

From Emoji Usage to Categorical Emoji Prediction

Gaël Guibon, Magalie Ochs, Patrice Bellot

▶ **To cite this version:**

Gaël Guibon, Magalie Ochs, Patrice Bellot. From Emoji Usage to Categorical Emoji Prediction. 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING 2018), Mar 2018, Hanoï, Vietnam. hal-01871045

HAL Id: hal-01871045
https://amu.hal.science/hal-01871045

Submitted on 10 Sep 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

From Emoji Usage to Categorical Emoji Prediction

Gaël Guibon[1,2], Magalie Ochs[1], and Patrice Bellot[1]

[1] LIS-CNRS UMR 7020, Aix-Marseille Université
[2] Caléa Solutions
firstname.lastname@lis-lab.fr

Abstract. Emoji usage drastically increased recently, they are becoming some of the most common ways to convey emotions and sentiments in social messaging applications. Several research works automatically recommend emojis, so users do not have to go through a library of thousands of emojis. In order to improve emoji recommendation, we present and distribute two useful resources: an emoji embedding model from real usage, and emoji clustering based on these embeddings to automatically identify groups of emojis. Assuming that emojis are part of written natural language and can be considered as words, we only used unsupervised learning methods to extract patterns and knowledge from real emoji usage in tweets. Thereby, emotion categories of face emojis were obtained directly from text in a fully reproductible way. These resources and methodology have multiple usages; for example, they could be used to improve our understanding of emojis or enhance emoji recommendation.

Keywords: emoji, recommendation, word embeddings, resource, clustering

1 Introduction

Research on emojis are growing steadily since a couple of years. Nowadays users of instant messaging applications have to scroll through huge libraries of emojis to select one: it would be useful to help the users by automatic recommendation. However, several emojis can be used to convey the same emotion, for instance 😀 and 😁, and it is not clear whether or not emojis can be considered as actual words, groups of metadata, or extra linguistic information for Natural Language Processing (NLP) approaches. Most of the recommendation systems tackle emoji as metadata. In this paper, we propose an approach considering emoji as words without any assumptions on their meanings.

Nowadays, few emoji recommendation systems have been proposed, and only very recently. Moreover, actual emoji recommendation systems consider each emoji as a single label to recommend. Those systems obtained perfectible results: 65% accuracy [20] and 65% f1-score [2]. These systems focused only on a very

limited number of emojis. Considering these works, instead of recommending emojis we propose to automatically recommend groups of emojis. Our work is based on the following hypotheses:

1. Emojis can be quite similar (💙❤️📷📸) and it is not clear whether we should recommend an emoji instead of another, thus we should predict categories of emojis instead of only considering specific ones
2. Emoji categories should be inferred from their real usage in order to adapt to changes and cultural differencies in order to obtain a good recommendation
3. Emotion-related emojis cannot be recommended only by matching keywords, as current smartphone systems implement for 🎄🍇 for instance. Emotion-related emojis can come from the whole feeling of the text. This is why we focus on them at first.

Considering these hypotheses, our purpose in this paper is to automatically regroup emojis based on their usage, focusing on the common subset of 63 face emojis. We propose an automatic emotion categorization of face emojis not by using metadata, but by using the real context usage of emojis in order not to assume predefined informations. The methodology used to obtain these resources can be applied on any emoji type, even though we applied it to face emojis: it consists in analyzing the similarity of context usage between emojis by using two unsupervised approaches: first, word embeddings of only tweets containing emojis (Section 3), then using these embeddings to apply a clustering algorithm (Section 4).

The paper is organized as follow: we summarize the related work on emojis and emoji embeddings (Section 2) before presenting our first resource: emoji embeddings (Section 3). Then, we present the second resource, a fine-grained clustering of face emojis (Section 4) before validating it with Ekman's theory on emotion-face expressions (Section 4.2).

2 Related Work

Emoticons (:-) , :P) and emojis (😍) are two different ways to represent facial cues. While the former are characters, the latter are pictures, and as such they can also represent different concepts and ideas, not only facial cues and expressions. Moreover, emojis tend to replace emoticons in social conversations [16]. The first 176 emojis were released in 1999 by the japanese telecom NTT DOCOMO[3], right now there are 2623 emojis according to the official Unicode list[4]. Their success is due to the support of emojis by the first iPhone from Apple, then other brand such as Google or Samsung started supporting them.

At first, some research work in socio-linguistics focused on the diverse understanding and usage of emojis, and the role they have in a textual conversation. According to these, emojis are used to improve the understanding of the message in 70% of cases [9]. Also, it has been demonstrated that emojis can serve

[3] `https://www.nttdocomo.co.jp/`
[4] `http://unicode.org/emoji/charts/full-emoji-list.html`

a number of roles in conversation [10] which are not necessarily related to the expression of emotions, such as maintaining a conversational connection, or engaging in playful interaction. Some of these roles can be linked to the Jackobson's functions of language [8]:

- Phatic function: "👋"
- Referential function: "Just bought it 🚲"
- Emotive function: "Seriously?! 😤"
- Conative function: "👆" for "call me"

Thus, it is necessary to have a good emoji recommendation system in social messaging application in order to enhance the conversation's quality.

Few research works focused on emoji recommendation, the main approach is based on neural networks to predict predict emojis. However, the performance of these models remains perfectible. Xie *et al.* [20] achieved 65% of accuracy for the 3 mostly used emojis in conversations from Weibo[5], the chinese Twitter, using Hierarchical LSTM [11]. Barbieri *et al.* [2] predicted the 20 most used emojis in 40 millions tweets using Long Short-term Memory Networks [7] and evaluated their prediction on the 5 most frequent emojis obtaining an average f1-score of 65%.

They are still only a few available resources for research on emoji, whether it is emoji embedding models or other kind of resources. Considering embedding models, there have been only a few work on emoji embeddings and all of them have been done recently. On one hand some work do not embed emojis directly. Emoji has been considered to be a group of meta-informations or words upon which the embedding will be based on. These meta-informations can be Unicode descriptions of emojis [4] or their possible meanings and senses associated to their description [1,19].

On the other hand, emoji have been embedded directly in context with all types of emojis from millions of tweets. In this paper, we used similar approach. However, in existing works, arbitary clusters have been defined [3] or hierarchial ones mixing all types of emojis either to detect sarcasm [6] or to find the best keyboard mapping [17]. The resources proposed in this paper are created totally automatically without predefined clusters and focusing on a particular type of emoji: the face emojis. The face emojis correspond in the Unicode list[6] to the "face positive", "face neutral", and "face negative" categories, which reflect faces with emotional expressions[7].

Our objective is to automatically obtain emoji clusters from real usage in order to further recommend them. We can compare our work with the one from Pohl *et al.* [17] which tackled hierarchical relations between emojis of any kind. In the contrary to our work, they did not tried to obtain clusters inside emojis specific types to enhance emoji recommendation systems, which is our main purpose by applying our methodology and obtaining these new resources for further emoji prediction models.

[5] `http://www.weibo.com/`

[6] `http://unicode.org/emoji/charts/full-emoji-list.html`

[7] As a first step, we did not considered animal faces

3 Emoji Embeddings

The most used emojis are those representing emotions or sentiments, according to their usage in social media such as Twitter[8]: 😂 💓 😍 ❤ 😭.

Considering this fact, we want to verify if face emoji usages implicitly follow existing face expression categorizations by observing the usage of the 63 face emojis, excluding cat 😸, demon 😈, alien 👽 emojis and so on 🙍🏽. We make an emoji embedding in order to obtain a more fine-grained categorization for these emojis. We exclude the very recent emojis which are not in our dataset such as the exploding face 🤯.

3.1 Dataset

Our dataset is made of 695 031 tweets emitted from the North American continent (United States and Canada), all of them containing at least one of the 800 emojis from our list, and all collected using the Twitter streaming API[9]. There is no topic filtering so all kind of topics are included. The dataset is composed of tweets in english using a language detection process made with the occurrence ratio of NLTK stopwords list [10]. Table 1 shows quantitative information on the dataset.

In the dataset, we consider emojis as words. Thus, we keep them as tokens. To ensure they are used as tokens we tokenized the text and separate each emoji from other word. Then we applied lemmatization to the words using NLTK. Hence, we obtained a corpus of tweets ready for applying unsupervised algorithms.

Table 1. Dataset of tweets containing emojis

Tweets	695 031
Emojis	901 669
Average tweet length	10.81 words
Distinct Emojis	844
Emoji/tweets	1.30

3.2 Embeddings

To embed tweets with their emojis we used two approaches using Word2Vec [18,14] in its gensim implementation[11]. We used a Continuous Bag-Of-Words

[8] http://www.emojitracker.com/
[9] https://dev.twitter.com/streaming/overview
[10] http://www.nltk.org/
[11] https://radimrehurek.com/gensim/models/word2vec.html

(CBOW) embedding to predict target (emoji) from context words (tweet) with hierarchical softmax [13], and another embedding using Skip-grams to predict a context using an emoji. For comparison, we also used the skip-grams embedding model from Pohl *et al.* [17]. The resulting vector spaces are made of 300 dimensions from words with a minimum of 5 occurrences. These different models are then used to compare their impact on the clustering (Section 4).

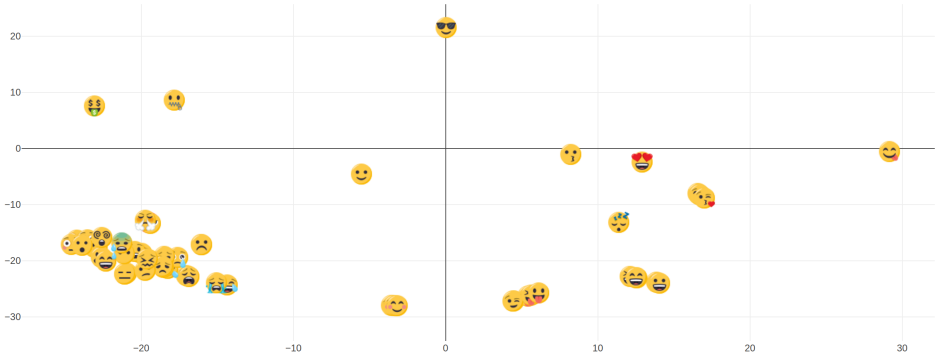


Fig. 1. Embeddings of 63 face emojis.

To display the dimensional space we selected the 63 emojis we are focusing on. To display a 300-dimensional space, we used the T-distributed Stochastic Neighbor Embedding (TSNE)[12] [12] in its Scikit-Learn implementation[13] to reduce this high dimensional space to a 2 dimensions space. The distribution of the emojis in the resulting 2 dimensions space is visible in Figure 1.

The complete 2 dimension visualization of the embedding space are available as supplementary materials and show different groups. However, the visualization can be quite different depending on the TSNE parameters, making the visualization not reliable to induce emoji groups. For the TSNE parameters we used a learning rate of 100, a perplexity of 30, and an early exageration of 2.0. Other parameters are the default ones from its Scikit-learn implementation[13]. Another approach would consist in defining an arbitrary threshold for the cosine distance to create clusters. To avoid arbitrary decisions, or group seection based on visual procimity, we decided to apply clustering on the produced emoji embeddings (Section 4), without assuming a number of clusters.

[12] `https://lvdmaaten.github.io/tsne/`

[13] `http://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html`

4 From Embeddings to Emoji Categories

4.1 Clustering

In the previous section we chose an emoji vector space of dimension 300. However, it does not gives clusters. As we wanted to automatically cluster face emojis from real usage data, we need to avoid any human interaction in the process. Hence, instead of directly using the cosine similarity we applied two clustering algorithms to automatically identify emoji clusters.

Unlike existing emoji clusterings in which the clusters were constructed based on the text and the emojis [3,17], we applied clustering only on emoji vectors, even if they have been embbeded using raw text. Plus, as we consider a sub type of emojis, this allows us to automatically retrieve good clusters directly from source, without mixing emoji types.

Clustering algorithms applied to emoji embeddings. We first applied the k-means algorithm to compute cluster centroids from the dataset considering n clusters $= n$ labels. So we did not assume a predefined number of clusters, each element could be contained into only one cluster. The k-means parameters were 63 possible clusters (for 63 emojis) with a maximum of 500 iterations and a number of 1000 initialisations. At the end, we removed the empty clusters and we finally obtained a reduced number of 18 clusters.

In order to compare the results with another algorithm, we used spectral clustering algorithm [15] considering 63 possible clusters. The hyper parameters were as follow: no eigenvalue decomposition strategy, a gaussian kernel, a gamma of 0.7, and discretization to assign label in order to differ from k-means. At the end, spectral clustering also gave us a total of 18 clusters.

K-means and spectral clustering results were close but we decided to keep the 18 clusters from spectral clustering because it avoided spliting intuitive emoji clusters such as kisses emojis 😙 😘 😗 😚. The different clusters are provided in the suplementary materials of the paper[14].

Using a skip-gram emoji embedding model we obtained 11 coarse clusters[15]. The resulting clusters are not statisfying since they mixed several emojis representing different emotions: anger, love, and tiring faces being in the same cluster for instance. With CBOW embeddings we obtained 18 emoji clusters[16], being more specific and fine-grained. Some of these clusters are visible in Table 2.

Comparison with existing related work. In order to compare the resulting clustering depending on the emoji embeddings, we also used Pohl *et al.* [17] embedding model, learnt using skip-gram, to extract face emoji clusters the same way we did using our embeddings. We used the same methodology, the desired number of clusters being equal to the number of elements, *i.e.* 63. The resulting clusters are not satisfying and can be seen in Figure 3. There were merely 6 coarse clusters being merely global separation between joy, anger, surprise and sadness. Considering the coarse clusters obtained using both skip-gram models and the

[14] See the html files in the "visualization" folder.

[15] "clusters_native63_skipgram.html" in supplementary materials of the paper.

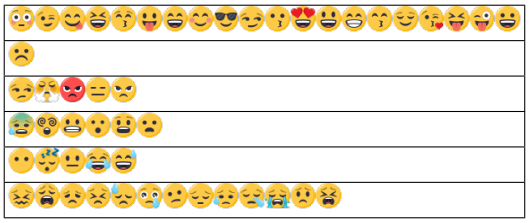
[16] "clusters_native63_cbow.html" in supplementary materials of the paper.

Table 2. Emoji clusters using Spectral Clustering on CBOW embeddings. Labelled using Ekman's categories of facial emotion expressions.

None (sleep)	**None (closed mouth)**
Contentment (Kisses)	**Sadness**
Contentment	**Shame**
Excitement	**None (not clear)**
Anger	**Amusement**
None (Greed)	**Sensory Pleasure**
Fear / Surprise	**Contentment / Amusement**
Satisfaction / Pride in Achievement	**None (Not happy)**
Contempt	**Excitement (Love)**

fine-grained clusters obtained using a CBOW embedding model, we came to the conclusion that CBOW embeddings are better to obtain more specific clusters of emojis. Hence, unlike existing emoji embeddings all using skip-gram models, this approach is better to retrieve latent information about real emoji usage in short messages.

Table 3. Clusters obtained using Pohl *et al.* embeddings



Because we consider the face emojis as our study material in this paper, we tried to obtain labels for these clusters considering existing theory (Section 4.2).

4.2 Cluster Validation Through Ekman's Theory

As the clusters represent face expressions, we can consider we did an face expression categorization of text messages. To verify if this categorization represents known emotions and evaluating their quality, we decided to labellize these clusters by Ekman's 16 basic face expressions of emotions [5] in order to compare and find the mistakes. For this purpose, we took each cluster automatically obtained and linked it to one of Ekman's categories by identifying the similarities on the emojis faces and the Ekman's facial expressions. For instance, the sadness category is represented by 😞😢😪😔😭😣😓😖😫.

The labelled clusters are described in Figure 2. Some categories are splitted by intensity, such as the joy wich have two clusters automatically extracted: one for the mild contentment 😄😃, and another one for a more intensive contentment 😁😆😊. Also, some Ekman's emotion categories can overlap in those clusters, such as fear/surprise.

Moreover, only the two unrealistic emojis and the sleeping emoji were alone in their clusters: 🤑, 😶 and 😴.

Note that the label attributed to the clusters may be discussed. However, the objective is not to find the precise label for each cluster but to identify different clusters of emoji. A comparison with the work of Ekman enabled us to relate emotional face expressions of humans to virtual facial expressions of emojis. This comparison shows that emojis are somewhat used the same way the face expressions of emotions are. However, some specific categories are inherent of emojis, such as 😶 and 🤑.

5 Conclusion and Perspectives

In conclusion, in this paper we presented two resources: an novel emoji embedding in real context in the scope of face emojis, and a set of face emoji automatic clusters from real usage only. Both can be used to improve emoji recommendation systems: instead of recommending specific emojis, groups of emojis (corresponding to clusters) will be used as elements to recommend. Recommending clusters of face emojis will positively impact the recommendation quality, as face emojis are some of the most used emojis. Moreover, the methodology we used to obtain these resources can be reproduced on different types of emojis to identify inherent categories from their usage. For instance, vehicule emoji unsupervised categorization could be done.

The methodology and resources can be used to recommend the emotion categories to express by an embodied conversational agent or in general dialog system, such as trending chatbots. With this work we want to change how to tackle emoji prediction by trying to generalize more, not only by parameter tuning, but also by changing the scope of the recommendation. Of course, because we focused on automatically retrieving emotion clusters of face emojis, theses resources could be helpful out of the scope of recommendation. For instance, embodied conversational agent could use them to determine which face expression is relevant to which emotion, and how to reproduce them without having to

necessarily regroup them from theories. Making the conversational agent more adaptative.

Finally, these resources with their visualisations and the python code used to produce them are available as supplementary materials. We plan to make them available to everyone in ORTOLANG[17], a platform that is part of CLARIN infrastructure[18]. The code and links to the models are available in the following repository: `https://github.com/gguibon/FaceEmojis`.

References

1. Ai, W., Lu, X., Liu, X., Wang, N., Huang, G., Mei, Q.: Untangling emoji popularity through semantic embeddings. In: ICWSM. pp. 2–11 (2017)
2. Barbieri, F., Ballesteros, M., Saggion, H.: Are emojis predictable? arXiv preprint arXiv:1702.07285 (2017)
3. 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 (May 2016)
4. Eisner, B., Rocktäschel, T., Augenstein, I., Bošnjak, M., Riedel, S.: emoji2vec: Learning emoji representations from their description. arXiv preprint arXiv:1609.08359 (2016)
5. Ekman, P.: Basic emotions in t. dalgleish and t. power (eds.) the handbook of cognition and emotion pp. 45-60 (1999)
6. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524 (2017)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. 9(8), 1735–1780 (Nov 1997)
8. Jakobson, R.: Closing statements: Linguistics and poetics, style in langage. TA Sebeok, New York (1960)
9. Kelly, C.: Do you know what i mean >:(: A linguistic study of the understanding of emoticons and emojis in text messages (2015)
10. Kelly, R., Watts, L.: Characterising the inventive appropriation of emoji as relationally meaningful in mediated close personal relationships. Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design (2015)
11. Li, J., Luong, M.T., Jurafsky, D.: A hierarchical neural autoencoder for paragraphs and documents. arXiv preprint arXiv:1506.01057 (2015)
12. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9(Nov), 2579–2605 (2008)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
15. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in neural information processing systems. pp. 849–856 (2002)

[17] `https://www.ortolang.fr/`

[18] `https://www.clarin.eu/`

16. Pavalanathan, U., Eisenstein, J.: Emoticons vs. emojis on twitter: A causal inference approach. arXiv preprint arXiv:1510.08480 (2015)
17. Pohl, H., Domin, C., Rohs, M.: Beyond just text: Semantic emoji similarity modeling to support expressive communication. ACM Transactions on Computer-Human Interaction (TOCHI) 24(1), 6 (2017)
18. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Citeseer (2010)
19. Wijeratne, S., Balasuriya, L., Sheth, A.P., Doran, D.: Emojinet: An open service and api for emoji sense discovery. In: ICWSM. pp. 437–447 (2017)
20. Xie, R., Liu, Z., Yan, R., Sun, M.: Neural emoji recommendation in dialogue systems. arXiv preprint arXiv:1612.04609 (2016)