

# SentiWorld: Understanding Emotions between Countries Based on Tweets

Sang-Jun Yea<sup>a,b</sup>, Sejin Kim<sup>a</sup>, John-Michaël TO<sup>a,c</sup>, Jae-Gil Lee<sup>a,\*</sup>

<sup>a</sup>KAIST, <sup>b</sup>KIOM, <sup>c</sup>Efrei  
{yeasangjun, ksj614, john-michael.to, jaegil}@kaist.ac.kr

## Abstract

In order to understand emotions between countries, we collected around 25 million tweets, analyzed them using statistical and network analysis methods, and visualized the analytic results as both a sentiment map and a sentiment network. Our web system, which we call *SentiWorld*, is accessible via <http://sentiworld.to.fr>.

## Introduction

Recognizing and understanding emotions in different cultures and countries is crucial for cross-cultural communication (Elfenbein and Ambady 2003). How, then, do we recognize the emotions that are directed at another country? How do we know which countries are emotionally closer to each other? How can we figure out what foreigners think about our country?

Many advanced studies have been performed to answer similar questions, typically based on tweets. These studies, however, have been limited to figuring out the sentiments of the tweets themselves, focusing on emotional reactions to special social events or using sentiment to mine political opinions. Recently, a few researchers have worked on sentiment in given destination countries without, however, looking at other countries' reciprocal sentiments for them (Mitchell et al. 2013).

Therefore, this study aims to analyze the sentiment *between countries* for a more precise understanding of worldwide emotions. In order to discover sentiments between countries, we collected tweets using Twitter API, calculated a sentiment score between the countries of each tweet, aggregated these scores, and visualized the scores intuitively on the world map.

## Implementation

Our web system, which we call *SentiWorld*, goes through eight steps, as shown in Figure 1. First, we selected the top 100 most populous countries based on the populations of 2014 year reported in the dataset of World Bank<sup>1</sup>. Next, we collected tweets written in English using the common English name of each country as the search term for Twitter

API. We collected 24,938,193 tweets between November 21 and December 15, 2015.

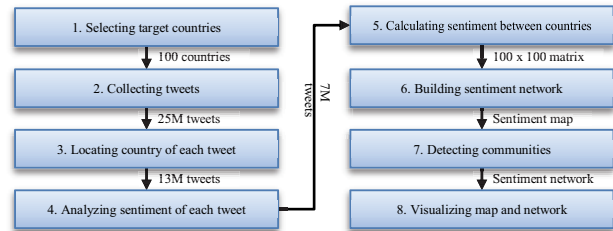


Figure 1: Overall procedure in *SentiWorld*.

In Step 3, to identify the country where each tweet was created, we looked into the geotag that is specified at ["place"]["country\_code"] in the metadata of a tweet. However, only about 1.6 percent of users have this function turned on (Leetaru et al. 2013). If it was not specified, we checked the ["user"]["location"] in the metadata of users' profiles. Nonetheless, because it is editable and not standardized, identifying the country of the user is quite tricky. To address the problem, we checked whether it matched the English official name, English common name, native official name, native common name, or ISO 3166-1 alpha-3 code for the country, all of which come from GitHub<sup>2</sup> and GeoNames<sup>3</sup>. If a match had not yet been achieved, we checked whether it matched a city name, using similar information to that provided above. We kept only the tweets for which we were able to figure out the origin. Thus, every tweet now specifies which country refers to which country—thus including a pair of countries.

In Step 4, to analyze the sentiment of a tweet, we adopted the LabMT<sup>4</sup> dataset, which consists of 10,222 English words and their sentiment scores. This corpus was extracted from Twitter, Google Books, music lyrics, and The New York Times and rated by the users of Amazon Mechanical Turk. 50 independent evaluations were done for every given word, each of which has a score from 1 to 9. The higher a score is, the happier a word is. Then, the score of each tweet is defined as the average of LabMT scores for the words appearing in the tweet.

\* Jae-Gil Lee is the corresponding author.

<sup>1</sup> <http://data.okfn.org/data/core/population#data>

<sup>2</sup> <https://github.com/mledoze/countries>

<sup>3</sup> <http://download.geonames.org/export/dump/>

<sup>4</sup> <http://trinker.github.io/qdapDictionaries/labMT.html>

In Step 5, to calculate the sentiment score between countries, we used the country identified in Step 3 as the source country and the query country from Step 2 as the destination country. Then, the country-wise score is defined as the average of the scores of relevant tweets. We discovered that the number of tweets between countries follows the power law distribution<sup>5</sup>. In addition, we provided a word cloud wherein the word sizes were proportional to the term frequency-inverse document frequency (TF-IDF).

In Step 6, we built a directed and weighted sentiment network, which is composed of countries as nodes and of the sentiment score as an edge weight. In Step 7, we adopted the *Louvain* and *Infomap* methods to detect communities. In order to apply Louvain, we transformed it into an undirected and unweighted network by keeping an edge between two countries if both directional sentiment scores were built from more than 100 tweets and if the difference between them was smaller than 0.3. Prior to running Infomap, in order to get a similar number of communities, we removed the edges built from less than 2500 tweets. Those network analysis results are visualized using D3.js<sup>6</sup>.

### System Demonstration

Our web system is accessible via <http://sentiworld.to.fr> using Firefox or Chrome browsers.

(1) As shown in Figure 2, the *sentiment map with chord diagrams* shows the results of Step 6 in a dynamic way. It includes *in-coming* and *out-going* maps in which the color of an arc represents the sentiment score and the width of an arc's line shows the number of tweets between countries. The in-coming map shows how other countries think about a given country in *terms of sentiment scores*, and the out-going map shows how a given country thinks about other countries. In addition, the actual score and count show up in a tooltip box when the mouse cursor rolls over the point of a country. In this way, users can estimate the reliability of the sentiment score by the count.

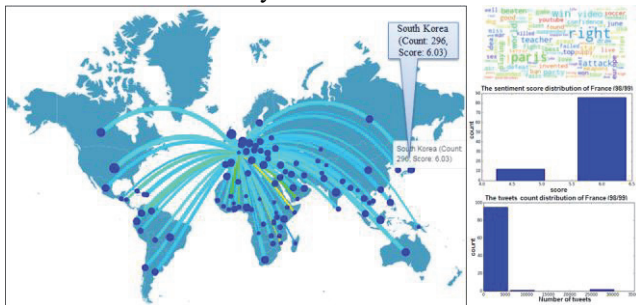


Figure 2: In-coming sentiment map for France.

(2) As shown in Figure 3, the *sentiment network of communities* shows the communities detected by the Louvain and Infomap methods in Step 7. The colors of the nodes and intra-community edges indicate the communities to which the countries belong. In order to show each community clearly, we removed inter-community edges, owing to the high density of the sentiments network. A communi-

ty in Infomap mainly represents a center country and the other countries that the center country thinks about similarly; a community in Louvain mainly represents a group of countries that think similarly to each other.

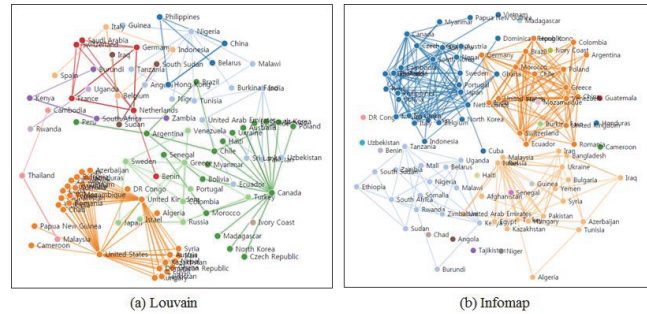


Figure 3: Sentiment network of communities.

### Discussion and Conclusion

France has “Paris,” “right,” and “attack” as big words in the word cloud of its in-coming sentiment map. The big word “right” appeared mostly from the developed countries that participated in the United Nations Climate Change Conference (UNCCC) held in Paris, whereas the big word “attack” came mostly from the Arabic countries which have continued to focus on the terrorist attack in Paris rather than the UNCCC. In order to drill down for the issues between countries, we compared the in-coming sentiments for France and other developed countries. Suppose that  $G1 = \{\text{France}\}$ ,  $G2 = \{\text{USA, United Kingdom, Germany, Netherland}\}$ ,  $G3 = \{\text{Iran, Iraq, Syria, Yemen}\}$ , and (A, B) denotes the sentiment score from A to B. It turned out that  $(G2, G1) - (G3, G1) = 0.50$  and  $(G2, G2) - (G3, G2) = 0.24$ . This observation implies that developed countries ( $G2$ ) and Arabic countries ( $G3$ ) have greater variances in the way they think about France than in their thoughts about other countries (e.g., Germany), possibly owing to the terrorist attack and the UNCCC in Paris.

As the future work, we plan to support time-evolving analysis and variable length of temporal windows.

### Acknowledgements

This research was partly supported by the maximum utilization of knowledge about herbal resource research program funded by the KIOM (Korea Institute of Oriental Medicine) (K16404) and the MOLIT (The Ministry of Land, Infrastructure and Transport), Korea, under the national spatial information research program supervised by the KAIA (Korea Agency for Infrastructure Technology Advancement) (15NSIP-B081011-02).

### References

- Elfenbein, H. A. and Ambady, N. 2003. Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science*, 12(5): 159–164.
- Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., and Shook, E. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, 18(5).
- Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., and Danforth, C. M. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5): e64417.

<sup>5</sup> [http://sentiworld.to.fr/statistical\\_analysis.html](http://sentiworld.to.fr/statistical_analysis.html)

<sup>6</sup> <http://d3js.org/>