SWiT: Emoji Annotation

Emily Braun
emilybraun@brandeis.edu

Joyce Guo
jguo@brandeis.edu

Kasey La
kaseyla@brandeis.edu

Keer Xu
keerxu@brandeis.edu

Abstract

Emojis, since their invention, have surged in popularity. They've become integral to text-based conversations, such as messages to friends and comments on YouTube videos, Tweets, and more. While plain text lacks the ability to convey emotions like facial expressions, tone, or intonation, emojis bridge that gap by allowing people to add feelings to their messages. However, much like words, emojis can also have their own unique connotation; the same emoji can have completely different meanings depending on the context in which it appears.

Our emoji annotation project aims to understand the nuanced connotations of three popular emojis: 1. 😭 (loudly crying face) 2. 🥲 (smiling face with tear) 3. 🥹 (face holding back tears)

This paper presents the annotation guidelines our group used to annotate the emoji emotions and the machine learning model we used to train on the annotated data.

1 Introduction

Before emoticons or emojis existed, computer-mediated communication was previously thought to be impersonal due to the lack of verbal queues. Since the invention of emojis and their widespread availability in digital spaces, this has drastically changed as people have increasingly utilized emojis to indicate emotion in plain text. They serve as a visual and semiotic token that carries semantic and/or tonal meaning.

While emojis are often used based on their most obvious emotional meaning (Shoeb and de Melo, 2020)—for instance, a smiling face emoji represents happiness—people can deploy the same emojis in various contexts because of its associated connotation. For example, some may use a smiling face emoji to convey irony or to inject humor into an awkward situation. In addition, previous research has already shown that emojis can have entirely different meanings based on the cultural subgroup using it (Danesi, 2017; La et al., 2022). These different meanings and connotations can also change and evolve over time, making emojis particularly interesting to observe in context.

In this project, our group aims to predict the potential meaning of an emoji given the context. The main task is to annotate and analyze the different meanings of three emojis and along with the contexts in which they appear in.

2 Guidelines

For the guidelines, we initially wanted to label each Tweet with either 'happy' or 'sad' and annotate them based on which emotion aligned better with the Tweet. This would allow us to denote the Tweet's context in which the emoji appeared. This annotation schema had a few issues.

Namely, there were many Tweets in our dataset that either included multiple emojis of the three that we were observing for this project and/or multiple spans of the same emoji that were separated by text. Tweets like these were more nuanced in sentiment. We noticed that certain emoji spans could have different sentiments depending on the context. For example, if a Tweet said, "My favorite team lost 🥲 but at least Peanut can fly home early to see his cats 🥹", it wouldn't make sense to label the entire Tweet as either 'happy' or 'sad' because one label isn't enough to capture the full context in which each emoji appears in.

Because of these issues, we decided it would be better to annotate and label pre-processed emoji spans. If a Tweet had multiple kind of emojis or multiple spans of the same emoji, there would be several pre-processed emoji spans within that one Tweet. This would allow us to label each span separately, solving the issue of multiple emoji spans having different sentiments in a Tweet. This would more precisely capture what sentiment each emoji was contributing and the immediate context. In addition to these changes, we also changed the labels to ‘positive’ and ‘negative’ to allow a wider range of emotions. We felt that many Tweets did not fit well into the labels of ‘happy’ or ‘sad’, and believed that the new labels of ‘positive’ and ‘negative’ would better capture other emotions such as anger, disgust, affection, and hope.

Our final guidelines had annotators labeling pre-processed emoji spans of the selected emojis depending on whether they are used with positive or negative sentiment. A third label of ‘unconfident’ were for emoji spans that were extremely difficult to label, had context or slang that the annotator was unfamiliar with, or if the primary language of the Tweet was not English.

3 Description of final dataset

The dataset we choose is from Kaggle https://www.kaggle.com/datasets/ericwang1011/tweets-with-emoji, which is a Twitter dataset of 43 types of emojis with 20k Tweets per emoji.

The raw dataset is noisy and contains irrelevant information. Our goal is to make each Tweet clear enough to only contain the plain text and the emojis. This way, it will be easier and more straightforward for annotators to understand the meaning behind each emoji given the context. We performed a series of data cleaning steps to pre-process the raw data.

First, we removed the URL and Twitter username in the Tweets. When users mention another Twitter user, they use the ‘@’ character with the other user’s username. When the users are replying to another Tweet, the URL of the original Tweet will be automatically appended on the replying Tweet. When users included an image or a video, a URL of the media will be appended to the Tweet. All of these pieces of information are irrelevant to the context, so they were removed. We also removed Tweets that only contain the three emojis but no context, because annotators will not be able to tell the underlying meaning of the emojis without context. We kept Tweets that featured emojis besides our three chosen ones because those emojis could potentially provide contextual information (i.e., “🥹❤️” appears to be more positive than negative simply with the addition of the heart). We opted not to filter out Tweets in foreign languages since many Tweets contained code-switching and many English parts can still be analyzed. Table 1 displays some Tweets before and after the pre-processing steps.

Pair #	Tweet before and after preprocessing
1	*@folastag A house doesn’t smell of food* 🥹 A house doesnt smell of food 🥹
2	*Coffee machines in office* 😂😭 *https://t.co/lxuZnGV1o1* Coffee machines in office 😂😭

Table 1: Examples of Tweets before and after data cleaning.

Emoji	Positive	Negative	Unconfident
🥹	376	119	27
😭	243	345	31
🥲	166	321	15
Total	**785**	**785**	**73**

Table 2: The number of times a sentiment was chosen for a span containing each emoji.

Our gold data contained 8 Tweets (10 spans) for the face holding back tears 🥹 emoji, 12 Tweets (14 spans) for the loudly crying face 😭 emoji, and 4 Tweets (11 spans) for the smiling face with tear 🥲 emoji.

For internal annotation, the final dataset contained 100 Tweets for each emoji, amounting to 300 Tweets total. There were 122 spans for the face holding back tears 🥹 emoji, 104 spans for the loudly crying face 😭 emoji, and 106 spans for the smiling face with tear 🥲 emoji.

For external annotation, the final dataset contained 200 Tweets for each emoji, amounting to 400 Tweets total. There were 438 spans for the face holding back tears 🥹 emoji, 419 spans for the loudly crying face 😭 emoji, and 419 spans for the smiling face with tear 🥲 emoji.

Overall, the face holding back tears 🥹 emoji had 570 spans, the loudly crying face 😭 emoji had 537 spans and the smiling face with tear 🥲 emoji had 536 spans. For the breakdown of sentiment counts per emoji, see Table 2.

4 IAA and adjudication

For this project, we had six annotators total. Our annotators were Emily Braun, Joyce Guo, Kasey La, Keer Xu, Ben Lambright, and Chris Tam. The four people in the SWiT group annotated 300 Tweets each, divided into 100 Tweets per selected emoji. Internally annotated Tweets were the only ones with multiple annotators. The two external annotators were assigned 2 random batches for each emoji. Each batch consisted of 100 Tweets. This amounted to 200 Tweets per emoji, and a total of 400 Tweets.

There were several steps to our adjudication process. First, we decided that a Tweet label required adjudication if only 2 out of 4 annotators agreed. After filtering out Tweets that matched these requirements, our group would look at these Tweets and reassess the label we each gave it. This would involve considering the text again and also discussing our reasoning for our label with each other. If reassessment resulted in 3 out of 4 agreement, the adjudication process would be finished. If reassessment still resulted in 2 out of 4 agreement, the final discussion and decision would be made between Guo and La.

The reason behind this is because many of these Tweets require a lot of cultural and contextual knowledge. Although our guidelines go over some of these subjects, we believe that Guo and La were best suited to decide the final label due to their familiarity with the various online spaces that these Tweets originated from and their understanding of general trends and terminology on Twitter. If reassessment resulted in 2 out of 4 agreement but Guo and La both agreed on a label, then that label would be chosen. Otherwise, there would be further discussion and research done to understand as much context as possible. For the few Tweets that Guo and La could not agree on, these were ultimately assigned 'unconfident'.

To evaluate the agreement between the 4 internal annotators, we calculated Fleiss's Kappa (κ). The following are the numbers that were computed to achieve κ, where $\bar{P}$ is the observed agreement and P_e is the expected agreement:

$$\begin{aligned} \bar{P} &= 0.7565 \\ \bar{P}_e &= 0.4366 \\ \kappa &= 0.5678 \end{aligned}$$

According to Landis and Koch (1977), the resulting κ can be interpreted to mean that the agreement level is moderate.

5 Machine learning baseline

Once we had our annotated data, we ran sklearn's Logistic Regression model and sklearn's Naive Bayes model on it. For Tweets that had multiple emojis that were annotated with different sentiment, we decided to split the Tweet into sections based on the text that we believed the emoji to be referring to, which generally was the text that came before the emoji. For example, for the Tweet "Eidrees is almost 8 months old 🥹 Why is he growing so fast 🥹", where the first emoji was tagged as 'positive', and the second was tagged as negative, we split it into "Eidrees is almost 8 months old 🥹", tagged as 'positive', and "Why is he growing so fast 🥹", tagged as 'negative'. There were less than 20 of these types of Tweets, so this splitting was done manually.

After that, we were able to use Naive Bayes and Logistic Regression to take a Tweet and predict its sentiment. We got similar results for each model, getting about 62% accuracy for Naive Bayes and about 66% accuracy for Logistic Regression. Both models especially struggled with predicting Tweets labeled 'unconfident' correctly.

Part of the model's inability to tag 'unconfident' data correctly may have had to do with the fact that we had much less 'unconfident' data than 'positive' or 'negative' data. When we ran the model again with all 'unconfident' tags removed from our dataset, the Logistic Regression model's accuracy was about 72%, and the Naive Bayes model's accuracy was about 69%.

The confusion matrices for each model are as follows:

97	28	0
60	90	0
11	16	1

Table 3: Confusion matrix for Naive Bayes including 'unconfident' tags

91	34	0
41	109	0
11	17	0

Table 4: Confusion matrix for Logistic Regression including 'unconfident' tags

103	31
59	97

Table 5: Confusion matrix for Naive Bayes excluding 'unconfident' tags

94	40
42	114

Table 6: Confusion matrix for Logistic Regression excluding 'unconfident' tags

6 Future directions

Some potential future directions for this project include revising our guidelines further, expanding our research to incorporate other emojis, inviting annotators with domain-specific knowledge, and performing our machine learning task on the span level rather than the Tweet level.

Revising our guidelines further include adjusting the way that the 'unconfident' label is used and breaking down the sentiment labels to account for the difficulties that came with ambiguous sentiments such as 'bittersweet' and 'hope'. Since the 'unconfident' label was created with the intent to find areas of improvement in our guidelines rather than including it as a sentiment label for the machine learning task, it was difficult to relay that reasoning to our annotators. The main use cases for the 'unconfident' label were for contexts where the sentiment could not be readily extracted, whether that be due to non-English text or the lack of context (which meant that the span could potentially be interpreted as 'positive' or 'negative' equally).

On the other hand, contexts that could be interpreted as 'bittersweet' or 'hope' tended to be difficult for annotators to identify as 'positive' or 'negative'. In these cases, it may be helpful to expand the two sentiment labels into distinct sentiments to account for those nuances. This idea is attributed to Shoeb and de Melo (2020), whose project features 8 emotion classes (i.e., joy, sadness, anticipation, surprise, disgust, fear, anger, and trust) for 150 emojis. Likewise, there are other emojis that we found to occur in similar contexts as the three we have chosen, but due to time constraints and the lack of data, we were unable to include them. Expanding the two existing sentiment labels to incorporate more distinct sentiments would also benefit from including other emojis.

Since our external annotators annotated individual batches that did not have overlaps with each other, the labels they chose for each emoji span could not be compared; thus, their annotations could not be checked for quality against one another. This proved to be an issue when we discovered 13 Tweets (16 spans) that were annotated both internally and externally, where 5 spans had differing annotated sentiments. For this particular issue, as well as acknowledging the many subcultures and terminologies that exist on Twitter and occurred in our dataset, the quality of the annotations would improve greatly if our annotators were more familiar with the online space. We were made aware that our guidelines were not utilized as much as we had hoped, so perhaps annotators with more interest about the online space would find referencing the guidelines to be more beneficial.

Our machine learning task in its current stage assigns a sentiment to a Tweet containing an emoji span. This means that the Tweets containing more than one emoji span with differing sentiments were split into sub-Tweets (discussed in greater detail in Section 5). To improve the task, we could assign sentiments to emoji spans instead, which would match our annotation task. One idea to combat the issue mentioned above, where a Tweet can contain emoji spans with different sentiments, would be to put more weight on the tokens closest to the emoji span compared to the rest of the Tweet, since, after performing annotations by hand, this is how we tended to perform the task. For example, consider a Tweet where the first sentence is negative while the second sentence, containing an emoji span at the end, is positive (e.g., someone is unhappy about a certain situation but expresses hope that things would get better). The emoji span most likely is used in a positive connotation and we want the model to prioritize the positive context closest to it rather than the negative context from the first part of the Tweet.

References

Marcel Danesi. *The Semiotics of Emojis*, chapter Emoji Semantics, pages 51–76. Bloomsbury, 2017.

Kasey La, Owen Fisher, and Nicole Wong. What's so funny about a skull and a statue?: A semantic analysis of sentence-final emojis. https://github.com/kla7/Skull-Moai-Emoji-Analysis, 2022.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2529310.

Abu Awal Md Shoeb and Gerard de Melo. EmoTag1200: Understanding the association between emojis and emotions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8957–8967, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.720. URL https://aclanthology.org/2020.emnlp-main.720.