

# Revealing Patterns of Twitter Emoji Usage in Barcelona and Madrid

Francesco BARBIERI <sup>a,1</sup>, Luis ESPINOSA-ANKE <sup>a,2</sup> and Horacio SAGGION <sup>a,3</sup>

<sup>a</sup> *Universitat Pompeu Fabra, Barcelona, Spain*

**Abstract.** Emojis are small sized images which are naturally combined with free text to visually complement or condense the meaning of a message. The set of available emojis is fixed, irrespective of a user’s location. However, their interpretation and the way they are used may vary. In this paper, we compare the meaning and usage of emojis across two Spanish cities: Barcelona and Madrid. Our results suggest that the overall semantics of the subset of emojis we studied is preserved over these cities. However, some of them are interpreted differently, which suggests there may exist cultural differences between inhabitants of Barcelona and Madrid, and that these are reflected in how they communicate in social networks.

**Keywords.** Emojis, Distributional Semantics, Barcelona, Madrid

## 1. Introduction

Today, we communicate with each other differently as we used to, mainly due to the advent of Social Media. Social networking platforms such as Twitter allow users to post short text messages to update followers and interested parties on current affairs, sentiments, emotions, or even for voicing opinions on matters of any kind. During the last few years, Twitter users have started to extensively use emojis in their posts. Emojis are pictures that can be naturally combined with plain text to create a new form of online language; a practice also adopted in other networking platforms such as Facebook, Whatsapp and Instagram. Emojis pose important challenges for researchers in multimedia information systems, since their meaning remains for the time being unexplored. Emojis meaning may change from language to language and culture to culture, in contrast to the common assumption that they represent the same meaning or idea irrespective of where or who uses them. Understanding the meaning of emojis with respect to their context of use is important in multimodal systems, human computer interaction and multimedia retrieval. In this paper we investigate the use of emojis from a natural language processing viewpoint. We adopt an empirical research methodology relying on current vector space representation modelling [17,13] to understand the “semantics” of these important elements of multimedia communication. More specifically, we investigate the use of emojis in the two largest Spanish cities, namely Madrid and Barcelona. To this end, we compiled

---

<sup>1</sup>Email: francesco.babieri@upf.edu

<sup>2</sup>Email: luis.espinosa@upf.edu

<sup>3</sup>Email: horacio.saggion@upf.edu

**Table 1.** 20 most frequent emojis used across the two cities studied.

Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Barcelona	❤️	😄	😂	🇪🇸	😍	👉	🎉	😎	🏀	👏	☀️	🌟	👏	❤️	💰	👑	👏	👏	👏	👏
Madrid	❤️	😄	😂	❤️	👉	🎉	👏	😍	😄	👏	🏀	👑	👏	👏	👏	👏	🇪🇸	🌟	👑	👏

a corpus of more than 1 million tweets sent from these two cities, and performed several experiments to compare emoji usage between them. We developed a comparison scheme that avoids the need to rely on language specific semantic space models to interpret the meaning of emojis, thus surmounting potential discrepancies in terms of the language used by inhabitants of Madrid and Barcelona. The case of the latter is particularly important because only leveraging the Spanish language, our model would not be sensitive to switching between Spanish and Catalan. Our results demonstrate that the semantics of the 150 most popular emojis is somehow preserved in both cities, but that for someone in Madrid, not all the emojis mean the same when compared with someone from Barcelona. For instance the emojis 🍷, 📷, and 🍷 seem to be used in different contexts across distinct languages, while there is a relative agreement on the use and meaning of 🇪🇸 and 🍷.

## 2. Related Work

Currently, emojis represent a widespread and pervasive global communication device largely adopted by almost any Social Media service and instant messaging platform [10,15,8]. Emojis (like the older emoticons) support the possibility to express diverse types of contents in a visual, concise and appealing way that is perfectly suited to the informal style of Social Media communications. Emoticons have also been often exploited to label and thus characterize the textual excerpts where they occur and several language resources have been built [16,4]. Go et al. [7] and Castellucci et al.[6] use distant supervision over emotional-labeled textual contents in order to train a sentiment classifier and build a polarity lexicon. Aoki et al.[1] describe a methodology to represent each emoticon as a vector of emotions, and Jiang [9] propose an approach that relies on word2vec [14] to build a distributional semantic vectorial space where to represent and compare emojis. Cappallo et al.[5] proposed Image2Emoji, a multimodal approach for generating emoji labels for images. Barbieri et al. [3] built distributional semantic spaces of emojis and studied the best settings to better learn these models (window of 6 words, cleaning stopwords and punctuation). Barbieri et al. studied, like in this research, the use of the emojis in Twitter over different languages (American English, British English, Spanish, and Italian) [2].

## 3. Dataset and Text Analysis

We construct a vector space model aiming at providing a common semantic ground in which emojis are naturally distributed according to idiosyncratic ways users may use them. Specifically, we gathered a dataset composed of more than 1 million tweets retrieved with the Twitter APIs. We took advantage of geo-localization in order to obtain tweets that were posted from the Madrid and Barcelona metropolitan areas.

Tweets were posted between October 2015 and May 2016. Our dataset is divided as follows: 400.000 tweets sent for Barcelona, and 600.000 from Madrid. In Table 1 we

**Table 2.** Experiment 1, emojis with high (H) and Low (L)  $sim_{c_1 c_2}$  (indicated as “Common” in the table)

H	Emoji																				
	Common	9	9	8	8	8	8	8	8	8	8	8	8	8	8	8	8	7	7	7	7
L	Emoji																				
	Common	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0

show the most frequent emojis for each city. We can see that 😄, ❤️ and 🍷 are the most common emojis for both cities.

Tweet texts were preprocessed with the CMU Tweet Twokenizer. We also removed, from each tweet, all hyperlinks, and lowercased all textual content in order to reduce noise and sparsity.

We generated our joint vector space models of tweet words and emojis by relying on the skip-gram neural embedding model introduced by Mikolov et al. [12], adopting the same methodology of Barbieri et al. [3] We built two skip-gram models (one per city) with 300 dimensions and a window size of 6 tokens.

#### 4. Experiments and Evaluation

We run two types of experiments. The first experiment was designed to explore the meaning of the emojis across different cities, and observe whether emoji meanings change over different population. The second experiment was run in order to see if pairs of emojis are used in the same way over citizens of Madrid and Barcelona. In order to study only frequent emojis, that are consistently modelled within the vectorial semantic spaces, we reduced the number of emojis by taking the 150 most frequent emojis across the two cities.

##### 4.1. Experiment 1

In experiment 1 we explore the meaning of the emojis. In order to represent an emoji we use other emojis with similar vectorial representations and thus presumably close in meaning. We define the Nearest Neighbours  $NN_l(e)$  of the emoji  $e$  in the cities  $c$ , as the set of the 10 nearest emojis to the emoji  $e$  in the semantic space of city  $c$ . We retrieve the nearest neighbours of each emojis with respect to its cosine similarity with other emojis. To this purpose, we exploit the vectorial semantic model built by considering both emojis and words of tweets, but we consider only emojis as valid neighbours. Since an emoji is defined by other similar emojis we are able to compare this representation over different languages, as the emojis vocabulary is the same.

























































In order to see if an emoji is similarly defined in two cities we look at the common elements in the  $NN$  representation of the emoji in the two cities. If the representations of the emoji in different cities share many elements it would mean that the emoji is defined and used in a similar way. If there are not common elements among the two representations, the emoji is more likely to mean something different in the two cities. More precisely, to determine if emoji  $e$  is similar in language  $c_1$  and  $c_2$  we measure the size of the intersection of the  $NN$  sets:

$$sim_{l_1 l_2}(e) = |NN_{l_1}(e) \cap NN_{l_2}(e)|$$

**Table 3.** Example of emoji used different. In the table are shown the 10 nearest neighbour of the paw prints emoji in the Barcelona and Madrid models.

10 Colosest Emojis to 🐾	
Barcelona	         
Madrid	         

**Table 4.** Experiment 2, pair of emojis with highest similarity disagreement between two languages.

N	Emoji Pair	Barcelona	Madrid	N	Emoji Pair	Barcelona	Madrid
1	 	0.194	-0.111	15	 	0.368	0.649
2	 	0.446	0.021	16	 	0.199	0.436
3	 	0.348	-0.036	17	 	0.352	0.651
4	 	0.474	0.11	18	 	0.032	0.323
5	 	0.226	0.58	19	 	0.046	0.302
6	 	0.528	0.083	20	 	0.164	0.399
7	 	0.503	0.13	21	 	0.137	0.422
8	 	0.433	0.075	22	 	0.174	0.446
9	 	0.413	0.05	23	 	0.247	0.467
10	 	0.22	0.579	24	 	0.162	0.405
11	 	0.402	0.051	25	 	0.203	0.452
12	 	0.204	0.525	26	 	0.131	0.376
13	 	0.226	0.545	27	 	0.14	0.38
14	 	0.218	0.447	28	 	0.312	0.517

We assume that if  $sim_{c_1c_2}$  is equal to 10, the emoji  $e$  has the same meaning in the cities  $l_1$  and  $l_2$  (respectively Barcelona and Madrid). On the other hand, if  $sim_{c_1c_2}$  is equal to 0 the emoji means something different in the two languages.

In Table 2 we report the results of Experiment 1 for the emojis with highest  $sim_{c_1c_2}$  on the left and with the lowest  $sim_{c_1c_2}$  on the right. The emojis that seem to keep the same meaning independently from the language are the music, nature, parties and food related ones. The emojis with the lowest  $sim_{all}$  have a score equal to 0,1,2. For example, the mouth, gun, school or ghost seem to be used in different ways in both cities, even if they seem to portray a clear-cut unambiguous meaning. Let us provide a more elaborate example of one of the most dissimilar emojis, namely the paw prints emoji. In Table 3, we observe that in Madrid, this emoji is very strongly associated to fauna, not only dogs, as we can find among its closest neighbours emojis of cats, crocodiles, rats or bees. However, in Barcelona, this emoji has a strong family component, as it is used in similar contexts as in-love couples, families with kids, and elderly people.

#### 4.2. Experiment 2

We explore the use and the similarity of emoji pairs between Barcelona and Madrid and compute a similarity matrix of the 150 emojis we consider. This is achieved by computing cosine similarity of an emoji with respect to the others, and normalising its value by dividing by the average of all the scores of the matrix.

#### 4.2.1. Correlation Between Similarity Matrices

We exploit the similarity matrices to analyse whether Barcelona and Madrid represent emojis in similar ways. The Pearson's correlation of the similarity matrices of both cities is 0.708. This is an interesting finding, since tweets vocabularies are not fully overlapping (for instance, due to the influence of the Catalan language, which amounts to more than 1/3 of the total).

#### 4.2.2. Studying Dissimilar Emojis

We previously observed that emojis semantics is somehow preserved. However, this does not hold for all of them. In this section, we explore disagreement in the two similarity matrices (in a similar fashion as [11]). Emojis pairs that show different similarity scores across both cities are shown in Table 4. We report numbers on those emojis with highest disagreement, i.e. high difference between the similarities in the two cities.

One remarkable example is the beer emoji. In the Madrid model the beer emoji is close to the heart emoji (25,26,27,28). This is certainly not the case in Barcelona. While it would be tempting to suggest that Madrid inhabitants may find a stronger correlation between love and beer, a more thorough analysis would be required to support this claim.

Another notable example is the lightning emoji. In Madrid, users associate it to bad and good weather indistinctly (22,23,24) and also with "agreement" emojis (18,19,20,21). In addition to the lightning emoji, another interesting example is the camera emoji. Users in Barcelona associate it frequently to the angry face (1), while users in Madrid do not. The only emoji showing higher similarity with the photo camera in Madrid is the basketball (5). The fact that Barcelona is a city with a very strong presence of tourists may trigger many complaining tweets about overcrowding, which would explain the higher correlation between photo camera and angry face emojis.

Finally, another interesting difference is the correlation that can be found between the gun emoji and high caloric food (7,8) by Barcelona Twitter users, in contrast with tweets coming from Madrid.

Let us finally mention the fact that there are inconsistencies that are difficult to explain, and which require further analysis.

## 5. Conclusions

In this paper we have explored the meaning and usage of emojis in Barcelona and Madrid tweets. This was achieved thanks to the use of distributional semantic models representing the semantics of the emojis in the two cities. Madrid and Barcelona models are compared, and analyzed both quantitatively and qualitatively. Our results suggest that, overall, emojis somehow preserve similar semantics across the cities we considered, but subtle differences may be spotted in the use of certain emojis. We have provided preliminary insights on possible explanations for the discrepancies in the usage of a narrow set of emojis between Barcelona and Madrid. These are only preliminary results, and we are planning to continue our research in this line, performing more extensive analyses of some of the patterns revealed in this paper.

## Acknowledgements

We acknowledge partial support from the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE) and the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502).

## References

- [1] Sho Aoki and Osamu Uchida, 'A method for automatically generating the emotional vectors of emoticons using weblogs articles', in *Proc. 10th WSEAS Int. Conf. on Applied Computer and Applied Computational Science, Stevens Point, Wisconsin, USA*, pp. 132–136, (2011).
- [2] Francesco Barbieri, Krzysztof German, Francesco Ronzano, and Horacio Saggion, 'How cosmopolitan are emojis? exploring emojis usage and meaning over different languages with distributional semantics', *Proceedings of the 24th Annual ACM Conference on Multimedia Conference*, (2016).
- [3] Francesco Barbieri, Francesco Ronzano, and Horacio Saggion, 'What does this emoji mean? a vector space skip-gram model for twitter emojis', *Proceedings of the International Conference on Language Resources and Evaluation*, (2016).
- [4] Marina Boia, Boi Faltings, Claudiu-Cristian Musat, and Pearl Pu, 'A:) is worth a thousand words: How people attach sentiment to emoticons and words in tweets', in *Social Computing (SocialCom), 2013 International Conference on*, pp. 345–350. IEEE, (2013).
- [5] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek, 'Image2emoji: Zero-shot emoji prediction for visual media', in *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pp. 1311–1314. ACM, (2015).
- [6] Giuseppe Castellucci, Danilo Croce, and Roberto Basili, 'Acquiring a large scale polarity lexicon through unsupervised distributional methods', in *Natural Language Processing and Information Systems*, 73–86, Springer, (2015).
- [7] Alec Go, Richa Bhayani, and Lei Huang, 'Twitter sentiment classification using distant supervision', *CS224N Project Report, Stanford*, 1, 12, (2009).
- [8] Alexander Hogenboom, Daniella Bal, Flavius Frasinca, Malissa Bal, Franciska De Jong, and Uzay Kaymak, 'Exploiting emoticons in polarity classification of text.', *J. Web Eng.*, **14**(1&2), 22–40, (2015).
- [9] Fei Jiang, Yi-Qun Liu, Huan-Bo Luan, Jia-Shen Sun, Xuan Zhu, Min Zhang, and Shao-Ping Ma, 'Microblog sentiment analysis with emoticon space model', *Journal of Computer Science and Technology*, **30**(5), 1120–1129, (2015).
- [10] Tanimu Ahmed Jibril and Mardziah Hayati Abdullah, 'Relevance of emoticons in computer-mediated communication contexts: An overview', *Asian Social Science*, **9**(4), 201, (2013).
- [11] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini, 'Representational similarity analysis—connecting the branches of systems neuroscience', *Frontiers in systems neuroscience*, **2**, 4, (2008).
- [12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781*, (2013).
- [13] Tomas Mikolov, Quoc V Le, and Ilya Sutskever, 'Exploiting similarities among languages for machine translation', *arXiv preprint arXiv:1309.4168*, (2013).
- [14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 'Distributed representations of words and phrases and their compositionality', in *Advances in neural information processing systems*, pp. 3111–3119, (2013).
- [15] Jaram Park, Young Min Baek, and Meeyoung Cha, 'Cross-cultural comparison of nonverbal cues in emoticons on twitter: Evidence from big data analysis', *Journal of Communication*, **64**(2), 333–354, (2014).
- [16] Duyu Tang, Furu Wei, Bing Qin, Ming Zhou, and Ting Liu, 'Building large-scale twitter-specific sentiment lexicon: A representation learning approach.', in *COLING*, pp. 172–182, (2014).
- [17] Peter D Turney, Patrick Pantel, et al., 'From frequency to meaning: Vector space models of semantics', *Journal of artificial intelligence research*, **37**(1), 141–188, (2010).