Interpretable Emoji Prediction via Label-Wise Attention LSTMs

Francesco Barbieri[♢] **Luis Espinosa-Anke**[♠] **Jose Camacho-Collados**[♠]
Steven Schockaert[♠] **Horacio Saggion**[♢]
[♢] Large Scale Text Understanding Systems Lab, TALN, UPF, Barcelona, Spain
[♠]School of Computer Science and Informatics, Cardiff University, UK

[♢]{francesco.barbieri,horacio.saggion}@upf.edu,
[♠]{espinosa-ankel,camachocolladosj,schockaerts1}@cardiff.ac.uk

Abstract

Human language has evolved towards newer forms of communication such as social media, where emojis (i.e., ideograms bearing a visual meaning) play a key role. While there is an increasing body of work aimed at the computational modeling of emoji semantics, there is currently little understanding about what makes a computational model represent or predict a given emoji in a certain way. In this paper we propose a label-wise attention mechanism with which we attempt to better understand the nuances underlying emoji prediction. In addition to advantages in terms of interpretability, we show that our proposed architecture improves over standard baselines in emoji prediction, and does particularly well when predicting infrequent emojis.

1 Introduction

Communication in social media differs from more standard linguistic interactions across a wide range of dimensions. Immediacy, short text length, the use of pseudowords like *#hashtags* or *@mentions*, and even metadata such as user information or geolocalization are essential components of social media messages. In addition, the use of *emojis*, small ideograms depicting objects, people and scenes (Cappallo et al., 2015), are becoming increasingly important for fully modeling the underlying semantics of a social media message, be it a product review, a *tweet* or an *Instagram* post. Emojis are the evolution of character-based emoticons (Pavalanathan and Eisenstein, 2015), and are extensively used, not only as sentiment carriers or boosters, but more importantly, to express ideas about a myriad of topics, e.g., mood (😘), food (🍴), sports (🏂) or scenery (⛱).

Emoji modeling and prediction is, therefore, an important problem towards the end goal of properly capturing the intended meaning of a social media message. In fact, emoji prediction, i.e., given a (usually short) message, predict its most likely associated emoji(s), may help to improve different NLP tasks (Novak et al., 2015), such as information retrieval, generation of emoji-enriched social media content or suggestion of emojis when writing text messages or sharing pictures online. It has furthermore proven to be useful for sentiment analysis, emotion recognition and irony detection (Felbo et al., 2017). The problem of emoji prediction, albeit recent, has already seen important developments. For example, Barbieri et al. (2017) describe an LSTM model which outperforms a logistic regression baseline based on word vector averaging, and even human judgement in some scenarios.

The above contributions, in addition to emoji similarity datasets (Barbieri et al., 2016; Wijeratne et al., 2017) or emoji sentiment lexicons (Novak et al., 2015; Wijeratne et al., 2016; Kimura and Katsurai, 2017; Rodrigues et al., 2018), have paved the way for better understanding the semantics of emojis. However, our understanding of what exactly the neural models for emoji prediction are capturing is currently very limited. What is a model prioritizing when associating a message with, for example, positive (😄), negative (😠) or patriotic (🇺🇸) intents? A natural way of assessing this would be to implement an attention mechanism over the hidden states of LSTM layers. Attentive architectures in NLP, in fact, have recently received substantial interest, mostly for sequence-to-sequence models (which are useful for machine translation, summarization or language modeling), and a myriad of modifications have been proposed, including additive (Bahdanau et al., 2015), multiplicative (Luong et al., 2015) or self (Lin et al., 2017) attention mechanisms.

However, standard attention mechanisms only tell us which text fragments are considered impor-

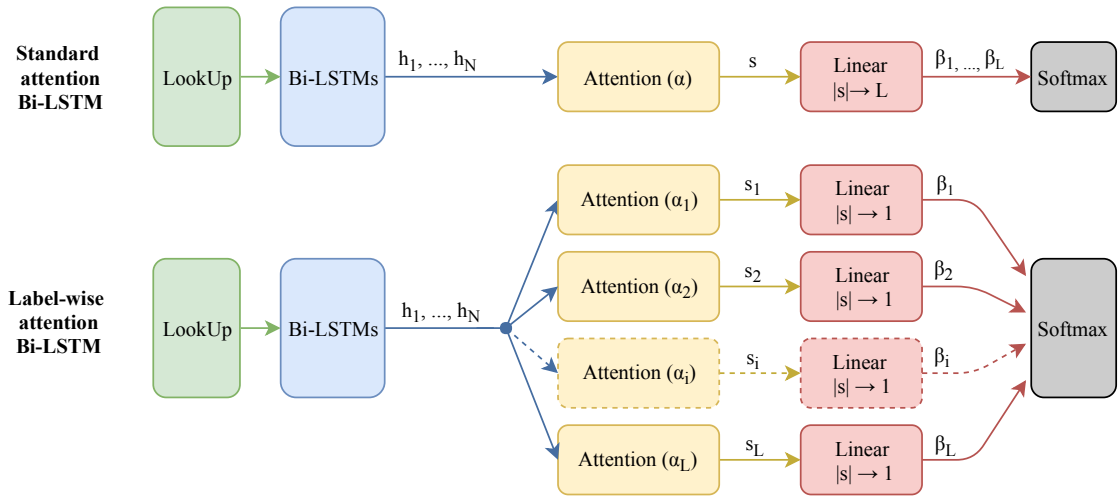

Figure 1: A classic attention network (top), and our attentive label-wise network (bottom), with a specific attention module for each label.

tant for the overall prediction distribution. While emoji prediction has predominantly been treated as a multi-class classification problem in the literature, it would be more informative to analyze which text fragments are considered important *for each individual emoji*. With this motivation in mind, in this paper we put forward a *label-wise* mechanism that operates over each label during training. The resulting architecture intuitively behaves like a batch of binary mini-classifiers, which make decisions over one single emoji at a time, but without the computational burden and risk of overfitting associated with learning separate LSTM-based classifiers for each emoji.

Our contribution in this paper is twofold. First, we use the proposed label-wise mechanism to analyze the behavior of neural emoji classifiers, exploiting the attention weights to uncover and interpret emoji usages. Second, we experimentally compare the effect of the label-wise mechanism on the performance of an emoji classifier. We observed a performance improvement over competitive baselines such as `FastText (FT)` (Joulin et al., 2017) and `Deepmoji` (Felbo et al., 2017), which is most noticeable in the case of infrequent emojis. This suggests that an attentive mechanism can be leveraged to make neural architectures more sensitive to instances of underrepresented classes.

2 Methodology

Our base architecture is the Deepmoji model (Felbo et al., 2017), which is based on two stacked word-based bi-directional LSTM recurrent neural networks with skip connections between the first and the second LSTM. The model also includes an attention module to increase its sensitivity to individual words during prediction. In general, attention mechanisms allow the model to focus on specific words of the input (Yang et al., 2016), instead of having to memorize all the important features in a fixed-length vector. The main architectural difference with respect to the typical attention is illustrated in Figure 1.

In Felbo et al. (2017), attention is computed as follows:

$$z_i = w_a h_i + b_a$$
$$\alpha_i = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}}$$
$$s = \sum_{j=1}^{N} \alpha_j h_j$$

Here $h_i \in \mathbb{R}^d$ is the hidden representation of the LSTM corresponding to the i^{th} word, with N the total number of words in the sentence. The weight vector $w_a \in \mathbb{R}^d$ and bias term $b_a \in \mathbb{R}$ map this hidden representation to a value that reflects the importance of this state for the considered classification problem. The values $z_1, ..., z_n$ are then normalized using a softmax function, yielding the attention weights α_i. The sentence representation s is defined as a weighted average of the vectors h_i. The final prediction distribution is then defined as follows:

$$\beta_l = w_{f,l} s + b_{f,l}$$

$$p_l = \frac{e^{\beta_l}}{\sum_{r=1}^{L} e^{\beta_r}}$$

where $w_{f,l} \in \mathbb{R}^d$ and $b_{f,l}$ define a label-specific linear transformation, with β_l reflecting our confidence in the l^{th} label and L is the total number of labels. The confidence scores β_l are then normalized to probabilities using another softmax operation. However, while the above design has contributed to better emoji prediction, in our case we are interested in understanding the contribution of the words of a sentence for each label (i.e., emoji), and not in the whole distribution of the target labels. To this end, we propose a label-wise attention mechanism. Specifically, we apply the same type of attention, but repeating it $|L|$ (number of labels) times, where each attention module is reserved for a specific label l:

$$z_{i,l} = w_{a,l} h_i + b_{a,l}$$
$$\alpha_{i,l} = \frac{e^{z_{i,l}}}{\sum_{j=1}^{N} e^{z_{j,l}}}$$
$$s_l = \sum_{j=1}^{N} \alpha_{j,l} h_j$$
$$\beta_l = w_{f,l} s_l + b_{f,l}$$
$$p_l = \frac{e^{\beta_l}}{\sum_{r=1}^{L} e^{\beta_r}}$$

3 Evaluation

This section describes the main experiment w.r.t the performance of our proposed attention mechanism, in comparison with existing emoji prediction systems. We use the data made available in the context of the SemEval 2018 Shared Task on Emoji Prediction (Barbieri et al., 2018). Given a tweet, the task consists of predicting an associated emoji from a predefined set of 20 emoji labels. We evaluate our model on the English split of the official task dataset. We also show results from additional experiments in which the label space ranged from 20 to 200 emojis. These *extended* experiments are performed on a corpus of around 100M tweets geolocalized in the United States and posted between October 2015 and May 2018.

Models. In order to put our proposed label-wise attention mechanism in context, we compare its performance with a set of baselines: (1) `FastText` (Joulin et al., 2017) (`FT`), which was the official baseline in the SemEval task; (2) 2 stacked Bi-LSTMs (`2-BiLSTMs`) without attention; and (3) 2 stacked Bi-LSTMs with standard attention (`2-BiLSTMs`$_a$) (Felbo et al., 2017). Finally, we denote as `2-BiLSTMs`$_l$ our proposed label-wise attentive Bi-LSTM architecture.

Lab	Syst	**F1**	**A@1**	**A@5**	**CE**
20*	`FastText`	30.97	42.57	72.45	4.56
	`2-BiLSTM`	33.52	45.76	75.54	3.88
	`2-BiLSTM`$_a$	**34.11**	**46.11**	75.68	3.86
	`2-BiLSTM`$_l$	33.51	45.94	**76.02**	**3.82**
50	`FastText`	18.04	22.33	48.13	14.27
	`2-BiLSTM`	19.07	25.35	53.38	9.37
	`2-BiLSTM`$_a$	19.83	25.52	53.51	9.35
	`2-BiLSTM`$_l$	**20.08**	**25.64**	**53.77**	**9.26**
100	`FastText`	16.25	20.29	42.65	26.04
	`2-BiLSTM`	17.44	**23.01**	**47.46**	15.24
	`2-BiLSTM`$_a$	17.56	22.77	46.93	15.51
	`2-BiLSTM`$_l$	**17.92**	22.80	47.41	**15.17**
200	`FastText`	13.31	18.80	38.99	51.06
	`2-BiLSTM`	16.16	21.05	42.64	24.68
	`2-BiLSTM`$_a$	16.30	21.13	42.50	24.60
	`2-BiLSTM`$_l$	**16.91**	**21.39**	**43.35**	**23.73**

Table 1: Experimental results of the two baselines, as well as single and label-wise attention modifications to the "vanilla" `2-BiLSTM` model.

Results. Table 1 shows the results of our model and the baselines in the emoji prediction task for the different evaluation splits. The evaluation metrics used are: **F1**, Accuracy@k (**A@k**, where $k \in \{1, 5\}$), and Coverage Error (**CE**[1]) (Tsoumakas et al., 2009). We note that the latter metric is not normally used in emoji prediction settings. However, with many emojis being "near synonyms" (in the sense of being often used almost interchangeably), it seems natural to evaluate the performance of an emoji prediction system in terms of how far we would need to go through the predicted emojis to recover the true label. The results show that our proposed `2-BiLSTMs`$_l$ method outperforms all baselines for F1 in three out of four settings, and for CE in all of them. In the following section we shed light on the reasons behind this performance, and we try to understand how these predictions were made.

[1] CE is computed as the average number of labels that need to be in the predictions for all true labels to be predicted.

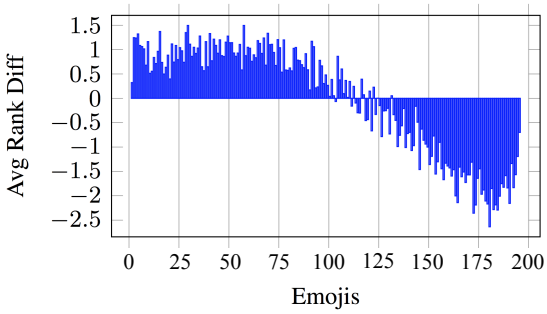

Figure 2: Difference in rank distributions. The x-axis represents emoji labels, ranked from most to least frequent. Lower scores indicate a higher average rank predicted by our proposed label-wise attention mechanism.

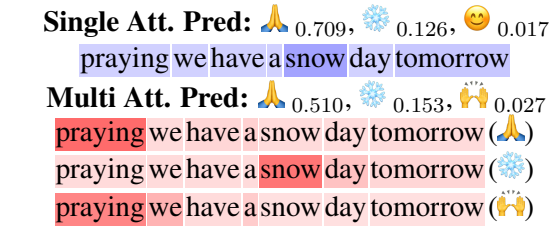

Figure 3: Attention weights α and α_l of single and label-wise attentive models. *Gold*: 🙏.

4 Analysis

By inspecting the predictions of our model, we found that the label-wise attention mechanism tends to be less heavily biased towards the most frequent emojis. This is reflected in the lower coverage error results in all settings, and becomes more noticeable as the number of labels grows. We verified this by computing the average difference between ranked predictions of the two attentive models in the 200-label setting (Figure 2). We can observe a sudden switch at more or less the median emoji, after which the label-wise attention model becomes increasingly accurate (relative to the standard attention model). This can be explained by the fact that infrequent emojis tend to be more situational (used in specific contexts and leaving less room for ambiguity or interchangeability), which the label-wise attention mechanism can take advantage of, as it explicitly links emojis with highly informative words. Let us illustrate this claim with a case in which the label-wise attention model predicts the correct emoji, unlike its single-attention counterpart:

> a friendship is built over time , but sisterhood is given automatically. *Gold:* 👯

For the above example[2], the predictions of the single attention model were all linked to the *general* meaning of the message, that is love and friendship, leading it to predict associated emojis (❤️, 💕 and 💗), failing to capture the most relevant bit of information. On the other hand, our proposed model "picks on" the word *sisterhood*, and with the added context of the surrounding words, ranks the gold label[3] in 4th position, which would be a true positive as per **A@5**.

Let us explore what we argue are interesting cases of emoji usage (ranging from highly explicit to figurative or situtational intent). Figure 3 shows how the word (*praying*) and emojis such as 🙏 and 🙌 are strongly correlated. In addition, the bond between the word *snow* and the ❄ emoji is also indisputable. However, a perhaps more surprising example is displayed in Figure 4, which is a negative example. Here, the ✓ emoji was predicted with rank 1, and we see it being strongly associated with the ordinal *second*, suggesting that the model assumed this was some kind of "ticked enumeration" of completed tasks, which is indeed regular practice in Twitter. Finally, we found it remarkable that the ambiguous nature of the word *boarding* is also reflected in two different emojis being predicted with high probability (❄ and ✈), each of them showcasing one of the word's senses.

As an additional exploratory analysis, we computed statistics on those words with the highest average attention weights associated with one single emoji. One interesting example is the 🎄 emoji, which shows two clear usage patterns: one literal (a tree) and one figurative (*christmas* and *holidays*). Finally, as a final (and perhaps thought-provoking) finding, the highest attention weights associated to the 🏈 emoji were given to the words *game*, *boys* and *football*, in that order. In other words, the model relies more on the word *boys* than on the actual description of the emoji. This is in line with a previous study that showed how the current usage of emojis in Twitter is in some cases associated with gender stereotypes (Barbieri and Camacho-Collados, 2018).

[2]The highlights show the α_l attention weights of 👯.

[3]Which is among the 10% most infrequent emojis in the dataset.

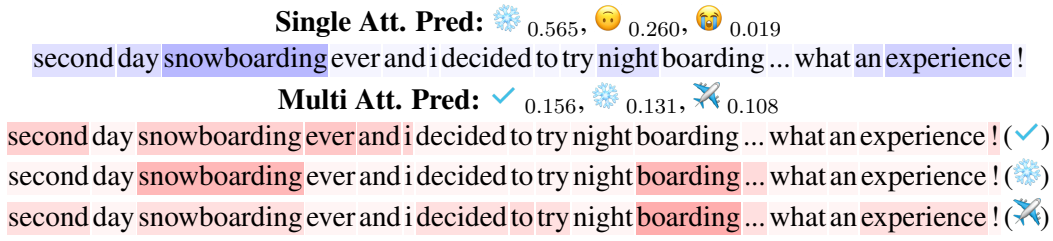

Figure 4: Attention weights α and α_l of single and label-wise attentive models. *Gold*: 😂.

5 Conclusion

In this paper we have presented a neural architecture for emoji prediction based on a label-wise attention mechanism, which, in addition to improving performance, provides a degree of interpretability about how different features are used for predictions, a topic of increasing interest in NLP (Linzen et al., 2016; Palangi et al., 2017). As we experimented with sets of emoji labels of different sizes, our proposed label-wise attention architecture proved especially well-suited for emojis which were infrequent in the training data, making the system less biased towards the most frequent. We see this as a first step to improve the robustness of recurrent neural networks in datasets with unbalanced distributions, as they were shown not to perform better than well-tuned SVMs on the emoji predicion task (Çöltekin and Rama, 2018).

As for future work, we plan to apply our label-wise attention mechanism to understand other interesting linguistic properties of human-generated text in social media, and other multi-class or multi-label classification problems.

Finally, code to reproduce our experiments and additional examples of label-wise attention weights from input tweets can be downloaded at `https://fvancesco.github.io/label_wise_attention/`.

Acknowledgments

F. Barbieri and H. Saggion acknowledge support from the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE). Luis Espinosa-Anke, Jose Camacho-Collados and Steven Schockaert have been supported by ERC Starting Grant 637277.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

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

Francesco Barbieri and Jose Camacho-Collados. 2018. How gender and skin tone modifiers affect emoji semantics in twitter. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 101–106. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. SemEval-2018 Task 2: Multilingual Emoji Prediction. In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.

Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016. What does this emoji mean? a vector space skip-gram model for Twitter emojis. In *Proc. of LREC 2016*.

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

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-oslo at semeval-2018 task 2: Svms perform better than rnns in emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 32–36, New Orleans, Louisiana. Association for Computational Linguistics.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *European Chapter of the Association for Computational Linguistics*, Valencia, Spain.

Mayu Kimura and Marie Katsurai. 2017. 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.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation.

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

Hamid Palangi, Paul Smolensky, Xiaodong He, and Li Deng. 2017. Deep learning of grammatically-interpretable representations through question-answering. *arXiv preprint arXiv:1705.08432*.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Emoticons vs. emojis on Twitter: A causal inference approach. *arXiv preprint arXiv:1510.08480*.

David Rodrigues, Marília Prada, Rui Gaspar, Margarida V Garrido, and Diniz Lopes. 2018. Lisbon emoji and emoticon database (leed): Norms for emoji and emoticons in seven evaluative dimensions. *Behavior research methods*, pages 392–405.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2009. Mining multi-label data. In *Data mining and knowledge discovery handbook*, pages 667–685. Springer.

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2016. Emojinet: Building a machine readable sense inventory for emoji. In *International Conference on Social Informatics*, pages 527–541. Springer.

Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017. A semantics-based measure of emoji similarity. In *Proceedings of the International Conference on Web Intelligence*, pages 646–653. ACM.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.