Am I Really Happy When I Write "Happy" in My Post?

Pavel Shashkin and Alexander Porshnev

National Research University Higher School of Economics
p-sh@live.ru, aporshnev@hse.ru

Abstract. Posts published on the Internet could serve as a valuable source of information regarding emotion. Recommendation systems, stock market forecast and other areas are likely to benefit from the advancement in mood classification. To deal with this task, researchers commonly rely on preassembled lexicons of emotional words. In this paper we discuss the possibility of extracting emotion-specific words from user-annotated blog entries. The study is based on analysis of the collection from 14800 Live Journal posts containing the "Current mood" tag, specified by the author. The analysis findings and possible applications are discussed.

Keywords: *sentiment analysis, user-annotated data, computational linguistics, emotional states, psycholinguistics*

Introduction

Over the last few years, a considerable amount of work has been done to reduce the overload of user-generated web content [1]. The possibility of grouping data in accordance with sentiment is repeatedly discussed in recent investigations [2, 3]. The system capable of extracting emotions inherent in the text is likely to assist both human-computer and human-human interactions, and help in various tasks. For example, automatic analysis of emotions could be used in some applications, such as: recommendation systems (personal emotions expressed during evaluation could be taken into account), monitoring of psychological user states (customer satisfaction or diagnostics of potential illness), business intelligence (evaluation of the emotional tone of comments circulating about one's company can be used to improve financial decisions).

Online diaries provide researchers with extremely diverse and manifold data. Blog entries are rich in deeply personal and subjective content. Unlike other corpora used in sentiment analysis, "Current Mood" is a text attribute directly specified by the author at the time of writing, rather than by some independent annotator. We expect that analysis of user-annotated data could provide new information about words people use to express their emotional states.

Related work

Over the past few decades there have been several projects devoted to analysis of emotions in the Internet posts. For example, a research project for measuring emotions is "Pulse of a nation" is based on analysis of Twitter messages from September 2006 to August 2009 [4]. In their research, Mislove and coauthors tried to find places where life is sweet, people are happier, and to reveal the unhappiest time of a day. Although, we were unable to find scientific articles, the authors of the "Pulse of a nation" project took part in several TV programs and published the results in newspapers and periodicals (including The Wall Street Journal and The New York Times).

To measure emotions in each tweet, Mislove and coauthors used the ANEW word list [5]. The methodology of emotion analysis was to calculate a sentiment score as a ration of the amount of positive messages to that of negative messages. A message is regarded as positive if it has at least one positive word and as negative if it has at least one negative word (the same message can be both negative and positive) [6].

The project focused on the expression of happiness in social media was developed by a group of researchers from the University of Vermont. They tried to measure happiness in Twitter posts [7].

First of all, Dodds and his coauthors conducted a survey using Amazon Mechanical Turks to obtain happiness evaluations of over 10,000 individual words, representing a tenfold size improvement over similar existing word sets (chosen by frequency of usage in collected samples of nearly 4.6 billion expressions posted over a 33 month span). The created words list contains ranks of their relation to happiness. For example, the top happiness words in their rank are laughter (rank=1) and happiness (rank=2). Next, they created on-line service hedonometer.org, which provides real time happiness analytics based on analysis of frequencies of the words from the list. It is worth mentioning that this service also has a rank for the word "birthday", so the expression "Happy Birthday" is not excluded from analysis.

Lansdall-Welfare, Lampos, and Cristianini tried to measure several emotions in twitter posts by counting the frequency of emotion-related words in each text published on a given day [8]. They also use a lexical approach and base their analytics on the word lists extracted from the WordNet Affect ontology [9].

After this pre-processing Lansdall-Welfare, Lampos, & Cristianini compiled four word lists containind 146 anger words, 92 fear words, 224 joy words and 115 sadness words. The evaluation of emotions in tweets was based on counting the amount of tweets containing each word from the compiled list. Lansdall-Welfare, Lampos, & Cristianini say they do not expect the high frequency of the word 'happy' to necessarily signify a happier mood in the population, as this can be due to expressions of greeting, like "Happy Birthday". Although they do not filter this and similar expressions in their analysis.

We can conclude that the projects running analysis of emotions and moods in social networks usually use the lexicon methodology based on expert-annotated words lists.

Application of the lexicon approach based on expert or naïve rating of emotions in the Internet posts can be supported by the findings made by Gill, Gergle, French and Oberlander. They examined the ability of naive raters of emotion to detect one of the eight emotional categories by asking participants to read 50 and 200 word samples of a real blog text and evaluate whether this message expresses one of the eight emotions: anticipation, acceptance, sadness, disgust, anger, fear, surprise, joy or being neutral [10]. Comparing the results of evaluation by expert raters and naive experts allowed the conclusion that rater agreement increased with longer texts, and was high for ratings of joy, disgust, anger and anticipation, but low for acceptance and 'neutral' texts.

Although raters show agreement in annotation of emotions, we can raise a question about its validity from the psychological perspective. We are not sure that all people express their emotions in a straightforward way, using words closest to the chosen mood category.

An interesting study of emotion in the context of a computer-mediated environment was conducted by Hancock, Landrigan, & Silver [11]. They organized an experimental study, in which some of the participants were asked to express either positive (happy) or negative (unhappy) emotions during a chat conversation, without explicitly describing their (projected) emotional state. Even though, their chat partners did not know about their instructions and their emotional state, they could accurately perceive their interlocutor's emotions. Linguistic analysis showed that the authors portraying positive emotions used a greater number of exclamation marks and more words overall. The participants portraying negative emotions used an increased number of affective words, words expressing negative feeling, and negations.

In this study the people understand emotions of their partner even if these emotions were not explicitly expressed. This raises a question: could we extend the lists of emotional words by analyzing data annotated with the current mood of an author?

Analysis of text semantics, therefore, can provide information about user emotions and we expect analysis of user-annotated data from LiveJournal to help extend the existing words lists related to emotions.

Data collection

We used DuckDuckGo[1] search engine in conjunction with "GoogleScraper"[2] Python module to make a list of English-speaking LiveJournal users who have at least once used the "Current Mood" functionality. The list of obtained URLs is passed down to the web crawler hosted on "import.io"[3] platform. Each visited page is parsed to extract user messages and links to other LiveJournal blogs to be added to crawling query (e.g. from the comment section). Data collecting

[1] `https://duckduckgo.com/`
[2] `https://github.com/NikolaiT/GoogleScraper`
[3] `https://import.io/`

continues until the specified maximum page depth is reached.

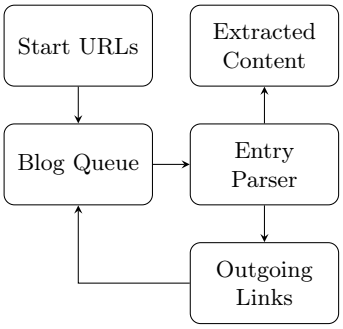

Fig. 1. Data collection system architecture

Data-set highlights

For each message in visited blogs we extract a web address, title, text content and mood tag (usually accompanied by "Current mood:", "Feeling rather:" or just "Mood:"). Although the presence of subjective content in a title or web address is questionable, they are needed to identify continuous entries (eg. stories divided into series of posts). Blog posts, especially the ones with a fair amount of text, are not as frequent as, say, twitter posts. For that reason we do not use time stamps and rarely present geolocation data. The acquired dataset contains 14,800 documents tagged with 800 unique mood labels. 6% of the labels were responsible for 60% of data entries (Figure 2). Average text length is 420 words. An approximate post count for the average author is 5 messages (Figure 3). The most popular mood tags are: "accomplished", "cheerful", "tired" and "amused".

Pre-Processing

The initial step is to clean data from invalid entries (non-Latin or comprising only media content). The dictionary is then reduced by transforming everything to lowercase, stemming words, removing punctuation, stopwords and numbers. If we find negations, like “don’t”, “didn’t” or “not”, the subsequent token is replaced by not_token. For example, “they didn’t come” includes three tokens: “they”, “didn’t”, “come”. We also keep negations as we can expect negative moods negotiations to carry some additional information. URLs are shortened to their respective domains and repeating letters (more than three) are reduced to three. Words and numbers representing time or date are replaced with "time_date".

Fig. 2. Label frequencies

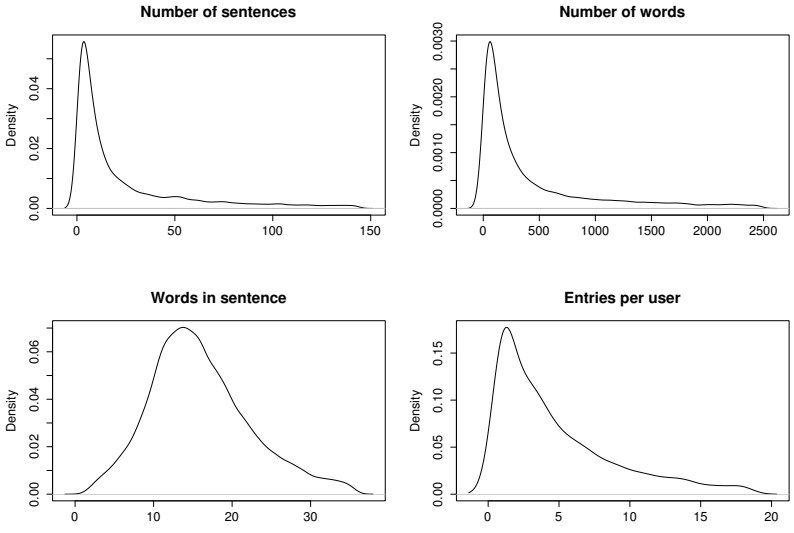

Fig. 3. Empirical distribution density for text statistics (sentences detected with openNLP, outliers removed)

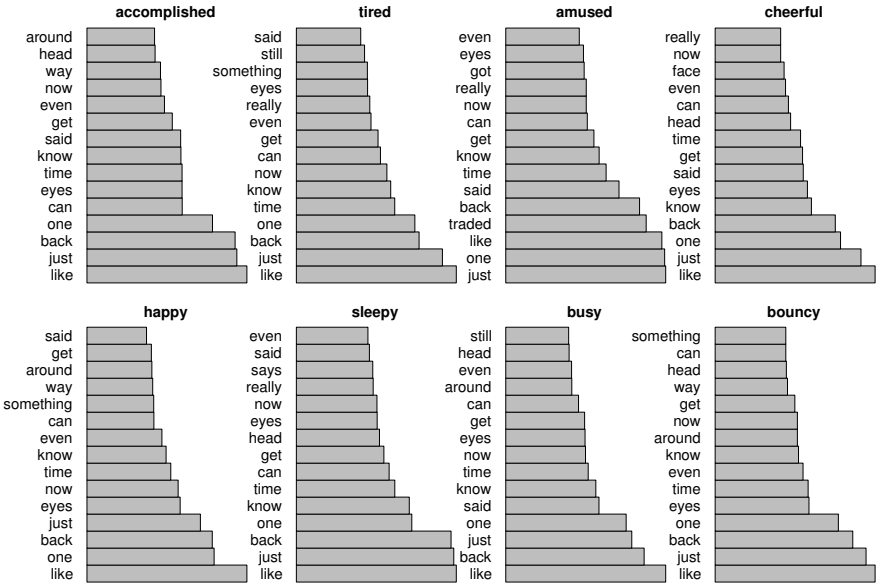

Fig. 4. Word frequencies distribution for popular mood labels

After pre-processing the portion of non-sparse terms doubled. Overall dimensionality of feature space was reduced by more than 3 times.

Fifteen highest word frequencies for 8 most used mood labels are very similar and do not provide any evidence that people use emotional words to mark their emotions (Figure 4). We can see that words highly associated with a mood are not included in the list with top 20 frequencies. For example, it is not often that messages tagged “Happy” contain “happy” in their body. The list of top 20 words does not contain many words from emotional lists. Words “like”, “one”, “back” are not put on the list of 10,000 words related to “happy” according to a Hedonometrics survey [7]. They are not included in the list of Affective Norms for English Words either [5].

The TF-IDF coefficient frequently used for document classification can provide more focused information about semantics of each emotion categories. To calculate TF-IDF, we joined all documents of the category into one document. First, the calculated TF-IDF allowed us to find most of the names used in posts. The words with the highest TF-IDF scores were “leo", "maes", "vampir", "jare","sandi","gaara", and "roger". After including the names in a list of stopwords, we received almost the same situation as with calculation of term frequencies (see Table 1).

accomplished		tired		cheerful		sleepy	
hand	505.7	hand	222.4	artwork	173.3	ghost	342.0
said	447.1	rift	216.7	hand	167.7	hand	170.2
eye	416.0	say	191.8	said	154.5	margin	153.3
back	389.8	f*ck	170.9	head	151.1	head	151.1
head	375.5	eye	156.9	eye	142.8	back	144.5
smile	368.7	back	155.6	smile	135.7	color	142.3
look	360.8	look	149.5	face	135.2	say	140.1
say	355.2	head	148.6	back	119.3	eye	130.8
robot	353.5	doesnt	145.5	look	117.9	superhero	130.0
mule	338.5	said	145.2	lip	110.8	said	125.4

amused		happy		busy		bouncy	
trade	692.7	array	279.9	dev	263.9	sampl	120.7
prize	379.7	hand	164.9	alt	178.5	hand	115.1
ward	143.9	eye	164.7	hand	148.6	introspect	106.3
claim	135.6	back	137.1	said	147.0	head	91.3
vote	128.4	lip	127.4	border	133.4	back	89.4
said	91.6	smile	125.5	back	118.3	said	84.8
hand	86.6	head	122.0	head	114.8	charact	80.6
bill	86.0	knew	122.0	eye	114.7	lip	76.5
materia	79.7	realis	121.2	knew	101.3	look	74.6
back	69.4	face	119.1	multi	101.1	kiss	74.4

Table 1. Words with highest TF-IDF score for eight most popular mood labels (with proper nouns removed)

Next, we introduced the TF-ICF coefficient. In order to identify important group-specific words, the term frequencies $TFij$ for word i in group j are multiplied by:

$$log \frac{||D||}{\sum_{j=1}^{||D||} \frac{TF_{ij}}{\max_{t \in T_j} TF_{tj}}} \tag{1}$$

where $||D||$ is the number of document groups and T_j - unique words in document group j.

The results produced by this transformation are listed in Figure 5 (apart from persons, locations and brands on top of the list) and provide more information regarding sentiment. For example, the word "finally" has a high value in documents tagged with "accomplished" or "tired". Although, some of these results are relatively counter-intuitive or even contradictory (e.g. "bed" is present in "accomplished", "bouncy", "cheerful", "busy", but absent in "sleepy").

This suggests that the distance between documents written in different emotional states could be shorter than that between documents written in the same emotional state by different authors. To test this hypothesis, we filtered documents by author and then, using vector representation of documents, we calcu-

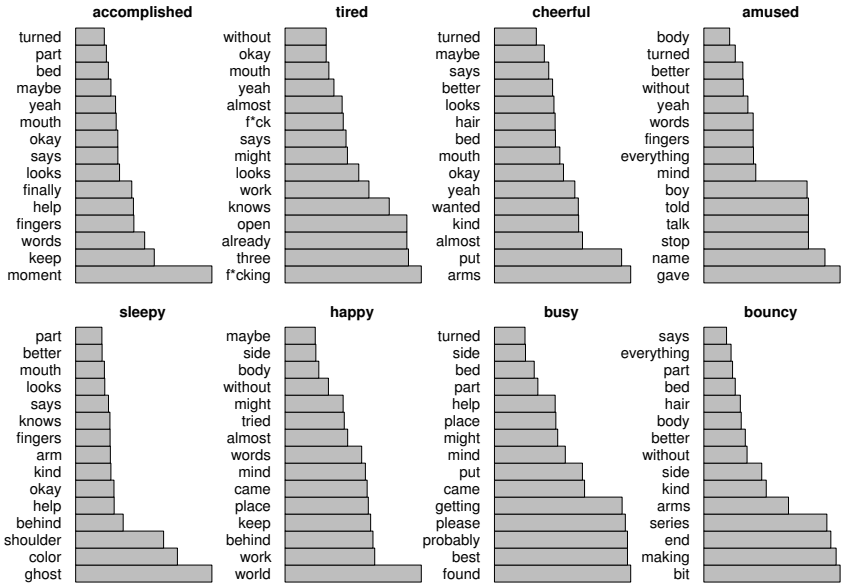

Fig. 5. Most important words according to TF ICF (with proper nouns removed)

lated cosine similarity between every pair of documents. The same procedure was carried out for documents filtered by current mood tag.

The only pair of tags "nervous" and "accomplished" has the distance between mood labels shorter than the average distance between different authors. This is probably because they carry a lot of objective content, which should have been filtered at earlier stages. The previously mentioned self-containing states of mind fall within the same group of labels, whose distances do not exceed the global average.

The vector model, therefore, contained enough information to distinguish emotions and what we needed was to find an approach to extracting words with maximum information. To solve this task, we used the Mutual Information feature selection algorithm [12].

Application of the mutual information feature selection algorithm showed that the word “happy” provided relevant information about the mood of an author. However, the top twelve terms for the “happy” category only contained two emotional words included in the Hedonometics or ANEW list (“happy” and “wonder”).

We saw that, according to mutual information feature selection, many of the categories were determined by the terms not included in emotional words lists. Then we checked whether or not category name synonyms obtained from WordNet were frequently encountered in the documents labeled with the same mood [13]. Most of the time mood was not specified in the text in an obvious

accomplished		**tired**		**cheerful**		**sleepy**	
eye	0.0084	three	0.0014	yes	0.0005	found	0.0009
head	0.0080	got	0.0011	feel	0.0004	name	0.0007
pull	0.0079	home	0.0010	still	0.0004	probabl	0.0007
turn	0.0074	day	0.0010	like	0.0004	knew	0.0007
smile	0.0072	lot	0.0009	way	0.0003	side	0.0006
arm	0.0071	tell	0.0009	right	0.0003	bad	0.0006
first	0.0070	realli	0.0009	see	0.0003	tri	0.0006
pair	0.0069	far	0.0008	guy	0.0003	rate	0.0006
behind	0.0069	think	0.0008	think	0.0003	time	0.0005
hand	0.0068	let	0.0008	girl	0.0003	walk	0.0005
away	0.0067	time	0.0008	hold	0.0002	mayb	0.0005
side	0.0063	part	0.0008	just	0.0002	find	0.0005

amused		**happy**		**busy**		**bouncy**	
knew	0.0008	happi	0.0020	thank	0.0012	your	0.0006
still	0.0008	dont	0.0006	set	0.0011	far	0.0006
way	0.0007	found	0.0005	pleas	0.0010	someon	0.0005
week	0.0007	ive	0.0005	use	0.0010	feel	0.0004
cant	0.0007	wonder	0.0004	everi	0.0010	ask	0.0004
pleas	0.0007	thank	0.0004	ask	0.0009	sinc	0.0003
feel	0.0007	wait	0.0004	man	0.0009	talk	0.0003
far	0.0006	also	0.0004	ill	0.0007	word	0.0003
will	0.0006	home	0.0004	leav	0.0007	dont	0.0003
tri	0.0005	one	0.0004	found	0.0007	guy	0.0003
found	0.0005	isnt	0.0003	comment	0.0006	way	0.0003
day	0.0005	day	0.0003	tag	0.0006	see	0.0003

Table 2. Most important words according to mutual information feature selection algorithm

way. Only 14 of 50 popular moods or their synonyms are frequently encountered in a text tagged with the same mood: crazy (stressed, crazy, sick), curious (good, curious, sore), depressed (depressed, hopeful, artistic), ecstatic (hopeful, ecstatic, productive), hopeful (hopeful, crazy, sad), pissed off (pissed off, nervous, okay), sad (sad, frustrated, pissed), sick (confused, crazy, sick), sleepy (stressed, sleepy, sick), sore (sad, ecstatic, sore), stressed (stressed, depressed, curious).

Surprised by such results, we tried to analyze the document using words from the Hedonometrics list. Analysis of frequencies of top twelve words from the Hedonometrics list in texts written in different moods showed that these words have the most common usage in emotional states different from “happy” (Table 3). Only one word “successful” is used more frequently by authors who tagged their message with the current mood “happy”.

	accomplished	tired	cheerful	sleepy	amused	happy	busy	bouncy
laughter	1.37	1.45	1.28	1.16	2.22	1.54	1.61	2.25
love	23.97	25.32	29.18	20.92	27.27	27.29	26.46	30.04
happy	7.28	6.74	8.90	7.98	7.83	11.95	7.71	12.14
laugh	10.83	12.03	12.43	7.28	9.79	7.48	8.88	9.01
excellent	0.50	0.33	0.56	0.54	0.26	0.46	0.54	0.38
joy	0.89	0.66	1.52	0.77	1.17	1.31	0.45	1.25
successful	0.64	1.06	0.72	0.62	0.65	1.39	0.45	0.13
win	1.42	1.19	1.20	0.85	3.26	0.92	1.97	0.75
rainbow	0.56	0.20	0.32	0.31	0.26	0.08	0.09	0.50
smile	27.29	25.45	28.86	21.23	23.10	25.67	20.63	24.53
won	0.73	0.60	1.20	0.46	0.91	0.39	0.90	1.25
pleasure	1.59	1.26	1.60	1.24	1.04	1.46	1.97	3.13
celebration	1.14	1.12	1.92	0.54	1.30	0.69	1.17	0.75

Table 3. Word frequencies $\times 10^5$

Conclusion

Analysis of user-annotated blog messages showed that connections between emotions and their linguistic expression could not necessarily be straightforward as is usually expected by compilers of emotional words lists. The most frequent words in each mood category are not included in the list of emotional terms. Application of TF-IDF and the calculated TF-ICF coefficient did not change the situation. Words with the highest scores continue not to be included in popular lists used for mood analysis. Application of the Mutual Information feature selection algorithm allowed us to find the most important words in each category, but only few of them are included in popular lists of emotional words. We can confirm that. according to the mutual information coefficient, the word "happy" has high discriminative power, while other words from the Hedonometrics list were not as successful.

People show a high ability to evaluate emotions of other persons even in a computer-mediated environment, although the way we can understand other people's emotions still raises questions. On the one hand, the ability to understand emotions also exists in situations where emotions are not explicitly expressed; on the other hand, our analysis showed a paradoxical situation when the terms used for evaluation of emotions are not among the top 20 frequent or discriminative words for each of mood categories. These facts raise a question about psychological validity of straightforward techniques for measuring emotions.

In our further research we plan to move in two different directions. One is to compare results of emotion analysis by applying the classical lexical approach with two dictionaries (ANEW and Hedonometrcs) and Naïve Bayes algorithm using the probabilities calculated in the current research. The other direction is to test agreement between naïve or expert annotators and authors of mood

labels. We also intend to develop more sophisticated procedures to filter objective content and detect invalid entries, establish a meaningful connection between content and label and further extend our database to improve validity of our study.

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, ACM (2008) 183–194
2. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 79–86
3. Yu, L.C., Wu, J.L., Chang, P.C., Chu, H.S.: Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Know.-Based Syst. (March 2013) 89–97
4. Mislove, A., Lehmann, S., Ahn, Y.Y., Onnela, J.P., Rosenquist, J.: Pulse of the Nation: U.S. Mood Throughout the Day inferred from Twitter (2010)
5. Bradley, M.M., Lang, P.J.: Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, Technical Report C-1, The Center for Research in Psychophysiology, University of Florida (1999)
6. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: Proceedings of the International AAAI Conference on Weblogs and Social Media. (2010) 122–129
7. Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M.: Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PloS one (2011)
8. Lansdall-Welfare, T., Lampos, V., Cristianini, N.: Effects of the Recession on Public Mood in the UK. In: Proceedings of the 21st international conference companion on World Wide Web, ACM (2012) 1221–1226
9. Strapparava, C., Valitutti, A., Stock, O.: The affective weight of lexicon. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation. (2006) 423–426
10. Gill, A.J., Gergle, D., French, R.M., Oberlander, J.: Emotion rating from short blog texts. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2008) 1121–1124
11. Hancock, J.T., Landrigan, C., Silver, C.: Expressing emotion in text-based communication. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM (2007) 929–932
12. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Volume 1. Cambridge university press Cambridge (2008)
13. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM (November 1995) 39–41