Micro-Blogging Content analysis via emotionally-Driven Clustering

Despoina Chatzakou, Vassiliki Koutsonikola, Athena Vakali
Department of Informatics
Aristotle University
54124 Thessaloniki, Greece
{deppych, vkoutson, avakali}@csd.auth.gr

Konstantinos Kafetsios
Department of Psychology
University of Crete
GR74100 Rethymno, Greece
k.kafetsios@psy.soc.uoc.gr

***Abstract*—Microblogging has become commonplace and created new methods of communication, contributing significantly to information sharing. This holds since microblogging focuses on sharing content while building social relations among people who share the same interests and/or activities. In this context, people's perception and emotions towards a specific subject is a valuable piece of information and sentiment and affective analysis play an important role. In this paper, an affective analysis methodology is proposed, which is a lexicon-based technique, for capturing the wisdom of crowds, as well as the social pulse and the trends, through the more accurate assessment of human emotion states. The methodology adopted involves the monitoring of the emotions' intensity, i.e. how strong or weak the emotional states of the published information are. The results suggest that the proposed approach manages to efficiently capture people's emotions as these were recorded in datasets derived from Twitter.**

I. Introduction

The emergence of the Social Web has brought significant changes in the way people communicate and share information since it provides many opportunities for publishing people's views, thoughts and information relevant to a variety of topics. Specifically, in social network applications the users are expressing their opinions through reviews, comments, ratings and other forms of online expressions. Since millions of users share opinions on different aspects of everyday life, social networks are considered as a credible source for exploring both factual and subjective information.

The social web growth and the desire to capture the wisdom of crowds have increased the interest in sentiment and affective analysis. Sentiment analysis focuses on opinion tracking, whereas in affective analysis the focus is on emotion tracking. Recently there has been recorded a significant research interest towards capturing people's emotions reflected in microblogging services, such as Twitter. The evolution of microblogs, which has resulted in user enriched information and opinionated personal commentary, offers a unique look into people's emotion-laden reactions and attitudes.

In our work, we perform a fine-grained analysis on users blog posts (tweets), based on six primary emotions (anger, disgust, fear, joy, sadness, surprise), which are in accordance with seminal theoretical frameworks such as Ekman's theory [4], and they are considered as valid in the description of the variety of humans' emotions. Along with capturing people's emotions, emotions' intensity is also recorded. An advanced methodology, based on semantic lexicons and emotional dictionaries, is proposed for mapping the similarity between the tweets and the primary emotions and for the estimation of the emotional strength and orientation of the tweets. Finally, for better understanding of users' emotional state, a clustering methodology is applied, organizing tweets into groups, with relevance to the emotions they express and their intensity.

The proposed analysis was applied on data derived from Twitter, which is a popular service that collects users' opinions for different events happening in different places all over the world. It is evident that the mining and analysis of twitter's content is considered to be important and necessary for recognizing interesting trends and opinions about various topics. This could be useful for companies, policy makers and the general public. For example, behavioural patterns and trends of the public can be recognized, which can facilitate the targeting of commercial communication as well as the identification of views on political issues.

The contribution of our work is summarized as follows:

- we propose an advanced method that takes into consideration several linguistic parameters, such as the intensifiers and the valence shifters, for understanding the emotions that are being expressed in microblogs;
- we propose a similarity measure which evaluates the relation of each tweet with any of the six primary emotions. The proposed approach improves significantly the results of previous works, since earlier methods that were based only in emotion valence (positive/negative scale) are limited and do not capture the spectrum of human emotions;
- we propose a clustering framework to organize tweets according to their intensity in relation to each primary emotion, resulting in a better understanding of the extracted results;

II. Related work

The interest in exploring people's emotions about certain topics has increased significantly and has inspired research in the area of automatic sentiment analysis. The challenge for leveraging sentiment analysis is tracking disparate information for accurately capturing the opinions that are being expressed in social networks. Typically, in document and sentence level sentiment analysis the objective is to determine the orientation (positive/negative/neutral) of each document and sentence, respectively, leading in a better understanding of people's opinions. For instance, in [15] an unsupervised learning algorithm is presented for classifying reviews into positive and negative. The semantic orientation of only those phrases containing adjectives and adverbs was estimated, through a sum of scores of each phrase. In [2] a methodology was proposed for estimating the orientation of texts, using linguistic analysis of adverbs.

In sentence level, sentiment analysis is usually applied in two steps. Initially, the subjectivity/objectivity score for each sentence is determined, and then, only for the subjective sentences, the orientation is studied. In [7], [16] some methodologies are presented for categorizing sentences based on how positive, negative or neutral they are. In [8], [9], the authors present sentiment classifiers that determine how positive, negative and neutral the Twitter messages are.

Affective analysis is considered as a valuable methodology for monitoring changes of people's emotions over time. It could be considered as a more involved version of sentiment analysis as it creates a more detailed categorisation of emotional opinions. It involves classification of texts into categories, based on the emotions that are being expressed. In [6] the classification is based on 8 primary emotions (fear, anger, disgust, sadness, acceptance, anticipation, joy, surprise) and a list of exemplar words, to represent each one of the eight primary emotions, was used. In [14], the previously referred emotions were used for creating groups of users that share the same emotions on specific topics in Twitter. Finally, in [3] the authors were based in the Profile of Mood States (POMS), a psychometric instrument, which makes use of 6 basic emotions, namely "tension, depression, anger, vigor, fatigue, confusion", for performing affective analysis of tweets.

For applying sentiment and affective analysis machine learning and lexicon-based techniques are adopted. For example, in [9], the Naive Bayes, the SVM and the CRF classifiers were examined. In [10], the authors tried to develop a sentiment analyzer for tweets using the Naive Bayes classifier and the Maximum Entropy Model. Instead in [7], [12], lexicon-based techniques were used. In [7], opinion sentences in customer reviews were identified and then the orientation of each sentence using a set of adjective words was determined. In [12], in addition to adjectives, the adverbs, the nouns and verbs were taken into consideration in the proposed procedure.

The above techniques are used widely in the literature, with the lexicon-based techniques presenting better performance in sentiment/affective analysis. As the number of existing lexicons has rapidly grown, there is an increasing number of researchers that use lexicon-based methods in order to evaluate the semantic orientation of context of microblogs. But the need of a more holistic approach is obvious. In this paper, a lexicon-based methodology was adopted and at the same time our analysis proceeds one step further by deepening even more in the emotional knowledge (affective analysis) and not being limited only to the calculation of twitter's content orientation.

III. Problem Formulation

The goal of our work was to evaluate and quantify the emotions expressed by users in their blog posts. Thus, in this section we provide the necessary definitions about all the information that is involved in the proposed affective analysis approach.

A. The emotional tweet structure

In Twitter and other microblogging services, users create new, shortened forms of expression due to the limitation on the number of characters that there are in tweets (140 characters). Thus, each tweet may carry useful emotional information for our analysis, but it can also involve information that is considered as noise and offers nothing useful towards our goal. In our approach we keep only emotion-relevant information. Next we provide some basic symbols that refer to the data structures used to describe the involved, in our approach, parameters.

Let $T = \{t_1, \ldots, t_m\}$ represent the set of m tweets and $E = \{e_1, \ldots, e_l\}$ denote the set of l primary emotions. The purpose of the proposed methodology is to estimate the relation between each tweet $t_i \in T$ and each one of the l primary emotions $e_l \in E$.

We consider the set $ER_i = \{r_{i1}, \ldots, r_{ip}\}$ of p words which act as representatives (secondary emotions) of emotion e_i. The ER_i set emerged from Wordnet-Affect [11] [1] which is an extension of Wordnet domains, that includes a subset of synsets suitable to represent affective concepts correlated with affective words. For a number of Wordnet synsets one or more affective labels were assigned. Moreover, in order to extend the list of representative words, Wordnet lexicon [5] was used. For each word the list of its synsets was retrieved, resulting in a new list with the total number of the representative words. As a consequence the total number of representative words increased significantly, covering a wider range of words.

Given the set ER_i of representative words for the e_i primary emotion, the set of emotional words ET_i of tweet t_i is the set of words that belong to tweet t_i as well as they have the same root with any of the words of the set ER_i. The common root between two words can be examined using an appropriate stemmer.

Definition 1 (The Tweet-EmotionalWords): The *Tweet-EmotionalWords* ET_i is defined as $ET_i = \{w_j : w_j \in t_i, stemmer(w_j, ER) \neq 0, 1 \leq j \leq n\}$

Example We consider the tweet t_i ="Few days before Christmas. I am too excited". If the ER_i set contained only the word "excited" from the words of the tweet t_i, then the ET_i set would be defined as $ET_i = \{excited\}$ as the word "excited" belongs in ERi. Thus, the t_i contains only one emotional word which is "excited".

Moreover, emoticons are the pictorial representation of facial expressions that visualize a person's mood. Specifically, the set $EL_i = \{el_{i1}, \ldots, el_{ir}\}$ contains the emoticons extracted from each tweet t_i.

Example In tweet "What a nice weather :)", the symbol ":)" is considered as emoticon, and thus $EL_i = \{:)\}$.

[1] Wordnet Affect: http://wndomains.fbk.eu/wnaffect.html

B. Tweet Semantic Refinement

The calculation of the sentiment orientation expressed by the tweets' words was based on the SentiWordnet dictionary [1] which assigns scores to each word on the basis of how positive or negative they are. Furthermore, the intensifiers that affected the meaning and intensity of a word were also considered. For example, in the expression "quite good", omitting the word "quite" would not express adequately the intensity of the sentiment expressed by a tweet containing this phrase. Thus, a list of intensifiers was used in which each intensifier is accompanied by a specific score (for example "hardly= -0.03", "very = 0.05", etc) [13].

The overall score of an emotional word et_{ir} that is combined with an intensifier $intens_j$ is calculated as:

$$SCI(et_{ir}) = (1 + score(intens_j)) \cdot score(et_{ir}) \tag{1}$$

where $intens_j$ is the intensity score of the intensifier and $score(et_{ir})$ is the sentiment score of the emotional word et_{ir} as defined in SentiWordnet lexicon. All scores are normalized in the interval $[0, 1]$.

In case there are valence shifters in a tweet (e.g. "Not", "Cannot", "Never","No", "Don't", etc) that reversed the meaning of a word, then the overall word's score is computed as:

$$SCI(et_{ir}) = 1 - score(et_{ir}) \tag{2}$$

Finally, another important parameter that was considered in our analysis, are the emoticons which are denoted as the EL set in our approach. Emoticons are used widely from users when they express their opinions/emotions. In our approach, we were based on a lexicon proposed by University of Maryland, Baltimore [2] which contains the 55 most used sentislangs, like ":-)", ":-(" scored into the interval $[0, \ldots, 1]$.

To define the overall score between a tweet t_i and a primary emotion e_j we have also used the *term frequency tf* of each emotional word et_{ir} of t_i tweet. The $tf(et_{ir}, e_j)$ is defined as the number of representatives of emotion e_j that matches the emotional word et_{ir} according to a stemmer.

Given the parameters previously discussed, the overall score $SC(t_i, e_j)$ between a tweet $t_i \in T$ and a primary emotion $e_j \in E$ is calculated as:

$$SC(t_i, e_j) = \frac{\sum_{\forall et_{ir} \in ET_i} tf(et_{ir}, e_j) \cdot SCI(et_{ir})}{\sqrt{\sum_{\forall et_{ir} \in ET_i} tf(et_{ir}, e_j)^2}} \tag{3}$$

Lemma 1: The score values SC (Equation 3) fluctuate in the interval $[0, \ldots, 1]$.

Proof: The SCI values that were defined in paper, fluctuate in the interval $[-0.4, \ldots, 0.4]$ and were rescaled in the interval $[0, \ldots, 1]$. Thus, in the SC definition, each $tf(et_{ir}, e_j) \cdot SCI(et_{ir})$ value fluctuates in the interval $[0, \ldots, 1]$. In Equation 3, the sum in the numerator is weighted in order to result in $[0, \ldots, 1]$ scale. ■

[2] http://www.csee.umbc.edu/

IV. The EmoGrabber framework

In our approach we use an emotional-driven clustering in order to analyse microblogging (tweets) datasets. The proposed approach is called EmoGrabber and it is a 3-step process, depicted in Figure 1, that results in the evaluation of sentiments expressed in tweets. Users' tweets are collected using the twitter streaming api and they are stored in a data repository.

Then, in the first phase of our approach, the preprocessing takes place. Appropriate linguistic analysis based on an English-term dictionary was used for the removal of words such as articles and numbers that bring no useful information for the affective analysis process. Moreover, words that are have no meaning and they are considered as noise, are being removed. Thus, at this phase of the process we are left with a set of tweets that contain words that may carry emotional information. To capture the exact emotional words we then use an emotional dictionary (in our case SentiWordnet) as well as the Lancaster Stemmer [3] to apply a further filtering and end up with words that are actually emotional and will contribute to our analysis.

Since we have isolated the emotional words for the whole set of tweets, then the affective analysis step follows. Following a process that considers existing valence shifters and intensifiers, the emotional dictionary and the primary emotions' representatives, a score of relation between each tweet and each one of the primary emotions is calculated, based on Equation 3.

At the last step of the EmoGrabber approach, the clustering algorithm step, the tweets are grouped using an emotion-driven clustering approach. Cluster centers are initialized in a way that tweets expressing the same high levels of a primary emotion are assigned to the same cluster. Therefore, useful conclusions can be derived, such as what is the prevailing emotion over a specific topic, which users share common feelings about the topic, what kind of expressions users use to express their emotions, etc.

Algorithm 1 EmoGrabberAlgo

Require: The set T of m tweets, the Emotional and Linguistic dictionaries, a set of primary emotions E and a vector C^* representing the clusters' centers.

Ensure: A set of k groups C consisting of tweets with scores in the same intensity levels with respect to specific primary emotions.

*/*Preprocessing*/*
$T^* = CleanData(T)$
$ER = FindRepresentatives(E)$
$ET = FindEmotionalWords(T)$
*/*Capturing Similarities*/*
$SCI(ET, intens) = CalculateScoreIntensifiers(ET, intens)$
$SC(T, E) = CalculateScore(T, E, SCI(ET, intens))$
*/*Clustering process*/*
$C = K\text{-}means(SC, k, C^*)$

Algorithm 1 is proposed in order to fulfill our goal and obtain a set of groups consisting of tweets that share the same levels of intensity for specific emotions. In the first step of the

[3] Lancaster Stemmer: http://www.comp.lancs.ac.uk/computing/research/stemming/

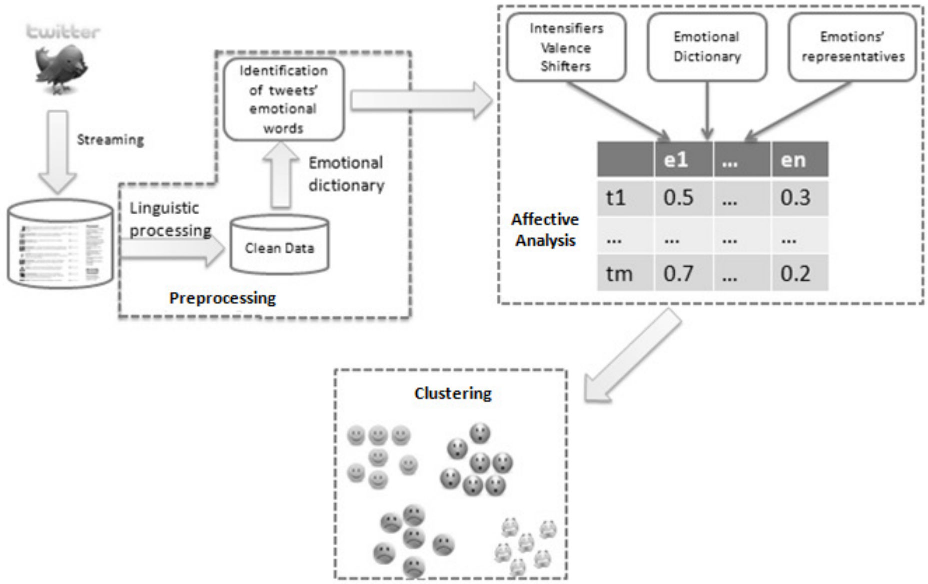


Fig. 1: The EmoGrabber Algorithmic Approach

EmoGrabberAlgo, a data preprocessing (line 1) takes place where the non valid words of the tweets and the words that carry no emotional information according to the emotional lexicon are removed (line 2). Moreover, the *ER* list (line 3) of the primary emotions' representatives as well as the *ET* list (line 4) of tweets' emotional words are formulated according to the process described in Section III. Next, in the second step of the proposed approach, the scores (Equation 3) between the tweets' emotional words and the set of primary emotions are calculated (line 7) while considering the way the intensifiers and valence shifters affect the words meaning (line 6). Thus, the result of the algorithm's two first steps is the formulation of a two-dimensional table which captures the relation between each tweet with each one of the primary emotions. This table along with the number of clusters and their centers are the input for the k-means algorithm that will drive the clustering process in a way that tweets having high level of the same emotion will be grouped together (line 8). The output of this step is the set of clusters C that contain tweets that share the same level of scores regarding specific primary emotions. The time complexity of the proposed approach is linear to the number of tweets, and that makes it scalable and affective.

V. Experimentation

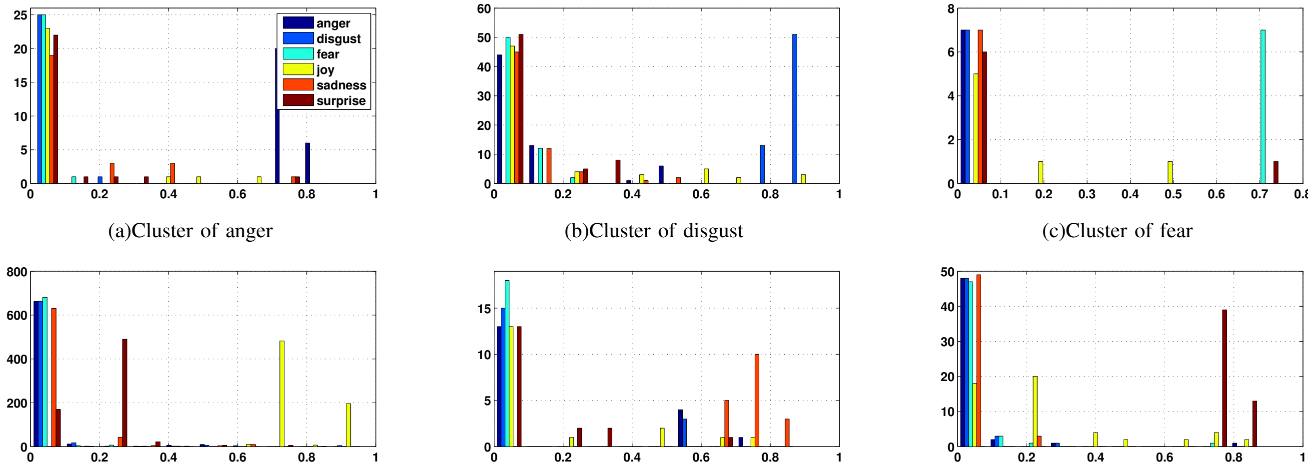
To evaluate the proposed approach we carried out experiments on various datasets derived from twitter (using the Twitter streaming api and a keyword-based filtering). Due to the lack of space, in this section we present the evaluation results for a dataset that contains tweets referring to the Christmas topic, located in the area of London. The dataset involved about 9500 tweets that were posted during the period December 2011 - February 2012. Although this is not a large scale dataset, the effectiveness of our approach is not affected since its complexity is linear to the number of tweets.

In Figure 2 we can see how tweets are assigned to clusters according to their prevailing emotion. The Y-axis corresponds to the number of tweets for each emotional score (X-axis). It is evident from all sub-figures, that in case the cluster contains tweets that express a negative emotion (e.g. fear, sadness, disgust), the degree of positive emotions they express (e.g. joy) is low and vice versa. For example, in Figure 2(a) that refers to the cluster containing tweets that mostly express anger, these tweets are characterized by low levels with reference to the emotion of joy. On the other hand, in Figure 2(d) which depicts the histogram of the cluster containing tweets with high level of joy, these tweets are characterized by low levels of intensity in terms of negative emotions such as anger, sadness and fear.

To validate our results we have used the factor analysis methodology, which is used to find latent variables or factors among observed variables. In our case we have six variables, one for each primary emotion, and the goal is to test if these variables are grouped together in a reasonable way. One possible "correct" categorization would be to group the negative emotions (anger, disgust, fear, sadness) and the positive (joy) and neutral (surprise) emotions in different classes, or it would be a further categorization where the neutral emotion arranged together with some of the other emotions.

Table I presents the correlation coefficients between each pair of variables (emotions). We observe that there is quite a strong relationship between the emotions of anger and disgust and between the emotions of joy and surprise. Additionally, there is a relationship, but it is not quite strong between the emotions of fear and sadness.

Table II shows the loadings of the six primary emotions on the three factors extracted. The higher the absolute value of the loading, the more the factor contributes to the variable. The gap on the table represents loadings that are less than 0.4 (Suppress small coefficients - Absolute value below: 0.40). Observing that table we see that there are three factors: joy-surprise, anger-disgust, and fear-sadness. This result is graphically depicted in Figure 3.



(a)Cluster of anger (b)Cluster of disgust (c)Cluster of fear

(d)Cluster of joy (e)Cluster of sadness (f)Cluster of surprise

Fig. 2: Tweets assignment to clusters

TABLE I: Correlation matrix

	Anger	Disgust	Fear	Joy	Sadness	Surprise
Anger	1.000	**0.347**	0.050	-0.125	0.045	-0.023
Disgust	**0.347**	1.000	0.049	-0.093	0.036	-0.036
Fear	0.050	0.049	1.000	-0.034	**0.051**	-0.002
Joy	-0.125	-0.093	-0.034	1.000	-0.031	**0.233**
Sadness	0.045	0.036	**0.051**	-0.031	1.000	-0.049
Surprise	-0.023	-0.036	-0.002	**0.233**	-0.049	1.000

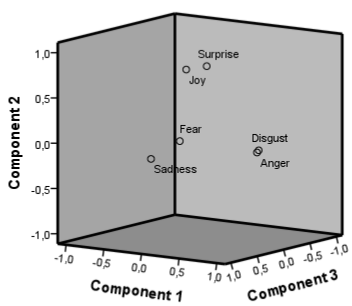


Fig. 3: Component Plot in Rotated Space

TABLE II: Rotated Component Matrix

	Comp 1	Comp 2	Comp 3
Anger	0.804		
Disgust	0.802		
Fear			0.637
Joy		0.750	
Sadness			0.800
Surprise		0.805	

Factor analysis was used in order to detect existing relations between the six primary emotions and thus to evaluate, based on the extracted structure, the proposed system. From the obtained results we conclude that our methodology successfully manages to capture the actual emotional nature of tweets, since the categorization of emotion is correct: joy-surprise, anger-disgust, and fear-sadness.

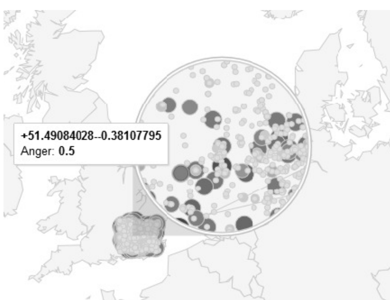
Finally, we have developed the EmoGrabber visualization tool which places the tweets as markers on a map and colors them according to the emotional intensity they express. In Figure 4, the visualization of tweets' intensity in terms of the six primary emotions is presented. The user can click on a marker to get the geo-coordinates of the user that posted the specific tweet as well as its emotional intensity. EmoGrabber is a web-based application and it was developed using the Geochart Google API [4]. In all sub-figures, the darker the color of a marker the higher the tweet's emotional intensity.

Such representation is of great importance in capturing branding success, diffusion in market and emotional states in relevance to different topics (such as ceremonies, events), as expressed by people. Such maps, for instance, might be of use to an authority's scenario where freeing fear policies, during Christmas period, might be suggested (in case the intensity of the emotion of fear is strong enough). For example, the city mayor might develop a campaign against fear and depression, starting earlier than Christmas time.

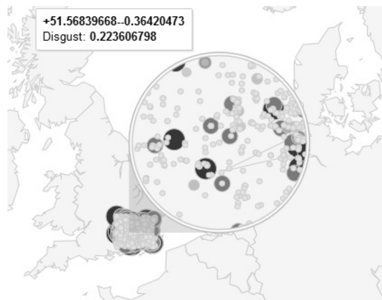
VI. Conclusion

In this paper a methodology is proposed for capturing people's emotions as these are recorded in users' posts in Twitter. Using an automated technique, via lexicon-based methodology,

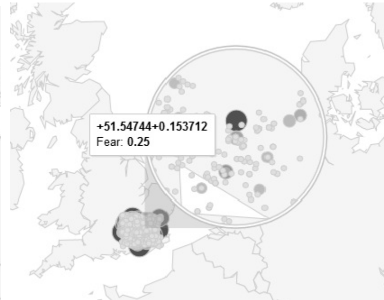
[4]Geochart: https://developers.google.com/chart/interactive/docs/gallery/geochart



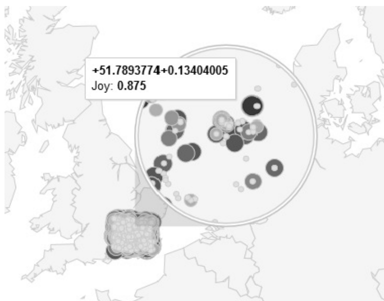
(a)Tweets' anger



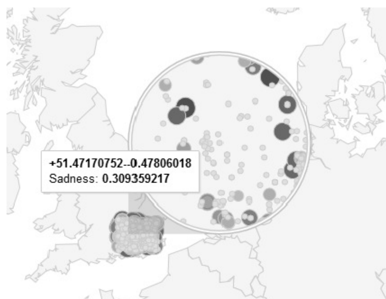
(b)Tweets' disgust



(c)Tweets' fear



(d)Tweets' joy



(e)Tweets' sadness

(f)Tweets' surprise

Fig. 4: The EmoGrabber visualizer

we provide the means for an automated real-time analysis, thus eliminating the need for manual analysis. This work addresses the challenge to go beyond the typical positive and negative analysis, by providing a more fine-grained analysis with a six class emotion spectrum along with the capturing of the emotions' intensity. The evaluation of the experimental results proves that the proposed method manages to capture efficiently the emotions expressed in tweets as well as their intensity.

The emotion patterns detected in microblogging media are of interest to a wide range of markets since an unprecedented level of analytics is majorly useful for a variety of stakeholders. Companies can track the current trends and adjust their communication strategies, increasing in this way their customers' level of satisfaction. Many people express their political views in microblogging services, so the analysis of social data can be quite important for capturing the public pulse.

In the future we aim to improve the proposed system by taking into consideration more advanced linguistic parameters, in order to capture in a more accurate way users' emotions. Moreover, we plan to try and improve the measure that was used to calculate the relations between the representative words and the emotion words of each tweet by including more semantic information. Finally, we plan on working towards a more attractive application framework for the visualization of the analysis results that will be more enriched with topics referring to various geographical regions.

References

[1] Baccianella S et al., *SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.* In P. of 7th Conference on Language Resources and Evaluation, Valletta, MT, 2200-2204, (2010).

[2] Benamara F. et al., *Sentiment Analysis: Adjectives and Adverbs are better than Adjectives Alone.* ICWSM 2007 Boulder, CO USA, (2007).

[3] Bollen, J. et al.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM11, arXiv: 0911.1583, (2011).

[4] Ekman, P., Friesen, W. V., Ellsworth, P., *What emotion categories or dimensions can observers judge from facial behavior?.* In P. Ekman, Emotion in the human face, 39-55. Cambridge University Press, (1982).

[5] Fellbaum,*WordNet, an electronic lexical database.* The MIT Press (1990).

[6] Gill A. J. et al., *Identifying Emotional Characteristics from Short Blog Texts.* Proc. of the 30th Annual Conference of the Cognitive Science Society, Washington DC, 2237-2242, (2008).

[7] Hu M. and Liu B.,*Mining and Summarizing Customer Reviews.* Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining, Seattle, WA, USA, (2004).

[8] O Connor B. et al., *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.* Proc. of the International AAAI conference on Weblogs and Social Media, Washington DC, (2010).

[9] Pak A., Paroubek P., *Twitter as a corpus for Sentiment Analysis and Opinion Mining.* Proc. of the 7th conference on Int Language Resources and Evaluation, 1320-1326, (2010).

[10] Parikh R., Movassatc M.,*Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques.* Available at: http://nlp.stanford.edu/courses/cs224n/2009/fp/19.pdf, 1-78, (2009).

[11] Strapparava C., Valitutti A.,*WordNet-Affect: an affective extension of-WordNet.* In Proceedings ofthe 4th International Conference on Language Resources and Evaluation, Lisbon, 1083-1086, (2004).

[12] Subrahmanian V.S., and Reforgiato D., *Ava: Adjective-verb-adverb combinations for sentiment analysis.* Intelligent Systems, 43-45, (2008).

[13] Taboada M., Brooke J., Tofiloski M., Voll K., Stede M., *Lexicon-based methods for sentiment analysis.* Comput. Linguist., 267-307, (2011).

[14] Tsagkalidou K., Koutsonikola V., Vakali A., and Kafetsios K.,*Emotional aware clustering on micro blogging sources.* In Proceedings of J. Spatial Information Science, 387-396, (2011).

[15] Turney, P.,*Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* Proc. 40th Annual Meeting of the Association for Computational Linguistics, 417-424, (2002).

[16] Yessenalina A., Cardie C.,*Compositional Matrix-Space Models for Sentiment Analysis.* Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing, 172-182, (2011).