The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

# Emoji-Word Network Analysis: Sentiments and Semantics

**S. M. Mahdi Seyednezhad,[‡] Halley Fede,[⋆] Isaiah Herrera,[†] Ronaldo Menezes[‡]**
[‡]BioComplex Lab, School of Computing, Florida Institute of Technology, Melbourne, USA
[⋆]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, USA
[†]Department of Mathematics, Westminster College, Lake City, USA
sseyednezhad2013@my.fit.edu, fedeh@rpi.edu, irh0612@westminstercollege.edu, rmenezes@cs.fit.edu

## Abstract

Emojis are very popular among online users. People use them in short texts they send or post because they are useful in conveying users' feelings. Despite their wide use, we still do not understand the patterns in their usage. In this paper, we present an important first step towards this understanding by using a modeling based on Network Science. We create a co-occurrence bipartite emoji-word network from 6 collections of tweets; each from a different topic of conversation. We then present results regarding the sentiment of an emoji as well as its semantics; its connection to words that define its meaning. Our results show that emojis are generally used in a positive sentiment but the semantics differ depending on the subject of the conversation.

## Introduction

The rapid growth of the online social media is probably unparalleled; never a social phenomena grew at the global scale as rapidly as social media, grabbing the attention of the population at all demographic environments. Such growth comes coupled with a need for better ways of expressing our feelings. With that in mind, several pictographs were introduced in 1999 by Japaneses telecommunication companies and later adopted by major tech companies. We know them now as *Emojis*.

Emojis today are standardized by the Unicode consortium (Consortium 2017b). They categorized emoji (version 5) into 8 major categories (Consortium 2017a) (shown in Table 1). The standardization boosted the use of Emojis because the pictographs can work at different mobile operating systems. To date, more than half of users on Instagram use emojis and messages with emojis attract 17% higher interaction (Gottke 2017). Emojis became so popular they constitute a system of symbols that have similarities to languages (Evans 2017).

Studies on emojis may unveil latent patterns of social media usage. As a result, we may be able to characterize users or datasets based on the emoji patterns or even recommend emojis to users. For instance, a social information-based recommender system may group users with similar similar emoji patterns, and use that information for collaborative filtering in which recommender systems first find similar users, then recommend items based on such similarity (Bobadilla et al. 2013).

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Table 1: Categories of emojis with samples and some sub-categories.

| Categories | Samples | Some sub-categories | # of emojis |
|---|---|---|---|
| Smiley & People | 😌😂😀 | face-positive, face-neutral, face-negative | 1507 |
| Animals & Nature | 🐵🦊🌵 | animal-mammals, animal-birds, plant-other | 113 |
| Food & Drink | 🍇🍷🍟 | food-fruit, drink, food-vegetable | 102 |
| Travel & Places | 🌍✈🚂 | place-map, transportation-ground, trans.-air | 207 |
| Activities | 🎃🚵⚽ | event, sport, game | 60 |
| Objects | 📽🏛🔋 | sound, phone, money | 222 |
| Symbols | ⛔🚙🔄 | transport-sign, arrow, warning | 427 |
| Flags | 🇺🇸🇧🇷🇮🇷 | flag, country-flag | 267 |

The information from emoji usage can be more accurate if we consider the words accompanying the emojis. In this paper, we generate an emoji-word bipartite network for different tweet collections from different subjects and look at usage patterns in these networks.

## Related Work

There are a number of attempts to understand patterns of emoji usage. For instance, Rodrigues et al. (Rodrigues et al. 2017) calculated the average aspects of each emoji and emoticon (e.g. aesthetic appeal, familiarity). They created a dataset of those emojis containing the average rating for 7 aspects. They show that emojis have similar aspects in Android and iOS. Surprisingly, Tigwell et al. (Tigwell and Flatla 2016) ran a similar study to understand user interpretations towards emojis as a function of the operating system. They mapped emojis according to their energy (high/low) and emotion (positive/negative). They concluded that the user interpretations are may differ depending on the operating system.

Jeager et al. (Jaeger and Ares 2017) tried to find how users interpret the meaning of 33 facial emojis. They found that emojis with similar facial expressions had considerably similar meanings. They concluded that for most emojis, consumers' interpretations corresponded to the meanings listed in the Internet resources. In contrast to our methodology in

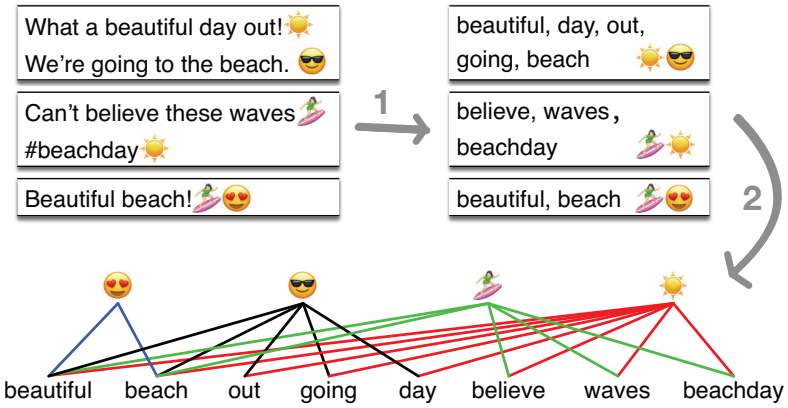


Figure 1: (1) Tweets are converted into a series of words and emojis. (2) from the converted tweets we generate a bipartite network linking words and emojis.

this paper, their approach do not extract the information automatically. They had participants evaluate a very limited number of emojis making their approach hard to be generalized and more susceptible to opinion of the chosen participants.

One of the first attempts to analyze the emojis used on social media was done by Novak et al. (Novak et al. 2015). They investigated the sentiment of the emojis and tried to find a meaningful correlation between the sentiment of emojis and their position and frequency in texts. Their results showed that most of the emojis are used towards the end of the text and most frequently associated with positive sentiments. They suggested that the rank of emoji sentiments could be considered as a resource for automated sentiment analysis. In this paper, we probe sentiments of emojis and their categories in a topic-based approach.

Emoji co-occurrence in social media messages was studied by Seyednezhad et al. (Seyednezhad and Menezes 2017). They build a co-occurrence network of emojis from two Twitter datasets. They found that emojis are not used randomly, but in fact, the edge-weight distribution follows a truncated power law in both datasets. Further Emoji network analyses were done by Fede et al. (Fede et al. 2017) on directed weighted network of emojis. They realized that the category-based entropy of communities of emojis reflects the users' will to use emojis from different categories (as defined in Table 1). In this paper we also use networks but we consider the words that accompany the emojis.

## Datasets and Methods

The data for this work comes from tweets collected about different subjects at different time periods. The reason to use this type of data is that we can cover a wider range of data and find potential differences due to the variety of subjects. Table 2 shows the datasets used.

We first extracted emojis and words from tweets containing at least one emoji. Then, we use the extracted information to create a *bipartite network* for each dataset linking words and emojis (see Figure 1). We used the networks to analyze the sentiment, semantic, and position of emojis.

Table 2: Six subject-based datasets from Twitter.

| Dataset | Characteristics | # tweets (millions) | % containing emojis | Collection period |
|---|---|---|---|---|
| *G20* | Surnames of G-20 countries' leaders | 10.6 | 7% | Aug. 24 - Sep. 24, 2014 |
| *Organ* | Organ transplantation terms | 2.5 | 9% | Oct. 2015 - Apr. 2017 |
| *WWC* | Women's World Cup 2015 | 10.7 | 1% | Jun. 06 - Jul. 05, 2015 |
| *rioSports* | Sports in the 2016 Rio Olympics | 1.8 | 1% | Aug. 05 – Aug. 21, 2016 |
| *rioTerms* | "Olympics" in different | 5.8 | 1% | Aug. 05 – Aug. 21, 2016 |
| *randSample* | 2 months samples from Twitter | 168.5 | $< 1\%$ | Dec. 13, 2016 - Jan. 31, 2017 |

In this paper we define the semantic of an emoji based on how many times we see that emoji appearing with words in its definition (according to the Unicode consortium). An emoji has high semantic meaning if it tends to appear a lot with words in its definition. We compute the semantics of each category (as in Table 1). The semantic meaning of a category is the combination of the semantic of each emoji in the category.

## Experimental Results

In our experiments we use the datasets in Table 2 which includes a random sample used as a ground truth (null model).

Our first experiment consists of extracting the sentiment of emojis using not only the sentiment of the tweets, but also the sentiment of the words accompanying each emoji. Then, we use TextBlob (Loria et al. 2014) library in Python to analyze sentiment of the tweets. Table 3 shows the portion of sentiments in each dataset. We also calculate the entropy of languages in each dataset ($d_i$) to investigate the impact of

Table 3: Sentiment analysis for each dataset.

| Dataset | % of neutral | % of positive | Positive to negative ratio | # of languages | Language entropy |
|---|---|---|---|---|---|
| *G20* | 48% | 40% | 4.0 | 47 | 1.18 |
| *Organ* | 46% | 45% | 5.9 | 47 | 0.52 |
| *WWC* | 73% | 22% | 5.9 | 43 | 2.13 |
| *rioSports* | 83% | 14% | 4.9 | 48 | 2.91 |
| *rioTerms* | 74% | 20% | 4.1 | 44 | 1.59 |
| *randSample* | 80% | 15% | 3.3 | 59 | 3.20 |

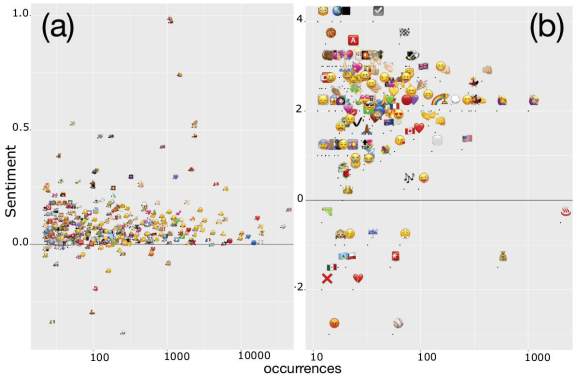


Figure 2: Tweet- and word-based sentiment of emojis versus their frequency of usage (occurrence) in the *WWC* dataset.

language variety on our results as defined in Equation 1.

$$\mathcal{S}(d_i) = -\sum_{\ell} p_{i\ell} log(p_{i\ell}) \quad (1)$$

where $p_{i\ell}$ is the probability of observing language $\ell$ in dataset $d_i$.

We find the sentiment of all tweets containing each emoji and found that the average sentiment for most of emojis is positive (Figure 2(a)). Similarly, we calculate the average sentiment of emojis based on the words that accompany them and confirmed that most of the emojis are used with words with positive sentiment (Figure 2(b)). This is true for all datasets.

In order to get a better insight about the sentiment of emojis, we calculated the average sentiment of the categories of emojis as well shown in Figure 3. The noticeable phenomenon is that we we see negative average sentiment only for the category *Flag* only in *WWC* dataset. The average sentiment distribution is almost uniform and it is not noticeably biased towards a certain category.

We try to study the semantics of emojis based on the words in their definition (as explained earlier). Figure 4 shows the average semantics of each category for all datasets. Unlike the uniformity in sentiment, shown in Figure 3, the semantic distributions is favored in a few categories. For example, in the *rioSports* and *rioTerms* datasets, *Flag* emojis are used a considerable number of times with the words in their definition. For instance, "Brazil" may be used with 🇧🇷 several times. In all datasets, the highest semantics is for *Flag* and *Activities*. Apparently, no matter what the subject of the tweet is, the users tend to use *Flag* emojis with the name of country, and *Activity* emojis with the name of the activity. As a final step in these analyses, we correlated semantic and sentiment and found no significant correlation. [1]

Recall that Novak et al. (Novak et al. 2015) discussed the fact that emojis tend to appear at the end of the text. Given that we are dealing with sentiment and semantics we decided

[1]The results are omitted here due to space restrictions. An interested reader may contact one of the authors if interested in these experiments.

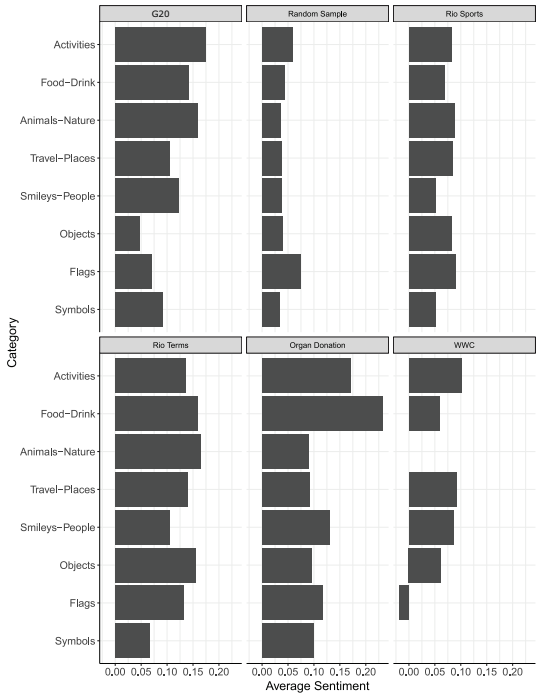


Figure 3: Average sentiment of categories based on sentiment of tweets.

to look for the sentiment as a function of position. In other words, does the position that an emoji appears relate to the sentiment of the tweet? Figure 5 shows the probability of an emoji appearing in a particular position (ECDF) for tweets with positive and negative sentiment. In general, emojis tend to appear slight earlier when the sentiment of the tweet is positive, except for the *organDonation* dataset.

## Conclusion

We extracted a bipartite emoji-word network from 6 different Twitter datasets. We showed that emojis are most likely used in a tweet with positive sentiments. This can help social media-based recommender system to consider the sentiment of users when suggesting emojis.

Semantics of emojis is considered as the percent of edges connect emojis to the words from their definition. The distribution of semantics of emojis is different from the sentiment one and it is not uniform through categories. This information can help us tune a typical emoji predictor.

Lastly, we looked at the positional analysis of the emojis in the tweets and found that although emojis tend to be placed at the end of the tweet, emojis tend to appear slightly earlier in positive tweets. This result is a first step in the understanding of regularities regarding where emojis are used as a function of the context (here the context is the sentiment of the tweet)

## Acknowledgments

We would like to thank the NSF grant No. 1560345 for partially supporting this research. We also thank Diogo

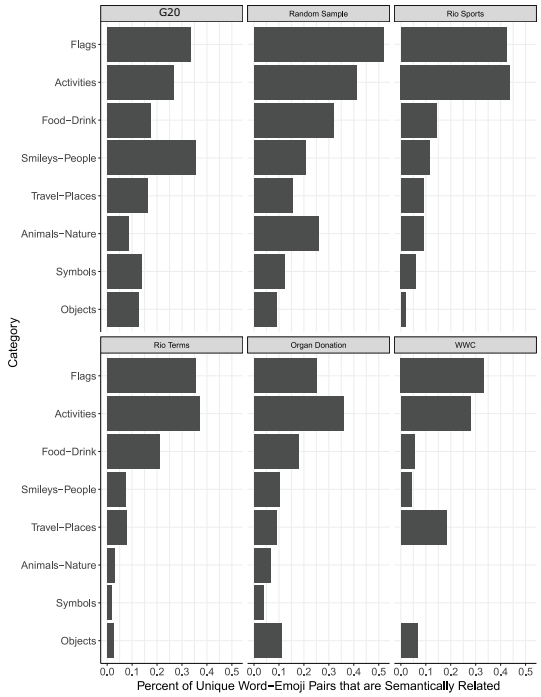


Figure 4: Semantic meaning of each emoji category.

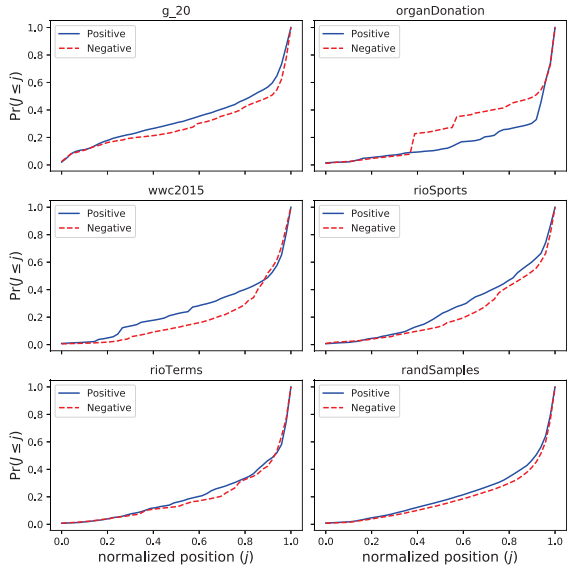


Figure 5: Empirical Cumulative Distribution Function (ECDF) of the position of the emojis in the text. The position is normalized with respect to the length of the tweets. The sharp increase at the end of all diagrams suggests most of the emojis are used in the end of the tweets with both negative and positive sentiment. The *organDonation* dataset is the only one in which emojis tend to appear earlier when the sentiment is negative.

Pacheco, Josemar F. da Cruz, and Diego Pinheiro for providing some of the datasets for this research.

## References

Bobadilla, J.; Ortega, F.; Hernando, A.; and Gutiérrez, A. 2013. Recommender systems survey. *Knowledge-based systems* 46:109–132.

Consortium, U. 2017a. Emoji ordering, v5.0.

Consortium, U. 2017b. Unicode emoji.

Evans, V. 2017. *The Emoji Code: The Linguistics Behind Smiley Faces and Scaredy Cats*. Picador USA.

Fede, H.; Herrera, I.; Seyednezhad, S. M.; and Menezes, R. 2017. Representing emoji usage using directed networks: A twitter case study. In *Proceedings of the the 6th International Conference on Complex Networks and Their Applications*, in press. Springer.

Gottke, J. 2017. Instagram emoji study – emojis lead to higher interactions.

Jaeger, S. R., and Ares, G. 2017. Dominant meanings of facial emoji: insights from chinese consumers and comparison with meanings from internet resources. *Food Quality and Preference*.

Loria, S.; Keen, P.; Honnibal, M.; Yankovsky, R.; Karesh, D.; Dempsey, E.; et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: Simplified Text Processing*.

Novak, P. K.; Smailović, J.; Sluban, B.; and Mozetič, I. 2015. Sentiment of emojis. *PloS one* 10(12):e0144296.

Rodrigues, D.; Prada, M.; Gaspar, R.; Garrido, M. V.; and Lopes, D. 2017. Lisbon emoji and emoticon database (leed): Norms for emoji and emoticons in seven evaluative dimensions. *Behavior Research Methods* 1–14.

Seyednezhad, S. M., and Menezes, R. 2017. Understanding subject-based emoji usage using network science. In *Workshop on Complex Networks CompleNet*, 151–159. Springer.

Tigwell, G. W., and Flatla, D. R. 2016. Oh that's what you meant!: reducing emoji misunderstanding. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, 859–866. ACM.