Analyzing Incorporation of Emotion in Emoji Prediction

Shirley Anugrah Hayati* and **Aldrian Obaja Muis***
Language Technologies Institute
Carnegie Mellon University
{shayati,amuis}@cs.cmu.edu

Abstract

In this work, we investigate the impact of incorporating emotion classes on the task of predicting emojis from Twitter texts. More specifically, we first show that there is a correlation between the emotion expressed in the text and the emoji choice of Twitter users. Based on this insight we propose a few simple methods to incorporate emotion information in traditional classifiers. Through automatic metrics, human evaluation, and error analysis, we show that the improvement obtained by incorporating emotion is significant and correlate better with human preferences compared to the baseline models. Through the human ratings that we obtained, we also argue for preference metric to better evaluate the usefulness of an emoji prediction system.

1 Introduction

Emoji is a set of pictograms that symbolize a lot of things from facial expressions to flags. Recently, research in emoji started to gain attention from Natural Language Processing (NLP) researchers due to its rising popularity in social media for users to express ideas, concepts, or emotion (Novak et al., 2015).

There has been some interest in tackling the task of emoji prediction (Barbieri et al., 2017, 2018a). Because of the rich expressiveness of emoji, understanding emojis will help other kinds of natural language understanding tasks such as sentiment analysis (Felbo et al., 2017) or generating or suggesting emoji for social media content (Novak et al., 2015).

Now, as noted by Wolny (2016), people use emojis to express diverse emotions. And intuitively we can see why certain emojis are used to convey certain emotions. For example, the 😭, which depicts "loudly crying face", seems highly correlated with the emotion of sadness. Figure 1 shows a tweet for which the user expresses their sadness about the event through the use of the emoji 😭. The individual words alone do not explicitly convey any sadness, but the readers will be able to get a sense of sadness from the tweet. In this case, a system that is able to recognize the emotion content of a tweet will be more likely to recommend emojis related to such emotion, hence providing better user experience.

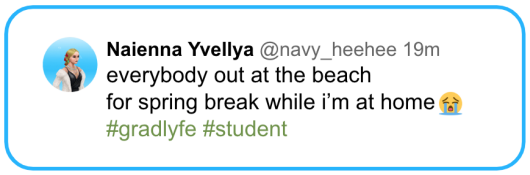

Figure 1: An example of a tweet with high emotional content (sadness) overall, while individual words do not really reflect any particular emotions.

Based on this intuition, in this work we aim to explore the incorporation of emotion content of a tweet to improve emoji prediction. Thus, the question that we would like to answer through this work is: "How can we make use of emotion content to guide emoji prediction models?"

Our contributions are as follows:

- We show more explicitly the link between certain emojis and certain emotions.
- We evaluate two simple methods to incorporate emotion information into an SVM model.
- We show, both through automatic and manual metrics, significant improvement of top emojis predicted by our emotion-aware models over the baseline models.
- We do an in-depth analysis of the dataset, the task, and give some recommendation for future directions.

* equal contributions

Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 91–99
Minneapolis, June 6, 2019.

- We release our crawled tweets the corpus containing human-rated tweet-emoji pairs for further analysis.[1]

2 Related Work

Barbieri et al. (2017) pioneered the task of emoji prediction by creating a dataset of 589,000 tweets containing a single mention of an emoji from the top-20 most frequent emojis. They also performed human evaluation by asking crowdworkers to give the emoji that best matches the tweet in a 5-emoji setting, and found that their systems are comparable to human performance in predicting emojis.

Cappallo et al. (2018) highlights the importance of having a balanced test set, in order to better evaluate the models' performance on rarer emojis. Bušić et al. (2018) also notice the imbalanced test set of the original dataset in Barbieri et al. (2017), and propose a more balanced dataset that is based on the top-5 and top-10 emojis in that dataset.

Çöltekin and Rama (2018) shows that SVM is better in emoji prediction than using bi-directional RNN. Wu et al. (2018) incorporated sentiment information in their neural models, and obtained small improvements in terms of overall F_1-score over the baseline models that do not use sentiment information.

Barbieri et al. (2018b) explores another metric, called *coverage error*, to account for the fact that some emojis are quite synonymous to each other (e.g., ❤️ and 💕).

3 Task Description and Data

In this paper, we begin by following Barbieri et al. (2017) on the definition of the task: given a tweet which initially contains a single emoji, predict the original emoji using just the text of the tweet. In our case, we would also like to offer a reinterpretation of the task as the task to suggest an appropriate emoji for a given tweet. This reinterpretation has a few benefits. First, it acknowledges that there is no a single correct emoji that can fit in a tweet. There could be (and there are, as we will see) multiple emojis that fit the tweet, depending on the context. Second, it makes it natural to use human ratings as evaluation metrics, instead of F_1-score, since now the systems are evaluated in how good their recommendations are.

[1] https://github.com/sweetpeach/semoji
As per Twitter policy, we release only the Tweet IDs, from which the actual texts can be queried using Twitter API.

Dataset	Train	Dev	Test
BARBIERI	580,271	4,359	4,370
UNION	597,995	74,747	75,000

Table 1: Training, development, and test set size for the two datasets in this paper.

3.1 Dataset

We use dataset from Barbieri et al. (2017), which consists of tweets retrieved between October 2015 and May 2016 containing exactly one emoji from the 20 most frequent emojis. We call this dataset BARBIERI.

As also observed by Cappallo et al. (2018) and Bušić et al. (2018), there are some limitations to this dataset, namely:

- The set of 20 emojis in the dataset are not all independent; some emojis have overlapping semantics or are ambiguous. For instance, ❤️ and 💕 arguably have similar semantics and we see people using them in similar context.
- The emoji distribution is imbalanced, as also mentioned by Bušić et al. (2018) and Cappallo et al. (2018). From Table 2, we can see that the tweets labeled with 😂, ❤️, and 😍 greatly outnumber the rest.
- It contains duplicate tweets that appear both in training and test data, diluting the model analysis. Moreover, the dataset is divided into train set, development set, and test set based on the timestamp of the tweets, resulting in more disparity in the dataset. For example, in the test set, it has 757 tweets labeled with 😂 but only 3 with 🎄.

To address the first issue, we collapsed some emojis and removed some others, and we combine the dataset from BARBIERI with the dataset from SEMEVAL (Barbieri et al., 2018a) to increase the diversity of the emojis. From BARBIERI, we removed 💙 and 😘 after analyzing the tweets with those labels because the context in which they appear tends to be too broad, and the emoji ❤️ covers similar semantics.

SEMEVAL has eight emojis which are not included in BARBIERI. We select ☀️, 😜, 📷, 📸, 😉, and 😁 to be included in our data. Then, we merge {❤️, 😍, 💕} as ❤️, {📷, 📸} as 📷, and {😊, 😁} as 😊. At the end, we have 20 emojis for our new dataset.

To address the label imbalance, we improve the number of tweets with low frequency emojis like

100.7	89.9	59	33.8	28.6	27.9	22.5	21.5	21	20.8
19.5	18.6	18.5	17.5	17	16.1	15.9	15.2	14.2	10.9

Table 2: The 20 emojis in Barbieri et al.'s (2017) dataset with their frequencies (in thousands).

All	75.0	75.0	75.0	75.0	60.4	59.5	51.9	41.0	27.8	27.6
Train	60.0	60.0	60.0	60.0	48.2	47.5	41.4	32.7	22.2	22.0
Dev	7.5	7.5	7.5	7.5	6.0	5.9	5.2	4.1	2.8	2.8
Test	7.5	7.5	7.5	7.5	6.0	5.9	5.2	4.1	2.8	2.8
All	25.9	22.4	19.4	19.2	17.9	16.9	14.8	14.7	14.1	13.0
Train	20.7	17.9	15.5	15.3	14.3	13.5	11.8	11.8	11.2	10.4
Dev	2.6	2.2	1.9	1.9	1.8	1.7	1.5	1.5	1.4	1.3
Test	2.6	2.2	1.9	1.9	1.8	1.7	1.5	1.5	1.4	1.3

Table 3: The 20 emojis in UNION dataset with their frequencies (in thousands).

❄ by including additional tweets that are crawled from February to April 2018. We follow Barbieri et al. (2017) in that we pick tweets that are geo-localized in the United States, and we pick only tweets that contain a single emoji that is in our set of 20 emojis. We also subsample the most frequent emojis so that they do not appear more than 75,000 times in this dataset.

Finally, the issue of duplicate tweets are handled in our preprocessing step, which will be described in more details in the next section.

We call our new dataset UNION. The way we construct UNION results in a much bigger validation and test data, as summarized in Table 1. The statistics for both datasets is also shown in Table 2 and Table 3.

3.2 Preprocessing

For BARBIERI dataset, we use their original dataset as is without any further modification. For UNION dataset, we preprocessed the tweets using NLTK Tweet tokenizer[2], normalizing user handles and URLs to special tokens. The tweets were tokenized and lowercased. Certain repeated punctuations are split, such as multiple exclamation marks, while others are kept, like ellipsis. Words with more than two same repeated characters are truncated into only 2 repeated characters, such as "*cooool*" becoming "*cool*". We also removed duplicate tweets and tweets with less than three tokens after tokenization. Unlike BARBIERI which was split based on timestamps, we randomly split the UNION dataset into training, validation, and test set with 80%, 10%, 10% ratios.

4 Emotion as Features

The objective of this work is to see how emotions can be incorporated into the models for predicting emoji, and whether they can be used to improve the models' performance.

For this study, we choose the more popular Ekman et al. (1969)'s six basic emotions: **anger**, **disgust**, **fear**, **joy**, **sadness**, and **surprise**. To label our tweets with emotion categories,[3] we used Twitter Emotion Recognition (Colnerič and Demsar, 2018), which is a character-based Recurrent Neural Network (RNN) model for predicting emotion categories from English tweets, to assign emotion scores to the Twitter texts. The model was trained on tweets distantly supervised by hashtags, and is reported to achieve 71.8% micro F_1-score for classifying Ekman's six emotions under multi-class setting. Distant supervision of emotion categories using hastags in tweets has been shown to correlate well with human judgments (Mohammad, 2012).

We suppose that some emojis such as 😭 and 😊 would have strong association with certain emotions. To validate this intuition, we extract from the emotion classifier the probabilities for each of the Ekman's six emotions for each tweet. We then aggregate these probability distributions based on the emoji labels, and measure the deviation of the probabilities from the average distribution over all tweets, representing the baseline probabilities for each emotion.[4]

More formally, let $X = \{x_1, \ldots, x_N\}$ be the collection of tweets with $Y = \{y_1, \ldots, y_N\}$ the corresponding emojis, where N is the number of tweets, and $y_i \in M = \{m_1, \ldots, m_{20}\}$, the set of 20 emojis. Let $X_{m_j} = \{x_i \mid y_i = m_j\}$ be the set of tweets that have m_j as the emoji label. Let $e_1, \ldots, e_K$ be the set of emotions, where K is the number of emotion categories, and let $p_{e_k}(x_i)$ be the probability of the emotion e_k assigned by

[2] http://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.casual.casual_tokenize

[3] Note that we cannot simply use emotion-labeled data as our dataset, since we also require the tweets to contain exactly one emoji.

[4] The emotion classifier seems to be biased towards the joy emotion, predicting on average 0.46 probability scores.

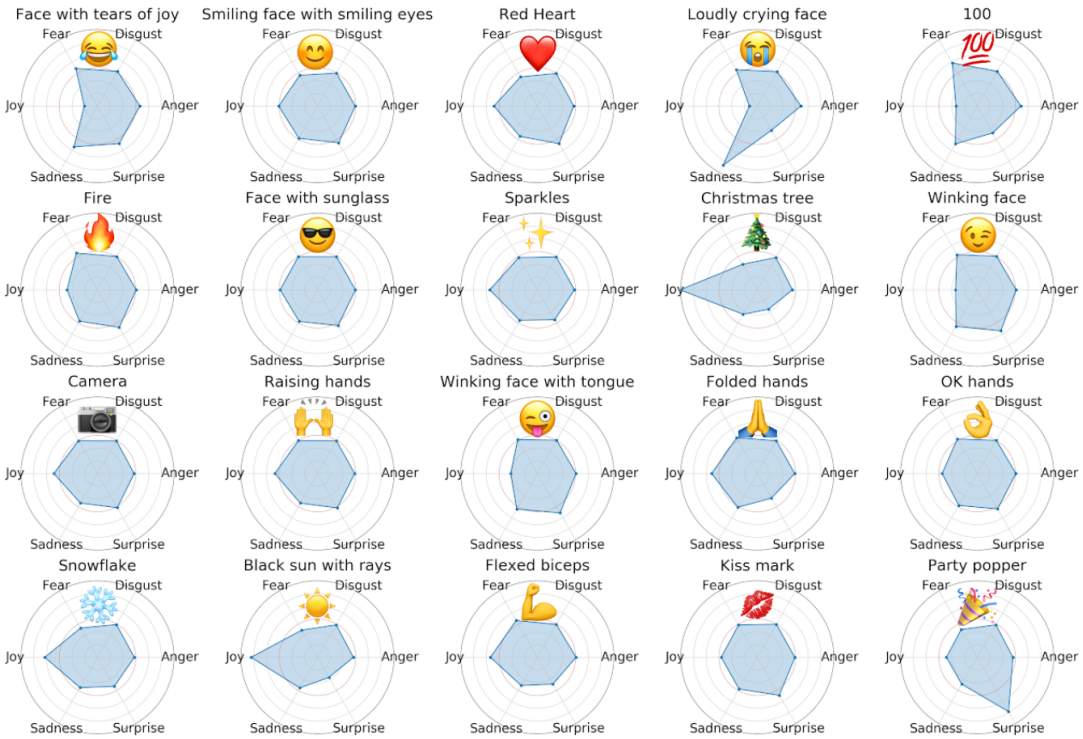

Figure 2: Emotion radar, showing the association of the 20 emojis in our UNION dataset with the six basic emotions defined by Ekman et al. (1969). These associations were calculated automatically by running emotion prediction model on tweets with emojis. Notice how 😭 has less joy and more sadness, and 🎉 has much more surprise. Ambiguous emojis, like 😂, which can appear both in positive (jokes) and negative (self-deprecation) context, have multi-peak distribution.

the emotion classifier to tweet x_i. Now, the **emotion score of an emoji** $S_{e_k}(m_j)$ as defined above is then:

$$S_{e_k}(m_j) = \frac{\sum_{x_i \in X_{m_j}} p_{e_k}(x_i)}{|X_{m_j}|} - \frac{\sum_{i=1}^{N} p_{e_k}(x_i)}{N}$$

We plot this deviation of probabilities in a radar chart, shown in Figure 2. In this chart, each emotion category is shown as separate polar axis from the center of the radar, with being closer to the center representing more negative value, and being closer to the perimeter representing more positive value over the baseline probabilities.

Some emojis are correlated with the emotions we anticipated, for example the 😭 emoji, which has very high sadness and very low joy although most emojis seem to be close to the average distribution. This might mean that the emotion classifier could not pick up the correct emotions in which those emojis are used, or simply that the emojis themselves are not particularly strongly associated with any of the six emotions. Nevertheless, we do see some intuitive trends in the emotion distribution, such as the high joy in 🎄 and the surprise in 🎉.

5 Models

To test the hypothesis that emotion information helps emoji prediction, we conduct experiments using Support Vector Machine (SVM) as the model of our choice.

As a baseline feature set for SVM, we use TF-IDF scores based on unigram bag-of-words features. This baseline model obtains 34.28% weighted F_1-score on BARBIERI dataset, which is comparable to the results in Barbieri et al. (2017) which uses bi-directional LSTM, which shows that this is a reasonable baseline model. This is also in line with the conclusion of Çöltekin and Rama (2018) that says SVM is a strong model for this task.

Inspired by our observation from the emotion radar, we consider two different ways to incorporate the emotion information produced by the emotion classifier, which we dub **basic** and **combi**. In **basic**, we use the probabilities of each emotion as features directly, resulting in 6 additional dense features on top of the unigram features. Since a single emotion might not capture the distribution of an emoji directly, in **combi** we also combine

Dataset	Emotion	P	R	F_1
BARBIERI	none	38.68	39.31	‡34.28
	basic	38.80	39.45	‡34.49
	combi	**39.25**	**39.73**	**34.86**
UNION	none	38.77	39.65	‡37.13
	basic	38.76	39.76	†37.22
	combi	**38.96**	**39.83**	**37.29**

Table 4: SVM performance with **weighted** F_1. Marked results are significantly different from the best in the respective dataset ($^{\dagger}p < 0.05$, $^{\ddagger}p < 0.01$) with bootstrap resampling ($n = 10,000$ for BARBIERI and $n = 1,000$ for UNION).

the emotion features by considering binary indicator features for all possible combination of emotion polarities. A tweet is considered to have positive polarity of an emotion if the probability of that emotion is higher than the average probability of that emotion in the training set, similarly for negative polarity. For example, a tweet might have a feature `-joy+sadness` describing the lack of joy and the abundance of sadness. This results in 722 sparse binary features.

6 Experiment Results and Analysis

To test the efficacy of the emotion features, we ran the models with the various feature combinations on BARBIERI and UNION datasets.[5] Our baseline for SVM uses only bag-of-word features. The results are shown in Table 4.

We can see that the emotion features consistently improve emoji prediction in the SVM model, with statistically significant results. The emotion combination features also consistently perform slightly better compared to the one using only the 6 basic emotions.

Based on our results, which show significant improvement coming from emotion features, we focus our analysis on the role of emotions in predicting emojis.

From the emotion radar in Figure 2, we expect that the model incorporating emotion features would get much improvement in recognizing the 😭 emoji, which has very distinct emotion distribution compared to other emojis, and we found that it is indeed the case. Table 5 shows the score

[5]For historical reasons we do not do in-depth analysis of SEMEVAL and subsequently do not report the results on SEMEVAL here. In general, we observe that the average weighted F_1-score is lower compared to BARBIERI.

🔥	-0.29	💪	-0.04	❤️	0.18	🙌	0.30
😎	-0.11	😜	0.00	🎄	0.20	☀️	0.35
😂	-0.10	😉	0.14	✨	0.20	👌	0.50
👄	-0.05	🙏	0.16	😊	0.20	📷	0.51
💯	-0.05	❄️	0.16	🎉	0.23	😭	**0.92**

Table 5: Change in F_1-score in UNION dataset from baseline SVM model to the model incorporating emotion combination features.

Emoji	Top Emotion Features
😭	Sad (0.3737), Dis (0.1162), Ang (0.0596)
🎉	Sur (0.0981), Joy (0.0400), +Ang+Dis+Fea+Joy-Sad-Sur (0.0112)
🔥	Fea (0.0867), Joy (0.0691), Sur (0.059)
😂	Dis (0.2752), +Ang+Dis+Joy+Sad+Sur (0.0133), +Ang+Dis-Fea+Joy+Sad+Sur (0.0133)
😊	Joy (0.0822), Sur (0.0421), +Ang-Dis+Joy+Sad+Sur (0.0181)

Table 6: The top features associated with the emojis. The emotions are truncated to their first three letters, so 'Ang' refers to 'Anger', 'Dis' refers to 'Disgust', and so on.

changes for each emoji in the UNION dataset from the baseline SVM model to the one with emotion combination features. We see that the 😭 has the highest positive change compared to other emojis. This shows the usefulness in distinguishing certain emojis through the emotion semantic space.

Some top features in the model show that the emotion *sadness* is the emotion feature given highest weight by the model to predict 😭 emoji. Some of which we display at Table 6. It is encouraging to see that the emojis with distinct emotion distribution, such as 😭, 🎉, and 🔥 have the corresponding emotion feature ranked the highest by the classifier.

7 Human Evaluation

In Barbieri et al. (2018b), they observe that many emojis are semantically close, and propose to use coverage error (Tsoumakas et al., 2009) to measure "how far we would need to go through the predicted emojis to recover the true label." While this is effective in measuring how the system rank the emoji in the original tweet, it does not measure the quality of the top-ranked emoji by the system. Given the possible application of an emoji predic-

tion system to recommend emojis to users, it is also important to see how well received are the top emojis predicted by the system.

To that end, we conduct human evaluation on the top-ranked emoji by each system (including the original emoji) to see which system is preferred by users. Note that this evaluation is different from the human evaluation performed in Barbieri et al. (2017). In that work, they ask human annotators to choose the best emoji from a list of 5 emojis, and compare the F_1-score with the system's predictions. In contrast, in this work we ask human raters to rate the predictions of several systems, enabling us to measure preferability of the emojis.

7.1 Methodology

We conducted human evaluation on the output of our baseline SVM model, our emotion-infused model, and also the original emoji in the tweets. We define emoji triple

$$< \text{emoji}_{\text{orig}}, \text{emoji}_{\text{bow}}, \text{emoji}_{\text{combi}} >$$

where $\text{emoji}_{\text{orig}}$ is original emoji, $\text{emoji}_{\text{bow}}$ is emoji predicted from baseline, and $\text{emoji}_{\text{combi}}$ prediction from our model infused with combination emotion features. From each dataset, we selected 1,000 pairs of (tweet, emoji triples); each pair has at least one different emoji in the emoji triple. Each pair is then given to *three raters* to be annotated. To ensure we will get sufficient data for distinguishing the preferability of different systems and of different emojis, we use the following criteria in sampling the (tweet, emoji triples):

1. The number of occurrences of all emojis in the original tweets should be approximately equal.
2. There should be at least one distinct emoji in the emoji triples.
3. There should be enough samples that have different emojis for all pairs of systems

Criterion 1 ensures that we can use the annotated data to gather baseline rating of each emoji in the original tweet. Criterion 2 ensures that we do not waste raters' time in annotating tweets that do not have distinguishing power. Criterion 3 ensures that for any two systems (e.g., BoW vs. Ours) we have enough samples to distinguish the presence or absence of preferability between them.

In the annotation interface, emojis in the triple are randomized so that the raters do not know if an emoji is the true label or a prediction from baseline or our model. Each emoji is rated in a 3-point Likert scale where 0 means that it does not makes sense to pair the emoji with the tweet, 1 means it is reasonable (there are some contexts where this would be applicable), 2 means it fits perfectly (something that they themselves would use). For this annotation task, we recruit English-speaking (not necessarily native speakers) university students and young professionals as our raters.

7.2 Result

We calculated inter-rater agreement using Fleiss's Kappa coefficient (Fleiss and Cohen, 1973), resulting in an agreement of 0.12 and 0.20 for BARBIERI and UNION, respectively. This rather low agreement shows that emoji use has quite a large variance between raters. Nevertheless, it is encouraging to see positive agreement between raters on this arguably very subjective task.

The average rating for each system is shown in Table 7. It is interesting to see that the output of our emotion-incorporated system is consistently more preferred compared to that of the baseline system, showing the benefit of emotions.

Another thing worth mentioning is the lower average rating of the original emoji compared to the system predictions. We believe this is due to the systems predicting more stereotypical emojis, thus have higher chance of being preferred by the raters given the tweets. This suggests that in a emoji prediction system, it would be better preferred by the users if the emojis are closer to the more stereotypical interpretation of the tweets, instead of second guessing the actual intent of the users.

We also note that even though the F_1-score of the systems are higher in UNION compared to BARBIERI (see Table 4), the average ratings for UNION is lower. This shows that in evaluating the quality of emoji prediction system, using F_1-score alone is not enough, as it might not give rise to a more preferred emojis.

Looking at the detailed ratings per emoji, shown in Table 8 and Table 9, we see that the emoji 🎄, 💯, and 💋 consistently get the lowest scores,. It is interesting to note that in Table 8, the four emojis ❤️, 😍, 💕, and 😘 which have similar meaning also have similar ratings, and all of them are in the top-5 emojis.

Emoji	BARBIERI	UNION
Original Tweet	1.43	1.02
Baseline Model	1.48	1.03
Our Model	**1.50**	**1.05**

Table 7: Average rating of emojis in 0-2 scale from original tweets, baseline, and our emotion-incorporated model. The difference in rating is statistically significant for BARBIERI with Wilcoxon signed rank test ($p < 0.005$ between original tweet and the two models, and $p < 0.025$ between the two models), while in UNION they are not ($p > 0.2$).

🎄	**0.87**	😂	1.35	🙌	1.45	❤️	1.57
💯	1.03	🙏	1.39	✨	1.48	😍	1.57
👄	1.31	❄️	1.39	😎	1.48	💕	1.67
👌	1.35	🔥	1.41	💪	1.54	😘	1.69
😭	1.35	💙	1.41	😊	1.56	🎉	**1.78**

Table 8: Average rating per emoji for BARBIERI

7.3 Discussion

To dig deeper into the ratings provided by the raters, we analyze some example tweets, shown in Table 10. We see that in some cases indeed there could be multiple emojis that fit in certain tweets. In the first example, the three emojis look reasonable according to the raters. 😎 may be used to explain if a user happily can go to the sold-out show. Meanwhile, if the user is unable to go to the sold-out show, depending on how the user feels, 😂 and 😭 may be used. The second example also shows that a tweet may have two different focuses. The more emotional phrase "*miss you*" can be described with ❤️ while 💪 may explain "*strong olympics*".

In the third example, the raters favor our model's prediction 😊 as the best emoji to describe the tweet. Even though the tweet contains a negative word "*can't*", the overall tone of the tweet is positive, and therefore 😊 or 😉 fits the tweet better than 😭.

It is interesting to see the fourth example which is considered as flirtatious by the tweet user through the use of 👄, but the raters consider ❤️ as a more suitable emoji since ❤️ is more versatile to be used in different contexts. Most raters do not prefer 💯 for the tweet as we see in Table 9 and Table 9 where 💯 average rate is lower than 👄.

The last example demonstrates the highest-rated emoji 😂 is the one output by the baseline. The raters also like our model's predicted emoji 😭 better than the original emoji 😊, suggesting that our model may be helpful to be used as emoji recommendation system since the model's emoji is more favorably stereotypical.

💯	**0.53**	😭	0.80	🎉	1.03	🙌	1.32
👄	0.62	❄️	0.83	✨	1.13	❤️	1.33
🎄	0.77	😂	1.00	😜	1.13	😎	1.38
👌	0.80	💪	1.02	🙏	1.16	😊	1.40
🔥	0.80	📷	1.02	☀️	1.20	😉	**1.41**

Table 9: Average rating per emoji for UNION

Tweet	Orig	BoW	Ours
sold out show at benedum center	😎 **1.33**	😂 **1.33**	😭 **1.33**
miss you but glad you are enjoying the strong olympics !	💪 **2.00**	💯 1.00	❤️ **2.00**
can't wait to use it too	😉 1.33	😭 0.33	😊 **1.67**
i'm always cute, wherever i'm going	👄 1.00	💯 0.67	❤️ **1.67**
i saw a gif of mrs smith on twitter	😊 0.00	😂 **1.67**	😭 0.67

Table 10: Sample tweets (lightly edited) with the original emoji (Orig), prediction from the bag-of-words baseline (BoW), and prediction from our emotion-infused combination model (Ours). Below the emojis are their average ratings from 3 raters. Pink highlight refers to phrase related to our system's prediction while blue highlight refers to original emoji.

8 Conclusion and Future Work

In summary, we first show the correlation between emojis and the emotion content of tweets from a large corpus of tweets. Then we make use of this correlation to improve the prediction of an emoji prediction model. Although we found that most of the emojis do not have much emotional content, for those emojis with strong emotional content, such as 😭, our experiments show significant improvements over the baseline models in terms of F_1-score.

We then further scrutinize the difference between the models through human evaluation. We confirm previous work that generally there could be more than one emojis that fit a given tweet, and conduct human rating experiments to see the

preferability of the systems' recommendation. We found that the output of the model with emotion features is generally more preferable over the baseline models and also the original tweet. This suggests that a more stereotypical emoji might be rated higher by users.

Our findings further emphasis the need for a better measure in emoji prediction task. That is, one that is more geared towards users' preferability instead of based on a single gold standard. In this work, we use a slightly labor-intensive method of collecting human ratings as a way to handle this multiple suitable emojis. A future direction would be to explore more automatic methods as proxies to users' preferability.

Another interesting venue for future work is to analyze the context of each emoji to determine how versatile an emoji is (e.g., it can appear in many different context). An emoji recommendation system should be aware of this versatility, so that it does not fall into the trap of always predicting versatile emojis due to relatively high suitability with any tweet.

Acknowledgements

The authors would like to thank Adeline Anugrah Raphani, Naoki Otani, Aditi Chaudhary, Emily Ahn, Jeffrey Wong, Mark Lauda Lauw, Christopher Bryant, Stephen Haniel Yuwono, Karinska Eunike Muis, Mercia Wijaya, Peter Phandi, Raymond Hendy Susanto, Stefanus Setiadi, Victoria Anugrah Lestari, and Raissa Eka Fedora for rating the emoji-tweet pairs.

The authors are also grateful for useful feedback from Eduard Hovy, Geoff Kaufman, Teruko Mitamura, people in Carnegie Mellon University Social Computing reading group, and the anonymous reviewers.

The first author was sponsored by US DOT FAST Act - Mobility National (2016 - 2022) - CMU 2017 Mobility21 UTC #31.

The second author was sponsored by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O), program: Low Resource Languages for Emergent Incidents (LORELEI), issued by DARPA/I2O under Contract No. HR0011-15-C-0114.

References

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, Valencia, Spain. Association for Computational Linguistics.

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 (SemEval-2018)*, New Orleans, LA, United States. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa-anke Jose, Steven Schockaert, and Horacio Saggion. 2018b. Interpretable Emoji Prediction via Label-Wise Attention LSTMs. In *EMNLP*, pages 4766–4771.

Hrvoje Bušić, Ante Spajić, and Nika Šućurović. 2018. How Much Context is Useful in Emoji Prediction? Technical report.

Spencer Cappallo, Stacey Svetlichnaya, Pierre Garrigues, Thomas Mensink, and Cees G. M. Snoek. 2018. The New Modality: Emoji Challenges in Prediction, Anticipation, and Retrieval. pages 1–13.

N. Colnerič and J. Demsar. 2018. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, pages 1–1.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 Task 2 : SVMs perform better than RNNs at Emoji Prediction. pages 34–38.

Paul Ekman, E. Richard Sorenson, and Wallace V. Friesen. 1969. Pan-Cultural Elements in Facial Displays of Emotion. *Science*, 164(3875):86–88.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Saif M. Mohammad. 2012. #Emotional Tweets. In *First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255.

Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PLoS ONE*, 10(12):1–22.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2009. Mining Multi-label Data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer.

Wiesław Wolny. 2016. Emotion Analysis of Twitter Data That Use Emoticons and Emoji Ideograms. In *International Conference on Information Systems Development (ISD)*, pages 476–483.

Chuhan Wu, Fangzhao Wu, Sixing Wu, and Zhigang Yuan. 2018. THU NGN at SemEval-2018 Task 2 : Residual CNN-LSTM Network with Attention for English Emoji Prediction. pages 410–414.