Detecting and Forecasting Local Collective Sentiment Using Emojis

Mei Fukuda[1, 2 *], **Kazuyuki Shudo**[1 †], **Hiroki Sayama**[2]

[1] Department of Mathematical and Computing Science, Tokyo Institute of Technology
[2] Center for Collective Dynamics of Complex Systems, Binghamton University, State University of New York

Abstract

The analysis of collective social sentiment using large-scale data obtained from the Internet, such as social media data, has been actively conducted in recent years, but not many of them considered geographical distributions of sentiments or their spatial dynamics. In this study, we analyzed tweets associated with location information to detect local collective sentiment of each prefecture in Japan, especially in response to societal events. To extract positive and negative sentiments, we used emojis as language-independent universal indicators of positive/negative sentiments. We found that negative sentiment increased nationwide on the day of a major typhoon hit and after the onset of a COVID-19 pandemic in Japan, while positive sentiment increased around Christmas and the announcement of university or high school admission decisions, with some geographical variations. Then, we computed the correlation coefficient of the number of positive tweets on the same day and observed the relationship between the prefectures. We also built a linear regression model to forecast the local positive sentiment of a prefecture from other prefectures' past values, which achieved a reasonable predictability with $R^2 = 0.5$–0.6. Based on the coefficient matrix of this sentiment forecast model, we constructed a causal network of prefecture sentiments in Japan. Interestingly, the relationships among prefectures and their centralities changed significantly before and after the COVID-19 pandemic.

Introduction

The large-scale analysis of collective social sentiment using data obtained from the Internet, such as social media data and blogs, has been actively conducted in recent years (Dodds et al. 2011; Sano et al. 2019). Most of the studies to date have been conducted in large units, such as nationwide, and not many have considered the geographical distribution or spatial dynamics. Also, many previous analyses of collective emotions have used the frequency of occurrence of a word in an emotion dictionary, a list of words that describe a certain emotion, as an indicator. There are several challenges with this approach. For example, it is not possible to apply the exact same analysis across multiple languages. Moreover, some languages require advanced morphological analysis techniques to find words, some words have multiple conjugations, or the same word is written using different characters, which can make the analysis difficult.

In this study, we analyzed tweets associated with location information to detect local collective sentiment of each prefecture in Japan, especially in response to societal events. In addition, to extract positive and negative sentiments, we used emojis as language-independent indicators. Through the analysis, we found that negative sentiment increased nationwide on days when a major typhoon hit, when the death of a celebrity was reported, and after the onset of a COVID-19 pandemic in Japan, while positive sentiment increased around Christmas and the announcement of university or high school admission decisions, with some geographical variations. We also found that the change in the number of tweets before and after the pandemic showed a characteristic of the prefecture well. Also, the co-occurrence of sentiments among the prefectures became stronger after the pandemic in the large cities, while the ones among the surrounding prefectures weakened. We also built a linear regression model to forecast the local positive sentiment of a prefecture from other prefectures' past values, which achieved a reasonable predictability with $R^2 = 0.5$–0.6. Based on the coefficient matrix of this sentiment forecast model, we constructed a causal network of prefecture sentiments in Japan. Interestingly, the relationships among prefectures and their centralities changed significantly before and after the pandemic.

Data Collection

We calculate the level of positive or negative sentiment based on the number of tweets that contain specific emojis.

We first select the representative emojis. We use the Emoji Sentiment Ranking (Kralj Novak et al. 2015) to select emojis. They computed the sentiment score of emojis by the sentiment of the tweets in which they occurred, and they published a list of 751 frequently used emojis with data such as their occurrence rates, sentiment scores, and neutrality. Based on the Emoji Sentiment Ranking, we exclude emojis with a neutrality value of 0.45 or higher, and selected the 300 most frequently used emojis. This is to eliminate emojis of things that are hard to imagine as carrying feelings, such as emojis of books and so on. Next, we classify emojis with a sentiment score higher than 0.3 as positive and lower

*Current affiliation is Google Japan G.K.
†Current affiliation is Kyoto University.

Workshop Proceedings of AAAI ICWSM-2022
(SocialSens 2022), June 2022

than 0.3 as negative, and obtained 211 positive and 89 negative emojis. 0.3 is the average sentiment score reported in (Kralj Novak et al. 2015). Finally, by a manual check, we exclude the obviously wrong ones, then reduce them to 60 emojis each. Fig. 1 shows the selected positive and negative emojis.

Next, we obtain the number of tweets using the endpoint `GET /2/tweets/counts/all` of the Twitter API for Academic Research. We retrieve tweets in Japanese, excluding retweets and replies, that contained at least one of the representative emojis selected for each of the positive and negative. We obtain the number of tweets which are associated with each 47 prefectures for each of the 366 days from October 1, 2019 to September 30, 2020.

Data Analysis

We analyze the obtained data in several ways and discussed results.

Number of Tweets in Tokyo

We visualize and observe the number of tweets for some major prefectures. Fig. 2 shows the number of positive and negative tweets for 366 days in Tokyo. We make the following observations. First, more positive tweets were posted than negative tweets, which is about twice as much. The previous study indicates that the popular emojis are mostly positive (Kralj Novak et al. 2015), so this result is reasonable. Second, the number of tweets increases on weekends, i.e, Saturdays and Sundays. Similar findings are obtained in the existing study (França et al. 2016). Third, there are significantly fewer positive tweets from Tokyo after COVID-19. We divide the tweets into two groups, one from October 2019 to March 2020 (referred to as "before COVID-19") and the other from April to October 2020 (referred to as "after COVID-19"), and conducted a t-tests. While there is no significant difference between the two groups for negative tweets, there is a significant difference for positive tweets with $p < 0.01$. Fig. 3 shows the violin diagrams. Fourth, changes in the number of tweets reflect some societal events. For example, negative tweets increased on the day that major typhoon Hagibis hit Japan (October 12, 2019) and on the day that the sudden death of famous Japanese comedian Ken Shimura was reported (March 30, 2020). In contrast, the percentage of positive tweets increased on Christmas Day (December 25, 2019) and on the day that national universities announced their acceptance decisions (March 6, 2020). This result is also observed in other prefectures, as can be seen in the heatmap shown in Fig. 4. The following section describes more about the heatmap.

Geographical Heatmap

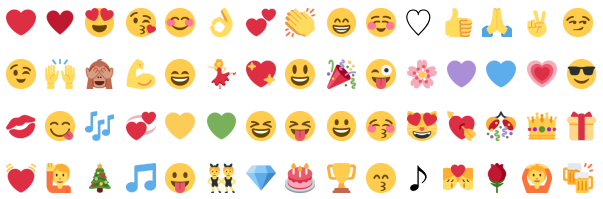
We calculate the value of the ratio of the number of positive tweets divided by the number of negative tweets for each day from October 1, 2019 to September 30, 2020, for each of the prefectures. We then create a geographic heatmap as shown in Fig. 4, with prefectures colored red if their value is large, i.e., more positive, white if it is about average (≈ 2.0), and blue if it is small, i.e., more negative. The heatmap for each day is also available at https://meipipo.github.io/emoji-sentiment/map.

As discussed in the previous section, several societal events are reflected in the sentiment throughout the country. Fig. 4(a) shows the days when the major typhoon landed on Japan. On October 12, 2019, the typhoon made landfall in the Kanto region in central Japan, where the capital Tokyo and other cities are located, causing storms, flooding, and other damage. Areas that were heavily damaged on this day show overall negative sentiment. On the following day, October 13, we can see positive sentiment beginning to return from southern Japan as the typhoon moves northward. Fig. 4(b)(c) shows the days before and after the days when the whole of Japan turned positive and negative, i.e., the days when national university acceptance announcements were made and the days before and after the death of a celebrity was reported. In both examples, sentiment is scattered until the day before the event occurs.

Changes in the Number of Tweets

Next, to observe whether people's sentiments change before and after COVID-19, we observe changes in the number of positive and negative tweets. We first divide the tweets into two groups, one from October 2019 to March 2020 (referred to as "before COVID-19") and the other from April to October 2020 (referred to as "after COVID-19").

Then, for each prefecture and for each of the positive and negative sentiments, we calculate the value of the change in the following way. First, we calculate the value by subtracting the average number of tweets before COVID-19 from the average number of tweets after COVID-19. Next, we normalize that value by dividing it by the average number of tweets for the entire 366 days. If this value is positive, it means that the number of tweets for that sentiment increased after COVID-19. The values for positive tweets are plotted as x and for negative tweets as y for each prefecture in Fig. 5. We make the following observations. First, the number of negative tweets increased in 39 of the 47 prefectures. And there are 36 prefectures out of 47 in which the number of negative tweets tended to increase more than the number of positive tweets. These results suggest that negative sentiments may have increased after COVID-19 due to anxiety over the pandemic and the impact of people refraining from going out. Second, the trend of this change differs depending on the characteristics of the prefectures. In prefectures that are large cities where many people commute to work or school, such as Tokyo, Osaka, and Aichi (indicated by the red stars in the figure), the number of negative tweets increased, while the number of positive tweets decreased. On the other hand, prefectures surrounding such large cities, such as Saitama, Chiba, Kanagawa, Nara, and Gifu (indicated by the green triangles in the figure), saw an increase in both negative and positive tweets. This can be attributed to the fact that working from home has become more common, and people who used to commute to large cities spend more time at home.



(a) Positive emojis.

(b) Negative emojis.

Figure 1: Representative emojis to extract sentiment from tweets.

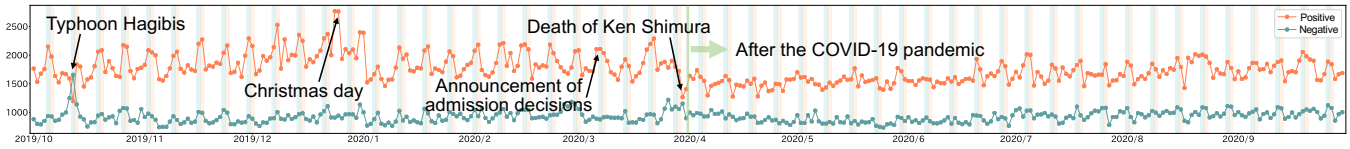
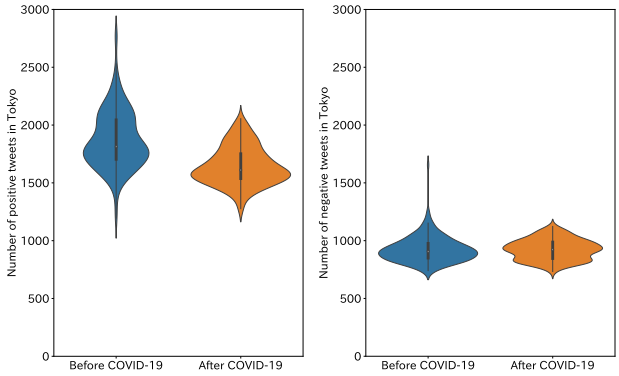


Figure 2: Number of positive and negative tweets in Tokyo from October 2019 to September 2020.



(a) Positive tweets. (b) Negative tweets.

Figure 3: Changes in the distribution of the number of tweets in Tokyo before and after COVID-19.

Co-Occurrence of Positive Sentiment Between Prefectures

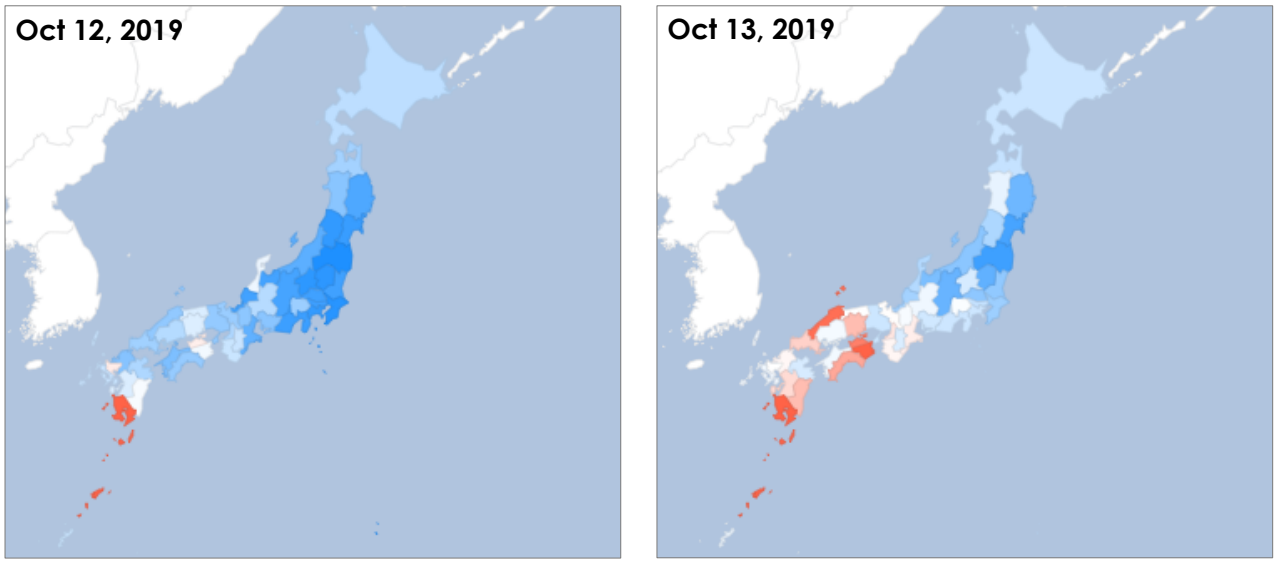
To observe the co-occurrence relationship between the prefectures, we calculate the correlation coefficient of the number of positive tweets on the same day for all pairs of prefectures. We discuss the differences in the values of the correlation coefficients for before and after COVID-19, respectively. This analysis reveals, first, that the correlation coefficients were generally smaller after COVID-19. The possible reason could be that the travel restrictions imposed by COVID-19 made it less likely that sentiment would be shared. On the other hand, the correlation coefficient between Tokyo and Osaka, the two largest cities in Japan, increased significantly from 0.77 to 0.85. Since the two cities are geographically located far apart, they may have shared fewer similar events before COVID-19, but after COVID-19, they may have been more affected by the pandemic because they are large cities, or they may have referenced the same national news more. These changes may have strengthened the co-occurrence of the positive sentiment.

We also build a network, by keeping the edges between prefectures that have high correlation coefficients (≥ 0.7). Fig. 6 visualizes the network before and after COVID-19. The number of nodes $|V|$ and edges $|E|$ in the networks are $(|V|, |E|) = (25, 95)$ before COVID-19 while $(|V|, |E|) = (16, 36)$ after COVID-19. Before COVID-19, the network was mainly formed by nodes of large cities and their surrounding prefectures with large populations, but after COVID-19, the number of nodes decreased, and the edges between the major cities basically remained.

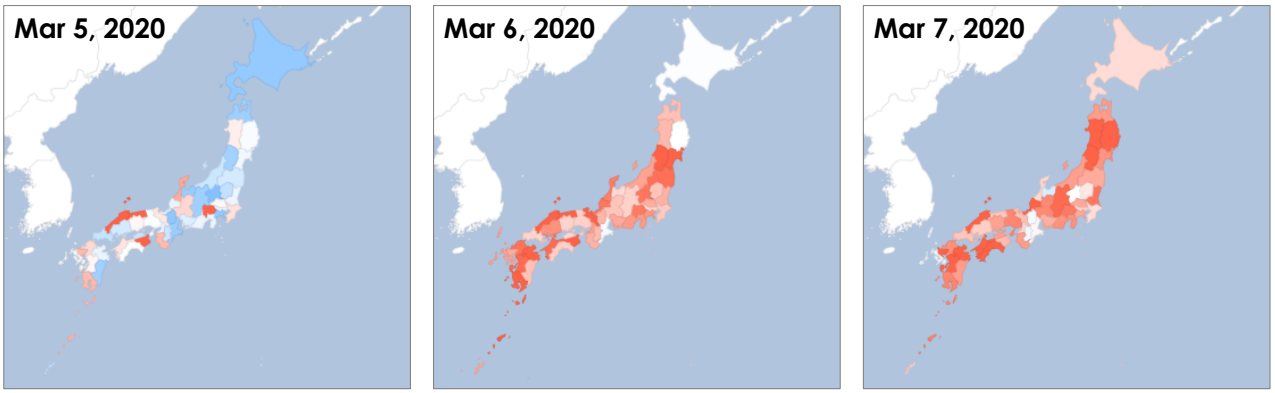
Forecasting Model

We build linear regression models to forecast future sentiment for each prefecture before and after the COVID-19. We use the number of positive tweets on a given day in a given prefecture as the dependent variable and the average number of positive tweets over the past five days for each of the 47 prefectures as the explanatory variables. The coefficient of determination R^2 for the model is about 0.5 to 0.6 on average. Then, we form a causal network from the coefficients of the model. If the contribution of the explanatory prefecture to prediction of the dependent prefecture was statistically significant ($p < 0.05$), we add a directed edge from the node of the dependent prefecture to the explanatory prefecture and use the absolute value of the coefficient as the weight of the edge. We calculate the PageRank centrality in that network and observe which prefectures have the larger influence on the other prefectures.

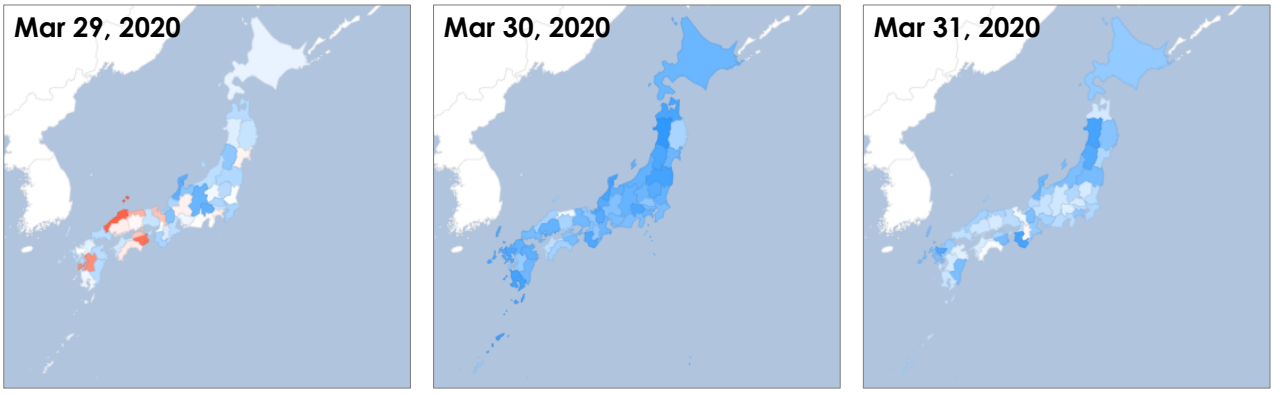
Fig 7 shows the geographical distributions and graph visualizations of centralities in the causal networks before and after COVID-19. In the geographical map (left), colors are darker for higher PageRank. In the visualized network (right), the size and color of the node indicates the value of the PageRank. We make the following observations. First, while there was a large gap between areas of



(a) Days when Typhoon Hagibis, which caused extensive damage, hit Japan and headed north.



(b) Around the day when the announcement of acceptance to national universities begins. The announcement began on March 6, with 60% of the universities announcing by March 6 and 80% by March 7.



(c) Around the day when it was reported that the famous comedian Ken Shimura passed away suddenly due to COVID-19. The news was reported on March 30.

Figure 4: Geographic heatmap of sentiment toward significant societal events in Japan for each prefecture.

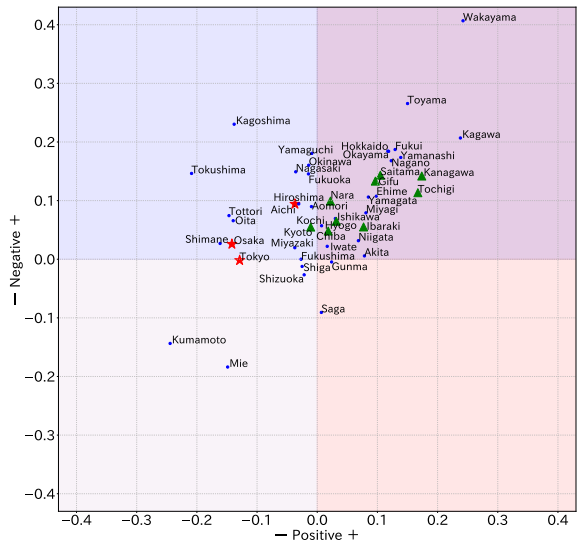
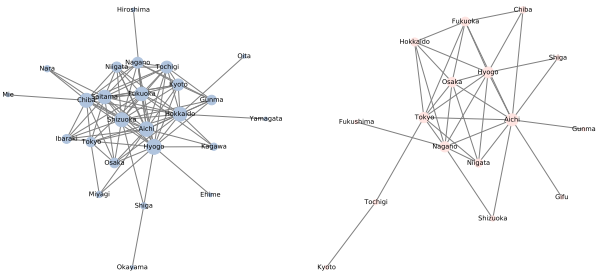


Figure 5: Changes in the total number of positive and negative tweets by prefecture before and after COVID-19.

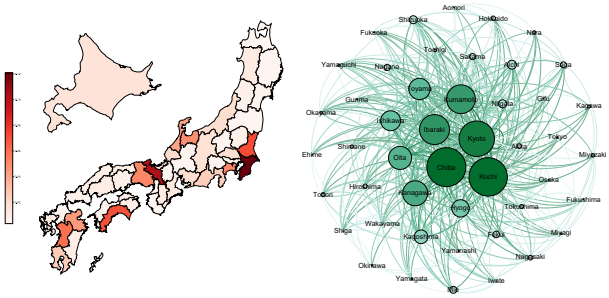
strong and weak influence before COVID-19, the size of influence is more dispersed after COVID-19. Second, prefectures that had a strong influence before COVID-19 are mainly those located around large cities such as Tokyo, Osaka, and Fukuoka, but after COVID-19, this regularity has disappeared. The reason why prefectures around large cities have a strong influence is that they have relatively large populations and large flows of people, and therefore are more likely to represent the national sentiment than the unique environment of the large cities themselves. After COVID-19, people moved less and the impact of the pandemic varied depending on prefectures, suggesting that there was no regularity in the influence of the sentiment.

These results were also observed for in-degree centrality and eigenvector centrality.

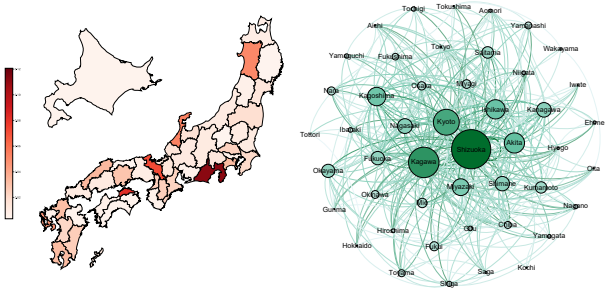


(a) Before COVID-19. (b) After COVID-19.

Figure 6: Networks between prefectures that have high correlation coefficients.



(a) Before COVID-19.



(b) After COVID-19.

Figure 7: Geographical distributions and graph visualizations of centralities in a causal network before and after the COVID-19 pandemic.

Conclusion

In this study, we observed collective sentiment in Japan using geo-tagged tweets to capture the geographical distribution and using emojis, which is a language-independent indicator of sentiments. Through the analysis of data for 366 days from October 2019 to September 2020, we examined various trends in collective sentiment in Japan. The results showed that the collective sentiment reflected societal events such as major disasters and the death of a celebrity. We also observed an increase in negative tweets after COVID-19 and a change in the co-occurrence relationship between prefectures. We also built a linear regression model to forecast the local positive sentiment of a prefecture from other prefectures' past values, and we indicated that the relationship of influence among prefectures also changed significantly before and after COVID-19.

Future work include the following. First, collect more data to carefully observe the impact of seasonal factors and to remove any bias toward individuals in less populated areas. Second, improve the forecasting model by using other methods, such as transfer entropy, autoregression model, LSTM and deep neural networks, and by adding other explanatory variables, such as known social events and weather. Third, apply the same analysis to other regions with different languages. This can be easily applied thanks to emojis and is expected to yield interesting results.

References

Dodds, P. S.; Harris, K. D.; Kloumann, I. M.; Bliss, C. A.; and Danforth, C. M. 2011. Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLOS ONE* 6(12).

França, U.; Sayama, H.; Mcswiggen, C.; Daneshvar, R.; and Bar-Yam, Y. 2016. Visualizing the "heartbeat" of a city with tweets. *Complexity* 21(6): 280–287.

Kralj Novak, P.; Smailović, J.; Sluban, B.; and Mozetič, I. 2015. Sentiment of emojis. *PLOS ONE* 10(12). URL http://kt.ijs.si/data/Emoji_sentiment_ranking/index.html.

Sano, Y.; Takayasu, H.; Havlin, S.; and Takayasu, M. 2019. Identifying long-term periodic cycles and memories of collective emotion in online social media. *PLOS ONE* 14(3).