ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάλυση Ελληνικής Κοινής Γνώμης στο Twitter Βασισμένη σε Λεξικό

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αχιλλεύς Ν. Σφακιανάκης

Επιβλέπων : Στέφανος Κόλλιας

Καθηγητής Ε.Μ.Π.

Αθήνα, Μάρτιος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Ανάλυση Ελληνικής Κοινής Γνώμης στο Twitter Βασισμένη σε Λεξικό

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Αχιλλεύς Ν. Σφακιανάκης

Επιβλέπων : Στέφανος Κόλλιας

Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 28η Μαρτίου 2016.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Στέφανος Κόλλιας	Κωνσταντίνος Καρπούζης	Γεώργιος Στάμου
Καθηγητής ΕΜΠ	Διευθυντής Ερευνών Ε.Π.Ι.Σ.Ε.Υ – Ε.Μ.Π	Επίκουρος Καθηγητής ΕΜΠ

Αθήνα, Μάρτιος 2016

(Υπογραφή)

.....

Αχιλλεύς Ν. Σφακιανάκης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Η τεράστια ανάπτυξη των διαδικτυακών εφαρμογών, όπως είναι τα μικροιστολόγια, τα φόρουμ και τα μέσα κοινωνικής δικτύωσης, έχει οδηγήσει στη δημιουργία ενός τεράστιου όγκου διαδικτυακών κειμένων. Η ανάγκη χειρισμού όλων αυτών των δεδομένων μέσω αυτοματοποιημένων τεχνικών, για την κατανόηση των συναισθημάτων, των επιθυμιών και των προθέσεων των χρηστών, προκειμένου να διευκολυνθεί η διαδικασία λήψης αποφάσεων, έχει οδηγήσει στη ραγδαία εξέλιξη του πεδίου της Ανάλυσης Συναισθήματος ή ισοδύναμα Εξόρυξης Γνώμης.

Η παρούσα εργασία εστιάζει το ενδιαφέρον της στο Twitter, το οποίο είναι μια πλατφόρμα κοινωνικής δικτύωσης που επιτρέπει στους χρήστες της να δημοσιεύουν ενημερώσεις και να ανταλλάσσουν σύντομα μηνύματα, τα οποία ονομάζονται tweets. Συγκεκριμένα, μελετώνται tweets Ελλήνων χρηστών που αφορούν θέματα κοινωνικής και πολιτικής φύσεως και αναπτύσσεται ένα αυτόματο σύστημα Ανάλυσης Συναισθήματος, το οποίο βασίζεται σε λεξικό συναισθήματος και ταξινομεί τα tweets σε μία απο τις τρείς πιθανές κλάσεις (αρνητικό, θετικό ή ουδέτερο), ανάλογα με την πολικότητα του συναίσθηματος που εκφράζουν. Το λεξικό δημιουργήθηκε με την προσθήκη τμήματος των συχνότερα εμφανιζόμενων λέξεων που εντοπίστηκαν σε ένα σύνολο 20000 tweets, σε ένα ήδη υπάρχον, ελεύθερα διαθέσιμο λεξικό συναισθήματος. Το σύστημα αποτελείται από τρία βασικά στάδια: το στάδιο της προεπεξεργασίας, στο οποίο τα tweets φιλτράρονται για να αφαιρεθεί όλη η άχρηστη πληροφορία, το στάδιο του μετασχηματισμού, κατά το οποίο οι λέξεις των φιλτραρισμένων tweets μετασχηματίζονται σε κατάλληλη μορφή για να μπορούν να εντοπιστούν στο λεξικό και τέλος το στάδιο της ανίχνευσης άρνησης, στο οποίο γίνεται χρήση κάποιων γλωσσολογικών κανόνων για να καθοριστεί αν κάθε λέξη που εντοπίστηκε εντάσσεται σε σχήμα άρνησης και συνεπώς πρέπει να αντιστραφεί η πολικότητά της.

Για την αξιολόγηση του συστήματος επιλέχθηκε ένα σύνολο tweets που αφορούσαν τα αποτελέσματα του δευτέρου γύρου των εσωκομματικών εκλογών της Νέας Δημοκρατίας, που ανακοινώθηκαν την 11η Ιανουαρίου 2016. Ο λόγος επιλογής του συγκεκριμένου θέματος, ήταν οτι οι εκλογικές διαδικασίες μονοπωλούν το ενδιαφέρον των χρηστών στο διαδίκτυο, τα σχόλια των οποίων σε τέτοιες περιπτώσεις έχουν πλούσιο συναισθηματικό περιεχόμενο, αφού εκφράζουν με ιδιαιτέρως καυστικό τρόπο είτε την επιδοκιμασία τους είτε την αποδοκιμασία τους προς τα εκλογικά αποτελέσματα.

Λέξεις κλειδιά: Ανάλυση συναισθήματος, εξόρυξη γνώμης, Twitter, μέσα κοινωνικής δικτύωσης, λεξικό συναισθήματος, συμφωνία μεταξύ κριτών

Abstract

The great development of internet applications, such as microblogs, forums and social networks has resulted into the generation of huge volumes of online textual data. The need for automated manipulation of such data in order to understand the feelings, the desires and the intentions of the users and ultimately facilitate the decision-making process has given rise to the field of Sentiment Analysis or Opinion Mining.

This thesis focuses on Twitter, which is an online social network platform that allows users to post updates and send short messages, called tweets. More specifically, tweets of Greek users that refer to political or social issues are taken into account and it is implemented an automatic system for Sentiment Analysis, that is based on a sentiment lexicon and classifies tweets into one of the three possible categories (negative, neutral, or positive), depending on the polarity of the sentiment expressed. The lexicon was created by adding part of the most frequent words that were detected in a data set of 20.000 tweets, into a pre-built greek sentiment lexicon that is publicly available. The system consists of three main stages: the pre-processing stage, in which tweets are filtered to remove noise, the transformation stage, in which the words of the filtered tweets are transformed into a proper form so that they can be traced in the lexicon and finally the negation-detection stage, in which linguistic rules are used to determine for each word of the transformed tweets, that it was found in the lexicon, if it is included in a negation scheme and so its polarity must be inverted.

To evaluate the system we used a data set concerning the results of the intraparty elections of New Democracy, that were announced on the 11th of January 2016. This choice was made because elections monopolize the interest of internet users, whose comments, in such occasions, have strong sentiments, because they express either their approval or their disapproval to the elective results in a penetrating way.

Keywords: Sentiment analysis, opinion mining, Twitter, social networks, sentiment lexicon, inter-rater agreement

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον καθηγητή Ε.Μ.Π κύριο Στέφανο Κόλλια για την εμπιστοσύνη που μου έδειξε και τη δυνατότητα που μου έδωσε να εκπονήσω τη διπλωματική μου εργασία σε ένα τόσο ενδιαφέρον θέμα, καθώς και για τη βοήθεια που μου έδωσε όποτε κρίθηκε απαραίτητο.

Στη συνέχεια, θα ήθελα να ευχαριστήσω το Διευθυντή Ερευνών Ε.Π.Ι.Σ.Ε.Υ - Ε.Μ.Π κύριο Κώστα Καρπούζη για την καθοδήγησή του και τις πολύτιμες και πάντα εύστοχες συμβουλές του, χωρίς τις οποίες δεν θα ήταν δυνατή η ολοκλήρωση αυτής της εργασίας.

Τέλος, δεν θα μπορούσα να μην ευχαριστήσω τους φίλους μου από τη σχολή με τους οποίους περάσαμε μαζί εύκολες και δύσκολες στιγμές αυτά τα έξι χρόνια, και περισσότερο από όλους την οικογένειά μου, που στέκεται δίπλα μου όλα αυτά τα χρόνια και μου παρέχει αγάπη και στήριξη σε κάθε μου βήμα.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1 Εισαγωγή

Η τεράστια ανάπτυξη των διαδικτυακών εφαρμογών, ιδιαίτερα κατά τη διάρκεια της τελευταίας δεκαετίας, όπως είναι τα μικρο-ιστολόγια, τα φόρουμ και τα μέσα κοινωνικής δικτύωσης, έχει οδηγήσει στη δημιουργία ενός τεράστιου όγκου διαδικτυακών κειμένων. Μέσω των παραπάνω εφαρμογών οι άνθρωποι συζητούν, σχολιάζουν διάφορα θέματα της επικαιρότητας, γράφουν κριτικές και προτάσεις για προϊόντα και υπηρεσίες που έχουν χρησιμοποιήσει και εκφράζουν τις πολιτικές και θρησκευτικές τους πεποιθήσεις. Τα δεδομένα που παράγονται από τους χρήστες «κουβαλάνε» πολύτιμες πληροφορίες καθώς αποτυπώνουν τα συναισθήματά τους, τις επιθυμίες τους, τις προθέσεις τους, αλλά και τις αντιλήψεις τους για διάφορα θέματα. Η ανάγκη χειρισμού όλων αυτών των δεδομένων μέσω αυτοματοποιημένων τεχνικών, έχει οδηγήσει στην ραγδαία εξέλιξη του χώρου της Ανάλυσης Συναισθήματος (Sentiment Analysis) ή ισοδύναμα Εξόρυξης Γνώμης (Opinion Mining), η οποία προσπαθεί να αξιοποιήσει το τι πιστεύουν οι άνθρωποι για να διευκολύνει τη διαδικασία λήψης αποφάσεων (decision-making process) σε διάφορα επίπεδα.

Στις μέρες μας η Ανάλυση Συναισθήματος χρησιμοποιείται από εταιρείες και οργανισμούς για να κατανοήσουν τις απόψεις των καταναλωτών για τα προϊόντα και τις υπηρεσίες τους, με απώτερο στόχο τον καθορισμό της στρατηγικής μάρκετινγκ και τη βελτίωση της εξυπηρέτησης πελατών (customer service) (business intelligence - business analytics), από κυβερνήσεις για να αντιληφθούν τις προθέσεις των ψηφοφόρων στα πλαίσια μιας εκλογικής διαδικασίας ή τη στάση των πολιτών σχετικά με τις πολιτικές που ακολουθούν (government intelligence), αλλά και από μεγάλες διαφημιστικές εταιρείες για την αξιολόγηση των διαφημιστικών εκστρατειών τους. Στα παραπάνω έρχεται να προστεθεί η χρήση σε συστήματα συστάσεων (recommendation systems) για την αποφυγή προτάσεων που έχουν λάβει αρνητικές κριτικές, καθώς και στον εντοπισμό ανεπιθύμητης αλληλογραφίας (opinion spam detection) η οποία στοχεύει στην παραπλάνηση των αναγνωστών ή αυτόματων συστημάτων [1]. Όπως μπορεί κανείς να φανταστεί λοιπόν, οι πιθανές εφαρμογές της Ανάλυσης Συναισθήματος είναι ουσιαστικά απεριόριστες, καθιστώντας την ως ένα από τα πιο ενδιαφέροντα και γεμάτα προοπτικές πεδία, τόσο σε ερευνητικό-ακαδημαϊκό επίπεδο όσο και στον ευρύτερο χώρο των επιχειρήσεων.

1.1 Σκοπός της διπλωματικής

Παρά το γεγονός οτι γίνεται όλο και πιο εκτεταμένη έρευνα στο χώρο της Ανάλυσης Συναισθήματος, ο κύριος όγκος της περιορίζεται στην ανάλυση κειμένων γραμμένων σε συγκεκριμένες γλώσσες, όπως τα αγγλικά και τα ισπανικά. Η σχετικά περιορισμένη έρευνα σε ό,τι αφορά την ελληνική γλώσσα

συνδυασμένη με τις εγγενείς δυσκολίες και ιδιαιτερότητες της ελληνικής, αλλά και την έντονη δραστηριότητα των Ελλήνων στα μέσα κοινωνικής δικτύωσης[1,2], μας έδωσαν το κίνητρο να υλοποιήσουμε ένα σύστημα εξόρυξης γνώμης, το οποίο θα εντοπίζει την πολικότητα των μηνυμάτων (των λεγόμενων tweets) που δημοσιεύουν οι Έλληνες χρήστες στη διάσημη πλατφόρμα κοινωνικής δικτύωσης *Twitter*, και θα εστιάζει κυρίως σε θέματα που αφορούν την ευρύτερη πολιτική και κοινωνική ζωή. Το σύστημα αυτό θα βασίζεται σε ένα λεξικό συναισθήματος, το οποίο συνδυαζόμενο με τα στάδια της προεπεξεργασίας και του μετασχηματισμού των δεδομένων, αλλά και με κάποιους απλούς γλωσσολογικούς κανόνες για το χειρισμό της άρνησης, θα ταξινομεί το κάθε μήνυμα σε μία από τις τρείς πιθανές κατηγορίες (θετικό, αρνητικό ή ουδέτερο) ανάλογα με την πολικότητα του συναισθήματος που εκφράζει.

1.2 Οργάνωση κειμένου

Η οργάνωση της διπλωματικής απο εδώ και πέρα έχει ως εξής:

Στο Κεφάλαιο 2, εστιάζουμε στο θεωρητικό υπόβαθρο της Ανάλυσης Συναισθήματος, δηλαδή στα προβλήματα και τις προκλήσεις που παρουσιάζονται γενικότερα, αλλά και ειδικότερα στα μικρο-ιστολόγια όπως το Twitter, καθώς και στους πιθανούς τρόπους με τους οποίους προσεγγίζεται το πρόβλημα. Αναφερόμαστε επίσης στις διάφορες μετρικές που χρησιμοποιούνται για την αξιολόγηση συστημάτων Ανάλυσης Συναισθήματος.

Στο Κεφάλαιο 3, εξηγούμε αναλυτικά τα διάφορα στάδια υλοποίησης του συστήματός μας. Αναφερόμαστε στις τεχνικές προεπεξεργασίας και μετασχηματισμού που εφαρμόσαμε στα δεδομένα μας, στον τρόπο με τον οποίο επεκτείναμε και τροποποιήσαμε το ήδη διαθέσιμο λεξικό καθώς και στους γλωσσολογικούς κανόνες που εφαρμόσαμε για τον εντοπισμό και χειρισμό της άρνησης.

Στο Κεφάλαιο 4, προχωρούμε σε αξιολόγηση του συστήματός μας, χρησιμοποιώντας τις μετρικές που επεξηγήσαμε στο Κεφάλαιο 2 και προσπαθούμε να κατανοήσουμε και να ερμηνεύσουμε τα αποτελέσματα προκειμένου να δούμε σε ποιά σημεία το σύστημα που υλοποιήσαμε είχε την επιθυμητή απόδοση και σε ποιά χρήζει βελτίωσης.

Τέλος στο Κεφάλαιο 5, παρουσιάζουμε τα συμπεράσματα της εργασίας καθώς και πιθανές τροποποιήσεις και προσθήκες που μπορούν να γίνουν στο εν λόγω σύστημα για βελτίωση της απόδοσής του.

---

[1] http://www.ekathimerini.com/201228/article/ekathimerini/community/half-of-greeks-engage-in-social-media

[2] http://www.statista.com/statistics/384378/social-network-penetration-in-greece/

Κεφάλαιο 2 Θεωρητικό Υπόβαθρο

2.1 Ανάλυση Συναισθήματος

Ας δούμε καταρχάς έναν τυπικό ορισμό της Ανάλυσης Συναισθήματος. Σύμφωνα με το λεξικό της Οξφόρδης[1], ως Ανάλυση Συναισθήματος ορίζεται:

«Η διαδικασία της ταυτοποίησης και κατηγοριοποίησης των απόψεων που εκφράζονται σε ένα κείμενο, ειδικά με σκοπό το να καθοριστεί εάν η άποψη του συγγραφέα σχετικά με ένα συγκεκριμένο θέμα, προϊόν κλπ είναι θετική, αρνητική ή ουδέτερη».

Ο Seth Grimes[2], ένας από τους κορυφαίους industry analysts σε παγκόσμιο επίπεδο, δίνει τον εξής εναλλακτικό ορισμό:

«Ως Ανάλυση Συναισθήματος ορίζουμε ένα σύνολο (συστηματικών) μεθόδων, οι οποίες τυπικά (αλλά όχι πάντα) υλοποιούνται σε λογισμικό υπολογιστή, και ανιχνεύουν, μετράνε και αξιοποιούν στάσεις, συμπεριφορές, γνώμες και συναισθήματα σε online κοινωνικές και επιχειρηματικές πηγές πληροφορίας».

Η Ανάλυση Συναισθήματος μπορεί να εφαρμοσθεί σε τρία επίπεδα:

- Σε επίπεδο εγγράφου (document-level SA), όπου υποθέτει οτι κάθε κείμενο εκφράζει μία και μοναδική άποψη για ένα συγκεκριμένο θέμα ή αντικείμενο.
- Σε επίπεδο πρότασης (sentence-level SA), όπου κάθε έγγραφο χωρίζεται σε προτάσεις, υποθέτοντας οτι καθεμιά από αυτές εκφράζει μια ξεχωριστή άποψη.
- Σε επίπεδο χαρακτηριστικών (feature-level SA), όπου κάθε έγγραφο χωρίζεται σε προτάσεις και κάθε πρόταση σε πολωμένες φράσεις που αντιστοιχούν σε ένα συγκεκριμένο χαρακτηριστικό του πραγματευόμενου θέματος ή αντικειμένου.

Δεδομένου οτι στην παρούσα διπλωματική υλοποιούμε ένα σύστημα Ανάλυσης Συναισθήματος στον ιστότοπο κοινωνικής δικτύωσης Twitter, εφαρμόζουμε στην ουσία Ανάλυση Συναισθήματος σε επίπεδο εγγράφου, αφού θεωρούμε οτι κάθε tweet εκφράζει μία και μοναδική άποψη.

2.2 Προβλήματα και προκλήσεις

Καθώς η ανάλυση συναισθήματος χρησιμοποιείται από όλο και περισσότερες

[1] https://www.socialmediaexplorer.com/social-media-monitoring/sentiment-analysis/

[2] http://altaplana.com/grimes.html

επιχειρήσεις, οργανισμούς, αλλά και ιδιώτες δημιουργείται η ανάγκη υλοποίησης όσο το δυνατόν πιο αξιόπιστων συστημάτων. Δεδομένης όμως της πολυπλοκότητας του γραπτού λόγου, αλλά και του γεγονότος οτι πολλές φορές είναι δύσκολο ακόμα και για τον ίδιο τον άνθρωπο να ερμηνεύσει το γραπτό λόγο και την άποψη του συγγραφέα, πόσο εύκολο είναι να διδάξουμε μια μηχανή να κατανοεί και να ερμηνεύει τον ανθρώπινο γραπτό λόγο και ποιά είναι τα εμπόδια που πρέπει να ξεπεράσουμε; Τα βασικά προβλήματα που έχουν να αντιμετωπίσουν όλα τα συστήματα Ανάλυσης Συναισθήματος είναι τα παρακάτω:

- *Αμφισημία – σημασιολογικό πλαίσιο*

  Μια λέξη μπορεί να έχει πολλαπλές έννοιες και μια φράση μπορεί να επιδέχεται πολλαπλών ερμηνειών, φαινόμενο γνωστό ως αμφισημία (*ambiguity*) ή πολυσημία. Το πώς τις ερμηνεύουμε κάθε φορά εξαρτάται από τα συμφραζόμενα, δηλαδή το εκάστοτε σημασιολογικό πλαίσιο(*context*). Για παράδειγμα η λέξη διαβήτης μπορεί να αναφέρεται είτε στο γεωμετρικό όργανο, είτε στην ιατρική ασθένεια, ενώ η φράση «απλά διαβάστε το βιβλίο» είναι θετική αν αναφέρεται σε κριτική βιβλίου, είναι όμως αρνητική αν αναφέρεται σε κριτική ταινίας. Μάλιστα η πλήρης έλλειψη περιεχόμενου ενδέχεται να καθιστά αδύνατη την κατηγοριοποίηση, αφού για παράδειγμα στη φράση «Η νέα ταινία ήταν τόσο καλή όσο και η προηγούμενη», ο μόνος τρόπος για να βρούμε το συναίσθημα είναι να γνωρίζουμε την άποψη του γράφοντος για την πρώτη ταινία. Επιπλέον, παρατηρείται και το φαινόμενο της λεγόμενης συντακτικής αμφισημίας, με χαρακτηριστικό παράδειγμα την πρόταση στα αγγλικά “flying planes can be dangerous”: Η φράση υποδηλώνει οτι τα ιπτάμενα αεροπλάνα μπορεί να αποδειχτούν επικίνδυνα ή οτι το να χειρίζεσαι-κάνεις πτήση με αεροπλάνα είναι επικίνδυνο [1]; Τέλος, υπάρχει η πιθανότητα η ίδια φράση να έχει θετικό συναίσθημα για μια ομάδα ατόμων και αρνητικό για μια άλλη, ανάλογα πχ με τις πολιτικές πεποιθήσεις, τις προσωπικές επιθυμίες κλπ. Για παράδειγμα η φράση «Νικητής ο <*όνομα υποψηφίου*> στις δημοτικές εκλογές» είναι θετική για τους ψηφοφόρους του υποψηφίου και κατά πάσα πιθανότητα αρνητική για τους υπόλοιπους.

  Ο χειρισμός της αμφισημίας αποτελεί, αν όχι το σημαντικότερο, ένα από τα πιο φλέγοντα ζητήματα της Ανάλυσης Συναισθήματος με τεράστιο πρακτικό και ερευνητικό ενδιαφέρον και εκτεταμένη βιβλιογραφία. Συγκεκριμένα, ανάμεσα σε πολλές άλλες, ο Cambria στην εργασία [2] κάνει μια εισαγωγή στην Ανάλυση Συναισθήματος σε επίπεδο έννοιας (concept), στο οποίο, σε αντίθεση με τις παραδοσιακές μεθόδους, η προσοχή εστιάζεται στην σημασιολογική (semantic) ανάλυση του κειμένου χρησιμοποιώντας διαδικτυακές οντότητες (web ontologies) ή σημασιολογικά δίκτυα (semantic networks), επιτρέποντας έτσι τη συσσώρευση εννοιολογικών και συναισθηματικών (affective)

πληροφοριών που σχετίζονται με τις απόψεις που εκφράζονται μέσω της φυσικής γλώσσας. Βεβαίως, η εγκυρότητα αυτών των μεθόδων βασίζεται στο βάθος και το εύρος της εφαρμοζόμενης γνωσιακής βάσης. Επιπλέον οι Aisopos et al. στην εργασία [3] προτείνουν μια μέθοδο που βασίζεται σε δύο ορθογώνιες αλλά συμπληρωματικές πηγές απόδειξης (evidence): χαρακτηριστικά βασισμένα στο περιεχόμενο (content-based features) τα οποία συλλαμβάνονται από γράφους ν-γραμμάτων (n-gram graphs) και χαρακτηριστικά βασισμένα στα συμφραζόμενα (context-based features) που συλλαμβάνονται από τον λόγο πολικότητας (polarity ratio). Τα αποτελέσματα υποδεικνύουν σημαντικές βελτιώσεις σε σχέση με τις παραδοσιακές μεθόδους.

- *Χειρισμός της Άρνησης*

Η άρνηση είναι ένα από τα πιο συνηθισμένα γλωσσικά φαινόμενα που επηρεάζουν την πολικότητα και συνεπώς πρέπει να ληφθεί υπόψιν στην Ανάλυση Συναισθήματος. Αν εξαιρέσουμε την απλή περίπτωση κατά την οποία είναι εμφανής η επιρροή της άρνησης (πχ Δεν μου αρέσει αυτό το αμάξι), δεν είναι πάντα εύκολο να διακρίνουμε την επιρροή της σε ένα κείμενο. Μια λέξη άρνησης μπορεί να χρησιμοποιηθεί χωρίς να αντιστρέψει την πολικότητα της εκφραζόμενης άποψης, όπως στην πρόταση «Το κινητό αυτό *όχι* μόνο είναι ακριβό, αλλά είναι επίσης βαρύ και δύσκολο στη χρήση», όπου η λέξη «όχι», όχι μόνο δεν αντιστρέφει την πολικότητα της πρότασης από αρνητική σε θετική, αλλά στην πραγματικότητα ενισχύει την αρνητική άποψη. Ακόμα, η ύπαρξη μιας λέξης άρνησης δεν σημαίνει οτι αντιστρέφεται η πολικότητα όλων των απόψεων που εκφράζονται· Στη φράση «Δεν μου αρέσει το σχέδιο του νέου Nokia κινητού, παρόλα αυτά έχει κάποιες ενδιαφέρουσες νέες λειτουργίες» η άρνηση δεν επηρεάζει τη λέξη «ενδιαφέρουσες» καθώς αυτή βρίσκεται στη δεύτερη πρόταση. Επιπλέον η άρνηση δεν εκφράζεται μόνο μέσω των συνηθισμένων λέξεων άρνησης (όχι, δεν, μην), αλλά και από άλλες λεκτικές μονάδες. Έρευνες στο συγκεκριμένο πεδίο έχουν δείξει οτι υπάρχουν πολλές άλλες λέξεις που μπορούν να αντιστρέψουν την πολικότητα μιας έκφρασης, όπως είναι οι μετατοπιστές-ολισθητές σθένους (valence shifters), οι σύνδεσμοι (connectives) ή τα βοηθητικά ρήματα (modals) [4]. «Βρίσκω τη λειτουργία του νέου κινητού *λιγότερο* πρακτική» είναι ένα παράδειγμα για ολίσθηση πολικότητας, «Μπορεί να είναι ένας σπουδαίος πολιτικός, *αλλά* αδυνατώ να δω το γιατί» μας δείχνει την επίδραση που μπορεί να έχουν οι (αντιθετικοί) σύνδεσμοι, ενώ ένα παράδειγμα πρότασης που χρησιμοποιεί βοηθητικό ρήμα είναι το «Θεωρητικά, το τηλέφωνο *θα έπρεπε* να λειτουργεί ακόμα και κάτω από το νερό».

Μια από τις πιο διάσημες εργασίες είναι της Jia et al. (2009) , στην οποία προτείνεται μια μέθοδος χειρισμού της άρνησης με χρήση στατικών και

δυναμικών χαρακτήρων οριοθέτησης (delimiters) και ευρετικών κανόνων (heuristic rules) που εστιάζουν σε πολικές εκφράσεις (polar expressions) [5]. Οι στατικοί χαρακτήρες οριοθέτησης είναι μονοσήμαντες λέξεις (πχ because, unless στα αγγλικά) που σημάνουν την αρχή μιας άλλης πρότασης. Οι δυναμικοί χαρακτήρες από την άλλη, είναι κανόνες που χρησιμοποιούν πληροφορία βασισμένη στο περιεχόμενο (contextual information) και χρησιμοποιούνται σε σύνθετους τύπους προτάσεων για να εντοπίσουν μόνο τις εκφράσεις στις οποίες υπάρχει άρνηση. Οι ευρετικοί κανόνες εστιάζουν σε περιπτώσεις κατά τις οποίες οι πολικές εκφράσεις προηγούνται λέξεων άρνησης, κάτι που μετατρέπει τις ίδιες τις πολικές εκφράσεις σε χαρακτήρες οριοθέτησης [5]. Άλλες ενδιαφέρουσες εργασίες σχετικά με τον εντοπισμό και τον χειρισμό της άρνησης έχουν γίνει από τους Wiegand et al και από τους Asmi και Ishaya [6],[7].

- *Ανάλυση αναφοράς ή συναναφορά (co-reference, anaphora resolution)*

  Είναι το πρόβλημα της απόφασης σε τι αναφέρεται ένα επίθετο ή μια αντωνυμία. Για παράδειγμα στη φράση «Είδαμε την ταινία και μετά πήγαμε μια βόλτα· ήταν απαίσια» το επίθετο απαίσια αναφέρεται στην ταινία ή στη βόλτα;

- *Υπονοούμενες δηλώσεις (implicitness – sarcasm)*

  Αφορά τον εντοπισμό της ειρωνείας, του χιούμορ και του σαρκασμού. Κάτι τέτοιο δεν μπορεί να συμβεί παρά μόνο εάν υπάρχει καλή κατανόηση του περιεχομένου της συζήτησης, της εκάστοτε κουλτούρας υπό την έννοια του κοινωνικοπολιτικού υποβάθρου και ίσως του συγκεκριμένου θέματος ή και των ανθρώπων που εμπλέκονται σε μια σαρκαστική δήλωση. Αυτού του είδους η γνώση, δηλαδή γνώση του πραγματικού κόσμου, είναι σχεδόν αδύνατο να χρησιμοποιηθεί από μια μηχανή. Επιπλέον, ακόμα και ο καθορισμός του εάν μια δήλωση είναι σαρκαστική, συνήθως είναι μη επαρκής για περαιτέρω ανάλυση, ιδίως σε επίπεδο συναισθήματος, γι'αυτό και σχεδόν όλη η σύγχρονη έρευνα στον εντοπισμό του σαρκασμού έχει μελετήσει το πρόβλημα της ταξινόμησης μια πρότασης «απλά» ως σαρκαστική ή όχι.

  Οι Maynard και Greenwood στην [8] πραγματοποιούν μια ανάλυση των επιπτώσεων του σαρκασμού (sarcasm scope) στην πολικότητα των tweets και φτιάχνουν ένα σύνολο κανόνων που επιτρέπει τη βελτίωση της ακρίβειας της Ανάλυσης Συναισθήματος όταν είναι γνωστή η παρουσία σαρκασμού, μελετώντας συγκεκριμένα τις επιπτώσεις του σαρκασμού που εντοπίζεται στα hashtags (βλέπε παράγραφο 2.3) και αναπτύσσοντας έναν θρυμματιστή (tokenizer) για hashtags που επιτρέπει τον ευκολότερο εντοπισμό συναισθήματος και σαρκασμού μέσα σε αυτά.

- *Αναγνώριση Ονομαστικών Οντοτήτων (Named Entity Recognition)*

  Αφορά την κατανόηση του σε τι πραγματικά αναφέρεται ένας άνθρωπος και στοχεύει στον εντοπισμό οντοτήτων όπως είναι οι άνθρωποι, οι περιοχές, οι οργανισμοί, οι ταινίες και τα προϊόντα· Για παράδειγμα στη φράση «Πήγαμε και είδαμε τους 300 Σπαρτιάτες» το «300 Σπαρτιάτες» αναφέρεται σε μια ομάδα ανθρώπων ή στην ομώνυμη ταινία; Τα περισσότερα συστήματα που ασχολούνται με αυτό το πρόβλημα παίρνουν σαν είσοδο ένα μη-επισημειωμένο κομμάτι κειμένου όπως «Ο Γιάννης αγόρασε 300 μετοχές της Acme Corp το 2006» και παράγουν ένα επισημειωμένο κομμάτι κειμένου που υπογραμμίζει τα ονόματα των οντοτήτων, όπως «Ο $[\text{Γιάννης}]_{όνομα}$ αγόρασε 300 μετοχές της $[\text{Acme Corp}]_{εταιρεία}$ το $[2006]_{χρόνος}$». Μια από τις πιο διάσημες υλοποιήσεις προέρχεται από το Πανεπιστήμιο του Stanford· ο λεγόμενος "Stanford NER"[3] τοποθετεί ταμπέλες σε μια σειρά από λέξεις σε ένα κείμενο, οι οποίες είναι οι ταμπέλες πραγμάτων όπως άνθρωποι, ονόματα εταιρειών, ονόματα γονιδίων ή πρωτεϊνών.

  Οι Derczynski et al. στην [9] ερευνούν το κατά πόσο είναι αποτελεσματικά και στιβαρά (robust) μια σειρά από τέτοια συστήματα της τελευταίας λέξης της τεχνολογίας (state-of-the-art systems) όταν εφαρμοστούν στο Twitter, ενώ οι Li, Chenliang, et al. παρουσιάζουν ένα καινοτόμο μη-επιβλεπόμενο σύστημα αναγνώρισης ονομαστικών οντοτήτων για ροές στο Twitter, που ονομάζεται TwiNER και το οποίο δεν εξαρτάται από αναξιόπιστα τοπικά γλωσσολογικά χαρακτηριστικά, αλλά αντιθέτως συσσωρεύει πληροφορίες από τον παγκόσμιο ιστό (World Wide Web) για να χτίσει στιβαρό τοπικό σημασιολογικό περιεχόμενο για tweets, με τα αποτελέσματα να είναι άκρως ενθαρρυντικά.

- *Υποθετικός λόγος (conditional sentences)*

  Οι υποθετικές προτάσεις είναι ένας από τους πιο συνηθισμένους γλωσσικούς σχηματισμούς, καθότι υπολογίζεται οτι σε ένα τυπικό κείμενο, περίπου το 8% των προτάσεων είναι υποθετικές. Εξαιτίας της υπόθεσης είναι δύσκολος ο καθορισμός του συναισθήματος μιας τέτοιας πρότασης καθώς μπορεί να έχουμε πολλές λέξεις με έντονο συναισθηματικό περιεχόμενο, αλλά στην πραγματικότητα να μην εκφράζεται καμία άποψη. Για παράδειγμα στη φράση «Αν κάποιος κατασκευάσει ένα όμορφο και αξιόπιστο αυτοκίνητο, θα το αγοράσω», αν και υπάρχουν οι συναισθηματικά θετικές λέξεις «όμορφο» και «αξιόπιστο», δεν εκφράζεται κάποιο συναίσθημα ή άποψη για κάποιο συγκεκριμένο αμάξι ενώ στη φράση «Αν το κινητό Nokia που πήρες δεν

[3] http://nlp.stanford.edu/software/CRF-NER.shtml

είναι καλό, αγόρασε αυτό το φανταστικό κινητό Samsung» ο συγγραφέας είναι θετικός ως προς το Samsung κινητό, αλλά δεν εκφράζει καμία άποψη για αυτό της Nokia.

Οι Narayanan et al. στην [11] παρουσιάζουν μια γλωσσολογική ανάλυση τέτοιων προτάσεων και στη συνέχεια χτίζουν κάποια επιβλεπόμενα μοντέλα μάθησης προκειμένου να καθορίσουν αν τα συναισθήματα που εκφράζονται σε διαφορετικά θέματα σε μια υποθετική πρόταση είναι θετικά, αρνητικά ή ουδέτερα.

- *Σύγκρουση πολικότητας (polarity conflict)*

  Οι λέξεις σε ένα κείμενο δεν πρέπει να ελέγχονται (μόνο) ξεχωριστά, αλλά και σε ομάδες με τις γειτονικές τους (bigrams, trigrams etc). Αυτό συμβαίνει γιατί μπορεί ένα αρνητικό επίθετο να μεταβάλλει την πολικότητα ενός θετικού ουσιαστικού και αντίστροφα, όπως στις φράσεις «δυσάρεστο όνειρο» ή «η αναγέννηση του φασισμού στην Ευρώπη».

- *Πολλαπλές ή/και αντικρουόμενες απόψεις*

  Είναι πολύ πιθανό μέσα στο ίδιο κείμενο ή ακόμα και στην ίδια πρόταση, να εκφράζονται τόσο θετικές όσο και αρνητικές απόψεις είτε για το ίδιο, είτε για διαφορετικά αντικείμενα, όπως στη φράση «Έχω δύο υπολογιστές: ένα Lenovo και ένα Dell. Ο Lenovo έχει τρομερό επεξεργαστή και χάλια κάρτα γραφικών, ενώ ο Dell έχει μεγάλο σκληρό δίσκο και προβληματική μπαταρία». Για την επίλυση του προβλήματος γίνονται προσπάθειες σε δύο επίπεδα: Στον εντοπισμό των αντικειμένων πάνω στα οποία εκφράζονται απόψεις (*object identification*), και στην εξαγωγή χαρακτηριστικών και την ομαδοποίηση των συνωνύμων (feature extraction and synonym grouping).

2.3 Μικρο-ιστολόγια και Twitter

Ένα τεράστιο μέρος των διαδικτυακών κειμένων δημιουργείται καθημερινά στα λεγόμενα μικρο-ιστολόγια (microblogs). Τα μικρο-ιστολόγια είναι μια ειδική μορφή ιστολογίων στα πλαίσια της οποίας το περιεχόμενο είναι περιορισμένο σε μέγεθος και δίνουν τη δυνατότητα στους χρήστες της υπηρεσίας να ανταλλάξουν μικρά μηνύματα (microposts) τα οποία μπορούν να αποτελούνται από σύντομες προτάσεις, εικόνες, ή υπερσυνδέσμους.

Ανάμεσα σε μια πλειάδα εξαιρετικά δημοφιλών μικρο-ιστολογίων, όπως τα Tumblr, LinkedIn, FriendFeed, Cif2.net, Plurk, Jaiku, η πιο δημοφιλής και επιτυχημένη πλατφόρμα στις μέρες μας είναι το Twitter. Το Twitter επιτρέπει στους χρήστες του να δημοσιεύσουν ενημερώσεις και να ανταλλάσσουν

μηνύματα με μέγεθος το πολύ 140 χαρακτήρες. Τα μηνύματα αυτά, που ονομάζονται *tweets*, μπορεί να περιέχουν αναφορές σε άλλους χρήστες, ή και συνδέσμους σε διαφόρων ειδών διαδικτυακό υλικό όπως ιστοσελίδες, εικόνες και βίντεο. Οι χρήστες μπορούν επιπλέον να «ακολουθούν» άλλους χρήστες (*follow*), καθώς και να προωθούν-αναδημοσιεύοουν τα μηνύματα άλλων χρηστών (*retweet*). Μια ακόμα σημαντική παράμετρος είναι οτι το Twitter επιτρέπει την αναζήτηση μηνυμάτων (tweets) και προσπαθεί να κάνει πιο εύκολο για τους χρήστες να δώσουν έμφαση στο θέμα των μηνυμάτων τους, δίνοντας τους τη δυνατότητα να εισάγουν το πρόθεμα # πριν από κάθε λέξη. Οι λέξεις αυτές, που ονομάζονται *hashtags*, γίνονται σύνδεσμοι που οδηγούν σε μια λίστα από άλλα tweets που περιέχουν το ίδιο hashtag. Τέλος το Twitter καταγράφει τις πιο «μοδάτες» (trending) λέξεις και hashtags, διευκολύνοντας τους χρήστες του να μάθουν τι συζητιέται κάθε στιγμή. Σύμφωνα με τα πιο πρόσφατα (επίσημα) στατιστικά στοιχεία, το τελευταίο τέταρτο του 2015[4], το Twitter αριθμούσε κατά μέσο όρο 305 εκατομμύρια ενεργούς χρήστες κάθε μήνα, ενώ υπολογίζεται οτι κάθε δευτερόλεπτο στέλνονται 6000 tweets, που αντιστοιχούν σε πάνω απο 350 χιλιάδες tweets το λεπτό, ή 500 εκατομμύρια tweets τη μέρα, ή 200 δισεκατομμύρια tweets κάθε χρόνο[5].

Αυτή η πραγματικά αδιανόητη ποσότητα δημόσιας πληροφορίας που βρίσκεται στο Twitter, με μια θεματολογία που αγγίζει όλες τις εκφάνσεις του καθημερινού βίου, και η πρόκληση της εκμετάλλευσής της για εξαγωγή χρήσιμων συμπερασμάτων, αποτέλεσαν την αφορμή για να εστιάσουμε το ενδιαφέρον μας, αποκλειστικά στην Ανάλυση Συναισθήματος στο Twitter.

2.3.1 Προβλήματα στο Twitter

Όπως θα μπορούσε κάποιος εύκολα να μαντέψει, η ίδια η φύση των μικρο-ιστολογίων και κατ'επέκταση και του Twitter, δημιουργεί επιπλέον προκλήσεις στο κομμάτι της Ανάλυσης Συναισθήματος. Τα βασικότερα προβλήματα τα οποία έχουμε να αντιμετωπίσουμε (πέραν των κλασσικών που αφορούν γενικότερα τον κλαδο της Ανάλυσης Συναισθήματος και αναφέρθηκαν στο κεφάλαιο 2.1), είναι τα εξής:

- *Μήκος κειμένου*

  Σε αντίθεση με τα συνηθισμένα κείμενα, όπως οι κριτικές προϊόντων ή ταινιών, τα οποία αποτελούνται από πολλές λέξεις που βοηθούν στην εξαγωγή πολύτιμων στατιστικών μεγεθών, τα κείμενα στο Twitter αποτελούνται από λίγες φράσεις ή 1-2 προτάσεις το πολύ. Αν και αυτό μπορεί να θεωρηθεί ως πλεονέκτημα από τη μία μεριά ([3]), καθώς οι συγ-

[4] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[5] http://www.internetlivestats.com/twitter-statistics/

γραφείς τείνουν να αναφέρονται απευθείας στο θέμα που τους ενδιαφέρει και αποφεύγουν (αναγκαστικά) την περιττολογία, θέτει το πρόβλημα της εξάρτησης της εκφραζόμενης άποψης από μία (ή ελάχιστες) και μόνο λέξη. Εάν η λέξη αυτή δεν βρίσκεται καταχωρημένη στο χρησιμοποιούμενο λεξικό, ή δεν έχει εμφανιστεί στο σύνολο δεδομένων εκπαίδευσης (βλέπε παράγραφο 2.4), τότε είναι αδύνατη η εξαγωγή του συναισθήματος. Επιπλέον το μικρό μέγεθος κειμένου συνεπάγεται συνήθως και έλλειψη σημασιολογικού περιεχομένου (context), καθιστώντας ακόμα πιο δύσκολη την Ανάλυση Συναισθήματος σε διφορούμενα ή ειρωνικά tweets.

- *Ποικιλία ορθογραφίας (spelling variation) και μη-τυποποιημένο λεξιλόγιο (non-standard vocabulary)*

  Εξαιτίας του περιορισμένου μεγέθους των μηνυμάτων η πλειονότητα των χρηστών γράφει σε *ανεπίσημο λόγο* (informal language) χρησιμοποιώντας *συντομογραφίες* ή και δημιουργώντας νέους τύπους λέξεων που φέρουν μικρή ομοιότητα με τις αυθεντικές (πχ χαίρομαι→ χρμ, σήμερα→σμρ), λέξεις και εκφράσεις της *αργκό-καθομιλουμένης* (slang) (πχ ψήσου, κούλαρε, τι φάση παίζει, μου τη βάρεσε), *νεολογισμούς* (πχ τρολλάρω, κλάιν, yolo), καθώς και μη συμβατικές εκφράσεις (πχ χαχαχαχα, 3ός αντί θεός). Επιπλέον ο αυθορμητισμός και η παρορμητικότητα των χρηστών, οδηγεί πολύ συχνά σε ανορθογραφίες, σε γραμματικά και συντακτικά λάθη, αλλά και στην ύπαρξη φαινομένων έμφασης η οποία υποδηλώνεται είτε μέσω της χρήσης κεφαλαίων γραμμάτων (πχ «ΜΑ ΕΙΝΑΙ ΣΟΒΑΡΟΣ;;;»), είτε μέσω της επανάληψης χαρακτήρων και συνεπώς της μεγένθυσης του μήκους μια λέξης (πχ «Η ταινία ήταν τέλειαααααα»). Μάλιστα τα παραπάνω, αναγκάζουν τις παρούσες μεθόδους να χρησιμοποιούν λεξικά συναισθήματος, τα οποία εξαρτώνται σε τεράστιο βαθμό από το εκάστοτε πεδίο που εφαρμόζονται (*domain-specific)*.

- *Ποικιλία θεμάτων*

  Τα θέματα που συζητιούνται στο Twitter δεν περιορίζονται, αντιθέτως εκτείνονται σε χιλιάδες θεματικά περιεχόμενα. Αυτό δημιουργεί πρόβλημα, γιατί πολλές λέξεις έχουν διαφορετική σημασία ανάλογα με το σημασιολογικό περιεχόμενο (context). Μάλιστα η τεράστια ποικιλία θεμάτων δημιουργεί και διαφορετικά είδη γραφής, καθώς μπορεί να συναντήσουμε από τελείως επίσημο κείμενο όπως θα παρουσιαζόταν σε μια εφημερίδα μέχρι τελείως ανεπίσημη γλώσσα, γεγονός που δημιουργεί τρομερά προβλήματα στην εκπαίδευση δεδομένων και στη δημιουργία λεξικών βοηθημάτων.

- *Πολυγλωσσικό περιεχόμενο*

  Στα περισσότερα tweets είναι εξαιρετικά πιθανή η ύπαρξη λέξεων που ανήκουν σε περισσότερες από μία γλώσσες. Σε ό,τι αφορά τα ελληνικά tweets είναι σχεδόν μόνιμη η χρήση αγγλικών όρων κυρίως λόγω των hashtags, ενώ παρουσιάζει τεράστιο ενδιαφέρον (και παράλληλα δημιουργεί αρκετά προβλήματα) η χρήση *greeklish*, δηλαδή λέξεων της ελληνικής γλώσσας γραμμένων με λατινικούς χαρακτήρες.

- *Πολυμεσικό περιεχόμενο*

  Η ύπαρξη εικόνων, βίντεο και υπερσυνδέσμων σε ένα tweet, μπορεί να υποδεικνύει το σημασιολογικό περιεχόμενο του μηνύματος το οποίο δεν υποδηλώνεται και δεν μπορεί να εξαχθεί μόνο από το κείμενο. Οι Borth, Damian, et al καταπιάνονται με την πρόκληση της Ανάλυσης Συναισθήματος από οπτικό περιεχόμενο, και σε αντίθεση με τις υφιστάμενες μεθόδους που εξάγουν το συναίσθημα απευθείας από χαμηλού επιπέδου οπτικά χαρακτηριστικά (low level features), προτείνουν μια νέα προσέγγιση βασισμένη στην κατανόηση των οπτικών εννοιών (concepts) που είναι στενά συνδεδεμένα με συναισθήματα [12]. Ακόμα οι Poria, Soujanya, et al προτείνουν μια νέα μεθοδολογία για πολυμεσική (multimodal) Ανάλυση Συναισθήματος, η οποία στοχεύει στη συλλογή συναισθημάτων από Web βίντεο, παρουσιάζοντας ένα μοντέλο που χρησιμοποιεί ήχο, οπτικές και κειμενικές βοήθειες (textual modalities) ως πηγές πληροφορίας [13].

- *Ιδιαίτερες λεκτικές μονάδες (special tokens)*

  Η χρήση υπερσυνδέσμων (url's) και χαρακτήρων emoticon[6] (δηλαδή γραφικών αναπαραστάσεων των ανθρώπινων εκφράσεων του προσώπου), μπορεί να οδηγήσουν σε δυσκολίες όταν προσπαθούμε να εφαρμόσουμε επεξεργασία φυσικής γλώσσας, όπως αναλυτές μερών του λόγου (part-of-speech taggers) ή συντακτικούς αναλυτές, καθώς οι τελευταίοι εκπαιδεύονται κυρίως σε κείμενα εφημερίδων, τα οποία διαφέρουν αισθητά σε σχέση με τα μικρο-ιστολόγια. Συνεπώς πρέπει να υπάρξει μια ιδιαίτερη μεταχείριση αυτών των μονάδων, είτε αυτή είναι η αφαίρεσή τους, είτε η αξιοποίησή τους (ειδικά στην περίπτωση των emoticon) για τον καθορισμό του συναισθήματος.

[6] https://www.piliapp.com/twitter-symbols/

2.4 Προσεγγίσεις

Αν και οι πρώτες ερευνητικές προσπάθειες στην Ανάλυση Συναισθήματος πάνε πίσω στα τέλη της δεκαετίας του 1970 με τη δουλειά του Jaime Carbonell στα belief models and systems ([14]), η πραγματική απογείωση έλαβε χώρα στις αρχές του 2000 ταυτόχρονα με την έκρηξη των διαδικτυακών εφαρμογών (Web 2.0) οι οποίες αύξησαν τη συναίσθηση της σημασίας του πεδίου και των προοπτικών του για τις επιχειρήσεις, τα οικονομικά, την πολιτική και το online marketing. Κατά τη διάρκεια των δύο τελευταίων δεκατιών έχουν γίνει εκτεταμένες ερευνητικές προσπάθειες στην Ανάλυση Συναισθήματος και έχει προταθεί μια τεράστια ποσότητα από πιθανά μοντέλα και προσεγγίσεις. Παρόλα αυτά, οι βασικές προσεγγίσεις είναι τρείς: της μηχανικής μάθησης (machine learning approach), της βασισμένης σε λεξικό μεθόδου (lexicon-based approach) και της υβριδικής μεθόδου (hybrid approach).

2.4.1 Μέθοδοι μηχανικής μάθησης

2.4.1.1 Ορισμός

Μηχανική μάθηση είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης (Artificial Intelligence) που ασχολείται με αλγορίθμους που επιτρέπουν στους υπολογιστές να μαθαίνουν. Σύμφωνα με έναν πιο τυπικό ορισμό είναι η αυτοματοποιημένη διαδικασία εξαξωγής προτύπων ή μοτίβων (patterns) από μεγάλους όγκους δεδομένων και χρησιμοποίησής των για τη διεξαγωγή προβλέψεων πάνω σε νέα δεδομένα. Συνήθως όταν υλοποιούμε συστήματα μηχανικής μάθησης χρησιμοποιούμε ένα σύνολο δεδομένων εκπαίδευσης. Το σύνολο αυτό είναι τα δεδομένα τα οποία χρησιμοποιεί ο αλγόριθμος για να κατανοήσει τη δομή των δεδομένων και να βρεί το υποβόσκον μοντέλο που θα χρησιμοποιηθεί αργότερα για το χειρισμό νέων, άγνωστων δεδομένων. Η ικανότητα του σωστού χειρισμού άγνωστων δεδομένων βάση του συνόλου εκπαίδευσης, ονομάζεται ικανότητα γενίκευσης (*generalization*). Η υπερπροσαρμογή (*overfitting*) από την άλλη μεριά, συμβαίνει όταν ένα σύνολο δεδομένων εκπαίδευσης έχει οδηγήσει σε ένα μοντέλο το οποίο είναι υπερβολικά προσαρμοσμένο πάνω στο συγκεκριμένο σύνολο δεδομένων, καθιστώντας αδύνατη τη γενίκευσή του. Η υπερ-εκπαίδευση του συνόλου δεδομένων μπορεί να αποφευχθεί αξιολογώντας το μοντέλο με κάποια δεδομένα (test data), τα οποία μπορούν να είναι ένα υποσύνολο των διαθέσιμων δεδομένων. Όταν έχουμε να επιλέξουμε μεταξύ πολλών ταξινομητών, αυτός που αποδίδει καλύτερα πάνω στο testing set έχει την μεγαλύτερη πιθανότητα να είναι αυτός με την καλύτερη ικανότητα γενίκευσης [15].

Τα μοντέλα μηχανικής μάθησης συνήθως χωρίζονται σε τρείς μεγάλες κατηγορίες ανάλογα με την επίβλεψη που απαιτείται κατά τη διάρκεια της εκπαίδευσης:

- Στην επιβλεπόμενη μάθηση (*supervised learning*) η εκπαίδευση των ταξινομητών γίνεται με επισημειωμένα δείγματα (labelled samples), δηλαδή, στην περίπτωση του Twitter, με tweets που έχουν επισημειωθεί με την κλάση που ανήκουν (αρνητική, θετική ή ουδέτερη). Οι Naïve Bayes, Maximum Entropy και Support Vector Machines (SVM) είναι οι πιο γνωστοί ταξινομητές αυτής της κατηγορίας [16].

- Στη μη-επιβλεπόμενη μάθηση (*unsupervised learning*) δουλεύουμε με μη-επισημειωμένα δεδομένα (unlabeled data). Διάσημοι μη-επιβλεπόμενοι αλγόριθμοι περιλαμβάνουν τους αλγόριθμους ομαδοποίησης (clustering algorithms) όπως k-means, k-methoids, hierarchical clustering, τα κρυφά μοντέλα Markov (HMM Hidden Markov Models) και κάποια μη-επιβλεπόμενα μοντέλα νευρωνικών δικτύων, όπως οι αυτοοργανούμενοι χάρτες (self-organising maps) [17].

- Στην ημι-επιβλεπόμενη μάθηση (*semi-supervised learning*) η εκπαίδευση των ταξινομητών γίνεται τόσο με επισημειωμένα όσο και με μη-επισημειωμένα δεδομένα. Τέτοιοι αλγόριθμοι είναι ο label propagation και τα διάφορα μοντέλα βασισμένα σε γράφους (graph-based models)[18].

2.4.1.2 Μηχανική μάθηση στην Ανάλυση Συναισθήματος

Αναλύοντας μια μεγάλη ποσότητα κειμένων, ένα μοντέλο μπορεί να εκπαιδευτεί να ταξινομεί νέα κείμενα με βάση τις ομοιότητες με τα κείμενα με τα οποία εκπαιδεύτηκε. Η ταξινόμηση απαιτεί τα κείμενα εκπαίδευσης να είναι ήδη επισημειωμένα (annotated) με κάποιο συναίσθημα, έτσι ώστε ο αλγόριθμος να προσπαθήσει να βρεί τι τα διαχωρίζει και να χρησιμοποιήσει αυτή τη γνώση σε νέα, άγνωστα κείμενα.

Το πλεονέκτημα των μεθόδων μηχανικής μάθησης είναι οτι δεν απαιτείται κάποια γνώση σχετικά με τη γλώσσα στην οποία είναι γραμμένα τα κείμενα. Ιδανικά, κάθε τέτοιο σύστημα θα μπορούσε να χρησιμοποιηθεί με πολλαπλές γλώσσες, κάτι που είναι πολύ χρήσιμο για Ανάλυση Συναισθήματος σε κείμενα γραμμένα σε γλώσσες για τις οποίες δεν υπάρχουν διαθέσιμα αξιόπιστα εργαλεία επεξεργασίας φυσικής γλώσσας (NLP tools) [19].

Αν και οι (επιβλεπόμενες (κυρίως)) τεχνικές μηχανικής μάθησης έχουν αποδειχτεί οι πιο χρήσιμες και αποτελεσματικές για την Ανάλυση Συναισθήματος, η ιδιαίτερη φύση του Twitter και των μηνυμάτων του θέτουν αρκετούς περιορισμούς σε αυτού του είδου την προσέγγιση. Καταρχάς, εξαρτώνται από την ύπαρξη εκτεταμένων επισημειωμένων συλλογών (corpora) από tweets για την εκπαίδευση των ταξινομητών, την ίδια στιγμή που η διαδικασία επισημείωσης κειμένων είναι συνήθως αρκετά ακριβή και χρονοβόρα [20], ειδικά για συνεχώς μεταβαλλόμενα και εξελισσόμενα θεματικά πεδία όπως στο Twitter. Δεύτερον, οι επιβλεπόμενες τεχνικές είναι συνήθως εξαρτώμενες από το εκάστοτε πεδίο

(domain dependent), δηλαδή οι ταξινομητές που εκπαιδεύονται σε δεδομένα από ένα συγκεκριμένο πεδίο (πχ tweets για πολιτικά γεγονότα) μπορεί να εμφανίσουν χαμηλή απόδοση όταν εφαρμοστούν σε δεδομένα από άλλο θεματικό πεδίο (πχ tweets για αθλητισμό) [21]. Τρίτον τα tweets τείνουν να είναι πολύ αραιά (sparse) εξαιτίας της συχνής χρήσης συντομογραφιών, ανορθογραφιών, γραμματικά λανθασμένων λέξεων και μη τυπικών εκφράσεων, γεγονός που επηρεάζει αρνητικά την απόδοση του ταξινομητή καθώς πολλοί όροι των δεδομένων εκπαίδευσης δεν εμφανίζονται στα δεδομένα αξιολόγησης [22].

Οι περισσότερες προσπάθειες για να ξεπεραστούν οι παραπάνω δυσκολίες έχουν να κάνουν με τη χρήση μη-επιβλεπόμενων ταξινομητών για τη μείωση της εξάρτησης από επισημειωμένα δεδομένα με τη χρήση διαφορετικών διαδικασιών (feature engineering processes) και την υϊοθέτηση τεχνικών μείωσης διάστασης (dimensionality reduction techniques) για τη μείωση της αραιότητας (sparsity) των tweets [23,24].

2.4.2 Μέθοδοι βασισμένες σε λεξικό συναισθήματος

Η προσέγγιση βασισμένη σε λεξικό (lexicon-based method) υποθέτει οτι ο συναισθηματικός προσανατολισμός ενός κειμένου μπορεί να συναχθεί από το συναισθηματικό προσανατολισμό των επιμέρους λέξεων και φράσεών του. Σε αντίθεση με τις μεθόδους βασισμένες σε μηχανική μάθηση, η προσέγγιση βάση λεξικού δεν απαιτεί την εκπαίδευση ενός ταξινομητή. Αντιθέτως, χρησιμοποιεί λεξικά συναισθήματος για να αποδώσει το συναίσθημα των συναισθηματικά φορτισμένων λέξεων στο κείμενο. Η απόδοση μιας μεθόδου Ανάλυσης Συναισθήματος βασισμένη σε λεξικό συνήθως καθορίζεται από τον τύπο του λεξικού συναισθήματος και από τον αλγόριθμο ανίχνευσης συναισθήματος (sentiment detection algorithms), δηλαδή από τον αλγόριθμο που χρησιμοποιείται για τον εντοπισμό των συναισθηματικά φορτισμένων λέξεων του κειμένου και τον υπολογισμό του συνολικού συναισθήματος.

Σε ό,τι αφορά το Twitter η πιο συνηθισμένη πρακτική είναι η χρήση ενός προ-ενσωματωμένου λεξικού συναισθήματος μαζί με έναν απλό αλγόριθμο ταιριάσματος λέξεων κλειδιών (keywords).

Οι βασισμένες σε λεξικό μέθοδοι έχουν το πλεονέκτημα της *απλότητας* και της *ταχύτητας*, στοιχεία που αποτελούν απαραίτητη προϋπόθεση ιδίως όταν απαιτείται η ανάλυση τεράστιων όγκων δεδομένων, όπως για παράδειγμα σε συλλογές που αποτελούνται από δεκάδες χιλιάδες tweets [25,26]. Από την άλλη, οι μέθοδοι αυτοί έρχονται συχνά αντιμέτωπες με δύο βασικούς περιορισμούς:

1) Οι συμβατικές μέθοδοι αδυνατούν να ανιχνεύσουν σύνθετους τύπους συναισθήματος σε tweets, όπως για παράδειγμα την άρνηση («Για κάποιο λόγο, ποτέ δεν μου άρεσε αυτό το τραγούδι) ή το ενισχυμένο συναίσθημα («Είμαι *πολύ* χαρούμενος για σένα φίλε», «Σήμερα νιώθω *εξαιρετικά*

κουρασμένος»). Επιπλέον εξαιτίας του γεγονότος οτι στα λεξικά συναισθήματος οι λέξεις επισημειώνονται με μια σταθερή πολικότητα, δεν λαμβάνεται υπόψιν το ευρύτερο σημασιολογικό πλαίσιο (context).

2) Τα παραδοσιακά λεξικά, όπως τα MPQA και SentiWordNet για την αγγλική γλώσσα, είναι σχεδιασμένα για να χειρίζονται επίσημα και καλογραμμένα κείμενα. Στο Twitter όμως, ο ιδιαίτερος και ανεπίσημος τρόπος γραφής εξαιτίας του περιορισμού των 140 χαρακτήρων, οδηγεί στον εντοπισμό-ανάκτηση ελάχιστων μη-συμβατικών ή ασυνήθιστων λέξεων, αφού αυτές σπάνια είναι καταχωρημένες στα παραδοσιακά λεξικά [25,27].

Συνεπώς οι μέθοδοι βασισμένες σε λεξικό μπορούν να εφαρμοστούν άμεσα σε δεδομένα από το Twitter, χωρίς να χάνεται πολύτιμος χρόνος για την εκπαίδευση ταξινόμητων, με το βασικό περιορισμό οτι η απόδοση και αποτελεσματικότητα όλου του συστήματος συνδέεται αναπόσπαστα με τη λεξική πηγή στην οποία βασίζεται· Μια μέθοδος βασισμένη σε λεξικό είναι τόσο καλή, όσο και το λεξικό που χρησιμοποιεί. Τέλος η απόδοση τέτοιων συστημάτων, σε επίπεδο ακρίβειας και χρονικής πολυπλοκότητας, επιδεινώνεται δραστικά με την εκθετική αύξηση του μεγέθους του λεξικού.

Στο σημείο αυτό θα πρέπει να σημειώσουμε οτι σε οποιαδήποτε εφαρμογή ανάλυσης κειμένου υπάρχει η δυνατότητα επιλογής μεταξύ στατιστικών ή συντακτικών τεχνικών. Οι συντακτικές τεχνικές (*syntactic techniques*) μπορούν να οδηγήσουν σε καλύτερη ακρίβεια γιατί κάνουν χρήση των συντακτικών κανόνων μιας γλώσσας με σκοπό να ανιχνεύσουν τα ρήματα, τα επίθετα και τα ουσιαστικά. Δυστυχώς αυτού του είδους οι τεχνικές είναι άρρηκτα συνδεδεμένες με τη γλώσσα του κειμένου και συνεπώς δεν μπορούν να εφαρμοστούν σε άλλες γλώσσες. Από την άλλη οι στατιστικές τεχνικές (*statistical techniques*) έχουν πιθανοτικό υπόβαθρο και εστιάζουν στις σχέσεις μεταξύ των λέξεων και των κατηγοριών. Οι τεχνικές αυτές έχουν δύο σημαντικά πλεονεκτήματα σε σχέση με τις συντακτικές: μπορούμε να τις χρησιμοποιήσουμε σε διάφορες γλώσσες τροποποιώντας τες ελάχιστα ή και καθόλου και μπορούμε να χρησιμοποιήσουμε μετάφραση μηχανής (machine translation) του αρχικού σετ δεδομένων και να έχουμε και πάλι ικανοποιητικά αποτελέσματα. Οι δύο προαναφερθείσες μέθοδοι μπορούν να συνδυαστούν τόσο με τεχνικές βασισμένες σε λεξικό, όσο και με τεχνικές μηχανικής μάθησης.

2.4.3 Υβριδικές μέθοδοι

Η όλο και αυξανόμενη πρόοδος που παρατηρείται στον τομέα της Ανάλυσης Συναισθήματος έχει παρακινήσει τους ερευνητές να εξετάσουν την πιθανότητα μιας υβριδικής προσέγγισης η οποία συνολικά θα παρουσίαζε την *ακρίβεια* των μεθόδων μηχανικής μάθησης και την *ταχύτητα* των μεθόδων βασισμένων σε λεξικά συνασισθήματος. Στην ουσία αυτό που προσπαθούν οι υβριδικές προσεγγίσεις είναι να εκμεταλλευτούν τα πλεονεκτήματα των δύο μεθόδων και να αποφύγουν τα μειονεκτήματα. Στη γενική περίπτωση, στις υβριδικές προσεγγίσεις, η μία από τις δύο βασικές προσεγγίσεις χρησιμοποιείται για να

τονώσει την απόδοση της άλλης προσέγγισης. Για παράδειγμα, από τη μιά μεριά οι επιβλεπόμενοι ταξινομητές μπορεί να κάνουν χρήση μεθόδων βασισμένων σε λεξικό για να μειώσουν την εξάρτηση από χειροκίνητες επισημειωμένες συλλογές δεδομένων εκπαίδευσης [28,34], ενώ από την άλλη οι αλγόριθμοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για να «στηρίξουν» (bootstrap) τα λεξικά συναισθήματος στις μεθόδους βασισμένες σε λεξικό [29].

2.5 Η δικιά μας προσέγγιση

Στην προηγούμενη ενότητα κάναμε μια ανάλυση των επικρατέστερων προσεγγίσεων στο πρόβλημα της Ανάλυσης Συναισθήματος. Είδαμε οτι οι μέθοδοι μηχανικής μάθησης εξαρτώνται άμεσα από το πεδίο στο οποίο εφαρμόζονται (domain-specific), γεγονός που είναι πλεονέκτημα και μειονέκτημα ταυτόχρονα: Από τη μια, οι ταξινομητές που εκπαιδεύονται με δεδομένα από ένα συγκεκριμένο πεδίο, έχουν εξαιρετικά υψηλές αποδόσεις όταν εφαρμόζονται σε δεδομένα από το ίδιο πεδίο, αδυνατούν όμως να παράξουν αντίστοιχα αποτελέσματα όταν εφαρμοστούν σε καινούριο πεδίο·σε μια τέτοια περίπτωση απαιτείται επανεκπαίδευση του ταξινομητή. Επιπλέον, η εκπαίδευση των ταξινομητών απαιτεί μεγάλες ποσότητες επισημειωμένων δεδομένων, τα οποία δεν είναι πάντα διαθέσιμα. Στην αντίπερα όχθη, οι τεχνικές βασισμένες σε λεξικό δεν απαιτούν ούτε εκπαίδευση ούτε επανεκπαίδευση, στοιχείο πολύ σημαντικό για διάφορες εφαρμογές στο Twitter, όπως για Ανάλυση Συναισθήματος σε πραγματικό χρόνο σε ροές tweets (real time Twitter streams), αλλά περιορίζονται από το λεξικό που χρησιμοποιούν, καθώς αυτό αφενός θα χρησιμοποιεί στατικές λέξεις με σταθερή πολικότητα ανεξαρτήτως περιεχομένου, και αφετέρου δεν θα μπορεί να καλύψει την τεράστια ποικιλία νέων όρων και νεολογισμών που συνεχώς ξεπροβάλλουν στο Twitter.

Σε ό,τι αφορά την Ανάλυση Συναισθήματος στο Twitter σε tweets γραμμένα στην ελληνική γλώσσα, ο σχετικά μικρός όγκος διαθέσιμων επισημειωμένων tweets , αλλά και η περιορισμένη πρόσβαση που είχαμε εμείς σε τέτοια δεδομένα καθώς δεν είναι ελεύθερα διαθέσιμα, καθιστά μια προσέγγιση βασισμένη σε μοντέλα μηχανικής μάθησης αρκετά επίφοβη και χρονοβόρα. Συνεπώς στην παρούσα διπλωματική, υλοποιούμε μια μέθοδο η οποία θα βασίζεται σε λεξικό συναισθήματος. Για το λεξικό αυτό θα χρησιμοποιηθεί ένα ήδη υπάρχον λεξικό, το οποίο θα επεκταθεί με κάποιες από τις πιο συχνά χρησιμοποιούμενες λέξεις στο Twitter, συμπεριλαμβανομένων λέξεων της καθομιλουμένης και διάφορων νεολογισμών, και οι οποίες θα προκύψουν μετά από στατιστική ανάλυση σε ένα μεγάλο συνόλο δεδομένων από tweets. Μάλιστα θα φροντίσουμε οι λέξεις αυτές να έχουν έντονο συναισθηματικό υπόβαθρο, δηλαδή να μην είναι κατά βάση ουδέτερης πολικότητας (πχ περπατάω, σκέφτομαι, πολιτική, εκλογές κλπ), και το αντίστοιχο score που θα ανατεθεί σε καθεμιά θα ακολουθεί μια κλίμακα από το -4 έως το +4, προκειμένου να κάνουμε μια διαβάθμιση της πολικότητας και να ξεχωρίσουμε τις περισσότερο από τις λιγότερο συναισθηματικά έντονες λέξεις.

2.6 Παράμετροι αξιολόγησης

Όπως έγινε φανερό και στις προηγούμενες παραγράφους οι εγγενείς δυσκολίες του προβλήματος της Ανάλυσης Συναισθήματος το καθιστούν ένα εξαιρετικά σύνθετο και πολυπρόσπωπο πεδίο. Είναι λογικό η αξιολόγηση ενός συστήματος Ανάλυσης Συναισθήματος να μην περιορίζεται σε μια μόνο μετρική, αλλά να απαιτεί το συνδυασμό και την ερμηνεία πολλών διαφορετικών μετρικών μεγεθών.

2.6.1 Συμφωνία μεταξύ κριτών

Η συμφωνία μεταξύ κριτών είναι ένα στατιστικό μέγεθος που αποτυπώνει την ομοιογένεια ή συμφωνία που υπάρχει στις αξιολογήσεις που γίνονται από μια ομάδα κριτών. Σύμφωνα με τον Stemler (2004) [30], σε όλες τις καταστάσεις στις οποίες εμπλέκεται η ανθρώπινη κρίση, είναι απαραίτητος ο υπολογισμός του βαθμού της συμφωνίας μεταξύ των κριτών[7] (inter-rater agreement ή inter-rater reliability), καθώς αυτή η τιμή έχει σημαντικές επιπτώσεις στην αξιοπιστία των αποτελεσμάτων της εκάστοτε μελέτης. Αν δύο (ή περισσότεροι) κριτές δεν παρουσιάζουν σημαντικό βαθμό συμφωνίας, τότε οποιαδήποτε περαιτέρω μελέτη του αντικειμένου που αξιολογείται από τους κριτές τίθεται εν αμφιβόλω.

Στα πλαίσια ενός συστήματος Ανάλυσης Συναισθήματος, αυτό που επιθυμούμε είναι να υπάρχει συμφωνία μεταξύ της εξόδου του συστήματος και της ανθρώπινης κρίσης. Ας υποθέσουμε λοιπόν οτι για τις ανάγκες αξιολόγησης ενός τέτοιου συστήματος χρησιμοποιούμε μια ομάδα n κριτών - βαθμολογητών. Τότε θεωρούμε οτι το σύστημα λειτουργεί αποδοτικά εάν αντικαθιστώντας κάποιον από τους n κριτές με την έξοδο του συστήματός μας, δεν υπάρχει σημαντική μεταβολή της συμφωνίας μεταξύ των κριτών. Με αυτόν τον τρόπο και επειδή σύμφωνα με έρευνες οι ανθρώπινοι κριτές δεν συμφωνούν σχεδόν ποτέ απόλυτα μεταξύ τους, δείχνουμε οτι το σύστημα δεν λειτουργεί χειρότερα από τον άνθρωπο, δηλαδή με άλλα λόγια κάνει τόσο καλή ταξινόμηση όσο θα έκανε και ένας άνθρωπος.

Υπάρχουν διάφορα στατιστικά μεγέθη που μπορούν να χρησιμοποιηθούν για τον υπολογισμό της συμφωνίας μεταξύ κριτών, καθένα από τα οποία είναι κατάλληλο για διαφορετικούς τύπους μέτρησης. Μερικά από αυτά είναι η από κοινού πιθανότητα συμφωνίας (joint-probability of agreement), οι συντελεστές κάππα (kappa statistics) του Cohen και του Fleiss, οι συντελεστές συσχέτισης (correlation coefficients) του Pearson, του Kendall και του Spearman και το άλφα

[7] https://en.wikipedia.org/wiki/Inter-rater_reliability

του Krippendorff (Krippendorff's alpha). Ακολουθεί μια σύντομη επισκόπηση των δύο βασικότερων μεγεθών που χρησιμοποιούνται στην Ανάλυση Συναισθήματος:

- *Cohen's Kappa coefficient*

  Θεωρείται πιο στιβαρό (robust) μέτρο από τον απλό ποσοστιαίο βαθμό συμφωνίας, καθώς λαμβάνει υπόψιν του την συμφωνία που συμβαίνει κατά τύχη. Συγκεκριμένα μετράει το βαθμό συμφωνίας μεταξύ *δύο* κριτών που ταξινομούν N αντικείμενα σε C αμοιβαίως αποκλειόμενες κατηγορίες. Η εξίσωση για τον υπολογισμό του είναι:

  $$\varkappa = \frac{p_0 - p_e}{1 - p_e} = 1 - \frac{1 - p_0}{1 - p_e}$$

  όπου p_0 είναι η σχετική παρατηρούμενη ακρίβεια μεταξύ των κριτών και p_e η υποθετική πιθανότητα τυχαίας συμφωνίας, χρησιμοποιώντας τα παρατηρούμενα δεδομένα για τον υπολογισμό της πιθανότητας της τυχαίας επιλογής κάθε μίας κατηγορίας από κάθε έναν από τους κριτές. Αν οι κριτές βρίσκονται σε τέλεια συμφωνία τότε $\varkappa = 1$, διαφορετικά αν δεν υπάρχει συμφωνία μεταξύ των κριτών, πέραν αυτής που αναμένεται λόγω τύχης (και δίνεται από το p_e), τότε $\varkappa \leq 0$.

- *Fleiss' Kappa coefficient*

  Σχετίζεται με τον συντελεστή κάππα του Cohen, με τη διαφορά οτι λειτουργεί για *οποιοδήποτε* πλήθος κριτών που αναθέτουν ένα πλήθος αντικειμένων σε συγκεκριμένες κατηγορίες. Στην ουσία εκφράζει το βαθμό κατά τον οποίο η παρατηρούμενη συμφωνία μεταξύ των κριτών ξεπερνάει αυτήν που θα αναμενόταν αν οι κριτές αξιολογούσαν τα δεδομένα με τελείως τυχαίο τρόπο. Είναι σημαντικό να σημειώσουμε οτι αν και ο συντελεστής κάππα του Cohen υποθέτει οτι οι ίδιοι δύο κριτές έχουν αξιολογήσει ένα σύνολο αντικειμένων, ο συντελεστής κάππα του Fleiss, αν και θεωρεί σταθερό αριθμό κριτών, επιτρέπει διαφορετικά αντικείμενα να αξιολογηθούν από διαφορετικούς κριτές. Ο συντελεστής ορίζεται ως εξής:

  $$\varkappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

  Ο παράγοντας $1 - \bar{P}_e$ δίνει το μέγιστο βαθμό συμφωνίας που είναι εφικτός ανεξαρτήτως τύχης και ο παράγοντας $\bar{P} - \bar{P}_e$ δίνει τον πραγματικό βαθμό συμφωνίας που τελικά παρατηρείται. Όπως και πριν, η τέλεια συμφωνία αντιστοιχεί σε $\varkappa = 1$, ενώ η πλήρης έλλειψη συμφωνίας, πέραν αυτής που αναμένεται λόγω τύχης (και δίνεται από το p_e), αντιστοιχεί σε $\varkappa \leq 0$.

  Η παραπάνω εξίσωση προκύπτει ως εξής: Έστω N ο αριθμός των αντικειμένων, n ο αριθμός των βαθμολογιών ανά αντικείμενο και k ο αριθμός των πιθανών κατηγοριών που μπορούν να ανατεθούν τα

αντικείμενα. Τα αντικείμενα υποδηλώνονται με το δείκτη $i = 1, \dots, N$ και οι κατηγορίες με το δείκτη $j = 1, \dots, k$. Έστω n_{ij} ο αριθμός των βαθμολογητών που ανέθεσαν το $i - οστό$ αντικείμενο στην $j - οστή$ κατηγορία.

Πρώτα υπολογίζουμε την πιθανότητα p_j, που υποδηλώνει το ποσοστό όλων των αναθέσεων στην $j - οστή$ κατηγορία:

$$p_j = \frac{1}{Nn} \cdot \sum_{i=1}^{N} n_{ij} \quad , \quad 1 = \frac{1}{n} \cdot \sum_{j=1}^{k} n_{ij}$$

Μετά υπολογίζουμε την πιθανότητα P_i , που υποδηλώνει το βαθμό στον οποίο συμφωνούν οι βαθμολογητές για το $i - οστό$ αντικείμενο (δηλαδή υπολογίζουμε πόσα είναι τα ζεύγη βαθμολογητών που συμφωνούν σε σχέση με όλα τα πιθανά ζευγάρια βαθμολογητών):

$$P_i = \frac{1}{n(n-1)} \cdot \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) = \frac{1}{n(n-1)} \cdot \left[\left(\sum_{j=1}^{k} {n_{ij}}^2 \right) - (n) \right]$$

Τέλος, υπολογίζουμε την πιθανότητα $\bar{P}$, όπου είναι ο μέσος όρος των P_i και την πιθανότητα $\bar{P}_e$:

$$\bar{P} = \frac{1}{N} \cdot \sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)} \cdot \left(\sum_{i=1}^{N} \sum_{j=1}^{k} {n_{ij}}^2 - Nn \right)$$

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2$$

Οι Landis και Koch πρότειναν τον παρακάτω πίνακα για την ερμηνεία των συντελεστών κάππα. Βεβαίως σε καμία περίπτωση δεν σημαίνει οτι ο παρακάτω πίνακας έχει καθολική αποδοχή.

Kappa	*Agreement*
<0.00	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

Πίνακας 1: Ερμηνεία συντελεστή κάππα του Fleiss

Για πρακτικούς λόγους δίνουμε ένα παράδειγμα στο οποίο φαίνεται πώς εφαρμόζονται οι παραπάνω τύποι.

- ✓ *Παράδειγμα 1ο*

  Έστω οτι έχουμε $n = 14$ βαθμολογητές οι οποίοι πρέπει να αναθέσουν $N = 10$ αντικείμενα σε μία από τις $k = 5$ πιθανές κατηγορίες. Στον παρακάτω πίνακα κάθε στήλη αντιστοιχεί σε μία κατηγορία και κάθε σειρά σε ένα αντικείμενο. Σε κάθε σημείο (i, j) του πίνακα βρίσκεται ο αριθμός των κριτών που συμφωνούν οτι το $i - οστό$ αντικείμενο ανήκει στην $j - οστή$ κατηγορία.

n_{ij}	***1***	***2***	***3***	***4***	***5***	P_i
1	0	0	0	0	14	1.000
2	0	2	6	4	2	0.253
3	0	0	3	5	6	0.308
4	0	3	9	2	0	0.440
5	2	2	8	1	1	0.330
6	7	7	0	0	0	0.462
7	3	2	6	3	0	0.242
8	2	5	3	2	2	0.176
9	6	5	2	1	0	0.286
10	0	2	2	3	7	0.286
Total	20	28	39	21	32	
p_j	0.143	0.200	0.279	0.150	0.229	

Ας υπολογίσουμε ενδεικτικά μερικά μεγέθη:

$$p_1 = \frac{0+0+0+0+2+7+3+2+6+0}{140} = 0.143$$

και

$$P_2 = \frac{1}{14(14-1)} \cdot (0^2 + 2^2 + 6^2 + 4^2 + 2^2 - 14) = 0.253$$

Για να υπολογίσουμε την πιθανότητα $\bar{P}$ θα πρέπει πρώτα να υπολογίσουμε το άθροισμα των P_i, δηλαδή το:

$$\sum_{i=1}^{N} P_i = 1.000 + 0.253 + 0.308 + \cdots + 0.286 = 3.780$$

Άρα:

$$\bar{P} = \frac{1}{10}(3.780) = 0.378$$

$$\bar{P}_e = 0.143^2 + 0.200^2 + 0.279^2 + 0.150^2 + 0.229^2 = 0.213$$

$$\kappa = \frac{0.378 - 0.213}{1 - 0.213} = 0.210$$

2.6.2 Μητρώο σύγχυσης, συνολική ακρίβεια, ακρίβεια και ανάκληση

Μητρώο σύγχυσης

Το μητρώο σύγχυσης (*confusion matrix*) είναι ένας πίνακας που επιτρέπει την οπτικοποίηση της απόδοσης ενός ταξινομητή. Κάθε στήλη του πίνακα αντιπροσωπεύει τα στιγμιότυπα μιας προβλεπόμενης κλάσης, ενώ κάθε γραμμή αντιπροσωπεύει τα στιγμιότυπα μιας πραγματικής κλάσης. Συνεπώς το στοιχείο στη θέση (i, j) αντιπροσωπεύει τον αριθμό των σημείων δεδομένων των οποίων η πραγματική ετικέτα κλάσης ήταν i και ταξινομήθηκαν στην κλάση j. Αν θεωρήσουμε την απλή περίπτωση ενός δυαδικού προβλήματος ταξινόμησης με κατηγορίες C_1 και C_2, τότε το μητρώο σύγχυσης αναπαρίσταται ως εξής:

Actual class	**Predicted class**	
	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$
$\boldsymbol{C_1}$	TP	FN
$\boldsymbol{C_2}$	FP	TN

Πίνακας 2: Μητρώο σύγχυσης στη γενική περίπτωση

όπου:

TP (True Positive): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C_1 και ταξινομήθηκαν στην κλάση C_1.

TN (True Negative): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C_2 και ταξινομήθηκαν στην κλάση C_2.

FP (False Positive): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C_2 και ταξινομήθηκαν στην κλάση C_1.

FN (False Negative): ο αριθμός των δειγμάτων που ανήκαν στην κλάση C_1 και ταξινομήθηκαν στην κλάση C_2.

Από το μητρώο σύγχυσης μπορούν να εξαχθούν άμεσα οι τιμές της συνολικής ακρίβειας (*accuracy*), της ακρίβειας (*precision*) και της ανάκλησης (*recall*). Σημειώνεται οτι στην περίπτωση πολλών κλάσεων η ακρίβεια και η ανάκληση υπολογίζονται για *κάθε μια* από τις κλάσεις ξεχωριστά.

Συνολική ακρίβεια

Η συνολική ακρίβεια (*accuracy*), είναι το ποσοστό των δεδομένων που ταξινομήθηκαν σωστά, δηλαδή:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Ακρίβεια

Είναι το ποσοστό των σημείων δεδομένων που ταξινομήθηκαν στην κλάση i, και των οποίων η πραγματική ετικέτα κλάσης ήταν i. Δηλαδή:

$$Precision(c_1) = \frac{TP}{TP + FP}$$

Στην πραγματικότητα η ακρίβεια μας δείχνει από όλα τα δείγματα που ταξινομήθηκαν στην C_1, πόσα πραγματικά ανήκαν στη C_1, δηλαδή απαντάει στην ερώτηση «Δοθείσας μιας εκτίμησης του ταξινομητή, ποιά η πιθανότητα να είναι σωστή;»

Ανάκληση

Είναι το ποσοστό των σημείων δεδομένων με πραγματική ετικέτα κλάσης i, τα οποία ταξινομήθηκαν επιτυχώς στην κλάση αυτή. Δηλαδή:

$$Recall(c_1) = \frac{TP}{TP + FN}$$

Η ανάκληση μας δείχνει από όλα τα δείγματα που ανήκαν πραγματικά στην κλάση C_1, πόσα από αυτά ταξινομήσαμε επιτυχώς, δηλαδή απαντάει στην ερώτηση «Δοθέντος ενός δείγματος με πραγματική ετικέτα C_1, ποιά η πιθανότητα να το ταξινομήσω σωστά; »

F-measure

Είναι μια μετρική που συνδυάζει την ακρίβεια και την ανάκληση. Συγκεκριμένα είναι ο αρμονικός μέσος των δύο μεγεθών και υπολογίζεται ως:

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Είναι κατά προσέγγιση ο μέσος όρος των δύο όταν είναι αρκετά κοντά και γενικότερα είναι το τετράγωνο του γεωμετρικού μέσου διαιρεμένου με τον αριθμητικό μέσο. Η μέγιστη τιμή του είναι 1 και η ελάχιστη 0. Είναι επίσης γνωστό σαν $F_1 - measure$, καθότι η ακρίβεια και η ανάκληση σταθμίζονται ισόποσα. Είναι μια ειδική περίπτωση του γενικού $F_\beta - measure$ (για μη αρνητικές τιμές του β):

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

Οι άλλες δύο συχνότερα χρησιμοποιούμενες εκδοχές είναι το $F_2 - measure$ που σταθμίζει περισότερο την ανάκληση σε σχέση με την ακρίβεια και το $F_{0.5} - measure$ που δίνει μεγαλύτερη έμφαση στην ακρίβεια σε σχέση με την ανάκληση.

Παρατηρήσεις

Αν και η συνολική ακρίβεια ενός συστήματος είναι ίσως η πιο συχνά χρησιμοποιούμενη μετρική απόδοσης, αφού αποτυπώνει την απόδοση του συστήματος σε έναν και μόνο αριθμό, μπορεί πολλές φορές να αποδειχτεί εξαιρετικά παραπλανητική. Το ίδιο ισχύει και για το F-measure. Αυτό συμβαίνει γιατί και οι δύο προαναφερθείσες μετρικές αδυνατούν να αποτυπώσουν την ελλατωματική συμπεριφορά που οφείλεται σε έναν προκατειλημμένο (biased) ταξινομητή (βλέπε παραδείγματα που ακολουθούν). Συνεπως για την καλύτερη κατανόηση και ερμηνεία ενός συστήματος είναι απαραίτητος ο υπολογισμός τόσο της ακρίβειας (precision) όσο και της ανάκλησης (recall) για *κάθε μία* από τις κλάσεις.

Προκειμένου να γίνει κατανοητή η σημασία του υπολογισμού της ακρίβειας και της ανάκλησης για κάθε κλάση, αλλά και ο τρόπος χρήσης των παραπάνω μετρικών σε ένα σύστημα Ανάλυσης Συναισθήματος, παραθέτουμε δύο παραδείγματα.

- ✓ *Παράδειγμα 2ο*

  Έστω ένα σύστημα Ανάλυσης Συναισθήματος το οποίο ταξινομεί πάντα ένα κείμενο ως θετικό. Ας υποθέσουμε οτι διαθέτουμε ένα σύνολο 100 κειμένων, εκ των οποίων τα 95 είναι θετικά και τα 5 είναι αρνητικά. Προφανώς ένα τέτοιο σύστημα θα ταξινομούσε όλα τα κείμενα ως θετικά και το αντίστοιχο μητρώο σύγχυσης θα ήταν:

Πραγματική κλάση	**Προβλεπόμενη κλάση**	
	$\boldsymbol{C_1}$ ($\Theta\varepsilon\tau\iota\kappa\acute{o}$)	$\boldsymbol{C_2}$ ($\Alpha\rho\nu\eta\tau\iota\kappa\acute{o}$)
$\boldsymbol{C_1}$ ($\Theta\varepsilon\tau\iota\kappa\acute{o}$)	95	0
$\boldsymbol{C_2}$ ($\Alpha\rho\nu\eta\tau\iota\kappa\acute{o}$)	5	0

Πίνακας 3: Μητρώο σύγχυσης - Παράδειγμα 1ο

Βάση του παραπάνω πίνακα προκύπτουν τα μεγέθη:

Accuracy	***Precision***		***Recall***		$\boldsymbol{F_1}$	
	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$
$\frac{95}{100} = 0.95$	$\frac{95}{100} = 0.95$	–	$\frac{95}{95} = 1$	$\frac{0}{5} = 0$	0.97	–

Πίνακας 4: Παράμετροι αξιολόγησης - Παράδειγμα 1ο

Αν και το παραπάνω σύστημα φαίνεται να είναι εξαιρετικά ακριβές, αφού ταξινόμησε σωστά το 95% των περιπτώσεων, κάτι τέτοιο απέχει παρασάγγας από την πραγματικότητα, καθώς δεν μπορεί να εντοπίσει τα αρνητικά κείμενα. Αυτό γίνεται σαφές αν κανείς παρατηρήσει την

ανάκληση για τη δεύτερη κλάση, η οποία λόγω της προκατάληψης του ταξινομητή ως προς τη θετική κλάση είναι 0. Αυτό το παράδειγμα καθιστά κατανοητό πως η συνολική ακρίβεια δεν θα πρέπει σε καμία περίπτωση να θεωρείται πανάκεια, αλλά αντιθέτως θα πρέπει να συνδυάζεται με τα υπόλοιπα μεγέθη για την εξαγωγή σωστών συμπερασμάτων.

- ✓ *Παράδειγμα 3ο*

Έστω οτι κατά την αξιολόγηση ενός συστήματος προκύπτει το παρακάτω μητρώο σύγχυσης:

Πραγματική κλάση	**Προβλεπόμενη κλάση**		
	$\boldsymbol{C_1}$ ($\Theta\varepsilon\tau\iota\kappa\acute{o}$)	$\boldsymbol{C_2}$ ($A\rho\nu\eta\tau\iota\kappa\acute{o}$)	$\boldsymbol{C_3}$ ($O\upsilon\delta\acute{\varepsilon}\tau\varepsilon\rho o$)
$\boldsymbol{C_1}$ ($\Theta\varepsilon\tau\iota\kappa\acute{o}$)	47	5	18
$\boldsymbol{C_2}$ ($A\rho\nu\eta\tau\iota\kappa\acute{o}$)	3	29	19
$\boldsymbol{C_3}$ ($O\upsilon\delta\acute{\varepsilon}\tau\varepsilon\rho o$)	17	21	41

Πίνακας 5: Μητρώο σύγχυσης - Παράδειγμα 2ο

Τότε βάση του μητρώου προκύπτουν τα παρακάτω μεγέθη:

Accuracy	***Precision***			***Recall***		
	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	$\boldsymbol{C_3}$	$\boldsymbol{C_1}$	$\boldsymbol{C_2}$	$\boldsymbol{C_3}$
$\frac{117}{200} = 0.585$	$\frac{47}{67} = 0.701$	$\frac{29}{55} = 0.527$	$\frac{41}{78} = 0.526$	$\frac{47}{70} = 0.671$	$\frac{29}{51} = 0.569$	$\frac{41}{79} = 0.519$

Πίνακας 6: Παράμετροι αξιολόγησης – Παράδειγμα 2ο

Τα αποτελέσματα αυτά δηλώνουν τα εξής:

1) Αν ένα κείμενο ταξινομηθεί ως θετικό υπάρχει πιθανότητα 70.1% το σύστημα να έχει κάνει σωστή πρόβλεψη. Από την άλλη, η πιθανότητα να ταξινομηθεί ένα θετικό κείμενο επιτυχώς είναι 67.1%.

2) Αν ένα κείμενο ταξινομηθεί ως αρνητικό υπάρχει πιθανότητα 52.7% το σύστημα να έχει κάνει σωστή πρόβλεψη. Από την άλλη, η πιθανότητα να ταξινομηθεί ένα αρνητικό κείμενο επιτυχώς είναι 56.9%.

3) Αν ένα κείμενο ταξινομηθεί ως ουδέτερο υπάρχει πιθανότητα 52.6% το σύστημα να έχει κάνει σωστή πρόβλεψη. Από την άλλη, η πιθανότητα να ταξινομηθεί ένα ουδέτερο κείμενο επιτυχώς είναι 51.9%.

Κεφάλαιο 3 Υλοποίηση

3.1 Περιγραφή αλγορίθμου

Όπως αναφέρθηκε και στα προηγούμενα κεφάλαια, στην παρούσα διπλωματική υλοποιούμε ένα αυτόματο σύστημα Ανάλυσης Συστήματος βασισμένο σε λεξικό συναισθήματος, το οποίο συλλέγει και ταξινομεί tweets κοινωνικοπολιτικής φύσεως, σε μιά από τις τρείς πιθανές κλάσεις: αρνητικό, θετικό ή ουδέτερο. Ο αλγόριθμος που αναπτύξαμε αποτελείται από 7 στάδια:

- ✓ *Συλλογή δεδομένων (data acquisition)*: Δημιουργούμε ένα σύνολο αποτελούμενο από σχεδόν 20000 tweets μέσω της εφαρμογής Twitter API.

- ✓ *Προεπεξεργασία δεδομένων (data pre-processing):* Φιλτράρουμε τα tweets προκειμένου να αφαιρέσουμε το θόρυβο – άχρηστη πληροφορία και κρατάμε μόνο τους σημαντικούς όρους, δηλαδή πρακτικά μόνο τις λέξεις που έχουν συναισθηματικό περιεχόμενο.

- ✓ *Μετασχηματισμός δεδομένων (data transformation):* Μετασχηματίζουμε τα δεδομένα μας, ώστε να μπορούμε να τα εντοπίσουμε στο λεξικό συναισθήματος.

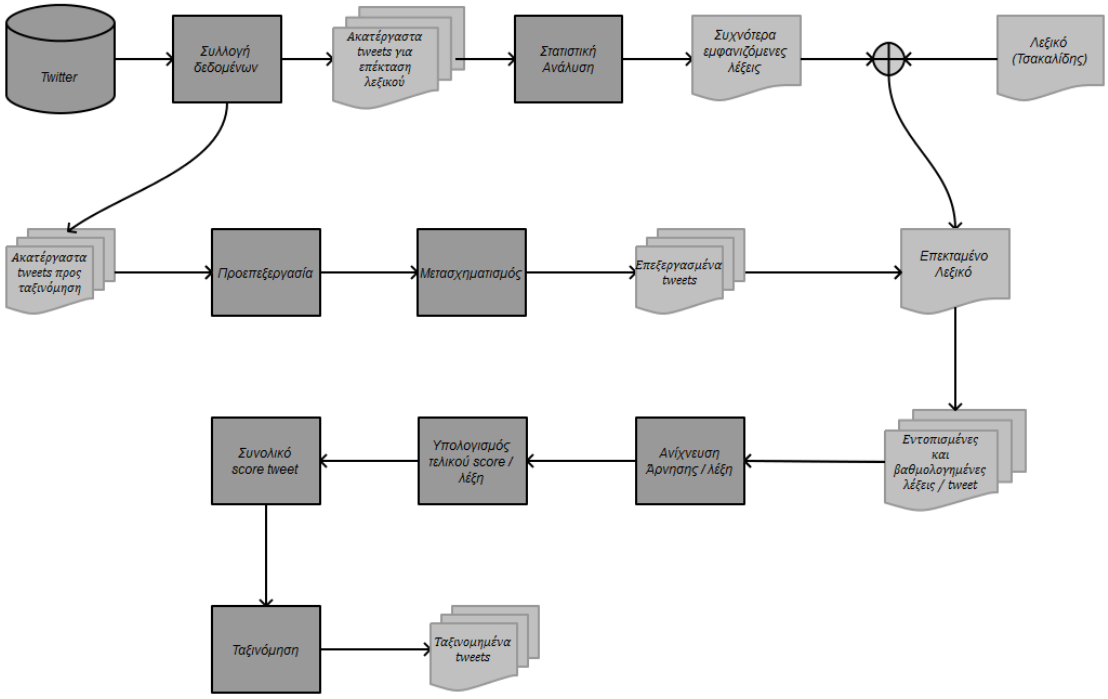
- ✓ *Δημιουργία – Επέκταση λεξικού:* Χρησιμοποιώντας ένα ήδη υπάρχον λεξικό συναισθήματος και διεξάγωντας στατιστική ανάλυση σε ένα τμήμα των tweets που συλλέξαμε στο πρώτο στάδιο, εντοπίζουμε τις συχνότερα εμφανιζόμενες λέξεις και δημιουργούμε ένα νέο επεκταμένο λεξικό συναισθήματος για όρους της ελληνικής γλώσσας, που επικεντρώνεται σε κοινωνικοπολιτικά θέματα.

- ✓ *Εντοπισμός λέξεων:* Αφού φιλτράρουμε και μετασχηματίσουμε τα tweets εντοπίζουμε τις λέξεις που έχουν αποθηκευτεί στο λεξικό και τους αναθέτουμε την αντίστοιχη βαθμολογία.

- ✓ *Ανίχνευση άρνησης (negation detection):* Για κάθε λέξη που εντοπίσαμε στο λεξικό, προσδιορίζουμε το κατά πόσο εντάσσεται σε σχήμα άρνησης και συνεπώς υπάρχει αντιστροφή πολικότητας.

- ✓ *Υπολογισμός συνολικού score και ταξινόμηση:* Υπολογίζουμε το συνολικό σκορ κάθε tweet και το ταξινομούμε αναλόγως σε μια από τις τρείς κλάσεις.

Στο παρακάτω διάγραμμα ροής φαίνονται συνοπτικά όλα τα βήματα του αλγορίθμου που υλοποιήσαμε:



Εικόνα 1: Διάγραμμα ροής αλγορίθμου

3.1.1 Υλοποίηση

Για την υλοποίηση του αλγορίθμου χρησιμοποιήσαμε δύο προγραμματιστικά εργαλεία. Η συλλογή δεδομένων έγινε μέσω Python scripts χρησιμοποιώντας τη βιβλιοθήκη Tweepy, η οποία είναι μια βιβλιοθήκη ελεύθερου λογισμικού[1] που επιτρέπει στην Python να επικοινωνήσει με την πλατφόρμα του Twitter και να χρησιμοποιήσει το API που προσφέρει. Απο εκεί και πέρα το κυρίως σώμα του αλγορίθμου, δηλαδή όλα τα στάδια από την προεπεξεργασία των δεδομένων μέχρι και την τελική ταξινόμηση, υλοποιήθηκαν στο προγραμματιστικό περιβάλλον του Matlab.

[1] https://github.com/tweepy/tweepy

3.2 Συλλογή δεδομένων

Το πρώτο κομμάτι του αλγορίθμου αφορά τη συλλογή δεδομένων, η πλειοψηφία των οποίων θα χρησιμοποιηθεί για την επέκταση του λεξικού συναισθήματος και τα υπόλοιπα για την αξιολόγηση της απόδοσης του ταξινομητή.

Το Twitter δίνει τη δυνατότητα στους χρήστες του να συνδέσουν τον ιστότοπό τους ή την εφαρμογή τους με την παγκόσμια συζήτηση που λαμβάνει χώρα, μέσω του API που παρέχει[2]. Στην πραγματικότητα παρέχει μια πλειάδα APIs, εκ των οποίων χρησιμοποιήσαμε τις παρακάτω:

- *REST API:* Παρέχει προγραμματιστική πρόσβαση ώστε να διαβάζεις και να γράφεις δεδομένα. Μπορείς να συντάξεις ένα νέο tweet (post), να διαβάσεις πληροφορίες του προφίλ διάφορων χρηστών, να διεξάγεις σύνθετες αναζητήσεις κ.α.

- *Streaming API:* Δίνει στους προγραμματιστές (developers) πρόσβαση χαμηλής καθυστέρησης (low latency access) στην παγκόσμια ροή δεδομένων Tweet μέσω επίμονων συνδέσεων HTTP (persistent HTTP connections). Επιτρέπει τη λήψη ενημερώσεων πάνω στα πιο πρόσφατα tweets που ταιριάζουν (match) με μια συγκεκριμένη αναζήτηση, το συγχρονισμό με τις ενημερώσεις των προφίλ των χρηστών κ.α. Γενικότερα η χρήση του Streaming API ενδείκνυται για παρακολούθηση και επεξεργασία tweets σε πραγματικό χρόνο (real time).

.

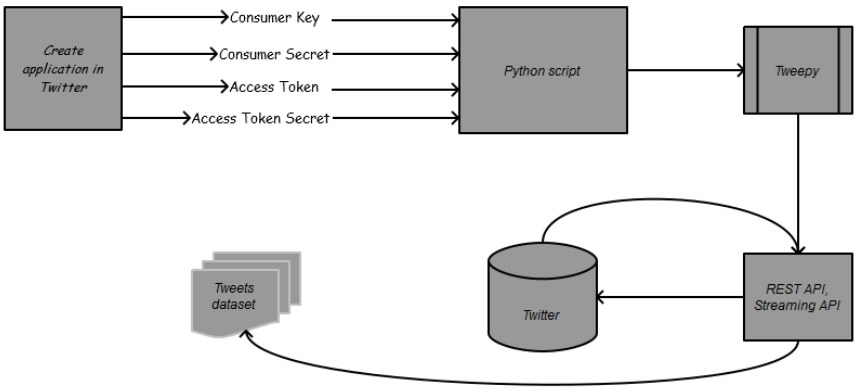
Σε όλες τις περιπτώσεις η επικοινωνία γίνεται μέσω του πρωτοκόλλου OAuth[3], το οποίο συνδέει τους χρήστες στο Twitter και τους επιτρέπει να στείλουν ασφαλείς, εξουσιοδοτημένες αιτήσεις (secure, authorized requests) στο Twitter API.

Στις παραγράφους που ακολουθούν δίνουμε, για λόγους πληρότητας, μια σύντομη περιγραφή του πρωτοκόλλου OAuth και της εφαρμογής που δημιουργήσαμε προκειμένου να μπορέσουμε να επικοινωνήσουμε μέσω κάποιων Python scripts με το API και να «τραβήξουμε» δεδομένα είτε σε πραγματικό χρόνο είτε αναζητώντας tweets συγκεκριμένων χρηστών και περιγράφουμε το σύνολο των δεδομένων που συλλέξαμε. Πριν από αυτό όμως παραθέτουμε ένα διάγραμμα ροής στο οποίο γίνεται αντιληπτή όλη η διαδικασία που ακολουθήσαμε για τη συλλογή δεδομένων μέσω του API. Τα επιμέρους στοιχεία που απαρτίζουν το διάγραμμα θα γίνουν κατανοητά στη συνέχεια.

---

[2] https://dev.twitter.com/overview/documentation

[3] http://hueniverse.com/oauth/



Εικόνα 2: Διάγραμμα ροής της διαδικασίας συλλογής δεδομένων

3.2.1 Twitter API

3.2.1.1 Πρωτόκολλο OAuth

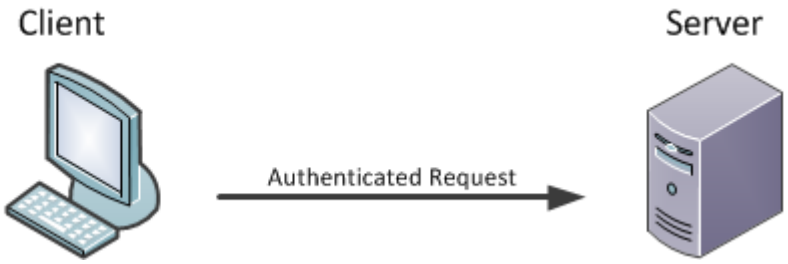
Καθώς το διαδίκτυο μεγαλώνει με συνεχώς αυξανόμενους ρυθμούς, όλο και περισσότεροι ιστότοποι βασίζονται σε κατανεμημένες υπηρεσίες (distributed services) και υπηρεσίες cloud (cloud computing), όπως για παράδειγμα ένα κοινωνικό δίκτυο που χρησιμοποιεί το λογαριασμό ενός χρήστη στη Google για να ψάξει για φίλους ή μια τρίτη εφαρμογή (third-party application) που αξιοποιεί APIs από διάφορες υπηρεσίες. Το πρόβλημα είναι οτι προκειμένου αυτές οι υπηρεσίες να έχουν πρόσβαση στα δεδομένα χρηστών που βρίσκονται σε άλλα sites, ζητάνε τα ονόματα χρηστών (usernames) και τους κωδικούς πρόσβασης(passwords). Αυτό οδηγεί στην έκθεση των κωδικών των χρηστών σε τρίτους, γεγονός που μπορεί δυνητικά να δημιουργήσει σοβαρά προβλήματα παραβίασης προσωπικών δεδομένων.

Το πρωτόκολλο OAuth παρέχει μια μέθοδο ώστε οι χρήστες να μπορούν να παραχωρήσουν πρόσβαση σε τρίτους (third-party access) στους πόρους τους (resources), χωρίς να μοιραστούν τους κωδικούς τους. Επιπλέον παρέχει έναν τρόπο παραχώρησης περιορισμένης πρόσβασης (limited access). Το πρωτόκολλο OAuth ορίζει τρείς ρόλους: του πελάτη (client), του εξυπηρετητή (server) και του ιδιοκτήτη πόρων (resource owner). Αυτοί οι τρείς ρόλοι είναι παρόντες σε κάθε συναλλαγή (transaction) OAuth.

Στο παραδοσιακό μοντέλο πιστοποίησης πελάτη-εξυπηρετητή (client server authentication model) ο πελάτης χρησιμοποιεί τα πιστοποιητικά *του* για να αποκτήσει πρόσβαση στους πόρους *του* που φιλοξενούνται από τον εξυπηρετητή.

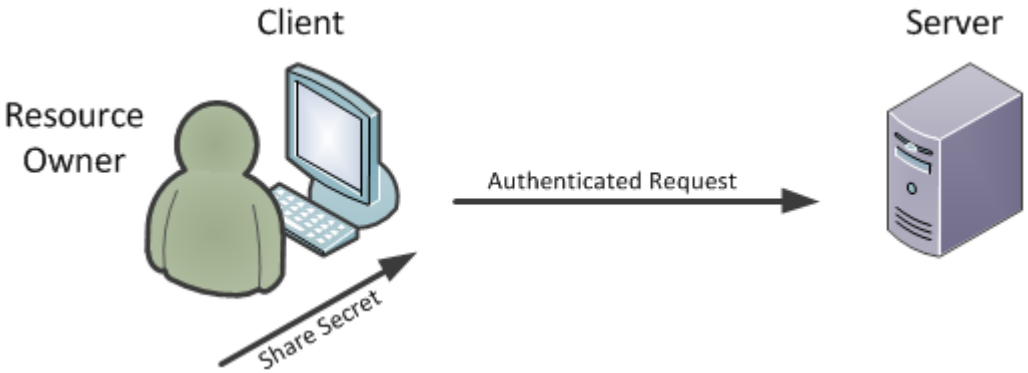
Σε ό,τι αφορά τον εξυπηρετητή, το κοινό μυστικό (shared secret) που χρησιμοποιείται από τον πελάτη ανήκει στον πελάτη και ο εξυπηρετητής δεν ενδιαφέρεται για το από πού προέρχεται ή για το εάν ο πελάτης δρα εκ μέρους

κάποιας άλλης οντότητας. Εφόσον το κοινό μυστικό ταιριάζει με αυτό που «περιμένει» ο εξυπηρετητής, το αίτημα τίθεται προς επεξεργασία.



Εικόνα 3: Μοντέλο πελάτη - εξυπηρετητή

Το πρωτόκολλο OAuth εισάγει έναν νέο ρόλο σε αυτό το μοντέλο: του ιδιοκτήτη των πόρων (resource owner). Στο μοντέλο του OAuth, ο πελάτης (client), ο οποίος δεν είναι ο ιδιοκτήτης πόρων αλλά δρα εκ μέρους του ιδιοκτήτη, αιτείται πρόσβαση στους πόρους που ελέγχονται από τον ιδιοκτήτη πόρων, αλλά φιλοξενούνται από τον εξυπηρετητή. Αντί λοιπόν ο πελάτης να χρησιμοποιήσει δικά του πιστοποιητικά, χρησιμοποιεί τα πιστοποιητικά του ιδιοκτήτη των πόρων, παριστάνοντας στην ουσία τον ιδιοκτήτη.



Εικόνα 4: Μοντέλο ιδιοκτήτη πόρων - πελάτη - εξυπηρετητή

Το πρωτόκολλο OAuth χρησιμοποιεί τριών ειδών πιστοποιητικά: πιστοποιητικά πελάτη (client credentials), προσωρινά πιστοποιητικά (temporary credentials) και πιστοποιητικά σκυτάλης (token credentials) τα οποία αρχικά είχαν οριστεί ως consumer key and secret (client credentials), request token and secret (temporary credentials) και access token and secret (token credentials).

Τα πιστοποιητικά πελάτη χρησιμοποιούνται για να πιστοποιήσουν τη γνησιότητα (authenticate) του πελάτη. Τα πιστοποιητικά σκυτάλης χρησιμοποιούνται στη θέση του ονόματος χρήστη και του κωδικού πρόσβασης του ιδιοκτήτη των πόρων. Αντί λοιπόν να χρειαστεί να μοιραστεί ο ιδιοκτήτης πόρων τα πιστοποιητικά του με τον πελάτη, εξουσιοδοτεί τον εξυπηρετητή να διανείμει μια ειδική τάξη (class) πιστοποιητικών στον πελάτη που αντιπροσωπεύουν την παραχώρηση πρόσβασης που δόθηκε στον πελάτη από τον ιδιοκτήτη πόρων. Ο πελάτης χρησιμοποιεί τα πιστοποιητικά σκυτάλης για να

αποκτήσει πρόσβαση στους προστατευόμενους πόρους χωρίς να χρειάζεται να μάθει τον κωδικό του ιδιοκτήτη.

Προφανώς όλες οι βιβλιοθήκες πελατών του Twitter (Twitter Client Libraries), όπως η βιβλιοθήκη Tweepy στην Python, έχουν ενσωματωμένο το πρωτόκολλο OAuth και συνεπώς ο προγραμματιστής δε χρειάζεται να ασχοληθεί με τις λεπτομέρειες υλοποίησής του. Παρόλα αυτά η κατανόηση του τρόπου λειτουργίας του πρωτοκόλλου μπορεί να βοηθήσει στη δημιουργία και την αποσφαλμάτωση (debugging) εφαρμογών που χρησιμοποιούν το API του Twitter.

3.2.1.2 Δημιουργία εφαρμογής

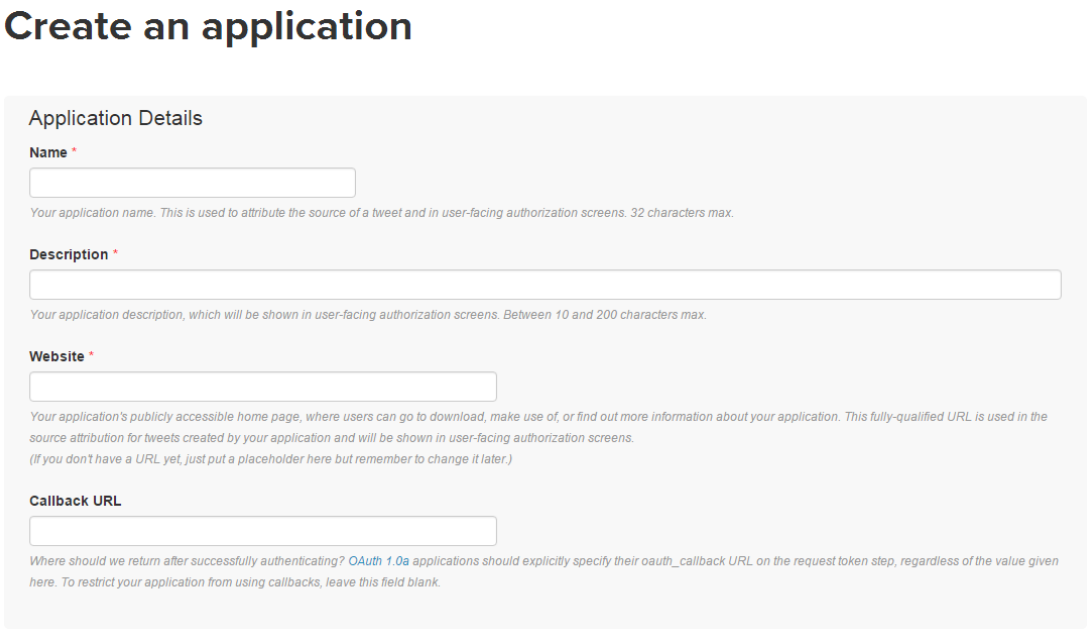
Προκειμένου να κάνεις εξουσιοδοτημένες (authorized) κλήσεις στα διάφορα API's του Twitter, πρέπει να δημιουργήσεις μια εφαρμογή στην οποία θα παραχωρηθεί μια σκυτάλη πρόσβασης OAuth (OAuth access token) έτσι ώστε να μπορεί να δρα εκ μέρους σου, με την προϋπόθεση οτι είσαι ήδη χρήστης του Twitter. Ο τρόπος με τον οποίο παραχωρούνται τέτοιες σκυτάλες (tokens) διαφέρει ανάλογα με την περίπτωση[4]. Εν προεκειμένω, επειδή το μόνο που θέλουμε είναι να έχουμε πρόσβαση στο API από τον προσωπικό μας λογαριασμό, αρκεί να δημιουργήσουμε μια εφαρμογή στο Twitter Developer Site - κέντρο ελέγχου εφαρμογών[5] (απαιτείται να είμαστε συνδεδεμένοι μέσω του λογαριασμού μας στο Twitter). Στην αρχική σελίδα του site εμφανίζονται όλες οι εφαρμογές που έχουμε δημιουργήσει μέχρι στιγμής:

Εικόνα 5: Κέντρο ελέγχου εφαρμογών στο Twitter

Αν επιλέξουμε πάνω δεξιά τη δημιουργία νέας εφαρμογής μας εμφανίζεται η παρακάτω φόρμα:

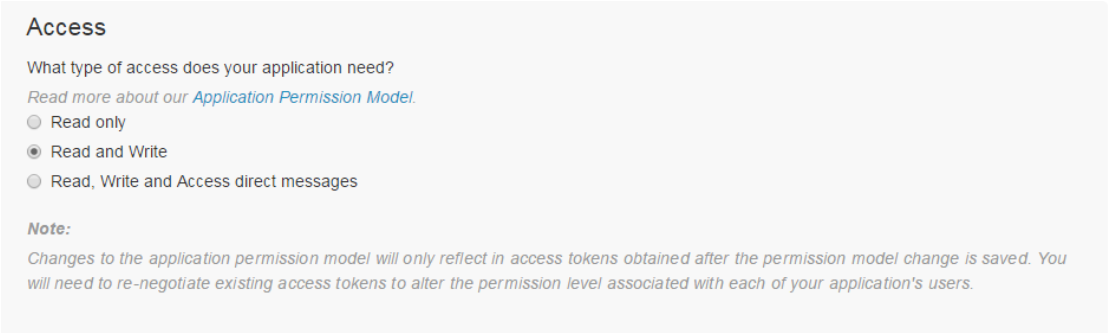
[4] https://dev.twitter.com/oauth/overview

[5] https://apps.twitter.com/



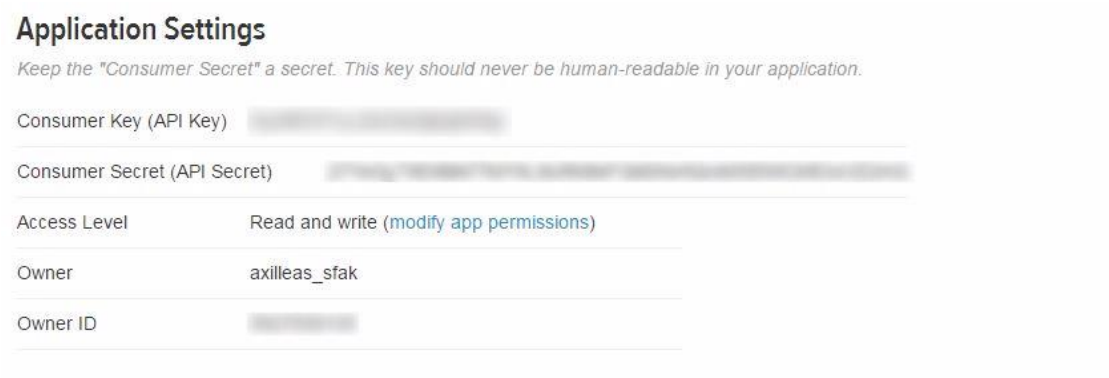
Εικόνα 6: Φόρμα δημιουργίας εφαρμογής στο Twitter

Αφού συμπληρώσουμε τη φόρμα και ολοκληρώσουμε τη δημιουργία της νέας εφαρμογής, ανακατευθυνόμαστε στη σελίδα ρυθμίσεων, στην οποία μπορούμε να ρυθμίσουμε τον τύπο της πρόσβασης που επιθυμούμε ανάλογα με το αν θα μπορεί η εφαρμογή απλά να διαβάζει μηνύματα εκ μέρους μας (read only), να διαβάζει και να γράφει μηνύματα (post) εκ μέρους μας (read and write) ή να έχει και πρόσβαση σε απευθείας μηνύματα (read, write and access direct messages).

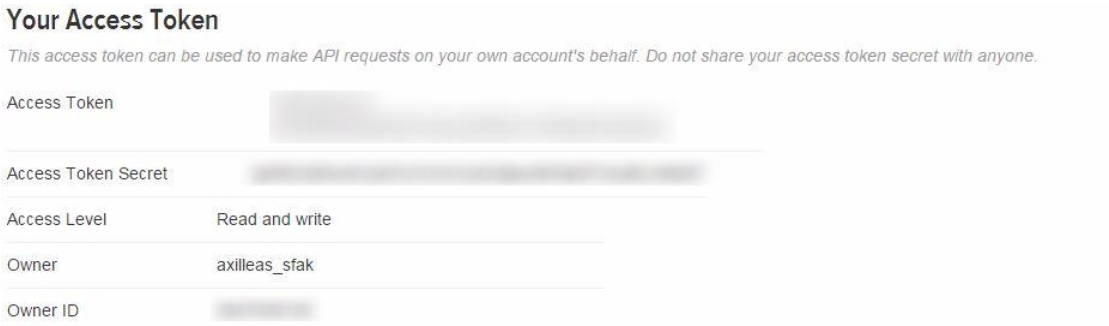


Εικόνα 7: Ορισμός τύπου πρόσβασης εφαρμογής

Μετά από όλα αυτά, είμαστε έτοιμοι να δημιουργήσουμε τα κλειδιά (Consumer key (API key), Consumer Secret (API secret)) και τις σκυτάλες πρόσβασης της εφαρμογής (Access Token, Access Token Secret), τα οποία είναι απαραίτητα για την πιστοποίηση της εφαρμογής μας από το Twitter και θα μας επιτρέψουν στη συνέχεια να κάνουμε αιτήσεις στο API (μέσω των Python scripts) εκ μέρους του λογαριασμού μας.



Εικόνα 8: Πιστοποιητικά πελάτη



Εικόνα 9: Πιστοποιητικά σκυτάλης

3.2.2 Σύνολο δεδομένων

Προκειμένου να εντοπίσουμε όσο το δυνατόν περισσότερες λέξεις με συναισθηματικό υπόβαθρο για την επέκταση του λεξικού, αλλά και για να συγκεντρώσουμε έναν ικανοποιητικό όγκο από tweets για να αξιολογήσουμε την απόδοση του ταξινομητή μας, δημιουργήσαμε ένα σύνολο που αποτελείται από περίπου 22.000 tweets (22.558 για την ακρίβεια).

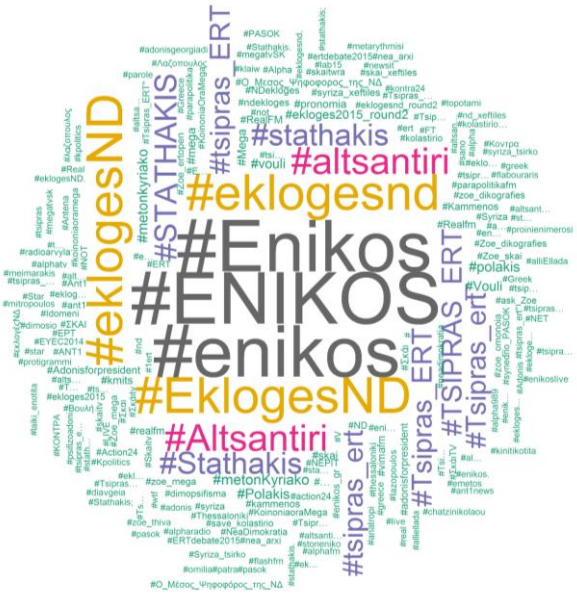
Το σύνολο αυτό είναι στη πραγματικότητα η ένωση πολλών επιμέρους υποσυνόλων tweets τα οποία δημιουργήθηκαν είτε κατεβάζοντας tweets συγκεκριμένων χρηστών, είτε tweets που αφορούσαν ένα συγκεκριμένο θέμα. Πιο συγκεκριμένα, έχοντας πάντα δεδομένο οτι το σύστημα που υλοποιούμε προσανατολίζεται σε tweets κοινωνικοπολιτικού περιεχομένου και συνεπώς πρέπει να συλλέξουμε tweets με αντίστοιχο περιεχόμενο και λεξιλόγιο κάναμε τα εξής:

1) Επιλέξαμε κάποιους Έλληνες πολιτικούς και δημοσιογράφους, οι οποίοι παρουσιάζουν αρκετά έντονη δραστηριότητα στο Twitter και κατεβάσαμε μέρος ή το σύνολο των tweets που έχουν ποστάρει. Η επιλογή δε βασίστηκε, προφανώς, σε πολιτικά κριτήρια, παρά μόνο στο πλήθος των tweets που είχε ποστάρει ο κάθε υποψήφιος. Μερικά παραδείγματα χρηστών είναι οι: Κυριάκος Μητσοτάκης

(@kmitsotakis), Γιάννης Βαρουφάκης (@yanisvaroufakis), Ζωή Κωνσταντοπούλου (@ZoeKonstant), Στέφανος Μάνος (@StefanosManos), Γιάννης Βρούτσης (@VroutsisGiannis) και Χάρης Θεοχάρης (@htheoharis).

2) Κάθε φορά που εμφανιζόταν στην επικαιρότητα ένα καινούριο φλέγον πολιτικό ζήτημα (πχ μεταναστευτικό, ζητήματα πολιτικής διαφθοράς, εσωκομματικές εκλογές κλπ) ή λάμβανε χώρα ένα γεγονός που απτόταν της πολιτικής και προκαλούσε έντονες αντιδράσεις και συζητήσεις (πχ κάποια δήλωση ή συνέντευξη πολιτικού προσώπου) στοχεύαμε συγκεκριμένα *hashtags* και μέσω του streaming API αποθηκεύαμε όλες αυτές τις μεγάλες ροές δεδομένων. Να σημειώσουμε στο σημείο αυτό οτι επειδή το API δεν επιτρέπει τη συλλογή δεδομένων παλαιότερων της μιας εβδομάδας (ή και πιο πρόσφατων αν υπάρχουν πολύ μεγάλες ροές για ένα θέμα), έπρεπε αναγκαστικά να φροντίσουμε να κατεβάσουμε τα tweets σε ένα διάστημα 1-2 ημερών από τη στιγμή που άρχισε η ροή τους. Τα δεδομένα αυτά συλλέχτηκαν λοιπόν σε ένα διάστημα σχεδόν τριών μηνών, από αρχές Νοεμβρίου του 2015 μέχρι τέλη Ιανουαρίου του 2016.

Για να πάρει μια αίσθηση ο αναγνώστης του ποιά ήταν τα πιο δημοφιλή hashtags που εμφανίζονταν στα tweets που συλλέξαμε, θα χρησιμοποιήσουμε τα λεγόμενα σύννεφα λέξεων (*word clouds*)[6]. Τα σύννεφα λέξεων είναι ένας τρόπος οπτικής αναπαράστασης δεδομένων από κείμενα και χρησιμοποιούνται συνήθως για μια γρήγορη επισκόπηση των συχνότερα εμφανιζόμενων λέξεων σε ένα κείμενο, όπου η σημασία κάθε λέξης αποτυπώνεται είτε με το μέγεθος είτε με το χρώμα της λέξης, είτε και με τα δύο.



Εικόνα 10: Word cloud των πιο συχνά χρησιμοποιούμενων hashtags στα tweets που συλλέξαμε (generated in R)

[6] https://en.wikipedia.org/wiki/Tag_cloud

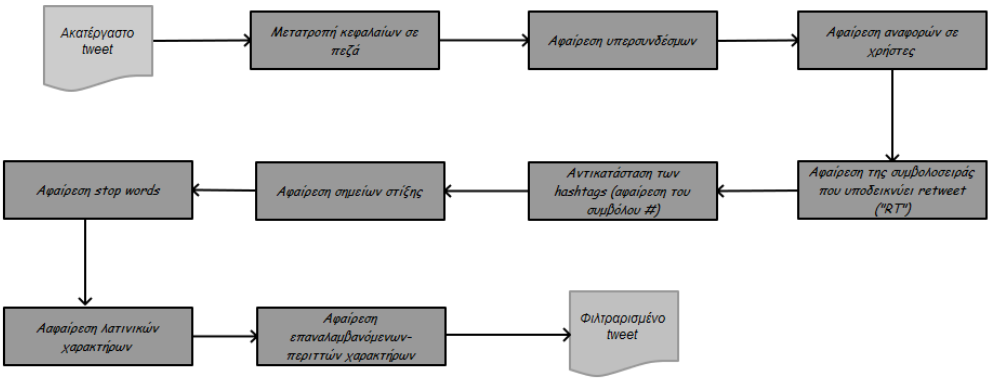
3.3 Προεπεξεργασία δεδομένων

Όπως αναφέρθηκε και στην παράγραφο 2.3 ο περιορισμός των 140 χαρακτήρων στο Twitter αποτελεί την αφορμή για ένα σύνολο ιδιαίτερων φαινομένων όπως συντομογραφίες, ακρωνύμια και γενικότερα χρήση μη συμβατικών λέξεων ή φράσεων, κάτι που σε συνδυασμό με την ιδιαίτερη φύση του και τη χρήση hashtags, emoticons και urls, δημιουργεί μηνύματα που πάσχουν από το πρόβλημα της αραιότητας (sparsity problem) [22]. Με λίγα λόγια τα tweets περιέχουν μεγάλες ποσότητες θορύβου και λέξεις που δεν συμβάλλουν στο συναισθηματικό περιεχόμενο της πρότασης.

Εάν κρατούσαμε όλες αυτές τις λέξεις θα αυξάναμε τη διάσταση του προβλήματος και θα δυσκολεύαμε την ταξινόμηση αφού θα χειριζόμασταν κάθε μια από τις λέξεις του μηνύματος ως μία διάσταση του προβλήματος. Εδώ έγκειται η ουσία της σωστής προεπεξεργασίας των δεδομένων: μειώνοντας το θόρυβο των tweets βελτιώνουμε την απόδοση του ταξινομητή και επιταχύνουμε τη διαδικασία ταξινόμησης, δηλαδή βελτιώνουμε τόσο τη χωρική όσο και τη χρονική πολυπλοκότητα του προβλήματος. Η αυξημένη πολυπλοκότητα θα καθιστούσε απαγορευτική τη χρήση ένος συστήματος Ανάλυσης Συναισθήματος σε προβλήματα πραγματικού χρόνου (real time sentiment analysis) που θα απαιτούσαν τον εξαιρετικά γρήγορο και αποτελεσματικό χειρισμό τεράστιων όγκων δεδομένων πραγματικού χρόνου.

Συνεπώς το στάδιο της προεπεξεργασίας είναι απαραίτητο για τον καθαρισμό και την προετοιμασία των δεδομένων για ταξινόμηση.

Ο αλγόριθμος προεπεξεργασίας δεδομένων που υλοποιήσαμε αποτελείται από 9 στάδια. Η διαδικασία προεπεξεργασίας φαίνεται σχηματικά στο παρακάτω διάγραμμα ροής:



Εικόνα 11: Διάγραμμα ροής του σταδίου προεπεξεργασίας δεδομένων

Ακολουθεί αναλυτική περιγραφή κάθε σταδίου:

- *<u>Στάδιο 1ο</u> Μετατροπή κεφαλαίων σε πεζά*

  Η χρήση κεφαλαίων γραμμάτων για λόγους έμφασης και πολλές φορές η ανακόλουθη χρήση αυτών (πχ ΚΑΛΗΜΕρα) είναι δύο πολύ συνηθισμένα φαινόμενα στο Twitter. Δεδομένου όμως οτι η έμφαση δεν επηρεάζει την πολικότητα των λέξεων παρά μόνο ενισχύει το συναίσθημα, και οτι η αντιστοίχιση των λέξεων των tweets με αυτές του λεξικού απαιτεί σταθερό casing, δημιουργείται η ανάγκη μετατροπής όλων των κεφαλαίων χαρακτήρων σε πεζούς.

- *<u>Στάδιο 2ο</u> Αφαίρεση των υπερσυνδέσμων*

  Οι υπερσύνδεσμοι σε ένα tweet συνήθως οδηγούν σε ένα νέο ιστότοπο που περιέχει ακατέργαστο κείμενο και ενδεχομένως και άλλου είδους πολυμεσικό περιεχόμενο όπως εικόνες και βίντεο. Εφόσον δεν υλοποιούμε ένα σύστημα που ανιχνεύει το ευρύτερο σημασιολογικό πλαίσιο του tweet ερμηνεύοντας πολυμεσικό περιεχόμενο, οι υπερσύνδεσμοι δεν μας προσφέρουν καμία απολύτως πληροφορία και συνεπώς ο αλγόριθμός μας όταν εντοπίσει μια συμβολοσειρά που ξεκινάει με *“www.”, “http://”* ή *“https://”*, αυτομάτως την αφαιρεί.

- *<u>Στάδιο 3ο</u> Αφαίρεση των αναφορών σε χρήστες (“@username”)*

  Όπως και οι υπερσύνδεσμοι, έτσι και οι αναφορές σε χρήστες δεν παρέχουν καμία πληροφορία και αφαιρούνται πάραυτα.

- *<u>Στάδιο 4ο</u> Αφαίρεση της συμβολοσειράς που υποδηλώνει retweet (“RT”)*

  Για τους ίδιους λόγους με τα στάδια 2,3 η συμβολοσειρά “RT” αφαιρείται οποτεδήποτε συναντάται.

- *<u>Στάδιο 5ο</u> Αντικατάσταση των hashtags – αφαίρεση του συμβόλου #*

  Σε αντίθεση με τις προηγούμενες περιπτώσεις, τα hashtags αποτελούν πολύτιμη πηγή πληροφοριών και είναι αδύνατον να αγνοηθούν, αφού είναι ενδεικτικά του θέματος που αναφέρεται το tweet, ενώ συνήθως μέσω αυτών οι χρήστες εκφράζουν με έντονο, πολλές φορές, τρόπο τα συναισθήματά τους για ένα συγκεκριμένο γεγονός. Συνεισφέρουν λοιπόν στο συνολικό *συναίσθημα* της πρότασης και αποτελούν και μια ένδειξη για το ευρύτερο *σημασιολογικό πλαίσιο* στο οποίο αυτή εντάσσεται. Για παράδειγμα από το σύνολο δεδομένων που συλλέξαμε, χαρακτηριστικά τέτοια παραδείγματα είναι τα hashtags: #save_kolastirio, #sano, #Idomeni,#skai_xeftiles, #syriza_xeftiles, #Ο_Μεσος_Ψηφοφορος_της_ΝΔ, #kolastirio, #Syriza_tsirko, #emetos, #nd_xeftiles, #geloios, #kolotoumpa, #σούργελα. Αντικαθιστούμε λοιπόν το γενικό τύπο #λέξη με τη

συμβολοσειρά «λέξη», δηλαδή στην ουσία απλά αφαιρούμε το σύμβολο # και κρατάμε την υπολειπόμενη λέξη ή φράση.

- *<u>Στάδιο 6ο</u> Αφαίρεση σημείων στίξης*

  Όπως και στην περίπτωση των κεφαλαίων γραμμάτων τα σημεία στίξης, κυρίως όταν χρησιμοποιούνται επαναλαμβανόμενα (πχ !!!!!, ;;;;;;;;;; κλπ), συμβάλλουν στην ενίσχυση του συναισθήματος και δεν επιδράνε, συνήθως, στην πολικότητα της λέξης, επομένως τα αγνοούμε. Σε ορισμένες περιπτώσεις βεβαίως, επιδρούν και στο συναίσθημα όλης της πρότασης (πχ «Μα είναι σοβαρός αυτός ο άνθρωπος;;;;;;;;;»), αλλά ο χειρισμός τέτοιων περιπτώσεων απαιτεί ενδελεχή γλωσσολογική ανάλυση.

- *<u>Στάδιο 7ο</u> Αφαίρεση stop words*

  Είναι οι πιο συνηθισμένες-κοινές λέξεις μιας γλώσσας οι οποίες δεν έχουν κανένα συναισθηματικό υπόβαθρο και για αυτό το λόγο αφαιρούνται. Τέτοιες λέξεις είναι συνήθως τα άρθρα , οι προθέσεις, οι σύνδεσμοι και οι αντωνυμίες. Μπορεί ακόμα να είναι ρήματα (πχ βλέπω, ακούω, περπατάω, μιλάω κλπ) και ουσιαστικά (μολύβι, αυτοκίνητο, υπολογιστής κλπ). Γενικότερα μιλώντας, δεν υπάρχει μια ενιαία λίστα stop words για κάθε γλώσσα, και το ποιές λέξεις θεωρούνται ως stop words μπορεί να εξαρτάται από την εκάστοτε εφαρμογή, ενώ από την άλλη κάποια εργαλεία επεξεργασίας φυσικής γλώσσας αποφεύγουν ρητά την αφαίρεση τέτοιων λέξεων, καθώς υποστηρίζουν ταίριασμα φράσεων. Στην πραγματικότητα, οποιαδήποτε ομάδα λέξεων μπορεί να επιλεχθεί ως stop words για δεδομένη εφαρμογή.

  Σε ό,τι αφορά την ελληνική γλώσσα, αν ψάξει κανείς στο διαδίκτυο θα βρεί αρκετές τέτοιες λίστες. Εμείς χρησιμοποιήσαμε την λίστα[7], την οποία τροποποιήσαμε ελαφρώς προσθέτοντας ορισμένες λέξεις και αφαιρώντας τις λέξεις άρνησης όπως «δεν» ή «μην», αφού θα τις αξιοποιήσουμε παρακάτω στο στάδιο της ανίχνευσης άρνησης. Τελικά δημιουργήσαμε μια λίστα από 638 τέτοιες λέξεις και οι οποίες φαίνονται στο Παράρτημα.

- *<u>Στάδιο 8ο</u> Αφαίρεση λατινικών χαρακτήρων και αριθμών*

  Το φαινόμενο της εμφάνισης ξενόγλωσσων λέξεων, χαρακτήρων ή συντομογραφιών αλλά και της χρήσης greeklish, είναι εξαιρετικά συνηθισμένο στα ελληνικά tweets. Για παράδειγμα, η πλειοψηφία των hashtags, είναι είτε στα αγγλικά είτε στα greeklish, ενώ φράσεις και λέξεις της καθομιλουμένης αγγλικής γλώσσας (όπως lol, omg) χρησιμοποιούνται

[7] http://www.translatum.gr/forum/index.php?topic=3550.0

κατά κόρον και από Έλληνες. Στα πλαίσια της παρούσας διπλωματικής, ασχολούμαστε αυστηρά και μόνο με όρους της ελληνικής γλώσσας, συνεπώς οποισδήποτε μη ελληνικός όρος δεν λαμβάνεται υπόψιν. Τέλος οι αριθμοί αφαιρούνται για προφανείς λόγους.

- *<u>Στάδιο 9ο</u> Αφαίρεση επαναλαμβανόμενων χαρακτήρων*

  Λόγω του ανεπίσημου λόγου που κυριαρχεί στο Twitter, οι χρήστες πολλές φορές επιμηκύνουν τις λέξεις (κυρίως με επανάληψη των φωνηέντων) προκειμένου να εκφράσουν δυνατά συναισθήματα. Για παράδειγμα μπορεί να γράψουν «καλημέρααααα» αντί «καλημέρα», «ντροπήηηηη» αντί για «ντροπή» ή «ρέεεεε» αντί για «ρε». Και πάλι δεν υπάρχει αλλαγή πολικότητας, επομένως κάθε φορά που εντοπίζουμε το *ίδιο* φωνήεν σε τρείς ή περισσότερες *διαδοχικές* θέσεις της ίδιας λέξεις αφαιρούμε τις περιττές επαναλήψεις .

Ακολουθεί ένα παράδειγμα στο οποίο φαίνονται αναλυτικά όλα τα στάδια της προεπεξεργασίας:

Βήμα	*Tweet*
0. Αρχικό tweet	Εντάξει τάισε μας σανό εμάς @atsipras.Ρίξε μια ματιά όμως σε #kolastirio γιατί πεθαίνουν άνθρωποι. https://t.co/z89duHvMyN
1. Μετατροπή κεφαλαίων σε πεζά	εντάξει τάισε μας σανό εμάς @atsipras.ρίξε μια ματιά όμως σε #kolastirio γιατί πεθαίνουν άνθρωποι. https://t.co/z89duhvmyn
2. Αφαίρεση υπερσυνδέσμων	εντάξει τάισε μας σανό εμάς @atsipras.ρίξε μια ματιά όμως σε #kolastirio γιατί πεθαίνουν άνθρωποι.
3. Αφαίρεση αναφορών σε χρήστες	εντάξει τάισε μας σανό εμάς .ρίξε μια ματιά όμως σε #kolastirio γιατί πεθαίνουν άνθρωποι.
4. Αφαίρεση του “RT”	εντάξει τάισε μας σανό εμάς .ρίξε μια ματιά όμως σε #kolastirio γιατί πεθαίνουν άνθρωποι.
5. Μετατροπή hashtags	εντάξει τάισε μας σανό εμάς .ρίξε μια ματιά όμως σε kolastirio γιατί πεθαίνουν άνθρωποι.
6. Αφαίρεση σημείων στίξης	εντάξει τάισε μας σανό εμάς ρίξε μια ματιά όμως σε kolastirio γιατί πεθαίνουν άνθρωποι
7. Αφαίρεση stop words	εντάξει τάισε σανό ρίξε ματιά kolastirio γιατί πεθαίνουν άνθρωποι
8. Αφαίρεση λατινικών χαρακτήρων	εντάξει τάισε σανό ρίξε ματιά γιατί πεθαίνουν άνθρωποι
9. Αφαίρεση επαναλαμβανόμενων χαρακτήρων	εντάξει τάισε σανό ρίξε ματιά γιατί πεθαίνουν άνθρωποι

Πίνακας 7: Παράδειγμα προεπεξεργασίας tweet – Παρουσίαση των επιμέρους σταδίων

Ακολουθούν δύο ακόμα παραδείγματα, χωρίς την εμφάνιση των ενδιάμεσων σταδίων, παρά μόνο του αρχικού ακατέργαστου και του τελικού φιλτραρισμένου tweet:

1) Το tweet

!! RT @a_m_papagiotis: Εξοργισμένος απ την απώλεια πολλών χρημάτων στην επένδυση @tzitzikostas φέρεται να είναι ο Ιβάν Σαββίδης. #eklogesnd

αφού περάσει όλα τα στάδια της προεπεξεργασίας, καταλήγει στην μορφή:

εξοργισμένος απώλεια πολλών χρημάτων επένδυση φέρεται είναι ιβάν σαββίδης

2) Το tweet

ΛΟΛ άκου τελετή! Και ο αγιασμός τί ώρα; #eklogesnd #σούργελα #metonKyrako https://t.co/lneKaXQJIU

αφού περάσει όλα τα στάδια της προεπεξεργασίας, καταλήγει στην μορφή:

λολ άκου τελετή αγιασμός τί ώρα σούργελα

3.4 Μετασχηματισμός δεδομένων

Αφού φιλτράραμε τα tweets και αφαιρέσαμε το θόρυβο προκειμένου να κρατήσουμε μόνον τις λέξεις που συνεισφέρουν στο συναισθηματικό υπόβαθρο του μηνύματος, θα πρέπει να τις μετατρέψουμε σε κατάλληλη μορφή ώστε να μπορέσουμε εν συνεχεία να τις εντοπίσουμε στο λεξικό συναισθήματος. Ο μετασχηματισμός αυτός πρέπει να γίνει γιατί μια λέξη μπορεί να εμφανιστεί με πολλές μορφές σε μία πρόταση ανάλογα με τον αριθμό (ενικός ή πληθυντικός) και την πτώση αν πρόκειται για ουσιαστικό, επίθετο ή μετοχή και ανάλογα με την έγκλιση (οριστική, υποτακτική, προστακτική), τον χρόνο (ενεστώτας, παρατατικός, αόριστος κλπ), την φωνή (ενεργητική, παθητική) και το πρόσωπο αν πρόκειται για ρήμα.

Είναι προφανές οτι στα λεξικά συναισθήματος μπορεί να αποθηκευτεί μόνο ένα «στιγμιότυπο» κάθε λέξης, αφού σε διαφορετική περίπτωση θα αυξανόταν απαγορευτικά το μέγεθος τους. Στα περισσότερα λεξικά, όπως και σε αυτό που θα δημιουργήσουμε, οι λέξεις αποθηκεύονται στο 1ο πρόσωπο ενικό αν πρόκειται για ουσιαστικά επίθετα ή μετοχές και στο 1ο πρόσωπο οριστικής ενεστώτα αν πρόκειται για ρήματα.

Συνεπώς αν στο φιλτραρισμένο tweet εμφανιστούν πχ οι λέξεις «ανθρώπων», «αξιόλογοι», «χαρούμενου» και «πιστεύετε» θα πρέπει να τις μετασχηματίσουμε με κάποιον τρόπο ώστε να μπορούν να εντοπιστούν στο λεξικό.

3.4.1 Συνήθης προσέγγιση

Στα περισσότερα συστήματα Ανάλυσης Συναισθήματος, μια από τις συνηθέστερες πρακτικές στο στάδιο εξαγωγής χαρακτηριστικών είναι αυτή της αποκοπής των καταλήξεων των λέξεων (*stemming*), δηλαδή στην ουσία της διατήρησης μόνο του προθέματος της λέξης. Αν εφαρμόσουμε για παράδειγμα αυτή τη λογική στις παραπάνω λέξεις θα πάρουμε τα προθέματα: «ανθρώπ», «αξιόλογ», «χαρούμεν» και «πιστεύ» αντίστοιχα. Η μέθοδος αυτή είναι ιδιαίτερα αποτελεσματική σε γλώσσες όπως τα αγγλικά, ενώ υπάρχουν διαθέσιμες στο διαδίκτυο αρκετές υλοποιήσεις και για την ελληνική γλώσσα[8,9,10,11].

Θα μπορούσε λοιπόν κάποιος να υποστηρίξει οτι μια πιθανή προσέγγιση θα ήταν πολύ απλά να βρώ τα προθέματα των λέξεων που έχω αποθηκεύσει στο λεξικό και να τα συγκρίνω με τα προθέματα των λέξεων του εκάστοτε tweet. Οι ιδιαιτερότητες όμως της ελληνικής γλώσσας καθιστούν μια τέτοια προσέγγιση πλήρως αναποτελεσματική, κυρίως εξαιτίας του φαινομένου της *αύξησης* στους παρελθοντικούς χρόνους και των διαφορετικών *καταλήξεων* της υποτακτικής και της προστακτικής σε σχέση με την οριστική έγκλιση. Χωρίς να μπούμε σε πολλές γλωσσολογικές λεπτομέρειες θα περιγράψουμε συνοπτικά τα δύο αυτά φαινόμενα:

- **Αύξηση**[12]

Αποτελεί βασικό μορφολογικό χαρακτηριστικό των παρελθοντικών χρόνων και στη νέα ελληνική διακρίνεται σε συλλαβική, φωνηεντική και εσωτερική.

1) Η συλλαβική αύξηση εμφανίζεται στους παρελθοντικούς χρόνους των ρημάτων που αρχίζουν από σύμφωνο και αυξάνει το θέμα του ρήματος σε δισύλλαβο, δηλαδή μεγαλώνει κατά μία συλλαβή τα μονοσύλλαβα (μόνο) θέματα (π.χ. έ-λυν-α, έ-λυσ-α, έ-γραφ-α, έ-γραψ-α. Όταν το ρηματικό θέμα έχει περισσότερες από μία συλλαβές, η αύξηση δε χρησιμοποιείται (π.χ. γέ/λασ-α, μί/λησ-α, ζω/γρά/φισ-α).

---

[8] http://deixto.com/greek-stemmer/

[9] http://people.dsv.su.se/~hercules/greek_stemmer.gr.html

[10] https://www.drupal.org/project/greekstemmer

[11] https://github.com/skroutz/elasticsearch-skroutz-greekstemmer

[12] http://www2.media.uoa.gr/language/grammar/pdf_files/GG12_2218.pdf

2) Η φωνηεντική αύξηση διατηρείται σε περιορισμένο αριθμό ρημάτων της νέας ελληνικής, συνήθως όταν τονίζεται το αρχικό φωνήεν του θέματος (π.χ. ελπίζω-ήλπιζα, ελέγχω-ήλεγχα/ήλεγξα, κατευθύνω-κατηύθυνα/κατηύθυνα, παράγω-παρήγα/παρήγαγα). Υπάρχει μια κατηγορία ρημάτων της νέας ελληνικής που παίρνουν αύξηση «η-». Αυτά είναι: θέλω με παρατατικό ήθελα, ξέρω με παρατατικό ήξερα, πίνω με αόριστο ήπια.

3) Στην κοινή νεοελληνική υπάρχει ένας μεγάλος αριθμός από σύνθετα ρήματα με προθέσεις (αρχαίες και νέες) που παίρνουν εσωτερική αύξηση στον παρατατικό και στον αόριστο (διαμένω [διά+μένω] - διέμενα, διέμεινα). Όταν υπάρχουν περισσότερες από μία προθέσεις, η αύξηση μπαίνει αμέσως μετά την τελευταία πρόθεση (π.χ. προεκβάλλω [προ+εκ+βάλλω] - προεξέβαλλα, προεξέβαλα).

Παρατηρούμε οτι ενώ το πρόθεμα του «λύνω» είναι «λύν», του έλυσα είναι «έλυσ», ενώ του «ελπίζω» είναι «ελπ» του «ήλπιζα» είναι «ήλπ» και ενώ του «διαμένω» είναι «διαμέν» του «διέμενα» είναι «διέμεν»

- **Καταλήξεις στην Υποτακτική και την Προστακτική**

Στη νέα ελληνική η χρήση ενός ρήματος στην υποτακτική ή την προστακτική αλλάζει σε σημαντικό βαθμό τη μορφή που παρουσιάζει το ρήμα σε σχέση με την οριστική έγκλιση. Για να γίνει κατανοητό αυτό, ας δούμε τα παρακάτω παραδείγματα:

Παράδειγμα 1ο

Ας θεωρήσουμε τις παρακάτω προτάσεις στην οριστική έγκλιση:

«*Παίζω ποδόσφαιρο* στο σχολείο»

«*Γυρνάω* στο σπίτι»

Εάν χρησιμοποιούσαμε τα ρήματα παίζω και γυρνάω στην υποτακτική, δύο πιθανές προτάσεις θα ήταν:

«Θέλω να παίξω μπάλα στο σχολείο»

«Θέλω να γυρίσω στο σπίτι»

Το πρόθεμα (stem) του «παίζω» είναι «παίζ», ενώ το πρόθεμα του «παίξω» είναι «παίξ». Αντίστοιχα το πρόθεμα του «γυρνάω» είναι «γυρνά» και του «γυρίσω» είναι «γυρίσ».

Παράδειγμα 2ο

Η προστακτική του ρήματος «τρέχω» είναι «τρέξε», «τρέξετε» και τα αντίστοιχα προθέματα είναι «τρέχ» και «τρέξ».

Συμπέρασμα

Από τα παραπάνω καθίσταται σαφές οτι το πρόθεμα της ίδιας λέξης διαφέρει ανάλογα με τον χρόνο και την έγκλιση στην οποία αυτή χρησιμοποιείται, δηλαδή δύο διαφορετικά στιγμιότυπα της *ίδιας* λέξης, ενδέχεται να έχουν *διαφορετικά* προθέματα. Συνεπώς μια τέτοια προσέγγιση δεν θα μπορούσε να εφαρμοστεί στην περίπτωσή μας αφού θα χάναμε λέξεις παρελθοντικών χρόνων και υποτακτικής ή προστακτικής έγκλισης, οι οποίες μπορεί μεν να βρίσκονταν στο λεξικό, αλλά θα διέφεραν τα προθέματά τους.

3.4.2 Η δικιά μας προσέγγιση

3.4.2.1 Ο ιστότοπος LEXIGRA*M*[13]

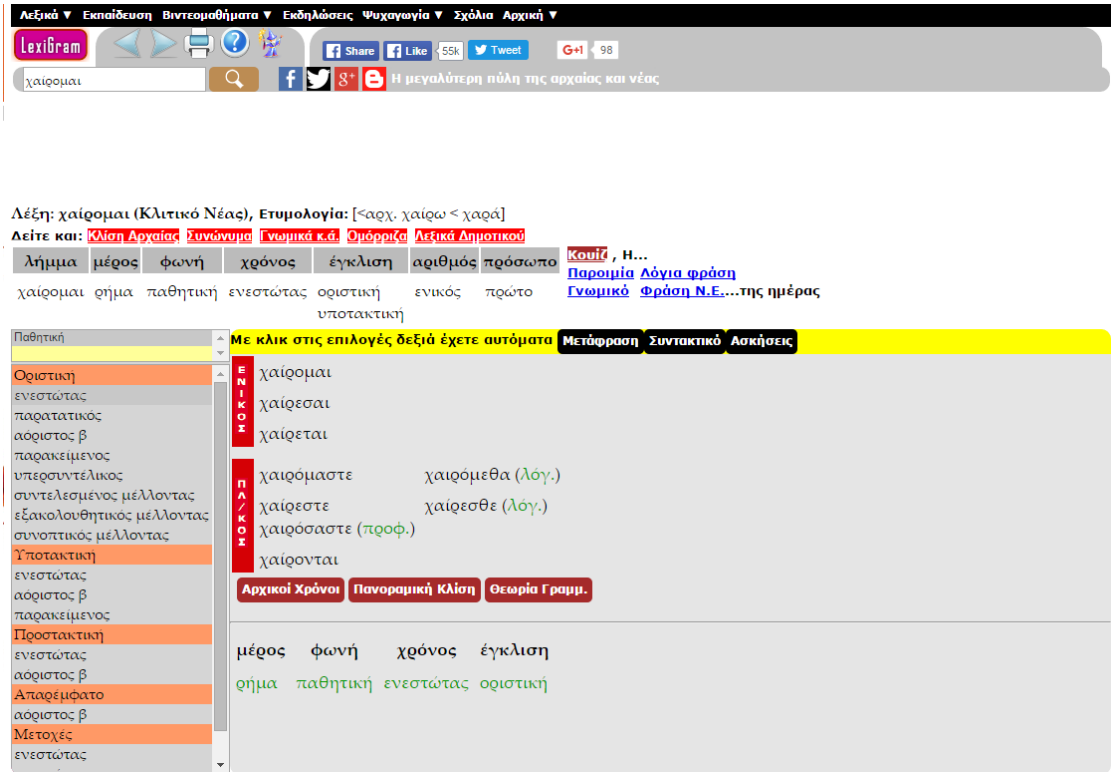
Προκειμένου να αποφύγουμε τα παραπάνω προβλήματα, θεωρήσαμε οτι ο ασφαλέστερος και πιο αποτελεσματικός τρόπος είναι η μετατροπή κάθε λέξης των φιλτραρισμένων tweets σε 1ο πρόσωπο. Το πρόβλημα της ενσωμάτωσης μιας τέτοιας διαδικασίας στο σώμα ενός προγράμματος αποδείχτηκε αρκετά δύσκολο, καθώς αφενός δεν βρήκαμε διαθέσιμη κάποια έτοιμη ρουτίνα, και αφετέρου, αν και υπήρχαν διαθέσιμες διάφορες online πλατφόρμες που επιτελούσαν αυτόν τον σκοπό, σχεδόν καμία τους δε διέθετε ένα API που θα διευκόλυνε την επικοινωνία με το πρόγραμμα - κώδικά μας. Στο σημείο αυτό αξίζει να σημειωθεί οτι η μόνη πλατφόρμα που διέθετε ένα τέτοιο API ήταν η Open Xerox[14], την οποία όμως δεν μπορέσαμε εν τέλει να αξιοποιήσουμε γιατί συναντήσαμε κάποια προβλήματα κωδικοποίησης (encoding) με την ελληνική γλώσσα.

Μετά από αρκετό ψάξιμο στο διαδίκτυο και πολλούς πειραματισμούς καταλήξαμε στο να χρησιμοποιήσουμε την πλατφόρμα LEXI$\mathcal{GRAM}$, η οποία προσφέρει online, δωρεάν, εφτά ηλεκτρονικά λεξικά και συγκεκριμένα κλιτικό λεξικό της αρχαίας ελληνικής, κλιτικό λεξικό της νέας ελληνικής, λεξικό συνωνύμων και αντιθέτων, λεξικό γνωμικών και παροιμιών, λεξικό ομόρριζων και παραγώγων, λεξικά δημοτικού και λεξικά λατινικών.

Εμείς χρησιμοποιήσαμε το κλιτικό λεξικό της νέας ελληνικής, η διεπιφάνεια (interface) του οποίου φαίνεται στην εικόνα:

[13] http://www.lexigram.gr/

[14] https://open.xerox.com/Services/fst-nlp-tools



Εικόνα 12: Διεπιφάνεια ιστοτόπου Lexigram

Το λεξικό αυτό λειτουργεί ως εξής:

1) Όταν εισάγουμε στην γραμμή εντολών ένα ρήμα σε οποιοδήποτε πρόσωπο, χρόνο και έγκλιση θα επιστραφεί το 1ο ενικό πρόσωπο οριστικής ενεστώτα.

Για παράδειγμα αν εισάγουμε τη λέξη «κατηύθυνα» επιστρέφει τη λέξη «κατευθύνω», αν εισάγουμε τη λέξη «περιέβαλλα» επιστρέφει τη λέξη «περιβάλλω», ενώ αν εισάγουμε τη λέξη «μιλήστε» επιστρέφει τη λέξη «μιλάω».

2) Όταν εισάγουμε στην γραμμή εντολών ένα επίθετο οποιουδήποτε γένους, αριθμού και πτώσης, θα επιστραφεί την ονομαστική πτώση ενικού του αρσενικού γένους.

Για παράδειγμα αν εισάγουμε τη λέξη «ντροπαλής» επιστρέφει τη λέξη «ντροπαλός», ενώ αν εισάγω τη λέξη «αδιάλλακτων» επιστρέφει τη λέξη «αδιάλλακτος».

3) Όταν εισάγουμε στην γραμμή εντολών μια μετοχή οποιουδήποτε γένους, αριθμού και πτώσης, θα επιστραφεί το 1ο ενικό πρόσωπο οριστικής ενεστώτα του ρήματος από το οποίο παράγεται η μετοχή.

Για παράδειγμα αν εισάγουμε τη λέξη «ευτυχισμένης» επιστρέφει τη λέξη «ευτυχώ».

4) Τέλος αν εισάγουμε στη γραμμή εντολών ένα ουσιαστικό σε οποιοαδήποτε πτώση και αριθμό, θα επιστραφεί η ονομαστική ενικού.

   Για παράδειγμα αν εισάγουμε τη λέξη «αδικίας» επιστρέφει τη λέξη «αδικία».

Φαίνεται λοιπόν από τα παραπάνω οτι η εν λόγω εφαρμογή αποτελεί ιδανική λύση στο πρόβλημά μας, αφού αποφεύγει όλα τα προβλήματα που θα δημιουργούνταν αν χρησιμοποιούσαμε αποκοπή κατάληξης.

3.4.2.2 Ενσωμάτωση πλατφόρμας στον κώδικα

Προκειμένου να ενσωματώσουμε τις λειτουργίες της online πλατφόρμας στον κώδικά μας, κάναμε χρήση της ρουτίνας urlread[15] του Matlab. Η ρουτίνα αυτή δέχεται σαν όρισμα μια διεύθυνση url και μεταφορτώνει (download) το διαδικτυακό HTML περιεχόμενο (HTML web content) σε μια συμβολοσειρά. Καλώντας δηλαδή τη συνάρτηση μπορούμε να αποθηκεύσουμε τον HTML κώδικα της σελίδας στην οποία «δείχνει» η url διεύθυνση, σε ένα string. Ο HTML κώδικας περιέχει πληροφορίες για τη δομή, τη μορφή και το περιεχόμενο της εκάστοτε ιστοσελίδας. Συνεπώς αν καλέσουμε την ρουτίνα urlread με όρισμα μια url διεύθυνση που αντιστοιχεί στην λέξη την οποία θέλουμε να μετασχηματίσουμε, θα μεταφορτώσουμε τον HTML κώδικα που αντιστοιχεί στη σελίδα που θα βλέπαμε εάν χρησιμοποιούσαμε την υπηρεσία μέσα από την online διεπιφάνειά της. Εν συνεχεία, αν φιλτράρουμε τον κώδικα και αφαιρέσουμε όλη την άχρηστη πληροφορία, μπορούμε να κρατήσουμε μόνο το κομμάτι του HTML κώδικα που μας ενδιαφέρει: το 1ο πρόσωπο της λέξης που δώσαμε σαν όρισμα.

Ας υποθέσουμε για παράδειγμα οτι θέλαμε να βρούμε το 1ο πρόσωπο της λέξης «ανθρώπων». Αν ενσωματώσουμε τη λέξη «ανθρώπων» στη url διεύθυνση "http://www.lexigram.gr/lex/newg/ ", δηλαδή καλέσουμε τη συνάρτηση urlread με όρισμα "http://www.lexigram.gr/lex/newg/ανθρώπων" θα δούμε να εμφανίζεται στην οθόνη μας ο παρακάτω HTML κώδικας:

[15] http://www.mathworks.com/help/matlab/ref/urlread.html

```
>> urlread('http://www.lexigram.gr/lex/newg/ανθρώπων')

ans =

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">

<html xmlns="http://www.w3.org/1999/xhtml">


<head>

    <!--- Copyright Andreas Tassos --->

    <meta http-equiv="Content-Language" content="el"/>

    <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>

    <meta http-equiv="Cache-Control" content="no-cache, private, no-store, must-revalidate, pre-c

    <meta http-equiv="Pragma" content="no-cache">

    <meta http-equiv="Expires" content="Mon, 26 Jul 1999 05:00:00 GMT">

    <meta http-equiv="Vary" content="*">


    <meta name="keywords" content="ανθρώπων, ανθρωπων, κλίση, νέα ελληνική, ορθογραφία, αρχικοί χ

    <meta name="description" content="ανθρώπων κλίση. ανθρωπων κλιση. ανθρώπων ορθογραφία. ανθρωπ


    <title>ανθρώπων - Νέα : Κλίση, Ορθογραφία, Αναγνώριση, Γραμματική (Νέα Και Λόγια Ελληνική) -

    <link rel="shortcut icon" type="image/x-icon" href="/favicon.ico"/>
```

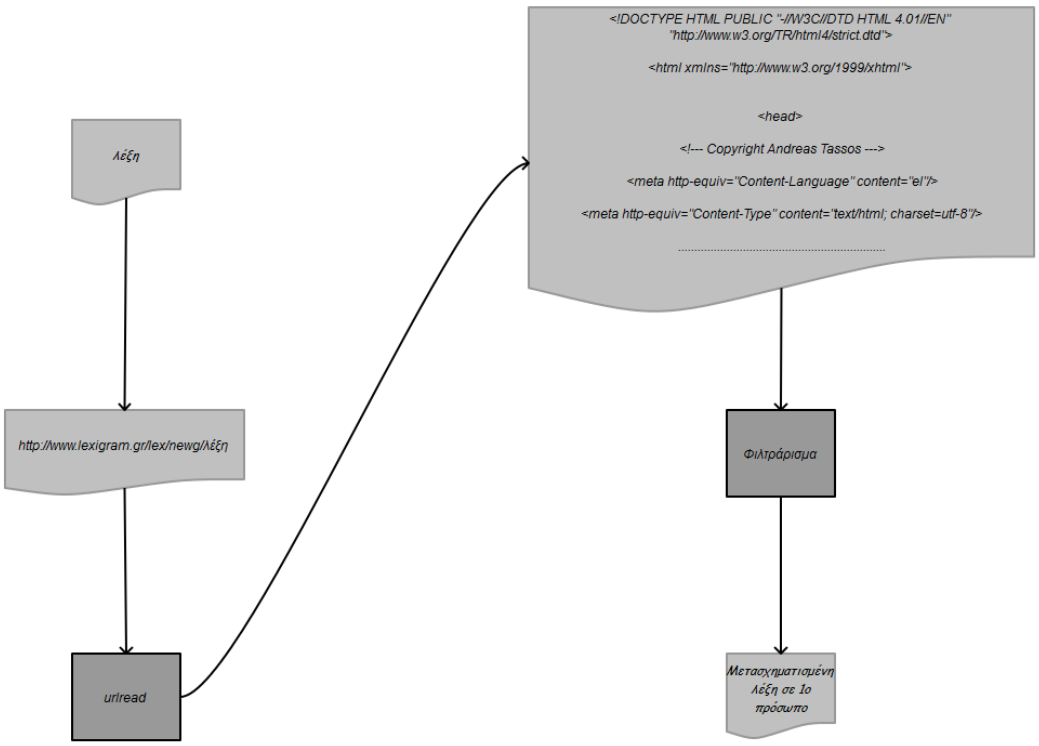
Εικόνα 13: Χρήση της συνάρτησης urlread για επικοινωνία με τον ιστότοπο Lexigram – Μέρος α

Κάπου μέσα στον κώδικα υπάρχει και η μετασχηματισμένη λέξη:

```
πτώση</th>

</tr>

<tr>

<td valign="top">

άνθρωπος</td>

<td valign="top">

ουσιαστικό</td>

<td valign="top">
```

Εικόνα 14: Χρήση της συνάρτησης urlread για επικοινωνία με τον ιστότοπο Lexigram – Μέρος β

Σχηματικά λοιπόν η διαδικασία που ακολουθούμε είναι η παρακάτω:



Εικόνα 15: Διάγραμμα ροής διαδικασίας μετασχηματισμού λέξης σε 1ο πρόσωπο μέσω του ιστοτόπου Lexigram

Ας δούμε τώρα ένα παράδειγμα στο οποίο φαίνεται η επίδραση των σταδίων της προεπεξεργασίας και του μετασχηματισμού των δεδομένων. Θα θεωρήσουμε δύο από τα tweets που είδαμε και στην προηγούμενη παράγραφο, μόνο που τώρα δεν θα δείξουμε όλα τα ενδιάμεσα στάδια της προεπεξεργασίας, παρά μόνο το αρχικό ακατέργαστο tweet, το φιλτραρισμένο και το μετασχηματισμένο σε 1ο πρόσωπο tweet.

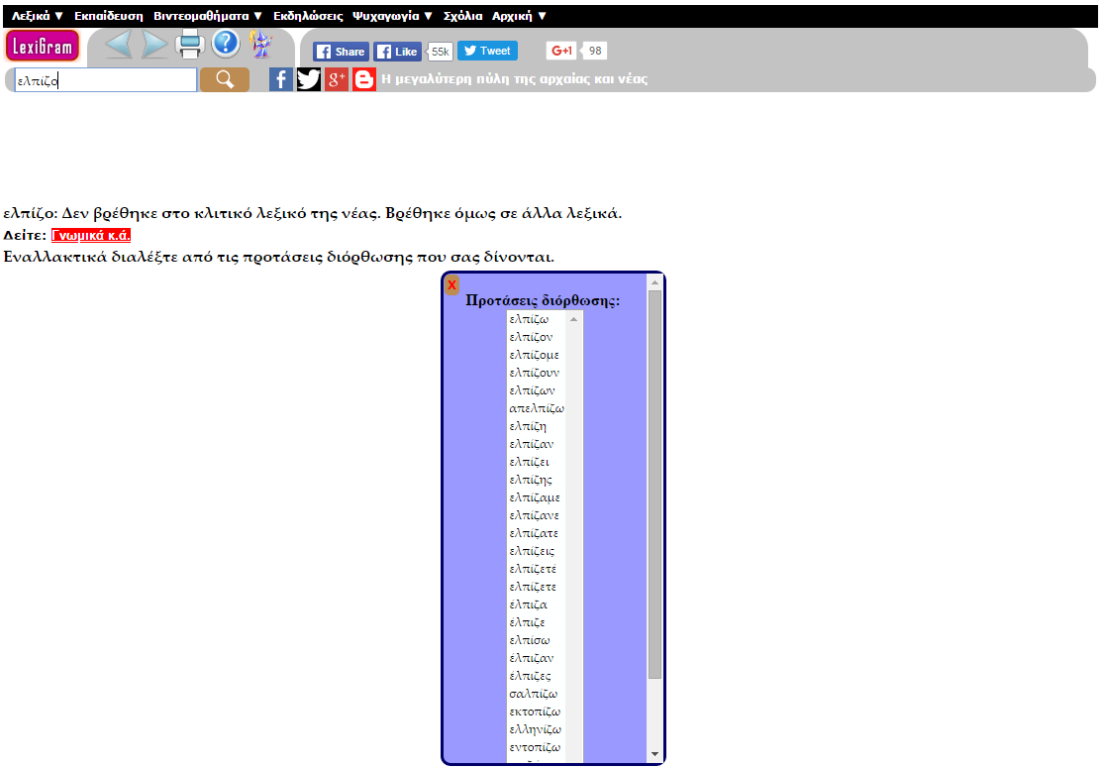
Αρχικό tweet	***Φιλτραρισμένο tweet***	***Μετασχηματισμένο tweet***
Εντάξει τάισε μας σανό εμάς @atsipras.Ρίξε μια ματιά όμως σε #kolastirio γιατί πεθαίνουν άνθρωποι. https://t.co/z89duHvMyN	εντάξει τάισε σανό ρίξε ματιά γιατί πεθαίνουν άνθρωποι	εντάξει ταΐζω σανό ρίχνω ματιά γιατί πεθαίνω άνθρωπος
!! RT @a_m_papagiotis: Εξοργισμένος απ την απώλεια πολλών χρημάτων στην επένδυση @tzitzikostas φέρεται να είναι ο Ιβάν Σαββίδης. #eklogesnd	εξοργισμένος απώλεια πολλών χρημάτων επένδυση φέρεται είναι ιβάν σαββίδης	εξοργίζω απώλεια πολύς χρήμα επένδυση φέρω είμαι ιβάν σαββίδης

Πίνακας 8: Παραδείγματα προεπεξεργασίας και μετασχηματισμού tweets

3.4.2.3 Πιθανά bugs

Σε ορισμένες περιπτώσεις όταν «επικοινωνούμε» με την πλατφόρμα μέσω του κώδικά μας δεν είναι δυνατόν, εκ των πραγμάτων, να πάρουμε το επιθυμητό αποτέλεσμα, δηλαδή τη μετασχηματισμένη σε 1ο πρόσωπο λέξη. Τα προγραμματιστικά αυτά σφάλματα, τα λεγόμενα bugs, συμβαίνουν για δύο λόγους:

- *Ανορθογραφία*: Όταν εισάγεται μια λέξη με λανθασμένη ορθογραφία, εμφανίζεται μια λίστα από πιθανές λέξεις, από τις οποίες μπορεί να επιλεχθεί *χειροκίνητα* και μόνον ποια είναι αυτή με τη σωστή ορθογραφία.



Εικόνα 16: Παράδειγμα ανορθογραφίας στο Lexigram

- *Πολυσημία:* Σε περιπτώσεις λέξεων που έχουν διπλή σημασία, όπως για παράδειγμα η λέξη «μπράβο» η οποία μπορεί να αναφέρεται είτε στο επιφώνημα, είτε στο ουσιαστικό «ο μπράβος», ή το ρήμα «καταλάβω» το οποίο μπορεί να προέρχεται είτε από το «καταλαβαίνω», είτε από το «καταλαμβάνω», δίνεται και πάλι η δυνατότητα χειροκίνητης επιλογής για τη λέξη που εννοούσε ο χρήστης.

Εικόνα 17: Παράδειγμα πολύσημης λέξης στο Lexigram

Όταν καλέσουμε τη συνάρτηση urlread με όρισμα μια λέξη που εντάσσεται στην 1η κατηγορία, δηλαδή είναι ανορθόγραφη (ή χωρίς τονισμό), τότε στην έξοδο δεν λαμβάνουμε καμία λέξη και συνεπώς ακόμα και αν η λέξη που υπονοείται υπάρχει στο λεξικό, δεν μπορεί να εντοπιστεί. Από την άλλη όταν η λέξη εντάσσεται στη 2η κατηγορία, δηλαδή μπορεί να επιδέχεται πολλαπλών ερμηνειών, τότε επιλέγεται αυτόματα μία από τις πολλές ερμηνείες, χωρίς όμως να υπάρχει έλεγχος από τη μεριά του προγραμματιστή με αποτέλεσμα να ενδέχεται να επιστραφεί είτε η σωστή λέξη, είτε μια λανθασμένη. Για παράδειγμα στη λέξη «μπράβο» επιστρέφεται το επιφώνημα «μπράβο», αν και θα μπορούσε να πρόκειται για την αιτιατική πτώση του ουσιαστικού «ο μπράβος», ενώ στη λέξη «καταλάβω» επιστρέφεται το ρήμα «καταλαμβάνω», αν και θα μπορούσε κάλλιστα η λέξη να προέρχεται από το ρήμα «καταλαβαίνω».

Σε αυτές τις περιπτώσεις λοιπόν δε δύναται να υπάρξει έλεγχος από τη μεριά μας, γι'αυτό και στο στάδιο της αξιολόγησης θα προσπαθήσουμε να θεωρήσουμε όσο το δυνατόν πιο ορθογραφημένα tweets. Σε ό,τι αφορά το θέμα της πολυσημίας δεν μπορούμε να επέμβουμε με κάποιο τρόπο και θα αρκεστούμε αναγκαστικά στη λέξη που επιστρέφεται αυτόματα.

3.5 Δημιουργία λεξικού

3.5.1 Αρχικό λεξικό

Ως βάση για το λεξικό που τελικά δημιουργήσαμε, χρησιμοποιήσαμε ένα λεξικό συναισθήματος που δημιουργήθηκε από τον Αδάμ Τσακαλίδη (Ινστιτούτο Τεχνολογιών Πληροφορικής και Επικοινωνιών) σε συνεργασία με τον Συμεών Παπαδόπουλο (Ινστιτούτο Τεχνολογιών Πληροφορικής και Επικοινωνιών) με αφορμή την [32] και τη συνεισφορά των Ουρανία Βοσκάκη (Κέντρο Ελληνικής Γλώσσας), Κυριακή Ιωαννίδου (Κέντρο Ελληνικής Γλώσσας) και Χριστίνα Βοϊδίδου (Κέντρο Ελληνικής Γλώσσας) . Το λεξικό αυτό δημιουργήθηκε με την υποστήριξη του προγράμματος SocialSensor[16] και βρίσκεται διαθέσιμο στο διαδίκτυο για κοινή χρήση στο κοινωνικό εναποθετήριο εργασιών ανοιχτού κώδικα GitHub.[17]

Το λεξικό αποτελείται από 2324 λέξεις οι οποίες έχουν επισημειωθεί από 4 διαφορετικούς βαθμολογητές σε 9 διαστάσεις. Οι πρώτες τρείς διαστάσεις περιλαμβάνουν το μέρος του λόγου της λέξης (Part-of-Speech Tag), την υποκειμενικότητα (ισχυρή/ασθενής/τίποτα) και την πολικότητα (θετική/αρνητική/ουδέτερη), και συνοψίζουν τις θεμελιώδεις προκλήσεις του προβλήματος της Ανάλυσης Συναισθήματος. Οι υπολοιπόμενες 6 διαστάσεις αναφέρονται στα συναισθήματα του θυμού (anger), της αηδίας (disgust), του φόβου (fear), της ευτυχίας (hapiness), της λύπης (sadness) και της έκπληξης (surprise) και για την επισημείωσή τους χρησιμοποιείται μια κλίμακα από το ένα έως το πέντε, αντιστοιχίζοντας κάθε λέξη με την πιθανότητα να εκφράζει το επικρατές συναίσθημα σε μια τυχαία πρόταση [31].

Στα πλαίσια της εργασίας μας, τροποποιήσαμε το λεξικό αυτό ως εξής: Αρχικά αφαιρέσαμε την πλειοψηφία των λέξεων που ήταν τελείως ουδέτερες (πχ ψυχοσύνθεση, αέρας, έδαφος κλπ) και συνεπώς δεν θα είχαν καμία επίδραση στο συναισθηματικό περιεχόμενο της πρότασης. Από τις 2324 αρχικές λέξεις καταλήξαμε σε 2244. Σε αυτές τις εναπομείνασες λέξεις επισημειώσαμε εξαρχής το συναίσθημά τους, χρησιμοποιώντας μια κλίμακα από το -4 έως το +4, όπου η βαθμολογία -4 αντιστοιχεί στη μέγιστη αρνητική βαθμολογία και η βαθμολογία + 4 αντιστοιχεί στη μέγιστη θετική βαθμολογία. Η επιλογή αυτή, αντί της τυπικής επισημείωσης πολικότητας με -1,0 ή +1 έγινε προκειμένου να υπάρξει μια διαβάθμιση μεταξύ του συναισθήματος των λέξεων, καθώς θεωρήσαμε οτι δεν έχουν όλες οι λέξεις, ακόμα και αν είναι της ίδιας πολικότητας, την ίδια επίδραση στο συναισθηματικό περιεχόμενο μιας πρότασης, αλλά *συνήθως* όταν υπάρχει έστω και μία λέξη με εξαιρετικά έντονο συναισθηματικό περιεχόμενο (πχ μασόνος, μισάνθρωπος, σκλάβος, στοργικός, αξιολάτρευτος), αυτή είναι που υπαγορεύει και το συνολικό συναίσθημα της πρότασης. Με αυτόν τον τρόπο λοιπόν δίνουμε μεγαλυτερο βάρος στις λέξεις με πολύ έντονο συναίσθημα και μικρότερο βάρος σε αυτές με λιγότερο έντονο συναίσθημα. Έτσι στον υπολογισμό του συνολικού σκόρ της πρότασης κάθε λέξη *σταθμίζεται* ανάλογα με το πόσο έντονο είναι το συναισθηματικό της περιεχόμενο και όχι απλά με βάση την πολικότητά της.

Να σημειωθεί οτι δεν ασχοληθήκαμε με τις υπόλοιπες διαστάσεις του αρχικού λεξικού (μέρος του λόγου, υποκειμενικότητα, θυμός, απέχθεια κλπ) καθώς απέκλειναν από τον σκοπό της εργασίας.

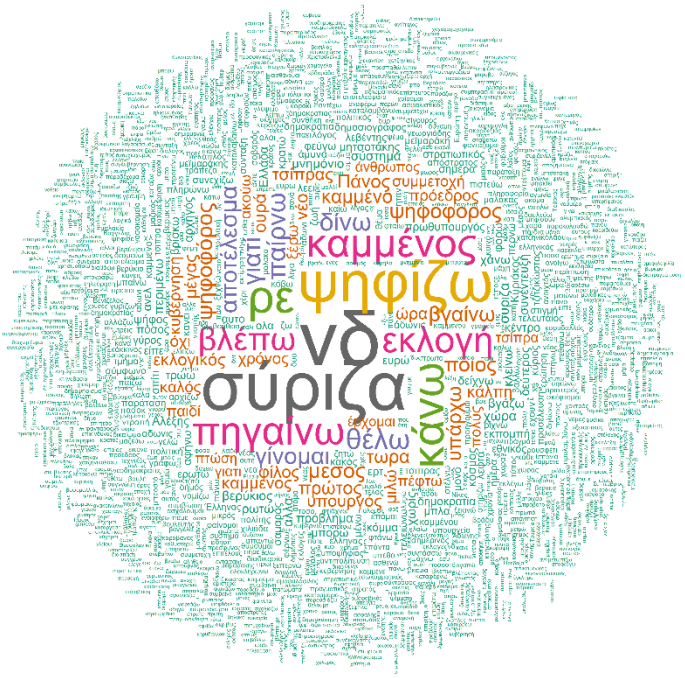
[16] http://www.socialsensor.eu/results/datasets/147-greek-sentiment-lexicon

[17] https://github.com/MKLab-ITI/greek-sentiment-lexicon

3.5.2 Επέκταση λεξικού

Χρησιμοποιώντας μέρος των tweets που συλλέξαμε κατά το 1ο στάδιο του αλγορίθμου (εξαιρουμένων αυτών που θα χρησιμοποιούνταν για την τελική αξιολόγηση), και αφού τα προεπεξεργαστήκαμε ώστε να αφαιρεθεί όλος ο θόρυβος και τα μετασχηματίσαμε στο 1ο πρόσωπο, κάναμε μια στοιχειώδη στατιστική ανάλυση και εντοπίσαμε τις πιο συχνά εμφανιζόμενες από αυτές. Συνολικά εντοπίσαμε κοντά στις 12000 μοναδικές λέξεις (δηλαδή κάθε μία από αυτές έχει εμφανιστεί από 1 φορά και πάνω) και τελικά προσθέσαμε στο λεξικό 828 από αυτές.

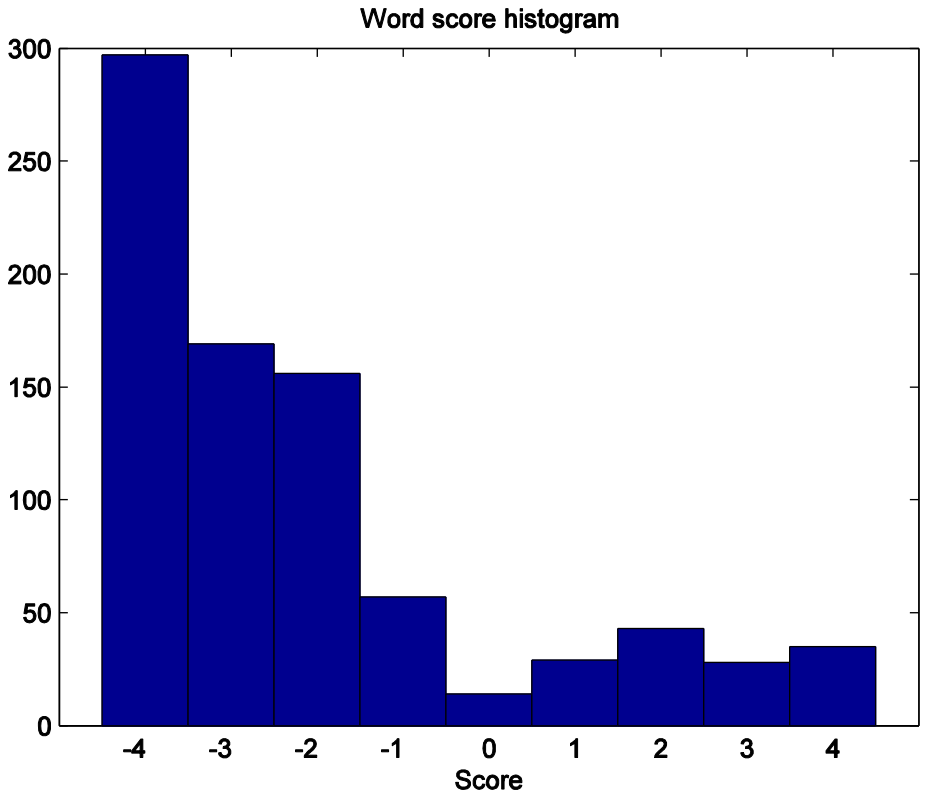
Όπως και στην παράγραφο 3.2.2, έτσι και τώρα, χρησιμοποιούμε τα λεγόμενα σύννεφα λέξεων (word cloud) για να οπτικοποιήσουμε τα αποτελέσματα και να δώσουμε στον αναγνώστη μια αίσθηση των συχνότερα εμφανιζόμενων λέξεων και συνεπώς και της ευρύτερης φύσης του λεξιλογίου που επικρατούσε στα tweets που συλλέξαμε.



Εικόνα 18: Word cloud των συχνότερα εμφανιζόμενων λέξεων στο σύνολο δεδομένων που συλλέξαμε (generated in R)

Τελικά τα λεξικό που δημιουργήσαμε αποτελείται από 3072 όρους. Οι λέξεις που προσθέσαμε είναι διαθέσιμες στο Παράρτημα στον Πίνακα 24.

Ένα ενδιαφέρον στοιχείο έχει να κάνει με την κατανομή της βαθμολογίας των λέξεων που προσθέσαμε στο λεξικό, δηλαδή από τις συνολικά 828 πόσες ήταν αρνητικές, πόσες θετικές και πόσες ουδέτερες. Αυτό γίνεται φανερό στο παρακάτω ιστόγραμμα:

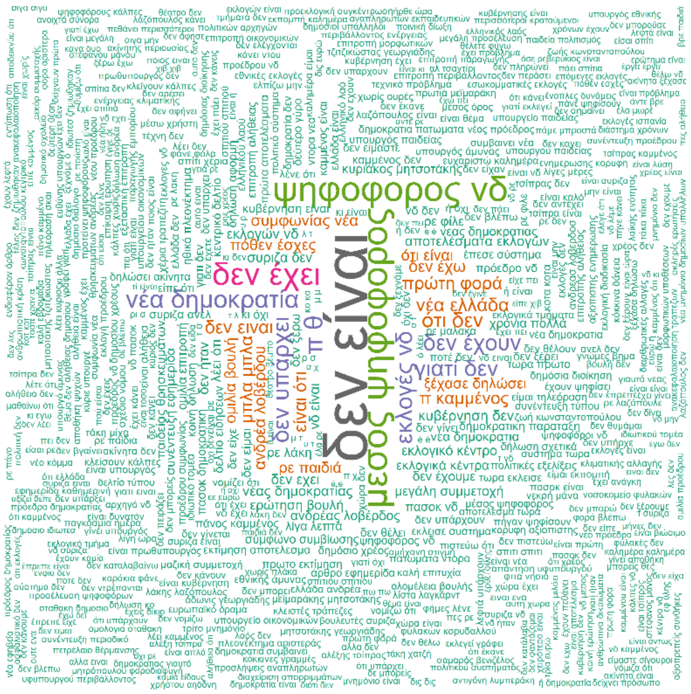


Εικόνα 19: Ιστόγραμμα βαθμολογιών των λέξεων που προστέθηκαν στο λεξικό

Από το ιστόγραμμα είναι φανερό οτι η συντριπτική πλειοψηφία των λέξεων που προσθέσαμε στο λεξικό είχε αρνητικό συναισθηματικό υπόβαθρο. Συγκεκριμένα το 82% των λέξεων που προστέθηκαν στο λεξικο είχαν αρνητικό συναισθηματικό υπόβαθρο, και μόλις το 13% είχαν θετικό συναισθηματικό υπόβαθρο. Αυτό βέβαια είναι απολύτως λογικό καθώς οι χρήστες του Twitter πολύ συχνότερα κριτικάρουν και αποδοκιμάζουν παρά επαίρουν και επιδοκιμάζουν πολιτικά πρόσωπα ή πολιτικές πρακτικές και αποφάσεις.

Τέλος να σημειώσουμε οτι, αν και τελικά δεν αξιοποιήθηκαν στα πλαίσια της εργασίας, εντοπίσαμε και τα πιο συχνά εμφανιζόμενα διγράμματα (bigrams). Συνολικά εντοπίσαμε κοντά στα 10.000 διγράμματα. Ο λόγος που δεν χρησιμοποιήθηκαν είναι οτι ήταν εξαιρετικά αραιά, δηλαδή καθένα παρουσίαζε αρκετά μικρή ποσοστιαία συχνότητα εμφάνισης, κάτι που σημαίνει οτι για να εντοπίσουμε τα πραγματικά χρήσιμα, δηλαδή αυτά με συναισθηματικό υπόβαθρο, θα έπρεπε να ελέγξουμε χειροκίνητα σχεδόν το σύνολο τους, κάτι που θα ήταν αδιανόητα χρονοβόρο. Παρόλα αυτά θα μπορούσαν να αξιοποιηθούν σε μελλοντικές εργασίες.

Ακολουθεί το αντίστοιχο word cloud:



Εικόνα 20: Word cloud των συχνότερα εμφανιζόμενων διγραμμάτων (bigrams) στο σύνολο δεδομένων που συλλέξαμε (generated in R)

3.6 Εντοπισμός στο λεξικό

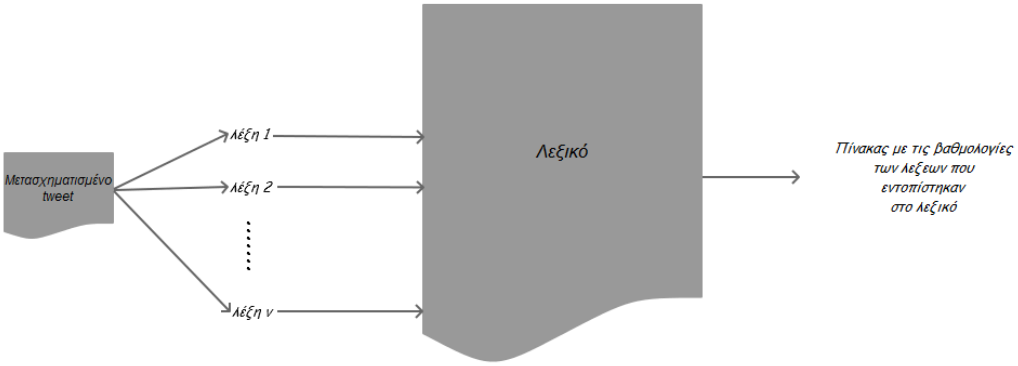
Έχοντας υλοποιήσει τα στάδια της προεπεξεργασίας αλλά και του μετασχηματισμού των δεδομένων και έχοντας δημιουργήσει και το λεξικό συναισθήματος που θα χρησιμοποιήσουμε, είμαστε έτοιμοι πλέον να αναφερθούμε στη διαδικασία εντοπισμού λέξεων στο λεξικό.

Χωρίζουμε λοιπόν κάθε (φιλτραρισμένο και μετασχηματισμένο) tweet στις επιμέρους λέξεις του και για κάθε μία από αυτές διατρέχουμε το λεξικό προκειμένου να δούμε αν αυτή βρίσκεται καταχωρημένη στο λεξικό ή όχι. Σε περίπτωση που μια λέξη εντοπιστεί στο λεξικό τότε δημιουργούμε δύο πίνακες: Στον έναν αποθηκεύουμε τη λέξη και στον άλλον τη βαθμολογία της λέξης. Κάθε φορά που εντοπίζουμε μια νέα λέξη, αυτή τοποθετείται στην αμέσως επόμενη θέση του 1ου πίνακα, και το ίδιο συμβαίνει με τη βαθμολογία της στο 2ο πίνακα. Εάν δεν εντοπίσουμε καμία λέξη, τότε ο 1ος πίνακας παραμένει άδειος και στον 2ο πίνακα τοποθετούμε τη βαθμολογία 0, αφού θεωρούμε, στην ιδανική περίπτωση, οτι ο αλγόριθμός μας δεν εντόπισε κάποια λέξη ακριβώς επειδή δεν υπήρχε καμία λέξη στην πρόταση με συναισθηματικό περιεχόμενο, δηλαδή ήταν όλες ουδέτερες.

Προφανώς κάτι τέτοιο δεν ισχύει στην πράξη, αφού στις περισσότερες των περιπτώσεων θα υπάρχουν τέτοιες λέξεις αλλά δεν εντοπίζονται είτε γιατί πολύ απλά δεν βρίσκονται στο λεξικό, είτε γιατί ναι μεν βρίσκονται στο λεξικό, αλλά

έχουν γραφτεί με λανθασμένη ορθογραφία και συνεπώς δεν μπορούμε να τις μετασχηματίσουμε σωστά.

Σχηματικά η παραπάνω διαδικασία έχει ως εξής:



Εικόνα 21: Διάγραμμα ροής διαδικασίας εντοπισμού των λέξεων των μετασχηματισμένων tweets στο λεξικό

3.7 Ανίχνευση άρνησης

Προτού προχωρήσουμε στον υπολογισμό του τελικού σκορ κάθε tweet είναι απαραίτητο να διερευνήσουμε την ύπαρξη άρνησης. Όπως αναφέρθηκε και στο Κεφάλαιο 2, η άρνηση είναι ένα από τα πιο συνηθισμένα και ταυτόχρονα σύνθετα γλωσσικά φαινόμενα, καθώς εκφράζεται μέσω μιας μεγάλης ποικιλίας συντακτικών και γραμματικών μοτίβων και συνεπώς ο χειρισμός όλων αυτών των μοτίβων θα απαιτούσε λεπτομερή γλωσσολογική ανάλυση.

Δεδομένου οτι η παρούσα διπλωματική δεν εστιάζει στο κομμάτι της επεξεργασίας φυσικής γλώσσας (NLP), επιλέξαμε να χρησιμοποιήσουμε κάποιους κανόνες οι οποίοι θα εντοπίζουν τις πιο συνηθισμένες μορφές άρνησης. Μετά από μια επισκόπηση των tweets που συλλέξαμε κατά το 1ο στάδιο του αλγορίθμου μας, καταλήξαμε στους τρείς παρακάτω θεμελιώδεις κανόνες:

- *Κανόνας 1ος : «Δεν» + ρήμα*

Είναι ίσως η πιο συνηθισμένη μορφή άρνησης που χρησιμοποιείται στην ελληνική γλώσσα, τόσο στον προφορικό όσο και στον γραπτό λόγο. Η ύπαρξη του μορίου δεν (ή δε) πριν από το ρήμα, είναι σαφές οτι, αντιστρέφει την πολικότητα του ρήματος. Μερικά χαρακτηριστικά παραδείγματα tweet που περιέχουν αυτήν την μορφή άρνησης είναι τα παρακάτω:

Δεν φοβάμαι την εσωστρέφεια γιατί είναι προϋπόθεση της εξωστρέφειας. Το θέμα είναι αν αφορά αρχές ή προσωπικά χαρακτηριστικά #vimafm

Απαράδεκτο κάθε σχόλιο που δεν σέβεται τους ποταμούς αίματος του ποντιακού Ελληνισμού. ΝΔ

Εκλογές δεν κερδίζονται χωρίς ιδέες και απαντήσεις στα χάλια που προκάλεσε η κυβέρνηση Τσίπρα #parapolitika901

Ο Τσίπρας για τον Σακελλαρίδη: Σέβομαι την απόφαση του αλλά δεν την κατανοώ - http://bit.ly/1NkGA4s

Τσίπρας: Οι Έλληνες δεν τρομοκρατούνται- Έρχεται το «Σαμαράς exit» #oikonomia http://bit.ly/1z4NAJj

- *Κανόνας 2ος : «Δεν» + «είμαι» + κατηγορούμενο*

  Το ρήμα «είμαι» αποτελεί το βασικότερο συνδετικό ρήμα της ελληνικής γλώσσας. Συνδετικά ρήματα είναι τα ρήματα που συνδέουν το υποκείμενο του ρήματος με ένα κατηγορούμενο (πχ «Η Άννα είναι μαθήτρια») και χρησιμοποιούνται συχνά στην *περιγραφή* προσώπων, αντικειμένων, καταστάσεων κλπ, επειδή με αυτόν τον τρόπο αποδίδουμε τις ιδιότητες της οντότητας ή ιδέας που περιγράφουμε (πχ «Ο πατέρας μου είναι ψηλός, αδύνατος .... »). Άλλα συνηθισμένα συνδετικά ρήματα είναι: γίνομαι, φαίνομαι, θεωρούμαι, εμφανίζομαι κ.α[18].

  Είναι προφανές λοιπόν οτι η προσθήκη του μορίου «δεν» πριν από ένα συνδετικό ρήμα και πιο συγκεκριμένα το «είμαι», αντιστρέφει την πολικότητα της ιδιότητας που αποδίδουμε στο υποκείμενο ή στο αντικείμενο της πρότασης (πχ Δεν είναι έξυπνος, Δεν είναι δίκαιο κλπ). Μερικά χαρακτηριστικά tweets στα οποία γίνεται χρήση του συγκεκριμένου μοτίβου είναι τα παρακάτω:

Συγγνώμη που δεν είμαι κόσμιος. Αηδιαστικός εμετικός Φίλης.

Ο επικοινωνιακός οίστρος του πρωθυπουργικού περιβάλλοντος κ η οξύτητα της προσφυγικής κρίσης δεν είναι ο καλύτερος σύμβουλος ενόψει Τουρκίας

Χρέος 120%: Το χρέος σας δεν είναι βιώσιμο, χρειάζεστε μνημόνιο. Χρέος 180%: Το χρέος σας είναι βιώσιμο πρέπει να συνεχίσετε το μνημόνιο.

Δεν είσαστε αγανακτισμένοι, ενώ θα έπρεπε με τη συμπεριφορά της κυβέρνησης και των ΜΜΕ απέναντι στη Δράση: http://www.athensvoice.gr/article/city-news-voices/

Επειδή δεν είναι ικανοί να φτιάξουν το μέλλον μας, ανασκαλεύουν το παρελθόν.

Απόδειξη ότι δεν είναι όλοι τρελοί στο ΣΥΡΙΖΑ! http://fb.me/7slR2T1Cf

Η ΝΔ δεν είναι σήμερα όσο ελκυστικό κόμμα θα μπορούσε να είναι.Δυστυχώς υπάρχουν ακόμα φεουδαρχικές δομές μέσα στο κόμμα.#metonKyriako #ert

Ο τρόπος χειρισμού της ΕΡΤ με το "μαύρο" δεν ήταν ο πλέον σωστός, αλλά ούτε και αυτό που προϋπήρχε ήταν σωστό #skaitv

- *Κανόνας 3ος : «Δεν» + «έχω» + αντικείμενο*

  Το βοηθητικό ρήμα «έχω» χρησιμοποιείται και αυτό ευρέως προκειμένου να αποδώσει μια ιδιότητα στο υποκείμενο ή στο αντικείμενο της πρότασης. Όπως και στις δύο προηγούμενες περιπτώσεις η ύπαρξη του «δεν» προκαλεί αντιστροφή της πολικότητας της αποδιδόμενης ιδιότητας.

Δεν έχει καμία λογική αυτό που έκανε. ΚΑΜΙΑ! http://fb.me/sKIzsjW9

Στην ΕΣΣΔ δεν ειχαμε παρανομη εισοδο μεταναστών στη χώρα. 70 χρονια δεν προσπαθησε να μπει ποτέ, κανεις. ΕΣΣΔ. Παντα μπροστα.

Δεν έχω εμπιστοσύνη στον @atsipras να διαχειριστεί τα εθνικά μας θέματα. #skai #metonKyriako

Επειδή δεν έχει σημασία μόνο το τι λες, αλλά και πως το λες. Δωρεάν επικοινωνιακά μαθήματα. http://fb.me/2Un9dZVob

Καλά δεν έχουν άδικο.. εκτός απ τον τσιπρα (το παλευει) ε ποιο άχρηστους υπουργούς μα το θεό δεν έχω ξαναδεί γτ δεν τους αλλάζει #vouli

Η τρομοκρατία δεν έχει πατρίδα, δεν έχει ιδεολογία, έχει μόνο το χρώμα του αίματος. #Turkey #AnkaraBombing

Μέχρι στιγμής λοιπόν τα αρχικώς ακατέργαστα tweet έχουν φιλτραριστεί και μετασχηματιστεί και με τη βοήθεια του λεξικού έχει υπολογιστεί η βαθμολογία των λέξεων (σε κάθε ένα από αυτά) που ήταν αποθηκευμένες στο λεξικό. Στο σημείο αυτό υπενθυμίζουμε πως στο στάδιο της προεπεξεργασίας και συγκεκριμένα σε αυτό της αφαίρεσης των stop words, δεν συμπεριλάβαμε σε αυτές, σκοπίμως, το μόριο «δεν» και τα ρήματα «είμαι» και «έχω», γιατί θα μας βοηθήσουν στον εντοπισμό της άρνησης βάση των κανόνων που μόλις εξηγήσαμε.

Η ανίχνευση της άρνησης λοιπόν υλοποιείται ω εξής: Αφού υπολογιστεί η βαθμολογία κάθε λέξης που εντοπίστηκε στο λεξικό, πηγαίνουμε στο φιλτραρισμένο και μετασχηματισμένο tweet και για κάθε μία από τις λέξεις που εντοπίσαμε στο λεξικό κάνουμε τις εξής ενέργειες:

[18] http://vprassas.blogspot.gr/2011/01/blog-post_1045.html

1) Εάν η ακριβώς προηγούμενη λέξη είναι «δεν», τότε αντιστρέφουμε τη βαθμολογία της λέξης.
2) Εάν η ακριβώς προηγούμενη λέξη είναι το ρήμα «είμαι» και η λέξη δύο θέσεις πιο δίπλα είναι το «δεν», αντιστρέφουμε τη βαθμολογία της λέξης.
3) Εάν η ακριβώς προηγούμενη λέξη είναι το ρήμα «έχω» και η λέξη δύο θέσεις πιο δίπλα είναι το «δεν» και πάλι αντιστρέφουμε τη βαθμολογία της λέξης.
4) Εάν δε συμβαίνει τίποτα από τα παραπάνω η λέξη διατηρεί την αρχική της βαθμολογία.

Όπως προείπαμε βεβαίως η άρνηση αποτελεί ένα εξαιρετικά πολύπλοκο φαινόμενο. Για να πάρει ο αναγνώστης μια ιδέα του πόσο σύνθετος μπορεί να γίνει ο εντοπισμός της άρνησης και ο μετέπειτα χειρισμός της, παραθέτουμε ορισμένα ενδεικτικά παραδείγματα:

Παράδειγμα 1ο

Το Σύγχρονο στην πολιτική δεν ταυτιζεται με ψέμα,λαϊκισμό με αυταρχισμο,εθνικό διχασμο, οικονομική κ κοινωνική ισοπέδωση!Το Νέο μπορει!

Στη συγκεκριμένη περίπτωση το μόριο «δεν», δεν αναφέρεται μόνο στη λέξη «ψέμα», αλλά και σε όλες τις επόμενες που ακολουθούν δηλαδή «λαϊκισμό», «αυταρχισμό», «διχασμό» και «ισοπέδωση».

Παράδειγμα 2ο

Δεν θα έλεγα ότι είναι ηθικός αυτουργός της χθεσινής αθλιότητας ο κ. Φίλης.Αλλά στο παρελθόν ο ΣΥΡΙΖΑ ανέχθηκε παρόμοιες συμπεριφορές #984

Ενώ η φράση «ηθικός αυτουργός» έχει ξεκάθαρα έντονο αρνητικό συναισθηματικό υπόβαθρο, η φράση που προηγείται αν και περιέχει το «δεν», δεν αντιστρέφει την πολικότητα του συναισθήματος, αλλά δίνει έναν ουδέτερο τόνο στην πρόταση.

Παράδειγμα 3ο

Κυριάκο αν δεν πεθάνει ο γκαντεμόσαυρος Πρωθυπουργό δεν σε βλέπω #eklogesneadimokratia

Όπως αναφέραμε και στην παράγραφο 2.2, ο υποθετικός λόγος χρήζει ιδιαιτέρου χειρισμού, καθώς εν προκειμένω αναιρεί την ύπαρξη του «δεν» και η λέξη «πεθάνει» διατηρεί το αρχικό αρνητικό συναίσθημα.

Ας δούμε λοιπόν μερικά παραδείγματα στα οποία φαίνεται ο χειρισμός της άρνησης. Θα θεωρήσουμε ένα παράδειγμα από κάθε κατηγορία κανόνων και θα παρουσιαστούν όλα τα μέχρι τώρα στάδια του αλγορίθμου, δηλαδή η προεπεξεργασία, ο μετασχηματισμος, ο εντοπισμός των λέξεων και οι αντίστοιχες βαθμολογίες και τέλος η ανίχνευση άρνησης και ο υπολογισμός του τελικού σκορ κάθε λέξης, ανάλογα με το αν υπήρχε άρνηση ή όχι.

Ακατέργαστο tweet	*Φιλτραρισμένο tweet*	*Μετ/νο tweet*	*Λέξεις*	*Αρχικό σκορ*	*Τελικό σκορ*
Τσίπρας: Οι Έλληνες δεν τρομοκρατούνται- Έρχεται το «Σαμαράς exit» #oikonomia http://bit.ly/1z4NAJj	τσίπρας έλληνες δεν τρομοκρατούνται έρχεται σαμαράς	τσίπρας Έλληνας δεν τρομοκρατώ έρχομαι σαμαράς	τρομοκρατώ έρχομαι	[-3,0]	[3,0]
Απόδειξη ότι δεν είναι όλοι τρελοί στο ΣΥΡΙΖΑ! http://fb.me/7slR2T1Cf	απόδειξη ότι δεν είναι τρελοί συριζα	απόδειξη ότι δεν είμαι τρελός συριζα	τρελός	-2	2
Δεν έχω εμπιστοσύνη στον @atsipras να διαχειριστεί τα εθνικά μας θέματα. #skai #metonKyriako	δεν έχω εμπιστοσύνη διαχειριστεί εθνικά θέματα	δεν έχω εμπιστοσύνη διαχειρίζομαι εθνικός θέμα	εμπιστοσύνη	4	-4

Πίνακας 9: Παραδείγματα χειρισμού άρνησης σε tweets

3.8 Υπολογισμός τελικού σκορ και ταξινόμηση

Πλέον αφού σε κάθε λέξη που εντοπίστηκε στο λεξικό έχει ανατεθεί το σωστό σκορ, αναλόγως με το αν εντοπίστηκε άρνηση ή όχι, είμαστε έτοιμοι να υπολογίσουμε το συνολικό σκορ ολόκληρου του tweet και να το ταξινομήσουμε σε μια από τις τρείς πιθανές κλάσεις (αρνητικό, θετικό, ουδέτερο) βάση αυτού.

Για τον υπολογισμό του συνολικού σκορ αθροίζουμε τις βαθμολογίες των λέξεων που εντοπίστηκαν και διαιρούμε με το πλήθος των, έτσι ώστε να υπολογίσουμε στην ουσία έναν σταθμισμένο μέσο όρο των επιμέρους βαθμολογιών, δεδομένου οτι όπως αναφέραμε και στην παράγραφο 3.5 κάθε λέξη παίρνει μια βαθμολογία στο έυρος [-4,4]. Στη συνέχεια κάθε tweet ταξινομείται ως θετικό αν η συνολική βαθμολογία είναι μεγαλύτερη του μηδενός, ως αρνητικό αν η συνολική βαθμολογία είναι μικρότερη του μηδενός και ως ουδέτερο αν η συνολική βαθμολογία ισούται με το 0.

Ακολουθεί ένα παράδειγμα στο οποίο φαίνονται όλα τα στάδια του αλγορίθμου που υλοποιήσαμε, από την προεπεξεργασία των δεδομένων μέχρι και την τελική ταξινόμηση.

Στάδιο	
Ακατέργαστο tweet	RT @Andrey_Vyshinsk: Ο ΠΑΝΟΣ ΑΚΑ Ψεκασμένος βρήκε συνωμοσία Σημιτικών Δημαριτων Ποταμιού Κ Ακροδεξιών Πια Αγκάθα ρε γατάκια #eklogesneadi…
Φιλτραρισμένο tweet	πανος ακα ψεκασμένος βρήκε συνωμοσία σημιτικών δημαριτων ποταμιού ακροδεξιών αγκάθα ρε γατάκια
Μετασχηματισμένο tweet	πανοσ ακα ψεκάζω βρίσκω συνωμοσία σημιτικός δημαριτων ποτάμι ακροδεξιός αγκάθα ρε γατάκι
Λέξεις που εντοπίστηκαν	ψεκάζω βρίσκω ακροδεξιός γατάκι
Αρχικό σκορ/λέξη	[-2,0,-4,-2]
Ύπαρξη Άρνησης	0
Τελικό σκορ/ λέξη	[-2,0,-4,-2]
Συνολικό σκορ	-2,00
Ταξινόμηση	Αρνητικό

Πίνακας 10: Εφαρμογή αλγορίθμου σε tweet – Παρουσίαση επιμέρους σταδίων

Κεφάλαιο 4 Πειραματικά αποτελέσματα

4.1 Σύνολο δεδομένων

Στο κεφάλαιο αυτό παρουσιάζονται τα αποτελέσματα της αξιολόγησης του ταξινομητή που υλοποιήσαμε, όπως αυτός περιγράφτηκε στο προηγούμενο κεφάλαιο. Το σύνολο δεδομένων που χρησιμοποιήσαμε αφορά τα αποτελέσματα του δεύτερου γύρου των εσωκομματικών εκλογών της Νέας Δημοκρατίας που ανακοινώθηκαν κατά την 11η Ιανουαρίου 2016. Η επιλογή της συγκεκριμένης θεματολογίας έγινε αφενός γιατί εκείνη την περίοδο ήταν ίσως το σημαντικότερο θέμα της πολιτικής επικαιρότητας και αφετέρου σχεδόν πάντα οι εκλογικές διαδικασίες μονοπωλούν το ενδιαφέρον των χρηστών στα κοινωνικά δίκτυα. Σε τέτοιες περιπτώσεις τα σχόλια των χρηστών έχουν ιδιαιτέρως πλούσιο συναισθηματικό περιεχόμενο, αφού εκφράζουν με έντονο και καυστικό τρόπο είτε την επιδοκιμασία τους είτε την αποδοκιμασία τους προς τα εκλογικά αποτελέσματα, καθιστώντας αυτά τα σύνολα δεδομένων γεμάτα σημαίνουσα πληροφορία και ιδανικά για την αξιολόγηση ενός συστήματος Ανάλυσης Συναισθήματος.

Η συλλογή του συνόλου δεδομένων έγινε αναζητώντας μέσω του Twitter API όλα τα tweets που περιείχαν το hashtag *"#eklogesnd".* Συνολικά συλλέχθηκαν κοντά στα 3000 tweets, εκ των οποίων επιλέχθηκαν τυχαία 200 για την τελική αξιολόγηση του ταξινομητή. Η *επισημείωση* αυτών των tweets έγινε από πέντε διαφορετικούς βαθμολογητές - επισημειωτές (raters - annotators). Στο σημείο αυτό θα πρέπει να διευκρυνιστεί το εξής: Στο εν λόγω σύνολο δεδομένων, αλλά και γενικότερα σε tweets που αφορούν πολιτικά γεγονότα, οι γνώμες των χρηστών δεν αποτυπώνονται, κατά κανόνα, με τον πλέον ευθή, σαφή και κατανοητό τρόπο και ο λόγος είναι οτι υπάρχει έντονο το φαινόμενο του σαρκασμού και της ειρωνείας. Είναι χαρακτηριστικό το γεγονός οτι σε αρκετά από τα tweets που συλλέξαμε, ήταν αρκετά δύσκολο ακόμα και για εμάς τους ίδιους να αποφανθούμε με απόλυτη σιγουριά για το συναίσθημα που εξέφραζε ο χρήστης, καθώς πολύ συχνά γινόντουσαν αναφορές σε τρίτα πρόσωπα, και απαιτούνταν βαθειά γνώση του ευρύτερου πολιτικού γίγνεσθαι για την κατανόηση των συχετισμών (πχ Τέρης Χρυσός, Χριστοφοράκος, Siemens). Για το λόγο αυτό ζητήθηκε από τους βαθμολογητές η επισημείωση των tweets να γίνει μόνο βάση του *λεξικολογικού περιεχομένου* και όχι του σημασιολογικού.

Το σύνολο δεδομένων και οι επισημειώσεις των πέντε βαθμολογητών είναι διαθέσιμα στο Παράρτημα στον Πίνακα 25.

4.2 Πορεία Αξιολόγησης

Για την αξιολόγηση του συστήματος χρησιμοποιήσαμε τις μετρικές που περιγράψαμε στην παράγραφο 2.6, δηλαδή τη συμφωνία μεταξύ κριτών (inter-rater agreement), τη συνολική ακρίβεια (accuracy), την ακρίβεια (precision), την ανάκληση (recall) και το F-measure. Συγκεκριμένα η αξιολόγηση έγινε ως εξής:

- Σε ό,τι αφορά τη συμφωνία μεταξύ κριτών δημιουργήσαμε 6 ομάδες. Η 1η ομάδα (Group 0) περιείχε τις βαθμολογίες που έδωσαν οι πέντε επισημειωτές. Οι υπόλοιπες ομάδες δημιουργήθηκαν αφαιρώντας κάθε φορά έναν από τους επισημειωτές και τοποθετώντας στη θέση του, την έξοδο του ταξινομητή μας. Για παράδειγμα στη 2η ομάδα (Group 1) αφαιρέσαμε τις βαθμολογίες του 1ου επισημειωτή και τοποθετήσαμε στη θέση τους τις εξόδους του ταξινομητή μας, στην 3η ομάδα (Group 2) αφαιρέσαμε τις βαθμολογίες του 2ου επισημειωτή και τοποθετήσαμε στη θέση τους τις εξόδους του ταξινομητή μας κ.ο.κ. Τελικά υπολογίστηκε η συνολική συμφωνία μεταξύ κριτών για κάθε ομάδα, καθώς και η συμφωνία μεταξύ κριτών για κάθε κλάση (θετική, αρνητική, ουδέτερη) μέσα σε κάθε ομάδα.

  Να σημειωθεί οτι για τον υπολογισμό της συμφωνίας μεταξύ χρηστών χρησιμοποιήσαμε τη συνάρτηση fleiss.m[1], οποία υλοποιήθηκε από τον Cardillo G. (2007) και είναι ελεύθερα διαθέσιμη στον ιστότοπο ανταλλαγής αρχείων του Mathworks.[2]

- Απο εκεί και πέρα, για κάθε έναν από τους βαθμολογητές, δηλαδή θεωρώντας ως «σωστή» ταξινόμηση την επισημείωση των βαθμολογητών, υπολογίστηκε η συνολική ακρίβεια του συστήματος ταξινόμησης, καθώς και η ακρίβεια, η ανάκληση και το F-measure για *κάθε κλάση* ξεχωριστά.

Για την αξιολόγηση του ταξινομητή χρησιμοποιήσαμε δύο διαφορετικές εκδοχές των δεδομένων επισημείωσης.

- Στην *1η εκδοχή* τα δεδομένα επισημείωσης χρησιμοποιήθηκαν ακριβώς όπως παρήχθησαν από τους πέντε βαθμολογητές. Δεν έγινε καμία απολύτως τροποποίηση.

- Στη *2η εκδοχή* αφαιρέσαμε μερικά tweets από το αρχικό σύνολο για τον εξής λόγο: Κατά την επισκόπηση των βαθμολογιών των επισημειωτών διαπιστώθηκε οτι μερικά tweets είχαν βαθμολογηθεί λαμβάνοντας υπόψιν όχι μόνο το λεξικολογικό, αλλά και το σημασιολογικό περιεχόμενο, ενώ σε

[1] http://www.mathworks.com/matlabcentral/fileexchange/15426-fleiss-es-kappa

[2] http://www.mathworks.com/matlabcentral/fileexchange/

κάποια άλλα ήταν απλά προφανής η λάθος επισημείωση. Επιπλέον σε κάποια άλλα υπήρχε εξαιρετικά μικρή συμφωνία μεταξύ των κριτών και ως εκ τούτου κρίθηκαν ακατάλληλα έως και παραπλανητικά για την αξιολόγηση του συστήματος. Μερικά ενδεικτικά παραδείγματα είναι τα παρακάτω:

Tweet	*Επισ. 1*	*Επισ. 2*	*Επισ. 3*	*Επισ. 4*	*Επισ. 5*
Οι χορηγοί πρέπει να ανταμείβονται! Άλλωστε πρέπει να στηρίξει και τη νύφη (Σία)! https://t.co/rBQS58xPvb	Neutral	Neutral	Neutral	Neutral	Neutral
Μπράβο @kmitsotakis. Δώσε τη μάχη για τη χώρα. #metonKyriako #eklogesneadimokratia #eklogesnd	Neutral	Positive	Neutral	Neutral	Neutral
RT @yioults1: Δηλαδή στην εκεί στην ΝΔ τα βάλανε κάτω και είπαν ότι θα πάνε μπροστά γυρνώντας πίσω #eklogesneadimokratia	Neutral	Negative	Neutral	Positive	Positive
Επικράτηση Μητσοτάκη στη γαλάζια μάχη για την Προεδρία #eklogesneadimokratia @kathimerini_gr https://t.co/0XwGClMgYG	Positive	Positive	Positive	Positive	Positive

Πίνακας 11: Παραδείγματα tweets με προβληματική επισημείωση

Στην πρώτη περίπτωση, αν ληφθεί υπόψιν μόνο το λεξικολογικό περιεχόμενο , το tweet έχει ξεκάθαρα θετικό συναισθηματικό υπόβαθρο, κάτι που ισχύει και για τη δεύτερη περίπτωση. Στην τρίτη περίπτωση είναι εξαιρετικά ασαφής η πρόθεση του συγγραφέα, εξού και το γεγονός οτι ταξινομείται και στις τρείς πιθανές κατηγορίες, ενώ στην τέταρτη εφόσον η δήλωση μπορεί να θεωρηθεί ως θετική για τον νικήτη και τους υποστηρικτές του, και αρνητική για τους αντιπάλους του θα έπρεπε να βαθμολογηθεί ως ουδέτερη αφού δεν εκφράζει γνώμη, αλλά απλά μεταφέρει ένα γεγονός.

Το σύνολο των tweets που αφαιρέθηκαν βρίσκεται και αυτό στο Παράρτημα στον Πίνακα 26.

4.3 Αξιολόγηση Συστήματος

Αφού κατανοήσαμε τη δομή και τις ιδιαιτερότητες των test δεδομένων, είμαστε έτοιμοι να περάσουμε στην τελική αξιολόγηση του συστήματος.

4.3.1 1ο πείραμα

Στο 1ο πείραμα χρησιμοποιήσαμε την 1η εκδοχή του συνόλου δεδομένων, δηλαδή η ταξινόμηση έγινε θεωρώντας αυτούσιο και χωρίς καμία παραλλαγή το σύνολο των επισημειώσεων που παρήχθει από τους πέντε επισημειωτές.

Στον πίνακα που ακολουθεί παρουσιάζεται, τόσο η συνολική όσο και η επιμέρους για κάθε κλάση, συμφωνία μεταξύ κριτών στις έξι ομάδες δεδομένων επισημείωσης (βλέπε παράγραφο 4.2). Στην τελευταία στήλη παρουσιάζονται οι αντίστοιχοι *μέσοι όροι* μόνο για τα Group 1-5, εφόσον μόνο σε αυτά εμφανιζόντουσαν τα αποτελέσματα του ταξινομητή.

	Group 0	***Group 1***	***Group 2***	***Group 3***	***Group 4***	***Group 5***	***Average (Gr 1-5)***
Total	83.24	62.80	65.41	65.90	63.31	65.03	64.49
Negative	90.14	74.69	76.80	76.26	74.69	75.79	75.65
Neutral	77.02	50.98	54.51	55.41	51.95	54.75	53.52
Positive	82.60	63.04	65.28	66.39	63.62	64.90	64.65

Πίνακας 12: Συμφωνία μεταξύ κριτών - Πείραμα 1ο

Στον επόμενο πίνακα φαίνεται η συνολική ακρίβεια του συστήματος και για τις πέντε περιπτώσεις βαθμολογητών:

	Annot 1	***Annot 2***	***Annot 3***	***Annot 4***	***Annot 5***
Accuracy	60.50	61.00	58.50	60.50	59.50

Πίνακας 13: Συνολική ακρίβεια συστήματος (ξεχωριστά για κάθε επισημειωτή) – Πείραμα 1ο

Τέλος, παρουσιάζεται η ακρίβεια, η ανάκληση και το F-measure για κάθε μια από τις κλάσεις και οι αντίστοιχοι μέσοι όροι και για τις πέντε περιπτώσεις βαθμολογητών:

	Annot 1			***Annot 2***			***Annot 3***			***Annot 4***			***Annot 5***		
	P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
Negative	70.15	67.14	68.61	73.13	67.12	70.00	68.66	66.67	67.65	71.64	68.57	70.07	73.13	70.00	71.53
Neutral	54.32	55.70	55.00	50.62	56.16	53.25	48.15	54.17	50.98	51.85	54.55	53.16	46.91	55.88	51.01
Positive	57.69	58.82	58.25	61.54	59.26	60.38	61.54	54.24	57.66	59.62	58.49	59.05	61.54	51.61	56.14

Πίνακας 14: Ακρίβεια, ανάκληση και F-measure - Πείραμα 1ο

	Average		
	P	***R***	***F1***
Negative	71.34	67.90	69.57
Neutral	50.37	55.29	52.68
Positive	60.39	56.48	58.23

Πίνακας 15: Μέσοι όροι ακρίβειας, ανάκλησης και F-measure - Πείραμα 1ο

4.3.2 2ο πείραμα

Στο 2ο πείραμα χρησιμοποιήσαμε τη 2η εκδοχή του συνόλου δεδομένων όπου αφαιρέσαμε κάποια tweets στα οποία θεωρήσαμε οτι η δοθείσα επισημείωση από τους βαθμολογητές δεν ανταποκρίνεται στην πραγματικότητα. Ακολουθούν τα αντίστοιχα αποτελέσματα:

	Group 0	***Group 1***	***Group 2***	***Group 3***	***Group 4***	***Group 5***	***Average (Gr 1-5)***
Total	83.99	67.09	69.45	70.14	67.39	69.29	68.45
Negative	90.07	76.25	78.50	77.70	76.25	77.42	76.96
Neutral	77.61	56.13	59.23	60.40	56.77	59.72	58.31
Positive	84.11	69.02	70.76	72.59	69.34	71.00	70.29

Πίνακας 16: Συμφωνία μεταξύ κριτών - Πείραμα 2ο

	Annot 1	***Annot 2***	***Annot 3***	***Annot 4***	***Annot 5***
Accuracy	64.86	64.86	61.62	64.86	64.32

Πίνακας 17: Συνολική ακρίβεια συστήματος (ξεχωριστά για κάθε επισημειωτή) – Πείραμα 2ο

	Annot 1			***Annot 2***			***Annot 3***			***Annot 4***			***Annot 5***		
	P	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***	***P***	***R***	***F1***
Negative	75.81	68.12	71.76	**79.03**	68.06	73.13	72.58	67.16	69.77	**77.42**	**69.57**	73.28	**79.03**	**71.01**	74.81
Neutral	56.58	62.32	59.31	52.63	62.50	57.14	50.00	59.38	54.29	53.95	62.12	57.75	50.00	**64.41**	56.30
Positive	63.83	63.83	63.83	**65.96**	63.27	64.58	65.96	57.41	61.39	65.96	62.00	63.92	**68.09**	56.14	61.54

Πίνακας 18: Ακρίβεια, ανάκληση και F-measure - Πείραμα 2ο

	Average		
	P	***R***	***F1***
Negative	**76.77**	**68.78**	72.55
Neutral	52.63	62.15	56.96
Positive	**65.96**	60.53	63.05

Πίνακας 19: Μέσοι όροι ακρίβειας, ανάκλησης και F-measure - Πείραμα 2ο

4.4 Σύνοψη - Ερμηνεία αποτελεσμάτων

Μετά την εξαγωγή των μετρικών για την αξιολόγηση του συστήματος, θα κάνουμε μια προσπάθεια να ερμηνεύσουμε τα πειραματικά αποτελέσματα προκειμένου να κατανοήσουμε σε ποιά σημεία ο αλγόριθμός μας είχε την επιθυμητή απόδοση και σε ποιά όχι, καθώς και ποιοί ήταν οι λόγοι που οδήγησαν σε μια τέτοια συμπεριφορά.

Υπενθυμίζουμε τα εξής: Ο βασικός στόχος για ένα σύστημα Ανάλυσης Συναισθήματος είναι να λειτουργεί τόσο καλά όσο και ο άνθρωπος, δηλαδή να ταξινομεί ένα tweet όπως θα το ταξινομούσε ένας άνθρωπος. Στο κομμάτι αυτό θα μας βοηθήσει η συμφωνία μεταξύ κριτών, όπου ουσιαστικά αυτό που ζητάμε είναι αν αντικαταστήσουμε έναν επισημειωτή με τα αποτελέσματα ταξινόμησης του συστήματός μας, να μην υπάρχει σημαντική μεταβολή της συμφωνίας μεταξύ κριτών. Από την άλλη μεριά, τα μεγέθη της ακρίβειας και της ανάκλησης θα μας βοηθήσουν να μελετήσουμε τη συμπεριφορά του συστήματός μας για κάθε μια από τις κλάσεις, και να διαπιστώσουμε ποιές περιπτώσεις tweets μπορούσε να χειριστεί με επιτυχία και σε ποιές δυσκολευόταν να κάνει σωστή ταξινόμηση.

4.4.1 Σύγκριση των πειραμάτων

Ας αρχίσουμε με μια σύγκριση των αποτελεσμάτων στα δύο πειράματα. Για λόγους ευκολίας συγκεντρώνουμε τα αποτελέσματα των δύο προηγούμενων παραγράφων στους εξής πίνακες:

	Πείραμα 1ο		***Πείραμα 2ο***	
	Group 0	*Average (Gr 1-5)*	*Group 0*	*Average (Gr 1-5)*
Total	83.24	64.49	83.99	68.45
Negative	90.14	75.65	90.07	**76.96**
Neutral	77.02	53.52	77.61	58.31
Positive	82.60	64.65	84.11	**70.29**

Πίνακας 20: Σύγκριση συμφωνίας μεταξύ κριτών στα δύο πειράματα

	Precision		***Recall***		***F-measure***	
	Πείραμα 1ο	*Πείραμα 2ο*	*Πείραμα 1ο*	*Πείραμα 2ο*	*Πείραμα 1ο*	*Πείραμα 2ο*
Negative	71.34	**76.77**	67.90	**68.78**	69.57	72.55
Neutral	50.37	**52.63**	55.29	62.15	52.68	56.96
Positive	60.39	**65.96**	56.48	60.53	58.23	63.05

Πίνακας 21: Σύγκριση ακρίβειας, ανάκλησης και F-measure στα δύο πειράματα

Είναι προφανές με μια πρώτη ματιά οτι η απόδοση του ταξινομητή στο 2ο πείραμα παρουσιάζει σημαντική βελτίωση σε όλα τα επίπεδα σε σχέση με το 1ο πείραμα.

Αυτό είναι λογικό καθώς η προβληματική επισημείωση μειώνει σημαντικά την απόδοση του ταξινομητή. Παρόλα αυτά η ποιοτική συμπεριφορά και στα δύο πειράματα ακολουθεί την ίδια λογική. Για το λόγο αυτό θα επικεντρώσουμε το ενδιαφέρον μας στο *2ο πείραμα*, καθώς τα συμπεράσματα που θα εξαχθούν ανταποκρίνονται και στα αποτελέσματα του 1ου πειράματος.

4.4.2 Ερμηνεία αποτελεσμάτων

Ας παρατηρήσουμε καταρχάς το Group 0, δηλαδή την ομάδα που αποτελείται μόνον από τις βαθμολογίες των επισημειωτών. Αν λάβουμε υπόψιν τον Πίνακα 1, το Group 0 παρουσιάζει σχεδόν τέλεια συμφωνία (almost perfect agreement), αφού $k = 83.99\% > 81\%$. Παρόλα αυτά η σχεδόν τέλεια συμφωνία επιτυγχάνεται σχετικά οριακά, γεγονός που αποτελεί ένδειξη οτι το συναισθηματικό υπόβαθρο των tweets που εξετάσαμε δεν ήταν σε καμία περίπτωση απολύτως ξεκάθαρο. Αυτό γίνεται ιδιαιτέρως εμφανές στην συμφωνία μεταξύ κριτών στην ουδέτερη κλάση, όπου οι βαθμολογητές παρουσιάζουν συμφωνία 77.61%, ποσοστό αρκετά μικρό. Αυτό είναι απολύτως λογικό να συμβαίνει στο συγκεκριμένο σύνολο δεδομένων καθώς αφενός η ουδέτερη κλάση αποτελεί έτσι και αλλιώς το όριο μεταξύ της θετικής και της αρνητικής κλάσης και συνεπώς για παράδειγμα ενα tweet που για κάποιον είναι ασθενώς θετικό για κάποιον άλλο είναι ουδέτερο, και αφετέρου η σχεδόν μόνιμη χρήση της ειρωνείας και οι υπαινιγμοί που απαιτούν γνώση του ευρύτερου σημασιολογικού πλαισίου καθιστούν ακόμα πιο δυσδιάκριτα τα όρια μεταξύ του πότε ένα tweet έχει ουδέτερη χροιά και του πότε είναι θετικό ή αρνητικό. Από την άλλη, σε ό,τι αφορά την αρνητική κλάση, εκεί παρουσιάζεται σχεδόν απόλυτη συμφωνία που αγγίζει το 90%, πράγμα που σημαίνει οτι όταν ένα tweet επισημειωθεί ως αρνητικό από κάποιον βαθμολογητή, τότε στις περισσότερες περιπτώσεις θα συμφωνήσουν και οι υπόλοιποι βαθμολογητές. Μια ερμηνεία αυτού του γεγονότος, θα μπορούσε να είναι ότι από τη μία μεριά, όταν οι χρήστες θέλουν πραγματικά να εκφράσουν την πλήρη αντίθεση ή διαφωνία τους σε σχέση με κάτι, προσπαθούν να είναι σαφείς, ξεκάθαροι και κατανοητοί στις διατυπώσεις τους, χωρίς να χρησιμοποιούν ειρωνεία και αμφιλεγόμενα νοήματα, ενώ από την άλλη, η αρκετά συχνή χρήση άσεμνων ή υβριστικών λέξεων και φράσεων καθιστά προφανές το αρνητικό συναισθηματικό υπόβαθρο. Τέλος αρκετά μεγάλη συμφωνία υπάρχει και στη θετική κλάση όπου βρίσκεται στο 84.11%.

Αν θεωρήσουμε τη συνολική συμφωνία μεταξύ κριτών ως το πλέον ενδεικτικό μέτρο της απόδοσης του συστήματος που υλοποιήσαμε τότε μπορούμε να πούμε οτι το σύστημά μας, παρουσιάζει αρκούντως ικανοποιητική συμπεριφορά και αυτό γιατί η συνολική συμφωνία μεταξύ κριτών για τα Group 1-5 βρίσκεται στο 68.45%, δηλαδή αντιστοιχεί σε σημαντική συμφωνία (substantial agreement). Συνεπώς αν και η συμφωνία «πέφτει» κατά ένα επίπεδο, δηλαδή υποβιβάζεται από «σχεδόν τέλεια» σε «σημαντική», σε καμία περίπτωση δεν μπορούμε να πούμε οτι το σύστημά μας παράγει παραπλανητικά αποτελέσματα, αλλά αντίθετα φαίνεται τις περισσότερες φορές να συμφωνεί με τους ανθρώπινους κριτές. Θα μελετήσουμε τώρα κάθε κλάση ξεχωριστά.

Το πρώτο πράγμα που παρατηρούμε μελετώντας τους πίνακες 20 και 21, είναι η πολύ μικρή συμφωνία μεταξύ κριτών στην ουδέτερη κλάση που βρίσκεται στο 58.31% και αντιστοιχεί σε μέτρια συμφωνία (moderate agreement). Ένα άλλο ενδιαφέρον στοιχείο είναι και η πολύ μικρή ακρίβεια στην ουδέτερη κλάση που βρίσκεται στο 52.63%, την ίδια στιγμή που η ανάκληση είναι σχεδόν δέκα ποσοστιαίες μονάδες πιο πάνω, σε ποσοστό 62.15%. Αυτό σημαίνει πρακτικά οτι από τα tweets που ταξινομούμε ως ουδέτερα, μόλις τα μισά είναι στην πραγματικότητα ουδέτερα, δηλαδή στην ουσία ταξινομούμε πολύ περισσότερα tweets ως ουδέτερα από οτι θα έπρεπε, γι'αυτό και τα σχετικά ψηλά ποσοστά ανάκλησης (συγκρινόμενα με την ακρίβεια). Η χαμηλή ακρίβεια στην ουδέτερη κλάση οφείλεται στον μη εντοπισμό συναισθηματικά φορτισμένων λέξεων, επειδή αυτές δεν ήταν καταχωρημένες στο λεξικό: Ακόμα λοιπόν και αν υπήρχαν θετικές ή αρνητικές λέξεις σε κάποια tweets, αυτές δεν εντοπίστηκαν, με αποτέλεσμα τα tweets να ταξινομήθηκαν ως ουδέτερα αντί ως θετικά ή αρνητικά, εξού και η χαμηλή ακρίβεια. Από την άλλη μεριά, η μικρή συμφωνία μεταξύ κριτών και ανάκληση, οφείλονται στο γεγονός οτι μερικά tweets που ήταν ουδέτερα, το σύστημά μας τα ταξινόμησε είτε ως θετικά είτε ως αρνητικά επειδή εντόπισε κάποιες θετικές ή αρνητικές λέξεις αντίστοιχα, όπως για παράδειγμα στο tweet «Εδώ στη Δράση αναζητούμε εθελοντές για να βοηθήσουν στη μετακόμηση. Πληροφορίες εντός #eklogesneadimokratia», όπου εντοπίστηκε η λέξη «εθελοντές» που έχει εν γένει θετικό υπόβαθρο, αλλά το tweet επισημειώθηκε ως ουδέτερο. Από όλα τα παραπάνω γίνεται φανερό πόσο «ασταθής» και δύσκολη στην πρόβλεψη είναι η ουδέτερη κλάση, αφού σε μια πρόταση μπορούν να υπάρχουν αρκετές λέξεις με συναισθηματικό υπόβαθρο, αλλά το συνολικό συναίσθημα να είναι ουδέτερο. Για το χειρισμό τέτοιων περιπτώσεων απαιτείται γνώση και εκμετάλλευση του ευρύτερου σημασιολογικού πλαισίου.

Στη συνέχεια ένα δεύτερο στοιχείο που είναι εμφανές είναι τα πολύ ενθαρρυντικά αποτελέσματα στην αρνητική κλάση, όπου ο μέσος όρος της συμφωνίας κριτών για τα Group 1-5 βρίσκεται στο 76.96% (δηλαδή για 4% χάνεται η σχεδόν τέλεια συμφωνία), ενώ η μέση ακρίβεια είναι 76.77% και η μέση ανάκληση 68.78%. Αυτό σημαίνει οτι το σύστημά μας χειρίζεται με αρκετά μεγάλη αποτελεσματικότητα τα αρνητικά tweets, και τις περισσότερες φορές τα ταξινομεί όπως θα τα ταξινομούσε και ένας άνθρωπος. Οι βασικοί λόγοι που δεν εντοπίζονται κάποια αρνητικά tweets είναι οτι αφενός, όπως και πριν, δεν εντοπίζονται οι αντίστοιχες λέξεις γιατί δεν υπάρχουν στο λεξικό, και αφετέρου χάνονται οι περιπτώσεις στις οποίες χρησιμοποιούνται ειρωνικά, θετικές λέξεις (πχ «Αφού είναι βέβαιο οτι δε θα απαλλαγούμε ποτέ από αυτό το σόι δν αλλάζουμε όνομα στη χώρα? Μητσοτακισταν *καλό* ακούγεται #eklogesneadimokratia») και οι περιπτώσεις αντικρουόμενων απόψεων όπου εντοπίζεται μόνο ο θετικός όρος (πχ «Οι *καλές* ειδήσεις συνεχίζονται!!! *ΞΥΔΑΚΙ ΨΕΚΑΣΜΕΝΟΙ*!!!! https://t.co/GqBpnqxR97»). Ο λόγος που παρουσιάζεται τόσο ικανοποιητική απόδοση στην αρνητική κλάση θεωρούμε οτι οφείλεται στην επέκταση του λεξικού, καθώς οι περισσότερες λέξεις που εντοπίστηκαν και προστέθηκαν είχαν αρνητική χροιά και επιπλέον βοήθησε η σωστή επισημείωση των ισχυρά «αρνητικών» λέξεων, καθώς φάνηκε να επαληθεύεται ο ισχυρισμός μας, οτι δηλαδή όταν υπάρχουν τέτοιες λέξεις σε

ένα tweet (πχ ακροδεξιός, φεουδαρχικός ή διάφορες βωμολοχίες), τότε υπαγορεύουν και το συνολικό συναίσθημα, στις περισσότερες των περιπτώσεων.

Τέλος τα αποτελέσματα στη θετική κλάση είναι αρκετά ικανοποιητικά, καθώς ο μέσος όρος της συμφωνίας κριτών για τα Group 1-5 βρίσκεται στο 70.29%, δηλαδή υπάρχει σημαντική συμφωνία (substantial agreement), ενώ η μέση ακρίβεια και ανάκληση είναι στο 65.96% και 60.53%. Η σχετικά μικρή ανάκληση οφείλεται σε δύο λόγους: Πρώτον, οι θετικά φορτισμένες λέξεις που εντοπίστηκαν και προστέθηκαν στο λεξικό ήταν πολύ λιγότερες από τις αρνητικές και δεύτερον ήταν πολύ συχνά τα ειρωνικά tweets, στα οποία δεν υπήρχε μεν κάποια θετική λέξη, αλλά είχαν θετικό συνολικό συναίσθημα όπως πχ στα tweets «RT @niemandsrose: Ανατέλλει ο Τέρυ Χρυσός αιώνας της ΝΔ #eklogesneadimokratia», «Ο Κυριάκος Μητσοτάκης είναι Χρυσός άνθρωπος.Ξέρει η Ντόρα.#eklogesneadimokratia» και «Τώρα που ο Κούλης βγήκε πρόεδρος ΝΔ,επιβάλλεται δωράκι σε όλους τους ψηφοφόρους της τηλέφωνο Siemens. #eklogesneadimokratia #eklogesnd».

Κεφάλαιο 5 Επίλογος

5.1 Συμπεράσματα

Στην παρούσα διπλωματική μελετήσαμε το πρόβλημα της ανάλυσης κοινωνικής γνώμης ή Ανάλυσης Συναισθήματος στην πλατφόρμα κοινωνικής δικτύωσης Twitter. Η ανάλυση εστιάστηκε σε ένα συγκεκριμένο θεματικό περιεχόμενο που αφορούσε tweets κοινωνικοπολιτικής φύσεως. Ο αλγόριθμος που υλοποιήσαμε βασίζεται σε ένα λεξικό συναισθήματος το οποίο δημιουργήθηκε επεκτείνοντας ένα ήδη υπάρχον λεξικό με κάποιες από τις πιο συχνά εμφανιζόμενες λέξεις που εντοπίσαμε σε ένα σύνολο δεδομένων που αποτελούταν από περίπου 20.000 tweets. Από όλες τις λέξεις που εντοπίστηκαν δόθηκε ιδιαίτερη σημασία στην επιλογή λέξεων με έντονο συναισθηματικό υπόβαθρο (είτε θετικό είτε αρνητικό), λέξεων που μπορεί στη γενική περίπτωση να μην εκφράζουν κάποιο συναίσθημα, αλλά όταν εντοπίζονται σε tweet πολιτικής φύσης τότε το κάνουν (πχ καρέκλα, υποβρύχια), αλλά και νεολογισμών και λέξεων που χρησιμοποιούνται στην καθομιλουμένη (πχ χρυσαύγουλα, φιλελέδες). Φροντίσαμε επιπλέον να δώσουμε τη μεγαλύτερη δυνατή βαθμολογία (θετική ή αρνητική) σε λέξεις με εξαιρετικά έντονο συναισθηματικό υπόβαθρο (πχ φορολαίλαπα, χαφιές, φιλοπατρία) καθώς θεωρήσαμε οτι σχεδόν πάντα θα καθορίζουν το συνολικό συναίσθημα, με εξαίρεση βεβαίως περιπτώσεις ειρωνείας, σαρκασμού ή σύγκρουσης πολικότητας. Απο εκεί και πέρα προκειμένου να ταξινομηθεί ένα tweet έπρεπε να περάσει το στάδιο της προεπεξεργασίας για αφαίρεση του θορύβου και διατήρηση μόνο της χρήσιμης πληροφορίας, του μετασχηματισμού κάθε λέξης σε 1ο πρόσωπο, χωρίς τον οποίο θα εντοπίζαμε στο λεξικό υποπολλαπλάσιες λέξεις ακόμα και αν αυτές ήταν καταχωρημένες (βλέπε παράγραφο 3.4.2), του εντοπισμού κάποιων βασικών μοτίβων άρνησης και τελικά του υπολογισμού του τελικού σκορ κάθε λέξης και του συνολικού σκορ του tweet.

Δεδομένων των ιδιαιτεροτήτων του συνόλου δεδομένων που μελετήσαμε, δηλαδή της σχεδόν μόνιμης ειρωνείας και των υπονοούμενων δηλώσεων (πχ Τέρης Χρυσός, Siemens, Χριστοφοράκος), μπορούμε να πούμε οτι συνολικά το σύστημά μας παρουσιάζει μια αρκούντως ικανοποιητική λειτουργία. Στις περισσότερες περιπτώσεις το σύστημα συμφωνούσε με την ανθρώπινη κρίση, αλλά υπήρχαν και περιπτώσεις διαφωνίας που οφείλονταν είτε σε προβληματική επισημείωση, είτε στον μη εντοπισμό των λέξεων που καθόριζαν το συναίσθημα της πρότασης ή σε φαινόμενα ειρωνείας, σύγκρουσης πολικότητας και ύπαρξης αντικρουόμενων απόψεων.

5.2 Προτεινόμενες βελτιώσεις

Μερικές από τις πιθανές τροποποιήσεις, αλλά και κάποιες γενικότερες ιδέες για τη βελτίωση της απόδοσης του ταξινομητή θα ήταν οι παρακάτω:

- *Εκπαίδευση επισημειωτών*

  Σε οποιοδήποτε σύστημα Ανάλυσης Συναισθήματος, είτε αυτό βασίζεται σε λεξικό συναισθήματος, είτε σε μηχανική μάθηση, είναι καθοριστικής σημασίας η ακριβής και σχετική με το θέμα επισημείωση, αφού πάνω σε αυτή θα βασίζεται η αξιολόγηση του συστήματος και συνεπώς πρέπει να υπάρχει εμπιστοσύνη στις ταμπέλες που αποδίδουν οι επισημειωτές. Δεν είναι τυχαίο, για παράδειγμα, οτι πολλές υπηρεσίες διαδικτυακών αγορών (crowdsourcing services) όπως το Amazon Mechanical Turk[1], οι οποίες αναζητούν χρήστες για online επισημείωση τεράστιων συνόλων δεδομένων, διαθέτουν εξελιγμένους αλγορίθμους που υπολογίζουν την ικανότητα του κάθε επισημειωτή και επιτρέπουν με αυτόν τον τρόπο τον εντοπισμό και αποκλεισμό αναξιόπιστων επισημειωτών [33]. Σε επίπεδο μιας διπλωματικής εργασίας, επειδή δεν υπάρχει προφανώς αντίστοιχη δυνατότητα, αυτό που θα μπορούσε να γίνει θα ήταν η αξιολόγηση της ικανότητας των επισημειωτών μέσα από ένα μικρό σύνολο δεδομένων, το οποίο θα μπορούσε εν προκειμένω να αποτελείται από 30-40 tweets. Στόχος θα ήταν ο εντοπισμός επισημειωτών που είτε θα ήταν προκατειλημμένοι ως προς μια κλάση, είτε λόγω παρανόησης δεν θα είχαν κατανοήσει το θέμα προς επισημείωση.

- *Επέκταση λεξικού*

  Δεδομένου οτι ένα σύστημα Ανάλυσης Συναισθήματος που βασίζεται σε λεξικό είναι (σχεδόν) τόσο καλό όσο και το λεξικό που χρησιμοποιεί, είναι απαραίτητο το λεξικό να περιέχει όσο το δυνατόν περισσότερους όρους. Στην περίπτωσή μας, οι συνολικά 3072 όροι δεν μπορούν μεν να θεωρηθούν αμελητέα ποσότητα, αλλά αν αναλογιστεί κανείς το πλήθος των όρων της ελληνικής γλώσσας, συμπεριλαμβανομένων των διάφορων νεολογισμών που έχουν δημιουργηθεί και συνεχίζουν να δημιουργούνται καθημερινά στο διαδίκτυο, τότε το λεξικό θα μπορούσε να εμπλουτιστεί με πολλούς επιπλέον όρους. Επειδή μια χειρωνακτική δημιουργία ίσως αποτελεί εμπόδιο λόγω χρονικών περιορισμών, μια ιδέα θα ήταν η χρήση online λεξικών συνωνύμων ή θησαυρών όρων της ελληνικής γλώσσας.

- *Χρήση διγραμμάτων (bigrams)*

  Πέραν του εντοπισμού μεμονωμένων λέξεων, θα μπορούσε να γίνει πειραματισμός με χρήση διγραμμάτων για πιθανή βελτίωση της απόδοσης του ταξινομητή. Τα διγράμματα θα μπορούσαν να αποδειχτούν ιδιαιτέρως χρήσιμα σε περιπτώσεις σύγκρουσης πολικότητας (πχ δυσάρεστο όνειρο, αύξηση της ανεργίας, μείωση των εισφορών, κροκοδείλια δάκρυα κ.α).

[1] https://www.mturk.com/mturk/help?helpPage=overview

- *Ενσωμάτωση αγγλικών όρων και greeklish – Μελέτη hashtags*

  Επειδή η χρήση αγγλικών χαρακτήρων και greeklish είναι εξαιρετικά συχνή στο ελληνικό Twitter, θα μπορούσε αφενός να δημιουργηθεί ένα λεξικό που να περιλαμβάνει τους πιο συχνά χρησιμοποιούμενους αγγλικούς όρους και επιπλέον να αξιοποιηθεί κάποιο σύστημα μετατροπής greeklish σε ελληνικά. Κάτι τέτοιο θα ήταν αρκετά βοηθητικό ιδιαίτερα σε ό,τι έχει να κάνει με τα hashtags, αφού όπως εξηγήσαμε αποτελούν πολύτιμη πηγή πληροφοριών και πολλές φορές συνεισφέρουν στο συναισθηματικό υπόβαθρο, αλλά τις περισσότερες φορές γράφονται στα αγγλικά ή σε greeklish. Μάλιστα θα μπορούσε να μελετηθεί και η πιθανότητα ταξινόμησης βασισμένης μόνο σε hashtags ή η ανάθεση μιας a-priori βαθμολογίας ανάλογα με το εάν εμφανίζονται συγκεκριμένα hashtags.

- *Αναλυτικότερη μελέτη της άρνησης και του υποθετικού λόγου*

  Και τα δύο φαινόμενα αποτελούν σημαντικότατες προκλήσεις στο πεδίο της Ανάλυσης Συναισθήματος. Για τον αποτελεσματικότερο χειρισμό τους, θα μπορούσε να συνδυαστεί με το παρόν σύστημα, ένα σύστημα που θα στρεφόταν πιο πολύ προς την επεξεργασία φυσικής γλώσσας (πχ με χρήση Κρυφών Μαρκοβιανών μοντέλων) μελετώντας με μεγαλύτερη λεπτομέρεια τους διάφορους συντακτικούς κανόνες και προσπαθώντας να εντοπίσει και να χειριστεί σύνθετους τύπους άρνησης και υποθετικού λόγου.

- *Συνδυασμός του παρόντος αλγορίθμου με μέθοδο βασισμένη σε μηχανική μάθηση*

  Θα ήταν πολύ ενδιαφέρουσα η υλοποίηση ενός υβριδικού συστήματος το οποίο θα χρησιμοποιούσε τον αλγόριθμο, βασισμένο σε λεξικό, που αναπτύξαμε, για την επισημείωση του συνόλου δεδομένων εκπαίδευσης ενός ταξινομητή μηχανικής μάθησης. Μια τέτοια προσέγγιση θα εκμεταλλευόταν την ταχύτητα του αλγορίθμου βασισμένου σε λεξικό προκειμένου να εξαλείψει ένα από τα βασικότερα μειονεκτήματα των μεθόδων μηχανικής μάθησης, δηλαδή την ανάγκη ύπαρξης μεγάλων όγκων επισημειωμένων συνόλων δεδομένων.

Βιβλιογραφία

[1] Sarah Schrauwen (July 2010).Machine Learning Approaches to Sentiment Analysis using the Dutch Netlog Corpus. University of Antwerp

[2] Cambria, Erik. "An introduction to concept-level sentiment analysis." *Advances in Soft Computing and Its Applications*. Springer Berlin Heidelberg, 2013. 478-483.

[3] Aisopos, F., Papadakis, G., Tserpes, K., & Varvarigou, T. (2012, June). Content vs. context for sentiment analysis: a comparative analysis over microblogs. In *Proceedings of the 23rd ACM conference on Hypertext and social media* (pp. 187-196). ACM.

[4] Vinodhini, G., & Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: a survey. *International Journal*, *2*(6).

[5] Jia, L., Yu, C., & Meng, W. (2009, November). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1827-1830). ACM.

[6] Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010, July). A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing* (pp. 60-68). Association for Computational Linguistics.

[7] Asmi, A., & Ishaya, T. (2012). Negation identification and calculation in sentiment analysis. In *The Second International Conference on Advances in Information Mining and Management* (pp. 1-7).

[8] Maynard, D., & Greenwood, M. A. (2014, May). Who cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis. In *LREC*(pp. 4238-4243).

[9] Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., ... & Bontcheva, K. (2015). Analysis of named entity recognition and linking for tweets. *Information Processing & Management*, *51*(2), 32-49.

[10] Li, C., Weng, J., He, Q., Yao, Y., Datta, A., Sun, A., & Lee, B. S. (2012, August). Twiner: named entity recognition in targeted twitter stream. In*Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (pp. 721-730). ACM.

[11] Narayanan, R., Liu, B., & Choudhary, A. (2009, August). Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (pp. 180-189). Association for Computational Linguistics.

[12] Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S. F. (2013, October). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In*Proceedings of the 21st ACM international conference on Multimedia* (pp. 223-232). ACM.

[13] Poria, S., Cambria, E., Howard, N., Huang, G. B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content.*Neurocomputing*, *174*, 50-59.

[14] Carbonell, J. G. (1979). *Subjective Understanding: Computer Models of Belief Systems* (No. RR-150). YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE.

[15] Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In*The Semantic Web–ISWC 2012* (pp. 508-524). Springer Berlin Heidelberg.

[16] Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques.

[17] Bishop, C. M. (2006). Pattern Recognition. *Machine Learning*.

[18] Zhu, X., & Ghahramani, Z. (2002). *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU-CALD-02-107, Carnegie Mellon University.

[19] Westling, A. Sentiment Analysis of Microblog Posts from a Crisis Event using Machine Learning.

[20] Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of natural language processing*, *2*, 627-666.

[21] Aue, A., & Gamon, M. (2005, September). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)* (Vol. 1, No. 3.1, pp. 2-1).

[22] Saif, H., He, Y., & Alani, H. (2012). Alleviating data sparsity for twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS. org).

[23] Aisopos, F., Papadakis, G., & Varvarigou, T. (2011, November). Sentiment analysis of social media content using N-Gram graphs. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media* (pp. 9-14). ACM.

[24] Asiaee T, A., Tepper, M., Banerjee, A., & Sapiro, G. (2012, October). If you are happy and you know it... tweet. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1602-1606). ACM.

[25] 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*, *11*(122-129), 1-2.

[26] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market.*Journal of Computational Science*, *2*(1), 1-8.

[27] Brody, S., & Diakopoulos, N. (2011, July). Cooooooooooooooollllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 562-570). Association for Computational Linguistics.

[28] Khan, A. Z., Atique, M., & Thakare, V. M. (2015). Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)*, 89.

[29] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, *61*(12), 2544-2558.

[30] Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. *Best practices in quantitative methods*, 29-49.

Stemler, S. E. (2010). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability 2004. *Practical Assessment, Research & Evaluation, Retrieved February*, *28*.

[31] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

[32] Tsakalidis, A., Papadopoulos, S., & Kompatsiaris, I. (2014). An Ensemble Model for Cross-Domain Polarity Classification on Twitter. In *Web Information Systems Engineering–WISE 2014* (pp. 168-177). Springer International Publishing.

[33] Welinder, P., & Perona, P. (2010). Online crowdsourcing: rating annotators and obtaining cost-effective labels.

[34] Mudinas, A., Zhang, D., & Levene, M. (2012, August). Combining lexicon and learning based approaches for concept-level sentiment analysis. In*Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining* (p. 5). ACM.

Παράρτημα

Α. Συχνότερα εμφανιζόμενες λέξεις

είμαι	0,1568	σκάι	0,0076	αριστερός	0,0047
δεν	0,1502	σχέδιο	0,0076	διαβάζω	0,0047
έχω	0,0777	ακούω	0,0074	ηταν	0,0047
κυβέρνηση	0,0440	αλλαγή	0,0074	μεγάλος	0,0047
λέω	0,0418	λοβέρδος	0,0074	πάνω	0,0047
βουλή	0,0404	οικονομία	0,0074	γνωρίζω	0,0046
συριζα	0,0362	θέτω	0,0073	μπαίνω	0,0046
κάνω	0,0354	μη	0,0072	αποτελώ	0,0045
Ελλάδα	0,0341	απόφαση	0,0071	κοινωνικός	0,0045
νέο	0,0292	λεφτά	0,0071	μείωση	0,0045
γίνομαι	0,0286	σύστημα	0,0071	τελευταίος	0,0045
εκλογή	0,0251	Έλληνας	0,0070	Κυριάκος	0,0044
νδ	0,0244	σχολείο	0,0070	αφήνω	0,0044
όχι	0,0242	ενδιαφέρω	0,0070	καλώ	0,0044
πολιτική	0,0238	κόσμος	0,0070	έλεγχος	0,0042
θέλω	0,0219	ποιος	0,0070	γραφείο	0,0042
υπάρχω	0,0210	έργο	0,0069	δημοκρατικός	0,0042
πολίτης	0,0208	κακός	0,0069	διοίκηση	0,0042
υπουργός	0,0208	φορώ	0,0069	εκπαιδευτικός	0,0042
χρόνος	0,0202	στηρίζω	0,0068	ευκαιρία	0,0042
δημόσιος	0,0201	συνεχίζω	0,0068	ποτέ	0,0042
κόμμα	0,0198	μήνας	0,0067	Αθήνα	0,0042
πασοκ	0,0197	ερτ	0,0066	άποψη	0,0042
γιατί	0,0196	άνθρωπος	0,0065	αφορώ	0,0042
μνημόνιο	0,0191	κοινός	0,0065	βγαίνω	0,0042
ψηφίζω	0,0191	μειώνω	0,0065	εφαρμόζω	0,0042
δίνω	0,0183	ξεκινώ	0,0065	μικρός	0,0042
πρόεδρος	0,0180	προεκλογικός	0,0065	προκαλώ	0,0042
βλέπω	0,0172	κυβερνηση	0,0064	συνάντηση	0,0042
πολύς	0,0170	ξέρω	0,0064	υπόσχομαι	0,0042
επιτροπή	0,0167	εφημερίδα	0,0063	ύφεση	0,0042
λαός	0,0160	σαμαράς	0,0063	ανάγκη	0,0041
ομιλία	0,0159	ασφαλιστικός	0,0062	ανακοίνωση	0,0041
χώρα	0,0157	εκπομπή	0,0062	δρόμος	0,0041
πρώτος	0,0156	οδηγώ	0,0062	εργάζομαι	0,0041
χρέος	0,0153	περνώ	0,0062	πραγματικότητα	0,0041
καλός	0,0145	πιστεύω	0,0062	ρωτώ	0,0041
βουλευτής	0,0145	μητσοτάκης	0,0061	συνεργασία	0,0041
μέγας	0,0143	φέρνω	0,0061	βημα	0,0040
δήλωση	0,0142	απαντώ	0,0061	γραμμή	0,0040
καλημέρα	0,0141	ευρωπαϊκός	0,0061	διεθνής	0,0040
συνέντευξη	0,0140	ζωή	0,0061	δυστυχώς	0,0040
άρθρο	0,0138	κρίση	0,0061	ενφια	0,0040
ελληνικός	0,0136	αυξάνω	0,0060	κοινωνία	0,0040

ζητώ	0,0134	γράφω	0,0060	μάχη	0,0040
χωρίς	0,0132	διαδικασία	0,0060	σημαίνω	0,0040
ευχαριστώ	0,0131	ιδιωτικός	0,0060	στιγμή	0,0040
συμφωνία	0,0126	τέλος	0,0060	τομέας	0,0040
μην	0,0125	χρήμα	0,0060	Νίκος	0,0039
μιλώ	0,0125	δραχμή	0,0058	έως	0,0039
πρωθυπουργός	0,0123	λοβέρδου	0,0058	δήμος	0,0039
τσίπρα	0,0118	περιβάλλω	0,0057	ειδικός	0,0039
θέμα	0,0116	φίλος	0,0057	εκδήλωση	0,0039
τωρα	0,0115	αλήθεια	0,0056	εκλογικός	0,0039
πηγαίνω	0,0112	διαπραγμάτευση	0,0056	επιλογή	0,0039
τσίπρας	0,0112	οχι	0,0056	επιτέλους	0,0039
ημέρα	0,0111	προτείνω	0,0056	καταστροφή	0,0039
οικονομικά	0,0110	σωστός	0,0056	μένω	0,0039
φόρος	0,0110	σύνταξη	0,0056	νομοσχέδιο	0,0039
ειναι	0,0109	παιδί	0,0055	παπανδρέου	0,0039
νόμος	0,0108	έπομαι	0,0054	ταμείο	0,0039
πρόταση	0,0108	βήμα	0,0054	υποψήφιος	0,0039
εθνικός	0,0108	παιδεία	0,0054	χώρος	0,0039
ευρώ	0,0108	αποφασίζω	0,0053	έσοδο	0,0038
πολιτικός	0,0107	υπάλληλος	0,0053	ανοίγω	0,0038
χρειάζομαι	0,0104	απάντηση	0,0052	ομάδα	0,0038
δημοκρατία	0,0103	περιμένω	0,0052	παραιτούμαι	0,0038
Ευρώπη	0,0100	σημερινός	0,0052	προηγούμαι	0,0038
παίρνω	0,0100	τρόπος	0,0052	συμμετέχω	0,0038
ανάπτυξη	0,0099	άλλος	0,0051	Μάνος	0,0037
τράπεζα	0,0099	ενέργεια	0,0051	ακίνητος	0,0037
χάνω	0,0098	κεντρικός	0,0051	δημιουργώ	0,0037
αλλάζω	0,0098	κερδίζω	0,0051	εξηγώ	0,0037
κράτος	0,0097	μεταρρύθμιση	0,0051	κόστος	0,0037
Βενιζέλος	0,0096	πολιτικά	0,0051	μεταρρυθμίζω	0,0037
πρόβλημα	0,0096	τρίτος	0,0051	μισθός	0,0037
πρόγραμμα	0,0096	φπα	0,0051	ξεχνώ	0,0037
μπορώ	0,0094	αντιπολίτευση	0,0051	παρέμβαση	0,0037
πληρώνω	0,0094	καταργώ	0,0051	στέλεχος	0,0037
βρίσκω	0,0090	μμε	0,0051	υπηρεσία	0,0037
οποίος	0,0089	σκαι	0,0051	φτάνω	0,0037
Ανδρέας	0,0089	σκληρός	0,0051	έλλειμμα	0,0036
λόγος	0,0087	τύπος	0,0051	βαρουφάκη	0,0036
έρχομαι	0,0086	ψήφος	0,0051	δείχνω	0,0036
υπουργείο	0,0086	ανοιχτός	0,0050	δικαίωμα	0,0036
δημοψήφισμα	0,0085	καθημερινός	0,0050	δικός	0,0036
δύναμη	0,0085	σύνταγμα	0,0050	είδηση	0,0036
λίγος	0,0085	δελτίο	0,0049	εκπαίδευση	0,0036
συζήτηση	0,0083	δύο	0,0049	εντολή	0,0036
συζητώ	0,0083	κατάργηση	0,0049	θυμούμαι	0,0036
ευθύνη	0,0082	καταθέτω	0,0049	σοβαρός	0,0036
εξεταστικός	0,0081	κλείνω	0,0049	σύμφωνο	0,0036
ώρα	0,0081	κυβερνώ	0,0049	ψέμα	0,0036
αρχή	0,0080	οικονομικός	0,0049	αποτέλεσμα	0,0035

δηλώνω	0,0080	συνέδριο	0,0049	ενημερώνω	0,0035
λάθος	0,0080	φαίνομαι	0,0049	ζήτημα	0,0035
νέος	0,0080	δεύτερος	0,0048	κυβερνητική	0,0035
ερώτηση	0,0078	τηλεόραση	0,0048	υπογράφω	0,0035
συμφωνώ	0,0078	ψέμματα	0,0048	αγορά	0,0034
σχετικός	0,0077	Κυριακή	0,0047	ανεργία	0,0034

Πίνακας 22: Συχνότερα εμφανιζόμενες λέξεις στο σύνολο δεδομένων που συλλέξαμε

Β. Συχνότερα εμφανιζόμενα bigrams

δεν είναι	λένε ότι	βουλή δεν
μεσος ψηφοφορος	μέσος ψηφοφόρος	βουλευτές συριζα
ψηφοφορος νδ	νέα δημοκρατια	δεν κάνουν
δεν έχει	νδ πασοκ	δεν ξέρουν
νέα δημοκρατία	προέδρου συμφωνίας	δεν πιστεύω
δεν υπάρχει	δημοσίου χρέους	δεν υπήρχε
εκλογές νδ	δημόσια διοίκηση	δεν υπαρχουν
γιατί δεν	εκλογές δεν	δημόσιας διοίκησης
δεν έχουν	εκλογικά τμήματα	είμαι εκπομπή
νέα ελλάδα	πιστεύω ότι	είναι κυβέρνηση
δεν ειναι	πολιτικό σύστημα	εθνικές εκλογές
πρώτη φορά	συνέντευξη τύπου	ελπίζω μην
μπλα μπλα	συστημα τωρα	επιτροπή περιβάλλοντος
π καμμένος	σχέδιο νόμου	ευρωπαϊκό όραμα
συμφωνίας νέα	υπουργείου οικονομικών	κυβέρνηση έχει
πόθεν έσχες	υπουργός άμυνας	λάκης λαζόπουλος
ότι είναι	χώρα δεν	λίγες μέρες
ομιλία βουλή	έχει κάνει	μέσω χρήστη
ρε παιδιά	έχουν ψηφίσει	μητσοτάκης γεωργιάδης
ξέχασε δηλώσει	ή όχι	παγκόσμια ημέρα
κυβέρνηση δεν	δεν μπορείς	παραγωγής εμπορίου
εκλογών νδ	δημοκρατια πατωματα	πλεονέκτημα αριστεράς
νέας δημοκρατίας	εθνικής άμυνας	προτείνει ψηφοφόρους
νδ είναι	ελλάδα ανδρέα	πρωτη φορά
αποτελέσματα εκλογών	ελληνικό λαό	τέχνη δεν
συριζα δεν	επιτροπής αλήθειας	τζιτζικώστας γεωργιάδης
εκλογικό κέντρο	καλή επιτυχία	υπουργείο παιδείας
ερώτηση βουλή	λαζόπουλος είναι	φήμες λένε
ανδρέας λοβέρδος	νομίζει ότι	φυλακών κορυδαλλού
μεγάλη συμμετοχή	ρε φιλε	χώρα έχει
δεν ξέρω	σπίτι χέρια	χώρα είναι
εκλογικά κέντρα	σπιτιού σπιτιού	αλήθειας δημοσίου
ρε λάκη	ακίνητα δεν	απάντηση υφυπουργού
δεν είμαι	δεν γίνεται	γίνει υπουργός
νεα δημοκρατια	είναι ρε	γιατί όχι
παιδείας θρησκευμάτων	είναι υπουργός	δεν έπρεπε
ρε φίλε	εκλογή προέδρου	δεν βγαίνει
δεν έχουμε	ευχαριστώ καλημέρα	δεν θυμάμαι
συνέντευξη εφημερίδα	ζωή κωνσταντοπούλου	δεν κανει
κυριάκος μητσοτάκης	κλείσουν κάλπες	δεν μπορώ
χρόνια πολλά	μαζική συμμετοχή	δεν πάει
δελτίο ειδήσεων	μεϊμαράκης μητσοτάκης	δημοκρατια συμβαινει
νεας δημοκρατιας	ολομέλεια βουλής	δημόσιο διάλογο
ομιλία επιτροπή	πήγαν ψηφίσουν	είναι ή
πασοκ νδ	πασοκ είναι	είναι δυνατόν
κυβέρνηση είναι	πρόεδρος νδ	είναι νέα
κι όχι	ρε παιδια	ειναι ενα

κοινή δήλωση	συμβαινει νέα	εκλογές είναι
λίγα λεπτά	τάκη μίχα	ενημερωσης κορυφη
πασοκ δημοκρατικη	ψηφοφόροι νδ	επιτροπή μορφωτικών
σύμφωνο συμβίωσης	ανελ δεν	εσωκομματικές εκλογές
δήλωση αφορμή	αξιοπιστης ενημερωσης	εφημερίδα καθημερινή
δημοκρατικη παραταξη	αποτελεσμα τωρα	κλιματικής αλλαγής
κεντρικό δελτίο	δεν έκανε	κύριε υπουργέ
πρόεδρο νδ	δεν αντέχει	κύριος δεν
συριζα ανελ	εκλογες νδ	λέει κύριος
έπεσε σύστημα	εξεταστική επιτροπή	λίγη ώρα
δεν γίνει	κάλπες χωρίς	μητσοτάκης τζιτζικώστας
δεν θέλει	καλημέρα είμαι	ξέχασα δηλώσω
δεν ξέρει	κορυφη αξιοπιστης	πει ότι
επιτροπή αλήθειας	νεκρή μάνα	περιβάλλοντος ενέργειας
ηθικό πλεονέκτημα	ποιος είναι	πού είσαι
καμμένος δεν	ρε λακη	προέδρου νδ
καμμένος είναι	συμφωνία νέα	ρε λαζόπουλε
πάνος καμμένος	τσίπρας δεν	σπιτιου σπιτιου
ψηφοφόρος νδ	χωρίς ουρές	τρίτο μνημόνιο
δεν είμαστε	έχει πάει	υπουργό άμυνας
δεύτερο γύρο	γνώμες βημα	φάνε φόλα
είπε ότι	δελτίο τύπου	φπα νησιά
εκτίμηση αποτελεσμα	δεν μπορει	χρέος δεν
πρωτο εκτίμηση	δεν ξεχνάμε	έχει ανάγκη
πρώτα αποτελέσματα	δεν πειράζει	έχει πει
τωρα πρωτο	είναι θέμα	αλ τσαντίρι
όχι δεν	είναι καλό	αλέξης τσίπρας
δεν βλέπω	είναι νδ	αλλα δεν
δεν υπάρχουν	είναι πρωθυπουργός	αμήχανη στιγμή
δεν υπαρχει	εκλογική διαδικασία	γιατί εκλεγεί
δηλώσει ακίνητα	επίκαιρη ερώτηση	γιαυτό νεας
είναι αλήθεια	θρησκευμάτων ανδρέας	γράφει γιατί
ελλάδα είναι	καμμένος έχει	δήλωση ανδρέα
ποτέ δεν	κλειστές τράπεζες	δεν αφήνει
άρθρο εφημερίδα	κόκκινα δάνεια	δεν ελέγχονται
δημόσιο χρέος	λίστα λαγκάρντ	δεν θέλω
εκλεισε συστημα	λεφτά υπάρχουν	δεν καταλαβαίνω
ελλάδα δεν	μέσος όρος	δεν υπάρξει
νδ έχει	νέος πρόεδρος	δημόσιοι υπάλληλοι
ντορα νεα	νδ ειναι	διαρθρωτικές αλλαγές
πατωματα ντορα	νομίζω ότι	δις ευρώ
πολιτικές εξελίξεις	πρωτιά μεϊμαράκη	είναι δεν
ρε μαλακα	σπιτι σπιτι	είναι πρώτη
τωρα εκλεισε	συριζα είναι	ενέργειας κλιματικής
υπουργός παιδείας	τεχνικό πρόβλημα	επιτροπή οικονομικών
χέρια τραπεζίτη	τσίπρα είναι	επιτροπή παραγωγής
δήλωση σχετικά	υπουργείο οικονομικών	ευθύνης υπουργών
δεν θέλουν	υπουργού παιδείας	κάνει ντου
δεν κάνει	άδωνις γεωργιάδης	καλή εβδομάδα
δημοκρατία δεν	αλήθεια είναι	καμμένος ότι

είμαι τηλεόραση	αποθήκη ψυχών	κι άλλα
ελληνικού λαού	αρχηγό νδ	κλείνουν κάλπες
ιδιωτικό τομέα	αυτο ειναι	κο κο

Πίνακας 23: Συχνότερα εμφανιζόμενα διγράμματα (bigrams) στο σύνολο δεδομένων που συλλέξαμε

Γ. Λέξεις που προσθέσαμε στο λεξικό

άγνοια	δοσίλογος	προπηλακίζω	αισχρότητα	κτηνοβασία	φακελώνω
άδικος	δραματικός	προπηλακισμός	αλαζονεία	κωλοβάρεμα	φελλός
άθλιος	δυσάρεστος	προσβάλλω	αλησμόνητος	κωλομπαράς	φιλελέ
άξιος	δυσκολεύω	προσβλητικός	αλισβερίσι	κωλοτούμπας	φιλελέδες
άτολμος	εγκληματικός	προχειρότητα	αλυσοπρίονο	κόκκινος	φοροκαταιγίδα
άψογος	εκβιάζω	πρόβλημα	αλώβητος	κόλαση	φορολαίλαπα
έγκλημα	εκβιασμός	πρόοδος	αμερόληπτος	κότα	φρενοβλαβής
έλλειμμα	εκπληκτικός	πτωχεύω	αμπελαλε	κώλος	φρικτός
έλλειψη	εκτιμώ	πτώχευση	αμπελοφιλοσοφία	κώμα	φόλα
ένδοξος	εκφοβισμός	ρήξη	ανάλατος	λαθρομετανάστης	χέζω
ένοχος	ελλιπής	ρεαλισμός	ανένταχτος	λαμόγια	χαράτσι
ήττα	ελπιδοφόρος	ρουσφέτι	ανέξοδος	λασπολόγος	χαρακίρι
αήθης	εμμονή	σάπιος	ανακριβής	λειψανδρία	χαραμοφάης
αίσχος	εμπαιγμός	σαμαροβενιζέλοι	αναξιοπρεπής	ληστοσυμμορίτες	χαφιές
αβέβαιος	εμπιστεύομαι	σανό	ανατριχιαστικός	λοβοτομή	χλευάζω
αγαθός	ενότητα	σημαντικός	ανενημέρωτος	λοβοτομημένος	χοντροκομμένος
αδίκημα	εξέγερση	σκάνδαλο	ανεπάγγελτος	λογικό	χοντρός
αδιάφορος	εξαθλίωση	σκανδαλοθηρικός	ανεπαρκής	λούγκρα	χουνέρι
αδιέξοδο	εξαιρετικός	σκανδαλώδης	ανθυγιεινός	λυκάνθρωπος	χουντικός
αδιέξοδος	εξαπάτηση	σκατόψυχος	ανιστόρητος	λυράτη	χούντα
αδικία	εξαπατώ	σκληρός	ανούσιος	μάγκας	χούφταλο
αδικαιολόγητος	εξευτελίζω	σκοτώνω	αντικοινωνικός	μάπα	χρεοκοπημένος
αδικώ	εξευτελισμός	σκουπίδι	αντρίκια	μέγας	χρεωκοπώ
αδράνεια	εξοργιστικός	σοβαρός	απαθής	μαγκιά	χρυσαύγουλα
αδρανώ	εξορθολογισμός	στήριξη	απατεωνιά	μαλακία	χρυσαύγουλο
αδυναμία	εξυγίανση	σταθερότητα	αποχή	μανα	χυδαίος
αδυνατώ	εξυγιαίνω	στεναχωρημένος	απώλεια	μοσχαροκεφαλή	χυδαιότητα
αθλιότητα	εξόντωση	στηρίζω	αργοπεθαίνω	μπαγλαμάς	ψεκάζω
αισιοδοξία	επίθεση	συγχαρητήρια	αρλούμπα	μπαρουφάκης	ψεκασμένος
αισιόδοξος	επιβαρύνω	συκοφαντία	αρουραίος	μπαρούφα	ψεύτικος
αισχρός	επικίνδυνος	συκοφαντικός	αρρενωπός	μπινελίκι	ψηφαλάκια
ακατανόητος	επιτίθεμαι	συκοφαντώ	αρχίδια	μπιχτή	ψοφάω
ακραίος	επιτυχία	συνέπεια	αρχιγλύφτης	μπορντέλο	ψυχοπάθεια
ακροδεξιός	ευνοϊκός	συναίνεση	αρχιδάκι	μπουμπούκος	ψυχοφάρμακο
αλήθεια	ευτελίζω	συναινώ	ασάφια	μπουχέσας	ψόφος
αλητεία	ευχαριστώ	συνεννόηση	ασθενώ	μπούρδα	ώρσε
αλληλεγγύη	εχθρός	συνεπής	ασχολίαστος	νεκρός	ομαλά
αλληλοσεβασμός	ζημιά	σωστός	αυλικός	νεοναζί	κάφρος
αμήχανος	ηλίθιος	σωτηρία	αυνανίζομαι	νεοφιλελέ	νεοφιλελευθερισμός
αμφιβολία	θάνατος	ταπείνωση	αυνανισμός	νεοφιλελεύθερος	άδικος
ανέντιμος	θίγω	ταπεινωτικό	αυτογνωσία	νοθεία	άμυαλος
ανήθικος	θαλασσοδάνεια	τερατούργημα	αχειραγώγητος	νταηλίκια	ένφια
ανίκανος	θράσος	τερατώδης	βαγγέλα	ντου	αγανακτισμένος
αναγέννηση	θύμα	τζιχαντιστές	βαγγέλας	ντόμπρος	ακροαριστερός
ανακρίβεια	ικανοποιημένος	τιμώ	βασανιστήριο	ξέπλυμα	αμάθεια
ανακριβές	ικανοποιώ	τολμηρός	βασιλόφρων	ξεβράκωμα	ανάκαμψη
αναποτελεσματικός	κέρδος	τούβλο	βαψομαλλιάς	ξεδιάντροπος	ανάπτυξη
αναρχοαριστερός	κίνδυνος	τραγικό	βλάχικος	ξεκάρφωμα	ανθέλληνες
ανεγκέφαλος	καθαρός	τραγωδία	βλήμα	ξεκαρδιστικός	ανθρωπισμός
ανεξέλεγκτος	κακώς	τραμπουκισμός	βλαβερός	ξεκατίνιασμα	αξιομνημόνευτος
ανεπάρκεια	καλημέρα	τραμπούκοι	βλαχομπαρόκ	ξεκατινιάζω	αξιοποιώ

ανεπρόκοπος	κατάθλιψη	τραμπούκος	βρισίδι	ξενοφοβικός	αποχαυνωμένος
ανεργία	κατάληψη	τρομοκράτης	βρισιά	ξετσίπωτος	αποχαύνωση
ανευθυνότητα	κατάντια	τρομοκρατία	βρομόστομα	ξεφτυλίζομαι	αριστερίλα
ανησυχία	κατάρρευση	τρομοκρατικός	βρυκόλακας	ξύλο	αριστεριστές
ανησυχώ	κατάσχω	τρομοκρατώ	βρωμοδουλειά	ομοφοβία	ασφαλιστικό
ανθέλληνας	κατήφεια	τροχοπέδη	βρωμόχερα	ομοφοβικός	αυτοκτονία
ανθρωπιστής	καταγγέλω	τσεκουριά	βυσματίας	ομοφυλοφιλία	βίαιος
ανιδιοτέλεια	καταγγελία	τυχοδιωκτισμός	βόδι	ομοφυλόφιλος	βανδαλίζω
ανικανότητα	κατακραυγή	υγεία	βόθρος	πανηλίθιος	βιασμός
ανοησία	καταπληκτικός	υγιής	βόλεμα	παπάντζα	γέλια
αντιδημοκρατικός	καταρρέω	υπέροχος	βύσμα	παπάρα	γελοιοδέστατος
αντιεξουσιαστικός	καταστρέφω	υπερφορολόγηση	γίδι	παπαριά	διαστρεβλώνω
αντιεπιστημονικός	καταστροφή	υποκρισία	γελοιοποίηση	παράνοια	διεφθαρμένος
αντικειμενικός	καταστροφικός	υπονομεύω	γελοιοποιώ	παράνομο	εγκληματικότητα
αντισυνταγματικότητα	καταψηφίζω	υπονόμευση	γελοιότητα	παρακράτος	εθνοκάθαρση
αντιφατικός	κατηγορώ	υποσιτισμός	γελωτοποιός	παρακρατικός	εισφορά
ανόητος	κινδυνεύω	ύφεση	γελώ	παραμυθάς	εισφορές
αξίζω	κλέβω	φασίστας	γκέι	παραμυθιάζω	ελευθερία
αξιοκρατία	κοροϊδία	φασισμός	γκέμπελς	παραμύθι	εμφύλιος
αξιοκρατικός	κοροϊδεύω	φασιστικός	γκεστάπο	παρανομία	ενφια
αξιοπιστία	κορυφαίος	φεουδαρχικός	γκεσταμπίτης	παρατράγουδο	επένδυση
αξιοποίηση	κουκουλοφόρος	φιάσκο	γλείψιμο	πασοκολόγος	επεισόδια
αξιοπρέπεια	κράζω	φιλελέρες	γλοιώδης	παταγώδης	επενδύω
αξιοπρεπής	κρατικοδίαιτος	φιμώνω	γλύφω	πατρωνία	ζαβός
αξιόπιστος	κωλοτούμπα	φοβούμαι	γλύψιμο	πατσάδες	ζώα
απάτη	λάθος	φοροδιαφεύγω	γυμνοσάλιαγκας	πατσοκοιλιά	ημιμάθεια
απαξίωση	λαθρέμπορος	φοροδιαφυγή	δάκρυ	πελατειακός	ηττημένος
απαράδεκτος	λαμογιές	φορολογώ	δεκανίκι	περηφάνεια	ιδανικός
απατεώνας	λαμόγιο	φοροφυγάδας	διαλλακτικός	πουρό	καθεστωτικός
απειλώ	λαϊκίζω	υφεσιακός	διεστραμμένα	πουτάνα	καινοτομία
απογοητευτικός	λαϊκισμός	φτώχεια	δικαίως	πουτανάκι	κακοδιαχείρηση
αποδοκιμάζω	λαϊκιστής	φόρος	δολοφονώ	πούλο	κακομαθημένος
απολύω	λαϊκιστικός	χάνω	δουλευταράς	πούτσα	κακοπληρωτής
αποτελεσματικός	λιποτακτώ	χάος	δούλεμα	πρήζω	κακοποιός
αποτυχία	λιτότητα	χαζός	δράμα	προδοσία	κολοτούμπα
αποτυχαίνω	λογοκρίνω	χαμόγελο	δραχμονικολάου	προικοθήρας	κομματικός
αποτυχημένος	μεθοδικός	χαροπαλεύω	δύσκολος	προικοθύρας	κοροιδία
απροκάλυπτος	μετριοπαθής	χαρούμενος	ειρωνικός	προοπτική	κουμουνιστές
απόγνωση	μιζέρια	χρήσιμος	ελαφρόμυαλος	προπαγανδίζω	κρίμα
απόλυση	μωρός	χρεοκοπία	ελεεινός	προπαγανδιστής	λαθρομετανάστες
απόρριψη	ναζί	χρεοκοπώ	ελπίδα	προπαγανδιστικός	λιποτάκτης
αρετή	ναζιστικός	χρυσαυγή	εντεταλμένος	πρόβατο	μερκελιστής
αριστούχος	νικώ	χρυσαυγίτης	εξαγριώνω	πτώση	μνημονιακοί
αρνητικός	ξεπουλώ	ψέμα	εξευτελιστικός	ρεμάλι	μνημονιακός
ασθένεια	ξεπούλημα	ψέματα	εξοπλιστικός	ρεύομαι	μολότωφ
αστείος	ξεφτίλα	ψέμματα	εξοργίζω	ρουσφετάκι	μορφωμένος
ασχετοσύνη	ξεχασιάρης	ψευδής	εξωγήινος	ρουφιανιά	μοσχάρια
αυθαίρετο	οδυνηρός	ψευτες	επανεκκίνηση	ρουφιανιλίκι	μούγκα
αυτοπεποίθηση	οδύνη	ψεύδομαι	επιδείνωση	ρόμπα	μπάτσοι
αφελής	οργή	ψεύδος	επιθετικός	σίχαμα	νίκη
αφελληνισμός	οχι	ψεύτης	επιπλήττω	σαμποτάζ	νικάω
βάλλω	παιδεραστής	όμορφος	επιτήδειος	σαπάκι	ξέφραγος
βάρβαρος	παλινωδία	ύποπτος	ερασιτέχνης	σαπάκια	οικουμενική
βία	παράλογος	ώριμος	ευεργέτης	σαπίλα	οικουμενικός

βεβηλώνω	παράνομος	αναρχικός	ευλαβικός	σκίζω	ορθολογικός
βελτίωση	παράσιτο	αριστεριστής	ημιμαθής	σκυλάδικο	παλαιοκομματικός
βιώσιμος	παραβιάζω	ρατσισμός	ηρωισμός	σούργελα	παραίτηση
βλάπτω	παραλήρημα	ρατσιστής	θαλασσοδάνειο	σούργελο	παρακμή
βλακώδης	παρανομώ	αμόρφωτος	θρηνώ	σπάω	παραποίηση
βρίζω	παραπλάνηση	βλαχοδήμαρχος	ισοπεδώνω	σπαζαρχίδας	πετυχημένος
γελοίος	παραπλανητικός	γίδια	κάγκουρας	σπαμάρω	πληστηριασμός
γενναίος	παραπλανώ	άξεστος	κάζο	στρατόκαυλος	πορτιέρης
γενοκτονία	παραπληροφορώ	αγράμματος	καημένος	στρουμφάκια	πρεζάκι
γερμανοτσολιάδες	παρωδία	αγενής	καθήκι	συριζαίος	σκάνδαλα
γκάφα	πασόκοι	αμορφωσιά	καθίκια	συριζανέλ	σκουπίδια
γραφειοκρατία	πασόκος	βλάκας	κακοποίηση	συριζοκαμμένος	συμπλέω
δήθεν	πατριώτης	ευγενής	κακοποιώ	σχιζοφρένεια	συμπλοκή
δίκαιο	πεθαίνω	ευγενικός	κακουργηματικός	σόιμπλε	συμφορά
δίκαιος	πειστικός	μπαχαλάκηδες	κακούργημα	ταλαιπωρία	συριζαίοι
δίκιο	περικοπή	ψευτόμαγκας	κακόγουστος	τοξικός	συριζανελ
δηλητηριάζω	πισωγύρισμα	θλίβομαι	κακόμοιρος	τρέμω	συριζοτρολ
δημαγωγία	πισώπλατος	άνεργος	καρότο	τραβέλι	συριζόπληκτος
δημαγωγός	πλαστός	άφθαρτος	καταδικασμένοι	τραπεζίτης	σύμπλευση
διάλυση	πλειστηριασμός	άχαρος	καταληψίας	τρολ	ταπεινωτικός
διαλύω	πλεονέκτημα	άχρηστος	καταστροφολογία	τσαμπουκάς	τρόμος
διαπλοκή	ποινικοποίηση	έρπομαι	κατρακύλα	τσιπροκαμμένοι	τσίρκο
διαφάνεια	ποινικός	αίσχιστος	κατσίκα	τυχάρπαστος	υποβαθμίζω
διαφθείρω	πολύτιμος	αβαβά	κλάμα	υβριστής	υπόδικος
διαφθορά	πονηρός	αβοήθητος	κλαψιάρης	υπαλληλάκος	φορολογία
διαφωνώ	πουθενάς	αγκάθι	κλόουν	υπερόπτης	φρικιαστικός
διαύγεια	πραξικόπημα	αγνώμων	κολαστήριο	υποβάθμιση	ιδιωτικοποιώ
δικτατορία	προβληματικός	αδίστακτος	κομματόσκυλα	υποβρύχιο	ιδιωτικοποίηση
διχάζω	προβληματισμός	αδαής	κομπλεξικός	υποκλοπή	επονείδιστος
διχασμός	προδότης	αδιάβαστος	κομπλεξισμός	υπομονή	απεχθής
διχαστικός	προκατάληψη	αείμνηστος	κοράκια	υποτιμητικός	ευφυής
δολοφονία	προκλητικός	αερολογία	κράξιμο	υστερικός	ευφυΐα
δολοφόνος	προπαγάνδα	αερολόγος	κτηνοβάτης	φίμωση	νικάω

Πίνακας 24: Λέξεις που προστέθηκαν στο λεξικό συναισθήματος

Δ. Σύνολο δεδομένων

Tweets	*Cl*	*1*	*2*	*3*	*4*	*5*
Ανατροπή δεν είναι να μαζεύονται ο 2ος, ο 3ος και ο 4ος για να φαν τον πρώτο #eklogesneadimokratia	0	-	0	-	-	-
RT @Andrey_Vyshinsk: Ο ΠΑΝΟΣ ΑΚΑ Ψεκασμένος βρήκε συνωμοσία Σημιτικών Δημαριτων Ποταμιού Κ Ακροδεξιών Πια Αγκάθα ρε γατάκια #eklogesneadi…	-	-	-	-	-	-
RT @niemandsrose: Ανατέλλει ο Τέρυ Χρυσός αιώνας της ΝΔ #eklogesneadimokratia	0	+	0	+	+	+
Οι ΝΔτες ξεθάρρεψαν άρχισαν να εύχονται απολύσεις και νομίζουν ότι θα κερδίσουν εκλογές!! Χαχαχαχαχααχα #eklogesneadimokratia #eklogesnd	0	+	0	+	+	+
Κοιτάξτε μια κωλόφατσα Και κοροιδεύει τη φάτσα του Κυριάκου λολ https://t.co/ByAInNboYo	0	-	-	-	-	-
@neademokratia @G_Plakiotakis: Συντεταγμένα θα βαδίσουμε όλοι μαζί https://t.co/4FBA6bUSrv … #Ysterografa #eklogesneadimokratia	0	+	+	+	+	+
RT @GMitakides: «Μέριασε, βράχε, να διαβώ!» (Βαλαωρίτης) #eklogesneadimokratia https://t.co/d7S7Uhdzkj	0	0	0	0	+	+
RT @decadenza35: Αυτοί που είναι βέβαιοι για τον πατέρα Τσίπρα και τη χούντα,δεν έχουν ακούσει κάτι για τον Κυριάκο και τη siemens.. #eklog…	-	+	+	+	0	+
Αφού είναι βέβαιο οτι δε θα απαλλαγούμε ποτέ από αυτό το σόι δν αλλάζουμε όνομα στη χώρα? Μητσοτακισταν καλό ακούγεται #eklogesneadimokratia	+	-	0	-	0	-
με την βαριά φωνή την μπάσα...έχει σκάσει σε κωλόμπαρο της συγγρού και πίνει τζονι μάυρο και κοιτάει μια χοντρη https://t.co/S2Wa4aSiIS	-	-	-	-	-	-
shοκολατάκια για την πίκρα της ντόρας #eklogesneadimokratia https://t.co/X8qlupsND6	-	-	-	-	-	-
Τα αποτελέσματα τα χέζω και τα γράφω στα αρχίδια μου..ο Άδωνις είναι ο πρόεδρος της καρδιάς μου!!❤ #eklogesneadimokratia	-	-	-	-	-	-
Σαμαράς: ενωμένοι να κερδίσουμε οριστικά τον λαϊκισμό https://t.co/VLJmZVAqaO via @AntennaNews @samaras_antonis #eklogesneadimokratia	0	+	-	+	+	0
Επικράτηση Μητσοτάκη στη γαλάζια μάχη για την Προεδρία #eklogesneadimokratia @kathimerini_gr https://t.co/0XwGClMgYG	0	+	+	+	+	+
RT @_bourdas: Κυριάκο αγόρι μου, 20 χρόνια ζωής (τουλάχιστον) κέρδισα απόψε #eklogesneadimokratia https://t.co/AHK3tQTcom	+	+	+	+	+	+
RT @ThodStr: Οικογενειακοί πανηγυρισμοί... #eklogesneadimokratia #eklogesnd #neadimokratia #ndekloges #nd https://t.co/7pYUzsLJIa	+	+	+	+	+	+
#eklogesneadimokratia Από τις πρώτες εξαγγελίες του Κυριάκου,είναι η ιδιωτικοποίηση των δημόσιων τουαλέτων...	-	0	+	0	0	+
Βρήκαμε νέο "Ανδρέα", νέο Μητσοτάκη, το Μακεδονία tv παίζει τόλμη και γοητεία, ε δε γίνεται αναβιώνουν τα 80΄s #eklogesneadimokratia	+	0	+	0	+	0
RT @fwnhlogikhs: Μπορεί να ψηφίσαμε Κούλη, αλλά εσάς προσμονούμε Μεγαλειότατε! #eklogesneadimokratia https://t.co/y0GxPHBL0K	0	+	+	+	+	+
#eklogesneadimokratia Μωρή πουτάνα αμυγδαλιά π' ανοίγεις το Γενάρη δεν καρτερείς την Άνοιξη ν' ανοίξουμ' όλοι αντάμα https://t.co/ooAK1sIpL0	-	-	-	-	-	-

RT @parallilos: Πόσο Όσκαρ μπορείς να γίνεις ρε Κυριάκο! Πρώτο ευχαριστήριο στη Μαρέβα που δεν σε χώρισε. Τον πούλο οι ψηφοφόροι. UFO! #ekl…	-	-	-	-	-	-
Συγχαρητήρια Τσίπρα στον Κυριάκο Μητσοτάκη https://t.co/7uCjn6YJTn #eklogesneadimokratia	+	+	+	+	+	+
Ο νεοφιλελευθερισμός είχε τα πρώτα θύματα. Είχε δίκιο ο Ευριπίδης #eklogesneadimokratia #eklogesnd https://t.co/WMArWvxG0R	-	-	-	-	-	-
RT @Saganium: Χειροκροτάνε μέχρι και το φελλό τον παπαμιμίκο τα ζωάδια. Μιλάμε για ανυπολόγιστη εγκεφαλική βλάβη. #eklogesnd #eklogesneadim…	-	-	-	-	-	-
RT @yioults1: Δηλαδή στην εκεί στην ΝΔ τα βάλανε κάτω και είπαν ότι θα πάνε μπροστά γυρνώντας πίσω #eklogesneadimokratia	0	0	-	0	+	+
Φαντάσου που καταντήσαμε... https://t.co/0A420xhQCp	0	-	-	-	-	-
RT @CoMoN99: Τώρα τι να πω; Ότι ήθελα να το δω και αυτό πριν πεθάνω; Θα ακουστεί σαν σαρκασμός #eklogesneadimokratia https://t.co/UvN2Z2EaYK	-	-	-	-	-	-
Με άνοδο ξεκίνησε το χρηματιστήριο στην Τρανσυλβανία #eklogesneadimokratia	0	0	0	0	0	0
Εγκεφαλικά επεισόδια #eklogesneadimokratia https://t.co/OanPRXIicQ	0	0	-	-	0	0
RT @doubamari: Καλύτερος ο Νεοφιλελευθερισμός από την σημερινή αριστεροδεξιά βλακεία!! https://t.co/93ytzmsfAN	-	+	+	+	+	+
Καημένε @PanosKammenos , δεν το ξερες ότι το ψάρι είναι λαχτάρα??? #metonKyriako #eklogesneadimokratia #eklogesnd https://t.co/Vv4yODM4Id	-	-	-	-	-	-
Ζαλίζομαι απ' τη μεγάλη ανατροπή. #eklogesneadimokratia	+	-	-	-	-	0
Αν γίνει πρωθυπουργός ο Κυριάκος θα έχουμε πρώτη κυρία τς χώρας μετά τν Περιστέρα την Μαρέβα.Που τα βρίσκουμε ρε σεις; #eklogesneadimokratia	0	0	+	+	0	+
Η οικογενειακή φωτογραφία του Κυριάκου Μητσοτάκη https://t.co/E4KZrfD6GB #metonKyriako #eklogesneadimokratia	0	0	0	+	0	+
RT @FirFirikos1: Πώς αντέδρασε ο Ανεπρόκοπος Φαυλόπουλος στη νίκη του #Kyriakos_Mitsotakis?? #eklogesneadimokratia #pakis #karamanlis https…	-	0	-	+	0	-
RT @bellaciaobella: Λογικά,συγχαρητήρια δέχτηκε κι απο Χριστοφοράκο,αλλά αυτά δεν λέγονται ανοιχτά #eklogesneadimokratia	+	0	+	0	0	+
Κυριάκο αν δεν πεθάνει ο γκαντεμόσαυρος Πρωθυπουργό δεν σε βλέπω #eklogesneadimokratia	+	-	-	-	-	-
Και όπως ήταν αναμενόμενο, "άνοιγμα" των ΑΝΕΛ στα "Καραμανλικά" στελέχη... #eklogesneadimokratia	0	0	0	0	0	+
Εδώ στη Δράση αναζητούμε εθελοντές για να βοηθήσουν στη μετακόμηση. Πληροφορίες εντός #eklogesneadimokratia	+	0	0	0	0	0
RT @stratisil: Προηγείται ο Κούλης. Πανηγυρισμοί στα γραφεία της Siemens! #eklogesnd #eklogesneadimokratia #Kyriakos_Mitsotakis #meimarak…	+	+	+	+	+	+
κατεβαίνουμε ομόνοια να το πανηγυρίσουμε; #metonKyriako #eklogesneadimokratia #eklogesneadimokratia #εκλογέςΝΔ https://t.co/bbjwqP5j0g	+	+	+	+	+	+
RT @mpikchri: Κάποιοι επιλέγουν πρόγραμμα Θεσσαλονίκης, κάποιοι βγάζουν αρχηγό που μιλάει για λιτότητα. Διάλεξε σε ποια μεριά θα'σαι #eklog…	-	-	-	-	-	-
Ποια είναι επιτέλους αυτή η Νίκη Μητσοτάκη? #eklogesneadimokratia	+	0	0	0	0	0
Καρούζος - Siemens Άσσο ημίχρονο Διπλό τελικό #eklogesneadimokratia	0	0	0	+	0	0

RT @GiorgosAlexaki3: Τελικα ουδεις άχρηστος με ονομα βαρυ χαχαχα κλαιω δε φτάνει ψηλά #eklogesneadimokratia #eklogesnd	-	-	-	-	-	-
Ε ξεκολλάτε... #Κυριάκος #eklogesneadimokratia https://t.co/hR0BwYa3BP	0	0	0	0	0	0
RT @avraamakis: Έτοιμη η εθνική ομάδα διαπραγμάτευσης #eklogesneadimokratia https://t.co/QL8qbKPYbX	0	0	0	0	0	0
Ασχημα μαντάτα για ΝΔ.Με την εκλογή Κούλη μειώνονται οι πιθανότητες επιστροφής Καμένου κ Νικολόπ.στο μαντρί #eklogesneadimokratia #eklogesnd	-	0	0	0	-	-
RT @Anergo_Teacher: Η Φιλελευθερη τάση της ΝΔ πήρε κεφάλι, για πρώτη φορά, μετά το 1993. Μένει να δούμε την πορεία της. #eklogesneadimokra…	0	0	+	0	0	+
RT @MrChristos: Έχουν ξεχυθεί οι Νεοδημοκράτες στους δρόμους και τραγουδάνε Τέρη Χρυσό. #eklogesneadimokratia	0	0	0	0	0	0
Μπράβο @kmitsotakis. Δώσε τη μάχη για τη χώρα.#metonKyriako #eklogesneadimokratia #eklogesnd	+	0	+	0	0	0
Να μην ξεχνάμε.... Ο Κυρ.Μητσοτάκης είναι ο μόνος βουλευτής της Ν.Δ. που δεν εψήφισε Παυλόπουλο για ΠτΔ. Ο ΜΟΝΟΣ #eklogesneadimokratia	-	0	0	0	0	0
RT @libertingr: Αυτό είναι το σχέδιο Κούλη ως ηγέτης της ΝΔ https://t.co/QWgbIWW6Rm #eklogesneadimokratia #metonKyriako https://t.co/gG749…	0	0	0	0	0	0
Ο Κυριάκος Μητσοτάκης είναι Χρυσός άνθρωπος.Ξέρει η Ντόρα.#eklogesneadimokratia	0	+	+	0	+	+
Οι #eklogesneadimokratia είχαν το ίδιο ενδιαφέρον με την παρακολούθηση ντοκιμαντέρ για την αναπαραγωγή της αμοιβάδας, σε 300ή επανάληψη.	0	0	0	+	0	0
Και κάποιοι αναρωτιώντουσαν μετά τον Α γύρο, γιατί δεν έμεινε αντιπρόεδρος παραδίδοντας τα σκήπτρα στον Βαγγέλη. #eklogesneadimokratia	0	0	0	0	0	0
RT @karydas_vasilis: Νεκρική σιγή από μερικούς... #eklogesneadimokratia #eklogesnd #metonkyriako	0	-	-	-	-	-
RT @akysmitsoulis: ... είναι τέτοια η απογοήτευση που η ελπίδα αναζητάται στον κυριάκο μητσοτάκη ... #eklogesneadimokratia	-	0	0	+	0	0
Στην αναμπουμπουλα των πανηγυρισμών στο στρατηγείο Μητσοτάτη πρώτη μούρη ο Πέτρος ο Μαντούβαλος...!!! χαχαχαχαχα #eklogesneadimokratia	+	+	+	+	+	+
Οι χορηγοί πρέπει να ανταμείβονται! Άλλωστε πρέπει να στηρίξει και τη νύφη (Σία)! https://t.co/rBQS58xPvb	+	0	0	0	0	0
RT @SirIliasM: Κωστάκη τα έχει αυτά ο δεύτερος γύρος, μπορεί να σου πέσει βαρύς. #eklogesneadimokratia	-	0	0	0	0	0
Η κόρη του Κυριάκου ειναι ίδια ο παππούς της #eklogesneadimokratia	0	0	0	0	0	0
Ο Βαγγέλης συνεχαρει τον Κυριάκο με επιστολή... Τα κυριακατικα τσίπουρα έχουν και τισ συνέπειες τους #eklogesneadimokratia #	+	+	+	+	+	+
Η χώρα που έκανε πρωθυπουργό τον ΓΑΠ.έχει και μία θέση για τον Κυριάκο! #eklogesneadimokratia	0	0	0	0	0	0
Ο Κυριάκος θα ιδιωτικοποιήσει,την εφορία,το ΥΠΕΞ,το ΥΠ.ΠΡΟ.ΠΟ και τις φυλακές.Μετά θα δούμε #eklogesneadimokratia	-	0	0	0	0	-
RT @leftgr: #eklogesneadimokratia: Νέος πρόεδρος της ΝΔ ο Κυριάκος Μητσοτάκης (συνεχής ενημέρωση) :: left.gr https://t.co/rcGSjusoHv	0	0	+	0	0	+

Ε, ρε και να βγει για δηλώσεις ο πατέρας του. Θα έρθει η συντέλεια. Άνετα. #eklogesneadimokratia	0	-	-	-	-	-
Στα παράξενα της κρίσης,γκρεμίζει η αρχιεπισκοπή Καραμανλίδων χάριν του νεοαποικιακού ναίσκου της σχολής Φράιμπουργκ https://t.co/HEhVGLy4x2	-	-	-	-	-	-
RT @AnatropiMegaTV: #anatropi στις #eklogesneadimokratia η πρώτη εκλογή που χάνει ο Τσιπρας από το 2014!	-	-	-	0	-	-
προφητεία #2: αύριο ο μειμαράκης αρχίζει τη διάλυση της γνωστής ΝΔ #eklogesneadimokratia και μετά έρχεται το ευρωπαικό μέτωπο του Κυριάκου	-	-	-	-	-	-
RT @immigrant_mind: Συγκινημένος και ο Τέρης Χρυσός για τη μεγάλη νίκη. #eklogesneadimokratia {SPY} https://t.co/jh0SlkQVoz	+	+	+	+	+	+
Οι καλές ειδήσεις συνεχίζονται!!! ΞΥΔΑΚΙ ΨΕΚΑΣΜΕΝΟΙ!!!! https://t.co/GqBpnqxR97	+	-	0	-	-	0
Κλίμα κηδείας στην ΕΡΤ, ή κάνω λαθος; #eklogesneadimokratia #eklogesnd	0	-	-	-	-	-
RT @steliosmats: Φήμες λένε ότι αν ανεβάσεις τώρα selfie με αυτοκόλλητο #metonkyriako θα σου βρει και δουλειά #eklogesneadimokratia	0	0	+	0	0	+
Ο Φλαμπουραρης εχει πάει τοσες φορές τουαλέτα με τι νικη του @kmitsotakis, που ετοιμάζει διακανονισμό με την ΕΥΔΑΠ https://t.co/QLLoa6veBS	0	0	-	+	0	0
ΑΝΕΛ συνέχεια στην κατρακύλα προς τον υπόνομο ... https://t.co/gkCm5h1EYc	-	-	-	-	-	-
Ο Κυριάκος κέρδισε τον Βαγγέλα και του έκανε και fatality,γούρλωσε τα μάτια και έβγαλε φωτιές. #eklogesneadimokratia	+	+	+	+	+	+
RT @Miropaki: Το θέμα είναι ότι πλήρωσαν κιόλας για να τον ψηφίζουν... #eklogesneadimokratia #eklogesnd	0	-	-	-	-	-
Μην πιστεύεις την κάθε ψυχανωμαλάρα εδώ μέσα. https://t.co/BKrncEk4FM	0	-	-	-	-	-
Η Νέα Δημοκρατία, ενωμένη, ανανεωμένη, δυνατή αλλάζει σελίδα στη χώρα. #metonKyriako #eklogesneadimokratia https://t.co/NsrAPCA2rS	+	+	+	+	+	+
RT @px75: Έγραψες λάθος το "θα χάσω την αργομισθία μου στο ΥΠΕΘΑ". https://t.co/rKJn90K5AM	-	-	-	-	-	-
Για πολλούς λόγους αυτό το τραγούδι να παίζει συνέχεια στην Συγγρού Τάκα τάκα τα-Τέρης Χρυσός https://t.co/BN5sab5CH5 #eklogesneadimokratia	+	0	0	0	0	0
Δεν ξέρω αν η γκαντεμιά ειναι κληρονομική& την κληροδοτήσει ο επίτιμος στο γιο του, η μαλακία του έθνους μας σίγουρα #eklogesneadimokratia	-	-	-	-	-	-
RT @HellenicTex: Σοβαρά λόγια, σοβαρών ανθρώπων, που ψηφίζουν πολιτικούς, οι οποίοι ΚΡΑΤΑΝΕ ΤΙΣ ΥΠΟΣΧΕΣΕΙΣ ΤΟΥΣ!!! https://t.co/gJvzDfWKZ2	+	+	+	+	+	+
Οι σημερινοί ηττημένοι! #eklogesneadimokratia https://t.co/o0m76jGGm2	0	-	-	-	-	-
ΛΟΛ άκου τελετή! Και ο αγιασμός τί ώρα; #eklogesnd #σούργελα #metonKyrako https://t.co/lneKaXQJIU	-	0	0	0	0	0
RT @filo333333: Ένα μπράβο στον @kmitsotakis για τη νίκη και ένα στους ΝΔκράτες που ψήφισαν αλλαγή πλεύσης #eklogesneadimokratia	+	+	+	+	+	+

RT @FirFirikos1: Η αμήχανη στιγμή που ο Μεϊμάρ Πασάς ανακοινώνει τα νέα στον ΚαραμάνΑλή #eklogesneadimokratia #meimarakis #karamanlis https…	-	-	-	-	-	-
RT @aytoproswpwsegw: Η MIZENS πότε και πως θα συγχαρεί το νέο πρόεδρο;#eklogesneadimokratia	0	+	+	+	+	+
Εμείς που δεν είμαστε αριστεροχαρούμενοι και ακροδεξιοβλαμένοι επιτέλους έχουμε κάποιον για να ελπίζουμε. #eklogesneadimokratia	0	-	-	-	-	-
Θα πεθάνουμε όλοι #eklogesneadimokratia	-	-	-	-	-	-
Οι φωτογραφίες νίκης του Κυριάκου με την οικογένειά του με καθησυχάζουν ότι η χώρα έχει παρόν και μέλλον. #eklogesneadimokratia	+	+	+	+	+	+
RT @BackbencherGR: Παιδιά, αύριο να ξυπνήσει κάποιος το Βαγγέλα γιατί έχει να πάει και στο Χατζηνικολάου αύριο #eklogesneadimokratia #eklog…	0	0	0	0	0	0
ΑΝΕΛ: Σκληρό νεοφιλελεύθερο κόμμα θα γίνει η Ν.Δ. https://t.co/YZpJU14OZn #eklogesneadimokratia #eklogesnd	-	0	0	-	0	0
οι #eklogesneadimokratia στο σχόλιο της μέρας. Οι #eklogesnd αφήνουν θετικό μήνυμα.. https://t.co/SLfwiPBIZS	+	+	+	+	+	+
#eklogesneadimokratiaΠήγε η άλλη η ξανθιά κ σκαρφάλωσε σε μια καρέκλα πίσω απτον Κυριάκο να βγει φωτο...φρενίτιδα στο μεγαλείο της!	-	0	0	0	0	0
RT @to_papagalaki: Δηλώσεις Καραμανλή για την εκλογή Μητσοτάκη: "Σήμερα κέρδισε το ποδόσφαιρο!" #eklogesneadimokratia	+	+	+	+	+	+
#eklogesneadimokratia Το νέο σύνθημα της ΝΔ, με τον νέο πρόεδρο: https://t.co/mvvuLdi7XH	0	0	0	0	0	0
Οι ψηφοφόροι που στήριξαν Μητσοτάκη σήμερα είναι κι εκείνοι που παραδοσιακά κρίνουν κυβερνήσεις.#eklogesneadimokratia #eklogesnd	+	+	+	+	+	+
ΕΚΛΟΓΕΣ ΧΤΕΣ!!! #syriza_xeftiles #metonKyriako #Kyriakos_Mitsotakis #eklogesnd #eklogesneadimokratia	0	-	-	-	-	-
Τον Παυλόπουλο τον προσέχει κάποιος απόψε που έχει νεύρα; Δεν ελέγχει και τη δύναμη του αυτός #eklogesneadimokratia	0	-	-	0	-	-
RT @Capitaloskylo: Συνωμοσιολάγνος απ' τα Lidl... https://t.co/LxX04V01WX	0	-	-	-	-	-
RT @BILLTI: '"Έλληνα ανθρωπάκι, θα ταΐζεις μια ζωη ρε τα παιδιά του Μητσοτάκη" #eklogesneadimokratia	-	-	-	-	-	-
Γιατί ρε γαμιόλες; #EklogesNeaDimokratia https://t.co/42kA9xEwfm	-	-	-	-	-	-
RT @wonderboycrete: Α ρε τσολιά, τον γκαντέμιασες τον άνθρωπο! Ο γέρος Μητσοτάκης σε έστειλε; #eklogesneadimokratia https://t.co/1CK80XhdY9	-	-	-	-	-	-
Ο Alpha δείχνει μια την Νάρνια μια τον Κυριάκο κ έχω μπερδευτεί ποιο είναι παραμύθι κ ποιο πραγματικότητα #eklogesneadimokratia #eklogesnd	-	-	0	-	-	0
RT @A_Morellas: Ήττα Μεϊμαράκη; Όχι. Ήττα Καραμανλή και φεουδαρχικής νοοτροπίας ΝΔ. Έπρεπε κάποτε να γίνει κι αυτό. #eklogesneadimokratia	-	-	-	-	-	-
Στον οικονόμο του χώρου που δεν του αρκούσε το #eklogesnd & το ήθελε #eklogesneadimokratia:γιατί όχι #eklogesStiMegaliKentrodexiaParataxi;	0	0	0	0	0	0
RT @lysigakis: Ο Τράγκας είναι ψυχοβγάλτης έχει Αντώναρο στο πάνελ. Κανένα έλεος στον πονεμένο. Μοιρολόι στο πάνελ #eklogesneadimokratia #e…	-	-	-	-	-	-
Όχι άλλο οικογένεια Μητσοτάκη σε fb κ Twitter..Το εμπεδώσαμε... Ένα μαύρο πέπλο γκαντεμιάς μας σκέπασε ομαδικώς✉🌜⚠⑱😂 #eklogesneadimokratia	0	-	-	-	-	-

RT @giatros7: Το να είναι ο Κυριάκος εν δυνάμει πρωθυπουργός δεν είναι μαγκιά...είναι αυτοταπείνωση #eklogesneadimokratia #eklogesnd	-	-	-	-	-	-
RT @Costakisand: Ο μπούλης σε κατάθλιψη.... #karamanlis #neadimokratia https://t.co/bpoY3ABbEX	-	-	-	-	-	-
Πράγματι το ζαβό τώρα αντιπροσωπεύει τους ψηφοφόρους του κόμματος. #eklogesneadimokratia #eklogesnd	-	-	-	-	-	-
RT @kaimporei: Αναλαμβάνει ο γιος του Αντίχριστου.θα ανοίξουν οι ουρανοί απόψε. ΜΕΤΑΝΟΕΙΤΕ. #eklogesneadimokratia	-	-	-	-	-	-
@neademokratia Συγχαρητήρια του Κώστα Καραμανλή στον @kmitsotakis https://t.co/cMqTadgX3u … #Ysterografa #eklogesneadimokratia	+	+	+	+	+	+
Financial Times και Γαλλικό Πρακτορείο για την επικράτηση Μητσοτάκη #eklogesneadimokratia #eklogesnd https://t.co/77tGVpVWAH	0	+	+	+	+	+
Είναι συγκινητικό να κερδίζει τις #eklogesneadimokratia όποιος σε μπλοκάρει μόλις αναφερθείς σε ηλεκτρικά σκεύη... https://t.co/9BWWAW3oJC	+	0	0	0	0	0
Ήδη οι πρώτες δημοσκοπήσεις δείχνουν τον Κυριάκο πέντε μονάδες μπροστά σαν καταλληλότερο πρωθυπουργό #eklogesneadimokratia	0	+	0	+	+	+
RT @_bourdas: άγιε βασίλη φέτος ήμουν καλό παιδί. θα με κάνεις πρόεδρο? #eklogesneadimokratia https://t.co/DGMAM2s5AS	+	+	0	+	+	+
Οικονομικά με τον Βαρουφάκη σπούδασε αυτος https://t.co/5xb61mDjii	0	0	0	0	0	0
Εντονη ενόχληση στους Ανεξάρτητους Ελληνες από την εκλογή Μητσοτάκη #eklogesneadimokratia #eklogesnd https://t.co/na5xfXeFhJ	-	-	-	-	-	-
RT @foititopateras: Μόνο ο Κυριάκος μπορεί να κοιτάξει τους Ευρωπαίους κατάματα. Όλους μαζί. #eklogesneadimokratia	0	0	0	0	0	0
#eklogesneadimokratia Ανδ. Παπαμιμίκος: Όλοι στο πλευρό του Κ. Μητσοτάκη https://t.co/CPoQn1Wuoy #eklogesnd	0	0	0	0	0	0
Η σημερινή εκπομπή του Τράγκα θα πρέπει να προβάλλεται σε πονεμένους για να γελάνε.Μαύρη κωμωδία το μοιρολόι Αντώναρου #eklogesneadimokratia	0	-	-	-	-	-
Τόση χαρά, επειδή στους τυφλούς επικράτησε ο μονόφθαλμος;Ρε, δεν πάμε καλά... #eklogesneadimokratia	+	-	-	-	-	-
Ελπίζω μονο να μην πανε οικογενειακός σε Καρρα η Τερζη να διασκεδάσουν. Τους συμπαθώ #eklogesneadimokratia	+	0	0	0	0	0
Η μεγάλη είδηση είναι ότι ο Καραμανλής τηλεφώνησε και ΜΙΛΗΣΕ στον Κυριάκο! #eklogesneadimokratia #eklogesnd	+	0	0	0	0	0
RT @kanaliotis: Δεν πρόλαβε να γίνει ένας Μητσοτάκης πρόεδρος της #ΝΔ και άρχισαν τα προβλήματα επικοινωνίας ρεπόρτερ - στούντιο #eklogesne…	-	-	-	-	-	-
Φαντάσου να είναι σαν το δημοψήφισμα και τελικά αυριο το πρωί να γίνει αρχηγός ο Βαγγέλης #eklogesneadimokratia #eklogesnd	0	0	0	0	0	0
Τέτοιες μαλακίες κάνατε και παλιά στην ΝΔ και κυβέρνησε 20 χρόνια το ΠΑΣΟΚ #eklogesneadimokratia	-	-	-	-	-	-
RT @Aristeroextrem: Συγκινητικά βρίσκω επίσης τα dm σας σχετικά με τους λαοφιλείς συναποφοίτους του νέου αρχηγού #eklogesneadimokratia http…	+	+	+	+	+	+
ούτε 2 ώρες πρόεδρος ο Μητσοτάκης και δεν μπορούν να σταυρώσουν σύνδεση τα κανάλια....#eklogesneadimokratia	0	0	0	0	0	0
Η χαρά του συριζαίου απόψε να εκτονώσει το αριστερό του απωθημένο πάντως..#eklogesneadimokratia	+	0	0	0	0	0

RT @ClarkGaybeul: Ο ΚΟΥΛΗΣ ΠΡΟΕΔΡΟΣ (μια νέα εποχή για τη σάτιρα, γλυκοχαράζει) #eklogesnd #eklogesneadimokratia https://t.co/0YSaMgF41Q	0	0	0	+	0	0
Τραγάκης: Τη Δευτέρα το τελικό αποτέλεσμα #eklogesneadimokratia https://t.co/OpUsrm1V1P	0	0	0	0	0	0
Θα δεις θα σε δικαιώσω κάποτε και ας σου λένε όλοι ότι είμαι ζαβό #eklogesneadimokratia https://t.co/szwV0FVQ3v https://t.co/oepbfEOqs7	-	0	0	0	0	0
RT @Lilasta: Κι ο Παπαμιμίκος χάρηκε, που δεν θα γίνει πορτιέρης #eklogesneadimokratia	+	0	0	+	0	+
#eklogesneadimokratia Πληροφορίες για συνάντηση Αλ. Τσίπρα - Κυρ. Μητσοτάκη εντός της εβδομάδας https://t.co/CPoQn1Wuoy #eklogesnd	0	0	0	0	0	0
Ο Ψινάκης στηρίζει Κυριάκο. "Κυριάκο θα σκίσεις" https://t.co/BBdxTJippe #eklogesneadimokratia #Mitsotakis	+	+	+	+	+	+
RT @GiaPich: Τελικά προφήτης ο Jo Di... ή η ιστορία επαναλαμβάνεται ως φάρσα! #eklogesneadimokratia https://t.co/O9Csa2eu8b	-	-	-	-	-	0
Αλέξη γάμα, για 10 χρόνια πρωθυπουργός και 32 μνημόνια θα σαι #eklogesneadimokratia #eklogesnd #mpeee https://t.co/2nCZi6TCKa	0	-	-	-	-	-
RT @philosofos: Στη χώρα με τη μεγαλύτερη ανεργία στην ΕΕ,το να πανηγυρίζεις για κάποιον που υποστηρίζει Μνημόνια είναι βλακώδες. #eklogesn…	-	-	-	-	-	-
RT @to_papagalaki: Έξαλλοι πανηγυρισμοί στο Σύνταγμα από οπαδούς Κ. Μητσοτάκη που γιορτάζουν γδέρνοντας ζωντανούς δημόσιους υπαλλήλους! #ek…	+	+	+	+	+	+
RT @LPapastergiou: Μα να να μην διαβάζουν οι νεοδημοκράτες την Αυγή; #eklogesneadimokratia	0	0	0	0	0	0
... από την μία ο τσίπρας από την άλλη ο κυριάκος θα περπατάμε και τα μνημόνια θα μας μπαίνουν στον κώλο ... #eklogesneadimokratia	-	-	-	-	-	-
RT @TheoNT1: Στην #ert παντως μπορείτε να νιώθετε ηδη μελλοντικοί απολυμένοι.Ο χρονος μετρα αντίστροφα συντροφοι. #eklogesneadimokratia	-	-	-	-	-	-
RT @dennisgreek: Τρίβει τα χέρια του ο Τσιπρας. Σου λέει, με τέτοιους μαλάκες αντιπάλους εγώ θα κυβερνάω άλλα 10 χρόνια. #eklogesneadimokr…	-	-	-	-	-	-
Όχι άλλοι Σωτήρες του τόπου ρεεε!! #eklogesneadimokratia	0	+	+	+	+	+
Για την εκλογή νέου προέδρου φωταγωγήθηκαν στα χρώματα της ΝΔ τα κεντρικά γραφεία της Siemens #eklogesneadimokratia	0	+	+	+	+	+
RT @protagongr: Κυριάκος, για τους νοήμονες ρε γαμώτο…, της Ρέας Βιτάλη https://t.co/IC8eBP8MgP #eklogesneadimokratia https://t.co/icLeEKaV…	-	+	+	0	+	+
Πάρει, δεν πάρει, η χλαπάτσα που έφαγε ο Βούδας της Ραφήνας κ οι υποτακτικοί είναι τεράστια #eklogesneadimokratia https://t.co/Faaebl1FDc	-	-	-	-	-	-
Θεοδωράκης #ToPotami σε Κυριάκο: Μπράβο σου.Ελπίζω να στρέψεις το καράβι σε προοδευτικές όχθες #eklogesneadimokratia https://t.co/Oa9hWVvzG8	+	0	0	0	0	0
Ούτε ο Βαγγέλης τέτοιο ζόρι! Μωρέ μπράβο! https://t.co/fMmqsnpekz	-	-	-	-	-	-
RT @MichGregor: Για τους "άπιστους θωμάδες"...#eklogesneadimokratia #metonKyriako https://t.co/vIlHgsjWDK	0	-	-	-	-	-
RT @Alexandros1909: Τσίπρα τρέμε. Σύντομα θα τελειώσεις και εσύ και η συμμορία των ακροδεξιών ψεκασμένων, κομμουνιστών και Πασόκων. #ekloge…	-	-	-	-	-	-

RT @naftemporikigr: #eklogesneadimokratia "Συγχαρητήρια φίλε" είχε γράψει νωρίτερα στο Twitter ο Αδ. Γεωργιάδης https://t.co/CPoQn1Wuoy ht…	+	0	0	0	0	0
Το σύστημα δεν έπεσε,παράταση δεν έδωσαν,ανησυχώ ότι κάτι αλλάζει #eklogesneadimokratia #eklogesnd	-	-	-	0	-	-
RT @LPapastergiou: Ο δρομος για τον Μητσοτάκη θα είναι ανηφορικός και δύσκολος.Στο εσωτερικό δε της ΝΔ, ακόμα δυσκολότερος #eklogesneadimok…	-	-	-	-	-	-
RT @iefimerida: Financial Times και Γαλλικό Πρακτορείο για την επικράτηση Μητσοτάκη #eklogesneadimokratia #eklogesnd https://t.co/77tGVpVWAH	0	0	0	0	0	0
RT @LiberalGr: Κυριάκο μην κάνεις πίσω! https://t.co/iVgvuGAAa0 #LiberalGr #eklogesneadimokratia	0	0	0	0	0	0
Αντικειμενικές τοποθετήσεις από αυτούς που νοιάζονται για τη ΝΔ https://t.co/nzLMTdwtmf	+	0	0	+	0	0
RT @PavlidisTheo: Μετά τη νίκη του αυτοδημιούργητου Μητσοτάκη η διαπλοκή έχει χεστεί πάνω της μιλάμε. #EklogesNeaDimokratia	-	-	-	-	-	-
RT @pseudo_chemeng: Πόσο ξεφτίλες? #fimotro #klaiw_goera #eklogesneadimokratia https://t.co/O61IbBa0kc	-	-	-	-	-	-
RT @Ouinston: Συριζανελληνικά: -Μεσαίων Υπαρκτού Νεοφιλελευθερισμού: δυτική Ευρώπη. -Αριστερός παράδεισος: πτωχευμένο Ελλαδιστάν. https://t…	-	0	-	0	0	-
Ο Καραμανλής δε κρύβεται... θα μιλήσει και θα το δείτε όταν δώσει δαχτυλίδι ο Τσίπρας... #eklogesneadimokratia #eklogesnd #metonKyriako	-	0	0	0	0	+
RT @CiaoVaso22: Φιλελέρες,συνήγοροι Ναζί,Ποταμίσιοι,γλοιώδεις φασίστες,όλοι μια ωραία ατμόσφαιρα εκεί στη ΝΔ #eklogesneadimokratia	-	-	-	-	-	-
Retweeted The TOC (@TheTOC_gr): Η μεγάλη νίκη Μητσοτάκη και η βαριά ήττα Καραμανλή #eklogesneadimokratia... https://t.co/G47g065iCH	-	0	0	0	0	0
Σιγά να μη...στο συνέδριο τον περιμένουν πολλοί στη γωνία... https://t.co/Tb2a2dnIpi	0	0	0	0	0	0
Κλείνει ο κύκλος της βεντέτας της πτώσης της κυβέρνησης Κων. Μητσοτακη με την στήριξη Σαμαρα στον Κυριάκο. #eklogesneadimokratia	+	0	0	0	0	0
Κάτι λίγο πιο όμορφο και αφοπλιστικό για να έχουμε ένα υπέροχο βράδυ. #metonkyriako #eklogesneadimokratia https://t.co/j0gNEEmLns	+	0	0	0	0	0
RT @ysterografa: @GKoumoutsakos Κουμουτσάκος: Τώρα, ΝΔ ενωμένη και ανανεωμένη! https://t.co/WLt04EZykZ #Ysterografa #eklogesneadimokratia	0	+	+	+	+	+
RT @pitsirikos: Αμέσως μετά την εκλογή του, ο Κυριάκος Μητσοτάκης πήγε και κατέθεσε λουλούδια στο μνημείο της Καισαριανής. #eklogesneadimok…	+	0	0	0	0	0
-Μπαμπά βγήκα! -Μπράβο ματάρες μου #eklogesneadimokratia	+	+	+	+	+	+
ΠΡΟΒΛΕΨΗ: Οι πολιτικές εξελίξεις με τις #eklogesneadimokratia θα φέρουν αύξηση στον αγροτικό τομέα. Νέα παραγωγή #sano	+	+	+	+	+	+
Ανανέωση-με γιό Μητσοτάκη που τον στήριξε γιός Τζιτζικώστα κόντρα σε ανηψιό Καραμανλή τον οποίο στήριζε κόρη Μητσοτάκη #eklogesneadimokratia	+	0	0	0	0	0

-Γιαγιά ο Μητσοτάκης είναι ο νέος πρόεδρος της Ν.Δ. -Άλλαξε κανάλι παιδάκι μου ολο επαναλήψεις έχει #eklogesneadimokratia	0	0	0	0	0	0
RT @mpikchri: Οι αγκαλιές με τον Μαντούβαλο πάντως, μεγάλη αλλαγή δε δείχνει. . Να τα λέμε κι αυτά. #eklogesneadimokratia	+	0	0	0	0	0
RT @seed30_Greek: Δημοψήφισμα θα ανακοινώσει αύριο ο Τσίπρας "Σίγουρα θέλετε Κυριάκο στη ΝΔ;" #eklogesneadimokratia	0	0	0	0	0	0
- Ποιος νίκησε; - Κανείς. Όλοι χάσαμε. #eklogesneadimokratia	+	0	0	0	0	0
Τελικά ποιός βγήκε...ο Κυριάκος,ο Μήτσος ή ο Τάκης ;; #eklogesneadimokratia	0	0	0	0	0	0
Πλακιωτάκης: απόψε δεν υπάρχουν ηττημένοι https://t.co/FJ56UpxUwx via @AntennaNews #eklogesneadimokratia @G_Plakiotakis @kmitsotakis	0	+	+	+	+	+
RT @Arhsx: #eklogesneadimokratia Αυτοδημιούργητος και χωρίς χορηγούς. https://t.co/zFnDdvXhdF	0	+	+	+	+	+
RT @SirIliasM: Με οριακό αποτέλεσμα θα έχει περισσότερο ενδιαφέρον η συνέχεια και το ξεκατίνιασμα, γιατί από ουσία δεν αλλάζει κάτι. #eklog…	-	0	0	0	0	0
RT @MarEyedoll: Η χαρά του πατέρα #eklogesneadimokratia https://t.co/Iy8PIKqPep	+	+	+	+	+	+
RT @News247gr: Κωνσταντίνος Μητσοτάκης σε Κυριάκο: Συγχαρητήρια και πρωθυπουργός! https://t.co/RqGelbl6qf #eklogesneadimokratia https://t.c…	+	+	+	+	+	+
Μ.Ο. ηλικίας εκλογέων: 60 έτη. Ανανέωση μαλάκες μου. #eklogesneadimokratia	-	0	-	0	0	0
RT @seed30_Greek: Το τέλος του αριστερού Καραμανλισμού και η αρχή για μία Ευρωπαική Δεξιά #eklogesneadimokratia	0	0	0	0	0	0
RT @iefimerida: Ο Ηλίας Ψινάκης ενθουσιάστηκε με τον Μητσοτάκη & ανέβασε πανέμορφη φωτό [εικόνα] #eklogesneadimokratia #eklogesnd https://t…	+	+	+	+	+	+
Τώρα που ο Κούλης βγήκε πρόεδρος ΝΔ,επιβάλλεται δωράκι σε όλους τους ψηφοφόρους της τηλέφωνο Siemens. #eklogesneadimokratia #eklogesnd	0	+	+	+	+	+
Ο άρχοντας των οικογενειοκρατιών #eklogesneadimokratia https://t.co/KCeipGXfaA	0	0	0	0	0	0
RT @Andrey_Vyshinsk: Έπεσε και η παράγκα του Πακησταν-Τσίπρας #metonKyriako #Kyriakos_Mitsotakis #syriza_xeftiles #eklogesneadimokratia	0	-	-	-	-	-
RT @el_politis: Καλά ξεκίνησε αυτή η χρονιά #eklogesneadimokratia #eklogesnd	0	+	+	+	+	+
Άλλο ένα γεγονός που συγκινεί είναι η αρωγή του λαϊκού @ToPotami στην εκλογή νέου αρχηγού... #eklogesneadimokratia https://t.co/M6cDFUiDOG	+	+	+	+	+	+
Δηλαδή για να καταλάβω την πρώτη μέρα που ο Κυριάκος γίνεται πρόεδρος ο Μειμαράκης δίνει συνέντευξη; Ως τι; #eklogesneadimokratia	-	0	0	0	0	0
#eklogesneadimokratia Δεν μπορεί ποτέ να ξυπνήσει ένας κοιμισμένος έναν άλλο κοιμισμένο https://t.co/e5AYgtQ4yV	-	-	-	-	-	-

RT @mar1agi: Ας αναγνωρίσει κάποιος τη πολιτική οξυδέρκεια του Νίκου Χατζηνικολάου να βιαστεί να κλείσει συνέντευξη με τον ηττημένο #ekloge…	0	+	+	+	+	+
Η ανανέωση στη ΝΔ ξεκίνησε ήδη: Fax τέλος! Με mail τα αποτελέσματα!! #eklogesnd #eklogesneadimokratia https://t.co/PusvjOS9Zi	0	+	+	+	+	+
RT @minouli7: Δείτε εδώ την αντίδραση του Κυριάκου όταν έμαθε ότι κέρδισε τις εκλογές #eklogesneadimokratia https://t.co/ae9mlxUYUw	+	+	+	+	+	+
RT @iefimerida: Η πρώτη δήλωση Κυριάκου Μητσοτάκη ως νέος αρχηγός της ΝΔ[βίντεο] #eklogesneadimokratia #eklogesnd #Mitsotakis #vouli https:…	0	0	0	0	0	0
Η δήλωση του Κυριάκου Μητσοτάκη μετά τη νίκη του https://t.co/UlVnWQrTai #eklogesneadimokratia #eklogesnd	+	+	+	+	+	+

Πίνακας 25: Σύνολο δεδομένων για αξιολόγηση του ταξινομητή, μαζί με αποτελέσματα ταξινόμησης και επισημειώσεις βαθμολογητών

Ε. Tweets με προβληματική επισημείωση

Tweets	*Annotation*				
RT @decadenza35: Αυτοί που είναι βέβαιοι για τον πατέρα Τσίπρα και τη χούντα,δεν έχουν ακούσει κάτι για τον Κυριάκο και τη siemens.. #eklog…	+	+	+	0	+
Επικράτηση Μητσοτάκη στη γαλάζια μάχη για την Προεδρία #eklogesneadimokratia @kathimerini_gr https://t.co/0XwGClMgYG	+	+	+	+	+
Μπράβο @kmitsotakis. Δώσε τη μάχη για τη χώρα.#metonKyriako #eklogesneadimokratia #eklogesnd	0	+	0	0	0
Οι χορηγοί πρέπει να ανταμείβονται! Άλλωστε πρέπει να στηρίξει και τη νύφη (Σία)! https://t.co/rBQS58xPvb	0	0	0	0	0
RT @SirIliasM: Κωστάκη τα έχει αυτά ο δεύτερος γύρος, μπορεί να σου πέσει βαρύς. #eklogesneadimokratia	0	0	0	0	0
RT @leftgr: #eklogesneadimokratia: Νέος πρόεδρος της ΝΔ ο Κυριάκος Μητσοτάκης (συνεχής ενημέρωση) :: left.gr https://t.co/rcGSjusoHv	0	+	0	0	+
RT @Miropaki: Το θέμα είναι ότι πλήρωσαν κιόλας για να τον ψηφίζουν... #eklogesneadimokratia #eklogesnd	-	-	-	-	-
ΛΟΛ άκου τελετή! Και ο αγιασμός τί ώρα; #eklogesnd #σούργελα #metonKyrako https://t.co/lneKaXQJIU	0	0	0	0	0
ΑΝΕΛ: Σκληρό νεοφιλελεύθερο κόμμα θα γίνει η Ν.Δ. https://t.co/YZpJU14OZn #eklogesneadimokratia #eklogesnd	0	0	-	0	0
Ήδη οι πρώτες δημοσκοπήσεις δείχνουν τον Κυριάκο πέντε μονάδες μπροστά σαν καταλληλότερο πρωθυπουργό #eklogesneadimokratia	+	0	+	+	+
RT @naftemporikigr: #eklogesneadimokratia "Συγχαρητήρια φίλε" είχε γράψει νωρίτερα στο Twitter ο Αδ. Γεωργιάδης https://t.co/CPoQn1Wuoy ht…	0	0	0	0	0
Αντικειμενικές τοποθετήσεις από αυτούς που νοιάζονται για τη ΝΔ https://t.co/nzLMTdwtmf	0	0	+	0	0
Κάτι λίγο πιο όμορφο και αφοπλιστικό για να έχουμε ένα υπέροχο βράδυ. #metonkyriako #eklogesneadimokratia https://t.co/j0gNEEmLns	0	0	0	0	0
RT @SirIliasM: Με οριακό αποτέλεσμα θα έχει περισσότερο ενδιαφέρον η συνέχεια και το ξεκατίνιασμα, γιατί από ουσία δεν αλλάζει κάτι. #eklog…	0	0	0	0	0
Financial Times και Γαλλικό Πρακτορείο για την επικράτηση Μητσοτάκη #eklogesneadimokratia #eklogesnd https://t.co/77tGVpVWAH	+	+	+	+	+

Πίνακας 26: Tweets με προβληματική επισημείωση

ΣΤ. Stop words

άλλες	ανωτέρω	εκείνων	κάπου	μόνο	ποιόν	τα	όλον
άλλη	απ	εκτός	κάπως	μόνοι	ποιός	ταχατα	όλος
άλλην	απέναντι	εμάς	κάτι	μόνος	ποιών	ταύτα	όλου
άλλης	απο	εμένα	κάτω	μόνου	πολύ	ταύτες	όλους
άλλο	από	εμείς	καθ	μόνους	που	ταύτη	όλων
άλλοι	απόψε	εμπρός	καθένα	μόνων	πουθε	ταύτην	όλως
άλλον	αργά	εν	καθένας	ν	πουθενά	ταύτης	όμως
άλλος	αργότερο	εντελώς	καθεμιά	να	πρέπει	ταύτο	όποια
άλλοτε	αριστερά	εντωμεταξύ	καθεμιάς	ναι	πριν	ταύτος	όποιαν
άλλους	αρκετά	εντός	καθενός	νωρίς	προ	ταύτου	όποιας
άλλων	αρχικά	ενός	καθετί	ξανά	προκειμένου	ταύτων	όποιες
άμα	ας	ενώ	καθόλου	ξαφνικά	προς	τελικά	όποιο
άμεσα	αυτά	εξ	καθώς	ο	προτού	τελικώς	όποιοι
άνω	αυτές	εξής	και	οι	προχθές	τες	όποιον
άξαφνα	αυτή	εξίσου	κανένα	ολονεν	προχτές	τη	όποιος
άρα	αυτήν	επ	κανέναν	ολωσδιόλου	πρωτύτερα	την	όποιου
άραγε	αυτής	επάνω	κανένας	ολόγυρα	πρόκειται	της	όποιους
έγκαιρα	αυτοί	επί	κανείς	ολότελα	πρόπερσι	τι	όποιων
έκαστα	αυτού	επίσης	κανεν	οποιαδήποτε	πως	τις	όπου
έκαστες	αυτούς	επειδή	κανενός	οποιανδήποτε	πόσες	τισ	όπως
έκαστη	αυτό	επι	κατ	οποιασδήποτε	πόση	το	όσα
έκαστην	αυτόν	επομένως	κατά	οποιδηποτε	πόσην	τοι	όσες
έκαστης	αυτός	εσάς	κατιτι	οποιεσδήποτε	πόσης	τον	όση
έκαστο	αυτών	εσένα	κατόπιν	οποιοδήποτε	πόσοι	τος	όσην
έκαστοι	αφού	εσείς	κιόλας	οποιονδήποτε	πόσος	του	όσης
έκαστον	αφότου	εσύ	κλπ	οποιοσδήποτε	πόσους	τουλάχιστο	όσο
έκαστος	αχ	ετέρες	κοντά	οποιουδήποτε	πότε	τουλάχιστον	όσοι
έκαστους	αύριο	ετέρου	κτλ	οποιουσδήποτε	σήμερα	τους	όσον
έκαστων	β	ετερης	κυρίως	οποιωνδήποτε	σαν	τούτα	όσος
ένα	βέβαια	ετερους	λίγο	οποτεδήποτε	σας	τούτες	όσου
έναν	βεβαιότατα	ετερων	λιγάκι	οπουδήποτε	σε	τούτη	όσους
ένας	γ	ετουτα	λιγότερο	οπότε	σεις	τούτην	όσων
έξω	γι	ετουτες	λοιπά	ορισμένα	σιγά	τούτης	όταν
έπειτα	για	ετούτη	λοιπόν	ορισμένες	σου	τούτο	ότου
έστω	γρήγορα	ετούτην	λόγω	ορισμένων	στα	τούτοι	ύστερα
έτερη	γύρω	ετούτης	μάλιστα	ορισμένως	στη	τούτοις	ώσαν
έτερο	δ	ετούτο	μάλλον	οσαδήποτε	στην	τούτον	ώσπου
έτεροι	δίπλα	ετούτοι	μέλει	οσεσδήποτε	στης	τούτος	ώστε
έτερον	δα	ετούτον	μέλλεται	οσηδήποτε	στις	τούτου	
έτερος	δείνα	ετούτος	μέσα	οσηνδήποτε	στο	τούτους	
έτσι	δεξιά	ετούτου	μέχρι	οσησδήποτε	στον	τούτων	
έφαφνα	δηλαδή	ετούτους	μήδε	οσοδήποτε	στου	τυχόν	
ήδη	δι	ετούτων	μήπως	οσοιδήποτε	στους	των	
ήτοι	δια	ευτυχώς	μήτε	οσονδήποτε	στων	τόσα	
ήττον	διαρκώς	εφεξής	μία	οσοσδήποτε	συ	τόσες	
ίδια	δικά	εχθές	μα	οσουδήποτε	συγχρόνως	τόση	
ίδιαν	δικοί	εχτές	μαζί	οσουσδήποτε	συν	τόσην	

ίδιας	δικού	εως	μακάρι	οσωνδήποτε	συνάμα	τόσης	
ίδιες	δικούς	η	μακριά	οτι	συνήθως	τόσο	
ίδιο	δικό	θα	μας	οτιδήποτε	συνεπώς	τόσοι	
ίδιοι	δικός	ι	με	ου	συνοί	τόσον	
ίδιον	δν	ιδίως	μείον	ουδε	συχνά	τόσος	
ίδιος	ε	ιι	μεθαύριο	ούτε	συχνές	τόσου	
ίδιου	εάν	ιιι	μεμιάς	πάλι	συχνή	τόσους	
ίδιους	είθε	κ	μεν	πάντοτε	συχνήν	τόσων	
ίδιων	είτε	κάθε	μερικά	πάντως	συχνής	τότε	
ίσαμε	εαυτού	κάμποσα	μερικές	πέρα	συχνου	τώρα	
ίσια	εαυτούς	κάμποσες	μερικοί	πέρσι	συχνού	υπ	
ίσως	εαυτό	κάμποση	μερικούς	πέρυσι	συχνούς	υπέρ	
α	εαυτόν	κάμποσην	μερικών	πίσω	συχνό	υπό	
αδιάκοπα	εαυτών	κάμποσης	μετ	παντού	συχνόν	υπόψη	
αι	εγκαίρως	κάμποσο	μετά	παρά	συχνός	υπόψιν	
ακριβώς	εγώ	κάμποσοι	μεταξύ	περί	συχνών	φέτος	
ακόμα	εδώ	κάμποσον	μιά	περίπου	συχνώς	χαμηλά	
ακόμη	ειδεμη	κάμποσος	μιάν	περι	σχεδόν	χθες	
αλλά	εις	κάμποσου	μια	περισσότερο	τάδε	χτες	
αλλάχου	εκ	κάμποσους	μιαν	πια	τάχα	χωριστά	
αλλιώς	εκάστου	κάμποσων	μιας	πιθανόν	τέτοια	ψηλά	
αλλιώτικα	εκεί	κάποια	μολονότι	πιο	τέτοιαν	ω	
αλλοιώς	εκείνα	κάποιαν	μονάχα	πλάι	τέτοιας	ως	
αλλοιώτικα	εκείνες	κάποιας	μονομιάς	πλέον	τέτοιες	ωστόσο	
αλλού	εκείνη	κάποιες	μου	πλην	τέτοιο	ωσότου	
αμέσως	εκείνην	κάποιο	μπορεί	ποιά	τέτοιοι	ωχ	
αν	εκείνης	κάποιοι	μπορούν	ποιάν	τέτοιον	όλα	
ανά	εκείνο	κάποιον	μπρος	ποιάς	τέτοιος	όλες	
ανάμεσα	εκείνοι	κάποιος	μόλις	ποιές	τέτοιου	όλη	
αναμεταξύ	εκείνον	κάποιου	μόνες	ποιοί	τέτοιους	όλην	
αντί	εκείνος	κάποιους	μόνη	ποιου	τέτοιων	όλης	
αντίπερα	εκείνου	κάποιων	μόνην	ποιους	τίποτα	όλο	
αντις	εκείνους	κάποτε	μόνης	ποιό	τίποτε	όλοι	

Πίνακας 27: Stop words που χρησιμοποίησε το σύστημά μας