Multi-resolution Annotations for Emoji Prediction

Weicheng Ma[1], Ruibo Liu[2], Lili Wang[3], and Soroush Vosoughi[4]

Minds, Machines, and Society Group
Department of Computer Science, Dartmouth College
[1,2,3]{first.last.gr}@dartmouth.edu
[4]soroush.vosoughi@dartmouth.edu

Abstract

Emojis are able to express various linguistic components, including emotions, sentiments, events, etc. Predicting the proper emojis associated with text provides a way to summarize the text accurately, and it has been proven to be a good auxiliary task to many Natural Language Understanding (NLU) tasks. Labels in existing emoji prediction datasets are all passage-based and are usually under the multi-class classification setting. However, in many cases, one single emoji cannot fully cover the theme of a piece of text. It is thus useful to infer the part of text related to each emoji. The lack of multi-label and aspect-level emoji prediction datasets is one of the bottlenecks for this task. This paper annotates an emoji prediction dataset with passage-level multi-class/multi-label, and aspect-level multi-class annotations. We also present a novel annotation method with which we generate the aspect-level annotations. The annotations are generated heuristically, taking advantage of the self-attention mechanism in Transformer networks. We validate the annotations both automatically and manually to ensure their quality. We also benchmark the dataset with a pre-trained BERT model.

1 Introduction

Emojis have become crucial components of written language. Emojis were initially designed to express emotions or feelings, e.g., 😊 for a smiley face, and they have grown to be a large family of over 2,000 icons over the years which can express not only emotions but a wide range of objects or actions, e.g., 🎁 for a gift and 🎉 for celebrations. Compared to words, emojis have the merit of preserving information more densely. For example, 😂 carries the same meaning as the phrase "laughing with tears in eyes". Additionally, the byte-level encoding of subtle linguistic expressions makes it easier to discriminate complicated feelings, e.g., the bond between 😂 and 😭 is clearly weaker than their phrasal explanations "laughing with tears in eyes" and "crying loudly" due to the similarity between "tear" and "crying". These characteristics of emojis aid in accurate summarization of text, thus benefiting natural language understanding (NLU) tasks.

Felbo et al. (2017) define the emoji prediction task by finding the most appropriate emoji(s) summarizing a piece of text. They also show with experiments that language representations learned on the emoji prediction task can boost the performance of emotion recognition, sentiment analysis, and sarcasm detection tasks. Consequently, using emoji prediction as a bridge to solve other natural language processing (NLP) tasks appears to be effective and promising. However, the emoji prediction task is yet far from being well established. First and foremost, as a classification task, there is not a set of labels agreed upon by previous research. To the best of our knowledge, all the existing papers on emoji prediction use either a handcrafted emoji set (Felbo et al., 2017) or the most frequent emojis in their individual datasets (Barbieri et al., 2018c,b). Handcrafted emoji sets are usually limited in size and topics (usually limited to emotional emojis), while frequency-based emoji sets are dataset-specific. The lack of a standard label set makes it difficult to evaluate and compare emoji prediction models, hampering the research on emoji prediction and its interactions with other NLP tasks. To solve this problem, we use an emoji list from the unicode office [1] as the label set for the emoji prediction task. This emoji list includes 1,467 emojis in total, ordered by the median frequency of their use from multiple resources. We believe using this emoji list is good

[1] https://home.unicode.org/emoji/emoji-frequency/

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 6684–6694,
November 16–20, 2020. ©2020 Association for Computational Linguistics

for standardizing the task since it is open to all researchers and is not influenced by how we sample the data.

The second problem with emoji prediction is that existing labeled datasets are either too small in scale or not publicly available. This often results from the policy of social media platforms on using their data and the constantly changing nature of posts on these platforms, e.g., post deletion and edits. To address the problem of data unavailability or expiration, we annotate the PAN-19 Celebrity Profiling corpus (Wiegmann et al., 2019), a tweet-based corpus, which is large and available to all researchers. We provide three types of annotations in this paper. Existing emoji prediction datasets are almost all annotated on the passage-level under the multi-class classification setting, which means each record contains exactly one tweet and one emoji. While we also release this type of annotation, we additionally provide passage-level multi-label and aspect-level multi-class classification annotations. Annotations for the passage-level multi-label classification setting are similar to the multi-class setting, but with possibly multiple emojis in each record (i.e., a tweet could be associated with multiple emojis). We introduce aspect-level labels to the emoji prediction task to enable a finer-grained analysis of the functions of emojis in tweets. Each emoji in these annotations points to a span of its corresponding text instead of the entire tweet. Text fractions associated with different emojis in the same tweet may overlap with each other.

Given the large size of our dataset, all three types of annotations are generated automatically using heuristics or with the help of a Transformer-based model. The assumption underlying the passage-level annotations is that the text fully covers the meanings of emojis in a tweet. Thus we extract the emojis appearing in the text as passage-level labels, as (Felbo et al., 2017) do. Under the multi-class classification setting, a record is duplicated and assigned different emojis if it contains multiple emojis. The aspect-level annotations are created based on passage-level multi-class classification labels. Since the attention maps in a Transformer-based model reflect the interrelations of each word pair, we are able to evaluate the contribution of each word to a predicted emoji under the multi-class classification setting. We then combine the labels based on tweets to form the aspect-level multi-class annotations for the dataset. We will introduce the annotation methods in more detail in Section 3.2.

The contributions of this paper are three-fold.
First, we provide a large emoji list to be used as a label set for the emoji prediction task. These emojis are all frequently-used and meaningful, benefiting further research on the emoji prediction task and its connections to other NLP tasks.
Second, we introduce a data annotation method based on the self-attention mechanism in Transformer networks (Vaswani et al., 2017). The method is designed specifically for annotating aspect-based labels and can potentially be used on any NLP task.
Third, we provide three types of annotations for emoji prediction based on a publicly available tweet dataset. Besides the commonly used tweet-level[2] multi-class classification labels, our annotations include passage-level multi-label and aspect-level multi-class classification labels for better understanding of the linguistic roles of emojis.

We release a carefully curated (both manually and automatically) emoji prediction dataset based on the 64 top-ranked emojis in our emoji list. [3]

2 Related Work

The study of emoji usage in textual data has seen a rise in recent years. Most related research are done over Twitter, Gab, or Microblog data since the use of emojis is more common on social media. Mahajan and Shaikh (2019) compared the way emojis were used in Twitter and Gab posts, and they claimed that emojis with negative sentiment scores were more frequently used on Gab than the other. The use of emojis on Twitter also appeared to be more balanced compared to Gab in posts related to the same event, i.e., the most frequent emoji counts for 19.79% of total emoji usage on Gab and 6.28% on Twitter. We base our research on a Twitter dataset for the balanced emoji usage and more neutral points of view.

Since the amount of unique emojis was large, early research treated emojis as special word-level tokens and examined the linguistic roles with coarse-grained classification objectives, for example, predicting whether an emoji was used redundantly in its context (Donato and Paggio, 2018)

[2] In this paper, we use tweet-level and passage-based annotations interchangeably.

[3] Available upon request.

or classifying the linguistic purposes of emojis (Na'aman et al., 2017). With the help of Recurrent Neural Networks (RNNs), Felbo et al. (2017) and Barbieri et al. (2018c) could predict proper emojis from tweet posts. Going one step further, Barbieri et al. (2018a) combined image and text features for predicting emojis. Due to the overly large amount of data required to train neural-network-based classifiers with tens or even hundreds of labels, most emoji prediction datasets were labeled automatically, using heuristics or pre-defined rules. The logic previous researchers used to label these datasets were simple, assuming that every emoji appearing in a social media post qualified as a label of the text. Our work also relies on this assumption since the data we use consists of tweets posted by authorized accounts, who are not likely to often use emojis arbitrarily or randomly. We also extend the annotation method to be able to generate more complex, aspect-level annotations automatically.

Additionally, it is commonly agreed that emojis are closely related to emotions, sentiments, sarcasm, irony, etc. Felbo et al. (2017) showed that the language representations learned from the emoji prediction task were useful in emotion recognition, sentiment analysis, and sarcasm detection tasks. Hayati et al. (2019) designed experiments to show the interconnections between emoji usage and ironic expressions. Singh et al. (2019) also evaluated the influence of emojis on irony detection and sentiment analysis tasks, but they replaced emojis with descriptive text in this process. Based on previous knowledge about emojis and the emoji prediction task, we also use sentiment analysis, emotion recognition, and formality classification tasks to validate the quality of our annotations in this paper.

3 Annotation Method

We annotate tweets for the emoji prediction task in this paper. For clarity, we refer to each tweet in the dataset by $t = \{w_1, w_2, ..., w_n\}$ of length n and an emoji set $E = \{e_1, e_2, ..., e_m\}$ of size m, where w_i is the $i-th$ word in t and e_j is the $j-th$ emoji in E. The three settings of this task are formally defined as follows.

Passage-level multi-class classification: Predict the best e_j most closely related to t.

Passage-level multi-label classification: Predict whether each $e_j \in E$ is associated with t closely enough.

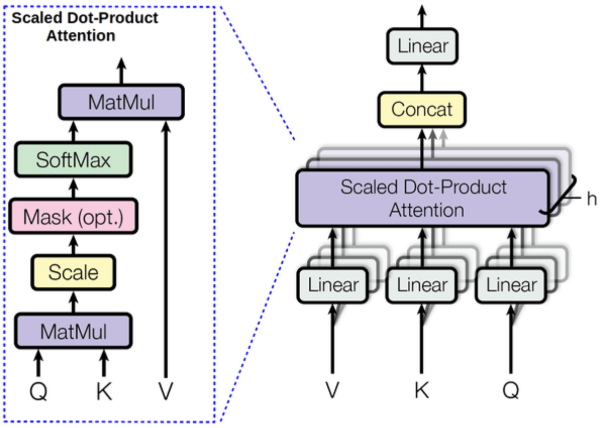


Figure 1: Interpretation of a self-attention head. Q, K, and V are query, key, and value matrices, respectively. The figure is cited from Vaswani et al. (2017).

Aspect-level multi-class classification: Given t and p subsets of t, each denoted as $s_q = \{w'_1, w'_2, ..., w'_k\}$ where $w'_l \in t$ for all $1 \leq l \leq k$ and $1 \leq q \leq p$, predict the best e_j mostly closely related to s_q.

This is an extension to our earlier work (Ma et al., 2020) in which tweets are labeled on passage-level only, and no official emoji list is used to standardize the annotations. The passage-level classification datasets are annotated directly using the emojis appearing in each tweet. Under the multi-class classification setting, we duplicate the tweets if they are bond to multiple emojis and use one emoji to label each copy. For the aspect-level annotations, each record is identified by a tweet and an aspect text piece. We generate the aspect-level annotations with the help of a Transformer-based model. Emojis are removed from the text in all three settings in case that neural models trained on our datasets directly copy and paste the emojis from text into predictions.

3.1 Transformer Networks

The core of Transformer networks is the self-attention mechanism. Figure 1 displays the structure of a self-attention head. In an attention head, each word gets its query, key, and value vectors by multiplying the Q, K, and V matrices with its representation vector. The attention score vector of a word is generated with the dot product between its query vector and the key vector of all the words in the same tweet. We get the attention map by stacking the attention vectors together. In a Transformer-based model, there are usually multiple attention heads on each layer (16 in the bert-large-cased model we use). According to past studies about Transformer networks, the highest layers of a Transformer model encode mainly task-specific features for predictions (Kovaleva et al., 2019), while shallower layers extract fundamental and low-level linguistic features (e.g., the middle Transformer layers attend mostly to syntactic features (Vig and Belinkov, 2019; Hewitt and Manning, 2019)). Thus we rely on the mean of all the attention maps on the last layer of a Transformer-based model to represent the token-level interrelations corresponding to each emoji label.

BERT (Devlin et al., 2019) is a family of pre-trained Transformer-based models. In BERT architecture, predictions are conditioned on the representation of the "[CLS]" token on its last layer. Thus, we are able to evaluate the contribution of each word to the final prediction by looking at how heavily the "[CLS]" token attends on the other words. To be specific, we use the pre-trained bert-large-cased model in all our experiments. It is worth noting that any self-attention-based neural model potentially fits our annotation framework. We pick the bert-large-cased model in our experiments because it performs the best on the 64-label single- and multi-label emoji prediction tasks, beating the bert-base models and two XLNet models.

3.2 Attention-based Automatic Annotation Method

Based on the observation that important tokens to a prediction made by BERT are heavily attended by the "[CLS]" token, we design the following steps in sequence, to annotate an aspect-level multi-class emoji prediction dataset.

3.2.1 Data Preparation

We use the 64 top-ranked emojis in our emoji list to annotate an aspect-level dataset. The selected emojis are shown in Table 1. We limit to the 64 emojis to allow for manual quality inspections of the annotations. As the first step, we remove tweets not containing any of the 64 emojis. Tweets shorter than five words are also discarded since most short tweets are formed only by mentions, retweets, URLs, and hashtags. This may unavoidably remove some meaningful short tweets, e.g., "Good night.". But generally, these tweets form one aspect as a whole, reducing the aspect-level annotations to tweet-level annotations. URLs and hashtags are replaced with #URL# and #HASHTAG# tokens as well to reduce noise. To avoid dominating classes in the dataset, we balance the number of records in each class by reducing the number of tweets associated with frequent emojis. We set the threshold to be 10,000 records. Tweets having labels in the classes with less than 10,000 instances are all preserved, while the rest are randomly pruned. Table 1 also displays the number of times each emoji appears in the balanced dataset. Duplicated emojis in the same tweet do not count

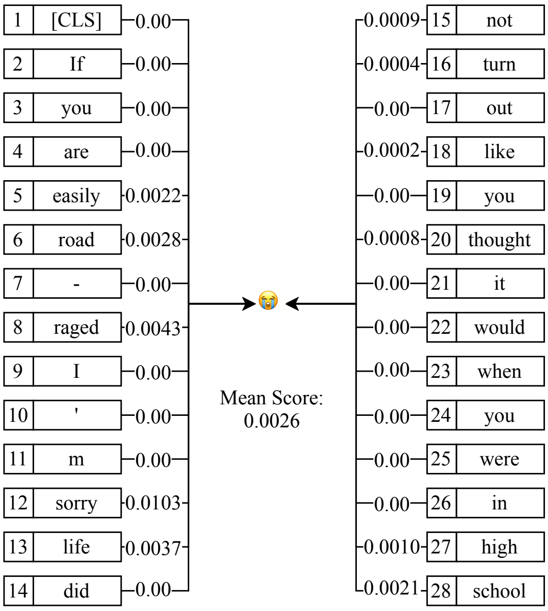


Figure 2: Attention score distribution in an example sentence regarding one specific emoji (the loudly crying face).

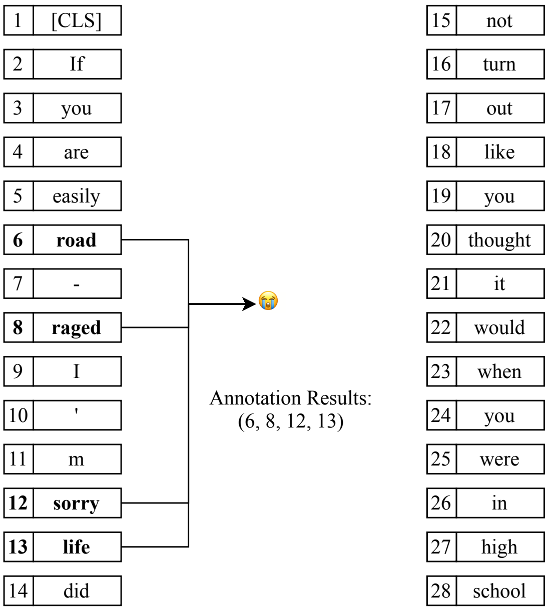


Figure 3: An example of aspect-level annotation results. Bold words connected to the emoji are annotated as the aspect for the given emoji.

😂	❤️	😍	🤣
45,846	76,301	17,369	16,260
😊	🙏	💕	😭
28,854	31,765	9,935	13,164
😘	👍	😅	👏
23,436	34,739	20,049	20,651
😁	♥️	🔥	💔
14,127	9,975	23,349	9,938
💖	💙	😢	🤔
9,939	15,670	9,937	19,471
😆	🙄	💪	😉
9,936	9,936	18,544	33,299
☺️	👌	🤗	💜
18,924	27,593	9,941	9,940
😔	😎	😇	🌹
9,935	14,134	9,943	12,378
🤦	🎉	‼️	💞
9,935	17,066	9,953	9,936
✌️	✨	🤷	😱
16,554	12,021	12,297	9,999
😌	🌸	🙌	😋
9,938	9,945	31,797	9,936
💗	💚	😏	💛
9,934	9,935	9,977	9,936
🙂	💓	🤩	😄
9,941	9,934	8,145	14,127
😀	🖤	😃	💯
9,936	9,939	9,936	11,433
🙈	👇	🎶	😒
9,934	9,934	9,935	10,006
🤭	❣️	❗	🤪
4,807	6,780	9,024	11,496

Table 1: The emoji list we use to annotate our dataset in this paper. The number of records related to each emoji is also noted.

😂	❤️	😍	🤣
81.63	97.14	88.91	90.06
😊	🙏	💕	😭
80.92	86.44	86.94	93.16
😘	👍	😅	👏
89.12	81.44	83.02	90.58
😁	♥️	🔥	💔
96.13	87.20	91.37	82.04
💖	💙	😢	🤔
88.12	86.39	87.18	84.57
😆	🙄	💪	😉
93.81	85.03	83.81	95.38
☺️	👌	🤗	💜
92.05	89.74	91.46	88.15
😔	😎	😇	🌹
81.94	83.40	91.66	83.99
🤦	🎉	‼️	💞
88.65	94.78	86.42	84.06
✌️	✨	🤷	😱
80.30	85.23	86.83	93.80
😌	🌸	🙌	😋
95.20	81.29	88.38	94.18
💗	💚	😏	💛
91.61	90.40	83.10	89.34
🙂	💓	🤩	😄
91.64	91.79	88.64	89.71
😀	🖤	😃	💯
88.49	90.56	97.32	95.43
🙈	👇	🎶	😒
92.43	95.13	93.46	87.36
🤭	❣️	❗	🤪
80.59	88.52	87.10	94.69

Table 2: The evaluation scores of a bert-large-cased model on 64 binary classification datasets we construct. The scores are in terms of F-1 score. The emojis are sorted by frequency.

for multiple occurrences.

The bert-large-cased model achieves 41.44% in F-1 score when evaluated on the entire dataset under the multi-class classification setting, which is too low to generate proper aspect-level annotations. To enable the model to learn better representations of the tweets in this dataset, we split our dataset into 64 binary classification subsets. In each subset, we choose all the tweets labeled with one specific emoji as positive examples. We generate equal numbers of negative examples by randomly sampling the same amount of tweets from the other emoji groups. For example, as there are 76,301 tweets labeled with the ❤️ emoji in our dataset, we sample 1,211 tweets randomly from tweets labeled with every of the rest 63 emojis to form the negative examples. This results in a binary classification dataset for ❤️ with 152,594 records. The same process is repeated to generate 64 binary classification datasets.

3.2.2 Model Fine-tuning

The bert-large-cased model is pre-trained on large text corpora. We fine-tune and evaluate the model on each binary classification dataset. The entire datasets are used as both the training and test data since the positive and negative instances in these datasets are perfectly balanced, and because we expect the model to overfit on the datasets. The experiments are done on one RTX Titan GPU for an average of 3 hours per experiment. As is shown in Table 2, the evaluation scores of the BERT model range from 80.30% to 97.32% in F-1 score in our experiments. The scores show that the BERT model is good enough for annotating a high-quality aspect-level emoji prediction dataset.

3.2.3 Word Scoring

After the model is fine-tuned on each binary classification dataset, we evaluate the model on the dataset again and use the attention map on the last layer of the model on positive instances to annotate them. Recall that for BERT models, the attention score between the "[CLS]" token and each other token reflects the token's contribution to the prediction. It is worth noting that BERT tokenizes words into tokens using Byte Pair Encoding (BPE) in its preprocessing step. For readability, we re-combine the subword tokens into words and average their scores to generate word-level attention scores. Assuming the model always makes correct predictions, the attention weights can model the relatedness of each word to the labeled emoji. Though our model cannot always generate correct predictions, we discard the annotations generated from wrong predictions in the final release of the dataset as the majority of data is annotated correctly.

3.2.4 Thresholding

As the last step of annotating the dataset, we generate the annotations from the attention scores. Figure 2 shows one example sentence with attention scores to an emoji attached to the words. We first set the scores of stopwords and punctuation marks to 0 to avoid including them in the annotations. After that, we use the mean attention score on the remaining words in each tweet as the threshold to select important words from the text. The tweet in Figure 2 is annotated as in Figure 3 after thresholding, for example. After the annotations are generated, we group the records based on tweets to form the aspect-level multi-class classification dataset.

4 Emoji Prediction Dataset

One of the goals of this paper is to annotate the PAN-19 Celebrity Profiling dataset for emoji prediction. We refer to the newly-annotated dataset as Multi-Resolution Emoji Prediction (MREP) Dataset since it contains both passage-level and aspect-level annotations. When releasing the data, we randomly split the dataset into train, dev, and test datasets with 80%, 10%, and 10% of the data amount, respectively. The random seed we use for this separation is 29936.

Table 3 displays one record in our dataset with its three sets of labels. The label set of our dataset is constructed by the 64 emojis in Table 1. Sentences are labeled by single emojis under the tweet-based multi-class classification setting. The tweet-based multi-label classification annotations are emojis separated by semicolons. An annotation under the aspect-level multi-class classification setting is formed by a list of emoji indices and their corresponding text spans in the tweet. The final annotated dataset contains 1,036,131 multi-class classification records and 500,114 multi-label or aspect-level records. We benchmark our dataset using a pre-trained bert-large-cased model and show the results in Table 4. We do not benchmark our dataset with other models since no existing emoji prediction model is designed for multi-label or aspect-based predictions.

5 Annotation Quality Validation

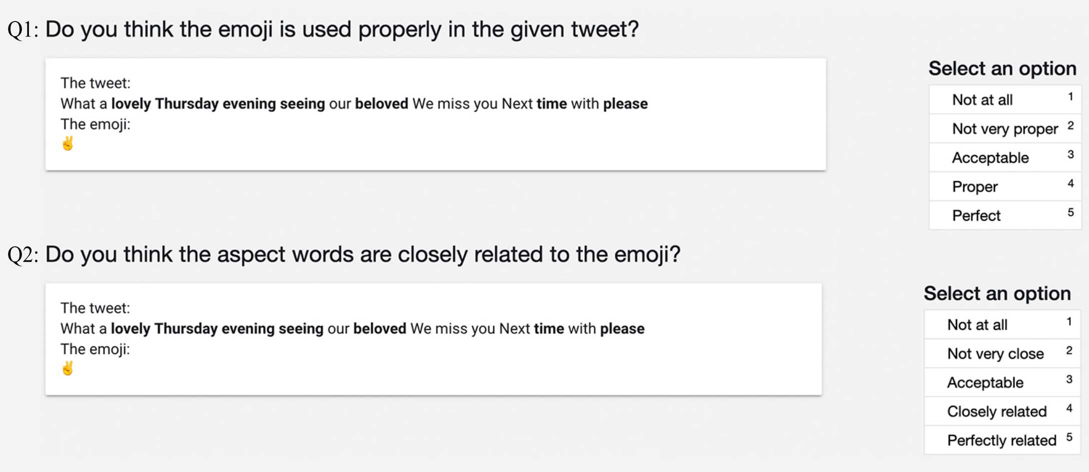
5.1 Automatic Validation

We validate the quality of our annotations in two ways. Since it is supported by (Felbo et al., 2017) that neural models trained on a high-quality emoji prediction dataset can help improve the performances of some NLU tasks, we train a BERT model jointly on our dataset and four other datasets to automatically validate our annotations. The four datasets we use are the Stanford Sentiment Treebank (SST) (Socher et al., 2013) for sentiment analysis, GYAFC (Rao and Tetreault, 2018) for formality classification, and MELD and MELD-Dyadic (Poria et al., 2019) for emotion recognition. In the experiments, we use the MT-DNN (Liu et al., 2019) codes with a batch size of 32. The pre-trained model we use is bert-large-cased with 24 layers and 16 attention heads on each layer. We fine-tune the model for seven epochs in each experiment, with a learning rate of 0.00005. The GYAFC, MELD, and MELD-Dyadic datasets are also partitioned into train/dev/test datasets by 80%/10%/10% of the entire data using a random seed of 29936, for consistency.

The evaluation results are displayed in Table 5. Our dataset brings noticeable improvements to all the tasks we choose. This is a strong validation

Tweet	Catch all the feels with me LIVE tonight on Instagram at 8p PT when @AmericanIdol is back for Hollywood Week solos I'll take YOUR questions at the commercial breaks! #HASHTAG# #URL#	
Multi-Class	🙏	
Multi-Label	🙏; 🎉	
Aspect-Level	Emoji	Aspect Text Span
	🙏	(catch, all, feels, LIVE, Instagram)
	🎉	(live, tonight, Hollywood, Week)

Table 3: An example record in our labeld dataset. Multi-Class, Multi-Label and Aspect-Level corresponds to the three types of annotations, respectively. We replace all the hashtags and URLs with #HASHTAG# and #URL# tokens at the preprocessing step.



Q1: Do you think the emoji is used properly in the given tweet?

The tweet:
What a **lovely Thursday evening seeing** our **beloved** We miss you Next **time** with **please**
The emoji:
✌

Select an option

- Not at all 1
- Not very proper 2
- Acceptable 3
- Proper 4
- Perfect 5

Q2: Do you think the aspect words are closely related to the emoji?

The tweet:
What a **lovely Thursday evening seeing** our **beloved** We miss you Next **time** with **please**
The emoji:
✌

Select an option

- Not at all 1
- Not very close 2
- Acceptable 3
- Closely related 4
- Perfectly related 5

Figure 4: A preview of our Amazon Mechanical Turk questionnaire. We ask three annotators to answer the questionnaire for each annotated record.

Task	ACC	ACC@5	F-1	ACCsub
PBMC	41.88	61.95	41.44	-
PBML	99.41	-	-	27.16
ABMC	82.16	96.07	79.91	-

Table 4: Benchmark results on our dataset under three different settings. PBMC, PBML, and ABMC correspond to passage-based multi-class, passage-based multi-label, and aspect-based multi-class classification settings, respectively. ACC denotes accuracy and ACC@5 refers to accuracy of the top-5 predictions. For multi-label classification, ACC refers to the average accuracy of predicting every single emoji while ACCsub counts only exact matches.

of the high quality of our annotations. Among the three types of annotations, the aspect-level classification setting helps the most. This is probably because the emojis are better associated with the aspects, not all the words in a tweet.

Additionally, we run experiments to explore the subjectivity and randomness using similar but nuanced emojis, e.g., the ten heart-shaped emojis with different colors in our emoji list. To be specific, we construct a multi-label classification dataset using the tweets associated with the heart-shaped emojis from our PBMC dataset. We then fine-tune a bert-large-cased model on the subsampled dataset. The model achieves an F-1 score of 43.47% in this experiment, indicating that these heart-shaped emojis are as distinguishable as the other emojis. Furthermore, we cluster the heart-shaped emojis in the PBMC dataset into one class and evaluate its influence on our four downstream tasks. The fine-tuned model produces 93.69%, 85.07%, 44.05%, and 44.82% F-1 scores on the SST, GYAFC, MELD, and MELD-Dyadic datasets, respectively, slightly lower than the original PBMC dataset without emoji clustering. These experiments make it clear that similar emojis are, though sometimes unconsciously, used in differently depending on the context. Since emoji clustering does not provide additional help to downstream tasks, and because clustering the emojis increases subjectivity in creating the datasets, we do not apply

	SST		GYAFC		MELD		MELD-Dyadic	
	ACC	F-1	ACC	F-1	ACC	F-1	ACC	F-1
Single-task	93.12	93.38	88.98	88.06	65.29	44.03	63.06	44.71
+ PBMC	93.46	93.74	87.02	86.59	65.80	44.08	64.18	44.97
+ PBML	94.88	94.05	89.27	88.94	72.48	46.70	70.22	46.03
+ ABMC	95.02	94.73	89.60	89.13	72.63	46.81	71.50	46.74
+ PBMC, PBML	94.63	94.57	89.11	88.37	69.05	45.50	68.25	45.86
+ PBMC, ABMC	94.50	94.46	88.60	88.44	70.22	45.86	69.41	46.01
+ PBML, ABMC	**95.44**	**95.28**	**90.77**	**89.62**	73.91	47.14	71.95	46.77
+ ALL	94.91	95.01	89.51	89.25	73.11	47.07	71.87	46.75
+ Emotional	95.18	95.19	89.94	89.31	**73.98**	**47.20**	**72.19**	**46.93**
+ Other	94.36	93.73	89.23	88.99	66.39	44.63	68.96	45.90

Table 5: Evaluation results on SST, MELD, MELD-Dyadic, and GYAFC datasets, and by jointly training these tasks with our emoji prediction datasets. PBMC, PBML, and ABMC refer to passage-based multi-class, passage-based multi-label, and aspect-based multi-class classification settings, respectively. For the "emotional" setting, we use the records bound to emojis not representing concrete items under all three settings for jointly training with the main tasks, while in "other" we use the emojis not expressing emotions only. ACC refers to Accuracy. Scores in bold are the best scores.

ID	Tweet	Emoji
1	Find someone who **looks** at you **like** @hashtagcatie **looks** @zachdonofrio #HASHTAG#	😍
2	On my second year as an Inquirer Read Along ambassador I always **look forward** to these interactive… #URL#	😊
3	**travel**... **work** mode #URL#	💕
4	Just an alround perfect **summers day**...! #URL#	🙈

Table 6: Examples of imperfect aspect-level annotations in our dataset. The words in bold are labeled aspects corresponding to the emoji.

😂	❤️	😍	🤣	😊	🙏	💕	😭
😘	👍	😅	👏	😁	♥️	😜	💔
💖	💙	😢	🤔	😆	🙄	💪	😉
☺️	👌	🤗	💜	😔	😎	😇	❣️
🤭	😒	😃	💞	✌️	😀	😄	😱
😌	🖤	🙌	😋	💗	💚	😏	💛
🙂	💓	🤩	💯	❗	‼️	👇	🙈
🤦‍♂️	🤷‍♀️						

Table 7: The list of emojis expressing abstract meanings in the top 64 emojis of our emoji set.

emoji clustering in our annotation framework or the annotated datasets.

We also run experiments by choosing the records labeled with "abstract" (e.g., emojis showing emotions) and "concrete" emojis (e.g., emojis representing objects) respectively. We arrange the "abstract" emojis in Table 7 and "concrete" emojis in Table 8. Results show that "abstract" emojis bring more improvements to the aforementioned tasks. This implies that sentiments, metaphors, and emotions may be abstract concepts, thus agreeing better with the predictions of "abstract" emojis.

🔥	🌹	🎉	🎶	✨	🌸

Table 8: The list of emojis expressing concrete meanings in the top 64 emojis of our emoji set.

5.2 Manual Validation

Since the use of emojis is very subjective in many cases, we sample 700 aspect-level annotations randomly for manual validation as well. Other than a few exceptions, the vast majority of the annotations (around 85%) look appropriate. We list four example imperfect annotations in Table 6. The most common problem with these annotations is that, since we removed stopwords from the annotations, some aspect-level labels are not complete and the meanings may change. In Tweets 1 and 2 in Table 6, for example, the words "at" and "off" are not chosen as parts of the labels. This does not affect Tweet 1 much, but for Tweet 2, the words "look forward" do not explain the usage of the 😊 emoji. We cannot leave all the stopwords as they are since they are usually heavily attended in attention maps of Transformer-based models. Being a language model, BERT unavoidably scores stopwords and punctuation marks high in attention maps due to their frequent co-occurrence with almost all the other words. Without removing the stopwords, the mean attention score will increase significantly, and useful words may not be correctly labeled without filtering them out. This can probably be avoided by using a list of phrases whose meanings change without all their component words. A postprocessing step removing only the stopwords outside the dependency path of any other aspect word may also be helpful. The annotation of Tweet 3 is not appropriate either since the text is too short to include any useful information associated with the emoji. It is possible that the user loves traveling or working, but the emotion cannot be inferred from this piece of text. Luckily, this does not happen often in our dataset (only 2 out of 700 in the sampled data fall under this category). The wise monkey emoji in Tweet 4 appears to be used randomly. This is the only instance in the sample that we do not know why it is used. The top-ranked emojis are mostly related to emotions or sentiments, the meanings of which are usually contained in the text. However, "concrete" emojis might pose more difficulty for annotation as they are sometimes used in place of words or phrases. This may cause annotation problems in the future if we expand our research to a broader range of emojis. A preprocessing step substituting all the "concrete" emojis with their descriptive texts can compensate for this problem.

To avoid bias, we also send the sampled aspect-level annotations for validation on Amazon Mechanical Turk. The questionnaire we design is shown in Figure 4. Each time a worker is given one tweet, one emoji associated with the tweet, and the text span annotated in our dataset corresponding to the emoji. We require the workers to answer two questions for each data point, namely, how well the emoji relates to the tweet and whether the selected span of text properly expresses the given emoji. Both questions are scored using a Likert scale (Joshi et al., 2015), in the range $\{1 \text{ (worst)}, 2, 3 \text{ (acceptable)}, 4, 5 \text{ (perfect)}\}$. Each record is validated by three different workers. The answers are aggregated together by averaging them. Question 1 mainly validates the quality of tweet-level annotations in our dataset. The average score for Question 1 is 2.9 (mainly acceptable), showing that the way emojis are used in our dataset is understandable, but does not perfectly reflect how our validators use emojis in their daily lives. The results also teach us that the patterns of emojis usage differ from person to person, as the unanimous agreement rate of our validators is 26.4% for Question 1. By categorizing the scores into Poor (1, 2), Acceptable (3), and High (4, 5), however, we get a unanimous agreement rate of 93.4% for Question 1. Question 2 is designed to validate the quality of our aspect-level annotations. The average scores are 2.9 on all the records and 3.5 on the tweets having an average score greater than or equal to 3 in Question 1 (i.e., for tweets where the emoji is deemed acceptable or higher). This shows that our aspect-level annotations for tweets where the emoji is appropriate are of acceptable quality, despite the issue discussed above.

6 Conclusion and Future Work

Emoji prediction has become a popular task in the NLP community, but the lack of publicly available large-scale datasets with high-quality annotations remains a bottleneck for this task. In this paper, we annotated a publicly available Twitter dataset for the emoji prediction task. We designed an annotation method for aspect-level annotations using the self-attention mechanism in Transformer networks. This method showed great performance in labeling our dataset, and can potentially be used in other tasks as well. Our dataset contains three types of annotations, namely the passage-level multi-class and multi-label classification labels, and the aspect-level multi-class classification annotations. We validated our annotations both automatically and manually to ensure their quality. We also benchmarked our dataset using a pre-trained bert-large-cased model. Our labeled datasets are available upon request. There are two main paths for extending this work. First, the aspect-level annotation method can be applied to other NLP tasks. Second, our annotations in the emoji prediction dataset can be enhanced by including an enriched label set.

Acknowledgment

This research was supported in part by the Dartmouth Burke Research Initiation Award.

References

Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. 2018a. Multimodal emoji prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 679–686, New Orleans, Louisiana. Association for Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa-Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018b. SemEval 2018 task 2: Multilingual emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 24–33, New Orleans, Louisiana. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa-Anke, Jose Camacho-Collados, Steven Schockaert, and Horacio Saggion. 2018c. Interpretable emoji prediction via label-wise attention LSTMs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4766–4771, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Giulia Donato and Patrizia Paggio. 2018. Classifying the informative behaviour of emoji in microblogs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Shirley Anugrah Hayati, Aditi Chaudhary, Naoki Otani, and Alan W Black. 2019. What a sunny day â˜”: Toward emoji-sensitive irony detection. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 212–216, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, 7(4):396.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Weicheng Ma, Ruibo Liu, Lili Wang, and Soroush Vosoughi. 2020. Emoji prediction: Extensions and benchmarking. *arXiv preprint arXiv:2007.07389*.

Khyati Mahajan and Samira Shaikh. 2019. Emoji usage across platforms: A case study for the charlottesville event. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 160–162, Florence, Italy. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Abhishek Singh, Eduardo Blanco, and Wei Jin. 2019. Incorporating emoji descriptions improves tweet classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2096–2101, Minneapolis, Minnesota. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Matti Wiegmann, Benno Stein, and Martin Potthast. 2019. Celebrity profiling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2611–2618, Florence, Italy. Association for Computational Linguistics.