Sentiment Analysis of Greek Tweets and Hashtags using Sentiment Lexicon

Dimitrios Mallis, Georgios Kalamatianos, Dimitrios Nikolaras, Symeon Symeonidis
Department of Electrical and Computer Engineering
Polytechnic School, Democritus University of Thrace, Xanthi 67 100, Greece
dimimall1@ee.duth.gr, georkala3@ee.duth.gr, diminiko4@ee.duth.gr, ssymeoni@ee.duth.gr

***Abstract*— The rapidly increasing growth of social media has rendered opinion and sentiment mining an important field of research. Within this project we examined the microblogging platform "Twitter" in order to extract the users' concerning different subjects (hashtags). Sentiment mining is achieved using a greek sentiment lexicon. The suggested process is capable of detecting the users' dominant sentiment while the conclusion it draws concerning the users' mood about the examined topics, appears to coincide with common knowledge. The results are presented both in total as well as over time intervals.**

Keywords — Sentiment Mining, Social Media, Twitter, Sentiment Lexicon, Opinion Mining

I. Introduction

Users' disposition towards topics of interest constitutes a valuable piece of information concerning both social as well as financial implications. Traditional opinion or sentiment mining methods consist of non-automated data evaluation sources such as researches or polls which are time consuming and fail to provide immediate results. Consequently, the need for an automated solution is apparent. The rapid increase in usage of social media has rendered automated sentiment mining a very important field of research in data mining and information retrieval.

This project examines text data, collected from the microblogging platform Twitter, as far as their sentimental content is concerned.

The present paper was accomplished within the course Advanced Databases, 2014-2015, in Electrical and Computer Engineering School of Democritus University of Thrace. The authors are:

Mallis Dimitrios, Kalamatianos Georgios, Nikolaras Dimitrios, undergraduate students and Symeon Symeonidis PhD candidate in Electrical and Computer Engineering School of Democritus University of Thrace.

The data (hereinafter referred to as tweets) are in Greek modern language. Our goals are the following:

- The implementation of a method which provides a sentiment rating for the Greek tweets, for a variety of sentiment such as anger, fear, happiness, surprise.
- The implementation of a method providing sentimental evaluation for different topics (hashtags) using rated tweets.
- The analysis of the change in sentiments over time, concerning certain hashtags.

The evaluations are accomplished using a Greek Sentiment Lexicon [3].

Our approach differs from existing research primarily in the use of Greek language which has not been examined, at least to our knowledge, for the purposes of sentiment analysis. Moreover, our method is fairly simple and efficient, since the ratings are a result of direct calculations derived from the words constructing the tweet, avoid the use of classification algorithms. This renders the method appropriate to be applied in massive datasets. Finally, we extract an overall conclusion about the use of Twitter from Greek users and determine the most frequently occurring sentiments.

The remainder of this paper is organized as follows. Related work is given in Section II. Section III consists of the analysis of the dataset, resources and method of work. The experiments and respective results are described in Section IV. In section V we discuss certain remark that arose during the run of the experiments. Finally, we present our conclusion and suggestions for future work in Section VI.

II. Related Research

A first approach to the problem of sentiment mining is "affective text", namely the sentiment analysis of segments of text. This method was used in SemEval-2007 [1] in purpose of determining the sentiment evoked in readers by different news headlines. Another tool used in sentiment mining is Latent Dirichlet Allocation (LDA) [2], which is a model attempting to extract the sentiment of each word according to the context and the topic of the text. Pang and Lee [9] presented an extensive overview of the problem in 2008.

The dominant approach, especially for Twitter, is the use of classification algorithms. Pak and Paroubek [10] use tweets

containing emoticons to attribute a sentiment rating to the words within them so as to build a training dataset. The collection of tweets was gathered from newspaper Twitter accounts (e.g. New York Times) and the classification is achieved using Naïve Bayes algorithm.

Kouloumpis, Wilson and Moore (2011) [11] theorize that the words which occur in certain hashtags have a specific sentiment value. For example, a highly rated positive sentiment is attributed to the words that occur in hashtag #thingsilike. In accordance with the above hypothesis, they train an AdaBoost classifier.

III. Sentiment Analysis of Tweets and Hashtags

A. Data Collection

The collection of data was achieved using the Streaming API Twitter via Python programming language. The approach we followed was a depth first search of the social graph of Twitter. Starting from a random user, we built a search list of users, with the "followers" of the first user. We then tested iteratively users contained in the list, collecting their tweets and the ids of their followers, to be examined in the same way as the process continues. Some comments about the process:

1. The selection of followers instead of following users, was made to avoid, as much as possible, the frequent recovery of public figures who are followed by a large number of users.
2. We only add users who use Greek characters in their tweets, so as to limit our search to Greek users only.
3. We do not add all of the users' followers, firstly because the number of users is very large and the size of the list would increase significantly, and secondly because it leads to more unnecessary requests to the API of Twitter, which is limited to 180 requests per 15 minutes for each application.
4. The tweets that we gather contain at least 4 Greek Unicode characters, to ensure the usage of Greek language.
5. Data collection lasted a long time (about a week) because of the limitation of API. For each user examined, only his 200 most recent tweets were recovered, including the timestamp of every one of them.

Table I provides statistical information for our dataset and in Figure 1 we present the cloud with the 100 most popular hashtags where the most frequently occurring hashtags are displayed in a larger font. Figure 2 provides the distribution of gathered tweets per day only for the period 22.11.2013 to 17.11.2014 whereas previous dates are not shown due to small size of data.

TABLE I
DATASET STATISTICS

Dataset Size	832.1 MB
Number of Tweets	4,373,197
Number of Users	30,778
Number of Hashtags	54,354
Hashtags with >1000 tweets	41
Time Span	24-04-2008 εώς 29-11-2014

Figure 1. Τα 100 πιο δημοφιλή Hashtags

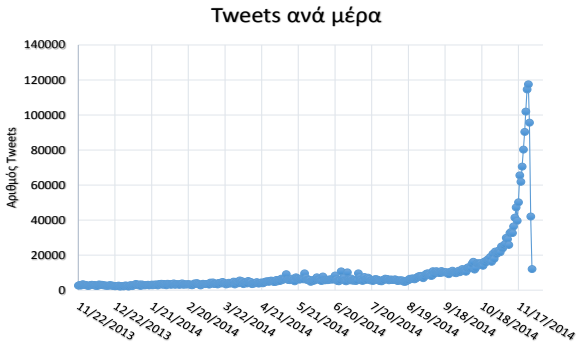

Figure 2 Tweets per day

The sharp increase observed in the number of tweets in the last days of the shown range, is due to the limitation of gathering only 200 tweets per user. This restriction together with the different posting frequencies of different users, is leading to a large number of tweets in the days preceding the start of gathering, and a smaller number in the more distant past. Posts of more active users is limited to a short period, while those of less systematic users is spread across the time range. Finally, due to the fact that the search lasted one week, there is a sharp drop in the number of tweets in the last 3 days of the search for which it was not possible to examine a large number of users.

B. Sentiment Lexicon

The sentiment dictionary that we use in this paper, is the Greek Sentiment Lexicon [3], which contains 2315 entries

evaluated for the following six emotions: anger, disgust, fear, happiness, sadness, surprise.

The dictionary includes emotional evaluation of entries by four independent raters. The rate of every entry is subjective due to individual rating and because of the fact that we use the average of the four scores to get the final score for each entry.

The dictionary also contained some linguistic information regarding the entries, as the part of speech, objectivity of each word as evaluated by each rater and also a field with comments that explain the use of the term. The above information is not taken into consideration in this work.

C. Data Preprocessing

The preprocessing of the data was performed using Hadoop [5] and MapReduce [6], due to the large size of data. Specifically:

- We divided the tweets in files according to their hashtags. We also merged similar hashtags by removing non-alphanumeric characters and substituting uppercase letters with lowercase ones. For example, the hashtag # wcgr14 and # WCgr14 were grouped in the same category. We chose to examine only the hashtags occurring in over 1000 tweets, so that we have enough data to assess in each thematic category. Due to the usual practice of twitter users to use many hashtags in their tweets, a tweet can be classified in more than one hashtags. Finally we chose to keep reposted tweets from other users (retweets) cause we theorize that they agree with the sentiment expressed by these users.
- We removed 627 Greek stop-words [8] from our data, to reduce the size and computational work.
- We replace intonated characters with corresponding non-intonated, and turned every letter to uppercase in order to have the same formatting as the dictionary and the stemmer that we used (in the next step).
- We applied a Greek stemmer [4] to both the data and the dictionary to increase the matching of the words.

D. Method of Tweet Sentiment Evaluation

For each entry of the lexicon which we identify in each tweet, we for a vector $\vec{W}$ with 6 components, one for each examined sentiment. We then have N vectors $\overline{W_j}$

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

where $\mathrm{j} = 1...\mathrm{N}$ and N is the number of entries that are identified in the tweet.

We then form a 6 component vector $\vec{T}$

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

of which, each component is a result of the following formula:

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

where i is the number of components of vector $\overline{T}$. Formula (1) is the quadratic mean of entries $\overrightarrow{W}$ that were identified in each tweet. Said formula was selected instead of the Arithmetic Mean, given its property to return higher values in cases of components with high variance. In this way it highlights the entries with a high value in one of their components.

E. Method of Hashtag Sentiment Evaluation

In the next step we combine the tweets vectors t_j for every hashtag $\overrightarrow{H}$,

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

using the quadratic mean. The final formula for every hashtag is,

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

where M is the number of tweets with sentiment value for each hashtag.

F. Method of Hashtag Sentiment evaluation over time

We sort the content of the hashtags file we chose to examine, according to time in incrementing values. Based on the process we propose in the previous sector, we calculate the average sentiment for one day intervals. We choose to examine only days for which we have gathered more than 60 tweets, in order to avoid the introduction of noise in our results.

IV. Experiments

A. Sentiments of individual Tweets

The following figure presents the performance of the algorithm in some hand-picked tweets, followed by loose translations, in order to better demonstrate its function.

***#kalokairipantou**: Καλημέρα αγαπημένοι μου! Μου λείψατε εχθές... Ετοιμαζόμαστε για το #KalokairiPantou και σας ταξιδεύουμε στους Παξούς.*
***#kalokairipantou**: Good morning my dears! I missed you yesterday... We're getting ready for #KalokairiPantou travel with you to Paxoi.*

***#panellinies2014**: προτιμω να χαραμισω τα μοια μου παρα τη ζωη μου #panellinies2014 #apotelesmata*
***#panellinies2014**: I prefer to waste my grades rather than my life #panellinies2014 #apotelesmata*

***#vouli**: Πεστε την αληθεια εκει στην #vouli κανετε ψηφοφοριες για να σχολιαζουμε εμεις.*
***#vouli**: You should tell the truth there at #vouli, you vote so that we have something to comment*
.
***#eurovisiongr**: Καλημέρα..... Καλή εβδομάδα.... Πάλι δουλειά... Αλλά... Το βράδυ έχει party... #madtv #eurovisiongr #eurosong*
***#eurovisiongr**: Good morning..... Have a good week.... Work again... But... Tonight we party... #madtv #eurovisiongr #eurosong*

Figure 3. Παραδείγματα tweets

The rates we calculate for the tweets in Figure 3 are the following:

Table II
Example Ratings

#	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

We can see that the algorithm is capable of determining the sentiment of the user. For the first and fourth tweet it extracts the sentiment of happiness and for the second and third the sentiment of anger and disgust. These results are consistent with the common perception.

B. Sentiments of Hashtags

Table III presents the total results for some of the hashtags tested.
These categories were selected because:

- They contain sufficient number of data (over 1000 tweets)
- They concern issues on which users may have expressed strong feelings.
- Their results can be evaluated based on common sense and experience.

Table III
Hashtags Ratings

#	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

We observe that the algorithm constructed is able to extract a result for the emotional content of the thematic categories which again corresponds to our intuition. Indeed, categories such as Football World Cup (# wc14gr), Summer Everywhere (#kalokairipantou) and Eurovision (#eurovisiongr) result in a happy feeling, as opposed to political issues such as the Parliament (#vouli), the closure of ERT (#ert) and the issue at Skouries (#skouries), where we observe higher value for the feelings of anger and disgust.

Compared to the various hashtags we can see that the sentiment of sadness and fear get great values for the hashtag # panellinies2014, which is a result that matches the topic.

C. Change of Hashtag sentiments over time

In this subsection we attempt to export sentiment of users over time on two characteristic hashtags for one day intervals. Here are the overall results for the sentiments of happiness and anger:

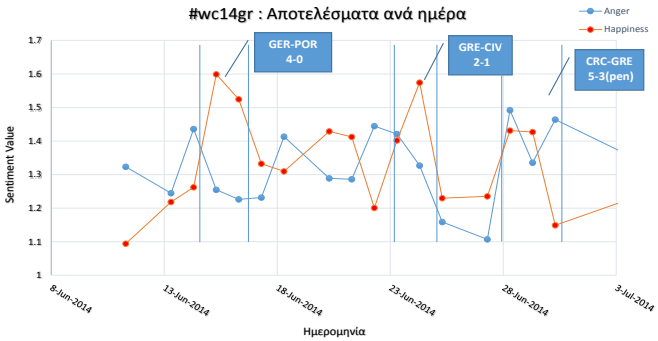

Figure 4. #wc14gr: Αποτελέσματα ανά ημέρα

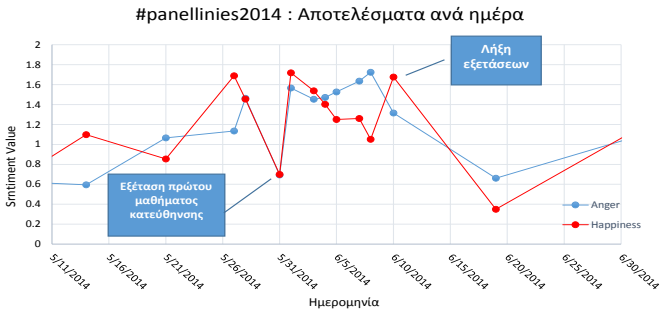

Figure 5. #panellinies2014: Αποτελέσματα ανά ημέρα

Regarding the experiments which examine the hashtags over time, we see that they are able to detect peaks in emotion values that can be associated with current events. For example, the positive result (for Greece) of the football match between Greece and Ivory Coast coincides with great happiness values and small values in anger. Even the game between Germany and Portugal, which attracted the interest of the Greek public displays great happiness values something that is apparent when we examine the tweets relevant to this event.

Finally, in the case of national exams, we can detect low values in both emotions measured before examining of the admittedly more difficult courses, and high values in the sentiment of joy on the day of the exam expiry.

V. Remarks

During the writing of this paper we made the following observations.

- This approach is not able to assess tweets that contain sarcastic comments and ambiguities, both of which can be found in abundance in Twitter, but only tweets with clear emotional content.
- It is observed that pairs of emotions like Anger - Disgust and Happiness - Surprise, indicated in Table IV with bold letters, receive similar values for the same categories, so they cannot be distinguished. We believe that this phenomenon is due to the large degree of correlation these sentiments have pairwise. This is evident in the table illustrating the values of the metric Pearson Correlation between all pairs of all emotions discussed.

TABLE IV
SENTIMENT PEARSON CORRELATION

	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	

To calculate the values of the above table, we form a vector containing the values of a particular emotion for each dictionary entry.

$$\overline{S} = [s_1 \quad s_2 \quad \dots \quad s_N]$$

Where $\overline{S}$ is a vector for each sentiment, s_i the value of the sentiment S for each entry i and N is the number of entries. The Pearson Correlation Coefficient can be calculated with the following formula:

$$\rho_{S_1,S_2} = \frac{\sum_{i=1}^{n}(s_{1_i} - \overline{s_1})(s_{2_i} - \overline{s_2})}{\sqrt{\sum_{i=1}^{n}(s_{1_i} - \overline{s_1})^2}\sqrt{\sum_{i=1}^{n}(s_{2_i} - \overline{s_2})^2}}$$

- An interesting observation can be made in the daily results for the hashtag #wc14gr. The feeling of happiness in Figure 4 seems to have inverse changes to the emotion of anger. Contrariwise, in the case of the hashtag #panellinies2014 fluctuations exhibit greater similarity. Generally, we can say that in the case of a football cup these sentiments do not manifest simultaneously, while in the occasion of national exams it is reasonable to observe mixed sentiment for the same time intervals.

- The dictionary which we used is not designed in a way that the entries coincide with the way the average user expresses himself through the social networks. It contains a large amount of entries that do not frequently appear in the tweets so it may not be the most ideal for this job. We measured that only 11.7% of the words that we examined are contained in the dictionary. However, the method proposed seem to work sufficiently.

- We generally observed that the sentiment of fear and sadness are receiving smaller values than the other emotions. This can be both because they are linguistically more difficult to identify through the colloquial language of the Internet, as well as the fact that the average user does not prefer to express such feelings in social media.

- All results are evaluated based on common sense and experience, since we are not able to calculate metrics for our results. This is due to the lack of a subjective sentiment evaluation of our data by independent users. Such an evaluation is generally difficult to be created.

VI. Conclusions – Suggestions for Improvements

The procedure that we propose, provides encouraging results and we can say that it is possible to extract the users feeling over different hashtags using a sentiment lexicon. Our results seem to be more accurate concerning Anger and Happiness.

As potential improvements of our method we propose the following:

- Use of a dictionary specialized for web applications.
- Utilization of linguistic data such as the part of speech that each entry is.
- Usage of tweets that contain Greek language written in Latin characters (greeklish).
- Creation of a testing set, with tweets that have been evaluated about their emotional content from independent raters. That way it will be possible to determine the effectiveness of our proposed technique using metrics.
- Expansion in a real time application so that it's possible to extract results on current events.

VII. Acknowledgments

We are mostly thankful to Avi Arampatzis, assistant professor of Electrical and Computer Engineering at Democritus University of Thrace, for his overall guidance and consultation.

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.

[8] Stop-words Version1.00 (20021106) Author: Dr. Holger Bagola (DIR-A/Cellule "Formats" 1 List of stopwords(ref. EURODICAUTOM, CELEX) Link: http://www.translatum.gr/forum/index.php?topic=3550.0

[9] Pang, B., and Lee, L. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2):1–135.

[10] Pak, A., and Paroubek, P. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In Proc. of LREC.

[11] Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg!. ICWSM, 11, 538-541