Towards the Creation of an Emotion Lexicon for Microblogging

Article *in* Journal of Systems and Information Technology · May 2018

DOI: 10.1108/JSIT-06-2017-0040

4 authors:

Georgios Kalamatianos
Uppsala University
10 PUBLICATIONS **77** CITATIONS
SEE PROFILE

Symeon Symeonidis
Democritus University of Thrace
34 PUBLICATIONS **495** CITATIONS
SEE PROFILE

Dimitrios Stefanos Mallis
University of Nottingham
6 PUBLICATIONS **34** CITATIONS
SEE PROFILE

Avi Arampatzis
Democritus University of Thrace
117 PUBLICATIONS **1,944** CITATIONS
SEE PROFILE

Towards the creation of an emotion lexicon for microblogging

Georgios Kalamatianos, Symeon Symeonidis, Dimitrios Mallis and Avi Arampatzis
Department of Electrical and Computer Engineering, Democritus University of Thrace, Xanthi, Greece

Received 1 June 2017
Revised 6 September 2017
2 February 2018
Accepted 10 April 2018

Abstract

Purpose – The rapid growth of social media has rendered opinion and sentiment mining an important area of research with a wide range of applications. This paper aims to focus on the Greek language and the microblogging platform Twitter, investigating methods for extracting emotion of individual tweets as well as population emotion for different subjects (hashtags).

Design/methodology/approach – The authors propose and investigate the use of emotion lexicon-based methods as a mean of extracting emotion/sentiment information from social media. The authors compare several approaches for measuring the intensity of six emotions: anger, disgust, fear, happiness, sadness and surprise. To evaluate the effectiveness of the methods, the authors develop a benchmark dataset of tweets, manually rated by two humans.

Findings – Development of a new sentiment lexicon for use in Web applications. The authors then assess the performance of the methods with the new lexicon and find improved results.

Research limitations/implications – Automated emotion results of research seem promising and correlate to real user emotion. At this point, the authors make some interesting observations about the lexicon-based approach which lead to the need for a new, better, emotion lexicon.

Practical implications – The authors examine the variation of emotion intensity over time for selected hashtags and associate it with real-world events.

Originality/value – The originality in this research is the development of a training set of tweets, manually annotated by two independent raters. The authors "transfer" the sentiment information of these annotated tweets, in a meaningful way, to the set of words that appear in them.

Keywords Social media, Sentiment mining, Emotion lexicon

Paper type Research paper

1. Introduction

The disposition of users toward 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 growing interest in extracting information about what people think regarding various products, services, public figures, political issues and many other things (Medhat *et al.*, 2014).

In the past, there has been a fair amount of research in the task of sentiment analysis on data originating from product reviews, news articles, blogging, etc. (Liu and Zhang, 2012; Hu and Liu, 2004). 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 allows researchers to investigate mining methods for different languages (Giachanou and Crestani, 2016). However, the

Journal of Systems and Information Technology
Vol. 20 No. 2, 2018
pp. 130-151

1328-7265
DOI 10.1108/JSIT-06-2017-0040

increasing popularity of microblogging has introduced new challenges in sentiment analysis, related to the informal tone used by its users, the increased variety of subjects referred to, the length constraints of the text, and the use of hashtags and emoticons (Giachanou and Crestani, 2016).

To our knowledge, the Greek language has not been examined sufficiently in tasks related to emotion analysis especially in relation to data from microblogging sources. This seems to be mainly due to a shortage of appropriate datasets. Emotion-annotated data sets in the Greek language have not yet been publicly available. An attempt to create appropriate resources for emotion evaluation in the Greek language was made by Tsakalidis *et al.* (2014) who created the first Greek Sentiment Lexicon (GSL). We use this lexicon for our research and improve upon it. Although Tsakalidis *et al.* (2014) have used the term Sentiment Lexicon, we are inclined to use the term emotion lexicon; this seems to be more appropriate, as we study six different emotions rather than positive versus negative tweets.

Our goals and contributions in this paper are the following:

- to create a benchmark data set with Greek tweets, along with a set of manually rated tweets for their emotion intensity, and make it publicly available;
- to develop automated methods for determining the emotion intensity of Greek tweets, for the six following emotions: anger, disgust, fear, happiness, sadness and surprise – the proposed methods are based on Greek emotion lexica;
- to develop automated methods for determining the emotion rating of different topics (hashtags) in the six aforementioned emotion dimensions, based on individual tweet emotion;
- to develop a new improved emotion lexicon, specialized for the task of sentiment analysis of Greek tweets that will enhance the performance of our tweet and hashtag evaluation methods; and
- to examine temporal aspects of emotions, such as changes in their intensity for hashtags over time.

This paper builds upon and extends the previous work of Kalamatianos *et al.* (2015); we can summarize the following contributions. We anonymize and make the benchmark Greek data set publicly available[1]; it could constitute a valuable resource for future research. The automated tweet emotion ratings are a direct result of calculations derived from the words occurring in the tweet, without using classification algorithms. Similarly, the automated hashtag emotion ratings are derived from the ratings of tweets where the hashtag occurs. 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 data set.

Also, as we show in the following sections, the emotion lexicon of Tsakalidis *et al.* (2014) is not specialized for emotion evaluation of internet related data, as its entries present a low matching rate to the terms appeared in the tweets. To deal with this issue, we develop a new emotion lexicon specifically for emotion analysis of Greek tweets that we will contribute as a resource for tasks related to emotion analysis in the Greek language (also online at the URL in previous page's Footnote). The new emotion lexicon, as we show later on, enhances the performance of our methods, bringing the results closer to the human intuition. Finally, we present an examination of temporal aspects for the emotion of *happiness* and *anger* and associate it with events that provoked intense emotions to the Greek population.

Traditional construction of sentiment and emotion lexicons are based on machine learning approaches where each term is represented with a binary label/polarity (Liu and Zhang, 2012; Chen and Skiena, 2014). A recent work (Xu *et al.*, 2013) has classified terms in

six different emotions. In this work, we aim to present the emotion intensity for each term by a scoring method for each emotion. This approach is a fairly new area of research in which we introduce some novelties, especially for the study of the Greek language.

The rest of this paper is organized as follows. Related research is given in Section 2. Section 3 describes the benchmark data set we developed. The emotion lexicon of Tsakalidis *et al.* (2014) and our automated emotion rating methods are described in Section 4. In Section 5, we present experiments evaluating the proposed methods, and in Section 6, we present the New Greek Emotion Lexicon and compare it with the one of Tsakalidis *et al.* (2014). In Section 7, we attempt to examine changes of emotion over time. Conclusions and directions for future research are given in Section 8.

2. Related research

In this section, we examine the related research and give an overview of the available resources that may be useful for dealing with the issues we tackle in the paper.

2.1 Sentiment analysis

Many definitions have been given about opinion mining and sentiment analysis. Recently, according to Serrano-Guerrero *et al.* (2015), opinion is defined as:

> A positive or negative sentiment, view, attitude, emotion, or appraisal about an entity (product, person, event, organization or topic) or an aspect of that entity from a user or group of users.

Khan *et al.* (2016) also define sentiment analysis as a research area which explains and extracts the attitude of a speaker toward a specific subject. An early approach on sentiment analysis is the "affective text", namely, the sentiment analysis of segments of text. This method was used in SemEval-2007 by Strapparava and Mihalcea (2007) for determining the sentiment evoked in readers by different news headlines for the six emotions that we also adopt in this paper. Pang and Lee (2008) present an extensive overview of the problem. A common target of sentiment analysis is polarity classification, based on opinion findings and sentiment identification Medhat *et al.* (2014). Many efforts have been made to categorize text according to sentiment and emotion with usage of a variety of techniques; lexicon-based and machine learning-based are the commonest approaches. Actually, sentiments and emotions are very close. This view is confirmed by the strain of six universal emotions which is combined to the way of measuring the intensity of an opinion (Serrano-Guerrero *et al.*, 2015).

Most approaches on Twitter data, use classification algorithms. Pak and Paroubek (2010) use tweets containing emoticons to attribute sentiment ratings in the words they contain, so as to build a training data set. The collection of tweets was gathered from Twitter accounts of newspapers (e.g. New York Times) and the classification is achieved by using the Naive Bayes algorithm. Kouloumpis *et al.* (2011) assume that words occurring 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 the hashtag #thingsilike. Based on this hypothesis, they train an AdaBoost classifier.

Severyn and Moschitti (2015b), dos Santos and Gatti (2014) and Severyn and Moschitti (2015a) present other approaches for Twitter sentiment classification based on deep neural networks, and they use character to sentence-level information to achieve sentiment extraction from text.

Tsakalidis *et al.* (2015) used a sentiment lexicon-based analysis on tweets, among other methods, to predict the results of the 2014 European Union elections, by assigning a polarity value (positive or negative sentiment) to every tweet. Then, they combined these results

with a fusion of various classification approaches, based on different features of tweets (emoticons, punctuation, repetitions, etc.).

We propose a system that distinguishes itself among the above methods in the following ways:

- It uses intensity scores for different emotions rather than a polarity measure.
- It uses a calculation scheme that could easily be applied to streaming data, requiring no form of training.
- It is applied at a microblogging platform dataset, rather than full text.

On the other hand it suffers from:

- accuracy; and
- the availability of resources to improve accuracy.

2.2 Benchmark data sets for sentiment analysis

To our knowledge, the only available Greek data set of microblogging related data is the aforementioned one by The Social Sensor Project Tsakalidis *et al.* (2015). The data set is of a relatively small size (368,547 tweets), as the tweets were collected in a span of 48 days and were only of political interest, thus domain-specific. The data set developed in the present paper is of a more general interest and longer timespan, resulting in a data set size of a larger order of magnitude (4,373,197 tweets).

An evaluation set for sentiment analysis methods is made by Paltoglou and Buckley (2013), who extend TREC's Microblog data set with manual subjectivity annotations about the relevance assessment, and discuss issues like inter-annotator agreement and the distribution of subjective tweets in relation to topic categorization. The resulting benchmark data set is in the English language and 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. In contrast, our benchmark data set focuses on the sentiment intensity of the tweets for six different emotions. This is the only related resource in the Greek Language, and, to our knowledge, the only one in general.

2.3 Sentiment lexica

A sentiment or emotion lexicon is composed by terms with strong emotional connotation, gathered from a lexical resource. Commonly, lexicon-based methods start with a small set of words, and are in need of human annotation and statistical methods. Lexicon-based approaches use statistical or semantic methods to find sentiment polarity, and split into dictionary-based and corpus-based approaches. On the one hand, a dictionary-based approach uses a primary set of terms, which are collected and annotated manually. Following, by searching in a dictionary about synonyms and antonyms, this set is growing. On the other hand, a corpus-based approach begins from a set of seed opinion terms, and then by using statistical or semantic techniques, searches other opinion words with context specific orientations in an extensive corpus (Medhat *et al.*, 2014).

Many theories and lexica have been developed for English and other languages. The theories about six basic emotions (Ekman, 1992) and pleasure-arousal-dominance emotional state mode (Mehrabian, 1995) are the most used in the research area of sentiment analysis. The most well-known sentiment lexica are:

- SentiWordNet, in which three numerical scores join with each WordNet synset and compute its correlate sentiment (Baccianella *et al.*, 2010);
- the LIWC lexicon (Pennebaker *et al.*, 2003);
- the NRC Emotion Lexicon Mohammad and Turney (2010) which used emotion-word hashtags; and
- Bing Liu's Opinion Lexicon (Hu and Liu, 2004).

Sentiment lexica have been developed in many languages, such as Spanish (Molina-Gonzalez *et al.*, 2013), German (Clematide and Klenner, 2010), French (Rao and Ravichandran, 2009), Arabian (Abdul-Mageed *et al.*, 2011; Mahyoub *et al.*, 2014), Chinese (Lu *et al.*, 2010), Dutch (Vossen *et al.*, 2007), hybrid approaches (Lo *et al.*, 2016a) and multilingual approaches such as EuroWordNet which is a multilingual database with wordnets (Vossen and Letteren, 1997) and (Lo *et al.*, 2016b) for formal, informal and scarce resource languages.

As far as the Greek Language is concerned, the only available sentiment lexicon is the one which was created for the Social Sensor Project (Tsakalidis *et al.*, 2015). This sentiment lexicon is based on a Greek dictionary and does not have a great coverage of terms that are frequently used on the internet. In the present work, a New Greek Emotion Lexicon is developed and made publicly available. We show that it can perform better in identifying tweet and hashtag sentiment, as it is specializes on microblogging applications.

3. A benchmark Greek emotion tweet data set

In this section, we describe the process via which we gathered our main tweet data set and collected manual human emotion judgments to develop our benchmark data set.

3.1 Data collection

The data set was collected via the Twitter API 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 iteratively processed users contained in the list, collecting their tweets and the IDs of their "following" users. In more detail:

- 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, first because the number of users is very large and the size of the list would increase significantly, and second due to the limitation of Twitter's API which allows only up to 180 requests per 15 min for each application.
- To enforce the collection of Greek tweets, we discarded tweets which did not contain at least four Greek Unicode characters. Another reason for this is that the minimum size of entries in the Greek Emotion Lexicon is four letters. The data were 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 their timestamps.

Table I provides statistical information for the Greek Emotion Tweets Dataset (GSTD) after pre-processing. In Figure 1, we present a cloud with the 100 most popular hashtags; the more frequent the hashtag, the larger its font.

3.2 Emotion judgments

To develop an emotion ground-truth for the data set, we asked two volunteers (undergraduate students at Democritus University of Thrace) to manually rate a sample of tweets. They rated each tweet (scale 0-5) for the six different emotions considered in this paper (anger, disgust, fear, happiness, sadness and surprise), judging the emotion that the user expresses through the tweet. The evaluation set that we created consists of 691 tweets, chosen randomly from ten hashtags, which were among the 20 most popular ones but also shared a potential of emergent dispute. For the latter quality, the hashtags were selected manually. These hashtags are presented in Table II. The sample size for each individual hashtag is roughly proportionate to its total size. To assess the "strength" of our ground-truth, 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'

Number of tweets	4,373,197
Number of users	30,778
Number of hashtags	54,354
Hashtags (occurring in more than 1,000 tweets)	41
Timestamps timespan	02/04/2008 - 29/11/2014

Table I.
Greek emotion tweets data set (GSTD) statistics

Figure 1.
The 100 most popular hashtags

Hashtag	No. of tweets
#wc14gr	344
#skouries	58
#kalokairipantou	55
#panellinies2014	42
#dwts	35
#tedxath	35
#feelfantastic	35
#stinigiamas	35
#gre	30
#gogreece	22
Total	691

Table II.
Number of tweets per hashtag in the evaluation set

kappa (both more "standard" for measuring inter-rater agreement) because Pearson's is scale- and shift-invariant, thus helping to remove individual rater biases. The results appear in Table III. We observe a fair/moderate inter-rater correlation for the emotions of *fear*, *happiness*, *sadness* and *surprise*, as opposed to the emotions of *anger* and *disgust* which present no correlation. We may attribute the disagreement of raters for *anger* and *disgust* to the large amount of sarcastic tweets; many of these can be perceived as either angry/hateful or cheerful/playful. Consequently, in the evaluation experiments that follow in Section 5, we focus on the aforementioned four emotions that users agree most.

3.3 Data preprocessing

We pre-processed the data set as follows:

- We retained re-posted tweets from other users (re-tweets), assuming that the act of sharing a piece of text written by someone else presupposes sharing the same feelings for the topic.
- To have enough data to assess in each thematic category, we chose to examine only the hashtags appearing in over 1,000 tweets. 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 merged similar hashtags by removing non-alphanumeric characters, and lowercasing everything. For example, the hashtags #wcgr14 and #WCgr14 were grouped in to #wcgr14.
- Intonated characters were replaced with non-intonated, and turned every letter to uppercase so that the data set has the same formatting as the GSL.
- We removed Greek stopwords from our data, using the CELEX stop-list of 627 words Bagola (2004), to reduce the size of the dataset and computational work.
- We applied the Greek stemmer of Ntais (2006) to both the data and the dictionary to increase the number of matching words.

For a recent extensive comparison of pre-processing techniques for microblogging sentiment analysis, see Effrosynidis *et al.* (2017).

4. Methods for tweet and hashtag emotion

In this section, we present our methodologies for determining tweet and hashtag emotion. Hashtag emotion is based on the emotion of the tweets it occurs in. Similarly, tweet emotion is based on the emotion of its occurring words. For determining word emotion, we use a publicly available Greek emotion lexicon.

First, in Section 4.1, we give some more details for the lexicon and our own deeper analysis. Then, in Sections 4.2 and 4.3, we present our methods for tweet and hashtag emotion, respectively.

Table III. Inter-rater Pearson's correlation

–	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Correlation	0.064	−0.034	*0.415*	*0.477*	*0.530*	*0.398*

4.1 The Greek emotion lexicon

The lexicon used in this paper is the Greek Emotion Lexicon (GSL) Tsakalidis *et al.* (2014), which contains 2,315 entries evaluated for the following six emotions according to theory of Ekman (1992): *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise*. The entries of this lexicon are a small subset of the electronic version of a Greek dictionary (electronic version of the Triantafyllides' lexicon). They were selected using search tools which allowed custom searches based on term metadata, specifically the "tone" and the "description". Tsakalidis *et al.* (2014) selected terms either because their "tone" was ironic, meiotic, abusive, mocking or vulgar, or because their "description" contained emotion related words (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 emotion[2].

Tsakalidis *et al.* (2014) do not provide an in-depth analysis of the quality of the emotion ratings of their lexicon. To determine the agreement between the raters, we again use Pearson's correlation coefficient, for each pair of raters (Table IV). We see that there is a fair correlation between all pairs of raters for all emotions. Although, rating tweets may sound like an easier task (as there is a context) than rating individual words (out of context), it seems it is not: GSL's ratings for *anger* and *disgust* are in agreement across raters in contrast to the ground-truth we developed for our benchmark data set in Section 3.2. 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 emotions of the lexicon terms (Table V) and observed the following. First, there exists a highly correlated pair, *anger/disgust*, suggesting that these two emotion dimensions may not be independent. This means that anger may cause disgust and/or the other way around. Second, *surprise* seems moderately correlated to all other five emotions. We will attempt to explain this as follows: a surprise is usually followed by a peak in another emotion. For example, "I am surprised by my good performance in the exams" (happy), or "I am surprised by my bad marks in the exams" (fearing that will never graduate from university).

Rater pairs	Anger	Disgust	Fear	Happiness	Sadness	Surprise
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

Table IV. Greek emotion lexicon inter-rater Pearson's 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	–

Table V. Pearson's correlation of emotion pairs in the Greek emotion lexicon

A characteristic of GSL that makes it less than ideal for our work 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, and misses many informal words used in colloquial speech/writing. The entries of the lexicon came from a Greek dictionary and as a result they correspond better to formal or literal speech, which does not appear frequently in tweets. We measured that only 42 per cent of the words occurring in our tweet benchmark dataset (not counting the ignored stopwords) are contained in the dictionary.

4.2 Methods for tweet emotion rating

Tweet emotion is determined by the emotion of its words, in the space of the six emotions/sentiments considered in this paper. Word emotion is taken from the GSL described in the previous section. As GSL has each word rated by four raters for all emotions, we use the average rating w per emotion, as a fair degree of inter-rater agreement (Table IV) allows us to do. In detail, for each tweet's words existing in the lexicon we form a vector W with six emotion dimensions, one for each examined emotion and presenting in equation (1). We then have N vectors W_j:

$$W_j = [w_{1,j}, w_{2,j}, w_{3,j}, w_{4,j}, w_{5,j}, w_{6,j}] \tag{1}$$

where $j = 1...N$, and N is the number of emotion words (i.e. the words occuring in GSL) identified in the tweet. We also form, as presenting in equation (2), a six-dimension vector T:

$$T = [t_1, t_2, t_3, t_4, t_5, t_6] \tag{2}$$

for every tweet. Each emotion dimension t_i of T is calculated with one of the following four formulas considered in this paper, presented in the equations below:

$$t_i = \frac{\sum_{j=1}^{N} w_{i,j}}{N} \tag{3}$$

Equation (3) is simply the arithmetic mean:

$$t_i = \sqrt{\frac{\sum_{j=1}^{N} w_{i,j}^2}{N}} \tag{4}$$

Equation (4) is the quadratic mean, which is selected due to its bias to the larger numbers. In this way, it highlights words with strong emotion:

$$t_i = \frac{\sum_{j=1}^{N} w_{i,j}}{N} \tag{5}$$

Another approach is equation (5), where we assign to the tweet the maximum emotion found in the words occurring in the tweet, per dimension. The assumption here is that the dominant emotion of a tweet is expressed in the words with the highest emotion intensity:

$$t_i = \frac{\sum_{j=1}^{N} w_{i,j}}{N} \tag{6}$$

Finally, equation (6), which is known as CombMNZ in combining rankings in meta-search Shaw and Fox (1995), returns a higher value for the tweets that contain multiple words with high intensity in a particular emotion.

Figure 2 presents an example of four tweets, accompanied by our own loose translations to English. The ratings we calculated through Formula 2 (quadratic mean) for these tweets are shown in Table VI. We can see that the quadratic mean produces emotion intensities which are mostly in-line with our intuition for the example tweets. For the tweets 3 and 4, we compute high *anger* values, which coincide with their content concerning the national exams for university entry (the student seems to have failed) and the way the Greek parliament functions, respectively. The same tweets also show some *disgust* in their readings, confirming that the high correlation between *anger* and *disgust* we found in GSL (Section 4.1) is reasonable, as these tweets may be considered as expressing some disgust as well. Concerning *fear*, no tweet seems to have any, and this seems to be captured accurately by the readings. For the tweets 1 and 2, *happiness* is accurately computed as their dominant emotion; their subjects regard the upcoming vacation and the expectation of the Eurovision

Figure 2.
Examples tweets

Table VI.
Ratings of the tweets of Figure 2, per emotion, using the quadratic mean of the emotion values of their words (rescaled at [0-1])

Tweet	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Tweet1	0.20	0.20	0.20	0.95	0.20	0.55
Tweet2	0.20	0.20	0.20	0.75	0.20	0.50
Tweet3	0.70	0.70	0.20	0.20	0.20	0.50
Tweet4	0.52	0.40	0.16	0.16	0.19	0.33

Music Contest, respectively. The only tweet that could imply some sadness is tweet 3, which is not captured well by the reading. This seems to be due to the low coverage of GSL. The high correlation of *surprise* to all other emotions that we found in GSL (Section 4.1) is occurring here as well: a high intensity in any of the other emotions produces a high intensity in *surprise*. Nevertheless, we do not see the surprise in any of these tweets.

In summary, at least the quadratic mean seems to work well by example. A more thorough/systematic evaluation of all four formulas, using the benchmark data set we developed in Section 3, follows in Section 5.

4.3 Methods for hashtag emotion rating

In the previous section, we introduced methods for determining tweet emotion. Hashtag emotion is determined by the emotion of the tweets it occurs in, in the space of the six emotions/sentiments considered in this paper. In this respect, we compute a hashtag's H vector as:

$$H = \left[h_{1,k}, h_{2,k}, h_{3,k}, h_{4,k}, h_{5,k}, h_{6,k}\right] \tag{7}$$

using either the arithmetic or the quadratic mean as shown in Table VII, where M is the number of tweets having the hashtag and $k = 1..M$. Here, we rejected the maximum formula, as its results would only depend on the most emotional tweet, not taking into account the rest of the tweets having the hashtag. We also rejected CombMNZ as it creates an unfair bias toward hashtags with a larger number of tweets.

To better demonstrate our method, we present in Table VIII the overall results for the top ten most frequent hashtags, using the quadratic mean for both the individual tweets (Section 4.2) and the hashtag ratings. Let us look at the highest and lowest readings

Table VII.
Two formulas for hashtag emotion rating

Arithmetic mean	$h_i = \frac{\sum_{k=1}^{M} t_{i,k}}{M}$
Quadratic mean	$h_i = \sqrt{\frac{\sum_{k=1}^{M} t_{i,k}^2}{M}}$

Table VIII.
Hashtag ratings

Hashtag	Anger	Disgust	Fear	Happiness	Sadness	Surprise
#wc14gr	0.2782	0.25724	0.19024	0.27208	0.16824	0.29104
#ekloges14	0.23254	0.23352	0.1636	0.22912	0.14438	0.2577
#kalokairipantou	0.1586	0.18316	0.15478	0.43712	0.1514	0.42168
#skouries	0.21216	0.2092	0.18798	0.21206	0.14674	0.22394
#panellinies2014	0.278	0.26748	0.1962	0.29042	0.16306	0.29318
#vouli	0.2608	0.25216	0.15664	0.23534	0.14838	0.26244
#ert	0.21784	0.21514	0.1613	0.20484	0.13388	0.22584
#eurovisiongr	0.26928	0.25914	0.15866	0.27066	0.15198	0.28184
#mb14gr	0.27896	0.25484	0.1902	0.26902	0.16082	0.2845
#enikos	0.26378	0.25732	0.1639	0.23836	0.15232	0.27102

per emotion. Summer Everywhere (#kalokairipantou) has the lowest *anger*, *disgust*, *fear* and the highest *happiness* and *surprise*. Basketball and Football World Cups (#mb14gr and #wc14gr) arouse the highest *anger*, and the highest *sadness* in football (the Greek national team did much worse than it did in Basketball). The national exams for university entry (#panellinies2014) produce the highest *disgust* and *fear*. The shutdown of the national radio and TV broadcaster (#ert) arise the lowest *happiness*. All the above are expected according to common intuition; nevertheless, the lowest *sadness* in #ert and lowest *surprise* in environmental activist actions against gold mining in Skouries #skouries are kind of borderline results in our opinion.

In summary, example results suggest that the proposed method is capable of producing reasonable emotion readings for the thematic categories. In any case, we will evaluate the method in a more systematic way in the following section.

5. Experimental evaluation of lexicon-based methods

In the previous section, we presented our methodologies for determining tweet and hashtag emotion in a space of six emotions, using the GSL, and we roughly checked the validity of our results by presenting some examples. In this section, we perform a systematic evaluation of those methods, using the benchmark data set (GSTD) we developed in Section 3.

5.1 Rating individual tweets

To evaluate the quality of the four proposed formulas for rating tweets, we calculate Pearson's correlation between the automated ratings we produce and those of the human raters in the benchmark data set. We choose this type of metric for this and all following evaluations, as this is not a typical retrieval task. There are neither explicit queries nor results of binary relevance judgments. Consequently, we set to discover a statistical correlation between the manually produced and the computed scores. We also introduce here Kendall's rank-correlation coefficient, to determine whether there is a non-linear relation. The data set has only 691 tweets rated by humans. For the evaluation, we use only 4 of the 6 available emotions, i.e. *fear*, *happiness*, *sadness* and *surprise*; the other two were deemed as less usable due to the low inter-rater agreement in the GSTD (Table III). The results are shown in Table IX; here, we use the average of the manual judgments of the two raters. Unfortunately, we see no correlation in most emotions and methods, except maybe in *happiness* which seems to pick up a bit.

The correlation values between our results and the evaluation set can be deemed more worthy considering the already marginally acceptable correlation values between the two annotators. In fact, the agreement between our system and the combined raters' scores is almost half the agreement between them (Table III) for *happiness*. The emotions of *fear*, *sadness* and *surprise* produce not acceptable results. In an attempt to evaluate the emotions of *anger* and *disgust*, we calculated the correlation of our results with each rater separately, which also got us nowhere (Table X).

Table IX. Pearson/Kendall correlation between the automated rating and the merged rating of the two individual raters

	Fear		Happiness		Sadness		Surprise	
Formula	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
Arithmetic	−0.008	−0.038	0.226	0.158	−0.042	−0.043	−0.064	−0.062
Quadratic	−0.032	−0.041	0.143	0.133	−0.048	0.028	−0.035	−0.047
Maximum	0.044	−0.004	0.198	0.150	−0.007	−0.026	−0.043	−0.049
Combmnz	0.004	−0.024	0.086	0.073	−0.010	−0.012	−0.049	−0.038

In a further failure analysis, we observed that a fair portion of the tweets in the evaluation set did not contain any terms of GSL. Consequently, we tried to re-evaluate our results using a subset of the evaluation data set with tweets that certainly contain terms included in GSL (524 tweets). Table XI presents the results of the other most promising emotions beyond *happiness*. Using the maximum formula for the aggregation, we calculated correlation higher than 0.1 for both *anger* and *fear*, for Rater2, which is some improvement in comparison with the results of the complete evaluation set. The other formulas did not present any promising results.

We conclude that the proposed methods are able to extract automated ratings for some of the examined emotions, especially in the case of the tweets that contain terms of the emotion lexicon. The results are fair for *happiness* and barely acceptable for *anger* and *fear* when the presence of GSL terms is ensured. As a result, we point again to the need for a new, extended, Greek Emotion lexicon, specialized for microblogging applications which will have a better coverage over the emotional terms for internet speech usage.

5.2 Evaluation of rating hashtags

As we do not have manual hashtag ratings (or ground-truth), we cannot determine the quality of our automated hashtag ratings. With that in mind, we assume the following (weaker) ground-truth: the emotion of each hashtag is the average manual emotion of all tweets in which the hashtag occurs. This manual emotion of each tweet is the average manual emotion of the two raters.

We calculated the emotion ratings for the ten hashtags contained in the evaluation dataset (Table II) only considering the sentiments *anger*, *fear* and *happiness* which were the only ones that produced promising results in Section 5.1. These ratings were calculated with the two formulas described in Section 4.3 and for the emotions that we consider worthy of examination, based on Section 5.1. We then calculated the Pearson correlation of each emotion for the ten different hashtags of Table VIII (Table XII).

The quadratic mean seems to be better for the task of hashtag rating. Especially in the case of *happiness*, which proves to be the easiest emotion to detect in both of our experiments, the correlation reaches a value of over 0.9 in Pearson. As a result, we are almost always able to detect the happy hashtags with our automated methods. Also, for the emotions of *fear* and *anger*, we manage to achieve a fair correlation of 0.4 using different

Table X.
Pearson/Kendall correlation between the automated rating and the rating of each one of the individual raters, for *anger* and *disgust*

	Rater1				Rater2			
	Anger		Disgust		Anger		Disgust	
Formula	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall	Pearson	Kendall
Arithmetic	−0.022	−0.036	0.006	0.005	0.051	0.000	0.008	0.001
Quadratic	−0.036	−0.045	−0.006	−0.020	0.010	−0.008	−0.019	−0.007
Maximum	−0.031	−0.034	−0.019	−0.010	0.063	0.014	0.033	0.010
Combmnz	0.031	0.019	0.031	0.040	−0.017	−0.021	0.003	−0.013

Table XI.
Pearson correlation for reduced evaluation set

	Rater1		Rater2	
Formula	Anger	Fear	Anger	Fear
Maximum	−0.047	0.035	0.133	0.131

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

6. A new Greek emotion lexicon

The Greek emotion lexicon (GEL) provides many terms with emotional content on which we have based our automated rating methods and obtained promising results. Nevertheless, as we have observed so far, its terms do not appear frequently in the examined tweets. We calculated that only 42 per cent of the words of the tweet collection appear as lexicon entries. In Section 5.1, we saw that when we ensure the existence of lexicon entries in the automatically rated tweets, the results extracted from our method came much closer to human intuition (Table XI). Having concluded that internet language/text is greatly different from ordinary text documents, we are led to seek a way to incorporate informal terms with emotional content in an emotion lexicon.

Thus, in this section, we introduce a New Greek Emotion Lexicon (NGEL), tailored for Greek micro-blogging text. First, we collect the requirements and present the method for developing the lexicon (Section 6.1). Then, we present the lexicon characteristics in Section 6.2, and finally, we evaluate the performance of NGEL in our experimental setup (Section 6.3).

6.1 Requirements and development

The most important requirement seems to be a larger coverage. The NGEL has to include entries that appear in Greek tweets, to provide higher matching rates with microblogging data. The second requirement tries to deal with word-emotion ambiguity. Words in GSL are rated out-of-context, although sometimes it is difficult to determine emotion like that. For example, the Greek word $\kappa\lambda\alpha\iota\omega$ (cry) would probably be assigned a high *sadness* in isolation, although in some contexts could denote something really funny (cry from laughter) and surprising, as it holds for the overwhelming majority of our tweets. In this respect, we examine word emotion in context. A third requirement is to collect words of a high emotional content, and not sentimentally neutral. To summarize, we target to develop an emotion lexicon with words:

- occurring in real tweets;
- emotion-rated in context; and
- of high emotional content.

To achieve these requirements, we selected a set of tweets, with the purpose of having them manually rated and then using these ratings to generate ratings for their words.

The set of tweets was created as follows. To ensure the presence of terms with a certain emotional content, we selected the 100 highest rated terms for each emotion, from the original GSL. Of those terms, we selected the ten most frequently encountered terms in our collection (i.e. a total of 60 for all 6 emotions), to ensure their frequent presence in the collection. Then, we retrieved ten unique random tweets containing each one of these terms

	Arithmetic			Quadratic		
Formula	Anger	Fear	Happiness	Anger	Fear	Happiness
Arithmetic	−0.09	−0.02	0.92	0.43	0.20	0.93
Quadratic	−0.12	−0.03	0.86	0.33	0.20	0.88
Maximum	−0.10	0.23	0.88	0.42	0.40	0.78

Table XII.
Pearson correlation for hashtag ratings (691 tweets)

for a total of 600 tweets. Two individuals raters (different individuals from the ones that created the benchmark data set) manually rated these tweets for the intensity of every emotion in a range of 0-5, thus fulfilling the requirement of in-context rating.

At this point, we had a set of manually rated tweets which certainly contain emotional content. We have made this evaluation set publicly available[3]. Our goal was to transfer this information to their occurring terms in a meaningful manner. For this purpose, we apply the following simple process:

- We stem all words in the set of rated tweets.
- We make a list of unique words ignoring words contained in the list of stopwords mentioned in Section 3.3.
- For every word, its intensity for an emotion is the average rating for that emotion of the tweets that contain that word.

Next, we present some statistics of the developed lexicon.

6.2 Lexicon characteristics

As described above, in the way it was constructed, the NGEL contains an overlap of terms from GSL and terms retrieved from our collection of Greek tweets.

Some statistics of the NGEL are presented in Table XIII.

A disadvantage of NGEL, is that as its terms derive from a relatively small number of tweets (600), there are plenty of terms that appear only once, and as a result, their emotional rating is affected only by the rating of one tweet. Also, we cannot exclude these terms from the lexicon because it will reduce its size by almost four times. We choose to keep the entries with only one appearance in the tweets, and as we will see in the next subsection, despite this problem, the new lexicon will provide better results.

Table XIV shows the inter-rater correlation between the two raters. Our raters have a fair degree of correlation in their ratings for all emotions. In Section 3.2, where we developed the benchmark data set, the other two raters there presented a practically zero correlation for the emotions of *anger* and *disgust*. This is not the case here; this may be due to the clear emotional content of the tweets selected to construct the NGEL. In the case of the benchmark data set, there were many neutral tweets or tweets with unclear emotional content, making the task of their rating difficult.

Table XIII.
New Greek emotion lexicon statistics

Number of entries	2,018
Number of tweets from which the entries originate	600
Number of raters for each tweet	2
Number of entries occurring in more than 10 tweets	111
Number of entries occurring in 3 to 10 tweets	227
Number of entries occurring in 2 tweets	266
Number of entries occurring in 1 tweet	1,413

Table XIV.
Inter-rater pearson correlation in NGEL

–	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Correlation	0.5614	0.3162	0.2326	0.5283	0.4018	0.1692

Comparing GSL to NGEL, the latter has slightly fewer entries (GSL:2315 [Section 4.1], NGEL: 2018 (Table XIII), but it should provide more coverage in a micro-blogging environments. Also, as we can observe from Tables IV and XIV, the inter-rater correlations of NGEL are comparable to the ones of GSL and better than certain pairs of GSL raters. We should note that GSL correlation regards to individual term ratings whereas the NGEL one regards tweet ratings.

6.3 Performance

To determine whether NGEL performs better than GSL for the task of emotion evaluation of internet data, we perform a series of experiments to compare the two. These experiments are along the same lines as earlier in this paper. We automatically evaluate the GSTD by using each lexicon as a base for our methods. We then compare the automated rating to the rater annotations in terms of Pearson correlation. We also extract automated ratings for hashtags based on the two lexicons. Subsequently, we calculate the correlation between these ratings and the ones produced by the accumulation of the manual annotations, as in previous sections.

Table XV presents the correlation of our results to the evaluation dataset, using the arithmetic mean formula for both lexica. We conclude that the arithmetic mean does not produce acceptable results in neither case. The correlation values are very low for the 691 examined tweets. Maximum, also performed well for more emotions, in the case of the reduced evaluation data set that was examined in Table XI.

As we can see in Table XVI, NGEL shows some correlation for the emotions of *anger*, *disgust*, *happiness* and *surprise*. In total, we can make the following remarks:

Table XV. Pearson correlation of automated results using both lexica with manual ratings of the evaluation dataset (formulae used: arithmetic mean)

	GSL			NGEL		
Emotion	Both raters	Rater 1	Rater 2	Both raters	Rater 1	Rater 2
Anger	0.0293	−0.0222	0.0512	0.0519	0.0030	0.0591
Disgust	0.0090	0.0056	0.0077	0.0605	0.0262	0.0449
Fear	−0.0080	−0.0047	−0.0081	0.0007	−0.0433	0.0244
Happiness	0.2263*	0.1912*	0.1982*	0.1539*	0.1580*	0.1146*
Sadness	−0.0416	−0.0516	−0.0257	0.0321	0.0244	0.0310
Surprise	−0.0637	−0.0391	−0.0633	0.0253	−0.0204	0.0484

Note: *Denotes statistical significance at the 5% level

Table XVI. Pearson correlation of automated results using both lexica with manual ratings of the evaluation dataset (formula: maximum)

	GSL			NGEL		
Emotion	Both raters	Rater 1	Rater 2	Both raters	Rater 1	Rater 2
Anger	0.0383	−0.0305	0.0633	*0.1015**	0.0098	0.1140*
Disgust	0.0178	−0.0192	0.0330	*0.0925**	0.0088	0.0944*
Fear	0.0449	0.0119	0.0529	−0.0166	−0.0344	−0.0032
Happiness	0.1987*	0.1828*	0.1633*	*0.1599**	0.1559*	0.1251*
Sadness	−0.0068	−0.0316	0.0125	*0.0650*	0.0675	0.0495
Surprise	−0.0433	−0.0171	−0.0490	*0.1047**	0.0497	0.1134*

Note: *Denotes statistical significance at the 5% level

- NGEL constitutes a better base for our automated methods. Once more, we mention that the correlation values of Tables XV and XVI can be taken as acceptable if we consider the low degree of inter-rater correlation (Table IV) between the rater annotations and our results.
- Once again our automated results approach the intuition of the second rater for both the examined lexica. This fact reminds us of the subjectivity that the sentiment analysis related tasks present.
- NGEL also presents a greater coverage of terms appeared in the evaluation data set. We measured that in the total data set, 42 per cent of the tweets contain terms of the GSL and 67 per cent of the tweets contain terms of NGEL[4]. This improvement in coverage is one of the main reasons the correlation results of Table XVI are improved. This renders NGEL a better resource for microblogging applications of sentiment analysis.
- Tables XV and XVI show a slightly weaker performance for the emotion of happiness compared to GSL. Overall, however, NGEL presents improvements to most other emotions.
- Our inability to achieve higher correlation values for the emotion of *fear* does not seem of great importance when we familiarize with the emotions expressed through the Twitter platform. From an empirical view of the data set, it seems that even terms usually related to *fear* do not express the expected emotion. The Greek users, when sharing their negative feeling through Twitter, prefer to do it in a sarcastic way, or by expressing pure *anger*. Only for very few tweets, the overall emotion is easily perceived as that of *fear*. This is apparent from Tables XV and XVI where *fear* values are consistently low, although statistically non-significant.

The results for the automated hashtag rating are shown in Table XVII. The arithmetic mean produces the best results, and we reach a fair degree of correlation in hashtag evaluation for the ten examined hashtags. Nevertheless, correlations for *Anger* are statistically insignificant. Once more, it seems that our methods produce more stable results with the accumulation of larger amounts of data.

7. Hashtag emotion over time

In this section, we demonstrate the usefulness of measuring emotion changes for hashtags over time. We choose the emotions of *anger* along with *happiness* because we can associate their changes with events in the timespan of the data set.

To calculate the emotion intensity, we use the quadratic mean both for individual tweets and for the accumulation of groups of tweets. Based on the method we proposed in the previous section, we calculate the average hashtag emotion for one-day intervals. We chose

Table XVII.
Pearson correlation NGEL hashtag evaluation

Emotion	Mean	Quadratic
Anger	0.4608	0.4646
Disgust	0.7702*	0.3979
Happiness	0.7206*	0.7149*
Surprise	0.7514*	0.8137*

Note: (*) denotes statistical significance at the 5% level

to examine only days for which we have gathered more than 60 tweets, to have more conclusive results.

Figure 3 depicts emotion changes for the Football World Cup '14 (#wc14gr) over time. We see that we are able to detect peaks in emotion 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.

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

An interesting observation can be made in the daily results for #wc14gr. The emotion of *happiness* in Figure 3 seems to have inverse changes to the emotion of *anger*. In contrast, in the case of the #panellinies2014, fluctuations exhibit greater similarity. Generally, we can say that in the case of a football cup, these emotions do not manifest simultaneously, while in the occasion of national exams, it is reasonable to observe mixed emotion for the same time intervals.

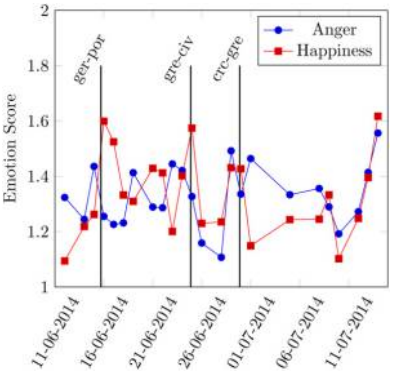

Figure 3.
#wc14gr: emotion intensity per day

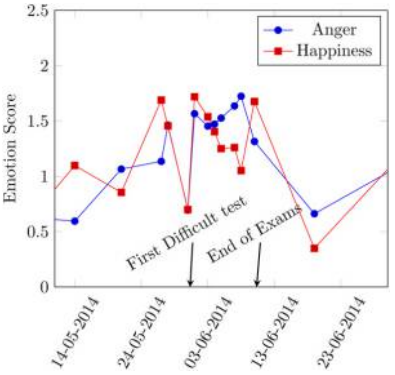

Figure 4.
#panellinies2014: emotion intensity per day

8. Conclusions and future work

Automated opinion and/or sentiment/emotion 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 or emotion 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. To demonstrate that let us consider the complexity of calculating emotion scores for a single tweet. For each tweet, the system has to find the emotion scores of every term in the tweet by searching for that term in the lexicon. Using a hashmap structure for the lexicon, this process would practically be of complexity $O(1)$. This must be done for all n terms of the tweet, so now we have a complexity of $O(n)$. In turn, the calculation of any of the formulas has a complexity of at most $O(n)$. As a result, the complexity of calculating the emotion scores of a single tweet is $O(n) + O(n) = O(n)$.

These initial experiments led to interesting remarks, which could guide further investigation and improvements:

- The emotion *happiness* seems to be the easiest one to detect throughout all the emotions in both of our experiments. For the emotion of *fear*, results are also promising.
- Different formulas appear to perform best for different emotions. For example, Formula 3 (max) returns better results than the others, for the emotion of *fear*. Formulas 1 and 2, on the other hand, perform better for the emotion of *happiness*.
- The presence of a large amount of tweets leads to a better assessment of the overall sentiment/emotions of the whole set, through the methods that we described, even in the cases where the individual tweet ratings do not appear as accurate.

We also investigated the performance of the same methods using an emotion lexicon that we created focusing on the particularities of internet speech. We found that its use is beneficial to the task of sentiment and emotion analysis on social media, as it achieves a higher coverage of our collection vocabulary and resolves some ambiguities of informal speech. These characteristics resulted in improved results in the metrics that we used.

As presented in Section 7, we may also be able to detect changes in emotion over time, and the results coincide with our intuition about real-world events. Furthermore, our data set of tweets together with the manual user ratings, the training set of manually rated tweets as well as the new emotion lexicon are publicly available at http://hashtag.nonrelevant.net, resources which could prove valuable for other researchers. As potential improvements of our methods or directions for further research, we propose the following:

- Utilization of linguistic data such as the part of speech that each entry is, and inclusion of other features of tweets such as emoticons and punctuation marks.
- The use of other statistical methods, such as keyword or co-occurrence analysis, to extract potential terms emotion terms.
- A sensitivity analysis to see whether the methods are robust to score shift or scale change (e.g. word rating in 0-7 instead of 0-5).
- Extension of the benchmark data set both in size and in number of raters.

- Extension of the New Greek Emotion Lexicon, by the use of a larger training set that will include more tweets/terms.
- Further examination of changes in emotion over time; this method is not evaluated in the present work.

We have examined the topic of Emotion Analysis using an emotion lexicon, providing a benchmark resource/data set together with baseline performance of several simple and efficient algorithms and a New Greek Emotion Lexicon. We hope that all these will be proven valuable for us and the community to build upon in future work.

Notes

1. http://hashtag.nonrelevant.net/
2. The Greek Emotion Lexicon also contains some linguistic information regarding the entries, such as part of speech and 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; we only use the emotion ratings.
3. http://hashtag.nonrelevant.net/
4. Percentages calculated after pre-processing.

References

Abdul-Mageed, M., Diab, M.T. and Korayem, M. (2011), "Subjectivity and sentiment analysis of modern standard arabic", *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT 11, Association for Computational Linguistics*, *Stroudsburg, PA*, pp. 587-591.

Baccianella, S., Esuli, A. and Sebastiani, F. (2010), "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining", *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 10), European Language Resources Association (ELRA)*, *Valletta*.

Bagola, H. (2004), "Informations utiles à l'intégration de nouvelles langues européennes", DIR/A-Cellule "Méthodes et développements", section "Formats et systemes documentaires".

Chen, Y. and Skiena, S. (2014), "Building sentiment lexicons for all major languages", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics*, *Baltimore*, pp. 383-389.

Clematide, S. and Klenner, M. (2010), "Evaluation and extension of a polarity lexicon for German", *Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, *Lisbon*, pp. 7-13.

dos Santos, C.N. and Gatti, M. (2014), "Deep convolutional neural networks for sentiment analysis of short texts", COLING 2014, *25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers,* 23-29 August, *Dublin*, pp. 69-78.

Effrosynidis, D., Symeonidis, S. and Arampatzis, A. (2017), "A comparison of pre-processing techniques for twitter sentiment analysis", *Proceedings of Research and Advanced Technology for Digital Libraries – 21st International Conference on Theory and Practice of Digital Libraries*, 18-21 September, *Thessaloniki*, pp. 394-406.

Ekman, P. (1992), "An argument for basic emotions", *Cognition & Emotion*, Vol. 6 Nos 3/4, pp. 169-200.

Giachanou, A. and Crestani, F. (2016), "Like it or not: a survey of twitter sentiment analysis methods", *ACM Computing Surveys*, Vol. 49 No. 2, pp. 28:1-28:41.

Hu, M. and Liu, B. (2004), "Mining and summarizing customer reviews", *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 22-25 August, *ACM, Washington*, pp. 168-177.

Kalamatianos, G., Mallis, D., Symeonidis, S. and Arampatzis, A. (2015), "Sentiment analysis of greek tweets and hashtags using a sentiment lexicon", *Proceedings of the 19th Panhellenic Conference on Informatics, PCI 15, ACM, New York, NY*, pp. 63-68.

Khan, F.H., Qamar, U. and Bashir, S. (2016), "SentiMI: introducing point-wise mutual information with SentiWordNet to improve sentiment polarity detection", *Applied Soft Computing*, Vol. 39, pp. 140-153.

Kouloumpis, E., Wilson, T. and Moore, J.D. (2011), *Twitter Sentiment Analysis: The Good the Bad and the OMG!*, AAAI Press, Palo Alto, CA, pp. 538-541.

Liu, B. and Zhang, L. (2012), *A Survey of Opinion Mining and Sentiment Analysis*, Springer, Berlin, pp. 415-463.

Lo, S.L., Cambria, E., Chiong, R. and Cornforth, D. (2016a), "A multilingual semi-supervised approach in deriving singlish sentic patterns for polarity detection", *Knowl.-Based Syst*. Vol. 105, pp. 236-247.

Lo, S.L., Cambria, E., Chiong, R. and Cornforth, D. (2016b), "Multilingual sentiment analysis: from formal to informal and scarce resource languages", *Artificial Intelligence Review*, Vol. 48 No. 4, pp. 499-527.

Lu, B., Song, Y., Zhang, X. and Tsou, B. (2010), "Learning chinese polarity lexicons by integration of graph models and morphological features", *Information Retrieval Technology', Vol. 6458 of Lecture Notes in Computer Science*, Springer, Berlin Heidelberg, pp. 466-477.

Mahyoub, F.H., Siddiqui, M.A. and Dahab, M.Y. (2014), "Building an arabic sentiment lexicon using semi-supervised learning", *Journal of King Saud University – Computer and Information Sciences*, Vol. 26 No. 4, pp. 417-424.

Medhat, W., Hassan, A. and Korashy, H. (2014), "Sentiment analysis algorithms and applications: a survey", *Ain Shams Engineering Journal*, Vol. 5 No. 4, pp. 1093-1113.

Mehrabian, A. (1995), "Framework for a comprehensive description and measurement of emotional states", *Genetic, Social, and General Psychology Monographs*, Vol. 121, pp. 339-361.

Mohammad, S.M. and Turney, P.D. (2010), "Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon", *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, Association for Computational Linguistics*, *Stroudsburg, PA*, pp. 26-34.

Molina-Gonzalez, M.D., Martinez-Camara, E., Martin-Valdivia, M.-T. and Perea-Ortega, J.M. (2013), "Semantic orientation for polarity classification in spanish reviews", *Expert Systems with Applications*, Vol. 40 No. 18, pp. 7250-7257.

Ntais, G. (2006), "Development of a stemmer for the greek language", Master's thesis, Stockholm University.

Pak, A. and Paroubek, P. (2010), "Twitter as a corpus for sentiment analysis and opinion mining", *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010*, 17-23 May, *European Language Resources Association*, *Valletta*.

Paltoglou, G. and Buckley, K. (2013), "Subjectivity annotation of the microblog 2011 realtime adhoc relevance judgments", *Advances in Information Retrieval, Vol. 7814 of Lecture Notes in Computer Science*, Springer, Berlin Heidelberg, pp. 344-355.

Pang, B. and Lee, L. (2008), "Opinion mining and sentiment analysis", *Foundations and Trends® in Information Retrieval*, Vol. 2 Nos 1/2, pp. 1-135.

Pennebaker, J.W., Mehl, M.R. and Niederhoffer, K.G. (2003), "Psychological aspects of natural language use: our words, our selves", *Annual Review of Psychology*, Vol. 54 No. 1, pp. 547-577.

Rao, D. and Ravichandran, D. (2009), "Semi-supervised polarity lexicon induction", *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, EACL 09, Association for Computational Linguistics, Stroudsburg, PA*, pp. 675-682.

Serrano-Guerrero, J., Olivas, J.A., Romero, F.P. and Herrera-Viedma, E. (2015), "Sentiment analysis: a review and comparative analysis of web services", *Information Sciences*, Vol. 311, pp. 18-38.

Severyn, A. and Moschitti, A. (2015a), "Twitter sentiment analysis with deep convolutional neural networks", *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 15, ACM, New York, NY*, pp. 959-962.

Severyn, A. and Moschitti, A. (2015b), "Unitn: training deep convolutional neural network for twitter sentiment classification", *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver*, pp. 464-469.

Shaw, J.A. and Fox, E.A. (1995), *Combination of Multiple Searches*, NIST Special Publication SP, Gaithersburg, p. 105.

Strapparava, C. and Mihalcea, R. (2007), "Semeval-2007 task 14: affective text", *Proceedings of the 4th International Workshop on Semantic Evaluations, Association for Computational Linguistics*, pp. 70-74.

Tsakalidis, A., Papadopoulos, S. and Kompatsiaris, I. (2014), "An ensemble model for cross-domain polarity classification on twitter", *Web Information Systems Engineering Wise 2014, Vol. 8787 of Lecture Notes in Computer Science*, Springer International Publishing, Berlin, pp. 168-177.

Tsakalidis, A., Papadopoulos, S., Cristea, A. and Kompatsiaris, Y. (2015), "Predicting elections for multiple countries using twitter and polls", *IEEE Intelligent Systems*, Vol. 30 No. 2, pp. 10-17.

Vossen, P. and Letteren, C.C. (1997), "Eurowordnet: a multilingual database for information retrieval", *Proceedings of the DELOS workshop on Cross-language Information Retrieval*, pp. 5-7.

Vossen, P., Hofmann, K., de Rijke, M., Sang, E.T. and Deschacht, K. (2007), "The cornetto database: architecture and user-scenarios", *Proceedings DIR 2007*, pp. 89-96.

Xu, J., Xu, R., Zheng, Y., Lu, Q., Wong, K.-F. and Wang, X. (2013), "Chinese emotion lexicon developing via multi-lingual lexical resources integration", *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing - Volume 2, CICLing 13, Springer-Verlag, Berlin, Heidelberg*, pp. 174-182.

Corresponding author
Georgios Kalamatianos can be contacted at: georkalamatianos@gmail.com