Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Μεταπτυχιακή Διατριβή
στα Πληροφοριακά και Επικοινωνιακά Συστήματα

Εξόρυξης Γνώμης για Πολιτικά Πρόσωπα από Αναρτήσεις στο Twitter

Θεοδώρα Κουτσού

Επιβλέπων Καθηγητής
Ιωάννης Κατάκης

Ιανουάριος 2015

Ανοικτό Πανεπιστήμιο Κύπρου

Σχολή Θετικών και Εφαρμοσμένων Επιστημών

Εξόρυξης Γνώμης για Πολιτικά Πρόσωπα από Αναρτήσεις στο Twitter

Θεοδώρα Κουτσού

Επιβλέπων Καθηγητής
Ιωάννης Κατάκης

Η παρούσα μεταπτυχιακή διατριβή υποβλήθηκε
προς μερική εκπλήρωση των απαιτήσεων για απόκτηση

μεταπτυχιακού τίτλου σπουδών
στα Πληροφοριακά Συστήματα

από τη Σχολή Θετικών και Εφαρμοσμένων Επιστημών
του Ανοικτού Πανεπιστημίου Κύπρου

Ιανουάριος 2015

Περίληψη

Τα κοινωνικά μέσα δικτύωσης κατά την περίοδο εκλογών, χρησιμοποιούνται για να προβάλλονται οι υποψήφιοι, να εκφράζονται οι απόψεις των πολιτικών αλλά και να τους παρακολουθούν οι χρήστες. Ταυτόχρονα οι χρήστες εκφράζουν και αυτοί την άποψη τους και έτσι αποτελεί μια σύγχρονη επικοινωνία με τους υποψηφίους. Η πλατφόρμα του Twitter είναι η τελευταία τάση στον παγκόσμιο ιστό και χρησιμοποιείται για διάφορα σενάρια από ένα ευρύ σύνολο χρηστών παγκοσμίως. Από τα tweets που αναρτούνται μπορούμε να ανακαλύψουμε πολύτιμες πληροφορίες, γιατί μέσα από αυτές εκφράζουν τις απόψεις τους για τα διάφορα προϊόντα αγοράς, πολιτικά θέματα που απασχολούν τους χρήστες. Περιλαμβάνοντας, ένα έντονο σημασιολογικό περιεχόμενο είναι δύσκολο να μπορέσει κανείς να τα διαβάσει αυτά τα πολυάριθμα σχόλια και συζητήσεις.

Η εξόρυξη γνώμης/ ανάλυσησυναισθήματος έχει τις δυνατότητεςνα ικανοποιήσει τις ανάγκες αφού η εξόρυξη γνώμης είναι ένας εξειδικευμένος κλάδος της επεξεργασίας φυσικής γλώσσας και στοχεύει να προσδιορίσει την υποκειμενική στάση του ομιλούντος ή του γράφοντος σχετικά με ένα ζήτημα δημιουργώντας ένα αυτοματοποιημένο σύστημα. Επίσης το ανάλυση συναισθήματος μεταφέρει τα λεγόμενα των χρηστών ταξινομώντας αν είναι θετικά, αρνητικά ή ουδέτερα.

Σε αυτή την διπλωματική εργασία, επικεντρωνόμαστε στις διάφορες αναρτήσεις των χρηστών που ασχολούνται με πολιτικές συζητήσεις στην Κύπρο και στην περιφέρεια της Αττικής (Ελλάδα) που είναι γραμμένες στην ελληνική γλώσσα κατά την περίοδο των Δημοτικών-Περιφερειακών Εκλογών και τις Ευρωεκλογές 2014. Η συλλογή των δεδομένων έγινε με το εργαλείο TwitterAPI για ένα χρονικό διάστημα επικεντρώνοντας στα δημοφιλέστερα κόμματα. Προτείναμε ένα τρόπο για να ταξινομούμε αυτόματα τα tweets ανά κόμμα,χτίζοντας ένα αλγόριθμο για να προσδιορίσουμε αν το συναίσθημα τους είναι θετικό αρνητικό ή ουδέτερο. Τέλος, τα αποτελέσματα συγκρίθηκαν με τα αποτελέσματα των δημοσκοπήσεων και τα πραγματικά αποτελέσματα.

Summary

During elections the social media networkingis used to show the relevant candidates, express their opinion and also to allow other parties to follow them. The fact that other parties can express their opinion gives asynchronous communication between them. The Twitter platform is the latest trend on the web and used for various scenarios from a wide range of users worldwide. The post of tweets we can discover valuable information because through them express their views on various products, political issues that concern them. Featuring a strong connotation is unlikely that one can read these numerous comments and discussions.

Opinion mining is a specialized course of natural language process and aims to identify the subjective attitude of the talking or writing about an issue by creating an automated system. Also, the sentiment analysis convey users'opinion given into classes of positive, negative or neutral opinion.

In this thesis, we focus on the various posts of users engaged in political discussions in Cyprus and in the region of Attica (Greece), which is written in the Greek language in the period of Municipal-Regional Elections and the European Parliament elections, 2014. The data collection was performed by the use of the "Twitter API" tool for a specific period of time focusing on the most popular parties.

We suggest a way to automatically classify the tweets per party, by building an algorithm to identify if sentiment is positive, negative or neutral. Finally, we compare them with the results of exit polls and regular results.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Κατάκη για την καθοδήγηση που μου έδωσε, τις πληροφορίες, τη συνεργασία που είχαμε σε όλη την διάρκεια της παρούσας εργασίας και την υπομονή που είχε μαζί μου μέχρι την ολοκλήρωση της εργασίας. Επίσης θα ήθελα να ευχαριστήσω την οικογένεια μου αλλά και τους φίλους μου που με στήριξαν μέχρι την ολοκλήρωση της εργασίας αλλά και των Θεματικών Ενοτήτων.

Περιεχόμενα

Κεφάλαιο 1ο
Εισαγωγή

«Προσπάθησε να μάθεις κάτι απ' όλα και όλα για κάτι.» Thomas Huxley

Η παρούσα μεταπτυχιακή διατριβή αφορά στον θεματικό χώρο της «Εξόρυξης Γνώμης» που εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος «Πληροφοριακά και Επικοινωνιακά Συστήματα».

1.1 Αντικείμενο- Σκοπός

Στην μεταπτυχιακή διατριβή έχει σκοπό να μελετηθεί η θεωρητική πλευρά της εξόρυξης γνώμης/ συναισθήματος ανάλυσης και να αναπτυχθεί/υλοποιηθεί μια μέθοδος(αλγόριθμος) για εξαγωγή ταξινόμησης γνώμης από την πλατφόρμα του Twitter. Επικεντρώνοντας, σε πολιτικές αναρτήσεις που είναι γραμμένες στα ελληνικά και να ερμηνεύσει τις αναρτήσεις αν είναι θετικές ή αρνητικές γνώμες. Η σχεδίαση του τεχνικού μέρους γίνεται με τη χρήση της γλώσσας προγραμματισμού R.

1.2 Μελέτη Περίπτωσης

Η συλλογή των δεδομένων αποτελεί ένα ευρύ πλαίσιο μελέτης και εξαγωγής συμπερασμάτων. Στην παρούσα περίπτωση τα διάφορα tweets από πολιτικές συζητήσεις από την Κύπρο και Ελλάδα έχουν ιδιαίτερο ενδιαφέρον για μελέτη. οι χρήστες μπορεί να επηρεάζονται ή να διαμορφώνουν καλύτερα την άποψη τους για ένα πολιτικό ζήτημα ,για τα πολιτικά άτομα ή για διάφορα πολιτικά θέματα που είναι επίκαιρα εκείνη την δεδομένη περίοδο(πχ νομοσχέδιο, περίοδος πολιτικών εκλογών, δημοψήφισμα κλπ).

1.3 Ενότητες Κεφαλαίων

Παρακάτω αναφέρονται περιληπτικά το περιεχόμενα του κάθε κεφαλαίου. Η μεταπτυχιακή διατριβή αναπτύσσεται σε επτά κεφάλαια.

Στο δεύτερο κεφάλαιο αναφέρεται η ανάγκη για την χρήση της εξόρυξης γνώμης μέσω της τεχνολογίας που αναπτύσσεται σε πολύ γρήγορους ρυθμούς έχοντας ένα τεράστιο πλούτο πληροφορίας. Αυτός ο όγκος της μεγάλης πληροφορίας γίνεται η ανάγκη για αυτόματη συλλογή και ανάλυση ώστε η πληροφορία να είναι χρήσιμη και κατανοητή από χρήστες και μελετητές. Η εξόρυξη γνώμης ασχολείται μέσα από τα κοινωνικά μέσα για ανάλυση και εξαγωγή συμπερασμάτων. Επίσης αντιμετωπίζει διάφορα προβλήματα γιατί οι γνώμες είναι γραμμένες στην φυσική γλώσσα.

Στο επόμενο κεφάλαιο γίνεται αναφορά στην εξόρυξη γνώμης μέσα από τα κοινωνικά δίκτυα επικεντρώνοντας στην πλατφόρμα Twitter. Στην εποχή μας που οι περισσότεροι πολίτες εκφράζουν τις γνώμες τους μέσα από τις διάφορες πλατφόρμες,blog, forum είναι δύσκολο κανείς να πάρει τα αποτελέσματα της γνώμης διαβάζοντας ένα προς ένα γιατί υπάρχει τόσο μεγάλος όγκος πληροφορίας. Ενδεικτικά θα αναφερθούμε την προεκλογική εκστρατεία που έκανε ο Obama το 2008, Ιρλανδικές Εκλογές 2011 που ήταν μέσω της πλατφόρμας Twitter εφαρμόζοντας το ανάλυση συναισθήματοςαντικαθιστώντας τις παραδοσιακές δημοσκοπήσεις. Επίσης υπάρχουν αρκετές εταιρίες που εφαρμόζουν λογισμικά στην εξόρυξη γνώμης και είναι δωρεάν.

Στο κεφάλαιο τέσσερα παρουσιάζεται η διαδικασία εξαγωγής αναρτήσεων στην Κύπρο και Ελλάδα από την πλατφόρμα Twitter χρησιμοποιώντας εργαλεία όπως το Twitter API. Η

υλοποίηση έγινε με την γλώσσα προγραμματισμού R έχοντας κάποια βήματα , επιλέγοντας ένα ικανοποιητικό αριθμό αναρτήσεων που είναι γραμμένα στην ελληνική γλώσσα. Ακολούθως ,δημιουργώντας λίστες με πολιτικά ονόματα με βάση κριτηρίων για την επιλογή και λέξεις κλειδιά για να πάρουν όσες αναρτήσεις αναφέρονται σε πολιτικές συζητήσεις.

Στο κεφάλαιο πέντε αναλύουμε τα δεδομένα που είχαμε στο προηγούμενο κεφάλαιο και συγκεκριμένα από αυτά της περιφέρειας Αττικής γιατί τα δεδομένα που πήραμε από Κύπρο δεν αναφέρονται σε πολιτικά θέματα. Εφαρμόζουμε δύο λίστες λεξικών με θετικές και αρνητικές λέξεις, μεταφρασμένες από το λεξικό Sentiment Lexicon των Hu and Liu καθώς επίσης κάνοντας αναζήτηση με την λέξη 'δεν' όπου μπορεί να αντιστρέψει το προσανατολισμό του συναισθήματος που έχει η λέξη λεξικού. Επίσης περιγράφουμε τα βήματα δημιουργίας του αλγορίθμου.

Στο κεφάλαιο έξι αναφέρουμε τι έγινε στις διάφορες δημοσκοπήσεις που έγιναν για τις Δημοτικές-Περιφεριακές Εκλογές και στις Ευρωεκλογές. Αναλύοντας τα αποτελέσματα που είχαμε από τον αλγόριθμο μας. Παρατηρώντας, σε κάθε κόμμα το ποσοστό του θετικού και του αρνητικού συναισθήματος καθώς επίσης και τα ουδέτερα σχόλια των χρηστών. Επίσης, επικεντρώθηκαμε σε τρία δημοφιλέστερα πολιτικά ονόματα αναλύοντας τι συναίσθημα εκφράζουν οι χρήστες προς αυτούς. Στο τέλος γίνεται μια σύγκριση των πειραματικών αποτελεσμάτων, των αποτελεσμάτων της εκλογικής διαδικασίαςκαι των δημοσκοπήσεων.

Φτάνοντας στο τελευταίο κεφάλαιο αναφέρονται τα συμπεράσματα και οι μελλοντικές προοπτικές της παρούσας μεταπτυχιακής διατριβής.

Κεφάλαιο 2ο
Εξόρυξη Γνώμης

«Πνιγόμαστε στις πληροφορίες, αλλά διψάμε για γνώση» John Naisbitt

2.1 Γενικά

Η αλματώδης αύξηση του τεράστιου όγκου δεδομένων που δημοσιεύονται καθημερινά στο διαδίκτυο μέσα από τα κοινωνικά μέσα μαζικής ενημέρωσης(κριτικές, φόρουμ συζητήσεων blogs, micro-blogs, Twitter, networks sites) εντάσσονται από τις καθημερινές καταστάσεις της ζωής των ανθρώπων. Είναι ένα κίνητρο προς έρευνα αλλά και η ανάπτυξη με τις τεχνικές της εξόρυξης γνώμης που μέχρι τώρα δεν μπορούσαν να μελετηθούν. Ωστόσο, υπάρχει μεγάλη ανάπτυξη στο ερευνητικό πλαίσιο της εξόρυξης γνώμης(opinion mining) με αρκετά θέματα να επικεντρώνονται σε κοινωνικά μέσα(social media) για τις απόψεις των ανθρώπων.

2.2 Εισαγωγή[1]

Η Εξόρυξη γνώμης και το συναίσθημα ανάλυσης βρίσκονται στο ίδιο πεδίο μελέτης που ασχολείται με τις ανθρώπινες απόψεις, τα συναισθήματα, τις αξιολογήσεις , τις συμπεριφορές των ανθρώπων και τις κριτικές. Δηλαδή, τις διάφορες απόψεις για προϊόντα, οργανισμούς, γεγονότα, πολιτικά θέματα, διάφορα πρόσωπα κλπ όπου είναι σε γραπτό λόγο. Είναι ένα ενεργό πεδίο μελέτης το οποίο συνδέεται με την επεξεργασία φυσικής γλώσσας καθώς μελετάτε και στα πεδία της επιστήμης υπολογιστών, της εξόρυξης πληροφορίας από κείμενο και το διαδίκτυο.

Η εξόρυξη γνώμης και η ανάλυση συναισθήματος μπορούμε και να το συναντήσουμε και αλλιώς δίνοντας διαφορετική ονομασία με ελαφριά διαφορά. Όπως : opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining κλπ. Ανάλογα σε ποια κατηγορία(πχ. εκπαίδευση, βιομηχανία) θα αναφερθεί η έρευνα που θα γίνει, αντιστοιχεί και η ανάλογη ονομασία. Η έννοια τους είναι ότι εστιάζονται στις απόψεις που εκφράζονται αν το συναίσθημα είναι θετικό ή αρνητικό.

Οι απόψεις για κάποια θέματα συζητήσεων είναι επίκεντρο σε όλες τις ανθρώπινες δραστηριότητες γιατί μέσα από αυτό εκφράζουν τη συμπεριφορά τους. Στις περιπτώσεις των επιχειρήσεων , οργανώσεων οι υπεύθυνοι πάντα θέλουν να γνωρίζουν την άποψη του καταναλωτή για τα προϊόντα που αγοράζουν και γενικά τις διάφορες κριτικές που γίνονται γύρω από τις επιχειρήσεις τους. Επίσης, ο καταναλωτής θέλει και αυτός να γνωρίζει τις κριτικές άλλων χρηστών για τα προϊόντα και θα ζητήσει τις γνώμες άλλων ανθρώπων προτού πάρουν μια απόφαση.

Ο παραδοσιακός τρόπος, ρωτώντας και αναλύνοντας τις απόψεις των πολιτών διεκπεραιώνεταιι σε δημοσκοπήσεις για το θέμα που ενδιαφέρονται Έτσι, οι επιχειρήσεις και οι οργανισμοί χρησιμοποιούν όλο και περισσότερο τα διάφορα μέσα λήψης αποφάσεων και όχι της δημοσκόπησης γιατί ήταν χρονοβόρες μέχρι να μέχρι να μαζευτούν αρκετές πληροφορίες. Αν και, η παρακολούθηση για τις διάφορες απόψεις και γνώμες στο διαδίκτυο είναι ένα δύσκολο έργο λόγω του ότι κάθε ιστοσελίδα περιέχει ένα τεράστιο όγκο από κείμενο και δεν είναι εύκολο να αποκρυπτογραφηθεί. Όμως, αποτελούνται πηγή θησαυρού πληροφοριών που αντιστοιχούν άμεσα οι απόψεις και γνώμες των χρηστών ή των πολιτών από μια ευρεία γκάμα θετικών

[1] Αναφορά από το βιβλίο [25] κεφάλαιο 1

περιοχών και η εξόρυξη γνώμης μπορεί να αντικαταστήσει τις παραδοσιακές έρευνες και δημοσκοπήσεις με επιτυχία. Μια σχετική έρευνα[30] έδειξε ότι :

- Το 60% των κατοίκων των Ηνωμένων Πολιτειών έχουν κάνει διαδικτυακή έρευνα αγοράς τουλάχιστον μια φόρα, ενώ το 15% αυτών κάνει κάτι τέτοιο καθημερινά.

- Ένα ποσοστό που κυμαίνεται μεταξύ 73%-87% ισχυρίζεται ότι οι διαδικτυακές κριτικές που διάβασε είχαν σημαντική επιρροή στην αγορά του.

- Το 30% των κατοίκων των Ηνωμένων Πολιτειών έχει γράψει μια διαδικτυακή κριτική ή έχει παραθέσει ένα αντίστοιχο σχόλιο.

- Όμως, το 58% των χρηστών του Internet στις Ηνωμένες Πολιτείες δηλώνει ότι οι διαθέσιμες πληροφορίες ήταν ελλιπείς, μπερδεμένες και πολυάριθμες.

Από την έρευνα αυτή βλέπουμε πως πρέπει να δημιουργηθεί τεχνολογία που να μπορεί να εξάγει και να αναλύει τις διάφορες γνώμες και απόψεις των πολιτών. Έτσι πολλές ερευνητικές εργασίες έχουν ασχοληθεί με την προβλέψη τις αποδόσεις των πωλήσεων αλλά και τις κριτικές των προϊόντων τους. Ακόμη αρκετές έρευνες έγιναν στις κοινές γνώμες στα blogs, forum και Twitter. Συγκεκριμένα στο Twitter μελέτησαν τις προβλέψεις των αποτελεσμάτων των εκλογών διάφοροι ερευνητές για τις πολιτικές απόψεις , για πρόβλεψη αποτελεσμάτων μέσα σε πολιτικά blogs, ακόμη και για τις κριτικές ταινιών.

Το δημοφιλέστερο ερευνητικό πρόβλημα επικεντρώνεται στις εφαρμογές τις πραγματικής ζωής που είναι γραμμένα σε φυσική γλώσσα όπως αναφέρθηκαν πιο πάνω για τη ανάλυση συναισθήματος.

2.3 Ορολογία Εξόρυξης Γνώμης

Η ορολογία της Εξόρυξη Γνώμης(Opinion Mining) αφορά την επεξεργασία κειμένων γραμμένων σε φυσική γλώσσα με τη χρήση υπολογιστικών και στατιστικών μεθόδων, με σκοπό την εξαγωγή υποκειμενικών πληροφοριών. Ο γενικός στόχος της Εξόρυξη Γνώμης είναι ο προσδιορισμός της άποψης του χρήστη για ένα συγκεκριμένο θέμα, προϊόν κλπ και η γενικότερη σημασιολογική πολικότητα του περιεχομένου ενός εγγράφου(contextual polarity). Η εξόρυξη γίνεται συνήθως από κείμενο το οποίο περιέχει έγκυρες πληροφορίες ή γνώμες, ενώ υπάρχει

πάντα υψηλό κίνητρο για την αυτοματοποίηση της διαδικασίας έστω και αν το αποτέλεσμα δεν είναι τέλειο.

Η άποψη που μπορεί να εκφράσει ο χρήστης προκύπτει είτε από την προσωπική κρίση, είτε από τη συναισθηματική του κατάσταση, είτε από τη συναισθηματική επιρροή την οποία ασκεί στο χρήστη ο συγγραφέας του κειμένου, είτε σαφώς από ένα συνδυασμό των παραπάνω.

Στις υπολογιστικές και στατιστικές μεθόδους οι οποίες χρησιμοποιούνται στην εξόρυξη γνώμης από κείμενα συμπεριλαμβάνονται : η επεξεργασία φυσικής γλώσσας(Natural Language Processing),η λεξικογραφική ανάλυση κειμένου μέσω υπολογιστικών αλγορίθμων(Computational linguistics),

2.4 Ιστορική Αναδρομή

Ο όρος opinion mining εμφανίστηκε για πρώτη φορά σε δημοσίευση των Kushal Dave, Steve Lawrence, David M. Pennock[08] στα πλαίσια του συνεδρίου WWW(WWW conference) κατά το έτος 2003. Αυτή η δημοσίευση στο συγκεκριμένο συνέδριο μπορεί εν μέρει να εξηγούσε την δημοτικότητα του όρου opinion mining μεταξύ των κοινοτήτων που είναι προσανατολισμένες στην κατεύθυνση αναζήτησης στο διαδίκτυο(Web search) ή ανάκτησης πληροφορίας(information retrieval). Επιπλέον αυτή η δημοσίευση με την ιδανική εφαρμογή για εξόρυξη συναισθήματος «θα μπορούσε να επεξεργαστεί ένα σύνολο από δεδομένα αναζήτησης, δημιουργώντας μία λίστα των κύριων χαρακτηριστικών αυτών και συνοψίζοντας τις απόψεις που επικρατούν για κάθε ένα από αυτά τα χαρακτηριστικά σε θετικές, ουδέτερες και αρνητικές». Επίσης έχουν διεξαχθεί αρκετές έρευνες στο πεδίο του opinion mining και διερευνούνται συνεχώς νέα ευρήματα για την ανάλυση κειμένου από διάφορες πλευρές. Ωστόσο, ο όρος της εξόρυξής γνώμης πρόσφατα έχει αναπτυχθεί και διακλαδωθεί καλύπτοντας μεγαλύτερο εύρος και περιέχοντας πολύ περισσότερους και ταυτόχρονα διαφορετικούς τύπους ανάλυσης των κειμένων προς αξιολόγηση των κειμένων.

Ας μελετήσουμε τώρα, την ιστορία του όρου ανάλυση συναισθηματος (sentiment analysis) που συμπίπτει με αυτή της εξόρυξης γνώμης από αρκετές πτυχές. Κατά αρχή, ο όρος «sentiment» εμφανίστηκε για πρώτη φορά την χρονολογία του 2001 στις δημοσιεύσεις[07],[37] που χρησιμοποιείται για την αυτόματη ανάλυση του κειμένου προς αξιολόγηση και την πρόβλεψη άποψης μέσα από αυτό. Το ενδιαφέρον ξεκίνησε για την ανάλυση συναισθήματος στην αγορά.

Την επόμενη χρονιά οι ερευνητές Turney[40] και Bo Pang, Lillian Lee, Shivakumar Vaithyanathan[32] στην ετήσια συνάντηση της Εταιρίας Υπολογιστικής Γλωσσολογίας(Association for Computational Linguistics(ACL))και επίσης στο ετήσιο συνέδριο Εμπειρικών Μεθόδων για την Επεξεργασία Φυσικής Γλώσσας(Empirical Methods in Natural Language Processing(EMNLP)) ασχολούνται με την ανάλυση συναισθήματος. Προσπαθούν να εξηγήσουν την έννοια του όρου sentiment analysis μεταξύ των κοινοτήτων με έμφαση στην Επεξεργασία Φυσικής Γλώσσας. Υπάρχει μια ποικιλία από αξιόλογες δημοσιεύσεις που αναφέρεται σε αυτό τον όρο εμβαθύνοντας κυρίως σε εφαρμογές κατηγοριοποίησης κριτικών σχετικά με την πόλωση συναισθήματος που περιέχεται σε αυτές. Επομένως, κάποιοι ερευνητές έχουν συνδέσει τον όρο sentiment analysis και με το opinion mining όπου εμπεριέχουν ποικιλία ανάλυσης.

2.5 Το πρόβλημα της Εξόρυξης Γνώμης[2]

Ακολουθούν οι πιο κάτω προτάσεις:

«Αγόρασα Ηλεκτρονικό Υπολογιστή HP πριν τέσσερεις μήνες.»

«Είναι ένας πολύ ωραίος Ηλεκτρονικός Υπολογιστής.»

«Ο επεξεργαστής είναι πολύ γρήγορος.»

«Τα ανεμιστηράκια(fans) του πύργου κάνουν πολύ θόρυβο.»

Η γνώμη αποτελείται από δύο συστατικά, τον στόχο(target) και το συναίσθημα του στόχου. Παρακάτω αναλύονται διάφοροι ορισμοί γύρω από την γνώμη των στόχο και το συναίσθημα.

Όπως έχουμε δει και πιο πάνω η γνώμη μπορεί να εκφραστεί για οτιδήποτε είτε για ένα προϊόν, μια υπηρεσία, ένα οργανισμό, ένα γεγονός ή ένα γενικό θέμα. Με τον όρο **αντικείμενο(object)** υποδηλώνουμε τον στόχο του αντικειμένου που έχει αναφερθεί. Το αντικείμενο κατατάσετε ιεραρχικά από ένα σύνολο συστατικών(components) ή μερών(part) και ένα σύνολο χαρακτηριστικών ή ιδιοτήτων(attributes or properties). Κάθε συστατικό μπορεί να έχει τα δικά του υπό-σύνολα(sub-components) και τα δικά του χαρακτηριστικά. Στην βιβλιογραφία αλλά και σε διάφορες έρευνες ο όρος αντικείμενο ορίζετε και ως οντότητα(entities).

[2] Αναφορά από το βιβλίο [25] κεφάλαιο 2

Παράδειγμα: Μια συγκεκριμένη μάρκα Ηλεκτρονικού Υπολογιστή είναι ένα αντικείμενο ,όπως είναι η πρόταση(1) HP. Έχει ένα σύνολο συστατικών όπως *οθόνη, πύργος, πληκτρολόγιο, ποντίκι* και ένα σύνολο από χαρακτηριστικά όπως μέγεθος οθόνης, μέγεθος πληκτρολόγιου. Το συστατικό *πύργος* έχει τα δικά του χαρακτηριστικά όπως σκληρός δίσκος, μητρική πλακέτα, CD-ROM.

Βασισμένοι στον ορισμό του αντικειμένου μπορεί να αναπαρασταθεί και ως ένα δέντρο ιεραρχίας. Στην ρίζα να βρίσκεται το όνομα του αντικειμένου και κάθε κόμβος κλαδί που να είναι ένα συστατικό ή υπό-συστατικό του αντικειμένου. Κάθε σύνδεση είναι ένα μέρος της σχέσης. Κάθε κόμβος είναι μια συσχέτιση με ένα σύνολο από χαρακτηριστικά ή ιδιότητες του κόμβου. Δηλαδή η ταχύτητα του επεξεργαστή είναι αργή. Έτσι κάποιος μπορεί να εκφράσει γνώμη σε οπουδήποτε από τα συστατικά ή χαρακτηριστικά του Ηλεκτρονικού Υπολογιστή.

Στην πράξη όμως είναι καλύτερα να απλοποιηθεί το πιο πάνω για τους παρακάτω λόγους. Η διαδικασία της φυσικής επεξεργασίας γλώσσας είναι δύσκολη και για ένα συνηθισμένο χρήστη για να χρησιμοποιήσει την πιο πανω ιεραρχία. Έτσι απλοποιούμε την ιεραρχία των δέντρων και χρησιμοποιούμε τον όρο **χαρακτηριστικά(features)**. Όπου χαρακτηριστικό είναι τα χαρακτηριστικά των υπό-συστατικών και των ιδιοτήτων που περιγράφουν τα υπό-συστατικά των αντικειμένων. Κάθε χαρακτηριστικό αναφέρεται σε μια γνώμη από ένα έγγραφο που ανήκει το αντικείμενο. Το χαρακτηριστικό(features) μπορούμε και να το συναντήσουμε και ως topic ή aspect.

Μια σειρά από προτάσεις σε ένα έγγραφο εκφράζουν γνώμη αντικειμένων ή χαρακτηριστικά στο αντικείμενο. Έτσι μια πρόταση μπορεί να εκφράσει περισσότερα από ένα χαρακτηριστικά. Για παράδειγμα *''Η κάρτα γραφικών είναι καλή αλλά το πληκτρολόγιο δεν είναι βολικό"*. Βλέπουμε στις προτάσεις υπάρχουν μια ποικιλία από εκφράσεις χαρακτηριστικών που συνήθως είναι ουσιαστικά ή φράσεις ουσιαστικών καθώς και ρήματα, επίθετα και επιρρήματα. Ορίζουμε ένα **ρητό χαρακτηριστικό(explicit feature)** τις εκφράσεις χαρακτηριστικών που είναι ουσιαστικά και φράσεις ουσιαστικών σε μια πρόταση. Δηλαδή *''Η διάρκεια ζωής της μπαταρίας του τηλεφώνου είναι μικρή"*. Η 'διάρκεια ζωής' είναι μια σαφής έκφραση χαρακτηριστικών. Επίσης ορίζουμε **υπονοούμενες/κρυφό χαρακτηριστικο(implicit feature)** τις εκφράσεις χαρακτηριστικών μέσα σε μια πρόταση. Υπονοούμενα χαρακτηριστικά είναι τα επίθετα και επιρρήματα. Για παράδειγμα ο όρος "μεγάλο" κρύβει ένα υπονοούμενο χαρακτηριστικό. Όπως για την πρόταση «Το πληκτρολόγιο είναι πολύ μεγάλο» εννοεί το χαρακτηριστικό για το μέγεθος. Ακόμη ένα παράδειγμα είναι «Η οθόνη είναι πολύ ακριβή» σε αυτή την περίπτωση το

‘ακριβή’ αναφέρεται στο κόστος. Ακόμη μπορεί να υπάρχουν και άλλα χαρακτηριστικά που είναι σύνθετα δηλαδή «Ο Η/Υ δεν χωράει εύκολα στο γραφείο», εδώ το ‘χωράει στο γραφείο’ εννοεί για χαρακτηριστικό μεγέθους ή σχήμα.

Υπάρχει μια κατηγορία που ονομάζεται **κάτοχος γνώμης(opinion holder)**. Ο κάτοχος γνώμης είναι το πρόσωπο ή οργανισμός που εκφράζει γνώμη. Ο ορισμός της γνώμης (opinion) είναι μια γνώμη από χαρακτηριστικά(feature) που είναι θετικές ή αρνητικές απόψεις, συμπεριφορές, συναισθήματα ή εκτιμήσεις από την κάτοχο γνώμης. Ο ορισμός του **προσανατολισμού γνώμης(opinion orientation)** είναι η γνώμη από τα χαρακτηριστικά που υποδεικνύει αν η γνώμη είναι θετική, αρνητική ή ουδέτερη. Ο προσανατολισμός γνώμης μπορεί και να τον συναντήσουμε και ως sentiment orientation, polarity of opinion or semantic orientation.

Ορισμός **Γνώμης(Opinion):** Η γνώμη περιγράφεται από μια πεντάδα(quintuple) (**O_i, F_{ij}, S_{ijkl}, H_k, T_l**) όπου :

O_i : έχουμε το όνομα του αντικειμένου

F_{ij} : Το χαρακτηριστικό του αντικειμένου O_i

S_{ijkl} : Το συναίσθημα της άποψης του αντικειμένου O_i

H_k : Κάτοχος γνώμης

T_l: Ο χρόνος που εκφράζεται η γνώμη από το H_k

Ο προσανατολισμός της γνώμης S_{ijkl} εκφράζεται ως θετική , αρνητική ή ουδέτερη με διαφορετικά επίπεδα μεγέθους/έντασης. Όταν μια γνώμη γίνεται για το αντικείμενο Oi χρησιμοποιούμε τον όρο «ΓΕΝΙΚΑ»(GENERAL) για να το υποδηλώσουμε.

Αναφέρουμε ορισμένες σημαντικές παρατηρήσεις για τον ορισμό της γνώμης(πεντάδα):

Τα πέντε πεδία που περιλαμβάνει ο ορισμός πιο πάνω πρέπει να αντιστοιχούν μεταξύ τους. Δηλαδή η γνώμη S_{ijkl} πρέπει να δίνεται από τον κάτοχο γνώμης H_k για το χαρακτηριστικό F_{ij} του αντικειμένου Oi στο σωστό χρόνο Ti αλλιώς δεν έχουμε σωστή εξαγωγή της πεντάδας.

Τα πέντε πεδία είναι απαραίτητα. Αν λείπει ένα από αυτά μπορεί η ανάλυση γνώμης να είναι προβληματική. Αν για παράδειγμα λείπει το πεδίο με το χρόνο είναι δύσκολο να αναλύσουμε την γνώμη γιατί δε θα μπορούσαμε να καταλάβουμε σε πια χρονική περίοδο έχει γίνει. Ο χρόνος πριν 7 χρόνια με τον χρόνο πριν 7 μήνες δεν είναι το ίδιο.

Στο ορισμό που δώσαμε πιο πάνω η γνώμη αναφέρεται για ένα είδος γνώμης όπου είναι η κανονική γνώμη(regular opinion). Ένα άλλο είδος γνώμης είναι η συγκριτική γνώμη(comparative opinion) όπου εκεί ορίζεται διαφορετικά η γνώμη.

Τα έγγραφα που παίρνουμε από την πεντάδα είναι αδόμητη πληροφορία ενώ τα δεδομένα έχουν μια βάση δεδομένων που αναφέρονται ως δοδημένα. Δηλαδή μια δοδομένη προσέγγιση μετατρέπει τα αδόμητα κείμενα σε παραδοσιακά εργαλεία διαχείρισης δεδομένων. Έτσι η πληροφορία που παρέχεται επιτρέπει στον χρήστη να αποκτήσει γνώσεις τόσο με ποιοτική όσο και με ποσοτική ανάλυση.

Μοντέλο αντικειμένου: Ένα αντικείμενο O εκπροσωπείται από ένα πεπερασμένο σύνολο χαρακτηριστικών $F=\{f_1,f_2,\ldots..,f_n\}$ που περιλαμβάνεται από το ίδιο το αντικείμενο σαν ένα ειδικό χαρακτηριστικό. Κάθε χαρακτηριστικό $f_i \in F$ του αντικειμένου μπορεί να εκφράσει και ένα σύνολο από πεπερασμένες λέξεις ή φράσεις $W_i=\{W_1,W_2,\ldots\ldots,W_n\}$ που είναι σύνολα των χαρακτηριστικών ή υποδεικνύουν από κάθε ένα από αυτά το σύνολο των εκφράσεων χαρακτήρων $I_i=\{I_1,I_2,\ldots..,I_n\}$.

Μοντέλο γνώμης κειμένου: Το μοντέλο αυτό περιέχει γνώμες από ένα σύνολο αντικειμένων $\{O_1,O_2,\ldots..,O_n\}$ μέσα από ένα σύνολο κατόχων γνώμης $\{h_1,h_2,\ldots..h_n\}$. Οι γνώμες χωρίζονται σε δύο είδη την **άμεση γνώμη(direct opinion)** και την **συγκριτική γνώμη(comparative opinion)**.

Η άμεση γνώμη εκφράζεται όπως έχει αναφερθεί πιο πάνω στον ορισμό της γνώμης. Προσθέτοντας στο πεδίο f_{jk} τα σχόλια της κατοχής γνώμης h_i να επιλέγει μια λέξη ή μια φράση από το αντίστοιχο συνώνυμο σύνολο w_{jk} ή λέξη ή φράση από το χαρακτηριστικό των συνόλων Ijk για να περιγράψει το χαρακτηριστικό όταν η έκφραση είναι θετική ή αρνητική. Επίσης στην άμεση γνώμη υπάρχουν δύο υπό-τύπων γνώμης. Το πρώτο είναι η άμεση έκφραση γνώμης από ένα αντικείμενο ή χαρακτηριστικό του αντικειμένου. Ο δεύτερος υπό-τύπος είναι η γνώμη του αντικειμένου που εκφράζεται με βάση τα αποτελέσματα των άλλων αντικειμένων. Αυτός ο υπό-τύπος συνήθως εκφράζεται για ιατρικά θέματα όπως για παράδειγμα «Μετά την λήψη αυτού

του φαρμάκου, το δεξί μου χέρι είναι καλύτερα». Δηλαδή βλέπουμε την επίδραση του φαρμάκου για τα χέρια του, ωστόσο είναι θετική γνώμη για το φάρμακο.

Η συγκριτική γνώμη είναι η γνώμη μεταξύ περισσοτέρων από δύο αντικειμένων και κατόχων γνώμης που βασίζεται σε κάποια χαρακτηριστικά των αντικειμένων. Η συγκριτική γνώμη συνήθως εκφράζεται με συγκριτική ή υπερθετική μορφή του επιθέτου ή του επιρρήματος αλλά όχι και πάντα(ενότητα 2.10).

Ακόμη μια πλευρά της γνώμης είναι ότι η γνώμη σε ορισμένες περιπτώσεις είναι ισχυρή και άλλοτε αδύνατη. Για παράδειγμα «Νομίζω πως ο ηλεκτρονικός Υπολογιστής είναι καλός» εκφράζει αδύνατη γνώμη ενώ η πρόταση «Ο Ηλεκτρονικός Υπολογιστής είναι καταπληκτικός» εκφράζει ισχυρή γνώμη. Έτσι, την γνώμη μπορούμε να την πάρουμε και κλιμακωτά είτε για ισχυρή και είτε για αδύνατη γνώμη. Για παράδειγμα η θετική γνώμη εκφράζεται με το ικανοποιημένος, χαρούμενος, εύθυμος, εκστατικός(ecstatic), για να είναι πιο ομαδοποιημένα χωρίζουμε τις λέξεις σε δυο επίπεδα. Το ένα επίπεδο να είναι οι ισχυρές θετικές λέξεις (ικανοποιημένος, χαρούμενος) και σε άλλο επίπεδο η πιο αδύνατες θετικές λέξεις. Από εδώ παρατηρούμε πως η γνώμη περιλαμβάνεται στο γεγονός των συναισθημάτων(emotions).

Το συναίσθημα(emotion) έχει μελετηθεί από διάφορα πεδία , όπως ψυχολογία, φιλοσοφία, κοινωνιολογία, βιολογία κλπ. Είναι τόσο πολύ το φάσμα που έχουν να καλύψουν σε αυτό το πεδίο και έτσι κάποιοι ερευνητές κατηγοριοποιούν τα συναισθήματα των ανθρώπων σε έξι συναισθήματα: αγάπη, χαρά, έκπληξη, θυμό, θλίψη και φόβο όπου το κάθε συναίσθημα μπορεί να υποδιαιρείται σε άλλα μεγαλύτερα συναισθήματα. Έτσι για να ορίσουμε το **συναίσθημα(emotion)** το ξεχωρίζουμε στα υποκειμενικά συναισθήματα(subjective felling) και στις σκέψεις .

Η υποκειμενικότητα(subjectivity) και το συναίσθημα(emotion) είναι δύο σημαντικές έννοιες που συνδέονται με το συναίσθημα (sentiment). Ο ορισμός της **υποκειμενικής πρότασης(sentence subjective)** είναι να εκφράζει κάποια προσωπικά συναισθήματα(feeling), απόψεις και πεποιθήσεις ενώ η **αντικειμενική πρόταση(objective sentence)** εκφράζει πληροφορίες γεγονότων για τον κόσμο.

Η πρόταση γνώμης(opinionated sentence) είναι η πρόταση που εκφράζετε άμεσα ή έμμεσα στην γνώμη όταν είναι θετική ή αρνητική και αυτό το συναντάμε τις υποκειμενικές και αντικειμενικές προτάσεις.

Παράδειγμα :

Αντικειμενική πρόταση : "Το MAC είναι ένα προϊόν της APPLE"

Υποκειμενική πρόταση : "Μου αρέσουν τα MAC"

Αυτές οι δύο έννοιες υποκειμενική πρόταση και δογματική πρόταση δεν είναι ισοδύναμες αλλά έχουν μια μεγάλη τομή που ονομάζεται ταξινόμηση υποκειμενικότητας(subjectivity classification).

2.6 Επίπεδα ανάλυσης[3]

Υπάρχουν τρία επίπεδα ανάλυσης από τα προβλήματα της επανάληψης όπου περιγράφονται εδώ και αναλύονται στις επόμενες ενότητες :

Επίπεδο εγγράφου(Document level): Σε αυτό το επίπεδο η ταξινόμηση εκφράζεται αν το έγγραφο στην συνολική γνώμη είναι θετικό ή αρνητικό μέσα από ένα ενιαίο αντικείμενο(πχ ένα προϊόν). Το επίπεδο εκφράζεται ως ταξινόμηση εγγράφου σε επίπεδο συναισθήματος «document level sentiment classification»

Επίπεδο πρότασης(Sentiment level): Σε αυτό το επίπεδο η ταξινόμηση εκφράζεται αν η πρόταση είναι θετική, αρνητική ή ουδέτερη. Ουδέτερη γνώμη είναι αυτή που δεν έχει απόψεις και δεν μπορεί να οριστεί αν είναι θετική ή αρνητική. Σύμφωνα με τους Wiebe ,Bruce ,O'Hora(1999)[43] στις υποκειμενικές προτάσεις δηλώνετε ποιος ενεργεί ή δέχεται μια ενέργεια ή αν βρίσκεται σε μια κατάσταση.

Επίπεδο χαρακτηριστικού(Feature level): Αυτό το επίπεδο μπορούμε και να το εντοπίσουμε και σαν Entity and Aspect level. Από τα προηγούμενα επίπεδα δεν μπορούμε να αντιληφθούμε τι ακριβώς αρέσει και τι δεν αρέσει. Έτσι σε αυτό το επίπεδο εντοπίζει απευθείας αν η γνώμη εκφράζεται από θετική ή αρνητική. Για παράδειγμα η πρόταση 'Αν και η εξυπηρέτηση ήταν χάλια, λατρεύω αυτό το εστιατόριο». Η πρόταση έχει θετική άποψη αλλά δεν κατατάσετε σαν απολύτως θετική γνώμη. Θετική είναι για το εστιατόριο αλλά είναι αρνητική για την εξυπηρέτηση.

[3] Αναφορά το βιβλίο[25]της βιβλιογραφίας

Επομένως ο στόχος της ανάλυσης σε ένα επίπεδο έγγραφου και πρότασης πρέπει να εντοπίζεται το συναίσθημα σε αντικείμενα από τους **στόχους(opinion targets)** και τις **πτυχές(aspects)**. Για παράδειγμα η πρόταση :"Η ποιότητα του τηλεφώνου είναι καλή αλλά η διάρκεια ζωής της μπαταρίας είναι πολύ μικρή". Βλέπουμε πως αξιολογεί δύο πτυχές, την ποιότητα του τηλεφώνου και την διάρκεια ζωής της μπαταρίας όπου είναι αρνητική. Η ποιότητα της κλήσης και η διάρκεια ζωής του τηλεφώνου είναι στόχοι. Παρατηρούμε πως το επίπεδο χαρακτηριστικό αποτελείται από πολλά υπό-προβλήματα σε αδόμητα ,δομημένα δεδομένα όπου θα μελετηθούν στις επόμενες ενότητες.

2.7 Συναίσθημα και Υποκειμενική ταξινόμηση[4]

Η ταξινόμηση συναισθήματος(sentiment classification) ταξινομεί ένα κείμενο γνώμης εγγράφου εκφράζοντας θετικές και αρνητικές γνώμες. Όπως αναφέραμε στην προηγούμενη ενότητα μπορούμε να το συναντήσουμε και σαν επίπεδο εγγράφου-ταξινόμηση συναισθήματος(document level sentiment classification). Επίσης την ταξινόμηση συναισθήματος δεν την εντοπίζουμε μόνο στα έγγραφα αλλά και στις μεμονωμένες προτάσεις. Η πρόταση της πρότασης αν είναι δογματική ή όχι λέγεται υποκειμενική ταξινόμηση(subjective classification) εκφράζοντας θετικές και αρνητικές γνώμες. Αλλιώς μπορούμε να το συναντήσουμε σαν επίπεδο πρότασης –ταξινόμηση πρότασης(sentence level sentiment classification). Στις πιο κάτω υποενότητες εξηγούμε αυτά τα δύο επίπεδα.

2.7.1 Επίπεδο Εγγράφου

Το επίπεδο αυτό αναφέρεται μόνο σε δογματικά έγγραφα γιατί εκφράζει μια απλή γνώμη είτε θετική είτε αρνητική. Δεν μπορεί να είναι στα forum και blogs γιατί έχει πολλές γνώμες και οι προτάσεις είναι συγκριτικές ή υπερθετικές. Ανάλογα με την επιθυμητή έξοδο που επιθυμούμε και το πρόβλημα που υπάρχει ,επιλέγουμε την κατάλληλη τεχνική όπου βασίζεται σε δυο τεχνικές μάθησης . Είναι η επιβλεπόμενη μάθηση(supervised learning) και η μη επιβλεπόμενη μάθηση(unsupervised learning).

Η επιβλεπόμενη μάθηση στηρίζεται στην κατηγοριοποίηση των αντικειμένων εισόδων ενώ στην μη επιβλεπόμενη μάθηση η μάθηση πραγματοποιείται σε δυο προσεγγίσεις τον παράγοντα(agents) και την ομαδοποίηση(clustering) στα αντικείμενα εισόδου. Με άλλα

[4] Αναφορά από το άρθρο[08] της βιβλιογραφίας

λόγια στην επιβλεπόμενη μάθηση υπάρχει ένα προκαθορισμένο(predefined) σύνολο κλάσεων και ο στόχος τους είναι να εξετάσουν τα αντικείμενα και μετά να τοποθετήσουν σε μια από τις κλάσεις. Ωστόσο, στην μη επιβλεπόμενη μάθηση δεν υπάρχουν προκαθορισμένες κλάσεις αλλά αντικείμενα που ομαδοποιούνται ανάλογα με την ομοιότητα που έχουν μεταξύ τους.

2.7.1.1 Επιβλεπόμενη Μάθηση

Το πρόβλημα της επιβλεπόμενης μάθησης αναφέρεται σε δυο κλάσεις, την θετική και την αρνητική. Όταν τα δεδομένα προέρχονται από διάφορες κριτικές προϊόντων η βαθμολογία περιλαμβάνεται από 1-5 αστέρια για κάθε θέμα που αναφέρεται. Η κριτική μεταξύ 4-5 αστέρια θεωρείται θετική κριτική ενώ η κριτική 1-2 αστέρια θεωρείται αρνητική κριτική. Κάπως έτσι γίνεται η ανάλυση συναισθήματος αλλά διαφορετικά από την κλασσική ταξινόμηση στο πεδίο όπως αθλητισμός, πολιτικές επιστήμες κλπ. Στην ταξινόμηση πεδίου οι σχετικές λέξεις θεωρούνται σημαντικές ενώ στην ανάλυση σημαντικές θεωρούνται οι λέξεις γνώμης του συναισθήματος που αντικατοπτρίζουν θετικές ή αρνητικές λέξεις. Όπως καταπληκτικός, κακός, υπέροχα, σπουδαίος κλπ. Η μέθοδος της επιβλεπόμενης μάθησης εφαρμόζεται στην ταξινόμηση συναισθήματος. Υπάρχουν διάφορα είδη ταξινομητών, όπου οι κυροί ταξινομητές είναι ο Naïve Bayesian, Μηχανές διανυσματικής Υποστήριξής(SVM) καθώς επίσης είναι γραμμικοί ταξινομητές. Περιλαμβάνουν ατομικές λέξεις με χαρακτηριστικά στην κάθε ταξινόμηση. Στις περισσότερες μηχανές μάθησης , η κύρια εργασία της ταξινόμησης συναισθήματος είναι η κατασκευή από ένα σύνολο χαρακτηριστικών(set of features). Παρακάτω αναφέρουμε μερικά παράδειγμα χαρακτηριστικών:

Όροι και η συχνότητα(Terms and their Frequency): Σε αυτό το χαρακτηριστικό περιλαμβάνεται από τις ατομικές λέξεις ή τις γλωσσικά πολύ-γραμμα(n-grams) και η συχνότητα τους. Το σύστημα στάθμης TF-IDF(Term Frequency-Inverse Document) μπορεί να χρησιμοποιηθεί σε αυτή την περίπτωση. Το σύνολο των χαρακτηριστικών έχουν αρκετά καλά αποτελέσματα στην ταξινόμηση συναισθήματος.

Λέξεις γνώμης και φράσεις(Opinion word and phrases): Οι λέξεις γνώμης είναι λέξεις που συνήθως εκφράζουν θετικό ή αρνητικό συναίσθημα ,για παράδειγμα όμορφος, ωραίος, καταπληκτικό όπου εκφράζουν θετική γνώμη. Οι λέξεις κακό, άσχημα, τρομερά, εκφράζουν αρνητική γνώμη. Παρατηρούμε πως πολλές λέξεις γνώμης είναι επιρρήματα, ουσιαστικά και ρήματα όπως ανοησίες, σκουπίδια, μίσος, που μπορούν να εκφράσουν γνώμη. Σε αυτή την περίπτωση υπάρχουν και εκφράσεις γνώμης που μπορούν να εκφράσουν γνώμη ,όπως "μου

κόστισε μια περιουσία". Έτσι οι λέξεις γνώμης και εκφράσεις είναι στοιχεία της ανάλυσης συναισθήματος.

Μέρος του λόγου(Part of Speech): Σε πολλές έρευνες έχουν βρεθεί τα επίθετα(adjectives) να είναι σημαντικός δείκτης στην υποκειμενική γνώμη. Άρα τα επίθετα χρησιμοποιούνται σαν ειδικό χαρακτηριστικό.

Συντακτική εξάρτηση(Syntactic dependency): Είναι οι λέξεις που εξαρτώνται από την στήριξη των χαρακτηριστικών ή είναι δέντρα εξαρτήσεως που έχουν χρησιμοποιηθεί από πολλούς ερευνητές. Αντί να χρησιμοποιηθεί μια συνηθισμένη μηχανή μάθησης κάποιοι ερευνητές προτείνουν διάφορες τεχνικές προς την ταξινόμηση συναισθήματος. Για παράδειγμα είναι η συνάρτηση βαθμολόγησης(score function) η οποία βασίζεται από θετικές ή αρνητικές κριτικές. Επίσης έχουμε την μέθοδο συνάθροισης(Aggregation method) η όποια χρησιμοποιεί λέξεις και φράσεις από το ίδιο το πεδίο.

Αρνήσεις(Negations): Ολοφάνερα οι αρνητικές λέξεις είναι σημαντικές γιατί μπορούν να αλλάξουν την προσανατολισμό της γνώμης. Για παράδειγμα 'Δεν μου αρέσει αυτός ο Ηλεκτρονικός Υπολογιστής' εκφράζει αρνητική γνώμη. Πρέπει να ελέγχουμε σωστά το έγγραφο αν υπάρχει αρνητική γνώμη γιατί δεν είναι σε όλες τις περιπτώσεις όταν υπάρχει το '*δεν*', '*όχι μόνο*' κλπ.

Μια ενδιαφέρουσα ερευνητική κατεύθυνση είναι η διερεύνηση της μεταφερόμενης μάθησης (transfer learning) ή προσαρμογή του πεδίου(Domain Adaptation) όπως έχει αποδειχθεί ότι η ταξινόμηση συναισθήματος είναι ιδιαίτερα σημαντικό στο πεδίο που εξάγονται τα δεδομένα. Ένας ταξινομητής που χρησιμοποιεί κείμενα από ένα πεδίο θα αποδίδει ελάχιστα όταν χρησιμοποιήτε ή ελέγχεται από ένα άλλο πεδίο εγγράφου. Ο λόγος είναι γιατί οι λέξεις και η γλωσσική δομή που χρησιμοποιείται είναι από διαφορετικά πεδία. Δηλαδή σε ένα πεδίο μια λέξη μπορεί να σημαίνει θετική ενώ σε ένα άλλο πεδίο μπορεί να είναι αρνητική. Για παράδειγμα το επίρρημα 'απρόβλεπτη' σε μια κριτική για τα αυτοκίνητα 'απρόβλεπτο οδήγημα ' θεωρείται αρνητικός προσανατολισμός ενώ σε μια κριτική για ταινίες 'απρόβλεπτη πλοκή' θεωρείται θετική.

2.7.1.2 Μη Επιβλεπόμενη Μάθηση

Το πρόβλημα της μη επιβλεπόμενης μάθησης είναι ότι προσπαθούν να βρουν κρυμμένη δομή σε μη ταξινομημένα δεδομένα. Η μη επιβλεπόμενη μάθηση περιλαμβάνει πολλές τεχνικές που επιδιώκουν να συνοψίζουν και να εξηγούν τα βασικά χαρακτηριστικά της.

Η μάθηση αυτή έχει δυο προσεγγίσεις. Η πρώτη προσέγγιση είναι να διδάξει τον πράκτορα (agents)χωρίς να δώσει σαφείς κατηγοριοποιήσεις αλλά με την χρήση κάποιου είδους συστήματος για να το επιτύχει . Αυτού του είδους προσέγγιση είναι κατάλληλο για τα προβλήματα απόφασης(decision problem)γιατί δεν έχουν στόχο να ταξινομούν αλλά να λαμβάνουν αποφάσεις . Αυτή η προσέγγιση ταιριάζει με τον πραγματικό κόσμο όπου οι πράκτορες θα μπορούσαν να ανταμειφθούν για να κάνει ορισμένες ενέργειες. Επίσης είναι ισχυρή γιατί δεν αναλαμβάνει καμία προ-ανακάλυψη για ταξινόμηση. Η δεύτερη προσέγγιση ονομάζεται ομαδοποίηση και στόχος της δεν είναι να μεγιστοποιηθεί μια λειτουργία χρησιμότητας αλλά να βρει ομοιότητες στα δεδομένα εκπαίδευσης. Δηλαδή δεν υπάρχουν προκαθορισμένες κατηγορίες αλλά προσδιορίζονται από τα δεδομένα. Δηλαδή η προσέγγιση στα δεδομένα εκπαίδευσης ομαδοποιούνται με βάση των συνόλων της ομοιότητας που παρουσιάζουν μεταξύ τους. Από εμάς εξαρτάται να καθορίσουμε την σημασία που θα έχει κάθε μια από τις ομάδες που προκύπτουν.

Η μέθοδος αυτή επιτυγχάνεται με ταξινόμηση στηριζόμενη σε ορισμένες φράσεις και λέξεις που χρησιμοποιούνται για να εκφράσουν γνώμης έχοντας τις λέξεις γνώμης και τις φράσεις. Ένας αλγόριθμος που μπορεί να χρησιμοποιηθεί για αυτή την προσέγγιση της φυσικής γλώσσας είναι η Αναγνώριση Μερών του Λόγου(Part of Speech (POS)Tagging).

2.7.2 Υποκειμενικότητα των προτάσεων και ταξινόμηση συναισθήματος

Το πρόβλημα του επιπέδου είναι πως δίνεται μια πρόταση και καθορίζεται εκφράζοντας αν είναι θετική, αρνητική ή ουδέτερη(ή καθόλου) γνώμη. Το επίπεδο αυτό είναι χρήσιμό σε πολλές περιπτώσεις αν γνωρίζουμε το αντικείμενο ή την πτυχή(aspects) που υπάρχει έτσι θα μπορέσουμε να καθορίσουμε τις γνώμες αν είναι θετικές ή αρνητικές. Επίσης το επίπεδο αυτό χωρίζεται σε δυο βήματα (ή αλλιώς προβλήματα):

Βήμα 1: Γίνεται ταξινόμηση αν μια πρόταση εκφράζει γνώμη ή όχι. Το βήμα αυτό ονομάζεται ταξινόμηση υποκειμενικότητας(Subjectivity Classification) η όποια προσδιορίζει αν μια πρόταση είναι υποκειμενική ή αντικειμενική.

Βήμα 2: Η ταξινόμηση γίνεται με τις προτάσεις όταν εκφράζουν θετικά, αρνητικά ή ουδέτερα. Το βήμα αυτό ονομάζεται ταξινόμηση συναισθήματος σε επίπεδο πρότασης(Sentence level-sentiment classification).

2.7.2.1 Ταξινόμηση Υποκειμενικότητας

Η ταξινόμηση Υποκειμενικότητας ταξινομείται σε δυο κλάσεις, την υποκειμενικότητα και το αντικείμενο. Στην Αντικειμενική πρόταση(Object Sentence) εκφράζεται με κάποιο πληροφοριακό γεγονός ενώ στην υποκειμενική πρόταση(Subjective Sentence) εκφράζεται μέσω διάφορων ειδών πληροφοριών όπως γνώμες , συναίσθημα, αξιολογήσεις, καταγγελίες κλπ. Κάποια από αυτά υποδεικνύουν θετικά ή αρνητικά συναισθήματα ή και καθόλου συναίσθημα.

Αρκετές έρευνες(Wiebe, Bruce and O'Hara 1999[43]) ασχολήθηκαν με την προσέγγιση της υποκειμενικής ταξινόμησης χρησιμοποιώντας επιβλεπόμενη μάθηση με την χρήση του ταξινομητή Naïve Bayes και μαζί με ένα σύνολο από χαρακτηριστικά.

Ο Wiebe(2000)[18] πρότεινε μη επιβλεπόμενη μάθηση για να χρησιμοποιηθεί στην υποκειμενική ταξινόμηση η οποία παρέχει υποκειμενικές εκφράσεις σε μια πρόταση για να καθοριστεί αν είναι υποκειμενική πρόταση. Δηλαδή ταξινομεί μια πρόταση ως αντικειμενική εάν δεν υπάρχουν δυνατά υποκειμενικά στοιχεία. Ο τρόπος αυτός υπολογίζεται με την χαμηλή ακρίβεια(low precision) και ψηλή ανάκλαση(high recall).

2.7.2.2 Ταξινόμηση Συναισθήματος σε Πρόταση

Θεωρούμε πως μια πρόταση εκφράζεται σαν απλή γνώμη από τον κάτοχο γνώμηςΑυτή υπόθεση είναι κατάλληλη για απλές προτάσεις και με απλή γνώμη. Ωστόσο μια τέτοια πρόταση μπορεί να εκφράσει περισσότερες από μια γνώμες.

Ο Yu και Hatzivassilog(2003)[44] χρησιμοποίησαν μια μέθοδο ίδια με αυτή που έκανε ο Turney(2002) [40]με βάση το Part- of- Speech(POS), ο αλγόριθμος αυτός γίνεται σε τρία βήματα. Η διαφορά εδώ είναι ότι πήραν μια σειρά από επίθετα αντί λέξεις και δεν χρησιμοποίησαν την

μέθοδο σημειακή αμοιβαία πληροφόρηση(PMI- Point wise Mutual Information) και με ένα σύνολο log-like lihood για τον προσδιορισμό αν είναι θετικό ή αρνητικό στα επίπεδα επιθέτου, επιρρήματος, ουσιαστικό και ρήμα. Έτσι για να οριστεί ένας προσανατολισμός σε κάθε πρόταση γίνεται η χρήση του log-likelihhod και καθορίζει αν η πρόταση είναι θετική , αρνητική ή ουδέτερη.

Ένα σημαντικό σημείο που πρέπει να λάβουν υπόψη σε αυτό το επίπεδο είναι τους διάφορους τύπους προτάσεων. Οι προτάσεις εκφράζονται με διαφορετικούς τρόπους και είναι αρκετά δύσκολες. Υπάρχουν οι προ-ϋποθετικές προτάσεις(Conditional Sentences) που περιγράφουν επιπτώσεις ή υποθετικές καταστάσεις και τις συνέπειες τους. Άλλο ένα είδος είναι οι ερωτηματικές προτάσεις(Question Sentences), που κάποιες φορές εκφράζουν γνώμη και άλλες φορές όχι.

2.8 Η προέλευση του λεξικού γνώμης [5]

Στις πιο πάνω ενότητες αναφέραμε αρκετές φορές τη φράση 'γνώμη λέξεων'(opinion words). Σε αυτό το σημείο θα εξηγήσουμε από που παράγονται αυτές οι λέξεις. Σύμφωνα με την βιβλιογραφία οι λέξεις γνώμης είναι :πολικές λέξεις, συναισθηματικές λέξεις και λέξεις συναισθήματος. Οι θετικές λέξεις γνώμης μπορεί να είναι οι ακόλουθες λέξεις: καταπληκτικό, καλά, εξαιρετικά, όμορφα δηλαδή αυτές που εκφράζουν επιθυμητές καταστάσεις ενώ οι αρνητικές γνώμες λέξεις είναι όπως το βαρετός, κακός, φτωχός κλπ δηλαδή είναι αυτές που εκφράζουν ανεπιθύμητες καταστάσεις. Επίσης εκτός από ατομικές λέξεις που αναφέραμε υπάρχουν και οι γνώμες φράσεων(opinion phrases). Έτσι όλες αυτές οι λέξεις γνώμης και γνώμες φράσεων ονομάζονται λεξικό γνώμης(opinion lexicon).

Στις λέξεις γνώμης υπάρχουν δυο τύποι, ο βασικός τύπος(base type) και ο συγκριτικός τύπος(comparative type). Στον δεύτερο τύπο εκφράζονται συγκριτικές και υπερθετικές γνώμες όπως για παράδειγμα οι λέξεις καλύτερα, χειρότερα. Έχοντας βάση τα επίθετα και επιρρήματα των λέξεων όπως η λέξη 'καλό' και 'κακό'. Σε αντίθεση με τον βασικό τύπο οι λέξεις δεν εκφράζουν κάποια γνώμη για το αντικείμενο αν είναι για σύγκριση. Όπως για παράδειγμα 'Ο ηλεκτρονικός υπολογιστή-Χ είναι καλύτερος από τον ηλεκτρονικό υπολογιστή-Υ', εδώ δεν εκφράζουν κάποια γνώμη για τους δύο ηλεκτρονικούς υπολογιστές.

[5] Αναφορά από το άρθρο[24] της βιβλιογραφίας

Για να συλλέξουμε και να δημιουργήσουμε λίστα από λέξεις γνώμης υπάρχουν τρεις προσεγγίσεις:

- Χειροκίνητη προσέγγιση(manual approach)
- Προσέγγιση βάση λεξικού(dictionary-base approach)
- Προσέγγιση βάση συλλογής(corpus-based approach)

Η πρώτη προσέγγιση είναι αρκετά χρονοβόρα και δεν μπορείς να το κάνεις μόνο του, αλλά γίνεται και με συνδυασμό με τις άλλες προσεγγίσεις που είναι αυτοματοποιημένες.

Προσέγγιση βάση λεξικού: Αυτή η προσέγγιση είναι μια απλή τεχνολογία που γίνεται χρησιμοποιώντας ένα μικρό σύνολο από λέξεις γνώμης και από ένα online λεξικό (για παράδειγμα το WordNet). Η διαδικασία της προσέγγισης αυτής είναι ότι πρέπει πρώτα να γίνει μια μικρή συλλογή χειροκίνητη για τις λέξεις γνώμης γνωρίζοντας τον προσανατολισμό και στην συνέχεια κάνουν αναζήτηση από το online λεξικό για συνώνυμα και αντωνυμίες. Η διαδικασία σταματά όταν βρει νέες λέξεις. Το μειονέκτημα της προσέγγισης αυτής είναι ότι δεν είναι σε θέση να βρει λέξεις γνώμης από ένα συγκεκριμένο πεδίο προσανατολισμού.

Προσέγγιση βάση συλλογής και συνέπειας συναισθήματος: Η μέθοδος αυτή βασίζεται σε συντακτική ή ταύτιση προτύπων και επίσης μια λίστα από λέξεις γνώμης για να βρουν άλλες γνώμες λέξεων από ένα μεγάλο πεδίο(corpus). Η διαδικασία περιλαμβάνει μια λίστα λέξεων από επίθετα και ένα σύνολο γλωσσικών που χρησιμοποιούνται για περιορισμούς ή συμβάσεις για αναζήτηση περισσοτέρων επιθέτων λέξεων γνώμης. Ένας περιορισμός για παράδειγμα είναι το 'ΚΑΙ' , για παράδειγμα σε μια πρόταση 'Ο ηλεκτρονικός υπολογιστής είναι εύχρηστος και όμορφος'. Εδώ έχουμε δυο θετικές γνώμες για το ίδιο θέμα αλλά μπορεί να συναντήσουμε σε μια πρόταση να υπάρχει θετική και αρνητική γνώμη. Αυτοί οι κανόνες περιορισμού ή σύνδεσμοί όπως το : 'ή', 'άλλα', 'είτε ', 'ούτε' κλπ ονομάζονται συναίσθημα συνέπειας(sentiment consistency). Η τεχνική εδώ εφαρμόζεται σε ένα μεγάλο σώμα για να προσδιορίσει αυτά τα δυο συνδεδεμένα επίθετα αν έχουν το ίδιο ή διαφορετικό προσανατολισμό. Από εδώ παράγει δυο σύνολα λέξεων το θετικό και αρνητικό όπου επιλέγεται ανάλογα.

Ο Qiu etal[34] προτείνει μια άλλη μέθοδο για να εξάγει λέξεις συναισθήματος από ένα συγκεκριμένο πεδίο για κριτικές. Η κεντρική ιδέα της μεθόδου είναι να εκμεταλλεύεται ορισμένες

συντακτικές σχέσεων της λέξεις γνώμης και αντικείμενα χαρακτηριστικών για εξαγωγή. Οι λέξεις γνώμης συνδέονται με τα χαρακτηριστικά του αντικειμένου με κάποιους τρόπους και ταυτολογούνται με αυτές. Η διαδικασία τελειώνει όταν δεν μπορούν να βρεθούν άλλες λέξεις γνώμης ή χαρακτηριστικά .

Όπως έχουμε δει η προσέγγιση βάση συλλογής χρησιμοποιείται για προσδιορισμό όλων των τύπων των λέξεων γνώμης αλλά δεν είναι τόσο αποτελεσματικό όσο η προσέγγιση βάση λεξικού. Είναι δύσκολο για αυτήν την προσέγγιση να προετοιμάσει ένα μεγάλο φάσμα από όλες τις αγγλικές λέξεις του λεξικού. Το θετικό της προσέγγισης είναι ότι δεν βασίζεται σε λεξικά. Μπορεί να βοηθήσει για ένα συγκεκριμένο εύρος των λέξεων γνώμης, περιεχόμενο και τον προσανατολισμό αν χρησιμοποιείται μια συλλογή συγκεκριμένου πεδίου. Επίσης σε μια πρόταση μπορεί να περιλαμβάνεται από τις λέξεις γνώμης λεξικού(opinion lexicon) και φράσεις που εκφράζουν γνώμη, όμως σε μια πρόταση μπορεί να μην εκφράζει κάποια γνώμη και θεωρείται πως είναι ουδέτερη. Για παράδειγμα : 'Ψάχνω μια καλή ασφάλεια ζωής για την οικογένεια μου', όπως βλέπουμε δεν εκφράζει κάποια γνώμη'.

2.9 Επίπεδο χαρακτηριστικού[6]

Το επίπεδο χαρακτηριστικού παρέχει κάποιες χρήσιμες πληροφορίες που δεν μπορούν να παρέχονται στην ταξινόμηση επιπέδου εγγράφου και στο επίπεδο πρότασης. Δεν έχουν την απαραίτητη λεπτομέρεια που είναι αναγκαία. Σε ένα τυπικό κείμενο γνώμης, ο συγγραφέας γράφει θετικές και αρνητικές γνώμες για τα χαρακτηριστικά ενός αντικειμένου ανεξάρτητα αν το γενικό συναίσθημα για το κύριο αντικείμενο είναι θετικό ή αρνητικό. Έτσι το επίπεδο εγγράφου και επίπεδο πρότασης δεν τα παρέχουν.

Ο προσδιορισμός των χαρακτηριστικών αντικειμένου(object features) γίνεται με την εξαγωγή από τα χαρακτηριστικά στα οποία περιγράφουν οι προτάσεις.Για παράδειγμα 'Η ποιότητα εικόνας αυτής της κάμερας είναι εκπληκτική', το χαρακτηριστικό του αντικειμένου εδώ είναι η 'ποιότητα της εικόνας'. Μπορεί να χρησιμοποιηθεί η επιβλεπόμενη μάθηση και η μη επιβλεπόμενη μάθηση. Η τεχνική βασίζεται σε ουσιαστικά και φράσεις ουσιαστικών των χαρακτηριστικών . Υπάρχουν πολλές τεχνικές που μπορούν να εφαρμοστούν, όπως Conditinal Radom field(CRF), Hidden Markov Models(HMM) κ.α.

[6] Αναφορά από το βιβλίο[25]της βιβλιογραφίας

Επίσης σημαντικό σημείο είναι ο προσανατολισμός της γνώμης του χαρακτηριστικού για να καθοριστεί αν το χαρακτηριστικό είναι θετικό, αρνητικό ή ουδέτερο. Για την πιο πάνω πρόταση που δόθηκε η 'ποιότητα εικόνας' είναι θετική. Η προσέγγιση βάση λεξικών χρησιμοποιεί λέξεις γνώμης και φράσεις στην πρόταση και καθορίζει τον προσανατολισμό της γνώμης του χαρακτηριστικού. Υπάρχουν πολύ τύποι επιβλεπόμενης μάθησης που είναι δυνατόν να χρησιμοποιηθούν στην προσέγγιση αυτή.

2.9.1 Εξόρυξη Χαρακτηριστικού

Η εξαγωγή έκφρασης χαρακτηριστικού εννοούμε τον στόχο της γνώμης στον οποίο απευθύνεται η λέξη γνώμης. Οι έρευνες που γίνονται για την εξαγωγή γνώμης γίνονται κυρίως σε online κριτικές προϊόντων. Υπάρχουν δυο μορφές κριτικών στο διαδίκτυο που χρειάζονται διαφορετικές τεχνολογίες για την εξαγωγή του χαρακτηριστικού. Είναι οι εξής :

Μορφή 1 : Εξαγωγή χαρακτηριστικού από θετικές και αρνητικές γνώμες

Η κριτική εδώ γίνεται πολύ σύντομη, περιγράφοντας μερικά χαρακτηριστικά για θετικά και αρνητικά σχόλια. Τα σχόλια μπορεί να είναι μικρές φράσεις ή σύντομες προτάσεις, διαχωρίζοντας με το κόμμα, και,& , άλλα, απόστροφο κλπ. Όπως αναφέραμε και πιο πάνω γίνεται χρήση των τεχνικών CRF [7], HMM. Στην τεχνική CRF το αντικείμενο του χαρακτηριστικού μπορεί να εκφράζει ουσιαστικά, επίθετα, ρήματα ή επιρρήματα. Η εξαγωγή των χαρακτηριστικών εκτελείται με το ταίριασμα προτύπων μαζί με κάθε τμήμα της πρότασης μέσα σε νέα κριτική για εξαγωγή αντικειμένων χαρακτηριστικών.

Μορφή 2: Ελεύθερη Μορφή

Η κριτική εδώ γίνεται ελεύθερη γράφοντας ολοκληρωμένες προτάσεις και η εξαγωγή του χαρακτηριστικού εφαρμόζεται με αλγόριθμο. Ωστόσο οι έρευνες δείχνουν ότι αυτές οι προτάσεις δεν είναι τόσο αποτελεσματικές γιατί περιέχουν ένα μεγάλο ποσοστό από θόρυβο.

μια μη κατευθυνόμενη μέθοδος περιγράφεται στην έρευνα[20]. Για εύρεση σαφών εκφράσεων χαρακτηριστικών. Η έρευνα αυτή αποτελείται από δυο βήματα. Τα βήματα αυτά σε πρώτη φάση εντοπίζουν να βρουν τα πιο συχνά ουσιαστικά και φράσεις ουσιαστικών(ή ομάδες)

[7] Είναι μια κατηγορία για μέθοδο στατιστική μοντελοποίηση και εφαρμόζεται στην αναγνώριση προτύπων και την μηχανική μάθηση.

αναγνωρίζοντας από ένα αναγνωριστεί POS(Part of Speech). Το πόσο συχνά εμφανίζονται τα ουσιαστικά και οι φράσεις ουσιαστικών αριθμούνται και κρατούνται . Με άλλα λόγια το βήμα αυτό είναι όταν οι άνθρωποι σχολιάζουν για διάφορα χαρακτηριστικά ενός προϊόντος ,με αποτέλεσμα το λεξικό να ποικίλει. Έτσι αυτά τα ουσιαστικά που εμφανίζονται πιο πολύ είναι και αυτά που είναι σημαντικά χαρακτηριστικά.

Το επόμενο βήμα είναι να ψάχνουν να βρουν τα σπάνια χαρακτηριστικά με την χρήση των λέξεων γνώμης. Οι λέξεις γνώμης συνήθως είναι επίθετα και επιρρήματα που εκφράζουν θετική και αρνητική γνώμη. Έτσι οι λέξεις γνώμης μπορούν να εφαρμοστούν στα μη συχνά χαρακτηριστικά και να χρησιμοποιηθούν για να εξάγουν μη συχνά χαρακτηριστικά.

Κατά την εξαγωγή των χαρακτηριστικών υπάρχουν δυο προβλήματα,όπου είναι τα πιο κάτω:

Πρόβλημα 1 : Συνώνυμες Ομάδες(Group Synonyms)

Όταν θέλουμε να περιγράψουμε κάτι μπορούμε να χρησιμοποιήσουμε διάφορες λέξεις και εκφράσεις αλλά να ανήκουν στο ίδιο χαρακτηριστικό. Για παράδειγμα η λέξη 'φωτογραφία' και 'εικόνα' είναι στην ίδια κατηγορία χαρακτηριστικών για τις κριτικές της ψηφιακής κάμερας. Αλλά οι λέξεις 'εικόνα ' και 'ταινία' μπορεί να είναι συνώνυμες στις κριτικές των ταινιών αλλά δεν είναι συνώνυμες για τις κριτικές της ψηφιακής κάμερας γιατί η ΄ εικόνα ΄ είναι για την φωτογραφία και η 'ταινία ' είναι το βίντεο. Έτσι ο αλγόριθμος κάνει συγχώνευση για κάθε νέο χαρακτηριστικό που ανακαλύπτει και το ταξινομεί σωστά. Όλα αυτά γίνονται με την χρήση του WordNet.

Πρόβλημα 2: Χαρτογράφηση σε σιωπηρά χαρακτηριστικά(Mapping to implicit features)

Η εξαγωγή χαρακτηριστικών μπορεί να γίνει σε πολλούς δέκτες χαρακτηριστικών. Δηλαδή τα επίθετα και τα επιρρήματα ως συνήθως έχουν κοινό τύπο δεικτών χαρακτηριστικών και μπορούν να τροποποιηθούν ή να περιγράψουν ένα χαρακτηριστικό ή μια ιδιότητα του αντικειμένου. Για παράδειγμα το επίθετο ' μικρό' περιγράφει το μέγεθος ενός αντικειμένου, οπότε κατατάσσεται στο χαρακτηριστικό του μεγέθους. Άλλος τρόπος είναι να γίνουν χειροκίνητα οι αντιστοιχίες των λέξεων για να χρησιμοποιηθούν μετά. Όμως πρέπει να γίνεται με πολύ προσοχή και με τους δυο τρόπους γιατί σε αρκετές περιπτώσεις το νόημα μιας λέξης δεν είναι πάντα το ίδιο. Για παράδειγμα η λέξη 'μικρή'στην παρακάτω πρόταση :'Η Μαρία είναι

μικρή' δεν αναφέρεται στο μέγεθος της Μαρίας αλλά στην ηλικία . Εν τούτοις, δεν είναι και τόσο αποτελεσματική αυτή η προσέγγιση.

2.9.2 Καθορισμός Προσανατολισμού Γνώμης

Όπως έχουμε δει σε κάθε πρόταση έχει κάποια χαρακτηριστικά, όπου αυτά τα χαρακτηριστικά παίρνουν προσανατολισμό γνώμης. Οι μέθοδοι που υπάρχουν έχουν δυσκολία στον εντοπισμό του προσανατολισμού σε ανάμικτες γνώμες στις προτάσεις γιατί στα blogs,forum κλπ γράφουν απεριόριστο κείμενο. Η προσέγγιση γίνεται βάσει λεξικού για να αποφεύγει προβλήματα πολυπλοκότητας. Η προσέγγιση αυτή βασίζεται σε ένα λεξικό γνώμης δηλαδή μια λίστα από λέξεις γνώμης και φράσεις για να προσδιοριστεί ο προσανατολισμός γνώμης της πρότασης. Η διαδικασία περιλαμβάνει τέσσερα επίπεδα θεωρώντας δεδομένο ότι τα χαρακτηριστικά είναι γνωστά.

Η προσέγγιση περιγράφεται εντοπίζοντας σε πρώτο βήμα όλες τις λέξεις γνώμης και φράσης μέσα σε μια πρόταση. Στην θετική γνώμη την βαθμολόγηση με +1 και στην αρνητική γνώμη με -1. Σε επόμενο βήμα πρέπει να εντοπιστούν οι μετατοπιστές γνώμης(ή μετατοπιστές σθένους) από λέξεις και φράσεις που αλλάζουν τον προσανατολισμό της γνώμης. Όπως το ' δεν', 'κανείς', 'πουθενά', ' ίσως', 'πρέπει', 'θα μπορούσε να', κλπ. Επίσης πρέπει να ελέγχεται όπου υπάρχει και η λέξη 'αλλά' . Ο προσανατολισμός εδώ θα μπορούσε να αντιστρέψει το συναίσθημα που υπάρχε, βέβαια εξαρτάται σε ποια θέση της πρότασεις βίσκεται. Ακόμη, στο τελευταίο βήμα γίνεται η χρήση της συνάρτησης της συνάθροισης γνώμης(Opinion Aggregation Function) για να προσδιοριστεί ο προσανατολισμός της γνώμης σε κάθε χαρακτηριστικό της πρότασης. Η συνάρτηση αυτή δεν είναι αποτελεσματική σε όλες τις περιπτώσεις.

2. 10 Συγκριτικές προτάσεις [8]

Το πρόβλημα των συγκριτικών προτάσεων είναι ότι βασίζονται σε ίδιες ή σε διαφορετικές από ένα αντικείμενο. Στην σύγκριση συνήθως χρησιμοποιούν τον συγκριτικό ή τον υπερθετικό βαθμό που μπορεί να είναι επίθετο ή επίρρημα. Δεν έχουν διαφορά στην έννοια αυτές οι προτάσεις αλλά είναι διαφορετική η σύνταξη τους. Πιο κάτω αναλύουμε τα παραθετικά των επιθέτων και επιρρημάτων. Παραθετικά εννοούμε τον συγκριτικό και υπερθετικό βαθμό.

[8] Αναφορά από το άρθρο[24]της βιβλιογραφίας

Τα παραθετικά επίθετα έχουν τον συγκριτικό βαθμό που φανερώνει μια ιδιότητα που έχει ένα ουσιαστικό τονίζοντας ότι αυτό το ουσιαστικό έχει τη συγκεκριμένη ιδιότητα σε μεγαλύτερο βαθμό από κάποιο άλλο ουσιαστικό που έχει επίσης την ίδια ιδιότητα.

Παράδειγμα : Ο Γιάννης είναι ψηλότερος(ή πιο ψηλός) από τον Κώστα.

Για να συγκρίνουμε δύο ουσιαστικά υπάρχουν δύο τρόποι :

1ος Τρόπος: Στα μονολεκτικά προσθέτουμε στο επίθετο την κατάληξη **–***ότερος, -ότερη, -ότερος, -ύστερος, -ύστερη, -ύστερο* και βάζουμε δίπλα του τη λέξη 'από'.

Παράδειγμα : Ο Κωνσταντίνος είναι μεγαλύτερος από τον Τάσο. Η Ειρήνη είναι ψηλότερη από την Χρυσούλα.

2ος Τρόπος: Στα περιφραστικά χρησιμοποιούμε το επίθετο όπως είναι χωρίς να το αλλάξουμε αλλά πριν από αυτό βάζουμε το επίρρημα '*πιο*'.

Παράδειγμα : Ο Μάριος είναι πιο ψηλός

Τα παραθετικά επίθετα στον υπερθετικό βαθμό υπάρχουν δύο τρόποι:

1ος τρόπος: Ο σχετικός υπερθετικός βαθμός που φανερώνει μια ιδιότητα που έχει ένα ουσιαστικό, τονίζοντας ότι αυτό το ουσιαστικό έχει την ιδιότητα αυτή στο μεγαλύτερο βαθμό απ' όλα τα ουσιαστικά που χαρακτηρίζονται από την ίδια ιδιότητα.

Παράδειγμα: Η Νίκη και όλοι οι συμμαθητές της είναι έξυπνοι. ή Η Νίκη είναι η πιο εξυπνότερη στην τάξη.

Δηλαδή ο υπερθετικός βαθμός σχηματίζεται από μονολεκτικά με το οριστικό άρθρο και το συγκριτικό βαθμό του επιθέτου(πχ ο πλουσιότερος). Και επίσης, περιφραστικά όπως και ο συγκριτικός βαθμός με το οριστικό άρθρο μπροστά(πχ ο πιο λεπτός)

2ος Τρόπος: Ο απόλυτος υπερθετικός βαθμός φανερώνει μια ιδιότητα που έχει ένα ουσιαστικό σε πολύ μεγάλο βαθμό χωρίς να γίνεται σύγκριση με άλλο ουσιαστικό.

Παράδειγμα: Η Χριστίνα είναι εξυπνότατη ή Η Χριστίνα είναι πολύ έξυπνη.

Δηλαδή ο απόλυτος υπερθετικός βαθμός σχηματίζεται αν προσθέσουμε στο επίθετο την κατάληξη *–τάτος, -τάτη, -τάτο* ή προσθέσουμε το ποσοτικό επίρρημα *πολύ ή πάρα πολύ*(πχ πλουσιότατος ή(πάρα)πολύ πλούσιος).

Τα παραθετικά επιρρήματα έχουν τα επιρρήματα που σχηματίζονται από επίθετα.

Παράδειγμα: Αγόρασα φτηνά την τσάντα.

Ο συγκριτικός βαθμός στα επιρρήματα σχηματίζεται από τα μονολεκτικά στον πληθυντικό των ουδέτερων του συγκριτικού βαθμού στο αντίστοιχο επίθετο(πχ ακριβά -> ακριβότερα). Ακόμη σχηματίζεται με περιφραστικά ποσοτικά επιρρήματα '*πιο*'(πχ ακριβά -> πιο ακριβά). Ο υπερθετικό βαθμός στα επιρρήματα σχηματίζεται σε μονολεκτικά από τον πληθυντικό του ουδέτερου υπερθετικού βαθμού του επιθέτου (πχ ακριβά -> ακριβότερα). Επίσης σχηματίζεται και περιφραστικά με ποσοτικά επιρρήματα με το πολύ ή πάρα πολύ και με το επίρρημα στο θετικό βαθμό (πχ πιο ακριβά ->(πάρα) πολύ ακριβά).

Αυτές οι προτάσεις που αναφέραμε πιο πάνω είναι προτάσεις που έχουν συνήθως λέξεις κλειδιά ή φράσεις που υποδεικνύουν σύγκριση. Σύμφωνα με τους N.Jindal και B.Liu χρησιμοποιήσαν ένα σύνολο 83 λέξεων κλειδιά και φράσεις κλειδιά. Μερικά παραδείγματα είναι τα παρακάτω:

- Συγκριτικά επίθετα και συγκριτικά επιρρήματα πχ *περισσότερο, λιγότερο, καλύτερος*

-Υπερθετικά επίθετα και υπερθετικά επιρρήματα πχ *περισσότερα, καλύτερο, τουλάχιστο*

-Διάφορες ενδεικτικές λέξεις όπως *ίδια, παρόμοια, διαφέρουν, νίκη, καθώς*

Η εξαγωγή αντικειμένων και χαρακτηριστικών αντικειμένων γίνονται με τις μεθόδους που αναφέρθηκαν στην ενότητα 2.9.1. Επομένως, οι συγκριτικές προτάσεις δεν μπορούν να εκφράσουν άμεσα αν είναι θετική ή αρνητική γνώμη από τις κανονικές προτάσεις. Επειδή, υπάρχουν πολλαπλά αντικείμενα για κατάταξη που βασίζονται σε κοινά χαρακτηριστικά. Έτσι οι συγκριτικές προτάσεις συγκρίνουν δυο σύνολα από αντικείμενα για ανάλυση μιας πρότασης γνώμης.

2.11 Διαδικασία εξαγωγής Γνώμης από τον Ιστό[9]

Πιο κάτω περιγράφεται η λειτουργία του συστήματος εξόρυξης γνώμης από ένα θορυβώδης κείμενο από τον ιστό(Web). Η λειτουργία απεικονίζεται στην εικόνα 2.1 περιγράφοντας τις πέντε σημαντικές λειτουργίες του.

Μονάδα συλλογής δεδομένων(Data acquisition module) : Είναι ένα σύνολο από προκαθορισμένες ιστοσελίδες αποθηκεύοντας τα περιεχόμενα τους χρησιμοποιώντας το προκαθορισμένο σχήμα. Κάθε καταχώριση(blogs κλπ) αποθηκεύεται μεμονωμένα σε ξεχωριστό στοιχείο δεδομένων και όλα συνδέονται με τα μετά-δεδομένα(metadata) για χρήσιμες πληροφορίες.

Δημιουργία βάση γνώμης(Creating knowledge –base opinion mining) : Σε κάθε περιοχή των λέξεων υπάρχει μια συλλογή από μονάδες λέξεων και φράσεων που χρησιμοποιούνται για την έκφραση των χαρακτηριστικών των προϊόντων ή τις απόψεις. Για παράδειγμα το περιεχόμενο των οχημάτων η φράση *χαμηλή τιμή* υποδηλώνει θετική γνώμη ενώ το *χαμηλή επίδοση* εκφράζει αρνητική γνώμη. Έτσι το σύστημα πρέπει να είναι καλά εκπαιδευμένο για διάφορους τομείς. Επιπλέον, το πλαίσιο αυτό βασίζεται στην προσέγγιση γνώσης δημιουργώντας γνώση με βάση την ανθρώπινη βοήθεια χρησιμοποιώντας λέξεις από το WordNet. Το σύστημα ακόμη μαθαίνει τα πρότυπα που εκφράζουν γνώμες από τα κείμενα.

Προ-επεξεργασία για καθαρισμό θορύβου από κείμενο(Pre-processor for cleaning noisy text):Το θορυβώδες κείμενο προσδιορίζεται από τέσσερα τμήματα έχοντας δικούς τους μεθόδους(εικόνα 2.2)και αναφέρονται πιο κάτω.

Καθαρισμός συμβόλων : Σε αυτό το πεδίο γίνεται η αφαίρεση των σημείων στίξεων για να εξασφαλιστεί ένα καλύτερο αποτέλεσμα. Σημεία στίξης εννοούμε όταν στο κείμενο υπάρχουν : ???, I), II), ==30, IV κλπ. Επίσης αυτά που επαναλαμβάνονται αφαιρούνται. Όσα σύμβολα δεν είναι ένα μέρος του καθορισμένου συνόλου του αλφαβήτου ή των σημείων στίξεων αφαιρούνται. Μετά, όταν αφαιρεθεί το σύμβολο γίνετε έλεγχος δεξιά και αριστερά και αν τα δύο τμήματα συνδέονται για να γίνει λέξη ή όχι. Αν δεν γίνεται λέξη υπάρχει κενό μεταξύ τους.

Καθορισμός όριου πρότασης : Τα όρια πρότασης δεν είναι θορυβώδες κείμενο αλλά υπάρχουν κάποιοι κανόνες βάση του συστήματος που εντοπίζουν τις λογικές προτάσεις και εισάγουν αυτά

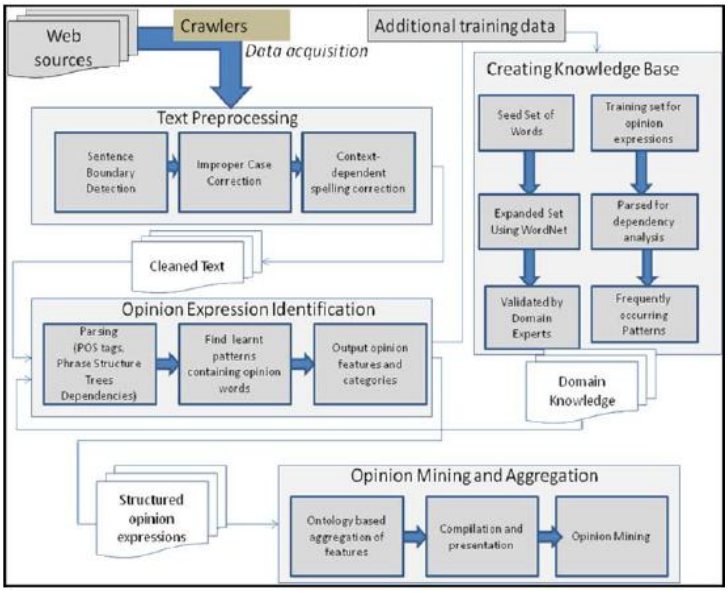
[9] Αναφορά από το άρθρο[10] της βιβλιογραφίας

τα όρια. Οι κανόνες είναι οι εξής. Το σύμβολο ‘ . ’ είναι ένα όριο πρότασης που καθορίζει ότι δεν έχει προηγηθεί ένα προκαθορισμένο σύνολο λέξεων(ltd. , Mr. , Προφεσόρε. κλπ) καθώς και οι μικρές λέξεις(οκ, σε κλπ) διατηρούνται. Για να συγχωνευθούν οι προτάσεις πρέπει να υπάρξουν κανόνες .Κάθε νέα πρόταση ξεκινά πάντα με το πρώτα γράμμα να είναι κεφαλαίο. Αν στην επόμενη γραμμή ξεκινά με μικρό αλφάβητο υπάρχει το ενδεχόμενο η προηγούμενη πρόταση να μη έχει ολοκληρωθεί και να είναι η συνέχεια της, έτσι συγχωνεύονται αυτές οι δύο γραμμές. Επίσης μπορεί στο τέλος τις γραμμής μια λέξη να μην είναι ολοκληρωμένη και στην επόμενη γραμμή να είναι η συνέχεια της και το πρώτο γράμμα να είναι μικρό. Τότε συγχωνεύονται για να ενωθεί η λέξη. Ακόμη αν σε μια νέα γραμμή ξεκινά με λέξη λεξικού τότε συγχωνεύεται με την προηγούμενη γραμμή έχοντας κενό μεταξύ τους.

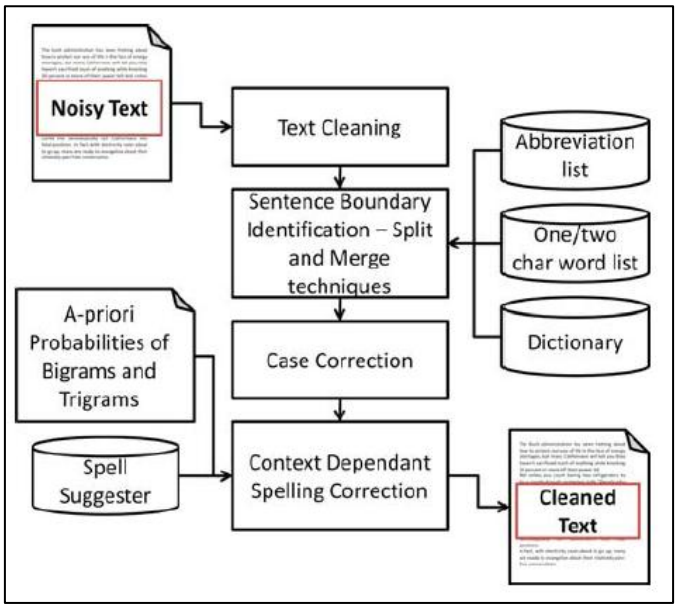
Πλαίσιο βάση διόρθωση ορθογραφίας : Το μεγαλύτερο πρόβλημα που αντιμετωπίζει η εξόρυξη γνώμης είναι τα ορθογραφικά λάθη μέσα στο κείμενο. Αν είναι γραμμένη λάθος τότε μπορεί να χαθεί από το πεδίο της εξόρυξης γνώμης. Το ίδιο ισχύει και με ένα χαρακτηριστικό γνώμης αν είναι λανθασμένο γιατί μπορεί να οδηγήθει σε αντίθετη γνώμη.

Προσδιορισμός έκφρασης γνώμης και εξαγωγή : Για τον καθορισμό του κειμένου υποβάλλεται το POS, την εξαγωγή δομής φράσης και την εξάρτηση ανάλυσης. Τα δυο τελευταία χρησιμοποιούνται για την ανάλυση γνώμης. Η εξαγωγή γνώμης είναι μια προσέγγιση με βάση γνώσης. Το σύστημα λαμβάνει πρότυπα με βάση γνώσης με νέα δεδομένα και κατά την διάρκεια μπορεί να εντοπιστούν πρότυπα που είναι ίδια και δεν είναι μέρος της βάσης γνώμης.

Αφομοίωση και συσσωμάτωση γνώμης: Στην διαδικασία αυτή το σύστημα αφομοιώνει το πεδίο της οντολογίας που εξάγει τις γνώμες και συσσωματώνει τις πληροφορίες σε πολλαπλά επίπεδα και λειτουργίες.



Εικόνα 2.1: Η αρχιτεκτονική της εξόρυξης γνώμης



Εικόνα 2.2 : Η 3η σημαντική λειτουργία για την διαδικασία εξαγωγής γνώμης η προ-επεξεργασία για καθορισμός θορύβου από κείμενο

2.12 Σχετικές Έρευνες

Οι Hatzivassiloglou and McKeown[17] αναλύουν και προτείνουν την χρήση ενός αλγορίθμου σε επιβλεπόμενη μάθηση για να συμπεραίνουν το σημασιολογικό προσανατολισμό για επίθετα από τους περιορισμούς που υπάρχουν. Χρησιμοποιώντας μια λίστα seedword για να προσδιορίσουν αν μια πρόταση περιέχει θετικά ή αρνητικά συναισθήματα. Ο Turney[40] πρότεινε μια

προσέγγιση για την επέκταση της λίστας με τις φράσεις έχοντας από έξι συντακτικά μοτίβα. οι Οι Υι και Nasukawa(2005) έφτιαξαν ένα λεξικό πολικότητας σαν αυτό που έκανε ο Turney για να εξάγει τα συναισθήματα από τις προτάσεις.

Οι Turney & Littman(2003)[6] υπολογίζουν το σημασιολογικό προσανατολισμό των λέξεων, με βάση τη σημασιολογική σύνδεσή τους, με προκαθορισμένες θετικές και αρνητικές λέξεις. Οι Kanayama & Nasukawa(2006)[21] προτείνουν με την μέθοδο της μη επιβλεπόμενης μάθησης, για την εξαγωγή συντακτικών δομών, που καθορίζουν τον προσανατολισμό της προτάσης στον τομέα σημασιολογική ανάλυση(domain oriented semantic analysis). Αυτές οι δυο εργασίες χρησιμοποιούν κανόνες σύνδεσης(conjunction rules) για να εξάγουν λέξεις - εξαρτώμενες από το περιεχόμενο(context-dependent opinion words) κάνοντας αναζήτηση σε μεγάλα κείμενα(large corpora).

Ο Pavel[35]προτείνει την χρήση του WordNet για την εξόρυξη γνώμης και οι Esuli και Sebastiani [11]πρότειναν το SentiWordNet για λεξικά με τρείς κατηγορίες συναισθήματος(θετικότητα, αρνητικότητα και αντικειμενικότητα).

Οι Pang & Lee(2008)[31] προτείνουν μεθόδους που χρησιμοποιούν τεχνικές μηχανικής μάθησης. Αυτοί οι μέθοδοι εκπαιδεύουν ένα ταξινομητή συναισθήματος(sentiment classifier) με σκοπό να καθορίσουν ποιες είναι οι θετικές, αρνητικές και ουδέτερες γνώμες μέσα σε ένα κείμενο. Τα χαρακτηριστικά που χρησιμοποιούνται για να εκπαιδεύσουν τους ταξινομητές είναι τα unigrams ή bigrams(n-grams μεγέθους 1 ή 2 αντίστοιχα)από τους ερευνητές de Kok & Brouwer, 2012[09]. Ωστόσο, το μειονέκτημα των μεθόδων αυτών είναι ότι χρησιμοποιούν ένα χειροκίνητο καθορισμό ετικετών(manual labeling) που απαιτείται σε μεγάλα σύνολα μηνυμάτων του tweeter(tweets). Ακόμη ο καθορισμός ετικετών πρέπει να εκτελεστεί σε κάθε ξεχωριστό πεδίο ενδιαφέροντος ώστε να επιτευχθούν ικανοποιητικά επίπεδα εκπαίδευσης για τον ταξινομητή του συγκεκριμένου πεδίου(Aue & Gamon, 2005[02]).

Ο προσανατολισμός της γνώμης βασίζεται στα χαρακτηριστικά(Feature-based opinion summarization) σύμφωνα με τους Hu & Liu(2004)[19]που επιτρέπει στον πελάτη να ξεχωρίζει στην αλυσίδα των κριτικών που αφορούν σε ένα συγκεκριμένο χαρακτηριστικό. Επίσης σε μια άλλη εργασία τους (την ίδια χρονική περίοδο) οι ίδιοι αυτοί συγγραφείς προτείνουν διάφορες τεχνικές εξόρυξης δεδομένων για να συνοψίσουν τις απόψεις των υφιστάμενων πελατών προβλέποντας τον σημασιολογικό προσανατολισμό των λέξεων. Παρατηρούν πως οι άνθρωποι χρησιμοποιούν συχνά διαφορετικές φράσεις ή λέξεις, για να περιγράψουν το ίδιο

χαρακτηριστικό. Η ομαδοποίηση(grouping) αυτών των χαρακτηριστικών είναι κρίσιμη για την αποτελεσματική περίληψη γνώμης(effective opinion summary). Οι λέξεις που περιγράφουν ένα χαρακτηριστικό του προϊόντος αναφέρονται ως 'Χαρακτηριστικό έκφραση' (feature expression). Η έρευνα του Zhai et al(2010)[45] παρουσιάζει την ομαδοποίηση των χαρακτηριστικών έκφρασης που χρησιμοποιείται για ένα συγκεκριμένο χαρακτηριστικό του προϊόντος με μια ημι-επιβλεπόμενη μέθοδο μάθησης. Αναφέρεται σαν Μεγιστοποίηση Προσδοκίας(Expectation Maximization).

2. 13 Περίληψη

Σε αυτό το κεφάλαιο παρουσιάσαμε τον ορισμό της εξόρυξης Γνώμης(Opinion Mining) και την ανάλυση συναισθήματος(sentiment analysis). Είδαμε την ιστορική αναδρομή των εννοιών , τις διάφορες έννοιες που μπορούμε να συναντήσουμε μέσα από μια ποικιλία ερευνητικών κατευθύνσεων αλλά και τα πολλά προβλήματα της έρευνας που έχει η φυσική επεξεργασία γλώσσας. Η ανάλυση συναισθήματος έχει αποτελέσει αντικείμενο συνεχούς ενδιαφέροντος σε πολλά συνέδρια τα τελευταία έτη. Επομένως, η εξόρυξης γνώμης στοχεύει στην προσδιορισμό της έκφρασης αν μια γνώμη είναι θετική ή αρνητική.

Επίσης είδαμε τα τρία επίπεδα ανάλυσης , το επίπεδο εγγράφου, επίπεδο πρότασης και το επίπεδο χαρακτηριστικού. Όπου το επίπεδο χαρακτηριστικού παρέχει την απαραίτητη λεπτομέρεια που δεν έχουν τα άλλα επίπεδα. Σε ένα έγγραφο ή σε μια πρόταση ο συγγραφέας γράφει θετικές και αρνητικές γνώμες ανεξάρτητα αν το γενικό συναίσθημα στο κύριο αντικείμενο είναι θετικό ή αρνητικό. Έτσι το χαρακτηριστικό επίπεδο εντοπίζει αυτά τα χαρακτηριστικά.

Γενικά όλες οι έρευνες για ανάλυση συναισθήματος είναι δύσκολες στην κατανόηση , στην γνώση του προβλήματος αλλά και στην λύση αφού είναι περιορισμένη. Όλα αυτά είναι δύσκολα διότι είναι το έργο της φυσικής επεξεργασίας γλώσσας αντιμετωπίζει πολλά προβλήματα. Ένας αναλυτής που έχει τόσες πολλές αναρτήσεις να διαβάσει σίγουρα θα του είναι πιο εύκολο να υπάρχει μια σύνοψη των αποτιμωμένων απόψεων σε θετικές ή αρνητικές αντί να διαβάζει ένα προς ένα τα τις απόψεις των χρηστών. Επίσης δεν είναι μόνο ο αναλυτής που μπορεί να επωφεληθεί από αυτή την αυτοματοποιημένη εξόρυξη γνώμης αλλά και διάφοροι οργανισμοί, εταιρίες για να βλέπουν τις κριτικές απόψεις των πελατών τους αλλά και τα παράπονα τους. Όμως και οι χρήστες με αυτό τον τρόπο να μπορούν να βρίσκουν πληροφορίες για διάφορα

θέματα που τους απασχολούν. Στο επόμενο κεφάλαιο επικεντρωνόμαστε στα κοινωνικά μέσα δικτύωσης και πως εφαρμόζεται μέσα από αυτό η εξόρυξη γνώμης.

Κεφάλαιο 3ο
Εξόρυξη Γνώμης μέσω των Κοινωνικών Δικτύων(Twitter)

«Σχολίασε σε περιορισμένο πλαίσιο έκτασης» Ανώνυμη

3.1 Εισαγωγή

Κάθε μέρα, δισεκατομμύρια άνθρωποι καταθέτουν τις απόψεις τους, γνώμες τους και τις σκέψεις τους για διάφορα θέματα σε κοινωνικές πλατφόρμες δικτύου Μέσα σε αυτές τις πλατφόρμες οι χρήστες μοιράζονται τα αληθινά συναισθήματα τους και τους προβληματισμούς τους για ένα συγκεκριμένο προϊόν/ μάρκα, όπως τα χαρακτηριστικά του, την εξυπηρέτηση των υπαλληλών, την τιμή σε διάφορα καταστήματα αλλά και την γνώμη τους για ένα πολιτικό θεμα. Σε αυτό τον δικτυωμένο κόσμο, η πληροφορία είναι κοινή για όλους και απαραίτητο να αναλυθεί σε διάφορες κλίμακες. Το περιεχόμενο της πληροφορίας εκφράζεται μέσα από τις θετικές, αρνητικές ή ουδέτερές απόψεις για να γίνεται πιο κατανοητή από τους ανθρώπους(κεφάλαιο 2).

Σε αυτή την εργασία θα επικεντρωθούμε στη χρήση του Twitter που είναι το πιο δημοφιλής micro-blogging πλατφόρμα σε εργασίες που έχουν να κάνουν με ανάλυση συναισθήματος. Όπου αυτή η μέθοδος σχετίζεται με τον σημασιολογικό προσδιορισμό της γνώμης και την υποκειμενική ταξινόμηση που έρχονται από το κοινωνικό δίκτυο.

, το Twitter είναι ένα παγκόσμιο δίκτυο επικοινωνίας που δεν έχει ακόμα πολλούς φανατικούς χρήστες στην Κύπρο και στην Ελλάδα, όμως έχει δισεκατομμύρια αφοσιωμένους χρηστές ανά τον κόσμο που το χρησιμοποιούν καθημερινά. Παρουσιάζει μια σημαντική διαφορά σε σχέση με άλλα γνωστότερα δίκτυα όπως το Facebook, Tumblr κ.α. Ο χρήστης μπορεί να δημοσιεύσει ένα κείμενο περιορισμένης έκτασης (140 χαρακτήρες) το οποίο ονομάζεται «tweet». Επομένως, αυτά τα tweets δεν είναι τόσο στοχαστικά με την έννοια ότι έχουν μια σύντομη εκφρασμένη άποψη σχετικά με ένα αντικείμενο όπως συνηθίζεται σε forums και blogs.

3.2 Κοινωνικό Δίκτυο "Twitter"

Πριν ξεκινήσουμε την μελέτη ,πρώτα θα αναφέρουμε πως έχει δημιουργηθεί η πλατφόρμα του Twitter. Το Twitter[10] είναι ένα online κοινωνικό δίκτυο και μια micro-blogging υπηρεσία, η οποία επιτρέπει στους χρήστες ,αφού εγγραφούν να δημοσιεύουν σύντομα μηνύματα και να διαβάζουν τα μηνύματα άλλων χρηστών της υπηρεσίας (τα γνωστά ως tweets). Μπορεί να χαρακτηριστεί το Twitter σαν ένα ενημερωτικό δίκτυο και μια πηγή ειδήσεων.

Το Twitter γράφτηκε πάνω σε μια πλατφόρμα ανοιχτού κώδικα τη Ruby on Rails με την γλώσσα Ruby και διαθέτει το δικό του API (Application Programming Interface). Δημιουργός της υπηρεσίας αυτής είναι ο Jack Dorsey ο οποίος το 2005 σκέφτηκε ότι θα ήταν πολύ ενδιαφέρον εάν μπορούσε να γνωρίζει τι κάνουν οι φίλοι του. Έτσι 'χτίστηκε' το Twitter, αρχικά από την εταιρεία ανάπτυξης «Obvious» που εδρεύει στο San Francisco. Το πρωτότυπο του δικτύου υλοποιήθηκε σε διάστημα δυο εβδομάδων (Μάρτιος 2006) και η πρώτη επίσημη εμφάνιση του στο παγκόσμιο ιστό έγινε τον Αύγουστο του 2006. Αυτή η υπηρεσία έγινε πολύ σύντομα δημοφιλής με αποτέλεσμα τον Μάιο 2007 να ιδρυθεί η εταιρία "Twitter Incorporated". Σύμφωνα με το Twiiter eMarketer[11], τον Ιανουάριος 2014 σε αυτή την πλατφόρμα είναι εγγεγραμμένοι περισσότεροι από 645,750,000 χρήστες , από τους οποίους 135.000 συνδέονται σε καθημερινή βάση ενώ καθημερινά ανταλλάσσονται πλέον των 58 εκατομμυρίων tweets. Για να γίνουν κατανοητά τι είναι αυτά τα tweets μπορούμε να τα χαρακτηρίσουμε ως ηλεκτρονικά μηνύματα

[10] https://twitter.com/
[11] http://www.statisticbrain.com/twitter-statistics/

παρόμοια των Short Message Service- SMS με την διαφορά ότι τα tweets έχουν δημόσια κοινοποίηση στην πλατφόρμα τους. Η λειτουργία αυτής της υπηρεσία είναι ο χρήστης θα κοινωνικοποίησει την κατάσταση που έχει σε άλλους, δηλαδή τις σκέψεις , επιθυμίες, προβληματισμούς , που βρίσκεται αυτή την στιγμή κ.α. Επίσης ο χρήστης έχει την δυνατότητα να παρακολουθεί τα μηνύματα άλλων χρηστών, να τα σχολιάζει και να ανατρέχει στο ιστορικό του κάθε χρήστη. Σημαντικό σημείο για να μπορέσουν οι χρήστες να λειτουργήσουν ως δίκτυο ανθρώπων πρέπει να δημιουργήσουν το κύκλο τους. Στην πλατφόρμα του Twitter υπάρχουν τα followers , αυτοί που ακολουθούν ένα χρήση και ειδοποιούνται για κάθε μήνυμα και οι following αυτοί που ακολουθεί ο χρήστης και ενημερώνεται για τις αναρτήσεις τους.

3.2.1 Για ποιούς λόγους χρημοποιούμε το Twitter[12]

Όπως αναφέραμε και πιο πάνω το Twitter περιέχει ένα πολύ μεγάλο αριθμό μικρών μηνυμάτων που δημιουργείται από τους χρήστες της πλατφόρμας micro- blogging. Τα περιεχόμενα αυτά περιέχουν τις διάφορες σκέψεις και γνώμες των χρηστών σε διάφορα ζητήματα. Η παρακάτω εικόνα(εικόνα3.1)δείχνει παραδείγματα τυπικών μηνυμάτων από ορισμένους χρήστες στο Twitter. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν στην εξόρυξη γνώμης και ανάλυση συναισθήματος. Ο κάθε οργανισμός/ επιχείρηση ή κατασκευαστικές εταιρείες μπορεί να τους ενδιαφέρει για διάφορους λόγους, έτσι χρειάζονται συγκεντρωτικές γνώμες. Βλέπουμε τις τρεις κύριες ερωτήσεις που μπορεί να ενδιαφέρει την κάθε εταιρία :

- Τι σκέφτονται οι άνθρωποι για το προϊόν μας(εταιρία, υπηρεσία, οργανισμός κλπ);
- Πόσο θετικοί(ή αρνητικοί)είναι οι άνθρωποι για τα προϊόντα μας ;
- Ποια προϊόντα προτιμούν περισσότερο ;

Δεν είναι μόνο οι εταιρίες, οργανισμοί που θέλουν να μάθουν πληροφορίες για τα προϊόντα τους αλλά είναι και τα πολιτικά κόμματα που θέλουν να μάθουν για την υποστήριξη των κομμάτων τους. Υπάρχουν οι κοινωνικές οργανώσεις που ζητούν την γνώμη των πολιτών για τα διάφορα θέματα που τρέχουν την κάθε περίοδο. Όλες αυτές οι πληροφορίες που μαθαίνουν μπορούν να ληφθούν από τις υπηρεσίες micro-blogging μαθαίνοντας και άλλες πληροφορίες για την καθημερινή ζωή των χρηστών.

[12] Αναφορά από το άρθρο[29] της βιβλιογραφίας

Χρησιμοποιούμε το micro-blogging και ειδικότερα το Twitter για τους ακόλουθους λόγους:

1. Οι πλατφόρμες micro-blogging χρησιμοποιούνται από διάφορους ανθρώπους για να εκφράσουν τις απόψεις τους για διάφορα θέματα, έτσι είναι σημαντικές οι απόψεις των ανθρώπων.

2. Το Twitter περιέχει ένα τεράστιο αριθμό θέσεων κειμένου και αυξάνεται καθημερινά. Η συλλογή μπορεί να είναι αυθαίρετα μεγάλη.

3. Το πλεονέκτημα που έχει το Twitter είναι ότι ποικίλλει από διάφορους χρήστες από διασημότητες, εκπροσώπους κομμάτων, εταιριών, μέχρι και τον πρόεδρο της χώρας. Με αυτή την διασημότητα μπορεί να συλλέγει μηνύματα κειμένου από διαφορετικά κοινωνικά και συμφέροντα ομάδων.

4. Το Twitter είναι παγκοσμίως αναγνωρισμένο και διαδεδομένο, και γίνεται η συλλογή δεδομένων σε διαφορετικές γλώσσες αλλά ταυτόχρονα να παρακολουθεί κανείς τι συμβαίνει στις άλλες χώρες μέσα από τις εκφράσεις των χρήστων.

funkeybrewster: @redeyechicago I think Obama's visit might've sealed the victory for Chicago. Hopefully the games mean good things for the city.
vcurve: I like how Google celebrates little things like this: Google.co.jp honors Confucius Birthday — Japan Probe
mattfellows: Hai world. I hate faulty hardware on remote systems where politics prevents you from moving software to less faulty systems.
brrooklyn: I love the sound my iPod makes when I shake to shuffle it. Boo bee boo
MeganWilloughby: Such a Disney buff. Just found out about the new Alice in Wonderland movie. Official trailer: http://bit.ly/131Js0 I love the Cheshire Cat.

Εικόνα 3.1 : Διάφορα παραδείγματα χρηστών του Twitter που εκφράζουν τις απόψεις τους

3.3 Ενδιαφέροντα στατιστικά στοιχεία

Σύμφωνα με τα τελευταία στατιστικά στοιχεία το Φεβρουάριο 2014 από την DMR(Digital Marketing Ramblings)από τον Cain Smith με θέμα «By the Numbers : 116 Amazing Twitters Statistics »(http://expandedramblings.com/)κυκλοφόρησαν ποία είναι τα κέρδη που έχει το Twitter. Παρακάτω αναφέρονται ενδεικτικά ορισμένα από τα ενημερωτικά στατιστικά στοιχεία:

1. Twitter Ημερομηνία έναρξης: 21 Μαρτίου, 2006(7 ετών)

2. Συνολικός αριθμός που έιναι εγγεγραμμένοι χρήστες τουTwitter : " Περίπου ένα δισεκατομμύριο". Τελευταία ενημέρωση 16/9/13

3. Μοναδικοί επισκέπτες κάθε μήνα στο Twitter.com: 36 εκατομμύρια επισκέπτες. Τελευταία ενημέρωση 24/9/13

4. Η χώρα με τους περισσότερους Twitter χρήστες : Κίνα, 35,5 εκατομμύρια χρήστες. Τελευταία ενημέρωση 5/1/13

5. Συνολικός αριθμός απεσταλμένων Tweets: 300 δις. Τελευταία ενημέρωση 3/10/13

6. Οι μηνιαίοι ενεργοί χρήστες του Twitter : 241 εκατομμύρια ενεργοί χρήστες. Τελευταία ενημέρωση 2/5/13

7. Τι ποσοστό ενεργών χρηστών βρίσκονται εκτός ΗΠΑ: 77%. Τελευταία ενημέρωση 10/3/13

8. Οι χρήστες που χρησιμοποιούν το Twitter από τις κινητές συσκευές: 184 εκατ. χρήστες. Τελευταία ενημέρωση5/2/14

9. Daily Active Users Twitter : 100 εκατ. Τελευταία ενημέρωση 10/3/13

10. Μέσος Αριθμός followers ανά χρήστη : 208 followers. Τελευταία ενημέρωση 10/11/12

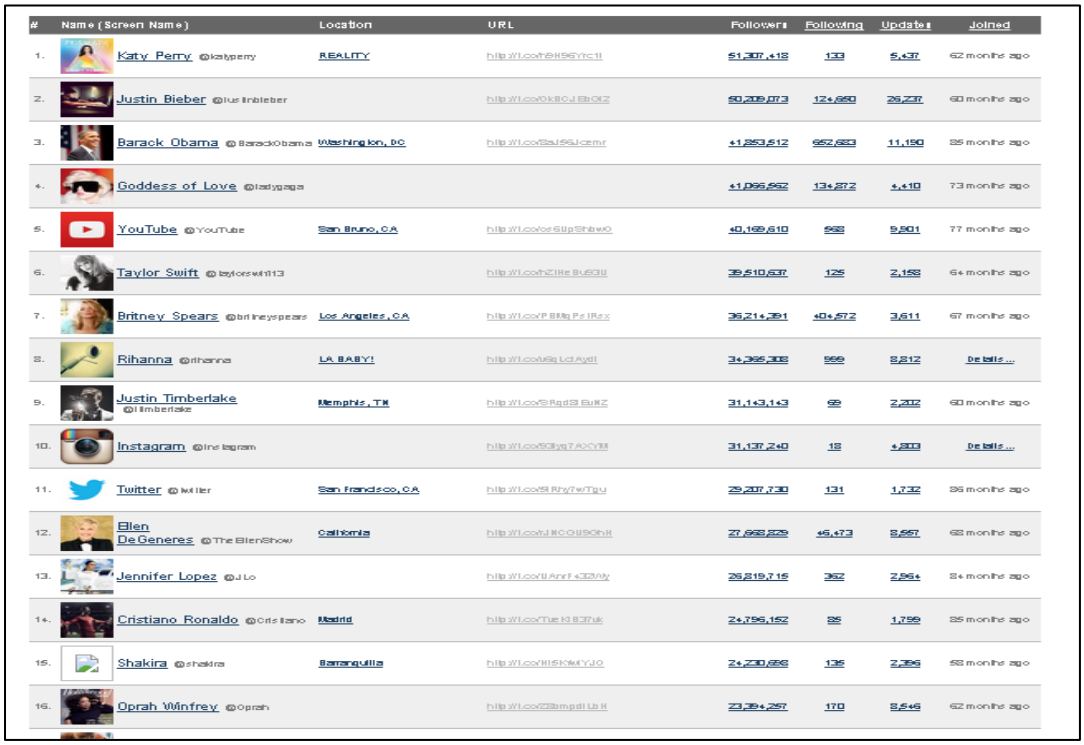
11. Μέσος αριθμός Tweets ανά Twitter χρήστη : 307 Tweets, Τελευταία ενημέρωση 11/1/13

12. Εγγραφή για Tweets ανά δευτερόλεπτο : 14199 για εγγραφή στις 3 Αυγούστου 2013.

13. Μέσος χρόνος ανά μήνα χρήστης του Twitter : 170 λεπτά. Τελευταία ενημέρωση 27/9/12

14. Ποσοστό των εφήβων που θεωρούν το Twitter αγαπημένο τους κοινωνικό δίκτυο : 26%, Τελευταία ενημέρωση 10/11/13

Επίσης από το Twitaholic(http://twitaholic.com/) είναι η κατάταξη του Twitter που παρουσιάζονται οι 100 πιο δημοφιλείς διάσημους χρήστες στην πλατφόρμα που κάνουν χρήση του Twitter. Στην πιο κάτω εικόνα(εικόνα3.2) εμφανίζονται ενδεικτικά ορισμένα ονόματα που βρίσκονται από τον χώρο της πολιτικής μέχρι και διάσημους τραγουδιστές.



#	Name (Screen Name)	Location	URL	Followers	Following	Updates	Joined
1.	Katy Perry @katyperry	REALITY		51,307,418	133	5,437	62 months ago
2.	Justin Bieber @justinbieber			50,209,073	124,650	29,237	60 months ago
3.	Barack Obama @BarackObama	Washington, DC		41,853,512	662,883	11,150	35 months ago
4.	Goddess of Love @ladygaga			41,096,562	134,872	4,410	73 months ago
5.	YouTube @YouTube	San Bruno, CA		40,169,610	668	9,501	77 months ago
6.	Taylor Swift @taylorswift13			39,510,637	125	2,158	64 months ago
7.	Britney Spears @britneyspears	Los Angeles, CA		36,214,391	404,672	3,611	67 months ago
8.	Rihanna @rihanna	LA BABY!		34,365,308	566	8,812	Details ...
9.	Justin Timberlake @Jtimberlake	Memphis, TN		31,143,143	69	2,202	60 months ago
10.	Instagram @Instagram			31,137,240	18	4,803	Details ...
11.	Twitter @twitter	San Francisco, CA		29,207,730	131	1,732	96 months ago
12.	Ellen DeGeneres @TheEllenShow	California		27,668,829	46,473	8,667	68 months ago
13.	Jennifer Lopez @JLo			26,819,715	362	2,564	84 months ago
14.	Cristiano Ronaldo @Cristiano	Madrid		24,756,152	85	1,756	35 months ago
15.	Shakira @shakira	Barranquilla		24,230,666	136	2,366	58 months ago
16.	Oprah Winfrey @Oprah			23,284,257	170	8,546	62 months ago

Εικόνα 3.2: Οι διάσημοι χρήστες που είχαν τα περισσότερα Followers από την καταμέτριση της Twitterholics

3.4 Δυσκολίες Ανάλυσης των Tweets

Σχεδόν όλες οι μελέτες που σχετίζονται με την σημασιολογική ανάλυση κειμένων εμφανίζουν να έχουν αρκετές δυσκολίες. Τα tweets πέρα αυτών των γνωστών προβλημάτων εμφανίζονται να παρουσιάζουν ακόμα κάποιες ιδιαιτερότητες. Παρακάτω αναφέρουμε τι εννοούμε με τις ιδιαιτερότητες.

Όπως αναφέραμε σε προηγούμενη ενότητα το αρχικό μέγεθος το κάθε μηνύματος tweets είναι 140 χαρακτήρες. Από το μέγεθός βλέπουμε πως δεν είναι αρκετά τα δεδομένα του μηνύματος για να αναλυθεί σημασιολογικά το περιεχόμενο. Επομένως, αυτή είναι η πρόκληση και η μεγαλύτερη δυσκολία για το σημασιολογικό προσδιορισμό των απόψεων των χρηστών.

Παρατηρώντας αρκετά μηνύματα, αφορούν προσωπικές συζητήσεις και δεν έχουν αρκετό ενδιαφέρον στο περιεχόμενο τους. Επίσης στα μηνύματα συνήθως εντοπίζονται η ασυνταξία και λάθος στους γραμματικούς κανόνες. Το πρόβλημα που υπάρχει εδώ είναι η υποκειμενικότητα και η αντικειμενικότητα μια άποψης μέσα στο ερευνητικό πεδίο, που το κάνει ακόμη πιο δύσκολο να εντοπισθεί λόγο ότι υπάρχει περιορισμένη έκταση του μηνύματος.

Ακόμη μια ιδιαιτερότητα , είναι ότι οι χρήστες καταγράφουν διάφορα προσωπικά συναισθήματα μέσα σε ένα μήνυμα. Ορισμένες εκφράσεις τους είναι με ιδιαίτερους χαρακτηρισμούς που δεν μπορούν να χρησιμοποιηθούν στον προφορικό λόγο. Πιο συγκεκριμένα, αναφερόμαστε στην νεαρή ηλικία(εφήβων)που δεν χρησιμοποιούν τον συντακτικό κανόνα. Παραδείγματος χάριν η λέξη «cu» σημαίνει «see you» ή "u2" σημαίνει "you too" ή «4u» σημαίνει «for you». Αυτή η συντομογραφία που χρησιμοποιούν οι νέοι στα μηνύματα κάνει αρκετά δύσκολη την ανάλυση γιατί οι τεχνικές στη σημασιολογική ανάλυση δεν μπορούν να βοηθήσουν και απαιτείται πειραματισμός και μελέτη αυτών των περιπτώσεων. Πρέπει να σημειωθεί κάποιες λέξεις έχουν διαφορετική ερμηνεία δηλαδή κάποτε χρησιμοποιούνται είτε ως ουσιαστικά, είτε ως ρήματα, πρέπει να ελέγχει και σε πιο σημείο της πρότασης βρίσκεται η λέξη μέσα στην πρόταση. Το πρόβλημα αυτό παρουσιάζεται σε όλες τις γνωστικές διαλέκτους, έτσι η δυσκολία ανεύρεσης της κατάλληλης ερμηνείας μεγαλώνει λόγου του περιορισμένου μηνύματος και της ελεύθερης γραμματικής δομής.

3.5 Χρήση του Twitter από τους Πολιτικούς[13]

Τα νέα μέσα στο διαδίκτυο έχουν γίνει ολοένα και πιο σημαντικά κατά τη διάρκεια της εκλογικής εκστρατείας. Η δύναμη που μπορεί να έχει το διαδίκτυο σήμερα για να κινητοποιήσουν τους ψηφοφόρους τους οι πολιτικοί είναι τεράστια και τους δίνει την ευκαιρία να προωθήσουν το προφίλ τους και να επικοινωνήσουν αμφίδρομα με το εκλογικό σώμα χωρίς να παρεμβάλλονται οι δημοσιογράφοι.

Επίσης, παρατηρώντας τα τελευταία χρόνια οι πολιτικές οργανώσεις έχουν αγκαλιάσει το διαδίκτυο χρησιμοποιώντας διάφορες πλατφόρμες όπως το Twitter, Facebook και άλλες online πλατφόρμες. Ορισμένοι ερευνητές όπως Gibson & Mc Allister(2006)[14]κάνουν έρευνες για αυτή την άνοδο του διαδικτύου για τα πολιτικά γεγονότα.

[13] Αναφορά από το άρθρο[22] της βιβλιογραφίας

Ακόμη ένα θετικό στοιχείο προς τους πολιτικούς που χρησιμοποιούν αυτές τις online πλατφόρμες είναι ότι έχουν διαδραστικές και εξατομικευμένες μορφές επικοινωνίας για να μπορούν να εκφράζουν κάθε τα προσωπικά τους συναισθήματα όπου και αν βρίσκονται, πληροφορίες για την προσωπική τους ζωή αλλά τα γεγονότα στην πολιτική. Ωστόσο, με αυτές τις μορφές επικοινωνίας μπορούν να επηρεάσουν αρκετό αριθμό υποψηφίων που τον υποστηρίζει αλλά και για άγνωστό ποσοστό υποψηφίων.

Στην παρούσα εργασία μελετούμε την χρήση του Twitter από τους διάφορους πολιτικούς κατά την διάρκεια της εκλογικής τους εκστρατεία σε συγκεκριμένες χώρες. Άλλωστε μια βασική λειτουργία του Twitter είναι ότι διευκολύνει την άμεση επικοινωνία μεταξύ των χρηστών, δηλαδή οι χρήστες αλλά κυρίως οι πολιτικές οργανώσεις και οι υποψήφιοι να σχολιάζουν δημοσιεύσεις από άλλους. Στην Κύπρο και στην Ελλάδα δεν υπάρχει μεγάλο ποσοστό που να χρησιμοποιούν την πλατφόρμα κατά την διάρκεια της πολιτικής εκστρατείας.

Η ενότητα αυτή επικεντρώνεται στα χαρακτηριστικά του περιεχομένου της πολιτικής εκστρατείας στην micro-blogging πλατφόρμα και στις συνέπειες της χρήσης των υποψηφίων μέσω αυτής της online πλατφόρμας και τι σχέση μπορεί να υπάρξει ανάμεσα στον υποψήφιο με ένα χρήστη.

3.5.1 Μορφές Επικοινωνίας

Υπάρχουν διάφορα στυλ και χαρακτηριστικά μορφών επικοινωνίας. Τα δυο πιο ισχυρά είναι η διαδραστικότητα και σε λιγότερο βαθμό η πολιτική εξατομίκευση.

Σύμφωνα με τους ερευνητές Sundan, Kalyanarman και Brown(2003)[36], μέσα από αρκετές έρευνες έχουν δείξει ότι η διαδραστικότητα αποτελεί βασικό στοιχείο στις χρήσεις των νέων μέσων τεχνολογιών. Γενικά η λέξη διαδραστικότητα μπορεί να εφαρμοστεί με διάφορους τρόπους άλλα επίσης και η αμφίδρομη επικοινωνία που αποτελεί ένα δυνατό σημείο στην επικοινωνία είναι μια έννοια με πολλούς ορισμούς. Ο Tedesco(2007)[37]έχει ορίσει την αμφίδρομη επικοινωνία ως την πληροφοριοδότηση για να επικοινωνούν απευθείας οι χρήστες. Αυτό είναι ένα νέο χαρακτηριστικό στα μέσα ενημέρωσης για να προσφέρει πληροφορίες χωρίς να παίρνουν πίσω πληροφορίες.

Η διαδραστικότητα παρατηρείται έντονα σε πολιτικά πλαίσια και συγκεκριμένα κατά τη διάρκεια πολιτικής εκστρατείας επειδή έχει την δυνατότητα για άμεση επικοινωνία. Η

πλατφόρμα του Twitter είναι ισχυρή και διαδραστική , όπως αναφέρθηκε και στην πιο πάνω ενότητα. .

Το άλλο σημαντικό χαρακτηριστικό της online επικοινωνίας είναι η εξατομικευμένη πολιτική. Δηλαδή μπορεί να χαρακτηριστεί ως μετατόπιση της εστίασης από τα πολιτικά κόμματα και θεσμικών οργανώσεων των υποψηφίων και των πολιτικών. Αυτή η μετατόπιση γίνεται αντιληπτή σε μεμονωμένους υποψηφίους αλλά και σε πολιτικούς που χρησιμοποιούν τα νέα μέσα επικοινωνίας με τους ψηφοφόρους τους. Αντιθέτως η αφοσίωση στους πολιτικούς αντί στα διάφορα κόμματα είναι διαφορετική από την εξατομίκευση που είναι κύριοι παράγοντες στο Twitter. Άλλωστε το Twitter είναι εξατομικευμένο εξ'ορισμού, δηλαδή ο κάθε υποψήφιος έχει το δικό το λογαριασμό και έτσι ο καθένας εστιάζει την ιδιωτική του ζωή. Αυτό περιλαμβάνει γύρω από τον υποψήφιο τα προσωπικά συναισθήματα, τις επαγγελματικές του δραστηριότητες. Σύμφωνα με τον Golbeck(2010)[15] όπου διαπίστωσε ότι οι πολιτικοί χρησιμοποιούν κατά κύριο λόγο το Twitter για διάδοση πληροφοριών και κυρίως άρθρα εφημερίδων κ.α. Θα μπορούσε κανείς να πει ότι το Twitter οι υποψήφιοι διαφημίζουν την ιδιωτική ζωή του και λιγότερο τα πολιτικά ζητήματα.

Παρατηρώντας, τους πολιτικούς που έχουν υιοθετήσει μια διαδραστική και εξατομικευμένη μορφή επικοινωνίας στην πλατφόρμα του Twitter υπάρχουν στοιχεία που δείχνουν πως αυτές του είδους μορφές κατά την προεκλογική εκστρατεία έχουν αποτελέσματα θετικά προς τους υποψηφίους.

3.5.2 Πιεστικές επιδράσεις στην προεκλογική εκστρατεία του Twitter

Όπως έχουμε παρατηρήσει η χρήση διαδικτύου προς τους υποψηφίους αποδεικνύει ότι έχει θετικά αποτελέσματα κάνοντας χρήση των νέων μέσω επικοινωνίας. Κάποιοι μελετητές υποστηρίζουν την ιδέα ότι η χρήση του διαδικτύου έχει ως αποτέλεσμα την κινητοποίηση για πολιτική συμμετοχή των πολιτών σε αντίθεση με τα παραδοσιακά μέσα ενημέρωσης. Τα νέα μέσα ενημέρωσης γίνονται σε απευθείας σύνδεση χωρίς κόστος και το σημαντικότερο είναι ότι ενθαρρύνει τους πολίτες να μαθαίνουν περισσότερα νέα για την πολιτική. Αυτό έχει σαν αποτέλεσμα οι πολίτες να είναι πιο δραστήριοι και να εκφράζουν την γνώμη τους πιο εύκολα στα διάφορα πολιτικά ζητήματα.

Ο D' Alessio[06] υποστηρίζει όταν οι υποψήφιοι έχουν μια ιστοσελίδα οδηγεί σε περισσότερους ψήφους ενώ όσοι δεν έχουν ιστοσελίδα έχουν σημαντικότερους λιγότερους ψήφους. Οι

ερευνητές Gibson και Mc Allister(2006)[14] μελέτησαν και διαπίστωσαν πως είναι αλήθεια με όσους έχουν ιστοσελίδα. Όταν έγιναν οι Αυστραλιανές εκλογές το 2004 όσοι υποψήφιοι χρησιμοποιούσαν ιστοσελίδα είχαν περισσότερες ψήφους. Φτάνει να έχουν και οι ιστοσελίδες άμεσο αποτέλεσμα στην πολιτική δέσμη σύμφωνα με τους Park και Perry 2008[33]. Δηλαδή να παρέχουν τις βασικές χρήσεις των δικτυακών τόπων , όπως να μπορούν να στέλλουν πολιτικά μηνύματα στο ηλεκτρονικό ταχυδρομείο για να πείσει τους πολίτες να ψηφίσουν.

Επομένως, οι Gibson και Mc Allister(2011)[13] αποδεικνύουν ότι η χρήση της απευθείας σύνδεσης σε εκλογικές πηγές και ιδιαίτερα από ιστοσελίδες έχει θετική επίδραση στην επιλογή του ψήφου.

Οι πιο πάνω μελέτες επικεντρώθηκαν στις πολιτικές ιστοσελίδες. Η παρούσα εργασία επικεντρωνότε στις επιπτώσεις της χρήσης της πλατφόρμας του Twitter κατά την διάρκεια εκλογική υποστήριξη. Η διαφορά που υπάρχει με τις ιστοσελίδες είναι ότι παρέχει μια πιο μεγάλη ποικιλία από περιεχόμενο όπως φωτογραφίες, tweets και μια πιο άμεση επικοινωνία με τον υποψήφιο. Ωστόσο, συγκρίνοντας τις ιστοσελίδες με το Twitter, οι ιστοσελίδες δεν χρησιμοποιούνται για την κοινωνική αλληλεπίδραση , άρα δεν είναι παρόμοια. Οι ομοιότητα που παρατηρείται είναι ότι έχουν θετικά αποτελέσματα προς την προεκλογική εκστρατεία.

Το Twitter έχει διάφορα εργαλεία επικοινωνίας που βοηθούν τους χρήστες να επικοινωνούν με τον άλλο. Αυτά τα εργαλεία συμβάλουν σε απευθείας σύνδεση με διάλογο με τον πολιτικό , παρέχοντας στους χρήστες την ευκαιρία να στείλουν, να διαβάζουν να απαντούν και να προωθήσουν μηνύματα άμεσα με τρίτους. Συγκεκριμένα, το Twitter έχει τρία εργαλεία που διευκολύνουν την αμφίδρομη επικοινωνία: mentions, retweets και hashtags. Οι χρήστες είναι σε θέση να ανταποκριθούν σε άλλους μέσω της χρήσης «@» . Ένα tweet με «@» ακολουθείται από όνομα, που είναι ότι ένας χρήσης στέλνει στο Twitter απευθείας ένα μήνυμα προς τον άλλο χρήστη. Μια άλλη μορφή στην διαδραστική επικοινωνία είναι η χρήση των retweets. Τα retweets διαβιβάζονται τα μηνύματα που στάλθηκαν από άλλους και επιτρέπει στους χρήστες να μεταφέρουν διάφορες πληροφορίες σε άλλους. Τέλος, το hashtags χρησιμοποιείται για να σηματοδοτήσει τα tweets από ένα συγκεκριμένο θέμα. Με αυτή την χρήση πολλοί χρήστες μπορούν να ακολουθήσουν συνομιλίες επικεντρώνοντας σε ένα συγκεκριμένο θέμα.

Ο στόχος και ο σκοπός των στρατηγικών των επιτελών του Obama κατά την προεκλογική του εξρατεία είχε ως αποτέλεσμα την ένταξη των social media και των κοινωνικών δικτύων. Έτσι οι υποψήφιοι υιοθετούν κυρίως το TWITTER ως εργαλείο προώθησης τους, γιατί είναι ένα

δυναμικό κοινωνικό δίκτυο με την μεγαλύτερη χρήση και αναγνωσιμότητα χρήσης από όλο σχεδόν τον κόσμο. Επομένως, η πολιτική επικοινωνία μετατρέπεται σε μια άλλη μορφή αντίληψης παρά του παραδοσιακού που γνωρίζουμε και μπορεί να έχει μεγάλα οφέλη ο υποψήφιος πολιτικός μέσω της χρήσης του.

3.6 Ηλεκτρονική Πολιτική Εκστρατεία

Η προεκλογική εκστρατεία του Barack Obama, η οποία διεξήχθη από τις 25-8-2008 μέχρι τις 4-11-2008 έχει αποτελέσει στόχο απόψεων και αναλύσεων τόσο για την οργάνωσή της, όσο και για τη στρατηγική πολιτικής επικοινωνίας, την οποίαν υιοθέτησε. Σημαντικό γεγονός στα πλαίσια της πολιτικής ιστορίας της Αμερικής, είναι πως ένας υποψήφιος χρησιμοποίησε σχεδόν αποκλειστικά το διαδίκτυο ως πεδίο ανάπτυξης της προεκλογικής του στρατηγικής και ως εργαλείο κινητοποίησης ανθρώπων και συλλογής οικονομικών πόρων.

Ο Barack Obama κέρδισε ένα μεγάλο στοίχημα, να προκαλέσει το ενδιαφέρον των Αμερικανών για την πολιτική, προσανατολίζοντάς τους σε πρώτη φάση να εμπλακούν στα πολιτικά πράγματα και στη συνέχεια να ψηφίσουν. Η στρατηγική του όμως πέτυχε σε μια τέτοια μαζική συμμετοχή των πολιτών, γιατί χρησιμοποίησε ρητορική και χαρακτηριστικά κοινωνικού κινήματος και απέφυγε τις κλασσικές μεθόδους διεξαγωγής προεκλογικής εκστρατείας. Η ρητορική που υιοθέτησε και χρησιμοποίησε είχε σκοπό να θεραπεύσει την έλλειψη ενδιαφέροντος των Αφρικανών για την πολιτική, ενώ η στρατηγική του ήταν να δικτυώσει μεταξύ τους ανώνυμους πολίτες και να τους κινητοποιήσει επ΄ ωφελεία του.

Αρκετοί ερευνητές αναφέρουν ότι όσες προεκλογικές εκστρατείες που έγιναν το 2008, το διαδίκτυο αποτέλεσε βασικό στοιχείο των στρατηγικών για την πολιτική τους επικοινωνία. Ωστόσο, ο Obama δεν διέθετε ισχυρή εκλογική βάση αλλά κατάφερε με τους επιτελείς του να εκμεταλλευτεί τη δύναμη των ηλεκτρονικών κοινωνικών δικτύων και των κοινωνικών μέσω επικοινωνίας. Ξεφεύγοντας από τα παραδοσιακά πρότυπα των προεκλογικών εκστρατειών, έτσι βάζοντας νέους μεθόδους. Με αυτό τον τρόπο είχε μια πιο άμεση επικοινωνία με τους πολίτες βλέποντας τις διάφορες απόψεις τους αλλά συμμετέχοντας τις συζήτησης που έγιναν και στο τέλος να κερδίζει όλο και περισσότερους ψηφοφόρους.

Η ενημέρωση και ανανέωση των σελίδων συμβάλλει ικανοποιητική ανταπόκριση για τους πολίτες που ενημερώνεται μέσα από αυτά. Αυτό αποτελεί θετικό γεγονός για τη χρήση των

ηλεκτρονικών κοινωνικών δικτύων και η τακτική ενημέρωση. Ωστόσο ο Obama είχε μια νέα προοπτική να εμπιστευτεί την διαχείριση των κοινωνικών δικτύων του σε νέους με στόχο να κερδίσει και άλλους ψήφους για να αποδίδουν νέες κοινωνικές αξίες. Την επόμενη χρόνια των εκλογών έγινε ο υπολογισμός της αξίας της προεκλογικής εκστρατείας του Obama, αναφέρονται κάποιοι ερευνητές πως εισήρθαν νέων ψηφοφόρων, αλλά και των μειονοτήτων(αφροαμερικανών), στην διαμόρφωση του εκλογικού αποτελέσματος. Ο Obama υπήρξε αδιαμφισβήτητα ο πολιτικός ηγέτης που ενέπνευσε νέους ψηφοφόρους, όσο κανείς άλλος κατά τις τέσσερις τελευταίες δεκαετίες. Επίσης και η νεολαία είχε αποστασιοποιηθεί από τις πολιτικές καταστάσεις των Η.Π.Α. λίγο μετά από την λήξη του πολέμου στο Βιετνάμμ.

Οι υποψήφιοι πολιτικοί απαντούν άμεσα μέσω των κοινωνικών δικτύων(Twitter)με τους πολίτες και να εκφράζουν τις γνώμες τους σε διάφορα ζητήματα , έχοντας κάποιά άτομα να διαχειρίζονται την προσωπική σελίδα τους αλλά όμως υπάρχουν και προγραμματιστές όπου θα αναλύσουν τα tweets γιατί είναι πάρα πολλά. Η ανάλυση αυτή περιλαμβάνει ταξινόμηση κατηγοριών(είτε θετικά, είτε αρνητικά)αλλά γενικότερα οι άνθρωποι που θα ασχοληθούν με το τεχνικό κομμάτι του κοινωνικού δικτύου αναφέρεται στην υπολογιστική εξαγωγή πληροφοριών από ένα δεδομένο δείγμα. Χρησιμοποιούν διάφορα συστήματα για την εξαγωγή των tweets(αναφορά στο προηγούμενο κεφάλαιο).

3.7 Ανάλυση των Tweets

Το Twitter έχει αυξηθεί εκρηκτικά μετά τις προεδρικές εκλογές του 2008 όπως έχουμε αναφέρει στην πιο πάνω ενότητα. Οι χρήστες αποστέλλουν τώρα περισσότερα tweets κάθε δέκα λεπτά κατά την διάρκεια των εκλογών. Το Twitter έχει πάρει πολιτικά στιγμιότυπα όπως η παρακολούθηση του tweets ανά λεπτό. Όλη αυτή η τεράστια αύξηση για τα πολιτικά ζητήματα δημιούργησε ένα τεράστιο σύνολο δεδομένων για ανάλυση. Το Twitter είναι σαν ένα πραγματικός χρόνος έκδοσης μιας δημοσκόπησης (exit poll) με μια μαζική αύξηση του μεγέθους του δείγματος σύμφωνα με τον αναλυτή Michael Fauscette[27].

Σύμφωνα με τον Frank Newport[26] αρχισυντάκτη της Galluo Pall λέει αυτοί που ασχολούνται με τις δημοσιοποιήσεις είναι πολύ ενδιαφέρον η ανάλυση των κοινωνικών μέσω μαζικής ενημέρωσης λόγο των μεγάλων όγκων των απόψεων και θα ήταν ανόητο να μην εξεταστεί αυτός ο τεράστιος όγκος απόψεων από 140 εκατομμύρια χρήστες από όλο τον κόσμο.

Όμως ο Marc Smith[26] ιδρυτής του ιδρύματος κοινωνικής Media Research τα μέσα κοινωνικής ενημέρωσης όπου είναι δημόσιες συζητήσεις παίρνουν νέους τρόπους, όμως αυτές οι συνομιλίες δεν μπορούν κατ' ανάγκη να προβλέψουν τα συγκεκριμένα αποτελέσματα των εκλογών.

3.7.1 Περιπτώσεις χρήσης του Twitter για ανάλυση [14]

Φαίνεται να υπάρχει μια αμφιβολία γύρω από την δυνατότητα ανάλυσης της ροής δεδομένων του Twitter σύμφωνα με τον αναλυτή Michael Fauscette[27] . Υπάρχουν πολλές χρήσης του Twitter όπως στο πλαίσιο μια πολιτικής εκστρατείας γύρω από συγκεκριμένες συζητήσεις και την ικανότητα προβλέψεις των αποτελεσμάτων. Όμως υπάρχουν κάποια ερωτήματα που πρέπει πρώτα να εξαταστούν:

1. Το συναίσθημα που εκφράζεται από το Twitter μέσω των χρηστών αντικατοπτρίζεται με ακρίβεια στα πολιτικά θέματα;

2. Το Twitter χρησιμοποιείται για να συζητούν πολιτικά θέματα ή είναι απλά μια πλατφόρμα για να εφράζουν τις διάφορες απόψεις τους οι χρήστες;

3. Το Twitter μπορεί να χρησιμοποιηθεί για να προβλέψει με ακρίβεια τα αποτελέσματα των εκλογών;

4. Υπάρχουν αρκετές μελέτες που έχουν εξετάσει πόσο αποτελεσματικό είναι το Twitter σαν εργαλείο κατά την διάρκεια πολιτικής εκστρατείας. Για παράδειγμα το 2010 μια μελέτη που διεξήχθη από το Τεχνικό Πανεπιστήμιο του Μοναχού και παρουσιάστηκε στο συνέδριο AAAI μελέτησαν το Twitter ως εργαλείο πρόβλεψης των αποτελεσμάτων των εκλογών. Η μελέτη χρησιμοποίησε τα tweets που οδηγούν μέχρι τις ομοσπονδιακές εκλογές του εθνικού κοινοβουλίου(27/9/2009) και χρησιμοποιήθηκαν τα τρία πιο πάνω ερωτήματα που αναφέραμε για να δουν αν υπάρχει συσχέτιση μεταξύ της ανάλυσης συναισθήματος και με τα αποτελέσματα των εκλογών. Τα αποτελέσματα της μελέτης είναι παρακάτω :

1. Το Twitter χρησιμοποιείται ως πλατφόρμα για πολιτική συζήτηση.

2. Το προσωπικό προφίλ των υποψηφίων και κομμάτων κατοπτρίζει από τα tweets και απεικονίζεται με ακρίβεια στο πραγματικό συναίσθημα του πληθυσμού.

[14] Αναφορά από το άρθρο [27]της βιβλιογραφίας

3. Ο αριθμός των tweets αντιστοιχούν με τα αποτελέσματα των εκλογών και ήταν ανάλογο με τις παραδοσιακές εκλογικές δημοσκοπήσεις.

Παρακάτω παρουσιάζονται σε ποιες περιπτώσεις μπορεί να γίνει η χρήση του Twitter :

Η πραγματική ανάλυση συναισθήματος περιλαμβάνεται από δημόσια ζητήματα. Για τα πολιτικά πρόσωπα είναι ένα ισχυρό εργαλείο που ξεπερνά τις παραδοσιακές δημοσκοπήσεις και έρευνες οι οποίες υστερούν σε μεγάλο βαθμό και ιδικά στο χρόνο αποκτήσεις αποτελεσμάτων. Δηλαδή για ένα πολιτικό να μπορεί να πάρει το συναίσθημα από τους χρήστες από μια δημόσια δήλωση του ή ένα πολιτικό θέμα και να ξεκινήσουν οι διάφορες απόψεις-μηνύματα σε πραγματικό χρόνο. Αυτό σημαίνει πως ο υποψήφιος θα παίρνει λεπτομερή μηνύματα σε όλη την διάρκεια της εκστρατείας του.

Το Twitter δεν είναι μόνο ένα κανάλι εκπομπής αλλά μπορεί να προωθήσει συζητήσεις γύρω από ένα θέμα. Για ένα πολιτικό αυτό θα μπορούσε να χρησιμοποιηθεί για να διευρύνει την κοινή γνώμη σχετικά με ένα συγκεκριμένο θέμα και να έχει μια συλλογή σχολίων και να μπορεί να πάρει κάποιες πιθανές λύσεις.

Το Twitter μπορεί με αρκετή ακρίβεια να προβλέψει τα αποτελέσματα των εκλογών όπου πιο σημαντικό είναι να χρησιμοποιηθεί ένα εργαλείο ανάλυσης συναισθήματος που μπορεί να ανταποκριθεί σε ικανοποιητική ακρίβεια.

3.7.2 Twitter Political Index [15]

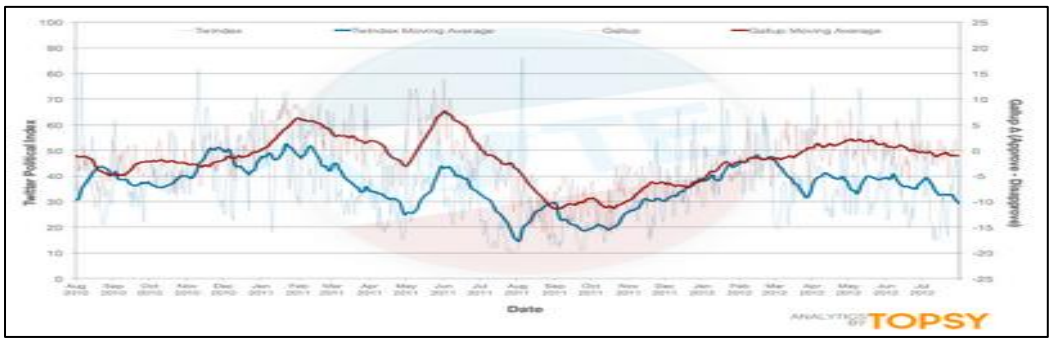
Το Twitter Political Index ξεκίνησε για να παρακολουθεί τα επίπεδα ενθουσιασμού για τον Obama και τον Romney με μέσο όρο δυο εκατομμύρια tweets την εβδομάδα το 2008. Ο οργανισμός αυτός στηρίζεται σε ανάλυση συναισθήματος, μιας αναζήτησης στον ηλεκτρονικό υπολογιστή μέσων των κοινωνικών μέσω μαζικής ενημέρωσης για να καθορίσει τη στάση των χρηστών. Επίσης είναι χτισμένο σε συνεργασία με την ομάδα ανάλυσης δεδομένων Topsy και με δυο σεβαστές εταιρίες δημοσκοπήσεων The Mellman Group και North Start Opinion Research.

Η Topsy αναφέρει ότι το σύστημα αναζήτησης που έχει μπορεί να εντοπίσει τα θετικά ή αρνητικά tweets να ερμηνεύσουν το σαρκασμό, το χιούμορ και ακόμα αν οι χρήστες είναι ενεργείς με τα tweets. Ακόμη έχει αναπτύξει ένα λογισμικό ανάλυσης χρησιμοποιώντας ένα

[15] Αναφορά από τα άρθρα [26,01]της βιβλιογραφίας

αρχείο δισεκατομμυρίων tweets από το 2008. Όπου περιλαμβάνει πληροφορίες σχετικά με το τι λένε οι χρηστές του Twitter για τους υποψηφίους.

Καθημερινά ο δείκτης Index αξιολογεί και ζυγίζει το συναίσθημα των tweets που παραπέμπουν ο Obama ή Romney με περισσότερα από 400 εκατομμύρια tweets που αποστέλλονται για όλα τα θέματα. Το Twitter Political Index μπορεί να οριστεί σαν μια νέα διορατικότητα για το συναίσθημα του εκλογικού σώματος αλλά δεν προορίζεται να αντικαταστήσει την παραδοσιακή δημοσκοπήσει. Επίσης βοηθά να συλλάβει τις αποχρώσεις της κοινής γνώμης.



Εικόνα3.3 : Για κάθε υποψήφιο και ένα χρώμα, για τις δημοσκοπήσεις από την ομάδα ανάλυσης δεδομένων Topsy για την προεκλογική εκστρατεία 2008.

3.7.3 Διάφορες απόψεις από ερευνητές[16]

Ο Tumasjan(2010)[39] εστιάζεται στην Γερμανική ομοσπονδιακή εκλογή 2009 και διευρύνει κατά πόσο το Twitter μπορεί να χρησιμοποιηθεί για την πρόβλεψη των εκλογών. Πάνω από 100 χιλιάδες tweets χρονολογούνται από τις 13 Αυγούστου μέχρι τις 19 Σεπτεμβρίου 2009 περιέχοντας έξι κόμματα. Έγινε η χρήση του λεξικού LIWC 2007 για να εξάγουν τα συναισθήματα από τα tweets. Το συμπέρασμα της μελέτης είναι ότι ο αριθμός των tweets ενός κόμματος είναι ευθέως ανάλογη προς τη πιθανότητα να κερδίσει το κόμμα.

Ο Connor(2010) [28]ερεύνησε την έκταση ποιες δημόσιες δημοσκοπήσεις συσχετίστηκαν με το πολιτικό συναίσθημα εκφράζοντας μέσω των tweets. Έγινε η χρήση του λεξικού Subjectivity Lexicon και εκτιμούν καθημερινά το συνολικό συναίσθημα για κάθε οντότητα. Ένα tweet ορίζεται ως θετικό αν περιέχει ένα θετικό και μια λέξη αντίστροφα. Τα αποτελέσματα του συναισθήματος συσχετίστηκαν με δημοσκοπήσεις σχετικά με τις προεδρικές εκλογές και λιγότερο με δημοσκοπήσεις για τα εκλογικά αποτελέσματα.

[16] Αναφορά από το άρθρο[03]της βιβλιογραφίας

Ο Choy(2011)[05]προσπάθησε να εφαρμόσει μια εφαρμογή σε απευθείας σύνδεση για ανίχνευση συναισθήματος ,ώστε να προβλέπει το ποσοστό ψηφοφορίας για κάθε ένα από τους υποψηφίους. Η μελέτη αυτή έγινε για τις προεδρικές εκλογές της Σιγκαπούρης 2011. Έχουν σχεδιάσει μια φόρμουλα για να υπολογίζει το ποσοστό ψήφου του κάθε υποψηφίου και θα λαμβάνει χρήσιμα στοιχεία που θα είναι μεταβλητές όπως ομάδα, φύλο, τοποθεσία κλπ. Θα συνδυάζεται αυτό με ένα λεξικό sentiment- lexicon-based που θα υπολογίζει την ανάλυση συναισθήματος για κάθε tweets αν είναι θετική ή αρνητική ψυχολογία για κάθε υποψήφιο. Το μοντέλο ήταν σε θέση τελικά να προβλέψει μεταξύ των δυο επικρατέστερων υποψηφίων, όμως απέτυχε να προβλέψει τον σωστό νικητή.

Ο Wang(2012)[42]πήρε δεδομένα από τις προεδρικές εκλογές του 2012 της Αμερικής και πρότεινε ένα real- time σύστημα για ανάλυση συναισθήματος για πολιτικούς σκοπούς από tweets. Μάζεψε πάνω από 36 εκατομμύρια tweets χρησιμοποιώντας το Amazon Mechanical Tunk για την συλλογή. Επίσης χρησιμοποίησε το μοντέλο ταξινόμησης Naïve Bayes με χαρακτηριστικά unigram. Είχε πετύχει το 59 % ακρίβειας της ταξινόμησης.

Οι Bermigham και Smeation(2011)[04] επικεντρώθηκαν στις εκλογές στην Ιρλανδία του 2011 με επιβλεπόμενη ταξινόμηση χαρακτηριστικού unigram, επιτυγχάνοντας το 65% ακρίβειας για την κατάταξη θετικό/αρνητικό/ουδέτερο συναίσθημα. Συμπεραίνουν ότι ο όγκος των δεδομένων είναι μια ισχυρή ένδειξη των εκλογών και το συναίσθημα έχει ένα σημαντικό ρόλο στην ανάλυση.

Ο Gayo Avello(2012)[12] θέτει το ερώτημα αν η χρήση του Twitter για τις εκλογές δίνει σωστή πρόβλεψη. Ο στόχος είναι η ακριβής αναγνώριση για κάθε συναίσθημα που εκφράζονται μέσα από τα tweets. Επίσης αν η ανάλυση συναισθήματος των πολιτικών σε tweets βελτιωθεί με περισσότερη ακρίβεια τότε θα έχει θετική επίδραση και εναλλακτική λύση από τις παραδοσιακές δημοσκοπήσεις.

3.8 Ανάλυση των Ιρλανδικών εκλογών 2011[17]

Αυτή η έρευνα επικεντρώνεται στις Ιρλανδικές Γενικές Εκλογές του Φεβρουαρίου 2011. Περιγράφεται το συναίσθημα σε πείραμα ανάλυσης σε μια πιο περίπλοκη εγκατάσταση. Δηλαδή το έργο κατατάσσει σε τρεις κατηγορίες την ταξινόμηση(θετικό, αρνητικό, ουδέτερο)και το

[17] Αναφορά από το άρθρο[03]της βιβλιογραφίας

συναίσθημα που ταξινομείτε δεν είναι σε ένα γενικό έγγραφο αλλά κατευθύνεται σε ένα συγκεκριμένο θέμα.

Η συλλογή δεδομένων έγινε από τα tweets που συλλέχθηκαν κατά την περίοδο των εκλογών που περιγράφονται στις τρεις κατηγορίες για κάθε ένα συγκεκριμένο πολιτικό κόμμα ή αρχηγό. Έγινε με την χρήση του Twitter API μεταξύ 20 Γενάρη και 25 Γενάρη και επέλεξαν τις πιο κυρίαρχες πολιτικές οντότητες για την συλλογή των tweets. Η συλλογή αυτή περιείχε 7,916 tweets εν των οποίων τα 4710 ήταν retweets ή αντίγραφα(duplicates)έτσι αφήνοντας 3,206 tweets συνολικά. Για τον προσδιορισμό του συναισθήματος των tweets προσέλαβαν δυο Ιρλανδικούς σχολιαστές που είχαν γνώση για το Ιρλανδικό πολιτικό τοπίο. Υπήρχαν όμως διαφωνίες και έφεραν τρίτο σχολιαστή και κατέληξαν σε έξι ειδών ετικέτες σχολιασμών όπου είναι τα παρακάτω:

Pos : τα tweets έχουν θετικό κλίμα προς το θέμα

Neg : τα tweets έχουν αρνητικό κλίμα προς το θέμα

Mix : τα tweets μεταφέρουν τόσο θετικό όσο και αρνητικό συναίσθημα προς το θέμα

Neu : τα tweets δεν έχουν κανένα συναίσθημα για το θέμα

Nen : τα tweets γράφτηκαν σε άλλες γλώσσες εκτός από αγγλικά

Non : τα tweets δεν έχουν καμία σχέση με το θέμα.

Επέλεξαν τις τρείς ετικέτες σχολιασμών όπου βρίσκονται στον συνοπτικό πίνακα πιο κάτω(πίνακας 3.1),περιλαμβάνοντας τα θετικά(Pos),αρνητικά(Neg) και ουδέτερα(Neu) συναισθήματα των tweets καταλήγοντας μόνο στα 2624 tweets. Κατά την διάρκεια του πειράματος χρησιμοποιήθηκαν δυο διαφορετικά λεξικά subjectivity lexicon για τον προσδιορισμό τις πολικότητας του συναισθήματος των λεξικών. Έγινε η χρήση του Subjectivity lexicon(SL) περιλαμβάνοντας 8221 λέξεις και με το λεξικό Sentiment Net 3.0(SWN)έχοντας πάνω από 100+ χιλιάδες λέξεις. Ακόμη, έχοντας ένα part-of –speech tagger χρησιμοποιώντας ένας ειδικά σχεδιασμένος tagger για τις αναρτήσεις και ένα πρόγραμμα ανάλυσης χρησιμοποιώντας το Stanford parser. 3.0(SWN).

Positive Tweets	256	9.75%
Negative Tweets	950	36.22%
Neutral Tweets	1418	54.03%
Total Tweets	2624	

Πίνακας 3.1 : Οι τρεις κατηγορίες ταξινόμησης από τις έξι κατηγορίες από τις Ιρλανδικές εκλογές 2011

3.8.1 Προσέγγιση Βάση Naïve Lexicon

Η προσέγγιση του συναισθήματος ταξινόμησης έγινε με το Naïve Lexicon για μη επιβλεπόμενη μάθηση(unsupervised lexicon-based approach). Η διαδικασία είναι ότι τρέχουν την κάθε λέξη από το λεξικό συναισθήματος κοιτάζοντας τα αθροιστικά αποτελέσματα απεικονίζοντας τα σε ένα πίνακα(πίνακας 3.2).

Method	Extended-SL		SWN		Combined	
3-Class Classification (Pos vs Neg vs Neu)	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
Baseline	1376	52.44%	1379	52.55%	1288	49.09%
Baseline + Adj	1457	55.53%	1449	55.22%	1445	55.07%
Baseline + Adj + S	1480	56.40%	1459	55.60%	1481	56.44%
Baseline + Adj + S + Neg	1495	56.97%	1462	55.72%	1496	57.01%
Baseline + Adj + S + Neg + Phrases	1511	57.58%	1479	56.36%	1509	57.51%
Baseline + Adj + S + Neg + Phrases + Than	1533	58.42%	1502	57.24%	1533	58.42%
Distance Based Scoring: Baseline + Adj + S + Neg + Phrases + Than	1545	58.88%	1506	57.39%	1547	**58.96%**
Sarcastic Tweets	87/344	25.29%	81/344	23.55%	87/344	25.29%

Πίνακας 3.2 : Τρεις κατηγορίες κατάταξης με προσέγγιση με βάση το Naïve λεξικό.

Η πρώτη στήλη από τον πίνακα δείχνει τα αποτελέσματα χρησιμοποιώντας το SL λεξικό. Η δεύτερη στήλη δείχνει τα αποτελέσματα χρησιμοποιώντας το SWN λεξικό. Η τρίτη στήλη δείχνει τα αποτελέσματα με τον συνδυασμό των δυο λεξικών.

Στην πρώτη σειρά του πίνακα βλέπουμε τα αποτελέσματα του πειράματος. Δηλαδή η κάθε λέξη του tweet που υπήρχε κοιταζόταν από το λεξικό συναισθήματος και βαθμολογούν το συναίσθημα και προσθέτονταν στο σύνολο. Επιτυγχάνοντας με κατηγοριοποίηση 52,44% με το

SL λεξικό. Το αποτέλεσμα έχει χαμηλή ακρίβεια γιατί πολλές λέξεις εμφανίζονται στο συναίσθημα λεξικό αλλά δεν περιλαμβάνονται στο συναίσθημα του θέματος που υπάρχει.

Για τον χειρισμό της άρνησης των tweets χρησιμοποιήθηκαν δυο διαφορετικές προσεγγίσεις.

1. Να οριστούν ποιες είναι οι αρνητικές οι λέξεις

2. Γίνεται χρήση της συντακτικής ανάλυσης για την έκταση των αρνητικών λέξεων με το πρόγραμμα Stanford.

Εντοπίστηκαν αρκετές φράσεις όπως «*god save us, wolf in sheep's clothing*». Στην μελέτη υπήρχαν 89 τέτοιες φράσεις. Είχαν άμεση υπόψη για αυτές τις φράσεις πετυχαίνοντας το 57,58% ακρίβειας ταξινόμησης(πέμπτη γραμμή για το SL λεξικό).

Μια άλλη μορφή έκφρασης γνώμης είναι συγκρίνοντας οντότητα με κάποια άλλη οντότητα. Για παράδειγμα *fast food sound like a better vote than Fianna Fail.* Αυτό είναι ένα αρνητικό tweet για το πολιτικό κόμμα Fianna Fail. Ακολούθησαν την παρακάτω διαδικασία να διαιρούν το tweeter σε δυο μέρη, δεξιά(right) και αριστερά(left).

Tweet: "X is better than Y"

Left: "X is better"

Right: "Y"

Μετά γίνεται ένας υπολογισμός από +1 και -1, αθροίζοντας για να εντοπιστέι η βέλτιστη ακρίβειας σε 0,8%(έκτη γραμμή για το SL λεξικό).

Η προτελευταία γραμμή από το SL Lexicon γίνεται ο υπολογισμός της απόστασης scoring για την ταξινόμηση προσανατολισμού συναισθήματος με την χρήση του πιο κάτω τύπου:

$$S(tweet) = \sum_{i=1}^{n} S(word_i)/dis(word_i).$$

Το dis είναι πόσες συνολικές λέξεις υπάρχουν για το θέμα και στην συνέχεια γίνεται υπολογισμός με μια ακρίβεια τις τάξεις 0,45% δηλαδή συνολική ακρίβεια 58,88% .

Παραδείγματα εσφαλμένης ταξινόμησης

Σε αυτή την υποενότητα αναθέτουμε κάποια παραδείγματα που είναι παγίδες- εσφαλμένης ταξινόμησης με την χρήση της πιο πάνω προσέγγισης(naïve lexicon based approach). Όπως αναφέρονται στον πιο κάτω πίνακα(πίνακας3.3).

Tweet	Topic	Manual Polarity	Calculated Polarity	Reason for misclassification
@username and u believe people in fianna fail . What are you a numbskull or a journalist ?	Fianna Fáil	neg	neu	Focus only on adjectives
@username LOL . A guy called to our house tonight selling GAA tickets . His first words were : I'm not from Fianna Fail .	Fianna Fáil	neg	neu	No sentiment words
@username Such scary words .' Sinn Fein could top the poll ' in certain constituencies . I feel sick at the thought of it .	Sinn Féin	neg	pos	Stemming and word distance order
@username more RTE censorship . Why are they so afraid to let Sinn Fein put their position across . Certainly couldn't be worse than ff	Sinn Féin	pos	neg	contribution of *afraid*
Based on this programme the winners will be Sinn Fein & Gilmore for not being there #rtefl	Sinn Féin	pos	neu	Focus only on adjectives
#thefrontline pearce Doherty is a spoofer ! Vote sinn fein and we loose more jobs	Sinn Féin	neg	pos	Focus only on adjectives & contribution of phrase *Vote X*
@username Tread carefully Conor . BNP endorsing Sinn Fin etc . etc .	Sinn Féin	neg	neu	No sentiment words
@username ah dude . You made me go to the fine gael web site ! :/	Fine Gael	neg	neu	No sentiment words

Πινάκας 3.3 : Παραδείγματα εσφαλμένης ταξινόμησης

Παράδειγμα 1 από τον πίνακα: σε αυτό το tweet βλέπουμε το αρνητικό συναίσθημα μεταδίνεται από μέσα στο σχόλιο αλλά δεν έχει καθόλου επίθετα συναισθήματος. Η λέξη numbskull περιλαμβάνεται στην ψυχολογία και είναι χαρακτηριστικό ουσιαστικού. Γενικά το σχόλιο ορίζεται αρνητικό από τους σχολιαστές και ουδέτερο από την ταξινόμηση lexicon-based.

Παράδειγμα 2από τον πίνακα: δεν υπάρχει συναίσθημα που να υπάρχει στο λεξικό συναισθήματος.

Παράδειγμα 8 από τον πίνακα: αντιμετωπίζει το ίδιο πρόβλημα με το παράδειγμα 2 γιατί γίνεται η χρήση emoticons(Lol, ☺, :P) και δεν γίνεται η χρήση αυτή.

3.8.2 Επιβλεπόμενη Μηχανική Μάθηση

Στην επιβλεπόμενη μηχανική μάθηση ο αλγόριθμος γίνεται με την χρήση του εργαλείου SUM light που είναι προσέγγιση βασισμένη σε λεξικό. Έχοντας δύο σύνολα από χαρακτηριστικά και το πρώτο να είναι τα unigram χαρακτηριστικά που έχουν χρησιμοποιηθεί και σε αλλαγές

αναλύσεις για την ταξινόμηση κειμένου. Το δεύτερο σύνολο είναι όπως περιγράφτηκε στην πιο πάνω υπό-ενότηρα στην προσέγγιση με βάση λεξικό.

3.8.3 Σύνοψη για τις Ιρλανδικές Εκλογές

Επομένως το πείραμα αυτό που έγινε δείχνει ότι μπορούν να χαρακτηρίσουν ένα tweet ως θετικό/αρνητικό/ουδέτερο προς ένα συγκεκριμένο πολιτικό κόμμα με 59% ακρίβεια χρησιμοποιώντας την απλή προσέγγιση βασίζοντας στην αναζήτηση στο λεξικό.

Φαίνεται η χρήση του Twitter έχει θετικές συνέπειες για τους πολιτικούς. Αυτό φαίνεται ότι το Twitter είναι στην πραγματικότητα μια σημαντική πλατφόρμα για τους πολιτικούς να επικοινωνούν με τους ψηφοφόρους τους. Επίσης η πλατφόρμα αυτή έχει πολύ διαδραστική δυνατότητα και σημαντικό ρόλο παίζει πόσο μεγάλη χρήση γίνεται του Twitter από τους πολιτικούς αφού ανάλογα της χρήσης έχει και τις επιπτώσεις. Ακόμη προσφέρει και πιο εξατομικευμένη επικοινωνία με τους πολίτες.

Για την ανάλυση του περιεχομένου χρησιμοποιούνται για την εξαγωγή ειδικά διαδραστικά χαρακτηριστικά από τα tweets(πχ retweets). Παίρνουν ένα συνολικό αριθμό δειγμάτων από τα tweets που κωδικοποιούνται και αναλύονται. Αυτό το έργο το αναλαμβάνουν προγραμματιστές-σχολιαστές και ταξινομούν τα tweets είτε ως θετικά, αρνητικά και ουδέτερα.

Επομένως, υπάρχουν πολλά θέματα για κάθε τύπο ανάλυσης που υπάρχουν ευκαιρίες να εφαρμοστούν στην ανάλυση συναισθήματος στην κοινωνική ζωή του Twitter.

3.9 Ομοσπονδιακές Γερμανικές Εκλογές 2013[18]

Στις Ομοσπονδιακές Γερμανικές Εκλογές έγινε μια έρευνα αξιολογώντας τις πολιτικές συζητήσεις μέσα από τους Γερμανούς πολιτικούς από το Twitter, εξάγοντας τις πληροφορίες από το εργαλείο του Twitter API. Η συλλογή των δεδομένων έγινε πριν, κατά τη διάρκεια και μετά τις εκλογές με σύνολο 13 εβδομάδες. Χρησιμοποίησαν μια διαφορετική προσέγγιση του Twitter απ' όσα αναφέραμε πιο πάνω, κοιτάζοντας τις Online ομιλητικές πρακτικές επικεντρώνοντας πόσοι πολίτικοι έκαναν Followed, Retweeted και Meniontεd. Ο σκοπός της έρευνας εντάσσετε μέσα σε πέντε κύριες ερωτήσεις: πόσες εβδομάδες ένα πολιτικό γεγονός αναφέρεται πολλές φορές, πόσο

[18] Αναφορά από το άρθρο[23]της βιβλιογραφίας

δημοφιλής είναι ένα γεγονός σε μια δεδομένη χρονική περίοδο σε σχέση με άλλα γεγονότα, πόσο σταθερό είναι ένα κόμμα κατά την πάροδο του χρόνου, όσο παρόμοια είναι τα μέρη των κομμάτων και πόσο έχει επικεντρωθεί ένα κόμμα σε πολιτιστικά γεγονότα.

Η θεωρητική και υπολογιστική προσέγγιση από τις ομιλητικές πρακτικές των πολιτικών κομμάτων στο Twitter που έγινε από τις κοινωνιολογικές σχέσεις από την πολιτιστική εστίαση(cultural focus),ομοιότητα(similarity) και η αναπαραγωγή των παραγόντων(reproduction) αλλά και η institutions, punctuation. Το μεγαλύτερο μέρος επικοινωνίας των πολιτικών έγινε μια εβδομάδα πριν τις εκλογές αλλά και στην 7η εβδομάδα όπου υπήρξε το debate στα κανάλια. Όλα τα κόμματα ασχολούνται με το hashtags προς τις εκλογές σε σχέση με τα similarity μέσω των retweeted ή το menionted σε πολικούς. Επομένως, πολλές online ομιλητικές πρακτικές διαφέρουν μεταξύ των κομμάτων μέσα από την πολιτική επικοινωνία του Twitter.

3.10 Εργαλεία και εμπορικές εφαρμογές

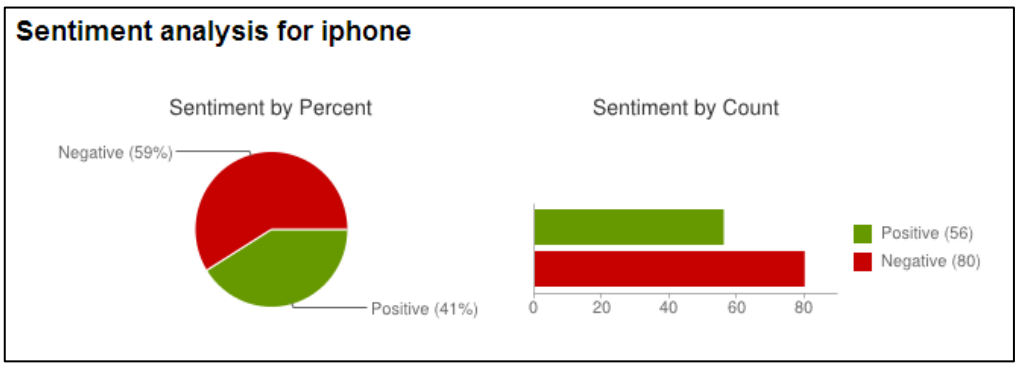
Τελειώνοντας, το παρόν κεφάλαιο παρουσιάζουμε κάποιες από τις πλέον ενδιαφέρουσες εμπορικές εφαρμογές στο χώρο της εξόρυξής γνώμης ή ανάλυσης συναισθήματος από κοινωνικά μέσα και τα οποία είναι διαθέσιμα δωρεάν στο διαδίκτυο. Οι εφαρμογές αυτές είναι εύχρηστες γιατί δεν απαιτούν καθόλου επεξεργασία από τον χρήστη αλλά ούτε και περίπλοκες γνώσεις κατά την χρήση τους. Τα αποτελέσματα που παίρνουν είναι άμεσα αλλά δεν μπορούν να επεξεργαστούν σε μεγάλο όγκο πλήθος πληροφορίας και προτείνεται μια πιο εξειδικευμένη εφαρμογή. Επίσης, τα αποτελέσματα ταξινομούνται σε θετικές ή αρνητικές απόψεις ανάλογα με τα θέμα. Ορισμένα από αυτά παρουσιάζονται παρακάτω και στο τέλος ένας συνοπτικός πίνακας.

Sentiment 140

Η εφαρμογή Sentiment 140 είναι ελεύθερη και λειτουργεί ως μηχανή αναζήτησης, επιστρέφει αποτελέσματα που συνδέονται μόνο από τα σχόλια των χρηστών του Twitter(γιατί συνδέονται πριν την αναζήτηση) και τα ταξινομεί είτε θετικά είτε αρνητικά.

Εικόνα 3.4: Το λογότυπο της Sentiment 140

Ένα χαρακτηριστικό screenshot της Sentiment 140 , μετά από αναζήτηση με θετικά και αρνητικά σχόλια απεικονίζονται πιο κάτω(εικόνα3.5).



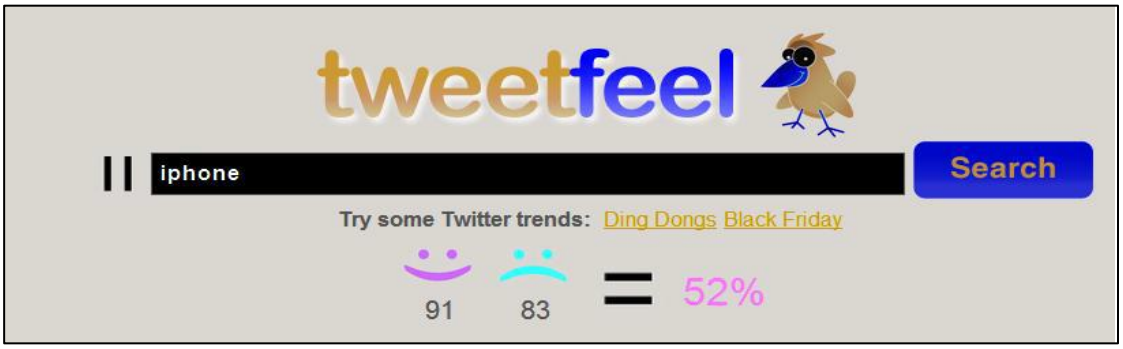
Εικόνα 3.5: Αναζήτηση με θετικά και αρνητικά σχόλια από το λογισμικό Sentiment 140

Tweetfeel

Το Tweetfeel χρησιμοποιεί συναίσθημα ανάλυσης από τις συνομιλίες του Twitter και επιτρέπει να πάρει μια πολύ σαφέστερη εικόνα του συναισθήματος. Υπάρχει σχετικό όριο αναζήτησης και λέξεις-κλειδία και ξεκινά η εφαρμογή. Το tweetfeel χρησιμοποιεί ένα πλήθος από πολύπλοκους αλγόριθμούς για να αξιολογήσει στα tweets αν είναι οι απόψεις θετικές ή αρνητικές.

Εικόνα 3.6 : Το λογότυπο της Tweetfeel

Ένα χαρακτηριστικό screenshot, μετά από αναζήτηση με θετικά και αρνητικά σχόλια από το πιο πάνω λογισμικό (εικόνα3.7).



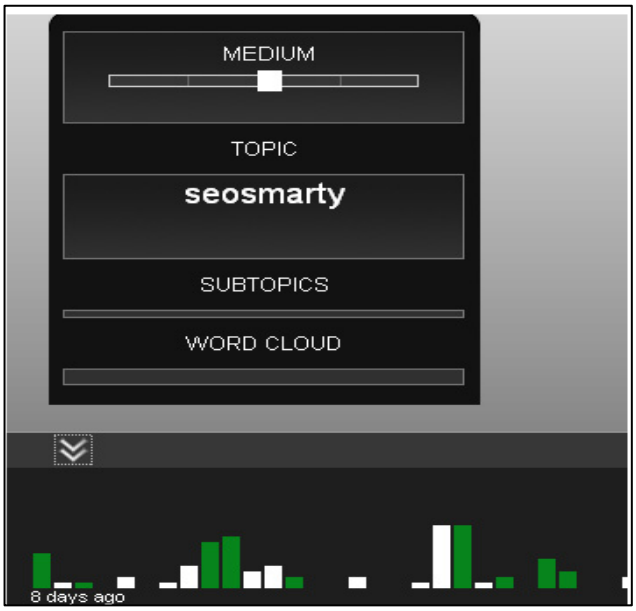
Εικόνα 3.7 : Αναζήτηση με θετικά και αρνητικά σχόλια από το λογισμικό Tweetfeel

Twendz

Το Twendz(εικόνα 3.8) είναι μια εφαρμογή σε ένα πραγματικό χρόνο αναζήτησης που ασχολείται με θέματα του Twitter. Βγάζει λέξεις –κλειδιά και συγκρίνει τις λέξεις κλειδιά σε μια συλλογή από χιλιάδες λέξεις που σχετίζονται με θετικά ή αρνητικά συναισθήματα.

Εικόνα 3. 8 : Το λογότυπο της Twendz

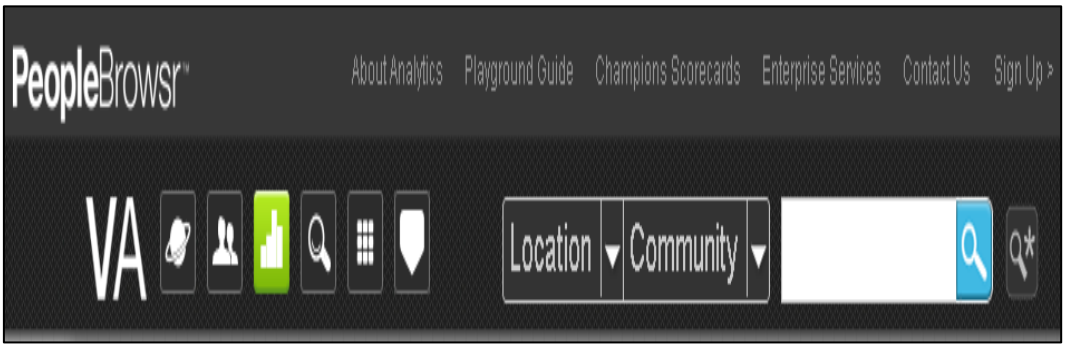
Τα αποτελέσματα της εφαρμογής αυτής εμφανίζονται στην πιο κάτω εικόνα(εικόνα 3.9).



Εικόνα 3. 9: Αναζήτηση με θετικά και αρνητικά σχόλια από το λογισμικό Twendz

PeopleBrowsr

Είναι μια εταιρία τεχνολογίας για επιχειρήσεις στον ιδιωτικό και στον δημόσιο τομέα. Αυτή η εφαρμογή έχει διάφορες λειτουργίες όπως το analytics, search, engagement, Grid. Για την ανάλυση υπάρχει ένα φίλτρο που εμφανίζει τα θετικά ή αρνητικά σχόλια. Στην εικόνα 3.11 εμφανίζονται τα αποτελέσματα μετά την αναζήτηση για θετικά και αρνητικά σχόλια.



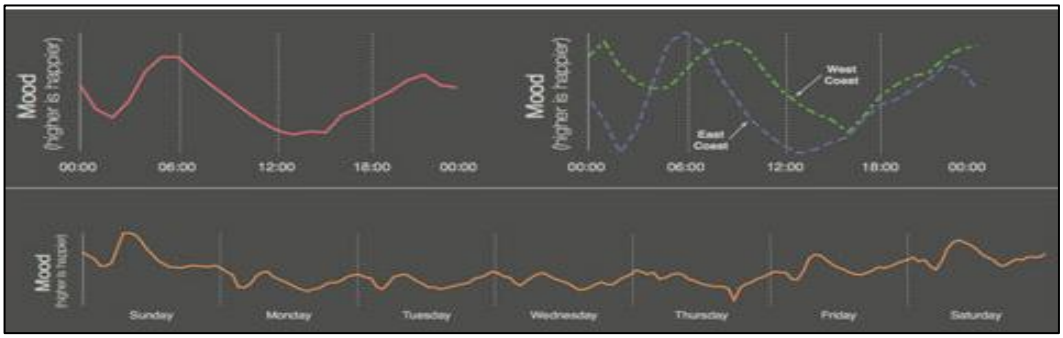
Εικόνα 3.10: Το λογότυπο της PeopleBrowsr

Εικόνα 3.11: Αναζήτηση με θετικά και αρνητικά σχόλια από το λογισμικό PeopleBrowsr

Pulse of the Nation

Αυτή η εφαρμογή(εικόνα 3.12)είναι μια αυτόματη ανάλυση κατά την διάρκεια της ημέρας στις ΗΠΑ από το Twitter(ανάλυση πάνω από 300 εκατομμύρια tweets που δημιουργούνται την ημέρα). Λειτουργεί με ένα αλγόριθμο επεξεργαστή φυσικής γλώσσας. Τα αποτελέσματα είναι θετικά ή αρνητικά και υπολογίζουν το μέσο όρο βαθμολογίας όλων των διαθέσιμών χρηστών. Ένα παράδειγμα της αναζήτησης εμφανίζεται πιο κάτω(εικόνα 3.13) με τις πιο κάτω γραφικές.

Εικόνα 3.12: Το λογότυπο της pulse of the Nation



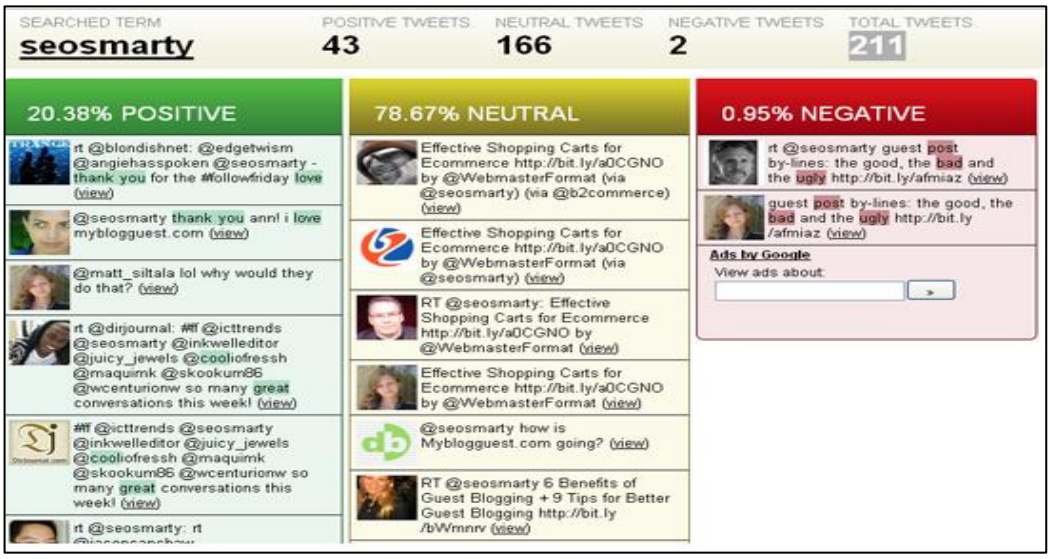
Εικόνα 3. 13: Αναζήτηση με θετικά και αρνητικά σχόλια από το λογισμικό Pulse of the Nation

Twitrratr

Από αυτή την εφαρμογή μπορείς να ανακαλύψεις τι πραγματικά λένε οι άνθρωποι στο Twitter. Μέσα από τα tweets θα διακριθούν αν είναι θετικά ή αρνητικά σχόλια κάνοντας μια αναζήτηση με λέξη- κλειδί στο Twitrratr. Εμφανίζονται τα tweets που ταιριάζουν σε τρεις στήλες από τα αρνητικά, θετικά και ουδέτερα σχόλια.

Εικόνα 3.14 : Το λογότυπο της Twitrratr

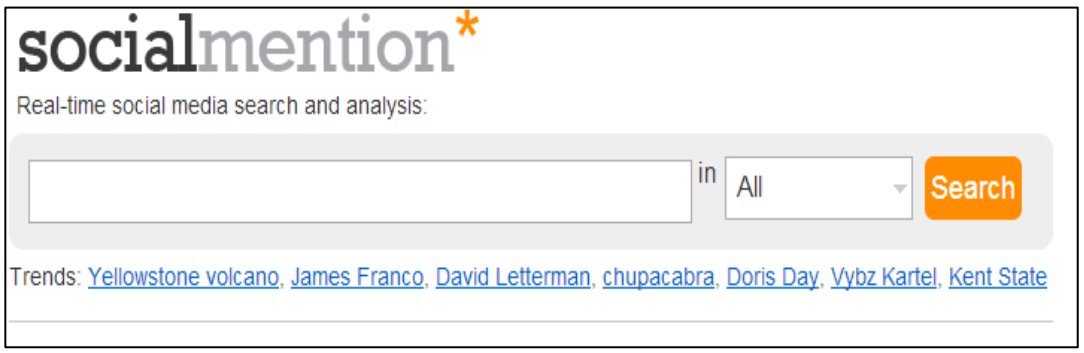
Στην εικόνα 3.15 εμφανίζονται οι 3 στήλες για την ανεύρεση ενός συγκεκριμένου θέματος.



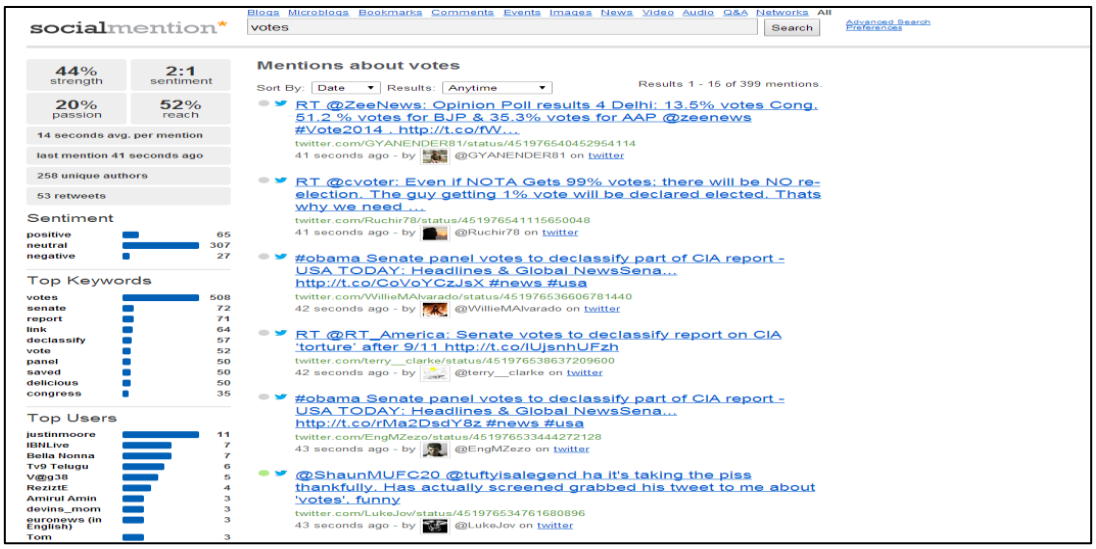
Εικόνα 3. 15: Αναζήτηση με θετικά, αρνητικά και ουδέτερα σχόλια από το λογισμικό Twitrratr

Social Mention

Η εφαρμογή αυτή είναι για τα κοινωνικά μέσα αναζήτησης και ανάλυσης πλατφόρμας. Επιτρέπει να παρακολουθήσεις εύκολα και να μετρήσεις τι λέει ο κόσμος για εσένα, για κάποιον προϊόν και γενικά οτιδήποτε άλλο κοινωνικό μέσο ενημέρωσης στο διαδίκτυο.



Εικόνα 3.16 : Το λογότυπο της Social Mention



Εικόνα 3.17: Αναζήτηση με θετικά και αρνητικά σχόλια του λογισμικού Social Mention

Ακολουθεί παρακάτω ένας συνοπτικός πίνακας(πίνακας 3.6)με όλα όσα αναφέρθηκαν πιο πάνω. Περιλαμβάνει μια σύντομη περιγραφή της εφαρμογή και το σύνδεσμο(url) που έχουν.

Όνομα	Περιγραφή	URL
Sentiment 140	Λειτουργεί ως μηχανή αναζήτησης και επιστρέφει σχόλια χρηστών του Twitter	http://www.sentiment140.com/
Tweetfeel	Η κατηγοριοποίηση γίνεται μέσω λέξεων-κλειδιών στην ελεύθερη έκδοση της εφαρμογής ενώ στην επαγγελματική χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης.	http://www.tweetfeel.com
Twendz	Είναι με ελεύθερη και επαγγελματική έκδοση	http://www.twtbase.com/twendz/
Peoplebrowsr	Αυτή η εφαρμογή είναι εστιασμένη σε κοινωνικά μέσα με πολλαπλές επιλογές αναζήτησης	http://www.peoplebrowsr.com/?option=com_pages&view=sentimently
Pulse of the Nation	Είναι μια αυτόματη ανάλυση της διάθεσης που επικρατεί σε κάθε πολιτεία των Η.Π.Α όως αυτή καταγράφεται μέσω του Twitter	http://www.ccs.neu.edu/home/amislove/twittermood/
Twitrratr	Εισάγοντας μια λέξη-κλειδί επιστρέφονται τα σχόλια ταξινομημένα βάση συναισθήματος σε θετικά. Αρνητικά και ουδέτερα	http://twitrratr.com
Social Mention	Είναι μια online εφαρμογή για συναίσθημα ανάλυσης για κοινωνική δικτύωση	http://socialmention.com/

Πίνακας 3.6: Συνοπτικός πίνακας με τα διαθέσιμα δωρεάν εργαλεία που εστιάζονται στο Twitter

Κεφάλαιο 4ο
Συλλογή Δεδομένων

«Ανά 15λέπτο συλλέγουμε δεδομένα»

Στα προηγούμενα κεφάλαια περιγράψαμε την θεωρία της εξόρυξης γνώμης(opinion mining)και τα διάφορα προβλήματα που έχει, καθώς επίσης την εξόρυξη γνώμης μέσα από τα κοινωνικά δίκτυα και συγκεκριμένα από τα πλατφόρμα του Twitter. Στο συγκεκριμένο κεφάλαιο παρουσιάζεται πως έχει γίνει η συλλογή δεδομένων από το Twitter παίρνοντας δεδομένα από την Κύπρο και από την Αττική - Ελλάδα που είναι γραμμένα στα ελληνικά.

4.1 Εισαγωγή

Η εφαρμογή η οποία έχει αναπτυχθεί επιτελείτε σε τέσσερις επιμέρους βασικές λειτουργίες ώστε να συλλέγουν τα κατάλληλα tweets. Για να αξιοποιήσουμε τις χρήσιμες πληροφορίες που υπάρχουν στο Twitter πρέπει να έχουμε πρόσβαση (δημιουργία λογαριασμού) στην πλατφόρμα Twitter και στην συνέχεια να λαμβάνουμε από εκεί όλα τα απαραίτητα στοιχεία. Το Twitter

παρέχει ένα περιβάλλον διεπαφής(application interface-API) για να το χρησιμοποιούν οι διάφοροι προγραμματιστές και να μπορούν να αναπτύξουν εφαρμογές που συνδέονται με το Twitter. Αυτές οι εφαρμογές μπορούν απλά να λαμβάνουν δεδομένα από το Twitter ή μπορούν να μορφοποιούν το προφίλ κάποιου χρήστη. Για τη δημιουργία τις εφαρμογής πρέπει να επισκεφτείς την ιστοσελίδα των προγραμματιστών του Twitter(Twitter Developers[19]) , εφόσον έχεις είδη λογαριασμό στην πλατφόρμα . Το επόμενο στάδιο είναι να δημιουργήσεις την δική σου εφαρμογή(My Applications- create new app)από τις εφαρμογές που έχει το Twitter (**https://apps.twitter.com/**) . Στην συνέχεια πρέπει να συμπληρώσεις τα κατάλληλα πεδία και τις κατάλληλες ρυθμίσεις εγγραφής, ανάγνωσης(read, write) που επιθυμείς. Στην περίπτωση μας το όνομα της εφαρμογής που δώσαμε είναι «oauth_tweets».

Σε αυτό το σημείο γίνεται η δημιουργία OAuth. Το OAuth είναι ένα ανοικτό πρωτόκολλο[16] ελέγχου ταυτότητας που επιτρέπει στους χρήστες να εγκρίνει την εφαρμογή του λογαριασμού χωρίς να μοιράζεται τον κωδικό του. Είναι μια ασφαλή χορήγηση άδειας, έχοντας μια απλή χρήση μεθόδους από το διαδίκτυο για κινητά και εφαρμογές στην επιφάνεια εργασίας(desktop applications). Δηλαδή μια διαδικτυακή εφαρμογή(web app)Χ προσφέρει ένα API και η συγκεκριμένη ιστοσελίδα(web site)Υ θέλει να χρησιμοποιήσει το API τότε δεν χρειάζεται να χρησιμοποιηθεί ο κωδικός πρόσβασης το Χ γιατί θα ελέγχεται από τον Υ.

Το OAuth χρησιμοποιεί ένα security token που ονομάζεται «access token»[16]. Αυτό το κλειδί είναι μοναδικό που επιτρέπεται μόνο σε όσους έχουν πρόσβαση στο token. Σημαντικό εδώ είναι πως το access token δεν είναι ίδιο με ένα κωδικό πρόσβασης το οποίο επιτρέπει σε όποιον γνωρίζει τον κωδικό πρόσβασης να συνδέεται με μια εφαρμογή. Στην πιο κάτω εικόνα(εικόνα 4. 1) εξηγούνται όλα τα πεδία που περιλαμβάνει το OAuth.

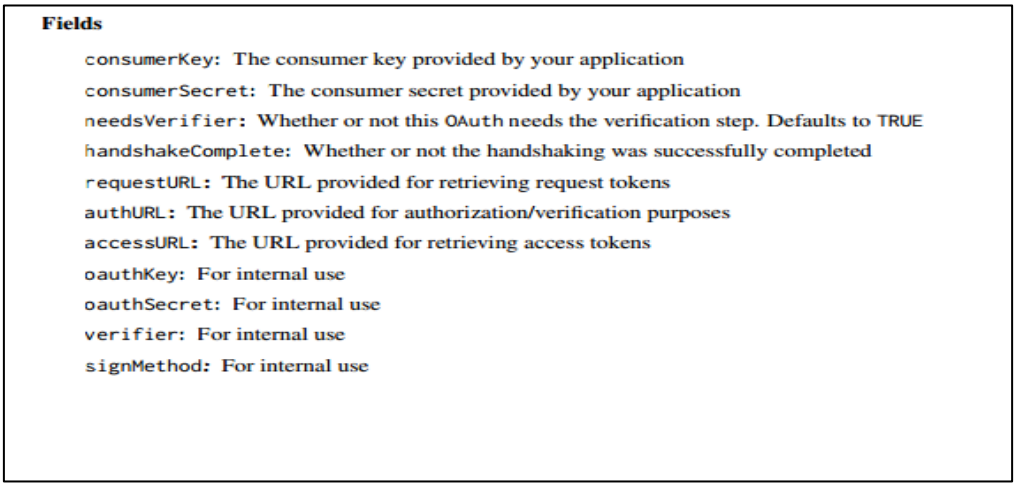
Για την συλλογή και γενικά όλα τα βήματα που θα αναπτυχθούν για την μελέτη μας, θα γίνουν με την γλώσσα προγραμματισμού R. Η γλώσσα προγραμματισμού R είναι ελεύθερα διαθέσιμη στην ιστοσελίδα **http://www.r-project.org/** και στηρίζεται στην ανάπτυξη προγραμμάτων μέσω πακέτων(packages) τα οποία διατίθενται ελεύθερα από χρήστες ανά τον κόσμο. Επιπλέον, χρησιμεύει στην επεξηγηματική ανάλυση δεδομένων και στην εφαρμογή διαφόρων στατιστικών μοντέλων. Επίσης, μπορεί να χρησιμοποιηθεί είτε με κατευθείαν εντολές είτε με προγράμματα τα οποία μπορούν να αναπτυχθούν και να δοθούν για εκτέλεση. Για να συνδέσουμε την γλώσσα προγραμματισμού R με το Twitter και με την εφαρμογή που δημιουργήσαμε πιο πάνω πρέπει να εγκαταστήσουμε τις κατάλληλες βιβλιοθήκες(library) .

[19] **https://dev.twitter.com/**

Μελετώντας από τον οδηγό της γλώσσας προγραμματισμού και ψάχνοντας στο διαδίκτυο βρήκαμε τις κατάλληλές βιβλιοθήκες και τις εγκαταστήσαμε(στην αρχή προσθέσαμε τις βιβλιοθήκες : ROAuth, StreamR). Η βιβλιοθήκη ROAuth παρέχει την διασύνδεση με την OAuth που δημιουργήσαμε πιο πάνω(access token key URL, API key, authentication)(εικόνα 4.2).

Η πιο σημαντική λειτουργία πριν την συλλογή δεδομένων είναι να βρούμε τις σωστές γεωγραφικές συντεταγμένες για την κάθε τοποθεσία που θέλουμε να μελετήσουμε(Κύπρο, περιφέρεια της Αττικής). Ρυθμίζοντας τον κώδικά μας, μετά από κάποιο χρονικό διάστημα που το έχουμε ορίσει εμείς, παίρνουμε ένα σύνολο από δεδομένα με την μορφή «.json» και με βάση κριτηρίων παίρνουμε τα κατάλληλα tweets.

Ακολουθεί αναλυτική περιγραφή της υλοποίησης τις εφαρμογής περιγράφοντας τα βήματα μέχρι να επιλέξουμε ένα ικανοποιητικό αριθμό αναρτήσεων που αφορούν πολιτικά πρόσωπα και να είναι γραμμένα στην ελληνική γλώσσα.



Εικόνα 4.1: Όλα τα πεδία του OAuth

```
library(ROAuth)
## Windows users need to get this file
download.file(url="http://curl.haxx.se/ca/cacert.pem",
              destfile="cacert.pem")
requestURL <- "https://api.twitter.com/oauth/request_token"
accessURL <- "https://api.twitter.com/oauth/access_token"
authURL <- "https://api.twitter.com/oauth/authorize"
consumerKey <-"xxxxxxxxxxxxxxx"
consumerSecret <-"xxxxxxxxxxxxxxxxxxxxxxxxx"
my_oauth <- OAuthFactory$new(consumerKey = consumerKey, consumerSecret = consumerSecret,
                             requestURL = requestURL, accessURL = accessURL, authURL = authURL)
my_oauth$handshake(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl"))
registerTwitterOAuth(my_oauth)
save(my_oauth, file = "my_oauth.Rdata")

library(streamR)
load("my_oauth.Rdata")
registerTwitterOAuth(my_oauth)
```

Εικόνα 4.2 : Δημιουργία OAuth στην R

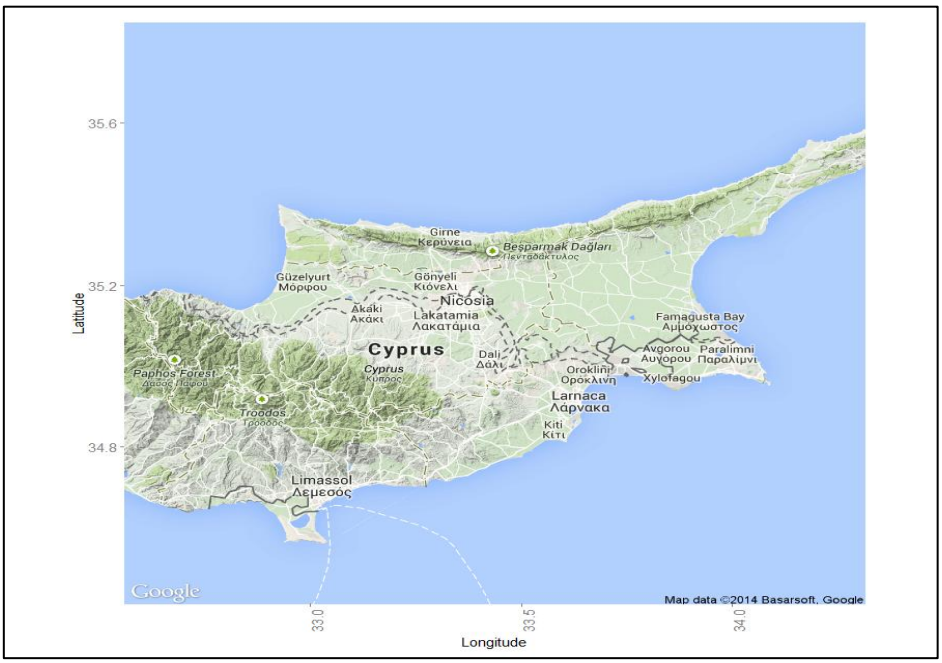
4.2 Διαδικασία Συλλογής Δεδομένων

Όπως αναφέραμε και πιο πάνω η συλλογή δεδομένων έγινε σε δυο φάσεις. Στην Κύπρο για τις Ευρωεκλογές και στην Ελλάδα λόγο των Ευρωεκλογών και των Δημοτικών -Περιφερειακών Εκλογών τον Μάιο 2014 αντίστοιχα.

Οι αναρτήσεις(tweets) των χρηστών συλλέγονται ανά 15 λεπτά και αποθηκεύονται με την μορφή «.json» με δοδημένο τρόπο κατά ώρα και ημέρα ώστε να επιτρέπεται η μελλοντική επεξεργασία και αξιοποίηση των δεδομένων για την ανάλυση τους. Τα Tweets αυτά προέρχονται από διάφορους χρήστες του Twitter από την Κύπρο και από την περιφέρεια της Αττικής. Η διαδικασία συλλογής δεδομένων δεν έγινε με κάποιο κριτήριο(ηλικία, συγκριμένα άτομα κλπ). Σε επόμενα στάδια φιλτράρουμε με συγκεκριμένα κριτήρια για να πάρουμε τα δεδομένα που θέλουμε να μελετήσουμε. Ο σκοπός της εργασίας αυτής είναι να συλλέξουμε ένα ικανοποιητικό αριθμό δεδομένων , ελληνικών αναρτήσεων που να αναφέρονται σε πολιτικές συζητήσεις. Οι πιο κάτω ενότητες περιγράφουν τα δεδομένα από Κύπρο και Αττική , τι περιλαμβάνουν τα αρχεία τους και τι αποτελέσματα προκύπτουν.

4.3 Δεδομένα από Κύπρο

Η περίοδος συλλογής δεδομένων έγινε από τις 26 Ιουνίου μέχρι τις 8 Σεπτεμβρίου του 2014, αναζητώντας από όλο το νησί περιλαμβάνοντας και τα κατεχόμενα εδάφη. Η αναζήτηση έγινε με τις κατάλληλες συντεταγμένες της Κύπρου(bounding box) όπου είναι [32.273090, 34.563511, 34.597919, 35.701542] και όπως απεικονίζεται στο πιο κάτω χάρτη(εικόνα 4.3).



Εικόνα 4.3 : Η περιοχή που μαζέψαμε τα δεδομένα από την Κύπρο

Η συλλογή δεδομένων είχε διάρκεια 75 μέρες και τα αρχεία json που δημιουργήθηκαν είχαν μέγεθος 732 MB , με το μεγαλύτερο μέγεθος να είναι στα 992 KB και το μικρότερο μέγεθος να αντιστοιχεί σε μηδέν(0). Δηλαδή έχουμε 287886 tweets.

Πρώτο στάδιο- Γλώσσες

Στο Twitter στο πεδίο γλώσσα 'lang' είναι το χαρακτηριστικό που καθορίζει την γλώσσα του tweet που συντάχθηκε από τους αλγόριθμους ανίχνευσης της γλώσσας μηχανής. Η γλώσσα 'lang' μπορεί να αντιπροσωπεύει οποιαδήποτε γλώσσα που υπάρχει στη διαθέσιμη λίστα γλωσσών του Twitter ή να είναι 'und' που δηλώνει πως δεν ανιχνευτικέ καμία γλώσσα.

Κατά τη διαδικασία συλλογής δεδομένων εξετάσαμε προσεκτικά πως υπάρχει μια ποικιλία από γλώσσες. Υπάρχουν 43 γλώσσες και οι έξι πιο κορυφαίες κατά αύξουσα σειρά(πίνακας 4.1) είναι

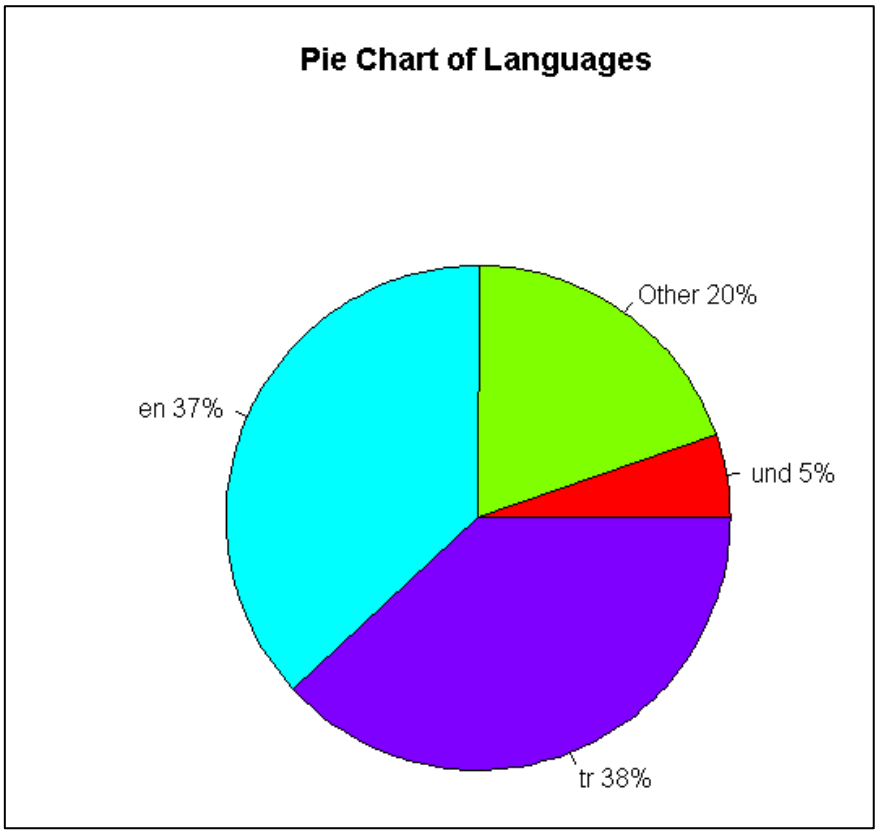
τα Τούρκικα ΄tr΄ , Αγγλικά ΄en΄, Καμία γλώσσα ΄und΄, Ρωσικά ‘ru’, Ελληνικά ΄el΄ και Αγγλικά της Ινδίας ΄in΄.

Παρατηρούμε πως τα Ελληνικά βρίσκονται στην πέμπτη θέση έχοντας μόνο 8049 tweets, πιο λίγα και από τα Ρωσικά που είναι στα 13949 tweets. Στην έκτη θέση βρίσκονται τα Αγγλικά της Ινδίας με 5620 tweets.

Α/Α	**Γλώσσα**	**Κωδικός**	**Tweets**
1	Τούρκικα	‘tr’	109717
2	Αγγλικά	‘en’	106415
3	Καμία γλώσσα	‘und’	15383
4	Ρωσικά	‘ru’	13949
5	Ελληνικά	‘el’	8049
6	Ινδιάνικα- Αγγλικά	‘in’	5620

Πίνακας 4. 1 : Οι έξι δημοφιλέστερες γλώσσες από την Κύπρο με αύξουσα σειρά

Μελετώντας τις τρείς δημοφιλέστερες γλώσσες όπου δίνονται στην εικόνα 4.4. Τα Τούρκικα βρίσκονται στην πρώτη θέση με 38% έχοντας 3302 tweets περισσότερα από την δεύτερη θέση όπου είναι τα Αγγλικά με 37%. Στην τρίτη θέση των δημοφιλέστερων γλωσσών είναι το ΄und ΄ δηλαδή καμία γλώσσα έχοντας 15383 tweets παίρνοντας το 5% όπου είναι κάποια βίντεο και εικόνες που αναρτούν οι χρήστες. Για να τα μάθουμε όλα αυτά εκτελέσαμε τις πιο κάτω εντολές(εικόνα 4.5) .Τις υπόλοιπες 40 γλώσσες τις ομαδοποιήσαμε στο τμήμα ‘other’ παίρνοντας το 19%. Παρατηρούμε ότι οι περισσότεροι χρήστες του Twitter στην Κύπρο είναι Τουρκοκύπριοι από τα κατεχόμενα εδάφη.



Εικόνα 4.4: Οι τρεις δημοφιλέστερες γλώσσες από τα δεδομένα της Κύπρου

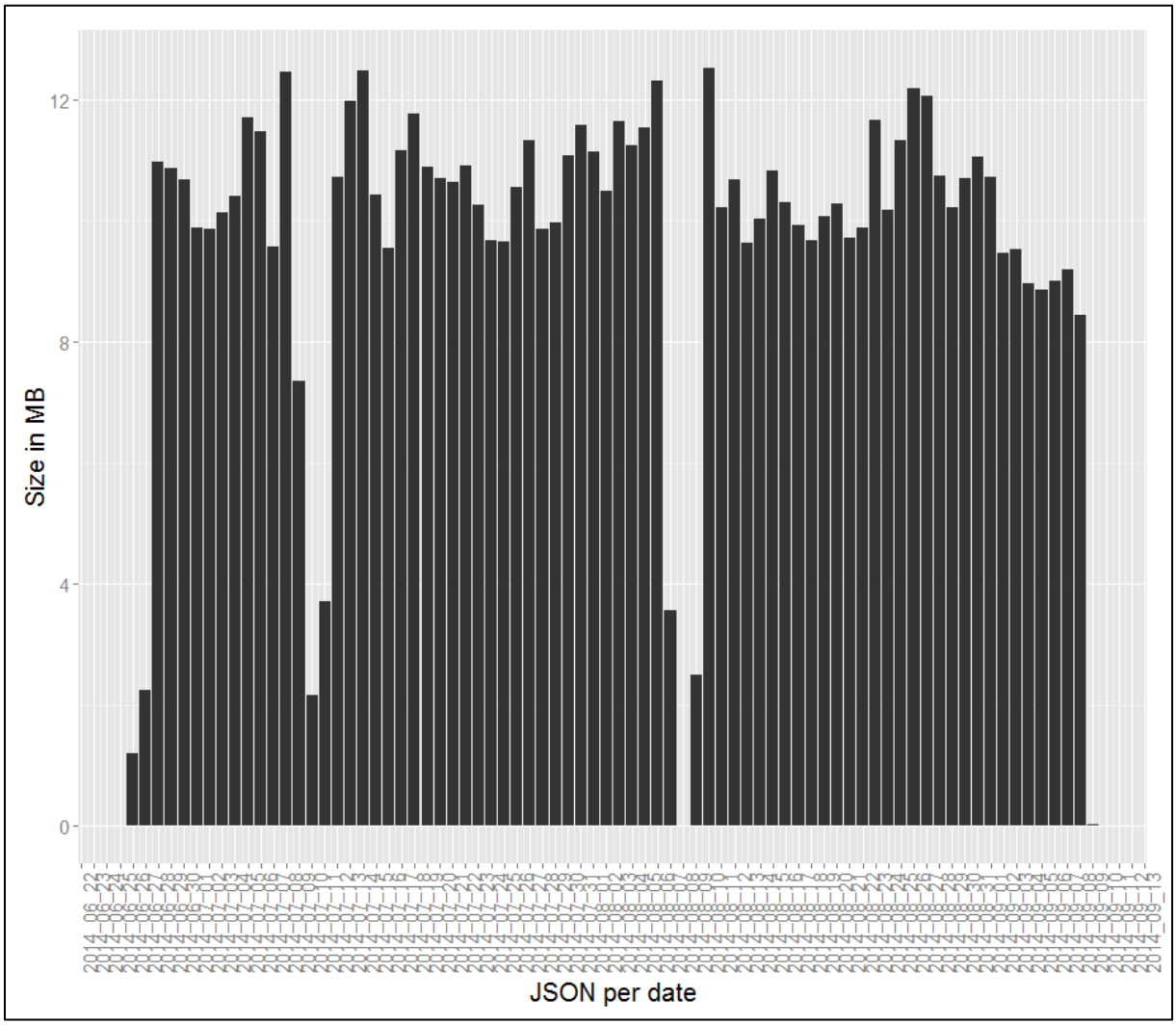
```
#save tweets with und language
new_df_tweets_und <- new_df_tweets[new_df_tweets$lang %in% 'und', ]
writeLines(new_df_tweets_und$text, "text_of_und.txt")

#-------------------------------------------------------------------------
```

Εικόνα 4. 5: Κώδικας αποθήκευσης 'und' language

Δεύτερο στάδιο- Όλα τα αρχεία json

Η γραφική παράσταση που ακολουθεί(εικόνα 4.6) απεικονίζει το συνολικό μέγεθος των αρχείων json ανά μέρα. Στον άξονα Χ είναι οι 75 μέρες έχοντας την μορφή «YY-MM-DD»(Έτος-Μήνας-Μέρα) και στον άξονα τον Υ το συνολικό μέγεθος(total size) με την μορφή MB.

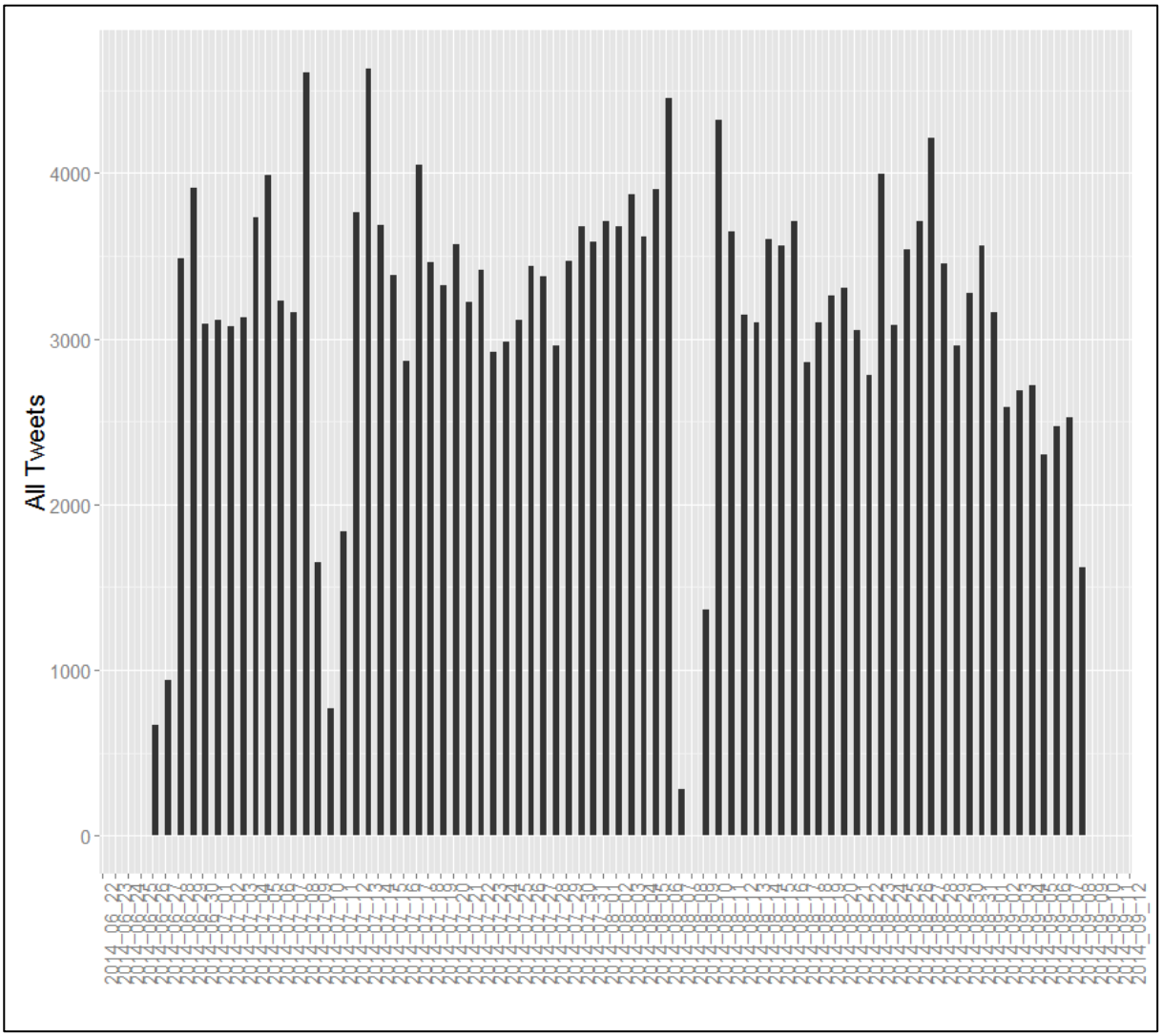


Εικόνα 4.6: Συνολικό μέγεθος των αρχείων json ανά μέρα, από τα δεδομένα της Κύπρου

Στις 26 Ιούνη όπου ήταν η έναρξη συλλογής δεδομένων ξεκίνησε το απόγευμα γι' αυτό υπάρχουν λίγα αρχεία json. Ακόμα κατά τις μέρες 27 Ιούνη, 9 ,10 Ιούλη και 6 Αυγούστου υπήρξε διακοπή ρεύματος για κάποιες ώρες ενώ στις 7 Αυγούστου υπήρξε διακοπή ρεύματος όλη την μέρα. Επίσης παρατηρούμε το μεγαλύτερο μέγεθος να βρίσκεται σε τρεις μέρες με σχεδόν το ίδιο μέγεθος αρχείων. Στις 7 , 13 Ιουλίου και στις 9 Αυγούστου με μέγεθος 12,5 MB. Ως γενικό συμπέρασμα για το μέγεθος των αρχείων είναι ότι στις περισσότερες μέρες το μέγεθος είναι περίπου 8 με 12 MB.

Τρίτο στάδιο- Όλα τα tweets

Η γραφική παράσταση που απεικονίζεται(εικόνα 4. 7)είναι οι συνολικές αναρτήσεις(tweets) ανά μέρα. Στον άξονα τον Χ είναι οι μέρες(26 Ιουνίου μέχρι 8 Σεπτεμβρίου) και στον άξονα Υ είναι πόσα συνολικά μηνύματα tweets έχουμε ανά μέρα.

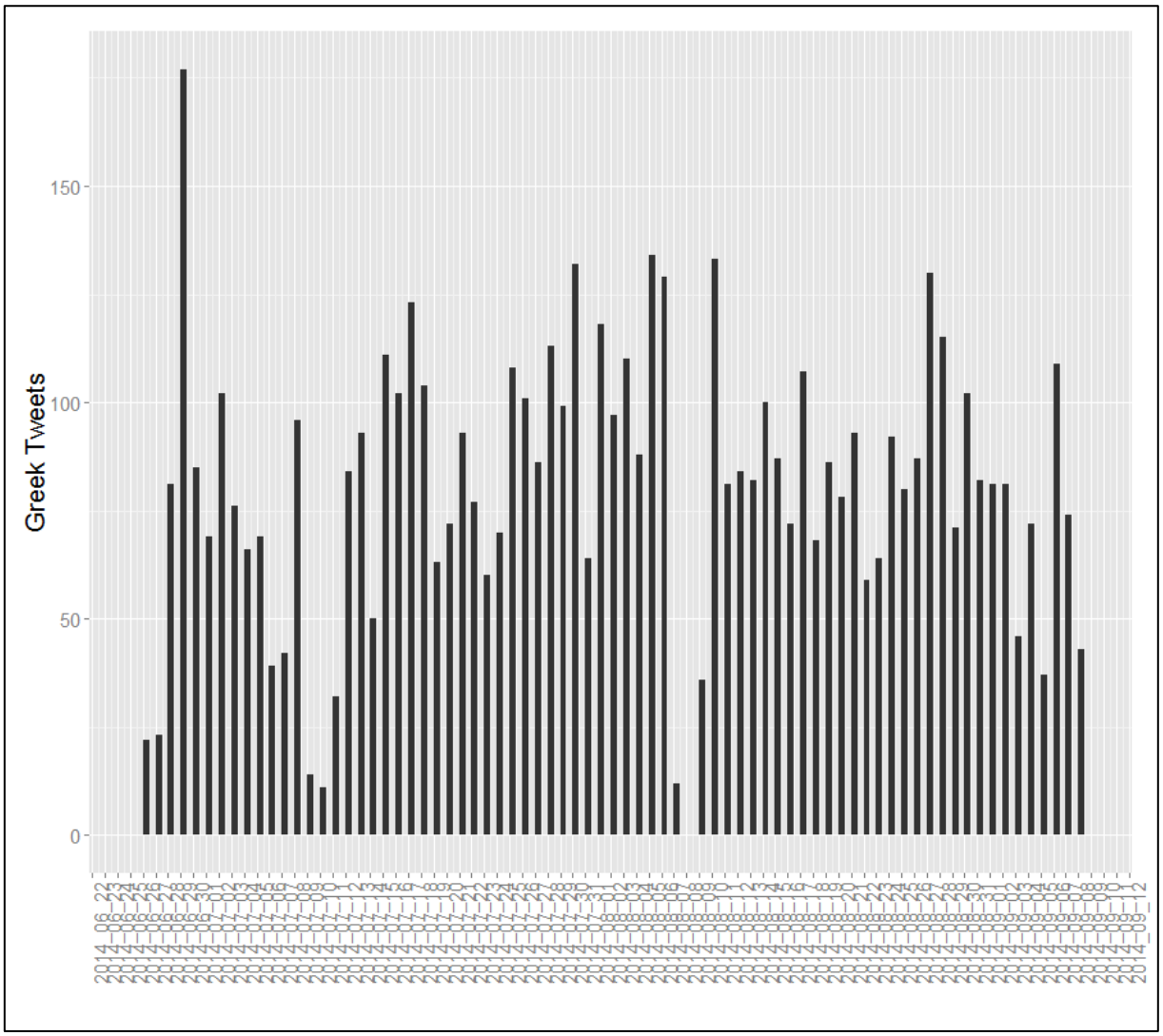


Εικόνα 4.7: Συνολικές αναρτήσεις(tweets)ανά μέρα από δεδομένα της Κύπρου

Για τις μέρες με τις πιο χαμηλές τιμές ή το κενό που υπάρχει αναφέραμε στην πιο πάνω υπό-ενότητα τον λόγο. Το πιο ψηλό σημείο με τις περισσότερες αναρτήσεις είναι στις 12 Ιουλίου με ελάχιστη διαφορά πιο κάτω να βρίσκεται στις 7 Ιουλίου. Τα tweets αυτές τις μέρες είναι περίπου στα 4520. Γενική παρατήρηση είναι ότι οι αναρτήσεις για όλες της μέρες δεν είναι σταθερές γι'αυτό είναι ασταθές διάγραμμα(μια πάνω μια κάτω).

Τέταρτο στάδιο- Επιλογή Ελληνικών αναρτήσεων

Στο πιο κάτω ιστόγραμμα(εικόνα 4.8)απεικονίζει τις αναρτήσεις που είναι μόνο γραμμένα στην Ελληνική γλώσσα. Στον άξονα Χ είναι ανά μέρα(26 Ιουνίου- 8 Σεπτεμβρίου) και ο άξονας Υ είναι το σύνολο των αναρτήσεων ανα μέρα που είναι γραμμένα στα ελληνικά.



Εικόνα 4. 8: Συνολικές αναρτήσεις(tweets) ανά μέρα από δεδομένα της Κύπρου

Η μεγάλη διαφορά που υπάρχει στο γράφημα είναι στις 28 Ιουνίου με τα περισσότερα tweets από όλη την διάρκεια συλλογής των δεδομένων. Τα επόμενα υψηλά σχόλια βρίσκονται γύρο στα 125 tweets με τις ακόλουθες ημερομηνίες 27 Ιουλίου, 4,9 και 26 Αυγούστου.

Στο πιο πάνω γράφημα οι συνολικές αναρτήσεις ανά μέρα βρίσκονται μεταξύ στα 50 και 125 tweets. Έτσι με αυτά τα ελληνικά tweets είναι πολύ δύσκολο να αναλυθούν αλλά και να εντοπιστούν πολιτικές συζητήσεις.

Πέμπτο στάδιο- Πολιτικά ονόματα και λέξεις κλειδιά

Σε αυτό το στάδιο δημιουργήσαμε δυο λίστες με τα πολιτικά ονόματα βουλευτών, υπουργών, ευρωβουλευτών, προέδρων κομμάτων, νυν και πρώην πρόεδρών της κυπριακής δημοκρατίας κλπ. Και μια λίστα με λέξεις κλειδιά που σχετίζονται με πολιτικές συζητήσεις(πίνακας 4.1) .

Λέξεις κλειδιά				
Αντιπολίτευση	Συνέδρια	Δημοσιογράφος	Υπουργός	Κόμμα
Πρόεδρος	Συνέντευξη	Κύπρος	Νομοσχέδιο	Προεκλογική
Κυβέρνηση	Ψήφος	Πολιτική	Ευρωβουλή	Βουλευτής
Κόμματα	Γραφείο	Δήμος	Ευρώπη	Γραμματεία
Θέμα	Πατρίδα	Κατεχόμενα	Τράπεζα	Κούρεμα
Επενδυτές	ΔΗΚΟ	ΔΗΣΥ	ΕΔΕΚ	ΑΚΕΛ

Πίνακας4.1 : Λέξεις –κλειδιά για τα δεδομένα της Κύπρου

Πρώτα κάναμε την αναζήτηση με την λίστα λέξεις κλειδιά και πήραμε 131 tweets. Στην συνέχεια έγινε αναζήτηση με τα υπόλοιπες λίστες(πίνακας 4.2) για τους πολιτικούς. Αυτές οι δυο λίστες είναι γραμμένα τα επίθετα των πολιτικών και στην μια δεν υπάρχουν τόνοι γιατί μπορεί κάποιοι χρήστες όταν αναρτούν να μην βάζουν τόνους στις λέξεις. Η επιλογή των πολιτικών είναι με βάση τα τρία δημοφιλέστερα κόμματα στην Κύπρο(ΔΗΣΥ, ΑΚΕΛ ,ΔΗΚΟ) . Επίσης υπάρχει ακόμα μια λίστα με όσους πολιτικούς έχουν λογαριασμό στο Twitter. Στον πίνακα 4.2 εμφανίζονται τα πολιτικά ονόματα και όσοι από αυτούς έχουν λογαριασμό στο Twitter . Ωστόσο, κάνοντας αναζήτηση και με αυτές τις λίστες δεν εντοπιστικέ καμία ανάρτηση με πολιτικές συζητήσεις.

Α/Α	Ονοματεπώνυμο, λογαριασμός του Twitter από πολιτικούς στην Κύπρο		
1	Ελένη Θεοχάρους @THEOCHAROUSE	Τάκης Χατζηγεωργίου @thadjigeorgiou	Χρήστος Στυλιανίδης @Stilianides
4	Νεοκλής Συλικιώτης @sylikiotis	Κώστας Μαυρίδης	Δημήτρης Παπαδάκης @DemPapadakis
7	Νίκος Αναστασιάδης @AnastasiadesCY	Ιωανάς Νικολάου @MJPO INicolaou	Χάρης Γεωργιάδης @Georgiades H
10	Χριστόφορος Φωκαίδης @cfokaides	Γιώργος Λακκοτρύπης @GLakkotrypis	Κώστας Κάδης @CostasKadis
13	Νίκος Τορναρίτης @nicostornaritis	Αβέρωφ Νεοφύτου @AverofCY	Πρόδρομος Προδρόμου @pprodromou
16	Ειρήνη Χαραλάμπους @Xaralambidou	Άριστος Δαμιανού @AristosDamianou	Αντιγόνη Παπαδοπούλου @AntPapadopoulou
19	Νικόλας Παπαδόπουλος @NicholasPapadop	Δημήτρης Συλλούρης @Syllouris	Δημήτρης Χριστόφιας @ChristofiasD
22	Σταύρος Μαλάς @MalasStavros	Ελένη Μαύρου @Eleni Mavrou	Χάρης Πολυκάρπου
25	Σταύρος Ευαγόρου	Κωνσταντίνος Αριστείδου	Γιώργος Γεωργίου
28	Μάριος Καρογιάν	Γιώργος Προκοπίου	Αντώνης Αντωνίου
31	Άγγελος Βότσης	Χριστίνα Ερωτοκρίτου	Άντρος Κυπριανού
34	Νίκος Κατσουρίδης		

Πίνακας 4.2: Τα πολιτικά ονόματα της Κύπρου και όσοι έχουν λογαριασμό στο Twitter

Γενικές παρατηρήσεις από δεδομένα την Κύπρου

Από την συλλογή των δεδομένων της Κύπρου σε 75 μέρες έχει μαζευτεί μικρό μέγεθος αναρτήσεων. Παρατηρούμε πως δεν υπάρχει γενική χρήση τουTwitter στην Κύπρο. Ακόμη και αυτές οι αναρτήσεις δεν ασχολούνται με πολιτικά θέματα, όταν κάναμε την αναζήτηση με τις λίστες των πολιτικών ονομάτων και λέξεις κλειδιά.

Οι περισσότεροι χρήστες είναι Τουρκοκύπριοι/ Τούρκοι όπως το είδαμε στο γράφημα με τις δημοφιλέστερες γλώσσες. Στα κατεχόμενα εδάφη γίνεται η περισσότερη χρήση του Twitter αντί στην ελεύθερη Κύπρο.

Υπάρχει το ενδεχόμενο οι Κύπριοι να χρησιμοποιούν άλλα μέσα κοινωνικών δικτύων(όπως το face book)αλλά γενικά δεν συζητούν τις πολιτικές του απόψεις στις διάφορες πλατφόρμες. Αφού από την μελέτη που είδαμε των αναρτήσεων οι κύπριοι αναρτούν τι κάνουν εκείνη την συγκεκριμένη στιγμή, εκφράζοντας τους προβληματισμούς τους, τα προβλήματα τους και απόψεις για αθλητικά θέματα. Πιο κάτω δείχνουμε ενδεικτικά και τυχαία τέσσερα παραδείγματα tweets από χρήστες της Κύπρου.

Παράδειγμα 1 : *«Καλη επιτυχια ΕΛΛΑΔΑ!!!!*

Το μπλε σου χρωμα το τιμας για χρονια!!!

Ο ουρανος η θαλασσα και συ...

ΕΛΛΑΣ ΕΛΛΑΣ..»

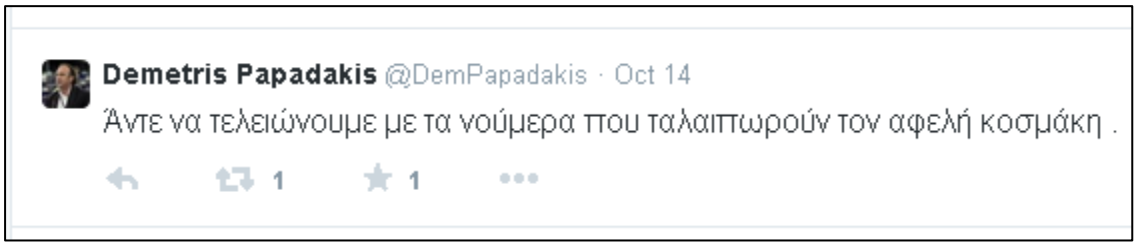
Παράδειγμα 2 : *«ΕΧΕΙ ΚΑΤΑΡΤΙΣΤΕΙ ΤΟ ΚΑΙΝΟΥΡΙΟ ΔΙΟΙΚΗΤΙΚΟ ΣΥΜΒΟΥΛΙΟ ΤΟΥ ΠΑΝ.ΣΥ.ΦΙ. ΑΠΟΕΛ. ΕΥΧΟΜΑΣΤΕ ΟΛΟΙ ΚΑΛΗ ΕΠΙΤΥΧΙΑ ΣΤΟ ΔΥΣΚΟΛΟ ΕΡΓΟ ΠΟΥ ΕΧΕΙ ΑΝΑΛΑΒΕΙ!"»*

Παράδειγμα 3 : *«Με αναγκάζουν να φύγω από παραλία για να πάω βαφτίσια. Έλεος ρε.»*

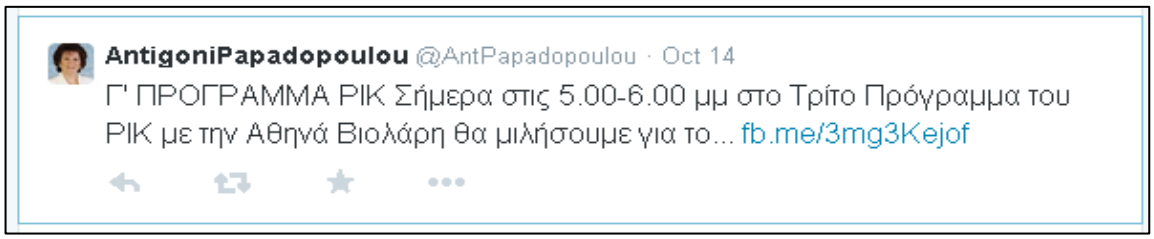
Παράδειγμα 4 : *«Μετά από ένα τόσο όμορφο Σάββατοκυριακο πως να αντέξω μία ανιαρή και βαρετή Δευτέρα;»*

Επίσης πιο κάτω παρουσιάζουμε τέσσερα τυχαία πολιτικά ονόματα που έχουν αναρτήσει στο Twitter στις 14 Οκτωβρίου 2014.

Εικόνα 4.9: Tweet από τον Πρόεδρο της Κυπριακής Δημοκρατίας κ. Αναστασιάδη



Εικόνα 4.10: Tweet από τον ευρωβουλευτή κ. Δημήτρη Παπαδάκη

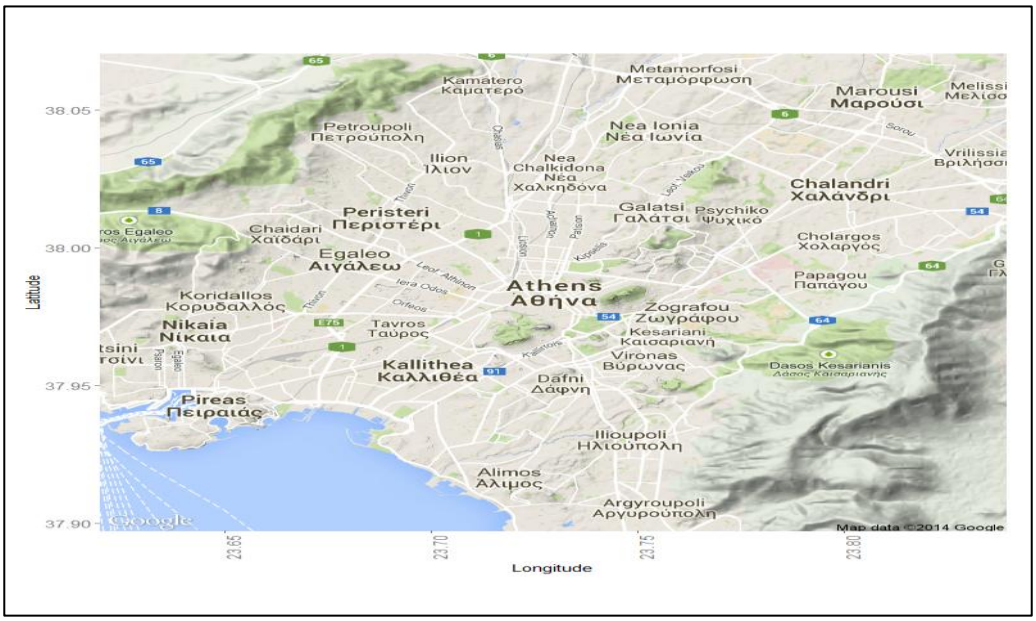


Εικόνα 4.11: Tweet από την βουλευτή κ. Αντιγόνη Παπαδοπούλου

Εικόνες 4 12 : Tweet από ευρωβουλευτή κ.Νεοκλη Συλικίωτη

4.4 Δεδομένα από την Αττική-Ελλάδα

Η περιόδος συλλογής δεδομένων έγινε από τις 12 Μαιου μέχρι τις 30 Ιουνίου 2014, αναζητώντας από την περιφέρια της Αττικής. Οι συντεταγμένες της Αττικής είναι [NE 38.033428, 23.789761 SW 37.948799, 23.686939] όπως φαίνεται και από την εικόνα 4.13 .



Εικόνα 4.13: Η περιφέρεια Αττικής με βάση τις συντεταγμένες που δώσαμε από το bounding box

Η διάρκεια των 50 ημερών είχε συνολικό μέγεθος 1.86 GB αρχείων json, με μεγαλύτερο μέγεθος 2410 KB και το μικρότερο να αντιστοιχεί σε μηδέν(0). Επίσης αυτά τα αρχεία json έχουν συνολικά 664714 tweets.

Πρώτο στάδιο- Γλώσσες

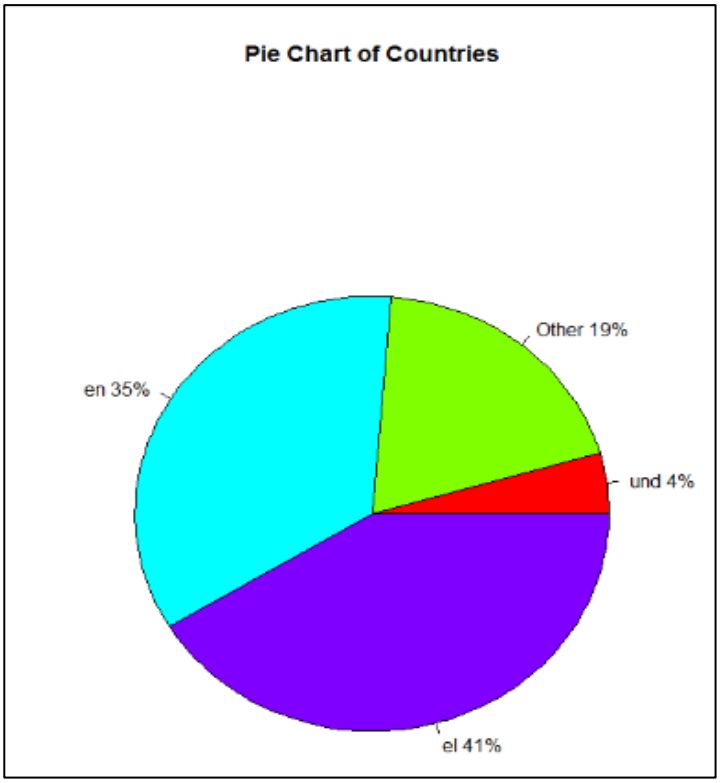
Μελετώντας τα tweets παρατηρούμε πως υπάρχουν πολλά που είναι γραμμένα σε διάφορες γλώσσες . Κάνοντας αναζήτηση πόσες γλώσσες περιλαμβάνονται από 43 γλώσσες και οι έξι πιο κορυφαίες ανά αύξουσα σειρά είναι τα Ελληνικά ΄el΄, Αγγλικά ‘en’, Καμία γλώσσα ‘und’, Τούρκικα ΄tr΄, Ρωσικά ΄ru΄ και τα Ισπανικά ΄es΄(πινάκας 4.4).

Α/Α	Γλώσσα	Κωδικός	Tweets
1	Ελληνικά	'el'	274835
2	Αγγλικά	'en'	232108
3	Καμία γλώσσα	'und'	29741
4	Τούρκικα	'tr'	12561
5	Ρωσικά	'ru'	10483
6	Ισπανικά	'es'	9284

Πίνακας 4.4: Οι έξι δημοφιλέστερες γλώσσες από την Αττική με αύξουσα σειρά

Τα Τούρκικα είναι περισσότερα κατά 2078 από τα Ρωσικά και από τα Ισπανικά κατά 3277. Ακόμη, με αυτά τα αποτελέσματα παρατηρούμε πως υπάρχουν αρκετοί Τούρκοι, Ρώσοι και Ισπανοί που μένουν στην Αττική είτε είναι για διακοπές είτε είναι μόνιμοι κάτοικοι καθ θέλουν να αναρτήσουν κάποιο θέμα που θέλουν να μοιραστούν με τους υπόλοιπους χρήστες του Twitter.

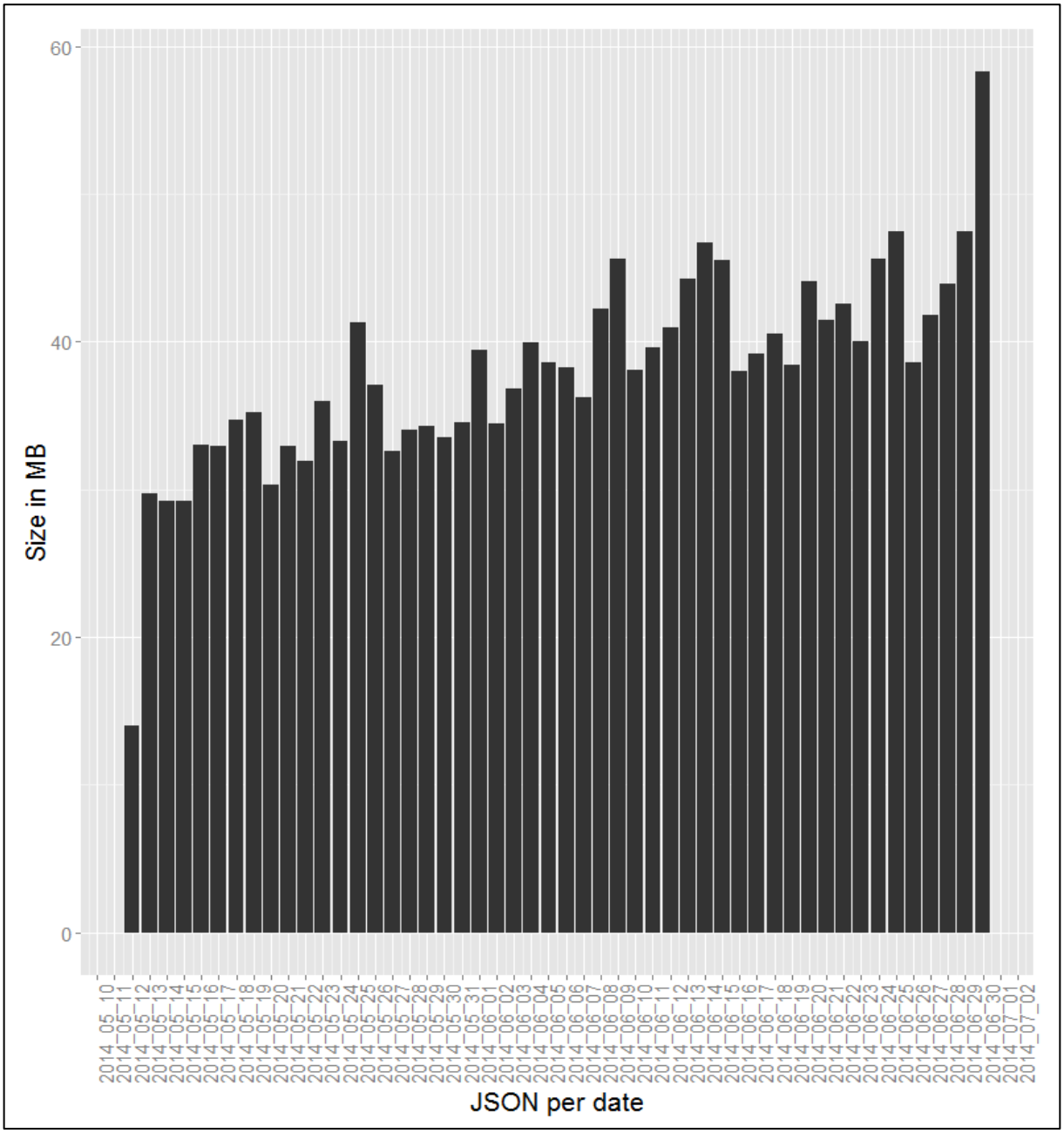
Τα συγκεντρωτικά αποτελέσματα των τριών δημοφιλέστερων γλωσσών δίνονται στην εικόνα 4.14. Τα Ελληνικά δεσμεύονται με το 41% , σε δεύτερη θέση έρχονται τα Αγγλικά με 35% ποσοστό. Η Τρίτη δημοφιλέστερη γλώσσα είναι το ΄und΄ το οποίο το αναφέραμε στην προηγούμενη ενότητα(ενότητα 4.3.1) που και στα Ελληνικά δεδομένα σημαίνει το ίδιο. Τις υπόλοιπες 40 γλώσσες τις ομαδοποιήσαμε στο τμήμα 'other' έχοντας 19%. Από τα ποσοστά του γραφήματος η χρήση των Ελληνικών αναρτήσεων είναι αρκετά ικανοποιητικό. Στα αγγλικά ωστόσο μπορεί να είναι έλληνες που γράφουν Greek- English .



Εικόνα 4.14: Οι τρείς δημοφιλέστερες γλώσσες από την συλλογή δεδομένων της Αττικής

Δεύτερο στάδιο- Όλα τα αρχεία json

Η γραφική παράσταση που ακολουθεί(εικόνα4.15) είναι το συνολικό μέγεθος των αρχείων json ανά μέρα. Στον άξονα Χ είναι οι 50 μέρες έχοντας την μορφή 'YY-MM-DD'(Έτος-Μήνας- Μέρα) και στον άξονα Υ είναι το συνολικό μέγεθος(total size) με την μορφή MB.

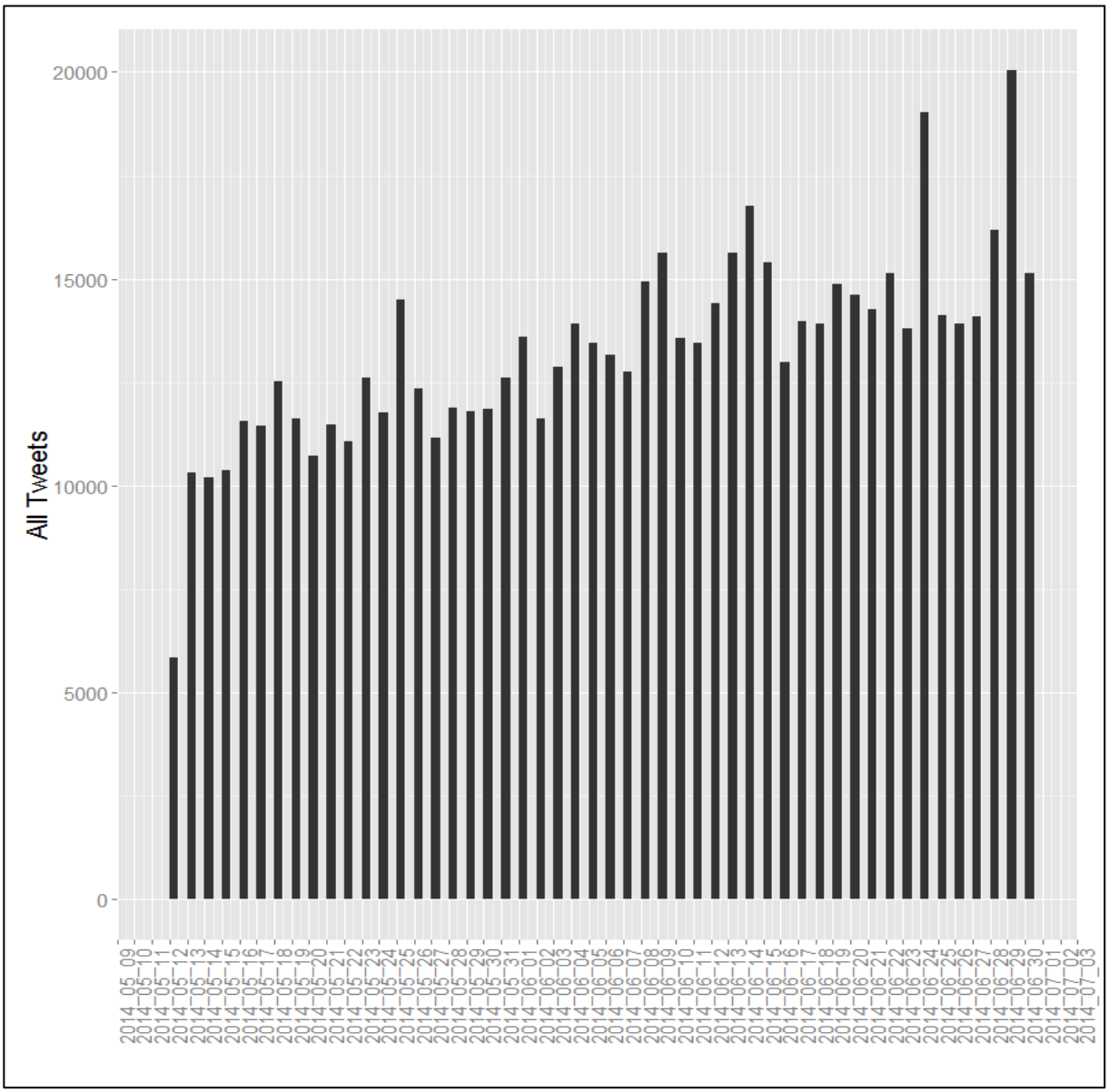


Εικόνα 4.15: Συνολικό μέγεθος αρχείων json ανά μέρα από δεδομένα της περιφέρειας Αττικής

Παρατηρούμε το μεγαλύτερο μέγεθος να είναι στις 30 Ιουνίου περίπου 58MB σε σύγκριση με τις 12 Μαΐου παίρνοντας το μικρότερο μέγεθος. Οι αναρτήσεις που υπάρχουν στις 30 Ιουνίου οφείλεται ότι αρκετοί φοιτητές τέλειωσαν τις εξετάσεις τους, ξεκινά το καλοκαίρι, συζητήσεις για το ποδόσφαιρο-Εθνική Ελλάδος. Ενώ στις 12 Μαΐου είναι το μικρότερο μέγεθος γιατί η έναρξη συλλογής δεδομένων ξεκίνησε το απόγευμα Την μέρα την δημοτικών εκλογών και Ευρωεκλογών(18 και 25 Μαΐου)με μεγέθη των αρχείων έφτασαν στα 36 MB και 41MB αντίστοιχα. Ως γενικό συμπέρασμα για το μέγεθος των αρχείων είναι ότι τα μεγέθη κυμαίνονται μέχρι τα 35 MB εκτός από ορισμένες μέρες που ξεπερνούν τα 42 MB.

Τρίτο στάδιο- Όλα τα tweets

Από την γραφική που είχαμε πιο πάνω με το σύνολο των αρχείων, στην πιο κάτω γραφική που απεικονίζεται(εικόνα 4.16) είναι οι συνολικές αναρτήσεις(tweets) ανά μέρα. Στον άξονα Χ έχουμε πάλι τις 50 μέρες(΄YY-MM-DD΄) και στην άξονα την Υ είναι πόσα συνολικά tweets έγιναν ανά μέρα.



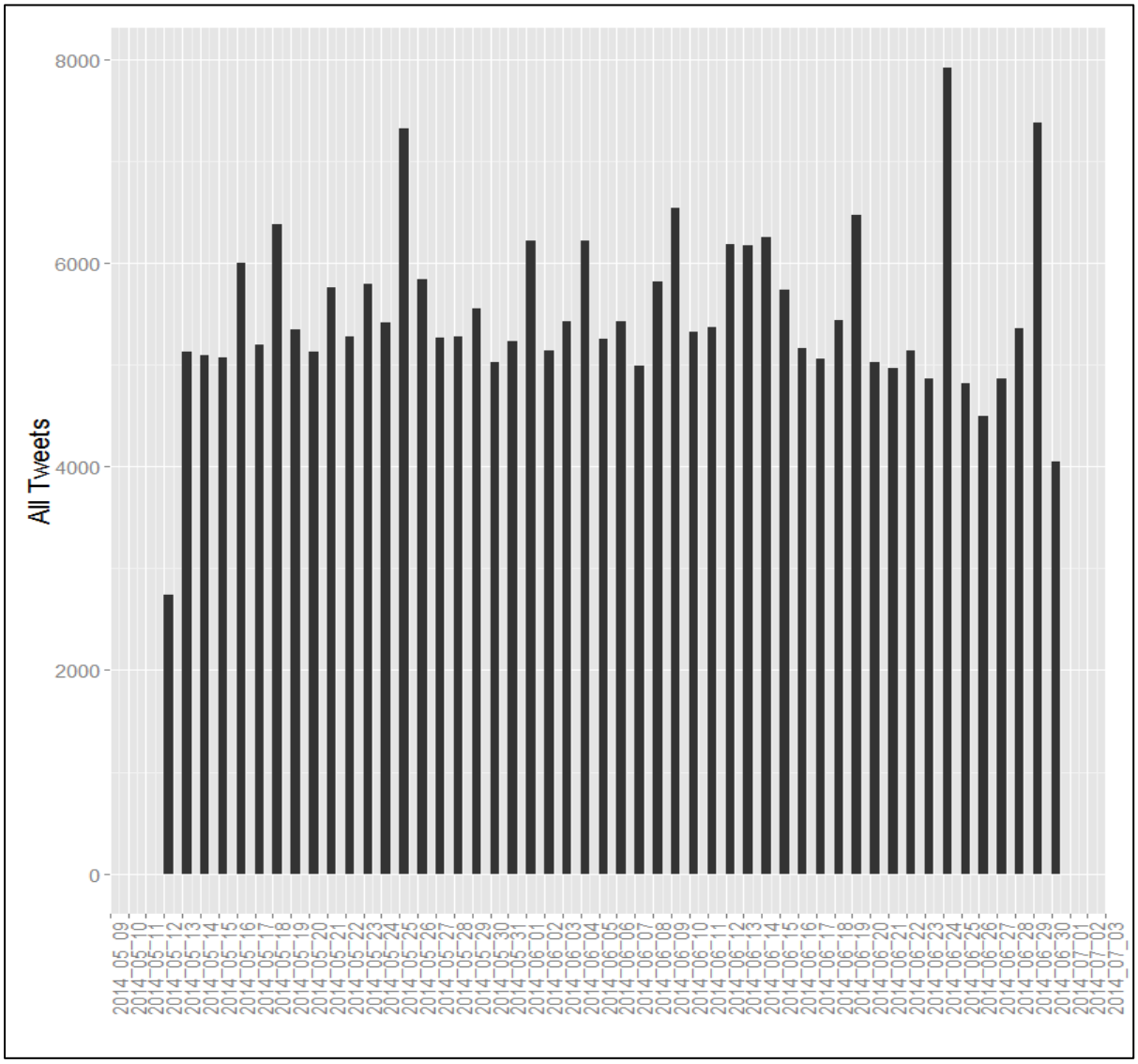
Εικόνα 4.16: Συνολικές αναρτήσεις(tweets)ανά μέρα από δεδομένα της Αττικής

Τα περισσότερα tweets έγιναν στις 29 Ιουνίου φτάνοντας στα 20000, με δεύτερη σειρά να καταφτάνει στις 24 Ιουνίου με 1980 tweets και τα λιγότερα να μαζεύονται στις 12 Μαΐου με 5030 tweets. Επίσης παρατηρούμε την μέρα των δημοτικών εκλογών και τον Ευρωεκλογών(18 και 25 Μαΐου) οι αναρτήσεις δεν ξεπερνούν τα 12500 tweets. Μπορούμε ακόμη να προσθέσουμε

πως μια μέρα πριν τις εκλογές και μια μέρα μετά τις εκλογές οι αναρτήσεις δεν ξεπερνούν τα 12450 tweets.

Τέταρτο στάδιο- Επιλογή Ελληνικών αναρτήσεων

Στο πιο κάτω ιστόγραμμα(εικόνα 4.17)απεικονίζονται οι αναρτήσεις που είναι μόνο γραμμένα στην ελληνική γλώσσα. Στον άξονα Χ είναι ανά μέρα(12 Μαΐου- 30 Ιουνίου)και ο άξονας Υ είναι το σύνολο των αναρτήσεων ανά μέρα όσα είναι γραμμένα στα ελληνικά.



Εικόνα 4.17: Όσες αναρτήσεις είναι γραμμένα στα ελληνικά από την συλλογή στην Αττική

Είναι ελάχιστές οι μέρες που ξεπερνούν τα 6000 tweets. Τα περισσότερα tweets έγιναν στις 24 Ιουνίου με σχεδόν 8000 tweets και τα λιγότερα στις 12 Μαΐου(όπου αναφέραμε πιο πάνω τον λόγο). Επίσης κατά την πρώτη Κυριακή των δημοτικών εκλογών(18 Μαΐου) οι αναρτήσεις αντιστοιχούν περίπου σε 6150 ενώ στην δεύτερη Κυριακή των εκλογών που είναι και οι Ευρωεκλογές μαζί(25 Μαΐου) οι αναρτήσεις πήγαν στα 7250. Πράγματι, υπάρχουν πολλές συζητήσεις στις 25 Μαΐου για τα πολιτικά θέματα. Οι χρήστες θα εκφράζουν τις γνώμες τους και την αγωνία που θα έχουν για τα αποτελέσματα των εκλογών.

Ελάχιστη διαφορά συνολικών αναρτήσεων έχουν στις 29 Ιουνίου και με τις 25 Μαΐου με περίπου 7250 tweets. Τα γεγονότα που συνέβησαν στις 29 Ιουνίου είναι ότι υπήρξε διάλογος με τν Κεντροδεξιά από υπουργούς, βουλευτές της ΝΔ και βρισκόταν εν όψει σοβαρών αποφάσεων, η εθνική Ελλάδος αγωνιζόταν την Κώστα Ρίκα κ.α. Έτσι οι χρήστες είχαν να συζητούν αυτά τα θέματα. Ωστόσο, στις 17 Μαΐου μια μέρα πριν τις εκλογές οι αναρτήσεις ήταν πιο χαμηλές από την μέρα των εκλογών. Επίσης το ίδιο σημειώθηκε και στις 24 Μαΐου που δεν ξεπερνά τα 5750 tweets.

Πέμπτο στάδιο- Πολιτικά ονόματα

Έχουμε 274835 tweets που είναι γραμμένα στα ελληνικά. Σε αυτό το στάδιο όπου είναι και ο σκοπός της εργασίας , να μελετήσουμε τις αναρτήσεις αν αναφέρονται σε πολιτικά άτομα και συζητήσεις όπου και θα αναλυθούν στο επόμενα κεφάλαια(κεφάλαιο 5,6).

Προσθέσαμε κάποιο κριτήριο για τη αναζήτηση αυτών των πολιτικών θεμάτων. Η αναζήτηση έγινε με πολιτικά πρόσωπα από την περιφέρεια Αττική των τριών δημοφιλέστερων κομμάτων(ΝΔ-ΠΑΣΟΚ-ΣΥΡΙΖΑ) όπως βουλευτές, ευρωβουλευτές, υποψήφιους δημάρχους, δημοτικούς συμβούλους. Επικεντρώνοντας στις δημοτικές εκλογές και ευρωεκλογές που έγιναν το Μάιο. Δημιουργήσαμε ένα dataframe στην R για να κάνουμε αναζήτηση των επιθέτων(surname) των πολιτικών αλλά και αν έχουν λογαριασμό στοTwitter, ψάχνοντας πόσες φορές υπάρχουν μέσα από τις αναρτήσεις που είναι γραμμένα στα ελληνικά. Η επιλογή των πολιτικών έγινε όπως αναφέρεται στα πιο κάτω παραδείγματα.

Για παράδειγμα ο πολιτικός κ. Γλέζος στην αναζήτηση έγινε ως εξής :

Γλέζος, Γλεζος, Γλέζο, Γλεζο = εμφανίζεται 83 φορές και μηδέν(0) που δεν έχει λογαριασμό στο Twitter, σύνολο δηλαδή 83 φορές στις αναρτήσεις.

Ενώ ο κ. Γεωργιάδης στην αναζήτηση έγινε ως εξής:

Γεωργιάδης, Γεωργιαδης, Γεωργιάδη, Γεωργιαδης, @AdonisGeorgiadi . Εντοπίστηκε 917 φορές μέσα από τις αναρτήσεις.

Δηλαδή η αναζήτηση έγινε με το επίθετο του πολιτικού να είναι τονισμένο και όχι , και με το θέμα του επιθέτου(surname). Παίρνουμε όλες τις πτυχές του επιθέτου του κάθε πολιτικού γιατί μπορεί κάποιοι χρήστες να απευθύνονται στο τρίτο πρόσωπο(πχ ο Σαμαράς, του Σαμαρά, Σαμαρά). Στην περίπτωση που κάποιοι πολιτικοί δεν έχουν λογαριασμό στο Twitter θεωρείται μηδέν και μετρούμε τις άλλες εμφανίσεις που έχει το επίθετο. Επίσης όταν ένα πολιτικό πρόσωπο δεν εμφανίζεται καθόλου στις αναρτήσεις παίρνουμε άλλο άτομο(πχ ο κ. Τέντομας κανένας χρήστης δεν τον αναφέρει). Στον πιο κάτω πίνακα(πίνακα 4.6) εμφανίζονται τα πολιτικά ονόματα που επιλέξαμε με βάση του κριτήριο που αναφέραμε πιο πάνω, έχοντας 15 πολιτικά πρόσωπα για κάθε κόμμα περιλαμβάνοντας τον λογαριασμό του Twitter όσοι έχουν.

Πολιτικά Ονόματα στην Αττική					
ΝΕΑ ΔΗΜΟ-ΚΡΑ-ΤΡΙΑ	Σαμαράς Αντώνης @PrimeministerGR	Άδωνις Γεωργιάδης @AdonisGeorgiadi	Σπηλιωτόπουλου Άρης @aris_spiliotop	Βορίδης Μάκης @MakisVoridis	Κικίλιας Βασίλης @Vkikilias
	Σπυράκη Μαρία @MariaSpyraki	Κεφαλογίαννη Όλγα @Okefalogianni	Ζαγοράκης Θοδωρής @zagorakis_euro	Πλεύρης Θάνος @thanosplevris	Κύρτσος Γιώργος @GiorgosKyrtsos
	Κεφαλλογιάννης Μανώλης @MKefaloyannis	Ανέστης Γιώργος	Παναγιωτοπού-λου Πηνελόπη @penelope_nd	Βοζενμπεργκ Ελίζα @vozemberg	Αμυράς Γιώργος
ΣΥΡΙΖΑ	Δούρου Ρένα @renadourou	Αλέξης Τσίπρας @atsipras	Σπανού Δέσποινα	Σακελλαρίδη Γαβριήλ @gabriel_athens	Παπαδημούλης Δημήτρης @papadimoulis
	Γλέζος Μανώλης	Κούνεβα Κωνσταντία	Δούκα Μαρία	Κατρούγκαλος Γιώργος @katrougkalos	Μηλιός Γίαννης @john_milios
	Σακοράφα Σοφία @SofiaSakorafa	Χουντής Νίκος	Χρυσόγονος Κώστας @Chrysogonos_K	Φωτίου Θεανώ @TheanoFotiou	Καρύδης Κωνσταντίνος @ntinoskaridis
ΠΑΣΟΚ	Βενιζέλος Ευάγγελος @EVenizelos	Καμίνης Γιώργος @KaminisG	Πάγκαλος Θεόδωρος @tpangalos	Καιλή Εύα	Κακλαμάνη Νικήτας @nkaklamanis
	Βασιλικός Βασίλειος	Ράπτη Συλβάνα @sylvanarapti	Δασκαλάκη Μελπομένη @MelinaDas	Ανδρουλάκης Νίκος	Παπαχέλα Νέλλη-Κανέλλα @nellyopanda
	Βρεττός Σπύρος	Αγανίδης Πασχάλης	Καφετζόπουλος Αντώνιος @Kafetzopoulos	Χρυσοχοΐδης Μιχάλης @chrisochoidis	Λοβέρδος Ανδρέας @a_loverdos

Πίνακας 4.6 : Οι πολιτικοί της Αττικής με βάση τα 3 δημοφιλέστερα κόμματα

Γενικές παρατηρήσεις από δεδομένα της περιφέρειας Αττικής

Κατά την διάρκεια συλλογής δεδομένων την χρονική περίοδο του ενάμιση μήνα, πριν και μετά τις δημοτικές εκλογές και τις Ευρωεκλογές, έχουμε τα πιο κάτω αποτελέσματα:

Αρκετοί Έλληνες χρησιμοποιούν το Twitter για συζητήσεις πολιτικών θεμάτων γράφοντας στα ελληνικά εκφράζοντας την άποψης τους, τον προβληματισμό που τους απασχολεί για τις τρέχουσες εκλογές, αναφέροντας και σε πολιτικά πρόσωπα.

Κατά την ημέρα των εκλογών(18 Μαΐου και 25 Μαΐου) υπάρχουν αρκετές συζητήσεις αλλά στις 25 ξεπερνούν τις αναρτήσεις από τις 18 Μαΐου γιατί είναι και οι Ευρωεκλογές. Μια μέρα πριν τις εκλογές δεν υπάρχουν πολλές αναρτήσεις.

Συγκρίνοντας με τα δεδομένα που είχαμε στην Κύπρο με αυτά της περιφέρειας Αττικής υπάρχει περισσότερη χρήση του Twitter για πολιτικές συζητήσεις κατά την προεκλογική περίοδο. Ενώ στην Κύπρο συζητούν περισσότερο για τους προσωπικούς τους προβληματισμούς και γενικά τι συμβαίνει στην προσωπική τους ζωή.

Η συλλογή δεδομένων στην Κύπρο ήταν σε 75 μέρες και είχαμε 287886 tweets με καθόλου πολιτικές συζητήσεις ενώ στην συλλογή δεδομένων στην Αττική ήταν σε 50 μέρες έχοντας 664714 tweets και από αυτά τα 274835 είναι γραμμένα στα ελληνικά και να έχουν πολιτικά θέματα.

Κεφάλαιο 5ο
Ανάλυση Δεδομένων

«Οι γνώμες τελικά διαμορφώνονται από τα αισθήματα και όχι από το μυαλό.» Herbert Spencer

Η επιλογή των Ελληνικών προσώπων έγινε στο προηγούμενο κεφάλαιο(κεφάλαιο 4 , ενότητα 4.4) επιλέγοντας από τα τρία δημοφιλέστερα κόμματα, 15 πολιτικούς ανά κόμμα. Σε αυτό το κεφάλαιο θα περιγράψουμε πως θα κατασκευαστεί ο αλγόριθμος για να αναλύσουμε τις ελληνικές αναρτήσεις που επιλέξαμε.

5.1 Γραφικές Πολιτικών Κομμάτων

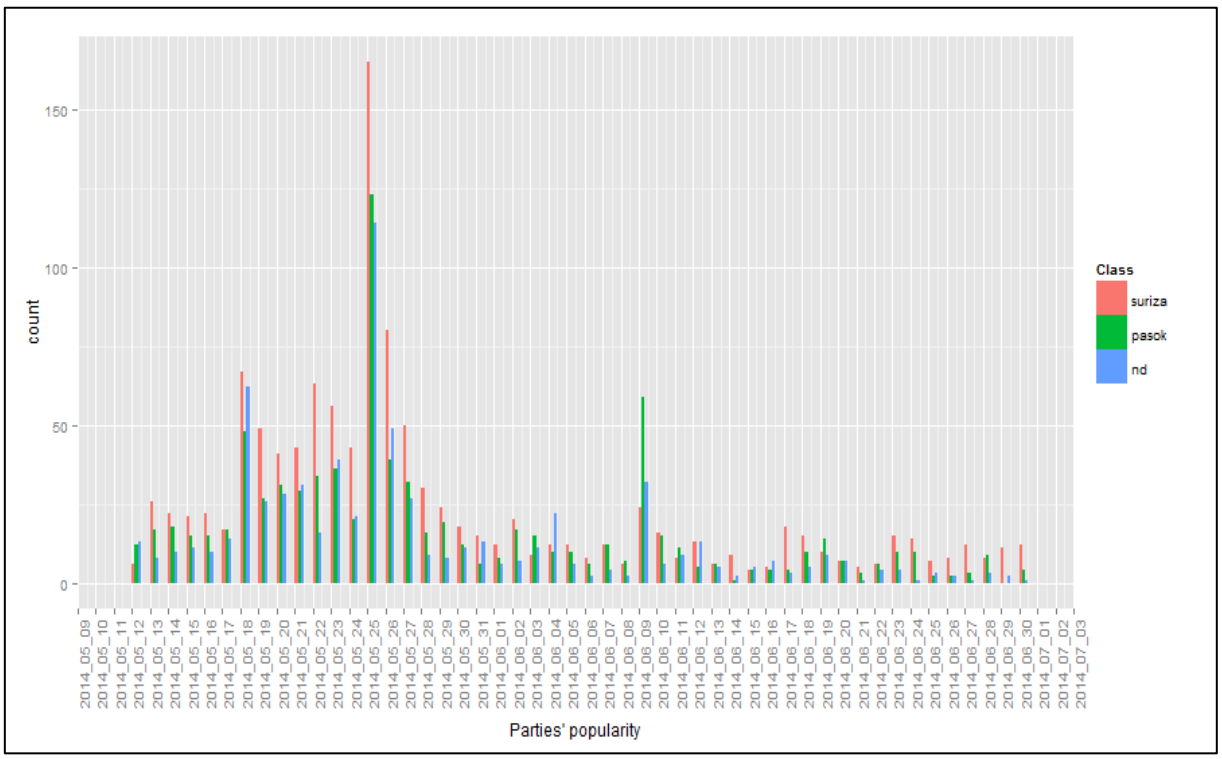
Στις πιο κάτω γραφικές παρουσιάζουμε τα τρία δημοφιλέστερα κόμματα με βάση την αναζήτηση που κάναμε για τα πολιτικά ονόματα(πίνακας 4.6) που επιλέξαμε αλλά και μια νέα αναζήτηση για τα ονόματα των κομμάτων.Το κάθε κόμμα απεικονίζετε με το αντιπροσωπευτικό χρώμα που έχουν(ΣΥΡΙΖΑ= κόκκινο, ΠΑΣΟΚ= πράσινο, ΝΕΑ ΔΗΜΟΚΡΑΤΙΑ= μπλε). Όσοι πολιτικοί είχαν δικά τους tweets τα διαγράψαμε γιατί μας ενδιαφέρουν μόνο τα

tweets των χρηστών. Εντοπίσαμε 24 tweets εκ τον οποίον τα 22 tweets ήταν του κ. Βενιζέλου και τα άλλα 2 tweets του κ.Σακελλαρίδη. Έτσι, συνολικά είναι 664688 tweets γραμμένα στα ελληνικά από τα 664712 που είχαμε αρχικά.

Αναζήτηση με το όνομα του κόμματος

Η αναζήτηση έγινε με το όνομα του κόμματος (ΝΔ,ΠΑΣΟΚ, ΣΥΡΙΖΑ) για να εντοπίσουμε πόσες φόρες οι χρήστες αναφέρουν τα πολιτικά ονόματα μέσα στις αναρτήσεις τους. Από την αναζήτηση βρήκαμε 2722 tweets και για τα τρία κόμματα. (ΝΔ = 705, ΠΑΣΟΚ= 835, ΣΥΡΙΖΑ=1182).

Στην γραφική(εικόνα 5.1) που απεικονίζονται πιο κάτω είναι πόσο συχνά χρησιμοποιούν τα ονόματα των κομμάτων οι χρήστες μέσα στις ελληνικές αναρτήσεις που μαζέψαμε. Στον άξονα Χ είναι οι μέρες της συλλογής δεδομένων 12 Μαΐου μέχρι 30 Ιουνίου και ο άξονας Υ είναι το συνολικό μέγεθος των μηνυμάτων.



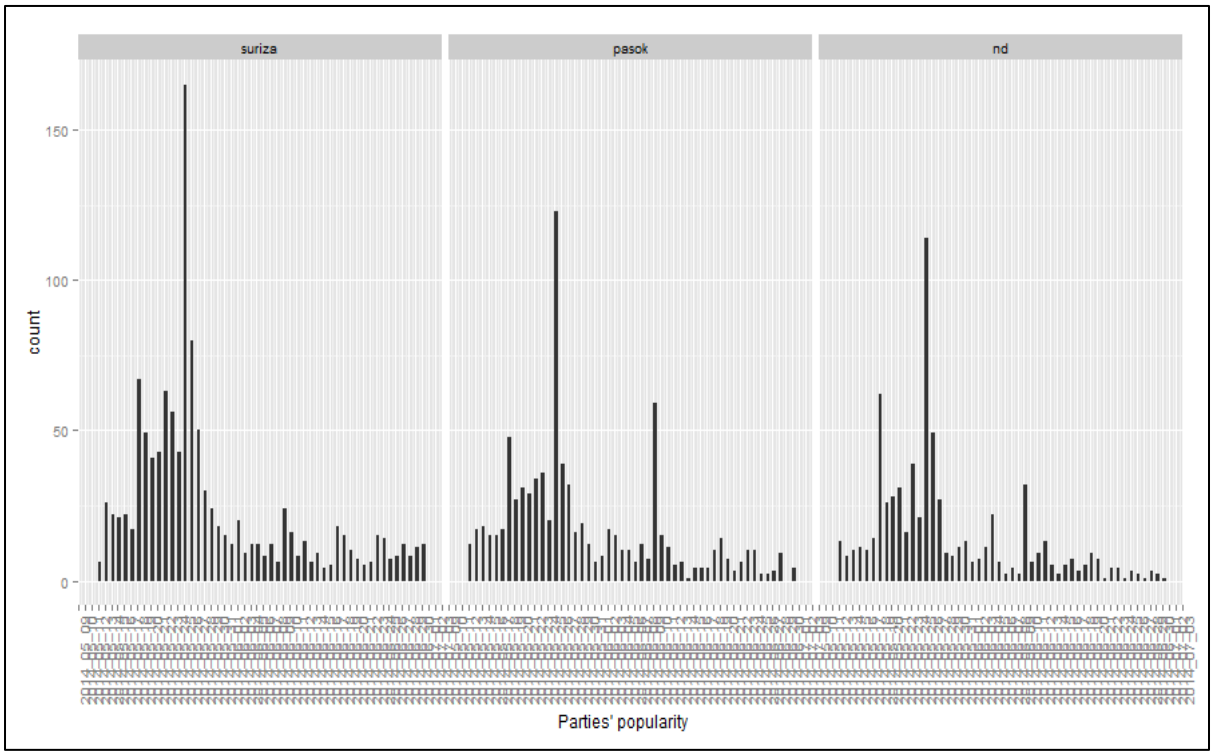
Εικόνα 5.1: Εμφάνιση των 3ων δημοφιλέστερων πολιτικών κομμάτων στην Αττική με βάση το πολιτικό κόμμα

Παρατηρούμε στις 25 Μαΐου το πολιτικό κόμμα Σύριζα έχει το μεγαλύτερο σύνολο αναφορών του κόμματος από όλες τις μέρες, που ξεπερνά τις 150 φορές εμφανίσεων ενώ την ίδια μέρα το κόμμα Πασοκ οι χρήστες το αναφέρουν 123 φορές περίπου και με την Νέα Δημοκρατία λίγο πιο

κάτω. Επίσης να αναφέρουμε στις 25 Μαΐου είναι η μέρα τον Ευρωεκλογών και η τελευταία Κυριακή των δημοτικών εκλογών. Έτσι βλέπουμε πως υπάρχουν έντονες συζητήσεις για την αναφορά του κόμματος Σύριζα γιατί κερδίζει για πρώτη φορά στην ιστορία τη νίκη στις Ευρωεκλογές ,κερδίζοντας και την περιφέρεια την Αττικής.

Στις 18 Μαΐου ήταν η πρώτη Κυριακή των δημοτικών εκλογών και υπάρχουν αρκετές αναφορές στα ονόματα των κομμάτων από τους χρήστες όπου ξεπερνά το πολιτικό κόμμα Σύριζα ελάχιστα από την Νέα Δημοκρατία και το Πασοκ φτάνοντας στις 50 αναφορές περίπου. Ακόμη παρατηρούμε κατά την διάρκεια των ημερών 18 Μαΐου μέχρι 27 Μαΐου που είναι η περίοδος των εκλογών οι χρήστες αναφέρουν στις αναρτήσεις του τα τρία δημοφιλέστερα πολιτικά ονόματα των κομμάτων μέσα στις συζητήσεις τους. Εκφράζοντας τις ανησυχίες τους, τις γνώμες τους και τους προβληματισμούς για τα κόμματα αλλά και τι πιστεύουν για τις τρέχουσες εκλογές.

Για τις ημερομηνίες 17 Μαΐου , 7,13,15 και 22 Ιουνίου τα κόμματα Σύριζα και Πασοκ έχουν τα ίδια ποσοστά μηνυμάτων αναφοράς, ενώ στις 20 Ιουνίου και τα τρία κόμματα βρίσκονται στην ίδια θέση περίπου. Ακόμη μια σημαντική μέρα είναι στις 9 Ιουνίου όπου το πολιτικό θέμα είναι που ο πρωθυπουργός της χώρας κ. Σαμαράς προχωρά σε ανασχηματισμό της κυβέρνησης του για 2η φόρα μετά από τις 24/6/2013. Χωρίς αμφιβολία ο ανασχηματισμός είχε εκπλήξεις έχοντας το Σύριζα να σχολιάζει πως η κυβέρνηση δεν έλαβε το μήνυμα από τις τρέχουσες εκλογές που είχαν. Έτσι οι χρήστες του Twitter σχολιάζουν τις απόψεις τους έχοντας με διαφορά το Πασοκ να πάνω από τις 50 φορές εμφανίσεις μέσα από τις αναρτήσεις. Τέλος στις 29 Ιουνίου κανένας χρήστης από την περιφέρεια Αττικής δεν αναφέρθηκε στο πολιτικό κόμμα του Πασοκ ενώ για την ΝΔ υπάρχουν λίγες εμφανίσεις μηνυμάτων. Στην πιο κάτω γραφική(εικόνα 5.2) παρουσιάζουμε τα αποτελέσματα που έχουμε πιο πάνω σε τρείς ξεχωριστές γραφικές για το κάθε κόμμα μόνο του.



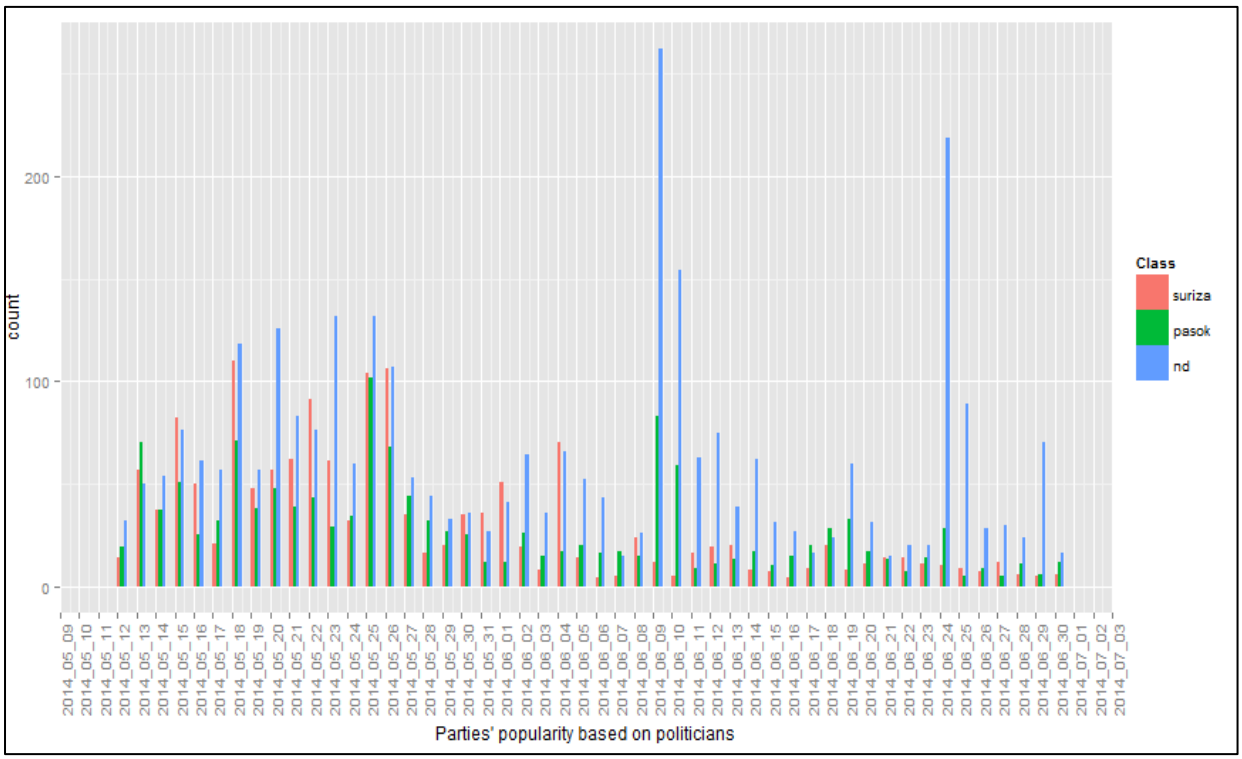
Εικόνα 5.2: Εμφάνιση των 3ων δημοφιλέστερων πολιτικών κομμάτων στην Αττική, σε ξεχωριστές γραφικές.

Και στις 3 γραφικές υπάρχουν αρκετές αναφορές μηνυμάτων για τα κόμματα πριν και μετά τις εκλογές με περισσότερο σύνολο μηνυμάτων στο Σύριζα την μέρα των Ευρωεκλογών, Ακόμη, συγκρίνοντας και τις γραφικές ξεχωριστά παρατηρούμε το Σύριζα γίνονται περισσότερες αναφορές για το κόμμα ενώ στα άλλα 3 κόμματα πιο λίγες αναφορές. Από τις 21/6 -30/6 στην γραφική για το πολιτικό κόμμα της ΝΔ οι αναφορές προς το όνομα του κόμματος είναι πιο λίγες σε σύγκριση με τις μέρες των εκλογών και την μέρα του ανασχηματισμού της κυβέρνησης.

Αναζήτση με τα ονόματα των πολιτικών προσώπων

Η επιλογή των πολιτκών προσώπων έγινε πιο πανω αλλα θέλουμε να εντοπίσουμε πόσες φορές αναφέρονται στα πολιτικά πρόσωπα των κομμάτων οι χρηστές. Εντοπίσαμε 6044 tweets από αυτά τα 3132 tweets ήταν για την ΝΔ, τα 1410 για το ΠΑΣΟΚ και τα 1502 για το ΣΥΡΙΖΑ.

Στην πιο κάτω γραφική παράσταση(εικόνα 5.3)παρουσιάζουμε τα 15 πολιτικά πρόσωπα(πίνακας 4.6) αντιστοιχώντας τα στο κόμμα τους. Δηλαδή πόσες φορές την μέρα αναφέρονται οι χρήστες στα πολιτικά ονόματα(σε σύνολο κόμματος).



Εικόνα 5.3: Εμφάνιση πολιτικών προσώπων μέσα από τα τρία δημοφιλέστερα κόμματα στην Αττική

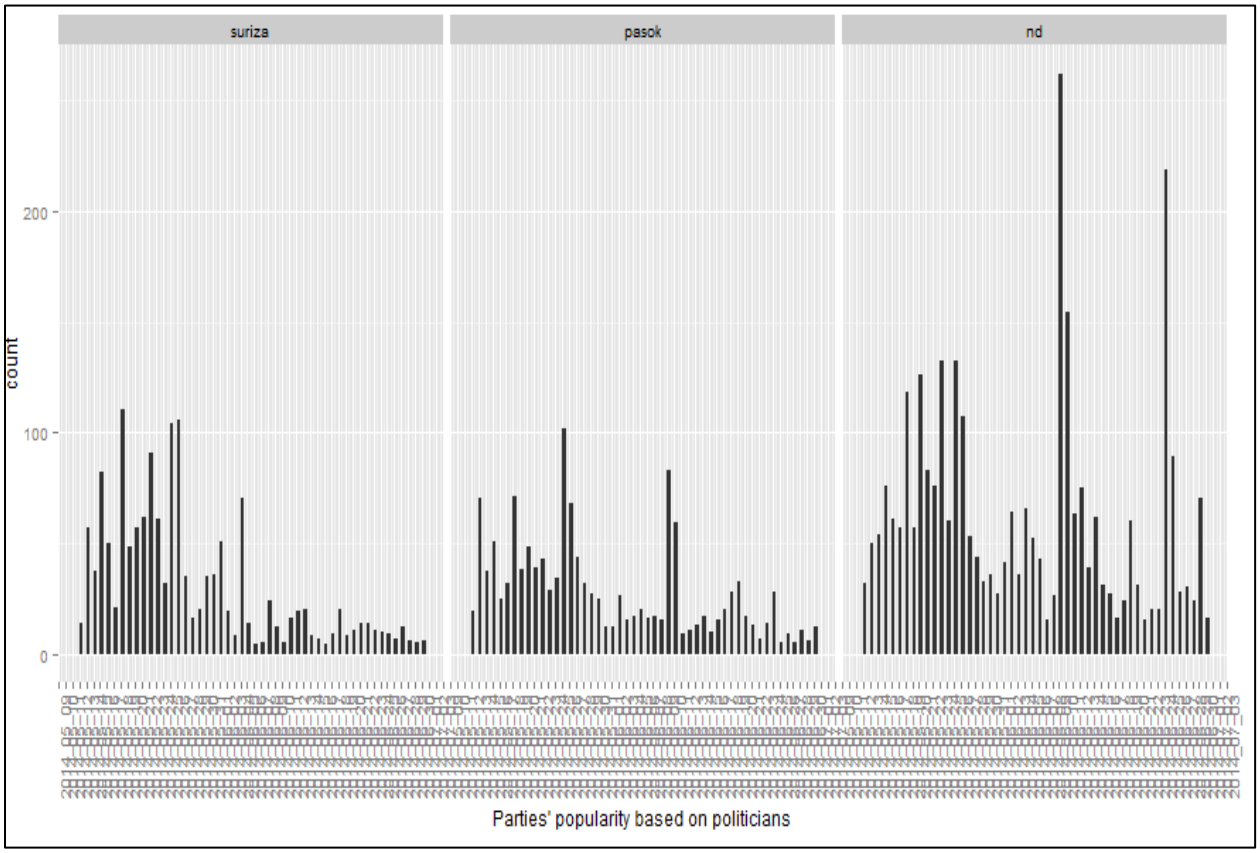
Το μεγαλύτερο ποσοστό μηνυμάτων ,πολιτικών προσώπων να βρίσκεται στην Νέα Δημοκρατία να ξεπερνά τις 250 εμφανίσεις μέσα από τις ελληνικές αναρτήσεις και απεικονίζονται στις 9 Ιουνίου όπου ανακοινώνεται το νέο κυβερνητικό σχήμα από την νέα κυβερνητική εκπρόσωπο. Την ίδια μέρα το Πασοκ δεν ξεπερνά τις 100 εμφανίσεις των πολιτικών ονομάτων ενώ το Σύριζα βρίσκεται πολύ χαμηλά,

Κατά δεύτερη σειρά συνολικών μηνυμάτων βρίσκεται πάλι η ΝΔ στις 24 Ιουνίου και το πολιτικό θέμα που επικρατεί είναι η έντονη ανησυχία στις τάξεις των πιστωτών για τις μεγάλες καθυστερήσεις που παρουσιάζει η υλοποίηση του οικονομικού προγράμματος ενώ την ίδια στιγμή βασικό θέμα συζήτησης στις πρεσβείες των μεγάλων χωρών στην Αττική-Ελλάδα είναι αν η χώρα θα οδηγηθεί σε πρόωρες εκλογές το φθινόπωρο. Την ίδια στιγμή ο Σαμαράς και Βενιζέλος είναι αποφασισμένοι να αντιστρέψουν τι κλίμα ανησυχίας που υπάρχει στο κύκλο των πιστωτών οι οποίοι μιλάνε για την 'πολιτική παράλυση' και να προχωρήσουν σε ταχύτατους ρυθμούς στην προώθηση των προαπαιτούμενων αλλά και στο κλείσιμο όλων των εκκρεμοτήτων. Η ΝΔ να ξεπερνά με μεγάλη διαφορά στις αναφορές των πολιτικών προσώπων στο Twitter σε σύγκριση με Πασοκ και Σύριζα που δεν φτάνουν ούτε μέχρι τιε 50 φορές εμφανίσεις του κόμματος.

Ακόμη στις 10 Ιουνίου το πολιτικό θέμα που έγινε είναι η ορκωμοσία της νέας κυβέρνησης και οι δηλώσεις του πρωθυπουργού Σαμαρά για μελλοντικά σχέδια κλπ. Οι αναρτήσεις που κάνουν τα πολιτικά πρόσωπα αλλά και οι χρήστες που αναφέρονται σα πολιτικά πρόσωπα της ΝΔ είναι περίπου 152 το συνολικό μέγεθος των μηνυμάτων ενώ το Σύριζα δεν αναφέρονται ούτε 25 φορές μέσα στις αναρτήσεις.

Κατά την περίοδο των εκλογών 18 Μαΐου και 26 Μαΐου και τα τρία δημοφιλέστερα κόμματα δεν ξεπερνούν τις 125 συνολικές αναφορές των πολιτικών ατόμων ανά κόμμα. Ακόμη και εδώ η ΝΔ έχει τα περισσότερα μηνύματα ενώ τα πολιτικά πρόσωπα του Σύριζα να βρίσκεται κατά δεύτερη σειρά των προτιμήσεων των συζητήσεων τους. Οι έντονες ανησυχίες των χρηστών τι θα ψηφίσουν στις εκλογές αλλά σχολιάζοντας τις διάφορες δηλώσεις των πολιτικών προσώπων.

Έχοντας σχολιάσει την πιο πάνω γραφική, στην εικόνα 5.4 βρίσκονται τα ίδια αποτελέσματα πιο πάνω αλλά ξεχωρίζοντας το κάθε κόμμα σε διαφορετική γραφική παράσταση,



Εικόνα 5.4: Εμφάνιση πολιτικών προσώπων μέσα από τα τρία δημοφιλέστερα κόμματα στην Αττική, σε τρείς ξεχωριστές γραφικές ανά κόμμα.

Το πολιτικό κόμμα Σύριζα κατά την περίοδο των δημοτικών εκλογών και των Ευρωεκλογών γίνονται αρκετές συζητήσεις στα πολιτικά πρόσωπα για της δηλώσεις που κάνουν οι χρήστες να τα σχολιάζουν ενώ μετά τις εκλογές το συνολικό μέγεθος των μηνυμάτων λιγοστεύουν μέχρι τέλος Ιουνίου. Από την άλλη πλευρά το κόμμα του Πασοκ την ίδια περίοδο των εκλογών βρίσκονται περίπου στα ίδια επίπεδα με το Σύριζα αλλά τις επόμενες μέρες μέχρι του Ιουνίου υπάρχουν αρκετές αναφορές των πολιτικών προσώπων του κόμματος. Ωστόσο, το κόμμα της ΝΔ κατά την διάρκεια της συνολικής περιόδου που έχουμε είναι πολύ ενεργό στις συζητήσεις/ αναφορές των πολιτικών προσώπων.

5.2 Αναλυτική Περιγραφή

Πριν την ανάπτυξη του ταξινομητή συναισθήματος πρέπει πρώτα να δημιουργηθούν δύο στάδια, ένα από λεξικό γνώμης(ενότητα 2.8) και βοηθητικές λέξεις για να προχωρήσουμε στο επόμενο στάδιο του αλγορίθμου. Αυτά τα δύο στάδια που θα αναπτυχθούν περιγράφουν λέξεις με θετικό και αρνητικό συναίσθημα που δημιουργείται αυτόματο ή ημι-αυτόματο λεξικό αλλά και κάποιες βοηθητικές λέξεις στον εντοπισμό του προσανατολισμού συναισθήματος της λέξης. Στις πιο κάτω υποενότητες περιγράφονται τα λεξικά που εντοπίσαμε και την βοηθητική λέξη.

5.2.1 Λεξικά Γνώμης

Η επιλογή του λεξικού έγινε από μια ποικιλία λεξικών βασισμένα στην πολικότητα ταξινόμησης και να μπορούμε να τα μεταφράσουμε τις λέξεις στην ελληνική γλώσσα. Αναζητώντας λεξικά που να έχουν θετικές και αρνητικές λέξεις. Πρώτα ψάξαμε λεξικά ταξινόμησης θετικών και αρνητικών λέξεων που είναι γραμμένα στα ελληνικά αλλά δεν εντοπίσαμε κανένα και έτσι ξεκινήσαμε την αναζήτηση σε αγγλικά λεξικά. Έχουμε εντοπίσει πέντε λεξικά βασισμένα στην πολικότητα ταξινόμησης που επικεντρώνονται στον προσανατολισμό των λέξεων γνώμης (opinion words) των θετικών και αρνητικών λέξεων όπου είναι τα παρακάτω :

The General Inquirer[20]

Αυτό το λεξικό γνώμης δημιουργήθηκε από τους Philips S Stone, Dexter CDunph, Marshal S. Smith και τον Daniel M. Ogilvie το 1996. Το General Inquirer είναι μια προσέγγιση υπολογιστών στην ανάλυση περιεχομένου και αποτελείται από τρεις κατηγορίες. Τις θετικές(1915 λέξεις) και

[20] http://www.wjh.harvard.edu/~inquirer/

αρνητικές(2291 λέξεις) λέξεις, των συνδυασμό ισχυρών έναντι αντίθετων λέξεων, τις ενεργούς έναντι τις παθητικές λέξεις. Άλλα και των συνδυασμό λέξεων που αναφέρονται για πόνο, ταλαιπωρία, λέξεις που περιγράφουν συγκεκριμένα συναισθήματα και που εκφράζουν εκτίμηση κλπ.

LIWC(Linguistic Inquiry and Word Count[21]) Γλωσσικά μηνύματα και λέξη ποσοστού

Είναι ένα πρόγραμμα λογισμικό ανάλυσης σχεδιασμένο από τους James W. Pennebaker, Roger j. Baoth και την Martha E. Francis. Το LIWC μπορεί να αναλύσει εκατοντάδες τυποποιημένα αρχεία κειμένων ASCII ή έγγραφα της Microsoft Word σε δευτερόλεπτα. Έχει 2300 λέξεις όπου κατηγοριοποιούνται σε 70 κατηγορίες. Αναφέρονται σε συναισθηματικές διεργασίες θετικό και αρνητικό συναίσθημα(αγάπη, όμορφος, μίσος, σκληρό) και στις γνωστικές διεργασίες όπως ίσως, υποθέτω, περιορισμός, στις αντωνυμίες και αρνήσεις.

MPQA(Multi- Perspective Question Answering)[22]

Το λεξικό αυτό δημιουργήθηκε από τους Theresa Wilson, Janyce και τον Paul Hoffmann το 2005 εντάσσοντας στο επίπεδο φράσης στην ανάλυση κειμένου. Περιλαμβάνει 6885 λέξεις από τις 8221 που είναι λήμματα. Έχει 2718 θετικές λέξεις και 4912 αρνητικές λέξεις όπου κάθε λέξη είναι σχολιασμός για τις εντάσεις, δηλαδή ισχυρές και ασθενείς λέξεις.

WordNet [23]

Το WordNet είναι ένα λεξικό με μια βάση δεδομένων από το αγγλικό λεξικό έχοντας ομάδες από ουσιαστικά, ρήματα, επίθετα και επιρρήματα. Η δομή του λεξικού αυτού είναι ένα χρήσιμο εργαλείο για την υπολογιστική γλωσσολογία και την επεξεργασία γλώσσας. Είναι ελεύθερο και δωρεάν για στην εγκατάσταση.

SentiWordNet [24]

Το SentiWordNet είναι ένα λεξικό για την εξόρυξη γνώμης και αντίστοιχη με το WordNet που έχουν τις τρείς βαθμολογήσεις την θετικότητα, αρνητικότητα και την αντικειμενικότητα. Η βαθμολόγηση γίνεται αυτόματα.

[21] http://www.liwc.net/
[22] http://mpqa.cs.pitt.edu/ ,
[23] http://wordnet.princeton.edu/
[24] http://sentiwordnet.isti.cnr.it/

Bing Liu Opinion Lexicon[25]

Είναι ένα λεξικό του Bing Liu και του Minqing Hu που δημιουργήθηκε τον Μάιο του 2004. Επίσης είναι συγγραφείς και του βιβλίου ' Sentiment Analysis and Opinion Mining ' . Το λεξικό περιλαμβάνει περίπου 6800 λέξεις με 2006 θετικές και 4783 αρνητικές λέξεις.

Από τα έξι λεξικά γνώμης που αναφέραμε πιο πάνω, επιλέξαμε το λεξικό '**Bing Liu Opinion Lexicon**'. Από τον σύνδεσμο(link) του λεξικού υπάρχουν δύο αρχεία .txt με τις θετικές και αρνητικές λέξεις. Αυτές οι λέξεις συναισθήματος είναι γραμμένα στην Αγγλική γλώσσα και τις μετατρέψαμε με το free Language Translator 3.4 [26] στην Ελληνική γλώσσα. Για όσες λέξεις στην μετάφραση ήταν φράσεις ή επαναλαμβάνεται η λέξη την διαγράψαμε γιατί το λεξικό μας θα διαβάζει μόνο λέξεις. Κάποιες ενδεικτικές θετικές και αρνητικές λέξεις βρίσκονται στον πιο κάτω πίνακα(πίνακας 5.1).

ΘΕΤΙΚΕΣ ΛΕΞΕΙΣ	**ΑΡΝΗΤΙΚΕΣ ΛΕΞΕΙΣ**
αυθεντικός	κακοποίηση
βραβείο	αγωνία
δημιουργικός	αδιέξοδο
εγκρίθηκε	ακυρώσει
ενεργητικός	δυσκολίες
ευχαριστημένος	εκβιασμός
θριαμβευτικά	λογομαχία
ικανοποιημένοι	παγιδευμένοι
ήρωες	χαμός

Πίνακας 5.1: Ένα δείγμα από τις θετικές και αρνητικές λέξεις από το λεξικό του Bing Liu

[25] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html
[26] http://languagetranslator.codeplex.com/releases/view/113319

5.2.2 Βοηθητικές Λέξεις

Κοιτάζοντας τις αναρτήσεις παρατηρούμε πως υπάρχει συχνά η λέξη 'δεν' που μπορεί να αλλάζει τη βαθμολογία του συναισθήματος από θετική σε αρνητική ή/και το αντίθετο όταν βρίσκεται στην 2 ή 3 θέση πριν την λέξη γνώμης. Για παράδειγμα η λέξη '*καταπληκτικό* ' είναι θετικό συναίσθημα, εάν όμως έχει μπροστά ' *δεν είναι καταπληκτικό* ' τότε από θετικό συναίσθημα γίνεται αρνητικό.

5.3 Περιγραφή Λειτουργίας Του Αλγορίθμου

Η ανάπτυξη του αλγορίθμου έγινε σε μια σειρά από τα εξής βήματα:

- Μετάφραση λεξικού στην ελληνική γλώσσα (ενότητα 5.2) σε θετικές και αρνητικές λέξεις με τονισμένες λέξεις και με χωρίς τόνους γιατί μπορεί οι χρήστες όταν αναρτούν να μην τονίζουν τις λέξεις τους.

- Έχοντας σε ένα ενιαίο dataframe για το κάθε κόμμα ξεχωριστά, όπου αυτά περιλαμβάνουν tweets με τα πολιτικά ονόματα (πίνακας 4.6) του κάθε κόμματος και το όνομα του κόμματος τους. Όσοι πολιτικοί έχουν αναρτήσει tweets τα διαγράφουμε γιατί επικεντρωνόμαστε μόνο για τις γνώμες των χρηστών που αναφέρονται σε πολιτικά πρόσωπα και συζητήσεις.

- Γίνεται αναζήτηση σε ποια θέση βρίσκεται η λέξη «δεν» γιατί επηρεάζει το συναίσθημα της λέξης γνώμης που βρίσκεται πριν από αυτό.

Ο αλγόριθμος διαβάζει όλες τις θετικές και αρνητικές λέξεις και τις προσθέτει σε ένα dataframe ξεχωριστά όπου ξεκινά να διαβάζει ένα προς ένα τα tweets κοιτάζοντας τις λέξεις μια προς μια κάνοντας έλεγχο που βρίσκεται η λέξη 'δεν' για να μπορεί να προσδιορίσει το συναίσθημα. Ο έλεγχος γίνεται στην $2^{η}$ ή $3^{η}$ θέση πριν από την λέξη γνώμη. Αυτό επαναλαμβάνεται τέσσερις φορές (θετικές τονισμένες λέξεις και μη, αρνητικές τονισμένες λέξεις και μη). Ο λόγος που δημιουργούμε λίστες με τις άτονες λέξεις είναι ότι στην R το κεφαλαίο και το μικρό γράμμα το διαβάζει ξεχωριστά, δεν το παίρνει να είναι ίδιο. Προσθέτει σε ένα πίνακα όλες τις θετικές λέξεις και σε ένα άλλο πίνακα τις αρνητικές λέξεις.

Σε κάθε tweet μετρά πόσα θετικά και αρνητικά υπάρχουν και μετά το κάθε tweet παίρνει το συναίσθημα με το μεγαλύτερο αριθμό(πχ 3 θετικές λέξεις και 2 αρνητικές λέξεις το αποτέλεσμα του tweet είναι θετικό συναίσθημα). Όσα tweets έχουν ίσο αριθμό θετικών και αρνητικών λέξεων το συναίσθημα ανήκει στην κατηγορία των ουδέτερων. Όλοι αυτή η διαδικασία επαναλαμβάνετε και για τια 3 κόμματα εμφανίζοντας στο τέλος σε μια γραφική παράσταση.

Παράδειγμα από κάποια tweets κάποιων χρηστών :

«Ο Πάγκαλος φοβήθηκε με τις απειλές κλπ, λέτε να κινδυνέψει το Πασοκ σε Ευρωεκλογές και Αττική και τρέμει……»

Πάγκαλος, Πασοκ = πολιτικό κόμμα, πολιτικός

Ευρωεκλογές, Αττική= σχετικά με τις πολιτικές συζητήσεις

Φοβήθηκε, απειλές, κινδυνέψει, τρέμει = αρνητικές λέξεις

Άρα το tweet παίρνει **αρνητικό συναίσθημα** για το ΠΑΣΟΚ

«Δικαιωματικά το πρώτο μεταλλιο απονέμεται στον Συριζα. Προσπάθησε αρκετά »

Συριζα= πολιτικό κόμμα

πρώτο, δικαιωματικά, προσπάθησε= θετικές λέξεις

Άρα το tweet παίρνει **θετικό συναίσθημα** για το ΣΥΡΙΖΑ

Κεφάλαιο 6ο
Εκτέλεση- Αξιολόγηση του αλγορίθμου

Φτάνοντας στην τελική φάση της διπλωματικής εργασίας και αφού έχουμε περιγράψει τα βήματα του αλγόριθμου στο προηγούμενο κεφάλαιο(κεφάλαιο 5). Σε αυτό το κεφάλαιο παρουσιάζουμε τα πειραματικά αποτελέσματα που περιγράφτηκαν στα προηγούμενα κεφάλαια αλλά και τις προβλέψεις που είχαν οι δημοσκοπήσεις για τις Εκλογές 2014.

6.1 Εισαγωγή

Από τη συλλογή των δεδομένων από την περιφέρεια Αττικής επιλέξαμε όσα :

-είναι γραμμένα στα ελληνικά

-επικεντρώνοντας στα τρία δημοφιλέστερα κόμματα(ΣΥΡΙΖΑ, ΝΔ, ΠΑΣΟΚ)

-αναφέρονται στο όνομα του κόμματος

- σε πολιτικά πρόσωπα(για την επιλογή των πολιτικών ατόμων έγινε με κριτήριο, περιγράψαμε στην ενότητα 4.4)

6. 2 Δημοσκοπήσεις vs Κοινωνικά μέσα

Οι δημοσκοπήσεις αποτελούνται από ερωτήσεις που επηρεάζουν την απόφαση της ψηφοφορίας περιλαμβάνοντας βασικές ερωτήσεις όπως φύλο, εθνικότητα, εκπαίδευση, ηλικία κλπ και ερωτήσεις αναλόγια με το θέμα τις δημοσκοπήσεις. Υπάρχει μια διαδικασία και μελέτη για τις ερωτήσεις που ακολουθούν για το θέμα, πρέπει να είναι σύντομες, στοχευόμενες και κατανοητές προς τις ηλικίες που αναφέρονται. Οι δημοσκοπήσεις γίνονται από μια ομάδα ατόμων και βάση στατιστικών παίρνουν την συνέντευξη από τους πολίτες. Τα αποτελέσματα δεν μπορείς να τα μάθεις από την ίδια χρονική στιγμή αλλά πρέπει να ολοκληρωθεί η διαδικασία της δημοσκοπήσεις από όλα τα άτομα που δουλεύουν σε όπιο σημείο και αν βρίσκονται. Μετά ξεκινά η καταγραφή των απαντήσεων από τα υπεύθυνα άτομα. Επομένως, οι δημοσκοπήσεις είναι χρονοβόρες προς τα αποτελέσματα, έχοντας μια δεδομένη στιγμή όταν πρέπει να ανακοινωθούν, Από την άλλη πλευρά, τα άτομα που κάνουν τις δημοσκοπήσεις έχουν την οπτική επαφή με τα τους πολίτες και μπορούν να αντιμετωπίσουν αρκετά προβλήματα γιατί κάποιοι πολίτες να μην θέλουν να λάβουν μέρος ή/και να ξεκινάνε να λένε τα προσωπικά τους προβλήματα, να προσπαθούν να στείλουν μηνύματα στους πολιτικούς μέσα από τα άτομα που δουλεύουν στις δημοσκοπήσεις. Γενικά πρέπει να μπορέσουν να ψυχολογήσουν τους πολίτες και να προσπαθήσουν να κερδίσουν τους πολίτες να λάβουν μέρος σε αυτή την δημοσκόπηση.

Από την αντίθετη πλευρά έχουμε τις απόψεις των πολιτών μέσα από τα κοινωνικά δίκτυα ενημέρωσης. Από εκεί μπορεί οποιοσδήποτε(φτάνει να έχει λογαριασμό στην ανάλογη πλατφόρμα), οποιαδήποτε στιγμή της ημέρας να εκφράσει την άποψη του στα διάφορα θέματα που τον απασχολούν. Για παράδειγμα όταν είναι περίοδος προεκλογικής εκστρατείας οι πολίτες/χρήστες παρακολουθούν τις κινήσεις και δηλώσεις που κάνουν οι διάφοροι πολιτικοί. Συγκεκριμένα μέσα από την πλατφόρμα του Twitter, μπορεί να ξεκινήσει ένας διάλογος που τους χρήστες εκφράζοντας τις απόψεις τους για τα πολιτικά πρόσωπα αλλά λέγοντας και την απόψεις τους για πολιτικά ζητήματα. Όμως, σε κάποιες περιπτώσεις που γίνεται διάλογος αρκετοί χρήστες βρίζονται μεταξύ τους για πολιτικά θέματα παρόλο που είναι άγνωστοι μεταξύ τους. Είναι ένα μέσο εκφράζοντας τις απόψεις τους δημόσια στην πλατφόρμα και να απαντάει

όποιος θέλει. Επίσης αξίζει να σημειωθεί πως οι χρήστες που κάνουν χρήση των κοινωνικών μέσων δικτύωσης είναι σε νεανικές ηλικίες σύμφωνα και με την έρευνα της Pew Internet Project's[27](εικόνα 6. 1)που έγινε τον Ιανουάριο 2014 για ποιες ηλικίες χρησιμοποιούν το Twitter και ελάχιστοι οι τρίτοι ηλικία. Άρα οι απόψεις από το μέσω τις κοινωνικής δικτυώσεις είναι συγκεκριμένες ηλικίες κυρίως νεανικές σε αντίθεση με της δημοσκοπήσεις που ορίζεις από την αρχή σε ποία ηλικία αναφέρεται.

Ωστόσο, υπάρχουν ομάδες αναλυτών που αναλύουν τις γνώμες των χρηστών οπουδήποτε στιγμή εκφράζοντας την ίδια στιγμή τα αποτελέσματα των χρηστών.

Twitter users

Among online adults, the % who use Twitter

All internet users	*19%*
a Men	22[b]
b Women	15
a 18-29	35[bcd]
b 30-49	20[cd]
c 50-64	11[d]
d 65+	5
a High school grad or less	15
b Some college	20
c College+	21
a Less than $30,000/yr	23[c]
b $30,000-$49,999	15
c $50,000-$74,999	13
d $75,000+	21

Pew Research Center's Internet Project January Omnibus Survey, January 23-26, 2014.
Note: Percentages marked with a superscript letter (e.g., [a]) indicate a statistically significant difference between that row and the row designated by that superscript letter, among categories of each demographic characteristic (e.g., age).

PEW RESEARCH CENTER

Εικόνα 6.1: Ποιες ηλικίες χρησιμοποιούν το Twitter,έρευνα από την εταιρία Pew Internet Project's

[27] http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

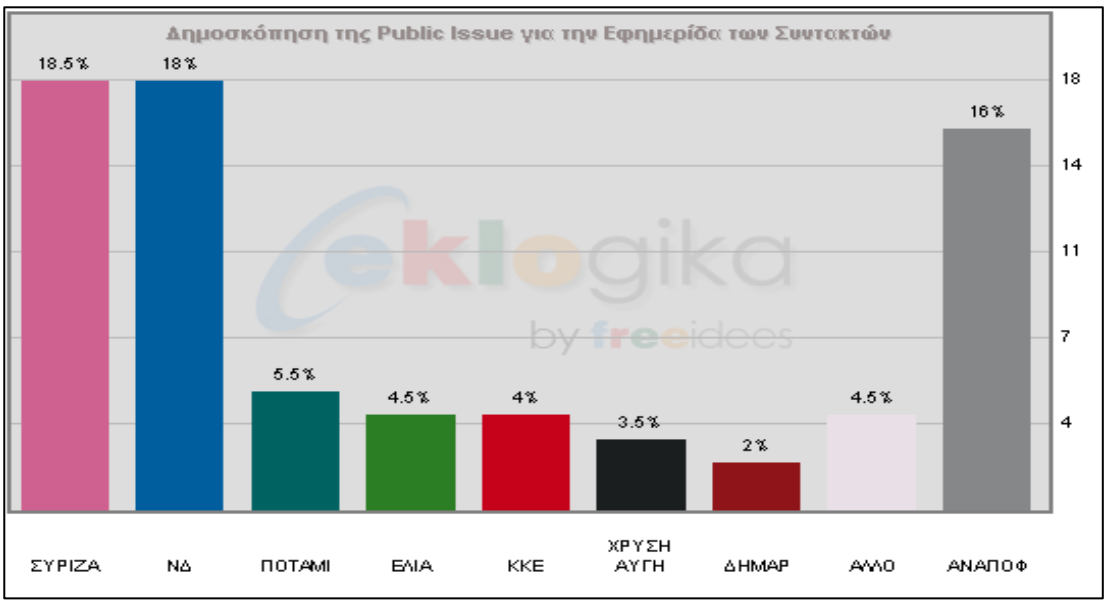
6.3 Προβλεπόμενα και Αποτελέσματα Εκλογών

Πιο κάτω παρουσιάζονται οι προβλέψεις που έγιναν στις Δημοσκοπήσεις από διάφορες εταιρίες για τις Δημοτικές-Περιφεριακές εκλογές και Ευρωεκλογές αλλά και τα αποτελέσματα που υπήρξαν στο τέλος.

Δημοσκοπήσεις

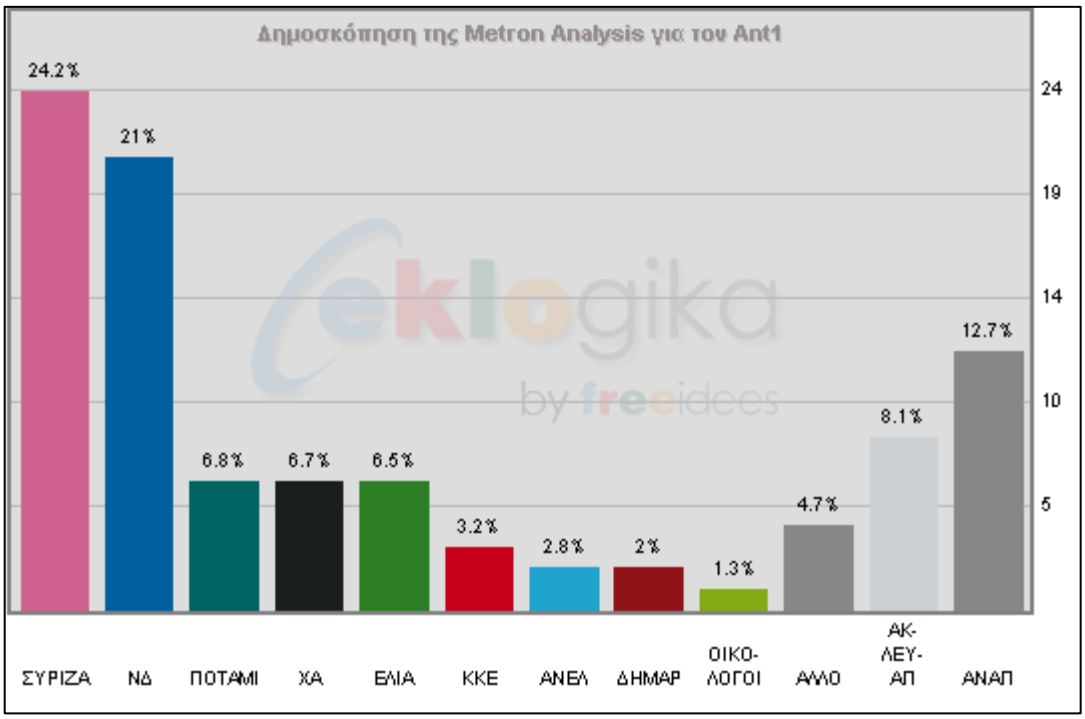
Δημοσκοπήσεις έγιναν από διάφορες εταιρίες όπως την GPO, MRB, Public Issue, Metro Analysis για τις διάφορες ηλεκτρονικές εφημερίδες αλλά και κανάλια κατά την διάρκεια τις προεκλογικής περιόδου των Δημοτικών Εκλογών και Ευρωεκλογών.

Η εταιρία Public Issue κάνει μια ποσοτική έρευνα για την πολιτική Συγκυρία από τις 29 Απριλίου μέχρι τις 6 Μάιου δημοσιεύοντας τα αποτελέσματα της στις 22 Μαΐου στην Εφημερίδα των Συντακτών. Η έρευνα αυτή αναφέρεται σε ηλικίες άνω των 18 χρονών και έγινε με τηλεφωνική συνέντευξη στα νοικοκυριά των ερωτώμενων και χρήση δομημένου ερωτηματολογίου χωρίς κάλπη. Τα θέματα που περιλάμβανε είναι για το γενικό πολιτικό και κοινωνικό κλίμα, τους αξιωματούχους του κράτους, την διακυβέρνηση και τα πολιτικά κόμματα, της Δημοτικές – Περιφερειακές Εκλογές και Ευρωεκλογές 2014 αλλά και τις Βουλευτικές εκλογές. Ήταν μια έρευνα με ποικιλία από ερωτήσεις για το τι πιστεύουν οι πολίτες. Τα αποτελέσματα απαρτίζονται πιο κάτω(εικόνα 6. 2) .



Εικόνα 6. 2: Δημοσκόπηση από την εταιρία Public Issue

Δυσαρέσκεια σχετικά με την κατάσταση που βιώνουν καθημερινά, αλλά και ανασφάλεια για το τι πρόκειται να ακολουθήσει τα επόμενα χρόνια, αισθάνεται η συντριπτική πλειοψηφία των πολιτών, όπως προκύπτει από το δεύτερο μέρος της έρευνας πολιτικής συγκυρίας της Public Issue[28] . Συγκεκριμένα, στο ερώτημα «πόσο ικανοποιημένοι είστε σήμερα από τη ζωή που ζείτε;» οι 7 στους 10 απαντούν «δυσαρεστημένοι», ποσοστό το οποίο είναι αυξημένο κατά 2% σε σχέση με αντίστοιχη έρευνα του προηγούμενου μήνα, ενώ «ικανοποιημένοι» δηλώνουν οι 3 στους 10(30%), ποσοστό μειωμένο αντίστοιχα κατά 2%. Σχετικά με το πώς αισθάνονται όταν σκέπτονται το μέλλον τους «τα επόμενα χρόνια», το 71% των πολιτών δηλώνουν «ανασφαλείς» και μόνο το 26% «ασφαλείς».



Εικόνα 6. 3: Δημοσκόπηση από την εταιρία Metro Analysis για τον Ant1 για τις Ευρωεκλογές

Πρωτιά στον ΣΥΡΙΖΑ, στην πρόθεση ψήφου για τις ευρωεκλογές, με 24,2% δίνει δημοσκόπηση της Metron Analysis [29]για τον τον τηλεοπτικό σταθμό «ΑΝΤ1». Στη δεύτερη θέση βρίσκεται η Ν.Δ. με ποσοστό 21%(εικόνα 6. 3).

[28] http://www.eklogika.gr/gallops/Public_Issue_efimerida_syntakton-12-05-2014

[29] http://www.eklogika.gr/gallops/MetronAnalysis-Ant1_22-5-14

Επίσης, δημοσκόπηση που έγινε από την εταιρία της GPO[30] για τα newsit.gr, iefimerida.gr και newpost.gr για τις Ευρωεκλογές δημοσιεύοντας τα αποτελέσματα στις 23 Μάιου δείχνει τη ΝΔ με το ΣΥΡΙΖΑ να έχουν απόσταση μόλις μια μονάδα. Συγκεκριμένα, το ΣΥΡΙΖΑ προηγείται της ΝΔ με 23,2% έναντι 22,2% και τρίτη θέση παίρνει η Χρυσή Αυγή, ενώ η ΕΛΙΑ με το Ποτάμι παίρνουν ακριβώς το ίδιο ποσοστό(7%). Πολύ χαμηλά ποσοστά έχει το ΚΚΕ(5,6%), οι ΑΝΕΛ(3,5%) η ΔΗΜΑΡ(2,5%), οι Οικολόγοι Πράσινοι(2%), ο ΛΑΟΣ(1,5%), ο ΑΝΤΑΡΣΥΑ(1,3%)ενώ οι αναποφάσιστοι 10,3%.

Από την πρώτη Κυριακή τον Δημοτικών Εκλογών τα exit polls των καναλιών[31], δείχνουν για τον Δήμο Αθηναίων ισοπαλία αναμεταξύ του κ. Καμίνη και του κ. Σακελλαρίδη ενώ στην περιφέρεια της Αττικής έχουμε την κ. Δούρου στο 24,5% και τον κ. Σγούρου στο 22 %. Ενώ το ποσοστό των αναποφάσιστων έρχεται στο 11,6%.

Αποτελέσματα εκλογών

Τα αποτελέσματα των δημοτικών εκλογών(περιφέρειας Αττικής) με πρώτη θέση να παίρνει το Σύριζα με την κ. Δούρου έχοντας το 50,82%, δεύτερη θέση να παίρνει το Πασοκ με τον κ. Σγούρου με 49,18% και στην 3η θέση τον κ. Κουμουτσάκο της ΝΔ με το 14,08% και για δήμαρχο Αθηναίων κερδίζοντας ο κ. Καμίνης.

Για τα αποτελέσματα των Ευρωεκλογών βρίσκονται στον πιο κάτω πίνακα(πίνακας 6.1) . Παρατηρούμε το κόμμα του Πασοκ να βρίσκεται στην έκτη θέση αποκτώντας 2 θέσεις στην ευρωβουλή, με τους ευρωβουλευτές Ανδρουλάκη Νικό και την Καηλή Εύα. Πρωτιά βέβαια είναι για το ΣΥΡΙΖΑ που για πρώτη φορά κερδίζει στις Ευρωεκλογές έχοντας το 28,70 % , ξεπερνώντας την ΝΔ.

[30] http://www.eklogika.gr/gallops/GPO_istoselides_23-05-2014

[31] http://www.nooz.gr/greece/ti-edeiksan-ta-exit-polls-ton-eklogon-ton-ota

Όνομα κόμματος	Ποσοστό (%)
ΣΥΡΙΖΑ	28,70
ΝΕΑ ΔΗΜΟΚΡΑΤΙΑ	21,16
ΧΡΥΣΗ ΑΥΓΗ	9,75
ΠΟΤΑΜΙ	7,34
ΚΚΕ	7,01
ΕΛΙΑ (ΠΑΣΟΚ)	6,38

Πίνακας 6. 1: Αποτελέσματα Ευρωεκλογών 2014

Σύγκριση πραγματικών αποτελεσμάτων

Παρατηρούμε πως το Σύριζα κερδίζει στις Εκλογές 2014 με μεγάλη πρωτιά όπως προέβλεπαν και οι δημοσκοπήσεις κατά όλη την διάρκεια της προεκλογικής περιόδου. Αυτό κάνει το Συνασπισμό Ριζοσπαστικής Αριστεράς(ΣΥΡΙΖΑ) να νιώθουν υπερήφανοι για το κόμμα τους αλλά και στείλουν πολλά μηνύματα στην κυβέρνηση το τι υποστηρίζουν οι πολίτες. Σε αντίθεση το Πασοκ βρίσκεται πολύ χαμηλά και η κυβέρνηση που υπάρχει (ΝΔ) τώρα να έρχετε σε δεύτερη θέση.

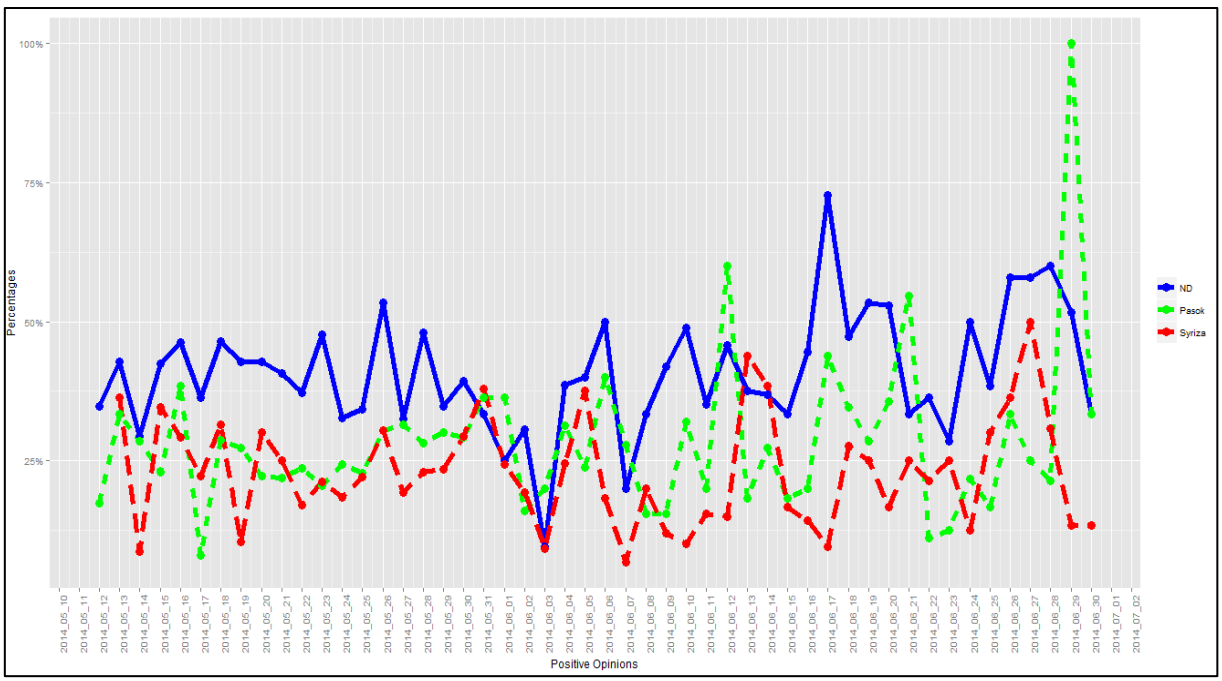
6.4 Πειραματικά Αποτελέσματα

Οι πιο κάτω γραφικές που απεικονίζονται είναι τα συνολικά θετικά, αρνητικά και ουδέτερα συναισθήματα που εντοπίστικαν για τα κόμματα. Ο άξονας Υ υπολογίζει πόσα τις εκατο (%)ανά μέρα είναι τα θετικά, αρνητκά και ουδέτερα tweets $\left(\frac{\alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}}{\alpha\rho\nu\eta\tau\iota\kappa\acute{\alpha}+\theta\epsilon\tau\iota\kappa\acute{\alpha}+ou\delta\acute{\epsilon}\tau\epsilon\rho\alpha}\right)$. Το κάθε κόμμα αντιστοιχή στο χρώμα του κόμματος του. Στον άξονα Χ είναι οι 50 μέρες συλλογής δεδομένων.

Συνολικός υπολογισμός θετικού συναισθήματος

Στην εικόνα 6.4 που βρίσκεται πιο κάτω, παρουσιάζονται όλα τα θετικά συναισθήματα που εντοπίστικαν από τα τρία κόμματα. Οι θετικές γνώμες της ΝΔ κυμένονται σε μέτριο βαθμό υπολογισμού από 23% μεχρι 65% αλλά στις 17 Ιουνίου να φτάνει στο 74%. Έρχετε δεύτερο σε σειρά υψηλότερο ποσοστό και από τα τρία κόμματα. Το πολιτικό γεγονός που υπήρξε στις 17 Ιουνίου είναι ότι πραγματοποιήθηκε γεύμα ηγετών της Ε.Ε συμμετέχοντας και ο προσθυπουργός Α. Σαμαράς. Επίσης, το πρακτορείο Reuters ανακεινόνει νέα έξοδο της Ελλάδας από τις αγορές πριν τον Αύγουστο.

Τα περισσότερα χαμηλότερα ποσοστά θετικότητας γνώμης έχει το Σύριζα και δεν ξεπερνούν το 50%. Την πρώτη Κυριακή των εκλογών οι χρήστες σχολίαζουν θετικά την ΝΔ σε σχέση με τα αλλά δύο κόμματα, όμως το Σύριζα έχει αρκετο χαμηλό ποσοστό και λίγο πιο πάνω το ΠΑΣΟΚ. Μια μέρα πριν τις εκλογές και τα τρία κόμματα κάνουν τους τελευτές τους υπολογισμούς και ο κ. Σαμαράς κάνει ορισμένες δηλώσεις. Για εκείνη την μέρα φαίνετε πως οι χρήστες δεν αναρτούν πολλα θετικά σχόλια για τα κόμματα συγκρίνοντας με τις υπόλιπες μέρες. Την 2η Κυριακή των εκλογών το ΠΑΣΟΚ και Σύριζα έχουν το ίδιο ποσοστό θετικών σχολιασμών από τους χρήστες έχοντας 25% ενώ η ΝΔ με 36%. Ακόμη στις 3 Ιουνιου η Σύριζα και η ΝΔ έχουν το ίδιο βαθμό θετικών αναρτήσεων, βέβαια δεν ξεπερνά το 15%.



Εικόνα 6. 4: Όλες οι θετικές γνώμες από τα τρία δημοφιλέστερα κόμματα

Τα πολιτικά γεγονότα που έγιναν εκείνη την μέρα είναι ο Τσίπρας βρίσκεται στις Βρυξέλλες για να μεταφέρει την αντιπαράθεση με την «Ευρώπη της Μέρκελ» έχοντας επαφές άλλους συνοψηφίους πεικεφαλής προεκλογικών εκστρατειών των Σοσιαλιστών-Δημοκρατκικών. Επίσης την ίδια μέρα ανακοινωνονται τα αποτελέσματα από την Ευρωπαική Υπηρεσία για την ανεργία έχοντας φτάσει στα 26,5% στην Ελλάδα.

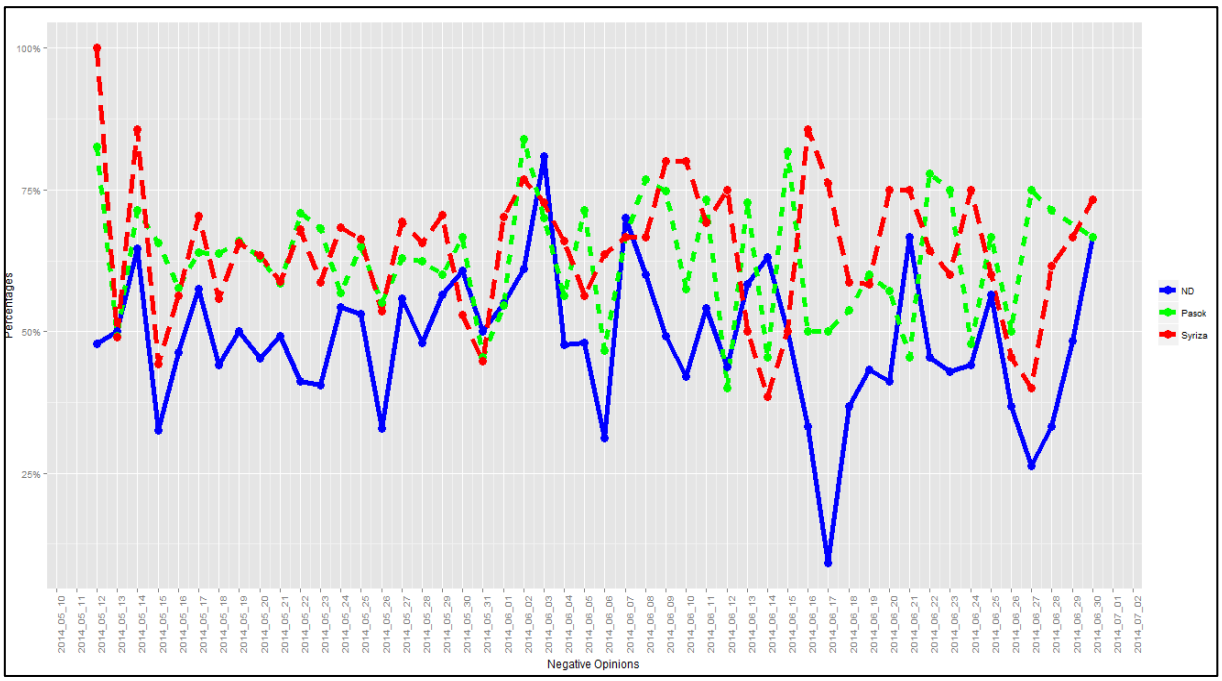
Το μεγαλύτερο θετικό ποσοστό αναρτήσεων έγινε στις 29 Ιουνίου αντιπροσοπόντας το ΠΑΣΟΚ, με το πολιτικό γεγονός οι 'γαλάζοι' να στέλνουν μήνυμα από νυν και πρώην υπουργοί για την στρατιγκή της επόμενης μέρας. Έχοντας θετικούς σχολιασμούς από τους χρήστες του Twitter. Επιπλέον, στις 30 Ιουνίου το ΠΑΣΟΚ και η ΝΔ έχουν το ίδιο αποτέλεσμα φτάνοντας στο 36% περίπου και το Σύριζα στα 14% περίπου. Εκείνη την μέρα ο κ.Βενιζέλος και ο προθυπουργός κάνουν συνάντηση και συζητούν για ιδιωτικοποίηση της «Μικρής ΔΕΗ» αλλά και άλλα θέματα της βουλής.

Συνολικός υπολογισμός αρνητικού συναισθήματος

Στην πιο κάτω γραφική παράσταση(εικόνα6.5)παρουσιάζονται οι αρνητικές γνώμες των χρηστών κατά την ανάλυση του συναισθήματος που είχαν κατά την περίοδο των εκλογών για τα τρία δημοφιλέστα κόμματα. Στις 12 Μαίου το Σύριζα φτάνει στο 100% αρνητικής γνώμης των χρηστών ενώ τα άλλα κόμματα να βρίσκονται σε πιο χαμηλά ποσοστά, ιδικά η ΝΔ να μην ξεπερνα το 50%. Οι χρήστες εκείνη την μέρα εκφράζουν τις γνώμες τους γύρο από τα πολιτικά θέματα που επικρατούν εκφράζοντας αρνητικές γνώμες για το Σύριζα. Την επόμενη μέρα και τα τρία κόμματα έχουν τον ίδιο βαθμό αρνητικότητων γνώμης από τους χρήστες κατά 50%. Αλλά τις επόμενες μέρες αυξάνεται τα ποσοστά στα κόμματα ξεπερνόντας με 65% το Σύριζα.

Την πρώτη Κυριακή των εκλογών οι χρήστες εκφράζουν περισσότερο αρνητικά για το ΠΑΣΟΚ κατά 62%, το Σύριζα καα 35% και την ΝΔ στα 45%. Ενώ την 2η Κυριακή των εκλογών οι χρήστες εκφράζουν αρνητικά tweets για τους πολιτικούς των τριών κομμάτων που σχεδόν φτάνουν στο 50%.

Στις 14 Ιουνίου τα αρνητικά tweets για την ΝΔ φτάνουν στο 60% ενώ τις επόμενες μέρες μειώνετε παρα πολύ φτάνοντας σχεδόν στα 20% στις 17 Ιουνίου έχοντας και το ελάχιστο ποσοστό αρνητικότητας για την ΝΔ.



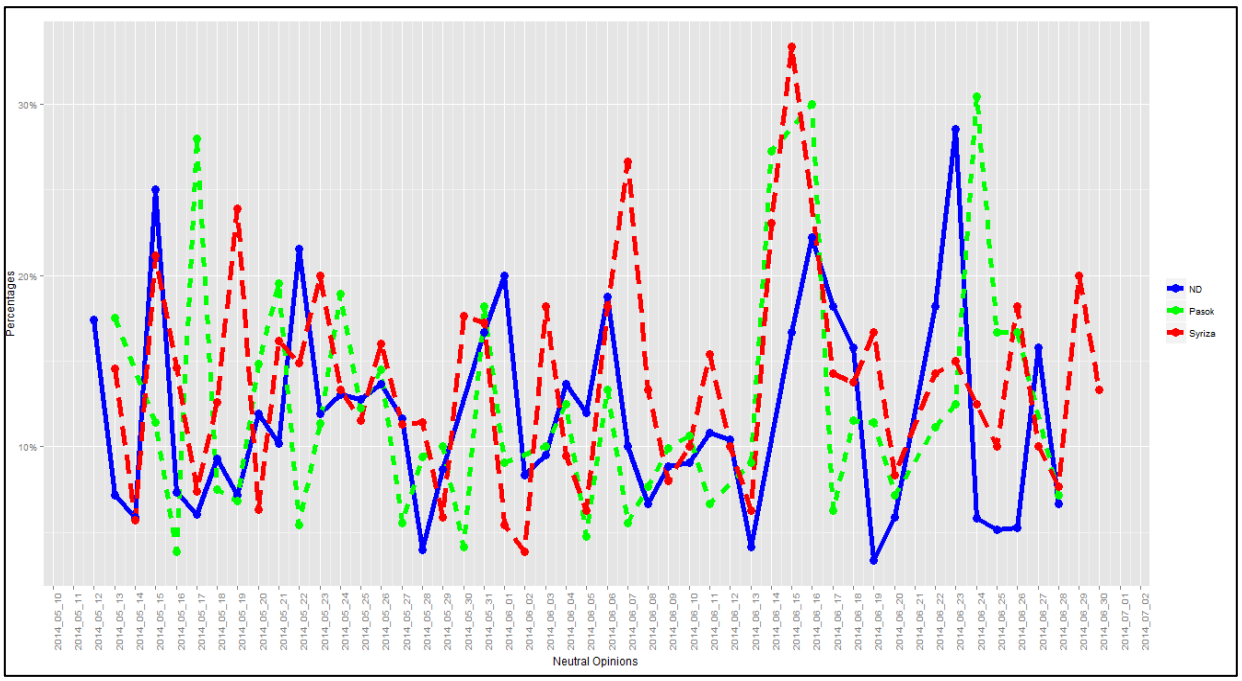
Εικόνα 6.5 : Όλες οι αρνητικές γνώμες από τα τρία δημοφιλέστερα κόμματα

Επίσης παρατηρούμε στις 27 Ιουνίου οι χρήστες στα tweets αναφέρονται και στα 3 κόμματα εκφράζοντας αρνητικές γνώμες με περισσότερη διαφορά για το ΠΑΣΟΚ στα 75%. Ολοκληρώνοντας στις 30 Ιουνίου ΝΔ και ΠΑΣΟΚ να συναντουνταί στις γνώμες των χρηστών και με λίγο πιο περισσότερο ποσοστό να εκφράζουν αρνητιτική γνώμη και για το ΣΥΡΙΖΑ.

Συνολικός υπολογισμός ουδέτερου συναισθήματος

Στην πιο κάτω γραφική παράσταση(εικόνα6.6)απεικονίζονται οι ουδέτερες γνώμες των χρηστών που εκφράζουν για τα τρία κόμματα. Οι γνώμες αυτές εμφανίστικαν όταν ο αλγόριθμος ταξινομούσε τις θετικές και αρνητικές γνώμες των χρηστών μέσα από τα tweets και είχαν σύνολο τον ίδιο αριθμό συναισθήματος από θετικά και αρνητκά. Αυτό προκύπτει ότι οι χρήστες δεν εκφράζουν ακριβώς το αυναίσθημα τους αν αυτό που λένε είναι θετικό ή αρνητικό είτε να μην έχουν μια ξεκάθαρη άποψη για το πολιτικό θέμα που αναφέρονται.

Σε αυτή την περίπτωση τα κόμματα σχεδόν έχουν το ίδιο αριθμό ποσοστού. Ξεχωρίζει το ΠΑΣΟΚ και λίγο πιο κάτω το Σύριζα. Στις 14 Μαίου μέχρι τις 16 Μαίου το κόμματα Σύριζα και ΝΔ σχεδόν έχουν τον ίδιο συναίσθημα εκφράσεων από τους χρήστες. Επίσης, στις 13 με 17 Ιουνίου ΠΑΣΟΚ και Σύριζα οι γνώμες των χρηστών είναι οριακά ίδιες.



Εικόνα 6. 6: Όλες οι ουδέτρες γνώμες από τα τρία δημοφιλέστερα κόμματα

Ακόμη στις 28 Ιουνίου εμφανίζει μικρό ποσοστό ουδέτερων γνώμης από τους χρήστες και τις επόμενες μέρες αυξανονται τα ουδέτερα tweets για την ΝΔ φτάνοντας στις 31 Μαίου.

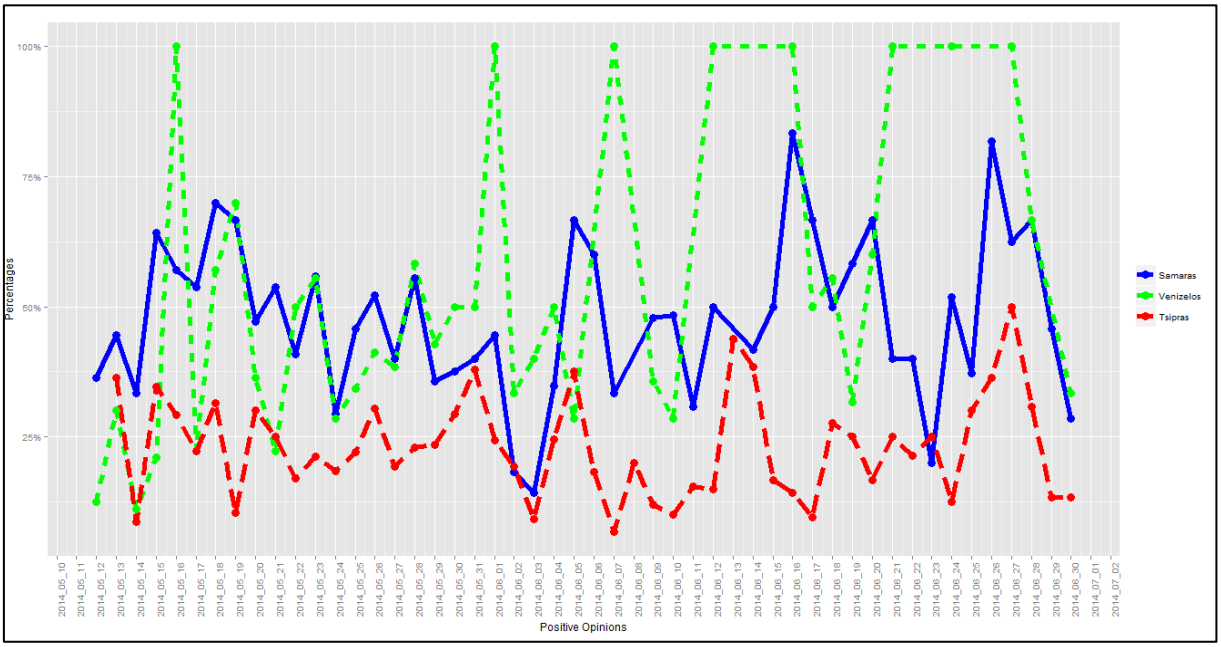
6.5 Ανάλυση Προέδρων Κομμάτων

Επικεντρωνόμαστε στους εκπροσώπους των τριών δημοφιλέστερων κομμάτων(Τσίπρας, Σαμαράς, Βενιζέλος),εντοπίζοντας αν εκφράζουν θετικές, αρνητικές ή ουδέτερες γνώμες οι χρήστες για αυτούς. Η διαδικασία εντοπισμού των προέδρων των κομμάτων έγινε με τον ίδιο τρόπο που αναλυθήκαν και πιο πάνω.

Συνολικός υπολογισμός θετικού συναισθήματος

Στην πιο κάτω γραφική(εικόνα 6.7)παρουσίαζουμε τις θετικές γνώμες των χρηστών του Twitter για τους τρείς εκπροσώπους των κομμάτων. Παρατητούμε πως ο Βενιζέλους του ΠΑΣΟΚ έχει τα πιο ψηλά ποσοστά θετικότητας από τους χρήστες. Ειδικά στις μέρες 16 Μαίου, 1, 7, 12 Ιουνίου και την περίοδο 12- 17 Ιουνίου και στις 20 με 27 Ιουνίου. Αυτές τις μέρες ο Βενιζέλος παίρνει 100% θετικά συναισθήματα από τους χρήστες. Τα γεγονότα που διαδραματίστηκαν εκείνες τις μέρες ο Βενιζέλος είχε συνάντηση με τον τούρκο ομόλογο στις Βρυξέλλες και ο απολογισμός για τα αποτελέσματα των εκλογών κ.α γεγονότα.

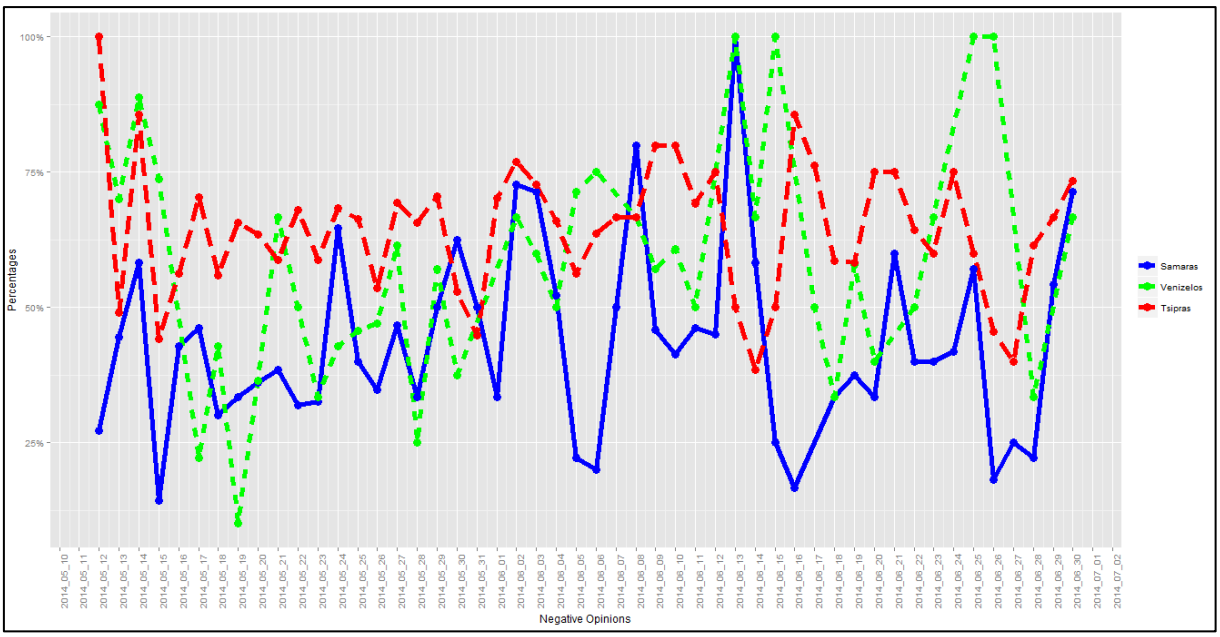
Στις πρώτες μέρες ο Βενιζέλος έχει χαμηλά ποσοστά θετικότητας συναισθημάτων από τους χρήστές ενώ τις επόμενες μέρες αυξάνεται αλλά μετα ξανα χαλημηνούν. Γενικά παρατηρούμε πως οι χρήστες του Twitter εκφράζουν τις προσσότερες θετικές γνώμες προς τον Βενιζέλο και μάλιστα αυξημένα ποσοστά, όπως αναφέρθηκαν και πιο πάνω. Οι θετικές γνώμες των χρηστών για τον προθυπουργό Σαμαρά είναι μέτριες με το υψηλότερο ποσοστό 80% με το χαμηλότερο στα 20% στις 3 Ιουνίου. Από την άλλη πλεύρα έχουμε τον Τσίπρα που οι χρήστες δεν εκφράζονται αρκετά θετικά για το πρόσωπο του. Συγκρίνοντας με τους άλλους δύο πολιτικούς ο Τσίπρας έχει τα πιο λίγα ποσοστά.



Εικόνα6.7: Θετικό συναίσθημα για τους τρείς εκπρόσωπων των κομάτων

Συνολικός υπολογισμός αρνητικού συναισθήματος

Στην πιο κάτω γραφική(εικόνα6.8)απεικονίζονται τα αρνητικά συναισθήματα των χρηστών προς τους τρίς προεδρούς των κομμάτων. Οι πέντε μέρες από τις 50 συνολικές που έχουμε οι χρήστες εκφράζουν 100% αρνητική γνώμη για τους εκπροσώπους των κομμάτων. Δηλαδή στις 12 Μαίου ο Τσίπρας παιρνει αρνητικά tweets, στις 13 Ιουνίου ο Βενιζέλος και ο Σαμαράς, 15 Ιουνίου ο Βενιζέλος και στις 25- 26 Ιουνίου ο Βενιζέλος. Στις 13 Ιουνίου μπορεί ο Σαμαράς να σχολιάζετε τόσο πολύ αρνητικά αλλά τις επόμενες μέρες μειώνεται κατακόριφα το πολιτικό γεγονός εκείνης της μέρα είναι πως ανακοινώνουν πως δεν θα δεχτούν διαφοραποιήσεις από υπουργούς.



Εικόνα6.8: Αρνητικό συναίσθημα για τις τρείς εκπροσώπων των κομμάτων

Το πιο χαμηλότερο ποσοστό αρνητικού συναισθήματος είναι την επόμενη μέρα τις 1ης Κυριακής των εκλογών(19 Μαίου) και οι εκπρόσωποι των κομμάτων να σχολίαζουν τα αποτελέσματα των Δημοτικών εκλογών και να στέλούν τα δίαφορα μηνύματα προς τους πολίτες. Οι χρήστες όμως του Twitter αναρτούν με 15% αρνητικά συναισθήματα προς τον Βενιζέλο. Την ίδια μέρα ομώς ο Τσίπρας παίρνει τα περισσότερα αρνητικά σχόλια από τους χρήστες με 62% περίπου.

Παρατηρούμε πως ο Βενιζέλος από τις αναρτήσεις των χρηστών προς το όνομα του σχολίαζεται αρκετά αρνητικά ενώ ο Τσίπρας από το Σύριζα να έχει περισσότερο αρνητικό σχολίασμο από τους χρήστες του Twitter. Ενώ οι χρήστες για τον Σαμαρά είναι ασταθής από μέρα σε μέρα είναι διαρορετικό το αποτέλεσμα.

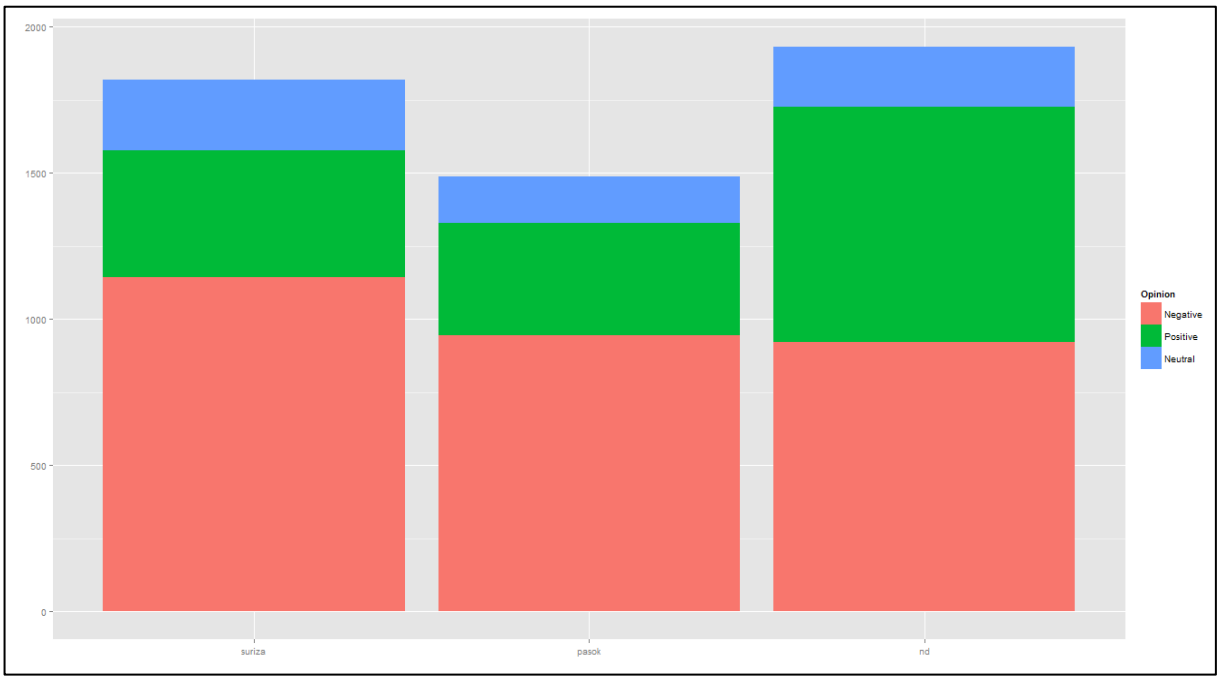
6.6 Ανάλυση Αποτελεσμάτων

Στην πιο κάτω γραφική παράσταση(εικόνα6.9)απεικονίζονται πόσα σε σύνολο θετικών, αρνητικών , ουδέτερων αναρτήσεων εντοπίστηκαν για το κάθε κόμμα συνολικά σε κάθε bar της γραφικής. Το κόκκινο αντίστοιχη στις αρνητικές αναρτήσεις το μπλε στα ουδέτερα και το

πράσινο στα θετικά. Στον άξονα Χ είναι το όνομα του κάθε κόμματος. Τα συνολικά θετικά tweets είναι 1623, τα αρνητικά στα 3004 tweets και τα ουδέτερα στα 606 tweets.

Παρατηρούμε πως έχουμε περισσότερο σχολιασμό για το πολιτικό κόμμα της ΝΔ φτάνοντας σχεδόν τα 2000 συνολικές αναρτήσεις από θετικές, αρνητικές και ουδέτερες. Το μέγεθος των αρνητικών είναι σχεδόν το ίδιο με το συνολικό του ΠΑΣΟΚ που είναι λίγο πιο χαμηλά. Βέβαια το ΣΥΡΙΖΑ σε σχέση και με τα δύο κόμματα έχει τις περισσότερες αρνητικές αναρτήσεις.

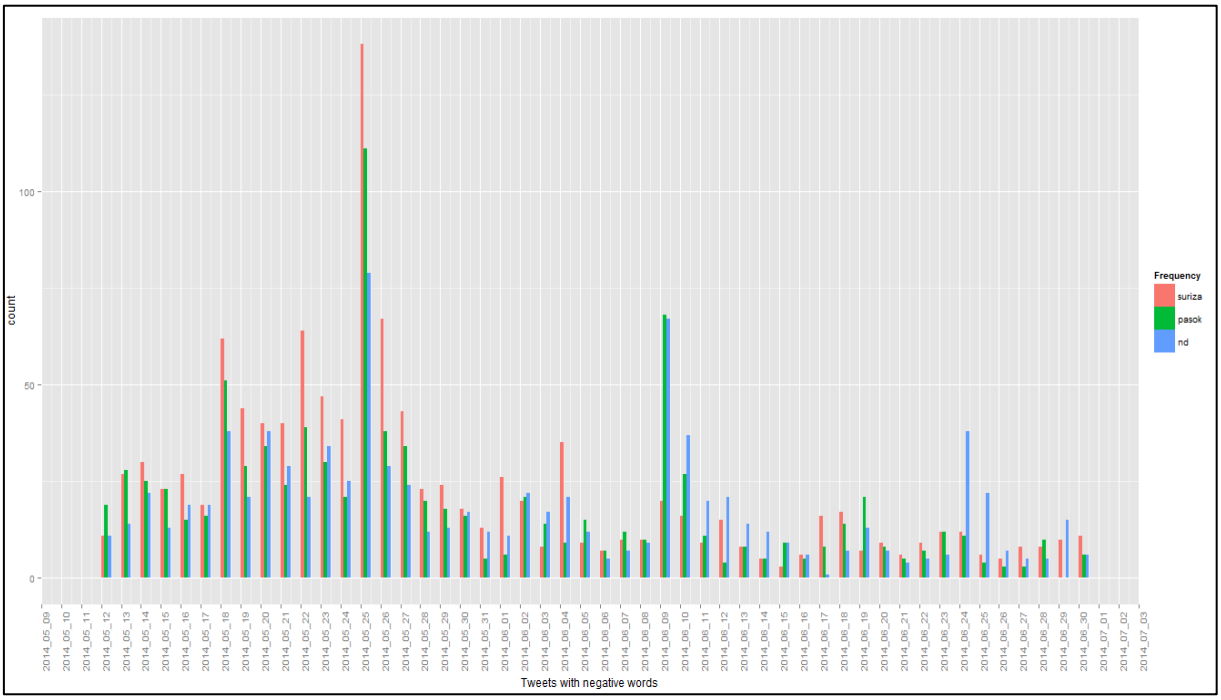
Συγκρίνοντας τα θετικά και τα αρνητικά της ΝΔ σχεδόν έχουν το ίδιο αποτέλεσμα, αλλά όμως ξεπερνούν με λίγη διαφορά τα αρνητικά tweets . Τα ουδέτερα του ΣΥΡΙΖΑ είναι τα διπλάσια από τα άλλα δύο κόμματα. Επίσης, το ΣΥΡΙΖΑ σε συνολικό αναρτήσεων φτάνει περίπου στα 1750 ενώ το ΠΑΣΟΚ σχεδόν στα 1500 tweets. Οι χρήστες μπορεί να σχολιάζουν αρκετά αρνητικά το ΣΥΡΙΖΑ αλλά οι περισσότεροι να ασχολούνται με την διοικούσα κυβέρνηση με τις δηλώσεις και τις ενέργειες που κάνει.



Εικόνα 6.9: Συνολικός αριθμός tweets για το κάθε κόμμα από τον εντοπισμό θετικών, αρνητικών και ουδέτερων tweets.

Συνολικά αρνητικά tweets

Από την πιο πάνω γραφική είδαμε τα συνολικά πόσα tweets είχαν τα κόματα εκφράζοντας θετικά, αρνητικά και ουδέτερα συναισθήματα. Τώρα επικεντρωνόμαστε μόνο στα αρνητικά tweets που παρουσιάζονται παρακάτω(εικόνα6.10)που έκαναν οι χρήστες και από τα τρία κόμματα γιατί είναι τα πιο πολλά.



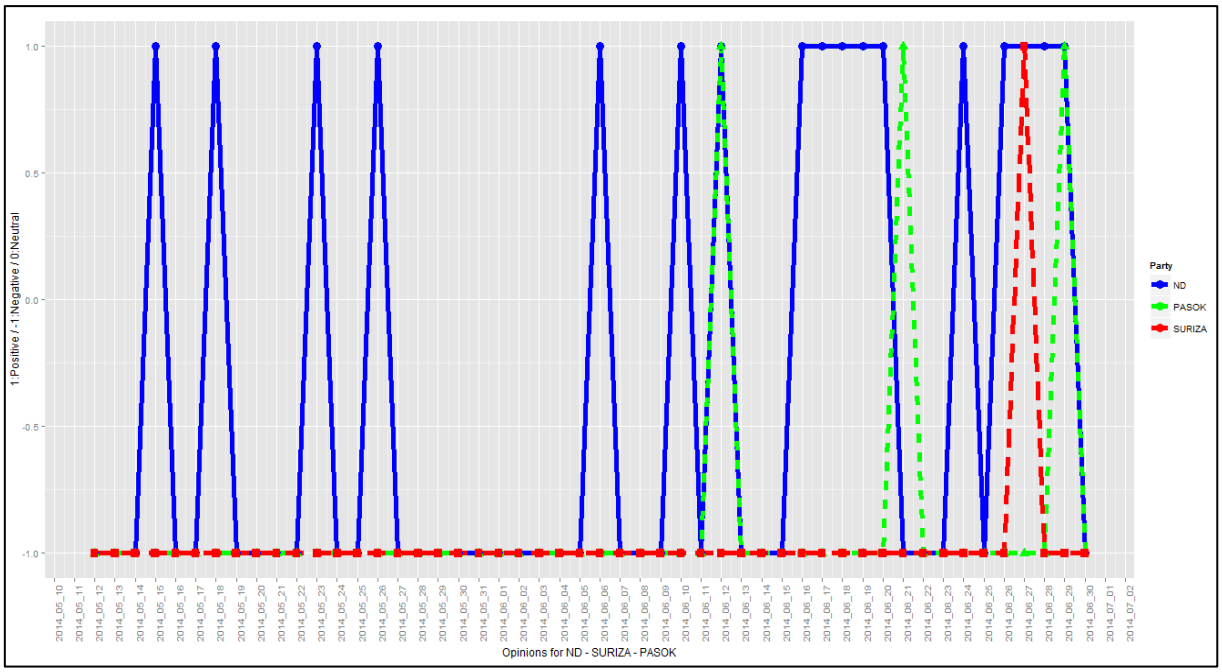
Εικόνα 6.10 : Τα αρνητικά tweets που έκανα οι χρήστες και για τα τρία κόμματα

Όπως αναφέραμε και πιο πανω τα περισσότερα αρνητικά tweets έγιναν για το κόμμα ΣΥΡΙΖΑ. Την 2η Κυριακή των εκλογών το ΣΥΡΙΖΑ με αρκετή διαφόρα από τα άλλα δύο κόμματα σχολιάζεται αρνητικά από τους χρήστες του Twitter,ενώ την ίδια μέρα ΠΑΣΟΚ και ΝΔ είναι πιο κάτω. Επίσης το ίδιο επίπεδο tweets υπάρχει και για την 1η Κυριακή των εκλογών αλλά με λιγότερα tweets. Στις 9 Ιουνίου ΝΔ και ΠΑΣΟΚ έχουν το ίδιο αριθμό tweets με σχεδόν τριπλάσια διαφόρα από το ΣΥΡΙΖΑ.

Συσχέτιση με τις εκλογές

Στην πιο κάτω γραφική παράσταση(εικόνα 6.10)παρουσιάζουμε το συνολικό συναίσθημα των χρήστων σε +1 τα θετικά συναισθήματα, -1 τα αρνητικά συναισθήματα και ουδέτερα συναισθήματα σε 0 ανά μέρα. Παρατηρούμε το κόμμα της ΝΔ να έχει τα περισσότερα θετικά

συναισθήματα σε σχέση με τα άλλα κόμματα σε αρκετές μέρες. Ενώ τα περισσότερα αρνητικά συναισθήμαα είχαν όσοι εκφράζοντας για το πολιτικό κόμμα του ΣΥΡΙΖΑ. Στις 11 Ιουνίου μέχρι τις 13 Ιουνίου και στις 30 Ιουνίου ΝΔ και ΠΑΣΟΚ είχαν θετικά συναισθήματα. Στις 27 Ιουνίου οι χρήστες εξέφρασαν θετικά συναισθήματα για το ΣΥΡΙΖΑ. Συνεχόμενα θετικά συναισθήματα για την ΝΔ ήταν στις 16 μέχρι τις 21 Ιουνίου. Επίσης την ημέρα των εκλογών οι χρήστες ανατρούσαν θετικά tweets για την ΝΔ ενώ για το ΣΥΡΙΖΑ αρνητικά.



Εικόνα 6.10: Συναίσθημα θετικό(+1),αρνητικό(-1),ουδέτερο(0) για τα τρία κόμματα

Γενικές παρατηρήσεις- Σύγκριση πραγματικών και πειραματικών αποτελεσμάτων

Από την ταξινόμηση που κάναμε για τα τρία κόμματα παίρνουμε τα εξής αποτελέσματα :

Όταν έγινε η αναζήτηση με το όνομα του κόμματος , οι περισσότερες αναρτήσεις ήταν για το ΣΥΡΙΖΑ, ένω όταν κάναμε την αναζήτηση σε πολιτικά πρόσωπα τα περισσότερα αναφέρονταν για την ΝΔ. Στην συνέχεια όταν ταξινομήσαμε την ανάλυση συναισθήματος των χρήστων για τα πολιτικά κόμματα και για τους πολιτικούς παρατηρήσαμε ότι εκφράζουν περισσότερες αρνητικές γνώμες για το ΣΥΡΙΖΑ ενώ για τα άλλα κόμματα πιο χαμηλά. Οι περισσότεροι χρήστες αναρτούσαν και εξέφραζαν την γνώμη τους για την ΝΔ. Επίσης το ΠΑΣΟΚ να έχει πιο χαμηλές αναρτήσεις από τους χρήστες αλλα εκφράζουν τις περισσότερες θετικές γνώμες προς τον κ.Βενιζέλο.

Συγκίνοντας τα πραγματικά με τα πειραματικά αποτελέσματα που είχαμε, δεδομένου ότι ο αλγόριθμος που δημιουργήθηκε δουλεύει σωστα δεν είναι αντιπροσωπεύτικά με τα πραγματικά αποτελέσματα. Πιθανό οι χρήστες να είναι από ένα συγκεκριμένο κόμμα και εκφράζοντας την γνώμη τους κάθως επίσης το δείγμα που είχαμε πιθανό να ήταν μικρό και να μην είχε πάρει όλο το φάσμα των πολιτικών πολιτών. Από την συλλογή των δεδομένων η ΝΔ συγκετρώνει τις περισσότερες θετικές γνώμες και ο Σύριζα να έχει τις περισσότερες αρνητικές γνώμες. Επίσης ο αριθμός των εγκεγραμένων υποψηφίων δεν αντιστιχοί στο ίδιο με τους χρήστες

6.7 Αξιολόγηση του Αλγορίθμου

Η αξιολόγηση του αλγορίθμου έγινε με δύο τρόπους:

Α' τρόπος

Δημιουργήσαμε 15 δικές μας πολιτικές προτάσεις και τις τρέξαμε στην R για να δούμε αν δουλεύει σωστά ο αλγόριθμος. Οι προτάσεις περιλάμβαναν μέσα και το 'δεν' έχοντας συζήτηση και για τα τρία κόμματα. Τα αποτελέσματα που πήραμε είναι αρκετά θετικά πως η αξιολόγηση του αλγορίθμου βγάζει σωστά το συναίσθημα. Παρατηρήσαμε πως σε αρκετές προτάσεις είναι συγκριτικές προτάσεις ή έχουν διάφορα emotions και σημεία στίξης, δεν μπορεί να ανάλυση ο αλγόριθμος σωστά το συναίσθημα. Παρακάτω παρουσιάζουμε τα αποτελέσματα(5 προτάσεις για το κάθε συναίσθημα): ΝΔ (4/5 θετικά , 3/5 αρνητικά , 3/5 ουδέτρα) , ΣΥΡΙΖΑ (3/5 θετικά ,4/5 αρνητικά , 3/5 ουδέτερα), ΠΑΣΟΚ (4/5 θετικά , 3/5 αρνητικά 3/5 ουδέτερα)

Β' τρόπος

Τυπώσαμε 15 tweets από το κάθε κόμμα όταν έγινε η ταξινόμηση των θετικών αρνητικών και ουδέτερων συναισθημάτων. Κάνοντας αντιστοίχιση με τις λέξεις του λεξικού κοιτάζοντας αν βγάζουμε το ίδιο αποτέλεσμα σε κάθε tweet όπως τον αλγόριθμο. Τα αποτελέσματα ήταν αρκετά ικανοποιητικά έχοντας τις ίδιες παρατηρήσεις με τον πιο πάνω τρόπο. Παρακάτω παρουσιάζουμε τα αποτελέσματα(5 προτάσεις για το κάθε συναίσθημα): ΝΔ (3/5 θετικά , 4/5 αρνητικά , 3/5 ουδέτρα) , ΣΥΡΙΖΑ (3/5 θετικά , 2/5 αρνητικά , 3/5ουδέτρα), ΠΑΣΟΚ (3/5 θετικά, 4/5 αρνητικά , 4/5 ουδέτερα).

Κεφάλαιο 7ο
Συμπεράσματα και προοπτικές

Σε αυτό το κεφάλαιο αναφέρονται περιληπτικά τα βήματα που έγιναν και τα αποτελέσματα που πήραμε από την υλοποίηση του αλγόριθμου. Στο τέλος αναφέρονται και ορισμένες μελλοντικές προοπτικές.

7.1 Συμπεράσματα

Η εξόρυξη Γνώμης μέσα από τα κοινωνικά μέσα(Twitter) αποτελεί ένα αναπτυσσόμενο επιστημονικό πεδίο που αποσκοπεί στην ανακάλυψη κρίσιμων 'κρυμμένων' πληροφοριών που αποτελεί σπουδαία εγχειρήματα της εποχής μας. Οι εφαρμογές και οι επιστημονικές μελέτες αυξάνονται συνεχούς λόγω της εκκριτικής πληροφορίας που παρατηρείται τα τελευταία χρόνια και την ανάγκη για εξαγωγή χρήσιμων πληροφοριών από την πληθώρα των διαθέσιμων δεδομένων που αναρτούν οι χρήστες κάθε λίγα δευτερόλεπτα. Το συναίσθημα ανάλυσης είναι σημαντικό μέσα από ένα εύρος αναρτήσεων να μπορέσουμε να δούμε σύντομα τη συναίσθημα(θετικό, αρνητικό) εκφράζει ο κάθε χρήστης.

Ο πρώτος στόχος της παρούσας διπλωματικής εργασίας ήταν να διερεύνηση τη έννοια της εξόρυξης γνώμης / συναίσθημα ανάλυσης μέσα από την σχετική βιβλιογραφία που μελετήθηκε καθώς επίσης στην εξόρυξη γνώμης μέσα από τα κοινωνικά μέσα δίνοντας έμφαση στην πλατφόρμα του Twitter. Βλέποντας σε ποιες περιπτώσεις χρησιμοποιείται αλλά και με ποιους μεθόδους γίνεται η εξαγωγή και η ταξινόμηση. Επίσης φαίνεται ο τεράστιος όγκος των δεδομένων να αυξάνεται λόγο της σύγχρονης ζωής και η ανάγκη για εξόρυξη γνώμης αφού οι χρήστες νιώθουν την ανάγκη να μοιραστούν τις απόψεις τους, τους προβληματισμούς και γενικά ότι τους συμβαίνει εκείνη την στιγμή.

Ο δεύτερος στόχος ήταν να συλλέξουμε τα πληθώρα δεδομένα για ένα χρονικό διάστημα και μετά να επιλέξουμε αυτά που είναι γραμμένα στα ελληνικά επικεντρώνοντας στις πολιτικές συζητήσεις. Στην συνέχεια ήταν η σχεδίαση και η υλοποίηση αλγορίθμου για κατηγοριοποίηση γνώμης στην εξόρυξη γνώμης μέσα από το πρόγραμμα της R. Προσπαθώντας να προσεγγίσουμε όσο το δυνατόν ακριβότερα και πληρέστερα.

Συγκρίνοντας τα αποτελέσματα των δημοσκοπήσεων, τα πραγματικά αποτελέσματα και τα πειραματικα που κάνουμε, βρήκαμε ένα έντονο σχολιασμό απόψεων των χρηστών αλλά δεν είναι αντιπροσωπευτικά με τα πραγματικά αποτελέσματα.

7.2 Προοπτικές

Ολοκληρώνοντας την παρούσα διπλωματική εργασία παρουσιάζουμε κάποιες προτάσεις που θα μπορούσαν να συνεισφέρουν στην βελτίωση της λειτουργίας του αλγορίθμου που κατασκευάστηκε αλλά και ανάπτυξη της εργασίας.

- Η διαμόρφωση του θετικού και αρνητικού λεξιλογίου ώστε να είναι προσαρμοσμένο κατάλληλα στην αντίστοιχη πολιτική κατηγορία για να γίνεται πιο σαφές η αναζήτηση λέξεων μέσα από τις ελληνικές αναρτήσεις και να είναι πιο χρησιμοποιούμενο λεξιλόγιο σχολιασμού που παρουσιάζονται στο συγκεκριμένο θέμα. Επίσης να γίνει λεξιλόγιο από φράσεις θετικών και αρνητικών λέξεων σχετικές με τα πολιτικά θέματα.

- Να υπάρξει συνεχόμενης ροής συλλογής δεδομένων και κάθε δύο ώρες να γίνεται η ανάλυση συναισθήματος ειδικότερα όταν είναι κάποιες εκλογές για να ενημερώνονται οι χρήστες αλλά και ειδικότερα οι πολιτικοί να γνωρίζουν τι γνώμες έχουν οι χρήστες για

αυτούς. Επίσης,να υπάρχει και ανάλυση συναισθήματος μόνο με τις αναρτήσεις των πολιτικών πως εκφράζονται κάθε μέρα μέσα από το Twitter.

- Η δυσκολία ερμήνευσες των αναρτήσεων από τις συντομογραφίες των λέξεων που είναι κωδικοποιημένα. Όπως το 'κ' που είναι το 'και', 'γτ' είναι το 'γιατί', 'πρπ' είναι το 'πρέπει' κ.α πολλά. Αξιοσημείωτη θα ήταν η ύπαρξη κάποιου μηχανισμού ταξινομητή για αυτή την κωδικοποίηση για καλύτερη κατανόηση της ανάλυσης των αναρτήσεων.

- Στον αλγόριθμο να αφαιρεθούν τα σημεία στίξης για να εξασφαλιστεί ένα καλύτερο αποτέλεσμα δηλαδή όπου υπάρχουν « ? , ! { } » γιατί επηρεάζεται και το συναίσθημα του tweet.

- Έλεγχος των συγκριτικών προτάσεων που υπάρχουν γιατί στην περίπτωση των πολιτικών συζητήσεων σε κάποιες αναρτήσεις χρηστών δεν μπορείς να καταλάβεις με πιο κόμμα ανήκουν καθώς επίσης δεν εκφράζονται άμεσα αν είναι θετική ή αρνητική η γνώμη τους.

- Η μελέτη αυτή ήταν περιορισμένη στο κοινωνικό δίκτυο του Twitter για τις αναρτήσεις των χρηστών για τα πολιτικά θέματα. Με αυτό το μοντέλο μπορεί να ενταχθεί και σε άλλες πηγές που αναρτούν γνώμες οι χρήστες, όπως μέσα στο YouTube και σε κριτικές προϊόντων στο Amazon.

Βιβλιογραφία

[01] Adam Sharp 2012, Wednesday, August 1, 2012-last update, A new barometer for the election. Available: https://blog.twitter.com/2012/a-new-barometer-for-the-election [2014, December,Friday 12].

[02] Aue, A. & Gamon, M. 2005, "Customizing sentiment classifiers to new domains: A case study", Proceedings of recent advances in natural language processing (RANLP)Citeseer, , pp. 2.1.

[03] Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L. & Hughes, M. 2013, "Sentiment analysis of political tweets: towards an accurate classifier", .

[04] Bermingham, A. & Smeaton, A.F. 2011, "On using Twitter to monitor political sentiment and predict election results", .

[05] Choy, M., Cheong, M.L., Laik, M.N. & Shung, K.P. 2011, "A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction", arXiv preprint arXiv:1108.5520, .

[06] D'alessio, D. 1997, "Use of the World Wide Web in the 1996 US election", Electoral Studies, vol. 16, no. 4, pp. 489-500.

[07] Das, S.R. & Chen, M.Y. 2001, "Yahoo! for Amazon: Sentiment parsing from small talk on the web", EFA 2001 Barcelona Meetings.

[08] Dave, K., Lawrence, S. & Pennock, D.M. 2003, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews", Proceedings of the 12th international conference on World Wide WebACM, , pp. 519.

[09] de Kok, D. & Brouwer, H. 2011, "Natural language processing for the working programmer", .

[10] Dey, L. & Haque, S.M. 2009, "Opinion mining from noisy text data", International Journal on Document Analysis and Recognition (IJDAR), vol. 12, no. 3, pp. 205-226.

[11] Esuli, A. & Sebastiani, F. 2006, "Sentiwordnet: A publicly available lexical resource for opinion mining", Proceedings of LREC, pp. 417.

[12] Gayo-Avello, D. 2012, "" I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data", arXiv preprint arXiv:1204.6441, .

[13] Gibson, R.K. & McAllister, I. 2011, "Do online election campaigns win votes? The 2007 Australian “YouTube” election", Political Communication, vol. 28, no. 2, pp. 227-244.

[14] Gibson, R.K. & McAllister, I. 2006, "Does cyber-campaigning win votes? Online communication in the 2004 Australian election", Journal of elections, public opinion and parties, vol. 16, no. 3, pp. 243-263.

[15] Golbeck, J., Grimes, J.M. & Rogers, A. 2010, "Twitter use by the US Congress", Journal of the American Society for Information Science and Technology, vol. 61, no. 8, pp. 1612-1621.

[16] Gregory Brail 2011, APR 08, 2011-last update, Why you need OAuth for your API or APP. Available: https://blog.apigee.com/detail/why_you_need_oauth_for_your_api_or_app [20 14, December,Friday 12].

[17] Hatzivassiloglou, V. & McKeown, K.R. 1997, "Predicting the semantic orientation of adjectives", Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational LinguisticsAssociation for Computational Linguistics, , pp. 174.

[18] Hatzivassiloglou, V. & Wiebe, J.M. 2000, "Effects of adjective orientation and gradability on sentence subjectivity", Proceedings of the 18th conference on Computational linguistics-Volume 1Association for Computational Linguistics, , pp. 299.

[19] Hu, M. & Liu, B. 2004, "Mining and summarizing customer reviews", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data miningACM, , pp. 168.

[20] Hu, M. & Liu, B. 2004, "Mining opinion features in customer reviews", AAAI, pp. 755.

[21] Kanayama, H. & Nasukawa, T. 2006, "Fully automatic lexicon expansion for domain-oriented sentiment analysis", Proceedings of the 2006 Conference on Empirical Methods in Natural Language ProcessingAssociation for Computational Linguistics, , pp. 355.

[22] Kruikemeier, S. 2014, "How political candidates use Twitter and the impact on votes", Computers in Human Behavior, vol. 34, pp. 131-139.

[23] Lietz, H., Wagner, C., Bleier, A. & Strohmaier, M. 2014, "When politicians talk: Assessing online conversational practices of political parties on twitter", International AAAI Conference on Weblogs and Social Media (ICWSM2014), Ann Arbor, MI, USA.

[24] Liu, B. 2010, "Sentiment analysis and subjectivity", Handbook of natural language processing, vol. 2, pp. 627-666.

[25] Liu, B. 2012, Sentiment analysis and opinion mining, Morgan & Claypool, San Rafael.

[26] Martha T. Moore 2012, 8/1/2012-last update, Twitter index tracks sentiment on Obama, Romney. Available: http://usatoday30.usatoday.com/news/politics/story/2012-08-01/twitter-political-index/56649678/1[2014, December,Friday 12].

[27] Michael Fauscette 2012, February 23, 2012-last update, Twitter, Politics, and Sentiment Analysis. Available: http://www.enterpriseirregulars.com/46060/twitter-politics-and-sentiment-analysis/ [2014, December,Friday 12].

[28] O'Connor, B., Balasubramanyan, R., Routledge, B.R. & Smith, N.A. 2010, "From tweets to polls: Linking text sentiment to public opinion time series.", ICWSM, vol. 11, pp. 122-129.

[29] Pak, A. & Paroubek, P. 2010, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining.", LREC.

[30] Pang, B. & Lee, L. 2008, "Opinion mining and sentiment analysis", Foundations and trends in information retrieval, vol. 2, no. 1-2, pp. 1-135.

[31] Pang, B. & Lee, L. 2004, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", Proceedings of the 42nd annual meeting on

Association for Computational LinguisticsAssociation for Computational Linguistics, , pp. 271.

[32] Pang, B., Lee, L. & Vaithyanathan, S. 2002, "Thumbs up?: sentiment classification using machine learning techniques", Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10Association for Computational Linguistics, , pp. 79.

[33] Park, H.M. & Perry, J.L. 2008, "Do campaign web sites really matter in electoral civic engagement? Empirical evidence from the 2004 post-election internet tracking survey", Social Science Computer Review, vol. 26, no. 2, pp. 190-212.

[34] Qiu, G., Liu, B., Bu, J. & Chen, C. 2009, "Expanding Domain Sentiment Lexicon through Double Propagation.", IJCAI, pp. 1199.

[35] Smrž, P. 2006, "Using WordNet for opinion mining", Proceedings of the Third International WordNet ConferenceMasaryk University}, , pp. 333.

[36] Sundar, S.S., Kalyanaraman, S. & Brown, J. 2003, "Explicating Web Site interactivity impression formation effects in political campaign sites", Communication research, vol. 30, no. 1, pp. 30-59.

[37] Tedesco, J.C. 2007, "Examining Internet interactivity effects on young adult political information efficacy", American Behavioral Scientist, vol. 50, no. 9, pp. 1183-1194.

[38] Tong, R.M. 2001, "An operational system for detecting and tracking opinions in on-line discussion", Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification, pp. 6.

[39] Tumasjan, A., Sprenger, T.O., Sandner, P.G. & Welpe, I.M. 2010, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.", ICWSM, vol. 10, pp. 178-185.

[40] Turney, P.D. 2002, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews", Proceedings of the 40th annual meeting on

association for computational linguisticsAssociation for Computational Linguistics, , pp. 417.

[41] Turney, P.D. & Littman, M.L. 2003, "Measuring praise and criticism: Inference of semantic orientation from association", ACM Transactions on Information Systems (TOIS), vol. 21, no. 4, pp. 315-346.

[42] Wang, H., Can, D., Kazemzadeh, A., Bar, F. & Narayanan, S. 2012, "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle", Proceedings of the ACL 2012 System DemonstrationsAssociation for Computational Linguistics, , pp. 115.

[43] Wiebe, J.M., Bruce, R.F. & O'Hara, T.P. 1999, "Development and use of a gold-standard data set for subjectivity classifications", Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational LinguisticsAssociation for Computational Linguistics, , pp. 246.

[44] Yu, H. & Hatzivassiloglou, V. 2003, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences", Proceedings of the 2003 conference on Empirical methods in natural language processingAssociation for Computational Linguistics, , pp. 129.

[45] Zhai, Z., Liu, B., Xu, H. & Jia, P. 2010, "Grouping product features using semi-supervised learning with soft-constraints", Proceedings of the 23rd International Conference on Computational LinguisticsAssociation for Computational Linguistics, , pp. 1272.

Παράρτημα Α

Πως γίνετε η αναζήτη του αλγορίθμου από τα tweets της Νέας Δημοκρατίας

```
keywords_pos<-t(positive)

keywords_found <- sapply(keywords_pos, regexpr, ND_all_tweets$text, ignore.case=TRUE)

keywords_coordinates<-which( keywords_found > -1, arr.ind=T )

rows_with_keywords <-as.vector(keywords_coordinates[,1])

cols_with_keywords <-as.vector(keywords_coordinates[,2])

if_case_nd <- NA

for (i in 1:length(rows_with_keywords)){

tmp<-str_split(ND_all_tweets[ rows_with_keywords[i],  ]$text,' ')

tmp<-unlist(tmp)

if_word <-which(apply(t(tmp), 2, function(x) any(grepl("δεν", x, ignore.case=TRUE))))

if(length(if_word)!=0){

pos_word_tmp<-
which(apply(t(tmp),2,function(x)any(grepl(keywords_pos[cols_with_keywords[i]],x,ignore.case=
TRUE))))

if(length(pos_word_tmp)!=0){

if(pos_word_tmp-if_word==2 | pos_word_tmp-if_word==3)

 {if_case_nd <- c(if_case_nd,rows_with_keywords[i])}}}}
```

```
if_case_nd <- if_case_nd[-1]

rows_with_keywords <- rows_with_keywords[!(rows_with_keywords%in%if_case_nd)]

pos_to_negs_1 <- if_case_nd

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

keywords_pos<-t(positive2)

keywords_found <- sapply(keywords_pos, regexpr, ND_all_tweets$text, ignore.case=TRUE)

keywords_coordinates<-which( keywords_found > -1, arr.ind=T )

rows_with_keywords2<-as.vector(keywords_coordinates[,1])

cols_with_keywords2 <-as.vector(keywords_coordinates[,2])

if_case_nd <- NA

for (i in 1:length(rows_with_keywords2)){

 tmp<-str_split(ND_all_tweets[ rows_with_keywords2[i],  ]$text,' ')

 tmp<-unlist(tmp)

 if_word<-which(apply(t(tmp), 2, function(x) any(grepl("δεν", x, ignore.case=TRUE))))

 if(length(if_word)!=0){



  pos_word_tmp<-
which(apply(t(tmp),2,function(x)any(grepl(keywords_pos[cols_with_keywords2[i]],x,ignore.case
=TRUE))))

  if(length(pos_word_tmp)!=0){
```

```
   if(pos_word_tmp-if_word==2 | pos_word_tmp-if_word==3)

    {if_case_nd <- c(if_case_nd,rows_with_keywords2[i])}  }}

if_case_nd <- if_case_nd[-1]

rows_with_keywords2 <- rows_with_keywords2[!(rows_with_keywords2%in%if_case_nd)]

pos_to_negs_2 <- if_case_nd

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

keywords_neg<-t(negative)

keywords_found <- sapply(keywords_neg, regexpr, ND_all_tweets$text, ignore.case=TRUE)

keywords_coordinates<-which( keywords_found > -1, arr.ind=T )

rows_with_keywords_negs<-as.vector(keywords_coordinates[,1])

 cols_with_keywords_negs<-as.vector(keywords_coordinates[,2])

if_case_nd <- NA

for (i in 1:length(rows_with_keywords_negs)){

 tmp<-str_split(ND_all_tweets[ rows_with_keywords_negs[i],  ]$text,' ')

 tmp<-unlist(tmp)

 if_word  <-which(apply(t(tmp), 2, function(x) any(grepl("δεν", x, ignore.case=TRUE))))

 if(length(if_word)!=0){

 pos_word_tmp<-
which(apply(t(tmp),2,function(x)any(grepl(keywords_neg[cols_with_keywords_negs[i]],x,ignore.
case=TRUE))))
```

```
  if(length(pos_word_tmp)!=0){

  if(pos_word_tmp-if_word==2 | pos_word_tmp-if_word==3)

   {if_case_nd <- c(if_case_nd,rows_with_keywords_negs[i])}} }}

if_case_nd <- if_case_nd[-1]

rows_with_keywords_negs<-
rows_with_keywords_negs[!(rows_with_keywords_negs%in%if_case_nd)]

negs_to_pos1 <- if_case_nd

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

keywords_neg<-t(negative2)

keywords_found <- sapply(keywords_neg, regexpr, ND_all_tweets$text, ignore.case=TRUE)

keywords_coordinates<-which( keywords_found > -1, arr.ind=T )

rows_with_keywords_negs2<-as.vector(keywords_coordinates[,1])

cols_with_keywords2 <-as.vector(keywords_coordinates[,2])

if_case_nd <- NA

for (i in 1:length(rows_with_keywords_negs2)){

 tmp<-str_split(ND_all_tweets[ rows_with_keywords_negs2[i],  ]$text,' ')

 tmp<-unlist(tmp)

 if_word     <-which(apply(t(tmp), 2, function(x) any(grepl("δεν", x, ignore.case=TRUE))))

 if(length(if_word)!=0){
```

```
 pos_word_tmp<-
which(apply(t(tmp),2,function(x)any(grepl(keywords_neg[cols_with_keywords2[i]],x,ignore.case
=TRUE))))

     if(length(pos_word_tmp)!=0){

   if(pos_word_tmp-if_word==2|pos_word_tmp-if_word==3){if_case_nd<-
c(if_case_nd,rows_with_keywords_negs2[i])}   } }}

if_case_nd <- if_case_nd[-1]

rows_with_keywords_negs2<-
rows_with_keywords_negs2[!(rows_with_keywords_negs2%in%if_case_nd)]

negs_to_pos2 <- if_case_nd

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

rows_with_pos_keywords_nd<-
c(rows_with_keywords,rows_with_keywords2,negs_to_pos1,negs_to_pos2)

new_df_tweets_pos_words_ND<-
cbind(ND_all_tweets[unique(rows_with_pos_keywords_nd),],Class2='Positive')

rows_with_neg_keywords_nd<-
c(rows_with_keywords_negs,rows_with_keywords_negs2,pos_to_negs_1,pos_to_negs_2)

new_df_tweets_neg_words_ND<-
cbind(ND_all_tweets[unique(rows_with_neg_keywords_nd),],Class2='Negative')

#----------------------------------  POS NEG NEU ------------------------------------------------

rows_with_pos_keywords_nd_backup <- sort(rows_with_pos_keywords_nd)

rows_with_neg_keywords_nd_backup <- sort(rows_with_neg_keywords_nd)
```

```
rows_with_pos_keywords_nd <- sort(rows_with_pos_keywords_nd)

rows_with_neg_keywords_nd <- sort(rows_with_neg_keywords_nd)

nd_intersect<-intersect(names(table(rows_with_pos_keywords_nd))                    ,
names(table(rows_with_neg_keywords_nd))    )

rows_with_neu_keywords_nd_red <- NULL

for (i in 1:length(nd_intersect)){

pos_coord <- which(names(table(rows_with_pos_keywords_nd_backup))==nd_intersect[i])

neg_coord <- which(names(table(rows_with_neg_keywords_nd_backup))==nd_intersect[i])

if(table(rows_with_pos_keywords_nd_backup)[pos_coord]<table(rows_with_neg_keywords_nd_b
ackup)[neg_coord]){

   rows_with_pos_keywords_nd<-
rows_with_pos_keywords_nd[!(rows_with_pos_keywords_nd%in%as.numeric(names(table(row
s_with_pos_keywords_nd_backup)[pos_coord])))] }

if(table(rows_with_pos_keywords_nd_backup)[pos_coord]>table(rows_with_neg_keywords_nd_b
ackup)[neg_coord]){

 rows_with_neg_keywords_nd<-
rows_with_neg_keywords_nd[!(rows_with_neg_keywords_nd%in%as.numeric(names(table(row
s_with_neg_keywords_nd_backup)[neg_coord])))]}

if(table(rows_with_pos_keywords_nd_backup)[pos_coord]==table(rows_with_neg_keywords_nd_
backup)[neg_coord]){

rows_with_neu_keywords_nd_red<c(rows_with_neu_keywords_nd_red,rows_with_pos_keywords
_nd_backup[(rows_with_pos_keywords_nd_backup%in%as.numeric(names(table(rows_with_po
s_keywords_nd_backup)[pos_coord])))])
```

```
 rows_with_pos_keywords_nd<-
rows_with_pos_keywords_nd[!(rows_with_pos_keywords_nd%in%as.numeric(names(table(rows_with_pos_keywords_nd_backup)[pos_coord])))]

rows_with_neg_keywords_nd<-
rows_with_neg_keywords_nd[!(rows_with_neg_keywords_nd%in%as.numeric(names(table(rows_with_neg_keywords_nd_backup)[neg_coord])))]}}

new_df_tweets_neg_words_ND_red<cbind(ND_all_tweets[unique(rows_with_neg_keywords_nd),],Class2='Negative')

new_df_tweets_pos_words_ND_red<cbind(ND_all_tweets[unique(rows_with_pos_keywords_nd),],Class2='Positive')

new_df_tweets_neu_words_ND_red<cbind(ND_all_tweets[unique(rows_with_neu_keywords_nd_red),],Class2='Neutral')

pos_neg_tweets_filt<rbind(new_df_tweets_neg_words_ND_red,new_df_tweets_pos_words_ND_red,new_df_tweets_neu_words_ND_red)

ggplot(pos_neg_tweets_filt,aes(x=created_format_date,fill=Class2))+geom_histogram(binwidth=.5,position='dodge') +  scale_x_date(breaks="1 day", labels=date_format("%Y_%m_%d"))+xlab('ND --- Tweets with positive/negative/neutral words')+   theme(axis.text.x = element_text(angle = 90, hjust = 1,size=10))+ scale_fill_discrete(name = "Frequency")

#-----------  ND ---------------------------------------

pos_neg_neu_ND<-aggregate(Class2~created_format_date, data=pos_neg_tweets_filt, FUN=paste0)

pos_neg_neu_split_class_ND <- data.frame(Date=as.Date(rep(NA,nrow(pos_neg_neu_ND))),


Opinion=(rep(NA,nrow(pos_neg_neu_ND))),Party=as.character(rep(NA,nrow(pos_neg_neu_ND)),stringsAsFactors=FALSE)
```

```
pos_neg_neu_split_ND <- NULL

for (i in 1:nrow(pos_neg_neu_ND)){

  pos_neg_neu_split_class_ND[i,1] <- as.Date(pos_neg_neu_ND[[i,1]])

 pos_neg_neu_split_class_ND[i,3] <- 'ND'

   tmp<-
t(data.frame(vars=as.vector((table(as.data.frame(pos_neg_neu_ND[[i,2]]))/dim(as.data.frame(po
s_neg_neu_ND[[i,2]]))[1])))

 names_tmp <- as.vector(names(table(as.data.frame(pos_neg_neu_ND[[i,2]]))))

   pos_neg_neu_split_ND <- rbind (pos_neg_neu_split_ND , data.frame( Party = 'ND' , date =
as.Date(pos_neg_neu_ND[[i,1]]) ,


as.data.frame(table(as.data.frame(pos_neg_neu_ND[[i,2]]))/dim(as.data.frame(pos_neg_neu_ND[[
i,2]]))[1] ) ) )

  if(names_tmp[which.max(tmp)] =='Negative')

 { pos_neg_neu_split_class_ND[i,2] <- -1

 }else if(names_tmp[which.max(tmp)] =='Neutral'){

  pos_neg_neu_split_class_ND[i,2] <- 0

 }else {  pos_neg_neu_split_class_ND[i,2] <- 1 } }

##########
```