How Cosmopolitan Are Emojis?

Exploring Emojis Usage and Meaning over Different Languages with Distributional Semantics

Francesco Barbieri
Universitat Pompeu Fabra
Barcelona, Spain
francesco.barbieri@upf.edu

German Kruszewski
University of Trento
Trento, Italy
germank@gmail.com

Francesco Ronzano
Universitat Pompeu Fabra
Barcelona, Spain
francesco.ronzano@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
Barcelona, Spain
horacio.saggion@upf.edu

ABSTRACT

Choosing the right emoji to visually complement or condense the meaning of a message has become part of our daily life. Emojis are pictures, which are naturally combined with plain text, thus creating a new form of language. These pictures are the same independently of where we live, but they can be interpreted and used in different ways. In this paper we compare the meaning and the usage of emojis across different languages. Our results suggest that the overall semantics of the subset of the emojis we studied is preserved across all the languages we analysed. However, some emojis are interpreted in a different way from language to language, and this could be related to socio-geographical differences.

Keywords

Emojis, Natural Language Processing, Distributional Semantics

1. INTRODUCTION

The information society has considerably changed the way in which we communicate with each other mainly due to the advent of Social Media. Social networking platforms such as Twitter allow users to post short text messages to update followers on current affairs, sentiments, emotions and express opinions on any topic. During the last few years, Twitter users have started to extensively use emojis in their posts[1]. Emojis are pictures that can be naturally combined with plain text to create a new form of language[2]; 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. In spite of their assumed universality, the sense of an emoji may change from language to language and culture to culture. Understanding the meaning of emojis with respect to their context of use is important for multimedia information indexing, retrieval, or content extraction systems.

In this paper we investigate the use of emojis across languages from a natural language processing viewpoint. We adopt an empirical research methodology relying on current vector space representation modelling [22, 15] to understand the "semantics" of these important elements of multimedia communication. More specifically, we collected a corpus of more than 30 million tweets in four languages, American English (USA), British English (UK), Peninsular Spanish (ESP), and Italian (ITA), and carried out various experiments to compare emojis. Despite the languages have different vocabularies and syntactic structures, we were able to find a way to compare the use of emojis across languages. Our results demonstrate that the semantics of the 150 most popular emojis is somehow preserved across different languages. Nevertheless, for some emojis we observed interesting language specific usage patterns. For instance, the emojis 🔥, 👋🏽, and 🙌🏻 seem to be used in different contexts across distinct languages, while there is a relative agreement on the cross-language use and meaning of 🎶 and 🌳.

In the next section we overview previous studies on emojis. In Section 3, we describe the dataset we have collected and the text processing tools we use to model the meaning of emojis across languages. By relying on this dataset, we run two experiments: in the first one (Section 4.1) for each emojis we compare the most similar ones in different languages. In our second experiment (Section 4.2) we focus on the comparison of the similarity of pairs of emojis across languages: in particular we analyse the language-dependent similarity matrices of the 150 most popular emojis. We find that the matrices that describe the four language variations considered are strongly correlated[3]. We conclude the pa-

[1]Realtime Emoji use on Twitter can be tracked at http://emojitracker.com/

[2]A complete list can be found at http://emojipedia.org/

[3]Detailed and complete results can be found at http://sempub.taln.upf.edu/tw/cosmopolitan/

Rank	USA		UK		ESP		ITA	
1	😂	350	😂	32	😍	46	❤️	55
2	❤️	301	❤️	27	❤️	38	😂	35
3	😍	213	😍	22	😂	34	😍	34
4	🔥	166	🎄	20	💕	24	🇮🇹	16
5	💯	104	😊	13	👌	23	😊	12
6	😊	101	🎅	11	😘	22	🔝	12
7	🙌	89	🎉	10	🎉	22	😘	11
8	😘	86	👍	10	💪	21	😁	10
9	🎄	85	👌	10	😊	19	☀️	10
10	💕	84	🇬🇧	9	👏	16	😎	10
11	🎉	80	😘	9	🔝	14	🎉	9
12	😭	79	✨	8	💃	12	💙	8
13	💙	77	❄️	8	💜	12	😉	8
14	✨	75	😁	8	👭	12	🌹	7
15	❄️	72	💕	8	💞	11	✌️	7

Table 1: The 15 most frequent emojis across the four languages studied. For each language, next to each emoji, we show the thousand of occurrences in the our dataset.

per with a summary of our findings and avenues for further research.

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 [12, 19, 18]. 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 communication. The meaning expressed by emoticons has been exploited to enable or improve several tasks related to the automated analysis of Social Media contents, like sentiment analysis [10, 9]. In this context, emoticons have also been often exploited to label and thus characterize the textual excerpts where they occur. As a consequence, by analyzing all the textual contents where a specific emoticon appears several sentiment and emotional lexicons have been build [23, 21, 21, 4]. Go et al. [8] and Castellucci et al. [7] use distant supervision over emotion-labeled textual contents in order to respectively 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 [11] proposed a sentiment and emotion classifier based on semantic spaces of emojis in the Chinese Website Sina Weibo.

Novak et al. [17] built a lexicons and drew a sentiment map of the 751 most frequently used emojis. Cappallo et al. [5] proposed Image2Emoji, a multimodal approach for generating emoji labels for images (they also presented a demo [6]). Miller et al. [16] explored whether emoji renderings or differences across platforms (e.g. Apple's iPhone vs. Google's Nexus phone) give rise to diverse interpretations of emojis. Pavalanathan and Eisenstein [20] used a matching approach from causal inference to test whether the adoption of emojis causes individual users to employ fewer emoticons in their text on Twitter.

Finally, we explored meaning of Twitter emojis in American English with Distributional Semantics [3]. We tested our models with semantic similarity experiments, comparing our models with human assessment. We also carried out a qualitative evaluation, exploring cluster of emojis and the most related words to each emoji (the models can be found online[4]). We also explored the usage of the emojis in Madrid and Barcelona [2].

3. DATASET AND TEXT ANALYSIS

To support the creation of the semantic vectorial models presented in this paper we gathered a dataset composed of more than 30 million tweets retrieved with the Twitter APIs. We retrieved geo-located tweets that were posted from United States of America, United Kingdom, Spain, and Italy. We collected tweets from October 2015 to April 2016. We used geo-located tweets in order to retrieve tweets from real user, filtering out spam and bot generated tweets. The total number of tweets is 28,8 millions for USA, 2.1 for United Kingdom, 1.56 for Spain and 1.63 for Italy. Table 1 shows, for each language, the 15 most frequent emojis together with the number of times each emoji occurs. We can see that 😂, ❤️ and 😍 are the most common emojis in each one of the four languages considered.

In order to preprocess the text of each tweet we follow the same procedure of Barbieri et al. [3]. We modelled in the same vectorial space both the words and the emojis of tweets by means of embeddings, by relying on the skip-gram embedding model introduced by Mikolov et al. [14] with 300 dimensions and a window size of 6 tokens (we previously found out that this is the best configuration to model emojis [3]). We built 4 models, one per language.

4. EXPERIMENTS AND EVALUATION

We run several experiments to compare the way the semantics of emojis varies across languages. In a first experiment (Section 4.1) we investigate if the meaning of single emojis is preserved across language variations. In a second experiment (Section 4.2), we compare the overall semantic models of the 150 most frequent emojis across languages.

4.1 Experiment 1

In our first experiment, we quantify how the meaning of an emoji A is preserved across different languages by measuring to what extent the emojis that are most similar to A overlap across languages. We exploit the vectorial representation of each emojis in a specific language to select the ones with similar vectors and thus presumably closest in meaning. We define the Nearest Neighbours $NN_l(e)$ of the emoji e in the language l, as the set of the 10 nearest emojis[5] to the emoji e in the semantic space of language l. We retrieve the nearest neighbours of each emojis with respect to its cosine similarity with other emojis. Note that the semantic vectors are derived from co-occurrence statistics extracted from both emojis and words. However, since an emoji is defined by other similar emojis, we are able to compare these representations across different languages. In order to see if an emoji is similarly defined in two languages,

[4]http://sempub.taln.upf.edu/tw/emojis/

[5]In average the cosine similarity drops after the 10^{th} closest emoji and for each emoji the cosine similarity of the ten most similar emojis is always greater than 0.4.

	Rank	USA UK	USA ESP	USA ITA	UK ESP	UK ITA	ESP ITA	s_{all}
🎶	23	9	9	9	9	9	9	9
🎤	91	9	9	9	9	9	9	9
🎵	118	9	9	9	9	9	9	9
🎼	126	9	9	9	9	9	9	9
🎧	146	9	8	9	8	9	8	8
🌳	150	8	9	9	9	8	8	8
🙋	128	8	8	8	8	8	8	8
🌸	74	8	9	8	8	7	9	7
🌿	77	7	7	8	9	9	8	7
🍁	80	7	7	7	9	9	9	7
🍂	72	7	7	7	9	9	8	7
🎃	79	8	8	7	8	7	7	7
🍴	84	8	7	9	7	9	8	6
🍔	144	8	7	8	8	7	6	6
🎂	60	6	9	8	6	6	8	6
👑	39	3	5	4	2	2	3	0
😅	55	4	1	5	1	4	3	0
🙌	14	5	3	0	4	0	1	0
😈	57	2	2	2	3	1	2	0
💎	85	3	3	2	2	1	1	0
😌	87	1	5	2	2	0	2	0
🔥	30	0	0	1	3	3	3	0
👋	95	0	4	2	0	1	2	0
🌈	149	3	4	0	1	0	1	0
😬	148	3	0	0	1	1	3	0
⚡	100	3	1	2	2	0	0	0
💯	69	3	0	0	0	0	3	0
🍀	97	1	0	0	0	4	0	0
👀	78	1	1	0	0	1	0	0
👅	110	1	0	1	0	0	0	0

Table 2: Experiment 1, emojis with high sim_{all} (indicated as s_{all} in the table) on the top, and emojis with low sim_{all} in the bottom.

we look at the common elements in the *NN* representation of that emoji in the two languages. If the representations of the emoji in different languages share many elements it would mean that the emoji is defined and thus 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 languages. More precisely, to determine if emoji e is similar in language l_1 and l_2 we measure the size of the intersection of the NN sets:

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

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

Moreover, we also measure whether an emoji means the same across all the languages by looking at the overlap of all the sets of emojis that are most similar to the emoji e in each language:

$$sim_{all}(e) = |NN_{l_1}(e) \bigcap NN_{l_2}(e) \bigcap ... \bigcap NN_{l_n}(e)|$$

where n is the number of languages.

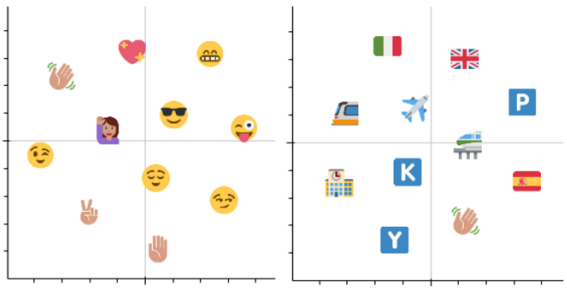

Figure 1: tSNE of nearest neighbours of the waving-hand emoji for USA (left) and UK (right)

	USA	UK	ESP	ITA	AVG
USA	1	**0.76**	0.743	0.698	0.734
UK	0.76	1	0.719	0.702	0.727
ESP	0.743	0.719	1	0.739	0.734
ITA	0.698	0.702	0.739	1	0.713

Table 3: Experiment 2, Pairwise Pearson's Correlation between similarity matrices.

In the top half of Table 2 we report the results of Experiment 1 for the emojis with highest sim_{all}. For each emoji are indicated the rank (where 1 is the most common emoji over the four languages and 150 is the least used emoji), the six combinations of $sim_{l_1 l_2}$, and the sim_{all}. The emojis that seem to keep the same meaning independently from the language are the music, nature and food related ones. In the bottom half of Table 2 are reported the emojis with the lowest sim_{all} (all emojis have a sim_{all} score equal to 0): these emojis meanings are probably language dependent. Looking at the bottom of the table we can see that the emojis 👀 and 👅 are used in a very different way across all the languages, and each language seems to have its own way to define them. Also the emojis 💯, 🙌, and 👋 do not seem to keep their meaning across different languages (the number 100 for example might be just a number or a excellent grade). For the waving-hand emoji 👋, we also plot the NN for USA and UK, (Figure 1) and we can observe that these two emojis are interpreted in different ways. In the case of American English the waving-hand seems to mean *bye/see you later* (smiles and people waving), while for British English the waving-hand emojis is related to travelling (countries flags, train and airplane are included in the NN). On the other hand, 👑, 💎, 🔥, and 🍀 are concrete objects but they are exploited to convey different meanings across different languages.

4.2 Experiment 2

In order to understand the use and the similarity of pairs of emojis across different languages we compute for each language the similarity matrix of the 150 emojis that we are studying. The value of each cell of this matrix is equal to the cosine similarity of the corresponding pair of emojis. We

	USA	UK		USA	ESP		USA	ITA		UK	ESP		UK	ITA		ESP	ITA
	0.54	0.06		0.39	0.73		0.35	0.67		0.2	0.68		0.29	0.62		0.22	0.57
	-0.01	0.4		0.54	-0.01		0.37	0.6		0.24	0.66		0.25	0.57		0.57	0.14
	0.26	0.62		0.35	0.66		0.35	0.59		0.35	0.75		0.27	0.58		0.57	0.16
	0.72	0.25		0.47	0.75		0.4	0.62		0.43	0.81		0.43	0.68		0.48	0.11
	0.25	0.61		0.46	0.72		0.38	0.59		0.22	0.61		0.43	0.68		0.5	0.13
	0.21	0.57		0.41	0.68		0.23	0.48		0.18	0.58		0.32	0.59		0.48	0.11
	0.58	0.9		0.26	0.56		0.42	0.61		0.47	0.82		0.57	0.11		0.47	0.11
	0.16	0.51		0.33	0.61		0.21	0.46		0.18	0.57		0.39	0.64		0.17	0.47
	0.39	0.71		0.29	0.57		0.6	0.73		0.25	0.59		0	0.33		0.75	0.36
	0.23	0.57		0.28	0.56		0.12	0.39		0.22	0.56		0.37	0.61		0.5	0.15
	0.44	0.03		0.47	0.71		0.36	0.56		0.41	0.72		0.47	0.68		0.43	0.09
	0.17	0.51		0.2	0.5		0.35	-0.06		0.42	0.72		0.69	0.85		0.68	0.31
	0.26	0.59		0.49	0.04		0.21	0.45		0.34	0.65		0.11	0.4		0.68	0.32
	0.37	0.69		0.53	0.08		0.43	0.6		0.27	0.59		0.25	0.51		0.39	0.06

Table 4: Experiment 2, pair of emojis with highest similarity difference between two languages.

normalise the cosine similarity values by the average cosine similarity of all the pairs of emojis.

4.2.1 *Correlation Between Languages*

We take advantage of the similarity matrices to analyse whether two languages represent emojis in similar ways. The Pearson's correlation of the similarity matrices of the four languages are reported in Table 3. We can see that most of the languages are strongly correlated to each other. This is an interest finding, as vocabularies of the languages are different, and the context words modelled by the semantic spaces too, but the semantic of the emojis we studied is in some way preserved. American English and Spanish are the languages that interpret emojis in the most universal way, with high correlation to all the other languages (both with an average of 0.734). British English has a lower similarity matrix correlation with other languages (average of 0.727) and Italian scores an even lower average correlation, 0.713. This suggests that these two latter languages interpret emojis in a slightly different way than the other languages, especially Italian.

Looking at the single emoji-pairs similarities, the strongest correlation is between USA and UK (0.760), probably supported by similar vocabularies, and the weakest is between USA and ITA (0.698). On the other hand, Italian has high correlation with Spanish (0.739), while Spanish correlates better with Italian and American English than British English.

4.2.2 *Emoji Differences*

Even if we observed that in most of the cases the semantics of emojis is somehow preserved across languages, we can also spot some interesting difference in the language-specific use of these pictograms. In particular, in this section we explore the disagreement in the similarity matrices of the four languages (using a method similar to [13]). We analyse the pairs of emojis that have different similarities across two languages. Table 4 shows the similarity matrix scores of pairs of emojis for all the possible language combinations. We report pairs of emojis which are semantically related in one language (e.g. USA) but unrelated in the other (e.g. UK).

The emojis with higher differences in American and British English are and that are similar in USA but not in UK. On the other hand the emoji seems to be used in a different ways in UK, as it scores higher than USA with the emojis , , and . It seems that the gift emoji is used mostly as Christmas gift in UK. Other interesting combinations are and that apparently are very similar in USA but not in Spain, and the pizza emoji which highly correlates with in Italian but not in USA, probably for the different way to eat pizza in these two countries.

An interesting pattern in the UK results are the emojis related to the beach and the good weather: emojis like , , , and are similar to each other in Italian and Spanish, but not in British English, and this is probably related to the countries geographical location.

Two emojis that are used in a different way in Spain are and . They both appear in various combinations in the disagreement table (Table 4) of the Spanish language. The sly smile and are similar in Spain but not in UK and Italy, suggesting that the combination is frequently used only in Spain. One of the differences between Spanish and Italian is the emoji that in Spanish is similar to , and , but not in Italian. One last interesting pattern is the emoji that in Italian means approvement as it is similar to emojis like and , but it is not the case in USA and UK (where probably the emoji is used instead).

5. CONCLUSIONS

In this paper we explore the meaning and usage of emojis across four languages: American English, British English, Peninsular Spanish and Italian. We use distributional semantic models to represent the semantics of the emojis in the four languages, and we compare the language-specific models of each emoji. Our results suggest that in spite of differences in use of emojis across the languages we studied, the overall semantics of the most frequent emojis is similar. These are only preliminary results, and we are planning to run further, more extensive analyses of the cross-language meaning of emojis in the near future.

Acknowledgements

We thank the three anonymous reviewers for their useful comments, especially for the future work. We received 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). First author also acknowledges the COST Action IC1307 iV&L Net (European Network on Integrating Vision and Language), supported by COST (European Cooperation in Science and Technology).

6. REFERENCES

[1] S. Aoki and O. Uchida. A method for automatically generating the emotional vectors of emoticons using weblog articles. In *Proc. 10th WSEAS Int. Conf. on Applied Computer and Applied Computational Science, Stevens Point, Wisconsin, USA*, pages 132–136, 2011.

[2] F. Barbieri, L. Espinosa-Anke, and H. Saggion. Revealing Patterns of Twitter Emoji Usage in Barcelona and Madrid. In *International Conference of the Catalan Association for Artificial Intelligence, CCIA*, Barcelona, Spain, October 2016.

[3] F. Barbieri, F. Ronzano, and H. Saggion. What does this Emoji Mean? A Vector Space Skip-Gram Model for Twitter Emojis. In *Language Resources and Evaluation conference, LREC*, Portoroz, Slovenia, May 2016.

[4] M. Boia, B. Faltings, C.-C. Musat, and P. 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*, pages 345–350. IEEE, 2013.

[5] S. Cappallo, T. Mensink, and C. G. Snoek. Image2emoji: Zero-shot emoji prediction for visual media. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 1311–1314. ACM, 2015.

[6] S. Cappallo, T. Mensink, and C. G. Snoek. Query-by-emoji video search. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 735–736. ACM, 2015.

[7] G. Castellucci, D. Croce, and R. Basili. Acquiring a large scale polarity lexicon through unsupervised distributional methods. In *Natural Language Processing and Information Systems*, pages 73–86. Springer, 2015.

[8] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12, 2009.

[9] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. de Jong, and U. Kaymak. Exploiting emoticons in sentiment analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 703–710. ACM, 2013.

[10] A. Hogenboom, D. Bal, F. Frasincar, M. Bal, F. De Jong, and U. Kaymak. Exploiting emoticons in polarity classification of text. *J. Web Eng.*, 14(1&2):22–40, 2015.

[11] F. Jiang, Y.-Q. Liu, H.-B. Luan, J.-S. Sun, X. Zhu, M. Zhang, and S.-P. Ma. Microblog sentiment analysis with emoticon space model. *Journal of Computer Science and Technology*, 30(5):1120–1129, 2015.

[12] T. A. Jibril and M. H. Abdullah. Relevance of emoticons in computer-mediated communication contexts: An overview. *Asian Social Science*, 9(4):201, 2013.

[13] N. Kriegeskorte, M. Mur, and P. A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4, 2008.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[15] T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*, 2013.

[16] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht. "Blissfully happy" or "ready to fight": Varying Interpretations of Emoji. *ICWSM'16*, 2016.

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

[18] J. Park, Y. M. Baek, and M. Cha. Cross-cultural comparison of nonverbal cues in emoticons on twitter: Evidence from big data analysis. *Journal of Communication*, 64(2):333–354, 2014.

[19] J. Park, V. Barash, C. Fink, and M. Cha. Emoticon style: Interpreting differences in emoticons across cultures. In *ICWSM*, 2013.

[20] U. Pavalanathan and J. Eisenstein. Emoticons vs. emojis on twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480*, 2015.

[21] D. Tang, F. Wei, B. Qin, M. Zhou, and T. Liu. Building large-scale twitter-specific sentiment lexicon: A representation learning approach. In *COLING*, pages 172–182, 2014.

[22] P. D. Turney, P. Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

[23] C. Yang, K. H.-Y. Lin, and H.-H. Chen. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. Association for Computational Linguistics, 2007.