Detecting the Likely Causes behind the Emotion Spikes of Influential Twitter Users

Calkin Suero Montero[1], Hatem Haddad[2], Maxim Mozgovoy[3], and Chedi Bechikh Ali[4]

[1] School of Computing, University of Eastern Finland, Joensuu, Finland
calkins@uef.fi
[2] Department of Computer Engineering, Mevlana University, Konya, Turkey
hhatem@mevlana.edu.tr
[3] Active Knowledge Engineering Lab, University of Aizu, Japan
mozgovoy@u-aizu.ac.jp
[4] LISI laboratory, INSAT, Carthage University, Tunisia
chedi.bechikh@gmail.com

Abstract. Understanding the causes of spikes in the emotion flow of influential social media users is a key component when analyzing the diffusion and adoption of opinions and trends. Hence, in this work we focus on detecting the likely reasons or causes of spikes within influential Twitter users' emotion flow. To achieve this, once an emotion spike is identified we use linguistic and statistical analyses on the tweets surrounding the spike in order to reveal the spike's likely explanations or causes in the form of keyphrases. Experimental evaluation on emotion flow visualization, emotion spikes identification and likely cause extraction for several influential Twitter users shows that our method is effective for pinpointing interesting insights behind the causes of the emotion fluctuation. Implications of our work are highlighted by relating emotion flow spikes to real-world events and by the transversal application of our technique to other types of timestamped text.

Keywords: emotion analysis · emotion fluctuation explanation · social media

1 Influence through Twitter

Word of mouth (WOM) has long been considered one of the most efficient information sources that influence customers' decision when purchasing products or services [15]. Due to its strong implications when designing targeted advertising and marketing or political campaigns, the suggestive influence that consumers exert on each other through the WOM phenomenon has been extensively studied also in online social media contexts, as a new form of communication spread [15, 27]. Particular attention has been devoted to single out influential users in social media as inexpensive seed catalyzers in the dissemination of news, opinions, and products or services attractiveness [26]. Specifically, the Twitter microblogging platform has enticed great deal of research on the analysis and identification of influential users due to ease that the platform presents in tracking the diffusion of information through its channels [4]. It is, therefore, not surprising that the analysis of the emotional state of influential users in Twitter reveals itself as a very important focal point of research.

adfa, p. 1, 2011.

Twitter, as a microblogging platform, receives over 500 million tweets worldwide every day as per 2016 [29]. This represents a well of information that is being exploited for journalism, business intelligence, monitoring natural and man-made disasters, terrorism and so forth [3, 10]. Hence, emotion analysis in Twitter has proven to be a valuable asset for marketing analysis including services consumers' satisfaction [7], and political candidate popularity [17], to name a few relevant examples. Since it has been shown that influential individuals in social media networks can positively diffuse their opinions and preferences to peers [2], this has implications for designing "intervention strategies, target advertising and policy making" [2]. However, the meaningful identification of the causes of fluctuation in the emotional flow dispositions of influential individuals through social media is a challenging and underexplored task.

Our application aims at identifying the likely explanations or causes of strong emotions fluctuations i.e., *emotion spikes*, within the temporal dimension of influential users' emotion flow in Twitter (see for example Kwak et al., 2010 [13], Twitter Counter [28] and Cha et al. 2010 [9] for influential users analysis and identification). These emotion spikes are presumably associated with a reaction to certain event. Hence, our system extracts *keyphrases,* phrases formed from linguistic patterns, adjacent to each identified emotion spike, and passes them to an analyst for subsequent examination. Our application goes beyond the detection of named-entities (e.g. "person", "company", "city", etc.) and events or topics identification since the extracted keyphrases are indicative of a change on user's sentiment, and represent the causes the emotion spike. This information is of value for business intelligence, designing targeted marketing campaigns and revamping brand image, for instance.

In this paper we present our contribution and ongoing work on the emotion analysis of influential Twitter users addressing two main issues:

- The visual representation and analysis of the temporal emotion flow of a user's tweets on which emotion spikes are identified.
- The extraction of the likely causes of the visualized emotion spikes using keyphrases for linguistic and statistical analyses.

The rest of the paper is organized as follows. Section 2 outlines existing work on emotional visualization and the identification of likely causes of emotion spikes. Section 3 describes our methodology. In section 4 we describe our experimental evaluation. Finally, discussion and future work are covered in section 5.

2 Related Work

2.1 Emotion Visualization

Emotion flow visualization in social media outlets has generated interest for a variety of applications. For instance, Mishne and De Rijke (2006) [19] developed a system called MoodViews, a collection of tools for analyzing, tracking and visualizing moods and mood changes in blogs posted by LiveJournal users. Similarly, Kempter et al.

(2014) [12] have proposed the EmotionWatch system to automatically recognize emotions and score tweets into 20 discrete emotion categories. Particularly relevant to our work is the TwitInfo tool proposed by Marcus et al. (2011) [16], a prototype system for monitoring events on Twitter, using a timeline graph to show major peaks of publication of tweets about a particular topic, the most relevant tweets, and the polarity of the opinions they express. These systems differ from our proposed approach in that their analysis is performed over the entire blogosphere instead of focusing on a singular user. Also, in our work we look at analyzing emotions alongside the likely explanations or causes that provoked the observed strong fluctuations over time in an emotion flow.

2.2 Likely Causes of Emotion Spikes

A relevant work to our research is that of Balog et al. [5] who proposed a method for identifying and explaining spikes in mood patterns in blogs. Balog et al. used empirical heuristics to identify spikes in users' reported moods within a LiveJournal blog corpus. From the identified spikes' period, a query formed by "overused words" was then extracted and used to search a news corpus in order to mine related news headlines. Similar to Balog et al., we effectively use empirical heuristics to detect and identify emotion spikes within the user's tweets timeline. However, our work differs from Balog et al. approach in that we use keyphrases instead of simple keywords to extract the likely causes of the spikes. By keyphrases we refer to a list of phrases formed from linguistic patterns that express the informative contents of tweets. Hence, in our approach the likely causes for the emotion fluctuation are extracted directly from the tweets themselves instead of an external source, which ensures that our system will return a likely explanation for the emotion spike.

Our work also differs from events detection ED [22], topic detection and tracking TDT [1], and topic classification TC [31] research in social media. That is, given an emotion spike our approach aims at exploring its likely causes, answering questions such as 'why does the user have this strong emotion fluctuation?' and 'what are the likely causes that provoked the spike?'

3 Methodology

Our framework is outlined in Fig. 1. The processes involved are detailed here.

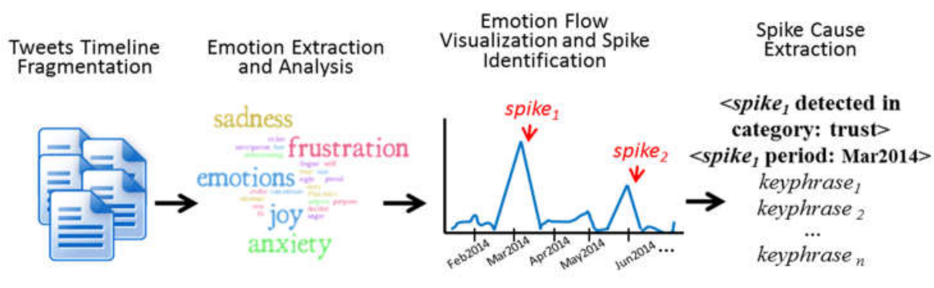

Fig. 1. Framework processes

3.1 Twitter Data and Timeline

We download automatically tweets from a user's account using the Twitter API. The API crawls up to 3,200 most recent tweets and retweets of a user for further processing. Our application discards retweets from our timeline analysis. Once the tweets are downloaded, the system fragments them time-wise using the time-stamp of each tweet.

3.2 Emotion Extraction and Analysis

After timeline fragmentation, the system extracts emotions from the input text using the NRC word-emotion association lexicon, EmoLex [20]. The lexicon has been manually annotated with the Plutchik's eight basic emotions i.e., joy, anger, sadness, fear, anticipation, surprise, disgust and trust [21]. An emotion score (*eScore*) is then calculated for each one of Plutchik's eight emotion categories represented in each text entry as follows:

$$eScore_{category} = \frac{eWords_{category}}{eWords_{all}} \tag{1}$$

where $eWords_{category}$ is the number of words in the analyzed text entry that have nonzero emotional score for the category according to the lexicon; and $eWords_{all}$ is the number of words in the text entry that have nonzero emotional score for any category according to the lexicon.

3.3 Emotion Spike Identification

The emotion flow visualization is given based on the calculated *eScore* on a monthly basis for easy access. A spike in the emotion flow is defined here as a sudden change in the *average* emotion flow of a user. That is, we calculate spikes in the emotion flow using the relative *eScore ratio* between adjacent months. A spike is detected if the ratio of eScore for an emotion category in one month to the eScore for the same emotion category in the month before and the month after is greater than an empirically predefined threshold. Formally:

$$Spike(c, m) = \left(\frac{eScore_{(c,m)}}{eScore_{(c,m-1)}} > \theta\right) AND \left(\frac{eScore_{(c,m)}}{eScore_{(c,m+1)}} > \theta\right) \tag{2}$$

where c is an emotion category, m is the month of the analysis and θ is the ratio threshold acceptable to define a spike. The value of θ was empirically set to *1.45* in our experiments, since it represented a balanced choice when identifying interesting spikes and filtering out noisy ones.

For our analysis, an emotion spike includes all the tweets from the month when the sudden change was detected. Fig. 2 shows the emotion flows extracted from Dalai Lama tweets (1113 tweets, from 2010 to 2015), alongside highlighted examples of detected spikes.

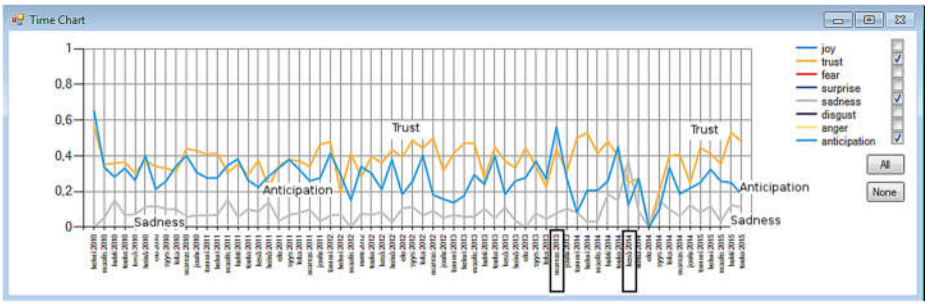

Fig. 2. Dalai Lama's emotion flow for *trust* (top), *anticipation* (middle) and *sadness* (bottom). Spikes were found for trust and anticipation in Nov 2013 and sadness in June 2014, among others

3.4 Extracting Likely Causes for Emotion Spikes

Aiming at uncovering the explanations or causes of a spike in the emotion flow, we extract keyphrases from the month period during which the spike is detected within the respective emotion category. Keyphrases have been used in several domains including text summarization [14], indexing [11] and searching [8], to describe and capture the main information enclosed in a given document [25]. Keyphrase extraction can be performed through linguistic analysis [8, 33], statistical analysis [32, 18] or through a combined linguistic and statistical analyses approach [6]. We use a combined linguistic and statistical analyses approach to keyphrase extraction as it achieves the best performance.

Linguistic Analysis

The system first performs part of speech tagging on the tweets in order to generate a tagged text where each word is labeled with a corresponding part-of-speech marker. Then, a set of manually defined syntactic rule patterns is used to extract keyphrases from the tagged text. Each rule is represented with a sequence of part-of-speech tags, describing the structure of the target keyphrase. Therefore, this method differs from n-gram filters and at the same time is general enough to allow creating patterns that describe keyphrases consisting of compound nouns, named-entities and events.

In our experiments, the length of the syntactic patterns varies from two to four part-of-speech tags, and all matching keyphrases are extracted. The syntactic patterns set consisted of twenty rules. For example, given the tweet *'The Senate already passed bipartisan immigration reform'*, its tagged text is: The/`determiner` Senate/`noun` already/`adverb` passed/`verb` bipartisan/`noun` immigration/`noun` reform/`noun`. Using two rule patterns "`Noun Noun`", and "`Noun Noun Noun`", for instance, the system extracted the following keyphrases:

- bipartisan_immigration (`Noun Noun`);
- immigration_reform (`Noun Noun`);
- bipartisan_immigration_reform (`Noun Noun Noun`)

Statistical Analysis

In order to select the most relevant keyphrases to represent the likely explanations or causes of an emotion spike, we use a two level filtering approach to reduce the number of candidates. Here we define relevancy based on the length as well as the frequency of the keyphrase. First, extracted keyphrases are filtered according to their length: longer keyphrases are preferred over shorter ones containing the same words [24]. Hence from the previous example, only the keyphrase *bipartisan_immigration_reform* is selected as a candidate.

Next, given that rare keyphrases are considered more informative than common ones [23], a second filter is applied to the spikes context based on a variation of the *tf-idf* weight. As a statistical measure, tf-idf provides scored weights to evaluate how important a keyphrase is to a spike within a set of detected spikes. We use tf-idf weights because this weighting scheme has been shown to achieve better performance than other algorithms like TextRank, SingleRank and ExpandRank [6]. With tf-idf we aim at scoring and ranking the extracted keyphrases in order to select *n*-best. In our application, the tf term analyzes the importance of a keyphrase against a spike, while the idf term analyzes the importance of the keyphrase against all the emotion spikes extracted. Hence, with *k* representing a keyphrase (given by the first filter), *s* an emotion spike (containing all the tweets in the analyzed month), and *S* the set of all the emotion spikes, the tf-idf is calculated as:

$$tf\_idf(k,s,S) = tf(k,s) * \ idf(k,S) \tag{3}$$

where $k \in s$ and $s \in S$. We adapted the classic tf-idf weighting scheme [23] to assign the weight of keyphrase *k* in a spike *s* as:

$$tf\_idf(k,s,S) = tf(k,s) * log \frac{number\ of\ spikes\ in\ S}{number\ of\ spikes\ in\ S\ containing\ k} \tag{4}$$

where $tf(k,s)$ is the number of times *k* occurs in emotion spike s. Hence, considering the following k_1 (a rare keyphrase) and k_2 (a common keyphrase) from the Barack Obama spikes collection (3 spikes) found in the emotion category *anticipation:*

- k_1= *anti-discrimination_law*, occurring once in the emotion spike period and only in this spike; then $tf\_idf(k_1) = 1 * \log(3/1) = 0.47$.
- k_2=*president_obama,* occurring 44 times in the emotion spike period and in 3 spikes; then $tf\_idf(k_2) = 44 * \log(3/3) = 0$.

Therefore, k_1 has higher importance based on its tf-idf score.

4 Experimental Evaluation and Results

In order to test whether the automatic detection of emotion spikes is feasible, we set a preliminary experiment where we extracted tweets of three influential Twitter users in the period from 2010 to 2015: Barack Obama (2979 tweets), Bill Gates (1512 tweets) and Dalai Lama (1110 tweets) [28]. Table 1 shows the statistics of the detected spikes and extracted keyphrases for these 3 influential Twitter users. For Bill Gates and Barack Obama tweets, there were 9 spikes discovered, whereas 15 spikes were extracted from the Dalai Lama tweets. The average number of tweets per spike detected is highest for Bill Gates which in turn yields the highest number of extracted keyphrases.

Influential user	Period	Tweets	Spikes	Tweets per spike	Extracted KPs	KPs per spike *(avg)*
Barack Obama	8.11.2013 to 6.6.2015	2979	9	8,8	235	21,1
Bill Gates	19.1.2010 to 9.5.2015	1512	9	21,6	769	85,5
Dalai Lama	22.2.2010 to 12.5.2015	1110	15	12,6	680	15,6

Table 1. Spikes detection and keyphrases (KPs) extraction descriptions

Table 2 shows the extracted keyphrases corresponding to emotion spike cause candidates for two emotion categories for each user. From the extracted keyphrases we could speculate on the reasons for the respective spikes. For instance, the spike detected in the joy category in Dec 2014 on Bill Gates' emotion flow, could be related to a breakthrough in tuberculosis research, or to positive advances in the global discussion about inequality. Our system is also able to mine named-entities and events as per the syntactic patterns used. For instance, for the spike detected in trust, Nov 2013 within Bill Gates' tweets, the keyphrase '`conversation_with_bill_clinton`' is extracted as one of the likely causes for the spike based on its tf-idf score. Conversely, the keyphrase '`wad2012_progress`' was also extracted as a likely cause of this emotion spike, however it was discarded from the list of *3-best* candidates presented in Table 2 as its tf-idf score was low.

Although it is not possible to guarantee completely accurate results, our system can however serve as a significant aid to analysts who otherwise would have to identify the causes of emotion spikes without computer support. A comparative baseline could be created by the analyst reading every tweet of the target user and employing brainpower to analyze the data. Although this is certainly possible to achieve, such a manual process is still a very resource-consuming task. Our preliminary experiments show that our system is indeed able to extract blocks of text that convey likely explanations or causes for the emotional reaction of a user. In addition, even though the analysis of emotions, opinions and polarity among other features of texts has attracted an increasing amount of research, it is also very important to take into account the analysis of such features within the temporal dimension, i.e., how these features change over time. In our work, the change of emotions in the temporal dimension is capture through the analysis of identified spikes and the extraction of likely explanations. We further confirmed that time-stamped texts such as blogs and microblogs are convenient input sources for the type of analysis needed when reactions change over time.

User	Emotion and spike month	Samples of extracted likely causes *(keyphrases)*	tf idf
Barack Obama	Anticipation – Dec 2013	comprehensive_immigration_reform bipartisan_immigration_reform anti-discrimination_law	0.97 0.47 0.47
	Anticipation – Aug 2014	economic_opportunity_for_all private_sector_job sector_job_creation	2.38 1.90 0.97
Bill Gates	Trust – Nov 2013	conversation_with_bill_clinton toilet_save_life future_of_vaccination	0.95 0.95 0.95
	Joy – Dec 2014	fantastic_global_discussion discussion_about_inequality breakthrough_in_tuberculosis	0.95 0.95 0.95
Dalai Lama	Trust – Mar 2012	main_tibetan_temple attitude_of_compassion deep_inner_satisfaction	2.35 1.17 1.17
	Sadness – June 2014	tibetan_childrens village_school condition_for_happiness	1.17 1.17 1.17

Table 2. Extracted candidates of likely causes of emotion spikes for influential Twitter users

5 Discussion and Future Work

While the mechanics of peer influence are complex and elusive to understand [2, 9], it has been long argued that influential people can catalyze peer behavior and opinions [30]. Hence, it is important to analyze influential users' emotion fluctuations in the temporal dimension, i.e., over a period of time, from the perspective of political voters' turnout, product demand or social unrest [2] and Twitter data shows value for these purposes.

Our contribution stands as the development and introduction of a very useful tool that works in detecting the possible reasons behind emotion spikes within a given user emotion flow timeline by using empirical methods heuristics. Our methodology works based on the combination of two algorithms: a lexicon-based algorithm that reacts to emotion-bearing keywords such as *'opportunity'*, and that can be considered as a coarse tool that shows likely places of emotion agitation or fluctuation; and a second algorithm that extracts complete keyphrases from where the emotion fluctuation was detected, such as *'economic opportunity for all'*.

It is important to observe that in our system the likely explanations for the emotion spikes are extracted in the form of keyphrases, only from within the tweets themselves. To further corroborate the validity of these extracted explanations or causes, external sources such as news corpora could also be utilized (see for example Balog et al. [5]). However, when using external sources of verification care should be taken since emotion spikes are subjective to the user and not necessarily have to be the result of an event reported elsewhere.

During the empirical testing for fine-tuning the value of the emotion spike identification threshold, we noticed that this value affects the successful retrieval of emotion spikes making the spikes noisier (if the threshold is lower) or more restricted (if the threshold is higher). This value can be dynamically set according to the analyst's needs and the individual user to be analyzed.

During our experiments we also observed that using our software the amount of some of the analyzed influential user tweets was rather small, against expectations. Noting this, we performed a meta-analysis on Bill Gates Twitter account and found that his tweeting frequency varied from one to three times a day during some weeks to no tweets at all for days during other weeks[1]. Following this pattern, it gave us a rough average of 360 tweets per year or 1800 tweets in a 5 years period available for retrieval. Hence, after discarding retweets, the amount of tweets from this user available for analysis in our application (1512) was made clear. Therefore, caution should be taken not to overestimate individual users' tweeting activity.

Our presented tool for emotion spike detection aims at assisting analysts when monitoring trends on the social media Twitter platform. The tool has been developed taking into account matters of simplicity and convenience of user interface and we are working towards releasing the software to the community as open source.

Our ongoing work includes performing a large-scale user evaluation of the system. Since there is no existing test collection for likely explanations or causes of strong emotions fluctuations, we plan to manually construct our test collection. Improvement of the spike detection procedure is also under consideration for future work, such that spikes are extracted on sliding averages instead of the current monthly basis. Also, while the core algorithm of the system for emotion spike detection is language-independent, emotions in documents are identified with English-based NRC lexicon. Thus, the adaptation of the system to languages other than English can be explored by using similar lexicons, customized for the given language.

References

1. Allan, J., editor. Topic detection and tracking: event-based information organization. Vol. 12. Springer Science & Business Media. (2002)
2. Aral, S. and Walker, D. Identifying influential and susceptible members of social networks. *Science* 337(6092): 337-341. (2012)
3. Atefeh, F., and Khreich, W. A survey of techniques for event detection in Twitter. *Computational Intelligence* 31(1):132–164. (2013)
4. Bakshy, E., Hofman, J.M., Mason, W.A., and Watts, D. J. Everyone's an influencer: quantifying influence on twitter. In Proceedings of *the fourth ACM international conference on Web search and data mining*, pp. 65-74. ACM. (2011)
5. Balog, K., Mishne, G., and De Rijke, M. Why are they excited?: identifying and explaining spikes in blog mood levels. Proceedings of the *Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pp. 207-210. ACL. (2006)

[1] Meta-analysis on Bill Gates tweeting activity, March 2016, https://twitter.com/BillGates

6. Bechikh Ali, C., Wang, R., & Haddad, H. A Two-Level Keyphrase Extraction Approach. In *Computational Linguistics and Intelligent Text Processing* pp. 390-401. Springer International Publishing. (2015)
7. Bougie, R., Pieters, R., and Zeelenberg, M. Angry Customers Don't Come Back, They Get Back: The Experience and Behavioral Implications of Anger and Dissatisfaction in Services. *Journal of the Academy of Marketing Science*, 31(4):377–393. (2003)
8. Bracewell, D.B., Fuji, R., and Kuriowa, S. Multilingual single document keyword extraction for information retrieval. In Proceedings of the *IEEE NLP-KE'05* pp. 517-522. (2005)
9. Cha, M., Haddadi,H., Benevenuto, F., and Gummadi, P. K. Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM* 10, 10-17: 30. (2010)
10. Cheng, T., and Wicks, T. Event Detection using Twitter: A Spatio-Temporal Approach. *PloS one* 9.6: e97807. (2014)
11. Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., and Frank, E. Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems* 27(1): 81-104. (1999)
12. Kempter, R., Sintsova, V., Musat, C., and Pu, P. EmotionWatch: Visualizing Fine-Grained Emotions in Event-Related Tweets. In *Eighth International AAAI Conference on Weblogs and Social Media.* (2014)
13. Kwak, H., Lee, C., Park, H., and Moon, S. What is Twitter, a social network or a news media? In Proceedings of the *19th International Conference on World Wide Web*, pp. 591-600. ACM. (2010)
14. Litvak, M., and Last, M. Graph-based keyword extraction for single-document summarization. Proceedings of the *workshop on Multi-source Multilingual Information Extraction and Summarization*. ACL. (2008)
15. Litvin, S.W., Goldsmith, R. E., and Pan, B. Electronic word-of-mouth in hospitality and tourism management. *Tourism management* 29(3): 458-468. (2008)
16. Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., and Miller, R.C. Tweets as data: demonstration of TweeQL and Twitinfo. In *Proceedings of the ACM SIGMOD* pp. 1259-1262. (2011)
17. Melville, P., Gryc, W., and Lawrence, R. D. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD*, pp 1275-1284. (2009)
18. Mihalcea, R., and Tarau, P. TextRank: Bringing order into texts. In *Proceedings of Conference on Empirical Methods in Natural Language Processing* pp. 404-411. (2004)
19. Mishne, G., and De Rijke, M. Moodviews: Tools for blog mood analysis. In *AAAI Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 153–154. (2006)
20. Mohammad, S. M., and Turney, P.D. 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*. ACL. (2010)
21. Plutchik, R. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1: 3-31. (1980)
22. Psallidas, F., Becker, H., Naaman,M., and Gravano, L. Effective Event Identification in Social Media. *IEEE Data Engineering Bulletin* 36(3): 42-50. (2013)
23. Salton, G., and Buckley, C. Term-weighting approaches in automatic text retrieval. Information processing & management 24(5): 513-523. (1988)
24. Song, Y. I., Lee, J. T., & Rim, H. C. Word or phrase?: learning which unit to stress for information retrieval. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* (pp. 1048-1056). ACL. (2009)

25. Turney, P.D. Learning algorithms for keyphrase extraction. *Information Retrieval* 2(4): 303-336. (2000)
26. Trusov, M., Bodapati, A.V., and Bucklin, R.E. Determining influential users in internet social networks. *Journal of Marketing Research* 47(4): 643-658. (2010)
27. Trusov, M., Bucklin, R.E., and Pauwels, K. Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of Marketing* 73(5): 90-102. (2009)
28. Twitter Counter - Twitter top 100 most followed (2015) http://twittercounter.com/
29. Twitter Usage Statistics (January 2016) http://www.internetlivestats.com/twitter-statistics/
30. Valente, T.W., and Davis, R.L. Accelerating the diffusion of innovations using opinion leaders. *The Annals of the American Academy of Political and Social Science* 566(1): 55-67. (1999)
31. Varga, A., Cano Basave, A.E., Rowe, M., Ciravegna, F., and He, Y. Linked knowledge sources for topic classification of microposts: A semantic graph-based approach. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 26: 36-57. (2014)
32. Wang, R., Liu, W., and McDonald, C. How Preprocessing Affects Unsupervised Keyphrase Extraction. In *CICLing* 163-176. (2014)
33. Zesch, T., and Gurevych, I. Approximate Matching for Evaluating Keyphrase Extraction. In *RANLP*, pp. 484-489. (2009)