Overview of the EVALITA 2018 Italian Emoji Prediction (ITAMoji) Task

Francesco Ronzano
Universitat Pompeu Fabra, Spain
Hospital del Mar Medical Research Center
Barcelona, Spain
francesco.ronzano@upf.edu

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

Endang Wahyu Pamungkas,Viviana Patti
Department of Computer Science
University of Turin, Italy
{pamungka,patti}@di.unito.it

Francesca Chiusaroli
Department of Humanities
Università di Macerata, Italy
f.chiusaroli@unimc.it

Abstract

English. The Italian Emoji Prediction task (ITAmoji) is proposed at EVALITA 2018 evaluation campaign for the first time, after the success of the twin Multilingual Emoji Prediction Task, organized in the context of SemEval-2018 in order to challenge the research community to automatically model the semantics of emojis in Twitter. Participants were invited to submit systems designed to predict, given an Italian tweet, its most likely associated emoji, selected in a wide and heterogeneous emoji space. Twelve runs were submitted at ITAmoji by five teams. We present the data sets, the evaluation methodology including different metrics and the approaches of the participating systems. We also present a comparison between the performance of automatic systems and humans solving the same task. Data and further information about this task can be found at: https://sites.google.com/view/itamoji/.

Italiano. *Il task italiano per la predizione degli emoji in Twitter (ITAmoji) viene proposto nell'ambito della campagna di valutazione di Evalita 2018 per la prima volta, dopo il successo del task gemello, il* Multilingual Emoji Prediction Task*, proposto a Semeval-2018 per stimolare la comunità di ricerca a costruire modelli computazionali della semantica delle emoji in Twitter. I partecipanti sono stati invitati a costruire sistemi disegnati per predire l'emoji piú probabile dato un tweet in italiano, selezionandola in uno spazio ampio e eterogeneo di emoji. In ITAmoji sono stati valutati i risultati di dodici sistemi di predizione di emoji messi a punto da cinque gruppi di lavoro. Presentiamo qui i dataset, la metodologia di valutazione (che include diverse metriche) e gli approcci dei sistemi che hanno partecipato. Presentiamo inoltre una riflessione sui risultati ottenuti in tale task da sistemi automatici e umani.*

1 Introduction

During the last decade the use of emoji has increasingly pervaded social media platforms by providing users with a rich set of pictograms useful to visually complement and enrich the expressiveness of short text messages. Nowadays this novel, visual way of communication represents a *de facto* standard in a wide range of social media platforms including fully-fledged portals for user-generated contents like Twitter, Facebook and Instagram as well as instant-messaging services like WhatsApp. As a consequence, the possibility to effectively interpret and model the semantics of emojis has become an essential task to deal with when we analyze social media contents.

Even if over the last few years the study of this new form of language has been receiving a growing attention, at present the body of investigations that deal with emojis is still scarce, especially when we consider their characterization from a Natural Language Processing (NLP) perspective. While there are notable exceptions which study the semantics of emojis and their usage (Barbieri et al., 2016a; Barbieri et al., 2018b; Aoki and Uchida, 2011; Eisner et al., 2016; Ljubešić and Fišer, 2016), reflecting also on their informative behaviour (Donato and Paggio, 2017; Donato and Paggio, 2018), or their sentiment (Novak et al., 2015), the interplay between text-based mes-

sages and emojis remains still explored only by a small number of studies. Among these investigations there is the analysis of emoji predictability by (Barbieri et al., 2017), which proposed a neural model to predict the most likely emoji to appear in a text message (tweet). The task resulted to be hard, as emojis encode multiple meanings (Barbieri et al., 2016b). Related to this, in the context of the International Workshop on Semantic Evaluation (SemEVAL 2018), the Multilingual Emoji Prediction Task (Barbieri et al., 2018a) has been organized in order to challenge the research community to automatically model the semantics of emojis occurring in English and Spanish Twitter messages. The task was very successful, with 49 teams participating in the English subtask and 22 in the Spanish subtask. This motivated us to propose the shared task also for the Italian language in the context of the Evalita 2018 evaluation campaign (Caselli et al., 2018), with the twofold aim to widen the setting for cross-language comparisons for emoji prediction in Twitter and to experiment with novel metrics to better assess the quality of the automatic predictions.

In general, exciting and highly relevant avenues for research are still to explore with respect to emoji understanding, since emojis represent often an essential component of social media texts: ignoring or misinterpreting them may lead to misunderstandings in comprehending the intended meaning of a message (Miller et al., 2016). The ambiguity of emojis raises also interesting questions in application domains, think for instance to a human-computer interaction setting: how can we teach an artificial agent to correctly interpret and recognize emojis' use in spontaneous conversation? The main motivation behind this question is that an AI system able to predict emojis could contribute notably to better natural language understanding (Novak et al., 2015) and thus to other Natural Language Processing tasks such as generating emoji-enriched social media content, enhancing emotion/sentiment analysis systems, improving retrieval of social network material, and ultimately improving user profiling.

In the following, we describe the main elements of the shared task (Section 3), after proposing a brief summary about previous projects reflecting on the semantics of emojis in Italian (Section 2). Then, we cover the data collection, curation and release process (Section 4). In Section 5 we detail the evaluation metrics, we describe the participants results and we propose a first comparison with performances of humans solving the same task. We conclude the paper with some reflections on the outcomes of the proposed task.

2 Emojis and Italian

We can observe a growing interest on the semantics of emojis in relation with Italian. In particular, some recent interesting projects have been carried out in the last years, which address the issue in a translation framework, investigating the possibility to translate from Italian literary texts into the universal visual language of emoji (Chiusaroli, 2015; Monti et al., 2016). In particular, the *Emojitaliano project* was launched as a translation project of the Italian novel *Pinocchio* in emoji (Chiusaroli, 2017) on Twitter. An original approach based on crowdsourcing was adopted, by involving for the translation task the Twitter community named as *Scritture Brevi*.

The Twitter community #scritturebrevi The community (#scritturebrevi, @FChiusaroli, 10,151 followers in November 2018) had previously been involved in experiments of creative writing, also in emojis: with the hashtag #inemoticon, on Twitter, experiments of mixed translation - words and emojis – have been carried out, experiencing the semantic versatility of emojis, and their values in rebus writings. Translating the whole *Pinocchio* book was a more complex and engaging task, especially for its focus on developing a common code base, in terms of glossary and grammar, which is absolute new with respect to previous projects. The translation of Pinocchio started on February 2016. Everyday, for 28 weeks, sentences taken from Pinocchio were tweeted, and the followers were invited to suggest their translations to emoji; at the end of each day, the official version of the translation was validated and published. An online tool *Emojiitalianobot* has been developed in order to support the community to memorize the semantic values assigned to each emoji during the collective translation process. Since its first beginning on Twitter, the project was an instant success, becoming a viral web phenomenon thanks to the *Scritture brevi* community. Therefore, it was a natural choice to involve the same Twitter community to reflect on the semantics of emoji from a different perspective, i.e. the one we

propose in the context of the ITAmoji shared task, thus helping us to understand how humans are good at predicting emojis (see Section 5.5.2).

3 Task Description

We invited participants to submit systems designed to predict, given a tweet in Italian, its most likely associated emoji, only based on the text of the tweet. As for the experimental setting, for simplicity purposes, we considered tweets including only one emoji (eventually repeated). After removing the emoji from the tweet, we asked users to predict it. We challenged systems to predict

> Innamorato sempre di più 😍 [URL]

Figure 1: Example of tweet with an emoji at the end, considered in the emoji prediction task.

emojis among a wide and heterogeneous emoji space. In particular, we selected the tweets that included one of the twenty five emojis that occur most frequently in the Twitter data we collected (see Table 1). Therefore, the task can be seen as a multi-class classification task where systems should predict one of 25 possible emojis from the text of a given tweet. Each participant was allowed to submit up to three system runs. Participants were allowed to use additional data to train the systems such as lexicons and pre-trained word embeddings. In order to have the possibility to perform a finer grained evaluation of results, we encouraged participants to submit, for each tweet, not only the most likely emoji predicted but also the complete *rank* from the most likely to the less likely emoji to be associated to the text of the tweet.

4 Task Data

The data for this task were retrieved from Twitter by experimenting with two different approaches: (i) gathering Twitter stream on (geolocalized) Italian tweets from October 2015 to February 2018; and (ii) retrieving tweets from the followers of the most popular Italian newspaper's accounts. We randomly selected $275,000$ tweets from these collections by choosing tweets that contained one and only one emoji over 25 most frequent emojis listed in Table 1. We split our data into two sets consisting of $250,000$ *training* samples and $25,000$ *test* samples.

Emoji	% Tweet in Train and Test set
❤️	20.27
😂	19.86
😍	9.45
😉	5.35
😊	5.13
😁	4.11
😀	3.54
😘	3.33
😎	2.80
👍	2.57
💯	2.18
🤔	2.16
💙	2.03
😜	1.94
😱	1.78
💪	1.67
😋	1.55
😅	1.52
😭	1.49
🔝	1.39
💕	1.37
💥	1.28
💋	1.12
✨	1.07
🌹	1.06

Table 1: The distribution (percentage) for each emoji in the train and test set

5 Evaluation

In this section we present the evaluation setting for the ITAmoji shared task.

5.1 Metrics

The evaluation of the emoji prediction systems has been based on the classic *precision* and *recall* metrics over each emoji. The final ranking of the participating teams of ITAmoji 2018 relies on the **Macro F1** score computed with respect to the most likely emoji predicted, given the text of each tweet of the test set, in line with the proposal in the twin task at Semeval 2018 for English and Spanish (Barbieri et al., 2018a). In this way we intend to

encourage systems to perform well overall, which would inherently mean a better sensitivity to the use of emojis in general, rather than for instance overfitting a model to do well in the three or four most common emojis of the test data.

In general, the identification of a coherent and effective approach to compare the performance of distinct emoji prediction systems is not an easy task. We have often the clear impression that the semantics of some sets of emojis can be similar, therefore it would be interesting to have a way to compare and evaluate at a finer grained level the emoji prediction quality of two distinct systems, when they both fail in predicting the right emoji to associate to a tweet. In such cases, indeed, it can be important to distinguish between the system that identifies the right prediction among the most likely emojis to be associated to that tweet and the one that characterizes the right prediction as an emoji that is unlikely to be associated to that tweet. In order to catch this aspect, we gave ITAmoji participants the possibility to submit as emoji predictions, the *ordered ranking* of the 25 emojis considered in ITAmoji. Systems providing the ranked list of emoji predictions were also compared by considering the following additional *emoji-rank-based metrics*: **Accuracy@5/10/15/20** and **Coverage Error**. All the submissions we received provided the ranked list of 25 emojis as predictions: as a consequence it was possible to compute the emoji-rank-based metrics considered for all of them.

A detailed description of all the evaluation metrics we considered to compare the quality of emoji prediction approaches is given below. The following three **standard metrics** are computed by considering only the emoji predicted as the most likely one to be associated to the text of a tweet:

- **Macro F1**: compute the F1 score for each label (emoji), and find their un-weighted mean (exploited to determine the final ranking of the participating teams);
- **Micro F1**: compute the F1 score globally by counting the total true positives, false negatives and false positives across all label (emojis);
- **Weighted F1**: compute the F1 score for each label (emoji), and find their average, weighted by support (the number of true instances for each label);

Regarding the **emoji-rank-based metrics**, we considered:

- **Coverage error**: compute how far we need to go through the ranked scores of labels (emojis) to cover all true labels;
- **Accuracy@n**: is the accuracy value computed by considering as right predictions the ones in which the right label (emoji) is among the top N most likely ones.

5.2 Baseline

In order to compare the performance of the ITAmoji participating systems with baseline approaches, we considered three different baselines:
- Majority baseline: for each text of a tweet we predict the ordered list of 25 most-likely emojis sorted by their frequency in the training set, that is, we always predict as first choice the red heart, and as last choice the rose emoji.
- Weighted random baseline: for each text of a tweet we predict the ordered list of the 25 most-likely emojis where the first prediction is randomly selected taking in consideration the label-frequency in the training set (in order to keep the same labels distribution) and the rest of the predictions (from the second to the last one) are generated by considering the rest of emojis sorted by label-frequency.
- FastText baseline: for each text of a tweet we predict the ordered list of the 25 most-likely emojis by relying on fasttext with basic parameters[1] and pretrained embeddings with 300 dimensions (Barbieri et al., 2016a).

5.3 Participating Systems and Results

We received 12 submissions in total from 5 different teams. The main approaches and features of participating teams are described below.
FBK_FLEXED_BICEPS (Andrei et al., 2018) This system exploit recurrent neural network architecture Bidirectional Long Short Term Memory (Bi-LSTM), together with user based features to deal with this task. They concatenate the output of Bi-LSTM network that take word sequence as input with the user history distribution in using emoji. Finally, the softmax activation is used to get the probability distribution of the 25 emoji labels.

[1] https://fasttext.cc/

GW2017 (Mauro and Xileny, 2018) This system based on ensemble of two models, Bi-LSTM and LightGBM[2]. The first model uses two different word2vec models based on the time creation, while the second model exploits several surfaces feature extracted from tweet text (e.g., number of words, number of characters).

CIML-UNIPI (Daniele et al., 2018) This system is based on ensemble composed of 13 models (12 basen on TreeESNs and one on LSTM over characters. Models based on TreeESN are built by varying the number of reservoir units, activation function, readout and parser.

sentim (Jacob, 2018) This system relies on a convolutional neural network (CNN) architecture which uses character embedding as input. 9 layers of residual dilated convolutions with skip connections are applied, followed by a ReLU activation to increase nonlinearity.

UNIBA (Lucia and Daniela, 2018) This system is built by using ensemble classifier based on WEKA[3] and scikit-learn[4]. Several features are exploited by using micro-blogging based feature, sentiment based feature, and semantic based feature.

Table 2 shows the official results of ITAmoji 2018 task, ordered by decreasing Macro F1. The best performing system was proposed by the *FBK_FLEXED_BICEPS* team, which achieves 0.365312 in Macro F1. Overall, we can see that systems which exploit neural network architecture obtained good performances in this task, especially when relying on Bi-LSTM model. Table 3 shows the performance of ITAmoji systems with respect to emoji-rank-based metrics.

5.4 Analysis

From Table 2 we can notice that the ranking order of the 5 system runs that obtained the best Macro F1 is substantially preserved when we consider Micro F1 or Weighted F1. Anyway, with respect to Macro F1, when we consider Micro F1 the differences among the scores obtained by the top-performing systems tend to be substantially smaller: for instance the Macro F1 of the best system is greater by a factor of 1.64 with respect to the fifth system, while the Micro F1 of the best system is greater by a factor of 1.18 with respect to the fifth system (ranked by Micro F1). This fact can be motivated by the trend, when we consider Micro F1, to favour systems that tend to overfit their prediction model to do well in the most common emojis of the test data with respect to systems with good performances over all emojis: this fact confirms our choice to select Macro F1 as the official metric to rank ITAmoji 2018 participating systems.

From Table 3 we can see how the order to the top-5 best performing systems in terms of Macro F1 is substantially preserved when we consider the emoji-rank-based metrics Coverage Error and Accuracy@5 (except for the switch between the fourth and fifth best performing approach).

If we consider the performance of our three baseline systems (described in Section 5.2) we can notice from Table 2 that, as expected, FastText is the best performing baseline approach: a FastText embedding based prediction system would have ranked as eight by Macro F1 in ITAmoji 2018.

Table 6 shows the highest F1 score for each emoji / label across all ITAmoji 2018 team submissions. We can notice that even if specific emojis like 😋, 💋, 🌹, or 🌟 are characterized by a small percentage of training samples (about 1%), prediction systems manage to obtain high Macro F1 scores. In contrast, when we consider emojis like 🤔 or 👍, even if there are more training samples available with respect to the previous set of emojis (more than 2%), we observe that the prediction systems do not manage to get high Macro F1 scores. This fact can be explained by the variability of the context of use that characterizes the latter set of emojis that makes it difficult for system to learn to predict.

To conclude our analysis, we have to notice that the three runs that obtained the highest Macro F1 scores, to predict the emojis exploited, besides the text of a tweet, the way the author of that tweet used emojis in previous tweets. This fact highlights that the choice of an emoji strongly depends on the preferences and writing style of each individual, both representing relevant inputs to model in order to improve emoji prediction quality.

5.5 Emoji prediction by humans

In this section we present a preliminary discussion of the results of two experiments designed in order to evaluate how humans perform when they are requested to identify the most likely emoji(s) to associate to the text of an Italian tweet. The

[2] https://github.com/Microsoft/LightGBM
[3] https://www.cs.waikato.ac.nz/ml/weka/
[4] http://scikit-learn.org/stable/

Rank	Team	Run Name	Macro F1	Micro F1	Weighted F1
1	FBK_FLEXED_BICEPS	base_ud_1f	36.53	47.67	46.98
2	FBK_FLEXED_BICEPS	base_ud_10f	35.63	47.62	46.58
3	FBK_FLEXED_BICEPS	base_tr_10f	29.21	42.35	39.57
4	GW2017	gw2017_p	23.29	40.09	37.81
5	GW2017	gw2017_e	22.21	42.19	36.90
6	CIML-UNIPI	run1	19.24	29.12	31.48
7	CIML-UNIPI	run2	18.80	37.63	34.101
-	**FastText baseline**		11.96	28.72	27.02
8	sentim	Sentim_Test_Run_3	10.62	29.43	23.24
9	sentim	Sentim_Test_Run_2	10.23	31.27	23.11
-	**Weighted random baseline**		3.94	10.36	10.36
10	GW2017	gw2017_pe	3.75	11.95	10.97
11	UNIBA	itamoji_uniba_run1	3.19	27.38	15.61
12	sentim	Sentim_Test_Run_1	1.95	6.48	3.99
-	**Majority baseline**		1.35	20.28	6.84

Table 2: Official Results of ITAmoji Shared Task: evaluation metrics computed by considering only the emoji predicted as the most likely one to be associated to the text of a tweet. Teams runs are ranked by Macro F1. The table shows also the performance of the three baselines considered in ITAmoji, ranked with respect to their Macro F1.

Rank	Team	Run Name	Coverage Error	Accuracy@5 / 10 / 15 / 20
1	FBK_FLEXED_BICEPS	base_ud_1f	3.47	81.67 / 92.14 / 96.86 / 99.10
2	FBK_FLEXED_BICEPS	base_ud_10f	3.49	81.53 / 91.94 / 96.82 / 99.17
3	FBK_FLEXED_BICEPS	base_tr_10f	4.35	74.54 / 87.50 / 94.34 / 98.00
4	GW2017	gw2017_p	5.66	67.18 / 81.49 / 89.42 / 92.99
5	GW2017	gw2017_e	4.60	71.30 / 85.90 / 94.30 / 98.25
6	CIML-UNIPI	run1	5.43	64.60 / 83.02 / 93.00 / 98.01
7	CIML-UNIPI	run2	5.11	68.46 / 83.86 / 92.38 / 97.28
-	**FastText baseline**		7.23	59.07 / 74.22 / 82.58 / 88.89
8	sentim	Sentim_Test_Run_3	6.41	58.53 / 76.93 / 88.52 / 95.74
9	sentim	Sentim_Test_Run_2	6.33	57.60 / 77.17 / 89.70 / 96.41
-	**Weighted random baseline**		6.92	59.06 / 76.11 / 86.42 / 94.10
10	GW2017	gw2017_pe	13.49	27.93 / 43.04 / 56.00 / 66.27
11	UNIBA	itamoji_uniba_run1	6.70	58.78 / 75.97 / 86.36 / 93.53
12	sentim	Sentim_Test_Run_1	12.45	29.20 / 48.78 / 64.38 / 74.04
-	**Majority baseline**		6.63	60.07 / 76.43 / 86.51 / 94.12

Table 3: Official Results of ITAmoji Shared Task: emoji-rank-based metrics (Coverage error and Accuracy@n). Teams runs ranked by Macro F1. The table shows also the performance of the three baselines considered in ITAmoji, ranked with respect to their Macro F1.

final purpose here is to explore if humans are better than automated systems in the emoji prediction task from text, or viceversa. In an attempt to consider an uniform set of emojis in our experimental settings, in both human emoji prediction experiments described in the rest of this section we decided to focus only on the 15 emojis shown in Table 4. This group of emojis includes all the yellow-face emojis considered in the ITAmoji task (Table 1).

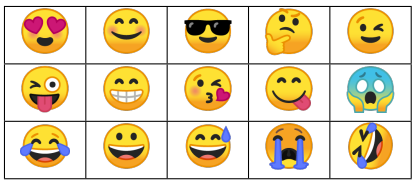

Table 4: The set of 15 face emoji considered in the human annotation experiments.

5.5.1 Figure 8 human annotation

We selected 1,005 tweets with one face-emojis from the ITAmoji test set and set up a collaborative annotation task in Figure Eight (F8)[5] by asking annotators to chose the first, second and third most likely face emoji they would associate to the text of each tweet [6]. The set of 1,005 tweets to annotate was perfectly balanced across the 15 face emojis considered. A total of 64 annotators from the F8

[5] https://www.figure-eight.com/

[6] Instructions provided to annotators (in Italian) here: http://bit.ly/itaMoji

platform provided 6,150 evaluations by spotting the 3 most likely face emojis to associate to the text of a tweet.

The Macro F1 of F8 annotators is 24.74. On the same set of 1,005 tweets, the emoji prediction performance of human annotators was better than 9 out of 12 systems submitted to ITAmoji. However, the the best performing system submitted to ITAmoji obtained a Macro F1 of 40.48 on those tweets, suggesting that computational models can perform better than humans in this task.

5.5.2 Twitter human annotation

Thanks to the support and collaboration of the #scritturebrevi Twitter community, we replicated the human annotation experiment carried out in F8 in a "crowdcourcing in the wild" setting. From the end of July to the beginning of September 2018, we posted 485 tweets on the Scritture Brevi Twitter account (*@FChiusaroli*), most of them selected from the same portion of the ITAmoji test set considered in our F8 experiment (see Section 5.5.1). Members of the Scritture Brevi Twitter community were called to participate to a sort of Twitter crowdsourcing game with slogan **#ITAmoji che passione** and hashtag #ITAmoji. Every day a set of tweets without emoji was posted on the Scritture Brevi Twitter account, and *ITAmojiers* had to post as a *reply* the most likely face emoj they would associate to the text of the posted tweet[7]. The game became viral. We managed to involve more than one hundred users with an average number of valid predictions/replies per tweet equal to 5.4. When the **#ITAmoji che passione** game ended, we were able to identify for each tweet posted on #scritturebrevi (485 tweets in total) the most-likely face-emoji that the Twitter community would associate. In general, the emoji prediction performance of people from Scritture Brevi Twitter community was better than 8 out of 12 systems submitted to ITAmoji (always on the same set of 485 tweets annotated by that community).

5.5.3 Comparing human and automated emoji predictions

In the two experiments just described, we asked humans to identify the face emoji(s) they would associate to the text of a tweet by exploiting different approaches to collect data: a controlled collaborative annotation environment in the case of F8 (Section 5.5.1) and a "crowdcourcing in the wild" setting in the case of the Scritture Brevi Twitter community (Section 5.5.2). In Table 5 we compare the emoji prediction performance of human annotators (from both F8 and Scritture Brevi Twitter community) with the performance of the emoji prediction systems submitted to ITAmoji. To perform this comparison we consider the set of 428 tweets of the ITAmoji test set annotated by F8 and the Scritture Brevi Twitter community.

We can notice that human predictions, both from F8 and Scritture Brevi, outperforms most of the automated systems. Moreover, F8 predictions obtain a Macro F1 (24.46) higher than Scritture Brevi Twitter community (22.94). This trend may be related to the fact that F8, in contrast to the #scritturebrevi Twitter community, represents a controlled annotation environment.

6 Conclusion

Considered the widespread diffusion of emojis as visual devices useful to provide an additional layer of meaning to social media messages, on one hand, and the unquestionable role of Twitter as one of the most important social media platforms, on the other, we proposed this year at Evalita 2018 ITAmoji, the Italian Emoji Prediction task.
Results of automated systems are in line with ones obtained in the twin shared task proposed for English and Spanish at Semeval 2018 (Barbieri et al., 2018a). The introduction of new experimental emoji-rank based metrics in ITAmoji allowed us to perform a finer-grained evaluation of the systems' emoji prediction quality. Moreover, comparing performances of humans and systems in the emoji prediction task confirms also in an Italian setting the outcomes of a similar experiment proposed for English (Barbieri et al., 2017), suggesting that computational models are able to better capture the underlying semantics of emojis.

[7]The announce of the "#ITAmoji che passione" game was published on the Scritture Brevi's blog and linked to every posted tweet: `https://www.scritturebrevi.it/2018/07/16/itamoji-che-passione/`

References

Catalin Coman Andrei, Nechaev Yaroslav, and Zara Giacomo. 2018. Predicting emoji exploiting multimodal data: Fbk participation in itamoji task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Team	Run Name	Macro F1	Micro F1	Weighted F1
FBK_FLEXED_BICEPS	base_ud_1f	35.70	34.81	35.94
FBK_FLEXED_BICEPS	base_tr_10f	35.03	34.81	35.36
FBK_FLEXED_BICEPS	base_ud_10f	34.73	34.11	34.83
Figure Eight predictions		24.46	26.40	24.57
CIML-UNIPI	run1	24.03	25.00	23.65
Scritture Brevi predictions		22.94	24.06	22.99
GW2017	gw2017_p	20.40	23.13	19.97
GW2017	gw2017_e	20.33	22.66	19.83
CIML-UNIPI	run2	19.45	21.26	18.80
sentim	Sentim_Test_Run_2	12.17	15.19	11.59
sentim	Sentim_Test_Run_3	11.07	14.49	10.82
GW2017	gw2017_pe	5.01	7.48	5.02
UNIBA	itamoji_uniba_run1	2.95	7.47	2.84
sentim	Sentim_Test_Run_1	2.74	4.90	2.83

Table 5: Performance of human (Scritture Brevi and Figure 8) and automated emoji prediction approaches, compared by considering the set of 428 tweets with face-emoji that are part of the ITAmoji test set and have been annotated by both Figure 8 platform and #scritturebrevi community. Emoji prediction approaches are ranked by decreasing Macro F1.

Emoji	Label	Macro F1	Num. Samples	% Samples
❤️	red heart	75.74	5069	20.28
😂	face with tears of joy	57.08	4966	19.86
💋	kiss mark	51.71	279	1.12
😋	face savoring food	48.34	387	1.55
🌹	rose	46.83	265	1.06
☀️	sun	44.69	319	1.28
😍	smiling face with heart eyes	42.93	2363	9.45
😘	face blowing a kiss	41.61	834	3.34
💙	blue heart	39.26	506	2.02
😊	smiling face with smiling eyes	38.92	1282	5.13
😀	grinning face	37.74	885	3.54
😉	winking face	34.98	1338	5.35
😁	beaming face with smiling eyes	34.47	1028	4.11
✨	sparkles	32.31	266	1.06
🤣	rolling on the floor laughing	31.79	546	2.18
👍	thumbs up	31.55	642	2.57
😎	smiling face with sunglasses	30.89	700	2.80
💪	flexed biceps	30.75	417	1.67
🤔	thinking face	29.06	541	2.16
💕	two hearts	27.48	341	1.36
😭	loudly crying face	25.62	373	1.49
🔝	top arrow	24.03	347	1.39
😅	grinning face with sweat	23.94	379	1.52
😜	winking face with tongue	23.66	483	1.93
😱	face screaming in fear	22.56	444	1.78

Table 6: Best F1 score for each emoji / label across all ITAmoji 2018 teams. The fourth and fifth columns respectively show, for each emoji, the number and percentage of test samples present in the test dataset.

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

Francesco Barbieri, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016a. How cosmopolitan are emojis?: Exploring emojis usage and meaning over different languages with distributional semantics. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 531–535. ACM.

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

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*, pages 105–111, Valencia, Spain, April. ACL.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018a. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33. ACL.

Francesco Barbieri, Luis Marujo, William Brendel, Pradeep Karuturim, and Horacio Saggion. 2018b. 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)*.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. Evalita 2018: Overview of the 6th evaluation campaign of natural language processing and speech tools for italian. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Francesca Chiusaroli. 2015. La scrittura in emoji tra dizionario e traduzione. In *Proceedings of 2nd Italian Conference on Computational Linguistics (CLiC-it 2015), Trento, Italy, December 3-4, 2015*. Aacademia University Press.

F. Chiusaroli. 2017. *Pinocchio in emojitaliano*. Apice Libri.

Di Sarli Daniele, Gallicchio Claudio, and Micheli Alessio. 2018. Itamoji 2018: Emoji prediction via tree echo state networks. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

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.

Giulia Donato and Patrizia Paggio. 2018. Classifying the Informative Behaviour of Emoji in Microblogs. In *Proc. of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. ELRA.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359*.

Anderson Jacob. 2018. Fully convolutional networks for text classification. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Nikola Ljubešić and Darja Fišer. 2016. A global analysis of emoji usage. In *Proceedings of the 10th Web as Corpus Workshop*, pages 82–89. Association for Computational Linguistics.

Siciliani Lucia and Girardi Daniela. 2018. The uniba system at the evalita 2018 italian emoji prediction task. In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Bennici Mauro and Seijas Portocarrero Xileny. 2018. The validity of word vectors over the time for the evalita 2018 emoji prediction task (itamoji). In Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso, editors, *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, Turin, Italy. CEUR.org.

Hannah Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. “Blissfully happy" or “ready to fight": Varying interpretations of emoji. *Proc. of ICWSM'16*.

Johanna Monti, Federico Sangati, Francesca Chiusaroli, Martin Benjamin, and Sina Mansour. 2016. Emojitalianobot and emojiworldbot - new online tools and digital environments for translation into emoji. In *Proc. CLiC-it 2016, Napoli, Italy, December 5-7, 2016.*, volume 1749 of *CEUR Workshop Proceedings*.

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