

Multimodal Emoji Prediction

Francesco Barbieri[◇] Miguel Ballesteros[♠] Francesco Ronzano[♡] Horacio Saggion[◇]

[◇] Large Scale Text Understanding Systems Lab, TALN, UPF, Barcelona, Spain

[♠] IBM Research, U.S

[♡] Integrative Biomedical Informatics Group, GRIB, IMIM-UPF, Barcelona, Spain

^{◇♡}{name.surname}@upf.edu, [♠]miguel.ballesteros@ibm.com

Abstract

Emojis are small images that are commonly included in social media text messages. The combination of visual and textual content in the same message builds up a modern way of communication, that automatic systems are not used to deal with. In this paper we extend recent advances in emoji prediction by putting forward a multimodal approach that is able to predict emojis in Instagram posts. Instagram posts are composed of pictures together with texts which sometimes include emojis. We show that these emojis can be predicted by using the text, but also using the picture. Our main finding is that incorporating the two synergistic modalities, in a combined model, improves accuracy in an emoji prediction task. This result demonstrates that these two modalities (text and images) encode different information on the use of emojis and therefore can complement each other.

1 Introduction

In the past few years the use of emojis in social media has increased exponentially, changing the way we communicate. The combination of visual and textual content poses new challenges for information systems which need not only to deal with the semantics of text but also that of images. Recent work (Barbieri et al., 2017) has shown that textual information can be used to predict emojis associated to text. In this paper we show that in the current context of multimodal communication where texts and images are combined in social networks, visual information should be combined with texts in order to obtain more accurate emoji-prediction models.

We explore the use of emojis in the social media platform Instagram. We put forward a multimodal approach to predict the emojis associated to an In-

stagram post, given its picture and text¹. Our task and experimental framework are similar to (Barbieri et al., 2017), however, we use different data (Instagram instead of Twitter) and, in addition, we rely on images to improve the selection of the most likely emojis to associate to a post. We show that a multimodal approach (textual and visual content of the posts) increases the emoji prediction accuracy compared to the one that only uses textual information. This suggests that textual and visual content embed different but complementary features of the use of emojis.

In general, an effective approach to predict the emoji to be associated to a piece of content may help to improve natural language processing tasks (Novak et al., 2015), such as information retrieval, generation of emoji-enriched social media content, suggestion of emojis when writing text messages or sharing pictures online. Given that emojis may also mislead humans (Miller et al., 2017), the automated prediction of emojis may help to achieve better language understanding. As a consequence, by modeling the semantics of emojis, we can improve highly-subjective tasks like sentiment analysis, emotion recognition and irony detection (Felbo et al., 2017).

2 Dataset and Task

Dataset: We gathered Instagram posts published between July 2016 and October 2016, and geolocalized in the United States of America. We considered only posts that contained a photo together with the related user description of at least 4 words and exactly one emoji.

Moreover, as done by Barbieri et al. (2017), we considered only the posts which include *one and only one* of the 20 most frequent emojis (the

¹In this paper we only utilize the first comment issued by the user who posted the picture.

most frequent emojis are shown in Table 3). Our dataset is composed of 299,809 posts, each containing a picture, the text associated to it and only one emoji. In the experiments we also considered the subsets of the 10 (238,646 posts) and 5 most frequent emojis (184,044 posts) (similarly to the approach followed by Barbieri et al. (2017)).

Task: We extend the experimental scheme of Barbieri et al. (2017), by considering also visual information when modeling posts. We cast the emoji prediction problem as a classification task: given an image or a text (or both inputs in the multimodal scenario) we select the most likely emoji that could be added to (thus used to label) such contents. The task for our machine learning models is, given the visual and textual content of a post, to predict the single emoji that appears in the input comment.

3 Models

We present and motivate the models that we use to predict an emoji given an Instagram post composed by a picture and the associated comment.

3.1 ResNets

Deep Residual Networks (ResNets) (He et al., 2016) are Convolutional Neural Networks which were competitive in several image classification tasks (Russakovsky et al., 2015; Lin et al., 2014) and showed to be one of the best CNN architectures for image recognition. ResNet is a feed-forward CNN that exploits “residual learning”, by bypassing two or more convolution layers (like similar previous approaches (Sermanet and LeCun, 2011)). We use an implementation of the original ResNet where the scale and aspect ratio augmentation are from (Szegedy et al., 2015), the photometric distortions from (Howard, 2013) and weight decay is applied to all weights and biases (instead of only weights of the convolution layers). The network we used is composed of 101 layers (ResNet-101), initialized with pretrained parameters learned on ImageNet (Deng et al., 2009). We use this model as a starting point to later finetune it on our emoji classification task. Learning rate was set to 0.0001 and we early stopped the training when there was not improving in the validation set.

3.2 FastText

FastText (Joulin et al., 2017) is a linear model for text classification. We decided to employ FastText as it has been shown that on specific classification tasks, it can achieve competitive results, comparable to complex neural classifiers (RNNs and CNNs), while being much faster. FastText represents a valid approach when dealing with social media content classification, where huge amounts of data needs to be processed and new and relevant information is continuously generated. The FastText algorithm is similar to the CBOW algorithm (Mikolov et al., 2013), where the middle word is replaced by the label, in our case the emoji. Given a set of N documents, the loss that the model attempts to minimize is the negative log-likelihood over the labels (in our case, the emojis):

$$loss = -\frac{1}{N} \sum_{n=1}^{n=N} e_n \log(\text{softmax}(BAx_n))$$

where e_n is the emoji included in the n -th Instagram post, represented as hot vector, and used as label. A and B are affine transformations (weight matrices), and x_n is the unit vector of the bag of features of the n -th document (comment). The bag of features is the average of the input words, represented as vectors with a look-up table.

3.3 B-LSTM Baseline

Barbieri et al. (2017) propose a recurrent neural network approach for the emoji prediction task. We use this model as baseline, to verify whether FastText achieves comparable performance. They used a Bidirectional LSTM with character representation of the words (Ling et al., 2015; Ballesteros et al., 2015) to handle orthographic variants (or even spelling errors) of the same word that occur in social media (e.g. *coooooool* vs *cool*).

4 Experiments and Evaluation

In order to study the relation between Instagram posts and emojis, we performed two different experiments. In the first experiment (Section 4.2) we compare the FastText model with the state of the art on emoji classification (B-LSTM) by Barbieri et al. (2017). Our second experiment (Section 4.3) evaluates the visual (ResNet) and textual (FastText) models on the emoji prediction task. Moreover, we evaluate a multimodal combination of both models respectively based on visual and

	top-5			top-10			top-20		
	P	R	F1	P	R	F1	P	R	F1
BW	61	61	61	45	45	45	34	36	32
BC	63	63	63	48	47	47	42	39	34
FT	61	62	61	47	49	46	38	39	36

Table 1: Comparison of B-LSTM with word modeling (BW), B-LSTM with character modeling (BC), and FastText (FT) on the same Twitter emoji prediction tasks proposed by Barbieri et al. (2017), using the same Twitter dataset.

textual inputs. Finally we discuss the contribution of each modality to the prediction task.

We use 80% of our dataset (introduced in Section 2) for training, 10% to tune our models, and 10% for testing (selecting the sets randomly).

4.1 Feature Extraction and Classifier

To model visual features we first finetune the ResNet (process described in Section 3.1) on the emoji prediction task, then extract the vectors from the input of the last fully connected layer (before the softmax). The textual embeddings are the bag of features shown in Section 3.2 (the x_n vectors), extracted after training the FastText model on the emoji prediction task.

With respect to the combination of textual and visual modalities, we adopt a middle fusion approach (Kiela and Clark, 2015): we associate to each Instagram post a multimodal embedding obtained by concatenating the unimodal representations of the same post (i.e. the visual and textual embeddings), previously learned. Then, we feed a classifier² with visual (ResNet), textual (FastText), or multimodal feature embeddings, and test the accuracy of the three systems.

4.2 B-LSTM / FastText Comparison

To compare the FastText model with the word and character based B-LSTMs presented by Barbieri et al. (2017), we consider the same three emoji prediction tasks they proposed: top-5, top-10 and top-20 emojis most frequently used in their Tweet datasets. In this comparison we used the same Twitter datasets. As we can see in Table 1 FastText model is competitive, and it is also able to outperform the character based B-LSTM in one of the emoji prediction tasks (top-20 emojis). This result suggests that we can employ FastText to represent Social Media short text (such as Twitter or Instagram) with reasonable accuracy.

²L2 regularized logistic regression

	top-5			top-10			top-20		
	P	R	F1	P	R	F1	P	R	F1
Maj	7.9	20.0	11.3	2.7	10.0	4.2	0.9	5.0	1.5
W.R.	20.1	20.0	20.1	9.8	9.8	9.8	4.6	4.8	4.7
Vis	38.6	31.1	31.0	26.3	20.9	20.5	20.3	17.5	16.1
Tex	56.1	54.4	54.9	41.6	37.5	38.3	36.7	29.9	31.3
Mul	57.4	56.3	56.7	42.3	40.5	41.1	36.6	35.2	35.5
%	2.3	3.5	3.3	1.7	8	7.3	-0.3	17.7	13.4

Table 2: Prediction results of top-5, top-10 and top-20 most frequent emojis in the Instagram dataset: Precision (P), Recall (R), F-measure (F1). Experimental settings: majority baseline, weighted random, visual, textual and multimodal systems. In the last line we report the percentage improvement of the multimodal over the textual system.



4.3 Multimodal Emoji Prediction



We present the results of the three emoji classification tasks, using the visual, textual and multimodal features (see Table 2).


The emoji prediction task seems difficult by just using the image of the Instagram post (**Visual**), even if it largely outperforms the majority baseline³ and weighted random⁴. We achieve better performances when we use feature embeddings extracted from the text. The most interesting finding is that when we use a multimodal combination of visual and textual features, we get a non-negligible improvement. This suggests that these two modalities embed different representations of the posts, and when used in combination they are synergistic. It is also interesting to note that the more emojis to predict, the higher improvement the multimodal system provides over the text only system (3.28% for top-5 emojis, 7.31% for top-10 emojis, and 13.42 for the top-20 emojis task).

4.4 Qualitative Analysis

In Table 3 we show the results for each class in the top-20 emojis task.

The emoji with highest F1 using the textual features is the most frequent one  (0.62) and the US flag  (0.52). The latter seems easy to predict since it appears in specific contexts: when the word USA/America is used (or when American cities are referred, like #NYC).

The hardest emojis to predict by the text only system are the two gestures  (0.12) and  (0.13). The first one is often selected when the gold stan-

³Always predict  since it is the most frequent emoji.

⁴Random keeping labels distribution of the training set

E	%	Tex	Vis	MM	E	%	Tex	Vis	MM
❤️	17.46	0.62	0.35	0.69	💙	3.68	0.22	0.15	0.29
😂	9.10	0.45	0.30	0.47	🙌	3.55	0.20	0.02	0.26
😍	8.41	0.32	0.15	0.34	😜	3.54	0.13	0.02	0.2
💕	5.91	0.23	0.08	0.26	🎯	3.51	0.26	0.17	0.31
🌟	5.73	0.35	0.17	0.36	💪	3.31	0.43	0.25	0.45
🔥	4.58	0.45	0.24	0.46	😊	3.25	0.12	0.01	0.16
🇺🇸	4.31	0.52	0.23	0.53	👍	3.14	0.12	0.02	0.15
☀️	4.15	0.38	0.26	0.49	🙏	3.11	0.34	0.11	0.36
😎	3.84	0.19	0.1	0.22	🎉	2.91	0.36	0.04	0.37
👏	3.73	0.13	0.03	0.16	👻	2.82	0.45	0.54	0.59

Table 3: F-measure in the test set of the 20 most frequent emojis using the three different models. “%” indicates the percentage of the class in the test set

standard emoji is the second one or 😍👏 is often mispredicted by wrongly selecting 😂 or 😎.

Another relevant confusion scenario related to emoji prediction has been spotted by Barbieri et al. (2017): relying on Twitter textual data they showed that the emoji ❤️ was hard to predict as it was used similarly to 🍷. Instead when we consider Instagram data, the emoji 🍷 is easier to predict (0.23), even if it is often confused with 😍.

When we rely on visual contents (Instagram picture), the emojis which are easily predicted are the ones in which the associated photos are similar. For instance, most of the pictures associated to 🐶 are dog/pet pictures. Similarly, ☀️ is predicted along with very bright pictures taken outside. 💪 is correctly predicted along with pictures related to gym and fitness. The accuracy of 🍷 is also high since most posts including this emoji are related to fitness (and the pictures are simply either selfies at the gym, weight lifting images, or protein food).

Employing a multimodal approach improves performance. This means that the two modalities are somehow complementary, and adding visual information helps to solve potential ambiguities that arise when relying only on textual content. In Figure 1 we report the confusion matrix of the multimodal model. The emojis are plotted from the most frequent to the least, and we can see that the model tends to mispredict emojis selecting more frequent emojis (the left part of the matrix is brighter).

4.4.1 Saliency Maps

In order to show the parts of the image most relevant for each class we analyze the global average pooling (Lin et al., 2013) on the convolutional

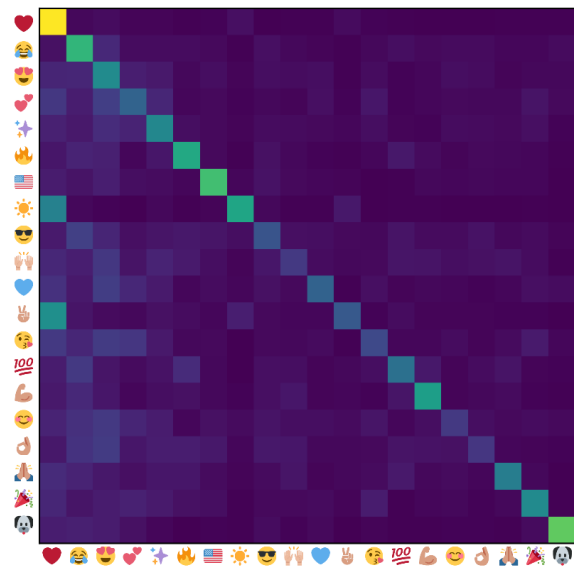


Figure 1: Confusion matrix of the multimodal model. The gold labels are plotted as y-axes and the predicted labels as x-axes. The matrix is normalized by rows.

feature maps (Zhou et al., 2016). By visually observing the image heatmaps of the set of Instagram post pictures we note that in most cases it is quite difficult to determine a clear association between the emoji used by the user and some particular portion of the image. Detecting the correct emoji given an image is harder than a simple object recognition task, as the emoji choice depends on subjective emotions of the user who posted the image. In Figure 2 we show the first four predictions of the CNN for three pictures, and where the network focuses (in red). We can see that in the first example the network selects the smile with sunglasses 😎 because of the legs in the bottom of the image, the dog emoji 🐶 is selected while focusing on the dog in the image, and the smiling emoji 😊 while focusing on the person in the back, who is lying on a hammock. In the second example the network selects again the 😊 due to the water and part of the kayak, the heart emoji ❤️ focusing on the city landscape, and the praying emoji 🙏 focusing on the sky. The same “praying” emoji is also selected when focusing on the luxury car in the third example, probably because the same emoji is used to express desire, i.e. “please, I want this awesome car”.

It is interesting to note that images can give context to textual messages like in the following Instagram posts: (1) “Love my new home ☀️” (associated to a picture of a bright garden, outside) and (2) “I can’t believe it’s the first day of school!!!



Figure 2: Three test pictures. From left to right, we show the four most likely predicted emojis and their correspondent class activation mapping heatmap.

I love being these boys’ mommy!!!! #myboys #mommy ❤️” (associated to picture of two boys wearing two blue shirts). In both examples the textual system predicts ❤️. While the multimodal system correctly predicts both of them: the blue color in the picture associated to (2) helps to change the color of the heart, and the sunny/bright picture of the garden in (1) helps to correctly predict ☀️.

5 Related Work

Modeling the semantics of emojis, and their applications, is a relatively novel research problem with direct applications in any social media task. Since emojis do not have a clear grammar, it is not clear their role in text messages. Emojis are considered function words or even affective markers (Na’aman et al., 2017), that can potentially affect the overall semantics of a message (Donato and Paggio, 2017).

Emojis can encode different meanings, and they can be interpreted differently. Emoji interpretation has been explored user-wise (Miller et al., 2017), location-wise, specifically in countries (Barbieri et al., 2016b) and cities (Barbieri et al., 2016a), and gender-wise (Chen et al., 2017) and time-wise (Barbieri et al., 2018).

Emoji semantics and usage have been studied with distributional semantics, with models trained on Twitter data (Barbieri et al., 2016c), Twitter data together with the official unicode description (Eisner et al., 2016), or using text from a popular keyboard app Ai et al. (2017). In the same

context, Wijeratne et al. (2017a) propose a platform for exploring emoji semantics. In order to further study emoji semantics, two datasets with pairwise emoji similarity, with human annotations, have been proposed: EmoTwi50 (Barbieri et al., 2016c) and EmoSim508 (Wijeratne et al., 2017b). Emoji similarity has been also used for proposing efficient keyboard emoji organization (Pohl et al., 2017). Recently, Barbieri and Camacho-Collados (2018) show that emoji modifiers (skin tones and gender) can affect the semantics vector representation of emojis.

Emoji play an important role in the emotional content of a message. Several sentiment lexicons for emojis have been proposed (Novak et al., 2015; Kimura and Katsurai, 2017; Rodrigues et al., 2018) and also studies in the context of emotion and emojis have been published recently (Wood and Ruder, 2016; Hu et al., 2017).

During the last decade several studies have shown how sentiment analysis improves when we jointly leverage information coming from different modalities (e.g. text, images, audio, video) (Morency et al., 2011; Poria et al., 2015; Tran and Cambria, 2018). In particular, when we deal with Social Media posts, the presence of both textual and visual content has promoted a number of investigations on sentiment or emotions (Bacchi et al., 2016; You et al., 2016b,a; Yu et al., 2016; Chen et al., 2015) or emojis (Cappallo et al., 2015, 2018).

6 Conclusions

In this work we explored the use of emojis in a multimodal context (Instagram posts). We have shown that using a synergistic approach, thus relying on both textual and visual contents of social media posts, we can outperform state of the art unimodal approaches (based only on textual contents). As future work, we plan to extend our models by considering the prediction of more than one emoji per Social Media post and also considering a bigger number of labels.

Acknowledgments

We thank the anonymous reviewers for their important suggestions. 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

- Wei Ai, Xuan Lu, Xuanzhe Liu, Ning Wang, Gang Huang, and Qiaozhu Mei. 2017. Untangling emoji popularity through semantic embeddings. In *ICWSM*. pages 2–11.
- Claudio Baecchi, Tiberio Uricchio, Marco Bertini, and Alberto Del Bimbo. 2016. A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications* 75(5):2507–2525.
- Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2015. Improved transition-based parsing by modeling characters instead of words with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 349–359.
- Francesco Barbieri, Miguel Ballesteros, and Horacio Saggion. 2017. Are emojis predictable? In *Proceedings of the 2017 Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain.
- Francesco Barbieri and Jose Camacho-Collados. 2018. 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.
- Francesco Barbieri, Luis Espinosa-Anke, and Horacio Saggion. 2016a. Revealing patterns of Twitter emoji usage in Barcelona and Madrid. *Frontiers in Artificial Intelligence and Applications*. 2016;(Artificial Intelligence Research and Development) 288: 239–44. .
- Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016b. How Cosmopolitan Are Emojis? Exploring Emojis Usage and Meaning over Different Languages with Distributional Semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, Amsterdam, Netherlands, pages 531–535.
- Francesco Barbieri, Luis Marujo, William Brendel, Pradeep Karaturim, and Horacio Saggion. 2018. Exploring Emoji Usage and Prediction Through a Temporal Variation Lens. In *1st International Workshop on Emoji Understanding and Applications in Social Media (at ICWSM 2018)*.
- Francesco Barbieri, Francesco Ronzano, and Horacio Saggion. 2016c. What does this emoji mean? a vector space skip-gram model for twitter emojis. In *LREC*.
- 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*. ACM, pages 1311–1314.
- Spencer Cappallo, Stacey Svetlichnaya, Pierre Garrigues, Thomas Mensink, and Cees GM Snoek. 2018. The new modality: Emoji challenges in prediction, anticipation, and retrieval. *arXiv preprint arXiv:1801.10253* .
- Fuhai Chen, Yue Gao, Donglin Cao, and Rongrong Ji. 2015. Multimodal hypergraph learning for microblog sentiment prediction. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, pages 1–6.
- Zhenpeng Chen, Xuan Lu, Sheng Shen, Wei Ai, Xuanzhe Liu, and Qiaozhu Mei. 2017. Through a gender lens: An empirical study of emoji usage over large-scale android users. *arXiv preprint arXiv:1705.05546* .
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
- Giulia Donato and Patrizia Paggio. 2017. Investigating redundancy in emoji use: Study on a twitter based corpus. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. pages 118–126.
- Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359* .
- 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. *EMNLP* .
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 770–778.
- Andrew G Howard. 2013. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402* .
- Tianran Hu, Han Guo, Hao Sun, Thuy-vy Thi Nguyen, and Jiebo Luo. 2017. Spice up Your Chat: The Intentions and Sentiment Effects of Using Emoji. In *In Proceeding of the International AAAI Conference on Web and Social Media (ICWSM)*. AAAI.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 2017 Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Valencia, Spain.
- Douwe Kiela and Stephen Clark. 2015. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *EMNLP*. pages 2461–2470.

- 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*. ACM, pages 1033–1036.
- Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in Network. In *International Conference on Learning Representations*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, pages 740–755.
- Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramon Fernandez, Silvio Amir, Luis Marujo, and Tiago Luis. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 1520–1530.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Hannah Miller, Daniel Kluver, Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. 2017. Understanding emoji ambiguity in context: The role of text in emoji-related miscommunication. In *11th International Conference on Web and Social Media, ICWSM 2017*. AAAI Press.
- Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, pages 169–176.
- Noa Na’aman, Hannah Provenza, and Orion Montoya. 2017. Varying linguistic purposes of emoji in (twitter) context. In *Proceedings of ACL 2017, Student Research Workshop*. pages 136–141.
- Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS one* 10(12):e0144296.
- Henning Pohl, Christian Domin, and Michael Rohs. 2017. Beyond just text: Semantic emoji similarity modeling to support expressive communication. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24(1):6.
- Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. pages 2539–2544.
- 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.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.
- Pierre Sermanet and Yann LeCun. 2011. Traffic sign recognition with multi-scale convolutional networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, pages 2809–2813.
- Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 1–9.
- Ha-Nguyen Tran and Erik Cambria. 2018. Ensemble application of elm and gpu for real-time multimodal sentiment analysis. *Memetic Computing* 10(1):3–13.
- Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017a. Emojinet: An open service and api for emoji sense discovery. *International AAAI Conference on Web and Social Media (ICWSM 2017)*. Montreal, Canada .
- Sanjaya Wijeratne, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2017b. A semantics-based measure of emoji similarity. *International Conference on Web Intelligence (Web Intelligence 2017)*. Leipzig, Germany .
- Ian Wood and Sebastian Ruder. 2016. Emoji as emotion tags for tweets. In *Emotion and Sentiment Analysis Workshop, LREC*.
- Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. 2016a. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, pages 1008–1017.
- Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016b. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. ACM, pages 13–22.
- Yuhai Yu, Hongfei Lin, Jiana Meng, and Zhehuan Zhao. 2016. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. *Algorithms* 9(2):41.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pages 2921–2929.