Identifying Emotions in Social Media:
Comparison of Word-emotion Lexicons

Ema Kušen[1], Giuseppe Cascavilla[4], Kathrin Figl[1], Mauro Conti[5], Mark Strembeck[1,2,3]

[1]*University of Economics and Business (WU Vienna), Austria*
[2] *Complexity Science Hub Vienna*
[3] *Security Business Austria (SBA) Research Center*
{*firstname.lastname*}*@wu.ac.at*
[4]*Sapienza - Università di Roma, Italy*
cascavilla@di.uniroma1.it
[5]*Università di Padova, Italy*
conti@math.unipd.it

*Abstract***—In recent years, emotions expressed in social media messages have become a vivid research topic due to their influence on the spread of misinformation and online radicalization over online social networks. Thus, it is important to correctly identify emotions in order to make inferences from social media messages. In this paper, we report on the performance of three publicly available word-emotion lexicons (NRC, DepecheMood, EmoSenticNet) over a set of Facebook and Twitter messages. To this end, we designed and implemented an algorithm that applies natural language processing (NLP) techniques along with a number of heuristics that reflect the way humans naturally assess emotions in written texts. In order to evaluate the appropriateness of the obtained emotion scores, we conducted a questionnaire-based survey with human raters. Our results show that there are noticeable differences between the performance of the lexicons as well as with respect to emotion scores the human raters provided in our survey.**

*Keywords***-word-emotion lexicon; emotions; social network.**

I. Introduction

Since the advent of the Web 2.0 era, people are publicly expressing their emotions as a form of interpersonal interaction on online social networks (OSNs). Even though prior studies have shown that such sharing of emotions may lead to beneficial effects [1], [2], there are also potential threats arising from emotionally-charged content that is shared via OSNs. In [1], Hidalgo et al. assert that online social sharing can influence the individuals' overall well-being and life satisfaction. With the advancement of online radicalization [3], spread of panic [4], and dissemination of misinformation [5] over OSNs, it has become important to study the effects that emotional messages invoke in OSN users. For example, understanding such effects can help to anticipate and abate the radicalization phenomenon. As noted in [6], emotions have been recognized as a mechanism to provoke and inspire (political) actions. For example, case studies have shown that jihadists use emotionally-charged content (e.g., martyr videos) to trigger a reaction in their prospective supporters

and evoke the feelings of anger and fear in those that oppose their cause [7].

While studying emotions in OSNs is an important research topic, it has also proven to be a challenging task due to the complexity and the ambiguity of natural language expressions [8]. In recent years, there has been a growing number of studies that utilize sentiment analysis tools to study the opinions of OSN users by analyzing written cues (texts) and multimedia content that people share online. While the scientific literature includes plenty of works focusing on polarizing positive and negative sentiments (see, e.g., [9], [10], [11]), identifying individual emotions such as *anger*, *joy*, *sadness*, or *disgust* has generally been understudied so far.

Previous studies have pointed to the importance of identifying actual emotions rather than sentiment polarities, stating that two emotions belonging to the same affective valence might induce different reactions and lead to different decisions [12], [13]. For example, two emotional reactions shared by two distinct OSN users to the same event "*I am crying tonight.*" (sadness) and "*I am furious - I shall have my revenge.*" (anger) are classified under the same category "negative polarity". However, the two messages convey prospectively different agendas. Thus far, researchers have approached the task of emotion recognition by utilizing machine-learning solutions [14], [15] or by relying on emotion-word lexicons [16], [17]. While both techniques have a potential to provide valuable insights, the question remains how well they are able to identify specific emotions in social media content.

In this paper, we compare the performance of three widely-used emotion-word lexicons on a sample of real-world English language Facebook and Twitter messages. In order to identify emotions, we used techniques from natural language processing (NLP) and implemented a corresponding R-script which considers a number of linguistic features

(see Section III-B).

The remainder of this paper is structured as follows. Section II summarizes related work. In Section III, we describe our research method. We report on the results in Section IV and conclude the paper in Section V.

II. RELATED WORK

While a number of approaches and lexicons focus on extracting sentiment polarities (positive, negative, and neutral [11], [10]), in this section we give an overview of studies that report on approaches to emotion detection.

Currently, two main approaches to sentiment analysis exist: 1) *machine learning approaches* and 2) *lexicon-based approaches*. While the former is assumed to be more accurate if properly trained [18], it highly depends on the quality of the training-set and the classifier being trained. In comparison, lexicon-based approaches are domain-independent and do not require training. However their accuracy depends on the availability of the word-emotion pairs in the respective lexicon [19].

Machine learning approaches generally rely on various classifiers. For example, [14] identified *joy*, *anger*, *sadness*, and *disgust* in Weibo (a Chinese social media platform) by relying on a Bayesian classifier. Another study [15] compared the performance of different classifiers (Bayesian, random forest, logistic regression, and support vector machine (SVM)) while classifying OSN texts according to Plutchik's wheel of emotions [20]. In addition to pure machine learning approaches, [21] compared a lexicon-based approach (NRC lexicon) to three machine learning algorithms (SVM, Naive Bayes and Decision Tree). The dataset used for the experimental phase was ISEAR[1] (International Survey on Emotion Antecedents and Reactions) which contains a large number of personal reports of people who were asked to write a short account on an event in which they experienced joy, fear, anger, sadness, disgust, shame, or guilt. The authors have shown that a combination of machine learning with a lexicon delivered the best performance compared to a pure machine learning and a pure lexicon-based approach.

Lexicon-based approaches (based on the NRC and DepecheMood lexicons) have been reported in [16], [17]. While the NRC lexicon was created via a crowdsourcing approach [16], DepecheMood was built by extracting Rappler's articles which were annotated according to the *Mood Meter* by Rappler's readers [17]. In addition, [22] reports on a lexicon that was built by using tweets that convey one of the Plutchik emotions [20]. In 2016, Song et al. [23] extended the concept of emotion-word lexicons by considering the semantics of emotion expression. In their paper, a word such as "*predictable*" may convey *disappointment* when it refers to a movie and *joy* when it refers to stock market.

[1]http://www.affective-sciences.org/system/files/webpage/ISEAR.zip

III. METHOD

For this paper, we compared the following publicly available lexicons: NRC[2], EmosenticNet[3], and DepecheMood[4]. A recent comparison of the three [24] has shown that they differ in the number of word-emotion pairs and emotion scores assigned to each word. In this paper, we explore the performance of the lexicons when used in a combination with NLP techniques that consider various linguistic features (see Section III-B) over a dataset containing 7,691 social media messages.

To set the *ground truth*, we used the ISEAR dataset along with the results obtained from a questionnaire-based survey in which we asked the participants to manually assign emotion scores to 25 Facebook posts and Twitter tweets (see Figure 1 and Section III-A1).
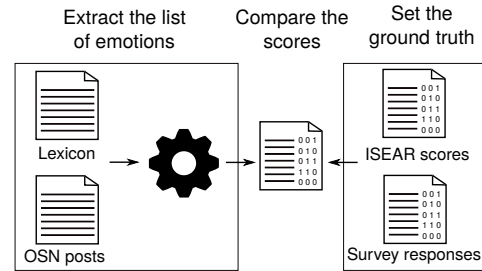

Figure 1: Approach overview.

A. Preparing the ground truth

1) Questionnaire design: In order to evaluate our emotion-extraction script (see Section III-B), we designed two online questionnaires. With the first questionnaire, we measured the impact that the context of a message has on the human perception of emotions. It included 15 real-word Facebook comments and 10 tweets. The second questionnaire was designed to compare the annotators' labeling at the sentence-level (thus, providing less information about the context of a message). Therefore, we extracted individual sentences from each of the Facebook messages (*n=36, mean=2.4, sd=1.8*) and tweets (*n=15, mean=1.5, sd=0.97*) that have been included in the first questionnaire.

Since in this paper we study the presence of emotions as well as their intensity in social media messages, we follow the method for affect measurement as suggested by Schimmack (see [25]). The method suggests to assess the intensity of the emotion on a 7-point Likert scale ranging from 0 = *not at all* to 6 = *strongest imaginable*. The wording of the scale anchors were inspired by a magnitude scale of Bartoshuk et al. (see [26]). The list of emotions used in

[2]http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[3]http://www.gelbukh.com/emosenticnet/
[4]http://github.com/marcoguerini/DepecheMood

	Correlation human-rated sentences (n=51 sentences, rated by 23 humans) and automatic ratings	Correlation human-rated comments (aggregated from sentence ratings by 23 humans) and automatic (n=25 comments)	Correlation human-rated comments, (n=25 comments, rated by 38 humans) and automatic ratings	Correlation human-rated comments (n=25 comments, rated by 38 humans) and human-rated comments (n=25 comments aggregated from sentence ratings by 23 humans)
Joy	0.39**	0.65*	0.97**	0.72*
Anger	0.09	−0.15	−0.08	0.95**
Sadness	0.13	−0.20	−0.15	0.94**
Fear	0.38**	0.06	0.05	0.90**

Table I: Spearman's rank correlations. ** *significant at the 0.01 level (2-tailed).* * *significant at the 0.05 level (2-tailed).*

our study is based on the Plutchik's wheel of emotions.[5] To address a potential impact of fatigue or learning effects on the results, we randomized the order of the Facebook comments and tweets that were presented to each respondent. Upon completing the questionnaire design, we first ran a pilot test with two independent reviewers, revised the survey according to their comments, and finally distributed both questionnaires.

Sample. Our survey was answered by 61 respondents (32 male, 29 female; *mean age=30.24, sd=8.05*), 38 of which completed the first and 23 the second survey.

Results. We computed Spearmans' rank correlations (r_s) to test whether there is a correspondence between human-rated sentences and automated ratings of emotions (based on the NRC lexicon). The results shown in Table I reveal that automated "joy" ratings correlated strongly with human ratings on the comment level ($r_s = 0.97$, $p. = 0.01$). Moreover, the results showed that upon removing the context of a sentence, the correlation between human ratings and automated ratings dropped to a weak positive correlation (joy $r_s = 0.39$, $p. = 0.01$ and fear $r_s = 0.38$, $p. = 0.01$). When comparing human ratings of comments with those aggregated from the mean sentence ratings, we found strong correlations for all emotions.

2) ISEAR dataset: In addition to our questionnaires, we also considered the ISEAR dataset including 7,666 human-annotated entries with respect to seven emotions (fear, shame, anger, disgust, guilt, joy, and sadness). However, this dataset does not include features that are characteristic for OSN messages, such as emoticons, URLs, recurrences of a single letter that serve as boosters (e.g., *soooo*), and common abbreviations (e.g., *LOL*). Thus, we extended ISEAR with our 25 real-world OSN messages and their corresponding annotations obtained through the questionnaire-based survey. Our ground truth finally comprised 7,691 annotated texts that better resemble real world OSN messages.

[5]Note that also other methods for *affect measurement* exist. For example, the affective slider (AS) [27] asks participants to choose between the bipolar affective states (e.g., happy - sad). However, we did not use the AS method because the set of emotions we used in our study (anger, fear, sadness, and joy) does not include pairs of bipolar emotions. Moreover, the PANAS method [28] typically does not distinguish between ratings such as *slightly present* or *not at all present*. Thus, we opted for a more fine-grained distinctions among the two ratings, as suggested in [25].

B. Emotion extraction

1) Issues with emotion extraction: In written texts, people express their emotions directly by using explicit words (such as *happy, sad*) as well as phrases that imply a certain emotion (such as "*I broke my toy*" → sadness). In order to obtain emotion scores from short texts, we used the NRC, DepecheMood, and EmoSenticNet word-emotion lexicons. Below, we give examples for some of the issues that arise if one only relies on looking up words in word-emotion lexicons (words available in a lexicon are printed in bold font):

1) "I am not **happy**." → joy = +1 (no negation)
2) "**Snakes**!!!" → fear = 0 (no direct lexicon match for words in their plural form)
3) "I am very **sad**!" → sadness = +1 (no amplifiers)
4) "I **broke** my **favourite toy**." → [broke] surprise = +1, [favourite] joy = +1, [toy] score 0 (no semantic association)

2) Emotion-extraction algorithm: In order to extract emotions from OSN messages and mitigate the aforementioned issues, we designed and implemented an algorithm (see pseudocode below) which applies NLP-techniques (lemmatization, part-of-speech (POS) tagging) and a number of heuristics (e.g., adverbs of degree, negation) which reflect the way humans assess emotions in texts.

In particular, we first clean links and HTML tags from the messages. Next, we perform POS-tagging and lemmatize the messages to prepare them for the lexicon matching (i.e. we split each message into sentences and each sentence into words). We then search for a lexicon match. As suggested in [29], we implemented a *weighting function* which amplifies the strength of an emotion if an emotion carrier is in a 3-step vicinity (backward and forward trigrams) of an adverb of degree – amplifiers (e.g., *awfully*), maximizers (e.g., *absolutely*), and downtoners (e.g., *slightly*). Note that the impact of adverbs of degree on sentiment polarity scores were extensively studied in [30] and also utilized in other related studies (e.g., [31]). Moreover, we also search for valence shifters (e.g., *not, none*) since previous studies have pointed to the importance of applying this heuristic [32]. In contrast to [15], we interpret emoticons as emotion intensifiers rather than the main emotion-identifiers. We assigned emoticons to three categories – happy (84 emoticons), sad (48 emoticons), and conditional (<3). Since "<3" can be

used to express love (e.g., *"I love this song <3"*) or sadness (e.g., *"I miss him so much <3"*), we treated "<3" as an amplifier of the dominant emotion in the sentence. Moreover, we treated recurrences of a single letter (e.g., *soooo happy*) as additional amplifiers and also extended the lexicons with the most common abbreviations (e.g., *OMG, LOL*).

PSEUDOCODE   1: Emotion extraction.
Data: C, Lexicon_dict, Secondary_dicts
Result: Emotions for each $c \in C$

1 $C\_clean = [\,]$
2 $Emoticons = [\,]$
3 $Lemmatized = [\,]$
4 **foreach** $c_i \in C$ **do**
    /* clean unnecessary HTML, tags */
5   $C\_clean.append(RemoveCode(c_i))$
6 **foreach** $c_i \in C\_clean$ **do**
    /* Extract emoticons */
7   $Emoticons.append(CodeSmiley(c_i))$

8 **foreach** $c_i \in Emoticons$ **do**
    /* Lemmatize the short texts */
9   $Lemmatized.append(Lemmatizer(c_i))$

10 $dict\_emot\_v = \{\,\}$
11 **foreach** $c_i \in Lemmatized$ **do**
12 **foreach** $sent_q \in sentiments$ **do**
13   $dict\_emot\_v[i][sent_q] = 0$
    /* Split text to sentences */
14 $dict\_emot\_v[i] = \{\,\}$
15 $sentences = SplitToSentence(c_i)$
16 **foreach** $s_j \in sentences$ **do**
     /* Split sentence to words */
17   $words = SplitToWord(s_j)$
18   **foreach** $w_k \in words$ **do**
      /* Find lexicon match */
19   **if** $w_k \in Lexicon\_dict$ **then**
       /* Update emotion score corresponding to w_k */
20     $dict\_emot\_v[i][Lexicon\_dict[w_k][0]]+ =[Lexicon\_dict[w_k][1]$
       /* Find negation and intensifiers */
21   **foreach** $w_r \in [w_{k-3} : w_{k-1}, w_{k+1} : w_{k+3}]$ **do**
22   **foreach** $d_t \in Secondary\_dicts$ **do**
         /* Updating emotions related to w_k */
         /* using negation and intensifiers */
23   **if** $w_r \in d_t$ **then**
24        $dict\_emot\_v[i][Lexicon\_dict[w_k]]+ = d_t[w_r]$

25 **return** $dict\_emot\_v$

The procedure implemented in our script is presented in Pseudocode 1, where we denote a list of OSN texts $c_1, c_2, ..., c_N$ with C. Each comment is composed of sentences $c_i = (s_{i1}, s_{i2}, ..., s_{im})$, and each sentence is composed of words $s_{ij} = (w_{ij1}, w_{ij2}, ..., w_{ijk})$. $List\_dicts = [\,]$ contains all the words in a sentence. $RemoveCode(i)$ is the function that removes HTML tags and links. Furthermore, $CodeSmiley(i)$ is the function that identifies emoticons and the $Lemmatizer(i)$ function lemmatizes messages. Finally, $FindMatch(i)$ is the function that identifies a lexicon match and assigns an emotion score e_i. *Secondary dictionary* refers to the dictionary of emoticons (e.g., *happy emoticons:* ":)", ":D") and the AFINN lexicon[6] which provides intensities with respect to a word's affective valence (e.g.,

overjoyed +5, troubled -3). For example, the NRC lexicon assigns a default intensity score of 1 if there is a matching word in a sentence. By utilizing the AFINN lexicon, we aimed to provide a more fine-grained score so that sentences such as "*I am cheerless*" and "*I am depressed*" have a different intensity for the emotion *sadness*.

IV. RESULTS OF THE COMPARISON

The first challenge we faced was the wide variety of emotions available in each word-emotion lexicon. While NRC relies on Plutchik's wheel of eight basic emotions, EmosenticNet includes Ekman's six basic emotions, and DepecheMood [17] includes eight mood-related words based on Rappler's mood meter. Thus, in order to compare the inter-lexicon agreement, we first extracted the intersecting emotions – anger, sadness, fear, and joy (see Figure 2).
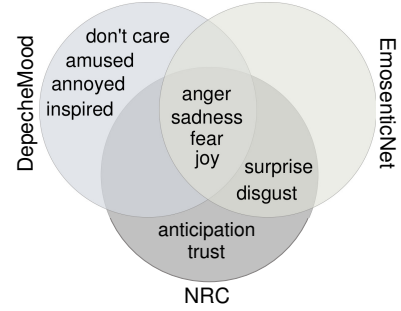

Figure 2: Intersection of the lexica.

After running the OSN messages through the script, we conducted feature scaling to normalize the scores to the closed interval $[0-1]$.

$$x' = \frac{x - min(x)}{max(x) - min(x)}.$$

To quantify the similarity among the three lexicons, we relied on a technique commonly used to quantify the similarity of user ratings in recommender systems. Thus, we consider each lexicon a rater and each text an item being rated. To quantify the similarity, we relied on the *cosine similarity* measure which takes two vectors (lexicon scores) as an input ($\vec{a}$ and $\vec{b}$) and obtains a dot product and a magnitude between the two vectors [33]. The similarity between each pair of lexicons is then found by obtaining the cosine angle between the two vectors.

$$cos(\theta) = \frac{\vec{a_i} \cdot \vec{b_i}}{||\vec{a}||_2 \times ||\vec{b}||_2}.$$

The results of the cosine similarity measure reveal a high degree of similarity between each pair of lexicons (see Table II). In particular, we observe a high similarity (close to perfect agreement) between the NRC and EmosenticNet for the emotions anger, fear, and sadness (each resulted in 11 distance between the two vectors). We received the lowest

similarity, albeit still high, when we compared NRC and EmosenticNet with DepecheMood (the results vary between [0.81, 0.91]).

	NRC	DepecheMood	EmosenticNet
NRC	1	.83, .82, .84, .83	.98, .98, .98, .93
DepecheMood		1	.81, .81, .84, .91
EmosenticNet			1

Table II: Results of the cosine similarity measure for anger, fear, sadness, and joy, respectively.

Next, we compared the performance of the three lexicons with respect to our ground truth. Since the ISEAR annotations are categorical, i.e. texts are sorted into a corresponding emotion category (e.g., anger, joy), we treat the analysis as a text classification problem. We limit the number of categories to four intersecting emotions – anger, sadness, fear, and joy (as seen in Fig. 2) and rely on the precision – recall metrics[7] [34] to quantify the respective classification performance.

$$precision = \frac{TP}{TP + FP} \quad recall = \frac{TP}{TP + FN}$$

We also obtained the F-measure [35] which combines precision and recall scores into a single value.

$$F - measure = 2 \times \frac{precision \times recall}{precision + recall}$$

	NRC			DepecheMood			EmosenticNet		
	P	R	F	P	R	F	P	R	F
Negative	**84%,**	**86%,**	**0.85**	83%,	70%,	0.76	88%,	20%,	0.32
Anger	**37%,**	**68%,**	**0.48**	50%,	28%,	0.36	29%,	7%,	0.11
Fear	**48%,**	**34%,**	**0.40**	41%,	12%,	0.18	56%,	1%,	0.03
Sadness	53%,	27%,	0.36	**34%,**	**56%,**	**0.42**	56%,	23%,	0.33
Positive (joy)	**56%,**	**52%,**	**0.54**	39%,	57%,	0.46	28%,	92%,	0.42

Table III: Precision (P), recall (R), and F-score (F) for each emotion and class of emotions with respect to the ground truth.

Table III shows the precision, recall, and F-measure for each emotion categorized alongside the corresponding valence dimension (*negative*: anger, fear, sadness; *positive*: joy). The highest scores are printed in bold font. Our results indicate a high precision [0.83, 0.88] for the category *negative emotions*. However, the recall measure also revealed a significant incompleteness of the results for EmosenticNet (only 20%), while NRC had the highest recall (86%).

When observing the scores at the level of a specific emotion, there is a noticeable decrease in the precision. Thus, our results reveal that the lexicons are generally able to correctly recognize the affective valence of a text, while the identification of a specific dominant emotion is not precise yet. One possible explanation is that words with

[7]TP refers to true positives and FP to false positives.

a negative meaning (e.g., *cry, death, break*) are generally assigned a score for anger, sadness, fear, or other emotions of a negative valence (e.g. *disgust*, depending on the lexicon). For example, false positive (FP) anger scores for the NRC lexicon included texts that were classified as fear (39% of FP) and sadness (30% of FP), as well as joy (31% of FP), according to ISEAR and our human annotators. Another explanation for such a number of FPs is the availability of word-emotion pairs in each lexicon (i.e., some words that might heavily influence the emotion score are not available in the lexicon). While NRC and DepecheMood provided comparative results, we noticed a weaker performance of the EmosenticNet lexicon (see Table III). Such a score can be attributed to the fact that EmosenticNet consists of a significant number of bigrams and trigrams (e.g., *watch_baseball_game*) which did not occur in our dataset.

V. CONCLUSION AND FUTURE WORK

If embedded in an NLP-based script, word-emotion lexicons have shown their ability to identify affective valence from a given social media message, exhibiting a satisfactory F-score. However, when assessing the correctness of the lexicons alongside specific emotions, our results indicate a relatively high number of false positives. While one explanation for such a decrease in the precision and completeness can be attributed to the similarity of scores in the word-emotion lexicons, another explanation might be the choice of words available in each lexicon. Based on our evaluation, the NRC lexicon performed better for identifying *anger*, *fear*, and *joy*, while DepecheMood performed better at identifying *sadness*.

In our future work, we plan to study the impact of emotions on the spread of information in OSNs. Moreover, since our algorithm has shown to be suitable for studying the linguistic features of how users express emotions over OSNs, another study will focus on analyzing the language of emotions that are characteristic for OSNs.

REFERENCES

[1] C. R. Hidalgo, E. Tan, and P. Verlegh, "The social sharing of emotion (sse) in online social networks: A case study in live journal," *Computers in Human Behavior*, vol. 52, 2015.
[2] E. Kušen, M. Strembeck, G. Cascavilla, and M. Conti, "On the influence of emotional valence shifts on the spread of information in social networks," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2017, (in press).

[3] M. Alarid, *Recruitment and Radicalization: The Role of Social Media and New Technology*, 2016.

[4] BBC, "#bbctrending: How panic about ebola is spreading faster than the virus," 2014. [Online]. Available: http://www.bbc.com/news/blogs-trending-29618224

[5] N. Woolf, "As fake news takes over Facebook feeds, many are taking satire as fact," 2016. [Online]. Available: https://www.theguardian.com/media/2016/nov/17/facebook-fake-news-satire

[6] D. Wright-Neville and D. Smith, "Political rage: terrorism and the politics of emotion." *Global Change, Peace and Security*, vol. 21, no. 1, 2009.

[7] A. Iyer, J. Webster, M. Hornsey, and E. Vanman, "Understanding the power of the picture: The effect of image content on emotional and political responses to terrorism." *Journal of Applied Social Psychology*, vol. 44, 2014.

[8] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," *Journal of King Saud University - Engineering Sciences*, 2016.

[9] I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," *Engineering Applications of Artificial Intelligence*, vol. 51, 2016, mining the Humanities: Technologies and Applications.

[10] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment in short strength detection informal text," *ournal of the American Society for Information Science and Technology*, 2010.

[11] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, "Comparing and combining sentiment analysis methods." S. M. Muthukrishnan, A. E. Abbadi, and B. Krishnamurthy, Eds., 2013, pp. 27–38.

[12] J. Berger, "Arousal increases social transmission of information," *Psychological Science*, vol. 22, no. 7, 2011.

[13] R. Raghunathan and M. T. Pham, "All negative moods are not equal: Motivational influences of anxiety and sadness on decision making," *Organizational Behavior and Human Decision Processes*, vol. 79, no. 1, 1999.

[14] R. Fan, J. Zhao, Y. Chen, and K. Xu, "Anger is more influential than joy: Sentiment correlation in Weibo," in *PloS one*, 2014.

[15] R. Meo and E. Sulis, "Processing affect in social media: A comparison of methods to distinguish emotions in Tweets," *ACM Transactions on Internet Technologies*, 2017.

[16] S. M. Mohammad, X. Zhu, S. Kiritchenko, and J. Martin, "Sentiment, emotion, purpose, and style in electoral tweets," *Information Processing Management*, vol. 51, no. 4, 2015.

[17] J. Staiano and M. Guerini, "Depechemood: A lexicon for emotion analysis from crowd annotated news," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014.

[18] M. Devika, C. Sunitha, and A. Ganesh, "Sentiment analysis: A comparative study on different approaches," *Procedia Computer Science*, vol. 87, 2016.

[19] A. Koumpouri, I. Mporas, and V. Megalooikonomou, "Evaluation of four approaches for "sentiment analysis on movie reviews": The Kaggle competition," in *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks*, 2015.

[20] R. Plutchik, "The nature of emotions," *American Scientist*, vol. 89, no. 4, 2001.

[21] S. Gievska, K. Koroveshovski, and T. Chavdarova, "A hybrid approach for emotion detection in support of affective interaction," in *2014 IEEE International Conference on Data Mining Workshop*, 2014.

[22] F. Koto and M. Adriani, "HBE: Hashtag-based emotion lexicons for twitter sentiment analysis," in *Proceedings of the 7th Forum for Information Retrieval Evaluation*. New York, NY, USA: ACM, 2015.

[23] K. Song, W. Gao, L. Chen, S. Feng, D. Wang, and C. Zhang, "Build emotion lexicon from the mood of crowd via topic-assisted joint non-negative matrix factorization," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2016.

[24] F. S. Tabak and V. Evrim, "Comparison of emotion lexicons," in *2016 HONET-ICT*, Oct 2016.

[25] U. Schimmack, "Affect measurement in experience sampling research," *Journal of Happiness Studies*, vol. 4, no. 1, 2003.

[26] L. M. Bartoshuk, V. B. Duffy, B. G. Green, H. J. Hoffman, C.-W. Ko, L. A. Lucchina, L. E. Marks, D. J. Snyder, and J. M. Weiffenbach, "Valid across-group comparisons with labeled scales: the glms versus magnitude matching," *Physiology behavior*, vol. 82, no. 1, 2004.

[27] A. Betella and P. Verschure, "The affective slider: A digital self-assessment scale for the measurement of human emotions," *PLOs One*, vol. 11, no. 2, 2016.

[28] J. Crawford and J. Henry, "The positive and negative affect schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample," *British journal of clinical psychology*, vol. 43, 2004.

[29] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, Jun. 2011.

[30] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. S. Subrahmanian, "Sentiment analysis: Adjectives and adverbs are better than adjectives alone," in *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2007.

[31] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[32] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo, "A survey on the role of negation in sentiment analysis," in *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 2010.

[33] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th International Conference on World Wide Web*, 2001.

[34] T. Fawcett, "ROC graphs: Notes and practical considerations for researchers," HP Laboratories, Palo Alto, Tech. Rep., 2004.

[35] K. M. A. Chai, "Expectation of F-measures: Tractable exact computation and some empirical observations of its properties," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2005.