Συναισθηματική Ανάλυση Ελληνικών Tweets και Hashtags με χρήση Λεξικού Συναισθημάτων

Δημήτριος Μάλλης, Γεώργιος Καλαματιανός, Δημήτριος Νικολαράς, Αυγερινός Αραμπατζής
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Πολυτεχνική Σχολή, Δημοκρίτειο Πανεπιστήμιο Θράκης, Ξάνθη 67 100, Ελλάδα
dimimall1@ee.duth.gr, georkala3@ee.duth.gr, diminiko4@ee.duth.gr, avi@ee.duth.gr

***Abstract*—Η ραγδαία ανάπτυξη των μέσων κοινωνικής δικτύωσης έχει καταστήσει την εξόρυξη απόψεων και συναισθημάτων των χρηστών (opinion and sentiment mining) ένα καινούργιο και σημαντικό πεδίο έρευνας της εποχής μας. Στα πλαίσια αυτής της εργασίας ασχοληθήκαμε με την πλατφόρμα microblogging Twitter με σκοπό να εξάγουμε την διάθεση χρηστών ως προς διάφορες θεματικές κατηγορίες (hashtags). Η εξαγωγή του συναισθήματος των χρηστών γίνεται με χρήση ελληνικού sentiment lexicon. Η διαδικασία που προτείνουμε είναι ικανή να εντοπίσει το επικρατέστερο συναίσθημα των χρηστών και να εξάγει συμπεράσματα για την διάθεση τους στις θεματικές κατηγορίες που εξετάζονται , τόσο συνολικά όσο και για διακριτά χρονικά διαστήματα.**

Index Terms*— *Εξόρυξη Συναισθήματος, Κοινωνικά Δίκτυα, Twitter, Λεξικό Συναισθημάτων, Εξόρυξη Άποψης

I. Εισαγωγη

Η διάθεση των ανθρώπων ως προς διάφορα ζητήματα ενδιαφέροντος αποτελεί σημαντική πληροφορία τόσο ως προς τις κοινωνικές όσο και ως προς τις πολιτικοοικονομικές της προεκτάσεις. Οι παραδοσιακές μέθοδοι εκτίμησης άποψης ή/και συναισθήματος περιλαμβάνουν μη αυτοματοποιημένη αξιολόγηση δεδομένων από διάφορες πηγές όπως δημοσκοπήσεις, έρευνες κ.α., διαδικασίες οι οποίες είναι χρονοβόρες και δεν δίνουν άμεσα αποτελέσματα. Κατά συνέπεια, είναι υπαρκτή η ανάγκη αυτοματοποιημένων λύσεων στο παραπάνω πρόβλημα. Η έξαρση χρήσης των κοινωνικών δικτύων στο Διαδίκτυο καθιστά τον τομέα της αυτόματης εξαγωγής συναισθήματος έναν σημαντικό και αναπτυσσόμενο κλάδο της εξόρυξης και ανάκτησης πληροφορίας.

Η παρούσα εργασία πραγματοποιήθηκε στα πλαίσια του μαθήματος Προχωρημένα Θέματα Βάσεων Δεδομένων, 2014-2015, του Δημοκριτείου Πανεπιστημίου Θράκης. Οι συντάκτες είναι οι:

Μάλλης Δημήτριος, Καλαματιανός Γεώργιος και Νικολαράς Δημήτριος, προπτυχιακοί φοιτητές του τμήματος Ηλεκτρολόγων μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Δημοκριτείου Πανεπιστημίου Θράκης.

Αραμπατζής Αυγερινός, Επίκουρος Καθηγητής του τμήματος Ηλεκτρολόγων μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών του Δημοκριτείου Πανεπιστημίου Θράκης, και διδάσκοντας του μαθήματος Προχωρημένα Θέματα Βάσεων Δεδομένων στα πλαίσια του οποίου υλοποιήθηκε η εργασία.

Μια κλασσική προσέγγιση στο πρόβλημα της εξαγωγής συναισθήματος αποτελεί το “affective text”, δηλαδή η ανάλυση του συναισθηματικού περιεχομένου λέξεων και τμημάτων κειμένου. Η λογική αυτή χρησιμοποιήθηκε στο SemEval-2007 [1] με σκοπό τον προσδιορισμό του συναισθήματος που προκαλούν στον αναγνώστη διάφοροι τίτλοι ειδήσεων. Μια πιο σύγχρονη προσέγγιση είναι το Latent Dirichlet Allocation (LDA) [2], το οποίο αποτελεί ένα θεματικό μοντέλο που επιχειρεί να εξάγει διαφορετικό συναίσθημα για κάθε λέξη ανάλογα με το συγκείμενο και το θεματικό πλαίσιο.

Ακολουθώντας τη λογική του “affective text”, σε αυτή την εργασία εξετάζονται δεδομένα κειμένου, που συλλέχθηκαν από την πλατφόρμα microblogging Twitter, ως προς το συναισθηματικό τους περιεχόμενο. Τα δεδομένα αυτά (στο εξής θα αναφέρονται ως tweets) είναι όλα στην ελληνική γλώσσα. Οι στόχοι μας είναι οι εξής:

- Η δημιουργία μιας μεθόδου η οποία βαθμολογεί συναισθηματικά τα ελληνικά tweets, για πολλαπλά είδη συναισθημάτων όπως θυμός, φόβος, χαρά, έκπληξη, κ.α.
- Η δημιουργία μιας μεθόδου για την συναισθηματική αξιολόγηση θεματικών κατηγοριών (hashtags), με χρήση των βαθμολογημένων tweets,
- Η ανάλυση της μεταβολής των συναισθημάτων που περιέχουν οι διάφορες κατηγορίες (hashtags) στο χρόνο.

Οι αξιολογήσεις των συναισθημάτων γίνονται με χρήση ελληνικού λεξικού συναισθημάτων (Greek Sentiment Lexicon) [3].

Η διαφοροποίηση μας από τις προϋπάρχουσες προσεγγίσεις έγκειται αρχικά στην χρήση της ελληνικής γλώσσας η οποία δεν έχει εξεταστεί μέχρι τώρα με παρόμοιες διαδικασίες εξόρυξης, τουλάχιστον στην βιβλιογραφία που εξετάσαμε. Ακόμα, η μέθοδος που προτείνεται είναι αρκετά απλή και αποδοτική κατά την εκτέλεση, γεγονός που την καθιστά εύκολα εφαρμόσιμη σε μεγάλες συλλογές δεδομένων. Τέλος, εξάγουμε γενικότερο συμπέρασμα για τον τρόπο που χρησιμοποιούν οι Έλληνες χρήστες το twitter και προσδιορίζουμε τα συναισθήματα που εκφράζουν συνηθέστερα μέσα από αυτό.

Στη συνέχεια της εργασίας ακολουθείται η παρακάτω δομή. Τα δεδομένα, οι πόροι που χρησιμοποιούνται και η μεθοδολογία αναλύονται στο μέρος II. Τα πειράματα και η εξαγωγή των αποτελεσμάτων παρουσιάζονται στο μέρος III. Στο μέρος IV συζητάμε διάφορες παρατηρήσεις που κάναμε κατά την εκτέλεση των πειραμάτων. Τέλος, παρουσιάζουμε τα συμπεράσματα και τις προτάσεις μας για περαιτέρω έρευνα στο μέρος V.

II. Συναισθηματικη Αναλυση Tweets και Hashtags

A. Συλλογή Δεδομένων

Για τη συλλογή των δεδομένων έγινε χρήση του Streaming API του Twitter μέσω της γλώσσας προγραμματισμού Python. Η τακτική που ακολουθήθηκε ήταν η σε βάθος αναζήτηση του social graph του twitter (Depth First Search). Ξεκινώντας από έναν αυθαίρετα ορισμένο χρήστη, επιλέξαμε να χτίσουμε μια λίστα χρηστών προς αναζήτηση, τοποθετώντας σε αυτή τους χρήστες που ακολουθούν ("followers") τον πρώτο χρήστη. Στη συνέχεια εξετάσαμε επαναληπτικά τους χρήστες που περιέχονται στην λίστα, συλλέγοντας τα tweets τους αλλά και τα ids των followers τους, ώστε να εξεταστούν στο μέλλον. Μερικές παρατηρήσεις είναι:

1. Η επιλογή των "followers" αντί των "following" χρηστών, έγινε με σκοπό να αποφευχθεί, όσο είναι δυνατόν, η συχνή ανάκτηση δημοσίων προσώπων οι οποίοι ακολουθούνται ("followed") από μεγάλο αριθμό χρηστών.
2. Στη λίστα αναζήτησης προστίθενται μόνο followers χρηστών με ελληνικούς χαρακτήρες στα tweets τους, ώστε να περιορίσουμε όσο το δυνατόν την αναζήτηση μας σε Έλληνες χρήστες.
3. Δεν προσθέτουμε στη λίστα προς αναζήτηση τους followers όλων των χρηστών που εξετάζονται. Αφενός γιατί το πλήθος των χρηστών είναι πολύ μεγάλο και το μέγεθος της λίστας θα αυξανόταν σημαντικά. Αφετέρου γιατί οδηγεί σε παραπάνω αχρείαστα αιτήματα στο API του twitter, το οποίο έχει περιορισμό σε 180 αιτήματα ανά 15 λεπτά από κάθε εφαρμογή.
4. Τα tweets που ανακτήθηκαν περιείχαν ελληνικούς χαρακτήρες του Unicode.
5. Η συλλογή των δεδομένων διήρκεσε μεγάλο χρονικό διάστημα λόγω πάλι του περιορισμού του API. Από κάθε χρήστη που εξετάστηκε ανακτήθηκαν τα τελευταία 200 tweets αλλά και η χρονική στιγμή που πραγματοποιήθηκε το καθένα από αυτά.

Στο Table I παρουσιάζονται στατιστικά της συλλογής των tweets που δημιουργήθηκε και στο Figure 1 το «σύννεφο» με τα 100 πιο δημοφιλή hashtags όπου τα πιο συχνά-εμφανιζόμενα hashtags εμφανίζονται με μεγαλύτερη γραμματοσειρά. Το Figure 2 δίνει την κατανομή των tweets που συλλέχθηκαν ανά μέρα.

TABLE I
ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ DATASET

Μέγεθος Dataset	832.1 MB
Αριθμός Tweets	4,373,197
Αριθμός Χρηστών	30,778
Αριθμός Hashtags	54,354
Hashtags με >1000 tweets	41
Χρονικό Διάστημα	24-04-2008 εώς 29-11-2014

Figure 1. Τα 100 πιο δημοφιλή Hashtags

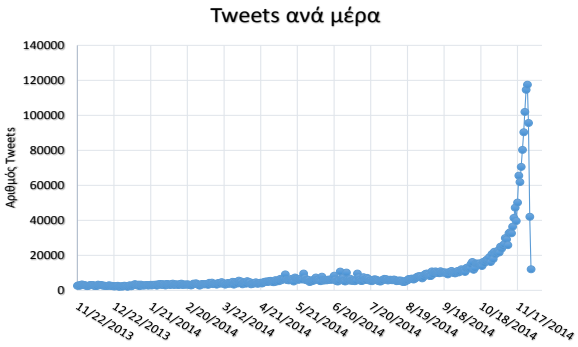

Figure 2. Tweets ανά ημέρα

B. Λεξικό Συναισθημάτων

Το λεξικό συναισθημάτων που χρησιμοποιήθηκε είναι το Greek Sentiment Lexicon [2], το οποίο περιέχει 2,315 λήμματα αξιολογημένα κατά τα εξής έξι συναισθήματα: anger (θυμός), disgust (απέχθεια), fear (φόβος), happiness (χαρά), sadness (λύπη), surprise (έκπληξη).

Το λεξικό περιλαμβάνει συναισθηματική αξιολόγηση των λημμάτων από τέσσερεις ανεξάρτητους βαθμολογητές. Είναι σαφές ότι το συναίσθημα που αντιλαμβάνεται κάποιος για μια λέξη είναι υποκειμενικό. Στην εργασία αυτή επιλέχθηκε ο μέσος όρος των τεσσάρων βαθμολογιών ώστε να προκύψουν οι τελικές βαθμολογίες κάθε λήμματος.

Στο λεξικό περιέχονται ακόμα κάποιες γλωσσολογικές πληροφορίες που αφορούν τα λήμματα ,όπως μέρος του λόγου, αντικειμενικότητα της κάθε λέξης όπως αξιολογήθηκε

από τον κάθε βαθμολογητή, ενώ περιέχεται ένα ακόμα πεδίο με σχόλια που διευκρινίζουν την χρήση του όρου. Οι παραπάνω πληροφορίες δεν λάβαμε υπόψιν στην παρούσα εργασία.

C. Προεπεξεργασία Δεδομένων

Η προεπεξεργασία των δεδομένων έγινε με χρήση Hadoop [5] και της προγραμματιστικής τεχνικής MapReduce [6], λόγω του μεγάλου όγκου των δεδομένων. Συγκεκριμένα:

- Χωρίσαμε τα tweets σε αρχεία ανάλογα με το hashtag που τα περιγράφει. Ακόμα ενώσαμε παρόμοια hashtags αφαιρώντας μη αλφαριθμητικούς χαρακτήρες και κάνοντας τα κεφαλαία γράμματα πεζά. Για παράδειγμα το hashtag #wcgr14 και #WCgr14 ομαδοποιήθηκαν στην ίδια κατηγορία. Επιλέξαμε να εξετάσουμε μόνο τα hashtags που εμφανίζονται σε πάνω από 1000 tweets, ώστε να έχουμε αρκετά δεδομένα προς αξιολόγηση σε κάθε θεματική κατηγορία. Λόγω της συνήθης πρακτικής των χρηστών του twitter να χρησιμοποιούν μεγάλο πλήθος hastags στα tweets τους, ένα tweet μπορεί να κατηγοριοποιηθεί σε παραπάνω από ένα hashtags. Ακόμα επιλέξαμε να κρατήσουμε και τις αναδημοσιεύσεις tweets από άλλους χρήστες εφόσον θεωρήσαμε ότι συμφωνούν με το συναίσθημα που εκφράζουν χρήστες αυτοί.
- Αφαιρέσαμε από τα tweets λέξεις από μια λίστα 627 ελληνικών stop-words, για μείωση του χώρου και υπολογιστικού φόρτου.
- Αντικαταστήσαμε τους τονισμένους χαρακτήρες με τους αντίστοιχους μη τονισμένους, και κάναμε κεφαλαία τα γράμματα για να έχουν την ίδια μορφοποίηση με το λεξικό και με τον stemmer που χρησιμοποιήθηκε (στο επόμενο βήμα).
- Εφαρμόσαμε τόσο στα δεδομένα όσο και στο λεξικό έναν ελληνικό stemmer [4] για να αυξήσουμε την αντιστοίχιση σε ομόριζες λέξεις.

D. Μεθοδολογία Συναισθηματικής Αξιολόγησης Tweet

Για κάθε λήμμα του λεξικού που εντοπίζεται σε ένα συγκεκριμένο tweet, δημιουργούμε ένα διάνυσμα $\overrightarrow{W}$ με 6 συνιστώσες, μία για κάθε ένα από τα συναισθήματα που μελετάμε. Τελικά έχουμε N διανύσματα $\overline{W_j}$

$$\overline{W_j} = [\mathrm{w}_{1_j} \quad \mathrm{w}_{2_j} \quad \mathrm{w}_{3_j} \quad \mathrm{w}_{4_j} \quad \mathrm{w}_{5_j} \quad \mathrm{w}_{6_j}]$$

όπου $\mathrm{j} = 1 \ldots \mathrm{N}$ και N είναι το πλήθος λέξεων που εντοπίζονται στο tweet.

Στη συνέχεια εξάγουμε ένα διάνυσμα 6 συνιστωσών $\overrightarrow{T}$

$$\overline{T} = [t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5 \quad t_6]$$

όπου κάθε συνιστώσα προκύπτει από την σχέση

$$t_i = \sqrt{\frac{\sum_{j=1}^{N} w_{i_j}^2}{N}} \qquad i = 1 \ldots 6 \qquad (1)$$

με i να αποτελεί το πλήθος των συνιστωσών του $\overline{T}$. Η σχέση (1) είναι ο τετραγωνικός μέσος (quadratic mean) των λημμάτων $\overrightarrow{W}$ που εντοπίστηκαν σε κάθε tweet. Ο μέσος αυτός επιλέχθηκε αντί του απλού αριθμητικού μέσου, λόγω της ιδιότητας του να δίνει μεγαλύτερες τιμές για περιπτώσεις συνιστωσών με μεγάλη διασπορά. Με αυτόν τον τρόπο πετυχαίνει να αναδεικνύει τα λήμματα που έχουν υψηλή τιμή σε κάποια από τις συνιστώσες τους.

E. Μεθοδολογία Συναισθηματικής Αξιολόγησης Hashtags

Στη συνέχεια συνδυάσαμε τα διανύσματα των tweets t_j κάθε hashtag $\overrightarrow{H}$,

$$\overline{H} = [h_1 \quad h_2 \quad h_3 \quad h_4 \quad h_5 \quad h_6]$$

πάλι με τον quadratic mean, ώστε να εξάγουμε τελικό συμπέρασμα για κάθε κατηγορία.

$$h_i = \sqrt{\frac{\sum_{j=1}^{M} t_{i_j}^2}{M}} \qquad i = 1 \ldots 6 \qquad (2)$$

και M είναι το πλήθος των tweets που εντοπίστηκαν να περιέχουν συναίσθημα για κάθε hashtag.

F. Μεθοδολογία Μέτρησης της Συναισθηματικής Μεταβολής Hashtag στον Χρόνο

Σε κάθε ένα από τα αρχεία hashtag που επιλέξαμε να μελετήσουμε, ταξινομήσαμε αρχικά τα περιεχόμενα τους κατά αύξοντα χρόνο.

Στην συνέχεια, τηρώντας διαστήματα μίας ημέρας υπολογίσαμε το μέσο συναίσθημα για εκείνη την ημέρα σύμφωνα με την διαδικασία που περιγράφεται στην προηγούμενη ενότητα.

Εξετάσαμε μόνο τα αποτελέσματα των ημερών που συγκέντρωναν αριθμό tweets μεγαλύτερο του 60, κρίνοντας ότι ημέρες με λιγότερα από 60 tweets ανά hashtag θα εισήγαν θόρυβο στο συνολικό χρονοδιάγραμμα.

III. ΠΕΙΡΑΜΑΤΑ

A. Συναισθήματα Μεμονωμένων Tweets

Στη συνέχεια παρουσιάζεται η επίδοση του αλγόριθμου σε συγκεκριμένα tweets ώστε να γίνει καλύτερα αντιληπτή η λειτουργία του σε μικρή κλίμακα.

***#kalokairipantou**: Καλημέρα αγαπημένοι μου! Μου λείψατε εχθές... Ετοιμαζόμαστε για το #KalokairiPantou και σας ταξιδεύουμε στους Παξούς.*

***# panellinies2014**: προτιμω να χαραμισω τα μορια μου παρα τη ζωη μου #panellinies2014 #apotelesmata*

***#vouli**: Πεστε την αληθεια εκει στην #vouli κανετε ψηφοφοριες για να σχολιαζουμε εμεις.*

***#eurovisiongr**: Καλημέρα..... Καλή εβδομάδα.... Πάλι δουλειά... Αλλά... Το βράδυ έχει party... #madtv #eurovisiongr #eurosong*

Figure 3. Παραδείγματα tweets

Η βαθμολόγηση όπως προέκυψε για τα παραπάνω τέσσερα tweets δίνεται στον Πίνακα ΙΙ.

TABLE II
ΒΑΘΜΟΛΟΓΗΣΗ ΠΑΡΑΔΕΙΓΜΑΤΩΝ

#	Anger	Disgust	Fear	Happiness	Sadness	Surprise
1	1.00	1.00	1.00	4.75	1.00	2.75
2	3.50	3.50	1.00	1.00	1.00	2.50
3	2.58	2.00	0.79	0.79	0.95	1.63
4	1.00	1.00	1.00	3.75	1.00	2.50

Βλέπουμε ότι ο αλγόριθμος είναι ικανός να εντοπίσει την διάθεση του χρήστη στις παραπάνω περιπτώσεις. Για το πρώτο και τέταρτο tweet έχει εξάγει πιο έντονο συναίσθημα χαράς, ενώ για το δεύτερο και τρίτο tweet συναίσθημα θυμού και απέχθειας, αποτελέσματα το οποία συμβαδίζουν με την κοινή αντίληψη.

B. Συναισθήματα των Hashtags

Στον Πίνακα ΙΙΙ παρουσιάζονται συνολικά αποτελέσματα για κάποιες από τις θεματικές κατηγορίες που εξετάστηκαν.

TABLE III
ΒΑΘΜΟΛΟΓΗΣΗ HASHTAGS

#	Anger	Disgust	Fear	Happi-ness	Sadness	Surprise
#wc14gr	1.3910	1.2862	0.9512	1.3604	0.8412	1.4552
#ekloges14	1.1627	1.1676	0.8180	1.1456	0.7219	1.2885
#kalokairipantou	0.7930	0.9158	0.7739	2.1856	0.7570	2.1084
#skouries	1.0608	1.0460	0.9399	1.0603	0.7337	1.1197
#panellinies2014	1.3900	1.3374	0.9810	1.4521	0.8153	1.4659
#vouli	1.3040	1.2608	0.7832	1.1767	0.7419	1.3122
#ert	1.0892	1.0757	0.8065	1.0242	0.6694	1.1292
#mb14gr	1.3948	1.2742	0.9510	1.3451	0.8041	1.4225
#eurovisiongr	1.3464	1.2957	0.7933	1.3533	0.7599	1.4092
#enikos	1.3189	1.2866	0.8195	1.1918	0.7616	1.3551

Παρατηρούμε ότι ο αλγόριθμος που κατασκευάστηκε είναι ικανός να βγάλει συμπέρασμα ως προς το συναισθηματικό περιεχόμενο των θεματικών κατηγοριών που πάλι ανταποκρίνεται στη διαίσθηση μας. Πράγματι, οι κατηγορίες όπως Παγκόσμιο Κύπελλο Ποδοσφαίρου (#wc14gr), Καλοκαίρι Παντού (#kalokairipantou) και Eurovision (#eurovisiongr) εξάγουν ένα συναίσθημα χαράς, σε αντίθεση με πολιτικά ζητήματα όπως η βουλή (#vouli), το κλείσιμο της ΕΡΤ (#ert) και το ζήτημα στις Σκουριές (#skouries), όπου μεγαλύτερη τιμή παίρνουν τα συναισθήματα του θυμού και της απέχθειας.
Συγκριτικά για τις διάφορες κατηγορίες μπορούμε να πούμε ακόμα, ότι το sadness και fear παίρνουν μεγάλες τιμές για το hashtag #panellinies2014, γεγονός που ταιριάζει στην συγκεκριμένη θεματική κατηγορία.

C. Μεταβολή Συναισθημάτων Hashtags στον Χρόνο

Στη συνέχεια εξάγουμε το συναίσθημα των χρηστών ως προς δύο χαρακτηριστικά hashtags, ανά χρονική περίοδο μίας μέρας. Παρακάτω παρουσιάζονται τα συνολικά αποτελέσματα για τις συνιστώσες της χαράς και του θυμού:

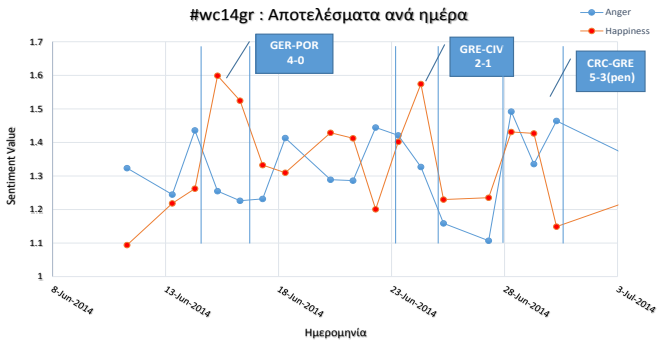

Figure 5. #wc14gr: Αποτελέσματα ανά ημέρα

Όσον αφορά στα πειράματα που εξετάζουν τα hashtags σε βάθος χρόνου, παρατηρούμε ότι είναι ικανά να εντοπίσουν κορυφές στις τιμές συναισθημάτων που μπορούν να συσχετιστούν με επίκαιρα γεγονότα. Για παράδειγμα, το θετικό αποτέλεσμα του αγώνα ποδοσφαίρου μεταξύ Ελλάδας και Ακτής Ελεφαντοστού συμπίπτει χρονικά με μεγάλες τιμές χαράς και μικρές τιμές θυμού. Ακόμα και το παιχνίδι ανάμεσα στην Γερμανία και την Πορτογαλία το οποίο συγκέντρωσε το ενδιαφέρον του ελληνικού κοινού εμφανίζει μεγάλες τιμές χαράς κάτι που γίνεται αντιληπτό όταν εξετάσει κάποιος τα

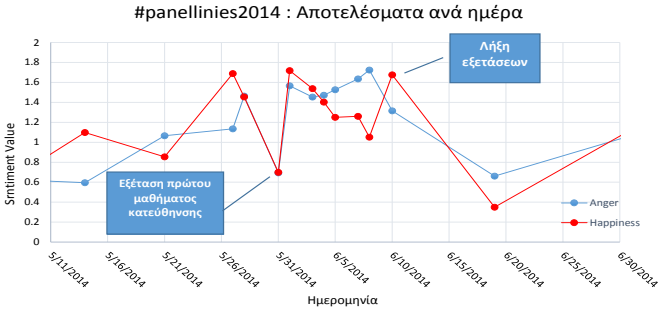

Figure 4. #panellinies2014: Αποτελέσματα ανά ημέρα

tweets που αναφέρονται σε αυτό.

Τέλος, στην περίπτωση των πανελληνίων εξετάσεων μπορούμε να εντοπίσουμε χαμηλές τιμές και στα δύο συναισθήματα που μετριούνται, πριν την εξέταση των---κατά κοινή ομολογία---δυσκολότερων μαθημάτων κατεύθυνσης, αλλά και υψηλές τιμές στο συναίσθημα της χαράς την ημέρα της λήξης των εξετάσεων.

IV. Παρατηρησεις

Κατά τη υλοποίηση της εργασίας αυτής έγιναν οι εξής παρατηρήσεις:

- Η προσέγγιση αυτή δεν είναι ικανή να αξιολογήσει tweets που περιέχουν ειρωνικά σχόλια και διφορούμενα νοήματα, τα οποία υπάρχουν σε πληθώρα στην συγκεκριμένη πλατφόρμα, αλλά tweets με ξεκάθαρο συναισθηματικό περιεχόμενο.
- Παρατηρείται ότι συναισθήματα όπως Anger και Disgust αλλά και Happiness και Surprise εμφανίζουν παραπλήσιες τιμές στις ίδιες κατηγορίες, με αποτέλεσμα να μην μπορούν να διαχωριστούν. Θεωρούμε ότι το φαινόμενο αυτό οφείλεται στο μεγάλο βαθμό συσχέτισης που έχουν ανά δύο τα παραπάνω συναισθήματα στο λεξικό που χρησιμοποιείται. Όντος κάτι τέτοιο είναι εμφανές στον πίνακα όπου παρουσιάζονται οι τιμές της μετρικής linear correlation μεταξύ όλων των ζευγαριών όλων των συναισθημάτων που εξετάζονται.

TABLE IV
LINEAR CORRELATION ΣΥΝΑΙΣΘΗΜΑΤΩΝ

	Anger	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger		**0.827**	0.500	0.002	0.384	0.465
Disgust	**0.827**		0.427	-0.105	0.370	0.403
Fear	0.500	0.427		0.205	0.530	0.549
Happiness	0.002	-0.105	0.205		0.196	**0.558**
Sadness	0.384	0.370	0.530	0.196		0.425
Surprise	0.465	0.403	0.549	**0.558**	0.425	

- Ενδιαφέρουσα παρατήρηση αποτελεί το γεγονός ότι στη περίπτωση των αποτελεσμάτων ανά ημέρα για το hashtag #wc14gr το συναίσθημα της χαράς έχει αντίθετες μεταβολές σε σχέση με το συναίσθημα του θυμού. Αντίθετα στο #panellinies2014 οι αυξομειώσεις εμφανίζουν μεγαλύτερη ομοιότητα. Γενικά θα μπορούσαμε να πούμε ότι στην περίπτωση ενός ποδοσφαιρικού κυπέλλου τα συναισθήματα αυτά δεν κυριαρχούν ταυτόχρονα, ενώ στα πλαίσια των πανελληνίων εξετάσεων είναι λογικό να είναι ανάμεικτα για ίδιες χρονικές στιγμές.

- Στη συλλογή μας δεν εμφανίζεται μεγάλος αριθμός από εικονίδια emoticons τα οποία περιέχουν ξεκάθαρη συναισθηματική πληροφορία. Από το υλικό που έχουμε μπορούμε να συμπεράνουμε ότι οι Έλληνες χρήστες δεν τα χρησιμοποιούν σε μεγάλο βαθμό.

- Το λεξικό που χρησιμοποιήσαμε δεν είναι σχεδιασμένο για διαδικτυακό λόγο και δεν περιέχει μεγάλο αριθμό λέξεων που εμφανίζονται συχνά στα tweets με αποτέλεσμα να μην είναι το ιδανικότερο για αυτή τη δουλειά. Μετρήσαμε ότι μόνο το 11.7% των λέξεων που εξετάζονται περιέχονται στο λεξικό. Παρόλα αυτά, οι μέθοδοι που προτείνουμε φαίνεται να δουλεύουν ικανοποιητικά.

- Γενικά παρατηρήθηκε ότι η συνιστώσες του φόβου και της λύπης παίρνουν μικρότερες τιμές σε σχέση με τις υπόλοιπες. Το γεγονός αυτό μπορεί να οφείλεται τόσο στο ότι γλωσσολογικά είναι πιο δύσκολο να προσδιοριστούν μέσω του κοινού λόγου του διαδικτύου, όσο και στο ότι ο μέσος χρήστης δεν προτιμά τα μέσα κοινωνικής δικτύωσης για να εκφράσει τέτοιου είδους συναισθήματα.

V. Συμπερασματα- Προτασεις για Βελτιωσεις

Η διαδικασία που προτείνουμε δίνει ενθαρρυντικά αποτελέσματα και μπορούμε να πούμε ότι είναι εφικτό να εξάγουμε το συναίσθημα χρηστών ως προς διάφορες κατηγορίες με τη χρήση λεξικού συναισθημάτων. Τα αποτελέσματά μας φαίνονται να είναι σωστά ως προς τα συναισθήματα Anger και Happiness ενώ για τα υπόλοιπα συναισθήματα η εικόνα δεν είναι πάντα τόσο ξεκάθαρη.

Ως προτάσεις για βελτίωση της μεθόδου που προτείνουμε αναφέρουμε:

- Χρήση λεξικού ειδικά κατασκευασμένου για διαδικτυακή εφαρμογή.
- Αξιοποίηση γλωσσολογικών δεδομένων όπως, πχ, το μέρος του λόγου κάθε λήμματος.
- Επέκταση σε εφαρμογή πραγματικού χρόνου ώστε να γίνει εφικτή η εξαγωγή αποτελεσμάτων για δεδομένα που αφορούν τρέχοντα ζητήματα.

VI. Ευχαριστιες

Ευχαριστούμε τον υποψήφιο διδάκτορα του Τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Δημοκριτείου Πανεπιστημίου Θράκης, Συμεών Συμεωνίδη, για τη καθοδήγηση του κατά την πραγματοποίηση της εργασίας.

References

[1] Carlo Strapparava and Rada Mihalcea (2007 June), SemEval-2007 Task 14: Affective Text, Presented at SemEval [Online]. Available: http://dl.acm.org/citation.cfm?id=1621487

[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[3] Adam Tsakalidis, "Greek Sentiment Lexicon". Available online: http://socialsensor.eu/results/datasets/147-greek-sentiment-lexicon

[4] Georgios Ntais, "Development of a Stemmer for the Greek Language", Master Thesis at Stockholm University / Royal Institute of Technology, Department of Computer and Systems Sciences, February 2006 Link: http://deixto.com/greek-stemmer/

[5] Tom White, “Hadoop: The definitive guide” 2nd editions, O’Reilly.

[6] Jeffrey Dean, Sanjay Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters”, 6th Symposium on Operating System Design and Implementation (OSDI 2004), San Francisco, California, December 6-8, 2004.

[7] Yanghui Rao, Qing Li, Xudong Mao, Liu Wenyin, “Sentiment topic models for social emotion mining”, *Information Sciences*, Vol 266, pages 90-100, 10 May 2010. Elsevier.