Understanding Emoji Ambiguity in Context: The Role of Text in Emoji-Related Miscommunication

Hannah Miller*, Daniel Kluver*, Jacob Thebault-Spieker*, Loren Terveen* and Brent Hecht[†]

*GroupLens Research, University of Minnesota, Minneapolis, MN 55455, USA
{hmiller, kluver, thebault, terveen}@cs.umn.edu

[†]People, Space, and Algorithms (PSA) Computing Group, Northwestern University, Evanston, IL 60208, USA
bhecht@northwestern.edu

Abstract

Recent studies have found that people interpret emoji characters inconsistently, creating significant potential for miscommunication. However, this research examined emoji in isolation, without consideration of any surrounding text. Prior work has hypothesized that examining emoji in their natural textual contexts would substantially reduce the observed potential for miscommunication. To investigate this hypothesis, we carried out a controlled study with 2,482 participants who interpreted emoji both in isolation and in multiple textual contexts. After comparing the variability of emoji interpretation in each condition, we found that our results do not support the hypothesis in prior work: when emoji are interpreted in textual contexts, the potential for miscommunication appears to be roughly the same. We also identify directions for future research to better understand the interplay between emoji and textual context.

Introduction

Emoji characters are extremely popular on the Web and in text-based communication (Dimson 2015; Medlock and McCulloch 2016). The ubiquity of emoji is in part enabled by the Unicode Consortium, which provides a worldwide text encoding standard for emoji characters just as it does for more traditional characters (e.g., letters, numbers, Chinese characters). The Unicode standard specifies both (1) a Unicode character for each emoji that identifies it across platforms and (2) a name that describes—but not prescribes—its appearance. The appearance of individual emoji is specified by a given font, just as for text characters.

However, there is an important difference between emoji characters and more traditional characters: emoji fonts are largely specific to individual technological platforms, so a given emoji character's appearance may vary extensively across platforms. For example, the Unicode character with code U+1F606 and name "smiling face with open mouth and tightly-closed eyes" renders as this pictograph 😆 on Microsoft Windows devices but as this pictograph 😆 on Apple devices. Emojipedia currently tracks 19 platforms with their own emoji fonts ("Emojipedia" 2017). Moreover, platforms update their emoji fonts just as they update their operating systems and, as such, emoji fonts are actually platform-version specific, not just platform-specific. For instance, this pictograph 😆 shows how emoji character U+1F606 was rendered in previous Microsoft implementations of the Unicode standard. This means, for example, that the emoji rendering a Twitter user chooses and sees when they compose a tweet (on one version of a platform) may very likely *not* be the emoji rendering many followers see when they read the tweet (as they may be using different platforms or versions).

Researchers have shown that this across-platform (and across-version) diversity, combined with varying interpretations of even the exact same pictograph, raises the risk of miscommunication when using emoji. Indeed, examining some of the most popular anthropomorphic (i.e., human-looking) emoji characters, Miller et al. (2016) and Tigwell and Flatla (2016) found that the perceived sentiment of a given emoji character varies extensively, even among people using the same platform. Psycholinguistic theory (Clark 1996) suggests that in order to avoid miscommunication incidents, people must interpret emoji characters in their exchanges in the same way (and they must know that they are interpreting them the same way). The research of Miller et al. and Tigwell and Flatla suggests that these interpretation pre-conditions may break down in certain cases.

There is, however, an important caveat to prior studies of miscommunication with emoji: they focused on people's interpretations of standalone emoji. Although emoji are sometimes used in isolation, they are most often accompanied by surrounding text (Medlock and McCulloch 2016). Indeed, Miller et al. (2016) recommended considering emoji in the context of surrounding text as a key direction of future work. In particular, they hypothesized that at least some of the potential for miscommunication that they observed would disappear in this more ecologically valid setting.

In this paper, we seek to test Miller et al.'s hypothesis directly. Specifically, we ask:

> ***RQ****: Does the presence of text reduce inconsistencies in how emoji are interpreted, and thus the potential for miscommunication?*

To address this question, we adopt an approach similar to that employed by Miller et al. in which we use an online survey to solicit people's interpretations of emoji. Emoji renderings were presented to participants either in isolation (*standalone*) or embedded in a textual context (*in-context*), and participants judged the sentiment expressed by each emoji rendering. Textual contexts were gathered by randomly selecting tweets containing the corresponding emoji character. For each condition, we computed how much people varied in their interpretations, estimating the potential for miscommunication of each emoji when it is presented with textual context and when it is presented without it.

Our results tell a clear story: the hypothesis of Miller et al. is not supported. In general, emoji are not significantly less ambiguous when interpreted in context than when interpreted standalone. In addition, all such differences are small relative to a baseline amount of ambiguity; roughly speaking, they are "just noise". Finally, our results do not trend in a particular direction: while some emoji are less ambiguous in context, others actually are *more* ambiguous in context.

We next discuss related work. Designing a robust experiment that controls for variation in types of textual contexts among other concerns was an involved process, and we outline this design following related work. We then discuss our statistical methods, followed by our results. We close by highlighting the implications of our results more broadly.

Related Work

We first give an overview of relevant psycholinguistic theory and how it relates to studying emoji in textual context. We next review work that has built emoji semantic and sentiment inventories and research that has more explicitly examined the consistency of emoji interpretation.

Linguistic Theory

Emoji serve a paralinguistic function in digital written text, substituting for nonverbal cues such as facial displays and hand gestures in face-to-face communication (Clark 1996; Medlock and McCulloch 2016; Pavalanathan and Eisenstein 2016; Walther and D'Addario 2001). More specifically, emoji usage can be understood as "visible acts of meaning" as defined by Bavelas and Chovil (2000):

> "(a) [Visible acts of meaning] are sensitive to a sender-receiver relationship in that they are less likely to occur when an addressee will not see them, (b) they are analogically encoded symbols (c) their meaning can be explicated or demonstrated in context, and (d) they are fully integrated with the accompanying words."

As part of their "Integrated Message Model", Bavelas and Chovil (2000) argue that audible and visible communicative acts (i.e., visible acts of meaning) should be considered as a unified whole, whereas previously these channels were often studied independently. By examining text and emoji together, this paper extends research on emoji-related communication towards this more "integrated" perspective.

Large-scale Emoji "Inventories"

A few recent research projects have sought to build "inventories" of meanings or senses associated with specific emoji characters. For instance, Novak et al. (2015) developed the first emoji sentiment lexicon, representing the sentiment of each emoji as the distribution of the sentiment of tweets in which it appeared. This work shows that emoji may be used in different ways and take on different meanings, but it does not address whether people agree on the meaning of an emoji in a given use case. Wijeratne et al. (2016) provide a similar resource to Novak et al. but for semantics. Wijeratne et al.'s emoji "dictionary" aims to help disambiguate emoji in context. Our results, however, suggest that doing so may be difficult, since people often do not agree on the meaning of specific emoji and specific emoji-bearing text snippets.

Consistency of Emoji Interpretation

Prior to emoji, Walther and D'Addario (2001) studied ambiguity of the emoticons ":-)", ":-(" and ";-)" and found that participants varied little in their sentiment interpretations. Recent research, however, has shown this is not the case for emoji. For instance, Miller et al. (2016) used a psycholinguistic lens to examine how much people vary in their interpretations of emoji and found that this variability can be extensive both in terms of sentiment and semantics. Tigwell and Flatla (2016) extended Miller et al.'s research to consider sentiment along two dimensions instead of one, finding similar results.

Miller et al. (2016) argued that the variance in emoji interpretation that they observed may be detrimental to the successful use of emoji in communication. Since two people must have the same interpretation of a signal (i.e., communicative act) in order for it to have been successful in an exchange (Clark 1996), when an addressee's interpretation differs from a sender's intended meaning, a *misconstrual* or miscommunication occurs. Miller et al. encode this psycholinguistic understanding of interpretation variability in their core metric – the emoji *misconstrual score* – and we use the same metric here.

More generally, both Miller et al. and Tigwell and Flatla studied interpretation of standalone emoji. But, as discussed above, the most common use case involves emoji characters embedded in surrounding text (Medlock and McCulloch 2016). Thus, our present study seeks to extend the literature on emoji interpretation and its relationship to emoji-related miscommunication by studying emoji in textual context.

Survey Design

To address our research question, we conducted a survey that solicited over two thousand people's interpretations of emoji in isolation and in context. Although we borrow the basics of our experimental design from Miller et al. (2016), the consideration of textual context required the addition of several complex components to our survey and analytical framework. In this section, we provide an overview of our survey design, and in the next section we highlight our statistical approach. We note that both sections feature rather detailed description of methods; this is to enable our work to be replicable. We also note that while Miller et al. examined both sentiment and semantic ambiguity, we focus on sentiment. As discussed below, considering both would have resulted in insufficient experimental power and, as noted by Miller et al. (2016), semantic differences have more limited interpretability.

Emoji and Platforms

Prior work (Miller et al. 2016) revealed variability in how people interpret emoji, identifying some as particularly subject to miscommunication. For our study, we selected the 10 emoji from that study that had the most potential for sentiment ambiguity. These "worst offenders" (see Table 1) are among the most frequently-used anthropomorphic emoji (Miller et al. 2016). Thus, by studying these ten emoji in context, we can determine whether the presence of surrounding text mitigates the problem where it is both impactful and most acute.

We considered the same five platforms as Miller et al. (Apple, Google, LG, Microsoft, and Samsung), as well as Twitter's emoji renderings (or "Twemoji") because we used Twitter as our source of text containing emoji (see the following sub-section). Importantly, all of these platforms have updated at least some of their emoji renderings since Miller et al. performed their work. Of the five platforms' renderings of our 10 emoji Unicode characters (50 renderings total), 30 have been updated[1] (all 10 of Apple's renderings, 6 of Google's, 2 of LG's, all 10 of Microsoft's, and 2 of Samsung's). Some of the updates are relatively minor, for example resolution changes (particularly in Apple's case) and changes to adhere better to emerging emoji norms (e.g., LG's updates to match emoji skin tone norms). However, other updates involve substantial modifications in rendering appearance and effectively result in new implementations of the emoji characters (e.g., Microsoft's changes).

To afford comparison to Miller et al.'s work while also ensuring that our results reflect the emoji state-of-the-art, we

UNICODE	NAME	Previous Apple	Current Apple	Previous Google	Current Google	Previous LG	Current LG	Previous Microsoft	Current Microsoft	Previous Samsung	Current Samsung	Twitter
1F606	SMILING FACE WITH OPEN MOUTH AND TIGHTLY-CLOSED EYES	😆	😆	😆	😆	😆		😆	😆	😆	😆	😆
1F601	GRINNING FACE WITH SMILING EYES	😁	😁	😁	😁	😁		😁	😁	😁	😁	😁
1F64C	PERSON RAISING BOTH HANDS IN CELEBRATION	🙌	🙌	🙌		🙌	🙌	🙌	🙌	🙌		🙌
1F605	SMILING FACE WITH OPEN MOUTH AND COLD SWEAT	😅	😅	😅	😅	😅		😅	😅	😅		😅
1F60C	RELIEVED FACE	😌	😌	😌	😌	😌		😌	😌	😌		😌
1F648	SEE-NO-EVIL MONKEY	🙈	🙈	🙈		🙈		🙈	🙈	🙈		🙈
1F64F	PERSON WITH FOLDED HANDS	🙏	🙏	🙏		🙏	🙏	🙏	🙏	🙏		🙏
1F60F	SMIRKING FACE	😏	😏	😏		😏		😏	😏	😏		😏
1F631	FACE SCREAMING IN FEAR	😱	😱	😱	😱	😱		😱	😱	😱		😱
1F602	FACE WITH TEARS OF JOY	😂	😂	😂	😂	😂		😂	😂	😂		😂

Table 1. The 10 emoji characters (Unicode and Name) in our study and their associated renderings for the six platforms in our study. The "Previous" column for each of the platforms shows the renderings at the time of Miller et al.'s (2016) work and the "Current" column shows the current renderings (as of Fall 2016). Merged cells indicate that no changes were made to a rendering. A white background indicates inclusion in our study (all current versions and previous versions we deem to be substantively different from the updated version, 77 renderings total). A gray background indicates exclusion (previous and current versions deemed not substantively different).

included in our study all current renderings of our 10 emoji characters, as well as all previous renderings whose current renderings substantively changed relative to the prior renderings. We determined whether a rendering underwent a substantive change by having two coders independently assess each update as substantive or not. A substantive change was defined as having nontrivial chance of affecting one's sentiment interpretation. The coders achieved 87% agreement (26/30 renderings), and resolved differences jointly. In the end, 17 renderings were determined to have substantively changed. Table 1 shows the full set of renderings that we considered; those with white backgrounds (77 total) were included in the study.

Building a Corpus of Emoji Textual Contexts

We chose Twitter as a corpus for text containing emoji (i.e. emoji-bearing tweets) for two key reasons. First, Twitter is a readily available source of communication that uses emoji. Second, most tweets are public and thus more likely to be interpretable without additional hidden interpersonal context. This would not be the case, for example, in a corpus of direct sender-receiver mobile text messages as such messages are often interpreted using established norms and shared knowledge between the two parties (Clark 1996; Cramer, de Juan, and Tetreault 2016; Kelly and Watts 2015), a point to which we return later. To maximize the likelihood that any participant would be able to interpret the tweets in our study (i.e., minimize the need for exogenous context), we also filtered tweets in the following ways:

- Tweets had to be written in English so that they would be readable by our participants.
- Tweets had to be original tweets, not retweets, so they appeared in their original context.
- Tweets could not contain user mentions, to reduce the chance that they were intended for a specific individual.
- Tweets could not contain hashtags, to reduce the chance that they were intended for a particular sub-community.
- Tweets could not be from a "verified" account (i.e., celebrity or public figure), to reduce the chance that the content (and interpretation) depended on context from popular culture, current events, and other exogenous information.
- Tweets could not contain URLs or attached media (e.g., photos, video), to reduce the chance that interpretation depends on external content rather than just the surrounding text.

We used the Twitter Streaming API to randomly collect approximately 64 million public tweets between September 27 and October 15, 2016. We then filtered these tweets according to the above criteria, leaving approximately 2 million tweets to select from for our study.

To ensure that our findings about emoji in context are not tweet-specific, we randomly sampled 20 unique tweets containing each emoji character (10x20 = 200 tweets total) from our filtered tweet dataset. When a Twitter user crafts a tweet on a specific platform (i.e. the tweet's "source" platform), the user is working with emoji as specifically rendered on that platform. Therefore, to minimize biased use cases of each emoji that may arise from differences between source platform renderings, we stratified the sampling of 20 tweets (for each character) to be from four identifiable rendering-specific sources. Specifically, we randomly sampled 5 tweets from each of the following[2]: (1) Twitter Web Client (originate with Twitter's emoji renderings, or Twemoji), (2) Twitter for iPhone, iPad, or Mac (originate with Apple's renderings), (3) Twitter for Android (cannot be sure of the origin of emoji renderings because Android is fragmented by manufacturer, and many use their own emoji fonts), and (4) Twitter for Windows Phone (originate with Microsoft's renderings). Finally, we also made sure that each tweet contained only a single emoji.

An emoji-bearing tweet is often read on platforms that have different emoji renderings than those from platform on which the tweet was authored. For example, this tweet from our dataset was shared from an Apple device:

Will be at work in the a.m 😆 (Apple)

But this same tweet is rendered differently for users of other platforms:

Will be at work in the a.m 😆 (Google)
Will be at work in the a.m 😆 (LG)
Will be at work in the a.m 😆 (Microsoft)
Will be at work in the a.m 😆 (Samsung)
Will be at work in the a.m 😆 (Twitter)

This example demonstrates emoji communication across platforms, in which people see different renderings of the same emoji character in the same tweet. Even people using the same platform but different versions of that platform may see different renderings of the same emoji:

Will be at work in the a.m 😆 (Current Microsoft)
Will be at work in the a.m 😆 (Previous Microsoft)

In other words, multiple versions of a given platform's renderings essentially creates another across-platform dimension.

To gain a cross-platform (and cross-version) understanding of the potential for miscommunication using emoji with

[2] For 5 of the 40 emoji-source pairs, we did not have enough tweets in our dataset due to limited data and low platform usage (Twitter Web Client and Twitter for Android), so we fulfilled this deficit by pulling tweets that satisfied the same criteria we outlined from a dataset that was collected using the Twitter API between August and September 2015.

text (as Miller et al. did on a standalone basis), we had to consider each sample tweet as it would be rendered on different platforms (and platform versions). As such, we replicated each of our 200 tweets for each rendering of the emoji they contained, as we did for the example above. In total, we gathered interpretations for 1,540 rendering-specific tweets (77 total emoji renderings x 20 tweets per rendering).

Experiment Design

We designed our experiment to capture the two types of data needed to make the comparison central to our research question: (1) interpretations of *standalone* emoji (replicating the work of Miller et al.) and (2) interpretations of emoji *in context*. We did this using a *between-subjects* experiment design; participants were randomly assigned to the standalone or context condition until the quota for each was met.

For the standalone emoji condition, we used the same survey design as Miller et al., except we collected only sentiment interpretations. We focused on sentiment interpretation because the sentiment rating scale lets us precisely compare interpretations, and differences between sentiment interpretations are easier to understand than open-response semantic interpretation differences. Importantly, considering semantics also would have affected our ability to recruit a sufficient number of participants, as the semantic component of the Miller et al. survey design requires a great deal more participant effort.

Participants in the *standalone* condition were randomly assigned 20 emoji renderings. Participants in the *in-context* condition were randomly assigned 20 of the emoji-containing tweets. For each tweet, we randomly showed one rendering of the emoji to display (simulating viewing the tweet on that platform-version). In both conditions, participants were instructed to judge the sentiment expressed by each emoji (standalone or in context) on an ordinal scale from Strongly Negative (-5) to Strongly Positive (5), mirroring the scale used in prior work (Miller et al. 2016; Taboada et al. 2011). For the *standalone* condition, we used the same intra-rater quality control as Miller et al. by having each participant interpret Apple's heart emoji (❤️, Unicode U+2764) both before and after their random sample of 20 emoji. For the *in context* condition, we used "love ❤️" to show before and after the sample of tweets.

Participants

We recruited participants via Amazon Mechanical Turk. Since geography and culture may influence interpretation (Barbieri et al. 2016; Park, Baek, and Cha 2014; Park et al. 2013), we recruited only participants from the United States (limiting our findings to this cultural context); we also required participants to have 97% of their work approved with at least 1,000 approved tasks completed. We estimated it would take participants roughly 10 seconds per interpretation. With each participant providing 22 interpretations (random sample of 20 plus the heart emoji twice), we compensated all participants $0.50 for completing the survey (prorating from a wage of $8 per hour and rounding up).

We established quotas to gather sufficient power for our statistical comparisons (see below) and to leave sufficient buffer for participants who might fail the intra-rater check. We aimed for 50 standalone evaluations of each of our 77 emoji renderings, and thus targeted 210 participants for the standalone condition and acquired 238[3]. We aimed for 30 interpretations for each of our 1,540 rendering-specific tweets, thus targeted 2,500 participants and acquired 2,356.

Following Miller et al., we used intra-rater reliability results as a filter: we excluded participants whose two ratings of the Apple heart emoji differed by more than 1.0 on the sentiment scale. This eliminated 4% of the initial participant pool, leaving 235 participants in the standalone condition, and 2,247 in the context condition. Of these 2,482 participants, 1,207 identified as male, 1,269 as female, and 6 as a different gender. The median age was 33 (SD = 11; min = 18; max = 79). For emoji usage, 92 said they "Never" use emoji, 346 "Rarely," 882 "Sometimes," 737 "Frequently," and 425 "Very Frequently." 37% of participants use Apple, 31% use Samsung, 9.5% use LG, 3.6% use Google, 1.1% use Microsoft, 12.7% use other platforms, and 4.5% do not have a smartphone.

The participants from the standalone condition provided a total of 4,700 interpretations, with a median of 61interpretations per rendering (min = 58; max = 64). The participants from the context condition provided 44,903 interpretations total, with a median of 30 interpretations per rendering (mins[4] = 12,19; max = 35).

Analytical Methods

To measure the potential for miscommunication associated with a particular emoji in and out of context, we used the same metric as Miller et al (2016): *average sentiment misconstrual score,* the average distance between all pairs of participant sentiment ratings. The motivation is that pairwise comparisons essentially simulate communication between two people, so the greater the average distance between interpretations the more likely people are to miscommunicate. Another benefit is that this metric can be computed for a single rendering or for two different renderings

[3] We exceeded this quota because it was met after other participants had already started taking the survey.

[4] We report two minimums because the first is due to a survey flaw: one single tweet for one single rendering was not recording interpretations for about half of the survey period, until we discovered and corrected the error to start collecting data. The next least amount of interpretations per context was 19.

of an emoji character, thus simulating both communication within and across platforms. By computing all pairwise distances between people's interpretations, we simulated the full communication space within and across platforms.

We aimed to compare the variability of interpretation for when each emoji was presented standalone versus in context, and for both within- and across-platform communication. We thus had to compute four (2x2) distinct sentiment misconstrual scores for each emoji character in our study:

- *Within-Standalone*: within-platform without textual context
- *Within-Context*: within-platform with textual context
- *Across-Standalone:* across-platforms without textual context
- *Across-Context:* across-platforms with textual context.

Within- and across-platform computation directly follows methods by Miller et al. (2016). For **within-platform** computations (with or without textual context), we computed pairwise comparisons between interpretations of the *same* emoji *rendering*. For **across-platforms** computations, we computed pairwise comparisons for interpretations of *different renderings* of an emoji character (e.g., the Apple and the Google renderings). For an across-platforms misconstrual score, we first computed the score for each possible pair of platforms (e.g., Apple-Google, LG-Samsung, etc.), and then averaged across these platform-pair scores to get the overall *across-platforms* sentiment misconstrual score.

Likewise, our approach to **standalone** computations (within or across platforms) was the same as that of Miller et al. We computed the misconstrual score for each standalone *rendering*, and then averaged these scores to get the misconstrual score for each standalone emoji *character*. For **context** computations (within or across platforms), we computed sentiment misconstrual scores for each tweet containing a given emoji rendering, and then averaged these misconstrual scores to get the sentiment misconstrual score for each *rendering* in context. Finally, we averaged the scores for all renderings of an emoji character to get the *in-context* misconstrual score for that emoji *character*.

Misconstrual scores are not conventional statistics, so we needed to employ statistical resampling in order to estimate their precision. To do so, we used *jackknifing* resampling, which involves repeatedly re-computing our metrics with one data point removed (Efron and Tibshirani 1994). This process allowed us to estimate statistical properties (e.g., standard deviation) of arbitrarily complex metrics. Typically, a bootstrapped resample might be used in this type of study, since it is a newer and better-studied resampling method. However, in the course of our evaluation we found that bootstrapping introduces a bias when used with pairwise difference metrics like our misconstrual score. Jackknife resampling does not have this problem.

We "jackknifed" our data by participant rather than by raw sentiment scores because ratings by the same participant cannot be assumed to be independent. Also, since a participant may not have interpreted every emoji, we performed jackknife resampling individually for each emoji, where each incorporated only those participants who had interpreted the given emoji. After completing the jackknifing, we computed the standard error of the four misconstrual scores for each emoji. These standard error values allow us to compute confidence intervals and perform statistical tests. As our metric is an average (of differences), the central limit theorem implies that the metric will follow an approximately normal distribution. Therefore, we used t-distribution based confidence intervals and statistical tests.

Finally, to directly answer our research question, we compared each emoji's standalone and context misconstrual scores, specifically *Within-Standalone* to *Within-Context*, and *Across-Standalone* to *Across-Context*. Thus, we tested the null hypothesis that the interpretation of each emoji character is equally ambiguous with or without textual context. We made these comparisons using a Welch's t-test (Welch 1947), modified to use the standard error of each score (from jackknifing) instead of standard deviations divided by the square root of the sample size. Finally, because we made these comparisons for each emoji separately, we apply the Holm method (Holm 1979) to adjust our *p*-values for multiple comparisons. With these adjusted *p*-values, we performed the statistical tests at a significance level of 0.05.

We included data for all of the 77 emoji renderings in our study (averaged across the renderings to get each emoji character's values). While this analysis combined previous and current renderings, we also performed our analyses on the current versions of emoji characters alone, as well as on the previous versions studied by Miller et al. (2016) alone. As we will discuss below, this analysis provided key insight into our high-level results.

Results

Table 2 presents the four misconstrual scores and associated 95% confidence intervals for each emoji character in our study. The "Difference" columns for the "Within" and "Across" platform conditions show the estimated *difference in misconstrual between a standalone emoji character versus the same character in textual context*. This is computed simply by subtracting each context score from the associated standalone score. If the resulting value is positive, then on average the emoji is less ambiguous in context. But if the result is negative, then on average the emoji actually is *more* ambiguous in context. Finally, we indicate the results of our hypothesis tests by highlighting in bold the differences that are statistically significant. We also display the confidence interval for each statistic.

Crucially, the lack of bold positive numbers in the "Difference" columns in Table 2 shows that *we found little to no support for the hypothesis that textual context reduces the potential for miscommunication when using emoji*. One emoji character – "person raising both hands in celebration" (Unicode U+1F64C) – had a significantly lower misconstrual score when considered in context (both within- and across-platform, both p<0.0001). However, another character – "relieved face" (Unicode U+1F60C) – has a significantly *higher* (p<0.001) misconstrual score (within-platform only), meaning that there is more potential for misconstrual with this emoji character when it is used with text.

Further, examining the non-significant results in Table 2 makes it clear that the differences between standalone and in-context misconstrual exhibit no clear directional tendency. Some emoji characters trend towards a lower misconstrual score when considered in context; others trend towards a higher misconstrual score considered in context.

While Table 2 examines our misconstrual results at the level of emoji characters, Figure 1 shows these results at the rendering level. Our basic finding at the emoji character level also holds at the rendering level: context does not consistently reduce misconstrual. In Figure 1, there are 20 subgraphs: one for each of our 10 emoji characters both within and across platforms. Each subgraph depicts the misconstrual of each rendering of the given emoji character in each tweet in which it appears (each • in the figure). Each triangle represents a rendering's (average) misconstrual score in all its tweets and relates this in-context score to its standalone misconstrual score (denoted as ○): a triangle points up △ for an in-context misconstrual score greater than the standalone score (▲ for statistically significant differences), and down ▽ if it is less (▼ if significant).

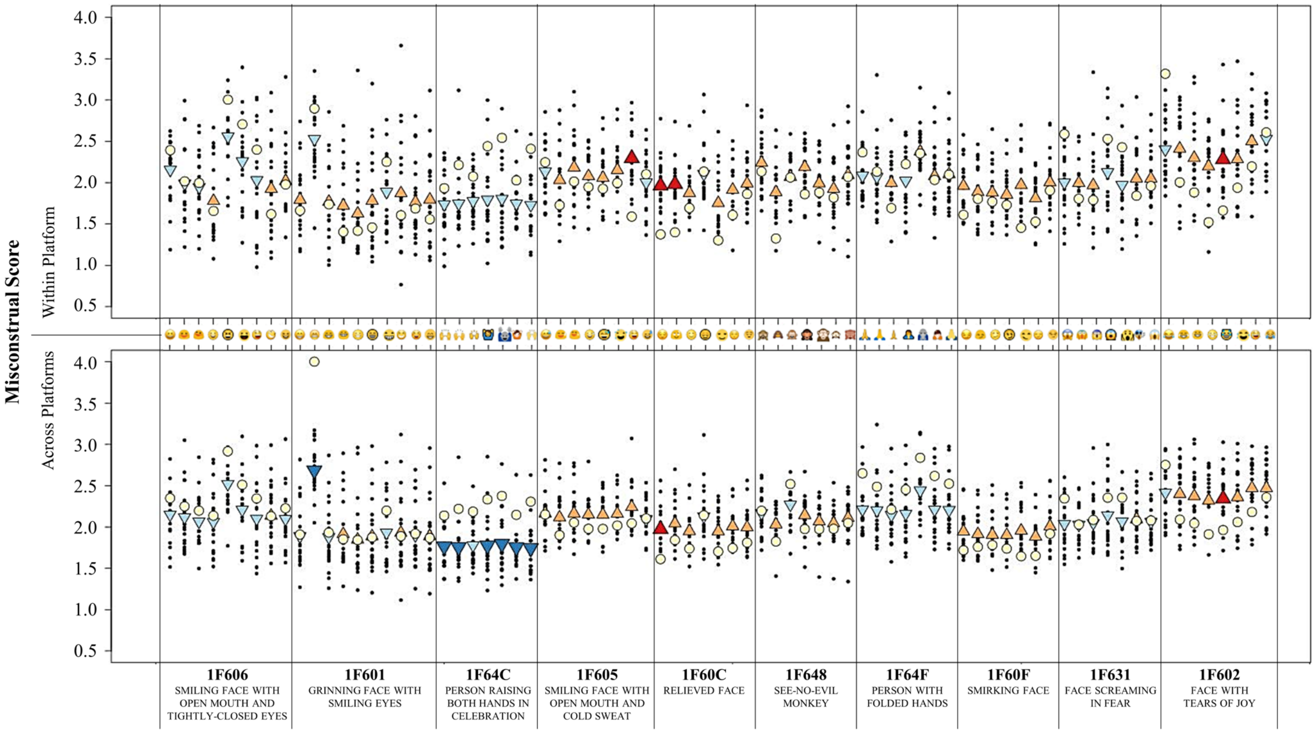
If the Miller et al. hypothesis were supported by our data – that is, if textual context reduces emoji's potential for miscommunication – we would see a trend of ▽and ▼ triangles. But this trend is not present in Figure 1. Further, like the character-level results, there are few statistically significant differences.

Figure 1 also lets us assess visually whether any outlier tweets might be driving our results. While there are some tweets where misconstrual was much higher or lower than most tweets with a given rendering, these outliers are few.

Returning to Table 2, the effect sizes for the difference in misconstrual between the two conditions (i.e., the values in the "Differences" column) can be difficult to interpret in isolation, so we sought to provide context by establishing a threshold below which any differences in misconstrual can

	WITHIN			ACROSS		
Emoji Unicode and Name	STANDALONE (Confidence Interval)	CONTEXT (Confidence Interval)	DIFFERENCE (Confidence Interval)	STANDALONE (Confidence Interval)	CONTEXT (Confidence Interval)	DIFFERENCE (Confidence Interval)
1F606 😆😆😆😆😆😆😆😆😆 SMILING FACE WITH OPEN MOUTH AND TIGHTLY-CLOSED EYES	2.197 (2.006, 2.389)	2.074 (2.028, 2.120)	0.124 (-0.111, 0.358)	2.314 (2.145, 2.537)	2.162 (2.115, 2.209)	0.179 (-0.061, 0.419)
1F601 😁😁😁😁😁😁😁😁😁😁 GRINNING FACE WITH SMILING EYES	1.769 (1.640, 1.897)	1.855 (1.813, 1.897)	-0.086 (-0.284, 0.075)	2.129 (1.994, 2.264)	1.976 (1.931, 2.020)	0.153 (-0.016, 0.323)
1F64C 🙌🙌🙌🙌🙌🙌🙌 PERSON RAISING BOTH HANDS IN CELEBRATION	2.235 (2.074, 2.397)	1.763 (1.718, 1.808)	**0.472*** **(0.273, 0.672)**	2.245 (2.091, 2.398)	1.767 (1.724, 1.811)	**0.477*** **(0.287, 0.668)**
1F605 😅😅😅😅😅😅😅😅 SMILING FACE WITH OPEN MOUTH AND COLD SWEAT	1.944 (1.785, 2.103)	2.118 (2.071, 2.165)	-0.174 (-0.372, 0.024)	2.029 (1.874, 2.184)	2.156 (2.109, 2.202)	-0.127 (-0.319, 0.066)
1F60C 😌😌😌😌😌😌😌 RELIEVED FACE	1.626 (1.509, 1.742)	1.941 (1.898, 1.985)	**-0.315*** **(-0.464, -0.167)**	1.799 (1.678, 1.920)	2.007 (1.963, 2.051)	-0.208 (-0.362, -0.054)
1F648 🙈🙈🙈🙈🙈🙈🙈 SEE-NO-EVIL MONKEY	1.879 (1.705, 2.053)	2.057 (2.010, 2.104)	-0.178 (-0.392, 0.037)	2.074 (1.894, 2.255)	2.120 (2.074, 2.166)	-0.046 (-0.268, 0.177)
1F64F 🙏🙏🙏🙏🙏🙏🙏 PERSON WITH FOLDED HANDS	2.129 (1.926, 2.332)	2.105 (2.056, 2.154)	0.024 (-0.225, 0.274)	2.541 (2.321, 2.761)	2.226 (2.175, 2.276)	0.315 (0.046, 0.584)
1F60F 😏😏😏😏😏😏😏 SMIRKING FACE	1.686 (1.540, 1.833)	1.911 (1.866, 1.911)	-0.224 (-0.407, -0.042)	1.745 (1.599, 1.891)	1.932 (1.888, 1.976)	-0.187 (-0.368, -0.005)
1F631 😱😱😱😱😱😱😱 FACE SCREAMING IN FEAR	2.135 (1.969, 2.301)	2.024 (1.970, 2.078)	0.111 (-0.097, 0.319)	2.189 (2.012, 2.367)	2.068 (2.015, 2.122)	0.121 (-0.100, 0.342)
1F602 😂😂😂😂😂😂😂😂 FACE WITH TEARS OF JOY	2.142 (1.947, 2.336)	2.364 (2.314, 2.414)	-0.222 (-0.461, 0.018)	2.170 (1.961, 2.379)	2.395 (2.346, 2.444)	-0.225 (-0.481, 0.031)

Table 2. The four sentiment misconstrual scores and associated confidence intervals for each emoji (renderings depicted with previous version underlined): standalone versus in context for both within- and across-platform analysis. The difference columns are the context scores subtracted from the standalone scores: when the value is positive, on average the emoji is less ambiguous in context, and vice versa. Differences that are bold are statistically significant at a level of 0.05; the lack thereof shows little support for the Miller et al. hypothesis.



Each • represents the misconstrual score of a tweet with the given rendering (renderings occupy the horizontal axis).
Each ○ represents the standalone misconstrual score of the given rendering.
A triangle represents the rendering's context misconstrual score:
▽ *if less than the standalone misconstrual score and* ▼ *if this relationship is statistically significant (p<0.05).*
△ *if greater than the standalone misconstrual score and* ▲ *if this relationship is statistically significant (p<0.05).*

Figure 1. Low-level visualization of misconstrual scores per emoji rendering, both within platform (top graph) and across platforms (bottom graph): The higher the point on the y-axis, the more potential there is for miscommunication, and vice versa. The variety of upward and downward pointing triangles illustrates the lack of a clear trend, in addition to the lack of statistically significant results.

be considered negligible. To do so, we compared the values in the "Differences" column to the misconstrual score of a minimally ambiguous emoji rendering, letting us check if any of the misconstrual differences are larger than one would expect in a minimally ambiguous context (i.e., larger than "interpretation noise"). We guessed that our control rendering ❤ would serve as a good minimally ambiguous rendering, and this hypothesis was supported: we computed the misconstrual score for each time participants interpreted this rendering—twice when presented standalone, and twice when presented in context ("love ❤"). This yielded four misconstrual scores for this rendering: 0.727 and 0.722 for its first and second standalone appearances, respectively, and 0.735 and 0.758 for its appearances in context. These values are all substantially below the standalone and context misconstrual values for the emoji in Table 2. As such, we conservatively choose 0.7 as a minimal threshold for differences in misconstrual to be considered meaningful, rather than just "interpretation noise."

Using our 0.7 threshold, we see that the effect sizes in the "Differences" column in Table 2 provide additional support for the conclusion that text has little to no disambiguating effect on emoji interpretation. The misconstrual differences between the standalone and context conditions, even for the few statistically significant results, are less than our "interpretation noise" threshold. Furthermore, the confidence intervals for each difference place a bound on how large of an impact context makes on emoji interpretation. None of the characters have differences that exceed the threshold of +/-0.7. In fact, we can be confident that more than half (12/20) of our differences are smaller than 0.4.

Finally, to understand our findings in more detail, we repeated our analyses separately for the Miller et al. renderings and for the current renderings. A standout result from these analyses was for the "Grinning Face with Smiling Eyes" emoji character (Unicode U+1F601). Miller et al. found that this character had high variation in interpretation across platforms and thus high potential for misconstrual, particularly due to Apple's previous rendering 😁 (this rendering has been substantially altered since the publication of Miller et al.; see Table 1). In our analysis using the Miller et al. renderings alone, we identified that there is a statistically significant reduction in the misconstrual score of this emoji character with textual context present for communication across platforms (p<0.01). Rendering-level results in Figure 1 verify that Apple's previous rendering is the main contributor to this effect (p<0.001). This suggests that in very extreme cases, there may be support for the hypothesis that text reduces the potential for emoji-related miscommunication. We return to this point in the Discussion section below.

Discussion

Our study suggests that text does not have the hypothesized disambiguation value for emoji. In this section, we discuss the implications of this finding more broadly.

An important question is *why* doesn't text reduce emoji ambiguity? One reasonable hypothesis is that sarcasm plays a role. Our survey contained an open-ended text box to gather feedback from participants, and several participants highlighted the role of sarcasm in their assessments:

> *"some of the emojis seemed sarcastic"*

> *"Wasn't sure how to analyze the sarcastic texts"*

Another insight as to why emoji were still ambiguous in context that was pointed out by a participant was that the texts containing the emoji were too short:

> *"A couple of the texts could use a little extra context to tell what the emoji is supposed to reflect. For instance, the "I didn't expect to see her unexpectedly" text could be either positive or negative based on context."*

With Twitter's 140 character length restriction, using tweets as our source of texts limited the amount of context accompanying emoji in our study, whereas many platforms for emoji usage are not limiting in that respect. Similarly, while using Twitter as we did (e.g., with the filtering steps outlined above) allowed us to maximize general interpretability and successfully examine general consistency of interpretation (as reflected in broadcast communication like Twitter), this approach limited the amount of *interpersonal* context (or *common ground* (Clark 1996)) in the simulated communication. Future work should seek to explore emoji ambiguity in longer-form texts and in longitudinal communication in more established relationships.

Interestingly, while our study controls for the presence or absence of text to study emoji ambiguity, the reverse relationship is also worthy of examination. In other words, future work should seek to investigate whether emoji affect the ambiguity of the text they accompany. Participants reflecting in the open-text box suggested that this could be the case. For example, one participant wrote:

> *"[emoji] do have their value in that they give you a sense of security that you've gotten across the right tone in an email. Whenever I feel I need to be most clear rather than risk a misunderstanding, I insert an emoji"*

This sentiment was reflected in some qualitative responses in Cramer et al.'s (2016) recent work on emoji as well.

Lastly, it is interesting to reflect on textual context's effectiveness in reducing the ambiguity of Apple's (former) rendering 😁 of the "grinning face with smiling eyes" character (U+1F601). Miller et al. identified a roughly bimodal distribution for sentiment interpretations for this rendering. Our results suggest that in these types of extreme ambiguity cases in which there are two clear senses that must be disambiguated, text may possibly help to distinguish between the two very different meanings. Examining this conjecture in detail would be a useful direction of future work.

Limitations

Although our study design was intentionally robust against a number of factors (e.g., idiosyncratic specific textual contexts, participant variation), it is not without limitations. First and foremost, to maximize ecological validity, we rendered the emoji images in the survey at a size that corresponds with their typical size in common use (rather than enlarged versions for easier viewing). This proved difficult for some participants that took the survey on desktop monitors. For instance, one participant wrote to us in an open feedback box at the end of the survey:

> *"The emojis were so small that it was difficult to determine what they were, even on a 17" monitor."*

This limitation suggests an interesting research question: how might the size of emoji affect interpretation? This could be an interesting and important direction of future work, particularly considering new ways emoji are being integrated into communication tools at different sizes. For example, in Slack and Apple Messages, when sending messages that solely contain emoji (standalone), the emoji appear larger than when you send them accompanied with text (in context).

Finally, as we mentioned above, even though we took precautions to limit the exogenous context required for interpreting tweets in our study, it is impossible to mitigate this concern entirely. For instance, some tweets may have been part of a larger series of tweets meant to be read in sequence (although the percentage of tweets in our study for which this was likely the case is very unlikely to have biased our results substantially).

Conclusion

When Miller et al. (2016) found extensive variation in the interpretation of some standalone emoji, it seemed natural that this variation would diminish, at least somewhat, if one considered the text that often accompanies emoji. However, analyzing the results of a survey with over two thousand participants, we found little to no support for this hypothesis. In fact, the preponderance of evidence suggests that text can increase emoji ambiguity as much as it can decrease it.

Open Data

The data we gathered for our experiment will be made available through ICWSM's data sharing initiative.

Acknowledgements

We wish to thank the Amazon Mechanical Turk workers who made this study possible by providing us with interpretations. This work was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 00039202. We would also like to thank our anonymous reviewers for their comments and suggestions.

References

Barbieri, Francesco, German Kruszewski, Francesco Ronzano, and Horacio Saggion. 2016. "How Cosmopolitan Are Emojis?: Exploring Emojis Usage and Meaning over Different Languages with Distributional Semantics." In *Proceedings of the 2016 ACM on Multimedia Conference*, 531–535. MM '16. New York, NY, USA: ACM. doi:10.1145/2964284.2967278.

Bavelas, Janet Beavin, and Nicole Chovil. 2000. "Visible Acts of Meaning An Integrated Message Model of Language in Face-to-Face Dialogue." *Journal of Language and Social Psychology* 19 (2): 163–94. doi:10.1177/0261927X00019002001.

Clark, H.H. 1996. *Using Language*. Cambridge University Press. https://books.google.com/books?id=DiWBGOP-YnoC.

Cramer, Henriette, Paloma de Juan, and Joel Tetreault. 2016. "Sender-Intended Functions of Emojis in US Messaging." In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 504–509. MobileHCI '16. New York, NY, USA: ACM. doi:10.1145/2935334.2935370.

Dimson, Thomas. 2015. "Emojineering Part 1: Machine Learning for Emoji Trends." *Instagram Engineering Blog*. http://instagram-engineering.tumblr.com/post/117889701472/emojineering-part-1-machine-learning-for-emoji.

Efron, Bradley, and R.J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability 57. Chapman & Hall/CRC. https://www.crcpress.com/An-Introduction-to-the-Bootstrap/Efron-Tibshirani/p/book/9780412042317.

"Emojipedia." 2017. Accessed January 14. http://emojipedia.org/.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2): 65–70.

Kelly, Ryan, and Leon Watts. 2015. "Characterising the Inventive Appropriation of Emoji as Relationally Meaningful in Mediated Close Personal Relationships." *Experiences of Technology Appropriation: Unanticipated Users, Usage, Circumstances, and Design*. https://projects.hci.sbg.ac.at/ecscw2015/wp-content/uploads/sites/31/2015/08/Kelly_Watts.pdf.

Medlock, Ben, and Gretchen McCulloch. 2016. "The Linguistic Secrets Found in Billions of Emoji." Technology presented at the SXSW. http://www.slideshare.net/SwiftKey/the-linguistic-secrets-found-in-billions-of-emoji-sxsw-2016-presentation-59956212.

Miller, Hannah J., Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "'Blissfully Happy' or 'Ready toFight': Varying Interpretations of Emoji." In *Proceedings of the 10th International AAAI Conference on Web and Social Media*. ICWSM '16. AAAI. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13167.

Novak, Petra Kralj, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. "Sentiment of Emojis." *PloS One* 10 (12). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4671607/.

Park, Jaram, Young Min Baek, and Meeyoung Cha. 2014. "Cross-Cultural Comparison of Nonverbal Cues in Emoticons on Twitter: Evidence from Big Data Analysis." *Journal of Communication* 64 (2): 333–54. doi:10.1111/jcom.12086.

Park, Jaram, Vladimir Barash, Clay Fink, and Meeyoung Cha. 2013. "Emoticon Style: Interpreting Differences in Emoticons Across Cultures." In *International AAAI Conference on Weblogs and Social Media (ICWSM 2013)*. AAAI. http://www.academia.edu/download/40936445/6132-30395-1-PB.pdf.

Pavalanathan, Umashanthi, and Jacob Eisenstein. 2016. "More Emojis, Less :) The Competition for Paralinguistic Function in Microblog Writing." *First Monday* 21 (11). http://firstmonday.org/ojs/index.php/fm/article/view/6879.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. "Lexicon-Based Methods for Sentiment Analysis." *Computational Linguistics* 37 (2): 267–307. doi:10.1162/COLI_a_00049.

Tigwell, Garreth W., and David R. Flatla. 2016. "Oh That's What You Meant!: Reducing Emoji Misunderstanding." In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, 859–866. MobileHCI '16. New York, NY, USA: ACM. doi:10.1145/2957265.2961844.

Walther, Joseph B., and Kyle P. D'Addario. 2001. "The Impacts of Emoticons on Message Interpretation in Computer-Mediated Communication." *Social Science Computer Review* 19 (3): 324–347.

Welch, B. L. 1947. "THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED." *Biometrika* 34 (1–2): 28–35. doi:10.1093/biomet/34.1-2.28.

Wijeratne, Sanjaya, Lakshika Balasuriya, Amit Sheth, and Derek Doran. 2016. "EmojiNet: Building a Machine Readable Sense Inventory for Emoji." In *Social Informatics*, 527–41. Springer, Cham. http://link.springer.com.ezp2.lib.umn.edu/chapter/10.1007/978-3-319-47880-7_33.