Exploiting Emoticons to Generate Emotional Dictionaries from Facebook Pages

Hanen Ameur, Salma Jamoussi and Abdelmajid Ben Hamadou

Abstract During the first events of the Tunisian revolution, the social network, Facebook, played a key role in Tunisia and everywhere in the world. It became the first political tool that allows the Tunisian people to share trending news in actual time. Facebook provides the opportunity for users to comment on the news by expressing their sentiments. In this paper, we focus on emotion analysis of Tunisian Facebook pages. To do this, we first collect comments from the Facebook pages in order to analyze sentiments written in Tunisian dialect. Then, we propose a new method for emotional dictionaries construction. In fact, we distinguish nine emotional classes: surprised, satisfied, happy, gleeful, romantic, disappointed, sad, angry and disgusted. At this step, we focus on the use of emotion symbols as indicators of sentiment polarity. Finally, we present the experimental results of our method. Our system achieves effective and consistent results.

Keywords Sentiment analysis · Emotion analysis · Emotional dictionaries · Tunisian dialect · Emotion symbols · Political lexicon

1 Introduction

Since the beginning of political upheavals and the triggering of the Tunisian revolution, social networks, especially Facebook, play a leading role in the political life in Tunisia. In fact, Facebook allows exchanging news in actual time, everywhere in the

H. Ameur (✉) · S. Jamoussi · A. Ben Hamadou
Multimedia InfoRmation Systems and Advanced Computing Laboratory,
MIRACL-Sfax University, Sfax-Tunisia Technopole of Sfax,
Av.Tunis Km 10 B.P. 242, 3021 Sfax, Tunisia
e-mail: ameurhanen@gmail.com

S. Jamoussi
e-mail: salma.jamoussi@isimsf.rnu.tn

A. Ben Hamadou
e-mail: abdelmajid.benhamadou@isimsf.rnu.tn

© Springer International Publishing Switzerland 2016

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies 2016*,
Smart Innovation, Systems and Technologies 57,
DOI 10.1007/978-3-319-39627-9_4

world. Today, we are witnessing the appearance of many popular Facebook pages that have political aspect. These pages allow the users to ask important questions about news and express their opinions and sentiments freely. Hence, the large quantity of exchanged political texts encouraged us to use sentiment analysis techniques to collect and treat the users' sentiments about the discussed political subjects. It seems pressing to develop tools for analyzing and investigating the conflicts of ideas and the variability of sentiments of Tunisian people. In addition, it is very interesting to know the lexicon of words used by commentators in some Facebook pages. For this reason, we propose, in this paper, to use Facebook comments as a source of textual data for the construction of emotional dictionaries (lexicon). Therefore, we first thought of collecting automatically Facebook comments to realize a fine sentiment classification task "emotion analysis".

The main difficulty of handling corpus collected from Facebook pages, is the wide variety of comments. Indeed, the comments present a great diversity, whatsoever, on the level of the writing style or their size. Furthermore, Facebook offers to users, another writing style which is closer to SMS language than to the language used by journalists and professionals. This new style is characterized by the presence of emoticons, elongated specific words and by a simplified syntax with misspellings and unpronounced characters. In addition, in the Tunisian pages, users write their comments using a mixture of three languages (French, standard Arabic and Tunisian dialect) and they are able to easily switch between them. Tunisian dialect is generally written by using Latin alphabets and numeric digits. All this diversity presents many challenges in sentiment analysis when it deals with Tunisian comments. The sentiment dictionaries "lexicon" creation is a very important and difficult task to achieve. In fact, most of the sentiments analysis methods are based on sentiment lexicons to classify subjective comments. However, from our best knowledge, there is no existing sentiment lexicon for the Tunisian language. Furthermore, we cant use English external resources (like WordNet affect[1]) because it requires a translation step which can affect the actual meaning of words. In this paper, we propose a new automatic method based on the emotion symbols for the construction of emotional dictionaries. Due to the richness of sentiments in the emotion symbols, we consider nine emotional classes (*surprised*, *satisfied*, *happy*, *gleeful*, *romantic*, *disappointed*, *sad*, *angry* and *disgusted*) based on expert judgments.

This paper is organized as follows. We review the related work in the next section. We then present our corpus acquired from the Tunisian Facebook pages and the step of its preprocessing in the Sect. 3. Next, in Sect. 4 we elaborate on the principle of our proposed method for automatic construction of emotional dictionaries in Tunisian dialect. Finally, we report the experimental results and conclude the paper with future works.

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

2 Related Work

There are many research studies based on using social networks to analyze sentiments and on the construction of sentiment dictionaries lexicon of words. To our best knowledge, there hasn't been any study conducted on Tunisian words and how to classify them according to the sentiments which they express in social networks, especially Facebook. In the literature, many works try to address the problem of sentiment dictionaries construction by extracting at least two vocabularies groups. One expresses the positive sentiments and the other expresses the negative sentiments. For this purpose, there are four techniques: the manual method, the dictionaries-based method, the corpora-based method and the method combining the last two ones.

Some sentiment lexicon has been constructed manually by experts [15]. As these lexicons often contain thousands of words, their manual creation is therefore very difficult, expensive and time-intensive. Other researchers (e.g. [9]) have proposed to construct the sentiment lexicon based on external linguistic resources that handle semantic relations (synonymy and antonymy), such as WordNet, SentiWordNet and ANEW. The idea is to classify words based on other words whose semantic orientation is known (called seed), by applying bootstrapping algorithms [14]. Kamps and Marx [9] have proposed a semantic distance to measure the shortest distance between the examined word and the seed words whose their valences are known. To deal with the lack of information about semantic relations between words, other researchers have proposed based on the information present in a corpus (annotated or not). Douglas and Christopher [6] have concentrated on the coordinating conjunctions present between the words such as: and, but, either-or, or, etc. In other words, if for example, two words are separated by the conjunction "and", they necessarily have a similar polarity. They consist of counting the number of times which the examined word appears beside the words already classified "seeds". Kim and Hovy [10] and Ameur and Jamoussi [3] have proposed a hybrid method combining techniques based on dictionaries and those based on corpus to construct a sentiment lexicon. Kim and Hovy [10] constructed a lexicon containing a large number of words (verbs and adjectives) carrying sentiments from three defined sets of words. Thereafter, they merged these three sets using an averaging method. Ameur and Jamoussi [3] used emoticons to differentiate between the positive sentiments (indicated by positive emoticons) and the negative sentiments (indicated by negative emoticons).

Instead of classifying the text into three classes only (i.e. positive, negative and neutral), other studies extended the sentiment analysis methods in order to treat and analyze emotions [12]. Mihalcea and Liu [11] have classified blog posts into two particular emotion classes (happiness and unhappiness). These blog posts are self-annotated by the blog writers with happy and sad mood labels.

Some researchers have analyzed the emotions of text focused on the six basic emotions identified by Ekman [8] (positive, negative, fear, joy, surprise, hate, disgust). Alena et al. [2] has used a rule-based method for determining Ekman's basic emotions in the sentences in blog posts. Balabantaray et al. [4] took into account all these basic emotions, with the addition of the neutral class in order to analyze the subjectivity of the text. Duyu et al. [7] have proposed a method which aims to classify emotion bearer in tweets as (happy, sad, angry or surprise) using pseudo-labelled data with emoticons. Solakidis et al. [13] applied the emotion classification task on multilingual data, focusing on documents written in Greek. They identified the polarity of the text (neutral, negative or positive); and the emotion expressed through the positive sentiments (joy and love) and the negative sentiments (anger and sadness).

Recently, there has been some works on Arabic sentiment analysis mainly concerning about the construction sentiment lexicons (e.g. [1]). Abdul-Mageed and Diab [1] presented SANA, a subjectivity and sentiment lexicon for Arabic. The lexicon combines pre-existing lexicons and involves automatic machine translation, manual annotations and gloss matching across several resources such as THARWA [5].

In this paper, we propose a new method allowed to distinguish nine emotional classes (emotional states), using Tunisian Facebook data. This method is based on the presence of emoticons in the corpus and without using external linguistic resources.

3 Tunisian Corpus Collection

In order to construct our Facebook corpus using the Tunisian dialect, we employed the APIs provided by Facebook.[2] We extracted the textual information from very active political Tunisian pages in the period [1-Jan-2010, 31-Dec-2013]. We used 13 political pages among the most popular in Tunisia. From these Facebook pages, we obtained 60,000 political comments and about 780 K words.

We present our collected corpus as a set of multilingual comments organized in a well-structured XML file to facilitate their handling. In The Tunisian Facebook pages, most users comment using free language as colloquial dialect. Furthermore, the Tunisian dialect is characterized by the presence of a mixture of languages such as French, Standard Arabic, Tunisian dialect which is an Arabic text written in Latin characters and numbers, etc.

Before performing processing on Facebook comments, we need a pretreatment and a shaping step to homogenize them. This step of corpus pre-processing aims to select relevant and the most significant words. Thereby, it allows us to facilitate the construction of our emotional dictionaries. In this step, we performed *character normalization* by replacing specific unpronounced characters with a space and removing accents, stars “*” and others. In order to avoid the presence of segments of words evoking no interest (such as hyperlinks and @target_user), we performed a *filtering* step. This step keeps only the words that reflect the semantic and sentimental content of the comments.

[2] https://developers.facebook.com/docs/reference/apis/.

In this paper, we proposed to construct a lexicon dictionary for each used language on the Tunisian Facebook. In order to distinguish between these languages, we use the language identification tool proposed by Cybozy Labs; hosted on Google Developers.[3] This tool allows us to identify the French and Standard Arabic comments. When the identified language is different from these two languages, we consider that the comment is written using the Tunisian dialect. Concerning the emotion symbols present in the dictionaries, we considered them as multilingual symbols. Therefore, they can also be added in the Tunisian dictionaries. Then, we performed a *lemmatization* step to encompass the words having the same primary entity "lemma". This step was applied only on the French text. However, it wasn't possible to apply lemmatization on the Arabic text because it was full of grammatical errors. For Tunisian dialect, the morphological analyser has not been available yet. To further remove all words deemed unnecessary and keep the interesting words, we eliminated the stop words. To do this, we prepared our own stop-list file containing the grammatical words and the linking words of the three languages (French, standard Arabic, Tunisian dialect). Then, we proceeded to apply a normalization technique for all lengthened words and emoticons, such as "::))))" which is replaced by ":)".

4 Emotional Dictionaries Construction

The goal of the automatic construction of sentiment dictionaries is to list the lexicon words in dictionaries by distinguishing between positive and negative.

Actually, the human sentiments are not limited to positive and negative expressions, but they contain several emotional states. In our case, we consider that there are ***four*** states of positive sentiments: satisfied, happy, gleeful, romantic; ***four*** other states of negative sentiments: disappointed, sad, angry and disgusted, and ***one*** surprised state (non-neutral and not positive or negative). In this paper, we generate a dictionary for each emotional state. In other words, we propose to train ***9 dictionaries***. However, the determination of these dictionaries is a very delicate and complicated task. Moreover, the distinction between emotional states is not well understood with used textual data.

Facebook users use, in their comments, emotion symbols (emoticons ":), :(, etc.", acronyms "lol, mdr, etc." and exclamation words "pf, hh, etc.") to emphasize on their sentiments. In fact, these emotion symbols provide an important contextual value to determine the general sentiment of the text. We take advantage of the integrity of these symbols to distinguish between 9 emotional states that we concluded from the used emoticons (see Fig. 1).

In order to create dictionaries pertaining to these emotional states, we elaborate two stages: the first one is the initial dictionaries construction and the second one is the enrichment of these dictionaries.

[3] https://code.google.com/p/language-detection/.

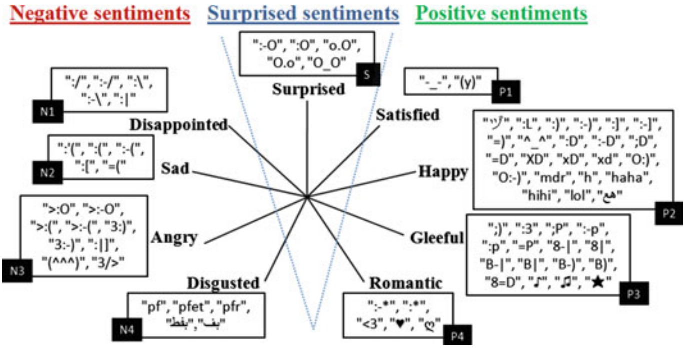


Fig. 1 The emotion symbols used on Facebook to express the nine emotions

4.1 Initial Construction of Emotional Dictionaries

We proposed a method to extract an initial version of the emotional dictionaries by using the comments that contain emotion symbols (4132 comments). This method is based essentially on the presence of emotion symbols in the comments. These emotion symbols reflect the sentiment expressed by the words that precede them in the comments. Indeed, a word has the same polarity as the first emotion symbol appears after it in a comment. In the case that the comment contains one emotion symbol, all words will be attached to this emotion symbol.

Each lexicon word can be attached to several dictionaries according to their appearance with the emotion symbols. Hence, we assigned to each word a valence value. This valence consists to divide the frequency j of the word by the sum of the frequencies of all words present in the dictionary j. The word valences are calculated using (1):

$$valence(w)_j = \frac{frequency(w)_j}{\sum_{i=0}^{n} frequency(w_i)_j} \times 1000 \qquad (1)$$

where: n is the number of words in the dictionary j. $j \in$ (surprised, satisfied, happy, gleeful, romantic, disappointed, sad, angry, disgusted). $frequency(w)_j$ denotes the cooccurrence of the word w in comments with emoticons expressing the sentiment j.

The negation words (e.g. not, no, never, non, ne, etc.) play a specific and important role in the sentimental orientation of words. For this reason, we have taken advantage the presence of negation particles in our comments, in order to reverse the polarity of all words directly preceded by one of these particles (e.g. Im not happy :)). Thereby, when we calculate the cooccurrence frequency of word and emotion symbols, we test if the word is preceded by a negation particle, we decrement its cooccurrence frequency by 1. Otherwise, we increment it by 1.

To determine the emotion brought by every lexicon word, we simply compare its valences in each dictionary. After the initial dictionaries construction step, we obtained *15,576 words* in the sentiment dictionaries. It is clear that in the initial construction step, we kept all the words, even the weakly appeared ones. In fact, these words are likely to be encountered in other comments. For this reason, we proceed to an enrichment step, the objective of the next section.

4.2 Emotional Dictionaries Enrichment

The dictionaries enrichment step aims to use the rest of the comments that do not contain emotion symbols (*50,917 comments*), for two objectives: (i) Settle and adjust the valences of words present in the dictionaries and (ii) Extend the initial dictionaries by adding new words appeared in other comments.

In order to achieve these two objectives, we must firstly determine the dictionary to enrich from a comment. To do this, we calculate the 9 valences of each comment C using the Eq. 2:

$$valence(C)_j = \frac{\sum_{i=0}^{p} frequency(m_i)_j}{\sum_{l=0}^{q} frequency(m_l)_j} \times 1000 \tag{2}$$

where: p is the number of known comment's words. q is the number of words in the dictionary j. $j \in$ (surprised, satisfied, happy, gleeful, romantic, disappointed, sad, angry, disgusted). However, sometimes we cannot calculate these valences, when the words of the comment are all unknown (do not exist in our dictionaries).

Then, we compare the values of these valences to identify the comment polarity. Thus, we enrich the dictionary having the same polarity of the comment by all words in the treated comment.

The principle of the enrichment step is to browse the words of the comment. If the word is inserted in the dictionary to enrich, we increase its frequency in this dictionary by the percentage of polarity of the comment, and then recalculate its valences according to formula (1). Whereas, if the word is new, we add it to the studied dictionary by initializing its frequency by the percentage of polarity of the comment and we then compute its valence (using (1)). As the negation particles occur in *5910 comments* that did not have any emotion symbols, we handle the presence of negation words. In fact, in the case where the treated word is preceded by a negation word, we use the inverse of the percentage ($-percentage(C)_j$). The percentage of polarity of the comment is calculated using (3).

$$percentage(C)_j = \frac{valence(C)_j}{\sum_{k=0}^{8} valence(C)_k} \tag{3}$$

where, $j \in$ (surprised, satisfied, happy, gleeful, romantic, disappointed, sad, angry, disgusted).

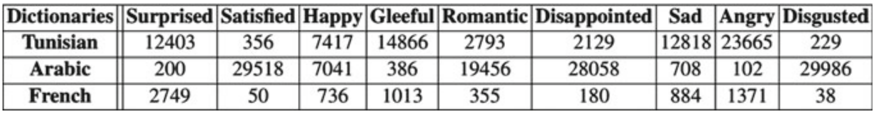
At this stage, we obtained nine dictionaries that cover the majority of the words of our corpus. In fact, the number of words in the enriched dictionaries equals to *131,937*. This shows that the enrichment step is allowed to widen the initial dictionaries.

5 Results and Interpretations

By applying our method of emotional dictionaries construction on our corpus, we have obtained encouraging results. Figure 2 shows for our studied corpus, the number of words attached to emotional dictionaries in each language. We notice that the number of Arabic words expressing disappointment and disgust is very important compared with the other languages. In fact, Tunisians use long Arabic texts in order to emphasize on their sentiments when they are disappointed or when they want to oppose something. However, they prefer to write short texts in Tunisian dialect for expressing fun and amusement.

In order to evaluate our method of emotional dictionaries construction, we used a test corpus containing *755 words* manually labelled by **three experts**. We applied the external evaluation techniques (recall, Accuracy and F-score) to measure the adequacy of the classification of words by our system and that made by the experts. In the Table 1, we presented the obtained results without taking into account the negation and with the negation handling at the level of generation of initial and enriched dictionaries. From these results, we obviously notice the interest of handling negation particles included in our comments and the usefulness of enrichment step. In fact, the best results are obtained with the enrichment method by taking account of the negation. The negation handling has an amplified effect when considered during the enrichment step. This allowed achieving well-adjusted valences and polarities of words. We obtain a F-score of **81.01** % compared to reference dictionaries “textbf-Expert 2”.

To analyze the performance of emotional dictionaries generated by our method, we propose to test the validity of the top-*n* words classified in each emotion. Indeed, we take the *n* words having the highest valance values in each emotional class and we annotate them manually depending on expert judgments. Thus, we compare them to the reference words in order to measure the success rate (see (4)) and the error rate (see (5)). In our experimentation, we try several *n* values [$n = 10, 20$ *and* 30] (see Fig. 3).



Dictionaries	Surprised	Satisfied	Happy	Gleeful	Romantic	Disappointed	Sad	Angry	Disgusted
Tunisian	12403	356	7417	14866	2793	2129	12818	23665	229
Arabic	200	29518	7041	386	19456	28058	708	102	29986
French	2749	50	736	1013	355	180	884	1371	38

Fig. 2 The number of words present in sentiment dictionaries in 3 languages: Tunisian dialect, Standard Arabic and French

Table 1 The precision, recall and F-score obtained with the method for initial construction and enrichment dictionaries

Expert	External evaluation measures	Sentiment dictionaries			
		Without negation		With negation	
		Initial (%)	Enriched (%)	Initial (%)	Enriched (%)
Expert 1	Precision	47.96	55.67	49.42	79.72
	Recall	21.71	60.72	21.80	78.89
	F-score	20.41	56.13	20.53	**80.36**
Expert 2	Precision	48.11	57.68	49.56	80.65
	Recall	21.76	62.48	21.85	79.45
	F-score	20.62	57.97	20.74	**81.01**
Expert 3	Precision	48.01	57.07	49.70	79.18
	Recall	21.83	62.92	21.92	78.10
	F-score	20.39	57.30	20.52	79.65

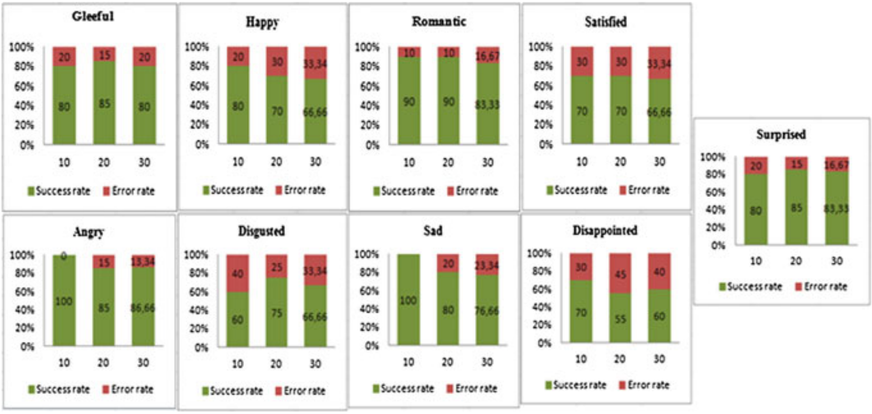


Fig. 3 The success rate and the error rate obtained by our classification method of the top-n words in the 9 emotional dictionaries (n = 10, 20 and 30)

$$Success\_rate = \frac{nbre_{Correct}}{n} \times 100 \tag{4}$$

where: $nbre_{Correct}$ is the number of words correctly classified. n is the chosen number for the test (the total number of words).

$$Error\_rate = 100 - Success\_rate \tag{5}$$

The goal of this evaluation is to verify the correctness of our method of calculating the valences of words to assign them to dictionaries. We note from Fig. 3 that, for every emotional class and whatsoever the value of *n*, the success rate remains high. This shows that the *n* words which have the highest valance values correctly express the emotion of the attached dictionary.

6 Conclusion

In this paper, we presented a new method to construct dynamically emotional dictionaries. Our method is essentially based on the emotion symbols which can be used to express sentiments when commenting on social networks. Our corpus of comments was collected from the Tunisian Facebook pages. Thus, we proposed to create initial dictionaries from the comments having emotion symbols. Moreover, using the rest of the comments, we proposed to enrich these dictionaries. Finally, we discussed the experimental results. In future work, we propose to use automatic processing tools of Tunisian dialect to improve obtained dictionaries, for example: manipulate the words semantically similar but different in writing, like: "ta7founa, tahfouna (wonderful)". Furthermore, we aim to use the obtained dictionaries for classifying a longer text carrier sentiment (comment).

References

1. Abdul-Mageed, M., Diab, M.: Sana: a large scale multi-genre, multi-dialect Lexicon for Arabic subjectivity and sentiment analysis. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). ELRA, Reykjavik, Iceland (2014)
2. Alena, N., Helmut, P., Mitsuru, I.: Analysis of affect expressed through the evolving language of online communication. In: Proceedings of the 12th International Conference on Intelligent User Interfaces, pp. 278–281. ACM, New York, NY, USA (2007)
3. Ameur, H., Jamoussi, S.: Dynamic construction of dictionaries for sentiment classification. In: 13th IEEE International Conference on Data Mining Workshops. ICDM Workshops, pp. 896–903. TX, USA (2013)
4. Balabantaray, R.C., Mohammad, M., Sharma, N.: Article: Multi-class twitter emotion classification: a new approach. Int. J. Appl. Inf. Syst. **4**(1), 48–53 (2012)
5. Diab, M., Albadrashiny, M., Aminian, M., Attia, M., Elfardy, H., Habash, N., Hawwari, A., Salloum, W., Dasigi, P., Eskander, R.: Tharwa: A large scale dialectal Arabic—standard Arabic—English Lexicon. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). ELRA, Reykjavik, Iceland (2014)
6. Douglas, R.R., Christopher, Z.: Corpus-based dictionaries for sentiment analysis of specialized vocabularies. In: New Directions in Analyzing Text as DataWorkshop (2013)
7. Duyu, T., Bing, Q., Ting, L., Zhenghua, L.: Learning sentence representation for emotion classification on microblogs. In: Natural Language Processing and Chinese Computing—Second CCF Conference, pp. 212–223. Chongqing, China (2013)
8. Ekman, P.: An argument for basic emotions. Cogn. Emot. **6**, 169–200 (1992)
9. Kamps, J., Marx, M.: Words with attitude. In: 1st International WordNet Conference, pp. 332–341. Mysore, India (2002)

10. Kim, S.M., Hovy, E.: Determining the sentiment of opinions. In: Proceedings of the 20th International Conference on Computational Linguistics. ACL, Stroudsburg, PA, USA (2004)
11. Mihalcea, R., Liu, H.: A corpus-based approach to finding happiness. In: Proceedings of the AAAI Spring Symposium on Computational Approaches to Weblogs (2006)
12. Mohammad, S.M.: Sentiment analysis: detecting valence, emotions, and other affectual states from text. In: Meiselman, H. (ed.) Emotion Measurement. Elsevier (2016)
13. Solakidis, G., Vavliakis, K., Mitkas, P.: Multilingual sentiment analysis using emoticons and keywords. In: 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), pp. 102–109. Warsaw, Poland (2014)
14. Taboada, M., Anthony, C., Voll, K.: Methods for creating semantic orientation dictionaries. In: Conference on Language Resources and Evaluation (LREC), pp. 427–432 (2006)
15. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 347–354. ACL, Stroudsburg, PA, USA (2005)