# Creating English and Japanese Twitter Corpora for Emotion Analysis

A. Danielewicz-Betz, H. Kaneda, M. Mozgovoy, and M. Purgina

***Abstract*—This paper describes the principles used to collect open English and Japanese Twitter corpora for emotion analysis. We have created a set of eight emotions, based on Ekman and Plutchik categories, applicable both to the 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 the 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.**

***Index Terms*—Emotion, corpus, microblogs, Twitter.**

## I. Introduction

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 Tumblr or Twitter, opened new challenges to sentiment analysis. Microblogs require separate treatment 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 characterised by an extensive use of acronyms, abbreviations, misspellings, and slang words [1]. Furthermore, as noted by [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 writings.

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 Empa Tweet corpus, containing microblog messages related to certain predefined topics and classified according to seven emotional categories [2].

Emoticons are widely used in microblog texts. They tend to emphasise a given emotion expressed, although they also might be used habitually, thus contradicting it in a sense. To our knowledge, the analysis of eastern-style emoticons in microblogs, such as tweets, has not attracted much academic interest so far. As part of our sentiment analysis of microblogs, we therefore also focus on a relationship between emotions and emoticons

In the present paper, we discuss our research efforts 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 insights into emotion in Twitter messages.

## II. 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 six basic emotions [5], i.e. *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*, overlapping with the corresponding emotions from Plutchik's wheel that are more or less of the same intensity. Thus, the original wheel (see Fig. 1) was simplified [6]. The two extra categories — *embarrassment* (a negative self-conscious emotion) and *pride in achievement* (associated with positive self-evaluation) — 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 (i.e. antecedents that bring about a given 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 micro blogosphere and should not be overlooked. Furthermore, our two corpora, created on

Manuscript received January 15, 2015; revised June 3, 2015.

A. Danielewicz-Betz, H. Kaneda, and M. Mozgovoy are with School of Computer Science and Engineering, the University of Aizu, Japan (e-mail: abetz@u-aizu.ac.jp, mozgovoy@u-aizu.ac.jp, mapurgina@gmail.com).

M. Purgina is with the Department of Computer Systems and Software Engineering, St. Petersburg State Polytechnic University, Russia (e-mail: rainbowdash7777@gmail.com).

the same basic principles, can be valuable for comparative studies of linguistic representation of emotion 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 a 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 an 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 micro-blogosphere reflected in the TREC dataset.

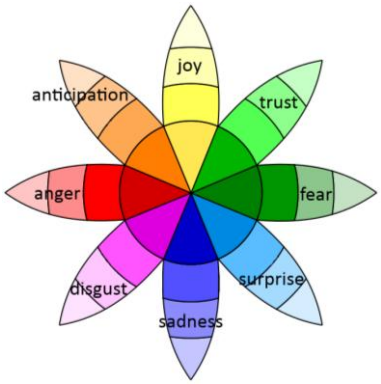


Fig. 1. Plutchik's wheel of basic emotions [simplified].

The use of TREC dataset has another, perhaps unobvious advantage. Due to Twitter's licence 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 by crawling 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 annotation instead of resorting to automated natural language processing instruments to separate English from non-English messages.

## III. Corpus Analysis

The status quo of our corpora as of November 2014 is summarized in Table I and Table II.

TABLE I: General Information about the Corpora

| Corpus | English | Japanese |
|---|---|---|
| **Tweets (coded by 3 people)** | 1634 | 5869 |
| **Identified as non-gibberish by the majority of coders** | 813 (49.8%) | 5335 (90.9%) |
| **Identified as non-gibberish or gibberish unanimously** | 1132 (69.3%) | 3902 (66.4%) |

TABLE II: Relative Contribution of Individual Coders

| | English | Japanese |
|---|---|---|
| **Graded 10+ tweets** | 57.4% | 82.4% |
| **Graded 100+ tweets** | 11.8% | 39.0% |

Table III and Table IV 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 at 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 AF = 86.5" means that 86.5% of the non-gibberish tweets were marked as "anger" or "non-anger" by all three coders.

TABLE III: ES and AF for Each of Eight Basic Emotions (English Corpus)

| | ES | AF |
|---|---|---|
| Anger | 5.6 | 86.5 |
| Disgust | 3.4 | 88.3 |
| Sadness | 5.9 | 86.0 |
| Surprise | 3.2 | 86.7 |
| Fear | 0.3 | 97.7 |
| Happiness | 26.7 | 65.9 |
| Pride | 3.9 | 81.2 |
| Embarrassment | 0.7 | 94.0 |

Approximately 22.8% of the non-gibberish tweets in the English corpus and 21.0% 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 identified 43.0% of the tweets in the English corpus and 53.0% of the tweets in the Japanese corpus as carrying no emotion.

TABLE IV: ES AND AF FOR EACH OF EIGHT BASIC EMOTIONS (JAPANESE CORPUS)

| | ES | AF |
|---|---|---|
| Anger | 1.3 | 91.6 |
| Disgust | 1.9 | 88.5 |
| Sadness | 6.3 | 83.1 |
| Surprise | 4.1 | 83.5 |
| Fear | 0.6 | 95.2 |
| Happiness | 10.7 | 67.2 |
| Pride | 1.8 | 79.9 |
| Embarrassment | 0.5 | 95.4 |

TABLE V: RELATIVE SHARE OF INDIVIDUAL EMOTIONS (ES / MAX ($ES_{ANG}$,…,$ES_{EMB}$))

| | English | Japanese |
|---|---|---|
| Anger | 0.21 | 0.12 |
| Disgust | 0.13 | 0.18 |
| Sadness | 0.22 | 0.59 |
| Surprise | 0.12 | 0.38 |
| Fear | 0.01 | 0.06 |
| Happiness | 1.00 | 1.00 |
| Pride | 0.15 | 0.17 |
| Embarrassment | 0.03 | 0.05 |

Table V depicts the normalised 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.

## IV. EMOTICON ANALYSIS

This section is devoted to emoticons that were detected in the Japanese corpus. The reasons why emoticons accompanying English tweets are excluded are: a) the limited scope of this thesis research; b) the fact that the author is Japanese and coded the Japanese tweets only; c) larger variety and higher frequency of occurrence of emoticons in the Japanese data set.

As can be seen on Twitter and other social media, various emoticons are frequently used in online communication [11]. Also, recent work has found that emoticons can emphasise emotion of a given text [12]. For this reason, we assume that analysing emoticons can assist sentiment analysis.

In general, a Japanese emoticon is composed of three parts and two areas: BORDERS, EYES, MOUTH, and INNER/OUTER additional area [13]. To make it simple, we define a combination of the left border BL and the right border BR as BORDER(BL,BR); the left eye EL and the right eye ER as EYES(EL,ER), etc. For example, emoticon "(o´_`o)/" consists of BORDER((,)), EYES(´,`), MOUTH(_), INNER(o,o) and OUTER(,/).

TABLE VI: TYPES OF BORDER COMBINATIONS AND NUMBER OF EMOTICONS THAT USE A GIVEN COMBINATION AS A BORDER

| Type of border | Number of emoticons |
|---|---|
| BORDER((,)) | 965 (18.1%) |
| BORDER([,]) | 16 (0.3%) |
| BORDER({,}) | 11 (0.2%) |
| BORDER(<,>) | 4 (0.1%) |

As depicted in Table VI, 18.09% of the Japanese tweets in our corpus include emoticons with BORDER((,)). By contrast, other bracket characters rarely form the emoticon border.

TABLE VII: ES FOR SPECIFIC EMOTION PARTS

| Ang. | Dis. | Sad. | Sur. | Fea. | Hap. | Pri. | Emb. |
|---|---|---|---|---|---|---|---|
| **All emoticons** | | | | | | | |
| 0.13 | 0.19 | 2.08 | 0.82 | 0.06 | 5.08 | 0.28 | 0.11 |
| **EYES(^,^)** | | | | | | | |
| 0.00 | 0.00 | 0.13 | 0.28 | 0.00 | 2.64 | 0.09 | 0.02 |
| **EYES(;,;)** | | | | | | | |
| 0.00 | 0.02 | 0.88 | 0.09 | 0.00 | 0.09 | 0.00 | 0.02 |
| **EYES(´, `) / INNER(´, `)** | | | | | | | |
| 0.00 | 0.04 | 0.86 | 0.15 | 0.06 | 1.41 | 0.07 | 0.04 |
| **INNER(*)** | | | | | | | |
| 0.02 | 0.04 | 0.11 | 0.04 | 0.00 | 1.61 | 0.09 | 0.04 |
| **INNER(;)** | | | | | | | |
| 0.06 | 0.13 | 0.51 | 0.39 | 0.02 | 0.32 | 0.04 | 0.02 |
| **OUTER(Ƹ)** | | | | | | | |
| 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 0.02 | 0.00 | 0.00 |

TABLE VIII: RELATIVE SHARE OF EACH EMOTICON PART (ES/max($ES_{ana}$, …, $ES_{emb}$)

| Ang. | Dis. | Sad. | Sur. | Fea. | Hap. | Pri. | Emb. |
|---|---|---|---|---|---|---|---|
| **All emoticons** | | | | | | | |
| 0.03 | 0.04 | 0.41 | 0.16 | 0.01 | 1.00 | 0.06 | 0.02 |
| **EYES(^,^)** | | | | | | | |
| 0.00 | 0.00 | 0.05 | 0.11 | 0.00 | 1.00 | 0.04 | 0.01 |
| **EYES(;,;)** | | | | | | | |
| 0.00 | 0.02 | 1.00 | 0.11 | 0.00 | 0.11 | 0.00 | 0.02 |
| **EYES(´, `) / INNER(´, `)** | | | | | | | |
| 0.00 | 0.03 | 0.61 | 0.11 | 0.04 | 1.00 | 0.05 | 0.03 |
| **INNER(*)** | | | | | | | |
| 0.01 | 0.02 | 0.07 | 0.02 | 0.00 | 1.00 | 0.06 | 0.02 |
| **INNER(;)** | | | | | | | |
| 0.11 | 0.26 | 1.00 | 0.78 | 0.04 | 0.63 | 0.07 | 0.04 |
| **OUTER(Ƹ)** | | | | | | | |
| 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.14 | 0.00 | 0.00 |

Table VII and Table VIII show some tendencies among each emoticon parts. The emoticon with EYES(^,^), INNER(*) or INNER(*,*), for instance, brings stronger *happiness* emotion than the other parts. On the other hand, we

can see that the emoticon with EYES(;,;) is often used to visualise *sadness*. In addition, EYES(´,`) and INNER(´,`) is used as sad or trustful eyes/eyebrows.

From Table VII and Table VIII we can detect that each emoticon part has a corresponding emotion. Also, comparing Table V with Table VIII, we can say that emoticons tend to be used to visually reinforce a given emotion, especially *happiness.*

## V. 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 represents 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, which may cause fear, for example. 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 becoming embarrassed. Maintaining dignity and avoiding embarrassment are of high importance in Japan (cf. [14]). Benedict [15] 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 [16]-[18].

Despite this, at the present stage, 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 emotion such as *embarrassment* is face threatening to such an 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.

## VI. Conclusion

While sentiment analysis and emotion analysis are 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 results 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 the English and Japanese microblogs. Further research is, therefore, needed to yield findings about emotions that are hardly present in the data sets coded so far.

As far as emoticons are concerned, it has been shown that approximately one in five Japanese tweets contains an emoticon. Furthermore, it has become clear that each emoticon part has different emotional effect. Additionally, in this study, emoticons have been identified particularly in messages falling under the broad category of *happiness*.

In the future, we intend 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.

## Acknowledgment

This 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 *Proc. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 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," *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," *ICWSM*, 2012.

[4] N. J. Sanders. Sanders-twitter sentiment corpus. [Online]. 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*, T. Dalgleish and M. Power, Eds., John Wiley & Sons, 1999, pp. 45–60.

[8] Semiocast SAS. (November 14, 2011). Arabic highest growth on Twitter. English expression stabilizes below 40%. [Online]. 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," presented at 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 *Proc. the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2012, pp. 1113–1114.

[11] M. Ptaszynski, R. Rzepka, K. Araki, and Y. Momouchi, "Research on emoticons: review of the field and proposal of research framework," in *Proc. the Seventeenth Annual Meeting of The Association for Natural Language Processing*, 2011, pp. 1159-1162.

[12] S. Aoki and U. Osamu, "A method for automatically generating the emotional vectors of emoticons using weblog articles," in *Proc. 10th WSEAS Int. Conf. on Applied Computer and Applied Computational Science*, Stevens Point, Wisconsin, USA, 2011.

[13] S. Bedrick, R. Beckley, B. Roark, and R. Sproat, "Robust kaomoji detection in Twitter," in *Proc. the Second Workshop on Language in Social Media*, Association for Computational Linguistics, 2012, pp. 56-64.

[14] 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.
[15] R. Benedict, *The Chrysanthemum and the Sword: Patterns of Japanese Culture*, Boston, USA: Houghton Mifflin Harcourt, 1967.
[16] 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.
[17] S. Ting-Toomey, *Communicating across Cultures*, New York: Guilford Press, 2012.
[18] 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.

**Anna Danielewicz-Betz** received her Ph.D. in applied linguistics, phonetics and English studies from the University of Saarland. She is currently an associate professor at the University of Aizu. Her interdisciplinary research interests include cross-cultural pragmatics, forensic linguistics, corporate discourse, and communication of emotion.

**Hiroki Kaneda** has recently received his BSc in computer science from the University of Aizu. He is particularly interested in sentiment analysis and expression of emotions in virtual communication.

**Marina Purgina** received her M.E. in computer science from St. Petersburg State Polytechnic. Her research interests include semantic information retrieval, natural language processing and computational aesthetics.

**Maxim Mozgovoy** received his Ph.D. degree in applied mathematics from St. Petersburg State University, and his Ph.D. in computer science from the University of Joensuu. He is currently an associate professor at the University of Aizu, where he conducts research in natural language processing and artificial intelligence.