Exploring Emoji Usage and Prediction
Through a Temporal Variation Lens

Francesco Barbieri∗ ♣ Luís Marujo♥ Pradeep Karuturi♥
William Brendel♣ Horacio Saggion♣
♣ Large Scale Text Understanding Systems Lab, TALN, UPF, Barcelona, Spain
♥ Snap Inc. Research, Venice, California, USA

Abstract
The frequent use of Emojis on social media platforms has created a new form of multimodal social interaction. Developing methods for the study and representation of emoji semantics helps to improve future multimodal communication systems. In this paper we explore the usage and semantics of emojis over time. We compare emoji embeddings trained on a corpus of different seasons and show that some emojis are used differently depending on the time of the year. Moreover, we propose a method to take into account the time information for emoji prediction systems, outperforming state-of-the-art systems. We show that, using the time information, the accuracy of some emojis can be significantly improved.

1 Introduction
Emojis are frequently used on social media (Snapchat, Twitter, Facebook, Instagram) and on communication platforms (Whatsapp, Messenger). In turn, they create a new form of multimodal communication, wherein images are used to enrich standard text messages. Over the past few years, the interest in emoji research has increased with several studies which contributed to emoji semantics [BRS16, ERA+16, WBSD17a, WBSD17b, BCC18], sentiment analysis [NSSM15, HGS+17, KK17, RPG+18], automatic emoji prediction [BBS17, FMS+17] and multimodal systems [CMS15, CSG+18, BBRS18]. However, to the best of our knowledge, the temporal dimension of emojis has not been addressed in past research. In this paper we explore the temporal correlation between emoji usage and events during the year, and we show that temporal information helps disambiguate emoji meanings. For example, the four leaf clover emoji ♣ is usually associated with good luck wishes, while in March, the same emoji indicates parties and drinking, due to St. Patrick day. In addition, some emojis are naturally associated with specific seasons. (e.g., during Christmas and ♣ in Summer), or specific hours (e.g., ♣ by night, and ♣ in the morning). We show that considering temporal information helps predict emojis, including those that are not time-specific such as ♣ and ♣.

2 Datasets
We first collected a corpus $C_{us}$ of more than 100 million English tweets, posted only in the U.S.\(^1\) from October 2015 to November 2017, and retrieved via the Twitter API\(^2\). We then extracted two datasets out of $C_{us}$.

| Season | Most Frequent Emojis |
|--------|----------------------|
| Spring | 🌄 ☀️ 🍃 🍂 🍁 🍀 🍙 🍄 🍊 🍆 🍃 |
| Summer | ☀️ 🌃 ☀️ ☀️ ☀️ ☀️ ☀️ ☀️ ☀️ ☀️ ☀️ |
| Autumn | 🍂 🍁 🍃 🍄 🍎 🍊 🍀 🍚 🌃 🌃 |
| Winter | 🍀 🍆 🍙 🍃 🍃 🍁 🍂 🍂 🍃 🍂 |

\(^1\)To remove spatial and cultural influence on data [BKRS16].

\(^2\)https://dev.twitter.com/streaming/overview
Table 2: Nearest-Neighbour (NN) results. Season/pair combinations, and all indicates the number of emojis that are common in all the season NN.

| Emoji | Spr-Sum | Spr-Aut | Spr-Win | Sum-Aut | Sum-Win | Aut-Win | All |
|-------|---------|---------|---------|---------|---------|---------|-----|
|       | 9 9 9 9 9 9 10 10 10 10 8 8 8 8 8 | 10 10 10 10 9 9 9 9 9 9 8 8 8 8 8 | 9 9 9 9 9 9 9 9 9 9 8 8 8 8 8 | 9 9 9 9 9 9 9 9 9 9 8 8 8 8 8 | 9 9 9 9 9 10 9 9 9 9 8 8 8 8 8 | 9 9 9 9 9 9 9 9 9 9 8 8 8 8 8 | 9 9 9 9 9 9 9 9 9 9 8 8 8 8 8 |

Table 3: The 10 NN emojis of the pine emoji 🌲 computed with respect to the semantic space of the four seasons. Emojis on the left are the closest ones.

| Season  | 🌲 | 🌸 | 🍂 | 🍁 |
|---------|----|----|----|----|
| Spring  | 🌲 | 🍂 | 🍀 | 🍁 |
| Summer  | 🌸 | 🌺 | 🍂 | 🍀 |
| Autumn  | 🌽 | 🍂 | 🍀 | 🍁 |
| Winter  | 🍂 | 🍀 | 🍁 | 🍂 |

study the variation of emojis usage across different seasons (Section 3). Table 1 shows the 15 most frequent emojis of each season. We can see that while emojis including 😊, ❤️ and 😎 are always the most common, other emojis are season-specific: 😯 in Autumn, 🎉 and 🌼 in Winter, and ☀️ in Spring and Summer.

2.2 Large Scale Emoji Prediction Dataset

We retain from 🌡 tweets containing only one emoji, and only if that emoji belongs to the set of top 300 most frequently occurring emojis. The final dataset for emoji prediction is composed of 900,000 tweets, with 3,000 tweets per class. In previous work, we experimentally observed that using more than 3,000 tweets per class does not significantly improve the prediction accuracy.

3 Does the Emoji Semantic and Usage Change Over Seasons?

Emoji semantics are difficult to analyze due to the subjective nature of emojis meanings, especially when it comes to describing emotions. Nevertheless, we study emoji semantic by association, i.e. we describe an emoji with a set of semantically close emojis, or by emoji pair co-occurrence in the same tweet. To this extent, we train skip-gram word embeddings models [MSC13] on the four different subsets (Spring, Summer, Autumn and Winter) of our seasonal dataset. Each model embeds emojis within a high dimensional space (300 dimensions, 6 tokens window) where distance metrics translate to semantic closeness and co-occurrence. Following [BKRS16] we first evaluate emoji semantics by describing each emoji with its k-Nearest-Neighbours (k-NN) for each season. Secondly, for each model, we produce a correlation matrix that encodes the semantic correlation of pairs of emojis appearing in the same tweet. We then compare the four matrices to see if their correlation statistics is preserved across different seasons.

3.1 Analysis From Emoji k-NN Description

For each season, emojis are associated to their k-NN, with k=10. We then look at the overlap of these NN for each emoji among different models. This way, we can investigate if a specific emoji shares the same set of NN across distinct seasons and thus if that emoji preserves its meaning across seasons. We also measure the number of NN that overlap in all the seasons to find emojis with smaller meaning variation across the year. Results are shown in Table 2. In the top half of the table, we notice that emojis related to music, animal, sweets, and emotions are not influenced by seasonality, i.e. they have the same set of Nearest-Neighbours in the four season-specific vector models (10-NN overlap ≥ 8).

The emojis for which their NN set varies the most across seasons are the ones listed in the bottom half of Table 2. Many are some sport-related emojis including 🏀 and 🏃‍♂️ probably due to seasons of the year that are include more sports events that other periods. Also the emoji 🍃, used in a school/university graduation context, seems to change meaning across seasons. The Nearest-Neighbours of these emojis are party and heart-related emojis in Spring, while school-related emojis in Autumn.

Another season-dependant example is the pine emoji 🌲 (Table 3). The pine tree is associated with vegetation, camping and sunrise-related emoji in Spring and Summer, while in Autumn and Winter it co-occurs with Christmas-related emoji. The gift emoji 🎁 present a similar behaviour: in Spring and Summer the three nearest neighbors are 🎉, 🎄 and 🎅, while in Winter the three closest emojis are 🎄, 🎅, 🎁.

3.2 Analysis From Emoji Correlation

We evaluate how the semantics of pairs of emojis is preserved across different seasons. To this extent, we compute for each season a 300 × 300 correlation matrix, where the correlation of emoji i and j is encoded as the cosine similarity between their 300-D feature vectors extracted from the season model. We then compare
pairs of seasons by evaluating the Pearson’s correlation between their respective matrices. The most correlated matrices are Spring and Summer (0.871) while the lowest correlation is between Spring and Winter (0.837). However, all the matrices are highly correlated, suggesting that only a small subset of emojis have their semantic varying across seasons.

Table 4 shows for each pair of seasons the pairs of emojis with the highest difference in similarity across those seasons. The differences between Spring and Summer (first column) does not look as significant as the differences between other seasons. We can spot few interesting cases. For example the pair 🌸 🌷 is more correlated during Autumn and Winter than during Spring and Summer. This is due to a famous case of doping in sport occurred during that Autumn/Winter of 2016. The pair 🎯 🕊 characterizes track related competitions happening during Spring and Summer. Another interesting case is the gift emoji 🎁 that in Autumn and Winter relates to a Christmas gift, as it is highly correlated to 🎀 and 🍂, while in Spring and Summer it is mostly used as a birthday gift as it is associated to emojis like 🎂 and 🎉. The case of the pair 🎆 🎇 can relate to either mass-shooting events or could simply suggest that in Autumn students have a hard time with the beginning of the new school year. One of the emojis that seems to be used differently in Autumn and Winter is the Skull 🎃, likely due to the usage of this emoji during Halloween time.

4 How Does Temporal Information Help Emoji Prediction?

In this section we evaluate how temporal information can improve the accuracy of emoji prediction models. We use the same experimental settings as [BBS17], except we predict 300 emojis classes in instead of 20. We use temporal information as an input to the classifier in addition to the tweet. The date is encoded as a vector of three dimensions (month, week day, and hour). For each of these dimensions we create a look up table of vectors of size 10, 12, 7, and 24 respectively. In this way, we can learn vectors of each month, day of the week and hour, using them for the final classification. These vectors are concatenated all together and incorporated in two different ways in the base system: at an early stage or at late stage. The early stage consists in concatenating this date embeddings to the word representation (char+word embeddings) and pass them to the Word LSTM. The late incorporation consists in concatenating the date embeddings with the output vector of the word attention and make the final predictions. We only use one of the methods to include date without combining them.

4.1 Emoji Prediction Model

We start from the state-of-the-art emoji prediction classifier [BBS17], and built two different methods – early and late temporal signal fusion – to incorporate the date information. The two entry points for fusing temporal information are evaluated in Section 4.2.
Table 4: Similarity-matrix results for every combination of two seasons. We show the pairs of emojis with biggest difference among their season specific similarity values.

| Spring Summer | Spring Autumn | Spring Winter | Summer Autumn | Summer Winter | Autumn Winter |
|---------------|--------------|--------------|--------------|--------------|--------------|
|               | 0.03         | 0.41         | 0.11         | 0.67         | 0.2          |
|               | 0.47         | 0.13         | 0.31         | 0.7          | 0.12         |
|               | 0.23         | 0.55         | 0.12         | 0.5          | 0.27         |
|               | 0.5          | 0.18         | 0.07         | 0.45         | 0.73         |
|               | 0.2          | 0.51         | 0.3          | 0.66         | 0.85         |
|               | 0.14         | 0.45         | 0.46         | 0.08         | 0.67         |
|               | 0.46         | 0.15         | 0.23         | 0.58         | 0.3          |
|               | 0.29         | 0.59         | 0.03         | 0.37         | 0.53         |
|               | 0.17         | 0.46         | 0.27         | 0.59         | 0.66         |
|               | 0.07         | 0.36         | 0.28         | 0.57         | 0.21         |
|               | 0.48         | 0.76         | 0.66         | 0.34         | 0.37         |
|               | 0.35         | 0.38         | 0.85         | 0.53         | 0.27         |
|               | 0.29         | 0.31         | 0.23         | 0.51         | 0.16         |
|               | 0.21         | 0.23         | 0.15         | 0.43         | 0.69         |

Table 5: Results for the three models: whitout date (W/O), Early date fusion, and Late date fusion. Precision, Recall, F1, accuracy at 1, 5, 10, and Coverage Error.

|       | P   | R   | F1  | a@1 | a@5 | a@10 | CE  |
|-------|-----|-----|-----|-----|-----|------|-----|
| W/O   | 21.97 | 23.22 | 21.89 | 23.13 | 38.22 | 45.70 | 44.29 |
| Early | 22.10 | 23.43 | 22.06 | 23.33 | 38.55 | 46.29 | 42.59 |
| Late  | 21.83 | 23.00 | 21.63 | 22.91 | 37.85 | 45.62 | 43.91 |

Table 6: Single emoji F1 when using Early date information, compared to the model without date information (i.e. W/O). We show the emojis with the biggest improvement. The Late fusion model variant is not incorporated as it produces worst performances (see Table 5).

| Emoji                  | W/O | Early |
|------------------------|-----|-------|
|                       | 0.54 | 0.63 |
|                       | 0.4  | 0.11  |
|                       | 0.34 | 0.22  |
|                       | 0.40 | 0.41  |
|                       | 0.40 | 0.41  |
|                       | 0.36 | 0.47  |
|                       | 0.36 | 0.47  |
|                       | 0.30 | 0.41  |
|                       | 0.36 | 0.47  |
|                       | 0.36 | 0.47  |

5 Conclusions

In the best of our knowledge, this is the first study to investigate if and how temporal information affects the interpretation and prediction of emojis. We studied whether the semantics of emojis change over different seasons, comparing emoji embeddings trained on a corpus of different seasons (Spring, Summer, Autumn, Winter) and show that some emojis are used differently depending on the time of the year, for example 🌞, 🌃, and 🌒. Moreover, we proposed a method to take in account the date information for emoji prediction systems, slightly improving the state-of-the-art. We show that, using the date information, the accuracy of some emojis can be improved. Some of them are clearly time dependent (e.g., 🌞, 🌃). Others are not directly associated to time but time information helps to predict them ( 🌘, 🌘, and 🌘).

In the future we plan to study the semantics of emojis over the day (morning/night) or over the week (weekdays/weekend) and improve the date information modules, trying the two methods we proposed together (early+late).

Acknowledgments

Part of this work was done when Francesco B. interned at Snap Inc. Francesco B. and Horacio S. acknowledge 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

[BBRS18] Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. Multimodal emoji prediction. In Proceedings of NAACL: Short Papers, New Orleans, US, 2018. Association for Computational Linguistics.

[BB17] Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. Are emojis predictable? In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 105–111, Valencia, Spain, April 2017. Association for Computational Linguistics.

[BCC18] Francesco Barbieri and Jose Camacho-Collados. How Gender and Skin Tone Modifiers Affect Emoji Semantics in Twitter. In Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM 2018), New Orleans, LA, United States, 2018.

[BRS16] F. Barbieri, F. Ronzano, and H. Saggion. What does this emoji mean? a vector space skip-gram model for t.emojis. In LREC, 2016.

[CMS15] Spencer Cappallo, Thomas Mensink, and Cees GM Snoek. Image2emoji: Zero-shot emoji prediction for visual media. In Proceedings of the 23rd ACM international conference on Multimedia, pages 1311–1314. ACM, 2015.

[CSG+18] Spencer Cappallo, Stacey Svetlichnaya, Pierre Garrigues, Thomas Mensink, and Cees GM Snoek. The new modality: Emoji challenges in prediction, anticipation, and retrieval. arXiv preprint arXiv:1801.10253, 2018.

[ERA+16] B. Eisner, T. Rocktäschel, I. Augenstein, M.B., and S. Riedel. emoji2vec: Learning emoji representations from their description. CoRR, abs/1609.08359, 2016.

[FMS+17] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain represent. for detecting sentiment, emotion and sarcasm. In EMNLP, 2017.

[HGS+17] Tianran Hu, Han Guo, Hao Sun, Thuy-vy Thi Nguyen, and Jiebo Luo. Spice up your chat: The intentions and sentiment effects of using Emoji. arXiv preprint arXiv:1703.02860, 2017.

[KK17] Mayu Kimura and Marie Katsurai. Automatic construction of an emoji sentiment lexicon. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pages 1033–1036. ACM, 2017.

[LLM+15] W. Ling, T. Luís, L. Marujo, R.F. Astudillo, S. Amir, C. Dyer, A.W. Black, and I. Trancoso. Finding function in form: Compositional character models for open vocabulary word representation. EMNLP, 2015.

[MSC+13] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In NIPS, 2013.

[NSSM15] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. PloS one, 10(12):e0144296, 2015.

[RPG+18] David Rodrigues, Marilía Prada, Rui Gaspar, Margarida V Garrido, and Diniz Lopes. Lisbon emoji and emoticon database (leed): norms for emoji and emoticons in seven evaluative dimensions. Behavior research methods, 50(1):392–405, 2018.

[WBSD17a] S. Wijeratne, L. Balasuriya, A. Sheth, and D. Doran. A semantics-based measure of emoji similarity. Web Intelligence, 2017.

[WBSD17b] Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. EmojiNet: An open service and api for emoji sense discovery. International AAAI Conference on Web and Social Media (ICWSM 2017). Montreal, Canada, 2017.

[YYD+16] Z. Yang, D. Yang, C. Dyer, X. He, A.J. Smola, and E.H. Hovy. Hierarchical attention networks for document classification. In HLT-NAACL, 2016.