Emotional Aware Clustering on Micro-blogging Sources

Katerina Tsagkalidou[1], Vassiliki Koutsonikola[1],
Athena Vakali[1], and Konstantinos Kafetsios[2]

[1] Department of Informatics
Aristotle University
54124 Thessaloniki, Greece
[2] Department of Psychology
University of Crete
GR74100 Rethymno, Greece

Abstract. Microblogging services have nowadays become a very popular communication tool among Internet users. Since millions of users share opinions on different aspects of life everyday, microblogging websites are considered as a credible source for exploring both factual and subjective information. This fact has inspired research in the area of automatic sentiment analysis. In this paper we propose an emotional aware clustering approach which performs sentiment analysis of users tweets on the basis of an emotional dictionary and groups tweets according to the degree they express a specific set of emotions. Experimental evaluations on datasets derived from Twitter prove the efficiency of the proposed approach.

Keywords: Microblogging services, sentiment analysis, web clustering.

1 Introduction

With the advent of Web 2.0 and social applications, millions of people broadcast their thoughts and opinions on a considerable variety of topics which are collectively called the *User Generated Content* (UGC) [10]. In particular, social networks and blogs provide an increasingly popular way of online communication by which users can interact and broadcast their personal thoughts. Therefore, these applications contain highly opinionated personal commentary and the new social media offer a unique look into people's emotion-laden reactions and attitudes.

In this paper we propose a method that employs a clustering technique in order to group users according to the affective content of opinions as expressed in a microblogging service and for this reason we have used datasets derived from the Twitter microblogging application. Our purpose was to test a tool that will automatically extract the affective orientation of users' posts and create groups of users that share common viewpoints towards a topic. Therefore, the affective element of people's attitudes, emotions and beliefs towards a specific topic could

S. D´Mello et al. (Eds.): ACII 2011, Part I, LNCS 6974, pp. 387–396, 2011.

be evaluated. Our method involves the creation of collections of users' posts that refer to specific topics and then based on an emotional dictionary we evaluated users' posts on the basis of a model of eight primary emotions [7]. Next, we applied a clustering algorithm which groups together users' tweets that present similar correlation to a set of primary emotions. Hence, we manage to categorize users tweets and therefore users according to their opinion in regard to a topic.

Given unique challenges in dealing with users' blog posts subjective manner of expression, appropriate pre-processing should be performed to result in suitable data structures for clustering analysis. The experimentation results show that the proposed framework managed to identify blog posts with similar evaluative or affective content towards a topic. To the authors' knowledge no emotional aware clustering approach has been proposed to analyze and identify users' emotions as expressed in blog posts.

The contribution of our work can thus be summarized to the following tasks:

- define the tweet-list, a convenient structure for further data analysis, to represent a user's blog post;
- create an extended emotional dictionary by enriching an opinion lexicon provided by the UMBC university[1] with synonymous words from WordNet;
- propose a similarity measure which evaluates the relation of a blog post to an emotion;
- propose a clustering framework to group blog posts according to the closeness they present to certain emotions.

2 Related Work

Typically, affective evaluations of internet content have been made through sentiment analysis. Sentiment Analysis (or Opinion Mining) is the computational study of opinion, sentiment and emotion of a text [10]. The research area of the sentiment analysis has significantly grown up mostly because of the web 2.0 technologies which have changed the way that people express their views and opinions in the web [12]. Most of the previous research work [14,9] proposed methods for the sentiment classification of product or movie reviews coming mostly from forums and blogs. Short messages from micro blogging sites, like Twitter, are different from reviews mainly because of their purpose and function. Reviews represent summarized thoughts of an author for a particular entity while tweets are more general, casual and abstract.

There is a number of interesting studies that have used the Twitter as a source for sentiment analysis. In [2], the authors found that the surveys' results about consumer confidence and political opinion correlate with sentiment word frequencies in tweets, and therefore they proposed text stream mining as a substitute for traditional polling. In [13], a machine learning algorithm was employed to compare multinomial naive Bayes, maximum entropy classifier, and a linear

[1] UMBC opinion lexicon: http://www.cs.umbc.edu/courses/331/spring10/2/hw/hw7/hw7/data/

support vector machine for the sentiment classification of tweets to positive and negative classes, while in [11], the classification based on the multinomial naive Bayes classifier that uses N-gram and POS-tags as features was evaluated. The above studies exhibited comparable accuracy on their test datasets, but they have certain differences concerning the features they used.

The method in [3] shows a methodology that demonstrates the temporal dynamics of sentiments in reaction to a live debate video. While viewers could only see opinions sequentially while watching, the proposed methodology offered a visual representation of tweets collection and provided the opportunity to understand the overall sentiment (positive and negative) of micro bloggers during the event. To accomplish this they proposed a method for an automatic collection of a corpus that has been used to train a sentiment classifier. The classifier also is based on the multinomial Naive Bayes classifier that uses N-gram and POS-tags as features.

In [1] a method was proposed which performs a sentiment analysis of tweets using an extended version of a well established psychometric instrument, the Profile of Mood States (POMS) and measures six individual dimensions of mood, Tension, Depression, Anger, Vigour, Fatigue, and Confusion. They extract a six-dimensional vector representing the tweet's mood and aggregate mood components on a daily scale comparing their results to the timeline of cultural, social, economic, and political events that took place in that defined period of time.

In this paper we adopted an approach following [7], that focuses on eight primary emotion dimensions: acceptance, fear, anger, joy, anticipation, sadness, disgust and surprise and their synonymous adjectives.This approach is based on the assumption that some basic emotional words which constitute the emotions representatives can determine the sentiment (or, at least some indication) of a sentence by analyzing how similar the sentence's words are with the emotion's representatives.

3 Problem Formulation

The main purpose of our analysis was to determine the orientation of an opinion expressed by web users in a blog post or else a tweet. Table 1 summarizes the basic symbols notation used in this paper.

Table 1. Basic Symbols Notation

Symbol	Definition
m, n, l, p	Number of tweets, tweet's words, primary emotions, emotion's representative words (respectively)
T	Tweets' set $\{t_1, \ldots, t_m\}$
TL_i	The set of words $\{w_1, \ldots, w_n\}$ contained in tweet t_i
E	The set of primary emotions $\{e_1, \ldots, e_l\}$
ER_i	The set of representative words $\{r_{i1}, \ldots, r_{ip}\}$ of the primary emotion e_i
ET_i	The set of emotional words $\{et_1, \ldots, et_d\}$ of tweet t_i

Let t_i denote a user tweet and $T = \{t_1, \ldots, t_m\}$ represents a set of m tweets that refer to a specific topic. Given that a tweet consists of words (140 at most) we define the notion of TWEET-LIST.

Definition 1 (THE TWEET-LIST). *The Tweet-List TL_i is a list used to represent the tweet t_i and it is defined as $TL_i = \{w_1, \ldots, w_n\}$: w_j is a word of t_i and $n \leq 140$.*

Example 1. If we assume that a user publishes the post "I really want a varsity-style jacket" then the respective Tweet-List will be $TL = \{I, really, want, a, varsity - style, jacket\}$.

In the proposed model our purpose is to group a set of tweets on the basis of their relation with specific primary emotions. Let $E_i = \{e_1, \ldots, e_l\}$ denote the set of l primary emotions and $ER_i = \{r_{i1}, \ldots, r_{ip}\}$ the set of p words which act as representatives of emotion e_i.

To calculate the similarity between a tweet t_i and a primary emotion e_j we define two types of similarities: the semantic similarity *Sema* and the sentiment similarity *Senti*.

For the estimation of the semantic similarity *Sema* between a tweet's word w_x and a primary emotion e_j we need to compute the semantic similarities *SeS* between the tweet's word and the emotion's representatives. In other words we need to define a measure that will capture semantic similarity between two words. Thus, we have to use external resources (i.e. web ontologies, thesauri, etc) and a mapping technique between the tweet and the emotion. In our work, we adopted the approach described in [15], due to its straightforward application to our data, according to which the semantic distance between two concepts is proportional to the path distance between them. For example, let w_x be a tweet's word and r_{yz} be the z-th representative of emotion e_y. Given that these are two words for which we want to find the semantic similarity, let $\overrightarrow{w_x}$ and $\overrightarrow{r_{yz}}$ be their corresponding mapping concepts via an ontology. Then, their *Semantic Similarity* SeS is calculated as:

$$SeS(w_x, r_{yz}) = \frac{2 \times depth(LCS)}{[depth(\overrightarrow{w_x}) + depth(\overrightarrow{r_{yz}})]} \tag{1}$$

where $depth(\overrightarrow{w_x})$ is the maximum path length from the root to $\overrightarrow{w_x}$ and LCS is the least common subsumer of $\overrightarrow{w_x}$ and $\overrightarrow{r_{yz}}$.

Then, the semantic similarity *Sema* between a tweet's word w_x and an emotion e_j is defined as the maximum *SeS* of the similarities between the word w_x and the emotion's representatives.

$$Sema(w_x, e_j) = max_{z=1,\ldots,p}(SeS(w_x, r_{jz})) \tag{2}$$

To compute the *Sentiment Score Senti* we have created an extended emotional dictionary, using two dictionaries provided by the University of Maryland. The first one contains 18.536 words and small phrases, as nouns, verbs, adjectives etc, scored with a value between $[-1, 1]$ that indicate their sentiment orientation. The second one contains the 55 most used sentislangs, like :-), :(, etc, scored into the same interval. We used the first dictionary as a seed word list and by looking on WordNet synsets, we derived all the synonyms and created an enriched opinion lexicon. We adopt the assumption of Esuli and Sebastiani [4] and we consider that if a synonymous word or phrase is found in WordNet, then it would have the same semantic meaning. So if a synonymous word exists then we add it into our lexicon assigning the same score value as the seed word. By this bootstrapping procedure we managed to create an extended emotional lexicon of 28.249 words and phrases.

Example 2. In the extended emotional lexicon the words "upstairs", "upstair", "up_the_stairs", ''on_a_higher_floor" are assigned the score value 0.625 while the words "overacting", "overact", "ham_it_up', "ham" the score value -0.75.

Thus, the *Sentiment Score* is directly extracted from the extended emotional lexicon and expresses a word's emotional intensity.

$$Senti(w_i) = FindScore(w_i, Emotional\ Lexicon) \tag{3}$$

Moreover, we define an *Emotional* word w_i as the word that presents an emotional intensity i.e. $Senti(w_i) \neq 0$. We then define the Tweet Emotional Set of tweet t_i, ET_i as the set of emotional words that belong to tweet t_i.

Definition 2 (The Tweet-Emotional-Set). *The Tweet-Emotional-Set ET_i is defined as $ET_i = \{w_j : w_j \in t_i, senti(w_j) \neq 0, 1 \leq j \leq n\}$*

Both the semantic similarity *Sema* (Equation 2) and the *Sentiment Score Senti* (Equation 3) are combined in order to compute the total similarity $Sim(t_x, e_y)$ between a tweet t_x and a primary emotion e_y. The $Sim(t_x, e_y)$ considers both the relation of a tweet's words to a primary emotion's representatives and the emotional intensity of the tweet's words. It is defined as:

$$Sim(t_i, e_j) = \frac{\sum_{x=1,\ldots,n}(Sema(w_x, e_j) \cdot Senti(w_x))}{|ET_i|} \tag{4}$$

Lemma 1. *The similarity values Sim fluctuate in the interval* $[-1, 1]$.

Proof. Given that the *SeS* similarity (Equation 1) fluctuates in the interval $[0, 1]$, the definition of *Sema* (Equation 2) indicates the same fluctuation interval. Moreover, the *Senti* similarities (Equation 3) range in the interval $[-1, 1]$ and therefore the numerator in the *Sim* definition (Equation 4) varies in $[-n, n]$. Dividing with the number of the tweet's emotional words we result in the $[-1, 1]$ fluctuation interval for the *Sim* similarity.

At this point, we can define the Emotion-aware Clustering problem as follows:

*Problem 1 (*Emotion-aware Clustering). Given a set T of tweets, a set E of primary emotions and their representatives ER and an integer value k, find k subsets $C_1, C_2, \ldots, C_k$ of tweets which result in the maximization of the quantity $\sum_{x=1}^{k} \sum_{t_i, t_j \in C_x} Sim(t_i, t_j)$, $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

4 The SentiTweetAlgo

The proposed clustering framework involves a three-step process which applies an emotional and semantic aware analysis on data retrieved from Twitter, based on a specific query/topic. At the first step, the preprocessing step, the data cleaning is performed which involves the removal of the tweets' words that are not semantically valid. Moreover, words that do not provide any useful information for our mining process, e.g. articles, numbers and links are removed. As a source of semantic information for terms concepts, we employ the lexicon WordNet [5], which stores english words in hierarchies, depending on their cognitive meaning.

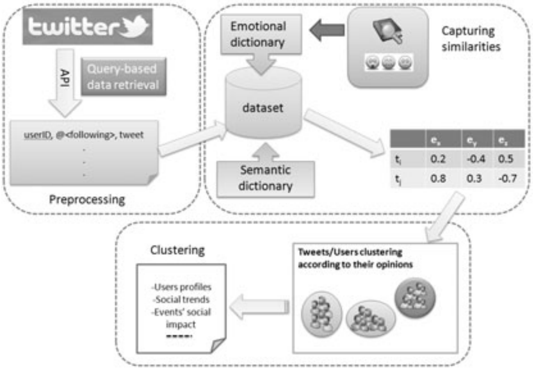


Fig. 1. The proposed framework

At the second step of the similarities capturing, the "clean tweets" derived from the previous step are evaluated in terms of a set of primary emotions. Given the semantic dictionary (Wordnet) and the extended emotional dictionary, which was created as described in Section 3, we calculate the relation of each tweet to a set of primary emotions. For this, we adopted an approach following [7], that focuses on eight primary emotion dimensions: acceptance, fear, anger, joy, anticipation, sadness, disgust and surprise and their synonymous adjectives. This approach is based on the assumption that some basic emotional words which constitute the emotions representatives can determine the sentiment (or, at least some indication) of a sentence by analyzing how similar the sentence's words are with the emotion's representatives. The model is in keeping with seminal activation-evaluative models of emotion [6] and has demonstrated to be particularly suited for computational and text analysis, with the emotions reliably discerned in asynchronous text.

Given the calculated relations between tweets and the eight primary emotions, we proceed to the third step of the algorithm, the clustering, where we apply the K-means clustering algorithm which assigns tweets (while the assignment could be extended to users) into k groups. K-means assigns tweets into groups in time linear on the number of tweets: O(m). Tweets of a group are characterized by similar expressions towards the set of primary emotions. For example, a cluster may contain those tweets that express fear and sadness about the earthquake in Fukushima, Japan. The analysis of the obtained clusters can be beneficial among others, in case of users' profile extraction and identification of social trends and social events' impact. Figure 1 presents the proposed clustering framework.

5 Experimentation

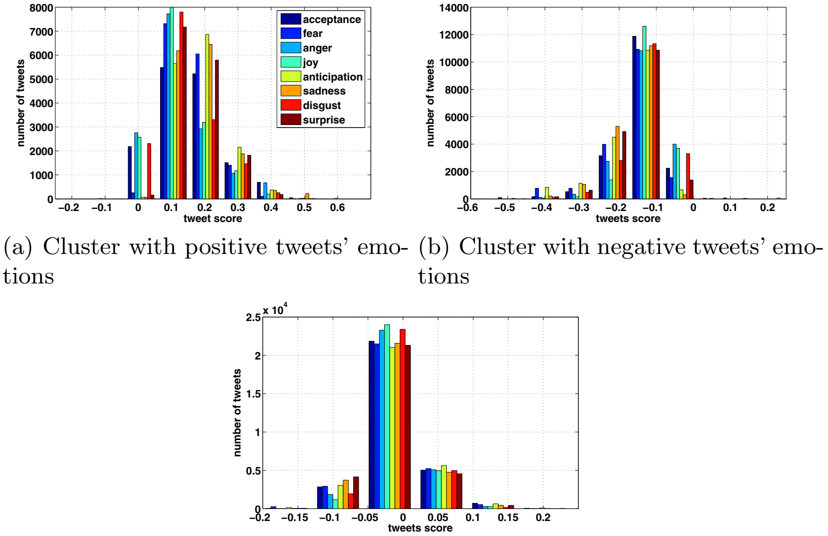
To evaluate the proposed approach we carried out experiments on various datasets derived from twitter (using the Twitter streaming api and a keyword-based filtering). A keyword-driven dataset collection normally results in better quality of data since chatty tweets without a specific meaning may be avoided. To this context, we have experimented with collections of tweets about topics that are expected to trigger different emotional behavior :"christmas" (for seasonal feelings), "lady gaga" (for idols followers) and "wikileaks" (for political opinions). Due to the lack of space we will present the results for the Christmas (of 65166 tweets) dataset which proved to contain quite emotional tweets. At the preprocessing phase tweets that contained no emotional words were removed resulting in a dataset with 63752. Thus, only 2% of the total tweets carried no emotional information. The experimental results presented below are indicative and were obtained for a number of clusters $k = 3$. This k value allows us to provide graphical representations of all clusters without exceeding the spatial restrictions.

Our main goal is to examine how well the proposed approach evaluates the tweets' emotional aspect, i.e. we want to examine the efficiency of our extended emotional dictionary and the proposed *Similarity Score Sim* (Equation 4). Thus, we proceed to a qualitative evaluation of the obtained clusters by examining the tweets assigned to each cluster in terms of the eight primary emotions.

To this context, Figure 2 depicts the number and the score of the tweets assigned to each cluster. Specifically, Figure 2(a) shows the distribution of tweets' scores in terms of the 8 primary emotions. As we can see this cluster contains tweets the majority of which is characterized by a positive score value for all of the emotions. On the contrary, Figure 2(b) depicts the second cluster whose members are tweets that denote negative score in terms of the emotions. Finally, Figure 2(c) includes the tweets with score near 0, which can be characterized as neutral tweets.

Next, we proceed to the creation of the clusters tag clouds (Figure 3) in order to provide an insight of the tweets content. Each tag cloud contains a set of the most frequent words that exist in the respective cluster's tweets. Thus, the first cluster (Figure3(a)) that contains positive tweets is represented by emotionally "strong" words such as *great, love, kiss* and *happy.* The second cluster (Figure 3(b)) is represented by words such as *separately, distance, working and little* which signify emotional negativeness. The third cluster (Figure 3(c)) contains words that do not exhibit a particular intensity.

Table 2 presents some sample tweets that exhibit high positive and negative scores in terms of the primary emotions. Our purpose is to examine whether the

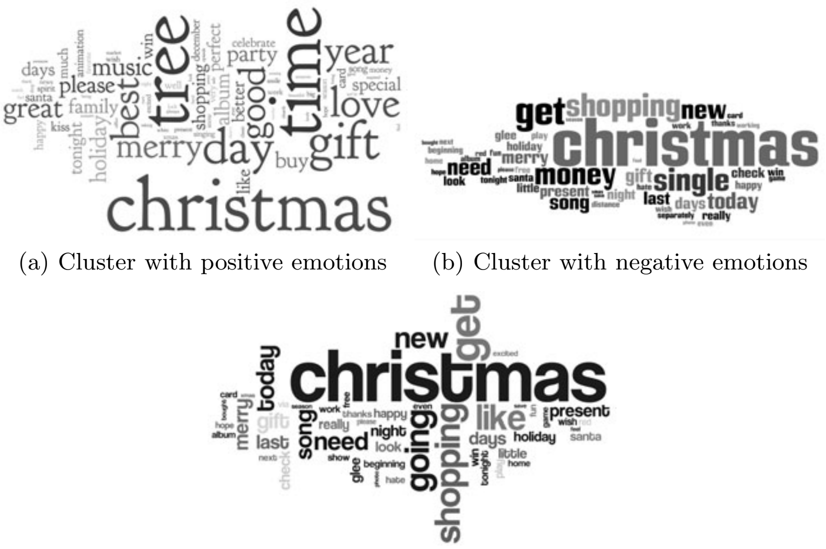


(a) Cluster with positive tweets' emotions

(b) Cluster with negative tweets' emotions

(c) Cluster with neutral tweets' emotions

Fig. 2. Tweets assignment to clusters



(a) Cluster with positive emotions

(b) Cluster with negative emotions

(c) Cluster with neutral emotions

Fig. 3. Clusters' tag clouds

Table 2. Sample tweets

Acceptance +	oh lovely christmas cake!?
Acceptance -	i really wish i understood christmas music goes annoying cheerful one week
Fear +	If I eat anymore chocolate oranges this Christmas I fear I will indeed become one!
Fear -	Who's brave enough to admit, that they secretly like to sing their favorite #Christmas songs, while driving alone in their car? Guilty ;)
Anger +	517/533 films on tv over Christmas will be repeats! Outrage!
Anger -	Investigation Reveals Christmas Cruelty Towards Reindeer
Joy +	At a Christmas concert supporting my niece. She is doing brilliantly.
Joy -	it is more that I hate people at Christmas. More than other times of the year that is
Anticipation +	merry christmas me! i've been listening the skillet tribute album and wow!! check it out! :)
Anticipation -	can't wait for new commercial merry christmas
Sadness +	sad i'm more excited ”in darkness and light” box set than i am christmas
Sadness -	one thing i love christmas time
Disgust +	i hate christmas they suck
Disgust -	one thing i love christmas time
Surprise +	I love surprises I can't wait for Christmas!!!
Surprise -	At this boring work Christmas party and we gotta go back to work!!!!

extended emotional dictionary and the proposed similarity *Sim* manages to guide efficiently the clustering process. The tweets which are indicatively presented show that words inside the tweets that carry emotional information significantly affect the overall tweet *Sim* score. For example words such as *lovely, annoying, favorite, sad, hate* and *boring* considerably determine the tweet-emotion relation.

The above indicative (due to the lack of space) discussion clearly shows that the proposed approach can be used in order to automatically retrieve a set of tweets that share common emotions towards a topic. To guide more the clustering process and result for instance to a cluster with positive tweets scores in terms of specific emotions, we could execute the kmeans algorithm by providing a set of seeds, i.e. the initial cluster centers.

6 Conclusions and Future Work

The social web growth and the corresponding rise in available emotional text over the past years has led to an increased interest in sentiment analysis. In this paper we propose an emotional aware clustering approach which aims to group tweets, and therefore users, according to their opinions as expressed in a microblogging application. The results of the proposed framework can be very useful for the efficient extraction of users profiles and the identification of social trends and events' impact. They can also provide an unprecedented level of analytics for companies, politicians and other public services. As shown in [8] micro blogging services are a potentially rich source for companies to explore as part of their overall branding strategy. Customer perceptions and purchasing decisions appear increasingly influenced by social networking services, since these act as trusted sources of information, insights and opinions.

In the future we aim to improve the overall process incorporating other sentiment lexicons to enhance the data cleaning process and eliminate noise. This work could expand to explore the relative impact of positive and negative affect [6] in internet attitudinal material, employing more advanced linguistic techniques. Moreover, we plan to apply more advanced mining techniques that will result in more automatic and accurate identification of users emotions.

References

1. Bollen, J., Pepe, A., Mao, H.: Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In: ICWSM 2011, arXiv: 0911.1583 (2011)
2. OĆonnor, B., Balasubramanyan, R., Routledge, B., Smith, N.: From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In: Int. AAAI Conf. on Weblogs and Social Media, Washington DC, pp. 122–129 (2010)
3. Diakopoulos, N., Shamma, D.: Characterizing Debate Performance via Aggregated Twitter Sentiment. In: ACM Conf. on Human Factors in Computing Systems (CHI), Atlanta Georgia, pp. 1195–1198 (2010)
4. Esuli, A., Sebastiani, F.: PageRanking WordNet Synsets: An Application on Opinion Mining. In: 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, CZ, pp. 424–431 (2007)
5. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
6. Feldman Barrett, L., Russell, J.A.: Independence and bipolarity in the structure of affect. J. Personality and Social Psychology 74, 967–984 (1998)
7. Gill, A.J., French R.M., Gergle, D., Oberlander, J.: Identifying Emotional Characteristics from Short Blog Texts. In: 30th Annual Conf. of the Cognitive Science Society, Washington DC, pp. 2237–2242 (2008)
8. Jansen, B., Zhang, M., Sobel, K., Chowdury, A.: Micro-blogging as online word of mouth branding. In: 27th Int. Conf. Extended Abstracts on Human Factors in Computing Systems, Boston, pp. 3859–3864 (2009)
9. Hu, M., Liu, B.: Mining and Summarizing Customer Reviews.In: 10th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining, Washington USA, pp. 168–177 (2004)
10. Liu, B.: In: Indurkhya, N., Damerau, F.J. (eds.) Handbook of Natural Language Processing, 2nd edn., Goshen, Connecticut, USA (2010)
11. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: 7th Conf. on Int. Language Resources and Evaluation, Malta, pp. 1320–1326 (2010)
12. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
13. Parikh, R., Movassate, M.: Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques
14. Turney, P.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: 40th Annual Meeting of the Association for Computational Linguistics, Philadephia, pp. 417–424 (2002)
15. Wu, Z., Palmer, M.: Verm semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, pp. 133–138 (1994)