Entropy in Network Community as an Indicator of Language Structure in Emoji Usage: A Twitter Study Across Various Thematic Datasets

Ryan Hartman[1], S.M. Mahdi Seyednezhad[1], Diego Pinheiro[2], Josemar Faustino[1], and Ronaldo Menezes[3]

[1] Department of Computer Engineering and Sciences
Florida Institute of Technology, USA
rhartman2014@my.fit.edu
sseyednezhad2013@my.fit.edu
cruzj2012@my.fit.edu

[2] Department of Internal Medicine
University of California, Davis, USA
pinsilva@ucdavis.edu

[3] BioComplex Laboratory
Department of Computer Science
University of Exeter, UK
r.menezes@exeter.ac.uk

Abstract. Emojis are emerging as an alternative way to interact and communicate online, and their large-scale adoption has the potential to reveal distinct patterns of human communication and social interactions. In this work, we investigate the hypothesis that emojis are a kind of language. By building networks of emoji co-occurrence, we examine the diversity of the community structure of such networks with regards to predefined categories of emojis. Using four different techniques of community detection, we validate our hypothesis on six Twitter datasets: five from specific topics and one random dataset. Our results demonstrate that the community structure of emojis is more diverse when they are used in non-random topics such as politics and sports, and that Stochastic Block Models appears to extract communities with higher diversity.

1 Introduction

Online social networks have attracted a significant number of users and have rapidly become people's main form of communication. In these social networks, users communicate their feelings and emotions mainly using short pieces of text. The message size limitation, either imposed by the platform or bolstered by the need to say more and type less, can be challenging to convey the right idea and can create misunderstandings. The need to convey meaning succinctly led to the usage of pictographical representations of emotions and ideas in the form of *emoticons* (e.g., "`;-)`", "`;p`"). The large scale adoption of *emoticons* led to the emergence of modern Emojis, which are small images used with the intent of expressing emotions on ideas within text. Nowadays, emoji usage is

widespread, and they can be found in many instant messaging apps as well as in social media sites such as Twitter, Instagram, Facebook, and many others.

Currently, more than half of the posts on Instagram contain emojis, and they attract a 17% higher interaction rate when compared to messages which do not include emojis [7]. In 2015, emojis were announced as the fastest growing communication method in the United Kingdom [2], and given the ever-increasing usage of emojis in social media, the assessment of emojis as a form of language deserves further investigation. However, this assessment cannot be made with classical linguistic techniques [21], as emojis inherently lack the features found in structured languages. Instead, the relationship of meaning and diversity of characters can be explored to asses the semantic patterns of the emoji usage phenomena.

One common way to analyze languages and their similarities is to probe the entropy of the words ordering in different families of the languages [14,12,9]. Furthermore, word maximum entropy approach is a reliable tool for linguistic disambiguation or part of speech text tagging [19,11]. In this paper, we consider emojis as words of a hypothetical language in which the entropy of the words will give us the information on how diverse emojis are used.

In this work, we analyze emojis by investigating its semantics through the sequences formed by their co-occurrence in social media posts. In order to communicate and express their feelings, users may use a more diverse set of emojis such that emojis of different categories will appear together with a higher likelihood. This diversity analysis is possible because emojis are originally grouped into seven categories, *Smiley-People, Animals-Nature, Food-drink, Activities, Objects, Symbols, and Flags*; the categories are used to gauge diversity. By building networks of co-occurring emojis, we can unveil their community structure and subsequently assess the diversity of their community structure with respect to categories of emojis to quantify the semantics. In this work, we use the assessment of communities proposed by Hartman *et al.* [8] to examine the richness of emoji semantics through the diversity of communities from six Twitter datasets collected based on different topics (as described in Table 1) using four techniques of community detection (as described in Table 2). By analyzing the structural properties of such communities as well as their resemblance to available metadata, our work sheds light on the idea that emojis are a form of language.

The contribution of this paper stems from the use of our previously proposed assessment tool for communities detection (in the sense of their ability to capture organizations known *a priori*) [8] to argue that emojis are being used in a way that resembles language structure (entropy) [12].

This paper starts with a description of related works in Section 2. We follow in Section 3 with the description of the data used in this paper, how we create emoji directed co-occurrence networks, and the community detection algorithms we use in this work. In Section 4 we present our main results, concluding the paper with some final considerations in Section 5.

2 Related Work

Recent studies on emojis can be divided into two major approaches. In the first approach, researchers aim to understand the meaning of emojis. Barbieri *et al.* [1] investigated the meaning of Twitter emojis by examining the likelihood of the pairwise appearance and measuring how often emojis convey the same meaning. Novak *et al.* [15] drew a sentiment map of the 751 most frequently used emojis and found high frequency of usage associated with positive tweets. Wijeratne *et al.* [20] created a dictionary to make a machine readable sense inventory for emoji. In order to create octuples representing the meaning of the emoji, they used the Unicode, description, image, and keywords attached to the meaning of the emoji.

The second approach attempts to analyze the collective behavior of users based on emoji usage. Novak *et al.* [15] found that the inter-annotator agreement of tweets containing emojis were higher than the ones without emojis. More interestingly, they acknowledged that users normally use emojis at the end of tweets, and the rank of emojis did not change between different languages. Seyednezhad *et al.* [18] extracted a network of emojis based on their co-occurrence in tweets from two different datasets. They stated the emoji with the maximum edge betweenness could give us a hint about the underlying subject in which the tweets were collected. This work was generalized by Fede *et al.* [4] by experimenting with more datasets which contained directed networks. They concluded that important emojis are topic dependent. Lu *et al.* [13] created a network of emojis by point-wise mutual information (PMI). Their findings pointed to a strong correlation between social indicators and patterns of emoji usage.

Using networks of emojis, we can extract the structure of related emojis using community detection techniques. Community detection techniques aim to identify the building blocks of networks and their structural properties. It has been applied to networks of protein interaction, food web, genetic disorders, gene expression, and social networks [5]. However, the most efficient techniques for exploring communities may yield different results [10,8]. Hence, recent works have used community detection techniques in a holistic approach [8], which includes a comparative analysis of multiple techniques as well as the the resemblance of extracted communities to available metadata. This is the direction we pursue in this work.

3 Data and Methods

3.1 Data Collection and Curation

For this study, we collected tweets from Twitter based on different topics at different time periods. The goal is to cover a diverse set of topics, allowing us to examine the effect of such diversity on communities, extracted by state-of-the-art community detection techniques. Moreover, we add to the analysis a topic-free dataset. This data contains tweets randomly sampled from the Twitter feed, without the use of tracking keywords. The random data allows us to observe any possible bias due to using topic-based data. Table 1 shows further information about the datasets used in this work.

In order to show the network statistics are correlated with the emojis' frequency of usage, we calculated the Spearman's rank correlation between the frequency and

weighted degree of the nodes. The highly correlated ranks suggest network characteristics such as weighted degree can explain the frequency of emoji usage.

Table 1. Six datasets collected from Twitter. The topics of the datasets covers several areas of interest. The Spearman's rank correlation is calculated for each dataset between the frequency and weight-degree of the nodes.

Label	Dataset	Characteristics	# Tweets (Millions)	% Containing emojis	Collection period	Spearman's rank correlation
D_1	*G-20*	Surnames of G-20 countries' leaders	10.6	7%	Aug. 24 - Sep. 24, 2014	0.94
D_2	*Organ*	Organ transplantation terms	2.5	9%	Oct. 2015 - Apr. 2017	0.85
D_3	*rioSports*	Sports in the 2016 Rio Olympics	1.8	1%	Aug. 05 – Aug. 21, 2016	0.95
D_4	*rioTerms*	"Olympics" in different	5.8	1%	Aug. 05 – Aug. 21, 2016	0.92
D_5	*WWC*	Women's World Cup 2015	10.7	1%	Jun. 06 - Jul. 05, 2015	0.91
D_6	***randSample***	**2 months samples from Twitter**	168.5	$< 1\%$	Dec. 13, 2016 - Jan. 31, 2017	0.97

In summary, we have data related to politics (D_1), health (D_2), sports (D_3, D_4, and D_5), as well as a random collection of tweets (D_6). The random sample D_6 has the lowest percentage of tweets containing emojis, while the organ transplantation D_2 collection has the greatest amount.

3.2 Network Construction

The main focus of this paper is on comparing prominent community detection techniques and the characteristics of the communities they uncover for a variety of datasets. Differing from previous works [18], here we consider that the order of emojis appearing in a tweet is fundamental and hence better represented using directed links.

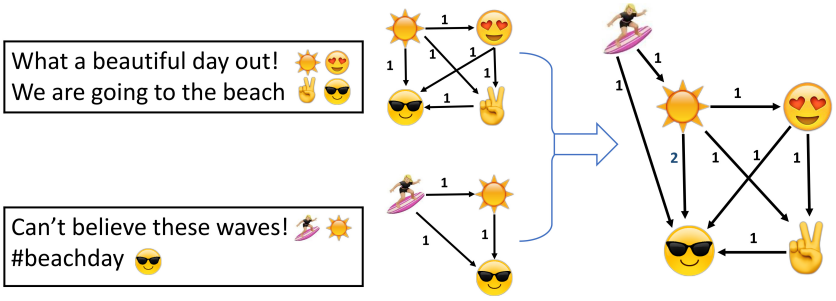

Fig. 1. Directed network of emojis. We create a connection from emoji to emoji in the order they appear in a tweet. This process is repeated for every tweet in the dataset. Then we accumulate all the sub-networks extracted from tweets into a main directed network of emojis.

A directed network of emojis gives us an opportunity to study the collective usage of emojis on social media. Additionally, different sequences of emojis may reflect different feelings expressed by users. For example, someone tweeting "I loved ❤️ this place until

that horrible 😔 incident happened", the meaning is different from another tweet such as "This place is sometimes horrible 😔, but I love ❤️ it anyway!". Note that the order of the emojis is related to the sentiment being expressed. In order to build the directed network, we connect each emoji to every subsequent one appearing in the same tweet. Figure 1 shows the process of making directed weighted links between emojis.

3.3 Community Detection Techniques and Evaluation Criteria

Since the emoji co-occurrence networks are weighted and directed, we used state-of-the-art techniques that support networks with these features [5]. Table 2 describes the selected techniques and their respective approach to identifying communities. For each emoji network that was constructed, all techniques are applied and the characteristics of the communities found are then analyzed.

Table 2. Community Detection Techniques

Acronym	Name	Approach	Description
$\mathscr{IM}$	InfoMap	Bottom-up	Builds a map of information flow in the network using a random walk. Finding a community is equivalent to minimizing the flow representation by applying a compression technique [17].
$\mathscr{BM}$	Stochastic Block Models	Top-down	Applies maximum likelihood estimation to infer the latent block division in the empirical network. Such inference is equivalent to the entropy minimization of the network ensemble [16].
$\mathscr{LP}$	Label Propagation	Hybrid	Based on belief propagation, where each node spreads its label to its neighbors. Convergence of labels uncovers the community structure [6].
$\mathscr{LM}$	Louvain modularity	Bottom-up	Works by optimizing network modularity, which is the tightness of node connectivity into modules/communities in the empirical network relative to a null model [3].

To begin, we examined the size characteristics of the communities found by these four techniques. Then, we apply an unsupervised evaluation by computing the communities' conductance. The conductance C of a community k, measures the ratio between the intragroup and intergroup connectivity of the communities [5] and is computed as shown in Equation 1.

$$C(k) = \frac{\sum_{i \in k, j \notin k} w_{ij}}{\sum_{i \in k, j} w_{ij}} \quad , \tag{1}$$

where w_{ij} is the weight of the link connecting nodes i and j. In this sense, well-structured communities exhibit a higher volume of edges between nodes within a community compared to edges going to the outside of the community.

We also conduct a supervised evaluation using the idea of rank stability for community detection [8] which was proposed as a way to measure the homogeneity of a particular community using the attribute values of nodes within that community as follows:

$$E(n) = -\sum_{t=1}^{L} p_{kt}(n) \log_2[p_{kt}(n)] \quad , \tag{2}$$

where $p_{kt}(n)$ is the proportion of nodes in community k that are associated with attribute t in their rank n, and L is the number of emoji categories. In this work, we only have one attribute ($n = 1$) which can be one of the six categories of emojis.

4 Results

The statistical and structural properties of extracted communities can vary depending on the community detection technique. For instance, the number of communities extracted on large-scale networks significantly vary depending on the community detection technique [8]. We organized our results in three parts. First, we characterize two structural properties of major importance, namely, community size and conductance. Then, we characterize the communities of emoji networks with regards to emoji categories known a priori to shed light in the context of emojis as language. Lastly, we present how these macroscopic characteristics are related and how a careful exploration of such relationship has the potential to help us gain insights on the structure and function of emojis.

Overall, emoji networks exhibit a well-defined community structure with regards to their size which is slightly shifted depending on the dataset (Fig. 2, left). Exceptionally, $\mathscr{LP}$ appears to find communities with typically greater size, it identifies the least number of communities, whereas other techniques find ten times more communities; this result is consistent with previous work [8]. Although the distribution of nodes within communities is an important aspect when identifying groups of interrelated emojis, we also need to quantify the extent in which extracted communities exhibit desirable structural properties.

Despite the lack of a general definition of a community, the number of links running between nodes within the community (i.e., internal edges) should be larger than the number of links running from nodes within the community to nodes outside the community (i.e., external edges). Conductance extends such definition for weighted networks (Equation 1). The conductance of communities vary depending on the technique and dataset (Fig. 2, middle). $\mathscr{IM}$ and $\mathscr{LM}$ show a more similar conductance distribution when compared to $\mathscr{BM}$. Precisely, $\mathscr{BM}$ has a higher likelihood of identifying communities with greater typical conductance. Similarly, $\mathscr{IM}$ and $\mathscr{LM}$ are likely to identify communities with moderate values of typical conductance. Lastly, $\mathscr{LP}$ is likely to identify communities with lower conductance.

Besides analyzing the structural properties of communities, we are also interested on evaluating communities to gain insight on the usage of emojis as a language. Here, we apply an entropy based metric to explore the levels of meaning in emoji usage. Our assumptions are that the higher the diversity of emoji categories within a community, the higher the level of meaning conveyed by these emojis.

We carry out this analysis by assessing the extent in which emojis within a community resemble the official emoji categories, using the rank entropy of communities [8]. The rank entropy (Equation 2) varies depending on the employed technique and underlying dataset (Fig. 2, right). Overall, the highest rank entropies are exhibited by communities extracted using $\mathscr{BM}$ as well as communities extracted from dataset D_1. Conversely, the lowest entropies are exhibited by communities extracted using $\mathscr{LP}$

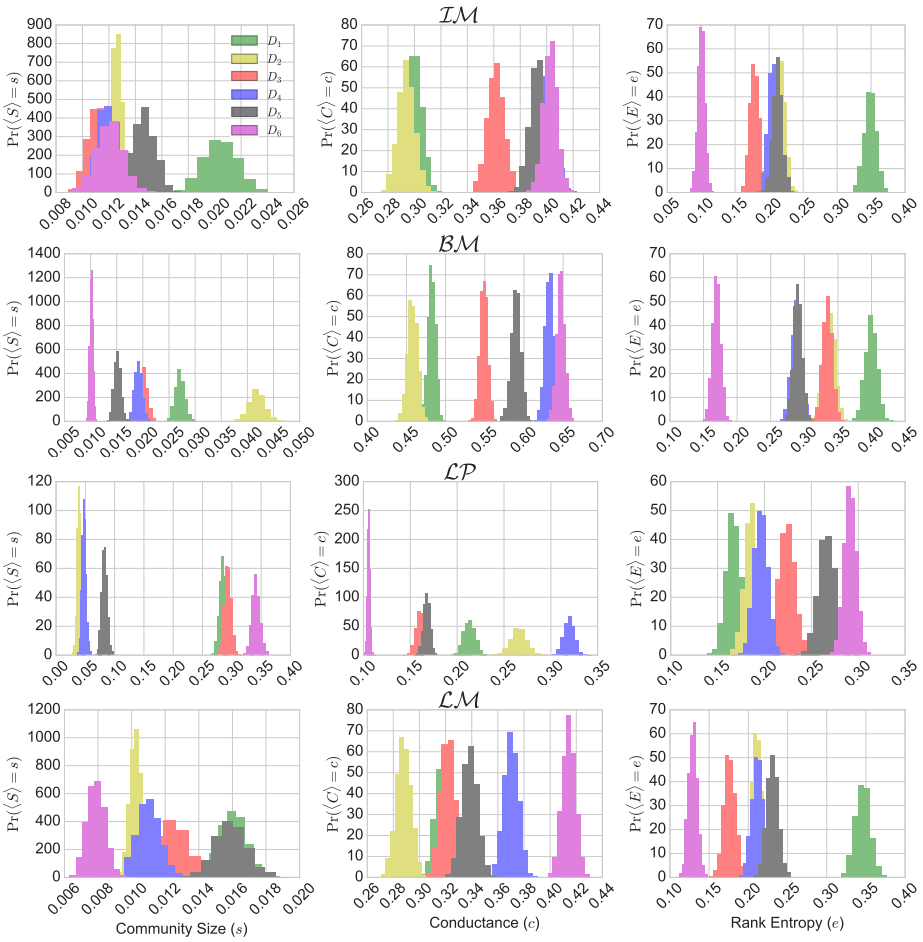

Fig. 2. Community size (S, left), conductance (C, middle), and rank entropy (E, right) of emoji networks characterized across six datasets of different thematics $D_{1...6}$ using four community detection techniques, namely, infomap $\mathscr{IM}$, block models $\mathscr{BM}$, label propagation $\mathscr{LP}$, and Louvain modularity $\mathscr{LM}$.

as well as communities extracted from the random dataset D_6. $\mathscr{LP}$ presents some exceptions to aforementioned statements.

We can characterize communities of emojis by their structural properties, such as community size and conductance, as well as by their resemblance with available metadata using the rank entropy. Besides the independent characterization of community size, conductance, and rank entropy, we can also examine the relationship between these characteristics and unveil additional properties of the extracted communities. Indeed, these characteristics are related to each other (Table 3). Overall, communities with greater size tend to be moderately associated with lower conductance and higher rank

entropy; however, there is a lack of a significant association between conductance and rank entropy. After controlling for community size, conductance appears to be associated with rank entropy mainly depending on the underlying dataset. For instance, it is moderate in datasets D_1 and D_2, and it is absent in the random dataset D_6.

Table 3. Relationship between community size S, conductance C, and rank entropy E as measured by Pearson correlation. Community size S is negatively correlated with conductance C and positively correlated with rank entropy E. Even after controlling for community size, there is a lack of significant correlation between conductance and rank entropy in all techniques, except for label propagation.($\hat{\rho}_{SE \cdot S} = 0.12,\ p > .1$). However, further looking such correlation in each dataset across multiple techniques, conductance is positively correlated with rank entropy, which is strongest in D_1 and weakest in the random dataset D_6.

	$\mathscr{IM}$	$\mathscr{BM}$	$\mathscr{LP}$	$\mathscr{LM}$	$D1$	$D2$	$D3$	$D4$	$D5$	$D6$
$\hat{\rho}_{SC}$	−0.14	−0.12	−0.24	−0.05	−0.19	−0.14	−0.14	−0.17	−0.13	−0.19
$\hat{\rho}_{SE}$	0.41	0.37	0.56	0.46	0.36	0.31	0.36	0.35	0.34	0.31
$\hat{\rho}_{CE}$	−0.03	0.01	−0.03	−0.03	0.11	0.12	0.10	0.05	0.01	−0.05
$\hat{\rho}_{CE \cdot S}$	0.01	0.05	0.12	−0.02	0.20	0.18	0.16	0.12	0.06	0.01

5 Conclusion

Emojis are a *de facto* form of communication; their simplicity and ease of use were fundamental for their wide adoption. When we build emoji networks, we can exhibit the structural properties of its usage with significant detail. These networks show the function of emojis and how they relate to language. In this work, we investigated the hypothesis that emojis are a form of language by building networks of emojis co-occurring on social media posts and subsequently analyzing the diversity of their community structure. To gain insights on how emojis are used, we compare the diversity of communities from specific topics such as politics and sports with that of randomly collected dataset. We found that users tend to communicate on social media using emojis of different categories. In this sense, the Stochastic Block Model would be more suitable way of doing community analysis on these networks because it is capable of finding more diverse (i.e., higher rank entropy) and well-formed communities (i.e., higher conductance). Yet, other possibilities to build emoji networks remain to be explored such as those based on risk ratio, pointwise mutual information and Φ-correlation.

Finding suitable datasets is quite hard because emojis are used in mostly personal communications. The usage on Twitter allowed us to perform this work but would be interesting to investigate the network of emojis on other social media such as Instagram to verify if the language characteristics we found here are also present. We have attempted to use another dataset from Reddit but unfortunately emoji is not widely used on Reddit.

Acknowledgements

Diego Pinheiro and Josemar Faustino would like to thank the Science Without Borders program (CAPES, Brazil) for financial support under grants 0624/14-4 and 1043-14-5, respectively. This material is based upon work supported by the National Science Foundation under Grant No. CNS 09-23050.

References

1. Barbieri, F., Ronzano, F., Saggion, H.: What does this emoji mean? a vector space skip-gram model for twitter emojis. In: Language Resources and Evaluation conference, LREC, Portoroz, Slovenia (2016). URL http://hdl.handle.net/10230/33776
2. Doble, A.: Uk's fastest growing language is... emoji (2015). URL http://www.bbc.co.uk/newsbeat/article/32793732/uks-fastest-growing-language-is-emoji
3. Dugué, N., Perez, A.: Directed Louvain : maximizing modularity in directed networks. Tech. rep., Université d'Orléans (2015). URL https://hal.archives-ouvertes.fr/hal-01231784
4. Fede, H., Herrera, I., Seyednezhad, S.M., Menezes, R.: Representing emoji usage using directed networks: A twitter case study. In: International Workshop on Complex Networks and their Applications, pp. 829–842. Springer (2017). DOI 10.1007/978-3-319-72150-7_67. URL https://dx.doi.org/10.1007/978-3-319-72150-7_67
5. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3-5), 75–174 (2010). DOI 10.1016/j.physrep.2009.11.002. URL http://dx.doi.org/10.1016/j.physrep.2009.11.002
6. Gaiteri, C., Chen, M., Szymanski, B., Kuzmin, K., Xie, J., Lee, C., Blanche, T., Chaibub Neto, E., Huang, S.C., Grabowski, T., Madhyastha, T., Komashko, V.: Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering. Scientific Reports **5**(1), 16,361 (2015). DOI 10.1038/srep16361. URL http://dx.doi.org/10.1038/srep16361
7. Gottke, J.: Instagram emoji study: Emojis lead to higher interactions (2017). URL https://www.quintly.com/blog/2017/01/instagram-emoji-study-higher-interactions/
8. Hartman, R., Faustino, J., Pinheiro, D., Menezes, R.: Assessing the suitability of network community detection to available meta-data using rank stability. In: Proceedings of the International Conference on Web Intelligence - WI '17, pp. 162–169. ACM Press, New York, New York, USA (2017). DOI 10.1145/3106426.3106493
9. Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F.K., Papageorgiou, H.: Word-length entropies and correlations of natural language written texts. Journal of Quantitative Linguistics **22**(2), 101–118 (2015)
10. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. Physical Review E **80**(5), 056,117 (2009). DOI 10.1103/PhysRevE.80.056117. URL http://dx.doi.org/10.1103/PhysRevE.80.056117
11. Le-Hong, P., Roussanaly, A., Nguyen, T.M.H., Rossignol, M.: An empirical study of maximum entropy approach for part-of-speech tagging of vietnamese texts. In: Traitement Automatique des Langues Naturelles-TALN 2010, p. 12 (2010). URL https://hal.inria.fr/inria-00526139
12. Levitin, L.B., Reingold, Z.: Entropy of natural languages: Theory and experiment. Chaos, Solitons & Fractals **4**(5), 709–743 (1994). URL https://doi.org/10.1016/0960-0779(94)90079-5
13. Lu, X., Ai, W., Liu, X., Li, Q., Wang, N., Huang, G., Mei, Q.: Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 770–780. ACM (2016). URL http://dx.doi.org/10.1145/2971648.2971724

14. Montemurro, M.A., Zanette, D.H.: Universal entropy of word ordering across linguistic families. PLoS One **6**(5), e19,875 (2011)
15. Novak, P.K., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS one **10**(12), e0144,296 (2015)
16. Peixoto, T.P.: Hierarchical block structures and high-resolution model selection in large networks. Physical Review X **4**(1), 1–18 (2014). DOI 10.1103/PhysRevX.4.011047. URL http://dx.doi.org/10.1103/PhysRevX.4.011047
17. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118–1123 (2008). DOI 10.1073/pnas.0706851105. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0706851105
18. Seyednezhad, S.M., Menezes, R.: Understanding subject-based emoji usage using network science. In: Workshop on Complex Networks CompleNet, pp. 151–159. Springer (2017). URL https://dx.doi.org/10.1007/978-3-319-54241-6_13
19. Suárez, A., Palomar, M.: A maximum entropy-based word sense disambiguation system. In: Proceedings of the 19th international conference on Computational linguistics-Volume 1, pp. 1–7. Association for Computational Linguistics (2002). URL https://doi.org/10.3115/1072228.1072343
20. Wijeratne, S., Balasuriya, L., Sheth, A., Doran, D.: Emojinet: Building a machine readable sense inventory for emoji. In: International Conference on Social Informatics, pp. 527–541. Springer (2016). URL https://dx.doi.org/10.1007%2F978-3-319-47880-7_33
21. Winograd, T.: Understanding natural language. Cognitive Psychology **3**(1), 1 – 191 (1972). DOI https://doi.org/10.1016/0010-0285(72)90002-3. URL http://www.sciencedirect.com/science/article/pii/0010028572900023