# Sentiment Analysis of Greek Tweets and Hashtags using a Sentiment Lexicon

Georgios Kalamatianos Dimitrios Mallis Symeon Symeonidis Avi Arampatzis
Department of Electrical and Computer Engineering
Democritus University of Thrace
Xanthi 67100, Greece
{georkala3,dimimall1,ssymeoni,avi}@ee.duth.gr

## ABSTRACT
The rapid growth of social media has rendered opinion and sentiment mining an important area of research with a wide range of applications. We focus on the Greek language and the microblogging platform "Twitter", investigating methods for extracting sentiment of individual tweets as well population sentiment for different subjects (hashtags). The proposed methods are based on a sentiment lexicon. We compare several approaches for measuring the intensity of "Anger", "Disgust", "Fear", "Happiness", "Sadness", and "Surprise". To evaluate the effectiveness of our methods, we develop a benchmark dataset of tweets, manually rated by two humans. Our automated sentiment results seem promising and correlate to real user sentiment. Finally, we examine the variation of sentiment intensity over time for selected hashtags, and associate it with real-world events.

## Categories and Subject Descriptors
H.3.3 Information [Search and Retrieval]: Retrieval models, Search process and Selection process

## General Terms
Algorithms, Measurement, Human Factors.

## Keywords
Sentiment Mining, Social Media, Twitter, Sentiment Lexicon.

## 1. INTRODUCTION
Users' disposition towards topics of interest constitutes a valuable piece of information that has social as well as financial implications. The rapid increase in usage of social media has rendered opinion and sentiment mining a promising area of research, as there is a growing interest in extracting information about what people think regarding various products, services, political issues, etc. The microblogging platform Twitter is especially appropriate for opinion mining and sentiment analysis, as it contains mostly textual information (very few other media), which is publicly available, and is therefore popular in related research. Additionally, the platform's international popularity and wide use of languages, allows researchers to investigate mining methods for different languages.

To our knowledge, the Greek language has not been examined sufficiently in tasks related to sentiment analysis. This seems to be mainly due to a shortage of appropriate datasets; sentiment-annotated datasets specialized in the Greek language have not yet been publicly available. Our goals in this paper are the following:

- To create a benchmark dataset with Greek tweets, along with a set of manually rated tweets for their sentiment intensity, and make it publicly available.
- To develop an automated method for determining the sentiment intensity of Greek tweets, for the six following dimensions: "Anger", "Disgust", "Fear", "Happiness", "Sadness", and "Surprise".
- To develop an automated method for determining the sentiment rating of different topics (hashtags) for the six aforementioned dimensions, based on tweet sentiment.
- To examine temporal aspects of sentiments, such as changes in their intensity for certain hashtags over time.

The automated sentimental rating of tweets is accomplished using a Greek Sentiment Lexicon [1].

The benchmark Greek dataset we contribute could be a valuable resource for future research and is available at: http://hashtag.nonrelevant.net/downloads.html. The automated tweet sentiment ratings are a direct result of calculations derived from the words occurring in the tweet, without using classification algorithms. Similarly, the automated hashtag sentiment ratings are derived from the ratings of tweets where the hashtag occurs in. Thus, the proposed methods are efficient and fairly simple to implement, and they can be used to provide baseline performance for future experimentation with the dataset. Finally, we present an examination of temporal aspects for the sentiment of "Happiness" and associate it with events that provoked intense emotions to the Greek population.

The rest of this paper is organized as follows. Related research is given in Section 2. Section 3 describes the benchmark dataset we developed. The sentiment lexicon and our methods are described in Section 4. In Section 5 we present experiments evaluating the proposed methods. In Section 6 we attempt to examine changes of sentiment over time. Conclusions and directions for future research are given in Section 7.


*PCI 2015*, October 01-03, 2015, Athens, Greece
 ISBN 978-1-4503-3551-5/15/10…$15.00
DOI: http://dx.doi.org/10.1145/2801948.2802010

## 2. RELATED RESEARCH

A dataset of Greek tweets was developed by The Social Sensor Project [2], in an attempt to predict the result of the upcoming elections. The dataset is of a relatively small size (368,547 tweets) since the collected tweets spanned a small period of time just before the elections and were only of political interest, thus, domain-specific. The dataset we develop is of a more general interest and wider timespan, resulting in an order of magnitude larger dataset (4,373,197 tweets).

Paltoglou & Buckley [3] extended TREC's Microblog dataset with manual subjectivity annotations about the relevance assessment and discussed issues like inter-annotator agreement and the distribution of subjective tweets in relation to topic categorization. The resulting benchmark dataset consists of 2,389 tweets that were annotated by multiple humans and 75,761 tweets that were annotated by one annotator according to the subjectivity of their relevance assessments, contrary to our benchmark dataset which focuses on the sentiment intensity of the tweets.

As far as sentiment analysis is concerned, an early approach is the "Affective Text", namely the sentiment analysis of segments of text. This method was used in SemEval-2007 [4] for determining the sentiment evoked in readers by different news headlines for the six sentiments that we also examine in this paper. Also, Pang and Lee [5] presented an extensive overview of the problem in 2008.

Most approaches for Twitter, use classification algorithms. Pak & Paroubek [6] use tweets containing emoticons to attribute sentiment ratings to the words they contain, so as to build a training dataset. The collection of tweets was gathered from Twitter accounts of newspapers (e.g. New York Times) and the classification is achieved using the Naïve Bayes algorithm. Kouloumpis, Wilson, & Moore [7] assumed that words that appear in tweets containing certain hashtags have a distinct sentiment value. For example, a highly-rated positive sentiment is attributed to the words that appear in hashtag #thingsilike. In accordance with the above hypothesis, they train an AdaBoost classifier.

The Social Sensor project [2] used sentiment analysis on tweets, among other methods, to predict the results of the upcoming elections. They used a sentiment lexicon in order to assign a polarity value (positive or negative sentiment) to every tweet. Then, they combined these results through a fusion of various classification approaches, based on different features of tweets (emoticons, punctuation, repetitions, etc.). This approach is more related to ours in the sense that the sentiment intensity of the tweets is measured as a score/rating, but is focused on polarity values instead of multiple sentiments.

## 3. A NEW BENCHMARK DATASET

In this section we describe the process via which we gathered our main dataset and subsequently constructed our benchmark dataset.

### 3.1 Data Collection

The dataset was collected via the Streaming API of Twitter using the Python programming language. The approach we followed was a width-first search of the social graph of Twitter. Starting from a random user, we built a search list with the "following" users of the first and every subsequent user. We tested iteratively users contained in the list, collecting their tweets and the IDs of their "following" users. Some comments about the process:

- The selection of "following" users instead of "followers" was made to avoid, as much as possible, the frequent occurrence of public figures who are "followed" by a large number of users.
- We did not request every user's "following" users, firstly because the number of users is very large and the size of the list would increase significantly, and secondly because it leads to unnecessary requests to the API of Twitter, which is limited to 180 requests per 15 minutes for each application. We decided to build a search list in which we only append new users when its size is less than 10,000.
- The tweets that we gather contain at least 4 Greek Unicode characters, to ensure the usage of the Greek language, as the minimum size of entries in the sentiment lexicon is 4 letters. The data was collected in the course of one week due to the API limitations. For each user examined, only the 200 most recent tweets were recovered, including the timestamp of every one of them.
- We only add users who use Greek characters in their tweets, so as to limit our search to Greek users only.

Table 1 provides statistical information for our dataset and in Figure 1 we present a cloud with the 100 most popular hashtags where the most frequently appearing hashtags are displayed in a larger font.

**Table 1. Dataset**

| | |
|---|---|
| **Dataset Size** | 832.1 MB |
| **Number of Tweets** | 4,373,197 |
| **Number of Users** | 30,778 |
| **Number of Hashtags** | 54,354 |
| **Hashtags (more than 1000 tweets)** | 41 |
| **Time Span** | 2-04-2008 until 29-11-2014 |


**Figure 1. The 100 most popular hashtags**

### 3.2 Sentiment Judgments

In order to evaluate the results of our methods (described in Section 4), we asked two volunteers (undergraduate students of Democritus University of Thrace) to manually rate a sample of tweets from our dataset. They rated each tweet (scale 0-5) for the six different sentiments examined in this paper ("Anger", "Disgust", "Fear", "Happiness", "Sadness" and "Surprise"), judging the sentiment that the user expresses through the tweet. The evaluation set that we created consisted of 681 tweets chosen randomly from 10 specific hashtags presented in Table 2.

**Table 2. Number of tweets per hashtag in the evaluation set**

| | | | | | |
|---|---|---|---|---|---|
| #wc14gr | 344 | #gogreece | 22 | #tedxath | 35 |
| #kalokairipantou | 55 | #gre | 30 | #feelfantastic | 35 |
| #skouries | 58 | #dwts | 35 | #stinigiamas | 35 |
| #panellinies2014 | 42 | | | | |

In order to assess the validity of our evaluation set, we calculated the inter-rater agreement between the two volunteers using Pearson's linear correlation coefficient. We selected Pearson's co-efficient instead of Cohen's or Fleiss' kappa (both more "standard" for measuring inter-rater agreement) because it is scaling and shift invariant, thus, helping to remove individual user biases. The results appear in Table 3:

**Table 3. Inter – Rater Correlation**

| | **Ang.** | **Disg.** | **Fear** | **Hap.** | **Sad.** | **Surp.** |
|---|---|---|---|---|---|---|
| **Rat.** | 0.064 | -0.034 | 0.415 | 0.477 | 0.530 | 0.398 |

We observe a fair/moderate inter-rater correlation for the sentiments: "Fear", "Happiness", "Sadness" and "Surprise". However, the sentiments "Anger" and "Disgust" present no correlation and are, therefore, rendered useless to evaluate our methods. We may attribute this phenomenon to the large amount of sarcastic tweets that can be perceived as either angry/hateful or cheerful/playful from different raters. Consequently, in the evaluation experiments that follow in Section 5, we focus on the aforementioned four sentiments that users agree in order to compare with our results.

# 4. SENTIMENT OF TWEETS / HASHTAGS

In this section we describe the sentiment lexicon, the preprocessing we did on our data, and present the aspects of the methodology we propose for our experiments.

## 4.1 Sentiment Lexicon

The sentiment lexicon that was used in this paper is the Greek Sentiment Lexicon [1], which contains 2,315 entries evaluated for the following six sentiments: "Anger", "Disgust", "Fear", "Happiness", "Sadness" and "Surprise". The entries of the sentiment lexicon were gathered through crawling, using the advanced search utilities of the electronic version of the Greek dictionary by Triantafyllides [8]. These specific entries were chosen using metadata that contain information concerning either the tone of the word (ironic, meiotic, abusive, mocking or vulgar) or the emotional content of their description (feel, love, etc.).

The dictionary includes emotional evaluation of entries by four independent raters who were asked to rate each entry according to the possibility of it expressing the corresponding sentiment. The lexicon 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.

In order to determine the agreement between the raters for each pair of raters we again used Pearson's correlation coefficient.

**Table 4. Lexicon Inter-Rater Correlation**

| | **Ang.** | **Disg.** | **Fear.** | **Hap.** | **Sad.** | **Surp.** |
|---|---|---|---|---|---|---|
| **Raters 1-2** | 0.345 | 0.378 | 0.333 | 0.318 | 0.349 | 0.206 |
| **Raters 1-3** | 0.650 | 0.701 | 0.611 | 0.780 | 0.604 | 0.566 |
| **Raters 1-4** | 0.474 | 0.444 | 0.320 | 0.449 | 0.460 | 0.270 |
| **Raters 2-3** | 0.365 | 0.447 | 0.358 | 0.346 | 0.379 | 0.290 |
| **Raters 2-4** | 0.445 | 0.532 | 0.294 | 0.462 | 0.460 | 0.371 |
| **Raters 3-4** | 0.567 | 0.542 | 0.335 | 0.476 | 0.456 | 0.325 |

Rating individual tweets may sound like an easier task than rating individual words, as in the former task there is a context while in the latter there is not. Nevertheless, it does not appear to be so in our experiments, as there is a fair correlation for all pairs of raters (Table 4), contrary to the manual evaluation of our benchmark dataset which presented no correlation for the sentiments "Anger", "Disgust" (Table 3). Although our remark in Section 3.2 about sarcastic tweets may be valid, this remains an interesting observation for further investigation.

We also examined the pairwise correlation of the sentiments in the terms of the lexicon (Table 5) and observed that there exist highly correlated pairs: Anger/Disgust and Happiness/Surprise. This characteristic will influence the results of the examined methods, as we explain in following sections.

**Table 5. Pearson Correlation of Sentiment Pairs**

| | **Ang.** | **Disg.** | **Fear** | **Hap.** | **Sad.** | **Surp.** |
|---|---|---|---|---|---|---|
| **Ang.** | | 0.827 | 0.500 | 0.002 | 0.384 | 0.465 |
| **Disg.** | 0.827 | | 0.427 | -0.105 | 0.370 | 0.403 |
| **Fear** | 0.500 | 0.427 | | 0.205 | 0.530 | 0.549 |
| **Hap.** | 0.002 | -0.105 | 0.205 | | 0.196 | 0.558 |
| **Sad.** | 0.384 | 0.370 | 0.530 | 0.196 | | 0.425 |
| **Surp.** | 0.465 | 0.403 | 0.549 | 0.558 | 0.425 | |

Another trait of this lexicon is that it is not designed in a manner that its entries coincide with the way the users express themselves through social networks. It contains a large amount of entries that do not frequently appear in 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.

## 4.2 Data Preprocessing

We applied some preprocessing steps to our data. Specifically:

- We divided the tweets in files according to their hashtags. We also merged similar hashtags by removing non-alphanumeric characters, and lowercasing everything. For example, the hashtags #wcgr14 and #WCgr14 were grouped in the same category.
- We chose to examine only the hashtags appearing 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 into more than one hashtag.
- We chose to keep reposted tweets from other users (retweets) because we assume that they agree with the sentiment expressed by these users.

- We removed a list of Greek stop-words from our data, to reduce the size and computational work
- We replaced intonated characters with non-intonated, and turned every letter to uppercase so that the dataset has the same formatting as the dictionary.
- We applied a Greek stemmer [9] to both the data and the dictionary to increase the matching of the words.

## 4.3 Methods for Tweet Sentiment Rating

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

$$\overline{W_{i,j}} = [\mathrm{w}_{1,j} \quad \mathrm{w}_{2,j} \quad \mathrm{w}_{3,j} \quad \mathrm{w}_{4,j} \quad \mathrm{w}_{5,j} \quad \mathrm{w}_{6,j}]$$

where $\mathrm{j} = 1...\mathrm{N}$ and $\mathrm{N}$ is the number of entries that are identified in the tweet, and also a 6 component vector $\overline{T}$

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

for every tweet.

Each component of $\overline{T}$ is calculated with the following 4 formulas that we examine in this paper, where $i$ is the number of components of vector $\overline{T}$.

**Table 6. Formulae for Tweet Sentiment Rating.**

| Formula 1: Arithmetic Mean | Formula 2: Quadratic Mean |
|---|---|
| $t_i = \frac{\sum_{j=1}^{N} w_{i_j}}{N} \; i = 1...6$ | $t_i = \sqrt{\frac{\sum_{j=1}^{N} w_{i_j}^2}{N}} \; i = 1...6$ |
| Formula 3: Maximum | Formula 4: CombMNZ |
| $t_i = \max_{j=1:N}(w_{i_j}) \; i = 1...6$ | $t_i = N\sum_{j=1}^{N} w_{i_j} \; i = 1...6$ |

Formula 1 is simply the arithmetic mean of entries $\overrightarrow{W}$ that were identified in each tweet. In Formula 2 we use the quadratic mean which is selected 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. Another approach is Formula 3, where we assign the maximum sentiment found in the words contained in the tweet, to the whole tweet. We base this on the assumption that the dominant sentiment of a tweet is expressed in the words with the highest sentiment intensity. Finally, Formula 4 (CombMNZ [10]) is a method mainly used in data fusion and returns a higher value for the tweets that contain multiple words with high intensity in a particular sentiment.

Figure 3 presents an example of hand-picked tweets, followed by loose translations.

The ratings we calculated through Formula 2 (quadratic mean) for the tweets of Figure 2 are the following:

***#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.]*

***#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 2. Example tweets**

**Table 7. Example Ratings**

| # | Ang. | Disg. | Fear | Hap. | Sad. | Surp. |
|---|---|---|---|---|---|---|
| **1** | 1.00 | 1.00 | 1.00 | 4.75 | 1.00 | 2.75 |
| **2** | 1.00 | 1.00 | 1.00 | 3.75 | 1.00 | 2.50 |

We can see in Table 7 that the algorithm agrees with our intuition and returns "Happiness" as the dominant sentiment for these tweets. A more thorough evaluation follows in Section 5.

## 4.4 Methods for Hashtag Sentiment Rating

In the next step we combine the tweet vector components $t_j$ for every hashtag $\overline{H}$ using both the arithmetic and the quadratic mean as described in the previous subsection. We rejected the maximum formula, as its results would only depend on the most sentimental tweet, not taking into account the rest of the data. We also rejected CombMNZ as it creates an unfair bias towards hashtags with a larger number of tweets.

To better demonstrate our method, we present, in Table 8, the overall results for some of the examined hashtags, using the quadratic mean for both the individual tweets and the hashtag ratings. We present the results of our method for the sentiments "Anger", "Disgust" as an example, although we are not able to evaluate them due to the lack of correlation in our rater judgments (Table 3).

**Table 8. Hashtag Ratings**

| Hashtag | Ang. | Disg. | Fear | Hap. | Sad. | Surp. |
|---|---|---|---|---|---|---|
| **#wc14gr** | 1.3910 | 1.2862 | 0.9512 | 1.3604 | 0.8412 | 1.4552 |
| **#kalokairipantou** | 0.7930 | 0.9158 | 0.7739 | 2.1856 | 0.7570 | 2.1084 |
| **#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 |
| **#eurovisiongr** | 1.3464 | 1.2957 | 0.7933 | 1.3533 | 0.7599 | 1.4092 |

We observe that the proposed algorithm is able to extract a result for the sentimental 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 happy sentiments, as opposed to political issues such as the Parliament (#vouli) and the shutdown of the national radio and TV broadcaster (#ert ), where we observe higher ratings for the sentiments of "Anger" and "Disgust". We can also see that the national exams for university entry (#panellinies2014) receive higher ratings for the sentiments

of "Sadness" and "Fear" compared to other hashtags. We generally observe that the sentiments "Sadness" and "Fear" receive smaller ratings than the other sentiments. This is characteristic for the lexicon, as these two sentiments receive lower values on average, compared to the others.

# 5. EXPERIMENTS

In this section we present experiments we performed to evaluate our methods, for both individual tweet and the hashtag ratings.

## 5.1 Evaluation of Rating Individual Tweets

In order to examine and compare the accuracy of the four proposed formulae for rating tweets, we calculate both the Pearson and the Kendall correlation between the ratings of the algorithm and those of the human raters. The results below and also in the next subsection only concern the evaluated tweets which contained terms of the sentiment lexicon (432 tweets). Also, due to the absence of correlation between our raters for the sentiments of "Anger" and "Disgust", the results for these sentiments are not presented here since they cannot be evaluated with this benchmark dataset. For the remaining four sentiments with higher correlation we chose to use the average of the four (lexicon) raters for every individual term, as allowed by the fair inter–rater correlation, calculated in Table 4.

**Table 9. Pearson/ Kendall Correlation**

| Formula | Fear | Happiness | Sadness | Surprise |
|---|---|---|---|---|
| Mean | 0.10 / 0.02 | 0.26 / 0.22 | 0.01 / 0.01 | -0.04 / -0.06 |
| Quad | 0.13 / 0.04 | 0.26 / 0.20 | 0.04 / 0.03 | -0.04 / -0.06 |
| Max | 0.15 / 0.10 | 0.20 / 0.17 | 0.09 / 0.09 | -0.01 / -0.03 |
| Comb | 0.02 / 0.05 | 0.05 / 0.07 | 0.08 / 0.09 | 0.04 / 0.01 |

For the cases of "Sadness" and "Surprise", the results are not acceptable and these two sentiments we won't be examined in the next subsection. This may be due to the high correlation pairs of Fear/Sadness and Happiness/Surprise for the terms of the lexicon which are not correlated in individual tweets (Happiness/Surprise correlation in individual tweets: -0.1329). Also, in an attempt to further examine our results, we calculate Kendall's rank-correlation coefficient in Table 10, to determine whether there is maybe a non-linear relation. As we can see, the results remain similar to Pearson Correlation.

Our method reaches a fair correlation value for the sentiment of "Happiness" and acceptable correlation value for the sentiment of "Fear". These results can be strengthened, considering the fact that we can reach almost half the correlation of our raters (Table 3) for the aforementioned sentiments.

We observe that different formulae are appropriate for different sentiments. For example, the arithmetic mean returns the best results for the sentiment of "Happiness", but seems to fail in the case of "Fear" where the arithmetic mean performs better. Finally, the CombMNZ does not give promising results in terms of correlation and it will not be further examined in this work.

## 5.2 Evaluation of Rating Hashtags

To evaluate the performance of our methods on hashtags, we calculated the sentiment ratings for the 6 hashtags contained in the benchmark dataset. These ratings are calculated with the two formulae described in Section 4.4 for both the manually rated tweets and the tweets rated by our method. Consequently, we calculate Pearson and Kendall correlation of each sentiment for the 6 different hashtags of Table 8.

**Table 10. Pearson/ Kendall Correlation for Hashtag ratings**

| | Arithmetic | | Quadratic | |
|---|---|---|---|---|
| | Fear | Happiness | Fear | Happiness |
| Mean | 0.24 / 0.28 | 0.90 / 0.78 | 0.26 / 0.11 | 0.80 / 0.38 |
| Quad | 0.36 / 0.32 | 0.87 / 0.78 | 0.35 / 0.24 | 0.77 / 0.33 |
| Max | 0.47 / 0.46 | 0.77 / 0.51 | 0.40 / 0.20 | 0.59 / 0.24 |

The arithmetic mean seems to be better for the task of hashtag rating. Especially in the case of "Happiness", which proves to be the easiest sentiment to detect in both of our experiments, the correlation reaches a value of up to 0.90 in Pearson. Also, for the sentiment of "Fear", despite the slightly weak results of Table 10, we get a fair correlation of almost 0.5.

Our methods seem to perform well through accumulation of a large amount of data and produce more accurate results than individual tweets.

# 6. HASHTAG SENTIMENT OVER TIME

After evaluating our methods, we applied it to examine the change of sentiment over time. We choose the sentiments of "Anger" along with "Happiness" because we can associate its changes with current events. To calculate the sentiment intensity, we use the quadratic mean both for individual tweets and for the accumulation of groups of tweets. The tweets of the hashtags we chose to examine, were sorted in an ascending order of time. Based on the method we proposed in the previous section, we calculate the average sentiment for one-day intervals. We chose to examine only days for which we have gathered more than 60 tweets, in order to have more conclusive results.

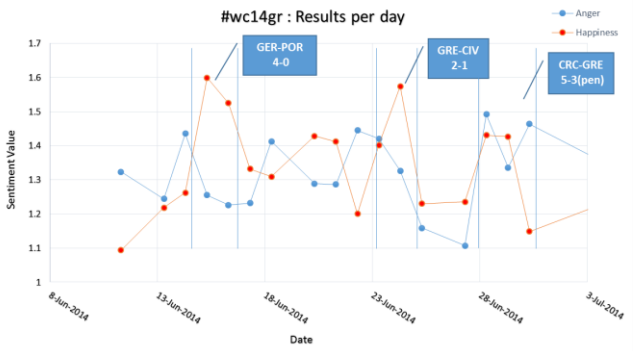


**Figure 3. #wc14gr: Results per day**

Figure 3 depicts sentiment changes for the football Word Cup '14 hashtag over time. We see that they are able to detect peaks in sentiment ratings that can be associated with current events. For example, the positive result (for the Greek fans) of the football match between Greece and Ivory Coast coincides with high ratings in happiness and low ones in anger. Also, the game between Germany and Portugal, which attracted the interest of the Greek public, displays high ratings in happiness. This is apparent when we examine the tweets relevant to this event.

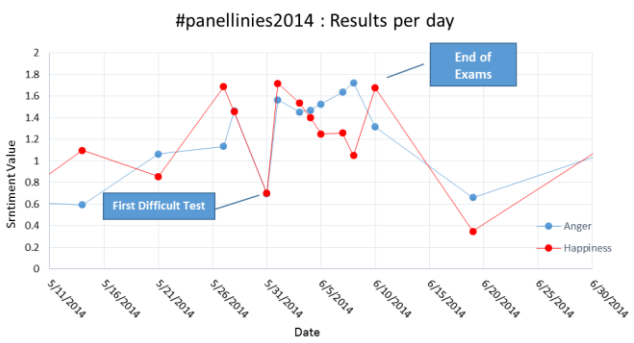


**Figure 4. #panellinies2014: Results per day**

Finally, in the case of national exams (#panellinies2014, Figure 4), we can detect low ratings in both sentiments measured before examining of the admittedly more difficult courses, and high values in the sentiment of happiness on the day of the exam completion.

An interesting observation can be made in the daily results for the hashtag #wc14gr. The sentiment of happiness in Figure 3 seems to have inverse changes to the sentiment 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.

## 7. CONCLUSIONS AND FUTURE WORK

Automated opinion and/or sentiment mining is a very promising topic with potential applications in social, political, marketing, financial, and other fields. We examined different methods to extract ratings for individual tweets as well as hashtags, based on a sentiment lexicon. The methods we propose, provided promising results.

Our approach uses direct calculations to aggregate ratings and can be implemented with a fairly low computational cost. These initial experiments led to interesting remarks which could guide further investigation and improvements, such as:

- The sentiment "Happiness" seems to be the easiest one to detect throughout all the sentiments in both of our experiments. For the sentiment of "Fear", results are also promising.
- Different formulae appear to perform best for different sentiments. For example, Formula 3 (max) returns better results than the others, for the sentiment of "Fear". Formulae 1, 2, on the other hand, perform better for the sentiment of "Happiness".
- The presence of a large amount of tweets leads to a better assessment of the overall sentiment of the whole set, through the methods that we described, even in the cases where the individual tweet ratings do not appear as accurate.
- As presented in Section 6, we may also be able to detect changes in sentiment over time and the results coincide with our intuition about real world events.

Furthermore, our dataset of tweets together with the manual user ratings are publicly available at http://hashtag.nonrelevant.net/downloads.html, a resource which could prove valuable for other researchers.

As potential improvements of our methods or directions for further research, we propose the following:

- Use/development of a dictionary specialized for web applications, in order to increase the matching terms between the lexicon and the tweets.
- Utilization of linguistic data such as the part of speech that each entry is, and inclusion of other features of tweets such as emoticons, punctuation marks, etc.
- Extension of the benchmark dataset both in size and in number of raters, in order to evaluate the performance of our methods for the sentiments "Anger", and "Disgust".
- Further examination of the observation that individual tweets seem to be more challenging than word judging, a fact that does not coincide with our intuition.
- Further examination of changes in sentiment over time the method of which, is not evaluated in the present work.

We have examined the topic of Sentiment Analysis using a sentiment lexicon, providing a benchmark resource/dataset together with baseline performance of several simple and efficient algorithms. We hope that all these will be proven valuable for us and the community to build upon in future work.

## 8. ACKNOWLEDGMENTS

We thank Dimitrios Nikolaras (undergraduate student) at Democritus University of Thrace, for his overall contribution, as well as Anastasia Filippou and Aristides Pantopikos, undergraduate students at Democritus University of Thrace, for their contribution in the manual evaluation of tweets.

## 9. REFERENCES

[1] Tsakalidis, A., et al. 2014. An Ensemble Model for Cross-Domain Polarity Classification on Twitter. *15th International Conference, Thessaloniki, Greece, Proceedings, Part II, (October 12-14, 2014,)* 168-177. DOI: 10.1007/978-3-319-11746-1_12.

[2] Burnside, G., Papadopoulos, S., and Petkos, G. 2014. D2.3 Social stream mining framework. *Social Sensor, Sensing User Generated Input for Improved Media Discovery and Experience. FP7-287975.*

[3] Paltoglou, G., and Buckley. K. 2013. Subjectivity annotation of the Microblog 2011 Realtime Adhoc relevance judgments. *ECIR 2013: 35th European Conference on Information Retrieval, pages 344-355.*

[4] Strapparava, C., and Mihalcea, R. 2007. Affective Text. *SemEval-2007 Task 14.*

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

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

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

[8] Triantafyllides G. 1998. Dictionary of Standard Modern Greek. *Institute for Modern Greek Studies of the Aristotle University of Thessaloniki.*

[9] Ntais, G., 2006. *Development of a Stemmer for the Greek Language*, Master Thesis. Stockholm University / Royal Institute of Technology, Department of Computer and Systems Sciences,

[10] Fox, E. A. and Shaw, J. Combination of multiple searches. *Second Text REtrieval Conference (TREC-2),* (Gaitherburg, MD, USA, August 1994), 243–252.