Creating English and Japanese Twitter Corpora for Emotion Analysis

Anna Danielewicz-Betz*, Hiroki Kaneda*,
Maxim Mozgovoy*, and Marina Purgina†

*School of Computer Science and Engineering,
The University of Aizu, Japan

†Department of Computer Systems & Software Engineering,
St. Petersburg State Polytechnic University, Russia

Abstract

This paper describes the principles used to collect open English and Japanese Twitter corpora for emotion analysis. We create a set of eight emotions, based on Ekman and Plutchik categories, and applicable both to English-speaking and Japanese cultures, ensuring that each tweet in our subset of TREC'2011 collection is coded independently by three individuals. We analyse emotions, contained in the resulting corpora, and briefly discuss obtained results. This work will provide valuable insights for researchers interested in emotion analysis of micro-blogosphere and comparative studies of English and Japanese tweets.

Keywords: microblogs, Twitter, corpus, emotion analysis.

1. Background

The analysis of emotions as depicted in blogosphere has a number of practical applications, ranging from social studies and forensics to business analytics and marketing. The rise of microblog platforms, such as Twitter, opened new challenges to sentiment analysis. Microblogs require separate since they differ significantly from blogs in terms of length, lexico-grammar, style, and content. For instance, the most popular microblogging platform Twitter has a limitation of 140 characters per message, thus effectively forcing the users to formulate what they wish to express in a very concise way. Researchers report that such *tweetspeak* is different from other written English genres in many respects, and characterized by an extensive use of acronyms, abbreviations, misspellings and slang words [1]. Furthermore, as noted in [2], the informal nature of microblogging encourages the users to write frequently, expressing their daily thoughts and emotions, which results in less polished text that is likely to be more emotionally charged than other written texts.

The study of emotions in text typically relies on the analysis of annotated corpora, providing samples of texts that contain traces of emotional manifestations, previously identified by human coders. However, to our knowledge, there have been few research activities aiming at creation of such corpora of microblog texts. Notable exceptions include a collection of tweets about people and/or film reviews] classified as *positive*, *negative*, *neutral*, or *objective* [3]; Sanders Corpus of tweets that contains the words Apple, Google, Microsoft or Twitter, classified as *positive*, *neutral*, *negative*, and *irrelevant* [4]; and the EmpaTweet corpus, containing microblog messages related to certain predefined topics and classified according to seven emotional categories [2].

In the present paper, we discuss our own research effort to create a corpus for automated emotion analysis in microblogs. Our project is similar to that of [2], but has a number of important distinctive features discussed below. While still a work-in-progress, our corpus is already large enough to provide valuable information on emotion in Twitter messages.

2. The Principles of Corpus Organization

Being interested specifically in analysing *emotions*, we see the primary goal of our corpus in providing reliable classification of Twitter messages according to a set of predefined emotional categories. In our system, the categories represent Ekman's eight basic emotions [5], i.e. *anger, disgust, fear, happiness, sadness, surprise* partly overlapping with those of Plutchik's wheel [6]. The two extra categories — *embarrassment* (a negative self-conscious emotion) *and pride in achievement* (associated with positive self-evaluations) — were selected for a specific reason from Ekman's extended (by 11 additional emotions) set [7], the difference being that they may not always be decoded via facial expressions. Our decision was made due to importance of those two emotions in Japanese culture and assumed cross-cultural differences as to their triggers (antecedents that bring about an emotion) and depiction in Japanese and English, respectively.

Each tweet is represented with a simple Boolean flag (i.e. emotion is present or absent). In addition, we have a "skip it" category, reserved for tweets containing gibberish or foreign (i.e. non-English or non-Japanese, respectively) language text.

In our project, we have been compiling two separate, English and Japanese, corpora. Being the second-popular language on Twitter and comprising 14% of all tweets (while the share of English is 39% [8]), Japanese represents a significant part of the world's microblogosphere, and should not be overlooked. Furthermore, our two corpora, created on the same basic principles, can be valuable for comparative studies of linguistic representation of emotions in different languages and cultures. Additionally, we assume that the Japanese would display more emotionally charged content in tweets rather than in face-to-face communication where overt expression of emotion is rare. This is due to culturally embedded emotion regulation and suppression, mostly as a result of direct socialisation of cultural values. This will be investigated further at later stage and falls outside the scope of this paper.

We should also note that the initial microblog collection given to the human coders for classification consists of unsorted and unfiltered subset of Twitter messages from the TREC'2011 dataset [9]. Therefore, our corpus is not biased towards any specific topic or emotion: it represents the actual state of microblogosphere reflected in the TREC dataset.

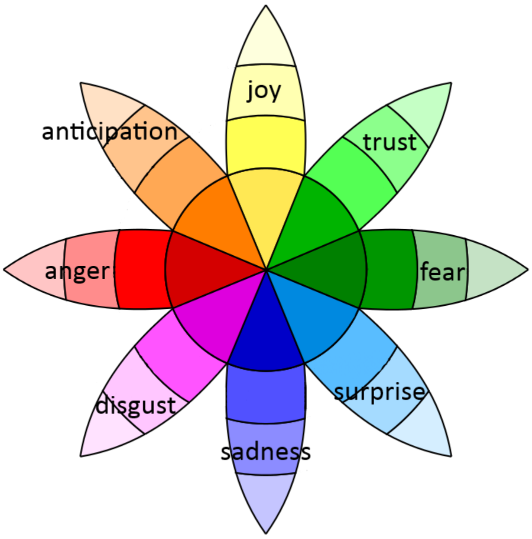


Fig.1: Plutchik's wheel of basic emotions [simplified]

The use of TREC dataset has another perhaps unobvious advantage. Due to Twitter's license restrictions, it is not possible to distribute complete collections containing actual text messages. Instead, available corpora contain unique tweet IDs that can be used to retrieve the corresponding messages from Twitter's servers. However, this approach has a significant drawback: the resulting corpora will degrade over time, since users can remove or protect their tweets from viewing. The study [10] reports that in early November 2011 around 21.2% tweets of the original TREC collection (containing approximately 16 million tweets posted between 23 January and 8 February 2011) were no longer available. We presumed that, over three years later, it would be rather unlikely for the Twitter users to start revising/deleting their messages written back in 2011; and indeed, for our subset of TREC collection (728,951 tweets) only 150,744 tweets were no longer available (20.7%). Therefore, a collection based on TREC'2011 data should be less susceptible to degradation than a collection obtained with a crawl of recent microblog posts.

Each tweet is coded by three individuals, so we can readily identify the final list of categories assigned to a tweet by means of a simple voting. The collection was initially separated into 'Japanese' and 'non-Japanese' parts. Our observations are generally consistent with [8]: the share of Japanese-language tweets in our collection is 12.78%. Fortunately, certain distinctive properties of Japanese writing (in particular, the presence of *hiragana* and *katakana* characters) allow isolating Japanese-language tweets automatically with high accuracy. Due to the unconventional nature of the English *tweetspeak*, we decided to treat all non-Japanese tweets as ‘English’, and rely on our coders’ manual work instead of resorting to automated natural language processing instruments to separate English from non-English messages.

3. Corpus Analysis

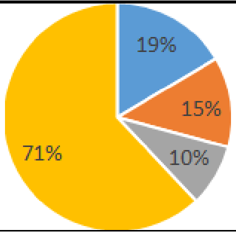
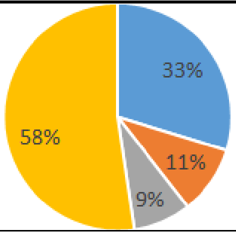
The status quo of our corpora as of November 2014 is summarized in Tables 1 and 2.

Table 1: General information about the corpora

Corpus	Tweets (coded by 3 people)	Identified as non-gibberish[1] by the majority of coders	Identified as non-gibberish or gibberish unanimously
English	1634	813 (49.8%)	1132 (69.3%)
Japanese	4338	3852 (88.8%)	2754 (63.5%)

Table 2: Relative contribution of individual coders

English		Japanese	
Graded 10+ tweets	Graded 100+ tweets	Graded 10+ tweets	Graded 100+ tweets
39 pers.	8 pers.	54 pers.	24 pers.

Tweets graded by the top three contributors and the rest of the group	
English	Japanese
	

Tables 3 and 4 have been calculated only for the tweets marked as non-gibberish by the majority of coders. Emotion score (ES) has been calculated as the percentage of tweets in which a given emotion was detected by least two of three coders. Agreement factor (AF) has been calculated as the percentage of tweets that have the presence or absence of a given emotion assigned by all three coders unanimously. For example, “Ang ES = 5.6” indicates that 5.6% of the non-gibberish tweets were marked as “anger” by 2 or 3 coders; whereas “Ang

[1] Note that we asked our coders to mark as gibberish all the tweets written in any language other than English or Japanese, respectively.

AF = 86.5” means that 86.5% of the non-gibberish tweets were marked as “anger” or “non-anger” by all three coders.

Approximately 22.8% of the non-gibberish tweets in the English corpus and 16.1% of the non-gibberish tweets in the Japanese corpus were unanimously marked as carrying no emotion by three independent coders. Voting by the majority of coders identifies 54.6% of the tweets in the English corpus and 72.7% of the tweets in the Japanese corpus as carrying no emotion.

Table 3: ES and AF for each of eight basic emotions (English corpus)

	Ang	Dis	Sad	Sur	Fea	Hap	Pri	Emb
ES	5.6	3.4	5.9	3.2	0.3	26.4	3.9	0.7
AF	86.5	88.3	86.0	86.7	97.7	65.9	81.2	94.0

Table 4: ES and AF for each of eight basic emotions (Japanese corpus)

	Ang	Dis	Sad	Sur	Fea	Hap	Pri	Emb
ES	1.7	2.0	6.6	5.0	0.6	11.1	1.6	0.5
AF	89.6	87.1	82.5	80.8	94.3	63.8	82.0	94.4

Table 5: Relative share of individual emotions (ES / max(ES_{ang}, …, ES_{emb})

	Ang	Dis	Sad	Sur	Fea	Hap	Pri	Emb
English	0.21	0.13	0.22	0.12	0.01	1.00	0.15	0.03
Japanese	0.15	0.18	0.59	0.45	0.05	1.00	0.14	0.05

Table 5 depicts the normalized emotions (scaled to the same factor), which allows for an analysis of the relative contribution of individual emotions to the pool of all the emotions detected. We can observe, for instance, that, in the English corpus, there are almost 5 times fewer occurrences of *sadness* than *happiness*, whereas in the Japanese corpus only 1.7 times, i.e. the relative proportion of sadness is much higher in the Japanese data set. The same applies to *surprise*, with the remaining emotions not exhibiting any considerable differences. We assume that *happiness* is the most commonly coded emotion because it also includes *joy*, *anticipation*, *excitement* and similar emotions.

4. Culture-specific source and display of emotion

Although basic emotions are considered universal, the meaning, circumstances, and the associated tasks related to their generation are culture-specific.

Japan is one of those cultures that greatly value face, and losing face in public is one of the worst things that can happen to a person, thus causes fear. Control over emotional display in public, including the emotions that we are tracing in our datasets, contributes to face management. Face has been equated with dignity, prestige and reputation.

It can also be said that the Japanese are generally very shy and despise being embarrassed. Maintaining dignity and avoiding embarrassment are very important in Japan (cf. [11]). Benedict [12] depicted Japanese culture as a “shame culture” relying on “external sanctions for self-respect”, claiming that the American culture was more of a “guilt culture” based on “internalised conviction of sin.”

It is also worth mentioning that cross-cultural differences have been reported regarding “feel good” emotions such as pride in achievement, whereby the Japanese subjects, representing a collectivist culture, tend to derive those emotions from social engagement (e.g. related to respect from friends), whereas for the American subjects (but also British), highly individualistic on the whole, successful achievement of goals is associated with personal recognition and pride and generally “feeling good” about themselves [13–15].

Despite this, we cannot assume to be able to detect any considerable differences in the content analysis of the Japanese and English tweets, respectively, as mapped against given emotional categories (mainly of highest respective frequency and corresponding with high coder agreement). What we have observed so far is that in the Japanese tweet corpus fewer emotions have been coded on the whole. As for the relative contribution of individual emotions, as mentioned above, *sadness* and *surprise* tend to prevail. One of the reasons for this might be that people do not necessarily express their true emotions (or anything verifiable for that matter) on Twitter. Moreover, expression of face-threatening emotions, such as *embarrassment*, is face threatening to the extent that even anonymous account owners will not tweet about it. So we can tentatively conclude that the Japanese and English micro-blogospheres are surprisingly similar.

5. Conclusion

While sentiment analysis and emotion analysis is a topic of numerous research efforts, there is a lack of open text corpora that can serve as a basis for emotion detecting systems and (micro-)blogosphere analysis. We address this issue by coding a fragment of TREC’2011 dataset with Boolean flags, corresponding to eight basic emotions, derived from Plutchik and Ekman emotional models.

Our preliminary result show that, in general, emotions are distributed very unevenly, with positive emotions (forming a wide category of *happiness* in our system) prevailing. At the same time, certain emotions, such as *fear* or *embarrassment* are virtually absent in the corpora. These observations hold both for English and Japanese microblogs.

In the future, we plan to continue coding the corpora, focusing on the most prevalent emotions by refining our set of emotional categories. We will also make the corpora openly accessible to support further research efforts in this area.

Acknowledgement

The research work is supported by JSPS KAKENHI Grant Number 25330410.

References

[1] E. Glennon, L. Sankar, and H. V. Poor, “Twitter vs. printed English: An information-theoretic comparison,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 3069–3072.

[2] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. M. Harabagiu, “EmpaTweet: Annotating and Detecting Emotions on Twitter,” in *LREC*, 2012, pp. 3806–3813.

[3] L. Chen, W. Wang, M. Nagarajan, S. Wang, and A. P. Sheth, “Extracting Diverse Sentiment Expressions with Target-Dependent Polarity from Twitter,” in *ICWSM*, 2012.

[4] N. J. Sanders, *Sanders-Twitter Sentiment Corpus.* Available: http://www.sananalytics.com/lab/twitter-sentiment/.

[5] P. Ekman, “An argument for basic emotions,” *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[6] R. Plutchik, “The nature of emotions,” *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[7] P. Ekman, “Basic Emotions,” in *Handbook of Cognition and Emotion*: John Wiley & Sons, 1999, pp. 45–60.

[8] Semiocast, *Arabic highest growth on Twitter English expression stabilizes below 40%.* Available: http://semiocast.com/publications/2011_11_24_Arabic_highest_growth_on_Twitter.

[9] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff, “Overview of the trec-2011 microblog track,” in *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, 2011.

[10] R. McCreadie, I. Soboroff, J. Lin, C. Macdonald, I. Ounis, and D. McCullough, “On building a reusable Twitter corpus,” in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 1113–1114.

[11] D. Y.-F. Ho, W. Fu, and S. M. Ng, “Guilt, shame and embarrassment: Revelations of face and self,” *Culture & Psychology*, vol. 10, no. 1, pp. 64–84, 2004.

[12] R. Benedict, *The chrysanthemum and the sword: Patterns of Japanese culture*: Houghton Mifflin Harcourt, 1967.

[13] S. Kitayama, H. R. Markus, and M. Kurokawa, “Culture, emotion, and well-being: Good feelings in Japan and the United States,” *Cognition & Emotion*, vol. 14, no. 1, pp. 93–124, 2000.

[14] S. Ting-Toomey, *Communicating across cultures*: Guilford Press, 2012.

[15] J. Stoeber, O. Kobori, and Y. Tanno, “Perfectionism and Self-conscious Emotions in British and Japanese Students: Predicting Pride and Embarrassment after Success and Failure,” *European Journal of Personality*, vol. 27, no. 1, pp. 59–70, 2013.