουμε να ξέρουμε και πόσες φορές εμφανίζονται οι παραπάνω όροι στα tweets του κάθε χρήστη. Αυτό είναι απαραίτητα καθώς θέλουμε να ταξινομήσουμε τους χρήστες ανάλογα με το βαθμό που ασχολήθηκαν με κάποιο συγκεκριμένο θέμα.

Για την πραγματοποίηση των παραπάνω συγκρίσεων, τα είκοσι tweets του κάθε χρήστη πρέπει να υποβληθούν σε επεξεργασία σαν και αυτή που περιγράφεται στην παράγραφο 3.2.2 όπου αναφέρεται αναλυτικά όλη η διαδικασία αφαίρεσης υπερσυνδέσμων (URLs), εικονιδίων συναισθήματος (emoticons) και αριθμών. Επιπλέον πραγματοποιούμε και πάλι αφαίρεση συγκεκριμένων λέξεων και συμβόλων.

Το επόμενο βήμα είναι να συγκρίνουμε τα tweets κάθε χρήστη με κάθε ένα από τα hashtags που αποτελούν τα σημαντικότερα θέματα και με τις τρεις πιο συχνά εμφανιζόμενες λέξεις για καθένα από αυτά. Θέλουμε να εντοπίσουμε πόσα και πια από τα παρελθοντικά tweets ενός χρήστη κυμαίνονται στο ίδιο θέμα για το οποίο μιλούσε όταν

τον εντοπίσαμε και το οποίο σε εκείνη την χρονική περίοδο αποτελούσε ένα από τα σημαντικότερα θέματα για τους χρήστες του Twitter.

Αφού πραγματοποιηθεί η εν λόγω σύγκριση, θα προκύψει για κάθε χρήστη μία λίστα από tweets τα οποία συμπεράναμε ότι σχετίζονται με το σημαντικό θέμα για το οποίο μιλούσε ο χρήστης όταν τον εντοπίσαμε.

Επειδή στη συνέχεια θέλουμε να εφαρμόσουμε Συναισθηματική Ανάλυση με χρήση Συναισθηματικού Λεξικού έτσι ώστε να βρούμε ποια είναι η γνώμη που έχει ένας χρήστης για το θέμα που εντοπίσαμε ότι τον απασχολεί (αν υπάρχει τέτοιο), το τελευταίο στάδιο της επεξεργασίας δεδομένων είναι να εφαρμόσουμε ξανά τον αλγόριθμο αφαίρεσης καταλήξεων του Καλοπόδη όπως περιγράψαμε και στην παράγραφο 3.2.4.

3.3.5. Συναισθηματική Ανάλυση για κάθε χρήστη

Όπως αναφέραμε και νωρίτερα, επιθυμούμε να εντοπίσουμε τη συναισθηματική διάθεση με την οποία ο κάθε χρήστης μιλάει για ένα από τα κορυφαία θέματα στο οποίο αναφέρεται. Αφού έχουμε εφαρμόσει τον αλγόριθμο του Καλοπόδη για αφαίρεση καταλήξεων στα tweets του χρήστη τα οποία εντοπίσαμε ότι αναφέρονται σε ένα διαδεδομένο θέμα, είμαστε πλέον σε θέση να βρούμε την πολικότητα των συγκεκριμένων tweets. Χρησιμοποιούμε το Συναισθηματικό Λεξικό που σχεδιάστηκε από το Ελληνικό Κέντρο Ερευνών (Information Technologies Institute of CERTH) όπως περιγράφουμε στην παράγραφο 3.2.5. Προσθέτοντας τα αποτελέσματα για κάθε tweet, έχουμε βρει κατά πόσο ο κάθε χρήστης μιλάει θετικά ή αρνητικά για κάποιο δημοφιλές θέμα στα τελευταία 20 tweets που δημοσίευσε στο προφίλ του στο Twitter. Επίσης, για τον κάθε χρήστη λαμβάνουμε υπόψη για την συνέχεια της υλοποίησής μας και το πόσα tweets από τα 20 που του αντιστοιχούσαν, αφορούν το θέμα που ψάχναμε.

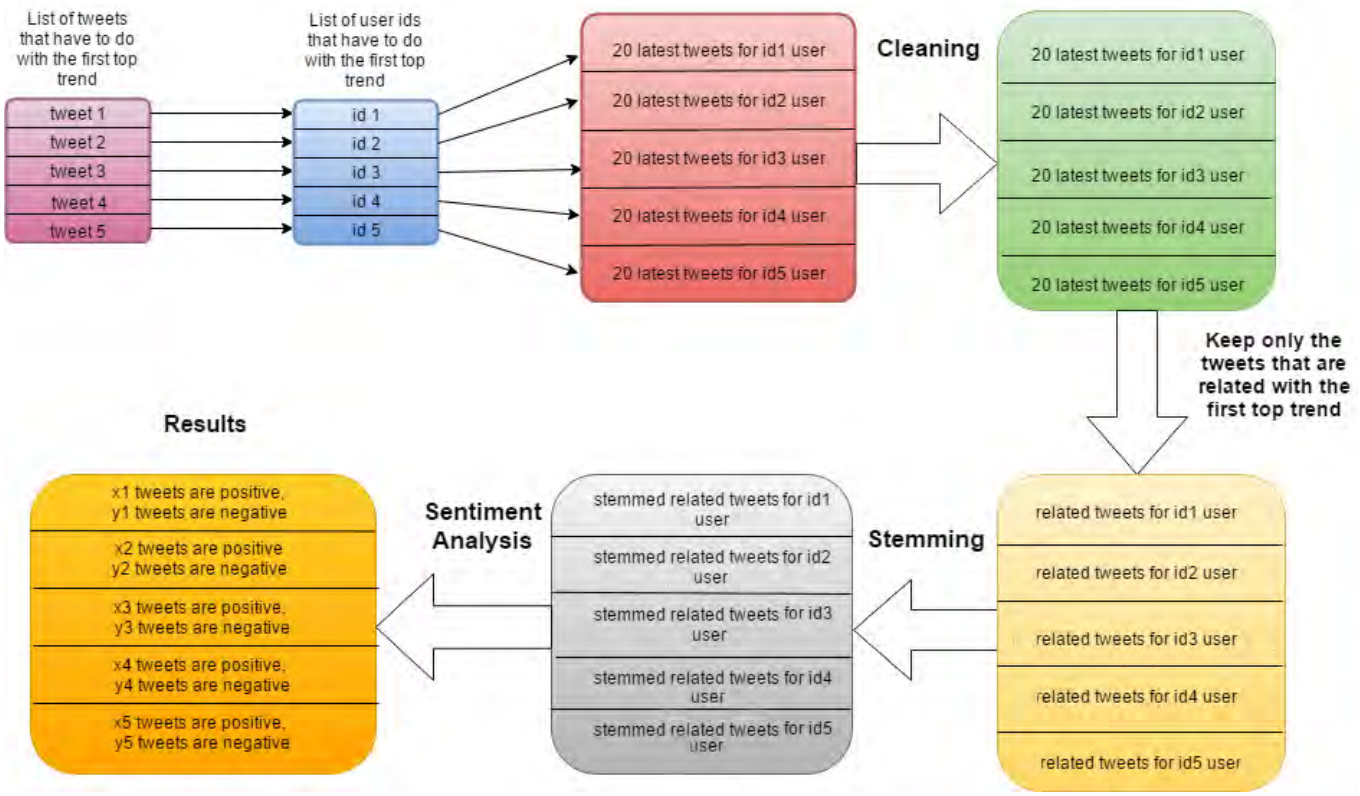

Figure 16 Ολόκληρη η διαδικασία της δεύτερης υπηρεσίας μας.

Στην παραπάνω εικόνα αναπαριστούμε ολόκληρη τη διαδικασία που περιγράψαμε νωρίτερα. Το παράδειγμα αυτό αναφέρεται στα βήματα που γίνονται για ένα μόνο δημοφιλές θέμα, το οποίο θεωρούμε για τις ανάγκες αυτού του παραδείγματος ότι αφορά πέντε διαφορετικά tweets. Αφού βρούμε ποιοι είναι οι πέντε χρήστες, παίρνουμε τα είκοσι τελευταία tweets τους και τα υποβάλουμε σε διαδικασία φιλτραρίσματος. Στη συνέχεια, κρατάμε από αυτά μόνο τα tweets που σχετίζονται με το σημαντικό θέμα που μελετάμε. Τέλος, αφού περάσουμε όλες τις λέξεις όλων των εναπομεινάντων tweets από τον αλγόριθμο αφαίρεσης καταλήξεων του Καλοπόδη, υποβάλλουμε το κάθε tweet σε Συναισθηματική Ανάλυση. Τα αποτελέσματα για το χρήστη με id1 είναι για παράδειγμα, ότι x1 tweets είναι θετικά και x2 tweets είναι αρνητικά.

3.3.6 Αποτελέσματα αλγορίθμου πρότασης φίλων

Το τελευταίο στάδιο υλοποίησης της υπηρεσίας στοχευμένης πρότασης φίλων στο κοινωνικό δίκτυο Twitter έχει να κάνει αρχικά με τον διαχωρισμό των χρηστών σε αυτούς που σχολίασαν θετικά ένα θέμα και σε αυτούς που το σχολίασαν αρνητικά. Εν συνεχεία, φροντίζουμε να ταξινομήσουμε τις δύο λίστες ανάλογα με πόσα από τα τελευταία είκοσι tweets ενός χρήστη αφορούσαν το θέμα που αναζητούσαμε. Έτσι, προκύπτουν δύο λίστες από χρήστες του Twitter η κάθε μία από τις οποίες περιέχει τους χρήστες που εκφράζουν την ίδια γνώμη για ένα θέμα και μάλιστα ξεκινώντας από τον χρήστη που αφιερώνει τις περισσότερες δημοσιεύσεις στο συγκεκριμένο θέμα.

Συμπερασματικά, αν ένας χρήστης παρατηρήσουμε ότι σχολιάζει αρνητικά ή θετικά ένα συγκεκριμένο θέμα που αποτελεί ένα από τα δέκα πιο πολυσυζητημένα θέματα στο Twitter για μία συγκεκριμένη χρονική περίοδο, είμαστε σε θέση να του προτείνουμε άλλους χρήστες οι οποίοι ασχολούνται με το εν λόγω θέμα. Οι χρήστες που θα του προτείνουμε θα έχουν εκφράσει στο παρελθόν την ίδια άποψη με αυτόν. Μάλιστα, θα ξεκινήσουμε τις προτάσεις φιλίας με έναν χρήστη που έχει ασχοληθεί όσο το δυνατόν περισσότερο με το θέμα που μας ενδιαφέρει. Σε περίπτωση που υπάρχει είδη δεσμός διαδικτυακής φιλίας μεταξύ των δύο χρηστών, θα προχωρήσουμε στον επόμενο χρήστη.

Στην παραπάνω εικόνα βλέπουμε πως έχουν διαμορφωθεί τα δεδομένα μας στο τέλος της υλοποίησης.

Figure 17 Αναπαράσταση της μορφής των αποτελεσμάτων μας για όλα τα hashtags.

Η υλοποίηση της υπηρεσίας της στοχευμένης πρότασης φιλίας φαίνεται στο Παράρτημα Γ και στο Παράρτημα Δ. Πιο συγκεκριμένα, στο Παράρτημα Γ βλέπουμε όλη τη διαδικασία εύρεσης των ταυτοτήτων των χρηστών που μας ενδιαφέρουν. Στη συνέχεια, για κάθε χρήστη καλούμε τη συνάρτηση που φαίνεται στο Παράρτημα Δ ώστε να πάρουμε τα είκοσι τελευταία tweets για τον κάθε έναν.

4. Πειραματική Αξιολόγηση

Σε αυτό το κεφάλαιο, θα αναφερθούμε στα πειράματα που διεξήγαμε αφού ολοκληρώσαμε την ανάπτυξη του συστήματος που περιγράψαμε στο Κεφάλαιο 3. Πρόκειται για προσπάθεια πειραματικής αξιολόγησής ολόκληρης της διαδικασίας όπως την αναλύσαμε μέχρι τώρα. Θα αναφερθούμε σε δύο διαφορετικά πειράματα που πραγματοποιήθηκαν με δεδομένα που συλλέξαμε από το κοινωνικό δίκτυο Twitter του μήνα Ιουνίου του έτους 2017.

4.1 Δεδομένα

Όταν τα δεδομένα ενός συστήματος προέρχονται από κοινωνικά δίκτυα, όπως το Twitter, είναι βέβαιο πως θα αντανακλούν την επικαιρότητα και θα έχουν να κάνουν με δημοφιλή θέματα για την περίοδο που τα συλλέξαμε. Επιπλέον, προέρχονται απευθείας από τους χρήστες και άρα το θέμα το οποίο πραγματεύονται δεν είναι ελεγχόμενο από εμάς.

4.2 Παρουσίαση πρώτου πειράματος

Σύνολο των tweets	7963
Tweets που περιέχουν hashtags	704
Tweets που δεν περιέχουν hashtags	7259
Tweets που κατηγοριοποιήσαμε εμείς σε κάποιο hashtag	439

Πίνακας 3 Τα δεδομένα του πρώτου πειράματος.

Στον παραπάνω πίνακα παρουσιάζουμε κάποια βασικά δεδομένα που αφορούν το συγκεκριμένο πείραμα. Αρχικά αναφερόμαστε στο πλήθος των tweets που συλλέξαμε και στη συνέχεια σε πόσα από αυτά περιείχαν ένα τουλάχιστον hashtag και πόσα δεν περιείχαν κανένα. Επίσης παραθέτουμε τον αριθμό των tweets τα οποία

κατηγοριοποιήσαμε εμείς σε κάποιο σημαντικό θέμα, δηλαδή με κάποιο από τα δέκα δημοφιλέστερα hashtags που εντοπίσαμε.

Από τα παραπάνω δεδομένα παρατηρούμε πως τα tweets που περιέχουν κάποιο hashtag είναι πολύ λιγότερα από τα tweets που δεν περιέχουν και άρα αν κάποιος μελετήσει μόνο τα πρώτα θα χάσει πολλή πληροφορία. Μπορέσαμε να αυξήσουμε τον αριθμό των tweets που ανήκουν σε κάποιο σημαντικό θέμα, αφού κατηγοριοποιήσαμε κάποια επιπλέον tweets με τη διαδικασία που αναλύσαμε στην παράγραφο 3.2.

Μία σημαντική παρατήρηση σε αυτό το στάδιο είναι ότι παρόλο που καταφέραμε να αυξήσουμε τον αριθμό τον tweets που συγκαταλέγονται σε κάποιο σημαντικό θέμα εφαρμόζοντας κάποιου είδους κατηγοριοποίηση, υπάρχουν πολλά tweets τα οποία μένουν ακόμη εκτός των σημαντικών θεμάτων. Αυτό προκύπτει λόγω του ότι στο συγκεκριμένο πείραμα έχουμε πολλά tweets που περιέχουν hashtags με χαμηλή δημοτικότητα. Συνεπώς αυτά τα tweets μένουν εκτός της ανάλυσης μας.

Στον παρακάτω πίνακα βλέπουμε τα δέκα σημαντικότερα θέματα που εντοπίσαμε στα tweets που συλλέξαμε σε συνδυασμό με τις τρεις πιο δημοφιλείς λέξεις που εντοπίσαμε ότι αντιστοιχούν σε κάθε ένα από αυτά.

Τα 10 Top Hashtags	**Δημοφιλής Λέξη 1**	**Δημοφιλής Λέξη 2**	**Δημοφιλής Λέξη 3**
#survivorgr	survivor	ντανο	τελικό
#hellass	ελλάδα	διάστημα	εκτόξευση
#hellassat	ελλάδα	διάστημα	εκτόξευση
#καυσωνας	τώρα	έξω	συμπολίτες
#cyprus	σομαλοί	παρανομοι	μεταναστεσ
#ειδήσεις	via	παγίδα	νέα
#news	ροδιακη	via	irodiaki
#greece	σκουπιδιαρηδες	σεβασμο	ους
#lykavitosgr	συριζα	σκορπιό	μαρίνος
#vouli	προσωπικού	επί	ενώ

Πίνακας 4 Εντοπισμός δημοφιλών λέξεων για κάθε hashtag.

Με την χρήση των παραπάνω δημοφιλών λέξεων, καταφέραμε να διευρύνουμε το σύνολο των tweets που αντιστοιχούν σε κάθε ένα από τα σημαντικά θέματα που εντοπίσαμε. Παρατηρούμε εδώ ότι τα hashtags #hellass και #hellassat τα οποία όπως είναι φανερό πραγματεύονται το ίδιο θέμα, έχουν και τις ίδιες δημοφιλέστερες λέξεις.
Συγκεκριμένα, τα tweets που προσθέσαμε για κάθε κατηγορία είναι τα ακόλουθα.

Τα 10 Top Hashtags	**Πλήθος Tweets που κατηγοριοποιήσαμε σε κάθε hashtag**
#survivorgr	37
#hellass	118
#hellassat	0
#καυσωνας	134
#cyprus	1
#ειδήσεις	46
#news	0
#greece	0
#lykavitosgr	61
#vouli	42

Πίνακας 5 Πόσα tweets κατηγοριοποιήσαμε για κάθε hashtag.

Από τον παραπάνω πίνακα παρατηρούμε αρχικά ότι σε κάποια συγκεκριμένα hashtags κατηγοριοποιήσαμε πολλά tweets ενώ για κάποια άλλα κανένα. Το γεγονός αυτό δεν είναι ανησυχητικό καθώς υπάρχει πιθανότητα να μην υπάρχουν κάποια tweets που να εντάσσονται σε αυτά που είναι προς κατηγοριοποίηση σύμφωνα με τον αλγόριθμό μας.

Παρατηρούμε επίσης ότι, στο #hellass κατηγοριοποιήθηκαν 118 tweets ενώ στο #hellassat δεν κατηγοριοποιήθηκε κανένα ενώ είχαν τις ίδιες ακριβώς δημοφιλείς λέξεις. Αυτό συμβαίνει επειδή αν ο αλγόριθμός μας εντοπίσει ισοψηφία στα ταιριάσματα δύο hashtags με ένα tweet, επιλέγει να τα ταξινομήσει στο πιο δημοφιλές. Η λογική που κρύβεται πίσω από αυτό, έχει να κάνει με την προσπάθεια μας να τονίσουμε τα δημοφιλέστερα hashtags σε μία προσπάθεια αντιπροσωπευτικής αναπαράστασης των προτιμήσεων της κοινής γνώμης.

Στο διάγραμμα που ακολουθεί βλέπουμε συγκεντρωτικά τα tweets που αντιστοιχούν σε κάθε hashtag. Με μπλε χρώμα αναπαριστούμε τα tweets που περιείχαν στο κείμενο τους το αντίστοιχο hashtag και με κόκκινο χρώμα αναπαριστούμε τα tweets που κατηγοριοποιήσαμε εμείς στη συνέχεια.

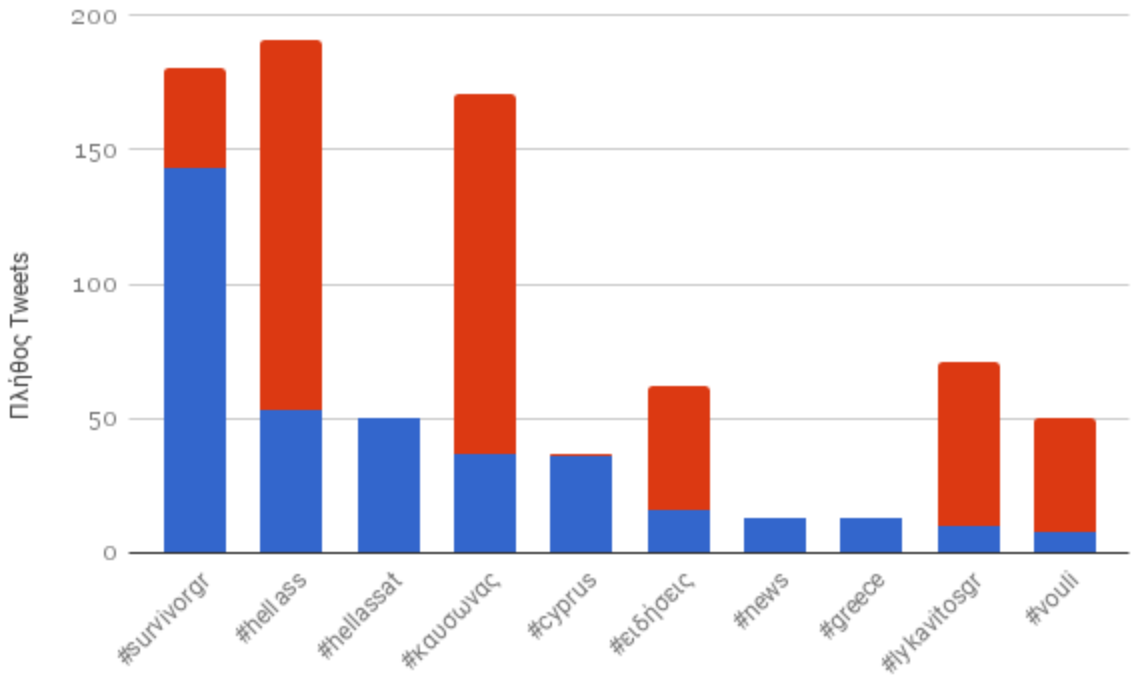

Figure 18 Hashtag σε συνδυασμό με πλήθος των tweets που τα περιέχουν.

Εύκολα παρατηρούμε ότι η κατάταξη των δημοφιλέστερων θεμάτων έχει αλλάξει μετά την κατηγοριοποίηση που εφαρμόσαμε. Η παρατήρηση αυτή αποτελεί χαρακτηριστική ένδειξη της σπουδαιότητας του αλγορίθμου μας.

Σε μία προσπάθεια να δούμε αν τα συγκεκριμένα θέματα που εντοπίσαμε σχετίζονται με τα θέματα που κυριαρχούν στα ειδησεογραφικές ιστοσελίδες, επισκεφθήκαμε κάποιες από αυτές. Διαπιστώσαμε την άμεση συνάφεια που εμφανίζουν τα θέματα που αναδείξαμε ως σημαντικότερα με αυτά που παρουσιάζονται σε ιστοσελίδες με ειδησεογραφικό περιεχόμενο.

Αφού δημιουργήσαμε νέες λίστες με τα tweets που ανήκουν σε κάθε ένα από τα σημαντικά θέματα, εφαρμόσαμε Συναισθηματική Ανάλυση χρησιμοποιώντας τη μέθοδο του Συναισθηματικού Λεξικού όπως αναλύσαμε στην παράγραφο 3.2. Τα αποτελέσματα αναπαρίστανται στο παρακάτω ραβδόγραμμα. Οι θετικές τιμές στις ράβδους μας δείχνουν ότι οι χρήστες του twitter στη χρονική διάρκεια που συλλέγαμε tweets εκφράζονταν θετικά για το συγκεκριμένο θέμα. Αντίστοιχα, οι αρνητικές τιμές μας δείχνουν

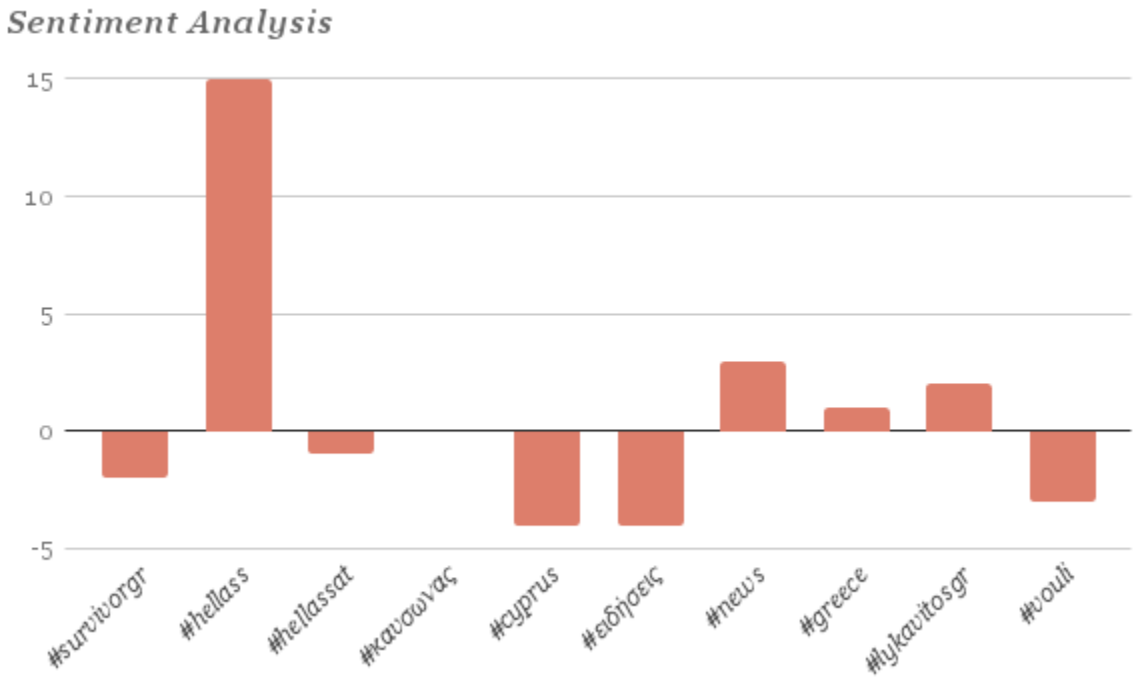

Figure 19 Αποτελέσματα ανάλυσης συναισθήματος για το πρώτο πείραμά μας.

ότι οι χρήστες εξέφραζαν αρνητικά συναισθήματα. Το μηδέν υποδηλώνει ότι εξετάζοντας στο σύνολο τους τα tweets τα οποία αντιστοιχούν σε ένα σημαντικό θέμα, τα συναισθήματα των επιμέρους tweets αλληλοεξουδετερώθηκαν.

Ωστόσο, όπως αναφέραμε και στην παράγραφο 3.2, η Ανάλυση Συναισθήματος μπορεί να επεκταθεί πολύ παραπάνω από την απλή αναφορά στην πολικότητα των δημοσιεύσεων των χρηστών. Συγκεκριμένα, μπορούμε για κάθε ένα από τα θέματα που αναδείξαμε ως σημαντικά να αναδείξουμε τα ακριβή συναισθήματα που του αντιστοιχούν επιλέγοντας ανάμεσα στα εξής συναισθήματα: ο θυμός, η αηδία, ο φόβος, η χαρά, η λύπη και η έκπληξη.

Στα παρακάτω διαγράμματα αναπαριστούμε κάποια από τα κορυφαία θέματα συζήτησης σε συσχέτιση με τα συναισθήματα που εξέφρασαν για αυτά οι χρήστες του Twitter κατά τη διάρκεια της συλλογής των δεδομένων μας. Για την εύρεση των συγκεκριμένων συναισθημάτων χρησιμοποιήσαμε και πάλι το Συναισθηματικό Λεξικό που μας βοήθησε να βρούμε την πολικότητα ενός θέματος στο προηγούμενο βήμα της ανάλυσης μας.

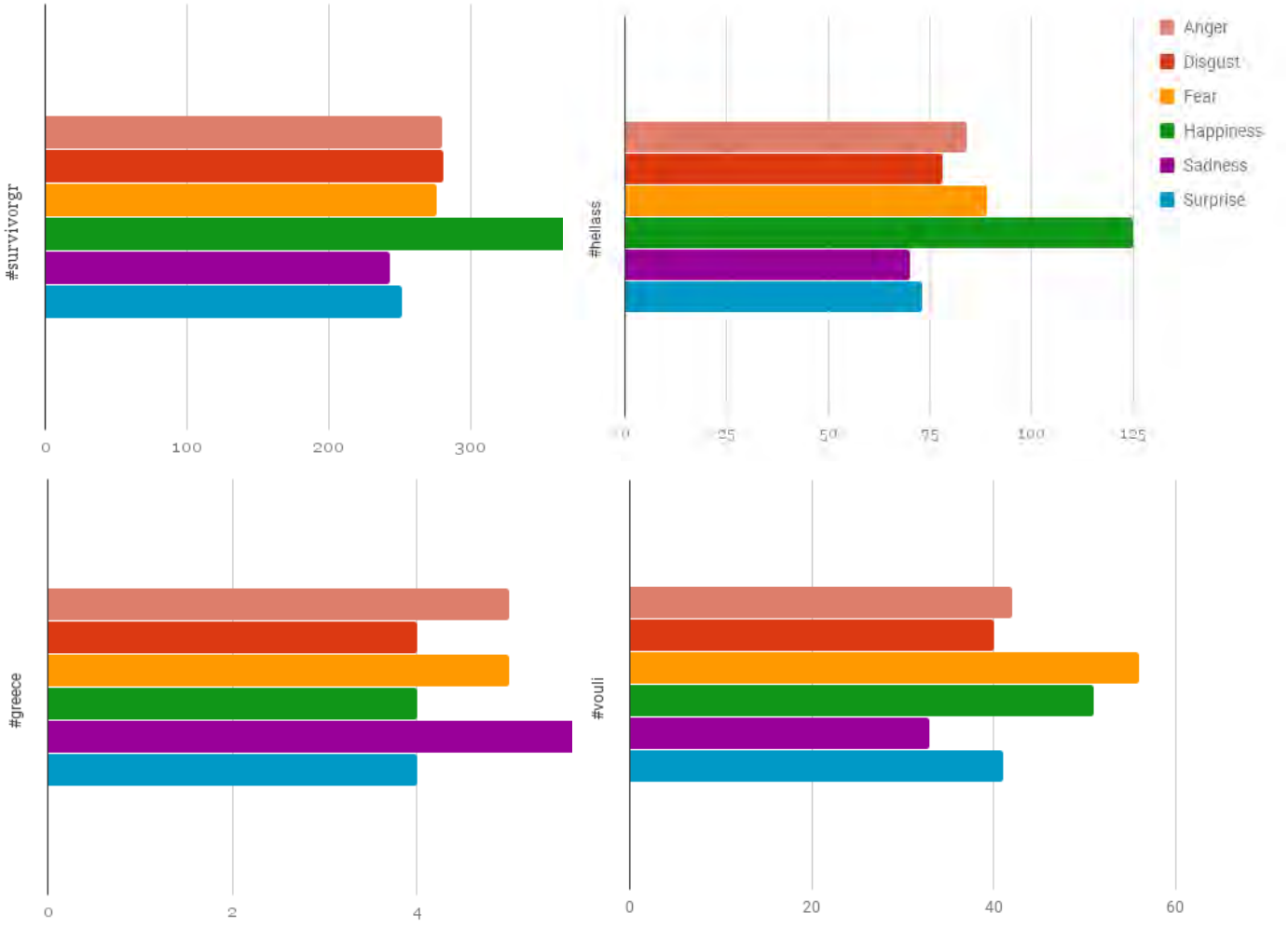

Figure 20 Αποτελέσματα εύρεσης συγκεκριμένου συναισθήματος για κάποια από τα hashtags.

Ας πάρουμε για παράδειγμα το διάγραμμα που αντιστοιχεί το #hellass. Όπως φαίνεται και στο ραβδόγραμμα που υποδηλώνει την πολικότητα κάθε σημαντικού θέματος, οι χρήστες κατά τη διάρκεια της συλλογής των δεδομένων μας είχαν θετική άποψη για το συγκεκριμένο θέμα. Το ίδιο διαπιστώνουμε και από το διάγραμμα των συγκεκριμένων συναισθημάτων του #hellass όπου τα συναισθήματα με φθίνουσα σειρά κατάταξης είναι: χαρά, φόβος, θυμός, αηδία, έκπληξη, λύπη.

Ένα άλλο παράδειγμα όπου η ανάλυση της πολικότητας ενός σημαντικού θέματος που υλοποιήσαμε στο προηγούμενο στάδιο έρχεται σε συνάφεια με τα συγκεκριμένα συναισθήματα που εκφράζονται από τους χρήστες, βλέπουμε στο #vouli. Για αυτή την περίπτωση, είχαμε εντοπίσει ότι η πολικότητα των tweet που αντιστοιχούν στο εν λόγω σημαντικό θέμα είναι αρνητική. Βλέπουμε ότι με αυτή τη διαπίστωση έρχεται να συμφωνήσει και η κατάταξη των συναισθημάτων των χρηστών η οποία είναι η εξής: φόβος, χαρά, θυμός, έκπληξη, αηδία, λύπη.

Τέλος, παρατηρούμε ότι το #survivor ενώ στην Συναισθηματική Ανάλυση παρουσιάζει ελαφρώς αρνητική πολικότητα, στο βήμα αυτό το κυρίαρχο συναίσθημα είναι η χαρά. Ο λόγος που συμβαίνει αυτό είναι ότι οι λέξεις με ουδέτερη πολικότητα, οι οποίες δηλαδή δεν συνεισφέρουν στην μέτρηση της πολικότητας, υπάρχει περίπτωση να έχουν συνεισφορά στα συναισθήματα που εξετάζουμε.

Περνώντας στην δεύτερη υπηρεσία που υλοποιήσαμε στην παρούσα διπλωματική εργασία, εντοπίσαμε αρχικά για κάθε δημοφιλές θέμα τις ταυτότητες των χρηστών που ασχολήθηκαν με αυτό στο διάστημα συλλογής των δεδομένων μας. Για την υπόλοιπη παρουσίαση της πειραματικής αξιολόγησης, δεν λάβαμε υπόψη τους χρήστες οι οποίοι έχουν κλειδωμένα τα προφίλ τους στο Twitter.

Για παράδειγμα, για το δημοφιλές θέμα #cyprus έχουν σχολιάσει 17 διαφορετικοί χρήστες των οποίων έχουμε συλλέξει τα 20 τελευταία tweets για τον κάθε ένα. Ψάχνοντας σε αυτά τα 20 tweets του κάθε χρήστη, εντοπίσαμε πόσα από αυτά ασχολούνται με το συγκεκριμένο θέμα και βγάζουμε ένα συμπέρασμα για την διάθεση που εκφράζουν οι χρήστες σε αυτά.

Ταυτότητα του κάθε χρήστη που ασχολείται με το θέμα #cyprus.	Πόσα από τα 20 tweets του ασχολούνται με το δημοφιλές θέμα #cyprus.	Πολικότητα που προκύπτει από τα σχετικά tweets κάθε χρήστη.
Id1	9	-2
Id2	20	-1
Id3	12	0
Id4	2	-2
Id5	13	-6
Id6	10	0
Id7	5	-1
Id8	20	0
Id9	2	-1
Id10	4	0
Id11	2	0
Id12	20	1
Id13	4	0
Id14	9	2
Id15	9	2
Id16	1	-1
Id17	6	0

Πίνακας 6 Τα tweets που εντοπισαμε για κάθε Id και ο\η πολικότητά τους.

Ένας τέτοιος πίνακας δημιουργείται για κάθε ένα από τα δέκα δημοφιλή θέματα. Επειδή ο στόχος μας είναι να δημιουργήσουμε δύο κατηγορίες προτάσεων φιλίας, μία για χρήστες με αρνητική άποψη πάνω σε ένα ζήτημα και μία για χρήστες με θετική άποψη, αφαιρούμε ότι πληροφορία έχουμε που μας δίνει πολικότητα 0.

Επίσης, σημαντικό ρόλο στη σειρά κατάταξης στις δύο αυτές κατηγορίες παίζει το πόσα tweets από τα τελευταία είκοσι δημοσιευμένα του χρήστη, ασχολούνται με το θέμα που

μας ενδιαφέρει. Για παράδειγμα, ανάμεσα στον χρήστη με ID1 και στον χρήστη με ID2 που έχουν και οι δύο αρνητική άποψη πάνω στο δημοφιλές θέμα #cyprus, θα προταθεί πρώτος ο χρήστης με ID2 επειδή η ενασχόληση του με το εν λόγω θέμα είναι εντονότερη. Ας μελετήσουμε τη γραφική παράσταση αναφορικά με τους χρήστες που ασχολήθηκαν με το θέμα #survivorgr.

Στο παρακάτω διάγραμμα παρουσιάζεται μία γραφική παράσταση των ταυτοτήτων των χρηστών σε συνάρτηση με το πόσα tweets έχει δημοσιεύσει ο καθένας, εντός των τελευταίων είκοσι δημοσιεύσεων του, σχετικά με το δημοφιλές θέμα #survivorgr.

Με το μπλε χρώμα αναπαριστούμε τους χρήστες που σχολίασαν θετικά σε σχέση με το #survivor, ενώ με κόκκινο τους χρήστες που σχολίασαν αρνητικά.

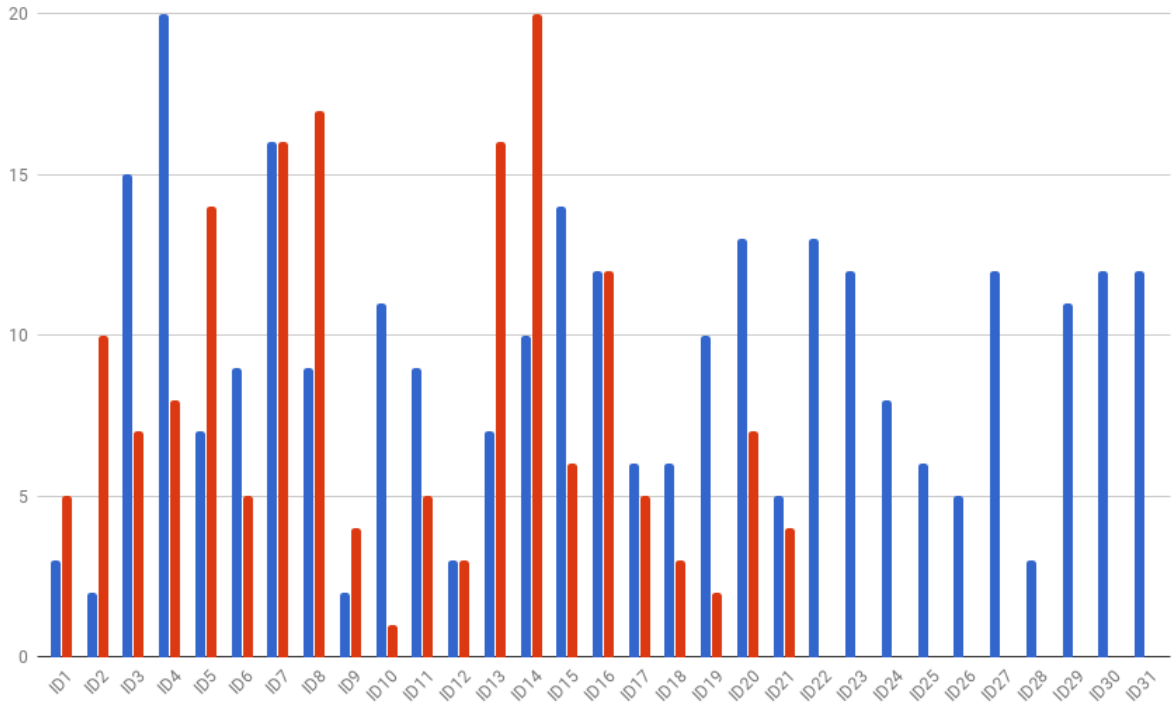

Figure 21 Χρήστες που σχολιάζουν θετικά ή αρνητικά κα σε πόα tweets.

Αρχικά, παρατηρούμε ότι από το σύνολο των 52 χρηστών οι 31 χρήστες σχολίασαν θετικά το θέμα #survivorgr και οι 21 το σχολίασαν αρνητικά. Ωστόσο, παραπάνω αναφέραμε πως το συγκεκριμένο θέμα έχει αρνητική πολικότητα στο σύνολο των χρηστών. Οδηγούμαστε λοιπόν στο συμπέρασμα ότι ενώ γενικά οι χρήστες που ασχολήθηκαν με

το εν λόγω θέμα το σχολίασαν θετικά, στην χρονική περίοδο που συλλέξαμε τα δεδομένα μας οι χρήστες εκφράστηκαν αρνητικά για το #survivor.

Συνοψίζοντας, μπορούμε πλέον να σχηματίσουμε ομάδες από χρήστες με κατάλληλα χαρακτηριστικά και να τους προτείνουμε σε άλλους χρήστες με κοινά ενδιαφέροντα και απόψεις. Για παράδειγμα, για το θέμα #cyprus έχουμε εντοπίσει τις παρακάτω κατηγορίες χρηστών.

Θετική διάθεση απέναντι στο #cyprus		Αρνητική διάθεση απέναντι στο #cyprus	
IDs	Πλήθος tweets	IDs	Πλήθος tweets
1127205486	20	814765508532781056	20
4217350108	9	53346239	13
4303523654	9	219110145	9
		4479760996	5
		223262701	2
		87170286	2
		37820726	1

Πίνακας 7 Αποτελέσματα για το #cyprus.

Επομένως, έχοντας ολοκληρώσει και αυτό το στάδιο είμαστε πλέον σε θέση να προτείνουμε σε κάποιον που ασχολείται με το #cyprus να αναπτύξει διαδικτυακή φιλία με κάποιον από τους παραπάνω χρήστες ξεκινώντας από την κορυφή της λίστας που αντιστοιχεί στο συναίσθημα του εν λόγω χρήστη.

4.3 Παρουσίαση δεύτερου πειράματος

Σε αυτό το στάδιο θεωρήσαμε σκόπιμο να παρουσιάσουμε περιληπτικά και κάποια από τα στοιχεία που προκύψανε από μία λίγο παλαιότερη διεξαγωγή πειράματος ώστε να τονίσουμε την παραμονή κάποιων θεμάτων στη λίστα με τα δέκα σημαντικότερα θέματα συζήτησης στο κοινωνικό δίκτυο Twitter.Τα δεδομένα μας για αυτό το πείραμα παρουσιάζονται στον ακόλουθο πίνακα..

Σύνολο των tweets	4686
Tweets που περιέχουν hashtags	904
Tweets που δεν περιέχουν hashtags	3782
Tweets που κατηγοριοποιήσαμε εμείς σε κάποιο hashtag	318

Πίνακας 8 Δεδομένα δεύτερου πειράματος.

Στο παρακάτω διάγραμμα παρουσιάζουμε τα δέκα δημοφιλέστερα θέματα για αυτό το πείραμα.

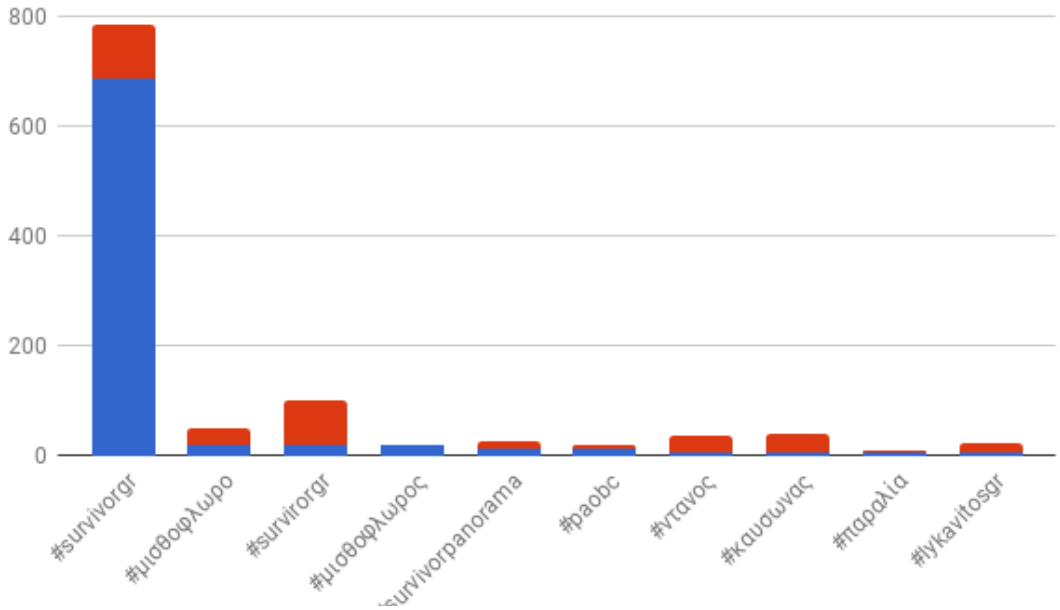

Figure 22 Hashtag σε συνδυασμό με πλήθος των tweets που τα περιέχουν.

Για ακόμα μία φορά με μπλε χρώμα αναπαριστούμε το πλήθος των tweets που αναφέρεται σε κάθε ένα από τα hashtags εξ' αρχής και με κόκκινο χρώμα το πλήθος των tweets που εμείς κατηγοριοποιήσαμε.

Παρατηρούμε πως υπάρχουν 3 σημαντικά θέματα κοινά με το προηγούμενο πείραμά μας. Το γεγονός αυτό μας δείχνει ότι υπάρχουν κάποια θέματα που ενδιαφέρουν τους χρήστες του twitter για μεγάλο χρονικό διάστημα ενώ ανοίγει και καινούργιες προοπτικές ανάλυσης. Επιπλέον, παρατηρώντας το παραπάνω διάγραμμα συνειδητοποιούμε ότι υπάρχουν 2 hashtags με ένα γράμμα διαφορά το οποίο οφείλεται σε ορθογραφικό λάθος. Το φαινόμενο των ορθογραφικών λαθών είναι ένας τομέας που χρήζει περαιτέρω μελέτης σε ότι αφορά κείμενα που συλλέγονται από τα κοινωνικά δίκτυα όπου οι χρήστες πολλές φορές γράφουν βιαστικά τις δημοσιεύσεις τους.

Στο παρακάτω ραβδόγραμμα παρουσιάζονται τα αποτελέσματα της Συναισθηματικής Ανάλυσης για το συγκεκριμένο πείραμα.

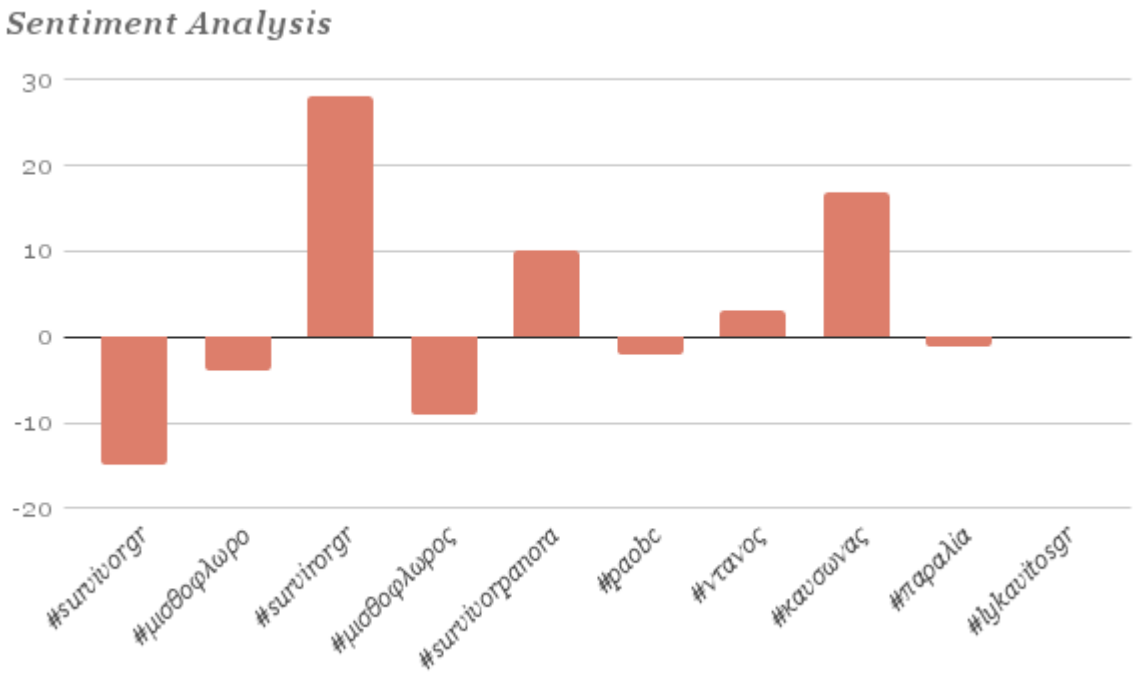

Figure 23 Αποτελέσματα Συναισθηματικής Ανάλυσης για το δεύτερο πείραμα.

Μπορούμε να δούμε πως για τα θέματα τα οποία διατηρούνται και στα δύο πειράματα, η πολικότητα των συναισθημάτων των χρηστών είναι διαφορετική. Συμπεραίνουμε επομένως ότι η κοινή γνώμη δεν διατηρεί συγκεκριμένη άποψη για τα εν λόγω θέματα.

Τέλος, παραθέτουμε και ένα παράδειγμα όπου εντοπίζουμε τους κατάλληλους χρήστες για να προτείνουμε νέους δεσμούς διαδικτυακής φιλίας. Στο συγκεκριμένο παράδειγμα ασχολούμαστε με το δημοφιλές θέμα #paobc.

Θετική διάθεση απέναντι στο #paobc		Αρνητική διάθεση απέναντι στο #paobc	
IDs	Πλήθος tweets	IDs	Πλήθος tweets
773531453615398913	5	589966757	8
397184989	1	2980341155	3
		171568558	2

Πίνακας 9 Αποτελέσματα για το #paobc.

Είμαστε πλέον σε θέση οδηγούμενοι από τον παραπάνω πίνακα να κάνουμε τις αντίστοιχες προτάσεις σε χρήστες που ασχολούνται με το δημοφιλές θέμα #paobc.

5.Σχετική Έρευνα

Η Επιστήμη των Δεδομένων (Data Science), αποτελεί στη σύγχρονη εποχή έναν από τους πιο δημοφιλείς τομείς τόσο για την ακαδημαϊκή έρευνα, όσο και για τη βιομηχανία. Το γεγονός αυτό έχει σαν αποτέλεσμα τη διαρκή ανάπτυξη αλγορίθμων και συστημάτων γύρω από τον τομέα των δεδομένων. Περικλείει πολλά επιμέρους επιστημονικά πεδία όπως την ανάλυση δεδομένων, την συναισθηματική ανάλυση και την εξόρυξη δεδομένων. Παρακάτω αναφέρονται μερικές έρευνες που έχουν διεξαχθεί ως τώρα που έχουν να κάνουν με την Επιστήμη των Δεδομένων και μας παρείχαν τις βάσεις για την ολοκλήρωση της παρούσας εργασίας.

Μία ενδιαφέρουσα προσέγγιση για ομαδοποίηση των λέξεων μετά την αφαίρεση των καταλήξεων τους περιγράφεται από τους Detorakis και Tambouratzis (2009). Η έρευνα αυτή αποσκοπούσε στην συσταδοποίηση (clustering) Ελληνικών λέξεων μετά την αφαίρεση των καταλήξεων τους, οι οποίες έχουν κάποια κοινά χαρακτηριστικά. Η κάθε ομάδα (cluster) από λέξεις που προκύπτει περιέχει μόνο λέξεις που βρίσκονταν για παράδειγμα στην παθητική φωνή πριν την αφαίρεση της κατάληξης τους. Το αποτέλεσμα ήταν η δημιουργία ενός μορφολογικού λεξικού στα Ελληνικά.

Από τους ερευνητές Pang και Lee (2002) οι οποίοι θεωρούνται πρωτοπόροι στο πεδίο της ανάλυσης συναισθήματος, μπορούμε να κατατοπιστούμε σχετικά με τα πειραματικά αποτελέσματα της κατηγοριοποίησης (classification) μέσω της εφαρμογής μίας σειράς από αλγορίθμους ταξινόμησης. Πρόκειται για τον τον Απλοϊκό Ταξινομητή κατά Bayes (Naive Bayes Classification), τον Ταξινομητή Μέγιστης Εντροπίας (maximum entropy classification) και τις Μηχανές Διανυσματικής Στήριξης (Support Vector Machines (SVM)).

Μία υβριδική προσέγγιση συναισθηματικής ανάλυσης δεδομένων πραγματοποιήθηκε από τους Zhang et al. (2011). Αρχικά, εφαρμόσανε ανάλυση συναισθήματος στο Twitter για συγκεκριμένες οντότητες, όπως για παράδειγμα για προϊόντα, ανθρώπους και οργανισμούς. Εν συνεχεία, εκπαίδευσαν έναν ταξινομητή ώστε να αναθέσει την κατάλληλα πολικότητα στις συγκεκριμένες οντότητες.

Μία ακόμη προσέγγιση ανάλυσης δεδομένων παρουσιάζεται από τους Hu and Liu (2004) οι οποίοι ακολουθούν, όπως και εμείς στην εργασία αυτή, μία προσέγγιση συναισθηματικής ανάλυσης βασισμένη σε συναισθηματικό λεξικό για τον εντοπισμό της πολικότητας και του συναισθήματος ενός κειμένου.

Οι Kolchyna et al. (2015) μελετήσαν και τις δύο προσεγγίσεις της ανάλυσης συναισθήματος. Ασχολήθηκαν τόσο με μεθόδους μηχανικής μάθησης, όσο και με μεθόδους που βασίζονται σε λεξικό. Μάλιστα, η έρευνα τους πραγματοποιήθηκε πάνω σε tweets και ο στόχος τους ήταν να τα κατηγοριοποιήσουν. Απέδειξαν μέσα από τη χρήση πολλών συναισθηματικών λεξικών, πως η χρήση συναισθηματικού λεξικού το οποίο περιέχει φράσεις αργκό και ανεπίσημης γλώσσας δίνει καλύτερα αποτελέσματα όσον αφορά τον τομέα της συναισθηματικής ανάλυσης με χρήση λεξικού. Στη συνέχεια, κατέληξαν στο γεγονός ότι οι αλγόριθμοι μηχανικής μάθησης δίνουν καλύτερα αποτελέσματα σε σύγκριση με τη χρήση συναισθηματικού λεξικού. Με βάση αυτές τις παρατηρήσεις, πειραματίστηκαν εφαρμόζοντας έναν συνδυασμό των παραπάνω αλγορίθμων. Κατέληξαν στο συμπέρασμα ότι ο συνδυασμός αυτός είχε τα καλύτερα αποτελέσματα.

Μία ακόμη έρευνα αναφορικά με την κατηγοριοποίηση μηνυμάτων στο Twitter διεξάχθηκε από τους Alec Go et al. (2009). Παρουσιάζουν τα αποτελέσματα της προσπάθειας τους για αυτοματοποίηση της κατηγοριοποίησης δεδομένων από το Twitter χρησιμοποιώντας αλγορίθμους μηχανικής μάθησης. Στόχος τους είναι να επιτύχουν όσο το δυνατό υψηλότερα ποσοστά σωστής ταξινόμησης των tweets με βάση την πολικότητα τους. Επιπλέον, αναλύουν όλα τα βήματα προεπεξεργασίας των tweets που εντόπισαν ότι βοηθούν στην επίτευξη καλύτερων αποτελεσμάτων ορθής ταξινόμησης.

Έχει ιδιαίτερο νόημα να αναφερθούμε στη δουλειά των συγγραφέων Maite Taboada et al. (2011), οι οποίοι αναλύουν εκτενώς όλη τη διαδικασία μέσω της οποίας μια προσέγγιση βασισμένη σε συναισθηματικό λεξικό μας οδηγεί στον εντοπισμό των συναισθημάτων ενός κειμένου. Ταυτόχρονα, παρουσιάζουν τη διαδικασία κατασκευής ενός λεξικού με βάση το οποίο υλοποιούνται τα παραπάνω.

Οι Cambria et al. (2013), αναλύουν τη χρησιμότητα της ανάπτυξης του τομέα της ανάλυσης συναισθημάτων και εξόρυξης γνώμης στη σύγχρονη εποχή. Αναφέρονται σε

αρκετές προσεγγίσεις ανάλυσης συναισθήματος. Μία από αυτές είναι η λεξικολογική προσέγγιση που εφαρμόζεται και στην παρούσα εργασία. Άλλες είναι η στατιστική προσέγγιση και η προσπάθεια ανάλυσης συναισθήματος με βάση λέξεις–κλειδιά. Καταλήγουν στη σπουδαιότητα της συνεργασίας της Επιστήμης των Υπολογιστών με τον Επιστημονικό τομέα μελέτης των συναισθημάτων με σκοπό την ανάπτυξη όσο το δυνατό πιο έξυπνων αλγορίθμων εξόρυξης γνώμης.

Τέλος, μία πιο στοχευμένη προσέγγιση του τομέα της ανάλυσης συναισθήματος πραγματοποίησαν οι Chamlertwat et al. (2011). Συγκεκριμένα, προτείνουν ένα σύστημα, το Micro-blog Sentiment Analysis System (MSAS), το οποίο είναι σε θέση να βγάλει συμπεράσματα για τη γνώμη των καταναλωτών προϊόντων από το Twitter βασιζόμενο στην Ανάλυση Συναισθήματος. Το σύστημά τους χωρίζεται σε μερικές επιμέρους λειτουργίες: τη συλλογή tweets, την επιλογή μόνο αυτών που εκφράζουν κάποια άποψη, τον εντοπισμός της πολικότητας τους και την οπτικοποίηση απόψεων των χρηστών του Twitter για ένα προϊόν. Πραγματοποίησαν πειράματα σχετικά με τη γνώμη των καταναλωτών για κάποια smartphones, τα αποτελέσματα των οποίων εγκρίθηκαν από ειδικούς στον τομέα των smartphones.

6.Συμπεράσματα

Στην παρούσα Διπλωματική Εργασία ερευνήσαμε τον τομέα της Ανάλυσης Δεδομένων και μελετήσαμε την σπουδαιότητα της Συναισθηματικής Ανάλυσης δεδομένων. Αφού διαπιστώσαμε τη χρησιμότητα των δεδομένων που προέρχονται από τα κοινωνικά δίκτυα, εφαρμόσαμε Ανάλυση Συναισθήματος στο κοινωνικό δίκτυο Twitter.

Καταφέραμε να δημιουργήσουμε ένα σύστημα το οποίο παρέχει δύο υπηρεσίες. Μέσω της πρώτης υπηρεσίας, εντοπίζουμε τα δέκα δημοφιλέστερα θέματα της χρονικής περιόδου που συλλέγουμε δεδομένα από το Twitter και τα αναλύουμε τόσο ως προς την πολικότητα (θετικά, αρνητικά, ουδέτερα), όσο και ως προς το συναίσθημα που εκφράζουν (θυμός, αηδία, φόβος, χαρά, λύπη και έκπληξη). Με την υλοποίηση της δεύτερης υπηρεσίας, αφού εντοπίσουμε τους χρήστες που αναφέρονται σε κάθε ένα από τα δέκα δημοφιλέστερα θέματα, αναλύουμε τα προφίλ τους ώστε να δούμε σε τι βαθμό ασχολούνται με το συγκριμένο θέμα και πια είναι η άποψη τους για αυτό. Με αυτόν τον τρόπο δημιουργούμε τη δυνατότητα στοχευμένης πρότασης φιλίας μεταξύ χρηστών που ενδιαφέρονται για το ίδιο θέμα και έχουν την ίδια άποψη για αυτό.

Επιπλέον, παρατηρήσαμε ένα θετικό της προσέγγισής μας που έχει να κάνει με το γεγονός ότι δίνουμε την δυνατότητα για ανίχνευση σημαντικών θεμάτων και εξελίξεων σε συγκεκριμένο χρόνο. Συνεπώς, μπορούμε έγκαιρα να εντοπίσουμε ένα σημαντικό θέμα που άρχισε πρόσφατα να διαδίδεται.

Μέσα από την υλοποίηση του παραπάνω συστήματος καταλήξαμε στο συμπέρασμα ότι η δυναμική που προσφέρει η απόκτηση γνώσης μέσα από τη διαδικασία την εξόρυξης και στη συνέχεια της ανάλυσης δεδομένων έχει τεράστιες διαστάσεις. Η επεξεργασία πληροφορίας που πηγάζει από τα μέσα κοινωνικής δικτύωσης, αν γίνεται με φαντασία και δημιουργικότητα έχει τη δυνατότητα να γεννήσει ενδιαφέρουσα γνώση.

6.1 Μελλοντική Έρευνα

Το σύστημα που περιγράψαμε παραπάνω μπορεί να επεκταθεί ποικιλοτρόπως. Μία επέκταση που αφορά το σύστημα στο σύνολο του είναι η χρήση βάσεων δεδομένων για την αποθήκευση των tweets που προκύπτουν από τα διάφορα στάδια επεξεργασίας. Μέσω της χρήσης βάσεων δεδομένων θα μπορούμε να αποθηκεύουμε περισσότερα tweets χωρίς να παίρνει πολύ χρόνο η επεξεργασία τους.

Μία άλλη επέκταση του συστήματος μας έχει να κάνει με την δημιουργία μίας εφαρμογής που θα ενσωματώνει τις υπηρεσίες που υλοποιήσαμε σε αυτή την εργασία. Έτσι, θα δίνεται η δυνατότητα σε όποιον το επιθυμεί να αποκτήσει μέσω μίας απλής εφαρμογής πρόσβαση στις υπηρεσίες που παρέχει το σύστημα μας.

Τέλος, όπως σε όλους τους αλγορίθμους έτσι και στους δικούς μας, υπάρχουν πάντα περιθώρια βελτίωσης. Επομένως, ένα από τα πεδία στα οποία θα μπορούσε η κινηθεί η μελλοντική έρευνα είναι η αλγοριθμική βελτίωση της υλοποίησης μας.

Αναφορές-Βιβλιογραφία

Tsakalidis, A., Papadopoulos, S. & Kompatsiaris, I. (2014). An Ensemble Model for Cross-Domain Polarity Classification on Twitter. In B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali & Y. Zhang (eds.), *WISE (2),* Springer, pages 168-177

Go, A., Bhayani, R. & Huang, L. (2009). Twitter Sentiment Classification using Distant Supervision. Processing, pages 1-6

Esuli, A. & Sebastiani, F. (2006). SentiWordNet: a high-coverage lexical resource for opinion mining. Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR).

Paul Ekman. (2003). Emotions Revealed. Times Books.

Cambria, E., Schuller, B., Xia, Y. & Havasi, C. (2013). New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems, volume 28, pages 15-21

Goldsmith, J. A. (2001). Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics, volume 27, pages 153-198

Rajaraman, A., Leskovec, J., Ullman, J. D. (2014). Mining Massive Datasets. Cambridge University Press

T.Z. Kalamboukis. (1995). Suffix stripping with modern Greek. MCB UP Ltd

Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, Bing Liu. (2011). Combining Lexicon-based and Learning-based Methods for Twitter. HP Laboratories, Technical Report HPL-2011

Liu, B. (2012). Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, volume 5, pages 1-167

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, volume 2, pages 1-135

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. D. & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, volume 37, pages 267-307

Giatsoglou, M., Vozalis, M. G., Diamantaras, K. I., Vakali, A., Sarigiannidis, G. & Chatzisavvas, K. C. (2017). Sentiment analysis leveraging emotions and word embeddings. *Expert Syst. Appl.*, pages 214-224

M.F. Porter, (1980). An algorithm for suffix stripping. Program, issue: 3, volume. 14, pages 130-137

Hu, M. & Liu, B. (2004). Mining and Summarizing Customer Reviews. pages 168-177

Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O'Reilly Media Inc

Ntais, G. (2006). Development of a stemmer for the Greek language. MSc Thesis, Department of Computer and Systems Sciences, Stockholm University / Royal Institute of Technology.

Olga Kolchyna, Tharsis T. P. Souza, Philip Treleaven, Tomaso Aste. (2015). Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination. Handbook of Sentiment Analysis in Finance. Mitra, G. and Yu, X. (Eds.)

Pang, B., Lee, L. & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification Using Machine Learning Techniques, emnlp2002, pages 79-86

Ngoc, P. T. & Yoo, M. (2014). The lexicon-based sentiment analysis for fan page ranking in Facebook. *ICOIN,* pages 444-448

Sergios Theodoridis, Konstantinos Koutroumbas. (2012). Pattern Recognition. Elsevier Inc.

Manning, C. D., Surdeanu, M., Baue,r J., Finkel, J., Bethard, S. J., McClosky D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55-60.

Tambouratzis, G, Carayannis, G. (2001). Automatic Corpora-based Stemming in Greek, LLC, volume 4, pages 445-466.

Twitter Inc. (2017). Twitter Developer Documentation. https://dev.twitter.com/overview/api

Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T. & Haruechaiyasak, C. (2012). Discovering Consumer Insight from Twitter via Sentiment Analysis, *J. UCS*, volume 18, pages 973-992.

Z. Detorakis, George Tambouratzis. (2012). Clustering Techniques for Establishing Inflectionally Similar Groups of Stems. International Journal of Computer Information Systems and Industrial Management Applications, volume 4, pages 219-227

Καλαπόδης, Α. (2013). Αλγόριθμοι Αποδοτικής Επιλογής Χαρακτηριστικών για Κατηγοριοποίηση Κειμένου στην Ελληνική Γλώσσα. Πτυχιακή Εργασία. Ελληνικό Ανοιχτό Πανεπιστήμιο

Παράρτημα Α

```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

#########################################################################
#Οι μεταβλητές που περιέχουν τα στοιχεία των χρηστών για σύνδεση με το
#Twitter API
#########################################################################

access_token = "..."
access_token_secret = "..."
consumer_key = "..."
consumer_secret = "..."

class StdOutListener(StreamListener):
    def on_data(self, data):
        print data
        return True

    def on_error(self, status):
        print status

if __name__ == '__main__':
#########################################################################
#Εξακρίβωση στοιχείων του χρήστη που θέλει να συνδεθεί στο Twitter
#Streaming API
#########################################################################

    l = StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)
    stream = Stream(auth, l)

    t = "\xCE\xBA\xCE\xB1\xCE\xB9".decode("utf-8")                    #και
    t0 = "\xCE\xB7".decode("utf-8")                                   #η
    t1 = "\xCE\xAE\x0A".decode("utf-8")                               #ή
    t2 = "\xCE\xBF".decode("utf-8")                                   #ο
    t3 = "\xCF\x84\xCE\xBF".decode("utf-8")                           #το
    t4 = "\xCF\x84\xCE\xB7".decode("utf-8")                           #τη
    t5 = "\xCF\x84\xCE\xBF\xCE\xBD".decode("utf-8")                   #τον
    t6 = "\xCF\x84\xCE\xB7\xCE\xBD\x0A".decode("utf-8")               #την
    ....

#########################################################################
#Φιλτράρισμα των δεδομένων που θέλουμε να συλλέξουμε με βάση τις παραπάνω
#λέξεις-κλειδιά
#########################################################################
    i = stream.filter(languages = ["el"],track=[t, t1, t2, t3, t4, t5, t6,
    t7, t8, t9, t10, t11, t12, t13, t14, t15, t16, t17, t18, t19, t20,
    t21, t22, t23, t24, t25, t26, t27, t28, t29, t30, t31, t32, t33, t34])
    print i
```

Παράρτημα Β

```
# -*- coding: utf-8 -*-

import json
import pandas as pd
import re
import nltk
import stemimport
from nltk.tokenize import TweetTokenizer
from nltk.probability import FreqDist
from collections import Counter
from pyexcel_ods import get_data

def removeStopwords(wordlist, flag):
    return [w for w in wordlist if w.encode("utf-8") not in flag]

def wordListToFreqDict(wordlist):
    wordfreq = [wordlist.count(p) for p in wordlist]
    return dict(zip(wordlist,wordfreq))

def sortFreqDict(freqdict):
    aux = [(freqdict[key], key) for key in freqdict]
    aux.sort()
    aux.reverse()
    return aux

#########################
#Ανάγνωση Δεδομένων
#########################

tweets_data_path= '.../tweets.txt'

tweets_data = []
tweets_file = open(tweets_data_path, "r")
for line in tweets_file:
      try:
            tweet = json.loads(line)
            tweets_data.append(tweet)
      except:
            continue

tweets = pd.DataFrame()
clean_tweets = []
tokens = []
clean_tokens = []
final_tokens = []
diction = {}
final_tokens_upper = []
final_tokens_upper_encoded = []
result = []
```

```
stopwords = ['ο', 'η', 'το', 'τον', 'τη', 'την', 'στο', 'στον', 'στη',
'στην', 'στα', 'στις','στους', 'οι', 'τα', '.', ',', '!', '(', ')', ':',
'rt', '«', '»', '"', ';', '?', '[', ']', 'τι', 'του', 'της', 'των', 'και',
'να', 'που', 'θα', 'ότι', 'εμείς', 'εσείς', 'αυτό', 'αυτά', 'αυτός',
'αυτή', 'αυτές', 'αυτοί', 'για', 'σε', 'με', 'πως', 'που', 'πότε', 'αν',
'ως', 'απ', 'τ', 'κ', '-', 'κι', 'μία', 'μια','τους', 'δεν', 'μα', 'μας',
'μου', 'δε', 'άλλα','αλλά', 'άλλος', 'άλλη', 'άλλο', 'ένα', 'από',
'γιατί', 'ένας', 'σου', 'σας', 'τις', 'είναι', '...', '..', '…', 'ρε',
'μόνο','πώς','χα','χ','κάτι', '|', 'όταν', '\'', 'οσο', 'ειναι',
'ποιος','σ', 'ν', 'τν','μ','μη','π','ν','', 'έχει']

tknzr = TweetTokenizer()

emoji_pattern = re.compile("["u"\U0001F600-\U0001F64F"u"\U0001F300-
\U0001F5FF"u"\U0001F680-\U0001F6FF"u"\U0001F1E0-\U0001F1FF""]+",
flags=re.UNICODE)

tweets['text'] = map(lambda tweet: tweet['text'], tweets_data)

for i in tweets['text']:
      stringwithout = i.lower()
      stringwithout = re.sub(r'http\S+', '', stringwithout)
      stringwithout = re.sub(r'@\S+', '', stringwithout)
      stringwithout = (emoji_pattern.sub(r'', stringwithout))
      stringwithout = re.sub(r'\d+', '', stringwithout)
      clean_tweets.append(stringwithout)

for j in clean_tweets:
      p = tknzr.tokenize(j)
      tokens.append(p)

for j in tokens:
      clean = removeStopwords(j, stopwords)
      clean_tokens.append(clean)

for i in clean_tokens:
      for j in i:
            final_tokens.append(j)

#######################################################
#εκτύπωση των tokens αφού αφαιρέσαμε τα stopwords
#######################################################

file = open('tokens.txt','w')

for i in final_tokens:
      file.write(i.encode("utf-8"))
      file.write("\n")

file.close()

#########################################################
#βρίσκω τα hashtags που υπάρχουν στα tweets μου
#########################################################
```

```
hashtaglist = []
hashtag_final = []
hashtag_final2 = []

tweets['entities'] = map(lambda tweet: tweet['entities']['hashtags'],
tweets_data)
for i in tweets['entities']:
      for j in range(0,len(i)):
            hashtaglist.append(i[j]['text'])

for i in hashtaglist:
      if i not in hashtag_final2:
            hashtag_final2.append(i)

for i in hashtag_final2:
      hashtag_final.append('#'+i)

for i in hashtag_final:
      print i.encode("utf-8")
print "ta hashtags einai" + str(len(hashtag_final))

#######################################################
#βρίσκω τα tweets που δεν περιέχουν τα hashtags
#######################################################
fl = 0
tweets_without_hashtags = []

for i in clean_tokens:
      for j in i:
            for m in hashtag_final:
                  if m in j:
                        fl = 1
                        break;
            if fl == 1:
                  break;
      if fl == 0:
            tweets_without_hashtags.append(i)
      fl = 0
print "ta tweets xwris hashtag einai"
print len(tweets_without_hashtags)

#########################################################
#βρίσκω τα tweets που περιέχουν τα hashtags μου
#########################################################

tweets_with_hash = []
list_of_lists_tweets_with_hash = []
all_diction = {}
sum_var = 0

for m in hashtag_final:
      all_diction[m] = []

for m in hashtag_final:
      for i in clean_tokens:
```

```
            for j in i:
                  if m in j:
                        all_diction[m].append(i)
                        tweets_with_hash.append(i)
                        break;

      sum_var = sum_var +len(tweets_with_hash)
      list_of_lists_tweets_with_hash.append(tweets_with_hash)
      tweets_with_hash = []

print "ta tweets me hashtag einai"
print sum_var

all_tokens = []
list_of_all_tokens = []
list_length = []
hashtags_with_frequency = {}
diction_list =[]
lista1 =[]
lista2 = []

for i in list_of_lists_tweets_with_hash:
      for j in i:
            for k in j:
                  all_tokens.append(k)

      list_of_all_tokens.append(all_tokens)
      all_tokens = []

for i in list_of_lists_tweets_with_hash:
      list_length.append(len(i))

hashtags_with_frequency = dict(zip(hashtag_final,list_length))

diction2 = sortFreqDict(hashtags_with_frequency)

for i in list_of_all_tokens:
      for j in i:
            stringwithout = re.sub(r'#\S+', "", j)
            if stringwithout != "":
                  lista1.append(stringwithout)
      lista2.append(lista1)
      lista1 = []

for i in lista2:
      diction = wordListToFreqDict(i)
      diction3 = sortFreqDict(diction)
      diction_list.append(diction3)

############################################################
#βρίσκω τα δημοφιλέστερα 10 hashtags
############################################################

top_ten_hashtags = []
for x in diction2:
```

```
    top_ten_hashtags.append(x[1])

top_ten_hashtags = top_ten_hashtags[0:10]

file = open('top_ten_hashtags.txt','w')

for i in top_ten_hashtags:
      stringwithout = re.sub(r'#([^\s]+)', r'\1', i)
      file.write(i.encode("utf-8"))
      file.write("\n")

file.close()

for i in top_ten_hashtags:
      print i.encode("utf-8") +' :'+ str(len(all_diction[i]))

#############################################################################
#βρίσκω τις 3 πιο δημοφιλείς λέξεις για κάθε hashtag
#############################################################################

list_col = []
listaaaa = []
every_hashtags_tokens ={}
every_top_ten_hashtags_3tokens = {}

for x in diction_list:
      for i in x:
            list_col.append(i[1])
      listaaaa.append(list_col)
      list_col =[]

every_hashtags_tokens = dict(zip(hashtag_final,listaaaa))

for i in top_ten_hashtags:
      if len(every_hashtags_tokens[i]) > 3:
            every_top_ten_hashtags_3tokens[i] =
every_hashtags_tokens[i][0:3]
            print i.encode("utf-8")
            for k in range(0,3):
                  print every_hashtags_tokens[i][k].encode("utf-8")
      else:
            every_top_ten_hashtags_3tokens[i] = every_hashtags_tokens[i]
            print i.encode("utf-8")
            for k in range(0,len(every_hashtags_tokens[i])):
                  print every_hashtags_tokens[i][k].encode("utf-8")

file = open('every_hashtags_tokens.txt','w')

file.write(json.dumps(every_top_ten_hashtags_3tokens))
file.write("\n")
file.close()

countr_list = [0] * 10
countr_list2 = [0] * 10
tweet_categ = 0
```

```
for_every_hash ={}

for_every_hash = dict(zip(top_ten_hashtags,countr_list2))
for j in tweets_without_hashtags:
      c = 0
      for k in top_ten_hashtags:
            for l in every_top_ten_hashtags_3tokens[k]:
                  if l in j:
                        countr_list[c] = countr_list[c] + 1
            c = c + 1
      if max(countr_list) != 0:
            res = top_ten_hashtags[countr_list.index(max(countr_list))]
            tweet_categ = tweet_categ + 1
            for_every_hash[res] = for_every_hash[res] + 1
            all_diction[res].append(j)
      countr_list = [0] * 10

print "Tweets categorized by us"
print tweet_categ
print for_every_hash

##################################
#Ανάλυση Συναισθήματος
##################################

tokens_single = []
clean_tokens_single = []
sentiment_dict = {}
keys = []
values = []
stem_keys = []
l = []
upper_tweet = []
upper_tweets = []
lista = []
newLista = []
stem_tweet = []
stem_tweets = []
lista_hashtag = []
stem_hashtags = []
total_hashtag_tweets = []
newLista2 =[]
new_stem_clean_tweets2 = []

data = get_data("lexeis.ods")

for i in range(0,2314):
      keys.append(data["Sheet1"][i][0])
      values.append(data["Sheet1"][i][1])

print "Stem keys"

for i in keys:
      stem_keys.append(stemimport.stem(i.encode("utf-8")))
```

```
sentiment_dict = dict(zip(stem_keys,values))

for i in top_ten_hashtags:
      for j in all_diction[i]:
            for k in j:
                  l.append(k.upper().encode("utf-8"))
            upper_tweets.append(l)
            l = []

      total_hashtag_tweets.append(upper_tweets)
      upper_tweets = []

print "replace start"
lista_tweet = []

for i in total_hashtag_tweets:
      for j in i:
            for p in j:
                  if "Έ" in p:
                        lista.append(p.replace("Έ","Ε"))
                  elif "Ό" in p:
                        lista.append(p.replace("Ό","Ο"))
                  elif "Ί" in p:
                        lista.append(p.replace("Ί","Ι"))
                  elif "Ώ" in p:
                        lista.append(p.replace("Ώ","Ω"))
                  elif "Ύ" in p:
                        lista.append(p.replace("Ύ","Υ"))
                  elif "Ά" in p:
                        lista.append(p.replace("Ά","Α"))
                  elif "Ή" in p:
                        lista.append(p.replace("Ή","Η"))
                  else:
                        lista.append(p)
            lista_tweet.append(lista)
            lista = []
      lista_hashtag.append(lista_tweet)
      lista_tweet = []

print "stem ta tweets"

for i in lista_hashtag:
      for j in i:
            for k in j:
                  stem_tweet.append(stemimport.stem(k))
            stem_tweets.append(stem_tweet)
            stem_tweet = []

      stem_hashtags.append(stem_tweets)
      stem_tweets = []

print "Sentiment Analysis"

total_counter = 0
counter = 0
```

```
apotelesmata = [0] * 10
n = 0

for i in stem_hashtags:
      for j in i:
            for k in j:
                  for m in stem_keys:
                        if k == m :
                              counter = counter + int(sentiment_dict[m])
                              break;
            total_counter = total_counter + counter
            counter = 0
      apotelesmata[n] = total_counter
      total_counter = 0
      n = n + 1
print "the top ten hashtags with polarity results"
print dict(zip(top_ten_hashtags,apotelesmata))

########################################
#euresi sunaisthimatwn
########################################

anger = []
disgust = []
fear = []
happiness = []
sadness =[]
surprise = []
emotion_list = []
second_emotion_list = []
final_emotion_list = []
sum_emot = []
total_emotion_list = []
token_counter = []

for i in range(0,2314):
      anger.append(data["Sheet1"][i][2])
      disgust.append(data["Sheet1"][i][3])
      fear.append(data["Sheet1"][i][4])
      happiness.append(data["Sheet1"][i][5])
      sadness.append(data["Sheet1"][i][6])
      surprise.append(data["Sheet1"][i][7])

emotion_list = zip(anger, disgust, fear, happiness, sadness, surprise)

print "Start emotions"
s = 0

for i in stem_hashtags:
      for j in i:
            total_counter = total_counter + len(j)
            for k in j:
                  for m in stem_keys:
                        if k == m :
```

```
            second_emotion_list.append(emotion_list[stem_keys.index(m)])
                                    break;

                  for a in range(0,6):
                        for b in second_emotion_list:
                              s = s + b[a]
                        sum_emot.append(s)
                        s = 0
                  final_emotion_list.append(sum_emot)

                  sum_emot = []
                  second_emotion_list = []

            for a in range(0,6):
                  for b in final_emotion_list:
                        s = s + b[a]
                  sum_emot.append(s)
                  s = 0
            token_counter.append(total_counter)
            total_counter = 0
            total_emotion_list.append(sum_emot)
            final_emotion_list = []
            sum_emot = []
            s =0

      print dict(zip(top_ten_hashtags,total_emotion_list))
      print "top_ten_hashtags"
      print top_ten_hashtags
      print "total_emotion_list"
      print total_emotion_list
```

Παράρτημα Γ

```
# -*- coding: utf-8 -*-

import json
import pandas as pd
import twitter_user_streaming

############################
#Ανάγνωση Δεδομένων
############################

tweets_data_path = '.../tweets.txt'
top_hashtags_data_path = '.../top_ten_hashtags.txt'
tweets_data = []
tweets_file = open(tweets_data_path, "r")

for line in tweets_file:
      try:
            tweet = json.loads(line)
            tweets_data.append(tweet)
      except:
            continue

tweets = pd.DataFrame()
tweets_file.close()

with open('top_ten_hashtags.txt') as f:
    content = f.readlines()
top_hashtags= [x.strip() for x in content]

tweets['entities'] = map(lambda tweet: tweet['entities']['hashtags'],
tweets_data)
tweets['id'] = map(lambda tweet: tweet['user']['id'], tweets_data)

ids = []
total_ids = [[],[],[],[],[],[],[],[],[],[]]
total_ids_tweets = []
total_emotion_list = []
top_hashtags_2 = []
positive = []
negative = []
for m in tweets['id']:
      ids.append(m)

c = 0

for i in tweets['entities']:
      for j in range(0,len(i)):
            temp = '#' + i[j]['text'].upper()
            for k in range(0,len(top_hashtags)):
                  if  temp.encode("utf-8")  == top_hashtags[k].upper():
```

```
                        total_ids[k].append(ids[c])
      c = c + 1

total_ids2 = []
s = []
for i in total_ids:
      for j in i:
            if j not in s:
                  s.append(j)
      total_ids2.append(s)
      s = []

for i in top_hashtags:
      top_hashtags_2.append(i.decode("utf-8"))

top_hashtags_dict = dict(zip(top_hashtags_2,total_ids2))
diction_pos = {}
diction_neg = {}
result1_list = []
result_list = []

for i in top_hashtags_2:
      print i.encode("utf-8")
      for j in top_hashtags_dict[i]:
            total_emotion_list = twitter_user_streaming.user_stream(j,i)
            if total_emotion_list != []:
                  print total_emotion_list
                  if total_emotion_list[1] > 0:
                        dict2 = {j: total_emotion_list[0] }
                        diction_pos.update(dict2)
                  elif total_emotion_list[1] < 0:
                        dict2 = {j: total_emotion_list[0] }
                        diction_neg.update(dict2)
      result1_list.append(diction_pos)
      result1_list.append(diction_neg)
      result_list.append(result1_list)
      result1_list = []
      diction_neg = {}
      diction_pos = {}

#########################################################################
#το πρώτο λεξικό περιέχει τα θετικά αποτελέσματα και το δεύτερο τα
αρνητικά
#########################################################################

print dict(zip(top_hashtags_2,result_list))
```

Παράρτημα Δ

```
#!/usr/bin/python
# coding=utf-8

import tweepy
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream
import json
import pandas as pd
import re
from nltk.tokenize import TweetTokenizer
import stemimport
from pyexcel_ods import get_data

def removeStopwords(wordlist, flag):
    return [w for w in wordlist if w.encode("utf-8") not in flag]

def wordListToFreqDict(wordlist):
    wordfreq = [wordlist.count(p) for p in wordlist]
    return dict(zip(wordlist,wordfreq))

def sortFreqDict(freqdict):
    aux = [(freqdict[key], key) for key in freqdict]
    aux.sort()
    aux.reverse()
    return aux

def user_stream(i,hash_param):
      result = []
      status_list = []
      json_list =[]
      all_ids_res = []
      list_of_tweets = []
      total_tweets = []
      clean_tweets = []
      tokens = []
      clean_tokens = []
      l = []
      total_hashtag_tweets = []
      lista = []
      lista_hashtag = []
      stem_tweet = []
      stem_hashtags = []

      stopwords = ['ο', 'η', 'το', 'τον', 'τη', 'την', 'στο', 'στον',
      'στη', 'στην', 'στα', 'στις','στους', 'οι', 'τα', '.', ',', '!',
      '(', ')', ':', 'rt', '«', '»', '"', ';', '?', '[', ']', 'τι', 'του',
      'της', 'των', 'και', 'να', 'που', 'θα', 'ότι', 'εμείς', 'εσείς',
      'αυτό', 'αυτά', 'αυτός', 'αυτή', 'αυτές', 'αυτοί', 'για', 'σε',
      'με', 'πως', 'που', 'πότε', 'αν', 'ως', 'απ', 'τ', 'κ', '-', 'κι',
      'μία', 'μια','τους', 'δεν', 'μα', 'μας', 'μου', 'δε', 'άλλα','αλλά',
```

```
'άλλος', 'άλλη', 'άλλο', 'ένα', 'από', 'γιατί', 'ένας', 'σου',
'σας', 'τις', 'είναι', '...', '..', '…', 'ρε',
'μόνο','πώς','χα','χ','κάτι', '|', 'όταν', '\'', 'οσο', 'ειναι',
'ποιος','σ', 'ν', 'τν','μ','μη','π','ν','', 'έχει']

#########################################################################
#Οι μεταβλητές που περιέχουν τα στοιχεία των χρηστών για σύνδεση με
#το Twitter API
#########################################################################

access_token = "..."
access_token_secret = "..."
consumer_key = "..."
consumer_secret = "..."

#########################################################################
#Εξακρίβωση στοιχείων του χρήστη που θέλει να συνδεθεί στο Twitter
#Streaming API
#########################################################################

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)

res = pd.DataFrame()
tknzr = TweetTokenizer()
emoji_pattern = re.compile("["u"\U0001F600-\U0001F64F"u"\U0001F300-
\U0001F5FF"u"\U0001F680-\U0001F6FF"u"\U0001F1E0-\U0001F1FF""]+",
flags=re.UNICODE)

try:
      user =  api.get_user(i)
      status_list = api.user_timeline(user.screen_name)
except tweepy.TweepError:
      print("Failed to run the command on that user, Skipping...")

for j in range(0,len(status_list)):
      status = status_list[j]
      json_str = json.dumps(status._json)
      tweet = json.loads(json_str)
      json_list.append(tweet)
res = map(lambda tweet: tweet['text'], json_list)

for line in res:
      stringwithout = line.lower()
      stringwithout = re.sub(r'http\S+', '', stringwithout)
      stringwithout = re.sub(r'@\S+', '', stringwithout)
      stringwithout = (emoji_pattern.sub(r'', stringwithout))
      stringwithout = re.sub(r'\d+', '', stringwithout)
      clean_tweets.append(stringwithout)

for j in clean_tweets:
      p = tknzr.tokenize(j)
      tokens.append(p)
```

```
for j in tokens:
      clean = removeStopwords(j, stopwords)
      clean_tokens.append(clean)

#####################################################################
#βρίσκουμε τα tweets που να περιέχουν έστω και μία από τις
#δημοφιλέστερες
#λέξεις για κάθε hashtag ή το ίδιο hashtag
#####################################################################

final_result = []
list_of_tweets = []
json_list = []
hash_values = []
tweets_with_3token = []
keys = []
values = []
stem_keys = []
values_data_path = '.../every_hashtags_tokens.txt'

with open(values_data_path) as json_data:
    d = json.load(json_data)
    hash_values.append(d)

hash_values[0][hash_param].append(hash_param)

for j in clean_tokens:
      for k in j:
            if k in hash_values[0][hash_param]:
                  tweets_with_3token.append(j)
                  break;

#####################################################################
#παίρνουμε τις δύο στήλες που μας ενδιαφέρουν για το λεξικό
#####################################################################

data = get_data("lexeis.ods")

for i in range(0,2314):
      keys.append(data["Sheet1"][i][0])
      values.append(data["Sheet1"][i][1])

for i in keys:
      stem_keys.append(stemimport.stem(i.encode("utf-8")))

sentiment_dict = dict(zip(stem_keys,values))

if tweets_with_3token != []:
      for i in tweets_with_3token:
            for k in i:
                  l.append(k.upper().encode("utf-8"))
            total_hashtag_tweets.append(l)
            l = []

      lista_tweet = []
```

```
      for i in total_hashtag_tweets:
            for p in i:
                  if "Έ" in p:
                        lista.append(p.replace("Έ","Ε"))
                  elif "Ό" in p:
                        lista.append(p.replace("Ό","Ο"))
                  elif "Ί" in p:
                        lista.append(p.replace("Ί","Ι"))
                  elif "Ώ" in p:
                        lista.append(p.replace("Ώ","Ω"))
                  elif "Ύ" in p:
                        lista.append(p.replace("Ύ","Υ"))
                  elif "Ά" in p:
                        lista.append(p.replace("Ά","Α"))
                  elif "Ή" in p:
                        lista.append(p.replace("Ή","Η"))
                  else:
                        lista.append(p)

            lista_hashtag.append(lista)
            lista = []

      for i in lista_hashtag:
            for k in i:
                  stem_tweet.append(stemimport.stem(k))
            stem_hashtags.append(stem_tweet)
            stem_tweet = []

      total_counter = 0
      counter = 0

      for i in stem_hashtags:
            for k in i:
                  for m in stem_keys:
                        if k == m :
                              counter = counter +
                              int(sentiment_dict[m])
                              break;
            total_counter = total_counter + counter
            counter = 0
      final_result.append(len(tweets_with_3token))
      final_result.append(total_counter)


else:
      return []


return final_result
```